Universidade Federal de Campina Grande

Centro de Engenharia Elétrica e Informática

Coordenação de Pós-Graduação em Ciência da Computação

Sistema Semiautomático de Reconhecimento de Identidade Vocal Forense

Danilo Coura Moreira

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande – Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Elmar Uwe Kurt Melcher

Joseana Macêdo Fechine Régis de Araújo

(Orientadores)

Campina Grande, Paraíba, Brasil

Setembro de 2013

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

M835s Moreira, Danilo Coura.
Sistema semiautomático de reconhecimento de identidade vocal forense / Danilo Coura Moreira. – Campina Grande, 2013.
134 f. : il.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática.

"Orientação: Prof. Dr. Elmar Uwe Kurt Melcher, Prof. Dr. Joseana Macêdo Fechine Régis de Araújo".
Referências.

1. Fonética Forense. 2. Identificação de Locutor. I. Melcher, Elmar Uwe Kurt. II. Araújo, Joseana Macêdo Fechine Régis de. III. Título.

CDU 004.7:343.95 (043)

"Ninguém é igual a ninguém. Todo o ser humano é um estranho ímpar."

(Carlos Drumond de Andrade)

AGRADECIMENTOS

Agradeço primeiramente a Deus, que me proporcionou saúde, determinação e uma família maravilhosa, que me estruturou para que eu pudesse ter chegado até aqui.

Ao meu pai, Antonimário (*in memorian*), por todo o amor, carinho e dedicação. Pelo exemplo de hombridade, responsabilidade e honestidade. E por sempre ter priorizado pela nossa educação, minha e de meus irmãos.

A minha mãe, Conceição, por todo amor, carinho e dedicação. Pelas preocupações e cuidados constantes. E por ser exemplo de determinação, mostrando-me que por mais difícil que fosse o caminho, eu não deveria desistir.

Aos meus irmãos, Toninho e Herlinha, pelo carinho e suporte contínuos, me fazendo sentir apoiado em todos os momentos.

A minha namorada, Marília, por todo o amor e incentivo, me fazendo acreditar que tudo daria certo no final.

Aos mestres, de toda minha trajetória escolar e acadêmica, pelo altruísmo por dividir o conhecimento conquistado. Em especial, aos Profs. Dr. Elmar Uwe e Dra. Joseana Fechine, pelo direcionamento e pela orientação dedicada, sendo indispensáveis para a realização deste trabalho.

Aos colegas/amigos do Instituto de Polícia Científica da Paraíba (IPC-PB), pelo apoio e incentivo.

Aos colegas/amigos de pós-graduação e do Laboratório de Arquiteturas Dedicadas (LAD), pelo convívio salutar e troca de experiências que contribuíram diretamente para o desenvolvimento deste trabalho.

Ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande (COPIN/UFCG), pela oportunidade que me foi dada.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

RESUMO

Devido ao desenvolvimento das redes de telefonia, o uso deste meio como auxílio para os criminosos cometerem seus delitos é cada vez mais frequente. Baseado, então, na possibilidade de individualizar uma pessoa a partir de suas características vocais, nesta pesquisa propõe-se utilizar técnicas para o reconhecimento semiautomático da identidade vocal de locutores em ambiente telefônico, buscando auxiliar as investigações criminais, direcionando a imputação de autoria de uma voz e, portanto, servindo como meio de prova, na área forense. Para tanto, são utilizada a remoção do nível DC, a detecção de atividade vocal, a subtração spectral, a normalização e a pré-ênfase como técnicas para o pré-processamento do sinal de voz, visando a minimizar os efeitos negativos que o ambiente telefônico oferece às elocuções transmitidas por este meio reduzindo, assim, os erros na extração das características e, posteriormente, na criação dos padrões de cada locutor. Visando à eficiência no processamento e à robustez ao ruído, em relação a outros métodos de extração de características, foram utilizados coeficientes mel-cepstrais (MFCC) para este fim. Para a criação e a classificação dos padrões dos locutores, utilizou-se o modelo de misturas gaussianas (GMM), por proporcionar resultados melhores quando não há dependência de texto, dado que os locutores são não colaborativos. Visando a encontrar a melhor configuração de parâmetros para o sistema semiautomático, foram realizados experimentos considerando um sistema automático de reconhecimento de identidade vocal. Desta forma, foi obtida uma taxa de identificação de até 87,80%, com nível de confiança de 98%. Por fim, o sistema semiautomático de reconhecimento de identidade vocal atingiu a probabilidade de 99,95% de que determinada elocução pertença a um locutor, dentre um conjunto de 30 suspeitos, utilizando um nível de confiança de 98%. Desta forma, a técnica proposta possibilitou fornecer, com uma taxa de acerto próxima a 100%, um subconjunto de locutores suspeitos para posterior análise pericial.

ABSTRACT

Due to the development of telephone networks, the use of this environment to support criminals to commit crimes is increasingly common. Based, then, on the possibility of individualizing one person from their vocal characteristics, this work proposes using techniques for the semiautomatic vocal identity recognition of speakers in telephone environment, aiming help in criminal investigations, directing the attribution of voice authorship and, thus, suiting as evidence in forensic. For that purpose, are used DC offset, vocal detector activity, spectral subtraction, normalization and pre-emphasis such as pre-processing techniques of speech signal, which aim to minimize the negative effects that provides the telephone environment utterances transmitted by these means, reducing the errors in feature extraction and subsequently, in the patterns creation of each speaker. In order to optimize the processing efficiency and robustness to noise compared to other methods for feature extraction, Mel-Frequency Cepstral Coefficients (MFCC) was employed. To create the speakers patterns and classification, it was used the Gaussian Mixture Model (GMM), because provide better results when there is no dependence of text, due to the speakers are non-cooperative. Aiming at finding the best parameter setting for the semi-automatic system, experiments were performed considering an automatic vocal identity recognition system. In this way, it was possible reach to correct identification rate of up to 87.80%, with a confidence level of 98%. Lastly, the semiautomatic speaker identification system reached the probability of 99.95% that a given utterance belongs to a given speaker from a set of 30 suspects, using a confidence level of 98%. Thus, the proposed technique has enabled to provide, with a tax rate close to 100%, a subset of speakers suspects for subsequent forensic analysis.

LISTA DE SIGLAS

CMN	*Cepstral Mean Normalization*
DAV	Detector de Atividade Vocal
DC	*Direct or Continuous Current*
DFT	*Discrete Fourier Transform*
DTW	*Dinamic Time Warping*
EM	*Expectatization Maximization*
FDP	Funções Densidade de Probabilidade
GMM	*Gaussian Mixed Models*
HMM	*Hidden Markov Model*
IDFT	*Inverse Discrete Fourier Transform*
IPC-PB	Polícia Científica da Paraíba
LBG	*Linde Buzo and Gray*
LPC	*Linear Predictive Coding*
LPCC	*Linear Predictive Cepstral Coefficients*
MCRA	*Mínima Controlled Recursive Average*
ML	*Maximum Likelihood*
MFCC	*Mel-frequency Cepstral Coeficients*
MP3	*MPEG-1/2 Audio Layer 3*
PCA	*Principal Component Analysis*
PFS	*Parametric Feature-sets*
RV	Razão de Verossimilhança
SNR	*Signal-to-noise ratio*
SVD	*Singular Value Decomposition*
TCD	Transformada do Cosseno Discreta
TFD	Transformada de Fourier Discreta
VQ	*Vector Quantization*

LISTA DE ABREVIATURAS

dB	Decibéis
Hz	Hertz
kHz	Quilohertz

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

CONTEÚDO

1 INTRODUÇÃO

A carga genética e a influência do meio em que vive é o que caracteriza cada indivíduo de forma única, individuando-o de todos os demais (STIGAR, 2012)

Individualizar é relacionar com tudo aquilo que não é. Dessa forma, a individualidade humana pode ser definida como o conjunto de características que distinguem um indivíduo do outro, características estas físicas, metabólicas e comportamentais (PENA, 2006).

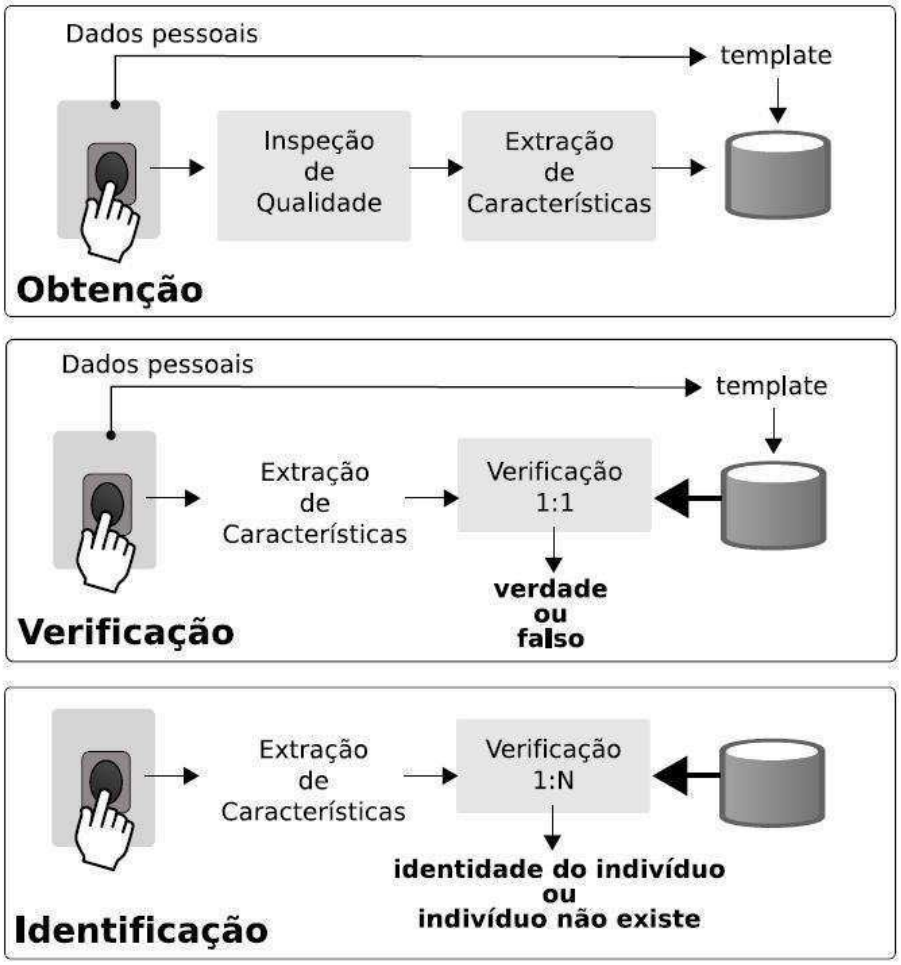
Portanto, a partir desses preceitos é possível caracterizar cada ser humano de forma ímpar. A assertiva de que cada indivíduo é único e conserva suas características físicas e comportamentais distintas e mensuráveis é a base para que a biometria seja utilizada para essa individualização.

1.1 Biometria

A biometria, como o próprio nome de origem grega sugere - *bio* (vida) + *metron* (medida), é a mensuração das características físicas e/ou comportamentais das pessoas, como forma de identificá-las unicamente. Dessa maneira, essa individualização vem sendo amplamente utilizada para esse fim (SINGH, PANDA, *et al.*, 2003; MARTINS, 2009).

Todo sistema biométrico é preparado para verificar ou identificar uma pessoa que foi previamente cadastrada. O procedimento de verificação ocorre quando o sistema confirma uma possível identidade comparando apenas parte da informação com o todo disponível referente àquela identidade. O processo de identificação confirma a identidade de um indivíduo, comparando o dado fornecido com todo o banco de dados registrado, como apresentado na Figura 1.1.

Figura 1.1 - Fases de obtenção (cadastramento), verificação e identificação do processo biométrico de impressão digital.



Fonte: BARBOSA JÚNIOR, 2007.

Tendo em vista a potencialidade de individualização dessas técnicas, é variado o número de aplicações que se utilizam desse reconhecimento, tais como sistemas de segurança para acesso, autenticação em sistemas telefônicos ou computacionais, investigações policiais baseadas em gravações de escutas telefônicas ou em outros ambientes.

O ser humano possui características, físicas e comportamentais, quem podem ser aferidas e/ou mensuradas. Esta distinção entre as características dá origem a duas subdivisões da biometria, conforme discriminado a seguir (MARTINS, 2009).

1.1.1 Biometria Física

A Biometria física está relacionada com a fisiologia do corpo humano, ou seja, características físicas, como pode ser observado em alguns exemplos a seguir:

- **Veias das mãos**: apesar do pouco tempo do seu descobrimento, apresenta alto grau de confiabilidade para o reconhecimento de indivíduos, pois suas características são imutáveis e é praticamente impossível a falsificação deste tipo de informação. Possui baixo custo financeiro para a obtenção das imagens para o reconhecimento (MARTINS, 2009).

- **Impressão digital**: É um dos métodos mais rápidos, tanto na fase de treinamento, quanto na fase de reconhecimento. Possui alta taxa de reconhecimento e baixo custo financeiro para ser implantado, contribuindo para que seja um dos método mais difundidos. É um método pouco intrusivo para o usuário (SILVA, E. C. L., 2007).

- **Reconhecimento da face**: A constante mudança da aparência facial humana faz com que este método não seja tão confiável quanto os demais, pois este método é baseado nos registros de vários pontos que delimitam o rosto, calculando proporções entre estes pontos. Apresenta rapidez de processamento e baixo custo financeiro para ser implantado. (SILVA, E. C. L., 2007).

- **Reconhecimento retinal**: apresenta maior confiabilidade em relação aos demais, devido as características da retina serem imutáveis com o passar dos anos, e o alto grau de dificuldade de realização de fraude. Sistema muito invasivo, possuindo uma leitura que incomoda o usuário. Apresenta alta complexidade na extração das características e alto custo financeiro para sua implantação (SILVA, E. C. L., 2007).

- **Reconhecimento irial**: apresenta boa confiabilidade em relação aos demais, devido as características da íris serem imutáveis com o passar dos anos.

Apresenta alto custo financeiro para sua implementação. A leitura das informações oriundas do usuário são menos invasivas em comparação ao reconhecimento retinal. (SILVA, E. C. L., 2007).

- **Reconhecimento de voz**: Este é um dos métodos menos invasivos e tem como o reconhecimento de fala, a sua forma mais natural de uso. Possui sensibilidade a ruídos oriundos do ambiente e problemas por mudança na voz do usuário causados por gripes ou estado emocional, porém possui baixo custo financeiro para sua implementação (SILVA, E. C. L., 2007).

- **Geometria da mão**: possui menor confiabilidade em relação as demais, necessita que o usuário encaixe a mão na posição correta, porém possui custo financeiro mediano para sua implantação, requer pouco espaço de armazenamento e pouco esforço ou atenção por parte do usuário durante o reconhecimento (SILVA, E. C. L., 2007).

1.1.2 Biometria Comportamental

Cada indivíduo reage de maneira distinta em determinadas situações. Essas diferentes reações são o que caracterizam, de forma única, cada ser (MARTINS, 2009).

Atualmente, sistemas de segurança utilizam o modo como usuário manuseia o teclado e mouse para detectar se realmente se trata do usuário autenticado.

Em um futuro muito próximo será possível destacar um indivíduo no meio de uma multidão apenas pelo seu jeito de andar, mexer as mãos ou identificando algum hábito desse indivíduo. Este tipo de análise é denominada biometria comportamental.

1.1.3 Biometria na Segurança Pública

O reconhecimento de suspeitos em aplicações de segurança pública, por meio do uso da biometria, é denominada identificação criminal. Tem como objetivo reconhecer

indivíduos baseado nas suas características únicas, auxiliando as funções da segurança pública, como investigações policiais e processos judiciais (CANEDO, 2010).

Sendo assim, a biometria tem um papel fundamental na Segurança Pública, pois por meio dela é possível reconhecer um suspeito ou associá-lo a um local de crime, mesmo o indivíduo sendo não colaborativo (CANEDO, 2010).

Por mais de um século a biometria vem sendo utilizada, por meio de métodos manuais, para identificação criminal. Porém as operações eram muito lentas e técnicas limitadas devido à alta complexidade. A automatização desses métodos proporcionaram a identificação em grande bases de dados, além da possibilidade da adoção de técnicas de alta complexidade (CANEDO, 2010).

1.2 Reconhecimento de Locutor

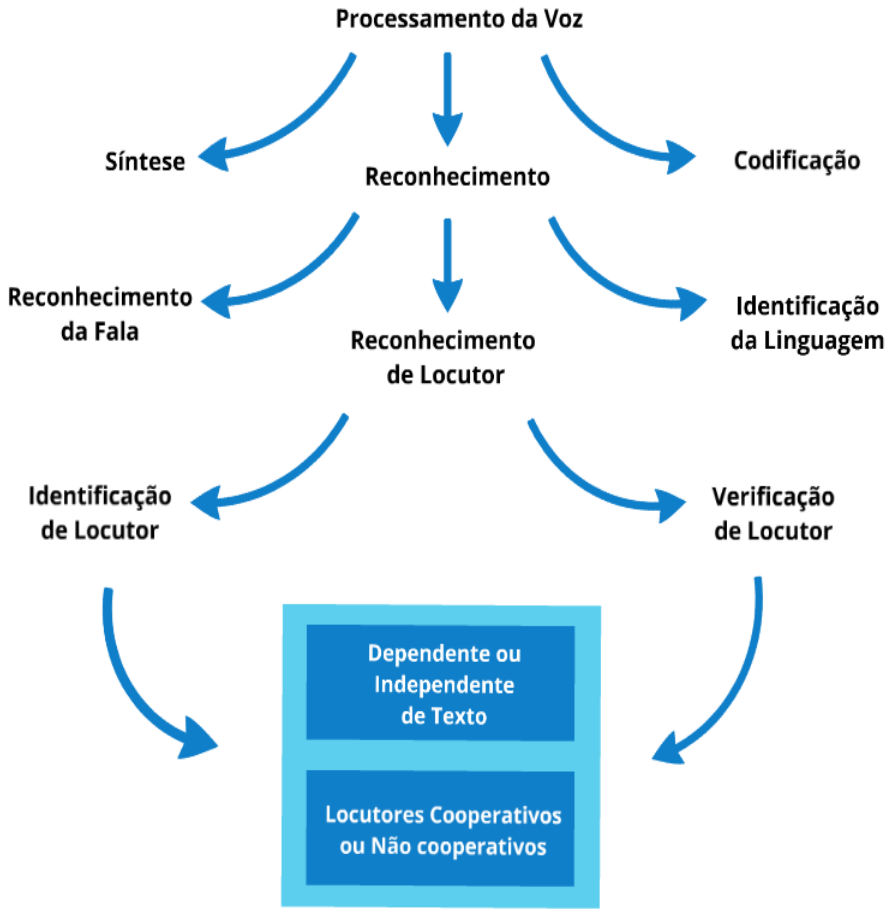
O objetivo de um sistema de reconhecimento de locutor é reconhecer um locutor a partir da sua voz, sendo bastante útil em aplicações de segurança como, por exemplo, o controle de acesso a ambientes restritos (utilização da voz para abrir e fechar portas) e o controle de acesso de dados em computadores. Em criminalística, pode ser utilizado com o mesmo propósito que hoje é dado às impressões digitais (RABINER e SCHAFER, 1978; FECHINE, 2000).

O reconhecimento de locutor consiste na capacidade de reconhecer um indivíduo com base em sua voz, por meio de suas características e peculiaridades, caracterizando-se, portanto, no reconhecimento de padrões da voz. A exemplo de outros sistemas de reconhecimento de padrões, este processo ocorre em duas fases, o treinamento e o reconhecimento propriamente dito. O treinamento se caracteriza pela criação de modelos, a partir da extração de características dos sinais de voz, individualizando cada locutor. O reconhecimento acontece a partir da comparação das características extraídas do sinal de voz de um locutor questionado com os modelos previamente gerados na fase de treinamento.

Conforme pode ser observado na Figura 1.2, o reconhecimento de locutor é uma das subáreas do reconhecimento de voz, que se divide também em reconhecimento de fala, cuja tarefa básica é reconhecer, em uma elocução, uma determinada sentença; e identificação da linguagem, que tem por objetivo identificar o idioma da elocução emitida. O processamento de voz é a área que abrange, além do reconhecimento, a síntese e a codificação de voz que, respectivamente, tratam da produção artificial da voz humana e do armazenamento compacto de sinais de voz.

Por fim, o reconhecimento de locutor subdivide-se em duas subáreas, a verificação e identificação de locutor, que podem ser dependentes ou não de texto, assim como possuir locutores colaborativos ou não.

Figura 1.2 - Descrição geral do processamento da voz.



Fonte: Adaptado de CAMPBEL, 1997.

1.2.1 Verificação e Identificação de Locutor

De acordo com Salman, Muhammad e Khurshid (2007), existem duas vertentes no reconhecimento de locutor, a verificação e a identificação. Na verificação, a averiguação da identidade é baseada apenas em um suposto emissor do sinal. Portanto, seu retorno é uma resposta binária, respondendo se aquele suposto emissor é ou não detentor daquela voz. No caso da identificação, a comparação é feita sobre um conjunto de locutores e seu retorno consiste em definir qual desses locutores possui a voz questionada.

Em se tratando de aplicações práticas, um sistema de verificação poderia ser empregado para confirmar a identidade de um indivíduo que tentasse acessar sua conta bancária, utilizando a voz como chave de acesso. Para o caso da identificação, pode-se citar uma situação em que se deseja descobrir a identidade de um suspeito, dentre diversos outros contidos em uma base de dados de vozes criminal. Na Figura 1.3, é ilustrada a diferença entre os dois modos de reconhecimento.

Figura 1.3 - Identificação e verificação de locutor.

Fonte: Adaptado de Boğaziçi University. Disponível em <http://www.busim.ee.boun.edu.tr/~speech/Speaker_recognition.html>.

1.2.2 Dependência e Independência de Texto

Segundo Che, Lin e Yuk (1996), as duas formas de reconhecimento, verificação e identificação, também são subdivididas em dois tipos. A forma dependente de texto, em que o reconhecimento é feito a partir de uma elocução fixa, pré-determinada, e a forma independente de texto, em que o locutor, para o seu reconhecimento, expressa qualquer elocução.

A forma independente é mais conveniente, pois o locutor pode falar livremente para o sistema. No entanto, requer mais treinamento e declarações de testes para se conseguir um bom desempenho (SUBRAMANYA *et al.*, 2007).

Vale observar que sistemas dependentes de texto tendem a apresentar desempenho superior àqueles independentes de texto. Tal característica se deve ao fato de que os sistemas dependentes de texto se apóiam, inicialmente, no reconhecimento da elocução para, então, a partir da divergência entre a elocução de teste e um modelo selecionado, reconhecer o locutor (CAMPBELL JR, 1997).

1.2.3 Locutores Cooperativos e Não-cooperativos

O foco central do reconhecimento de locutor é o indivíduo que pretende ser reconhecido. Tal locutor, conforme seu interesse, pode ou não ser cooperativo. No caso de sistemas de autenticação, em que é do interesse do usuário ser reconhecido, ele se caracteriza como um locutor cooperativo e o sistema é, usualmente, dependente de texto.

Na área da criminalística, por exemplo, é de maior interesse o uso de sistemas de reconhecimento de locutor independente de texto, uma vez que, na maioria das aplicações os locutores a serem identificados são não cooperativos (FECHINE, 2000; KINNUNEN e HAIZHOU, 2010).

A não cooperação dos locutores nas aplicações voltadas à Criminalística é justificável, visto que de acordo com a Constituição da República Federativa do Brasil, ninguém é obrigado a produzir provas conta si mesmo (BRASIL, 1988).

Portanto, ser colaborativo ou não está fortemente associado à escolha do sistema ser dependente ou não de texto, pois caso colabore, o locutor estará predisposto a citar frases pré-determinadas, dando subsídios para escolha de um sistema dependente de texto. Caso contrário, deverá optar por um sistema que não dependa de citações determinadas anteriormente.

1.2.4 Degradação de Desempenho

De acordo com Campbell (2009), o desempenho de um sistema de reconhecimento de voz pode ser influenciado por diversos fatores externos que degradam a qualidade do resultado. Esses fatores podem ter origem humana ou ambiental.

Fatores de origem humana podem ser causados por algum erro de elocução ou leitura de sentenças pré-definidas; pela variação do estado emocional, que provoca mudança na voz; por problemas de saúde que podem alterar as características do trato vocal, *e.g.* rouquidão, gripe ou resfriado; e, por fim, pela variação das gravações ao longo do tempo, visto que a idade pode alterar a forma do trato vocal.

No que diz respeito aos fatores ambientais, a mudança do posicionamento do microfone no decorrer das capturas das elocuções está sujeita a variações indesejadas nas gravações. Vale ressaltar, também, que um ambiente acústico pobre ou inconsistente pode alterar as características originais do sinal de voz, pois quando estas são extraídas de um sinal contaminado por ruído, não refletem de forma fidedigna o locutor que as gerou (TOGNERI e PULLELLA, 2011).

No tocante ao canal, sua variabilidade na captura das elocuções (*e.g.* microfone, canal telefônico) torna heterogênea a condição de gravação, devendo ser empregado o mesmo canal de comunicação tanto na fase de treinamento quanto na fase de reconhecimento. Modelos quando são treinados a partir de elocuções captadas do

ambiente com o emprego de determinado microfone podem se mostrar incompatíveis com sinais amostrados por outros equipamentos com características distintas na fase de teste ou ambientes distintos quanto às propriedades acústicas (CARDOSO, 2009).

A qualidade do canal influencia de forma importante, o desempenho do reconhecimento. Muitas vezes, visando a economia no armazenamento ou na transmissão, como ocorre na telefonia, é fixada uma frequência de amostragem reduzida, porém não comprometendo a inteligibilidade do sinal, sendo definida em 8 kHz. Esse valor também foi escolhido baseado no Teorema de Nyquist[1], pois a faixa de frequências do sinal de voz utilizada varia de 300 a 3400 Hz. Desta forma, a comunicação é realizada, porém com perda de informação e, consequentemente, de qualidade, mas garantindo a inteligibilidade, o que pode ser suficiente para a comunicação telefônica (MATEUS, 2012).

Por fim, um fator que influencia de forma considerável apenas os sistemas de identificação é a escalabilidade em função do número de locutores na base de dados de voz. Isto acontece em virtude da menor variabilidade de características que distinguem os locutores, tornando mais difícil a tarefa de destacá-los de forma exclusiva (CARDOSO, 2009).

1.2.5 Processo de Reconhecimento

De forma geral, o processo de reconhecimento de locutor se divide em 5 etapas descritas em seguida.

1. Aquisição do Sinal de Voz

Para o processamento do sinal por um computador, é preciso que este esteja na forma digital. Para tanto, torna-se necessário capturar, por meio de um microfone ou, até

[1] **Teorema de Nyquist**: a frequência de amostragem de um sinal analógico, para que possa posteriormente ser reconstituído com o mínimo de perda de informação, deve ser igual ou maior a duas vezes a maior frequência do espectro desse sinal.

mesmo, de um aparelho telefônico, a onda acústica emitida pelo trato vocal do locutor, convertendo-a, assim, em um sinal analógico. De posse do sinal analógico, um conversor é utilizado para transformá-lo em um sinal digital (OPPENHEIM e SCHAFER, 2009).

Esta etapa acontece tanto na fase de treinamento quanto na fase de reconhecimento. Na fase de treinamento, são capturadas elocuções da voz para que seja criado um padrão que individualize o locutor. Na fase de reconhecimento, uma elocução de voz é capturada para ser comparada com os padrões já existentes na base de dados do sistema.

2. Extração de Características

A partir de uma elocução, é possível extrair características e armazená-las, de tal maneira que estas possam ser utilizadas na construção dos padrões (etapa de treinamento), assim como na análise entre os padrões de voz armazenados e a voz questionada (etapa de reconhecimento) (VIBHA, 2010).

A voz possui diversas peculiaridades que podem ser exploradas. Uma delas é a frequência fundamental. Esta característica varia de acordo com o gênero e a idade do locutor, possibilitando distinguir entre vozes masculinas, femininas, adultos e crianças. Não é um método de reconhecimento, porém pode ser usado como um filtro, direcionando de uma forma mais adequada a busca do resultado (FECHINE, 2000).

3. Criação dos Padrões (Treinamento)

Nesta etapa, a partir das elocuções de treinamento, são criados padrões que caracterizam individualmente cada locutor. Sendo assim, na fase de correspondência de padrões, as características extraídas das elocuções questionadas são comparadas com estes (LIMA, 2001).

4. Correspondência de Padrões

Nesta etapa, as características extraídas da elocução são comparadas com a base de dados que contém os padrões/modelos de cada locutor. A partir dessa comparação, um nível de similaridade é obtido (LIMA, 2001).

5. Tomada de Decisão

Após obter a similaridade entre as características extraídas da voz e o modelo, se faz necessária uma decisão para aceitar ou não o locutor. No caso da verificação de locutor, a similaridade é comparada com um limiar de decisão: caso seja maior, o resultado é positivo; caso contrário, não. Na identificação de locutor, a elocução que tiver maior similaridade, dentre todos os modelos, é dita a originadora da elocução (CARDOSO, 2009).

Em métodos estocásticos, utiliza-se uma medida de probabilidade como parâmetro de decisão. Em métodos paramétricos, utiliza-se uma medida de distância, sendo comum o uso da Distância Euclidiana (FECHINE, 2000).

1.3 Caracterização do Problema

Devido ao grande desenvolvimento das redes de telefonia, que abrangem grande parte dos territórios habitáveis, o uso desse meio como auxílio para os criminosos cometerem seus delitos é cada vez mais frequente (DRYGAJLO, 2007).

Baseado, então, na possibilidade de individualizar uma pessoa a partir de suas características vocais, a fonética forense vem utilizando o reconhecimento da identidade vocal de locutores para auxiliar nas investigações criminais. Trata-se de um exame direcionado a determinar a autoria de uma voz e, portanto, serve como meio de prova, na área forense, para atribuir a alguém a autoria de um crime, sua participação nele ou a ele relacioná-lo. Também é capaz de desvincular o envolvimento de um

inocente em um crime que lhe possa estar sendo imputado, o que, no teor jurídico, é mais importante que incriminar um culpado (BRAID, 2009).

Portanto, foi percebida a relevância da implementação de técnicas automatizadas para a identificação de locutor no âmbito do Instituto de Polícia Científica da Paraíba (IPC-PB), visando ao apoio na tomada de decisão para a inferência da identidade vocal, devido ao fato de, atualmente, ser utilizada na instituição apenas uma técnica manual de verificação da identidade vocal de locutores, a partir da extração de características, tais como a frequência fundamental[2], os formantes[3] e a energia das elocuções questionadas e de elocuções padrão. De posse das características, gráficos são comparados manualmente para que possa ser atestado se aquela voz realmente foi produzida pelo suspeito em questão.

Dado este cenário, o objetivo desta pesquisa consiste na concepção de uma técnica que seja capaz de identificar, a partir de sinais de voz obtidos de ligações telefônicas, quem é o detentor da voz, em meio a um universo pré-determinado de locutores. Também se torna necessário considerar as peculiaridades do ambiente em questão, que possui ruídos que dificultam esse reconhecimento. Portanto, busca-se também uma técnica robusta, que venha minimizar os efeitos adversos do ambiente ruidoso de aquisição do sinal.

1.4 Objetivos

Nesta seção, serão expostos o objetivo geral da pesquisa em questão, assim como os objetivos específicos necessários para atingir a finalidade proposta.

1.4.1 Objetivo Geral

O principal objetivo desta pesquisa consiste em desenvolver um sistema para o reconhecimento de locutor voltado à fonética forense.

[2] **frequência fundamental:** primeira frequência produzida na glote (espaço entre as cordas vocais).
[3] **formantes:** podem ser definidos como picos de energia em uma região do espectro sonoro.

Por ser uma área em que a grande maioria dos sinais de voz é oriunda de gravações telefônicas, esse sistema deverá ser robusto ao ruído proveniente desse ambiente de aquisição do sinal, tratando-o ou eliminando-o.

Por se tratar de um sistema voltado para investigações criminais, é relevante que o reconhecimento de locutor seja eficiente na extração de características capazes de identificar, em meio a vários suspeitos, quem é o responsável pela elocução questionada. Sendo assim, trata-se de um identificador de locutor que, por meio da extração das características de uma elocução questionada, seja capaz de compará-la com uma gama de padrões de voz (modelos) de eventuais suspeitos e determinar de quem é a voz em questão.

Como os atores do procedimento são não cooperativos, ou seja, não é de interesse de cada locutor reproduzir uma frase ou um texto pré-definido, visto que poderá ser usado para sua acusação, é importante também que o sistema seja independente de texto, isto é, que não haja necessidade de que o locutor questionado reproduza um texto pré-determinado para ser identificado ou não.

1.4.2 Objetivos Específicos

A fim de concretizar o objetivo principal, foram formulados os objetivos secundários a seguir.

1. Realizar um levantamento do estado da arte na área, de forma a aprofundar os conhecimentos no âmbito do Processamento Digital de Sinais de Voz e da Fonética Forense;
2. Utilizar as técnicas mais adequadas para remoção ou redução dos ruídos provenientes do ambiente de gravação;
3. Determinar qual a técnica mais adequada para extração das características, considerando as condições de aquisição do sinal intrínsecas ao ambiente;

4. Escolher o modelo que melhor se adeque às características do sistema para identificação da identidade vocal de locutores, independente de texto;

5. Definir qual o método de tomada de decisão a ser adotado.

1.5 Relevância

Devido à dificuldade do processo de identificação de locutor, independente do texto, em ambientes com baixa qualidade do sinal de voz e presença constante de ruídos, esta pesquisa torna-se relevante, visto que objetiva a obtenção de resultados significativos, mesmo nessas condições, utilizando, para tanto, algoritmos de remoção e/ou tratamento de ruídos e técnicas para a extração de características e para a classificação de padrões que sejam adequadas a esses ambientes.

A pesquisa visa a também auxiliar a fonética forense do IPC-PB, visto que esta não possui um sistema para identificação de locutor. Atualmente, são realizadas apenas comparações manuais de características extraídas a partir de amostras de elocuções para simples verificação de locutor, implicando um baixo índice de eficiência no reconhecimento.

1.6 Organização do Trabalho

Este documento foi estruturado em sete capítulos. O presente capítulo teve por objetivo permitir ao leitor uma visão mais ampla do reconhecimento de locutor, suas variações e o que caracteriza os sistemas biométricos, e ao mesmo tempo procura direcionar texto para o objeto de estudo deste trabalho. Além disto, nesta seção é apresentada uma breve descrição dos conteúdos dos demais capítulos deste documento.

No Capítulo 2, é descrito o mecanismo de produção da voz e seu modelo correspondente, por meio da descrição da fisiologia humana e seus subsistemas, que são responsáveis pela geração dos parâmetros necessários para representar os sinais de voz, visando o seu reconhecimento. Também é feita a descrição de um sinal de voz,

seus processos de digitalização e pré-processamento. Por fim, é apresentada uma descrição de técnicas utilizadas na extração de características vocais e modelagem dos sinais de voz de cada locutor.

No Capítulo 3, é apresentada a relação entre o reconhecimento de locutor e a criminalística, com destaque para as áreas de atuação da Fonética Forense e os requisitos para que esse reconhecimento seja satisfatório. Também é apresentada uma gama de pesquisas que tratam do reconhecimento de locutor, aplicados à área forense e/ou voltados à identidade vocal em ambientes telefônicos e com presença de ruídos, com o propósito de encontrar as melhores técnicas para implementação do sistema de reconhecimento.

No Capítulo 4, é descrito o conjunto de técnicas e parâmetros utilizados para a construção do Sistema Semiautomático de Identificação Vocal Forense, desde o pré-processamento, seguido da extração de características e modelagem dos padrões dos locutores, até a tomada de decisão.

No Capítulo 5, são reportados, analisados, discutidos e comparados os principais experimentos realizados ao longo da pesquisa. Também estão inclusos neste capítulo as particularidades do ambiente de desenvolvimento e simulação, além das características da base de vozes e metodologia aplicada.

No Capítulo 6, são comentados os resultados finais, as conclusões obtidas, contribuições, limitações e as sugestões para pesquisas futuras.

No Apêndice A, são detalhados os 2 tipos de verificadores de atividade vocal, por energia e cruzamento de zero, que embasaram a construção de verificadores mais sofisticados.

No Apêndice B, está detalhada a análise estatística referente aos resultados obtidos nos experimentos.

2 RECONHECIMENTO AUTOMÁTICO DE IDENTIDADE VOCAL

A voz humana é o meio para o reconhecimento automático de identidade vocal. A maneira única como a voz é produzida é o que individualiza cada ser dos demais. Essa unicidade se dá devido as particularidades da fisiologia de cada indivíduo, refletindo nas características extraídas dos sinais de voz, e consequentemente na criação dos modelos que representam dada indivíduo.

2.1 Voz Humana

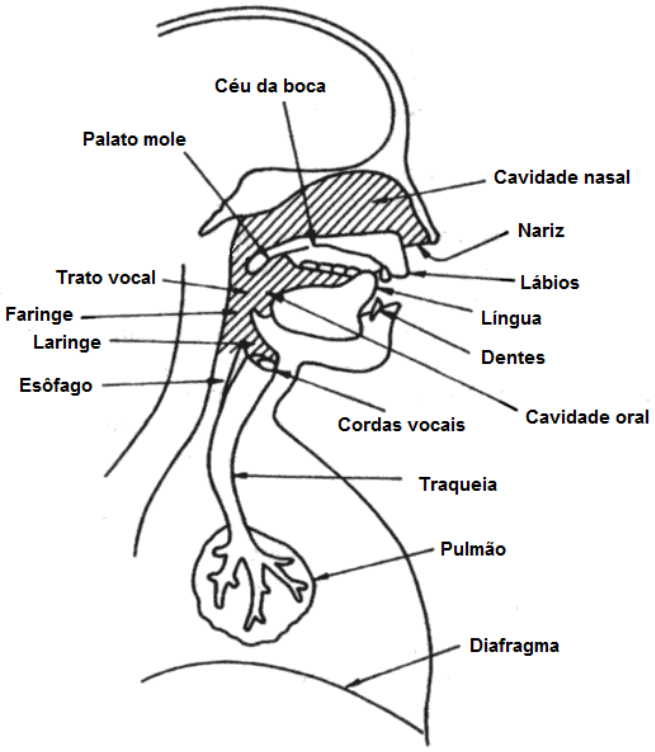
A voz humana é produzida desde o nascimento e se apresenta de diversas formas, tais como o choro, o grito, o riso e os sons da fala. É um dos meios de comunicação do indivíduo com o exterior, particularmente com seus semelhantes. É um elemento de grande valor nas relações interpessoais, porque declara, revela e anuncia o que o próprio indivíduo é, suas potencialidades e suas deficiências, refletindo a personalidade e exercendo enorme influência no aprendizado do outro (GABANINI, 2003).

Portanto, a voz é o principal veículo de comunicação do ser humano. A partir da voz, qualquer indivíduo é capaz de manifestar ideias, emitir ordens, solicitar o outro, expressar emoções e até mesmo fazer arte. Mesmo que um indivíduo não tenha aprendido a falar, a voz é capaz de causar reações no ouvinte.

2.1.1 Fisiologia da Voz

Para simplificar a fisiologia do aparato vocal do ser humano, este aparato é subdividido em três subsistemas: o respiratório, o laríngeo e o supralaríngeo (BRAID, 2009), conforme apresentado na Figura 2.1.

Figura 2.1 - Aparelho fonador humano.



Fonte: SCATENA, 2010.

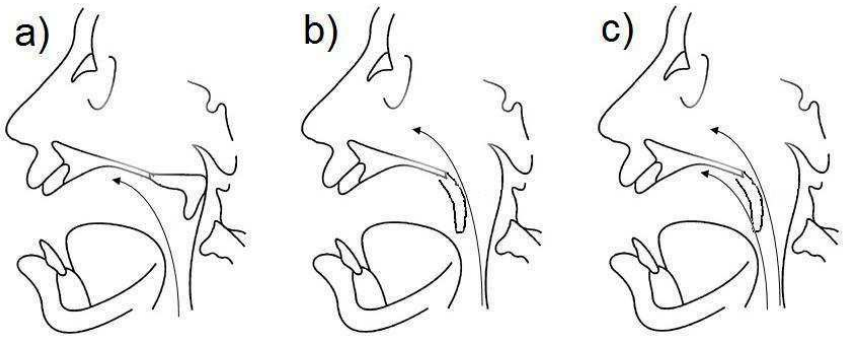
O subsistema respiratório é composto pelos pulmões, os músculos respiratórios, os brônquios e a traqueia. Tem como função produzir as correntes de ar que impulsionam a fala. A maioria dos sons é gerada pelas correntes de ar egressiva[4]. Porém, existem sons formados pelas correntes de ar ingressiva[5], o que não acontece na língua portuguesa (BRAID, 2009).

O subsistema laríngeo é formado por um conjunto de músculos, ligamentos e cartilagens que controlam a posição das dobras vocais. Este subsistema responsável pela fonação, que são as modificações da corrente de ar emitida pelo sistema respiratório, transformando esta corrente em pulsos de ar (BRAID, 2009).

[4] **corrente de ar egressiva**: O ar se dirige para fora dos pulmões e é expelido por meio da pressão exercida pelos músculos do diafragma.
[5] **corrente de ar ingressiva**: O ar se dirige de fora para dentro dos pulmões, devido à contração da musculatura do diafragma e dos músculos intercostais.

O subsistema supralaríngeo compreende as regiões faríngea, bucal e nasal. Este subsistema é responsável pela modulação do som, por meio da disposição do palato mole, que determina por onde a corrente de ar irá passar (boca, nariz ou ambos), e dos articuladores ativos[6] e passivos[7], moldando o som (BRAID, 2009), conforme a Figura 2.2.

Figura 2.2 - Fluxo da corrente de ar quando da realização de sons a) orais, b) nasais e c) nasalizados.



Fonte: O Mundo Ortodôntico. Disponível em <http://omundoortodontico.blogspot.com.br/2010/05/protese-de-palato.html>.

2.2 Processamento Digital de Sinais de Voz

A voz humana é um sinal de pressão acústica que varia com o tempo. Esse sinal, analógico, pode ser convertido em um sinal digital de modo a possibilitar seu processamento a partir de programas de computador. O processo de digitalização começa com a captura do sinal de áudio, por um microfone para converter o sinal de voz em sinal elétrico.

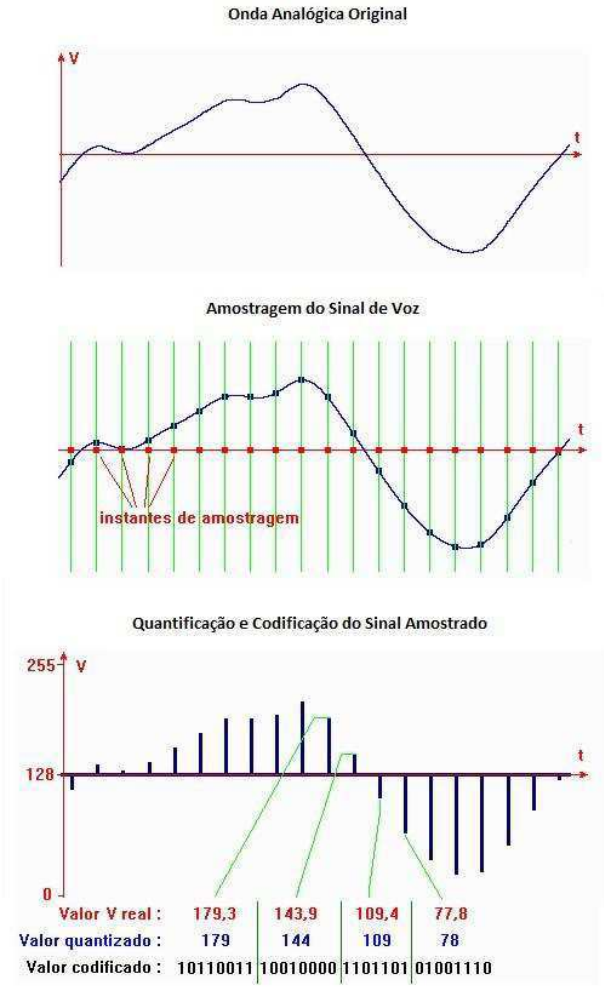
Primordialmente, o sinal de voz é submetido a um filtro passa-baixas[8] analógico denominado *anti-aliasing*, que visa eliminar frequências altas não controladas, que

[6] **articuladores ativos**: movimentam-se em direção ao articulador passivo e esse é um fator importante para a diferenciação dos sons. A língua, o lábio inferior, o véu palatino e as cordas vocais constituem o quadro dos articuladores ativos.
[7] **articuladores passivos**: o lábio superior, os dentes superiores e o céu a boca (alvéolos, palato duro, palato mole e úvula). São chamados passivos, porque não se movimentam, apenas recebem movimentação dos articuladores ativos.
[8] **filtro passa-baixas:** permite a passagem das frequências abaixo de um limite e rejeita (atenua) as frequências acima desse limite.

podem ter sido geradas por distorções, interferências ou ruídos (MENDOZA, 2009). Dessa forma, há um controle sobre a frequência máxima da amostra, para que, de acordo com o Teorema de *Nyquist*, a frequência de amostragem seja ser maior que o dobro da maior frequência contida no sinal a ser amostrado, sendo possível reproduzir integralmente sem erro de *aliasing*[9] (erro de atribuição).

Em seguida, o sinal de voz é submetido ao processo de amostragem, que consiste em obter amostras do sinal de voz que o representem, estes com frequência baseada no Teorema de *Nyquist*. Essas amostras são quantizadas em valores discretos e por fim, codificados em bits, conforme ilustrado na Figura 2.3.

Figura 2.3 - Processo de Digitalização de Sinal de Voz.



[9] **aliasing**: erro de atribuição que acontece quando há subamostramento do sinal, podendo este não ser representado com fidedignidade.

2.2.1 Pré-processamento do Sinal de Voz

Anterior à etapa da extração das características da elocução, se faz necessário o pré-processamento do sinal de voz, a fim de deixá-la apta a esse procedimento, disponibilizando, assim, um conjunto de dados com informação útil e minimização da presença de ruídos.

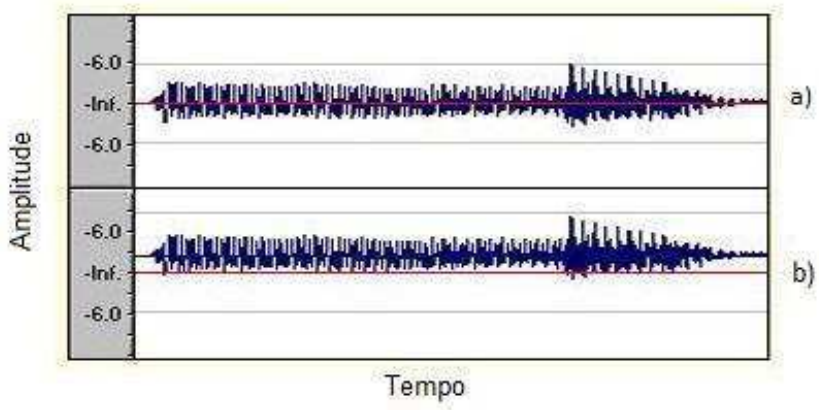
2.2.1.1 Remoção do Nível DC (*DC Offset*)

Segundo Al-Hassani e Kadhim (2012), o deslocamento ocorre quando um dispositivo de hardware, tal como uma placa de som, acrescenta uma componente contínua a um sinal de áudio gravado. Esta componente produz uma forma de onda gravada que não está centrada na linha de base.

O deslocamento pela corrente contínua pode ter efeito sobre a qualidade da informação extraída, prejudicando a extração das características e a modelagem (WATTS, 2006).

Um exemplo ilustrativo de remoção de desvio pela corrente contínua a partir de uma elocução é mostrado na Figura 2.4.

Figura 2.4 - Remoção do desvio pela corrente contínua de uma elocução a) depois da remoção b) antes da remoção.



Fonte: AL-HASSANI E KADHIM, 2012.

Portanto, a remoção desse deslocamento força o sinal de entrada para a linha de base média por meio da subtração da média das *p* amplitudes *a* para cada elocução, conforme a Equação 2.1.

$$a[n] = a[n] - \bar{a} \quad para\ n = 1, 2, \dots, p, \tag{2.1}$$

2.2.1.2 Detector de Atividade Vocal (DAV)

Segundo Patra (2007), a Detecção de Atividade Vocal é uma etapa crucial no processo de reconhecimento de locutor, pois as informações mais importantes em uma elocução geralmente estão contidas na porção sonora do sinal de voz. Quando se trata de ambiente ruidoso, essa importância se torna ainda maior, pois o ruído afeta com mais intensidade as porções surda e silenciosa da elocução, devido às características de baixa energia e alta frequência que possuem. Como consequência, há uma redução na dimensão do sinal de voz, o que auxilia na redução da complexidade computacional das etapas posteriores do reconhecimento de voz.

Devido à difícil distinção dos sons surdos do silêncio, geralmente esses dois tipos são classificados em um único grupo, como surdo/silencioso, o que de certa forma reduz a eficácia da extração da parte útil da elocução. Visando minimizar esse efeito, segundo Burielanu *et al.* (2000), pode-se utilizar uma estimação estatística dos quadros de transição entre as regiões sonoras e surdas/silenciosas, verificando a similaridade por meio de medidas de distância destes com as regiões definidas anteriormente.

Os Detectores de Atividade Vocal são geralmente baseados na energia ao longo do sinal, verificação do número de cruzamentos por zero ou modelos estatísticos, como proposto em (SOHN, KIM e SUNG, 1999), que sugere que é possível determinar com acurácia as regiões com voz de um sinal se as estatísticas do ruído tiverem menor variabilidade que as estatísticas da voz que se deseja localizar.

Portanto, o modelo em questão utiliza uma métrica estatística, que determina um limiar de decisão para distinguir os trechos que possuem voz e os trechos que possuem apenas ruídos e/ou silêncio, conforme Equação 2.2.

$$\begin{cases} H_0: sem\ voz \rightarrow\ Y = N \\ H_1: com\ voz \rightarrow\ Y = N + S\ ' \end{cases} \tag{2.2}$$

em que H_0 é a hipótese que o quadro do sinal de voz seja composto apenas por ruído (*N*) e H_1 é a hipótese que o quadro do sinal de voz seja composto por ruído (*N*) e por voz (S).

Dessa forma, as FDP (Funções Densidade de Probabilidade) condicionadas por H_0 e H_1 são dadas por:

$$P(X|H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi\lambda_N(k)} \exp\left(-\frac{|X_k|^2}{\lambda_N(k)}\right), \tag{2.3}$$

$$P(X|H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi[\lambda_N(k) + \lambda_S(k)]} \exp\left(-\frac{|X_k|^2}{\lambda_N(k) + \lambda_S(k)}\right), \tag{2.4}$$

em que λ_N e λ_S representam a variância do espectro do ruído e do sinal de voz, respectivamente. De posse dessas FDP, o estimador de máxima verossimilhança é definido, este sendo responsável por determinar se existe ou não de voz em quadros do sinal, como:

$$\Lambda_k \triangleq \frac{P(X_k|H_1)}{P(X_k|H_0)} = \frac{1}{1+\xi_k} \exp\left(\frac{\gamma_k \xi_k}{1+\xi_k}\right), \tag{2.5}$$

em que $\xi_k \triangleq {\lambda_S(k)}/{\lambda_N(k)}$ e $\gamma_k \triangleq {|X_k|^2}/{\lambda_N(k)}$ são denominados de relação sinal-ruído *a priori* e *a posteriori*, respectivamente. Por fim, a média geométrica das razões de

probabilidade para as bandas de frequências individuais é o que estabelece a regra de decisão, o que é dado por:

$$\log \Lambda = \frac{1}{L}\sum_{k=0}^{L-1} \log \Lambda \underset{H_1}{\overset{H_0}{\gtrless}} \eta, \tag{2.6}$$

2.2.1.3 Subtração Espectral

Segundo Silva, V. R. V. G. (2007), o método da Subtração Espectral é eficaz para redução do ruído de linha telefônica, que trata-se de um ruído de fundo estacionário.

Primeiramente, baseia-se na premissa que o espectro na frequência do sinal de voz $x(n)$ é dado pela soma do espectro de voz $s(n)$ e do espectro de ruído $v(n)$, conforme a Equação 2.7.

$$x(n) = s(n) + v(n), \tag{2.7}$$

Devido o procedimento ser realizado no domínio da frequência, calcula-se a transformada de Fourier de ambos os termos da Equação 2.7, obtendo-se:

$$X(j\omega) = S(j\omega) + V(j\omega), \tag{2.8}$$

Assim:

$$|X(j\omega)|^2 = |S(j\omega)|^2 + |V(j\omega)|^2 + 2|S(j\omega)||V(j\omega)| \cos\Theta, \tag{2.9}$$

$$S^2(\omega) + V^2(\omega) = 2S(\omega)V(\omega)\cos\Theta, \tag{2.10}$$

em que $S(\omega)$ e $V(\omega)$ são, respectivamente, as magnitudes de $S(j\omega)$ e $V(j\omega)$, e Θ é a subtração de fase entre o sinal de voz e o sinal de ruído. Com a premissa de que $s(n)$ e $v(n)$ são processos não correlacionados e aleatórios, é feita aproximação:

$$X^2(\omega) = S^2(\omega) + V^2(\omega), \tag{2.11}$$

Para que haja a recuperação do quadrado da magnitude ou da potência instantânea do espectro, subtraí-se uma estimativa de $V^2(\omega)$ de $X^2(\omega)$, conforme a Equação 2.12.

$$\hat{S}^2(\omega) = X^2(\omega) - \hat{V}^2(\omega) = S^2(\omega) + \left[V^2(\omega) - \hat{V}^2(\omega)\right], \tag{2.12}$$

Então, a magnitude do espectro da voz é dada por:

$$\hat{S}(\omega) = \sqrt{\hat{S}^2(\omega)} = \sqrt{X^2(\omega) - \hat{V}^2(\omega)}, \tag{2.13}$$

E a estimação da magnitude do espectro de voz é obtida pela Equação 2.14.

$$\hat{S}(\omega) = X(\omega) - \hat{V}(\omega) = S(\omega) + \left[V(\omega) - \hat{V}(\omega)\right], \tag{2.14}$$

A B-*ésima* potência do espectro de potência do sinal de voz pode ser expressa como:

$$\hat{S}^B(\omega) = X^B(\omega) - n\hat{V}^B(\omega), \tag{2.15}$$

em que B é um inteiro (igual a 1 ou a 2) e *n* é um parâmetro que controla a quantidade de ruído que será diminuída. Logo, a estimativa do espectro de voz é obtida pela equação 2.16:

$$\hat{S}(j\omega) = \left[X^B(\omega) - n\hat{V}^B(\omega)\right]^{1/B} e^{j\Psi(\omega)}, \tag{2.16}$$

sendo $\Psi(\omega)$ a fase de $\mathrm{X}(j\omega)$.

Os valores $B = 2$ e $n = 1$ retornam a subtração instantânea da potência de espectro. No caso dos valores $B = 2$ e $n = 1$, há o retorno subtração da magnitude de espectro.

Para que haja a estimação do espectro do ruído, o sinal de voz de ruído $x(n)$ passa por um processo de segmentação, resultando em N blocos. Em seguida, cada bloco passa pela transformada rápida de Fourier, resultando em um bloco de N amostras espectrais. O conjunto desses blocos de amostras espectrais compõem uma matriz bidimensional (contendo informação tempo-frequência), denotada por $X_T(j\omega)$, em que T é o índice do número do bloco e designa a dimensão do tempo.

O ruído é estimado como:

$$\hat{V}_T^B(\omega) = \alpha_A \hat{V}_{T-1}^B(\omega) + (1-\alpha_A)X_T^B(\omega), \qquad se\ X_T^B \geq V_{T-1}^B(\omega), \tag{2.17}$$

$$\hat{V}_T^B(\omega) = \alpha_1 \hat{V}_{T-1}^B(\omega) + (1-\alpha_1)X_T^B(\omega), \qquad se\ X_T^B < V_{T-1}^B(\omega), \tag{2.18}$$

Nas Equações 2.17 e 2.18, α_1 e α_A são parâmetros responsáveis pelo controle das constantes de tempo nas recursões, havendo uma atualização mais rápida nos casos em que o ruído é predominante no sinal, e mais lenta nos casos em que ele não é tão presente.

Na Figura 2.5 visualiza-se o diagrama em blocos da Subtração Espectral.

Figura 2.5 - Diagrama de blocos do algoritmo de subtração espectral.

Fonte: SILVA, V. R. V. G. DA, 2007.

2.2.1.4 Normalização

As amplitudes da elocução são normalizadas visando-se sua equiparação com as demais elocuções. Este passo é muito importante, pois as amplitudes possuem intensidades diferentes, devido à intensidade do locutor, à distância do locutor para o microfone e ao nível de gravação (AL-HASSANI e KADHIM, 2012). A normalização é feita dividindo-se cada amplitude a pelo valor máximo do sinal, conforme indicado na Equação 2.19.

$$a[n] = {a[n]}\Big/_{\max(a)} \quad para\ n = 1, 2, \dots, p, \tag{2.19}$$

2.2.1.5 Pré-ênfase

A filtragem da pré-ênfase serve para atenuar as componentes de baixa frequência e incrementar as componentes de alta frequência do sinal de voz, prevenindo contra instabilidade numérica, também, minimizando o efeito dos lábios (AL-HASSANI e KADHIM, 2012).

A pré-ênfase das frequências altas é necessária para que se obtenham amplitudes mais homogêneas das frequências formantes, porque informações importantes sobre a locução também estão presentes nas altas frequências.

Na Equação 2.20, apresenta-se a função de transferência mais usada para um filtro de pré-ênfase.

$$\mathrm{H(z)} = 1 - \mathrm{az}^{-1}, \qquad 0 \leq \mathrm{a} \leq 1, \tag{2.20}$$

Neste caso, a saída do sistema de pré-ênfase $\tilde{s}(n)$ está relacionada à entrada $s(n)$ pela equação de diferenças, conforme Equação 2.21.

$$\tilde{s}(\mathrm{n}) = \mathrm{s(n)} - \mathrm{as(n-1)}, \tag{2.21}$$

cujo valor de a usualmente varia entre 0,95 e 0,98 (CARDOSO, 2009).

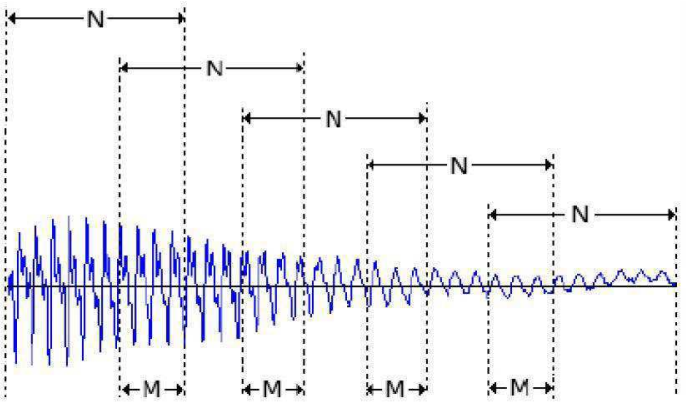
2.2.1.6 Segmentação

De acordo com Campbell Jr (1997), as características do sinal de voz são invariantes no tempo para curtos intervalos na ordem de 10 a 30 ms. Levando-se em consideração esta propriedade, é realizada a análise do sinal de voz em tempo curto, que se trata da divisão do sinal em quadros de tamanho fixo com um número de amostras obtidas num período de tempo escolhido dentro deste intervalo, o que permite caracterizar a cada instante o trato vocal como um filtro digital.

Visando-se a facilitar o cálculo da transformada de Fourier, necessária em outras fases do processo de reconhecimento, busca-se escolher um número de amostras que seja da forma 2^n, sendo *n* inteiro (CARDOSO, 2009).

Para que não haja perdas de informações nas descontinuidades bruscas nas extremidades dos quadros, os quadros adjacentes possuem uma sobreposição M apresentando assim, um determinado número de amostras em comum entre quadros vizinhos. Dessa forma, um quadro de comprimento N terá M amostras sobrepostas a cada quadros que lhe é adjacente (RABINER e JUANG, 1993), conforme pode ser observado na Figura 2.6.

Figura 2.6 - Segmentação do sinal de voz.

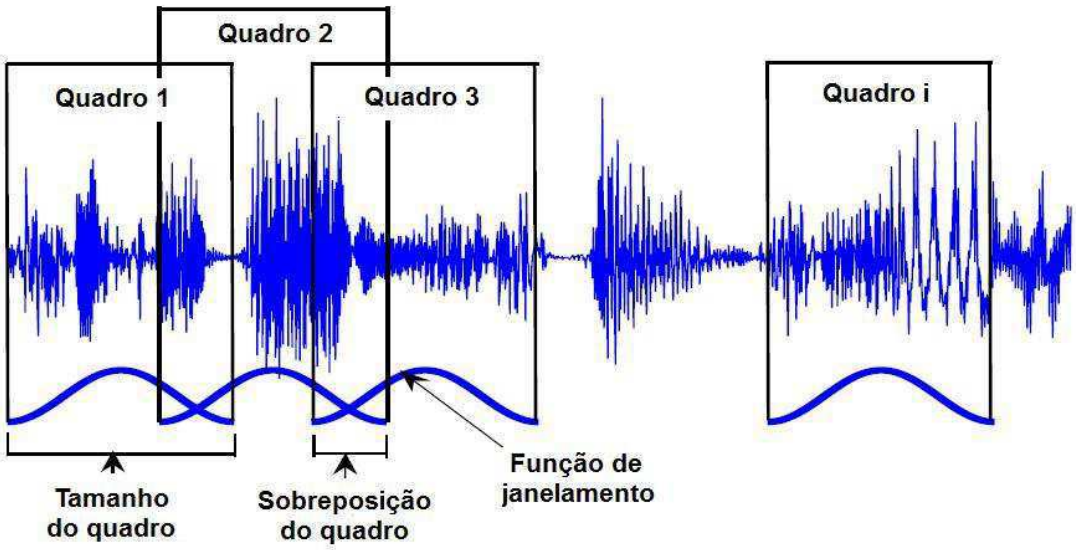


Fonte: MENDOZA, 2009.

2.2.1.7 Janelamento

De posse dos quadros obtidos no processo de segmentação do sinal de voz, faz-se necessário minimizar os efeitos das variações abruptas de amplitude encontradas nas extremidades de cada quadro, como mostra a Figura 2.7. Portanto, cada quadro é multiplicado por uma função janela, atenuando-se o valor das amostras que se localizam no início e final de cada quadro. Sendo representado pela Equação 2.22, em que o quadro apresenta amostras $x(n)$ e a função janela é dada por $w(n)$, resultando-se no quadro $y(n)$, possuindo redução das variações bruscas das suas extremidades (CARDOSO, 2009).

Figura 2.7 - Janelamento do sinal de voz.



Fonte: Adaptado de KINNUNEM, 2004.

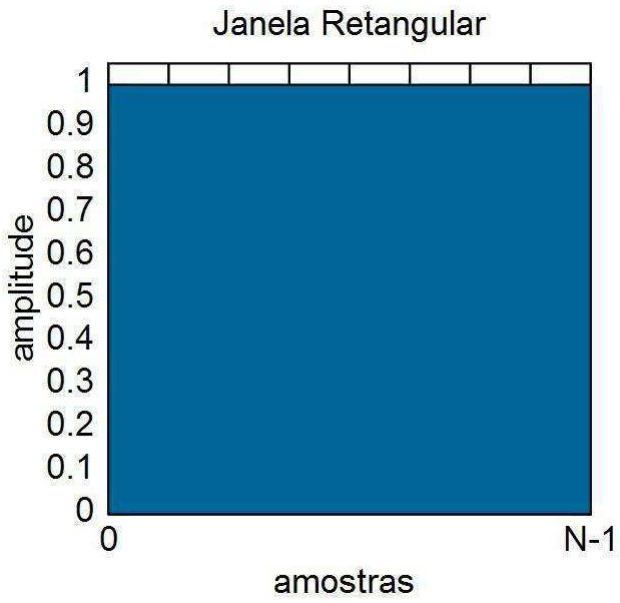
$$y(\mathrm{n}) = \mathrm{x}(\mathrm{n})\mathrm{w}(\mathrm{n}), \tag{2.22}$$

Existem diversas funções de janelamento, porém as mais utilizadas são as janelas Retangular, de Hanning e de Hamming, descritas a seguir:

- **Retangular**

Aplicar uma janela retangular é o mesmo que não utilizar qualquer janela. Utilizada para a análise de quadros que possuem um tamanho menor do que a da janela em análise. Na Figura 2.8 é apresentada a área de atuação da janela Retangular (PATRA, 2007).

Figura 2.8 - Janela Retangular.



Fonte: Adaptado da WIKIPEDIA. Disponível em <http://en.wikipedia.org/wiki/Window_function>. Acesso em 12 mai. 2013.

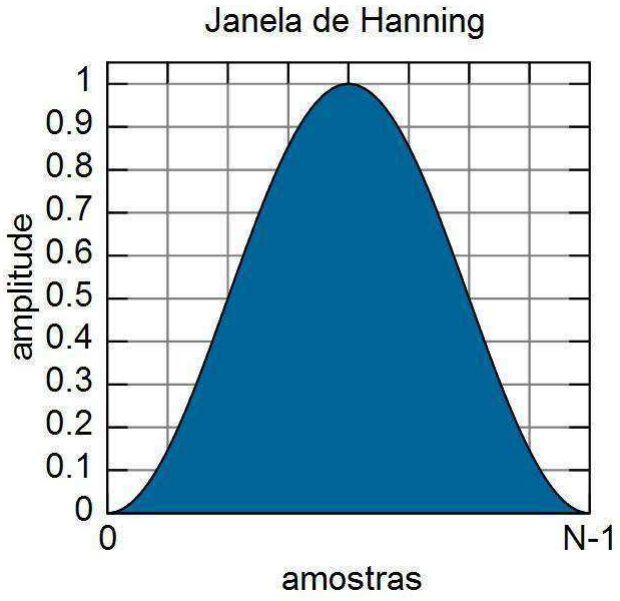
A janela retangular possui o valor igual a 1 em todo o seu intervalo de tempo. Uma janela de tamanho N pode ser definida matematicamente por meio da Equação 2.23:

$$W(\mathrm{n}) = 1, \quad \text{para } \mathrm{n} = 0, 1, 2, \ldots, \mathrm{N} - 1, \tag{2.23}$$

- **Hanning**

Utilizada para a análise de quadros maiores que o tempo de duração da janela (PATRA, 2007). A Figura 2.9 ilustra a área de atuação da janela de Hanning.

Figura 2.9 - Janela de Hanning.



Fonte: Adaptado da WIKIPEDIA. Disponível em <http://en.wikipedia.org/wiki/Window_function>. Acesso em 12 mai. 2013.

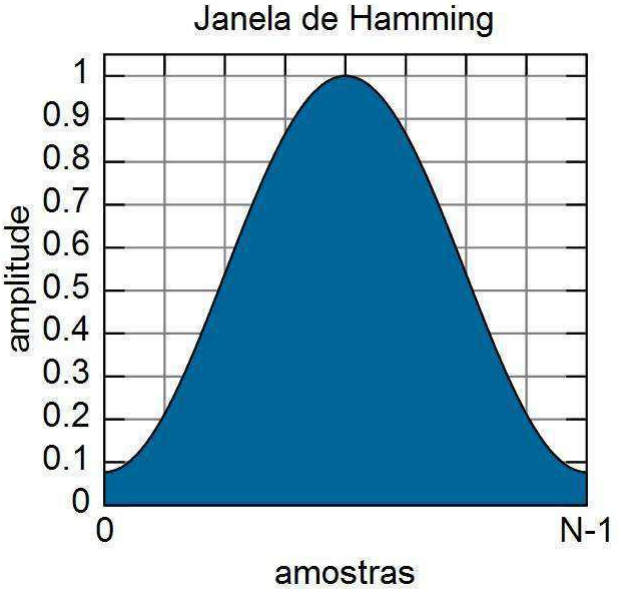
Possui o formato semelhante àquela de meio ciclo de uma forma de onda cossenoidal. A janela em questão de tamanho *N* está definida por meio da Equação 2.24.

$$W(\mathrm{n}) = 0{,}5\left(1 - 0{,}5\cos\left(\frac{2\mathrm{n}\pi}{\mathrm{N}}\right)\right), \quad \text{para } \mathrm{n} = 0, 1, 2, \ldots, \mathrm{N} - 1, \tag{2.24}$$

- **Hamming**

A janela de Hamming possui características semelhantes à de Hanning. Porém, deve-se observar que a janela de Hamming não se aproxima do zero como a janela de Hanning, em relação ao domínio do tempo (PATRA, 2007). Na Figura 2.10 é apresentada a área de atuação da janela de Hamming.

Figura 2.10 - Janela de Hamming.



Fonte: Adaptado da WIKIPEDIA. Disponível em <http://en.wikipedia.org/wiki/Window_function>. Acesso em 12 mai. 2013.

Sua forma também é semelhante àquela de meio ciclo de uma forma de onda cossenoidal. Uma janela de tamanho N está definida por meio da Equação 2.25:

$$W(\mathrm{n}) = 0{,}54 - 0{,}46\cos\left(\frac{2\mathrm{n}\pi}{\mathrm{N}}\right), \quad \text{para } \mathrm{n} = 0, 1, 2, \ldots, \mathrm{N} - 1, \tag{2.25}$$

2.2.2 Extração de Características

A extração de características é um problema de categorização de redução de um determinado objeto, mantendo o poder de discriminação desta. Usualmente é uma transformação com perdas, ou seja, que não leva a reconstrução perfeita da entrada.

No caso específico do reconhecimento de locutor, trata-se da representação do sinal de voz por um vetor de características de tamanho reduzido comparado ao sinal. A razão principal para permitir essa perda de informações é de reduzir a complexidade computacional no momento da criação dos modelos representativos dos locutores.

Porém, faz-se necessário que o número de amostras reservadas para o treinamento dos modelos seja suficientemente grande, em comparação com a dimensionalidade das medições. A quantidade de vetores de treinamento necessários cresce exponencialmente com a dimensionalidade do objeto (KINNUNEN, 2003).

De acordo com Singh *et al.* (2003), é desejável que o conjunto de características possua as seguintes propriedades:

I) Preservar ou realçar a informação e as variações na expressão que é relevante para a base a ser utilizada para o reconhecimento de voz e, ao mesmo tempo minimizar ou eliminar qualquer variação irrelevante para essa tarefa;
II) Ser relativamente compacto, a fim de permitir um treinamento mais simples dos modelos, a partir de quantidades finitas de dados;
III) Poder ser utilizada sem grandes restrições na maior parte dos casos, como em ambientes acusticamente pobres ou com qualidade reduzida;
IV) Ter um processamento de cálculo computacionalmente barato. O atraso no processamento é um fator significativo em alguns contextos, como no reconhecimento em tempo real.

Para que haja um treinamento e correspondência eficiente de padrões de sinais de voz, é necessário escolher adequadamente sua representação. Desta forma, cada sinal de voz observado é representado por um conjunto de características relevantes (BIMBOT *et al.*, 2004).

2.2.2.1 Análise Cepstral

De acordo com a teoria fonte-filtro[10], a voz humana é formada pelos sons sonoros[11] e surdos[12] oriundos da fonte, que sofrem alteração dos filtros, visando à produção de

[10] **Teoria Fonte-filtro**: considera que a produção da fala é dividida em duas partes: a primeira é a fonte de sons, em que se produz o sinal de voz (subsistemas respiratório e laríngeo) e a segunda é um sistema de filtros em série que modificam o sinal (subsistema supralaríngeo ou trato vocal).
[11] **Sons sonoros**: composto por sinais de características periódicas, gerado pela passagem forçada do ar pelas dobras vocais, vibrando-as.

diferentes sons. Desta forma, o sinal de voz que representa a fala $x(t)$ é formada pela convolução entre a fonte $e(t)$ e o filtro $h(t)$, conforme explicitado na Equação 2.26.

$$x(t) = e(t) * h(t), \tag{2.26}$$

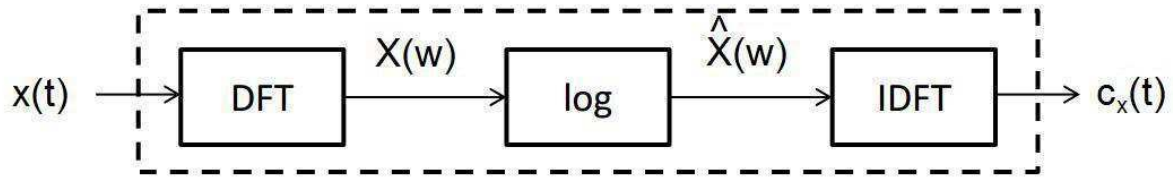
De acordo com Sanchez (2008) e Oppennheim e Schafer (2009), para que haja uma boa representação para o reconhecimento de locutor é necessário eliminar a influência mútua da fonte e do filtro. Porém, é necessário que haja a deconvolução entre eles. A obtenção do *cepstrum* é uma alternativa para esse procedimento.

Cepstrum é um anagrama da palavra *spectrum*, tendo suas primeiras quatro letras invertidas. O cepstrum pode ser definido como a inversa da transformada discreta de Fourier do logaritmo da magnitude da transformada discreta de Fourier de um sinal, conforme explicitado na Equação 2.27.

$$c_x(t) = \mathcal{F}^{-1}\{\log|\mathcal{F}\{x[n]|\}, \tag{2.27}$$

na qual $\mathcal{F}$ é a DFT e $\mathcal{F}^{-1}$ é a IDFT. Na Figura 2.11, ilustra-se o processo de transformação do sinal.

Figura 2.11 - Diagrama do processo de transformação de um sinal no domínio do tempo para o domínio cepstrum.



Fonte: Adaptado de RABINER e SCHAFER, 1978.

[12] **Sons surdos**: composto por sinais de características não periódicas, gerada pela passagem do ar pelas dobras vocais, quando a glote está totalmente aberta.

Ao aplicar a transformada discreta de Fourier, o sinal passa do domínio do tempo para o domínio da frequência, e a convolução entre a fonte $e(t)$ e o filtro $h(t)$ passa a ser um produto entre esses fatores, conforme explicitado na Equação 2.28.

$$X(w) = E(w).H(w), \tag{2.28}$$

Portanto, como a adição apresenta maior independência entre dois fatores em comparação à multiplicação destes, aplica-se o logaritmo no sinal e em seguida faz-se o uso da propriedade algébrica mostrada, como pode ser visto em seguida:

$$\log\big(X(w)\big) = \log\big(E(w).H(w)\big), \tag{2.29}$$

$$\log\big(X(w)\big) = \log\big(E(w)\big) + \log\big(H(w)\big), \tag{2.30}$$

$$\hat{X}(w) = C_e(w) + C_h(w), \tag{2.31}$$

Por fim, para o sinal retornar ao domínio do tempo, aplica-se a inversa da transformada discreta de Fourier, que atua individualmente em cada um dos fatores, conforme explicitado na Equação 2.32.

$$c_x(\mathrm{t}) = \mathcal{F}^{-1}\{C_e(w) + C_h(w)\} = \mathcal{F}^{-1}\{C_e(w)\} + \mathcal{F}^{-1}\{C_h(w)\}, \tag{2.32}$$

$$c_x(\mathrm{t}) = \mathrm{c_e} + \mathrm{c_h}, \tag{2.33}$$

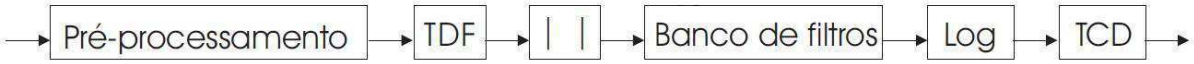
Desta forma, é possível haver a separação, por meio de filtro (*lifter*) linear, entre a fonte e o filtro, no domínio do tempo (*quefrency*). Na prática, são utilizados apenas os primeiros coeficientes componentes do *cepstrum*. Tais coeficientes contêm a informação relativa ao trato vocal, que está intimamente relacionada ao locutor (PETRY, ZANUZ e BARONE, 2000).

2.2.2.2 Coeficientes de Frequência Mel-cepstral (MFCC)

A técnica extração de características de sinais de voz baseados nos coeficientes de frequência é baseado nos estudos na área de psicoacústica [13]. A psicoacústica baseia-se em que a percepção da audição humana, em relação às frequências de tons puros ou de sinais de voz, não segue uma escala linear. Dessa forma, foi criada uma escala que se aproxima desta percepção, denominada escala mel (MENDOZA, 2009). O algoritmo foi originalmente introduzido por Davis e Mermelstein (1980) em uma publicação tradicional de discurso de processamento de voz.

O algoritmo pode ser resumido da seguinte forma: um sinal de voz, que após seu pré-processamento é provido por meio de janelas recebe a transformada discreta de Fourier, passando do domínio do tempo para o domínio da frequência. A soma dos coeficientes do módulo da potência espectral reduzida por cada um dos filtros triangulares igualmente espaçadas num eixo de frequências logarítmica é comprimido logaritmicamente e transformado a partir da Transformada do Cosseno Discreta para coeficientes cepstrais, como mostra a Figura 2.12.

Figura 2.12 - Processo de obtenção dos coeficientes mel-cepstrais.



Fonte: CARDOSO, 2009.

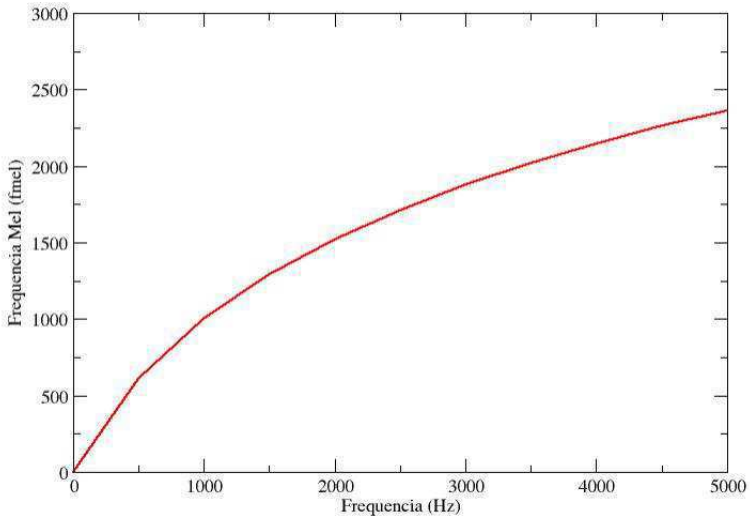
- **Escala Mel**

A escala mel foi desenvolvida por Stevens e Volkman (1940) e baseia-se no sistema auditivo humano, que percebe os sons em uma escala não linear de frequências. Dessa forma, trata-se da percepção de cada nível de tom relacionado com a frequência. O primeiro tom foi definido como 1.000 mels e tinha frequência de 1 kHz. Os tons subsequentes recebiam os valores 2.000 mels, 3.000 mels, 4.000 mels e $n\,x$ 1.000 mels, de acordo com testes auditivos em que os participantes ajustavam a frequência

[13] **Psicoacústica**: ciência que estuda a percepção auditiva humana.

física de um tom até a que frequência percebida fosse duas, três, quatro, *n* vezes a frequência de referência, como é mostrado na Figura 2.13.

Figura 2.13 - Escala mel.



Fonte: VASCONCELOS, 2011.

A partir de então, obteve-se a equação que determina a escala mel em função da frequência (Equação 2.34).

$$f_{mel} = 1127 \ln(1 + \frac{f}{700}), \tag{2.34}$$

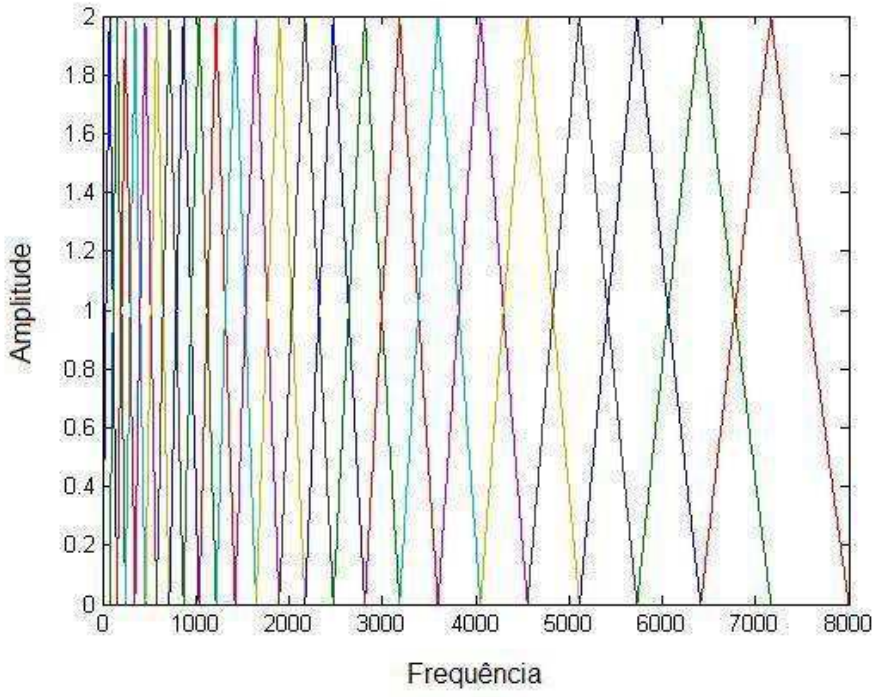
- **Banco de Filtros**

Trata-se de um conjunto de filtros de passa-faixa[14], de formato triangular com altura uniforme, espaçados uniformemente na escala mel, de forma que o inicio do filtro subsequente tenha a frequência central de seu antecessor. Assim, a largura de banda do filtro na MFCC é determinada pela gama de frequências do banco de filtro, bem

[14] **filtro passa-faixa:** permite a passagem das frequências de uma certa faixa e rejeita (atenua) as frequências fora dessa faixa

como pelo número de filtros existentes no banco (SKOWRONSKI e HARRIS, 2004), conforme ilustrado a Figura 2.14.

Figura 2.14 - Banco de Filtros baseado na escala mel contendo 24 filtros e taxa de amostragem de 16kHz.



O crescimento logarítmico da largura de banda dos filtros implica maior concentração destes filtros nas frequências mais baixas, dando maior ênfase a estas frequências (VASKAS, ESFANDIYARI e SHAMSHIRBAND, 2010).

Essa distribuição permite que as frequências centrais de cada filtro correspondam àquelas em que se percebe a mudança de tom, segundo os princípios que levaram a criação da escala mel. Além do mais, o formato de resposta triangular do filtro permite enfatizar as componentes presentes nas frequências centrais, atenuando linearmente as demais (MENDOZA, 2009).

De acordo com Vaskas, Esfandiyari e Shamshirband (2010), cada um dos *M* filtros pode ser representado pela função:

$$H_i(w) = \begin{cases} 0, & para\ \ k < fb_{i-1} \\ \frac{k - fb_{i-1}}{fb_i - fb_{i-1}}, & para\ fb_{i-1} \leq k \leq fb_i \\ \frac{fb_{i+1} - k}{fb_{i+1} - fb_i}, & para\ fb_i < k \leq fb_{i+1} \\ 0, & para\ \ k > fb_{i+1} \end{cases}, \quad (2.35)$$

na qual $i = 1,2, \dots, M$ representa o i-*ésimo* filtro, $fb_{(i)}$ são os pontos de contorno do filtro e $w = 1,2, \dots, N$ corresponde ao w-*ésimo* bloco de N-pontos que é aplicado ao filtro.

- **Cálculo dos Coeficientes**

Após ser submetido à etapa de pré-processamento, o sinal de voz encontra-se dividido em *N* segmentos, estes tendo sido submetidos a uma função janelamento.

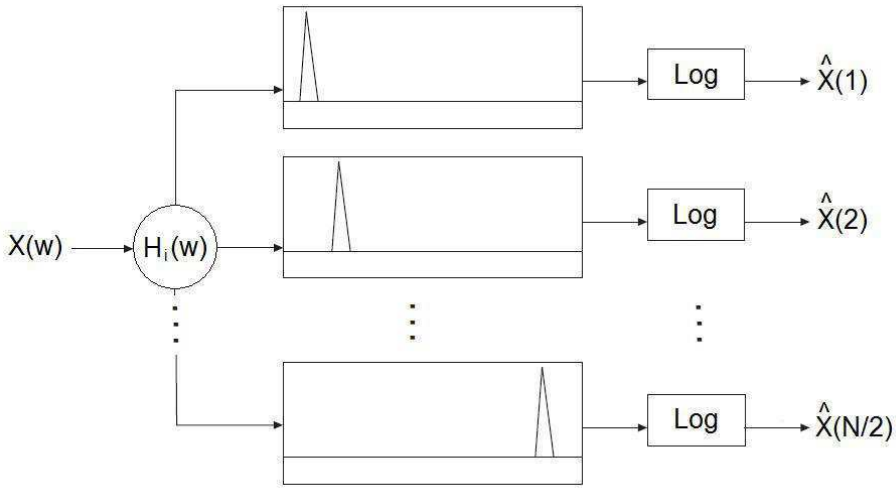
Deve-se converter o sinal de voz do domínio do tempo para o domínio da frequência. Essa conversão é realizada por meio da aplicação transformada discreta de Fourier em cada janela *n*, conforme explicitado na Equação 2.36.

$$X(w) = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn}, \quad (2.36)$$

Com o sinal no domínio da frequência é possível aplicar o banco de filtros de frequências mel-cepstrais no sinal, contendo *L* filtros e, em seguida, o logaritmo, conforme explicitado na Equação 2.37 e ilustrado na Figura 2.15.

$$\hat{X}(w) = \log\left(\sum_{w=1}^{\frac{N}{2}} |X(w)|.H_i(w) \right), \quad para\ i = 1,2, \dots, L, \quad (2.37)$$

Figura 2.15 - Ação da função do banco de filtros juntamente com a aplicação do logaritmo.



Fonte: Adaptado de CARDOSO, 2009.

De acordo com Cardoso (2009), a variação de *w* de 1 até $\frac{N}{2}$ é dada devido à redundância dos módulos dos valores de $X(w)$, devido à seguinte propriedade:

$$|X(w)| = |X(M - w - 1)|, \qquad para\ w = 1,2, \dots, M, \qquad M\ é\ par\ , \tag{2.38}$$

Por fim, é aplicada a transformada do cosseno discreta que, de forma análoga à transformada discreta de Fourier inversa, converte o sinal do domínio da frequência para o domínio do tempo, diferenciando-se desta última pelo fato de só ser aplicável a sequências reais (OPPENHEIM e SCHAFER, 2009; CARDOSO, 2009), conforme explicitado na Equação 2.39.

$$C_{mel}(j) = \sum_{i=1}^{L} X_i \,.\cos\left(j.\left(i - \frac{1}{2}\right)\frac{\pi}{M}\right), \quad para\ j = 0,1,2, \dots, L - 1, \tag{2.39}$$

Portanto, o conjunto de *j* coeficientes mel-cepstrais é denominado vetores acústicos, que são dispostos conforme a Equação 2.40.

$$C_{mel} = c_0, c_1, c_2, \dots, c_{j-1}, \tag{2.40}$$

O coeficiente c_0 contém informação do meio de transmissão, razão pela qual, muitas vezes, é desconsiderado no processo de reconhecimento de locutor.

2.2.3 Classificação de Padrões

Independentemente dos recursos empregados na extração de características, faz-se a suposição de que um modelo vai se adequar aos vetores gerados e que o mecanismo de classificação vai funcionar perfeitamente. Obviamente, isto depende da complexidade do modelo (que é proporcional ao tempo de execução do algoritmo). Em qualquer caso, é preciso tomar uma decisão de como modelar um locutor e qual o esquema de classificação apropriado (SANDOUK, 2012).

A extração das características dos sinais de voz são a base para a criação dos modelos dos locutores. A cada locutor é representado por um modelo construído a partir das características extraídas de seus sinais de voz, modelo este armazenado na base de dados do sistema (LIMA, 2001).

De acordo com Campbell Jr (1997), existem dois tipos principais de modelos: o modelo estatístico e o modelo baseado em casamento de padrões característicos, denominado paramétrico.

O modelo estatístico, posteriormente a observação de todos os padrões da sua base de dados, possui um sistema de classificação baseado em razão de verossimilhança ou probabilidades condicionadas. O modelo paramétrico compara os modelos questionados com os padrões, como se fossem uma cópia do modelo contido na base de dados. A diferença entre a cópia e o original é dada por medidas de distâncias, *e.g.* Euclidiana, Mahalanobis (LIMA, 2001).

Nos métodos estatísticos, modelos ocultos de Markov (HMM) e modelos de misturas gaussianas (GMM) são os mais utilizados. Os métodos paramétricos mais conhecidos são o *Dynamic Time Warping* (DTW) e a quantização vetorial (VQ) (CHEN,

HSIEH e HSU, 2008).

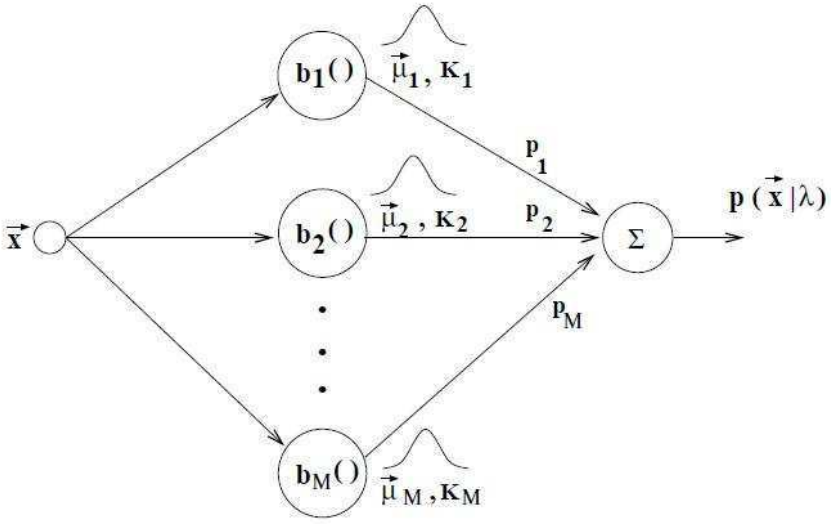
2.2.3.1 Modelo de Misturas Gaussianas (GMM)

Os modelos de misturas gaussianas (GMM) foram primeiramente utilizados para reconhecimento de locutor por Reynolds (1992). Desde então são considerada referência para modelagem de locutores.

Trata-se de um modelo composto por um conjunto de funções de densidade de probabilidades gaussianas, que podem modelar várias classes fonéticas, desconsiderando a evolução temporal do sinal, sendo próprio para sistemas de reconhecimento de locutor independente de texto (LIMA, 2001).

Uma mistura de densidades de probabilidades gaussianas é representada pelo somatório ponderado de *M* densidades, conforme mostrado na Figura 2.16, e dada pela Equação 2.41:

Figura 2.16 - M densidades de probabilidade formando um GMM.



Fonte: REYNOLDS, 1992.

$$P(\overrightarrow{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}), \tag{2.41}$$

em que $\vec{x}$ é um vetor de características de dimensão *N*, $b_i(\vec{x})$, *para i=1,2,...,M,* são as densidades componentes e p_i, *para i=1,2,...,M* é o peso das misturas. As densidades componentes são representadas por uma função gaussiana de dimensão *N* da forma:

$$b_i(\vec{x}) = \frac{1}{2\pi^{\frac{N}{2}}|\Sigma_i|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(\vec{x} - \overrightarrow{\mu_i})' \Sigma_i^{-1}(\vec{x} - \overrightarrow{\mu_i})\right\}, \qquad para\ i = 1, 2, \ldots, M, \tag{2.42}$$

com vetor de média $\overrightarrow{\mu_i}$ e matriz de covariância Σ_i. Os pesos das misturas devem obedecer à seguinte condição:

$$\sum_{i=1}^{M} p_i = 1, \tag{2.43}$$

A densidade de mistura gaussiana possui como parâmetros um vetor de médias, uma matriz de covariância e coeficientes da mistura ponderada de todas as densidades componentes (modelo λ). A notação que representa estes parâmetros é dada por:

$$\lambda = \{p_i, \overrightarrow{\mu_i}, \Sigma_i\}, \quad \text{para i} = 1, 2, \ldots, \text{M}, \tag{2.44}$$

- **Treinamento**

O vetor de características extraído nas fases preliminares $X = \{\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_T}\}$ contém as propriedades inerentes a cada locutor. Tomando como base a independência entre os vetores característicos $\overrightarrow{x_t}$, segundo Reynolds (1995), a probabilidade de observação do conjunto X dado o modelo λ é representado pela Equação 2.45.

$$P(X|\lambda) = P(\overrightarrow{x_1}|\lambda)P(\overrightarrow{x_2}|\lambda)\dots P(\overrightarrow{x_T}|\lambda) = \prod_{t=1}^{T} P(\overrightarrow{x_t}|\lambda), \quad (2.45)$$

Normalizando-se o somatório da probabilidades condicionadas pelo número total de vetores e usando-se o logaritmo, tem-se:

$$log\ P(X|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log P(\overrightarrow{x_t}|\lambda), \quad (2.46)$$

Como cada locutor é representado por um modelo λ, se faz necessária a estimação de cada parâmetro que compõe o modelo definido na Equação 2.41.

Existem vários métodos de estimação de parâmetros para modelos de misturas gaussianas. Um método bastante difundido e que apresenta bons resultados é a estimação da máxima verossimilhança (ML). O algoritmo de máxima expectativa (EM) implementa esse método de forma iterativa, utilizando condição de parada que leva em consideração o número de interações e limiar de decisão (REYNOLDS, 1995).

- **Máxima expectativa (EM)**

Trata-se de um algoritmo, tendo como modelo inicial λ^0, que a cada iteração estima um novo modelo λ^{n+1}, objetivando que este represente mais fielmente o conjunto de características X, comparado ao modelo criado na iteração anterior λ^n, de acordo com a Equação 2.47.

$$P(X|\lambda^{n+1}) \geq P(X|\lambda^n), \quad (2.47)$$

O modelo inicial λ^0 pode ser gerado de duas formas: aleatoriamente ou por clusterização.

Aleatoriamente, conforme o próprio nome sugere, as médias $\vec{\mu}$ são representadas por componentes aleatórios de cada vetor de características. A matriz de

covariância Σ é dada por uma matriz identidade e os pesos das misturas p são inicializados de forma uniforme, dividindo-se o valor 1 pela quantidade de misturadas adotadas (NERI, 2012).

Por clusterização, as médias $\vec{\mu}$ são obtidas por meio dos centros de cada grupo, sendo a quantidade de grupos igual a quantidade de misturas gaussianas adotadas. A matriz de covariância Σ é obtida por meio da variância entre os dados e o centro de cada grupo e os pesos das misturas p são inicializados de forma uniforme, dividindo-se o valor 1 pela quantidade de misturadas adotadas (NERI, 2012). Os algoritmos mais comuns para inicialização por clusterização são o k-*means* e LBG (*Linde Buzo and Gray*).

O EM é um algoritmo de maximização que visa o máximo local, possuindo duas etapas, a E (*Expectation*) e a M (*Maximization*). Estas se repetem até que uma de duas de suas condições de parada sejam satisfeitas, quando o algoritmo atinge um número máximo de iterações ou quando a diferença relativa do logaritmo da verossimilhança entre o modelo atual e o anterior atinge um determinado limiar, indicando que a etapa de treinamento não consegue mais melhorar os parâmetros do modelo (NERI, 2012). A diferença relativa é expressa pela Equação 2.48.

$$\frac{\mathrm{Log}\,P\left(X\middle|\bar{\lambda}\right) - \log P(X\,|\,\lambda)}{\log P(X\,|\,\lambda)} < \theta, \tag{2.48}$$

Na etapa E (*Expectation*), calcula-se a razão de verossimilhança entre o um modelo específico e o somatório de todos os modelos, para cada vetor de treinamento $\overrightarrow{x_t}$, conforme explicitado na Equação 2.49.

$$P(i|\overrightarrow{x_T},\lambda) = \frac{p_i b_i(\overrightarrow{x_t})}{\sum_{i=1}^{M} p_i b_i(\overrightarrow{x_t})}, \tag{2.49}$$

Na etapa seguinte, M (*Maximization*), os parâmetros do modelo são atualizados a partir da razão obtida na fase E (*Expectation*). O objetivo desta etapa é a criação de

um modelo que represente com maior fidelidade os dados de treinamento em comparação ao modelo da iteração anterior. As Equações que representam o novo modelo $\bar{\lambda}$ a partir do modelo antigo λ, são as seguintes:

$$\bar{p_i} = \frac{1}{T}\sum_{t=1}^{T} P(i|\overrightarrow{x_t}, \lambda), \tag{2.50}$$

$$\bar{\mu_i} = \frac{\frac{1}{T}\sum_{t=1}^{T} P(i|\overrightarrow{x_t}, \lambda)\, x_t}{\frac{1}{T}\sum_{t=1}^{T} P(i|\overrightarrow{x_t}, \lambda)}, \tag{2.51}$$

$$\bar{\Sigma_i} = \frac{\frac{1}{T}\sum_{t=1}^{T} P(i|\overrightarrow{x_t}, \lambda)\, (x_t - \bar{\mu_i})(x_t - \bar{\mu_i})'}{\frac{1}{T}\sum_{t=1}^{T} P(i|\overrightarrow{x_t}, \lambda)}, \tag{2.52}$$

- **Tomada de decisão para identificação**

No sistema automático de identificação utiliza-se um classificador baseado na máxima-verossimilhança. Cada modelo de mistura gaussiana $\lambda_1, \lambda_2, \dots, \lambda_S$ representa, de forma respectiva, o grupo de *S* locutores $S = \{1, 2, \dots, S\}$. Por fim, o classificador encontra o locutor que apresenta maior similaridade com o sinal de voz questionado (LIMA, 2001). Este classificador é representado matematicamente pela Equação 2.53.

$$\hat{S} = \arg \max_{1 \leq \mathrm{k} \leq \mathrm{S}} P(\lambda_k|X) = \arg \max_{1 \leq \mathrm{k} \leq \mathrm{S}} \frac{P(X|\lambda_k)P(\lambda_k)}{P(X)}, \tag{2.53}$$

A parcela da Equação 2.54 correspondente à regra de Bayes pode ser simplificada para a Equação 2.53, inferindo-se que haja igualdade de condições para a classificação de cada locutor $P(\lambda_k) = {}^{1}/_{S}$ e que a probabilidade da elocução questionada $P(X)$ seja a mesma para todos os casos.

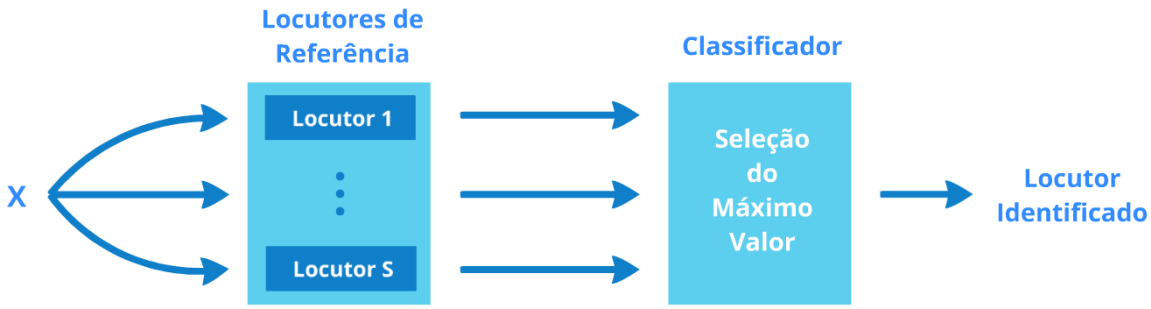
$$\hat{S} = \arg \max_{1 \leq k \leq S} P(X|\lambda_k), \tag{2.54}$$

De acordo com a propriedade logarítmica logaritmo e a independência entre os vetores característicos, a regra de decisão para a identificação é dada pela Equação 2.55:

$$\hat{S} = \arg \max_{1 \leq k \leq S} \sum_{t=1}^{T} \log P(\overrightarrow{x_t}|\lambda_k), \tag{2.55}$$

em que $P(\overrightarrow{x_t}|\lambda_k)$ é dada pela Equação 2.41. O diagrama em blocos do classificador de um sistema de identificação de locutor é apresentado na Figura 2.17.

Figura 2.17 - Classificador de um sistema de identificação de locutor com S locutores.



Fonte: Adaptado de LIMA, 2001.

2.3 Considerações Gerais

Apresentaram-se, neste capítulo, os principais conceitos sobre a voz humana e sua fisiologia. No que diz respeito ao sinal de voz, foram descritas as etapas de digitalização, pré-processamento, extração de características, modelagem e classificação dos padrões, e por fim, tomada de decisão. Dessa forma, proporcionou-se uma base para o entendimento do problema levantado nesta pesquisa.

Diante do exposto, foi possível compreender o funcionamento da voz humana, assim como sua digitalização, a minimização dos efeitos de ruídos, a extração de peculiaridades dos sinais de voz podendo assim, individualizar cada locutor a partir desta característica humana.

No próximo capítulo será apontado o papel do reconhecimento de locutor na Criminalística, mostrando onde encontra-se inserido na Fonética Forense, além da apresentação de diversas pesquisas aplicados à área forense e/ou voltados à identidade vocal em ambientes telefônicos e com presença de ruídos.

3 RECONHECIMENTO DE LOCUTOR APLICADO À CRIMINALÍSTICA

É recorrente o anseio de poder identificar uma pessoa com base em sua voz. Por muitos anos, juízes, advogados, detetives e autoridades policiais têm interesse em usar a autenticação de voz forense para investigar um suspeito ou confirmar uma sentença de culpa ou inocência (CAMPBELL, SHEN, *et al.*, 2009).

Registros de vídeos ou de áudio estão cada vez mais presentes no mundo atual e são muito utilizados como provas de delitos. São provas bastante fortes e, na maioria das vezes, incontestáveis. Entretanto, para que tenham poder probante, é necessário a elaboração de laudo pericial, a fim de se constatar o conteúdo do material gravado e sua autenticidade. Também, quando questionado, é necessário comprovar que a voz do interlocutor pertence realmente à pessoa que está sendo investigada.

Ainda levando em consideração as peculiaridades deste tipo de reconhecimento, uma característica importante é a baixa qualidade das gravações, devido ao ruído gerado, por se tratarem, em sua maioria, de interceptações telefônicas. Este fator é determinante e deve ser levado em consideração, pois influencia negativamente o reconhecimento.

De acordo com Espindula e Toccheto (2005), o Perito Criminal faz uso de vários conhecimentos de diversas ciências, tendo como objetivo elucidar ilícitos penais. A junção de todas essas ciências, voltadas para o estudo sistemático dos objetos e locais envolvidos em um crime é denominada Criminalística. Uma dessas ciências é a Fonética Forense, que tem como principal objetivo determinar a autoria e/ou autenticidade de sinais de voz.

3.1 Fonética Forense

Uma das atribuições da fonética é o reconhecimento humano por meio das

características contidas na voz de determinado indivíduo. Essas características podem ser relacionadas à fisiologia humana, como também à questões regionais e sociais, semântica ou estado emocional (BRAID, 2009).

A Fonética Forense é uma subária da linguística voltada para a elucidação de delitos na qual existam registro de vozes em qualquer tipo de mídia, sendo praticamente insubstituíveis nesses casos (BRAID, 2009).

Desta forma, a Fonética Forense se divide em subáreas: reconhecimento de locutor, verificação de edição e análise de conteúdo fonográfico (RIBEIRO *et al.*, 2008).

3.1.1 Reconhecimento de locutor

Reconhecer determinado locutor a partir de elocuções questionadas oriundas de qualquer tipo de mídia. Esse reconhecimento é realizado a partir da comparação das características extraídas do sinal de voz questionado com todos os modelos contidos na base de dados do sistema, no caso de identificação, ou com apenas um modelo, no caso de verificação. O exame pericial resulta um laudo técnico, apontando se o sinal de voz questionado pertence ou não a determinado indivíduo, funcionando como prova material [15]. Em um caso real, a prova pericial tem papel importante e subsidia fortemente a convicção do juiz em sua decisão final (RIBEIRO *et al.*, 2008).

3.1.2 Verificação de edição

É a atividade pericial dentro da Fonética Forense que tem como objetivo examinar a autenticidade de arquivos de áudio contidos em uma mídia, verificando a ocorrência de edições, *e.g.* modificação, supressão ou acréscimo (RIBEIRO *et al.*, 2008).

[15] **prova material**: Provas produzidas a partir de vestígios encontrados no local do crime. Sendo uma das provas admitidas no processo penal brasileiro

3.1.3 Análise de conteúdo fonográfico

É a descrição e filtragem por relevância do conteúdo de áudio contido em uma mídia. Possui maior valor em relação à transcrição textual, visto que neste método não estão inclusas características como entonação, velocidade da fala, regionalismos e som ambiente (RIBEIRO *et al.*, 2008).

3.2 Requisitos dos sinais de voz para reconhecimento de locutor

O exame de reconhecimento de locutor depende de fatores como: a autenticidade e espontaneidade do material padrão; a adequabilidade do padrão ao questionado e do questionado aos exames; a contemporaneidade dos registros de voz confrontados; e a quantidade de material (repetições de segmentos fonológicos) (SCATENA, 2010).

3.2.1 Autenticidade

Os sinais de voz voltados para o treinamento dos locutores, ou seja, as amostras de dados padrão, devem ser coletadas e manipuladas pelo perito realizador do exame pericial. A comparação deve ser realizada com o sinal de voz questionado original (SCATENA, 2010).

3.2.2 Adequabilidade

Os sinais de voz que compõem as amostras de dados padrão e de dados questionados devem ser produzidas sob as mesmas condições, ou as mais similares possíveis, havendo isonomia no método de gravação, no que diz respeito à qualidade e presença de ruído ambiente. (SCATENA, 2010).

3.2.3 Contemporaneidade

Os sinais de voz que compõem as amostras de dados padrão e de dados questionados devem ser produzidas em data próximas, pois a diferença de tempo entre a produção do sinal de voz questionado em relação ao sinal de voz padrão pode prejudicar o resultado final, visto que a ação do tempo modifica o trato vocal, que é o responsável por grande parte das características relativas à voz (SCATENA, 2010).

3.2.4 Espontaneidade

É de bom grado que o suspeito forneça os sinais de voz, que compõem a amostra de dados padrão, de forma espontânea, não alterando a autenticidade do material (SCATENA, 2010).

3.2.5 Quantidade

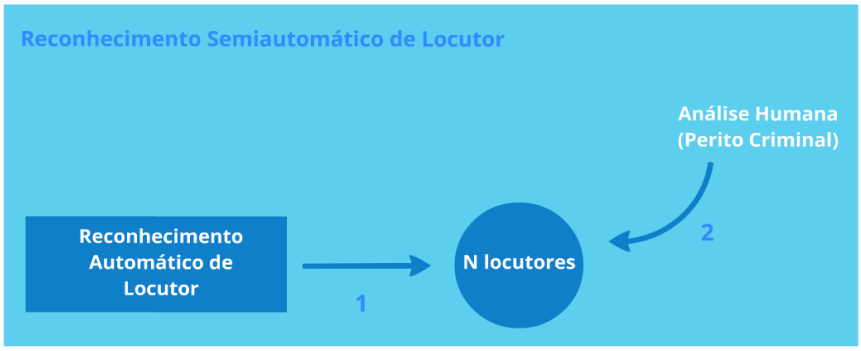
O perito deve responsabilizar-se pela coleta do material padrão, no mínimo, em duas oportunidades, prevenindo-se de situações fortuitas que fogem do seu controle, *e.g.* erros de gravação, perda do material (SCATENA, 2010).

3.3 Reconhecimento semiautomático de locutor

Essencialmente, o resultado de um sistema de reconhecimento automático de locutor baseia-se apenas nas características extraídas das elocuções de teste e de treinamento. Essa fonte de subsídio única, de certa forma, não é suficiente para uma conclusão com um fim de tamanha importância, podendo levar, de forma equivocada, a desvinculação de um criminoso a um fato delituoso e, principalmente, incriminar um inocente. Além do panorama em questão, o ambiente telefônico, que possui fatores que minimizam o desempenho do reconhecimento, conforme explicitado previamente no item 1.2.4, também deve ser considerado.

Também é importante citar a principal conclusão de Rose (2002) que, em casos forenses, diferentes técnicas devem ser utilizadas conjuntamente em comparações de voz, a partir do uso tanto de informação auditiva (escuta), quanto visual (espectrograma, forma de onda) e estatística, conforme ilustrado na Figura 3.1.

Figura 3.1 - Diagrama do Sistema Semiautomático de Locutor



No caso da identificação de locutor, a elocução que tiver maior similaridade entre todos os modelos é dita como a originadora da elocução. Porém, para Svirava (2009), no âmbito jurídico, este resultado não é suficiente. O autor afirma que, em diversos países, inclusive na Austrália, a força da evidência não está apenas no resultado ou na similaridade apresentada pela comparação das características da elocução de teste com a de treinamento e sim na razão de duas probabilidades: (i) a probabilidade de que a elocução tenha sido originada pelo suspeito; e (ii) a probabilidade de que a elocução não tenha sido originada pelo suspeito. Esta relação é denominada razão de verossimilhança (RV).

3.3.1 Razão de Verossimilhança

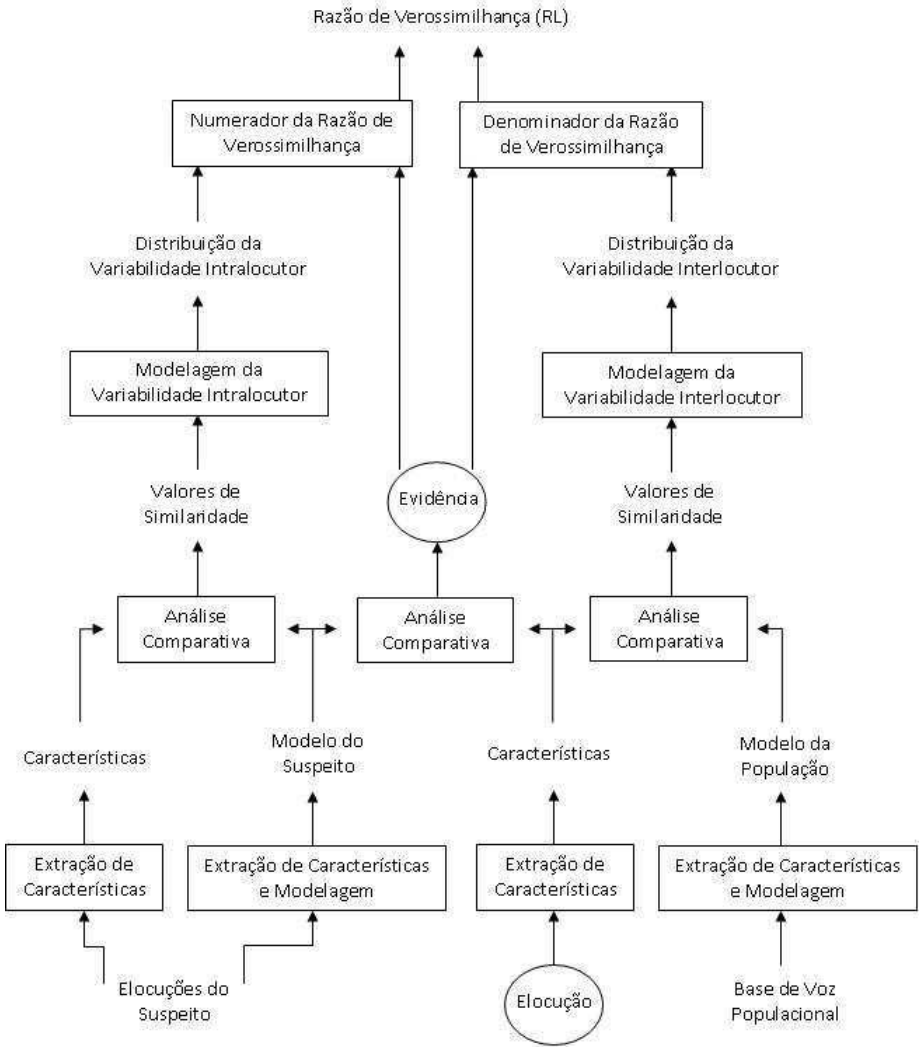
A razão de verossimilhança (RV) expressa, de forma quantitativa, o número de vezes em que é mais provável (ou menos) a ocorrência de uma hipótese (H_0), em comparação com sua negativa (H_1), conforme a Equação 3.1.

$$RV = \frac{P(H_0)}{P(H_1)}, \tag{3.1}$$

Essa metodologia, baseada em razões de verossimilhança, foi pioneira na Criminalística na área de Genética Forense. Dessa forma, verificou-se sua viabilidade de implementação e que é passível de aplicação nas demais ciências que compõem a Criminalística (VALENTE, 2012).

Drygajlo (2007) propõe que, para encontrar as probabilidades supramencionadas, sejam levadas em consideração as variações interlocutores (variações das elocuções de todos os locutores) e intralocutores (variação das elocuções do mesmo locutor), como apresentado na Figura 3.2.

Figura 3.2 - Diagrama do processamento da evidência e interpretação do sistema.



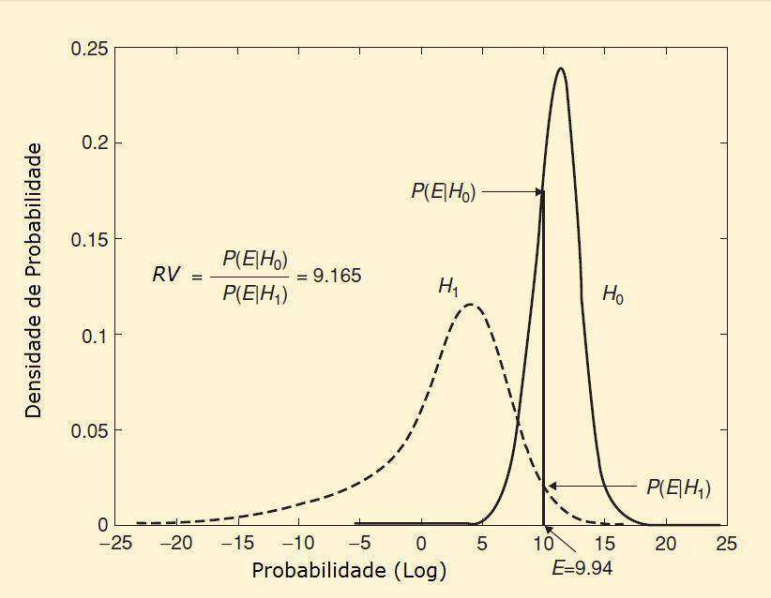
Fonte: Adaptado de DRYGAJLO, 2007.

Primeiramente, é obtida a similaridade entre a elocução questionada e a base de dados de referência do suspeito, denominada *E*.

O segundo passo consiste em obter as distribuições das variações interlocutores e intralocutores. A distribuição intralocutor é obtida pela similaridade entre as elocuções do suspeito. A distribuição interlocutor é obtida pela similaridade da elocução questionada e uma base de dados populacional, contendo diversos sinais de vozes.

Conforme visto na Figura 3.3, de posse da similaridade *E,* verifica-se qual é probabilidade desse valor nas distribuições intralocutor e interlocutor, as quais correspondem, respectivamente, ao numerador e denominador da razão de verossimilhança apresentada na Equação 3.2.

Figura 3.3 - A Razão de Verossimilhança (RV) obtida por meio do valor da Evidência (E) nas distribuições intralocutor e interlocutor.



Fonte: Adaptado de DRYGAJLO, 2007.

$$\frac{P(E|H_0)}{P(E|H_1)}, \tag{3.2}$$

em que H_0 é a hipótese de que a elocução tenha sido originada pelo suspeito e H_1 de que a elocução não pertença ao suspeito em questão.

Como o objetivo é encontrar a razão de verossimilhança entre as hipóteses de que a elocução tenha sido originada pelo suspeito, dada a existência da evidência ($P(H_0|E)$) e que a elocução não tenha sido originada pelo suspeito, dado a existência da evidência ($P(H_1|E)$), utiliza-se o Teorema de Bayes, conforme explicitado na Equação 3.3.

$$\frac{P(H_0|E)}{P(H_1|E)} = \frac{P(E|H_0)}{P(E|H_1)} \times \frac{P(H_0)}{P(H_1)}, \tag{3.3}$$

O modelo sugere que, para se obter o resultado final, a probabilidade *a posteriori*, seja usada uma probabilidade subjetiva denominada probabilidade *a priori*, baseada em outras provas do âmbito jurídico (depoimentos, acariações vestígios, entre outros), que é a probabilidade de que a elocução tenha sido originada pelo suspeito ($P(H_0)$) e de que não tenha sido originada por esse suspeito ($P(H_1)$).

Desta forma, a razão de verossimilhança é igual à razão entre a probabilidade de que a elocução tenha sido originada pelo suspeito, dada a existência da evidência e a probabilidade de que a elocução não tenha sido originada pelo suspeito, dada a existência da evidência.

A partir do numerador e denominador que compõem probabilidade *a posteriori*, obtém-se uma razão que pode ser transformada em uma probabilidade x, de acordo com Drygajlo (2007), conforme Equação 3.4.

$$\left(\frac{P(H_0|E)}{P(H_1|E)}\right) x + x = 1, \tag{3.4}$$

Por outro lado, Svirava (2009) faz uma relação qualitativa da taxa obtida, a qual possui vários níveis de força da evidência, variando de 10.000 até 0,0001. O autor afirma que se essa relação for maior que 10.000, tem-se uma evidência muito forte a favor e sendo menor que 0,0001 implica uma evidência muito forte contra, conforme sintetizado no Quadro 3.1.

Quadro 3.1 - Valores equivalentes com relação à taxa de probabilidade.

Taxa de probabilidade	Valor equivalente
> 10.000	Evidência muito forte a favor
1.000 – 10.000	Evidência forte a favor
100 – 1.000	Evidência moderadamente forte a favor
10 – 100	Evidência moderada a favor
1 – 10	Evidência limitada a favor
1 – 0,1	Evidência limitada contra
0,1 – 0,01	Evidência moderada contra
0,01 – 0,001	Evidência moderadamente forte contra
0,001 – 0,0001	Evidência forte contra
< 0,0001	Evidência muito forte contra

Fonte: Adaptado de SVIRAVA, 2009.

3.4 Trabalhos Relacionados

O processo de reconhecimento de locutor é dividido em 5 etapas: (i) pré-processamento; (ii) extração de características; (iii) criação dos padrões (treinamento); (iv) correspondência dos padrões; e (v) tomada de decisão. Assim sendo, são revisadas, nas próximas subseções, publicações relacionadas aos temas supramencionados, voltados para identificação de locutor em ambiente telefônico e com presença de ruídos.

A pesquisa de Hanilci e Ertas (2011) visou à comparação da Quantização Vetorial (VQ) e dos Modelos de Misturas Gaussianas (GMM) como técnicas de criação e correspondência de padrões quando destinados à identificação de locutor independente de texto. A extração de características foi realizada a partir do uso de coeficientes de frequências mel-cepstrais (MFCC). Os autores utilizam a base de dados de vozes NTIMIT (JANKOWSKI *et. al* 1990), constituída por sinais de voz da base TIMIT transmitidos pelo ambiente telefônico e, assim, regravados com a respectiva qualidade (8 kHz). Foram utilizados 168 locutores, sendo 112 do sexo masculino e 56

do sexo feminino. Para o treinamento de cada locutor foram usadas 8 elocuções, com aproximadamente 24 segundos de duração. Para os testes foram utilizadas apenas 2 elocuções. O Modelo de Misturas Gaussianas apresentou maior eficiência em relação à Quantização Vetorial, atingindo as taxas de identificação de 73,21% e 68,15%, respectivamente.

No estudo de Chen, Hsieh e Hsu (2008), houve a verificação da identidade de locutores a partir da voz, independente do texto que venha a ser mencionado. As características extraídas dos sinais de voz são os coeficientes cepstrais, obtidos a partir dos coeficientes LPC (LPCC). Anterior à extração das características, tanto para treinamento, quanto para teste, foi feita a normalização da média cepstral (CMN), visando eliminar o viés do canal produzido pela aquisição do sinal de voz em diferentes microfones. A criação e classificação dos padrões deu-se por meio da quantização vetorial em dois estágios. Para cada locutor, foram criados n modelos, baseados na faixa de frequência total e em diversas sub-bandas do sinal de voz. Desta forma, na tomada de decisão há um somatório das similaridas dos n modelos de cada locutor.

Para a realização dos experimentos de identificação de locutor, foi utilizado o banco de dados de voz KING. Trata-se de uma coleção de amostras de elocuções de 51 locutores do sexo masculino. Para cada locutor, existem 10 elocuções da mesma frase que foram gravadas em momentos distintos, tendo cada uma aproximadamente 30 segundos de duração Os sinais de voz foram digitalizados a uma taxa de 8 kHz e 16 bits/amostra, ou seja, qualidade telefônica. As locuções ruidosas foram geradas pela adição de ruído gaussiano em elocuções sem ruído, de acordo com a SNR[16] (relação sinal-ruído) desejada. A partir de um conjunto de experimentos, variando o número de sub-bandas utilizadas e número de vetores nos 2 dicionários do modelo, chegou-se no melhor resultado, obtendo taxa de acerto de 95,36% em ambiente limpo, 93,02% com 20 dB de SNR, 86,38% com 15 dB de SNR, 75,08% com 10 dB de SNR e 53,49% com 5 dB de SNR.

[16] **Relação sinal-ruído** ou **razão sinal-ruído** é um conceito de telecomunicações, também usado em diversos outros campos que envolvem medidas de um sinal em meio ruidoso, definido como a razão da potência de um sinal e a potência do ruído sobreposto ao sinal.

A pesquisa de Aldhaheri e Al-Saadi (2003) também é fundamentou-se na identificação de locutor independente de texto. Utilizaram-se dois métodos, em separado, para a extração de características: (i) Coeficientes Cepstrais por Predição Linear (LPCC) e (ii) Coeficientes obtidos a partir da Predição Linear (LPC). Aliado à extração de características, fez-se uso de um modelo de classificação denominado SVD (*Singular Value Decomposition*) para correspondência dos padrões. Foi utilizada uma base de dados de voz própria, gravada em ambientes limpo e ruidoso. Os sinais de voz foram digitalizados à taxa de 8 kHz e 8 bits/amostra, ou seja, qualidade telefônica. Foram utilizados 10 locutores, sendo 7 do sexo masculino e 3 do sexo feminino. Para o treinamento de cada locutor, foram usadas 80 elocuções, com aproximadamente 3 segundos de duração. O procedimento apresentou melhores resultados com o uso da extração de características por meio dos coeficientes LPC cepstrais, com taxa de acerto de 99,5% em ambiente limpo, 95,5% com 20 dB de SNR, 93,5% com 15 dB de SNR, 88,5% com 10 dB de SNR e 82,5% com 5 dB de SNR. Utilizando os coeficientes LPC, os resultados foram 91,5% em ambiente limpo, 82,5% com 20 dB de SNR, 76,0% com 15 dB de SNR, 68,0% com 10 dB de SNR e 57,0% com 5dB de SNR.

Na pesquisa de Chetouani *et al.* (2009), compararam-se os coeficientes mel-cepestrais (MFCC) e os coeficientes cepestrais de predição linear (LPCC) como técnicas de extração de características, métodos estes voltados para a identificação de locutor independente de texto. Para a criação e a correspondência de padrões, foi utilizado um método estatístico, por meio da matriz de covariância, para medir a similaridade. Os autores utilizaram a base de dados de vozes NTIMIT, constituída por sinais de voz da base TIMIT transmitidos pelo ambiente telefônico e, assim, regravados com a respectiva qualidade (8 kHz). Foram utilizados 630 locutores, sendo 438 do sexo masculino e 192 do sexo feminino. Para o treinamento de cada locutor foram usadas 5 elocuções, com aproximadamente 15 segundos de duração. Para os testes foram utilizadas 5 elocuções. O uso dos coeficientes de frequência mel-cepstrais (MFCC) apresentou maior eficiência em relação ao uso dos coeficientes cepestrais de predição linear (LPCC), atingindo as taxas de identificação de 27,3% e 24,6%, respectivamente.

Nas pesquisas de Ming *et al.* (2007) e de Hsieh e Hsu (2008), foram realizadas extração das características subdivisões da faixa total de frequência para a identificação de locutor independente de texto. Foram utilizadas, em conjunto, a extração de características por meio de coeficientes de frequências mel-cepstrais (MFCC) e a correspondência de padrões por modelo de misturas gaussianas (GMM), com variações na quantidade de misturas gaussianas por locutor, essas sendo de 32, 64 e 128. Para a realização dos experimentos de identificação de locutor, foi utilizado o banco de dados de voz TIMIT, que são sinais de vozes obtidos em ambientes limpos. Os sinais de voz foram digitalizados a uma taxa de 16 kHz. Foram utilizados 630 locutores, sendo 438 do sexo masculino e 192 do sexo feminino. Para o treinamento de cada locutor foram usadas 8 elocuções, com aproximadamente 3 segundos de duração. Para os testes, foram utilizadas apenas 2 elocuções. As locuções ruidosas foram geradas pela adição de ruído de diversos ambientes reais em elocuções sem ruído, como sons de motores, toques de celular, música, noticiário, entre outros. Esses ruídos foram adicionados para as SNR de 10 dB, 15 dB e 20 dB. Por fim, as taxas de acerto obtidas em ambiente limpo com a quantidade de misturas gaussianas por locutor de 32, 64 e 128 foram, respectivamente, 90,64%, 94,84% e 96,51%. Com a adição de ruído, na grande maioria dos casos, assim como aconteceu no ambiente limpo, quanto maior o número de misturas gaussianas por locutor, melhor foi o resultado da identificação.

Hanilci *e* Ertas (2009) mencionaram que a classificação baseada na análise de componentes principais (PCA) tem sido utilizada mais recentemente na literatura para aplicações de identificação de locutor independente de texto. Porém, os resultados obtidos foram apenas para ambientes livres de ruídos. Na referida pesquisa, foi aplicada esta técnica de classificação, combinada com quantização vetorial (VQ), utilizando elocuções oriundas de ambiente telefônico. Para ambos os métodos de classificação, foi utilizada a mesma técnica para extração de características, fazendo uso dos coeficientes de frequências mel-cepstrais (MFCC). Os autores utilizaram as bases de dados de vozes TIMIT, composta por sinais de vozes obtidos em ambientes limpos e NTIMIT, constituída por sinais de voz da base TIMIT transmitidos pelo ambiente telefônico e, assim, regravados com a respectiva qualidade (8 kHz). Foram

utilizados 168 locutores, sendo 112 do sexo masculino e 56 do sexo feminino. Para o treinamento de cada locutor foram usadas 8 elocuções, com aproximadamente 24 segundos de duração. Para os testes foram utilizadas apenas 2 elocuções. Os resultados dos métodos de classificação foram observados de forma isolada. A técnica PCA, em ambientes limpos, apresentou taxas de identificação tão boas quanto à técnica de VQ (97,8% e 99,7%, respectivamente). Porém, em se tratando de ambientes ruidosos, os resultados do PCA se apresentam inferiores aos obtidos com VQ (38,9% e 75,0%, respectivamente). Porém, com os dois métodos em conjunto, utilizando uma técnica de decisão denominada fusão de opiniões (*Opinion fusion*), que consiste no uso das pontuações das duas técnicas (d_1 e d_2), além de um fator de combinação (β) que determina a importância de cada pontuação no resultado final, os resultados apresentaram melhorias, com taxas de acerto de 100% em ambientes limpos e 78,2% em ambientes telefônicos.

Cetnarowicz, Drgas e Dabrowski (2010) apresentaram uma técnica de reconhecimento de locutor voltada para elocuções provenientes de sinais telefônicos. Nos experimentos, são utilizados coeficientes mel cepstrais (MFCC) e o modelo de classificação baseado em misturas gaussianas (GMM), justificando-se seu uso a partir da afirmação de que o estado da arte no reconhecimento de locutor foi atingido a partir dessa modelagem. Utilizou-se a base de dados de vozes CORPORA, que são sinais de vozes obtidos em ambientes limpos convertidos em qualidade telefônica por meio de transmissão nesse meio. Foram utilizados 45 locutores, sendo 28 do sexo masculino, 11 do sexo feminino e 6 crianças. Para cada locutor, foram utilizadas 114 sentenças de aproximadamente 2 segundos. Os autores, em busca da melhor taxa de aceitação na identificação de locutor, fizeram diversas variações nos parâmetros que representavam as características, tais como modificações no número de sub-bandas e uso de coeficientes mel cepstrais, além do uso ou não da normalização da média cepstral (CMN), que visa eliminar o viés do canal de comunicação. Por fim, em ambientes limpos, conseguiram uma taxa de aceitação de 87,0% utilizando 32 sub-bandas, 12 coeficientes mel cepstrais e descartando o uso da CMN. Em ambiente telefônico, foi

obtida a taxa de aceitação de 73,4% utilizando 28 sub-bandas, 20 coeficientes mel cepstrais e também descartando o uso da CMN.

Nakasone e Steven (2001) afirmaram que, em uma avaliação de diversos sistemas de reconhecimento automático de locutor, verificou-se que um modelo baseado em GMM, aliado a uma técnica de extração de características robusta a variações de canal, apresentou os melhores resultados. Por isso, a escolha pelo uso do modelo de misturas gaussianas (GMM) como classificador de padrões em seu estudo. Os autores não mencionaram qual a técnica que utilizaram no processo de extração de características, afirmando apenas o uso de características de baixo nível e de um banco de voz próprio digitalizado a 16 kHz, com 16 bits/amostra que fora convertido para taxa de 7,5 kHz, ou seja, próxima a qualidade telefônica. A amostra considerada foi de 50 locutores, sendo todos do sexo masculino. Para o treinamento de cada locutor, foram coletadas 10 elocuções, com aproximadamente 57 segundos de duração. Para transmissões em telefone fixo foi obtida a excelente taxa de aceitação de 100%, havendo o uso de normalização de canal. Sem a normalização, esta taxa caiu para 97,5%. Na transmissão utilizando telefone móvel, os resultados não foram bons, apresentando taxas de aceitação de 58,3% e 30% com e sem o uso de normalização de canal, respectivamente.

Cardoso (2009) apresentou uma técnica generalista que propõe o reconhecimento da identidade vocal tanto para ambientes limpos, quanto para telefônicos, utilizando locutores das bases de vozes TIMIT e NTIMIT, respectivamente. Para a extração das características, fez-se uso dos coeficientes mel cepstrais (MFCC), estes obtidos de duas formas distintas: pelo banco de filtros ou coeficientes por predição linear. No pré-processamento, sugeriu-se o uso de um detector de atividade de voz que utiliza um estimador da relação sinal-ruído baseado no método *mínima controlled recursive average* (MCRA), o qual se baseia na observação da razão entre a energia local de determinado quadro, pela energia mínima do ruído de fundo presente numa sequencia de quadros que precedem o quadro atual. Utilizou-se o modelo de misturas gaussianas (GMM) como gerador e classificador de padrões. Foram utilizados 472 locutores, apresentando proporção aproximada de 70% para o sexo masculino e

30% para o sexo feminino. Para o treinamento de cada locutor, foram coletadas 8 elocuções, com aproximadamente 24 segundos de duração. Foram realizados testes curtos[17], com apenas 2 elocuções. O processo apresentou ótimo desempenho em ambiente limpo (TIMIT), tendo sido alcançada uma taxa de identificação de 97% a partir do uso de banco de filtros, porém péssimo resultado em ambiente telefônico, com 13% e 17%, a partir do uso de banco de filtros e de predição linear, respectivamente.

Skosan e Mashao (2004) propuseram a repetição do experimento de Reynolds (1995), cujo objetivo foi a identificação de locutor em ambientes telefônicos utilizando um detector de atividade vocal no pré-processamento e coeficientes de frequência mel-cepstrais (MFCC) para a extração de características, na faixa de amostragem telefônica e o modelo de misturas gaussianas (GMM) para a criação e a classificação dos padrões. Foram utilizados 630 locutores da base de vozes com qualidade telefônica NTIMIT, *i.e.* 8 kHz, sendo 438 do sexo masculino e 192 do sexo feminino. Para o treinamento de cada locutor, foram usadas 8 elocuções, com aproximadamente 24 segundos de duração. Foram realizados testes curtos com apenas 2 elocuções. Skosan e Mashao (2004) obtiveram uma taxa de identificação de 58,1%, próxima daquela atingida por Reynolds (1995), 60,7%.

Skosan e Mashao (2006) apresentaram uma técnica a identificação de locutor em ambientes telefônicos combinando o uso de dois classificadores, estes baseados nos coeficientes mel-cepstrais (MFCC) e no algoritmo paramétrico de conjunto de características (PFS). Foram utilizados 630 locutores da base de vozes com qualidade telefônica NTIMIT, com qualidade de 8 kHz, sendo 438 do sexo masculino e 192 do sexo feminino. Para o treinamento de cada locutor, foram usadas 8 elocuções, com aproximadamente 24 segundos de duração no total. Foram realizados testes longos[18] com apenas 2 elocuções, chegando a taxa de identificação de locutor de 79,0%.

Beritelli e Spadaccini (2012) recomendaram que, em sistemas de reconhecimento de locutor voltados para aplicações forenses, sejam utilizados métodos

[17] **Testes curtos**: não há a concatenação das elocuções destinadas ao reconhecimento, sendo realizado um teste por elocução.

[18] **Testes longos**: quando há a concatenação das elocuções destinadas ao reconhecimento, aumentando a elocução, porém reduzindo o número de testes.

de extração derivados de banco de filtros, coeficientes cepestrais ou por predição linear. No pré-processamento, sugeriram a utilização da normalização das amplitudes do sinal, assim como o uso de detector de atividade vocal.

Meuwly e Veldhuis (2012) afirmaram que no contexto forense, a necessidade de rapidez no resultado não é de alta prioridade, corroborando, assim a utilização de sistema semiautomáticos de identificação de locutor. Por fim, os autores alegaram que, no âmbito em questão, ainda há a necessidade de melhorias na escalabilidade, no ajuste da sensibilidade ao ruído, faixa de frequência de canal restrita e na variação da idade dos locutores.

No Quadro 3.2 são apresentados os experimentos com maior similaridade com a pesquisa proposta.

Quadro 3.2 - Síntese dos experimentos com maior similaridade com a pesquisa proposta.

Autor	Abordagem	Sinais de Voz	Resultados
Skosan e Mashao (2004)	MFCC (filtros faixa telefônica) para extração e GMM para classificação	NTIMIT 8kHz, 16 bits/amostra, 630 locutores, 10 elocuções/locutor, testes curtos	Taxa de Identificação de 58,1%
Reynolds (1995)	MFCC (filtros faixa telefônica) para extração e GMM para classificação	NTIMIT 8kHz, 16 bits/amostra, 630 locutores, 10 elocuções/locutor, testes curtos	Taxa de Identificação de 60,8%
Skosan e Mashao (2006)	Classificadores baseados em MFCC e PFS	NTIMIT 8kHz, 16 bits/amostra, 630 locutores, 10 elocuções/locutor, testes longos	Taxa de Identificação de 79%

No Quadro 3.3 é apresentada uma síntese dos demais estudos contidos nesta seção.

Quadro 3.3 - Síntese das demais pesquisas relacionadas

Autor	Abordagem	Sinais de Voz	Resultados
Hanilçi *e* Ertas (2011)	MFCC para extração e QV e GMM para classificação	8 kHz, 168 locutores, 8 elocuções/locutor de 24s.	68,15% VQ e 73,21% GMM
Chen, Hsieh e Hsu (2008).	LPCC com uso de sub-bandas para extração e QV para classificação,	8 kHz, 16 bits/amostra, 51 locutores do sexo masculino, 10 elocuções/locutor de 30s.	95,36% limpo, 93,02% 20 dB SNR, 86,38% 15 dB SNR, 75,08% 10 dB SNR e 53,49% 5dB SNR.
Aldhaheri e Al-Saadi (2003)	LPC e LPCC para extração e SVD para classificação	8 kHz, 8 bits/amostra, 10 locutores 80 elocuções/locutor de 3s.	99,5% limpo, 95,5% 20 dB SNR, 93,5% 15 dB SNR, 88,5% 10 dB SNR e 82,5% 5 dB SNR.
Chetouani *et al.* (2009)	MFCC e LPCC para extração e matriz de covariância para classificação	8 kHz, 630 locutores, 8 elocuções/locutor de 24s.	27,3% MFCC e 24,6% LPCC
Hanilçi e Ertas (2009)	MFCC para extração e QV e PCA para classificação	8 kHz, 168 locutores, 8 elocuções/locutor de 24s.	100% limpo e 78,2% em telefone
Ming *et al.* (2007)	MFCC com uso de sub-bandas para extração e GMM para classificação	16 kHz, 630 locutores 8 elocuções/locutor de 3s.	96,51% limpo

Autor	Abordagem	Sinais de Voz	Resultados
Cetnarowicz, Drgas e Dabrowski (2010)	MFCC para extração e GMM para classificação	45 locutores. 114 elocuções/locutor de 2s	87% limpo e 73,4% em telefone
Nakasone e Steven (2001)	GMM para classificação	7,5 kHz, 16 bits/amostra, 50 locutores, 10 elocuções/locutor de 57s	100% limpo e 58,3% em telefone
Cardoso (2009)	MFCC (filtros ou predição) para extração e GMM para classificação	TIMIT 16kHz e NTIMIT 8kHz, 16 bits/amostra, 472 locutores, 10 elocuções/locutor, testes curtos	97% para TIMIT e 13% e 17% para NTIMIT

Ficou claro, de acordo com as pesquisas supramencionadas, a discrepância nos resultados dos experimentos em ambiente telefônico ruidoso e sem ruído, mostrando que a degradação das taxas de identificação é proporcional ao aumento do ruído. Dessa forma, faz-se necessário a utilização de técnicas de pré-processamento que minimizem esses ruídos, reduzindo os seus efeitos negativos no resultado final.

Nas pesquisas de Aldhaheri e Al-Saadi (2003) e Chetouani *et al.* (2009) percebeu-se a melhor adaptação da técnica de extração por meios dos coeficientes de frequência mel-cepstrais (MFCC) em relação aos métodos de coeficientes de predição linear (LPC e LPCC), no que diz respeito ao ambiente telefônico, possuindo banda de frequência reduzida e presença de ruído.

Os experimentos realizados por Hanilçi e Ertas (2009, 2011), todos submetidos a condições telefônicas e com presença de ruído, mostram a superioridade dos modelos de misturas gaussianas comparado aos modelos de quantização vetorial (VQ) e de análise do componente principal (PCA).

Portanto, na abordagem proposta, visa-se reconhecer a identidade vocal de locutores, independente do texto, fazendo-se uso de um pré-processamento do sinal de voz, que visa a minimização dos efeitos dos ruídos; de coeficientes de frequências mel-cepstrais (MFCC) como técnica de extração de características e; de modelos de misturas gaussianas (GMM) para a criação e a classificação dos padrões. As elocuções de treinamento e teste são oriundas da base de dados de vozes NTIMIT, com qualidade telefônica.

3.5 Considerações Gerais

Neste capítulo, foi apresentado o papel da Fonética Forense na Criminalística, suas ramificações e onde o reconhecimento de locutor está inserido neste meio, assim como requisitos e metodologias para aplicações com este fim, sendo possível perceber a importância da Fonética Forense para a Criminalística e como o reconhecimento de locutor pode ser utilizada como meio probante.

Por fim, foram apresentadas diversas pesquisas destinadas ao reconhecimento de locutor e aplicadas à Criminalística e/ou a ambientes telefônicos e com presença de ruídos, possibilitando avaliar o estado da arte, no que diz respeito as técnicas de minimização de ruídos, extração de características e classificação de padrões aplicadas ao reconhecimento de identidade vocal, com enfoque em ambientes telefônicos ruidosos e voltados à Fonética Forense.

No próximo capítulo será descrito o conjunto de técnicas e parâmetros adotados para a criação do Sistema Semiautomático de Identificação Vocal Forense, iniciando pelo pré-processamento, seguido em sequência pela extração de características e modelagem dos padrões dos locutores, e por fim, a tomada de decisão.

4 SISTEMA SEMIAUTOMÁTICO DE IDENTIFICAÇÃO VOCAL FORENSE

Para o reconhecimento de locutor, voltado à área forense, fica clara a necessidade de que seja um identificador, pois se visa buscar o originador da elocução questionada dentre um universo de suspeitos (DRYGAJLO, 2007; SVIRAVA, 2009). Além disto, segundo Kinnunen e Haizhou (2010), como os participantes são não colaborativos, ou seja, não têm intenção alguma de cooperar, faz-se necessário o uso de um sistema independente de texto, em que as elocuções de teste e de treinamento não precisam corresponder ao mesmo texto.

Cetnarowicz, Drgas e Dabrowski (2010) afirmaram que, no reconhecimento de voz, as características extraídas podem ser de baixo nível[19] ou alto nível[20]. Por se tratar de uma identificação independente de texto, é recomendado pelos autores o uso de características de baixo nível.

Os denominados “sistemas de reconhecimento semiautomático de locutor”, voltados especificamente para a comparação forense de locutor não apresentam um resultado binário, como acontece nos sistemas automáticos de controle de acesso (positivo ou negativo), mas sim um direcionamento ao Perito Criminal, por meio da razão de verossimilhança entre as probabilidades elocução tenha sido originada pelo suspeito, dada a existência da evidência ($P(H_0|E)$) e que a elocução não tenha sido originada pelo suspeito, dado a existência da evidência ($P(H_1|E)$) (VALENTE, 2012).

De acordo com o cenário proposto na pesquisa ora descrita, se faz necessário que o reconhecimento de locutor seja semiautomático e pertencente à subárea de identificação, que seja independente de locutor e, por essa particularidade, trabalhe com características de baixo nível.

[19] **baixo nível**: características derivadas dos aspectos acústicos da estrutura anatômica do trato vocal e nasal (*e.g.* frequência, magnitude ou potência espectral).
[20] **alto nível**: características influenciadas pela personalidade, condição socioeconômica e região (*e.g.* prosódia, semântica, sotaque, dicção, idiossincrasias)

O sistema pode se subdividido em 4 etapas principais: Pré-processamento do Sinal de Voz, Extração das Características, Criação ou Classificação dos Padrões e Tomada de Decisão. Estas etapas são descritas detalhadamente em seguida, estando presentes nas fases de treinamento e reconhecimento de locutor, representadas respectivamente por meio das Figura 4.1 e Figura 4.2.

Figura 4.1 - Etapas da fase de treinamento de locutor.

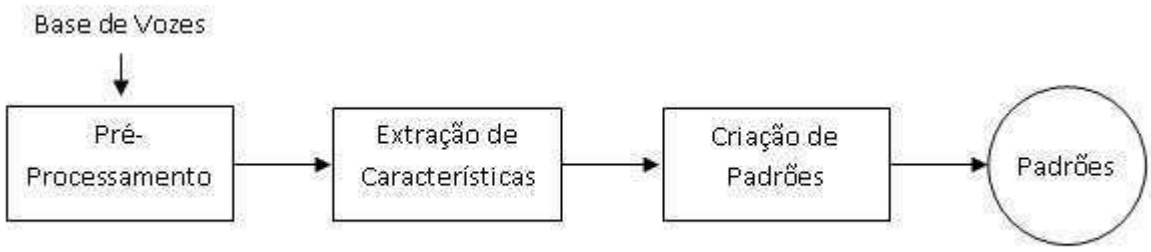
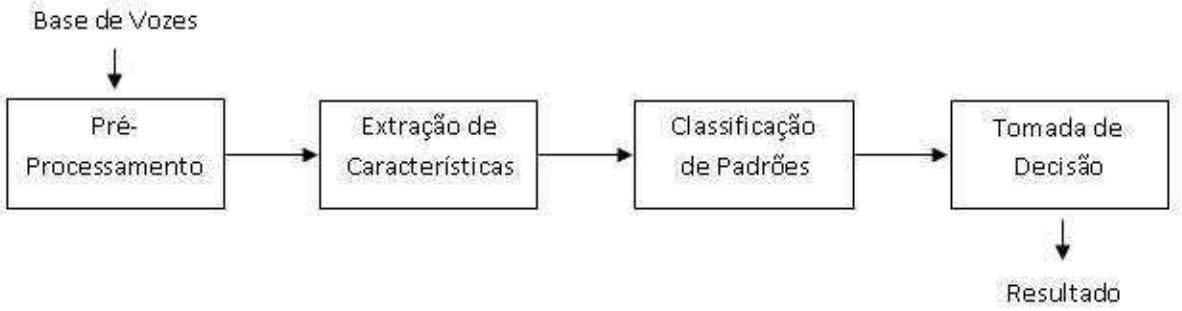


Figura 4.2 - Etapas da fase de reconhecimento de locutor.

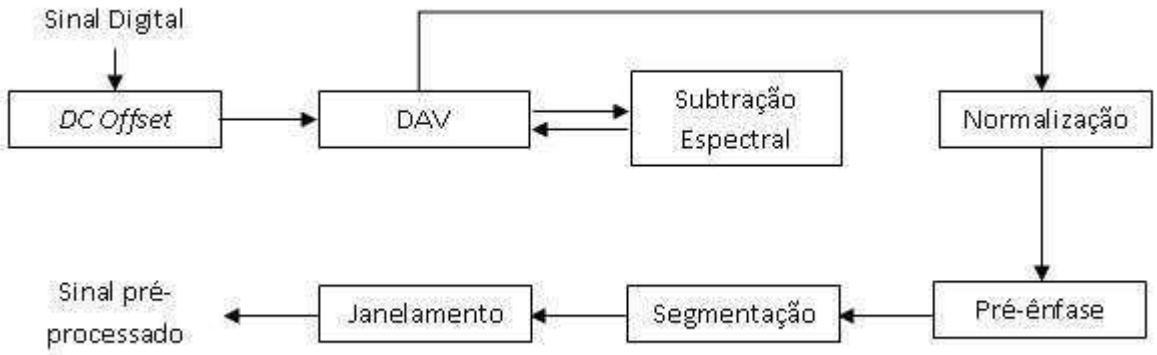


4.1 Pré-processamento do Sinal de Voz

Nesta pesquisa, propõe-se que, antes da extração das características de cada elocução, seja de teste ou treinamento, haja o pré-processamento do sinal de voz, preparando-o de forma adequada para o procedimento subsequente e disponibilizando, assim, um conjunto de dados com informação útil e com o mínimo ruído.

O pré-processamento é subdividido em diversas etapas, estas sequenciadas e discriminadas em seguida e representadas pelo diagrama em blocos da Figura 4.3.

Figura 4.3 - Diagrama do pré-processamento do sinal digital.



1. Remoção do Deslocamento pela Corrente Contínua (DC *Offset*)

Ao receber o sinal digitalizado, efetua-se a remoção da corrente contínua acrescentada por meio de dispositivo de *hardware* à gravação do sinal de áudio, aplicando-se a subtração da média das *p* amplitudes para cada elocução, conforme Equação 2.1.

2. Detecção da atividade de Voz:

Sanado o deslocamento pela corrente contínua no sinal, é aplicada a detecção da atividade de voz, utilizando-se o modelo estatístico proposto por Sohn, Kim e Sung (1999), descrito na seção 2.2.1.2, levando em consideração a potência do sinal.

3. Subtração Espectral

Ainda sem a remoção dos blocos silenciosos/ruidosos, é aplicada a subtração espectral do sinal, conforme descrito na seção 2.2.1.3, que objetiva amenizar os ruídos das elocuções. Este procedimento também foi aplicado sobre a potência do sinal.

4. Remoção da Inatividade de Voz

O sinal, após submetido à subtração espectral, teve seus blocos silenciosos/ruidosos removidos, de acordo com detecção realizada no item 2.

5. Normalização

Em seguida, foi realizada a normalização do sinal de voz, que tem por objetivo equalizar a magnitude das elocuções, minimizando os efeitos da variabilidade de canal, dividindo cada amplitude pelo valor máximo do sinal, conforme indicado na Equação 2.19.

6. Pré-ênfase

Posteriormente, foi aplicado o filtro de pré-ênfase ao sinal de voz. Este é representado matematicamente pela Equação 2.21 e nesta pesquisa foi utilizada o coeficiente $a = 0{,}97$.

7. Segmentação

A segmentação foi realizada em blocos de 320 amostras e sobreposição de 160 amostras. Como se trata de elocuções gravadas em 16 kHz, estas correspondem à 20 ms e 10 ms, respectivamente. Portanto, foram obedecidas as proposições estabelecidas em Campbell Jr (1997), que afirma que os parâmetros do sinal de voz podem ser considerados invariantes no tempo para curtos intervalos de tempo da ordem de 10 a 30 ms.

8. Janelamento

Para cada bloco resultante da segmentação, foi utilizada uma função, conforme Equação 2.25, denominada Janela de Hamming, visando a reduzir o efeito das

variações bruscas de amplitude presentes no início e no término de cada quadro (segmento).

4.2 Extração de Características

A extração de característica por meio dos coeficientes de frequência mel-cepstrais tem sido o método mais utilizado em reconhecimento automático de locutor, devido a sua eficiência computacional e robustez ao ruído, em comparação a outros métodos de extração (SKOWRONSKI e HARRIS, 2004; SAHIDULLAH e SAHA, 2012).

Sendo assim, a extração de características por meio dos coeficientes de frequência mel-cepstrais foi utilizada nesta pesquisa, pois apresenta as particularidades exigidas no panorama proposto.

Foi utilizado um banco de 24 filtros mel-cepstrais, conforme apresentado na Figura 2.14, abrangendo a faixa de frequência total do sinal, tendo-se como centro de cada filtro as seguintes frequências: 74, 156, 247, 348, 459, 582, 718, 868, 1.034, 1.218, 1.422, 1.646, 1.895, 2.171, 2.475, 2.812, 3.184, 3.596, 4.052, 4.556, 5.113, 5.730, 6.412, 7.166.

Por se tratar de um sistema de reconhecimento de locutor forense voltado para gravações em ambiente telefônico, se fez necessário o uso de uma faixa de frequência adequada para este cenário, variando em torno de 300 Hz a 3.400 Hz, visando à filtragem adequada da potência do sinal. Para tanto, foram escolhidos os filtros limítrofes, de acordo com a menor distância entre sua frequência central e a faixa de frequência utilizada, sendo estes os de centros de frequência 348 e 3.596. Portanto, foram utilizados 15 dos 24 filtros propostos inicialmente, tendo estes as frequências centrais 348, 459, 582, 718, 868, 1.034, 1.218, 1.422, 1.646, 1.895, 2.171, 2.475, 2.812, 3.184, 3.596, como apresentado na Figura 4.4.

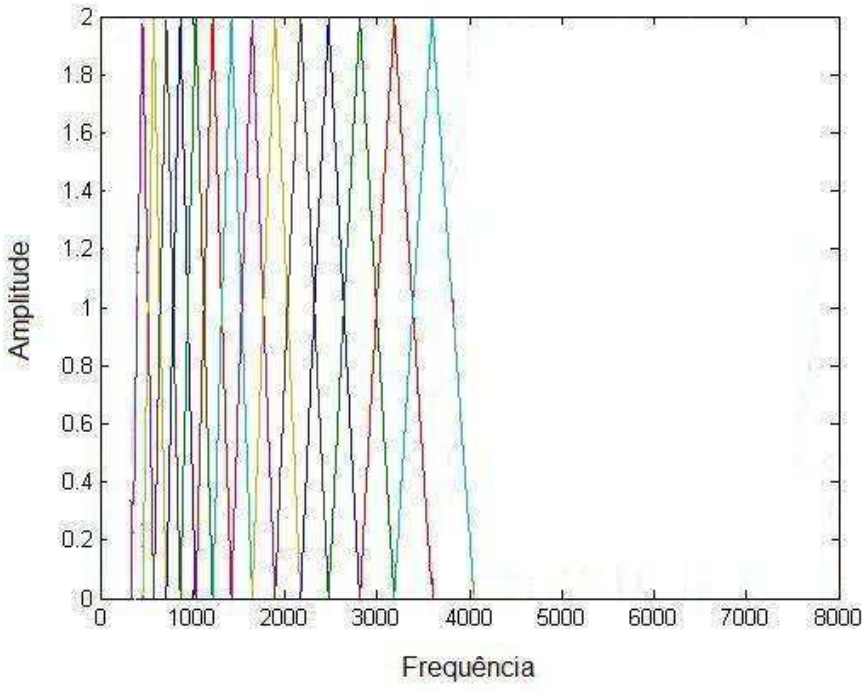
Desta forma, foram gerados 15 coeficientes mel-cepstrais dispostos como segue:

$$C_{mel} = c_0, c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}, c_{11}, c_{12}, c_{13}, c_{14}, \quad (4.1)$$

sendo o coeficiente c_0, desnecessário por possuir informações do meio de transmissão, desconsiderado, restando os demais 14 coeficientes:

$$C_{mel} = c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}, c_{11}, c_{12}, c_{13}, c_{14}, \quad (4.2)$$

Figura 4.4 - Banco de Filtros baseado na escala mel contendo 15 filtros e faixa de frequência aproximada de 300-3.400Hz.



4.3 Classificação de Padrões

Segundo Barisevičius (2008) e Campbell (2009), cada modelo se adapta melhor a um tipo de reconhecimento. No caso de reconhecimento dependente do texto, os HMM são os mais recomendados. Os modelos GMM proporcionam melhores resultados quando não existe dependência de texto.

Reynolds, Quatieri e Dunn *et. al.* (2000), afirmaram que os modelos de misturas gaussianas apresentam baixo custo de processamento e representam com fidelidade

os aspectos ligados ao locutor, estando estritamente relacionados ao trato vocal.

Diante do exposto, foi utilizada a modelagem de misturas gaussianas para criação e classificação dos modelos referentes a cada locutor. Utilizou-se a estimação de verossimilhança para treinamento, por meio do algoritmo de máxima expectativa (EM). Esta metodologia necessita de um modelo inicial, com a geração dos vetores médios a partir de algoritmo de clusterização k-*means* aplicado sobre estes.

Para a posterior estimação e maximização de seus componentes, foi utilizada como condição de parada 10 iterações ou limiar convergência menor que 0,001, como mostra a Equação 2.48.

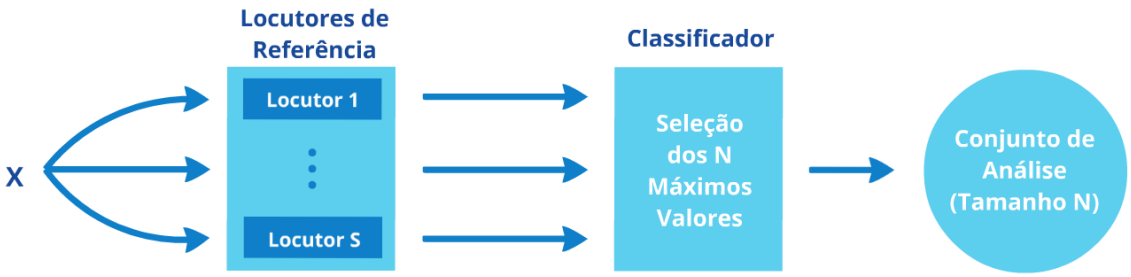
A pesquisa ora descrita utilizou 32 misturas gaussianas para cada modelo representativo de locutor, afirmado como ideal por Reynolds (1995) e repetido por Skosan e Mashao (2004; 2006).

4.4 Tomada de Decisão

Por se tratar de um sistema de reconhecimento semiautomático de identidade vocal, a partir do qual o Perito Criminal não somente tem como subsídios para sua decisão as características extraídas das elocuções de teste e treinamento, a tomada de decisão adotada é uma variação daquela usualmente utilizada, por exemplos, nos sistemas de controle de acesso.

Análogo aos Sistemas Automatizados de Reconhecimento de Impressões Digitais Criminais, que realizam um grande trabalho de aceleração da busca em grandes bases de dados e retorna uma lista curta de candidatos, por exemplo 10 candidatos dentro de banco de dados de diversos indivíduos, conforme descrito em Canedo (2010), a pesquisa em questão visou aos n mais prováveis resultados de um universo de S locutores, de acordo com uma variação do classificador simples de máxima verossimilhança descrito na Equação 2.55. Foram realizadas n iterações para cada uma destas o locutor $\hat{S}$ foi removido do universo de locutores e adicionado a um conjunto resultante, conforme ilustrado na Figura 4.5.

Figura 4.5 - Classificador do Sistema Semiautomático de Locutor



A partir de então, baseado na razão de verossimilhança, descrita anteriormente na seção 3.3.1, entre as probabilidades das hipóteses, dada a existência da evidência, de que a elocução tenha sido originada pelo suspeito e de que a elocução não pertença ao suspeito, conforme explicitado na Equação 3.3, foi obtida de cada locutor do conjunto resultante uma classificação qualitativa que o caracteriza como detentor da elocução questionada, de acordo com o Quadro 3.1.

Estas probabilidades foram geradas a partir da criação de densidades de probabilidades das características intralocutor e interlocutor. A densidade de probabilidade intralocutor, ou seja, que representa que a elocução tenha sido originada pelo suspeito é obtida pelas similaridades de uma combinação cruzada (*cross-over*) de testes realizados com as elocuções destinadas ao treinamento, como mostrado na Equação 4.3.

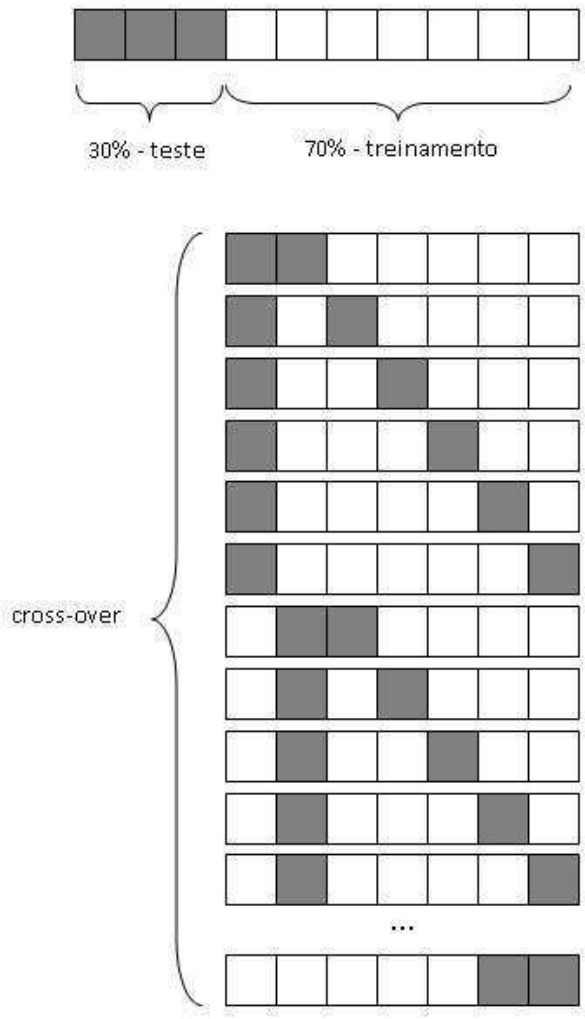
$$C_s^n = \binom{n}{s} = \frac{n!}{s!\,(n-s)!}, \tag{4.3}$$

Como exemplo, para cada locutor possuindo 10 elocuções e utilizando uma proporção 70% / 30% para treinamento/teste, coletaram-se 7 elocuções para treinamento e 3 para teste. Como as elocuções destinadas à realização dos testes são apenas suposições de que sejam ou não originadas do locutor questionado, fez-se necessário que a densidade seja criada a partir das elocuções de treinamento, que certamente foram geradas por aquele locutor. Desta forma, foi realizada a combinação

cruzada das 7 elocuções de treinamento, utilizando a mesma proporção de 70% / 30% para treinamento/teste, teríamos 4,9 elocuções para treinamento e 2,1 para testes que, no arredondamento por aproximação se obtêm 5 elocuções para treinamento e 2 para teste, conforme mostrado na Figura 4.6. Aplicando-se essa combinação, para $n = 7$ e $s = 5$, obteve-se:

$$C_5^7 = \binom{7}{5} = \frac{7!}{5!\,2!} = \frac{7x6x5!}{5!\,x2} = \frac{7x6}{2} = 21, \tag{4.4}$$

Figura 4.6 - Divisão das elocuções realização do *cross-over* de testes para criação da densidade intralocutor.



Desta forma, existem 21 diferentes combinações de testes para o cenário especificado. Portanto, o resultado subsidia a geração desta densidade de probabilidades.

Por fim, a densidade de probabilidade interlocutor, ou seja, aquela que representa que a elocução não pertença ao suspeito, foi obtida por meio do conjunto das similaridades da elocução de teste com todos os modelos na base de dados.

Portanto, o sistema semiautomatizado de reconhecimento de identidade vocal visou a reduzir o universo de locutores contidos na base de vozes criminal para um conjunto viável de ser analisado pelo Perito Criminal, além de proporcionar o direcionamento, por meio da classificação qualitativa de cada elocução do conjunto resultante.

4.5 Considerações Gerais

Neste capítulo, foram justificadas as escolhas das técnicas utilizadas no pré-processamento, na extração de características, na criação/classificação dos padrões e na tomada de decisão, assim como a descrição detalhada da metodologia, parâmetros e métricas utilizadas.

Sendo assim, foi possível compreender todo o processo de reconhecimento de identidade vocal voltado à Criminalística.

No próximo capítulo será detalhada toda a metodologia experimental realizada ao longo da pesquisa, além das particularidades do ambiente de desenvolvimento e simulação, e das características da base de vozes.

5 EXPERIMENTOS, ANÁLISES E VALIDAÇÃO DOS RESULTADOS

Neste capítulo, são reportados, analisados, discutidos e comparados os principais experimentos realizados ao longo da pesquisa.

Neste sentido, considerou-se como experimento um conjunto de testes, cujo modelo é o mesmo para todos os casos, variando apenas parâmetros utilizados, como tipo de teste, proporção treinamento/teste e quantidade de locutores.

Adicionalmente, também estão inclusas neste capítulo as particularidades do ambiente de desenvolvimento e simulação, além das características da base de vozes e metodologia aplicada.

5.1 Ambiente de Desenvolvimento e Simulação

Nesta seção, são descritos os ambientes, tanto de hardware quanto de software, de desenvolvimento e simulação do sistema de reconhecimento de identidade vocal forense.

5.1.1 Hardware

Para a realização dos experimentos foi utilizado um notebook ASUS K45VM, com barramento de 64 bits, processador INTEL CORE I7-3610QM, memória RAM principal de 8 GB e disco rígido de 1 TB.

5.1.2 Software

O desenvolvimento do Sistema Semiautomático de Reconhecimento de Identidade Vocal Forense foi realizado utilizando-se o MATLAB R2012a (7.14.0.739) 64 bits.

Além das bibliotecas padrão integradas ao *software* em questão, foi utilizada a biblioteca externa de processamento de voz VOICEBOX (BROOKS, 2006).

5.2 Base de Vozes

Em todos os experimentos, foi utilizada a base de vozes NTIMIT (JANKOWSKI *et. al* 1990). Esta base foi obtida pela reprodução de cada elocução da base de vozes TIMIT a partir de diferentes linhas telefônicas para cada sentença, ou seja, trata-se de uma versão corrompida por ruído de canal telefônico da base de vozes TIMIT. Este procedimento se deu pela instalação de uma boca e ouvido artificiais em frente a dois telefones fixos, respectivamente, o transmissor e o receptor.

Trata-se de uma base de vozes do inglês norte-americano, contendo 630 locutores, dos quais 438 do sexo masculino e 192 do sexo feminino. Existe uma subdivisão por região de origem do locutor, feita em 8 seções de tamanho heterogêneo, conforme indicado no Quadro 5.1. Cada locutor possui 10 elocuções com conteúdo distinto entre si, cada uma das quais com aproximadamente 3 segundos de duração.

Quadro 5.1 - Divisão geográfica dos locutores

Região	Quantidade de Locutores
New England	49
Northern	102
North Midland	102
South Midland	100
Southtern	98
New York City	46
Western	100
Army Brat	33

O sinal é amostrado a 16 kHz, mas sua largura de banda útil é limitada à largura telefônica (aproximadamente 300-3.400 Hz). Cada uma das amostras é representada

em 16 bits.

5.3 Metodologia Experimental

A realização dos experimentos foi dividida em dois grandes grupos: Identificação Automática de Locutor e Identificação Semiautomática de Locutor.

O primeiro grupo de experimentos visou a analisar a taxa de acerto da identificação, utilizando o classificador simples, em que é retornado apenas o locutor que apresenta a maior verossimilhança dentre um universo de S locutores. Esta estratégia proporciona mais possibilidade de validação dos resultados, comparando-os com trabalhos que utilizam esta técnica de classificação.

Este grupo de experimentos foi subdividido em grupos menores que visaram a analisar questões particulares, tais como proporção e tamanho treinamento/teste, escalabilidade, gênero e sotaque.

Em relação à proporção e tamanho treinamento/teste, o objetivo foi alcançar a melhor taxa de identificação, por meio da análise de sensibilidade do tamanho do teste, podendo ser curto ou longo, assim como da proporção dos conjuntos de treinamento e de teste.

No que diz respeito à escalabilidade, os experimentos revelaram os efeitos do aumento do número de locutores, tanto para testes curtos, quanto para testes longos.

No tocante ao gênero dos locutores, observou-se o comportamento da taxa de identificação de cada grupo, de forma isolada, assim como a influência de uma possível separação dos grupos antes da realização dos testes.

Por fim, verificou-se a sensibilidade do sistema em relação aos sotaques de diferentes regiões, analisando-se os efeitos desse fator na taxa de identificação.

Em seguida, o segundo grupo de testes foi realizado, tendo como subsídio resultados obtidos nos experimentos do primeiro grupo. Esta gama de experimentos

utiliza um classificador complexo, a partir do qual são escolhidos os n mais prováveis resultados de um universo de S locutores, de acordo com uma variação do classificador simples de máxima verossimilhança descrito na Equação 2.56. Em seguida, cada locutor do subconjunto de tamanho n, recebe uma razão de verossimilhança que classifica qualitativamente sua potencialidade.

5.4 Resultados e Análise Estatística

O resultado e a análise estatística dos experimentos realizados nos grupos de Identificação Automática de Locutor e Identificação Semiautomática de Locutor estão discriminados a seguir.

5.4.1 Grupo Experimental 1: Identificação Automática de Locutor

O grupo experimental em questão analisou a taxa de acerto da identificação automática de locutor, sendo dividido em subgrupos menores que analisaram questões específicas, tais como proporção e tamanho treinamento/teste, escalabilidade, gênero e sotaque.

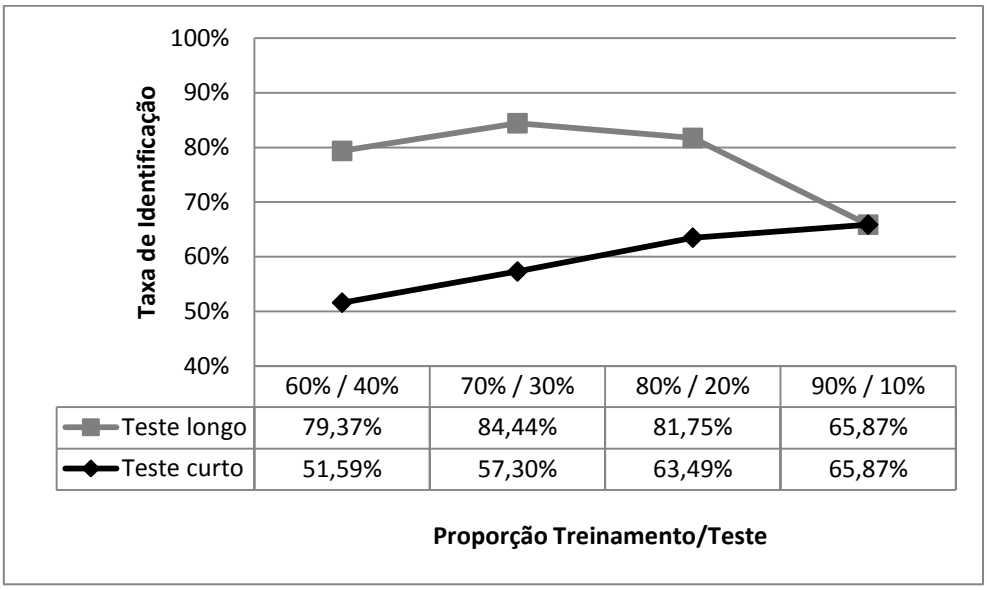
- **Divisão e tamanho treinamento/teste**

Neste subgrupo experimental, utilizaram-se todas as elocuções dos 630 locutores da base de vozes NTIMIT. Parte das 10 elocuções de cada locutor foi designada para o treinamento, sendo concatenadas umas às outras, de modo a formarem, ao final, uma única elocução. As elocuções residuais deste conjunto de 10 foram utilizadas para a realização dos testes. As elocuções destinadas aos testes, quando concatenadas, gerando apenas uma elocução, são denominadas de testes longos; quando as elocuções são testadas individualmente, são denominadas de testes curtos.

Portanto, teve-se como objetivo, por meio de análise de sensibilidade, observar a taxa de acerto de acordo com a variação da proporção utilizada para o treinamento e o teste do locutor, além de fazer uso de testes curtos ou longos.

Para tanto, foram utilizadas as divisões treinamento/teste de 60% / 40%, 70% / 30%, 80% / 20% e 90% / 10%, tanto para testes curtos, quanto para testes longos. As taxas de acerto são sumarizadas na Figura 5.1.

Figura 5.1 - Taxa de acerto variando proporção e tamanho treinamento/teste.

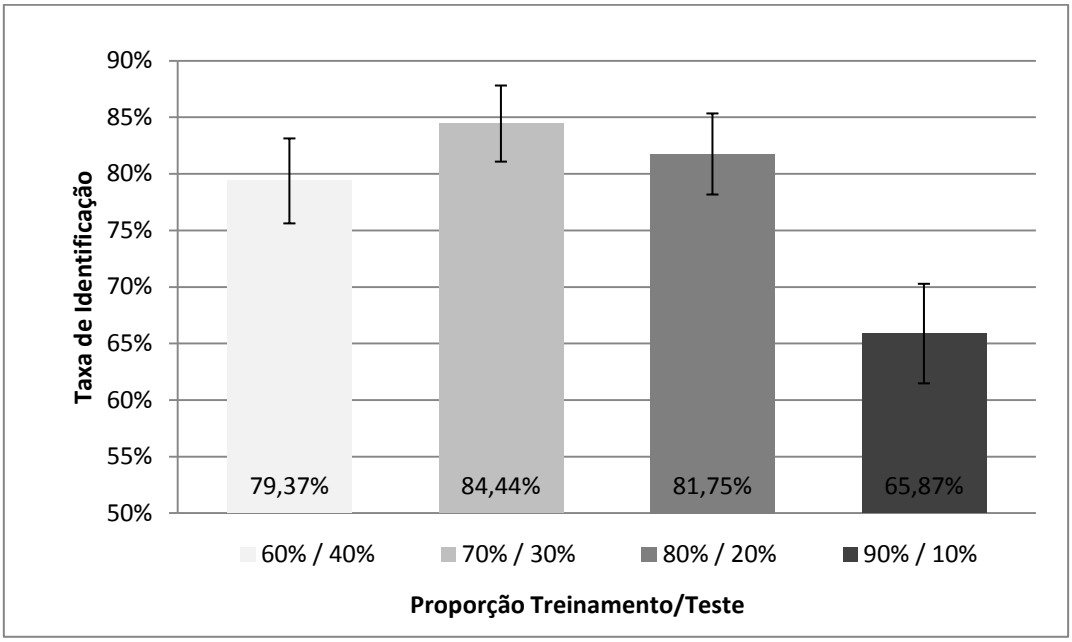


De posse da taxa de identificação e da quantidade de testes realizados, utilizando-se um nível de confiança $\alpha = 98\%$, foi possível definir o intervalo de confiança de proporções dos experimentos realizados, conforme sumarizados na Tabela 5.1 e na Figura 5.2 para testes longos, e na Tabela 5.2 e na Figura 5.3 para testes curtos.

Tabela 5.1 - Taxa de identificação e intervalo de confiança (α=98%) variando a divisão treinamento/teste longo.

Divisão Treinamento/Teste longo	**Taxa de Identificação ± Desvio**	**Número de Locutores**	**Quantidade de Testes**
60% /40%	79,37% ± 3,76%	630	630
70% / 30%	84,44% ± 3,36%	630	630
80% / 20%	81,75% ± 3,59%	630	630
90% / 10%	65,87% ± 4,40%	630	630

Figura 5.2 - Taxa de identificação e intervalo de confiança (α=98%) variando a divisão treinamento/teste longo.

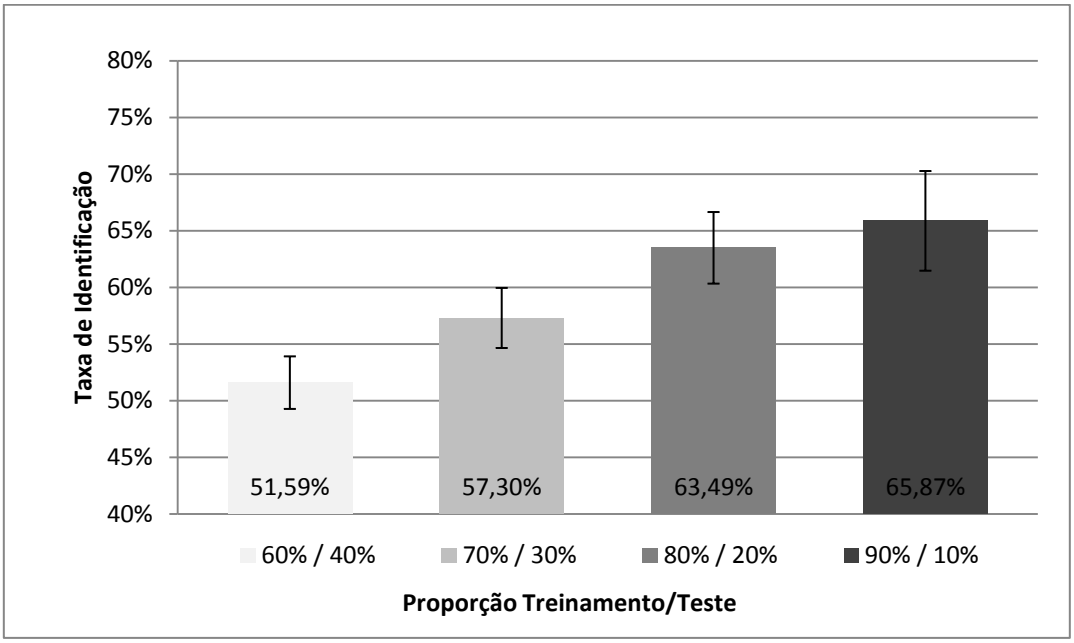


Para os testes longos, concluiu-se que as divisões 60% / 40% e 90% / 10% possuem desempenho abaixo das demais proporções e que a divisão 80% / 20% possui desempenho equivalente à divisão 70% / 30%, sendo estas duas as melhores opções a serem adotadas.

Tabela 5.2 - Taxa de identificação e intervalo de confiança (α=98%) variando proporção treinamento/teste curto.

Divisão Treinamento/Teste curto	**Taxa de Identificação ± Desvio**	**Número de Locutores**	**Quantidade de Testes**
60% /40%	51,59% ± 2,32%	630	2520
70% / 30%	57,30% ± 2,66%	630	1890
80% / 20%	63,49% ± 3,16%	630	1260
90% / 10%	65,87% ± 4,40%	630	630

Figura 5.3 - Taxa de identificação e intervalo de confiança (α=98%) variando proporção treinamento/teste curto.



Nos testes curtos, concluiu-se que as divisões 90% / 10% e 80% / 20% possuem desempenhos equivalentes entre si, sendo estas duas as melhores opções a serem adotadas.

Por fim, após uma análise estatística (Apêndice B) dos dados em questão, que constatou as melhores divisões treinamento/teste longos e curtos, pôde-se concluir que a utilização de testes longos forneceu intervalos de confiança com taxas de identificação bastante superiores em relação aos resultantes dos testes curtos, mesmo que tenha havido a redução de número de testes, que implica em uma menor precisão na geração do intervalo de confiança. Em suma, é vantajosa a utilização de testes longos em contrapartida da redução do número de testes aplicados.

- **Escalabilidade**

Visando a averiguar a escalabilidade da identificação de locutor, foram utilizadas diferentes quantidades totais de locutores da base de vozes NTIMIT. Assim, similarmente às pesquisas de Campbell (1995) e Skosan (2004), foram utilizados 7

subconjuntos, contendo 30, 100, 200, 300, 400, 500, 600 e 630 locutores escolhidos aleatoriamente, porém obedecendo a proporção original de gênero.

Essa análise de sensibilidade à quantidade total de locutores foi realizada tanto para os testes curtos quanto para os longos, utilizando as divisões 70% / 30% para testes longos e 90% / 10% para testes curtos, fazendo uso de um nível de confiança $\alpha = 98\%$.

Sendo assim, foram obtidos os seguintes resultados, com as divisões treinamento/teste supramencionadas, expostos na Tabela 5.3.

Tabela 5.3 - Taxas de identificação para testes curtos e longos variando o número de locutores treinados.

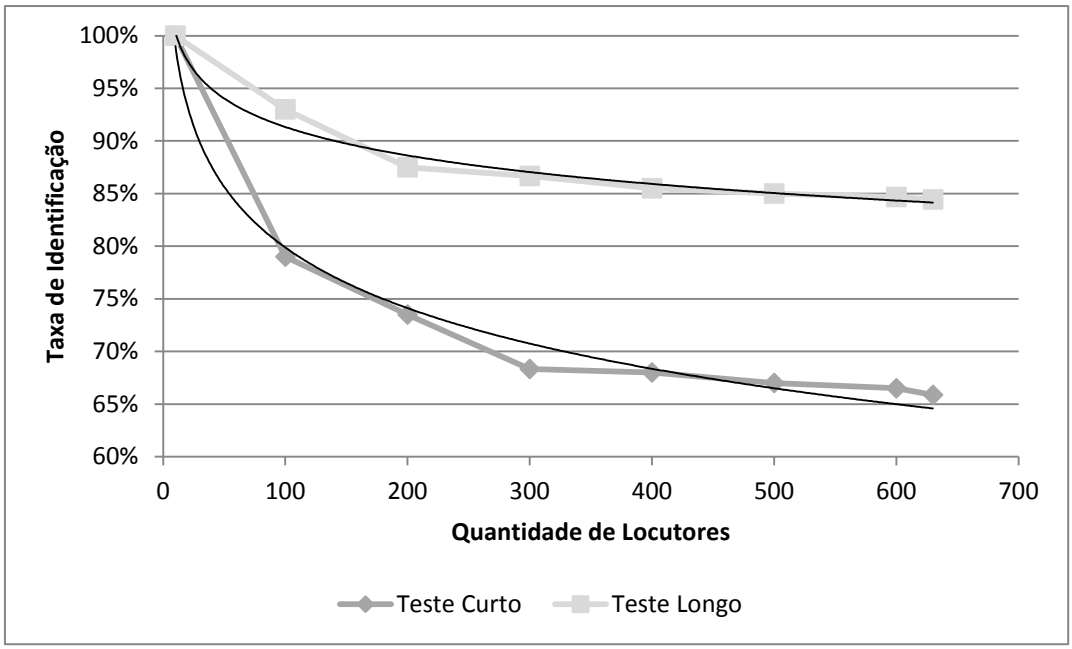
Número de Locutores	Teste Curto	Teste Longo
10	100,00% ± 0,00%	100,00% ± 0,00%
100	79,00% ± 9,49%	93,00% ± 5,94%
200	73,50% ± 7,42%	87,50% ± 5,45%
300	68,33% ± 6,26%	86,67% ± 4,57%
400	68,00% ± 5,43%	85,50% ± 4,10%
500	67,00% ± 4,90%	85,00% ± 3,72%
600	66,50% ± 4,49%	84,67% ± 3,43%
630	65,87% ± 4,40%	84,44% ± 3,36%

De acordo com os resultados obtidos, é possível concluir que se trata de um sistema não-escalável, pois a taxa de identificação não se mantém com o aumento da quantidade de locutores totais utilizados, de modo que se tratam de duas grandezas inversamente proporcionais entre si.

Baseado no gráfico da Figura 5.4, é possível visualizar que os pontos que representam as taxas de identificação nos testes longos apresentam uma estabilização próxima de linear a partir da utilização de 200 locutores, o que nos testes curtos acontece apenas a partir do uso de 300. Desta forma, percebeu-se que os testes

longos apresentaram precocidade na imposição aos efeitos do aumento do número de locutores em relação aos testes curtos.

Figura 5.4 - Gráfico de dispersão e funções de regressão das taxas de identificação dos testes curtos e longos variando o número total de locutores.



Por fim, conclui-se que a função logarítmica natural regressiva dos testes longos apresenta menor decaimento comparada à dos testes curtos.

Sendo assim, foi possível concluir que os resultados obtidos nos experimentos realizados com testes longos, em comparação aos testes curtos, receberam menor influência negativa do aumento da quantidade de locutores.

- **Gênero**

Inicialmente, nesta etapa experimental, a base de vozes NTIMIT foi dividida em dois subgrupos, de acordo com o gênero, estes contendo 438 locutores do sexo masculino e 192 do sexo feminino, fazendo uso de um nível de confiança $\alpha = 98\%$..

De acordo com os experimentos realizados nos dois subtópicos anteriores, concluiu-se que a taxa de identificação apresentava melhor rendimento quando

utilizada a divisão treinamento/teste longo de 70% / 30%. Portanto, esta foi a configuração adotada para a realização dos experimentos em questão.

Dessa forma, foram obtidos os resultados, mostrados na Tabela 5.4.

Tabela 5.4 - Taxas de identificação para subdivisão por gênero.

Gênero	Taxa de Identificação	Número de Locutores	Quantidade de Testes
Masculino	84,93% ± 3,98%	438	438
Feminino	83,33% ± 6,34%	192	192

Obtendo-se, então, uma média ponderada oriunda dos testes dos dois subgrupos isoladamente, chegou-se à taxa de identificação de 84,49%:

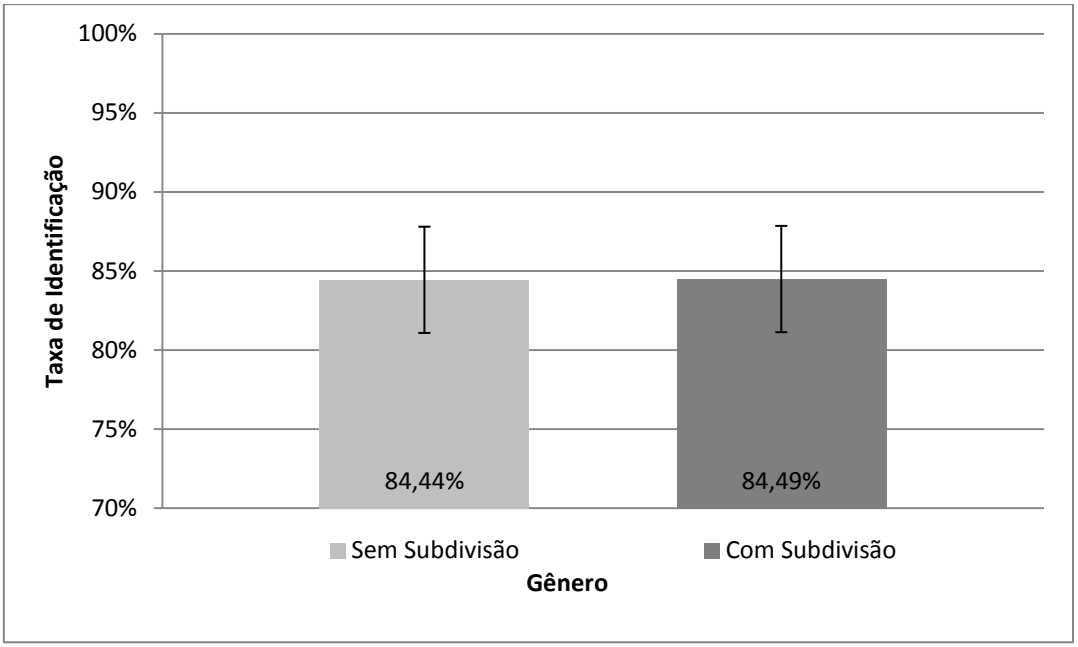
De posse da taxa de identificação, aliada a um nível de confiança de $\alpha = 98\%$ e o número de testes realizados $n = 630$, foi obtido o seu intervalo de confiança. Dessa forma foi possível confrontar com a taxa de identificação e intervalo de confiança da mesma configuração de treinamento/teste, porém sem a subdivisão por gênero, conforme Tabela 5.5 e Figura 5.5.

Tabela 5.5 - Taxa de identificação e intervalo de confiança (α=98%) sem e com subdivisão por gênero.

Divisão Treinamento/Teste longo	Taxa de Identificação ± Desvio	Número de Locutores	Quantidade de Testes
Sem subdivisão por gênero	84,44% ± 3,36%	630	630
Com subdivisão por gênero	84,49% ± 3,36%	630	630

Desta forma, conclui-se que as configurações ora propostas possuem desempenho equivalente, tornando-se desnecessária uma divisão prévia de gênero.

Figura 5.5 - Taxa de identificação e intervalo de confiança (α=98%) sem e com subdivisão por gênero.



Em um segundo estágio, visando analisar o desempenho separadamente de cada gênero, mais um subgrupo de testes foi criado, composto por 192 locutores do gênero masculino, escolhidos de forma aleatória, para que se pudesse comparar em igualdade numérica com o subgrupo de 192 locutores do gênero feminino.

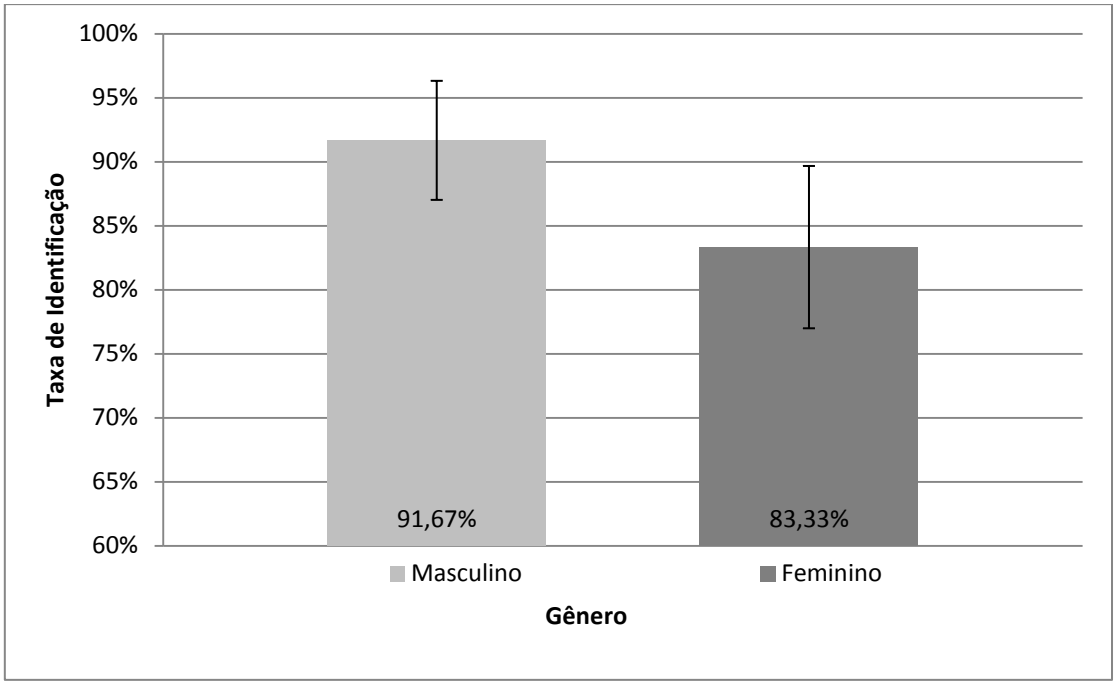
De posse da taxa de identificação e quantidade de testes realizados ($n = 192$), utilizando um nível de confiança $\alpha = 98\%$, foi possível chegar ao intervalo de confiança de proporções dos experimentos realizados, como mostra a Tabela 5.6 e a Figura 5.6.

Tabela 5.6 - Taxa de identificação e intervalo de confiança (α=98%) sem e com subdivisão por gênero.

Gênero	Taxa de Identificação ± Desvio	Número de Locutores	Quantidade de Testes
Masculino (n=192)	91,67% ± 4,65%	192	192
Feminino (n=192)	83,33% ± 6,34%	192	192

Portanto, concluiu-se que a taxa de identificação apresenta desempenho superior no subgrupo do gênero masculino em comparação ao feminino, por possuírem desempenho distintos, de acordo com o nível de confiança adotado.

Figura 5.6 - Taxa de identificação e intervalo de confiança (α=98%) para gêneros distintos.



- **Região**

Nesta etapa experimental do primeiro grupo de experimento, a base de vozes NTIMIT foi dividida em oito subgrupos, de acordo com a região dos locutores, como mostra a Tabela 5.1.

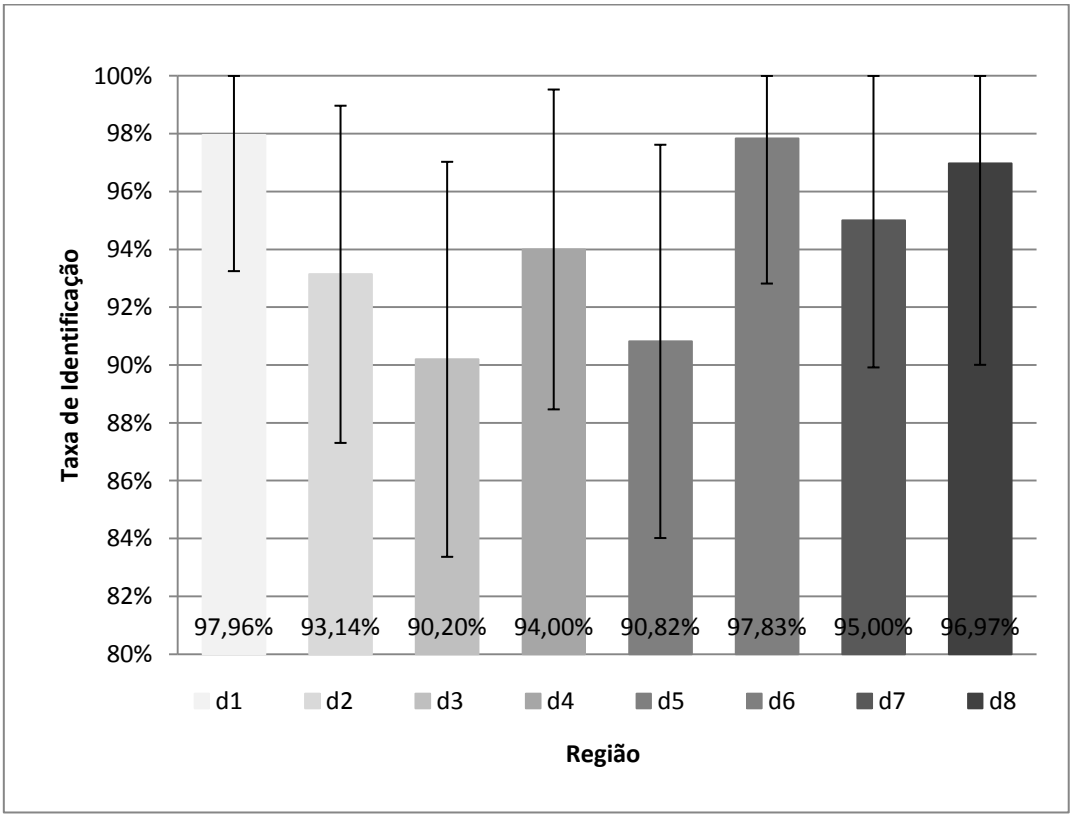
O critério para utilização da configuração treinamento/teste e o tipo de teste utilizado foi a mesma adotada no tópico referente ao gênero, sendo a divisão treinamento/teste longo de 70% / 30%.

De posse da taxa de identificação e quantidade de testes realizados de cada subgrupo, utilizando um nível de confiança $\alpha = 98\%$, foi possível chegar ao intervalo de confiança de proporções dos experimentos realizados, conforme mostrado na Tabela 5.7 e sumarizado na Figura 5.7.

Tabela 5.7 - Taxa de identificação e intervalo de confiança (α=98%) para diferentes regiões.

Legenda	Região	Taxa de Identificação ± Desvio	Número de Locutores	Quantidade de Testes
d1	***New England***	97,96% ± 4,71%	49	49
d2	***Northern***	93,14% ± 5,83%	102	102
d3	***North Midland***	90,20% ± 6,83%	102	102
d4	***South Midland***	94,00% ± 5,53%	100	100
d5	***Southtern***	90,82% ± 6,80%	98	98
d6	***New York City***	97,83% ± 5,01%	46	46
d7	***Western***	95% ± 5,08%	100	100
d8	***Army Brat***	96,97% ± 6,96%	33	33

Figura 5.7 - Taxa de identificação e intervalo de confiança (α=98%) para regiões distintas.



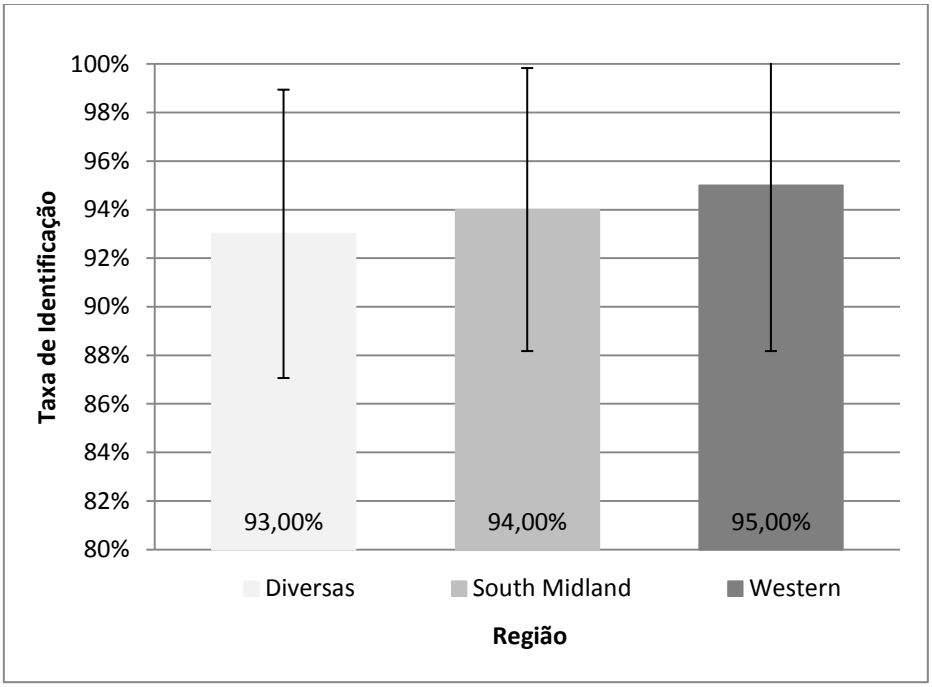
Por fim, foram selecionados 3 subgrupos contendo mesma quantidade de locutores para a comparação de desempenho entre si, visando a verificar qual o efeito da diversificação de regionalidade de locutores. Estes são compostos por um subgrupo contendo 100 locutores de diferentes regiões, escolhidos de forma aleatória, e outros 2 de mesma quantidade das regiões *South Midland* e *Western*.

De posse da taxa de identificação e da quantidade de testes realizados de cada subgrupo, utilizando-se um nível de confiança $\alpha = 98\%$, foi possível chegar ao intervalo de confiança de proporções dos experimentos, conforme sumariam a Tabela 5.7 e a Figura 5.8.

Tabela 5.8 - Taxa de identificação e intervalo de confiança (α=98%) para diferentes regiões.

Região	Taxa de Identificação ± Desvio	Número de Locutores	Quantidade de Testes
Diversas	93,00% ± 5,94%	100	100
South Midland (d4)	94,00% ± 5,53%	100	100
Western (d7)	95% ± 5,08%	100	100

Figura 5.8 - Taxa de identificação e intervalo de confiança (α=98%) comparativo entre locutores de diversas regiões e as específicas *South Midland* e *Western*.



De acordo com os resultados obtidos, pôde-se concluir que, neste caso, a diversidade da regionalidade dos locutores não influencia no desempenho do sistema.

5.4.2 Grupo Experimental 2: Identificação Semiautomática de Locutor

Este grupo experimental utilizou em sua plenitude as elocuções dos 630 locutores da base de vozes NTIMIT. Uma porção das 10 elocuções de cada locutor foi destinada ao treinamento, sendo estas concatenadas uma as outras, que por fim formaram uma única elocução. As elocuções restantes deste conjunto de 10 foram utilizadas para a realização dos testes.

De acordo com as conclusões obtidas no Grupo Experimental 1, optou-se por adotar o uso de testes longos com a divisão treinamento/teste 70% / 30% para a realização dos experimentos deste grupo.

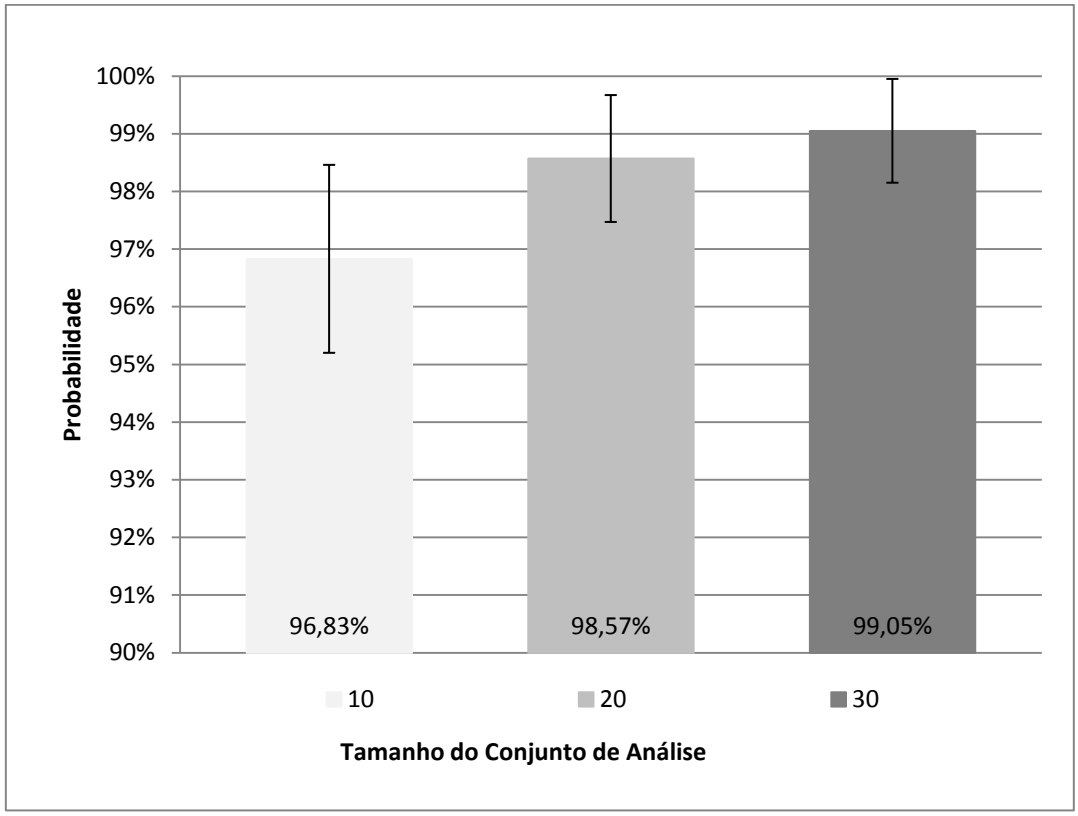
Visando a observar os resultados para um sistema semiautomático de identificação de locutores, foi utilizado o modelo de tomada de decisão que retorna os n mais prováveis resultados de um universo de S locutores, de acordo com uma variação do classificador simples de máxima verossimilhança descrito na Equação 2.57. De posse desta proporção e de quantidade de testes realizados, utilizando um nível de confiança $\alpha = 98\%$, foi possível chegar ao intervalo de confiança de proporções dos experimentos realizados.

Sendo assim, para $n = 10, 20$ e 30, tem-se as seguintes probabilidades de a verdadeira elocução pertencer ao conjunto de análise, visto na Tabela 5.9 e Figura 5.9.

Tabela 5.9 - Probabilidade e intervalo de confiança (α=98%) para verdadeira locução contida no conjunto de análise.

Tamanho do Conjunto de Análise	Probabilidade ± Desvio	Número de Locutores	Quantidade de Testes
10	96,83% ± 1,63%	630	630
20	98,57% ± 1,10%	630	630
30	99,05% ± 0,90%	630	630

Figura 5.9 - Probabilidade e intervalo de confiança (α=98%) para verdadeira locução contida em cada conjunto de análise.



Por fim, foi realizado um teste direcionado, com a elocução de teste sendo sempre a verdadeira, observando a classificação qualitativa que lhe foi atribuída, baseada na razão de verossimilhança entre as probabilidades de ser e de não ser o locutor em questão.

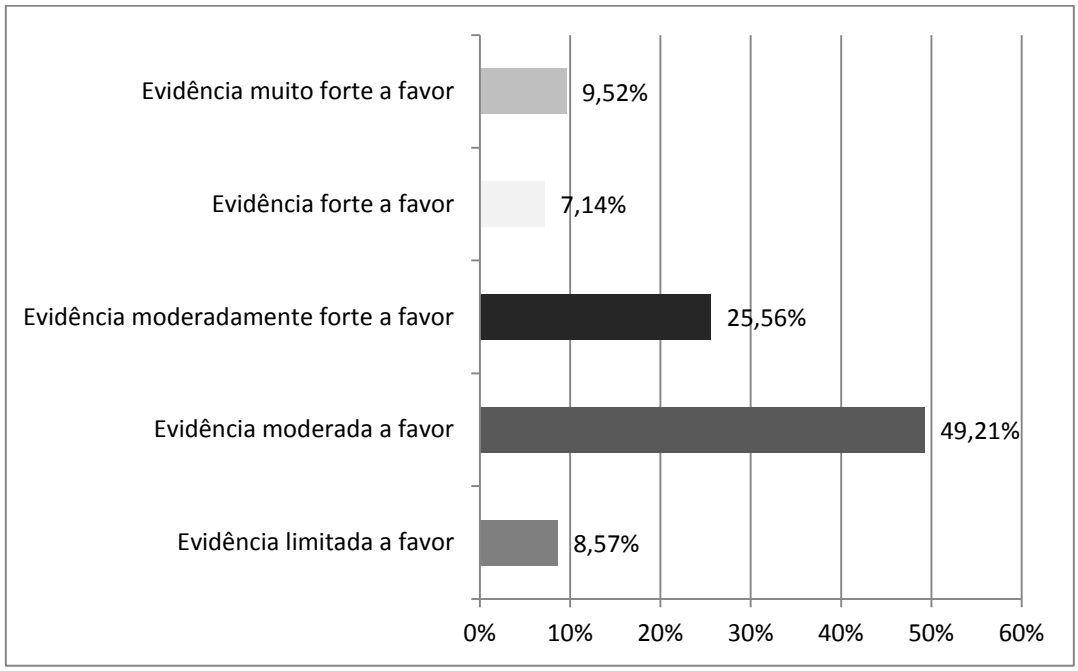
Assim, foram obtidas as seguintes proporções das classificações qualitativas destas elocuções, sumariadas na Tabela 5.10 e na Figura 5.10.

Tabela 5.10 - Proporções da classificação qualitativa das elocuções verdadeiras.

Classificação Qualitativa	**Quantidade**	**Proporção**
Evidência muito forte a favor	60	9,54%
Evidência forte a favor	45	7,14%
Evidência moderadamente forte a favor	161	25,26%

Classificação Qualitativa	Quantidade	Proporção
Evidência moderada a favor	310	49,21%
Evidência limitada a favor	54	8,57%
Evidência limitada contra	0	0%
Evidência moderada contra	0	0%
Evidência moderadamente forte contra	0	0%
Evidência forte contra	0	0%
Evidência muito forte contra	0	0%

Figura 5.10 - Proporções da classificação qualitativa das elocuções verdadeiras.



Foi possível observar que em nenhum dos casos a classificação qualitativa apresentou relevância desfavorável, sendo sempre direcionado a favor, tendo como moda a classificação *evidência moderada a favor*.

Conforme exposto, concluiu-se que a classificação qualitativa baseada na razão de verossimilhança entre as probabilidades da elocução teste ser e de não ser do locutor questionado apresentou sempre inclinação favorável em seus resultados,

apresentando proporção majoritária como sendo moderada a favor. Esta inclinação não tão incisiva pode ser dada pela pequena quantidade de combinações geradas mostradas na Equação 4.4, estes que resultaram na criação das densidades de probabilidades intralocutor e interlocutor, que, respectivamente, representam as probabilidades da elocução teste ter sido originada pelo locutor em questão e da elocução teste não ter sido originada pelo locutor em questão.

5.5 Comparação de Modelos

Para os experimentos realizados adotando-se os testes curtos, foi possível comparar os resultados obtidos nesta pesquisa com aqueles divulgados por Reynolds (1995) e Skosan e Mashao (2004), devido à similaridade das técnicas adotadas para extração de características e de criação e classificação dos modelos, além do uso da mesma quantidade de locutores da base de vozes NTIMIT. Vale ressaltar que nas pesquisas mencionadas foi utilizada a divisão de treinamento/teste de 80% / 20%

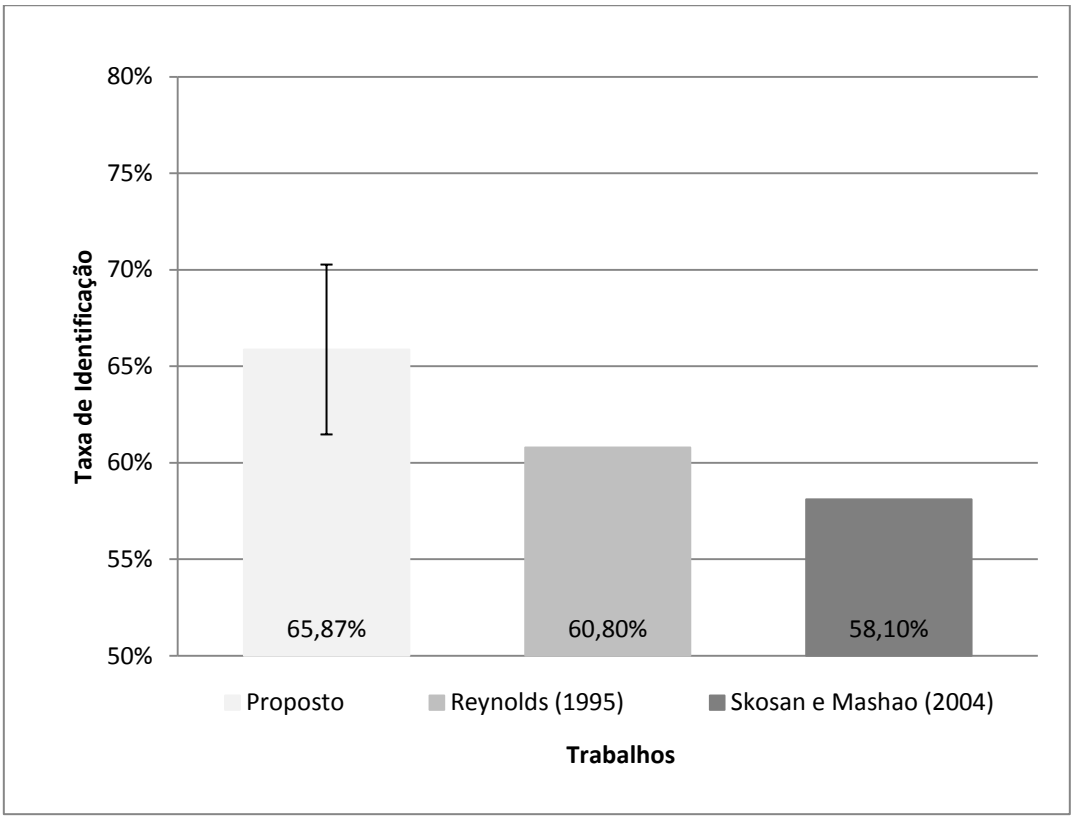
Para mensurar o intervalo de confiança da taxa de identificação do trabalho foi utilizada a divisão treinamento/teste curto de 90% / 10% e um nível de confiança de $\alpha = 98\%$. As pesquisas supracitadas não apresentam resultados de análises estatísticas, apenas a proporção de acertos na identificação de locutor, conforme apresentado na Tabela 5.11.

Tabela 5.11 - Comparação entre trabalhos com testes curtos.

Trabalhos (Teste Curto)	**Taxa de Identificação ± Desvio**	**Número de Locutores**	**Quantidade de Testes**
Proposto	65,87% ± 4,40%	630	630
Reynolds (1995)	60,8% ± 0%	630	1260
Skosan e Mashao (2004)	58,1% ± 0%	630	1260

Desta forma, devido às taxas de identificação das pesquisas supracitadas não estarem presentes no intervalo de confiança da taxa de identificação do trabalho proposto, foi possível concluir que este possui desempenho superior ao demais, tendo 98% de confiança, conforme sumariado na Figura 5.11.

Figura 5.11 - Taxa de identificação e intervalo de confiança (α=98%) comparativo entre o Trabalho proposto, Reynolds (1995) e Skosan e Mashao (2004), para testes curtos.



Para os experimentos realizados, adotando-se os testes longos, foi possível comparar os resultados obtidos nesta pesquisa com àqueles divulgados por Skosan e Mashao (2006), por ambos adotarem metodologia de concatenação das elocuções para realização dos testes, além do uso da mesma quantidade de locutores da base de vozes NTIMIT. Vale ressaltar que na pesquisa mencionada foi utilizada a divisão de treinamento/teste de 80% / 20%

Para mensurar o intervalo de confiança da taxa de identificação nesta pesquisa, foi utilizada a divisão treinamento/teste curto de 70% / 30% e um nível de confiança de $\alpha = 98\%$. Os resultados divulgados por Skosan e Mashao (2006) não incluem análise

estatística, apenas a proporção de acertos para identificação de locutor, conforme sumariada na Figura 5.12 e na Tabela 5.12.

Figura 5.12 - Taxa de identificação e intervalo de confiança (α=98%) comparativo entre o Trabalho proposto e Skosan e Mashao (2006), para testes longos.

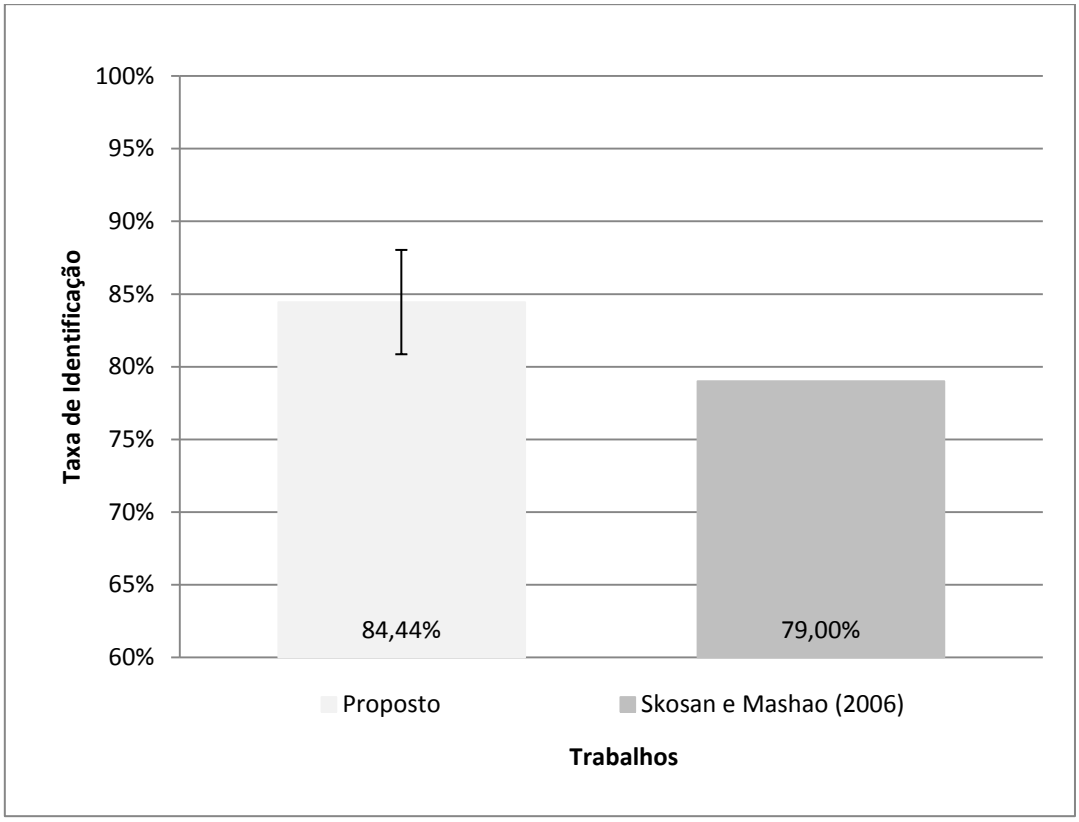


Tabela 5.12 - Taxa de identificação e intervalo de confiança (α=98%) comparativo entre o Trabalho proposto e Skosan e Mashao (2006), para testes longos.

Trabalhos (Teste Longo)	**Taxa de Identificação ± Desvio**	**Número de Locutores**	**Quantidade de Testes**
Proposto	84,44% ± 3,36%	630	630
Skosan e Mashao (2006)	79% ± 0%	630	630

Portanto, devido à taxa de identificação apresentada em Skosan e Mashao (2006) não estar presente no intervalo de confiança da taxa de identificação do trabalho

proposto, podemos concluir que este possui desempenho superior, tendo 98% de confiança.

Por fim, levando em consideração a divisão dos locutores por região, Cardoso (2009) conseguiu as seguintes taxas de identificação, utilizando locutores da base de vozes NTIMIT, utilizando a divisão de treinamento/teste de 80% / 20%, coeficientes de frequência mel-cepstrais (MFCC), para a extração das características, o modelo de misturas gaussianas (GMM) como gerador e classificador de padrões em seu estudo e o uso de testes longos, conforme sumariado mostra a Tabela 5.13.

Tabela 5.13 - Taxa de identificação para diferentes regiões (CARDOSO, 2009).

Região	**Taxa de Identificação**	**Número de Locutores**	**Quantidade de Testes**
***New England* (d1)**	50%	38	38
***Northern* (d2)**	41%	76	76
***North Midland* (d3)**	45%	76	76
***South Midland* (d4)**	62%	68	68
***Southtern* (d5)**	50%	70	70
***New York City* (d6)**	66%	35	35
***Western* (d7)**	52%	77	77
***Army Brat* (d8)**	91%	22	22

Mesmo dispondo da teórica desvantagem do número aumentado de locutores, como pode ser visualizado na Tabela 4.7, em comparação à pesquisa de Cardoso (2009), esta pesquisa apresentou melhor desempenho em todos os subgrupos das regiões, exceto *Amy Brat*, que se mostrou equivalente, com 98% de confiança nos resultados.

5.6 Considerações Gerais

No capítulo em questão, foram descritos o ambiente utilizado no desenvolvimento do sistema, a base de dados de vozes, a metodologia experimental aplicada, e por fim, a análise estatística e validação dos resultados obtidos.

Portanto, subsidiado pelos resultados, os quais foram submetidos à análise estatística, foi possível tecer conclusões, contribuições e sugestões para futuras pesquisas, estas contidas no próximo capítulo, a respeito do reconhecimento de identidade vocal, aplicados área forense e/ou em ambientes telefônicos com presença de ruídos.

6 CONSIDERAÇÕES FINAIS E SUGESTÕES PARA PESQUISAS FUTURAS

Nesta pesquisa, avaliou-se o desempenho de um sistema de identificação de locutor em ambiente telefônico com presença de ruído, nas formas automática, em que apenas as características extraídas dos sinais de voz são utilizadas, resultando em apenas um único locutor; e semiautomática, em que um conjunto de locutores mais prováveis é retornado, baseado nas características extraídas dos sinais de voz. Em seguida, uma gama de outros fatores pode ser levada em consideração para se chegar ao resultado final, com vistas à adequação às necessidades de um sistema designado à área forense, partindo-se do pressuposto de que uma única fonte de informações não deve ser considerada suficiente para uma conclusão com um fim de tamanha importância.

A escolha da técnica de extração de características baseadas nos coeficientes de frequência mel-cepstrais (MFCC), do modelo de criação dos padrões, baseada em misturas gaussianas (GMM) e da combinação de técnicas no pré-processamento do sinal, teve como objetivo minimizar os efeitos da baixa qualidade do canal telefônico, assim como dos ruídos presentes no sinal de voz.

A avaliação do sistema automático teve como objetivo verificar a acurácia da taxa de acerto na identificação de locutores em diferentes cenários, servindo como base para a escolha da melhor configuração para o sistema semiautomático. Foram analisados a melhor proporção treinamento/teste a ser adotada; o tipo de teste, curto ou longo; os efeitos do aumento da quantidade de locutores, a escalabilidade do sistema; e as particularidades em relação ao gênero e à região dos locutores.

O sistema semiautomático avaliou a probabilidade da elocução verdadeira estar presente no conjunto de análise posterior. Também foi verificada a relevância das elocuções ditas como verdadeiras, de acordo com a qualificação recebida baseada na razão de verossimilhança dentre as probabilidades da elocução questionada pertencer ao locutor e da elocução questionada não pertencer ao locutor.

6.1 Conclusões

Após a realização dos experimentos e análise estatística dos resultados obtidos, foi possível chegar a diversas conclusões a respeito do sistema automático de identificação de locutor, no tocante à proporção e tamanho treinamento/teste, escalabilidade, gênero e sotaque, assim como sobre o sistema semiautomático de identificação de locutor, como relatado em seguida:

- Para testes curtos, as proporções treinamento/teste de 80% / 20% e 90% / 10% apresentaram os melhores resultados e, para um nível de confiança de $\alpha = 98\%$, equivalência em seus desempenhos;

- Para testes longos, as proporções treinamento/teste de 70% / 30% e 80% / 20% apresentaram os melhores resultados e, para um nível de confiança de $\alpha = 98\%$, equivalência em seus desempenhos;

- Verificou-se que, mesmo havendo a redução da quantidade de testes, que diminui a acurácia do intervalo de confiança do resultado final, é extremamente vantajosa a utilização de testes longos em comparação aos testes curtos;

- Ficou evidente o decréscimo da taxa de identificação com o aumento da quantidade total de locutores, porém o uso de testes longos, em comparação aos testes curtos, é menos influenciado por esta ação;

- Uma divisão prévia por gênero não altera a taxa de identificação final, sendo o desempenho equivalente para testes com ou sem divisão;

- A identificação de locutor do gênero masculino apresenta resultado superior em comparação ao gênero feminino;

- Experimentos com conjuntos de locutores de apenas uma região, em comparação a outro experimento envolvendo um conjunto de locutores oriundos de diversas regiões, apresentaram resultados equivalentes, concluindo assim, que a identificação não sofre sensibilidade à diferença de sotaques;

- Nos testes realizados para o sistema semiautomático, utilizando o conjunto de análise de tamanho $n = 30$, a probabilidade de o locutor verdadeiro pertencer ao conjunto em questão pode chegar a 99,95%, com um nível de confiança de $\alpha = 98\%$.

- Verificou-se a necessidade de maior quantidade de elocuções por locutor para criação das densidades de probabilidade intra e interlocutor, visando o não comprometimento da relevância qualitativa dos resultados.

6.2 Contribuições

Com o término da pesquisa em questão, foi possível elencar algumas contribuições atingidas, estas discriminadas em seguida:

- A metodologia de organização das etapas de pré-processamento, visando à minimização dos efeitos negativos da presença de ruídos;

- O uso de 15 dos 24 filtros triangulares da escala mel, filtros estes apenas contidos na faixa de frequência telefônica (300-3.400 Hz);

- A disponibilização do código fonte do sistema e da base de dados de vozes, auxiliando a pesquisas similares ou até mesmo possibilitando a continuação e aprofundamento do trabalho ora relacionado;

- A possibilidade de implementação do Sistema de Reconhecimento de Identidade Vocal Forense no Instituto de Polícia Científica do Estado da Paraíba (IPC-PB).

6.3 Sugestões para Pesquisas Futuras

Com base na análise dos resultados, juntamente com as contribuições e conclusões obtidas nesta dissertação, foi possível observar a necessidade de se desenvolver novos estudos envolvendo os seguintes quesitos:

- Utilizar uma base de vozes com maior quantidade de elocuções, visando a uma maior confiabilidade dos resultados e da densidade de probabilidade para as variações intra e interlocutor;

- Buscar por novas técnicas de minimização dos efeitos de ruídos, extração de características e modelagem dos padrões que maximizem as taxas de identificação no cenário escolhido;

- Procurar por metodologias que reduzam *a priori* a quantidade de locutores que serão analisados, tendo em vista a sensibilidade do sistema em relação ao aumento da quantidade de locutores,;

- Analisar com maior ímpeto as razões de verossimilhança do conjunto de análise, e de alguma forma, procurar identificar relações entre elas, verificando a existência de alguma potencialidade de classificação, que até então não é conhecida, pois serve apenas para mostrar a relevância do resultado por meio de qualificação qualitativa;

- Avaliar o sistema proposto em um cenário real, utilizando base de dados de vozes criminais;

- Buscar metodologias de análise de sensibilidade dos parâmetros envolvidos nas etapas de pré-processamento, extração de características e criação dos modelos;

- Criar de um sistema que receba as probabilidades *a priori* de que a elocução tenha sido originada ou não pelo suspeito. Probabilidades essas fora da alçada computacional, sendo oriundas no âmbito jurídico (depoimentos, acariações vestígios, entre outros). Desta forma, seria possível integrar várias fontes para a tomada de decisão final;

REFERÊNCIAS BIBLIOGRÁFICAS

ALDHAHERI, R. W.; AL-SAADI, F. E. Text-Independent Speaker Identification in Noisy Environment Using Singular Value Decomposition. In: Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Singapore. **Anais...** Singapore, 2003. p. 1624-1628.

AL-HASSANI, M. D.; KADHIM, A. A. Design A Text-Prompt Speaker Recognition System Using LPC-Derived Features. In: The 13th International Arab Conference on Information Technology (ACIT'2012). Balamand. **Anais...** Balamand, 2012. p. 555-562.

BARBOSA JÚNIOR, A. A. **Estudo e Desenvolvimento de Aplicação Biométrica em Ambiente de Larga Escala - Reconhecimento de Impressões Digitais**. Monografia (Graduação em Ciência da Computação), Universidade Federal da Bahia. 2007.

BARISEVIčIUS, G. **Text-Independent Verification**. 2008.

BERITELLI, F.; SPADACCINI, A. Performance Evaluation of Automatic Speaker Recognition Techniques for Forensic Applications. In: YANG, J.; XIE, S. J. **New Trends and Developments in Biometrics**. InTech, 2012.

BIMBOT, F. et al. A Tutorial on Text-Independent Speaker Verification. **EURASIP Journal on Applied Signal Processing**, 1, 2004. 430-451.

BIMBOT, F.; MAGRIN-CHAGNOLLEAU, I.; MATHAN, L. Second-order statistical measures for text-independent speaker identification, 17, ago. 1995. 177-192.

BRASIL. *Constituição Federal de 1988*. Promulgada em 5 de outubro de 1988. Disponível em <http://www.planalto.gov.br/ccivil_03/constituicao/constituição.htm>.

BROOKS, M. VOICEBOX: A speech processing toolbox for MATLAB, 2006. Disponivel em: <http://www.ee.imperial.ac.uk/hp/staff/dmb/ >

BRAID, A. C. M. **Fonética Forense**. 2ª. ed. São Paulo: Millennium, 2009. 144 p.

BURILEANU, D. et al. An Adaptive and Fast Speech Detection Algorithm. In: Third International Workshop (TSD 2000). Brno. **Anais...** Brno, 2000. p. 177-182.

CAMPBELL JR, J. P. Speaker recognition: a tutorial. **Proceedings of the IEEE**, v. 85, n. 9, p. 1437-1462, set. 1997.

CAMPBELL, J. P. et al. Forensic speaker recognition. **IEEE Signal Processing Magazine**, v. 26, n. 2, p. 95-103, mar. 2009. ISSN 1053-5888.

CANEDO, J. A. Biometria na segurança pública. **Fórum Biometria**, 2010. Disponivel em: <http://www.forumbiometria.com/fundamentos-de-biometria/218-biometria-na-seguranca-publica.html>. Acesso em: 15 mai. 2013.

CARDOSO, D. P. **Identificação de locutor utilizando modelos de misturas gaussianas**. 86 f. Dissertação (Mestrado em Ciência da Computação), Universidade de São Paulo, São Paulo, 2009.

CETNAROWICZ, D.; DRGAS, S.; DABROWSKI, A. Speaker recognition system and experiments with head / torso simulator and telephone transmission. In: Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings (SPA 2010). Poznan. **Anais...** Poznan, 2010. p. 99-103.

CHE, C.; LIN, Q.; YUK, D.-S. An HMM Aproach to Text-Prompted Speaker Verification. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96). Atlanta. **Anais...** Atlanta, 1996. p. 673-676.

CHEN, W.-C.; HSIEH, C.-T.; HSU, C.-H. Robust Speaker Identification System Based on Two-Stage Vector Quantization. **Tamkang Journal of Science and Engineering**, 11, 2008. 357-366.

CHETOUANI, M. et al. Investigation on LP-residual representations for speaker identification. **Pattern Recognition**, 42, n. 3, Mar 2009. 487-494.

DAVIS, S. B.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. **IEEE Transactions on Acoustics, Speech and Signal Processing**, v. 28, n. 4, p. 357-366, ago. 1980. ISSN 0096-3518.

DRYGAJLO, A. Forensic Automatic Speaker Recognition. **IEEE SIGNAL PROCESSING MAGAZINE**, v. 24, n. 2, p. 132-135, mar. 2007. ISSN 1053-5888.

ESPINDULA, A.; TOCCHETO, D. **Criminalística, Procedimentos e Metodologia**. 2. ed. Porto Alegre: Millennium, 2005. 516 p.

FECHINE, J. M. **Reconhecimento Automático de Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística**. 212 f. Tese (Doutorado em Engenharia Elétrica), Universidade Federal de Campina Grande. Campina Grande, 2000.

FURUI, S. Speaker-independent isolated word recognition using dynamic features of speech spectrum. **IEEE Transactions on Acoustics, Speech and Signal Processing**, v. 1, n. 34, p. 52-59, fev. 1986. ISSN 0096-3518.

GABANINI, A. P. N. A Voz Humana. **Profala**, 25 set. 2003. Disponivel em: <http://www.profala.com/arttf57.htm>. Acesso em: 16 mai. 2013.

HALNIÇI, C.; ERTAS, F. Comparison of the impact of some Minkowski metrics on VQ/GMM based speaker recognition. **Computers and Electrical Engineering**, 37, 2011. 41-56.

HANILCI, C.; ERTAS, F. **Principal Component Based Classification for Text-Independent Speaker Identification**. In: Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control (ICSCCW 2009). Famagusta, **Anais...** Famagusta, 2009.

JANKOWSKI C.; A. KALYANSWAMY; BASSON S.; SPITZ J. NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP) **Anais...** Abr. 1990

JUANG, B.-H.; FURUI, S. Automatic recognition and understanding of spoken language - a first step toward natural human-machine communication. **Proceedings of the IEEE**, v. 88, n. 8, p. 1142-1165, ago. 2000. ISSN 0018-9219.

KINNUNEN, T. **Spectral Features for Automatic Text-Independent Speaker Recognition**. University of Joensuu. Joensuu, p. 144. 2003.

KINNUNEN, T.; HAIZHOU, L. An overview of text-independent speaker recognition: From features to supervectors. **Speech Communication**, v. 52, n. 4, p. 12-40, Abr 2010.

LIMA, C. B. **Sistemas de Verificação de Locutor Independente do Texto Baseado em GMM e AR-Vetorial Utilizando PCA**. 126 f. Dissertação (Mestrado em Engenharia Elétrica), Instituto Militar de Engenharia. Rio de Janeito, p. 126. 2001.

MARTINS, E. O Que é Biometria. **TECMUNDO**, 2009. Disponivel em: <http://www.tecmundo.com.br/o-que-e/3121-o-que-e-biometria-.htm>. Acesso em: 15 mai. 2013.

MATEUS, G. R. Conceitos Básicos Transmissão e Comutação. **UFMG**, 2012. Disponivel em: <homepages.dcc.ufmg.br/~mateus/compmovel/aula4.pdf>. Acesso em: 27 fev 2012.

MATLAB 8.0 and Statistics Toolbox 8.1, The MathWorks, Inc., Natick, Massachusetts, Estados Unidos

MENDOZA, L. A. F. **Redes Neurais e Máquinas de Vetores de Suporte no reconhecimento de locutor usando coeficientes MFC e características do sinal glotal**. 129 f. Dissertação (Mestrado em Engenharia de Telecomunicações), Universidade Federal Fluminense. Niterói, 2009.

MEUWLY, D.; VELDHUIS, R. Forensic Biometrics: From two communities to One Discipline. In: International Conference of the Biometrics Special Interest Group (BIOSIG). Darmstadt, **Anais...** Darmstadt, 2012. p. 1-12.

MING, J. et al. Robust Speaker Recognition in Noisy Conditions. **IEEE Transactions on Audio, Speech, and Language Processing**, v. 15, n. 5, p. 1711-1723, jul. 2007. ISSN 1558-7916.

NAKASONE, H.; STEVEN, D. B. Forensic Automatic Speaker Recognition. In: The Speaker Recognition Workshop. Creta, **Anais...** Creta, 2001. p. 18-22.

OPPENHEIM, A. V.; SCHAFER, R. W. **Discrete-Time Signal Processing**. 3ª. ed. Englewood Cliffs: Prentice Hall, 2009. 1120 p. ISBN 0131988425.

PATRA, S. **Robust Speaker Identification System**. Indian Institute Science. Bangalore, p. 118. 2007.

PENA, S. Pequena História da Individualidade Genética Humana. **Ciência Hoje**, 2006. Disponivel em: <http://cienciahoje.uol.com.br/colunas/deriva-genetica/pequena-historia-da-individualidade-genetica>. Acesso em: 4 jun. 2013.

PETRY, A.; ZANUZ, A.; BARONE, D. A. C. **Reconhecimento Automático De Pessoas Pela Voz Através de Técnicas de Processamento Digital De Sinais**. Universidade do Rio Grande do Sul. Porto Alegre, p. 4. 2000.

PREGAS Vocais. **Wikipedia**, 2013. Disponivel em: <http://pt.wikipedia.org/wiki/Pregas_vocais>. Acesso em: 14 mai. 2013.

RABINER, L. R.; JUANG, B.-H. **Fundamentals of Speech Recognition**. Prentice-Hall, v. 1, 1993. ISBN 0130151572.

RABINER, L. R.; SCHAFER, R. W. **Digital Processing of Speech Signals**. US. ed. Upper Saddle River, New Jersey: Prentice Hall, 1978. ISBN 978-0132136037.

REYNOLDS, D. A. **A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification**. Tese (Ph.D em Engenharia Elétrica), Georgia Institute of Technology. 1992.

REYNOLDS, D. A. Large population speaker identification using clean and telephone speech. **IEEE Signal Processing Letters**, v. 2, n. 3, p. 46-48, mar. 1995. ISSN 1070-9908.

REYNOLDS, D. A.; QUATIERI, T. F.; DUNN, R. B. Speaker Verification Using Adapted Gaussian Mixture Models. **Digital Signal Processing**, v. 10, n. 1-3, p. 19-41, 2000.

RIBEIRO, J. F. et al. Exames periciais em fonética forense: Recomendações técnicas para a padronização de procedimento em metodologias. **Associação Brasileira de Criminalística**, 2008. Disponivel em: <http://www.abcperitosoficiais.org.br/hotsites/seminariopara/Criminal-12-fonetica.pdf>.

ROSE, P. **Forensic Speaker Identification**. London: Taylor & Francis, 2002. ISBN 0-415-27182-7.

SAHIDULLAH, M.; SAHA, G. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. **Speech Communication**, 54, mai. 2012. 543-565.

SALMAN, A.; MUHAMMAD, E.; KHURSHID, K. Speaker Verification Using Boosted Cepstral Features with Gaussian Distributions. In: Multitopic Conference (INMIC). Lahore, **Anais...** Lahore, 2007. p. 1-5.

SANCHEZ, F. L. **Análise Cepstral Baseada em Diferentes Famílias de Transformada Wavelet**. 98 f. Dissertação (Mestrado em Engenharia Elétrica), Universidade de São Paulo. São Carlos. 2008.

SANDOUK, U. **Speaker Recognition - Speaker Diarization and Identification**. The Univesity of Manchester. p. 101. 2012.

SCATENA, H. J. **A Física Aplica à Perícia Criminal: Fonética Forense**. Universidade Católica de Brasília. Brasília, p. 32. 2010.

SILVA, E. C. L. Tecnologias biométricas: Comparação. **Grupo de Teleinformática e Automação**, 4 nov. 2007. Disponivel em: <http://www.gta.ufrj.br/grad/07_2/eric/Tecnologiasbiomtricas.comparao.html>. Acesso em: 29 mar. 2013.

SILVA, V. R. V. G. D. **Algoritmos para redução de ruído em sinais de áudio**. 49 f. Monografia (Graduaçao em Engenharia Elétrica), Universidade Federal do Rio de Janeiro. Rio de Janeiro. 2007.

SINGH, G. et al. Vector Quantization Techiniques for GMM Based Speaker Verification. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03). Honk Kong, **Anais...** Honk Kong, 2003. p. 65-72.

SKOSAN, M.; MASHAO, D. **Improving Speaker Identification Performance for Telephone-based Applications**. University of Cape Town. Rondebosch, p. 6. 2004.

SKOSAN, M.; MASHAO, D. Rapid and brief communication: Combining classifier decisions for robust speaker identification. **Pattern Recognition**, 39, jan. 2006. 147-155.

SKOWRONSKI, M. D.; HARRIS, J. G. Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition. **The Journal of the Acoustical Society of America**, 3, n. 116, set. 2004. 1774-1780.

SOHN, J.; KIM, N. S.; SUNG, W. A Statistical Model-Based Voice Activity Detection. **IEEE Signal Processing Letters**, 6, n. 1, jan. 1999. 1-3.

STEVENS, S. S.; VOLKMAN, J. The Relation of Pitch to Frequency: A Revised Scale. **American Journal of Psychology**, 53, jul. 1940. 329–353.

STIGAR, R. A Pessoa Humana na Perspectiva Humanista, 2012. Disponivel em: <http://meuartigo.brasilescola.com/filosofia/a-pessoa-humana-na-perspectiva-humanista.htm>. Acesso em: 04 jun. 2013.

SUBRAMANYA, A. et al. A Generative-Discriminative Framework Using Ensemble Methods For Text-Dependent Speaker Verification. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007). Honolulu, **Anais...** Honolulu, 2007. p. 225-228.

SVIRAVA, T. **The use of statistical methods in forensic speaker identification in Russian Federation**. Statistical Methods. 2009.

TOGNERI, R.; PULLELLA, D. An Overview of Speaker Identification: Accuracy and Robustness Issues. **IEEE Circuits and Systems Magazine**, v. 11, n. 2, 2011. ISSN 1531-636X.

VALENTE, C. R. Perspectivas da fonética forense num cenário de quebra do dogma da unicidade. In: Conferência Internacional de Ciências Forenses em Multimídia e Segurança Eletrônica. Brasília, **Anais...** Brasília, 2012. p. 7-27.

VASKAS, A. S.; ESFANDIYARI, A.; SHAMSHIRBAND, S. Modified Mfcc for Speaker Recognition. **Australian Journal of Basic and Applied Sciences**, 4, 2010. 4357-4364.

VIBHA, T. MFCC and its applications in speaker recognition. **International Journal on Emerging Technologies**, 2010. 19-22.

VOZ Humana. **Wikipedia**, 2013. Disponivel em: <http://pt.wikipedia.org/wiki/Voz_humana>. Acesso em: 14 mai. 2013.

WATTS, D. M. G. **Speaker Identification - Prototype Development and Performance**. 100 f. Monografia (Graduação em Engenharia Elétrica), University of Southern Queensland. p. 100. 2006.

APÊNDICE A

Os Detectores de Atividade Vocal são podem ser baseados na energia ao longo do sinal e também na verificação do número de cruzamentos por zero, como descritos com mais detalhes em seguida:

- **Energia**:

As técnicas que utilizam essencialmente detecção de energia são bem simples, porém bastante sensíveis à presença de ruído. O sinal é pré processado por meio de segmentação e janelamento, resultando em *M* blocos e a energia é dada pelo somatório da magnitude ou potência do sinal de cada segmento *x*, este variando entre 10 a 30ms, como mostram as Equações A.1 e A.2 para magnitude e potência, respectivamente (BURILEANU *et al.*, 2000):

$$E = \sum_{1}^{M} x_i, \ \ para\ i = 0, 1, 2, \dots, M, \tag{A.1}$$

$$E = \sum_{1}^{M} (x_i)^2, \ \ para\ i = 0, 1, 2, \dots, M, \tag{A.2}$$

Por fim, para a utilização desta técnica é preciso definir um limiar, ou conjunto destes, para que seja possível classificar a energia medida como silêncio (menor energia) ou atividade de voz (maior energia), como mostra a Figura A. **1**.

- **Cruzamento por Zero**:

A taxa de cruzamento por zero indica o número de vezes que as amostras de um sinal, em um determinado segmento, cruzam o zero (limiar) tomado como referência. Ao

contrário da energia, altas taxas de cruzamento por zero caracterizam os sons surdos e taxas mais reduzidas indicam a presença de sons sonoros (FECHINE, 2000).

Assim como na técnica baseada na energia, o sinal é pré processado por meio de segmentação e janelamento entre 10 e 30 ms, resultando em *M* blocos e a taxa de cruzamento por zero é obtida por meio das Equações A.3 e A.4 (PATRA, 2007):

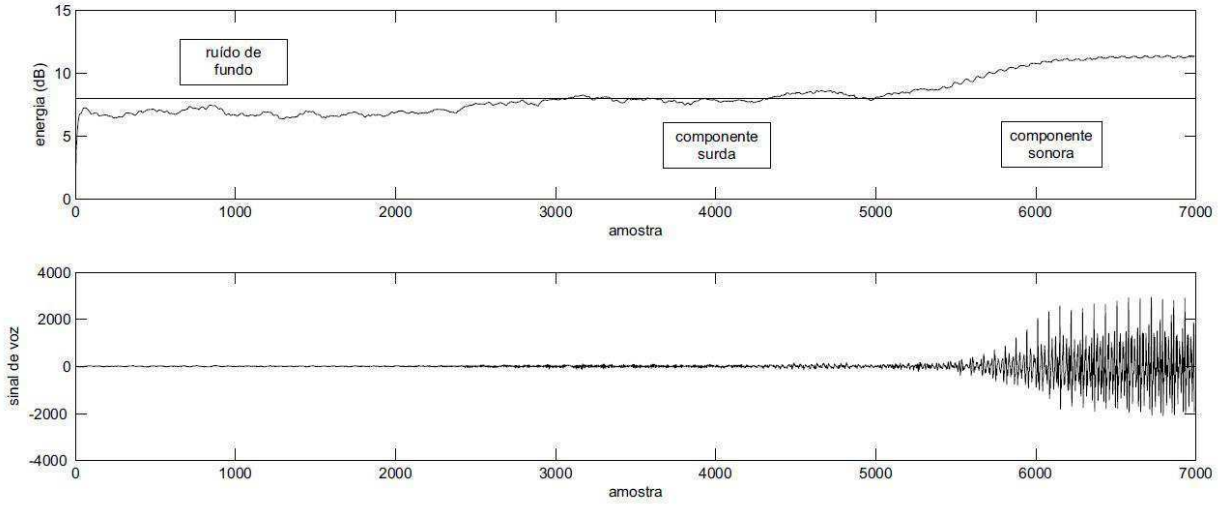
$$TCZ = \sum_{1}^{M} |sgn[s(n)] - sgn[s(n-1)]|, \tag{A.3}$$

em que:

$$sgn[s(n)] = \begin{cases} 1, se\ s(n) \geq 0 \\ -1, se\ s(n) < 0 \end{cases}, \tag{A.4}$$

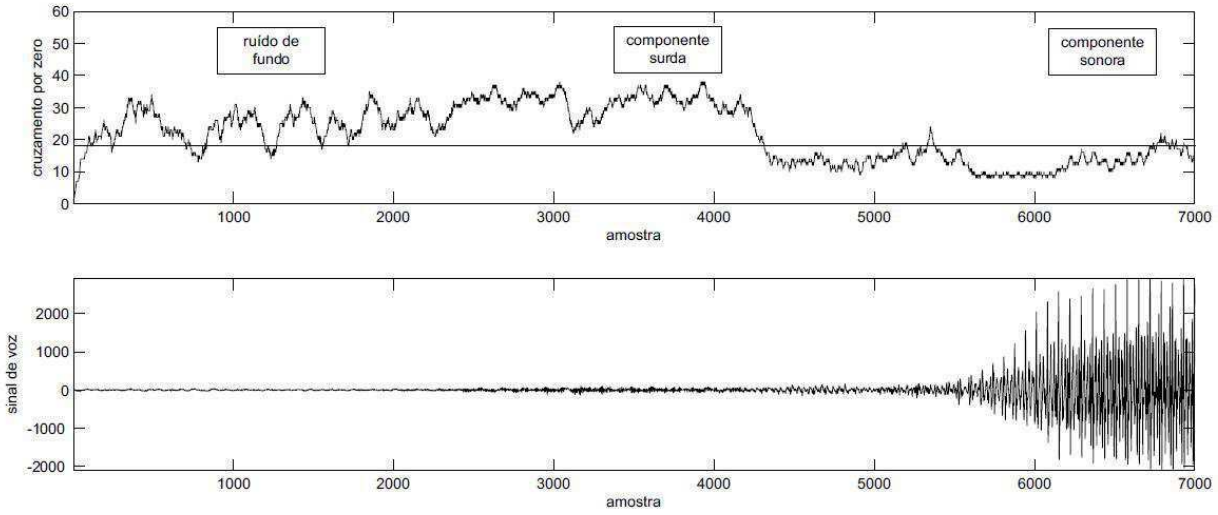
A partir de então, definir um limiar, ou conjunto destes, para que seja possível classificar a taxa de cruzamento por zero como silêncio ou atividade de voz, como mostra a Figura A. **2**.

Figura A. 1 - Diagrama de energia do sinal e limiar de classificação, seguido da respectiva elocução.



Fonte: CARDOSO, 2009.

Figura A. 2 - Diagrama de cruzamento por zero do sinal e limiar de classificação, seguido da respectiva elocução.



Fonte: CARDOSO, 2009.

APÊNDICE B

B.1 Conceitos Básicos

Na realização de uma pesquisa o ideal seria proceder-se a determinação das medidas sobre toda a população em estudo. Sendo esse procedimento impraticável (ou difícil), em alguns levantamentos, recorre-se ao processo de amostragem e, a partir da Estatística Indutiva (ou Inferência Estatística), pode-se tirar conclusões sobre a população com base nos resultados observados na amostra. O que é necessário garantir, em suma, é que a amostra seja representativa da população. Isso significa que, a menos de certas discrepâncias inerentes à aleatoriedade sempre presente, em maior ou menor grau, no processo de amostragem, a amostra deve possuir as mesmas características básicas da população, no que diz respeito à(s) variável(is) de interesse da pesquisa (COSTA NETO, 1977; FECHINE, 2000).

De acordo com Fechine (2000), *lei empírica do acaso* ou *lei dos grandes números* consagra o princípio de que a aproximação relativa aumenta à medida que cresce o número de determinações. A amostra deve incluir um número suficiente de casos, escolhidos aleatoriamente, para oferecer certa segurança estatística em relação à representatividade dos dados. Assim, o tamanho de uma amostra deve alcançar determinadas proporções mínimas, estabelecidas estatisticamente. Além disso, as necessidades práticas de tempo, custos, etc. recomendam não ultrapassar o tamanho mínimo determinado pela estatística.

Portanto, como os experimentos são baseados em uma amostra populacional, é de bom grado que se faça uso do intervalo de confiança para indicar a confiabilidade dessa estimativa. No caso da identificação automática de locutor, onde o resultado de cada identificação possui um retorno binário, e por fim resulta em uma taxa de aceitação proporcional ao número total de testes realizados é recomendada a utilização do intervalo de confiança para proporções que utiliza o método de aproximação normal.

Consideremos X a variável aleatória que representa a identificação (ou não) de cada locutor dentre a população. Assim temos que X tem distribuição de Bernoulli com parâmetro p, no qual p representa a probabilidade de um determinado elemento da amostra ter a característica de interesse. Retiramos uma amostra aleatória $X_1, X_2, \ldots, X_n$ desta população. Cada $X_i,\ i = 1, 2, \ldots,\ n$ tem distribuição de Bernoulli com parâmetro p, isto é,

$$X_1, X_2, \ldots, X_n \sim Bernoulli(p), \tag{B.1}$$

com média $\mu = p$ e variância $\sigma^2 = p(1\ -\ p)$.

Neste caso, o estimador de máxima verossimilhança $(\hat{p})$ para o parâmetro populacional p é dado por:

$$(\hat{p}) = \frac{Nº\ de\ locutores\ identificados}{Nº\ de\ testes\ realizados} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}, \tag{B.2}$$

Consideremos $\hat{p}$ a proporção amostral. Pelo Teorema Central do Limite temos que para um tamanho de amostra grande podemos considerar a proporção amostral $\hat{p}$ como tendo aproximadamente distribuição Normal com média p e variância $p(1-p)/n$. Desse modo segue que:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right), \tag{B.3}$$

Observemos que a variância $\hat{p}$ de depende do parâmetro desconhecido p. No entanto, pelo fato de n ser grande, podemos substituir p por $\hat{p}$. Com isso temos que:

$$IC(p, 1-\alpha) = \left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right), \tag{B.4}$$

Para fins de comparação entre duas proporções, pode ser utilizado o Intervalo de confiança para a diferença entre proporções, que resulta em um erro padrão referente aos resultados das duas proporções, como mostra a Equação em seguida:

$$EP = (\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \quad \text{(B.5)}$$

B.2 Grupo Experimental 1: Identificação Automática de Locutor

- **Divisão e tamanho treinamento/teste**

Para experimentos com testes longos, devido a concatenação das elocuções destinadas ao teste, foram realizados 630 testes em todas as variações propostas das divisões treinamento/teste. Portanto, os intervalos de confiança, a um nível de confiança $\alpha = 98\%$, foram todos calculados tendo $n = 630$, como mostrado a seguir:

- Teste longo com proporção de treinamento/ teste 60% / 40%

$$IC(0{,}793{,}1-0{,}98) = \left(0{,}7937 - 2{,}33\sqrt{\frac{(0{,}7937)(0{,}2063)}{630}}, 0{,}7937 + 2{,}33\sqrt{\frac{(0{,}7937)(0{,}2063)}{630}}\right),$$

$$IC(0{,}793{,}1-0{,}98) = ((0{,}7937) - (2{,}33)(0{,}016122)\ ,\ (0{,}7937) + (2{,}33)(0{,}016122)),$$

$$IC(0{,}793{,}1-0{,}98) = \big((0{,}7937) - (0{,}037563)\ ,\ (0{,}7937) + (0{,}037563)\big),$$

$$IC(0{,}793{,}1-0{,}98) = \big((0{,}7937) \pm (0{,}037563)\big)$$

- Teste longo com proporção de treinamento/ teste 70% / 30%

$$IC(0{,}8444{,}1-0{,}98) = \left(0{,}8444 - 2{,}33\sqrt{\frac{(0{,}8444)(0{,}1556)}{630}}, 0{,}8444 + 2{,}33\sqrt{\frac{(0{,}8444)(0{,}1556)}{630}}\right),$$

$$IC(0{,}8444{,}1-0{,}98) = ((0{,}8444) - (2{,}33)(0{,}014441)\ ,\ (0{,}8444) + (2{,}33)(0{,}014441)),$$

$$IC(0{,}8444{,}1-0{,}98)=\big((0{,}8444)-(0{,}033648)\ ,\ (0{,}8444)+(0{,}033648)\big),$$

$$IC(0{,}8444{,}1-0{,}98)=\big((0{,}8444)\pm(0{,}033648)\big)$$

- Teste longo com proporção de treinamento/ teste 80% / 20%

$$IC(0{,}8175{,}1-0{,}98)=\left(0{,}8175-2{,}33\sqrt{\frac{(0{,}8175)(0{,}1825)}{630}},0{,}8175+2{,}33\sqrt{\frac{(0{,}8175)(0{,}1825)}{630}}\right),$$

$$IC(0{,}8175{,}1-0{,}98)=((0{,}8175)-(2{,}33)(0{,}015389)\ ,\ (0{,}8175)+(2{,}33)(0{,}015389)),$$

$$IC(0{,}8175{,}1-0{,}98)=\big((0{,}8175)-(0{,}035856)\ ,\ (0{,}8175)+(0{,}035856)\big),$$

$$IC(0{,}8175{,}1-0{,}98)=\big((0{,}8175)\pm(0{,}035856)\big)$$

- Teste longo com proporção de treinamento/ teste 90% / 10%

$$IC(0{,}6587{,}1-0{,}98)=\left(0{,}6587-2{,}33\sqrt{\frac{(0{,}6587)(0{,}3413)}{630}},0{,}6587+2{,}33\sqrt{\frac{(0{,}6587)(0{,}3413)}{630}}\right),$$

$$IC(0{,}6587{,}1-0{,}98)=((0{,}6587)-(2{,}33)(0{,}01889)\ ,\ (0{,}6587)+(2{,}33)(0{,}01889)),$$

$$IC(0{,}6587{,}1-0{,}98)=\big((0{,}6587)-(0{,}044015)\ ,\ (0{,}6587)+(0{,}044015)\big),$$

$$IC(0{,}6587{,}1-0{,}98)=\big((0{,}6587)\pm(0{,}044015)\big)$$

Portanto, para determinar qual a melhor divisão a ser adotada, tomou-se como base a divisão detentora da maior taxa de identificação e verificou-se quais demais divisões possuíam intervalo de confiança que se sobrepunha ao intervalo de confiança desta. Sendo assim, a divisão 70% / 30% possuía a maior taxa de identificação e as divisões 60% / 40% e 80% / 20% possuíam esta sobreposição. Desta forma, a divisão 90% / 10% foi descartada, pois seu intervalo de confiança não possuía intersecção com o intervalo de confiança da divisão 70% / 30%.

Em seguida aplicou-se o Intervalo de Confiança para Diferença entre Duas Proporções entre a divisão 70% / 30% e as demais, de forma individual, como relatado a seguir:

- Comparação entre testes longos com proporção de treinamento/ teste 70% / 30% e 60% / 40%

$$EP = (\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

$$EP = (0{,}8444 - 0{,}8175) \pm 2{,}33\sqrt{\frac{(0{,}8444)(0{,}1556)}{630} + \frac{(0{,}7937)(0{,}2063)}{630}},$$

$$EP = 0{,}0507 \pm 2{,}33\sqrt{(0{,}0002085533968253968) + (0{,}0002599052539682 54)},$$

$$EP = 0{,}0507 \pm (2{,}33)(0{,}021643905627071),$$

$$EP = 0{,}0507 \pm 0{,}0504,$$

$$EP = [0{,}0003; 0{,}1011]$$

- Comparação entre testes longos com proporção de treinamento/ teste 70% / 30% e 80% / 20%

$$EP = (\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

$$EP = (0{,}8444 - 0{,}8175) \pm 2{,}33\sqrt{\frac{(0{,}8444)(0{,}1556)}{630} + \frac{(0{,}8175)(0{,}1825)}{630}},$$

$$EP = 0{,}0329 \pm 2{,}33\sqrt{(0{,}0002085533968253968) + (0{,}0002368154761904762)},$$

$$EP = 0{,}0329 \pm (2{,}33)(0{,}0211037644276033),$$

$$EP = 0{,}0329 \pm 0{,}0492,$$

$$EP = [-0{,}0183; 0{,}0821]$$

Portanto, foram obtidos os intervalos de erro padrão $[0{,}0003; 0{,}1011]$ e $[-0{,}0183; 0{,}0821]$ para 60% / 40% e 80% / 20%, respectivamente. A presença do 0 (zero) no intervalo mostra que é plausível a igualdade das proporções, o que acontece para a divisão 80% / 20%, descartando assim a divisão 60% / 40% por não possuir o 0 (zero) no intervalo.

Nos experimentos com testes curtos, a não concatenação das elocuções destinadas à teste possibilita um maior número de realizações destes, na maioria dos casos. O número de testes é proporcional a proporção que lhe é destinada. Portanto os intervalos de confiança, a um nível de confiança de $\alpha = 98\%$, foram todos calculados tendo $n = 2520$, $n = 1890$, $n = 1260$ e $n = 630$ para as respectivas proporções de treinamento/teste 60% / 40%, 70% / 30%, 80% / 20% e 90% / 10%, como mostrado a seguir.

- Teste curto com proporção de treinamento/ teste 60% / 40%

$$IC(0{,}5159{,}1-0{,}98) = \left(0{,}5159 - 2{,}33\sqrt{\frac{(0{,}5159)(0{,}4841)}{1520}}, 0{,}8444 + 2{,}33\sqrt{\frac{(0{,}5159)(0{,}4841)}{1520}}\right),$$

$$IC(0{,}5159{,}1-0{,}98) = ((0{,}5159) - (2{,}33)(0{,}009955)\ ,\ (0{,}5159) + (2{,}33)(0{,}009955)),$$

$$IC(0{,}5159{,}1-0{,}98) = \big((0{,}5159) - (0{,}023196)\ ,\ (0{,}5159) + (0{,}023196)\big),$$

$$IC(0{,}5159{,}1-0{,}98) = \big((0{,}5159) \pm (0{,}023196)\big)$$

- Teste curto com proporção de treinamento/ teste 70% / 30%

$$IC(0{,}573{,}1-0{,}98) = \left(0{,}573 - 2{,}33\sqrt{\frac{(0{,}573)(0{,}427)}{1890}}, 0{,}573 + 2{,}33\sqrt{\frac{(0{,}573)(0{,}427)}{1890}}\right),$$

$$IC(0{,}573{,}1-0{,}98) = ((0{,}573) - (2{,}33)(0{,}011378)\ ,\ (0{,}573) + (2{,}33)(0{,}011378)),$$

$$IC(0{,}573{,}1-0{,}98) = \big((0{,}573) - (0{,}02651)\ ,\ (0{,}573) + (0{,}02651)\big),$$

$$IC(0{,}573{,}1-0{,}98) = \big((0{,}573) \pm (0{,}02651)\big)$$

- Teste curto com proporção de treinamento/ teste 80% / 20%

$$IC(0{,}6349{,}1-0{,}98) = \left(0{,}6349 - 2{,}33\sqrt{\frac{(0{,}6349)(0{,}3651)}{1260}}, 0{,}6349 + 2{,}33\sqrt{\frac{(0{,}6349)(0{,}3651)}{1260}}\right),$$

$$IC(0{,}6349{,}1-0{,}98) = ((0{,}6349) - (2{,}33)(0{,}013564)\ ,\ (0{,}6349) + (2{,}33)(0{,}013564)),$$

$$IC(0{,}6349{,}1-0{,}98) = \big((0{,}6349) - (0{,}031603)\ ,\ (0{,}6349) + (0{,}031603)\big),$$

$IC(0{,}6349{,}1-0{,}98) = \big((0{,}6349) \pm (0{,}031603)\big)$,

- Teste curto com proporção de treinamento/ teste 90% / 10%

$$IC(0{,}6587{,}1-0{,}98) = \left(0{,}6587 - 2{,}33\sqrt{\frac{(0{,}6587)(0{,}3413)}{630}}, 0{,}6587 + 2{,}33\sqrt{\frac{(0{,}6587)(0{,}3413)}{630}}\right),$$

$IC(0{,}6587{,}1-0{,}98) = ((0{,}6587) - (2{,}33)(0{,}01889)\ ,\ (0{,}6587) + (2{,}33)(0{,}01889))$,

$IC(0{,}6587{,}1-0{,}98) = \big((0{,}6587) - (0{,}044015)\ ,\ (0{,}6587) + (0{,}044015)\big)$,

$IC(0{,}6587{,}1-0{,}98) = \big((0{,}6587) \pm (0{,}044015)\big)$

Como realizado na análise para os testes longos, para determinar qual a melhor divisão a ser adotada, tomou-se como base a divisão detentora da maior taxa de identificação e verificou-se quais demais divisões possuíam intervalo de confiança que se sobrepunha ao intervalo de confiança desta. Sendo assim a divisão 90% / 10% possuía a maior taxa de identificação e apenas a divisão 80% / 20% possuía esta sobreposição. Dessa forma as divisões 60% / 40% e 70% / 30% foram descartadas, pois seus intervalos de confiança não possuíam intersecção com o intervalo de confiança da divisão 90% / 10%.

Em seguida aplicou-se o Intervalo de Confiança para Diferença entre Duas Proporções entre as divisões 90% / 10% e 80% / 20%, como relatado a seguir:

- Comparação entre testes longos com proporção de treinamento/ teste 90% / 10% e 80% / 20%

$$EP = (\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

$$EP = (0{,}6587 - 0{,}6349) \pm 2{,}33\sqrt{\frac{(0{,}6587)(0{,}3413)}{630} + \frac{(0{,}6349)(0{,}3651)}{1260}},$$

$EP = 0{,}0238 \pm 2{,}33\sqrt{(0{,}0003563605396825397) + (0{,}0001880554761904762)}$,

$EP = 0{,}0238 \pm (2{,}33)(0{,}0233327241417066)$,

$EP = 0{,}0238 \pm 0{,}0544,$

$EP = [-0{,}0266; 0{,}0742]$

Obteve-se então o intervalo de erro padrão $[-0{,}0266; 0{,}0742]$. A presença do 0 (zero) no intervalo mostra que é plausível a igualdade das proporções, possuindo desempenho equivalente entre si.

- **Escalabilidade**

A escolha das divisões de treinamento/teste adotadas nessa fase foram subsidiadas pelos resultados que mostram que as melhores configurações de treinamentos/teste são 90% / 10% e 80% / 20% para testes curtos e divisões 70% / 30% e 80% / 20% para testes longos.

Apesar de, em suas devidas categorias, apresentarem equivalência de desempenho entre si em um nível de confiança $\alpha = 98\%$, é possível afirmar de acordo com intervalo de erro gerado entre as comparações de proporções, que é mais provável a obtenção de uma maior taxa de identificação utilizando as divisões 70% / 30% para testes longos e 90% / 10% para testes curtos.

- Teste longo com proporção de treinamento/ teste 70% / 30% (n=100)

$$IC(0{,}9300{,}1 - 0{,}98) = \left(0{,}9300 - 2{,}33\sqrt{\frac{(0{,}9300)(0{,}0700)}{100}}, 0{,}9300 + 2{,}33\sqrt{\frac{(0{,}9300)(0{,}0700)}{100}}\right),$$

$$IC(0{,}9300{,}1 - 0{,}98) = ((0{,}9300) - (2{,}33)(0{,}025515)\ ,\ (0{,}9300) + (2{,}33)(0{,}025515)),$$

$$IC(0{,}9300{,}1 - 0{,}98) = \big((0{,}9300) - (0{,}059449)\ ,\ (0{,}9300) + (0{,}059449)\big),$$

$$IC(0{,}9300{,}1 - 0{,}98) = \big((0{,}9300) \pm (0{,}059449)\big)$$

- Teste curto com proporção de treinamento/ teste 90% / 10% (n=100)

$$IC(0{,}7900{,}1-0{,}98)=\left(0{,}7900-2{,}33\sqrt{\frac{(0{,}7900)(0{,}2100)}{100}},0{,}7900+2{,}33\sqrt{\frac{(0{,}7900)(0{,}2100)}{100}}\right),$$

$$IC(0{,}7900{,}1-0{,}98)=((0{,}7900)-(2{,}33)(0{,}040731)\ ,\ (0{,}7900)+(2{,}33)(0{,}040731)),$$

$$IC(0{,}7900{,}1-0{,}98)=\big((0{,}7900)-(0{,}094903)\ ,\ (0{,}7900)+(0{,}094903)\big),$$

$$IC(0{,}7900{,}1-0{,}98)=\big((0{,}7900)\pm(0{,}094903)\big)$$

- Teste longo com proporção de treinamento/ teste 70% / 30% (n=200)

$$IC(0{,}8750{,}1-0{,}98)=\left(0{,}8750-2{,}33\sqrt{\frac{(0{,}8750)(0{,}125)}{200}},0{,}8750+2{,}33\sqrt{\frac{(0{,}8750)(0{,}1250)}{200}}\right),$$

$$IC(0{,}8750{,}1-0{,}98)=((0{,}8750)-(2{,}33)(0{,}023385)\ ,\ (0{,}8750)+(2{,}33)(0{,}023385)),$$

$$IC(0{,}8750{,}1-0{,}98)=\big((0{,}8750)-(0{,}054488)\ ,\ (0{,}8750)+(0{,}054488)\big),$$

$$IC(0{,}8750{,}1-0{,}98)=\big((0{,}8750)\pm(0{,}054488)\big)$$

- Teste curto com proporção de treinamento/ teste 90% / 10% (n=200)

$$IC(0{,}7350{,}1-0{,}98)=\left(0{,}7350-2{,}33\sqrt{\frac{(0{,}7350)(0{,}2650)}{200}},0{,}7350+2{,}33\sqrt{\frac{(0{,}7350)(0{,}2650)}{200}}\right),$$

$$IC(0{,}7350{,}1-0{,}98)=((0{,}7350)-(2{,}33)(0{,}031837)\ ,\ (0{,}7350)+(2{,}33)(0{,}031837)),$$

$$IC(0{,}7350{,}1-0{,}98)=\big((0{,}7350)-(0{,}074181)\ ,\ (0{,}7350)+(0{,}074181)\big),$$

$$IC(0{,}7350{,}1-0{,}98)=\big((0{,}7350)\pm(0{,}074181)\big)$$

- Teste longo com proporção de treinamento/ teste 70% / 30% (n=300)

$$IC(0{,}8667{,}1-0{,}98)=\left(0{,}8667-2{,}33\sqrt{\frac{(0{,}8667)(0{,}1333)}{300}},0{,}8667+2{,}33\sqrt{\frac{(0{,}8667)(0{,}1333)}{300}}\right),$$

$$IC(0{,}8667{,}1-0{,}98)=((0{,}8667)-(2{,}33)(0{,}019624)\ ,\ (0{,}8667)+(2{,}33)(0{,}019624)),$$

$$IC(0{,}8667{,}1-0{,}98)=\big((0{,}8667)-(0{,}045724)\ ,\ (0{,}8667)+(0{,}045724)\big),$$

$$IC(0{,}8667{,}1-0{,}98)=\big((0{,}8667)\pm(0{,}045724)\big)$$

- Teste curto com proporção de treinamento/ teste 90% / 10% (n=300)

$$IC(0{,}6833{,}1-0{,}98)=\left(0{,}6833-2{,}33\sqrt{\frac{(0{,}6833)(0{,}3167)}{300}},0{,}6833+2{,}33\sqrt{\frac{(0{,}6833)(0{,}3167)}{300}}\right),$$

$$IC(0{,}6833{,}1-0{,}98)=((0{,}6833)-(2{,}33)(0{,}026858)\ ,\ (0{,}6833)+(2{,}33)(0{,}026858)),$$

$$IC(0{,}6833{,}1-0{,}98)=\big((0{,}6833)-(0{,}062578)\ ,\ (0{,}6833)+(0{,}062578)\big),$$

$$IC(0{,}6833{,}1-0{,}98)=\big((0{,}6833)\pm(0{,}062578)\big)$$

- Teste longo com proporção de treinamento/ teste 70% / 30% (n=400)

$$IC(0{,}8550{,}1-0{,}98)=\left(0{,}8550-2{,}33\sqrt{\frac{(0{,}8550)(0{,}1450)}{400}},0{,}8550+2{,}33\sqrt{\frac{(0{,}8550)(0{,}1450)}{400}}\right),$$

$$IC(0{,}8550{,}1-0{,}98)=((0{,}8550)-(2{,}33)(0{,}017605)\ ,\ (0{,}8550)+(2{,}33)(0{,}017605)),$$

$$IC(0{,}8550{,}1-0{,}98)=\big((0{,}8550)-(0{,}041020)\ ,\ (0{,}8550)+(0{,}041020)\big),$$

$$IC(0{,}8550{,}1-0{,}98)=\big((0{,}8550)\pm(0{,}041020)\big)$$

- Teste curto com proporção de treinamento/ teste 90% / 10% (n=400)

$$IC(0{,}6800{,}1-0{,}98)=\left(0{,}6800-2{,}33\sqrt{\frac{(0{,}6800)(0{,}3200)}{400}},0{,}6800+2{,}33\sqrt{\frac{(0{,}6800)(0{,}3200)}{400}}\right),$$

$$IC(0{,}6800{,}1-0{,}98)=((0{,}6800)-(2{,}33)(0{,}023324)\ ,\ (0{,}6800)+(2{,}33)(0{,}023324)),$$

$$IC(0{,}6800{,}1-0{,}98)=\big((0{,}6800)-(0{,}054344)\ ,\ (0{,}6800)+(0{,}054344)\big),$$

$$IC(0{,}6800{,}1-0{,}98)=\big((0{,}6800)\pm(0{,}054344)\big)$$

- Teste longo com proporção de treinamento/ teste 70% / 30% (n=500)

$$IC(0{,}8500{,}1-0{,}98)=\left(0{,}8500-2{,}33\sqrt{\frac{(0{,}8500)(0{,}1500)}{500}},0{,}8500+2{,}33\sqrt{\frac{(0{,}8500)(0{,}1500)}{500}}\right),$$

$$IC(0{,}8500{,}1-0{,}98) = ((0{,}8500) - (2{,}33)(0{,}015969)\ ,\ (0{,}8500) + (2{,}33)(0{,}015969)),$$

$$IC(0{,}8500{,}1-0{,}98) = \big((0{,}8500) - (0{,}037207)\ ,\ (0{,}8500) + (0{,}037207)\big),$$

$$IC(0{,}8500{,}1-0{,}98) = \big((0{,}8500) \pm (0{,}037207)\big)$$

- Teste curto com proporção de treinamento/ teste 90% / 10% (n=500)

$$IC(0{,}6700{,}1-0{,}98) = \left(0{,}6700 - 2{,}33\sqrt{\frac{(0{,}6700)(0{,}3300)}{500}}, 0{,}6700 + 2{,}33\sqrt{\frac{(0{,}6700)(0{,}3300)}{500}}\right),$$

$$IC(0{,}6700{,}1-0{,}98) = ((0{,}6700) - (2{,}33)(0{,}021029)\ ,\ (0{,}6700) + (2{,}33)(0{,}021029)),$$

$$IC(0{,}6700{,}1-0{,}98) = \big((0{,}6700) - (0{,}048997)\ ,\ (0{,}6700) + (0{,}048997)\big),$$

$$IC(0{,}6700{,}1-0{,}98) = \big((0{,}6700) \pm (0{,}048997)\big)$$

- Teste longo com proporção de treinamento/ teste 70% / 30% (n=600)

$$IC(0{,}8467{,}1-0{,}98) = \left(0{,}8467 - 2{,}33\sqrt{\frac{(0{,}8467)(0{,}1533)}{600}}, 0{,}8467 + 2{,}33\sqrt{\frac{(0{,}8467)(0{,}1533)}{600}}\right),$$

$$IC(0{,}8467{,}1-0{,}98) = ((0{,}8467) - (2{,}33)(0{,}014708)\ ,\ (0{,}8467) + (2{,}33)(0{,}014708)),$$

$$IC(0{,}8467{,}1-0{,}98) = \big((0{,}8467) - (0{,}034270)\ ,\ (0{,}8467) + (0{,}034270)\big),$$

$$IC(0{,}8467{,}1-0{,}98) = \big((0{,}8467) \pm (0{,}034270)\big)$$

- Teste curto com proporção de treinamento/ teste 90% / 10% (n=600)

$$IC(0{,}6650{,}1-0{,}98) = \left(0{,}6650 - 2{,}33\sqrt{\frac{(0{,}6650)(0{,}3350)}{600}}, 0{,}8467 + 2{,}33\sqrt{\frac{(0{,}6650)(0{,}3350)}{600}}\right),$$

$$IC(0{,}6650{,}1-0{,}98) = ((0{,}6650) - (2{,}33)(0{,}019269)\ ,\ (0{,}8467) + (2{,}33)(0{,}019269)),$$

$$IC(0{,}6650{,}1-0{,}98) = \big((0{,}6650) - (0{,}044897)\ ,\ (0{,}8467) + (0{,}044897)\big),$$

$$IC(0{,}6650{,}1-0{,}98) = \big((0{,}6650) \pm (0{,}044897)\big)$$

Os pontos que representam as taxas de identificação dos testes curtos e longos podem ser representados regressivamente pelas funções logarítmicas naturais a seguir, de forma respectiva:

$$y_L = -0{,}039ln(x) + 1{,}093,$$

$$y_C = -0{,}083ln(x) + 1{,}1814,$$

Representações estas fidedignas, pois apresentam coeficientes de regressão próximos do ideal $R^2 \cong 1$, respectivamente, ${R_L}^2 = 0{,}9777$ e ${R_C}^2 = 0{,}9869$.

As funções são dispostas no formato $y = \mathrm{a}.ln(x) + \mathrm{b}$, em que a é a constante responsável pela ascendência (para valores positivos) ou descendência (valores negativos) da função e b pela translação vertical da função. Dessa forma, percebe-se que a função logarítmica natural regressiva dos testes longos apresenta menor decaimento quando comparada à dos testes curtos, por apresentar maior valor da constante a.

- **Gênero**

De acordo com os experimentos realizados nos dois subtópicos anteriores, concluiu-se que a taxa de identificação apresentava melhor rendimento quando utilizada a divisão treinamento/teste longo de 70% / 30%. Portanto, esta foi a configuração adotada para a realização dos experimentos em questão.

- Teste longo com proporção de treinamento/ teste 70% / 30% (n=438, masculino)

$$IC(0{,}8493{,}1 - 0{,}98) = \left(0{,}8493 - 2{,}33\sqrt{\frac{(0{,}8493)(0{,}1507)}{438}}, 0{,}8493 + 2{,}33\sqrt{\frac{(0{,}8493)(0{,}1507)}{438}}\right),$$

$$IC(0{,}8493{,}1 - 0{,}98) = ((0{,}8493) - (2{,}33)(0{,}017094)\ ,\ (0{,}8493) + (2{,}33)(0{,}017094)),$$

$$IC(0{,}8493{,}1 - 0{,}98) = \big((0{,}8493) - (0{,}039830)\ ,\ (0{,}8493) + (0{,}039830)\big),$$

$$IC(0{,}8493{,}1 - 0{,}98) = \big((0{,}8493) \pm (0{,}039830)\big)$$

- Teste longo com proporção de treinamento/ teste 70% / 30% (n=192, feminino)

$$IC(0{,}8281{,}1-0{,}98)=\left(0{,}8281-2{,}33\sqrt{\frac{(0{,}8281)(0{,}1719)}{192}},0{,}8281+2{,}33\sqrt{\frac{(0{,}8281)(0{,}1719)}{192}}\right),$$

$$IC(0{,}8281{,}1-0{,}98)=((0{,}8281)-(2{,}33)(0{,}027229)\ ,\ (0{,}8281)+(2{,}33)(0{,}027229)),$$

$$IC(0{,}8281{,}1-0{,}98)=\big((0{,}8281)-(0{,}063443)\ ,\ (0{,}8281)+(0{,}063443)\big),$$

$$IC(0{,}8281{,}1-0{,}98)=\big((0{,}8281)\pm(0{,}063443)\big)$$

A média (M) proporcional da taxa de identificação oriunda dos testes dos dois subgrupos, masculino (T_M) e feminino (T_F), isoladamente é dada de acordo com os cálculos a seguir:

$$M=(T_M\times P_M)+(T_F\times P_F)$$

$$M=\big(0{,}8493\times {}^{438}/_{630}\big)+\big(0{,}8333\times {}^{192}/_{630}\big)=$$
$$=0{,}5909+0{,}2540=0{,}8449$$

Em seguida, aplicou-se o intervalo de confiança de forma individual e para diferença entre as duas proporções, no que diz respeito as taxas de identificação sem e com divisão por gênero, como mostrado a seguir:

- Teste longo com proporção de treinamento/ teste 70% / 30% (sem divisão por gênero)

$$IC(0{,}8444{,}1-0{,}98)=\left(0{,}8444-2{,}33\sqrt{\frac{(0{,}8444)(0{,}1556)}{630}},0{,}8444+2{,}33\sqrt{\frac{(0{,}8444)(0{,}1556)}{630}}\right),$$

$$IC(0{,}8444{,}1-0{,}98)=((0{,}8444)-(2{,}33)(0{,}014441)\ ,\ (0{,}8444)+(2{,}33)(0{,}014441)),$$

$$IC(0{,}8444{,}1-0{,}98)=\big((0{,}8444)-(0{,}033648)\ ,\ (0{,}8444)+(0{,}033648)\big),$$

$$IC(0{,}8444{,}1-0{,}98) = \big((0{,}8444) \pm (0{,}033648)\big)$$

- Teste longo com proporção de treinamento/ teste 70% / 30% (com divisão por gênero)

$$IC(0{,}8449{,}1-0{,}98) = \left(0{,}8449 - 2{,}33\sqrt{\tfrac{(0{,}8449)(0{,}1551)}{630}}, 0{,}8449 + 2{,}33\sqrt{\tfrac{(0{,}8449)(0{,}1551)}{630}}\right),$$

$$IC(0{,}8449{,}1-0{,}98) = ((0{,}8449) - (2{,}33)(0{,}014422)\ ,\ (0{,}8449) + (2{,}33)(0{,}014422)),$$

$$IC(0{,}8449{,}1-0{,}98) = \big((0{,}8449) - (0{,}033603)\ ,\ (0{,}8449) + (0{,}033603)\big),$$

$$IC(0{,}8449{,}1-0{,}98) = \big((0{,}8449) \pm (0{,}033603)\big)$$

- Comparação entre testes longos, com proporção de treinamento/ teste 70% / 30%, com divisão por gênero e sem divisão por gênero

$$EP = (\hat{p}_1 - \hat{p}_2\,) \pm Z_{\alpha/2}\sqrt{\tfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \tfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

$$EP = (0{,}8449 - 0{,}8444) \pm 2{,}33\sqrt{\tfrac{(0{,}8449)(0{,}1551)}{630} + \tfrac{(0{,}8444)(0{,}1556)}{630}},$$

$$EP = 0{,}0005 \pm 2{,}33\sqrt{(0{,}0002080063333333333) + (0{,}0002085533968253968)},$$

$$EP = 0{,}0005 \pm (2{,}33)(0{,}0204097949563104),$$

$$EP = 0{,}0005 \pm 0{,}0476,$$

$$EP = [-0{,}0471; 0{,}0481]$$

Na diferença entre as duas proporções, obteve-se o intervalo de erro padrão $[-0{,}0471; 0{,}0481]$. A presença do 0 (zero) no intervalo mostra que é plausível a igualdade das proporções.

Em um segundo momento, visando analisar o desempenho separadamente de cada gênero, mais um subgrupo de testes foi criado, este composto por 192 locutores

do gênero masculino escolhidos de forma aleatória para que se pudesse comparar em igualdade numérica com o subgrupo de 192 locutores do gênero feminino.

De posse da taxa de identificação e quantidade de testes realizados ($n = 192$), utilizando um nível de confiança $\alpha = 98\%$, foi possível chegar ao intervalo de confiança de proporções dos experimentos realizados, como mostrado a seguir:

- Teste longo do gênero masculino (n=192)

$$IC(0{,}9167{,}1-0{,}98) = \left(0{,}9167 - 2{,}33\sqrt{\frac{(0{,}9167)(0{,}0833)}{192}}, 0{,}9167 + 2{,}33\sqrt{\frac{(0{,}9167)(0{,}0833)}{192}}\right),$$

$$IC(0{,}9167{,}1-0{,}98) = ((0{,}9167) - (2{,}33)(0{,}019943)\ ,\ (0{,}9167) + (2{,}33)(0{,}019943)),$$

$$IC(0{,}9167{,}1-0{,}98) = \big((0{,}9167) - (0{,}046467)\ ,\ (0{,}9167) + (0{,}046467)\big),$$

$$IC(0{,}9167{,}1-0{,}98) = \big((0{,}9167) \pm (0{,}050221)\big)$$

- Teste longo do gênero feminino (n=192)

$$IC(0{,}8281{,}1-0{,}98) = \left(0{,}8281 - 2{,}33\sqrt{\frac{(0{,}8281)(0{,}1719)}{192}}, 0{,}8281 + 2{,}33\sqrt{\frac{(0{,}8281)(0{,}1719)}{192}}\right),$$

$$IC(0{,}8281{,}1-0{,}98) = ((0{,}8281) - (2{,}33)(0{,}027229)\ ,\ (0{,}8281) + (2{,}33)(0{,}027229)),$$

$$IC(0{,}8281{,}1-0{,}98) = \big((0{,}8281) - (0{,}063443)\ ,\ (0{,}8281) + (0{,}063443)\big),$$

$$IC(0{,}8281{,}1-0{,}98) = \big((0{,}8281) \pm (0{,}063443)\big)$$

Posteriormente foi aplicado o Intervalo de Confiança para Diferença entre Duas Proporções entre as taxas de identificação dos gêneros masculino e feminino, nessa ordem, como pode ser visualizado a seguir:

- Comparação entre os testes do gênero masculino e feminino

$$EP = (\hat{p}_1 - \hat{p}_2\,) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

$$EP = (0{,}9167 - 0{,}8333) \pm 2{,}33\sqrt{\frac{(0{,}9167)(0{,}0833)}{192} + \frac{(0{,}8333)(0{,}1719)}{192}},$$

$$EP = 0{,}0834 \pm 2{,}33\sqrt{(0{,}0003977141145833333) + (0{,}00074606390625)},$$

$$EP = 0{,}0834 \pm (2{,}33)(0{,}033819787415555),$$

$$EP = 0{,}0834 \pm 0{,}0796,$$

$$EP = [0{,}0038; 0{,}163]$$

Por fim, obteve-se o intervalo de erro padrão $[0{,}0038; 0{,}163]$ na comparação entre os testes do gênero masculino e feminino. A ausência do 0 (zero) no intervalo mostra que não é plausível a igualdade das proporções, possuindo desempenho distinto entre si, podendo afirmar então que a taxa de identificação apresenta desempenho superior no subgrupo do gênero masculino em comparação ao feminino.

- **Região**

O critério para utilização da configuração treinamento/teste e o tipo de teste utilizado foi a mesma adotada no tópico referente ao gênero, sendo a divisão treinamento/teste longo de 70% / 30%.

De posse da taxa de identificação e quantidade de testes realizados de cada subgrupo (região), utilizando um nível de confiança $\alpha = 98\%$, foi possível chegar ao intervalo de confiança de proporções dos experimentos realizados, como mostrado em seguida:

- New England (n=49)

$$IC(0{,}9796{,}1 - 0{,}98) = \left(0{,}9796 - 2{,}33\sqrt{\frac{(0{,}9796)(0{,}0204)}{49}}, 0{,}9796 + 2{,}33\sqrt{\frac{(0{,}9796)(0{,}0204)}{49}}\right),$$

$$IC(0{,}9796{,}1 - 0{,}98) = ((0{,}9796) - (2{,}33)(0{,}020194)\ ,\ (0{,}9796) + (2{,}33)(0{,}020194)),$$

$$IC(0{,}9796{,}1 - 0{,}98) = \big((0{,}9796) - (0{,}047054)\ ,\ (0{,}9796) + (0{,}047054)\big),$$

$$IC(0{,}9796{,}1-0{,}98)=\big((0{,}9796)\pm(0{,}047054)\big),$$

- Northern (n=102)

$$IC(0{,}9314{,}1-0{,}98)=\left(0{,}9314-2{,}33\sqrt{\frac{(0{,}9314)(0{,}0686)}{102}},0{,}9314+2{,}33\sqrt{\frac{(0{,}9314)(0{,}0686)}{102}}\right),$$

$$IC(0{,}9314{,}1-0{,}98)=((0{,}9314)-(2{,}33)(0{,}025028)\ ,\ (0{,}9314)+(2{,}33)(0{,}025028)),$$

$$IC(0{,}9314{,}1-0{,}98)=\big((0{,}9314)-(0{,}058316)\ ,\ (0{,}9314)+(0{,}058316)\big),$$

$$IC(0{,}9314{,}1-0{,}98)=\big((0{,}9314)\pm(0{,}058316)\big),$$

- North Midland (n=102)

$$IC(0{,}902{,}1-0{,}98)=\left(0{,}902-2{,}33\sqrt{\frac{(0{,}902)(0{,}098)}{102}},0{,}902+2{,}33\sqrt{\frac{(0{,}902)(0{,}098)}{102}}\right),$$

$$IC(0{,}902{,}1-0{,}98)=((0{,}902)-(2{,}33)(0{,}029436)\ ,\ (0{,}902)+(2{,}33)(0{,}029436)),$$

$$IC(0{,}902{,}1-0{,}98)=\big((0{,}902)-(0{,}068592)\ ,\ (0{,}902)+(0{,}068592)\big),$$

$$IC(0{,}902{,}1-0{,}98)=\big((0{,}902)\pm(0{,}068592)\big)$$

- South Midland (n=100)

$$IC(0{,}94{,}1-0{,}98)=\left(0{,}94-2{,}33\sqrt{\frac{(0{,}94)(0{,}06)}{100}},0{,}94+2{,}33\sqrt{\frac{(0{,}94)(0{,}06)}{100}}\right),$$

$$IC(0{,}94{,}1-0{,}98)=((0{,}94)-(2{,}33)(0{,}023749)\ ,\ (0{,}94)+(2{,}33)(0{,}023749)),$$

$$IC(0{,}94{,}1-0{,}98)=\big((0{,}94)-(0{,}055334)\ ,\ (0{,}94)+(0{,}055334)\big),$$

$$IC(0{,}94{,}1-0{,}98)=\big((0{,}94)\pm(0{,}055334)\big)$$

- Southern (n=98)

$$IC(0{,}9082{,}1-0{,}98)=\left(0{,}9082-2{,}33\sqrt{\frac{(0{,}9082)(0{,}0918)}{98}},0{,}9082+2{,}33\sqrt{\frac{(0{,}9082)(0{,}0918)}{98}}\right),$$

$$IC(0{,}9082{,}1-0{,}98)=((0{,}9082)-(2{,}33)(0{,}029167)\ ,\ (0{,}9082)+(2{,}33)(0{,}029167)),$$

$$IC(0{,}9082{,}1-0{,}98)=\big((0{,}9082)-(0{,}067960)\ ,\ (0{,}9082)+(0{,}067960)\big),$$

$$IC(0{,}9082{,}1-0{,}98)=\big((0{,}9082)\pm(0{,}067960)\big)$$

- New York City (n=46)

$$IC(0{,}9783{,}1-0{,}98)=\left(0{,}9783-2{,}33\sqrt{\tfrac{(0{,}9783)(0{,}0217)}{46}},0{,}9783+2{,}33\sqrt{\tfrac{(0{,}9783)(0{,}0217)}{46}}\right),$$

$$IC(0{,}9783{,}1-0{,}98)=((0{,}9783)-(2{,}33)(0{,}021483)\ ,\ (0{,}9783)+(2{,}33)(0{,}021483)),$$

$$IC(0{,}9783{,}1-0{,}98)=\big((0{,}9783)-(0{,}050054)\ ,\ (0{,}9783)+(0{,}050054)\big),$$

$$IC(0{,}9783{,}1-0{,}98)=\big((0{,}9783)\pm(0{,}050054)\big)$$

- Western (n=100)

$$IC(0{,}95{,}1-0{,}98)=\left(0{,}95-2{,}33\sqrt{\tfrac{(0{,}95)(0{,}05)}{100}},0{,}95+2{,}33\sqrt{\tfrac{(0{,}95)(0{,}05)}{100}}\right),$$

$$IC(0{,}95{,}1-0{,}98)=((0{,}95)-(2{,}33)(0{,}021794)\ ,\ (0{,}95)+(2{,}33)(0{,}021794)),$$

$$IC(0{,}95{,}1-0{,}98)=\big((0{,}95)-(0{,}050781)\ ,\ (0{,}95)+(0{,}050781)\big),$$

$$IC(0{,}95{,}1-0{,}98)=\big((0{,}95)\pm(0{,}050781)\big),$$

- Army Brat (n=33)

$$IC(0{,}9697{,}1-0{,}98)=\left(0{,}9697-2{,}33\sqrt{\tfrac{(0{,}9697)(0{,}0303)}{33}},0{,}9697+2{,}33\sqrt{\tfrac{(0{,}9697)(0{,}0303)}{33}}\right),$$

$$IC(0{,}9697{,}1-0{,}98)=((0{,}9697)-(2{,}33)(0{,}029838)\ ,\ (0{,}9697)+(2{,}33)(0{,}029838)),$$

$$IC(0{,}9697{,}1-0{,}98)=\big((0{,}9697)-(0{,}069525)\ ,\ (0{,}9697)+(0{,}069525)\big),$$

$$IC(0{,}9697{,}1-0{,}98)=\big((0{,}9697)\pm(0{,}069525)\big)$$

Por fim, foram selecionados 3 subgrupos contendo mesma quantidade de locutores para comparação de desempenho entre si, visando verificar qual o efeito da diversificação de regionalidade de locutores. Sendo estes compostos por um subgrupo contendo de 100 locutores de diferentes regiões, escolhidos de forma aleatória, e outros 2 de mesma quantidade das regiões *South Midland* e *Western*. O grupo composto por diversos locutores possuía o seguinte intervalo de confiança:

- Locutores de diversas regiões (n=100)

$$IC(0{,}93{,}1-0{,}98) = \left(0{,}93 - 2{,}33\sqrt{\frac{(0{,}93)(0{,}07)}{100}}, 0{,}93 + 2{,}33\sqrt{\frac{(0{,}93)(0{,}07)}{100}}\right),$$

$$IC(0{,}93{,}1-0{,}98) = ((0{,}93) - (2{,}33)(0{,}025515)\ ,\ (0{,}93) + (2{,}33)(0{,}025515)),$$

$$IC(0{,}93{,}1-0{,}98) = \big((0{,}93) - (0{,}059449)\ ,\ (0{,}93) + (0{,}059449)\big),$$

$$IC(0{,}93{,}1-0{,}98) = \big((0{,}93) \pm (0{,}059449)\big)$$

Em seguida aplicou-se o Intervalo de Confiança para Diferença entre Duas Proporções entre o subgrupo contendo os locutores de diversa regiões e os subgrupos *South Midland* e *Western*, de forma individual.

- Comparação entre os testes do subgrupo de diferentes regiões e *South Midland*

$$EP = (\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

$$EP = (0{,}93 - 0{,}94) \pm 2{,}33\sqrt{\frac{(0{,}93)(0{,}07)}{100} + \frac{(0{,}94)(0{,}06)}{100}},$$

$$EP = 0{,}01 \pm 2{,}33\sqrt{(0{,}000651) + (0{,}0000564)},$$

$$EP = 0{,}01 \pm (2{,}33)(0{,}0255147016443461),$$

$$EP = 0{,}01 \pm 0{,}056,$$

$$EP = [-0{,}055; 0{,}057]$$

- Comparação entre os testes do subgrupo de diferentes regiões e *Western*

$$EP = (\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

$$EP = (0{,}93 - 0{,}95) \pm 2{,}33\sqrt{\frac{(0{,}93)(0{,}07)}{100} + \frac{(0{,}95)(0{,}05)}{100}},$$

$$EP = 0{,}02 \pm 2{,}33\sqrt{(0{,}000651) + (0{,}0000475)},$$

$$EP = 0{,}02 \pm (2{,}33)(0{,}026429150572805),$$

$$EP = 0{,}02 \pm 0{,}062,$$

$$EP = [-0{,}060; 0{,}064]$$

Resultaram os intervalos de erro padrão $[-0{,}055; 0{,}057]$ e $[-0{,}060; 0{,}064]$ para *South Midland* e *Western*, respectivamente. A presença do 0 (zero) em ambos os intervalos mostra que é plausível a igualdade das proporções.

B.3 Grupo Experimental 2: Identificação Semiautomática de Locutor

De acordo com as conclusões obtidas no Grupo Experimental 1, optou-se por adotar o uso de testes longos com a divisão treinamento/teste 70% / 30% para a realização dos experimentos deste grupo.

Visando observar os resultados para um sistema semiautomático de identificação de locutores, foi utilizado o modelo de tomada de decisão que retorna os n mais prováveis resultados de um universo de S locutores, de acordo com uma variação do classificador simples de máxima verossimilhança descrito na Equação 2.57. De posse desta proporção e quantidade de testes realizados, utilizando um nível de confiança $\alpha = 98\%$, foi possível chegar ao intervalo de confiança de proporções dos experimentos realizados.

- *(n=10)*

$$IC(0{,}9683{,}1-0{,}98)=\left(0{,}9683-2{,}33\sqrt{\frac{(0{,}9683)(0{,}0317)}{630}},0{,}9683+2{,}33\sqrt{\frac{(0{,}9683)(0{,}0317)}{630}}\right),$$

$$IC(0{,}9683{,}1-0{,}98)=((0{,}9683)-(2{,}33)(0{,}006980)\ ,\ (0{,}9683)+(2{,}33)(0{,}006980)),$$

$$IC(0{,}9683{,}1-0{,}98)=\big((0{,}9683)-(0{,}016264)\ ,\ (0{,}9683)+(0{,}016264)\big),$$

$$IC(0{,}9683{,}1-0{,}98)=\big((0{,}9683)\pm(0{,}016264)\big)$$

- *(n=20)*

$$IC(0{,}9857{,}1-0{,}98)=\left(0{,}9857-2{,}33\sqrt{\frac{(0{,}9857)(0{,}0143)}{630}},0{,}9857+2{,}33\sqrt{\frac{(0{,}9857)(0{,}0143)}{630}}\right),$$

$$IC(0{,}9857{,}1-0{,}98)=((0{,}9857)-(2{,}33)(0{,}004730)\ ,\ (0{,}9857)+(2{,}33)(0{,}004730)),$$

$$IC(0{,}9857{,}1-0{,}98)=\big((0{,}9857)-(0{,}011021)\ ,\ (0{,}9857)+(0{,}011021)\big),$$

$$IC(0{,}9857{,}1-0{,}98)=\big((0{,}9857)\pm(0{,}011021)\big)$$

- *(n=30)*

$$IC(0{,}9905{,}1-0{,}98)=\left(0{,}9905-2{,}33\sqrt{\frac{(0{,}9905)(0{,}0095)}{630}},0{,}9905+2{,}33\sqrt{\frac{(0{,}9905)(0{,}0095)}{630}}\right),$$

$$IC(0{,}9905{,}1-0{,}98)=((0{,}9905)-(2{,}33)(0{,}003865)\ ,\ (0{,}9905)+(2{,}33)(0{,}003865)),$$

$$IC(0{,}9905{,}1-0{,}98)=\big((0{,}9905)-(0{,}009005)\ ,\ (0{,}9905)+(0{,}009005)\big),$$

$$IC(0{,}9905{,}1-0{,}98)=\big((0{,}9905)\pm(0{,}009005)\big)$$