UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE ENGENHARIA MECÂNICA
PROGRAMA DE PÓS-GRADUAÇÃO
MESTRADO ACADÊMICO EM ENGENHARIA MECÂNICA

MATEUS LICHFETT MACHADO

IMPLEMENTAÇÃO DE UM SISTEMA DE RECONHECIMENTO AUTOMÁTICO DE VOZ UTILIZANDO AS TÉCNICAS MFCC E QUANTIZAÇÃO VETORIAL COM ATRIBUTOS DINÂMICOS, DE NORMALIZAÇÃO E DETECÇÃO DE VOZ ATIVA

UBERLÂNDIA
2016

MATEUS LICHFETT MACHADO

IMPLEMENTAÇÃO DE UM SISTEMA DE RECONHECIMENTO AUTOMÁTICO DE VOZ UTILIZANDO AS TÉCNICAS MFCC E QUANTIZAÇÃO VETORIAL COM ATRIBUTOS DINÂMICOS, DE NORMALIZAÇÃO E DETECÇÃO DE VOZ ATIVA

Dissertação de Mestrado apresentada ao Curso de Pós-Graduação em Engenharia Mecânica da Universidade Federal de Uberlândia, como requisito parcial para a obtenção do título de Mestre em Engenharia Mecânica.

Orientador: Prof. Dr. Marcus Antonio Viana Duarte

UBERLÂNDIA

2016

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

M149i 2016

Machado, Mateus Lichfett, 1989
Implementação de um sistema de reconhecimento automático de voz utilizando as técnicas MFCC e quantização vetorial com atributos dinâmicos, de normalização e detecção de voz ativa / Mateus Lichfett Machado. - 2016.
148 f. : il.

Orientador: Marcus Antonio Viana Duarte.
Dissertação (mestrado) - Universidade Federal de Uberlândia, Programa de Pós-Graduação em Engenharia Mecânica.
Disponível em: http://dx.doi.org/10.14393/ufu.di.2018.82
Inclui bibliografia.

1. Engenharia Mecânica - Teses. 2. Reconhecimento automático da voz - Teses. 3. Normalização - Teses. 4. Sistemas de processamento de voz - Teses. I. Duarte, Marcus Antonio Viana. II. Universidade Federal de Uberlândia. Programa de Pós-Graduação em Engenharia Mecânica. III. Título.

CDU: 621

Marcello Mundim Rodrigues – CRB-6/3275

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE ENGENHARIA MECÂNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA MECÂNICA

ALUNO: Mateus Lichfett Machado

NÚMERO DE MATRÍCULA: 11412EMC009

ÁREA DE CONCENTRAÇÃO: Mecânica dos Sólidos e Vibrações

LINHA DE PESQUISA: Dinâmica de Sistemas Mecânicos

PÓS-GRADUAÇÃO EM ENGENHARIA MECÂNICA: NÍVEL MESTRADO

TÍTULO DA DISSERTAÇÃO:

"Implementação de um Sistema Robusto de Reconhecimento Automático de Voz Utilizando as Técnicas MFCC e VQ com Atributos Dinâmicos, de Normalização e Detecção de Voz Ativa"

ORIENTADOR: Prof. Dr. Marcus Antonio Viana Duarte

A Dissertação foi **APROVADA** em reunião pública, realizada no Anfiteatro B do Bloco 5 O, Campus Santa Mônica, em 18 de abril de 2016, às 14:00 horas, com a seguinte Banca Examinadora:

NOME	ASSINATURA
Prof. Dr. Marcus Antonio Viana Duarte (orientador)	UFU
Prof. Dr. Elias Bitencourt Teodoro	UFU
Prof. Dr. Sérgio Lima Netto	UFRJ

Uberlândia, 18 de abril de 2016

Dedico este trabalho aos meus pais,
Álisson e Roseli,
aos meus irmãos, Tomás e Vitor e;
Ao meu amor, Caroline
Amo vocês!

AGRADECIMENTOS

Primeiramente gostaria de agradecer os meus pais pelo enorme incentivo para o cumprimento dessa trabalhosa etapa, pela educação concedida ao longo dos anos, pelos conselhos, valores e sabedoria transmitidos. Não há palavras para mensurar o quão grato sou e o quanto são importantes em minha vida. Obrigado por sempre acreditarem em mim.

Aos meus irmãos, Tomás e Vitor, pelas contribuições diárias, pela força, companheirismo e apoio incondicional durante esta jornada.

Agradeço prestigiosamente à minha namorada, Caroline, razão da minha inspiração. Agradeço pelo companheirismo, pelas palavras de carinho e de encorajamento, pela presença constante em minha vida, pelo amparo e amor.

Agradeço ao meu orientador, Dr. Marcus Antonio Viana Duarte pelos conselhos, apoio, incentivo, coragem, entusiasmo e sobretudo sabedoria dispensada para o desenvolvimento desta pesquisa.

Aos colegas do Laboratório de Acústica e Vibrações (LAV) pelas diversas contribuições e conselhos.

Agradeço também a Universidade Federal de Uberlândia, a Faculdade de Engenharia Mecânica e ao Programa de Pós-Graduação pela oportunidade de realizar esse trabalho.

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo fomento à pesquisa científica e pelo apoio financeiro.

“I must create a system, or be enslaved by another man's. I will not reason and compare, my business is to create.”

(William Blake)

RESUMO

A presente pesquisa pauta-se na investigação e elaboração de um sistema de reconhecimento automático de voz robusto. Para tanto, utilizou-se *Mel Frequency Cepstral Coefficients* (MFCC) como técnica para extração das propriedades acústicas de sinais de voz e Quantização Vetorial (VQ) para classificação e reconhecimento de padrões. Foram também incorporadas ferramentas dinâmicas, de normalização e detecção de voz ativa com intuito de aperfeiçoar o sistema. Testaram-se dois tipos de coeficientes dinâmicos: *Delta-delta Coefficients* (DDC) e *Shifted-Delta Coefficients* (SDC); três tipos de ferramentas para normalização dos vetores cepstrais: *Cepstral Mean and Variance Normalization* (CMVN), *Windowed Cepstral Mean and Variance Normalization* (WCMVN) e *Short-Time Gaussianization* (STG); além da técnica de detecção de voz ativa: *Voice Activity Detection* (VAD), que fora implementada segundo o algoritmo desenvolvido por Qiang He, combinando as metodologias *Short-Time Energy* (STE) e *Zero Crossing Rate* (ZCR). A pesquisa realizada analisa a capacidade do sistema desenvolvido em operar segundo uma pluralidade de tarefas: reconhecer palavras ou comandos; identificar o locutor; e a combinação das duas primeiras. Além disso, investigou-se qual a melhor combinação, dentre as técnicas e atributos em escopo, para realização das tarefas citadas analisando a eficiência do sistema. Foram realizados cinco experimentos em ambiente de ruído controlado, dos quais participaram oito indivíduos. Destes, quatro tiveram suas vozes treinadas para criação de bancos de dados, e os demais participaram da fase de testes com os primeiros. Foram captadas ao total 144 amostras para realização do experimento. Destas, 24 foram utilizadas para construção de bancos de dados e 120 utilizadas durante a fase de testes. Para garantir a integridade dos experimentos, as amostras de treinamento e testes foram espelhadas para serem processadas segundo a configuração de cada experimento. Os resultados obtidos aprovaram o uso destas técnicas como ferramentas aptas à execução das tarefas para o qual o sistema fora proposto e apontaram a melhor configuração como combinação das MFCC e VQ, os atributos VAD, *Shifted-Delta Coefficients* e a ferramenta de normalização *Short-Time Gaussianization*.

Palavras chave: Reconhecimento Automático de Voz, *Mel Frequency Cepstral Coefficients*, Quantização Vetorial.

ABSTRACT

The present research investigates and elaborates an automatic and robust voice recognition based system using Mel Frequency Cepstral Coefficients (MFCC) as a technique for extracting the acoustic properties of speech signals and Vector Quantization (VQ) for classification and pattern recognition. Combined to these techniques it was added dynamic tools, normalization techniques and active voice detection in order to improve the system. Two dynamic coefficients were tested: Delta-Delta Coefficients (DDC) and Shifted Delta-Coefficients (SDC); as well as three different normalization techniques: Cepstral Mean and Variance Normalization (CMVN), Windowed Cepstral Mean and Variance Normalization (WCMVN), and Short-Time Gaussianization (STG); and also the Voice Activity Detection (VAD) tool, which was implemented according to the algorithm developed by Qiang He, combining the Short-Time Energy (STE) and Zero Crossing Rate (ZCR) methodologies. The research examines the ability of the designed system to operate according to a plurality of tasks: recognition of words or commands; speaker identification; and the combination of the two first tasks. In addition, the research investigates the best configuration of the system among the tested techniques for performing the tasks mentioned, analyzing its efficiency. Five experiments were conducted in a noise controlled environment, with the participation of eight persons. Four of them had their voices trained to create databases, and the others participated only in the test phase together with the ones that had trained the system. It was captured 144 speech samples for the experiments, 24 of them were used for building the database and the 120 others used during the test phase. To ensure the integrity of the experiments, the training and the testing samples were mirrored to be processed according to the configuration of each experiment. The use of these techniques was aprooved as tools capable of performing the tasks for which the system was proposed and the best configuration found was the combination of the MFCC and VQ techniques with VAD, *Shifted-Delta Coefficients* and the Short-Time Gaussianization normalization technique.

Keywords: Automatic Speech Recognition, Mel Frequency Cepstral Coefficients, Vector Quantization

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ADC	Conversor analógico-digital
AM	Modelo acústico
ANN	Redes Neurais Artificiais
ARPA	*Advanced Research Projects Agency*
ATIS	*Air-Travel Information Systems*
CDF	*Cumulative Distribution Function*
CMN	*Cepstral Mean Normalization*
CMU	Carnegie Mellon University
CMVN	*Cepstral Mean and Variance Normalization*
DC	*Delta-coefficients*
DCT	Transformada Discreta do Cosseno
DDC	*Delta-delta Coefficients*
DFT	Transformada Discreta de Fourier
DLSF	*Delta Line Spectral Frequencies*
DNN	*Deep Neural Networks*
DTW	*Dynamic Type Warping*
FFT	Transformada Rápida de Fourier
FIR	*Finite Impulse Response*
GUI	*Graphical User Interface*
HHT	*Hilbert-Huang Transform*
HMM	*Hidden Markov Models*
ICA	*Independent Component Analysis*
LM	Modelo Linguístico
MFCC	*Mel Frequency Cepstral Coefficients*
ML-EM	*Maximum-Likelihood Expectation-Maximization*
MIT	*Massachusetts Institute of Technology*
NIST	*National Institute of Standards and Technology*
PLP	*Perceptual Linear Prediction*
RBF	*Radial-Basis Function*
SDC	*Shifted Delta Coefficients*
SOM	*Self-Organizing Map*
STE	*Short-Time Energy*

STG	*Short-term Gaussianization*
STMSN	*Short-time Mean and Scale Normalization*
STMVN	*Short-time Mean and Variance Normalization*
TI	*Texas Instruments*
VAD	*Voice Activity Detection*
VQ	Quantização Vetorial
WCMVN	*Windowed Cepstral Mean and Variance Normalization*
ZCR	*Zero Crossing Rate*

LISTA DE VARIÁVEIS

A	Amplitude do sinal de fala
Cd	*Codeword*
$\widetilde{c}_n$	*Mel Frequency Cepstral Coeficients*
d	Diferencial temporal
d_i	*Delta-coefficient* do *i-ésimo frame*
D_l	Distância média entre um vetor acústico e um centróide
E	Energia transportada pela onda sonora
f	Frequência [Hz]
I	Número de *frames* do sinal
i	Número inteiro indicador do *frame* do sinal
K	Número de coeficientes
k	número de blocos para concatenar *Delta-coefficients*
l	l *-ésimo* centróide
L	Limiar de Distorção VQ
M	Número total de centróides
m	Número inteiro que representa onúmero de coeficientes cepstrais por *frame*
mel	Escala Mel
N	Número de amostras por *frame*
n	Número inteiro indicador de amostras do sinal
$N\widetilde{c}_{n\,i}$	Vetor de coeficientes cepstrais do *i-ésimo frame*
p	Diferença temporal entre blocos consecutivos
S	Sinal analógico
$\widetilde{S}_k$	Coeficientes de espectro de potência Mel
T	Período de amostragem
t	Tempo do sinal de fala [s]
w	Função janela
x	Sequência numérica de sinais temporalmente discretos
X_i	*i-ésimo frame* do sinal transformado para o domínio da frequência
$\widetilde{Y}$	Sinal de voz discretizado e com janelamento

y_i	Janelamento do sinaldiscreto por *frame*
ZCR	*Zero Crossing Rate*
ε	Parâmetro de repartição do algoritmo LBG
σ	Desvio padrão dos coeficientes cepstrais
σ^2	Variância dos coeficientes cepstrais
μ	Média dos coeficientes cepstrais

SUMÁRIO

1 INTRODUÇÃO

Reconhecimento automático de voz tem sido um campo atrativo de pesquisas há mais de cinco décadas, e é considerado uma importante ponte nas relações de interação *homem-homem* e *homem-máquina*. Avanços tecnológicos emergentes e consolidados desenvolvidos nas áreas da computação, metalurgia, mecânica e eletroeletrônica em importantes laboratórios de empresas multinacionais como IBM, Microsoft, Google, Mercedes-Benz e Laboratórios Bell, além de conceituadas universidades como a Carnegie Mellon University, Oxford University e University of Texas, possibilitaram o aprimoramento e pluralização de atividades realizadas utilizando-se a voz como comando de acionamento de mecanismos diversos (GOLD et al., 2011).

Antes da década de 70 a fala não era considerada uma modalidade importante, tratando-se de meios de comunicação *homem-máquina*. Esta decorrência pode ser justificada pela carência de tecnologias da época, incapaz de processar elevadas quantidades de dados em poucos segundos. Além disso, outros mecanismos de iteração com máquinas, como a utilização de controles remotos, teclados e *mouse*, por exemplo, apresentavam desempenhos superiores à fala em termos de eficiência de comunicação, ou eram preferíveis em algumas situações.

O emprego globalizado desta insigne ferramenta em atividades modernas foi fomentado principalmente pela demanda de velocidade, conectividade e melhoramento de técnicas de interação social compelidos pela sociedade contemporânea.

Segundo os renomados pesquisadores da Microsoft, Xuedong Huang, Alex Acero, e Hsiao-Wuen Hon, muitos avanços estão por vir, tratando-se de ferramentas de reconhecimento de voz. Eles consideram que os estágios atuais da empregabilidade de sistemas de reconhecimento de voz, técnicas, modelos e ferramentas estão ainda em um nível prematuro de desenvolvimento. Isso se torna mais evidente, analisando as recentes descobertas, ainda em escala de projeção comercial no ramo da ciência computacional, que prometem revolucionar este setor, como a computação quântica (desenvolvida pela IBM); construção de nano-processadores programáveis (por pesquisadores da Harvard University e MITRE Corporation); introdução de estruturas do tipo *Deep Learning* em arquiteturas de redes neurais (pelo pesquisador Igor Aizenberg, professor da Texas A&M University); e fabricação de chips e microprocessadores com materiais semicondutores orgânicos.

A diversidade de aplicações de sistemas de reconhecimento de voz e a gama de benefícios que possibilitam, assim como, agilizando, integrando e facilitando a rotina diária das

pessoas, são seus grandes trunfos. Além disso, esses sistemas podem ser instrumentos facilitadores do dia-a-dia de pessoas com deficiência física. Seu uso propicia ambientes seguros e permitem comodidade a esses usuários, ensejando-os realizar tarefas cotidianas sem necessidade de auxílio (caso de pessoas com tetraplegia, por exemplo).

A inclusão desta tecnologia está presente, por exemplo, em sistemas de roteamento ativado por voz; discagem por voz em telefones móveis; pesquisas por voz em endereços, portais eletrônicos e *menus*; sistemas de edição de texto; acionamento de alarmes; exames biométricos para identificação de indivíduos; sistemas de acessibilidade de informações restritas; navegadores de bordo; sistemas de direção guiada por satélites; sistemas de tradução automática; e, em aplicações no ramo da domótica.

Sistemas de reconhecimento de voz são empregados na área de segurança, garantindo elevado grau de eficiência. Isto pode ser explicado através de um simples fenômeno físico: vozes diferentes apresentam características acústicas diferentes, e por consequência soam de maneira distinta. Esta singularidade é decorrente dos diferentes formatos entre os aparelhos de produção da fala humana (cavidade oral e nasal, traqueia, entre outros) e da variação na quantidade de pressão utilizada para enunciação de palavras. Fonemas, palavras, sentenças e mesmo locutores podem ser identificados através do processamento e análise de correspondência dos sinais de fala. Além disso, outros fatores favorecem o uso destes sistemas: são fáceis de serem implementados, são sistemas de baixo custo, e apresentam elevado grau de eficiência.

Todavia, a fala humana, assim como sua interpretação são fenômenos de elevada complexidade e importância, o que condiciona a viabilidade desta tecnologia a vários fatores, como: desempenho do sistema de captação do sinal de fala; capacidade do sistema de segregação do sinal de fala de ruídos externos; capacidade do sistema de segregação de duas ou mais vozes misturadas em um único sinal; desempenho do sistema digitalizador; desempenho das técnicas de extração das propriedades acústicas dos sinais de fala; desempenho do modelo classificador e de reconhecimento de padrões; distâncias relevantes entre o sistema de captação do sinal de voz e do locutor e efetividade de técnicas dinâmicas, de normalização e detecção de voz ativa. Sistemas adequados aos critérios citados podem apresentar até 99% de efetividade.

O reconhecimento de sinais de voz consiste em um processo complexo e sequencial. Nesta pesquisa, empregou-se um sistema composto pelas técnicas *Mel Frequency Cepstral Coefficients* (MFCC) e Quantização Vetorial (VQ). Primeiramente, é necessário converter o sinal de fala de sua forma de onda em um sinal digitalizado e então extrair as características acústicas úteis que serão agrupadas em uma matriz composta por uma série de '*vetores*

acústicos'. A matriz é lançada em um espaço vetorial, do qual ocupará pequenas regiões denominadas *clusters*. Estas regiões designarão espaços particulares relacionados a cada palavra, sentença ou locutor, servindo de pontos de referência para o efetivo reconhecimento da fala através de análises de correlações com novas matrizes lançadas no mesmo espaço vetorial.

A técnica MFCC apresenta algumas vantagens que favorecem sua aplicação para a tarefa de extrair atributos acústicos dos sinais de fala. Essa técnica tem a capacidade de capturar as relações de tempo, frequência e energia em um conjunto de coeficientes chamados coeficientes cepstrais. Utilizando uma escala chamada Mel, os coeficientes são calculados em faixas de frequências logarítmicas e correspondentes àquelas dos tons puros percebidos pelo sistema auditivo humano. Os coeficientes formulados carregam consigo informações acústicas úteis de seus respectivos sinais de voz. Estes são agrupados para formarem os '*vetores acústicos*' que serão lançados em um espaço vetorial para comparação, usando VQ.

A técnica Quantização Vetorial é utilizada inicialmente para a construção de um banco de dados. Trata-se de um espaço vetorial constituído de '*vetores acústicos*' de treinamento da voz. Assim, é possível mensurar distâncias Euclidianas entre novas matrizes de vetores acústicos lançadas no mesmo espaço vetorial e as regiões designadas para cada vetor de treinamento. Essa distâncias determinarão o efetivo reconhecimento de padrões.

O objetivo geral da presente pesquisa contempla a análise da aplicabilidade da voz como instrumento de acionamento mecânico para atividades cotidianas diversas, a partir do emprego das técnicas *Mel Frequency Cepstral Coefficients* (MFCC) utilizadas para extração das propriedades acústicas dos sinais de fala e Quantização Vetorial (VQ) empregadas para classificação e reconhecimento de padrões. Para tanto, implementou-se um sistema de reconhecimento automático de voz destinado à realização de múltiplas tarefas: reconhecimento de comandos ou palavras; identificação do locutor; e a combinação das duas primeiras. Experimentou-se ainda a inserção de atributos dinâmicos, de normalização cepstral e detecção de voz ativa para aprimoramento do sistema de reconhecimento de voz.

Os atributos dinâmicos estudados foram Delta-delta Coefficients (DDC) e Shifted-Delta Coefficients (SDC); as técnicas de normalização dos vetores cepstrais foram Cepstral Mean and Variance Normalization (CMVN), Windowed Cepstral Mean and Variance Normalization (WCMNV) e Short-Time Gaussianization (STG); e usou-se ainda uma ferramenta de detecção de voz ativa: Voice Activity Detection (VAD), implementada segundo o algoritmo desenvolvido por Qiang He, constando de duas metodologias: Short-Time Energy (STE) e Zero Crossing Rate (ZCR).

Verificada a eficácia das técnicas elegidas neste trabalho, o estudo pautou-se especificamente na investigação da capacidade do sistema em reconhecer padrões de fala simples e identificar locutores quando adicionados ao sistema atributos dinâmicos, de normalização e de detecção de voz ativa. Por fim, apurou-se qual a melhor configuração do sistema para o aprimoramento das capacidades supracitadas.

Para tanto, essa Dissertação foi segmentada em outros quatro capítulos: Fundamentos Teóricos, Capítulo 2; Arquitetura de Sistemas de Reconhecimento de Voz, Capítulo 3; e Experimentos e Análise de Resultados, Capítulo 4 e as conclusões, Capítulo 5.

No capítulo 2, serão introduzidos os principais conceitos associados às tecnologias da voz com analises do funcionamento do sistema de reprodução da fala, bem como do sistema auditivo humano, correlacionando-os com modelos de reconhecimento de voz e discutindo as nuances entre as fases de treinamentos e testes. Ainda neste capítulo serão abordados conceitos de processamento de sinais, pré-ênfase e os principais modelos extratores de propriedades acústicas, técnicas de classificação e reconhecimento de padrões. No Capítulo 3, será apresentado as metodologias que governam as técnicas de extração das características de sinal (MFCC) e reconhecimento de padrões (VQ), bem como dos atributos DDC, SDC, CMVN, WCMVN, STG e VAD.

No Capítulo 4, será apresentada a interface gráfica desenvolvida para recolhimento e avaliação de amostras de sinais de fala, além da descrição dos experimentos realizados, detalhando as condições de ensaio. Também, serão apresentados os resultados dos testes realizados.

Finalmente, no Capítulo 5, serão apresentadas as conclusões acerca dos resultados obtidos através dos experimentos com observações ponderadas.

2 FUNDAMENTOS TEÓRICOS

A palavra comunicação deriva da palavra *'communis'* do Latim e significa compartilhar (ROSENGREN, 2000). Trata-se de uma ação em que um indivíduo transmite informações através da troca de mensagens que podem ser escritas, faladas ou gesticuladas.

Historiadores acreditam que uma das primeiras formas de comunicação entre nossos ancestrais ocorreu na África através da fala, há cerca de 350-150 mil anos atrás (ATKINSON, 2011). Com a evolução do cérebro humano, diversos tipos de artes, símbolos, escritas e dialetos foram criados. Este último possibilitou maior velocidade e facilidade de compreensão das informações que eram trocadas entre indivíduos da mesma civilização, promovendo desenvolvimento acelerado da mesma, principalmente no âmbito agrícola, nas caçadas mais complexas, na construção de abrigos e na forma de interação social (DRAKE, 2015).

O desenvolvimento de métodos de comunicação jamais interrompeu-se desde então. A comunicação através da fala foi, e continua sendo, um dos modos dominantes de conexão social e de troca de informações da humanidade. Dos dialetos antigos, novas linguagens surgiram, assim como novas formas de comunicação. As mediações tecnológicas como a telefonia, o rádio, a televisão, o cinema e o advento da era dos computadores e da internet promoveram o rompimento de barreiras entre fronteiras, o que impulsionou o fenômeno da globalização, aumentando drasticamente a troca de informações entre nações. Proporcionou ainda a otimização do setor industrial e fomentou a troca de conhecimentos e culturas. Métodos arcaicos, como as cartas escritas em papel e tinta tornaram-se obsoletos.

A sociedade moderna tem todas as suas relações e conexões sociais ancoradas nos meios de comunicação e exige destes meios cada vez mais velocidade, facilidade de acesso, e interatividade com pessoas nas mais diversas regiões do globo. Portanto, a relação entre comunicação e sociedade atingiu níveis extraordinários de desenvolvimento.

Neste contexto, a notável preferência e usabilidade da comunicação através da fala enraizada na sociedade moderna começa a ser diversificada nas relações *homem-máquina*. A maioria dos computadores atualmente utiliza interfaces gráficas (*Graphical User Interface,* GUI) baseadas em representações gráficas de objetos e funções como: janelas, ícones, menus, barras de comando e de acessibilidade e ponteiros. A maior parte dos sistemas operacionais de computadores ainda dependem de mecanismos externos como o *mouse*, o teclado e monitores para efeito retroativo. Os computadores modernos têm carência de habilidades humanas

fundamentais: a capacidade de fala, de audição e interpretação (HUANG et al., 2001). Essa exiguidade, no entanto, tende a ser modificada.

Sistemas inteligentes com capacidade de resposta automática para comandos através da voz já estão sendo estudados desde a década de 50 e implantados há mais de três décadas com resultados expressivos (GOLD et al., 2011). A julgar pela gama de investimentos em pesquisas neste campo de inteligência; pelo desenvolvimento de super processadores; pela criação, modificação e aprimoramento de métodos para reconhecimento de padrões e técnicas de extração de características acústicas de sinais de fala; além da pluralidade de tarefas que podem ser executadas através do princípio de sistems de reconhecimento de voz, é natural a predição de que, em um futuro próximo, o mecanismo da fala será um dos meios primários de comunicação *homem-máquina* (McLOUGHLIN, 2009). No entanto, ainda há alguns obstáculos.

A fala é um sinal complexo produzido como resultado de várias transformações em vários níveis diferentes: semânticos, linguísticos, articulatórios e acústicos. Diferenças nestes fundamentos aparecem como diferenças nas propriedades acústicas de cada sinal. A combinação das diferenças anatômicas dos aparelhos de reprodução da fala humana, os hábitos de dicção e sotaque, e a quantidade de pressão utilizada para expelir o ar durante a locução são fatores que permitem a distinção entre os indivíduos locutores, carregando diferentes propriedades acústicas quando palavras ou uma frases são enunciadas (GOLD et al., 2011). A complexidade envolvida na tarefa de extrair as propriedades acústicas, além da eficiente capacidade de relacionar tais propriedades com outras previamente obtidas e salvas em um banco de dados representam as dificuldades no projeto destes sistemas.

O desenvolvimento da arquitetura dos sistemas de reconhecimento de voz foi inspirado no próprio sistema humano, ou seja, no sistema auditivo para interpretação do sinal de entrada e no aparelho de reprodução da fala para reprodução de vozes sintetizadas (*text-to-speech*). Portanto, o entendimento da funcionalidade do sistema auditivo humano, bem como do aparelho de reprodução da fala são pontos importantes ao estudar estes sistemas.

Embora haja características espectrais e temporais semelhantes em sons produzidos através de meios diversos, sinais de fala apresentam peculiaridades que permitem sua diferenciação dos demais sinais. Trata-se de um conjunto estruturado de sons contínuos que apresenta conjuntos espectrais variantes de 20 Hz a 20 KHz. Sons '*vozeados*' ou nasais (vogais e as consoantes *j*, *l*, *m*) têm um espectro discreto com uma frequência funamental variante de 100 Hz a 200 Hz para homens e de 200 Hz a 400 Hz para mulheres. Sons não '*vozeados*' (consoates *f*, *s*, *p*, e o dígrafo *ch*) são gerados pelo fluxo de ar na boca modulados pelos

maxilares, língua e lábios, e, apresentam espectros contínuos (GOLD et al., 2011). Assim, é necessário estudar quais distorções afetam a inteligibilidade e a qualidade da voz no sinal captado, para execução de técnicas capazes de realizar eficazmente a interpretação de que se trata de um sinal de voz, bem como, avaliar, em alguns casos, o que fora enunciado e além disto, separar o sinal correspondente à voz enunciada de ruídos presentes durante a captação do mesmo.

2.1 Produção e Percepção da fala humana

Como qualquer outro som da natureza, a onda sonora produzida através da fala apresenta características sob as quais condicionamos nossos cérebros a captar e interpretar. Esta importante capacidade permite-nos distinguir a origem e o significado de diferentes tipos de sons. O sistema humano com relação ao reconhecimento e interpretação de fala é considerado por pesquisadores o único exemplo de um sistema robusto existente, cuja performance é insensitiva a variações, sob a ótica de fatores não-linguísticos, no sinal de fala (McLOUGHLIN, 2009). Os mecanismos dos sistemas de reconhecimento de voz são, portanto, construídos de maneira a mimicar o aparelho de reprodução e percepção da fala humano.

2.1.1 Produção da Fala

O processo de produção da fala inicia-se com uma mensagem semântica em nosso pensamento. Após a criação da mensagem, o passo seguinte é convertê-la em uma sequência organizada de palavras. Cada palavra consiste em uma sequência de fonemas que designarão a maneira como estas serão pronunciadas. Na condição da mensagem ser um conjunto organizado de palavras formando uma ou mais sentenças, cada sentença apresentará padrões prosódicos definindo a duração da enunciação de cada fonema, a entonação da sentença e a intensidade sonora (HUANG et al., 2001). Uma vez que o sistema linguístico finaliza o mapeamento destas características, uma série de sinais de excitações neuromusculares são executados para realizarem um mapeamento articulatório que controlará as cordas vocais, o movimento dos lábios, da mandíbula, da língua e do palato, além da contração dos pulmões para expelir o ar.

As propriedades acústicas particulares dos sinais de voz produzidos são dependentes de uma série de fatores, como a modulação da quantidade de potência e pressão de ar expelido pelos pulmões, a compressão da glote, a tensão nas cordas vocais, a anatomia do aparelho bucal e da disposição dos dentes e língua. A capacidade articulatória sincronizada dos órgãos responsáveis pela modulação da voz permite a locução dos fonemas (McLOUGHLIN, 2009),

que na língua portuguesa, como na maioria das línguas do mundo, são divididos entre vogais e consoantes, criando conjuntos de palavras que formam a mensagem a ser transmitida através da fala.

2.1.2 Percepção da Fala

O processo de percepção da fala é antagônico à produção da mesma. O ouvinte recebe a informação armazenada no sinal de fala na forma de onda sonora e processa a informação através de seu sistema neurológico, para que o cérebro interprete aquela mensagem.

Primeiramente, o sinal de pressão acústico é recebido pelo sistema auditivo periférico, qual seja o ouvido externo. Este comporta-se como uma antena, captando e transmitindo o sinal ao canal auditivo externo, por onde a onda sonora percorre. O comprimento deste canal é tal que funciona como um ressonador acústico, no qual o sinal oscila naturalmente com amplitudes maiores que as demais nas frequências de ressonância. Em seguida, a onda sonora chega ao tímpano, vibrando-o na mesma frequência que o sinal da onda de pressão sonora. As vibrações são transmitidas ao ouvido médio que se comunica com o ouvido interno. A estrutura principal do ouvido interno é a cóclea, que atua como um banco de filtros (HUANG et al., 2001). Os filtros próximos à base da cóclea representam as altas frequências e, os distantes, as baixas frequências. Estas informações são então moduladas em uma série de pulsos pelo nervo auditivo e transmitidas ao cérebro pelo sistema nervoso auditivo, responsável por extrair as informações úteis do sinal, permitindo a interpretação deste pelo nosso cérebro (GOLD et al., 2011).

2.2 A fala em sua forma de onda acústica

Diversos mecanismos de transmissão de informações tiveram importantes papéis em sistemas de comunicação sofisticados, no entanto, considera-se a fala em sua forma de onda acústica, ou algum modelo paramétrico da mesma, um dos métodos mais eficientes de comunicação para aplicações práticas já desenvolvidos (RABINER e SCHAFER, 1978).

Em vários sistemas de comunicação *homem-máquina*, a informação a ser transmitida está codificada na forma de um sinal continuamente variante que pode ser transmitido, gravado, manipulado e decodificado (RABINER e SCHAFER, 2007). No caso da fala, sua forma analógica fundamental é um sinal acústico na forma de onda sonora, conforme exemplificado pela Figura 2.1, o qual é convencionalmente chamado de sinal de voz.

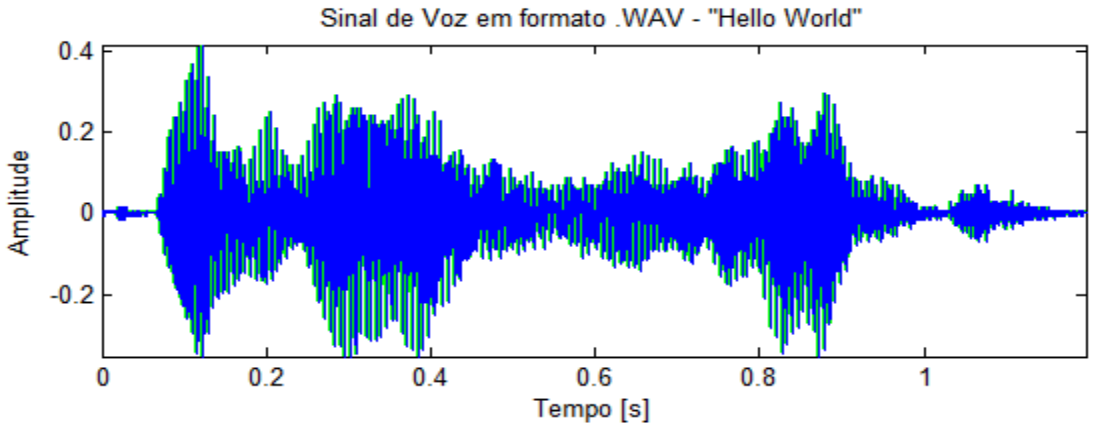

Figura 2.1 - Representação de um sinal de voz no formato *WAV*

O sinal de voz pode ser convertido para um sinal elétrico através de um microfone e, posteriormente, manipulado utilizando processamento de sinais analógicos e digitais. Esse sinal pode ser novamente convertido em uma forma acústica através de um alto-falante, por exemplo.

Esta forma de processamento do sinal de fala constituiu a base das ideias para a invenção do telefone em 1876 por Alexander Graham Bell, assim como dos atuais dispositivos de gravação, transmissão, produção e manipulação de sinais de áudio e, particularmente, de fala (RABINER e SCHAFER, 2007).

Sistemas de comunicação através da fala apresentam maneiras variadas de transmissão, armazenamento e processamento de dados, e devem seguir duas premissas: (i) serem capazes de preservar o conteúdo da mensagem do sinal de fala; (ii) representar adequadamente o sinal de fala em uma forma que seja conveniente para sua transmissão ou armazenamento sem que haja uma degradação do mesmo (RABINER e SCHAFER, 1978). Um diagrama representando a manipulação e processamento de sinais é mostrado na Figura 2.2.

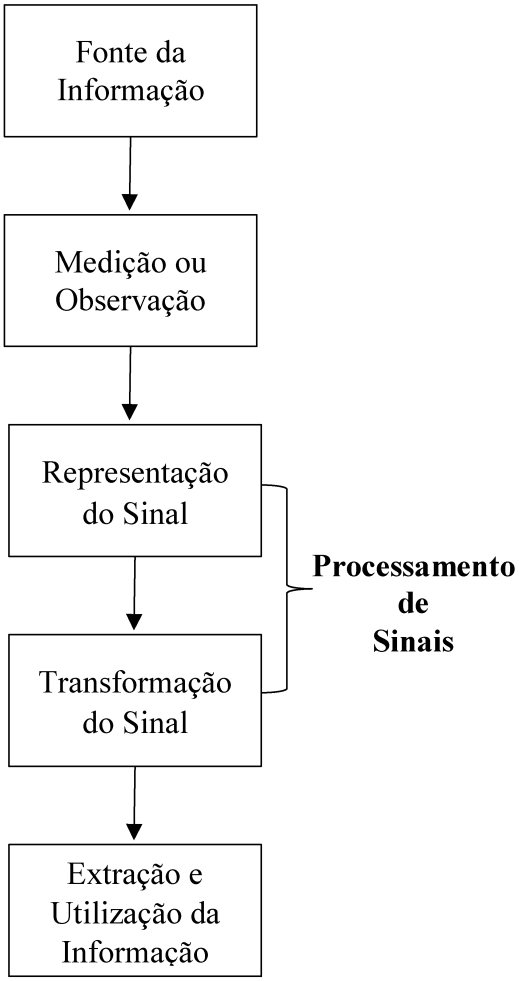

Figura 2.2 - Diagrama de manipulação e processamento de informação (adaptado de RABINER, L.; SCHAFER, R., 1978).

2.2.1 Processamento de Sinais Digitais

O termo sinal é comumente aplicado a algo que contém ou transmite informações acerca do estado ou comportamento de um sistema físico. Embora possa ser representado de diversas maneiras, em todos os casos o sinal contém a informação na forma de algum padrão que pode ser decodificado. Sinais são representados matematicamente como funções de uma ou mais variáveis independentes, que podem ser contínuas ou discretas (GOLD et al.; 2011). Particularmente, um sinal de fala digitalizado, por exemplo, é representado matematicamente em função do tempo de maneira discreta.

Sinais de tempo discretos são originados através da amostragem de um sinal de tempo contínuo ou diretamente por meio de um processo temporalmente discreto. Sua construção é bastante flexível, podendo ser gerados por uma variedade de tecnologias como: dispositivos de ondas acústicas superficiais, computadores digitais, e microprocessadores de alta velocidade

(OPPENHEIM e SCHAFER, 1999). O sinal de áudio digitalizado é uma sequência de amostras capturadas sob uma taxa de amostragem de frequência definida que conjuntamente representam som. A obtenção discreta do sinal em sua forma de onda acústica digitalizada é a primeira etapa do processo de digitalização de sinais de fala (OPPENHEIM e SCHAFER, 1999).

Sistemas temporais discretos podem ser usados para simular sistemas analógicos ou realizar transformações no sinal de funções que não podem ser implementadas com *hardwares* operantes com parâmetros temporais contínuos. A representação de sinais na forma discreta é frequentemente desejada, principalmente, ao se tratar de processamentos de sinais modernos.

Qualquer sinal limitado por banda pode ser representado por amostras tomadas periodicamente no tempo, desde que em uma amostragem elevada o suficiente, impedindo a degradação do sinal (OPPENHEIM e SCHAFER, 1999). O processo de amostragem embasa toda a teoria e aplicação de processamento de sinais digitalizados.

Em termos puramente físicos, o som é uma onda longitudinal que se propaga através do ar pelas vibrações entre moléculas. A variação de pressão da propagação do som determina sua frequência, e o grau desta variação determina a amplitude do sinal. A captação das ondas sonoras é realizada utilizando-se de um microfone, que é constituído por uma pequena membrana fina e sensível que deflete-se na ordem do estímulo provocado pelo som. Este estímulo é proporcionalmente convertido em voltagem ou corrente elétrica (McLOUGHLIN, 2009). O sinal elétrico é filtrado, amplificado e transformado em uma sequência de códigos digitais por meio de um conversor analógico-digital (ADC) e então processado, transmitido ou armazenado. Um procedimento inverso converte o sinal digital novamente em uma onda sonora utilizando um conversor digital-analógico (DAC) e um amplificador.

Sinais temporalmente discretos, lineares e invariantes são representados matematicamente como sequências numéricas, conforme apresentado na Eq. 2.1, em que x representa esta sequência numérica e n é um número inteiro.

$$x = \{x[n]\}, \qquad -\infty < n < \infty, \tag{2.1}$$

A sequência matemática dada pela Eq. (2.1) pode representar uma amostragem periódica de um sinal analógico, em que o *enésimo* valor desta sequência é igual ao valor do sinal analógico, $S(t)$, no tempo nT, com T sendo o período de amostragem.

$$x[n] = S(nT), \quad -\infty < n < \infty, \tag{2.2}$$

Sinais temporalmente discretos podem ser representados graficamente como mostra a Figura 2.3. Apesar da abscissa estar disposta como uma linha contínua, deve-se entender que a sequência $x[n]$ está definida apenas para valores inteiros de n, sendo indefinida para valores não inteiros de n.

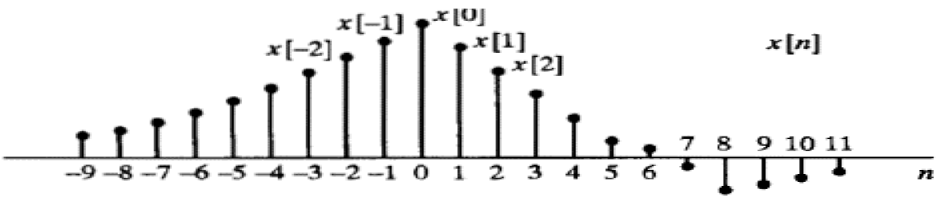

Figura 2.3 - Representação gráfica de um sinal discreto (OPPENHEIM e SCHAFER, 1999).

A Figura 2.4 mostra um segmento de um sinal de fala em sua forma analógica e em sua forma temporalmente discretizada. Em sua forma analógica, o segmento do sinal é representado como a variação de pressão sonora ao longo de 32 ms, e em sua forma discretizada, o mesmo sinal fora particionado em 256 amostras. Cada amostra do sinal representa um conjunto de propriedades estatísticas de uma pequena parte do sinal analógico. Assim, embora o sinal original esteja definido em todo o período de tempo de 32 ms, a sequência discreta do sinal contém apenas informações nos instantes discretos de tempo. Notadamente, quanto maior o número de amostras utilizadas, maior a correspondência entre os dois sinais.

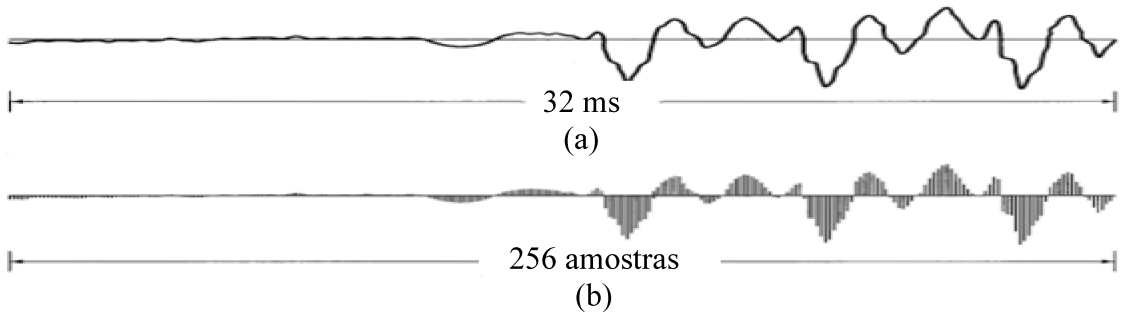

Figura 2.4 - (a) Segmento de um sinal de fala contínuo – sinal analógico; (b) Sequência de amostras obtidas a partir do sinal analógico sinal discreto (OPPENHEIM e SCHAFER, 1999).

2.3 Sistema de Reconhecimento Automático de Voz

A discretização de sinais de voz para estudo de suas propriedades acústicas passou a ter elevada importância a partir da década de 50. A reunião e interpretação das informações

contidas nos sinais de voz como componentes de frequência, amplitude e fase do sinal, além de suas propriedades fonéticas, de entonação, inflexão e modulação da voz, começaram a despertar novas possibilidades no campo da ciência que estuda análise e processamento de sinais. Com a evolução tecnológica, os resultados obtidos utilizando abordagens estatísticas orientadas à sinais de voz atingiram resultados que possibilitaram conferir à voz a possibilidade de atuar como mecanismo acionador de tarefas. Assim, foram criados os primeiros sistemas de reconhecimento de voz, que tinham na fala humana seu principal recurso de funcionamento.

Sistema de reconhecimento Automático de voz é uma reunião de complexas ferramentas computacionais que captam, filtram, processam e interpretam sinais de fala. Estes sistemas são dotados de técnicas utilizadas para extrair as prorpiredades acústicas de um sinal de fala; e de técnicas utilizadas para reconhecimento de padrões.

Os sistemas que utilizam o processamento da voz como mecanismo de ação podem ainda ser classificados em: sistemas *text-to-speech*, que convertem textos em sinais de fala sintetizados; sistemas *speech-to-text*, que converte o sinal de fala em textos; sistemas de codificação da fala; e sistemas de compreensão da fala, que mapeia comandos através do sinal de voz em ações computacionais ou mecânicas (HUANG et al., 2001).

Uma vasta classe de aplicações de processamento de sinais de fala trabalha com extração de informações dos sinais de voz, e a maioria utiliza alguma forma de combinação de padrões. Tais aplicações incluem: reconhecimento da fala, na qual extrai-se a mensagem do sinal de fala para posterior execução de comandos; reconhecimento do locutor, que identifica o indivíduo que enuncia o comando; verificação do locutor, que analisa se a voz capturada do indivíduo que enunciou o comando está catalogada em um banco de dados; pesquisa de palavras, que envolve o monitoramento do sinal de fala investigando a ocorrência de palavras ou frases específicas e indexação automática de gravações de voz com base no reconhecimento de palavras-chave. A Figura 2.5 ilustra os módulos de sistemas de combinação de padrões.

Figura 2.5 - Diagrama de blocos de Sistemas de Combinação de Padrões (adaptado de RABINER; SCHAFER, 2007).

O primeiro módulo de um sistema de combinação de padrões converte o sinal analógico em digital utilizando um conversor AD. Então, o segundo módulo, de análise de

características, converte o sinal discreto em um conjunto de vetores acústicos. E, finalmente, o último módulo, combinação de padrões, alinha temporalmente os vetores acústicos extraídos do segundo módulo com um conjunto concatenado de padrões armazenados. Por fim, o sistema escolhe a identidade associada com o padrão cuja correspondência é a mais próxima ao alinhamento de vetores acústicos do sinal de fala. O resultado pode ser uma ou mais palavras, ao se tratar de um sistema de reconhecimento de fala; ou da identidade do locutor, referindo-se a um sistema de reconhecimento automático do locutor; ou da verificação ou não da identidade do locutor tratando-se de sistemas de verificação.

A Figura 2.6 ilustra algumas áreas que utilizam processamento de sinais de voz digitais.

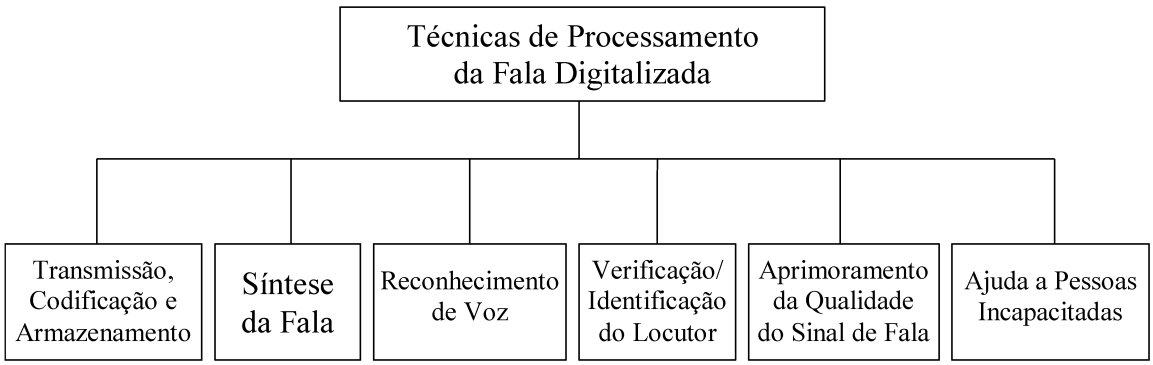

Figura 2.6 - Áreas que utilizam processamento do sinal de voz (adaptado de RABINER e SCHAFER, 2007).

As técnicas de processamento de sinal de fala são indispensáveis para a funcionalidade de diversas ferramentas de análise de sinal. Uma das importantes tarefas relacionada a transmissão de sinais de voz é reduzir a largura da banda efetiva de transmissão. Por conseguinte, existe a necessidade de sistemas que digitalizem o sinal de voz a uma taxa reduzida de bits, facilitando a transmissão e armazenamento do mesmo (RABINER e SCHAFER, 2007). Além disso, sistemas de transmissão digital de dados operam inclusive com sinais de voz encriptados, instigando a importância de sistemas de digitalização de sinais sofisticados. Nesta vertente, sistemas sintetizadores de voz apreendem um grande interesse por sua econômica necessidade de armazenamento digital do sinal de voz para computadores que procedem esta função de comunicação (RABINER e SCHAFER, 1978).

Sistemas verificadores do locutor envolvem a autenticação do indivíduo locutor de um conjunto de indivíduos que tiveram suas vozes catalogadas em um banco de dados. Estes

sistemas devem decidir se o indivíduo tem acesso ou não a determinados serviços, informações diversas ou acesso a áreas restritas. Estão diretamente relacionados a exames de biometria.

Técnicas sofisticadas de processamento de sinais digitais também podem ser utilizadas para aprimorar a qualidade dos sinais de áudio e voz. Em muitas situações, sinais de voz são degradados por meio do canal de recepção do sinal, ou por perdas de conteúdo durante amostragem, impedindo a compreensão da mensagem a ser transmitida. Estas técnicas podem ser utilizadas para remoção da reverberação sonora do sinal, remoção de ruídos e restauração do sinal.

2.3.1 Modelo de Sistemas de Reconhecimento de Voz

A arquitetura típica de sistemas de reconhecimento automático da voz conta com quatro componentes principais: processamento de sinais e extração de propriedades acústicas do mesmo; modelo acústico; modelo linguístico; e pesquisa por hipóteses (HUANG et al., 2011). A Figura 2.7 esquematiza os componentes do sistema.

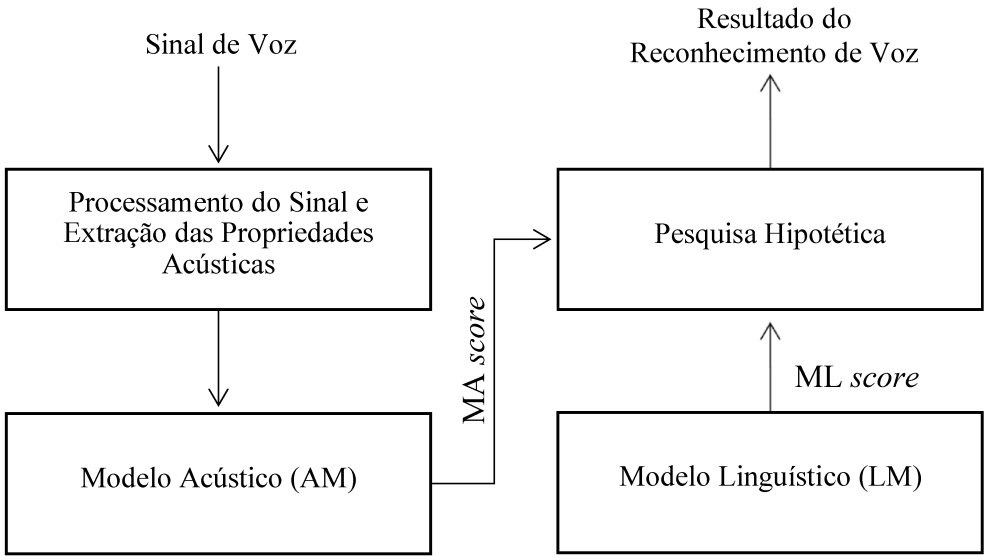

Figura 2.7 - Arquitetura básica de Sistemas de Reconhecimento Automático da Voz (adaptado de YU e DENG, 2015).

O primeiro componente destes sistemas recebe o sinal de voz como comando de entrada. O sinal é processado e suas propriedades acústicas extraídas. Este componente é ainda responsável por aprimorar a qualidade do sinal removendo ruídos e distorções do canal de entrada. No componente seguinte as informações acústicas e fonéticas são interpretadas e integradas a um modelo acústico através de parâmetros graduais, chamado *score*. Este modelo é então comparado com um modelo linguístico, que estima as probabilidades da formação de

uma palavra hipotética, através de um corpora de treinamento, tipicamente um texto, e avalia o entendimento das correlações linguísticas existentes entre os dados, seguindo também um balanço gradual. O resultado da comparação entre os *scores* dos modelos ocorre no último componente deste sistema, chamado de pesquisa hipotética. Neste componente, a sequência de dados referentes à palavra que obteve os mais altos índices dos balanços realizados é escolhida, efetuando o trabalho de reconhecimento do sinal de voz (YU e DENG, 2015).

O processamento da voz é ramificado em uma variedade de aplicações, conforme apresentado na Figura 2.8. Dentre elas, exames de biometria em sistemas de reconhecimento do locutor.

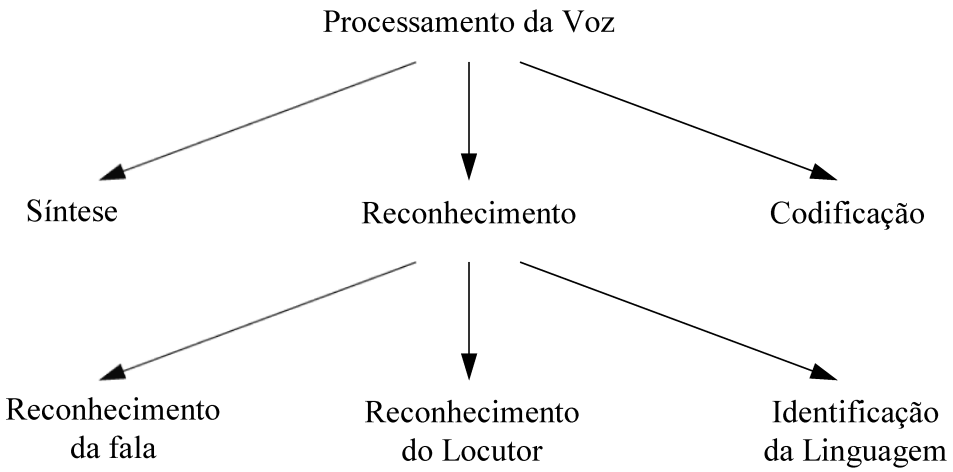

Figura 2.8 - Representação esquemática das modalidades de Processamento de Voz.

2.3.2 Sistemas de Reconhecimento Automático do locutor

Sistema de reconhecimento automático do locutor é uma modalidade biométrica que utiliza o sinal de voz para propósitos de reconhecimento e autenticação de indivíduos. As particularidades do aparelho humano de reprodução da fala, as características de locução, como amplitude da voz e sotaque, bem como a quantidade de pressão utilizada para expelir o ar durante a enunciação de palavras, permitem que a voz seja uma característica forense, ou seja, que a voz seja utilizada como instrumento de identificação de indivíduos.

A usabilidade de sistemas de reconhecimento do locutor em aplicações biométricas tornou-se popular, dentre outros métodos, em função da disponibilidade de dispositivos para coletar as amostras de voz, como: aparelhos telefônicos, celulares e microfones conectados a

computadores; pela facilidade de implementação e rapidez de processamento das informações; pela sua acuracidade; e pelo baixo custo comparado a outros sistemas biométricos, como análise da íris, por exemplo.

2.3.2.1 Taxonomia de Sistemas de Reconhecimento do Locutor

Sistemas de reconhecimento do locutor são classificados em: verificação do locutor; e identificação do locutor. O primeiro trata da tarefa de verificar a identidade alegada de um indivíduo pela sua voz. Este processo envolve apenas decisões binárias sobre a identidade reivindicada, aceitando-a ou rejeitando-a (PATRA, 2007). O sistema deve comparar uma locução dada pelo usuário com o padrão associado à identidade dentro de uma margem de segurança, confirmando se a locução realmente pertence ao locutor ao qual o indivíduo alega ser.

Identificação do locutor é uma tarefa na qual o sistema deve analisar o sinal de entrada perante a variedade de locuções de diferentes indivíduos em um banco de dados e definir a quem pertence a locução. Essa classe de sistemas de reconhecimento do locutor é dividida em duas subclasses: conjunto aberto, na qual a decisão deve ser feita mediante a quem a amostra de fala desconhecida mais se assemelha, e se nenhuma correspondência satisfatória for cabível, o sistema analisa discursos externamente à base de dados; e conjunto fechado, em que o processo de identificação é restrito ao banco de dados e, consequentemente, o sistema tem de responder à qual indivíduo a locução mais se assemelha (MAFRA, 2002).

Em sistemas de identificação do locutor de conjunto fechado há ainda outras duas subdivisões: dependente de texto; e independente de texto. A diferença entre estas tarefas é que, no primeiro caso, o sistema apenas reconhece o locutor se este enunciar o texto, palavra ou sentença que fora catalogado no banco de dados. No segundo caso, o sistema procede a determinação da identidade do locutor com base em quaisquer comandos ou palavras enunciadas como comando e entrada (MAFRA, 2002). A Figura 2.9 apresenta um esquema abordando as diferentes classes sobre sistemas de reconhecimento do locutor.

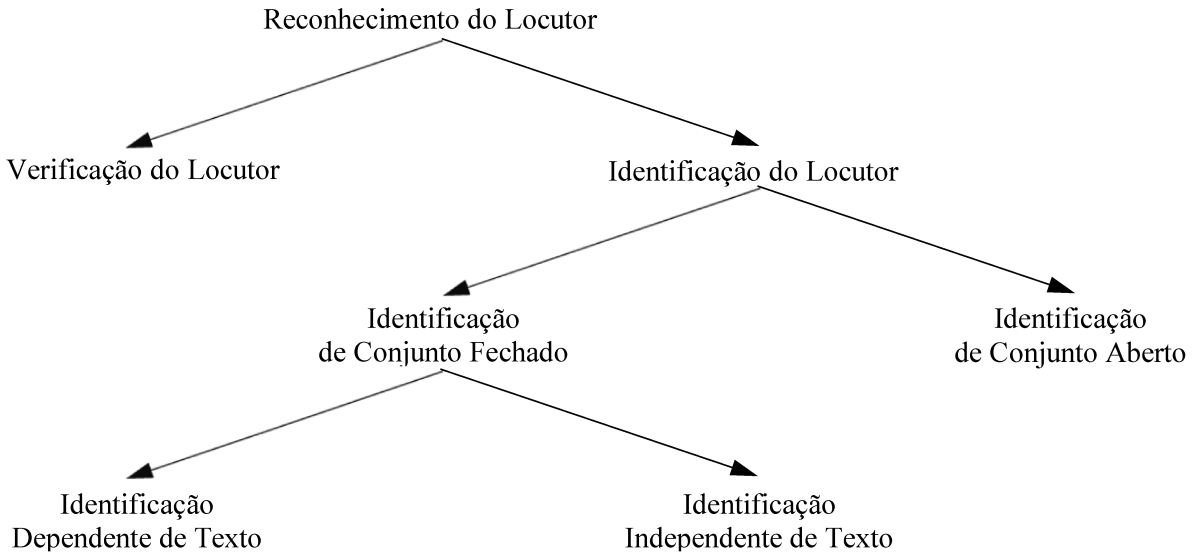

Figura 2.9 - Classificações de sistemas de reconhecimento do locutor (PATRA, 2007).

2.3.2.2 O Processo de identificação do locutor

O processo de identificação do locutor é dividido em duas fases principais: a fase de treinamento e a fase de teste. Durante a fase de treinamento, os indivíduos que participarão do processo de reconhecimento de locutor devem ter suas vozes indexadas no banco de dados do sistema. Nesta etapa, amostras de fala são recolhidas, digitalizadas para a geração de modelos correspondentes aos usuários locutores e então armazenadas em um banco de dados (PATRA, 2007). A segunda etapa, fase de testes, consiste em interpretar sinais de fala e identificar o locutor por meio da comparação dos dados acústicos deste sinal com sinais presentes no banco de dados. Se houver alguma combinação satisfatória, o sistema irá informar a identificação do indivíduo ao qual a voz compreendida como sinal de entrada pertence. Caso contrário, o sistema fornece como resposta a não identificação do locutor.

Ambas as fases envolvem um procedimento técnico comum, ou seja, a extração de características acústicas. O principal objetivo deste procedimento é reduzir o número de testes, associando a maior quantidade de informações ao indivíduo locutor, criando vetores acústicos que são modelados e armazenados no banco de dados do sistema. Então, na fase de testes, o sinal de voz de entrada é também modelado em um vetor acústico e comparado aos pré-existentes no banco de dados. Finalmente, baseado nos resultados obtidos destas comparações, o sistema realiza ou não a identificação do locutor. A Figura 2.10 ilustra um exemplo de discriminação entre as etapas deste sistema.

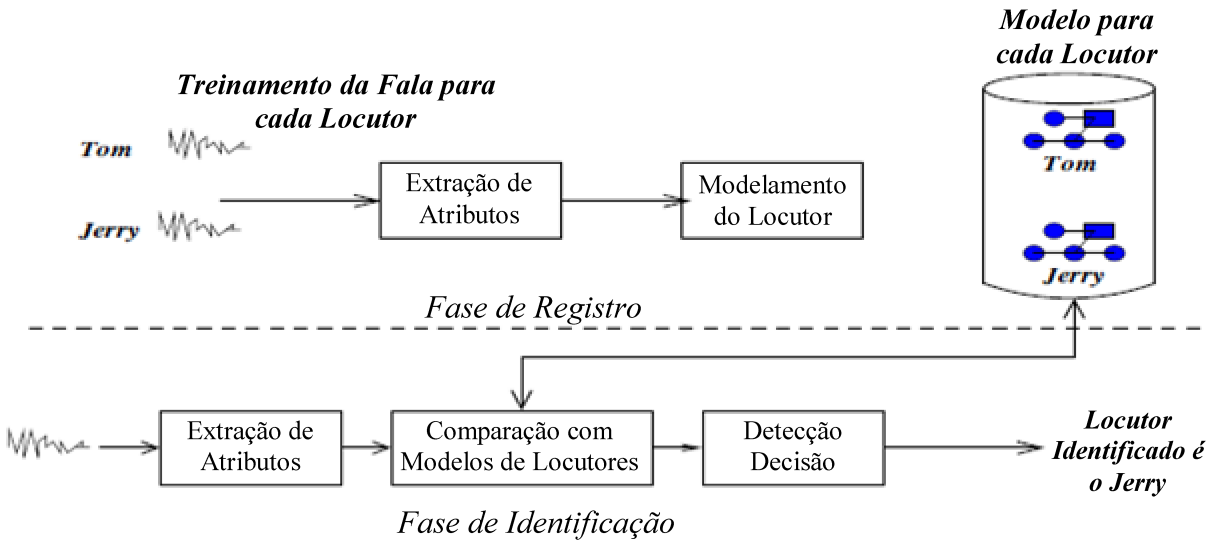

Figura 2.10 - Fases distintas do processo de identificação do locutor (adaptado de PATRA, 2007).

2.3.3 Aplicações de Sistemas de Reconhecimento Automático de Voz

Muitos fatores contribuíram para diversificação de sistemas fundamentados em processamento de linguagem e voz,entre eles, inovações na área da computação, o surgimento da internet como uma massiva fonte de informações, o aprimoramento e aumento de dispositivos conectados em redes sem fio e a crescente necessidade da sociedade moderna de conectividade, facilidade e velocidade de obtenção de informações (YU e DENG, 2015).

As principais aplicações de sistemas de reconhecimento automático da voz incluem por exemplo, sistemas de tradução automáticos; pesquisas na rede de internet; assistência digital personalizada; interatividade em plataformas de jogos; automatização residencial e sistemas de informação embarcados em veículos (RABINER e SCHAFER, 2007). Alguns serão detalhados na sequência.

1. *Aplicações com Pesquisa de voz*: permitem que usuários pesquisem por diversas informações, como localização de estabelecimentos, instruções de direção, avaliações de produtos, e pesquisas em *menus* (ANUSUYA e KATTI, 2009). Esses sistemas reduzem significantemente o esforço e tempo dos usuários. Eles são também populares em dispositivos móveis como iPhone, Windows Phone e Android.
2. *Assistência Digital Personalizada*: apreende informações acerca de conteúdos diversificados, como informações básicas guardadas em dispositivos móveis e preferências do usuário. Com estas informações, o sistema é capaz de interagir,

provendo informações adequadas e agilizando tarefas. Estes sistemas são capazes de organizar tarefas, agendar reuniões, discar números telefônicos, procurar por músicas e outras preferências do usuário. Tornou-se popular através do sistema Siri da Apple.

3. *Interatividade em Plataforma de Jogos*: aprimoramento das plataformas de *video*games e computadores, transformando os jogos em uma nova realidade, na qual o jogador pode comunicar-se com objetos ou personagens do jogo por meio da voz, presente em sistemas como o Xbox da Microsoft.
4. *Automação Residencial*: permite que usuários sejam capazes de realizar tarefas do dia-a-dia por meio de comandos de voz, como ligar e apagar luzes, abrir e fechar cortinas, baixar e levantar persianas, fechar e abrir janelas, ativar e desativar alarmes, ligar e desligar aparelhos televisores e de ar condicionado, por exemplo. Estes sistemas enquadram-se no ramo da domótica e provém diversos benefícios, principalmente para pessoas com necessidades especiais (McLOUGHLIN e SHARIFZADEH, 2009).
5. *Sistemas de Informação Embarcados em Veículos*: pressentes em aeronaves e veículos automotivos. Esses sistemas permitem que o motorista se comunique com um navegador de bordo interativo por meio da voz, para adquirir diferentes informações, pesquisar músicas, localizações, e controlar sistemas interno do veículo.

A modernização de sistemas provedores de dados com aplicações de processamento de voz agilizou a obtenção de informações e serviços, reduziu custos de operacionalidade de sistemas e aderiu maior conforto aos usuários.

Hoje é possível, por exemplo, agendar reservas de passagens aéreas e adquirir informações de voo, interagindo com agentes virtuais em companhias como a Amtrak e United Airlines; adquirir informações a bordo de veículos, controlar a temperatura do ambiente interno, escolher músicas, e fazer ligações sem retirar as mãos do volante, em carros de luxo, como os da companhia Mercedes-Benz; crianças podem aprender a ler através de tutores interativos baseados em personagens animados, e até mesmo, pessoas podem receber tratamento médico através de consultas virtuais (ANUSUYA e KATTI, 2009). Companhias de busca de vídeos, como a Blinkx, utilizam tecnologias de reconhecimento de voz para encontrar palavras em trilhas sonoras de vídeos; a companhia Google fornece recuperação de informações multilíngue

e serviços de tradução no qual usuários podem fazer consultas em suas línguas nativas para obterem informações em outras línguas.

A vasta aplicabilidade de sistemas de reconhecimento de voz justifica os elevados investimentos em pesquisas na área que conduzem à melhoria e inovação de sistemas de captação, digitalização e armazenamento do sinal de áudio, bem como, sistemas extratores de propriedades acústicas de sinais de voz e sistemas de reconhecimento de padrões (YU e DENG, 2015).

2.4 Breve Histórico de Sistemas de Reconhecimento de Voz

Os diferentes tópicos e áreas do conhecimento envolvidos em processamento de linguagem e de voz como conhecimento de linguística computacional na área de linguística; processamento de linguagem natural na área de ciências da computação; reconhecimento de voz na área de engenharia eletroeletrônica e acústica; e psicolinguística computacional na área de psicologia, tiveram seus surgimentos em momentos diferentes na história e todas estas ciências contribuíram para a evolução dos sistemas de reconhecimento de voz (GOLD et al., 2011).

Muitos trabalhos relevantes em sistemas de reconhecimento de voz começaram a ser desenvolvidos nas décadas de 1930 e 1940. Na década de 1940, no período pós Segunda Guerra Mundial, surgiram os primeiros computadores, advindos da invenção de Alan Turing, e com eles, os primeiros modelos teóricos de informação. Contudo, apenas no início da década de 1950 surgiu um sistema completo relacionado a reconhecimento de voz (HUANG et al., 2001; GOLD et al., 2011).

O primeiro sistema capaz de reconhecer palavras foi desenvolvido nos Laboratórios Bell, no início da década de 1950. Esse sistema foi elaborado para reconhecer dígitos de um único locutor e operava medindo uma simples função do espectrograma temporal de energia em duas bandas largas, aproximando as primeiras duas ressonâncias do trato vocal. Apesar de ser um sistema simples, que perdia informações temporais, sua capacidade de reconhecimento superava tecnologias desenvolvidas posteriormente à essa (RABINER e JUANG, 1993). Pesquisadores relatam que mesmo em 1952, sistemas de reconhecimento de palavras operavam com 98% de acuracidade apesar das carências tecnológicas daquele período (GOLD et al., 2011).

Outras inovações ocorreram na década de 50 em termos de reconhecimento de voz e processamento de sinais. Pesquisadores foram capazes de realizar monitoramentos espectrais, detectar algumas palavras e sons, realizar testes com um pequeno grupo de pessoas e desenvolver algoritmos probabilísticos para processamento de linguagem e voz. Em 1958, foi construído um classificador que analisava espectros de maneira contínua (GOLD et al., 2011), substituindo aqueles que analisavam informações particionadas. Ainda nesta década, surgiram modelos probabilísticos de processos discretos de Markov para automação de linguagem. Estes modelos primários conduziram aos primeiros modelos formais de linguagem usando álgebra e definindo linguagens formais como uma sequência de símbolos (HUANG et al., 2001).

Na década de 60, propriedades fonéticas foram exploradas e originaram-se as primeiras abordagens para estimar espectros de curto prazo, por meio de bancos de filtros: Transformada Rápida de Fourier (*Fast Fourier Transform*, FFT), desenvolvido pelos pesquisadores James Cooley e John Tukey (HUANG et al., 2001), uma forma computacionalmente eficiente da Transformada Discreta de Fourier (*Discrete Fourier Transform*, DFT); análises cepstrais de sinais, técnica definida em 1963 em um artigo escrito por Bogert, Healey e Tukey e aplicada posteriormente a sinais de áudio e voz pelos pesquisadores Oppenheim, Schaffer e Stokham, como uma forma alternativa para bancos de filtros estimar envelopes espectrais; *Linear Predictive Coding* (LPC) (HUANG et al., 2001), uma abordagem matemática para modelos de voz baseada no modelo de tubo acústico do trato vocal (RABINER e JUANG, 1993).

Em 1963 e 1964, sistemas utilizando redes neurais para reconhecimento de fonemas e sistemas de reconhecimento de dígitos foram criados alcançando ótimos resultados para operações com múltiplos locutores (RABINER e JUANG, 1993). Foram desenvolvidos métodos para reconhecimento de padrões como: *Dynamic Time Warp* (DTW) e um método estatístico, *Hidden Markov Models* (HMM). DTW utiliza uma abordagem determinística com programação dinâmica para otimização sequencial de sinais, operando como um método de normalização temporal dos mesmos. Utilizar esta técnica resolve o problema de potenciais não correspondências entre o sinal de entrada e o armazenado durante a fase de treinamento, em função de diferentes entonações para a mesma palavra ou sentença (GOLD et al., 2011).

Muitos aspectos de HMM foram desenvolvidos de maneira pioneira pelo grupo IBM na década de 1970 (RABINER e JUANG, 1993). O grupo desenvolveu um sistema de reconhecimento automático de voz baseado em HMM utilizado para reconhecimento de voz contínua, chamado de *New Raleigh Language*. Outros sistemas foram desenvolvidos utilizando

HMM na década de 1970, destacando-se o Dragon, construído por Baker, na época estudante de graduação da Carnegie Mellon University (CMU) (GOLD, et al., 2011).

Aos poucos, pesquisadores começavam a usar vetores com características espectrais, LPC e características fonéticas em seus sistemas de reconhecimento de voz, incorporando ainda informações semânticas e de sintaxe. Quantização Vetorial começava a emergir como uma técnica de compressão de dados e de reconhecimento de padrões, mas apenas em meados da década de 1980 fora largamente explorada em conceitos de reconhecimento de voz, devido ao algoritmo LBG desenvolvido por Linde, Buzo e Gray.

Em 1971, em virtude dos financiamentos da ARPA (*Advanced Research Projects Agency*), o primeiro sistema capaz de reconhecer 1000 palavras utilizando alto-falantes e gramática restrita foi desenvolvido, obtendo ótimos resultados com menos de 10% de erros semânticos. Este projeto foi realizado com o trabalho de três diferentes instituições: a *System Development Corporation*, a CMU e a Bolt, Beranek & Newman (BBN *Technologies*). No entanto, foi na CMU que o projeto denominado Harpy, desenvolvido pelo então estudante de graduação Bruce Lowerre, obteve sucesso. Este sistema utilizou segmentos LPC, incorporou conhecimento de alto escalão e técnicas modificadas do sistema Dragon e de outro sistema desenvolvido na CMU, chamado Hearsay (GOLD et al., 2011).

Durante a década de 80, pesquisadores concentraram-se no desenvolvimento das técnicas já existentes e na construção de interfaces gráficas (*front-end*) para coletar sinais de voz de usuários e processá-los para exibição de respostas do sistema (HUANG et al., 2001). Sistemas de reconhecimento de voz foram treinados com uma maior quantidade de dados e estendidos a tarefas mais complexas, requerendo maiores investimentos em engenharia, sistemas computacionais robustos e pesquisas linguísticas (GOLD et al., 2011).

A comunidade científica que trabalhava com sistemas de reconhecimento de voz naquele período não tinha um banco de dados comum e específico para realizar treinamentos de locuções, impossibilitando ou dificultando troca de informações entre laboratórios. Assim, foi necessário criar um corpora padrão de linguagem de treinamento. Pesquisadores do grupo *Texas Systems* e *Dragon Systems* que trabalharam em conjunto com o NIST (*National Institute of Standards and Technology*) encarregaram-se desta tarefa. O primeiro corpora foi coletado no TIMIT, união da *Texas Instruments* (TI) com o *Massachusetts Institute of Technology* (MIT). Os dados eram gravados na TI e foneticamente segmentados no MIT e então inspecionados e reparados manualmente. O corpora continha 61 fonemas alfabéticos e contava com 630 locutores que enunciavam dez sentenças cada, incluindo duas que eram comuns a todos os

locutores. Este permanece um dos maiores e mais utilizados corpora fonético marcado à mão (GOLD et al., 2011).

Outros corpora gramaticalmente enriquecidos em relação aos sistemas antigos foram desenvolvidos pela ARPA na década de 1980, constando de um modelo linguístico de 1000 palavras. Estes sistemas eram capazes de minimizar as incertezas da próxima palavra em um contexto gramatical e ainda continham um banco de dados de fala coletados a partir da leitura. Alguns destes sistemas incluíam reconhecimento do locutor independente de texto.

Muitos subsistemas extratores de características do sinal de voz foram inventados na década de 1980. Destacam-se os seguintes: *Mel Frequency Cepstral Coefficients* (MFCC); *Perceptual Linear Prediction* (PLP); e *Delta-cepstral Coefficients*. Além disto, sistemas de informações aéreas (*Air-Travel Information Systems*, ATIS) operadas com atendentes virtuais já eram capazes de transmitir informações aos usuários, bem como proceder reservas em certas companhias aéreas.

No final da década de 1980, HMM havia se tornado o método dominante utilizado para reconhecimento de padrões em sistemas de reconhecimento de voz. Estava presente nos principais sistemas desenvolvidos pela SRI (*Stanford Research Institute*), MIT-Lincoln, CMU e *Bell Laboratories*. Na década de 1990, na Universidade de Cambridge, na Inglaterra, foi desenvolvido o HTK, (*HMM Tool-kit*), um conjunto de algoritmos relevantes para sistemas de reconhecimento de voz, que, inclusive podem ser acessados de maneira gratuita até os dias de hoje (GOLD et al., 2011).

No início da década de 1990 houve continuidade das pesquisas envolvendo elementos de sistemas de reconhecimento de voz, no entanto, poucas técnicas revolucionaram este setor com a mesma significância do que as técnicas de programação dinâmica como LPC, HMM, MFCC e Quantização Vetorial (GOLD et al., 2011). Apesar disso, importantes desenvolvimentos ocorreram nesse período, como o aperfeiçoamento de *front-ends*; adoção de técnicas de normalização do canal receptivo do sinal de entrada; técnicas de normalização do trato vocal; adoção de coeficientes cepstrais utilizando escalas de Mel e de Bark; adoção dos coeficietes *Delta-cepstral*; e estimativas probabilísticas associadas a combinação de padrões (HUANG et al., 2001).

Atualmente, há uma recorrente preocupação para que sistemas de reconhecimento automático de voz expandam suas capacidades para várias línguas e dialetos de maior complexidade por meio de novos modelos linguísticos, como o inglês indiano, por exemplo, no qual a distinção entre os fonemas é mais difícil (GOLD et al., 2011). Outros pontos de investigação tratam, por exemplo, de vozes simultâneas sendo coletadas, capacidade do sistema

em operar em ambientes com ruido de fundo, capacidade do sistema em identificar dialetos de forma automática e ferramentas de tradução instantânea.

Neste âmbito, pesquisadores investigam e desenvolvem mecanismos capazes, por exemplo, de segregar dois sinais de fala coletados simultâneamente e de trata-los de maneira separada, caso da técnica ICA (*Independent Component Analysis*); novas transformadas de sinais são adaptadas para aumentar o desempenho computacional e testar a eficiência de processamento de sinais, como a Transformada HHT (*Hilbert-Huang Transform*); diversos testes analisando a eficiência de distinção de pronúncias são realizados; medidas Euclidianas ponderadas foram incorporadas em sistemas de reconhecimento de padrões; sistemas inteligentes capazes de interpretar o contexto textual foram desenvolvidos, facilitando consultas diversas; linguagens não formais, como expressões da internet (*Web language*), foram adaptadas e incorporadas a alguns bancos de dados e agendas de reconhecimento de voz foram criadas (sistemas que detectam autores das elocuções e gravam os momentos das elocuções).

Ao longo da história, percebe-se a evolução de tarefas associadas a sistemas de reconhecimento de voz diretamente relacionadas à complexidade das tarefas, disposição de tecnologias avançadas e novos conhecimentos adquiridos. A Figura 2.11 resume de maneira concisa o processo histórico associando as tecnologias de reconhecimento de voz.

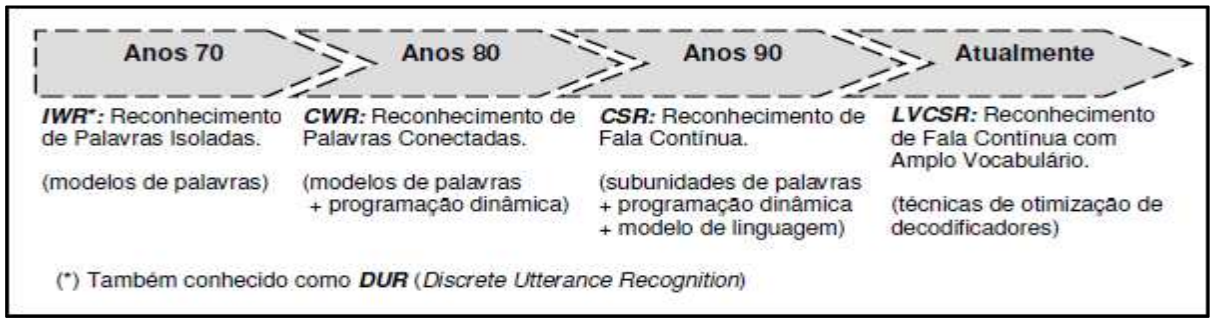

Figura 2.11 - Esquema descritivo do processo histórico de Sistemas de Reconhecimento Automático de Voz (TEVAH, 2006).

A diversidade de aplicações embasadas em sistemas de reconhecimento automático de voz aumentou consideravelmente a partir do século XXI. Com a geração dos modernos processadores *multi-core*, ampliação da memória de dispositivos móveis, constante necessidade de conectividade e comforto do mundo moderno, tais sistemas passaram a ocupar um espaço importante nas tarefas do *dia-a-dia*. Estes sistemas estão presentes em ações de consultas online (sistemas de comando de voz utilizados no Google Chrome, por exemplo); acessibilidade a dados bancários (utilizado no banco Abu Dhabi Commercial Bank, ADCB, como método

biométrico e informativo); exames de biometria; sistemas embarcados em veículos automotores (presente em modelos da Mercedes-Benz, por exemplo); consultas em dispositivos móveis (como o comando de voz integrado ao iPhone da empresa Apple); sistemas residenciais automatizados por controle de voz; e traduções em tempo real (como o Google Translate).

A comercialização em maior escala de mecanismos com componentes de reconhecimento automático de voz está impulsionando pesquisas e financiamentos nesta área do conhecimento por todo o mundo, capacitando assim, a construção de sistemas mais robustos e inteligíveis. Apesar de ser considerada uma história recente, as pesquisas indicam que muitos avanços ainda estão por ocorrer nesta área, que revolucionará a maneira como o ser humano se comunicará com as máquinas.

2.5 Técnicas de Extração das Propriedades Acústicas do sinal de voz

Existem dois principais módulos de reconhecimento de voz, quais sejam: (i) a extração de características do sinal; (ii) reconhecimento de padrões.

Basicamente, a extração de características do sinal pode ser descrita como um processo de construção de vetores que carregam consigo informações acústicas úteis para a descrição do conteúdo vocalizado do sinal chamados de vetores acústicos. Estes são utilizados para representar cada um dos locutores, bem como representar comandos específicos que transmitirão ordens ao sistema para reprodução de alguma ação.

Particularmente, durante o módulo de extração das propriedades acústicas do sinal, elimina-se diversas fontes de informação, como segmentos que contém apenas ruído ou ausência de som audível (normalmente regiões inicias e finais de uma gravação), efeitos de periodicidade, amplitude do sinal de excitação, reverberação, componentes de ruído e frequências fundamentais.

As principais técnicas utilizadas em sistemas modernos de reconhecimento automático de locutor e voz são: *Mel Frequency Cepstral Coefficients* (MFCC); *Linear Predictive Coding* (LPC) e *Perceptual Linear Predictive Coefficients* (PLP) (GOPI, 2014; DAVE, 2013). MFCC e PLP são técnicas que consideram a natureza da elocução ao extrair as propriedades importantes do sinal, enquanto LPC opera com modelos probabilísticos '*prevendo*' as próximas características do sinal.

2.5.1 Mel Frequency Cepstral Coefficients (MFCC)

O método prevalecente em sistemas modernos de reconhecimento de voz usado para extrair características espectrais é MFCC. Trata-se de uma das técnicas de extração de características acústicas de sinais de voz mais populares usadas no reconhecimento da fala com base no domínio da frequência, sendo considerada mais precisa que técnicas que concatenam dados no domínio do tempo (TIWARI, 2010). Essa ferramenta utiliza uma escala chamada Mel, que se baseia na escala do ouvido humano (LOGAN, 2000).

Mel Frequency Cepstral Coefficients são representações da parte real do cepstro de um janelamento em curto período de tempo de sinais acústicos, derivados de uma Transformada Rápida de Fourier (FFT) de um sinal. A diferença destes coeficientes com coeficientes cepstrais reais é que uma escala logarítmica é utilizada, aproximando os coeficientes ao comportamento do aparelho auditivo humano (MOLLA e HIROSE, 2004). Estes coeficientes são robustos e confiáveis às variações de alto-falantes e condições de gravação.

Esta técnica, além de extrair as características úteis de um sinal de voz para posterior criação dos vetores acústicos, também reduz a quantidade de informações não úteis, retirando partes sem registro de voz, além de reduzir efeitos de reverberação. Para calcular tais coeficientes, deve-se seguir alguns passos sequenciais, dentre eles o enquadramento e janelamento do sinal, utilização dos dados no domínio da frequência usando-se de uma Transformada Rápida de Fourier, aplicação da escala logarítmica (Mel) para construção de um conjunto de coeficientes cepstrais e, finalmente, utilização de uma Transformada Discreta do Cosseno (DCT) sobre os dados adquiridos no passo anterior (TIWARI, 2010).

A mais notável desvantagem desta técnica está em sua sensibilidade ao ruído devido a sua caracterização espectral. Assim, é comum utilizar-se paralelamente ferramentas que operam com periodicidade do sinal de voz.

2.5.2 Linear Predictive Coding (LPC)

Linear Predictive Coding é uma ferramenta extratora de propriedades acústicas de sinais de fala que utiliza modelos probabilísticos. Esta técnica aproxima uma amostra específica em um dado momento temporal, por meio de uma combinação linear de amostras do sinal de fala avaliadas anteriormente. É considerada robusta e também desejável para comprimir sinais de áudio, obtendo condições de transmissão e armazenamento eficientes (GOPI, 2014).

Em sistemas de reconhecimento de voz, a técnica LPC é responsável por formular coeficientes calculando o espectro de potência dos sinais e efetuando análises de ressonâncias

acústicas e amplitude dos tons que representam os fonemas enunciados em cada sílaba. Usando a teoria do Mínimo Erro Quadrado, minimiza-se a soma das diferenças quadradas entre o sinal original e o estimado sobre uma duração finita. O resultado é uma combinação de coeficientes preditivos, cujas análises envolvem um processo de decisão marcando regiões de voz e com ausência de voz, por meio de detecção de picos de ressonância e ganho (DAVE, 2013).

LPC pode também ser utilizado em técnicas de codificação do sinal de fala. Isto porque, ao passar o sinal de fala em um filtro para remover as redundâncias do mesmo, é gerado um erro residual na saída do processo que pode ser quantizado em um número reduzido de *bits* quando comparado ao sinal original de fala. Assim, ao invés de transferir o sinal completo, é possível transferir apenas o erro residual gerado pelo filtro e os parâmetros acústicos extraídos do sinal de fala para gerar o sinal original. Este processo consome um número inferior de *bits*, tornando tal método efetivo para codificação de palavras.

2.5.3 Perceptual Linear Predictive Coefficients (PLP)

Outro importante e usual modelo de extração de características acústicas do sinal de voz é o PLP. É similar ao método LPC, mas utiliza espectros em curtos períodos de tempo do sinal de fala. Esta técnica modela o aparelho auditivo humano baseado nos conceitos psicofísicos da audição. Em contraste com analises LPC, PLP modifica o espectro de curto período de tempo do sinal de fala através de várias transformações psicofísicas (DAVE, 2013).

O processo de computação da técnica PLP ocorre em três passos principais de aproximações perceptuais: curvas de resolução de banda crítica, curva de igualdade de intensidade sonora e relação de potência e intensidade sonora, conforme pode ser observado na Figura 2.12.

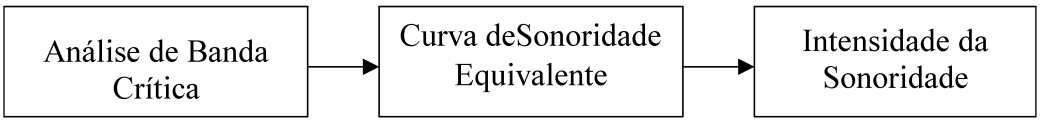

Figura 2.12 - Diagrama do processo de computação da técnica PLP (adaptado de DAVE, 2013).

Uma particularidade desta técnica é o fato de utilizar a escala de Bark que tem boa representação da resolução em frequências da audição humana. O sinal digitalizado apresentado em frequências na escala Bark passa por uma convolução com o espectro de potência da curva de banda-crítica e é segmentado em amostras ainda menores. O resultado é ponderado por meio

de uma pré-ênfase da curva de intensidade de igualdade sonora, simulando a sensibilidade da audição humana e então equalizando-a. Finalmente, aplica-se a ferramenta preditiva linear ao espectro linear resultante, obtendo-se os coeficientes cepstrais PLP (DAVE, 2013). A Figura 2.13 apresenta um diagrama dos passos para computação dos coeficientes.

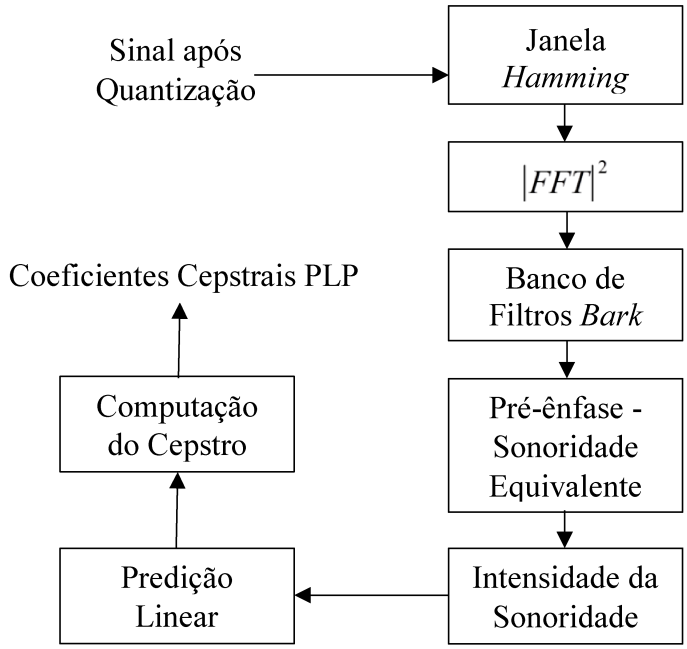

Figura 2.13 - Diagrama esquemático da técnica PLP (adaptado de DAVE, 2013).

2.6 Técnicas de Reconhecimento de Padrões

A combinação de características de sinais de voz ou reconhecimento de padrões é um processo que pontua o sinal de entrada, segundo uma graduação específica, determinando o grau de correspondência entre o sinal de voz de entrada e sinais salvos em um banco de dados. Havendo uma graduação que caracterize similitude entre os dois sinais, o reconhecimento de padrões é efetivo. Esta associação pode corresponder a um fonema, palavra, sentença, ou até mesmo a um indivíduo, particularmente para sistemas de reconhecimento de locutor.

Os principais modelos de reconhecimento de padrões utilizados em sistemas de reconhecimento de voz modernos são: Quantização Vetorial (VQ); *Gaussian Mixed Models* (GMM); *Hidden Markov Models* (HMM); *Dynamic Time Warping* (DTW) e Redes Neurais Artificiais (*Artificial Neural Networks*, ANN) (GOPI, 2014; YU e DENG, 2015).

2.6.1 Quantização Vetorial (VQ)

Quantização Vetorial é um processo que mapeia vetores acústicos de um espaço vetorial para um número finito de regiões naquele espaço. Estas regiões são chamadas de *clusters* e representadas por seu centroide, as *codewords* (GERSHO e GRAY, 1992). O conjunto de centroides é chamado de *codebook*. Em sistemas de reconhecimento do locutor, por exemplo, é formulado um *codebook* para cada indivíduo que treina sua voz, aplicando VQ no conjunto de vetores acústicos extraídos do sinal de fala. Tal *codebook*, referido como modelo acústico, é significantemente menor que o tamanho do vetor acústico pelo qual este fora formado (GILL et al., 2010).

A construção de um *codebook* é um procedimento recursivo que converge quando uma das distâncias Euclidianas (chamada nesta ferramenta de distorção espectral) entre o centroide do vetor acústico de entrada e os demais centroides presentes no espaço vetorial é inferior a um valor anteriormente estabelecido. Teoricamente, é possível que cada *cluster* possa modelar componentes particulares da fala. Matematicamente, a tarefa de Quantização Vetorial é definida como segue: dado um conjunto de vetores acústicos, define-se a partição do espaço vetorial em um número específico de regiões e cada vetor dentro desta região é representado por seu centroide correspondente. Este pode representar um comando, um fonema, uma palavra, ou ainda caracterizar um indivíduo (GERSHO e GRAY, 1992).

Após a construção do modelo acústico (*codebook*), o qual contém as informações de sinais de fala treinados, seccionadas em regiões particulares no espaço vetorial criado, é possível realizar testes de reconhecimento. Depois de extrair as propriedades acústicas necessárias do sinal de fala desconhecido por meio da ferramenta VQ, constrói-se novos *clusters* que são lançados no mesmo espaço vetorial do banco de dados. Calcula-se a distância deste novo *cluster* aos outros centroides e investiga-se a mínima distância encontrada. Se esta distância Euclidiana for inferior a um determinado limite (adotado experimentalmente) associa-se a elocução (sinal de entrada) à característica definida por este *cluster* (MAKHOUL et al., 1985; LANDELL et al., 1984).

Quantização Vetorial é um método vastamente empregado em sistemas de reconhecimento de voz modernos, por apresentar acuracidade, rapidez de processamento e ótimo desempenho, principalmente em sistemas de reconhecimento do locutor (GILL et al., 2010).

2.6.2 Gaussian Mixture Model (GMM)

Gaussian Mixture Model é um método estocástico baseado na modelagem de variações estatísticas das características do sinal. Esta técnica fornece uma representação estatística de como um locutor produz diferentes tipos de sons. Em sistemas de reconhecimento do locutor, cada indivíduo é representado pelo seu próprio GMM, que é parametrizado pelos vetores acústicos, matrizes de covariância e ponderações mistas de todos os componentes de densidade dos sinais de voz (YU e DENG, 2015).

GMM é um importante classificador de sinais de fala que pode ser segmentado em duas partes: um classificador uni-modal Gaussiano; e um modelo que separa espacialmente classes acústicas. Esta técnica combina a robustez do modelo paramétrico Gaussiano com modelos arbitrários que utilizam funções probabilísticas para gerar classificadores acústicos (GOPI, 2014). Usualmente, utiliza-se o algoritmo *K-means* ou algoritmo *Maximum-Likelihood Expectation-Maximization* (ML-EM) para realizar esta tarefa (GÂTA e TODEREAN, 2006).

2.6.3 Hidden Markov Models (HMM)

Hidden Markov Models é uma ferramenta popular em aplicações para modelagem de séries temporais. É utilizada em vários segmentos da ciência como: em sistemas de reconhecimento de voz; compressão de dados; modelagens de sequências de imagens; rastreamento de objetos e sistemas inteligentes com aplicações de visão computacional (VIRTANEN et al., 2013). A vasta aplicabilidade desta técnica é garantida pelo fato do modelo HMM apresentar uma estrutura matemática robusta e operar em situações que envolvem fenômenos aleatórios (JUANG e RABINER, 1991).

HMM é uma ferramenta utilizada para representar distribuições de probabilidades sobre sequências de observações. O nome dado a esta ferramenta é justificado porque assume que a observação realizada em um dado momento qualquer é gerada através de um processo, cujo estado é '*escondido*' do usuário. Assume-se ainda que o estado deste processo '*escondido*' satisfaz a propriedade de Markov, que afirma que o processo é independente de qualquer outro processo em tempos anteriores ou posteriores. Assim, o estado em dados momentos de um processo engloba todo o conhecimento histórico para predizer o futuro daquele mesmo processo (VIRTANEN et al., 2013).

HMM é uma poderosa técnica estatística para processos de variação temporal quase-estacionários. A estrutura HMM, portanto, fornece um meio eficiente para caracterizar a distribuição de um sinal de voz e ainda apresenta facilidade de treinamento que pode ser

realizada de maneira automática, além de ser um método simples e robusto matematicamente (CHOU; JUANG, 2003).

2.6.4 Dynamic Time Warping (DTW)

Dynamic Time Warping é uma técnica de combinação de parâmetros baseada em programação dinâmica. O algoritmo DTW é implementado para medir similitudes entre duas series temporais que podem variar no tempo. Calcula-se a mínima distância entre vetores de propriedades acústicas capturadas do sinal de voz correspondente a elocução com parâmetros de referência em um banco de dados. A detecção de um fonema, palavra ou sentença é determinada através do mínimo valor encontrado dentre todas as características salvas em um banco de dados (MUDA et al., 2010).

Em sistemas de reconhecimento de voz DTW pode-se atuar de maneira isolada para reconhecimento de padrões, ou combinada a outras técnicas como por exemplo, a Quantização Vetorial. Uma das vantagens de DTW é o alinhamento temporal entre dois sinais diferentes para uma mesma locução (MOHAN e BABU, 2014). Em sistemas de reconhecimento de voz, isto ocorre quando as gravações de treinamento se diferem das locuções de testes, embora a frase, comando ou palavra enunciada esteja presente no banco de dados do sistema. Este fato pode ser causado por: maneiras distintas de elocução; diferenciação da qualidade de captação dos sinais; diferenciação no período de tempo das locuções e ruidos de fundo. A Figura 2.14 ilustra a aplicação da técnica DTW para duas amostras de sinais.

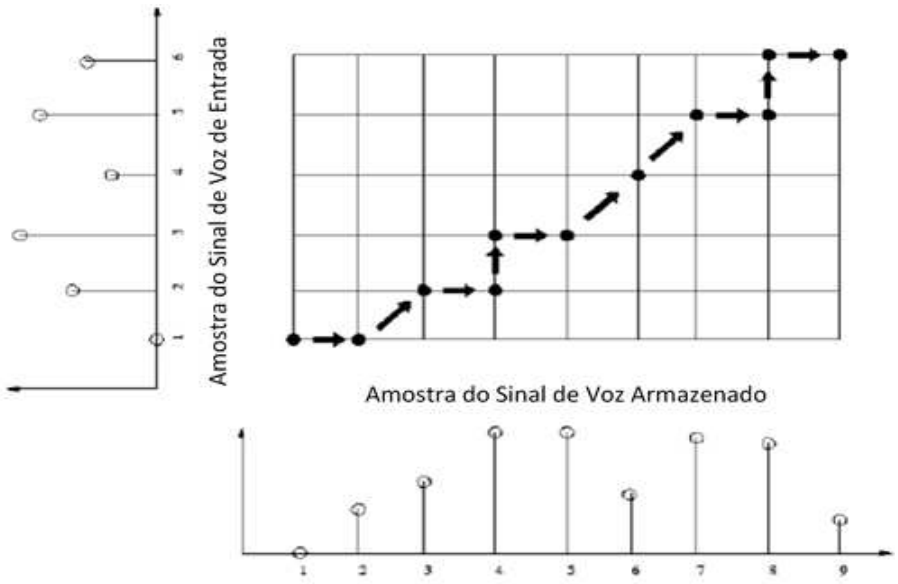

Figura 2.14 - *Dynamic Time Warping* de duas amostras de sinais de voz (MOHAN e BABU, 2014).

2.6.5 Redes Neurais Artificiais (Artificial Neural Networks, ANN)

Redes Neurais Artificiais têm a habilidade de aprender relações de entrada e saída de sinais não-lineares e complexos, utilizar procedimentos de treinamento sequenciais e adaptar-se aos dados de trabalho.

ANN é um modelo de processamento de informações inspirado no sistema nervoso humano, especificamente no processo de sinapses que ocorre entre dois ou mais neurônios, transmitindo impulsos nervosos que serão interpretados pelo cérebro. Este modelo é composto por uma grande quantidade de elementos de interconexão (similares aos neurônios) que operam em união para resolver problemas específicos. Cada ANN é configurada para uma aplicação específica, e o processo de treinamento da rede envolve ajustes às conexões entre seus elementos (YU e DENG, 2015).

As redes mais usadas para classificação de padrões em sistemas de reconhecimento de voz, as chamadas *Deep Neural Networks* (DNN) são: a rede neural *Feedforward*, que inclui multicamadas das redes *Perceptron* (MLP) e *Radial-Basis Function* (RBF); e a rede *Self-Organizing Map* (SOM), também conhecida como Kohonen-Network, que é utilizada para agrupamento de dados e combinação de padrões (RUSSEL e NORVIG, 1995).

O procedimento de aprendizagem das redes neurais envolve a atualização da arquitetura da mesma, assim como dos pontos de conexão ponderados, tornando a rede eficiente para realização de tarefas de comparação e classificação. A Figura 2.18 enfatiza a arquitetura da taxonomia de Redes Neurais Artificiais.

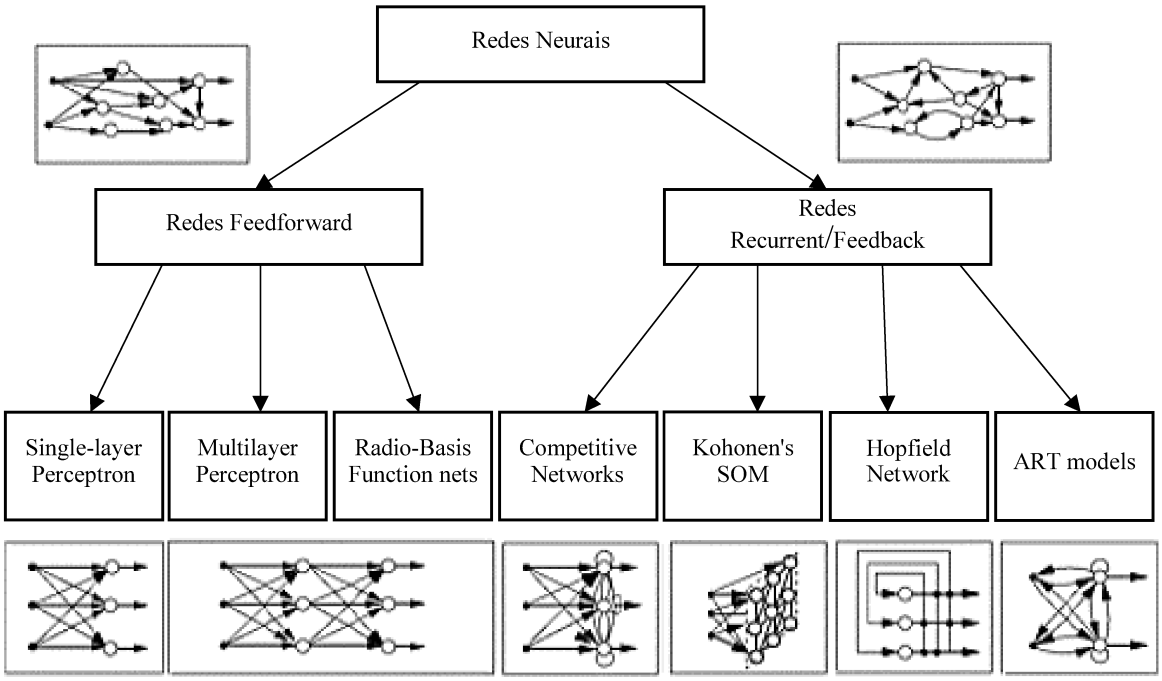

Figura 2.15 - Taxonomia de estruturas de Redes Neurais Artificiais (adaptado de RUSSEL e NORVIG, 1995).

Redes Neurais Artificiais fornecem um novo conjunto de algoritmos não-lineares para classificação de padrões acústicos. Em adição, sua capacidade adaptativa permite que este modelo seja configurado de maneira eficiente para diferentes tipos de sistemas de reconhecimento de voz como identificação de fonemas e comandos, bem como de locutores (YU e DENG, 2015).

3 ARQUITETURA DO SISTEMA DE RECONHECIMENTO DE VOZ

A presente pesquisa abrange o desenvolvimento de um sistema de reconhecimento automático de voz, que opera em três vertentes distintas: reconhecimento do locutor; reconhecimento de palavras e frases; e a combinação das duas tarefas anteriores. As técnicas empregadas ao sistema foram: *Mel Frequency Cepstral Coefficients* (MFCC) para extração das propriedades acústicas dos sinais de voz e Quantização Vetorial (VQ) para classificação de padrões.

A opção da técnica MFCC para extração de características acústicas do sinal de voz se deu pela robustez e eficiência deste modelo que tenta reproduzir o aparelho auditivo humano. Em adição, este modelo opera com espectros de frequência, sendo mais preciso que modelos que operam no domínio do tempo. Sua vasta aplicabilidade em sistemas modernos de reconhecimento de voz destaca-se também por sua compatibilidade de processamento com outros mecanismos como por exemplo, técnicas de normalização do sinal.

Embora a extração de propriedades acústicas de um sinal de áudio seja geralmente um processo irreversível, pois há perdas de informações do sinal ao se utilizar o método MFCC com frequência de amostragem entre 20 e 40 ms e janelamento do sinal com deslocamento de aproximadamente 10 ms, as perdas de importantes partes do sinal são reduzidas, dessa forma o método pode ser utilizado para análise de sinais.

Como ferramenta de classificação e reconhecimento de padrões optou-se por Quantização Vetorial, a qual apresenta eficiência na classificação de parâmetros, capacidade de comprimir dados, alta velocidade de processamento e facilidade de implementação. O modelo de Quantização Vetorial adotado nesta pesquisa foi o LBG, nome que leva as iniciais dos pesquisadores que inventaram este eficiente e intuitivo algoritmo, *Linde, Buzo* e *Gray*. Este modelo projeta vetores quantizados para treinamento de uma longa cadeia de dados, sendo vastamente aplicados a sistemas de reconhecimento de voz.

3.1 Digitalização do sinal de voz

Previamente à extração das propriedades acústicas de um sinal de áudio, faz-se necessário a captação, digitalização e pré-processamento do mesmo. O objetivo do pré-processamento é reproduzir atributos paramétricos dos sinais de áudio reduzindo redundâncias,

ruídos, e mantendo a integridade das informações úteis correspondentes às elocuções contidas no sinal. Estes atributos devem ser facilmente transformados em vetores acústicos; ser robustos; facilmente mensuráveis; conter características naturais da voz como frequências, harmônicos e tons; ser pouco afetados pelas características de fala do locutor, sua idade ou sexo e apresentar o mínimo possível de ruído, seja este ruido branco ou externo. Tais características são melhores atendidas utilizando o modelo MFCC (PICONE, 1993).

O digitalizador é responsável pela captura e digitalização do sinal de voz e, normalmente, é integrado a uma placa de som. É constituído por um microfone, um pré-amplificador, um sistema de *anti-aliasing*, um *sampler*/*holder*, um conversor A/D e um banco de polifones (TEXAS INSTRUMENTS INC., 1995). A Figura 3.1 ilustra a estrutura de um digitalizador.

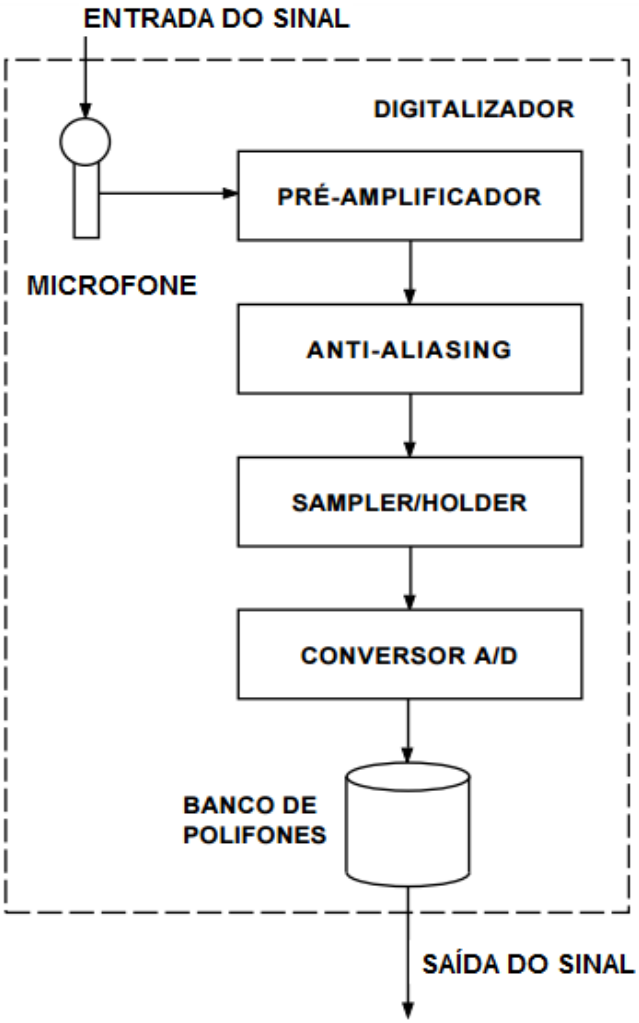

Figura 3.1 - Diagrama de um Digitalizador (adaptado de MAFRA, 2002).

A seguir são descritos os componentes do sistema.

1. **Microfone**: instrumento que capta as ondas sonoras emitidas pelo locutor através de vibrações de uma sensibilíssima membrana e as converte em sinais de tensão elétrica analógicos.
2. **Pré-amplificador**: filtro analógico de ganho positivo ao sinal de entrada.
3. **Anti-aliasing**: filtro analógico (*Butterworth, Chebyshev, Inverse Chebyshev, Cauer, Bessel-Thomson*) que elimina altas frequências (frequências superiores à da largura da banda relevante) evitando que estas estejam presentes nas análises espectrais e desta forma, introduzindo erro de *aliasing* (TEXAS INSTRUMENTS INC., 1995).
4. **Sampler/Holder**: etapa do digitalizador que amostra o sinal filtrado em pequenos intervalos. O sinal é estabilizado através de um *holder* durante a conversão do mesmo de analógico para digital. Neste trabalho usou-se uma taxa de amostragem igual a 22.050 Hz.
5. **Conversor A/D**: converte o sinal discreto analógico em digital, quantizando-o com uma determinada resolução variável em *bits*. Neste trabalho optou-se por uma resolução igual a 8 *bits*.
6. **Banco de Polifones**: armazena os sinais em memórias permanentes através de um *hard-drive* em arquivos digitais em formatos variados (*.wav*,.*mp3*).

O sinal discreto e armazenado em um *hard-drive* pode então ser processado, e dele extraídos importantes informações acústicas úteis para reconhecimento de voz.

3.2 Extração de propriedades acústicas - *Mel Frequency Cepstral Coefficients*

Modelos extratores de propriedades acústicas de sinais de áudio são responsáveis por interpretar as informações úteis para sistemas de reconhecimento de voz, ou seja, informações correspondentes à fala humana do sinal e agrupar estes atributos. É portanto necessário, que estas ferramentas sejam capazes de distinguir o que é fala do sinal de voz do que não é sem degradá-lo, isto é, com o mínimo de perda de informações úteis possível.

Mel Frequency Cepstral Coefficients são representações na qual a forma do trato vocal se manifesta no envelope do espectro de potência em um curto espaço de tempo. A técnica MFCC tenta mimicar o aparelho de reprodução da fala humana que é constituido pelas

cavidades orais e nasais, língua, dentes e traqueia. Este aparelho atua de maneira similar a um filtro, sendo seu formato quem determina qual som é produzido (DHINGRA et al., 2013).

Os coeficientes cepstrais podem ser compreendidos como variáveis que carregam informações acerca de taxas de variação em diferentes bandas do espectro do sinal de áudio. *Mel Frequency Cepstral Coefficients* representam uma ligeira mudança a outros coeficientes cepstrais, incorporando a estes a escala Mel, que aproxima as bandas do espectro aos diferentes tons compreendidos pelo sistema auditivo humano. Podem ser descritos como uma espécie de processador de bancos de filtros adaptados às especificidades do sinal de fala.

A Figura 3.2 apresenta um diagrama que esquematiza o processo de obtenção dos *Mel Frequency Cepstral Coefficients* que serão explicados a seguir.

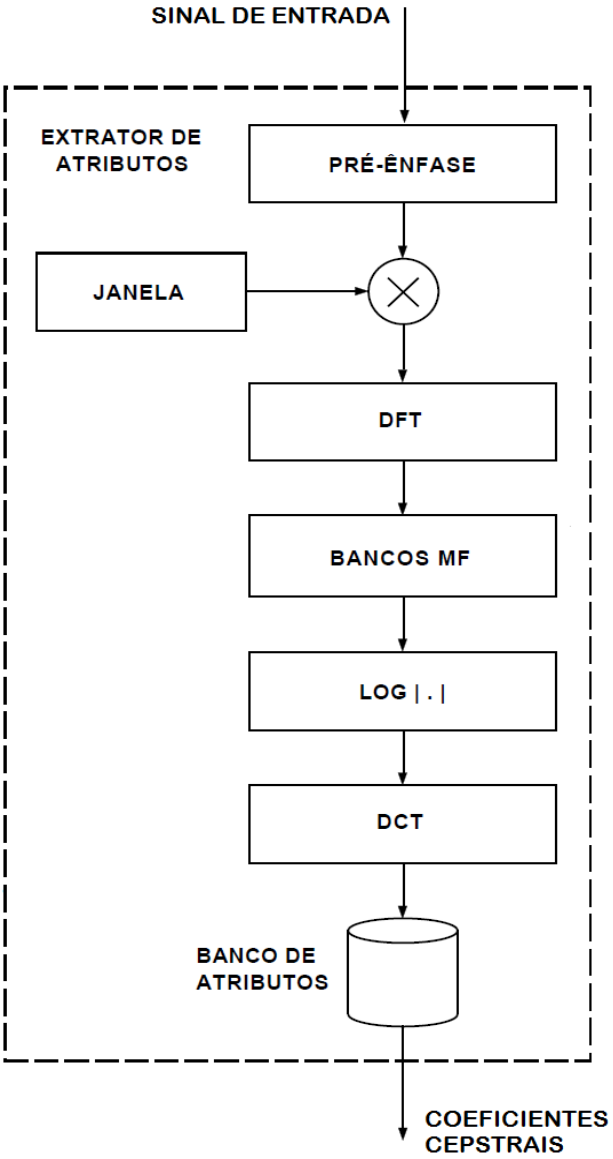

Figura 3.2 - Diagrama do processamento de extração de coeficientes cepstrais de um sinal de voz (adaptado de MAFRA, 2002).

3.2.1 Pré-ênfase

Em sinais de voz, a energia contida nas altas frequências é bastante inferior às baixas frequências (curva de ponderação A). Assim, anteriormente à análise espectral, é necessário uma pré-ênfase do sinal por meio da aplicação de um filtro digital. O objetivo é compensar regiões de alta frequência do sinal que foram suprimidas durante a produção do som, obtendo amplitudes mais homogêneas das frequências do sinal. Normalmente, para este fim, utiliza-se um filtro digital Passa-alta FIR (*Finite Impulse Response*) (OPPENHEIM e SCHAFER, 1999).

A Figura 3.3 apresenta um sinal de voz previa e posteriormente à aplicação do filtro digital FIR com seus respectivos espectros de potência.

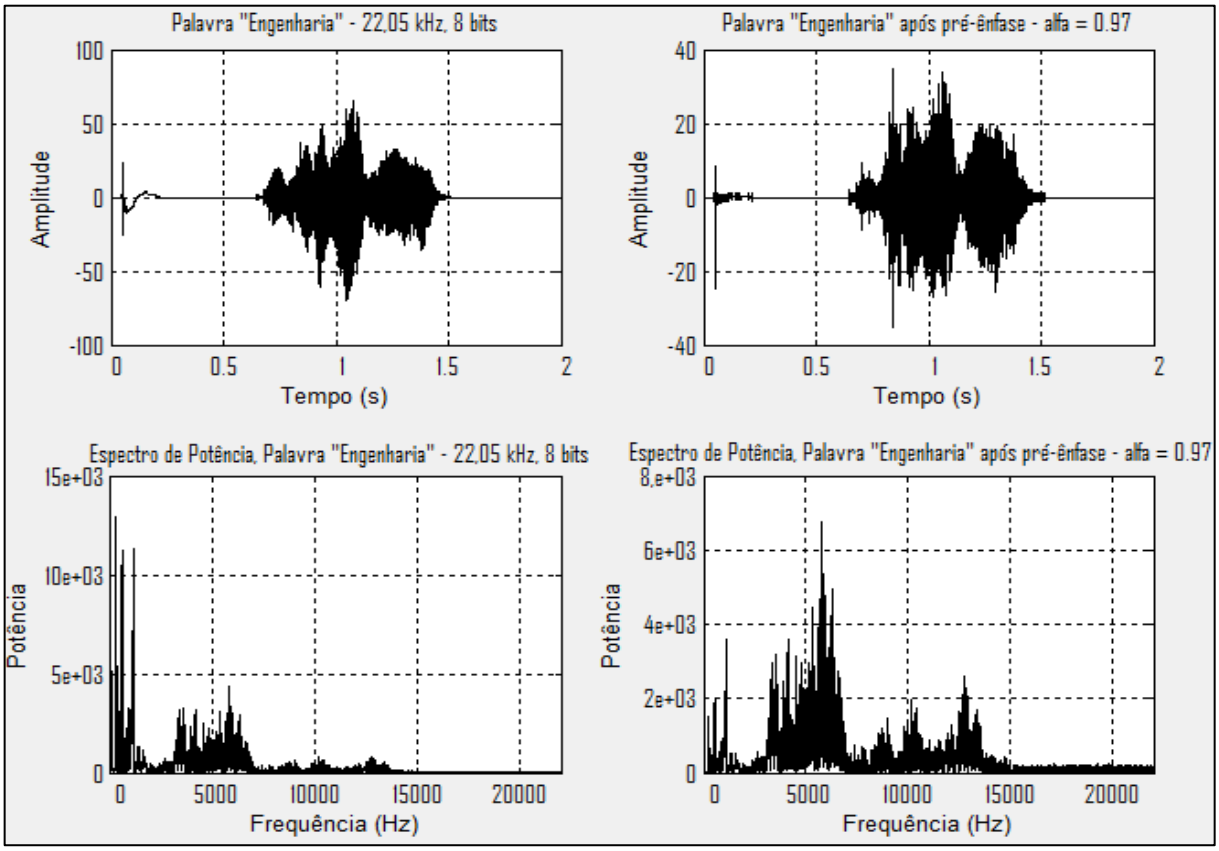

Figura 3.3 - Comparação de sinais no domínio do tempo e de seus respectivos espectros de potência antes e após aplicação da pré-ênfase.

3.2.2 Enquadramento do Sinal (*Frame Blocking*) e Janelamento

O próximo passo do processamento é enquadrar o sinal digitalizado em pequenas amostras. Um sinal de voz está constantemente variando no tempo, por isto, é necessário

estabelecer que em um curtíssimo período de tempo ele seja quase estacionário. Para isto divide-se o sinal em segmentos de curto intervalo de tempo e igualmente espaçados.

O sinal digitalizado é então discreto em *frames* que podem variar de 20 a 40 ms (OPPENHEIM e SCHAFER, 1999). Nesta pesquisa, optou-se por enquadrar o sinal em frames de 30 ms, com o sinal contendo 256 amostras sobrepostas a cada 100 amostras. Cada frame será convertido em 12 *Mel Frequency Cepstral Coefficients*. A Figura 3.4 é uma representação do enquadramento de sinais.

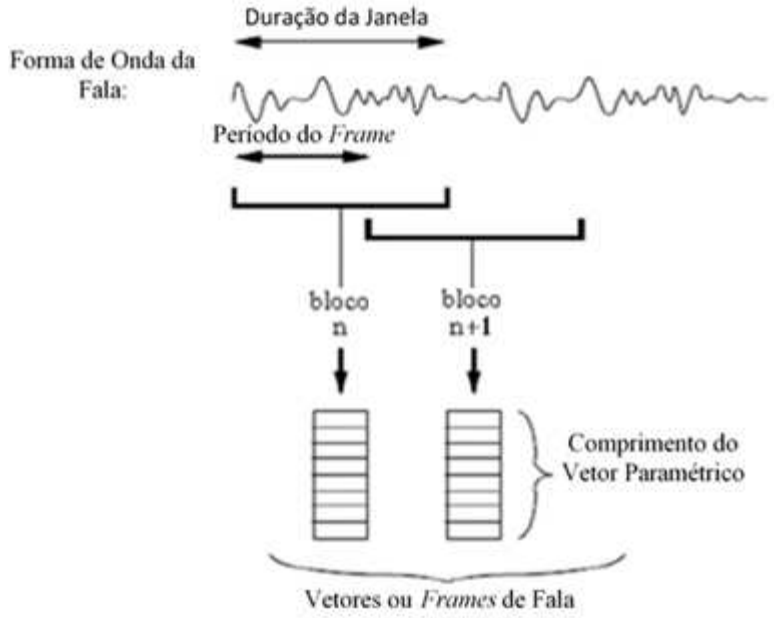

Figura 3.4 - Enquadramento de sinais digitalizados (adaptado de YOUNG et al., 2006).

A segmentação do sinal pode promover efeitos de *leakage* (introdução de ruídos em frequência), sendo assim, necessário realizar um janelamento em cada frame, minimizando as descontinuidades e distorções espectrais do sinal. Este tipo de ruído pode ocorrer tanto no início quanto no fim de cada frame, sendo a magnitude igual a zero nesses limites (PROAKIS e MANOLAKIS, 1996).

Definindo a janela como *w(n),* sendo *n* e *i* números inteiros, o resultado do janelamento a cada *frame* (y_i) é a multiplicação do sinal discretizado $(x(n))$ por essa função, conforme mostra a Eq. 3.1:

$$y_i(n) = x_i(n) \cdot w(n) \qquad 0, \leq n \leq N-1 \text{ e } 1 \leq i \leq I \tag{3.1}$$

Onde N é o número de amostras a cada frame e I indica o *frame* para o janelamento. O conjunto de *frames* multiplicados pela função janela forma o sinal discretizado, com janelamento denotado por $\widetilde{Y}$.

$$\widetilde{Y} = \{y_i\}, \qquad 0 < i < I - 1 \tag{3.2}$$

As funções de janela mais utilizadas para sistemas de reconhecimento de voz são a *Hamming*, e a *Hanning*. Optou-se por utilizar nessa pesquisa a janela do tipo *Hamming* (definida pela Eq. 3.3), pois essa não anula os valores em suas extremidades, o que ocorre no caso da janela *Hanning*, conforme mostra a Figura 3.5.

$$w(n) = 0{,}54 - 0{,}56\cos\left(\frac{2\pi n}{N-1}\right), \qquad 0, \leq n \leq N-1 \tag{3.3}$$

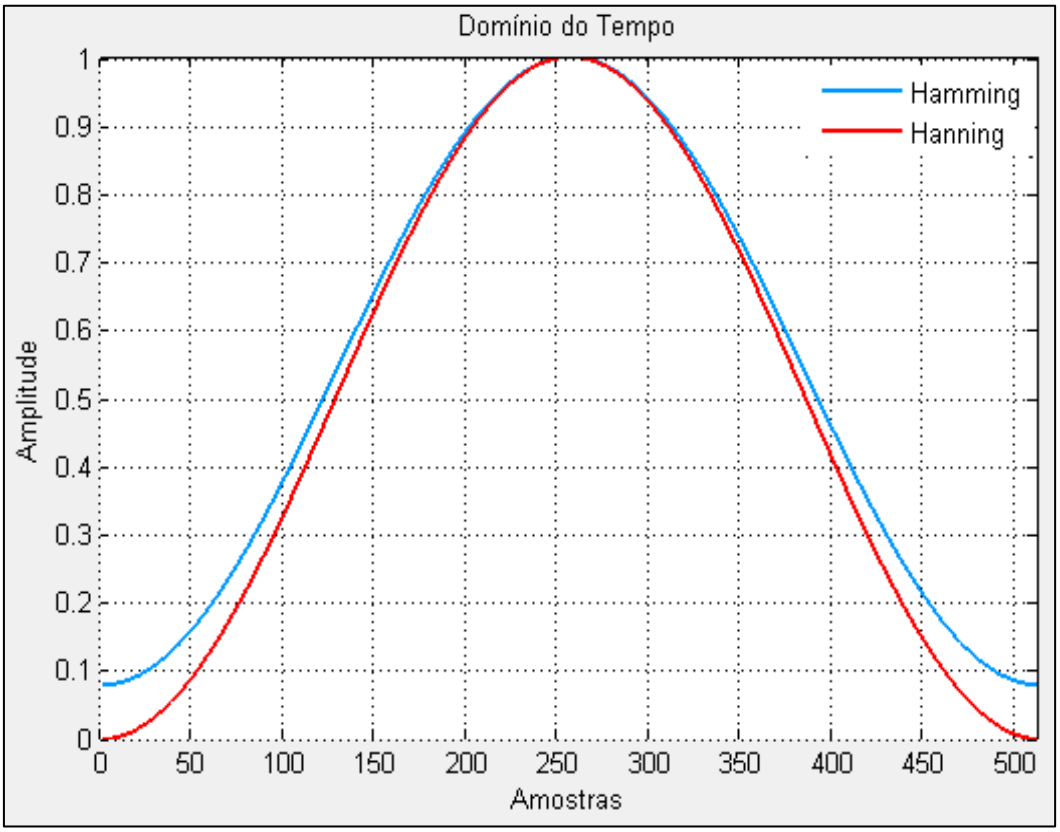

Figura 3.5 - Funções *Hamming* e *Hanning* de janelamento para 512 amostras.

3.2.3 Transformada Rápida de Fourier (FFT)

Após o janelamento do sinal, é necessário calcular o espectro de potência de cada frame por meio de um periodograma, que identifica quais frequências estão presentes naquele sinal, operação similar à tarefa dedicada à cóclea humana. Cada *frame* é convertido do domínio do tempo para o da frequência, utilizando uma Transformada Rápida de Fourier (FFT) (DUHAMEL e VETTERLI, 2010). A FFT é um algoritmo eficiente baseado na Transformada Discreta de Fourier (DFT) (JENKINS, 2010) e pode ser definida sobre o conjunto de N amostras em cada frame, segundo a Eq. 3.4 (NGOC et al., 2011):

$$X_i = \left\{ \sum_{n=0}^{N-1} \left(y_i(n) \cdot e^{-j2\pi i/N} \right) \right\}, \qquad i = 0,1,2,...,I-1 \tag{3.4}$$

Dos resultados obtidos, calcula-se a magnitude das frequências, uma vez que X_i contém valores Reais e Imaginários. O conjunto de *frames* transformados para o domínio da frequência é chamado de periodograma

espectral e seus componentes ainda contêm informações inúteis para sistemas de reconhecimento automático de voz.

3.2.4 Banco de Filtros Mel

O próximo passo do processamento consiste em adaptar a resolução de frequência utilizando uma escala que satisfaça as propriedades do sistema auditivo humano, uma vez que este não interpreta propriedades perceptuais do som e sua altura (tom sonoro) em uma relação linear de frequências ao longo do espectro audível (MOLAU et al., 2001). Em 1937, os pesquisadores Stevens, Volkmann, e Newman, desenvolveram uma escala chamada de frequências Mel com objetivo de compensar este problema. O nome Mel deriva da palavra '*melodia*'.

Por meio de um estudo comparativo com o aparelho de audição humano foi definido essa escala psicoacústica igualmente espaçada caracterizada pela sensibilidade do ouvido para diferentes frequências do espectro audível, isto é, caracterizada pela capacidade de identificação da frequência fundamental, e, portanto, da altura de um som. O experimento realizado mostrou que existe uma relação linear entre essa escala pscicoacústica e a escala Hertz apenas na amplitude de 0 a 1000 Hz. Acima deste intervalo a escala torna-se logarítmica e pode ser

aproximada segundo a fórmula apresentada pela Equação (3.5), em que $mel(f)$ é o resultado de frequência em *mels* e f é frequência nominal medida em Hz.

$$mel(f) = 1127 \cdot \ln(1 + f/700) \tag{3.5}$$

O ponto de referência entre a escala Mel e a medida da frequência nominal é definida atribuindo-se uma altura perceptiva de 1000 mels a um tom de 1000 Hz, 40 dB abaixo do limiar de audição humano. Ouvintes julgaram que acima de 500 Hz, intervalos cada vez maiores produzem incrementos de altura equivalentes. Como resultado, quatro oitavas na escala de Hertz acima de 500 Hz compreendem a aproximadamente duas oitavas na escala Mel (STEVENS et al., 1937). A Figura 3.6 mostra a relação entre a escala Mel e a escala de frequênciais nominais em Hertz.

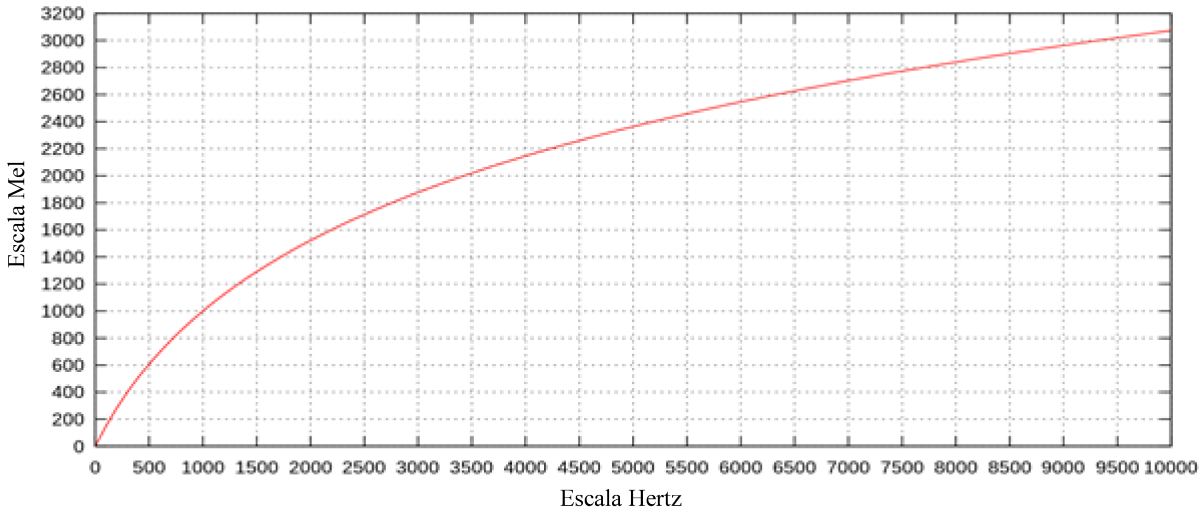

Figura 3.6 - Relação entre a escala Mel e Hertz.

A percepção auditiva humana é influenciada pela quantidade de energia em uma banda crítica de uma frequência particular. A largura das bandas críticas variam com as frequências, e quando são combinadas com a escala Mel, as distribuições de frequências dessas bandas tornam-se linear. Assim, as bandas críticas podem ser comparadas a um conjunto de filtros passa-banda justapostos ajustados em torno da frequência central.

Como no periodograma espectral há excesso de informações, os bancos de filtros eliminam a maior parte dos ruídos do sinal que não têm funcionalidade para a operação de reconhecimento de voz, além de concentrar toda a energia contida nessas bandas de frequência,

uma vez que o periodograma é incapaz de discernir frequências estreitamente espaçadas (GUPTA et al., 2013). Este cálculo é feito através de bancos de filtros Mel somando os coeficientes do periodograma das respectivas frequências através de uma função triangular.

A integração da escala Mel com o uso de bancos de filtros com funções triangulares comprovadamente amplia o desempenho de sistemas de reconhecimento de voz (O'SHAUGHNESSY, 2000).

Nessa pesquisa usou-se 26 filtros com envelopes triangulares. O primeiro filtro é bastante estreito e sua função é calcular a energia contida em regiões de pequenas frequências, próximas a 0 Hz. Conforme as frequências aumentam, os filtros subsequentes tornam-se mais largos, uma vez que não há grande importância com as frequências mais elevadas para questões de reconhecimento de voz (GUPTA et al., 2013).

Os filtros são espaçados através de suas frequências em aproximadamente 100 *mel*, e apresentam uma extensão que normalmente varia de 200 a 7 kHz, região onde há maior concentração de energia. A frequência superior é limitada pela frequência de Nyquist (metade da frequência de amostragem), neste trabalho igual a 11,025 kHz (TIWARI, 2010). A Figura 3.7 mostra o banco de filtros Mel construído para essa pesquisa.

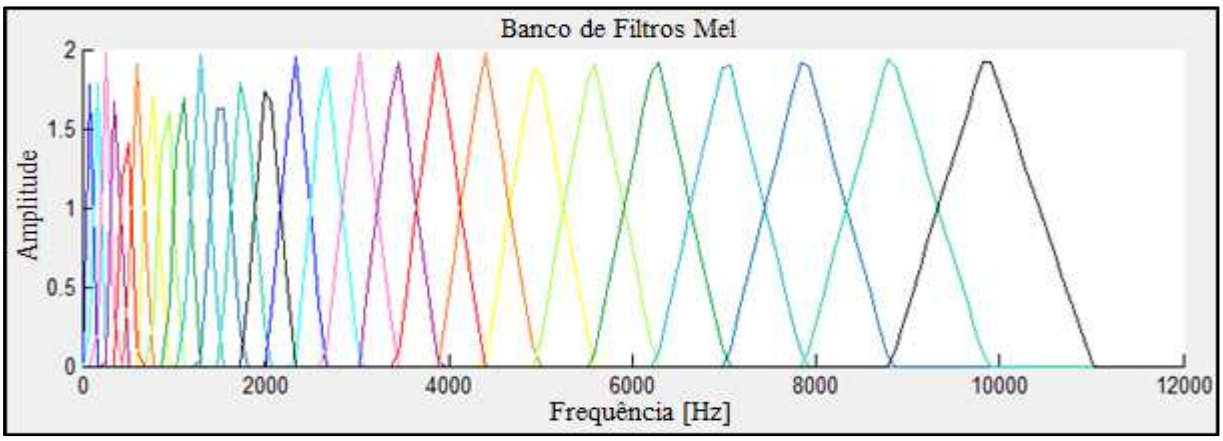

Figura 3.7 - Banco de filtros Mel com 26 filtros.

3.2.5 Mel Frequency Cepstral Coefficients

Sinais de voz são formados pela convolução no tempo entre o sinal de excitação produzido pela traqueia e a resposta impulsiva instantânea do trato vocal. Estes sinais, convoluídos no domínio do tempo manifestam-se multiplicados no domínio da frequência. Assim, faz-se necessário a aplicação de uma função logarítmica sobre o espectro obtido, transformando a multiplicação em uma soma, e assim, permitindo a segregação linear dos

componentes do sinal de excitação e do sinal de modulação do trato vocal. A Figura 3.8 apresenta o Espectro de Potência e o Espectro de Potência Logarítmico referente à elocução da palavra "*Engenharia*".

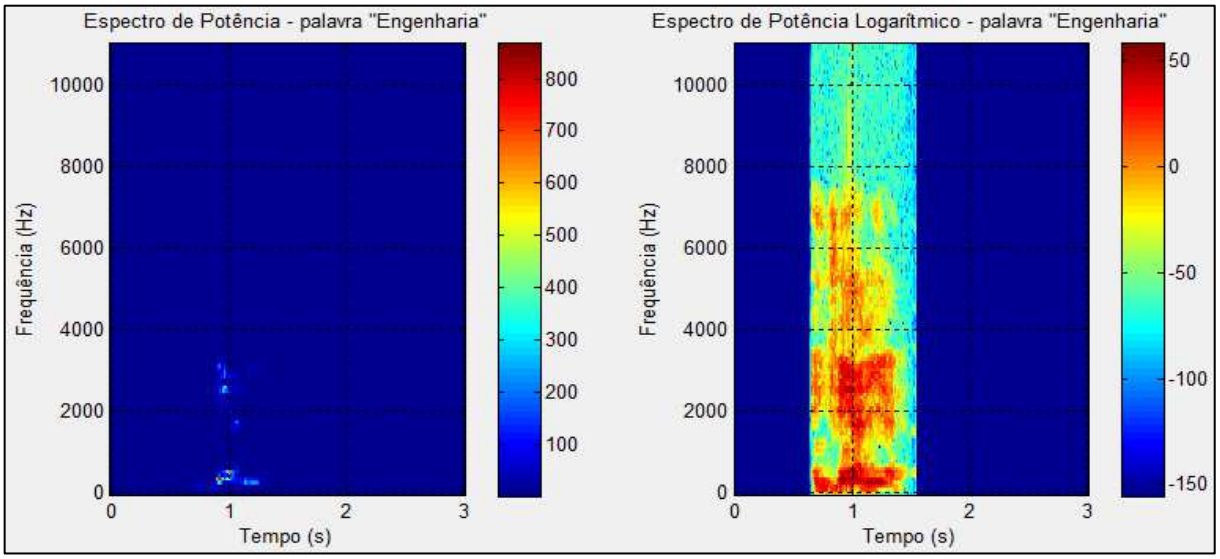

Figura 3.8 - Espectros de potência (à esq.) e com escala logaritmica (à dir.) - palavra *Engenharia*.

A escala logaritmica ajusta os parâmetros associados a não lienaridade da escala Mel, uma vez que para frequências superiores a 1000 Hz, não há linearidade das amostras (GUPTA et al., 2013).

O passo final é converter os dados até aqui obtidos de volta para o domínio do tempo, aplicando uma Transformada Discreta do Cosseno (DCT), obtendo os *Mel Frequency Cepstral Coefficients* (PROAKIS e MANOLAKIS, 1996). Se denotarmos os coeficientes de espectro de potência Mel, calculados no passo anterior como $\tilde{S}_k$, $k = 0,2,...,K-1$, em que K é o número total de coeficientes ($K=26$), os *Mel Frequency Cepstral Coefficients*, denotados como $\tilde{c}_n$, podem ser calculados conforme a Eq. 3.6 (HUANG et al., 2001).

$$\tilde{c}_n = \sum_{k=1}^{K} \log\left(\tilde{S}_k\right) \cdot \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \qquad n = 0,1,2,...,K-1 \tag{3.6}$$

O espectro logarítmico gerado pela resposta impulsiva do trato vocal apresenta-se com uma variação suave e de baixa frequência, e em contrapartida, o espectro associado à excitação produzido pela traqueia é altamente variável e apresenta periodicidade, especialmente em

regiões de enunciação de vogais. Assim, a resposta do trato vocal pode ser obtida retendo-se apenas os 13 primeiros *Mel Frequency Cepstral Coefficients* (MAFRA, 2002). A Figura 3.9 ilustra o espectro de potência, bem como o espectro de potência logarítmico modificado através do banco de filtros Mel referentes à palavra "Engenharia".

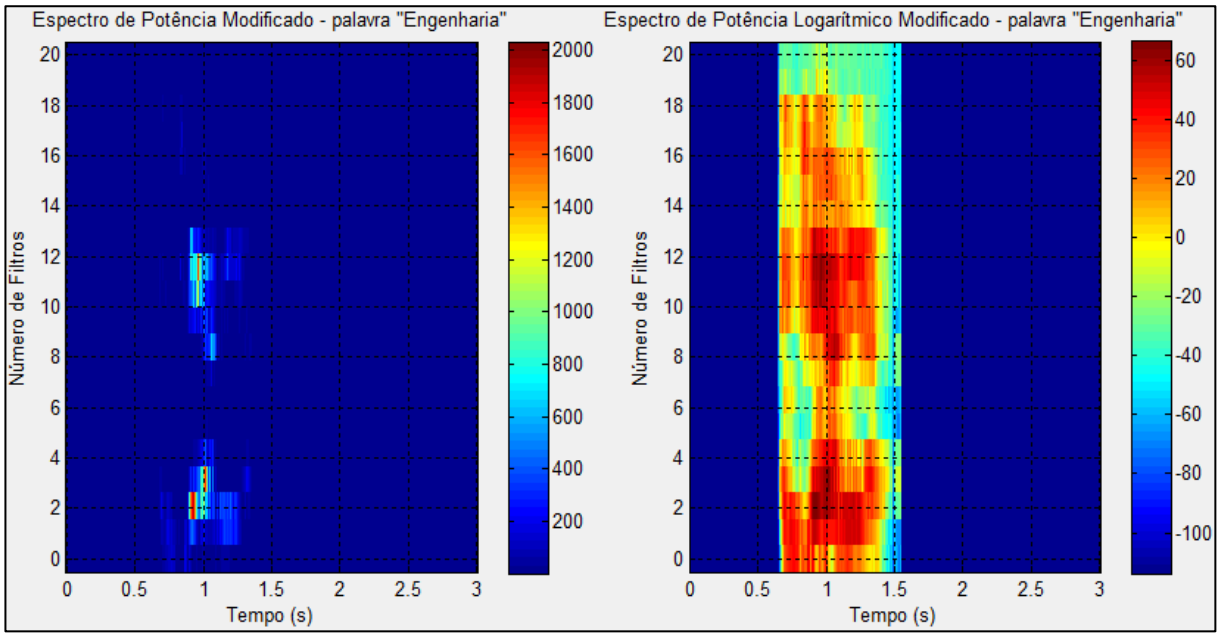

Figura 3.9 - Espectro de Potência (à esq.) e Espectro de Potência Logarítmico (à dir.) modificados através do banco de filtros Mel da palavra "*Engenharia*".

O primeiro coeficiente do espectro de potência também é excluído do cálculo para Transformada Discreta do Cosseno, pois representa o nível DC do sinal de entrada, que contém pouca informação sobre o Locutor (TIWARI, 2010).

A *DCT* dos coeficientes espectrais de potência é calculada por duas razões: para descorrelacionar as energias dos diferentes bancos de filtros, que estão correlacionadas em função do *overlap* entre eles, e assim modelar as características do sinal em um sistema de reconhecimento de padrões; além disso, é uma maneira de aumentar a eficiência do sistema, uma vez que apenas 12 dos 26 coeficientes obtidos pela *DCT* são mantidos (LOGAN, 2000). Coeficientes mais elevados da *DCT* representam rápidas mudanças nas energias dos bancos de filtros, degradando o desempenho de sistemas de Reconhecimento Automático de Voz.

Após a realização dos passos supracitados, obtém-se os cepstros correspondentes ao sinal de fala contendo informações acústicas representativas daquele sinal. O conjunto de cepstros integram o vetor acústico, que será processado utilizando Quantização Vetorial. A Figura 3.10 apresenta o vetor acústico correspondente à elocução da palavra "*Engenharia*".

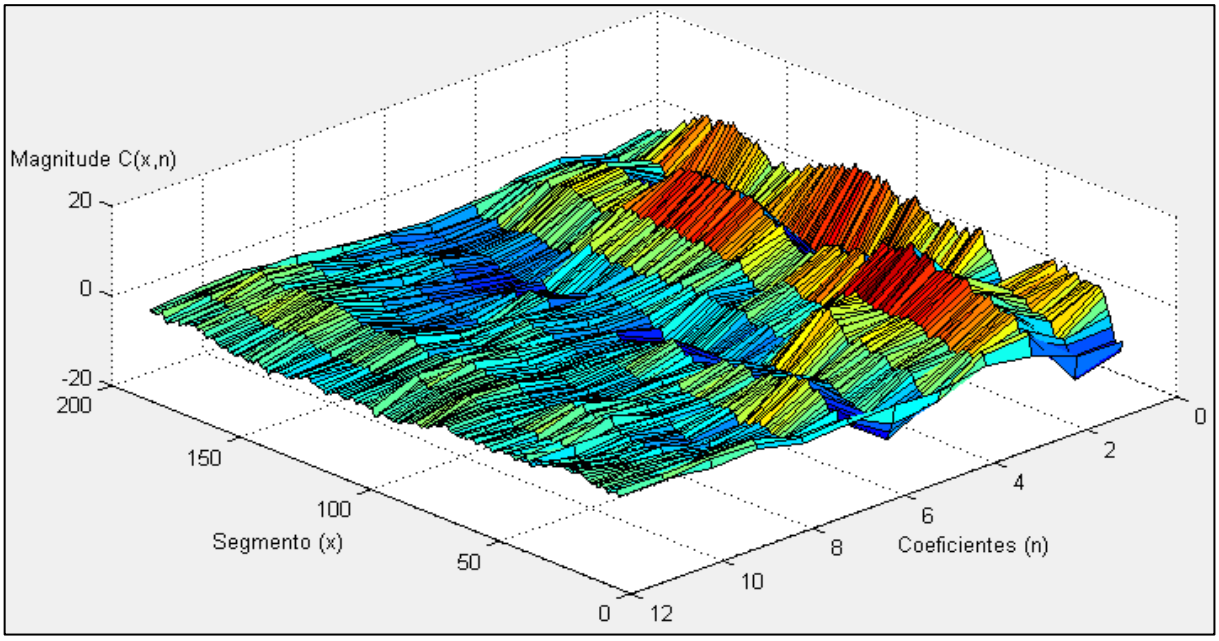

Figura 3.10 - Vetor acústico correspondente à palavra "*Engenharia*".

3.3 Classificação e Reconhecimento de Padrões – Quantização Vetorial (VQ)

A construção do sistema de reconhecimento de voz ainda depende de outra etapa, o reconhecimento de padrões. Trata-se, para o sistema em escopo, de um mecanismo de análise de correspondência entre dois vetores acústicos.

O reconhecimento de padrões de voz é dividido em dois momentos. O primeiro é a classificação dos vetores acústicos e o segundo é a comparação desses com outros vetores presentes em um banco de dados (LANDELL et al., 1984). Desta forma, anteriormente ao funcionamento do sistema, é necessário classificá-lo. Isso é feito, criando-se um banco de dados composto pelos vetores acústicos de diferentes sinais de voz que carregam consigo informações fonéticas de interesse. Essas informações serão posteriormente utilizadas como padrões para comparação com outros sinais de voz. Assim, novos sinais de entrada no sistema serão processados, classificados, e então comparados com o banco de dados do sistema. Se houver similitude entre os vetores, o sistema fornecerá uma resposta com identificação ou reconhecimento daquele sinal. Caso contrário, o sistema não identificará o sinal em análise como proveniente de outro que foi utilizado para classificar o mesmo.

O banco de dados é criado durante a fase de treinamento do sistema. Nessa fase, fonemas, palavras e sentenças de interesse são processadas, transformadas em vetores acústicos e gravadas no banco de dados. O mesmo acontece com sinais de voz de locutores que desejarem

ser reconhecidos (ARYA et al., 1993). Após a criação do banco de dados, o sistema pode ser testado, isso é, novos sinais de voz serão processados e comparados com aqueles previamente armazenados.

Como ferramenta para realização da classificação dos sinais e comparação dos mesmos, usou-se a Quantização Vetorial. Essa técnica foi escolhida, pois a mesma apresenta rápida compilação por operar com dados comprimidos, por ser fácil de implementar computacionalmente e devido ao seu elevado índice de acertos quando utilizada como mecanismo comparador de padrões (MAKHOUL et al., 1985).

A técnica funciona como um mecanismo de mapeamento de um espaço vetorial para classificação do sistema e com modelagem de funções de densidade probabilística para estudar os diferentes graus de similitude entre os vetores acústicos.

Durante a etapa de classificação, cada vetor acústico dos sinais para armazenamento ocupa pequenas regiões do espaço vetorial. Essas regiões são governadas por um centroide. As regiões são chamadas de *clusters* e seu centroides de *codewords*. Uma coleção de *codewords* forma um *codebook* (GERSHO e GRAY, 1992), ou seja, um conjunto de regiões que caracterizam o espaço vetorial do sistema.

O método utilizado para construção dos *codebooks* é o algoritmo LBG (LINDE et al., 1980). Esse algoritmo opera agrupando conjuntos de vetores acústicos em uma matriz formando *codebooks*. A Figura 3.11 ilustra um espaço vetorial bidimensional implementado por esse algoritmo. Na figura há 18 células correspondentes a regiões específicas do espaço vetorial. Cada célula representa um *cluster* que é governada por uma *codeword*, o centroide, que está representado pelos pontos pretos. Destaca-se o *cluster* C_1, governado pela *codeword* y_i, cujo conglomerado forma um *codebook*.

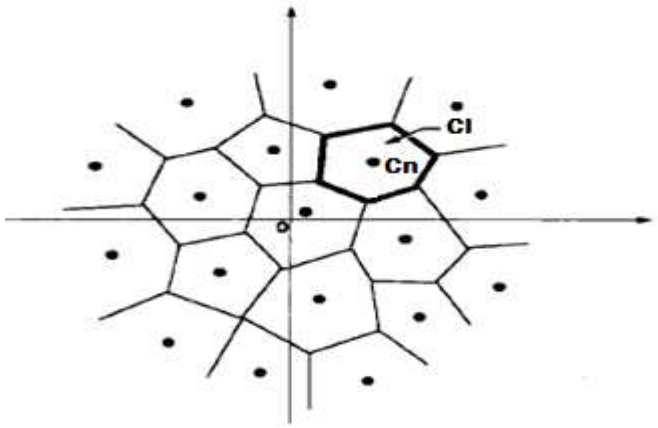

Figura 3.11 - Representação de um *codebook* bidimensional construído utilizando a ferramenta Quantização Vetorial (adaptado de MAKHOUL et al., 1985).

O processo de implementação do algoritmo LBG ocorre de maneira recursiva e está apresentado pelo fluxograma da Figura 3.12. explicado na sequencia.

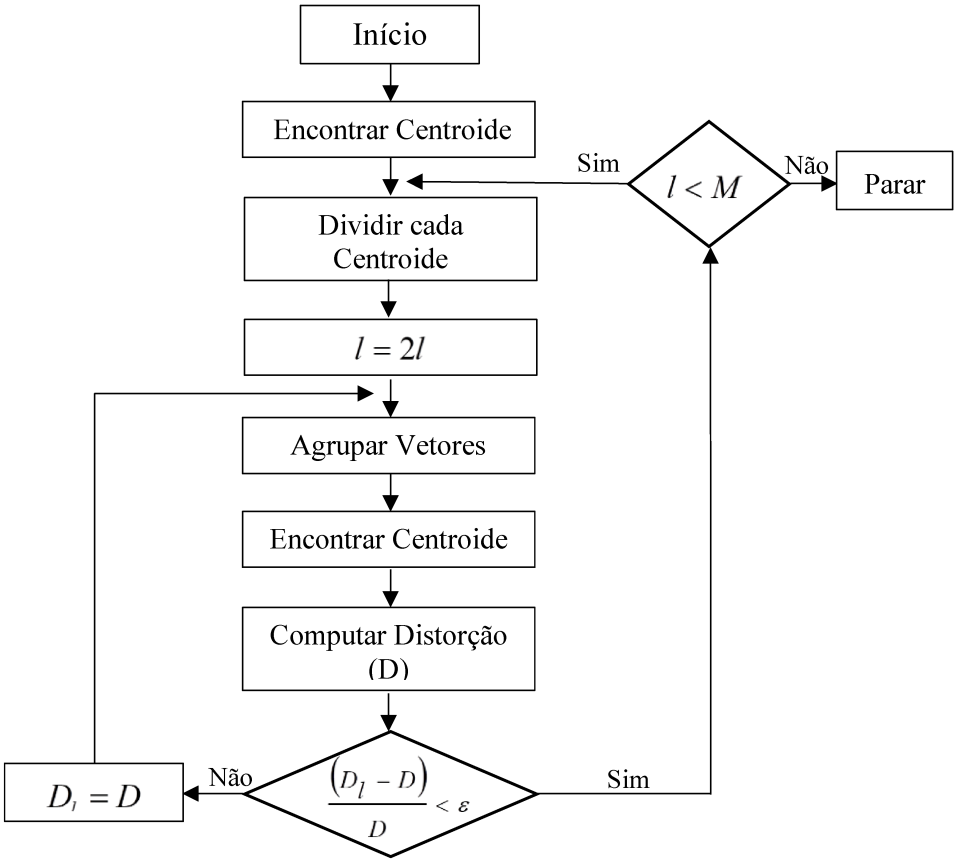

Figura 3.12 - Fluxograma do algoritmo LBG (adaptado de KABIR e AHSAN, 2007)

1. Primeiramente, utiliza-se um vetor acústico aleatório que será lançado no espaço vetorial vazio, e então calcula-se seu centroide. Tem-se assim, um *codebook* composto por 1 vetor. No fluxograma a variável l representa o número de centroides presentes no *codebook*, ao passo que a variável M indica o número máximo de centroides. Usou-se a recomendação dos autores LINDE, BUZO e GRAY, escolhendo-se $M = 16$.
2. Em seguida, dobra-se o tamanho do *codebook*, repartindo cada centroide, ou *codeword* (Cd) utilizando o parâmetro ε, também chamado de parâmetro de repartição do algoritmo LBG. As equações (3.7) apresentam o resultado.

$$Cd_l^+ = Cd_l(1+\varepsilon) \tag{3.7.a}$$

$$Cd_l^- = Cd_l(1-\varepsilon) \tag{3.7.b}$$

Seguindo orientação dos autores LINDE, BUZO e GRAY, usou-se $\varepsilon = 0{,}01$.

3. O terceiro passo é determinar o vizinho mais próximo, ocorrendo o agrupamento de vetores, ou seja, para cada vetor, procura-se o centroide no presente *codebook* mais próximo e o assimila àquela correspondente célula (RAMACHANDRAN, 2010).
4. O quarto passo, refere-se à atualização do centroide em cada célula, uma vez que novos vetores foram atribuídos à elas.
5. O quinto passo consiste da primeira iteração. Repete-se os passos 3 e 4 até que a distância média entre vizinhos (D), isso é, entre um vetor acústico e um centroide, seja encontrado com valor inferior ao parâmetro ε, ou seja, inferior a 0,01.
6. Todo o processo à partir do passo 2 é repetido até que a matriz representativa do *codebook*, matriz M, adquira o número de centroides desejado (BHARTI e BANSAL, 2015).

Todos os vetores acústicos formados durante a fase de treinamento são inseridos no processo iterativo do algoritmo LBG acima descrito, gerando novos *codebooks*, assumindo regiões específicas do espaço vetorial correspondendo à cada fonema, palavra ou sentença. O conjunto de *codebooks* formará o banco de dados pretendido.

A Figura 3.13 apresenta um diagrama simples composto de dois locutores para exemplificar o processo de criação do *codebook*. Nela, os círculos verdes referem-se ao locutor 1 e os triângulos vermelhos ao locutor 2. Durante a fase de treinamentos, um *codebook* específico para cada locutor é gerado, agrupando os vetores acústicos correspondentes aos sinais de voz de cada locutor. As *codewords* de cada indivíduo estão representadas através de círculos pretos para o locutor 1 e triângulos pretos para o locutor 2. A distância Euclidiana ponderada entre um vetor acústico e o *codebook* mais próximo é denominada distorção VQ, destacada na figura.

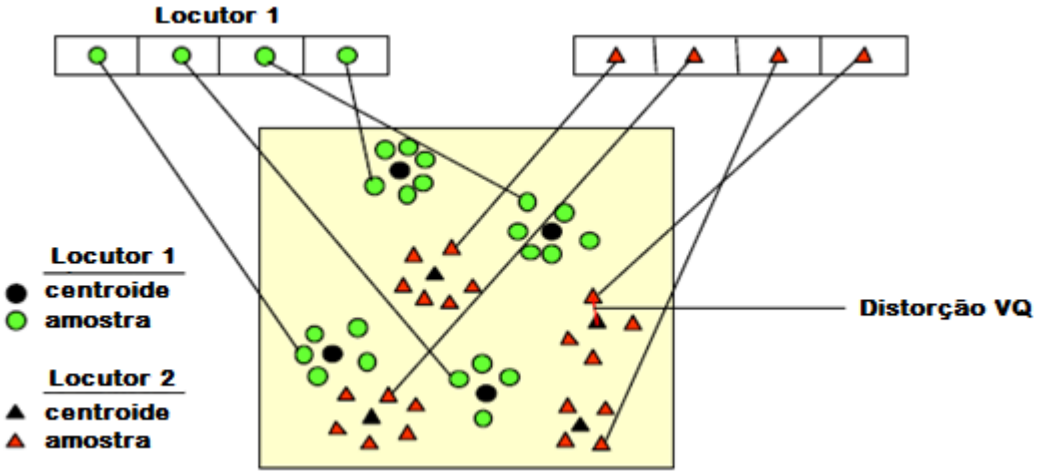

Figura 3.13 - Diagrama VQ com dois locutores (SONG et al., 1987).

Na etapa de reconhecimento de padrões, analisa-se a menor distorção VQ calculada entre o vetor acústico correspondente à um sinal de entrada no sistema e todos os *codebooks* do espaço amostral. Se essa distância for inferior a um limiar (L) previamente definido (usou-se nessa pesquisa $L = 0.2$), associa-se o sinal de entrada àquele *codebook*, havendo portanto, a identificação entre os dois sinais. Caso essa distância seja superior ao limiar, o sistema conclui que não há similitudes significativas para associar o sinal de entrada à qualquer outro sinal armazenado e, portanto, não há reconhecimento (GERSHO e GRAY, 1992).

3.4 Técnicas para robustez do sistema de reconhecimento de voz

A maioria do estado-da-arte de sistemas de reconhecimento de voz que operam com *Mel Frequency Cepstral Coefficients* e Quantização Vetorial apresentam uma ótima performance com projeções de acerto superiores a 90% em ambientes de ruído controlado (KABIR e AHSAN, 2007; PATRA, 2007; RAJSEKHAR, 2008; JAIN e SHARMA, 2013). No entanto, diferenças acústicas durante as fases de treinamento e testes podem afetar diretamente esses resultados (ALAM et al., 2011) deteriorando o desempenho destes sistemas.

A presente pesquisa almeja o desenvolvimento de um modelo de reconhecimento de voz robusto e efetivo, capaz de minimizar fatores causadores de erros que possam degradar os resultados na configuração adotada com MFCC e VQ. Com esse objetivo, implementou-se no sistema: a técnica *Voice Activity Detection* (VAD) que remove regiões do sinal que não contenham vocalização, isso é, informações referentes à fala do locutor; análise dinâmica dos

coeficientes cepstrais, para a qual foi testada as três seguintes diferentes ferramentas, *Delta-Coefficients* (DC), *Delta-Delta-Coefficients* e Shifted-Delta Cepstra (SDC); além de técnicas de normalização dos coeficientes cepstrais para compensar defeitos de captação dos sinais. Para isso também foi testado três diferentes técnicas, *Cepstral Mean Normalization* (CMVN); *Windowed Cepstral Mean and Variance Normalization* (WCMVN); e *Short-term Gaussianization* (STG).

A Figura 3.14 apresenta um diagrama correspondente às diferentes técnicas utilizadas neste trabalho.

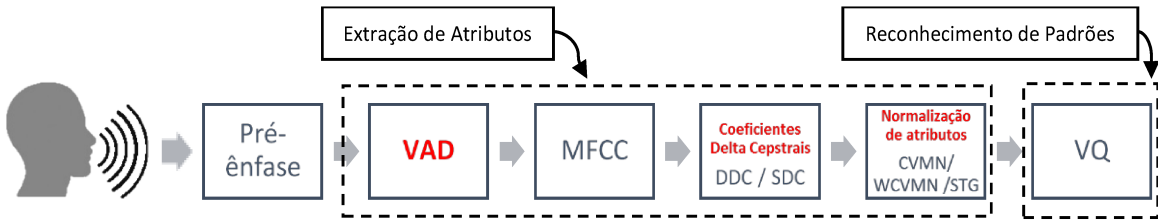

Figura 3.14 - Diagrama do Sistema de Reconhecimento de Voz Robusto implementado nessa pesquisa.

3.4.1 Voice Activity Detection - *VAD*

A técnica *Voice Activity Detection*, ou detecção de voz ativa, como o próprio nome sugere, remove porções do sinal silenciadas ou que apresentam ruídos. A técnica considera ruídos não apenas regiões aperiódicas do sinal, mas também ressonâncias (RAMIREZ, 2007).

Um sinal de fala é caracterizado por três regiões: porções silenciosas; porções não vocalizadas; e porções vocalizadas. Nenhuma fala é produzida nas porções silenciosas. As regiões não vocalizadas ocorrem quando pronunciamos as letras '*e*', 'f', 'g', 's', 'p' e 'ch' no alfabeto da língua potuguesa. Nessa região de espectro contínuo, a forma da onda sonora é aperiódica e aleatória, gerada pelo fluxo de ar na boca modulado pelos maxilares, língua e lábios. Diferentemente, nas partes vocalizadas, as cordas vocais são tensionadas e vibram quando há passagem de ar, resultando em uma forma de onda quase-periódica do sinal de voz (PATRA, 2007; AKILA e CHANDRA, 2014).

O processo de isolamento das informações redundantes, ou seja, regiões afetadas pelo ruído, partes com pronunciamento não vocalizado e regiões silenciadas, é convenienete para reduzir a dimensão do sinal digitalizado e, portanto, reduzir a complexidade bem como o tempo computacional nas etapas subsequentes. Este processo não degrada as partes vocalizadas

(RAMIREZ, 2007). Além disto, sistemas de reconhecimento de voz demandam eficiência na extração das propriedades acústicas mesmo em ambientes ruidosos (PATRA, 2007).

A segmentação do sinal de fala nessas três regiões é laboriosa e nem sempre exata com relação à distinção entre regiões silenciadas e não vocalizadas, no entanto, não ocorre deteriorização das características acústicas do sinal. Garantindo, portanto ser uma técnica confiável, robusta por atender qualquer tipo de sistema de reconhecimento de voz e que demanda pouco tempo computacional, permitindo processamento em tempo real dos sinais de entrada (MOATTAR e HOMAYOUNPOUR, 2009).

A implementação da técnica VAD pode ser realizada através de diversas metodologias. As mais utilizadas são citadas a seguir (AKILA e CHANDRA, 2014):

1. *Short-Time Energy* (STE);
2. *Zero Crossing Rate* (ZCR);
3. *Mel Frequency Cepstral Coefficients*;
4. *Delta Line Spectral Frequencies* (DLSF);
5. *Higher order statistics*.

Nessa pesquisa, implementou-se a ferramenta VAD usando-se um algoritmo que combina as metodologias STE e ZCR, modelo desenvolvido por Qiang He em 2001 e que está demonstrado pelo fluxograma apresentado pela figura 3.15. Este foi escolhido para ser utilizado nessa pesquisa por sua simplicidade, facilidade de implementação e resultados satisfatórios quando utilizados sobre um sinal de voz (AKILA e CHANDRA, 2014).

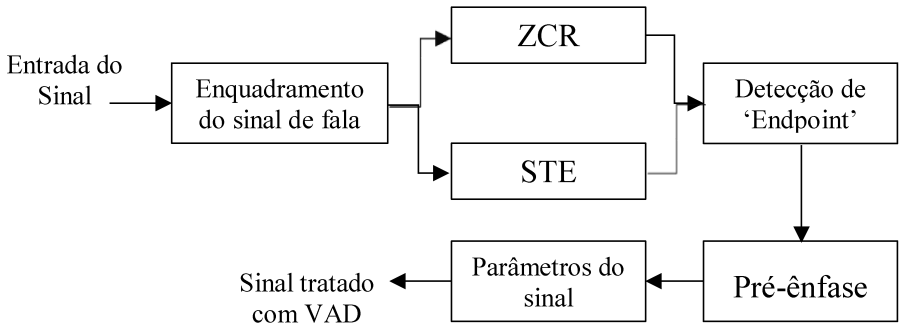

Figura 3.15 - Diagrama do algoritmo de Qiang He (adaptado de QIANG e YOUWEI, 1998).

O algoritmo de Qiang He recebe o sinal de voz de entrada e reparte-o em *frames* sobrepostos a cada 80 ms. Em seguida, calcula o espectro de energia (STE) e as taxas de cruzamentos por zero (ZCR) de cada *frame*. Os resultados de STE e ZCR são então comparados com valores limiares para caracterização de regiões silenciadas, vocalizadas e não vocalizadas.

AKILA e CHANDRA (2014) propõem que os valores para os limiares sejam iguais a 2 e 10. Para o caso dos valores de STE e ZCR serem ambos inferiores a 2, implica que o *frame* em análise é uma região silenciada. No caso de ambos apresentam valores entre 2 e 10, pode se tratar de uma região de ruído (pronunciação sem vibrações das cordas vocais) ou região de vocalização. Essa diferenciação é realizada através de uma contagem de *frames*. Isso é, se o número de *frames* sucessivos, cujos valores de STE e ZCR estão dentro do limiar estabelecido for inferior a 150 ms, considera-se essa região do sinal como ruidosa, e se superior, considera-se uma região vocalizada. E, para o caso de ambos os valores serem superiores a 10, trata-se de uma região vocalizada (AKILA e CHANDRA, 2014).

3.4.1.1 Short-Time Energy (STE)

A amplitude do sinal de fala varia com o tempo e está diretamente relacionada à quantidade de energia do sinal. As porções compostas por pronunciações não vocalizadas apresentam baixas amplitudes e, consequentemente, energias muito inferiores às porções vocalizadas, regiões de altas amplitudes (ENQING et al., 2002). A energia (E) transportada pela onda sonora é diretamente proporcional ao quadrado da amplitude (A) da onda. Assim, a energia em um curto espaço de tempo (*Short-Time Energy*) pode ser calculada utilizando-se a fórmula especificada pela Eq. (3.8) (PATRA, 2007; AKILA e CHANDRA, 2014).

$$E = \sum_{i=1}^{I} \left| A(i)^2 \right| \tag{3.8}$$

Em que $A(i)$ é a amplitude de cada *frame*. I é o número de *frames* do sinal de fala.

3.4.1.2 Zero Crossing Rate (ZCR)

Taxa de Cruzamentos por Zero (ZCR) é comumente utilizada em processamento de sinais, referindo-se a um ponto onde a polaridade de um sinal muda ao se cruzar um determinado eixo, conforme ilustrado pela figura 3.16.

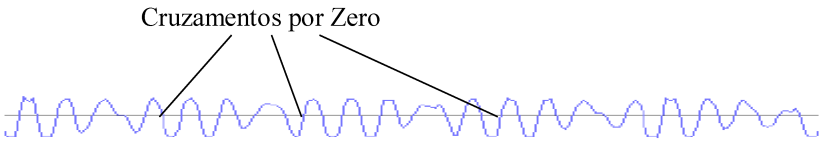

Figura 3.16 - Cruzamentos por zero de um sinal de fala (adaptado de AKILA; CHANDRA, 2014)

ZCR é uma medida da quantidade de vezes que a amplitude de um sinal passa através do eixo das abscissas (valor zero) em um dado *frame*, isso é, é uma medida do conteúdo de frequências de um sinal de banda estreita (LOKHANDE et al., 2001). ZCR é maior em regiões de pronunciação sem vibração do sinal e é expressa de acordo com a Eq. (3.9) (PATRA, 2007).

$$ZCR = \frac{1}{I-1}\sum_{i=1}^{I-1}\left|\frac{\mathrm{sgn}(A(i)) - \mathrm{sgn}(A(i-1))}{2}\right| \tag{3.9}$$

Onde $\mathrm{sgn}(A(i))$ refere-se à polaridade da amplitude do *i-ésimo frame*, recebendo o valor de 1 para o caso positivo $(A(i) > 0)$, ou -1 para o caso negativo $(A(i) < 0)$. I é o número de *frames* do sinal de fala.

3.4.2 Coeficientes Cepstrais Dinâmicos

Os vetores acústicos obtidos usando MFCC descrevem apenas o envelope de potência espectral de cada *frame*, embora o sinal de fala apresente propriedades dinâmicas em sua composição, como trajetórias dos coeficientes ao longo do tempo. Para conservar essa propriedade dos sinais de fala, foram analisadas as técnicas *Delta-Delta Coefficients* (DDC) e *Shifted Delta Coefficients* (SDC) (KUMAR et al., 2011).

3.4.2.1 Delta-Delta Coefficients (DDC)

Delta-Delta Coefficients são atributos que contém informações dinâmicas extraídas de coeficientes estacionários, como no caso de *Mel Frequency Cepstral Coefficients*. DDC são chamados de coeficientes cepstrais de aceleração e obtidos à partir dos *Delta-Coefficients* (DC). Esses são calculados segundo a Eq. (3.10).

$$d_i = \frac{\sum_{n=1}^{2} n\left(\widetilde{c}_{n\,i+n} - \widetilde{c}_{n\,i-n}\right)}{2\sum_{n=1}^{2} n^2} \tag{3.10}$$

Em que d_i é um *Delta-Coefficient* extraído do *i-ésimo* frame; $\widetilde{c}_{n_{i+n}}$ e $\widetilde{c}_{n_{i-n}}$ são coeficientes estáticos, ou *Mel Frequency Cepstral Coefficients*.

Delta-Delta Coefficients são calculados da mesma forma como apresentado na Eq. (3.10), substituindo os coeficientes estáticos pelos *Delta-Coefficients* (d_{i+n} e d_{i-n}) na fórmula (MUDA et al., 2010).

Ao final serão obtidos 24 coeficientes dinâmicos: 12 coeficientes DC; e 12 coeficientes DDC. KUMAR, KIM, e STERN (2011) em seus estudos comparativos dos coeficientes dinâmicos, obtiveram melhores resultados quando utilizavam ambos os coeficientes DC e DDC anexados aos coeficientes MFCC. O mesmo foi feito nessa pesquisa resultando em um vetor acústico com 36 coeficientes.

A Figura 3.17 apresenta um gráfico produzido pelos autores KUMAR, KIM, e STERN (2011), o qual relaciona a performance dos coeficientes em análise em um sistema de reconhecimento de voz. Para tanto, foi utilizado a *World Error Rate* (WER), uma medida comum da performance de um sistema de reconhecimento de voz comparado à relação sinal-ruído (SNR) expressa em decibéis.

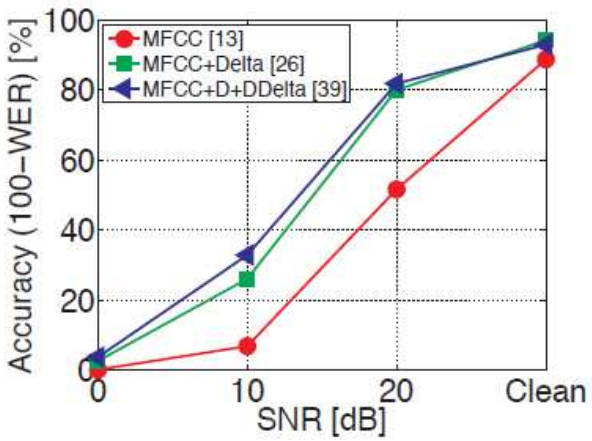

Figura 3.17 - Gráfico comparativo de atributos cepstrais (KUMAR et al., 2011).

Como se observa, em condições da razão SNR igual a 10 dB, a adesão dos coeficientes DDC aos coeficientes MFCC e DC apresenta a melhor performance, que também pode ser notada para a condição SNR igual a 20 dB, ainda que com menos significância com relação ao conjunto MFCC e DC (linha verde do gráfico), mas com razoável diferença para os coeficientes MFCC (linha vermelha do gráfico).

Em condições de um sinal limpo, isso é, sem ruído, os três atributos apresentaram ótima performance, não havendo grande diferença nos resultados WER.

3.4.2.2 Shifted-Delta Coefficients (SDC)

Shifted-Delta Coefficients (SDC) são coeficientes dinâmicos utilizados em sistemas de reconhecimento de voz, com índices de alta performance, como apuraram os estudos de TORRES-CARRASQUILLO et al. (2002). Essa técnica incorpora informações temporais adicionais sobre a fala aos vetores acústicos abrangendo um grande número de *frames*, o que retrata com maior naturalidade as características fonéticas da fala. Esses vetores com características pseudo-prosódicas não necessitam da modelagem de uma estrutura linguística do sinal de fala (ALLEN et al., 2005).

Vetores acústicos SDC são construídos empilhando-se *Delta Coefficients* ao longo de múltiplos *frames* de um sinal de fala. Quatro parâmetros são utilizados para o cálculo de SDC: K, número de coeficientes cepstrais computados por *frame*; d, diferencial temporal de avanço e retardo para a computação dos *Delta Coefficients*; k, número de blocos nos quais os *Delta Coefficients* serão concatenados ao longo do tempo do sinal de fala (t); e p, diferença temporal entre blocos consecutivos.

Nessa pesquisa, utilizou-se 12 coeficientes cepstrais computados por *frame*. Os parâmetros d, k e p foram adotados iguais a 1 ms, 7 blocos e 3 ms; valores sugeridos segundo o estudo de SINGER et al. (2003). A Figura 3.18 mostra um esquema dos *Delta Coefficients* concatenados em blocos formando os vetores acústicos SDC.

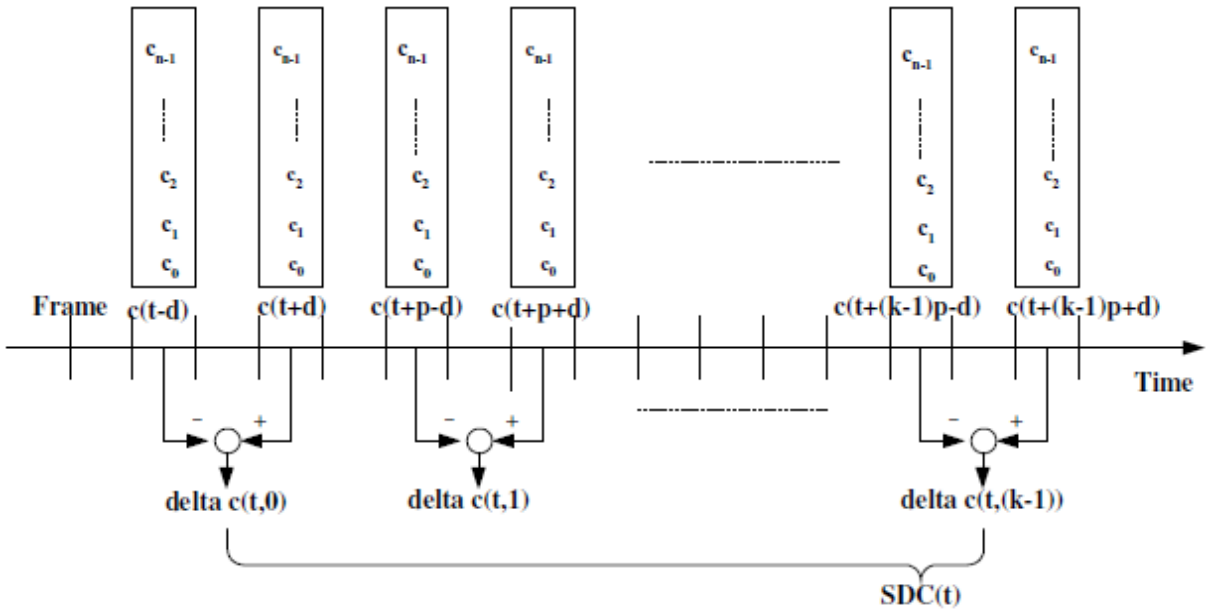

Figura 3.18 - Representação esquemática da computação dos atributos SDC (RONG, 2006).

3.4.3 Normalização dos Coeficientes Cepstrais

Modelos de normalização dos coeficientes cepstrais foram incorporadas ao sistema para compensar efeitos de ruídos externos ou provenientes do canal de entrada. Essas técnicas não necessitam de conhecimento a *priori* dos atributos e são facilmente adaptáveis aos sistemas de reconhecimento de voz (ALAM et al., 2011).

As técnicas de normalização podem ser diferenciadas entre dois grupos. O primeiro grupo atua normalizando propriedades estatísticas, como a média, a variância e o desvio padrão de um sinal discretizado. Nessa categoria estão técnicas como: *Cepstral Mean Normalization* (CMN); *Cepstral Mean and Variance Normalization* (CMVN); *Windowed Cepstral Mean and Variance Normalization* (WCMVN); *Short-time Mean and Variance Normalization* (STMVN); e *Short-time Mean and Scale Normalization* (STMSN).

No segundo grupo, estão as técnicas baseadas na normalização da distribuição de dados, e portanto destinam-se a normalizar a distribuição do vetor acústico. As técnicas de normalização do Histograma e *Short-term Gaussianization* (STG) são exemplos dessa classe. Nessa pesquisa investigou-se três diferentes técnicas de normalização: CMNV, WCMNV e STG. Essas técnicas são consideradas robustas e eficientes (XIANG et al.; 2002; ZHENG et al., 2006; ALAM, et al., 2011).

3.4.3.1 Cepstral Mean and Variance Normalization (CMVN)

A técnica *Cepstral Mean and Variance Normalization* é robusta contra os ruídos aditivos nos canais de entrada e apresenta resultados com mínima distorção produzida por contaminação de ruído ambiente (ZHENG et al., 2006).

CMVN transforma linearmente os coeficientes cepstrais para que todos tenham as mesmas propriedades estatísicas: média igual a zero e variância unitária. Esse método não requer conhecimento *a priori* das estatísticas do ruído do sinal, pode ser implementado independentemente do uso de técnicas VAD, e adapta-se rapidamente à condições de variações de ruído, o que a torna uma ferramenta importante em sistemas de reconhecimento de voz (PRASAD e UMESH, 2013).

Para o cálculo dos coeficientes cepstrais normalizados com CMVN, primeiramente computa-se a média (μ) e a variância (σ^2) de cada coeficiente ao longo de todos os *frames* do sinal. Isso é, seja I o número de *frames* do sinal, e K o número de coeficientes cepstrais por

frame, calcula-se a média e a variância de cada coeficiente, segundo as Equações (3.11.a) e (3.11.b) respectivamente (PRASAD e UMESH, 2013).

$$\mu(m) = \frac{1}{I}\sum_{i=1}^{I} \widetilde{c}_{n_i}(m) \qquad 1 \le m \le K \tag{3.11.a}$$

$$\sigma^2(m) = \frac{1}{I-1}\sum_{i=1}^{I}\left(\widetilde{c}_{n_i}(m) - \mu(m)\right)^2 \qquad 1 \le m \le K \tag{3.11.b}$$

Em que $\widetilde{c}_{n_i}(m)$ é o *m-ésimo* coeficiente cepstral do *i-ésimo frame*.

A normalização dos coeficientes é o próximo passo para adequar as propriedades estatísticas de média igual a zero e variância unitária de cada coeficiente cepstral. Para tanto, será utilizado o desvio padrão (σ) de cada coeficiente, que é calculado á partir da variância dos mesmos. A Equação (3.12) mostra como é feita a normalização.

$$N\widetilde{c}_{n_i}(m) = \frac{\widetilde{c}_{n_i}(m) - \mu(m)}{\sigma(m)} \qquad 1 \le m \le K \tag{3.12}$$

Em que $N\widetilde{c}_{n_i}$ é o vetor de coeficientes cepstrais do *i-ésimo frame*.

O número de coeficientes $N\widetilde{c}_n$ por *frame* varia, podendo ser igual a 12, quando os coeficientes são MFCC, ou SDC; ou 36, no caso de DDC. O conjunto de vetores $N\widetilde{c}_{n_i}$ forma a matriz $X = \left[N\widetilde{c}_{n_1}, N\widetilde{c}_{n_2}, \cdots, N\widetilde{c}_{n_i}, \cdots, N\widetilde{c}_{n_I} \right]$, vetores acústicos que serão posteriormente analisados usando Quantização Vetorial.

A Figura 3.19 apresenta uma comparação dos histogramas calculados antes e posteriormente a aplicação da técnica de normalização CMVN para o segundo atributo cepstral do vetor acústico da enunciação da palavra “Engenharia”. Como esperado, não há mudanças bruscas no formato generalizado da distribuição comparando-se ambos os casos.

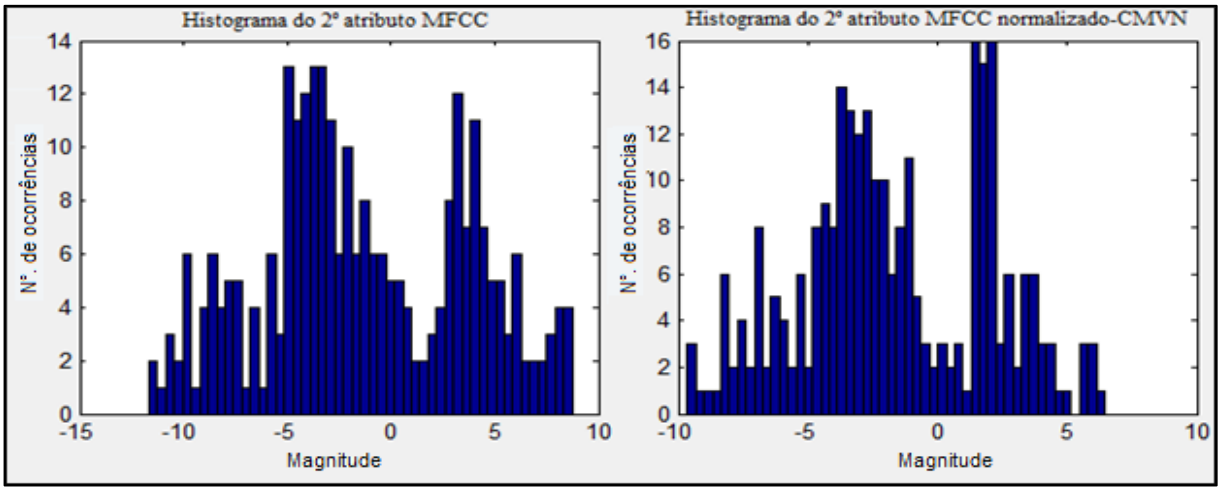

Figura 3.19 - Histogramas MFCC (à esq.) e CMVN (à dir.) do 2° atributo cepstral.

3.4.3.2 Windowed Cepstral Mean and Variance Normalization (WCMVN)

A técnica de normalização *Windowed Cepstral Mean and Variance Normalization* é implementada aplicando-se uma janela deslizante à normalização CMVN. O *frame* médio na janela é normalizado baseado nas mesmas propriedades estatísticas de média igual a zero e variância unitária, como no caso de CMVN (ZHENG et al., 2006).

Nessa pesquisa usou-se a janela deslizante a cada 301 amostras do sinal, valor superior à quantidade de amostras por *frame* (256 amostras) e inferior à soma de amostras de dois *frames*, garantindo que a janela deslize por todos os *frames* do sinal.

A Figura 3.20 compara os histogramas de MFCC e WCMVN do 2º atributo cepstral do vetor acústico da enunciação da palavra “Engenharia”.

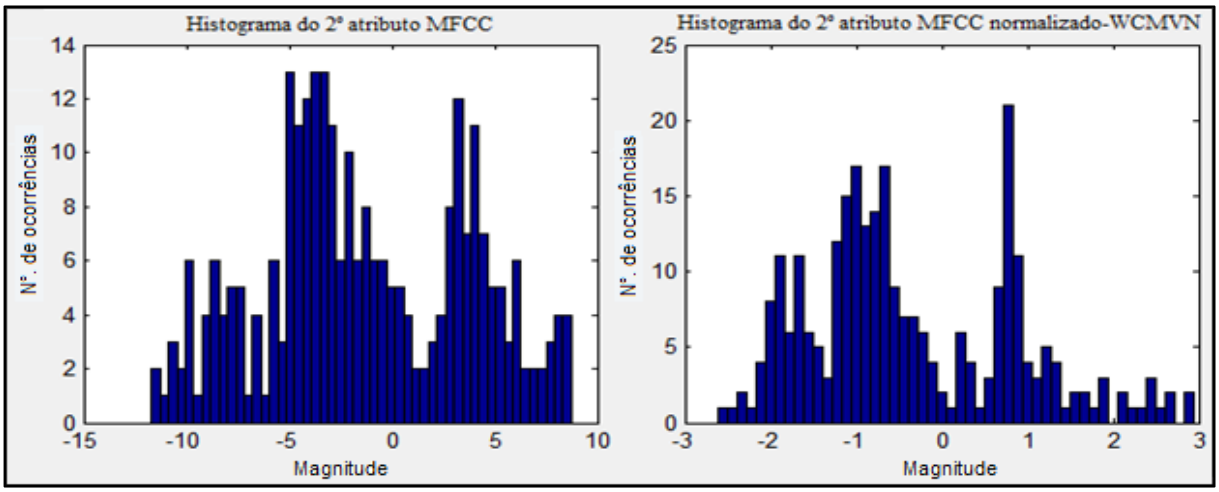

Figura 3.20 - Histogramas MFCC (à esq.) e WCMVN (à dir.) do 2° atributo cepstral.

3.4.3.3 Short-Time Gaussianization (STG)

Short-time Gaussianization é uma técnica de normalização aplicada em sistemas de reconhecimento de voz com o objetivo de compensar distorções causadas pelo canal de entrada. É baseada na modificação da distribuição em um curto espaço de tempo do vetor acústico em uma distribuição Normal padrão, cuja função densidade da distribuição tem o formato da curva de Gauss.

Inicialmente, o vetor acústico é transformado de maneira linear para que cada *frame* seja localmente independente, isso é, para que sejam não correlacionados. Em seguida, usando uma função de distribuição cumulativa (*Cumulative Distribution Function*, CDF), computada através da soma da função densidade de probabilidade sobre um janelamento ao longo do sinal (nessa pesquisa usou-se novamente o janelamento a cada 301 amostras do vetor acústico) obtém-se coeficientes cepstrais normalizados, cujas distribuições em curtos períodos de tempo apresentam funções densidades de uma distribuição Normal padrão (ALAM et al., 2011).

A Figura 3.21, compara os histogramas MFCC e STG do 2º atributo cepstral do vetor acústico da enunciação da palavra "Engenharia", a título de observação da performance de normalização dessa técnica. Nota-se que a distribuição global do conjunto de atributos foi modificada, aproximando-se da distribuição normal padrão.

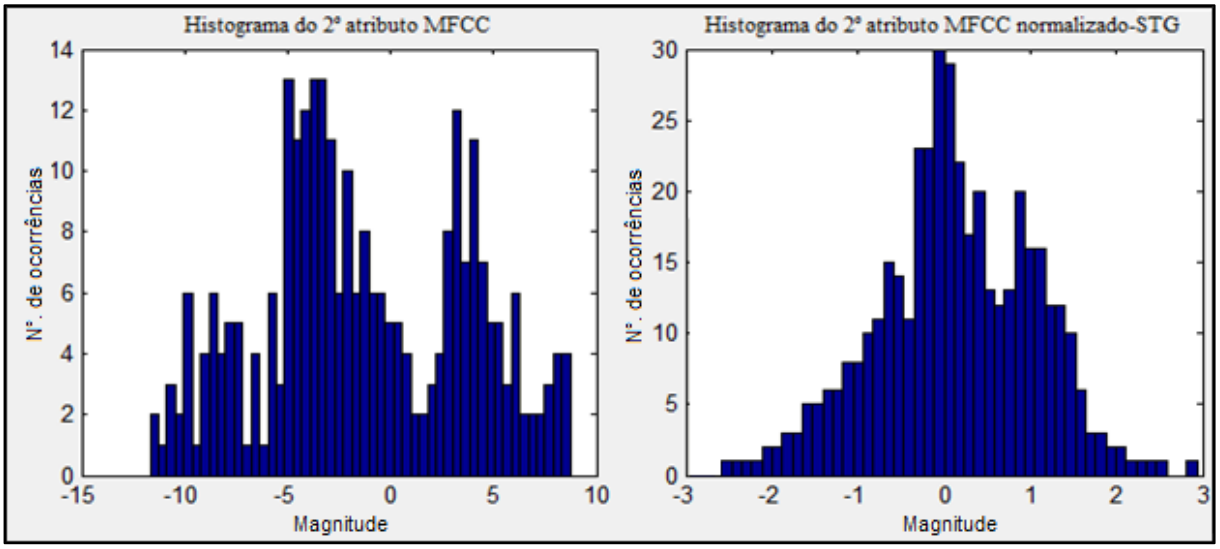

Figura 3.21 - Histogramas MFCC (à esq.) e STG (à dir.) do 2° atributo cepstral.

4 EXPERIMENTOS E ANÁLISE DE RESULTADOS

Os estudos realizados nessa pesquisa objetivaram, mormente, uma análise criteriosa acerca da eficiência do sistema de reconhecimento de voz desenvolvido, utilizando a combinação das técnicas MFCC para extração de propriedades acústicas dos sinais de fala e Quantização Vetorial para classificação e reconhecimento de padrões. O sistema foi analisado com base na capacidade do mesmo em operar com multitarefas, quais sejam: reconhecimento de palavras, comandos e frases; identificação do locutor; e identificação tanto do locutor quanto do comando por ele expresso.

A pesquisa averiguou a implementação de atributos ao sistema para incrementar sua performance. São eles: atributo de detecção de voz ativa (VAD); coeficientes dinâmicos (DDC e SDC); e critérios de normalização de coeficientes cepstrais usando as técnicas CMVN, WCMVN e STG. Adicionalmente, investigou-se a melhor configuração do sistema de reconhecimento de voz, combinando as técnicas e atributos experimentados com relação à eficiência do mesmo para realização das tarefas anteriormente indicadas.

Foram realizados no total, cinco diferentes experimentos, dos quais participaram oito indivíduos. Quatro indivíduos participaram tanto da fase de treinamento do sistema, elaborando assim um banco de dados para o mesmo, quanto da fase de testes. Os demais participaram apenas da fase de testes.

Para diversificar as análises comparativas das diferentes características fonéticas presentes na voz humana, escolheu-se indivíduos de ambos os gêneros. Três indivíduos participantes dos experimentos são do sexo feminino, das quais duas treinaram o sistema.

A pesquisa foi conduzida com auxílio de um programa elaborado na plataforma Matlab v.7.12.0 (R2011a), da qual utilizou-se ainda, funções e ferramentas de processamento de sinais embutidas. Foi construído uma interface gráfica (GUI), com intuito de ser um instrumento facilitador para recolhimento e processamento de informações.

O tempo das gravações foi padronizado para comandos e frases em ambas as etapas, com 3 segundos para o primeiro caso e 5 segundos para o segundo. O ambiente de captação dos sinais de voz para treinamento e testes também foi controlado acusticamente, impedindo interferências ou distorções sonoras que pudessem degradar as amostras de sinais de voz recolhidos e, consequentemente, afetar os resultados comparativos das amostras de treinamento e testes.

As amostras de sinais de voz captados durante as fases de treinamento e testes foram gravadas no sistema e submetidas a análises variadas para cada experimento, segundo a configuração e uso de atributos dos mesmos. Ressalta-se que cada indivíduo gravou apenas uma vez cada comando ou frase. O sinal digitalizado gerado à partir da captura da voz do locutor foi incorporado aos bancos de dados de cada experimento e processado segundo a configuração do mesmo experimento. Assim, o mesmo sinal é tratado e processado de maneiras distintas com objetivo de averiguar qual das combinações estudadas apresenta a melhor performance.

4.1 Equipamentos e Acessórios computacionais utilizados

Os equipamentos e acessórios utilizados para implementação do sistema de reconhecimento automático de voz e realização dos experimentos em pauta foram:

- Notebook ASUS modelo X53sv, com processador Intel(R) Core(TM) i7-2630QM CPU @ 2.00 GHz;
- Sistema operacional Windows 7 *Home Premium*;
- Memória RAM 6 Gb;
- Microfone embutido do notebook ASUS modelo X53sv;
- Software Matlab v.7.12.0 (R2011a).

4.2 Dados de Experimentação

Os cinco experimentos realizados nessa pesquisa são citados a seguir em ordem cronológica de avaliação:

1. Experimento para identificação do locutor e comandos usando MFCC e VQ;
2. Experimento para avaliação do atributo VAD;
3. Experimento para avaliação dos atributos dinâmicos (DDC e SDC);
4. Experimento para avaliação dos atributos de normalização (CMVN, WCMVN e STG);
5. Experimento para avaliação da melhor configuração do sistema.

Em todos os experimentos foi utilizado uma taxa de amostragem para os sinais de voz igual a 22.050 Hz, com 8 bits por amostra e 1 canal (mono). Os sinais de voz digitalizados

foram enquadrados com 256 amostras por *frame* ($N = 256$), que foram sobrepostos a cada 100 amostras. Em seguida, usou-se a janela Hamming para a etapa do janelamento.

Para garantir boa qualidade dos ensaios, algumas precauções foram tomadas durante as gravações:

A. Distância entre o locutor e o microfone de captação tomada para ensaios igual a 40-50 cm;

B. Altura dos sinais de fala de cada indivíduo, associado às condições normais de fala daquele mesmo, impedindo que esse modificasse sua maneira de falar (avaliação feita de maneira perceptiva, sem a utilização de um sonômetro);

- Nível de Pressão Sonora (NPS) utilizada por cada indivíduo, obedecendo limiar entre 35 – 55 dB.

Os indivíduos que participaram da fase de treinamentos realizaram as gravações para fase de testes seguidamente à conclusão das gravações correspondentes à etapa de treinamentos, evitando grandes variações na maneira de enunciar os comandos e frases durante os ensaios. Além disso, foi orientado aos mesmos, que tentassem assemelhar a maneira de enunciar os comandos ou frases que foram treinados, enquanto executaram a fase de testes.

Locuções que não obedeceram ao protocolo de ensaio foram descartadas e refeitas, ou seja, quando a distância do locutor em relação ao microfone de captação era incorreta, ou haviam variações na maneira de enunciar determinado comando ou frase, ou ainda, quando haviam diferenças sonoras perceptíveis durante as gravações de treinamentos e testes.

A Figura 4.1 apresenta um exemplo de um sinal digitalizado. O sinal está associado à locução "*Seu sonho é tão palpável quão grande sua vontade de realizá-lo*", enunciada pelo Indivíduo 1 durante a etapa de treinamento.

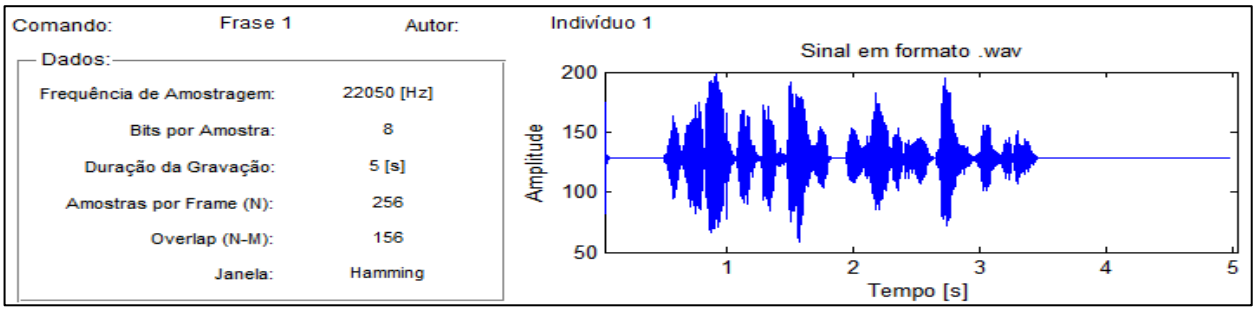

Figura 4.1 - Ilustração de um sinal de voz digitalizado utilizando-se a ferramenta Matlab.

Destaca-se na figura, os dados constando as particularidades do processo de pré-ênfase e janelamento do sinal em questão, como a frequência de amostragem, o número de bits por

amostra, a duração da gravação, o número de amostras por *frame*, o número de amostras a cada *overlap* e o tipo de janela utilizado.

Para realização dos experimentos, orientou-se aos 4 indivíduos escolhidos para treinar o sistema (denotados Indivíduos 1 - 4), que enunciassem 5 comandos e 1 frase. Ao menos um comando deveria ser comum a todos, e todas as frases diferentes. O comando comum escolhido para enunciação durante a fase de treinamentos fora a palavra 'Elo', adotada arbitrariamente. Salienta-se que 3 indivíduos participantes do estudo são irmãos e têm vozes parecidas. Portanto, a percepção da altura, atributo auditivo associado à identificação de tons, entre os irmãos é bastante parecida. Assim, para explorar a capacidade de reconhecimento do sistema, um dos irmãos foi escolhido dentre os indivíduos que tiveram suas vozes treinadas, e os outros 2 participaram apenas da fase de testes.

Durante a fase de testes, todos os indivíduos que participaram dos experimentos foram compelidos a enunciar 10 comandos e 5 frases. Os indivíduos cujas vozes foram treinadas enunciaram os 5 comandos que usaram para treinar o sistema, 2 comandos do banco de dados que foram enunciados por outro indivíduo, 2 comandos arbitrários (comandos presentes ou não no banco de dados do sistema) e 1 comando diferente dos demais. Quanto às frases, esses indivíduos enunciaram a própria frase de treinamento, 1 frase do banco de dados treinada por outro indivíduo e 3 frases arbitrarias (frases presentes ou não no banco de dados do sistema).

Os indivíduos que participaram apenas da fase de testes (Indivíduos 5 - 8) enunciaram ao menos 5 comandos e 2 frases do banco de dados. Os demais comandos e frases foram arbitrariamente escolhidos.

A Tabela 4.1 mostra as frases que foram enunciadas pelos indivíduos que treinaram o sistema. As gravações tiveram duração de 5 segundos, nesse caso.

Tabela 4.1 Frases enunciadas para treinamento do sistema

	Indivíduos	**Frases**
Fase de treinamento	1	"Seu sonho é tão palpável, quão grande sua vontade de realizá-lo"
	2	"Tudo que você precisa é de amor" (John Lennon & Paul McCartney)
	3	"Felicidade, só é real, quando compartilhada" (H. David Thoreau)
	4	"Poesia, beleza, romance e amor. Por estas razões, continuamos vivendo" (Robin Williams, *Death Poets Society*)

A Tabela 4.2 mostra os cinco comandos enunciados por cada indivíduo que participou do treinamento do sistema. Nesse caso, cada gravação durou 3 segundos.

Tabela 4.2 Comandos enunciadas para treinamento do sistema

	Indivíduos	Comandos
Fase de treinamento	1	Elo
		Sinestesia
		Terapêutico
		Cosmologia
		Cruzeiro
	2	Elo
		Descanso
		Natureza
		Endorfina
		Musicalidade
	3	Elo
		Insólito
		Multipotencialidade
		Poesia
		Rudimentar
	4	Elo
		Genealogia
		Orgânico
		Evolucionismo
		Herbáceo

4.3 Interface Gráfica

Os experimentos realizados nessa pesquisa foram conduzidos em um programa computacional de reconhecimento de voz desenvolvido na platafoma Matlab v.7.12.0 (R2011a). O programa integra as técnicas MFCC e VQ, além dos atributos de detecção de voz ativa (VAD), coeficientes dinâmicos (DDC e SDC) e ferramentas de normalização dos coeficientes cepstrais (CMVN, WCMVN e STG).

Foi construída uma interface gráfica para facilitar a interação dos usuários com o programa. Esta está apresentada pela Figura 4.2.

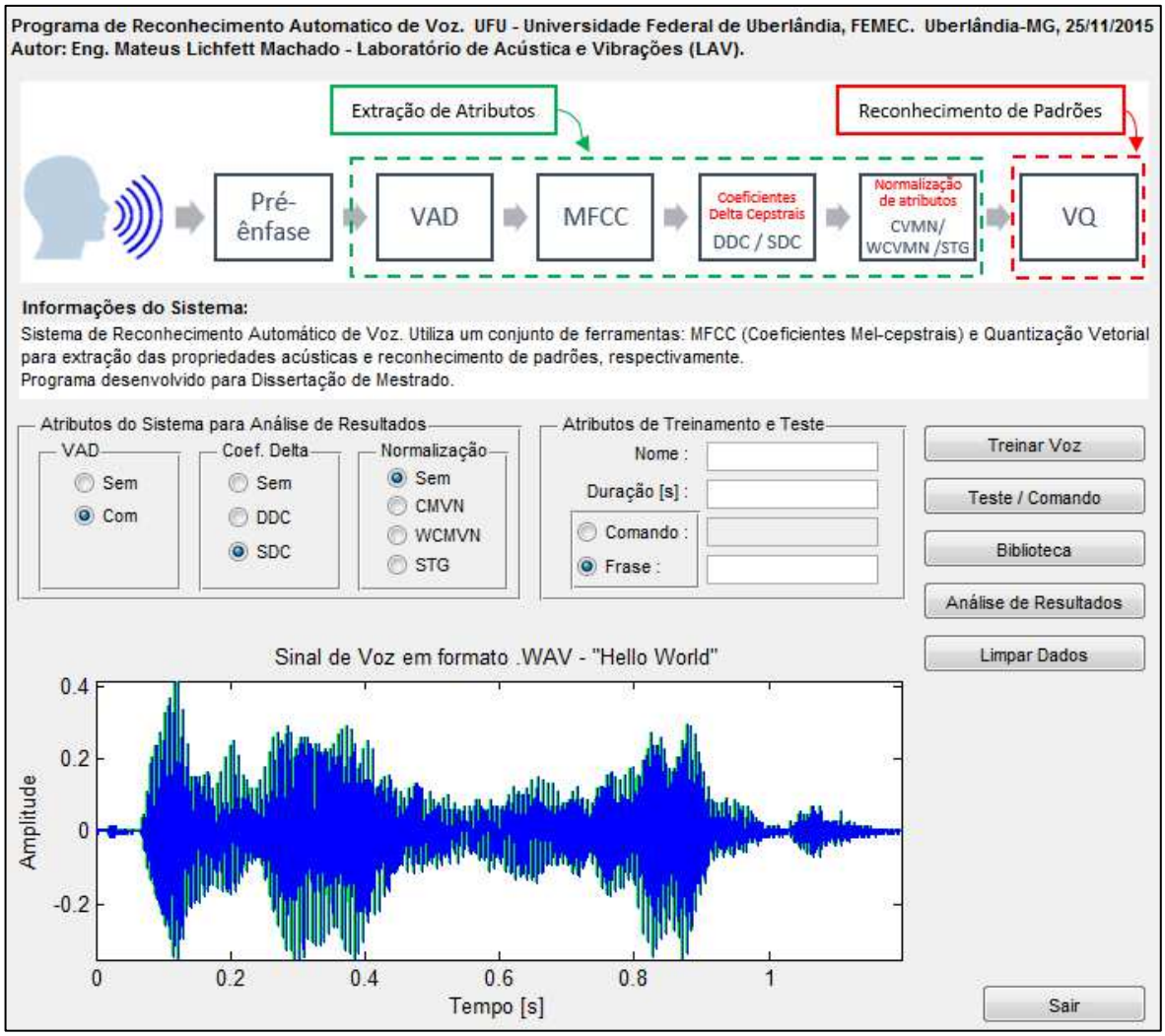

Figura 4.2 - Interface gráfica do sistema de reconhecimento de voz desenvolvido.

A interface criada permite a configuração do sistema segundo o experimento a ser tratado. No painel "*Atributos do Sistema para Análise de Resultados*", mostrado na Figura 4.2, o sistema pode ser configurado com ou sem o atributo VAD; com o tipo de coeficiente cepstral desejado (MFCC, DDC ou SDC); e ainda com o tipo de ferramenta de normalização dos coeficientes cepstrais (CMVN, WCMVN ou STG).

Para construção do banco de dados, o usuário deve primeiramente inserir os dados do painel "*Atributos de Treinamento e Teste*" com seu nome, duração da gravação, escolher treinar um comando ou frase, e escrever qual comando ou frase irá enunciar. Após o preenchimento desses campos, o usuário deve apertar o botão "*Treinar Voz*" para ter sua voz captada e processada em vetores acústicos que serão lançados no banco de dados. Essa tarefa foi realizada pelos Indivíduos 1 a 4.

A etapa de testes do sistema também exige o preenchimento dos campos do painel "*Atributos de Treinamento e Teste*". Em seguida, o usuário deve apertar o botão "*Teste/Comando*" para ter sua voz captada e processada em vetores acústicos que serão comparados com sinais de voz presentes no banco de dados do sistema. Essa tarefa foi realizada por todos os indivíduos que realizaram os experimentos.

Quando o usuário seleciona os botões para treinar ou testar o sistema, a figura mostrada na interface, cujo título é "*Sinal de Voz em formato WAV – Hello World*" se modifica para mostrar a captação do sinal de voz do locutor em tempo real.

Além de treinar e testar o sistema através dos botões "*Treinar Voz*" e "*Teste/Comando*", o usuário tem acesso à biblioteca, à análise de resultados e pode limpar os dados do painel "*Atributos de Treinamento e Teste*", bem como os dados da figura mostrada na interface, apertando o botão "*Limpar Dados*". Para encerrar o programa, o usuário deve apertar o botão "*Sair*".

A biblioteca do sistema consta de todos os comandos e frases que foram treinados ou testados pelos indivíduos. A biblioteca consta de duas partes: "*Biblioteca de Treinamento*" e "*Biblioteca de Teste*". O usuário deve marcar no painel "*Atributos de Treinamento e Teste*" se deseja acessar os comandos ou frases de cada biblioteca, e então apertar o botão "*Biblioteca*".

A Figura 4.3 apresenta os comandos testados pelo Indivíduo 1 no sistema de reconhecimento de voz desenvolvido nesa pesquisa.

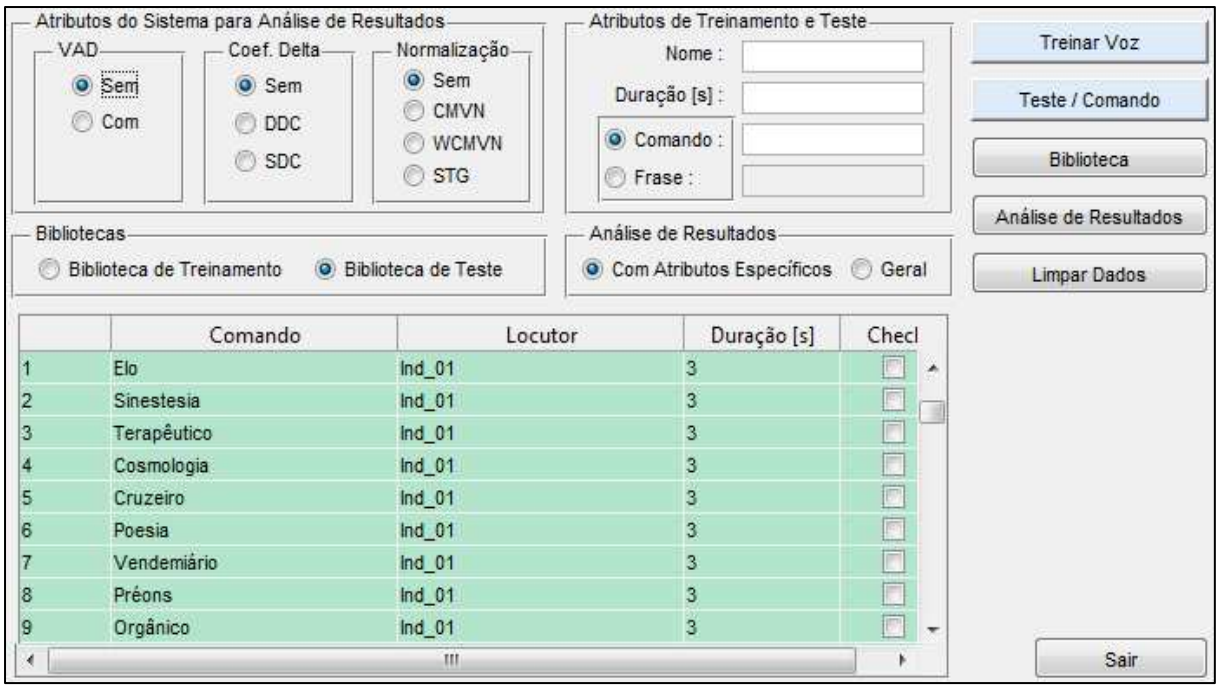

Figura 4.3 - Biblioteca de dados da interface do sistema de reconhecimento de voz desenvolvido.

Na biblioteca, o usuário pode acessar as características de cada locução, como a identificação de quem é o locutor, qual o comando ou frase enunciada, e qual a duração da gravação. A disposição dos comandos ou frases na biblioteca obedece a ordem cronológica de gravação dos sinais de voz.

Ao selecionar um comando ou frase na "*Biblioteca de Teste*", o usuário pode analisar os resultados dessa enunciação com relação ao banco de dados criado durante a fase de treinamento. No painel "*Análise de Resultados*", o usuário deverá optar pela apresentação dos resultados para uma configuração especificada no painel "*Atributos do Sistema para Análise de Resultados*", ou para todas as configurações possíveis, permitindo analisar qual das configurações é mais eficaz para uma determinada tarefa (identificação do locutor; identificação do comando/frase; ou identificação do locutor e do comando/frase).

Ao apertar o botão "*Análise de Resultados*", o sistema exibe as distâncias Euclidianas entre o vetor acústico do sinal em análise com vetores correspondentes à todos os comandos ou frases do banco de dados, em ordem crescente. Além disso, identifica se o comando ou frase em questão foi reconhecido pelo sistema, bem como, se o locutor foi corretamente identificado.

4.4 Experimentos

Os 5 experimentos realizados nessa pesquisa foram avaliados para comandos (enunciações com pouca variabilidade fonética) e frases (enunciações com maiores variações fonéticas). Participaram dos experimentos tanto indivíduos do sexo masculino (Indivíduos 1, 4, 5, 6 e 8) quanto feminino (Indivíduos 2, 3, e 7).

Cada experimento foi estudado cruzando-se informações dos sinais de fala referentes à etapa de testes com o banco de dados gerados na etapa de treinamento. Durante os testes foram realizadas 120 gravações, das quais 80 referentes aos comandos e 40 às gravações de frases. Essas foram processadas e transformadas em vetores acústicos que foram lançados nos espaços vetoriais correspondentes à cada experimento, gerando 1600 amostras de distorções VQ relativas à comandos e 160 amostras relativas às frases.

A Tabela 4.3 mostra os comandos enunciados durante a fase de testes. Nela, os comandos numerados de 1 a 5 estão presentes no banco de dados do sistema de reconhecimento de voz desenvolvido.

Tabela 4.3 Comandos de Testes

	Comandos de Teste			
	Ind. 1	**Ind. 2**	**Ind. 3**	**Ind. 4**
Nº. 1	Elo	Elo	Elo	Elo
Nº. 2	Sinestesia	Descanso	Insólito	Genealogia
Nº. 3	Terapêutico	Natureza	Multipotencialidade	Orgânico
Nº. 4	Cosmologia	Endorfina	Poesia	Evolucionismo
Nº. 5	Cruzeiro	Musicalidade	Rudimentar	Herbáceo
Nº. 6	Poesia (3)	Terapêutico (1)	Placebo	Descanso (2)
Nº. 7	Vendemiário	Meditação	Matéria	Facultativo
Nº. 8	Préons	Rudimentar (3)	Sinestesia (1)	Alquimia
Nº. 9	Orgânico (4)	Corpo	Musicalidade (2)	Cruzeiro (1)
Nº. 10	Imaginação	Verde	Panaceia	Carbono
	Comandos de Teste			
	Ind. 5	**Ind. 6**	**Ind. 7**	**Ind. 8**
Nº. 1	Insólito (3)	Sinestesia (1)	Cosmologia (1)	Orgânico (4)
Nº. 2	Descanso (2)	Multipotencialidade (3)	Musicalidade (2)	Natureza (3)
Nº. 3	Poesia (3)	Cruzeiro (1)	Endorfina (2)	Sinestesia (1)
Nº. 4	Herbáceo (4)	Orgânico (4)	Rudimentar (3)	Evolucionismo (4)
Nº. 5	Cosmologia (1)	Insólito (3)	Poesia (3)	Cruzeiro (1)
Nº. 6	Minas	Mente	Porto	Estrutura
Nº. 7	Jovem	Liberdade	Romantismo	Carvão
Nº. 8	Guitarra	Virtude	Renascença	Indelével
Nº. 9	Sinapses	Destreza	Saúde	Pimenta
Nº. 10	Energia	Medicina	Atração	Brasil

Os números destacados em azul na Tabela 4.3 identificam a autoria dos comandos treinados, quando estes são citados por outros indivíduos. Por exemplo, Rudimentar (3), indica que o comando "*Rudimentar*" foi treinado pelo Indivíduo 3, embora testado pelo Indivíduo 2.

A Tabela 4.4 mostra as frases enunciadas pelos indivíduos durante a fase de testes.

Tabela 4.4 Frases de Testes (continua)

		Frases de Teste
Ind. 1	**Nº.** 1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo
	Nº. 2	"Não existe um caminho para felicidade. Felicidade é o caminho" (Thich Nhat Hanh)
	Nº. 3	"O sucesso é ir de fracasso em fracasso sem perder o entusiasmo" (Winston Churchill)
	Nº. 4	"Temos de nos tornar a mudança que queremos ser" (Mahatma Gandhi)
	Nº. 5	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (4) (Robin Williams, *Death Poets Society*)
Ind. 2	**Nº.** 1	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)
	Nº. 2	"O segundo nada mais é que o primeiro dos últimos" (Ayrton Senna)
	Nº. 3	"Cada um de nós compõe sua própria história" (Almir Sater)
	Nº. 4	"Felicidade só é real quando compartilhada" (3) (H. David Thoreau)
	Nº. 5	"É melhor ser alegre que ser triste, alegria é a melhor coisa que existe" (Vinicius de Moraes)

Tabela 4.5 Frases de Testes (conclusão)

		Frases de Teste
Ind. 3	**Nº. 1**	"Felicidade só é real quando compartilhada" (H. David Thoreau)
	Nº. 2	"Ser feliz sem motivo é mais autêntica forma de felicidade" (Carlos Drummond de Andrade)
	Nº. 3	"Prefiro o paraíso pelo clima, o inferno pela companhia" (Mark Twain)
	Nº. 4	"Algumas pessoas nunca cometem o mesmo erro duas vezes. Descobrem sempre novos erros para cometer" (Mark Twain)
	Nº. 5	"Seu sonho é tão palpável quão grande sua vontade de realiza-lo" (1)
Ind. 4	**Nº. 1**	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)
	Nº. 2	"Quem não ama o sorriso feminino desconhece a poesia de Cervantes" (Zé Ramalho)
	Nº. 3	"Liberdade, Igualdade e Fraternidade" (Lema da Revolução Francesa)
	Nº. 4	"Cuidado com seu caráter, ele controla o seu destino" (Paulo Coelho)
	Nº. 5	"Tudo o que você precisa é de amor" (2) (John Lennon & Paul McCartney)
Ind. 5	**Nº. 1**	"Felicidade só é real quando compartilhada" (3) (H. David Thoreau)
	Nº. 2	"Seu sonho é tão palpável quão grande sua vontade de realiza-lo" (1)
	Nº. 3	"A melhor lembrança da minha infância: eu não precisava trabalhar" (Vanessa Pimentel)
	Nº. 4	"Nós fomos além de onde devíamos ter ido" (Jack Johnson)
	Nº. 5	"Seja um homem simples. Seja algo que ame e que entendas" (Ronnie Van Zant & Gary Rossington)
Ind. 6	**Nº. 1**	"Tudo o que você precisa é de amor" (2) (John Lennon & Paul McCartney)
	Nº. 2	"Seu sonho é tão palpável quão grande sua vontade de realiza-lo" (1)
	Nº. 3	"Vocês riem de mim por eu ser diferente e eu rio de vocês por serem todos iguais" (Bob Marley)
	Nº. 4	Qual o tamanho de sua vontade de conquistar o mundo?
	Nº. 5	"Saber como pensar torna a pessoa muito mais capaz do que aquele que apenas sabe o que deve pensar" (Neil deGrasse Tyson)
Ind. 7	**Nº. 1**	"Felicidade só é real quando compartilhada" (3) (H. David Thoreau)
	Nº. 2	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (4) (Robin Williams, *Death Poets Society*)
	Nº. 3	"E que a minha loucura seja perdoada. Porque metade de mim é amor, e a outra metade também" (Osvaldo Montenegro)
	Nº. 4	"Te vejo do lado escuro da Lua" (Roger Waters)
	Nº. 5	"A criança cresceu, o sonho acabou. E eu fiquei confortavelmente entorpecido" (David Gilmour & Roger Waters)
Ind. 8	**Nº. 1**	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (4) (Robin Williams, *Death Poets Society*)
	Nº. 2	"Tudo o que você precisa é de amor" (2) (John Lennon & Paul McCartney)
	Nº. 3	"O Rio de Janeiro continua lindo" (Gilberto Gil)
	Nº. 4	Cruzeiro, o maior time de Minas
	Nº. 5	"Por tanto amor, por tanta emoção, a via me fez assim, doce ou atroz, manso ou feroz, eu caçador de mim" (Milton Nascimento)

Os números destacados em azul na Tabela 4.4 identificam a autoria das frases treinadas, quando essas são citadas por outros indivíduos.

Em função da grande quantidade de dados gerados pelos experimentos nessa pesquisa, serão apresentados resultados (distorções VQ) referentes a 2 comandos e 2 frases (escolhidos aleatoriamente) de cada indivíduo, totalizando 320 amostras de comandos e 64 amostras de

frases, para cada experimento. A título de comparação das diferentes configurações utilizadas no sistema de reconhecimento de voz desenvolvido, serão apresentados os resultados referentes aos mesmos 2 comandos e 2 frases escolhidos por indivíduo para todos os experimentos.

Foram enunciados ao total 80 comandos na etapa de testes, 20 dos quais representam comandos de treinamento enunciados pelos seus próprios autores, 28 dos quais referem-se à comandos de treinamento, mas não enunciados pelos próprios autores, e 32 comandos quaisquer, isso é, não presentes no banco de dados. Além disso, foram analisadas 40 frases, das quais 4 foram treinadas e enunciadas pelos próprios autores, 12 das quais referem-se à frases de treinamento, mas não enunciadas pelos próprios autores e 24 das quais não estão presentes no banco de dados.

Os resultados obtidos para os comandos e frases de testes escolhidos para os 8 indivíduos que participaram dos experimentos estão apresentados em tabelas e gráficos a seguir. Nas tabelas, as distâncias entre os comandos e frases testadas são comparados com relação aos comandos e frases treinados, respectivamente. Quão menor a distorção VQ entre o *codebook* gerado pelo comando ou frase de teste e os *codebooks* do banco de dados, maior similitude haverá entre eles. Se essa distância calculada for inferior à um limiar específico, o *codebook* do banco de dados com menor distância será associado ao comando ou frase de entrada.

O limiar das distânciais Euclidianas que separa a região de identificação de comandos e locutores da região de não identificação dos mesmos foi definido diferentemente para cada experimento segundo verificações experimentais, uma vez que os autores LINDE et al. (1980), indicam o valor desse limiar apenas para o caso em que o sistema utiliza as técnicas MFCC e VQ sem o uso de atributos.

As tabelas mostram a correta identificação do locutor através da coluna de acertos, isso é, a marca assertiva indica se o locutor é corretamente identificado (Indivíduos 1 a 4, indivíduos que treinaram o sistema), ou se o locutor é corretamente não identificado (Indivíduos 5 a 8, indivíduos que apenas testaram o sistema).

As figuras apresentam os resultados gráficos das distâncias Euclidianas mostradas nas tabelas de cada experimento, para todos os indivíduos. Nas figuras, a região de identificação, delimitada pelo limiar das distâncias Euclidianas está destacada em azul. Amostras que estiverem contidas nessa região, representam amostras cujos locutores foram identificados. Além disso, nas figuras é possível identificar com facilidade a menor distância Euclidiana associada a cada comando de teste em análise. A menor distância revela com qual comando de treinamento o comando de teste em estudo apresenta maior similitude.

4.4.1 Experimento 01 – Ensaio de Identificação do Locutor e Comandos Usando MFCC e VQ

O primeiro experimento realizado nessa pesquisa tem como objetivo estudar a capacidade do sistema de reconhecimento de voz composto pelas técnicas MFCC e VQ em realizar três tarefas: identificar os locutores; reconhecer comandos e frases; e, por fim, identificar o locutor e o comando ou frase por ele enunciado. O limiar das distâncias Euclidianas adotado nesse experimento é igual a 5.

Para esse experimento, não foi utilizado nenhum atributo adicional com intuito de avaliar a eficiência das técnicas MFCC para extração das propriedades acústicas dos sinais de fala, e Quantização Vetorial para o reconhecimento de padrões, quando essas são utilizadas em um sistema de reconhecimento de voz.

Para o cumprimento das tarefas, o sistema deve ser capaz de assimilar as vozes dos locutores quando esses forem indivíduos que treinaram o sistema e não assimilar as vozes dos indivíduos que apenas participaram da fase de testes, ou seja, que não possuem registros nos bancos de dados do sistema; além disso, o sistema será avaliado se é capaz de reconhecer corretamente o comando ou frase do banco de dados, quando esses são expressos tanto por um indivíduo que treinou o sistema, quanto por um indivíduo que apenas testou o sistema; e, finalmente, o sistema será avaliado com relação à capacidade em identificar corretamente o locutor e o comando ou frase por ele enunciado.

Os resultados para os comandos de testes do primeiro experimento estão apresentados a seguir nas Tabelas 4.5, a 4.7, e graficamente nas Figuras 4.4, a 4.7. Cada uma das figuras consta as 2 locuções de comandos para 2 indivíduos, para evitar poluição gráfica. Nos gráficos, o eixo das abscissas corresponde aos 20 comandos treinados, apresentados na mesma ordem das tabelas, isto é, os comandos de números 1 a 5 referem-se ao Indivíduo 1, os comandos de 6 a 10, ao Indivíduo 2, os comandos de 11 a 15, ao Indivíduo 3 e os comandos de 16 a 20, ao Indivíduo 4; e o eixo das ordenadas indica as distâncias Euclidianas entre os vetores acústicos do comando de teste e os vetores acústicos dos comandos treinados.

Tabela 4.6 Resultados do Experimento 01 – Comandos de Testes (Ind. 1-2-3)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 1		Ind. 2		Ind. 3		
Locutor (Ind.) ↓	Comandos de Treinamentos ↓	Sinestesia	Vendemiário	Descanso	Multipotencialidade	Poesia	Evolucionismo	Acertos
1	Elo	3,248	3,454	6,340	7,148	6,381	7,021	✓
	Sinestesia	2,718	4,052	6,972	7,322	6,487	6,863	✓
	Terapêutico	3,781	4,524	6,511	6,921	7,251	6,607	✓
	Cosmologia	3,524	3,933	6,883	7,036	6,822	6,459	✓
	Cruzeiro	3,693	4,109	7,117	7,259	7,115	7,168	✓
2	Elo	7,228	7,437	2,926	3,462	5,862	6,152	✓
	Descanso	7,011	7,154	2,785	3,716	6,164	6,302	✓
	Natureza	7,271	7,012	3,256	3,612	6,583	6,415	✓
	Endorfina	7,124	7,355	3,529	3,852	6,138	6,208	✓
	Musicalidade	7,519	7,786	3,428	3,768	6,512	5,851	✓
3	Elo	7,257	6,884	6,104	6,602	2,874	3,248	✓
	Insólito	7,601	7,552	5,873	7,126	3,152	3,782	✓
	Multipotencialidade	8,177	7,426	6,246	6,203	3,496	4,119	✓
	Poesia	7,560	7,653	5,922	7,012	2,517	3,517	✓
	Rudimentar	7,622	7,006	6,608	6,109	3,291	3,358	✓
4	Elo	5,853	6,392	6,661	6,583	6,953	6,783	✓
	Genealogia	5,451	6,445	7,153	7,438	6,526	6,542	✓
	Orgânico	6,147	6,528	6,425	6,708	6,852	6,603	✓
	Evolucionismo	6,766	5,817	7,294	7,391	7,194	6,922	✓
	Herbáceo	6,314	6,296	6,848	6,644	6,993	6,538	✓

Tabela 4.7 Resultados do Experimento 01 – Comandos de Testes (Ind. 4-5-6)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 4		Ind. 5		Ind. 6		
Locutor (Ind.) ↓	Comandos de Treinamentos ↓	Alquimia	Elo	Cosmologia	Sinapses	Evolucionismo	Destreza	Acertos
1	Elo	6,758	6,229	6,317	6,560	6,568	6,970	✓
	Sinestesia	6,446	6,741	6,201	6,333	6,919	6,805	✓
	Terapêutico	7,081	6,848	6,715	6,631	7,107	6,827	✓
	Cosmologia	6,562	6,912	5,791	6,576	6,684	7,254	✓
	Cruzeiro	6,837	6,754	6,898	7,011	6,492	6,772	✓
2	Elo	7,254	6,521	7,479	7,264	6,856	7,119	✓
	Descanso	7,680	6,871	7,586	6,916	7,149	6,750	✓
	Natureza	7,847	6,904	7,110	7,082	7,582	6,802	✓
	Endorfina	7,775	7,165	7,243	7,470	7,378	7,157	✓
	Musicalidade	8,260	7,570	8,106	7,370	7,750	7,366	✓
3	Elo	7,616	6,658	7,227	6,627	6,759	6,831	✓
	Insólito	7,230	7,228	7,624	7,369	7,236	7,139	✓
	Multipotencialidade	8,096	7,748	7,165	7,457	7,246	7,664	✓
	Poesia	7,652	7,125	6,818	7,140	7,194	6,768	✓
	Rudimentar	7,397	7,676	7,276	7,438	7,560	7,168	✓
4	Elo	3,866	2,118	6,535	6,635	6,891	6,662	✓
	Genealogia	3,751	3,114	6,270	6,706	6,549	7,218	✓
	Orgânico	4,074	3,323	6,656	6,583	6,971	7,102	✓
	Evolucionismo	4,462	3,259	6,948	7,189	6,705	7,578	✓
	Herbáceo	4,127	2,874	6,613	6,367	6,837	7,191	✓

Tabela 4.8 Resultados do Experimento 01 – Comandos de Testes (Ind. 7-8)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)				
		Ind. 7		Ind. 8		
Locutor (Ind.) ↓	Comandos de Treinamentos ↓	Natureza	Poesia	Cruzeiro	Orgânico	Acertos
1	Elo	7,030	6,619	6,407	6,784	✓
	Sinestesia	7,226	6,449	6,579	6,816	✓
	Terapêutico	7,310	7,177	6,723	6,641	✓
	Cosmologia	7,389	6,812	6,804	6,962	✓
	Cruzeiro	6,970	7,223	6,205	6,714	✓
2	Elo	6,304	6,191	7,391	6,960	✓
	Descanso	6,299	6,613	7,618	6,828	✓
	Natureza	6,172	6,590	7,308	7,197	✓
	Endorfina	6,281	6,388	7,578	7,161	✓
	Musicalidade	6,822	6,989	7,804	7,252	✓
3	Elo	6,675	6,211	7,364	6,904	✓
	Insólito	6,878	6,520	7,926	7,176	✓
	Multipotencialidade	6,445	6,941	8,150	7,685	✓
	Poesia	6,183	6,029	7,423	7,735	✓
	Rudimentar	6,524	6,736	7,616	7,179	✓
4	Elo	7,458	7,197	6,665	6,544	✓
	Genealogia	7,184	6,931	6,762	6,828	✓
	Orgânico	6,820	7,576	7,054	6,307	✓
	Evolucionismo	7,386	7,489	7,180	7,246	✓
	Herbáceo	7,203	7,768	6,831	7,038	✓

Destaca-se nas figuras o limiar das distâncias Euclidianas igual a 5. As amostras inseridas na região demarcada por distâncias Euclidianas ponderadas inferiores a 5 correspondem a regiões de identificação dos locutores.

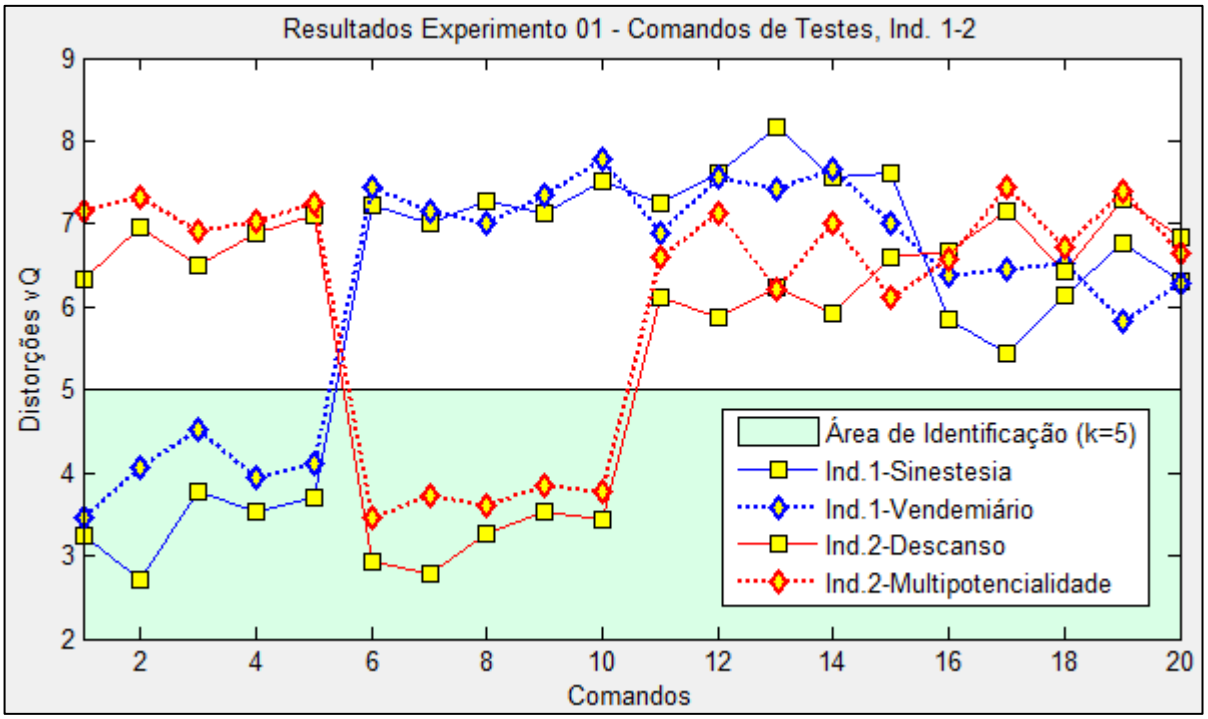

Figura 4.4 - Representação gráfica dos resultados para Comandos de Testes - Indivíduos 1 e 2.

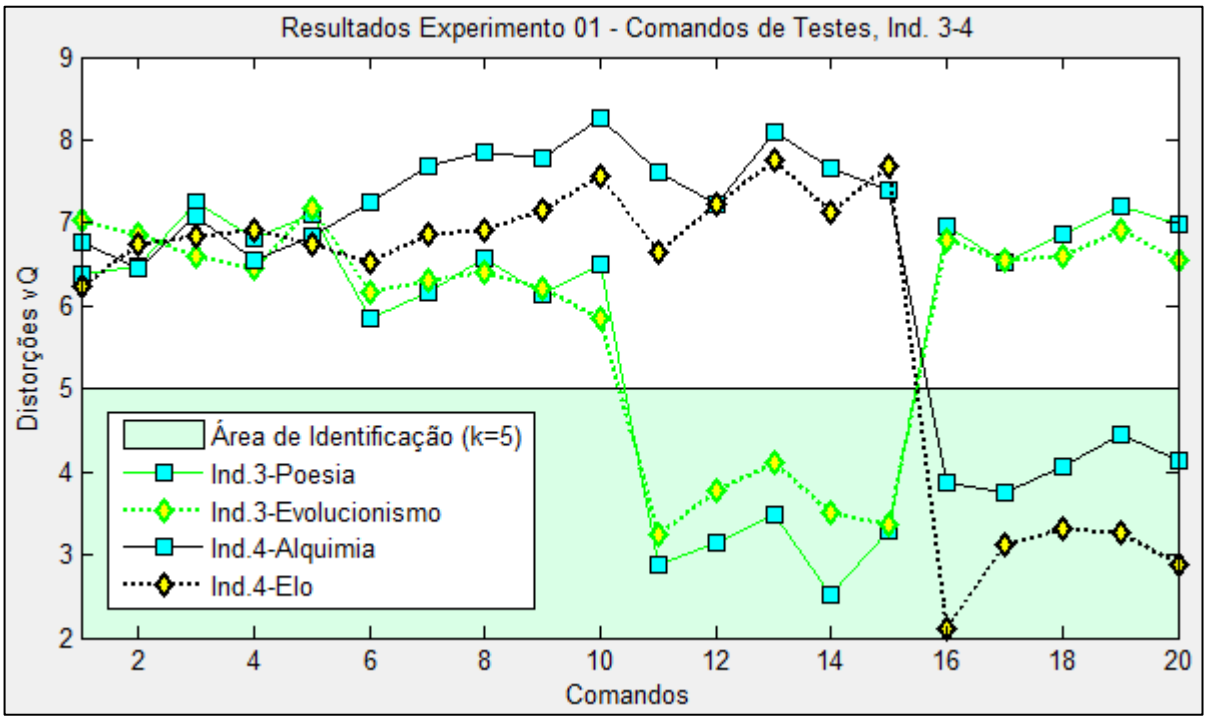

Figura 4.5 - Representação gráfica dos resultados para Comandos de testes - Indivíduos 3 e 4.

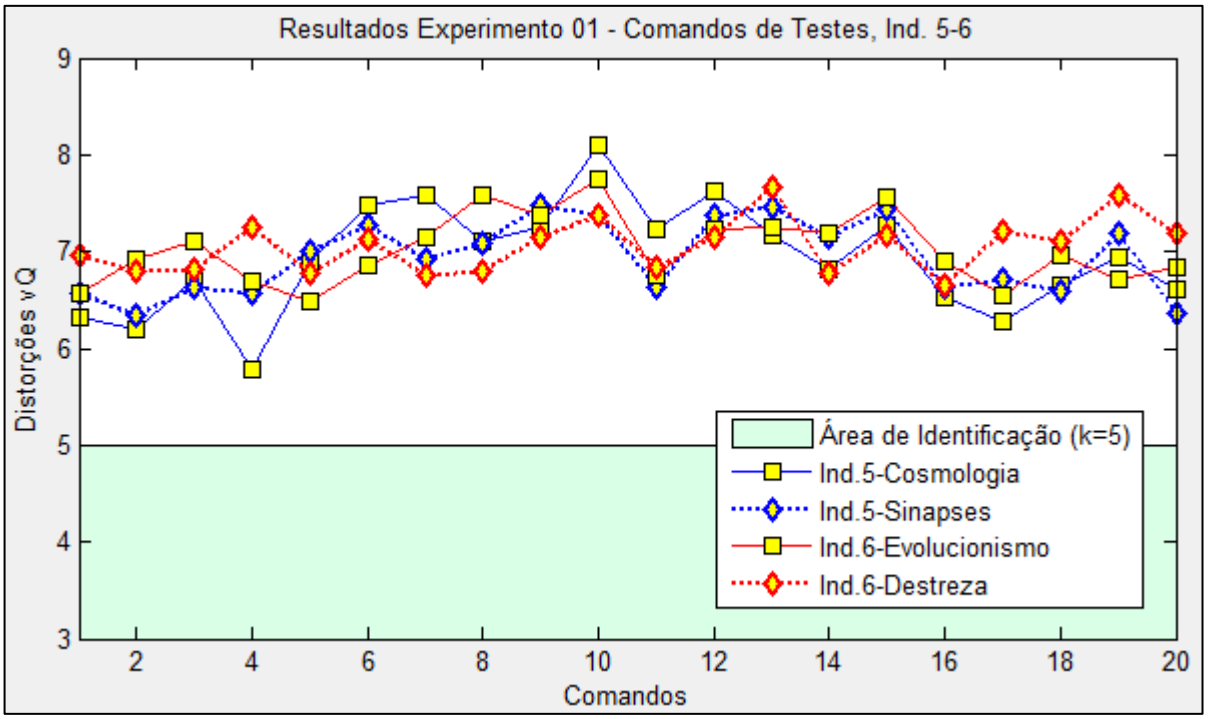

Figura 4.6 - Representação gráfica dos resultados para Comandos de testes - Indivíduos 5 e 6.

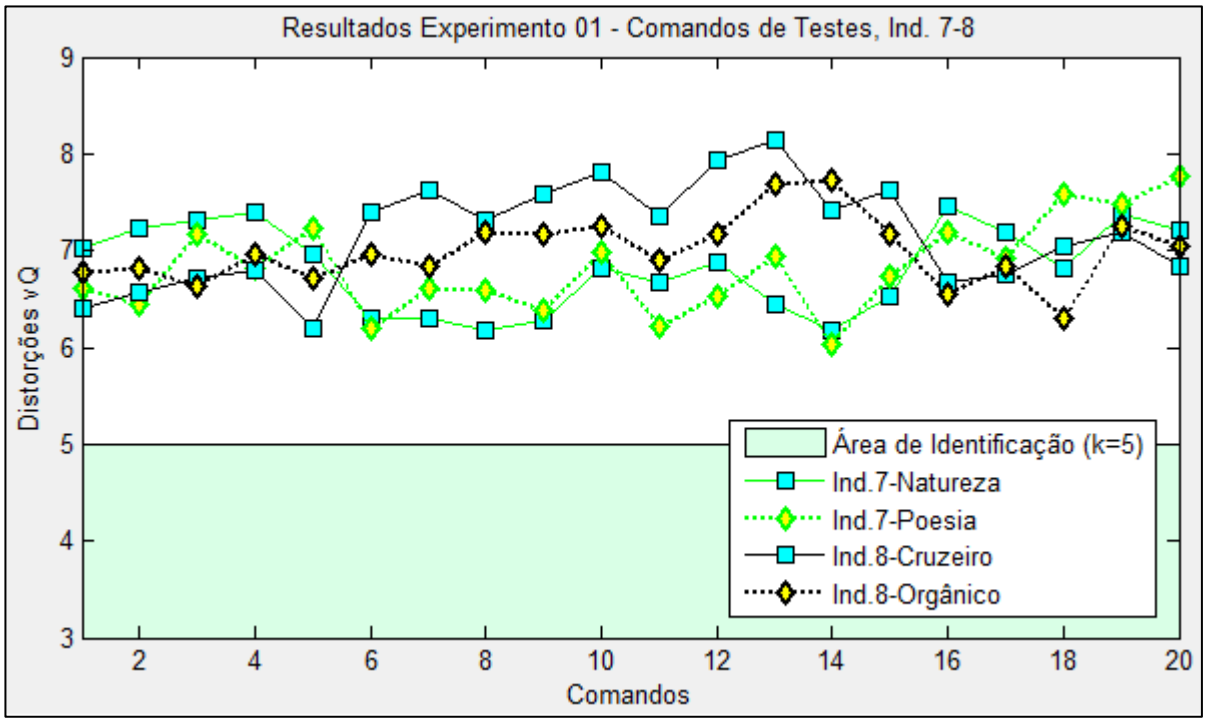

Figura 4.7 - Representação gráfica dos resultados para Comandos de testes - Indivíduos 7 e 8.

Os resultados apresentados nas tabelas e gráficos desse experimento apontaram conclusões interessantes quanto ao sistema de reconhecimento de voz construído. O sistema

mostrou-se altamente eficaz em reconhecer o indivíduo locutor, como pode ser notado com relação aos resultados de similitude para os Indivíduos 1 a 4. Nesses casos, para os 2 comandos testados por esses indivíduos, os resultados apresentaram-se dentro da região de identificação do locutor, ou seja, as distâncias Euclidianas foram inferiores ao limiar definido igual a 5, como pode ser constatado nos gráficos das Figuras 4.4 e 4.5. Ao passo que, para os indivíduos que apenas testaram o sistema, nenhuma amostra apresentou-se dentro da região de identificação dos locutores, conforme pode ser notado nas Figuras 4.6 e 4.7.

Quanto à capacidade do sistema em identificar corretamente os comandos de testes apresentados nas tabelas, quando estes estão presentes no banco de dados, notou-se que o comando de teste foi corretamente identificado para todos os casos em que o indivíduo treinara tal comando. Esse fato é apurado para o caso da palavra '*Sinestesia'*, treinada e testada pelo Indivíduo 1; da palavra '*Descanso*', treinada e testada pelo Indivíduo 2; do comando '*Poesia*', treinado e testado pelo Indivíduo 3; e da palavra '*Elo*', que embora tenha sido treinada por todos os Indivíduos, fora corretamente identificada quando testada pelo Indivíduo 4. Esse registro conclui que o sistema apresenta grande capacidade de reconhecer o locutor e o comando por ele enunciado.

Os resultados das tabelas com relação à correta identificação de um comando presente no banco de dados, quando este foi testado por um indivíduo que não treinou o sistema, foram satisfatórios. As menores distâncias Euclidianas foram registradas para um comando testado presente no banco de dados para os casos da palavra '*Cosmologia*', testada pelo Indivíduo 5 e treinada pelo Indivíduo 1; da palavra '*Poesia*', testada pelo Indivíduo 7 e treinada pelo indivíduo 3; da palavra '*Natureza*' testada pelo indivíduo 7 e treinada pelo Indivíduo 2; da palavra '*Cruzeiro*', testada pelo Indivíduo 8 e treinada pelo indivíduo 1; e da palavra '*Orgânico*', testada pelo Indivíduo 8 e treinada pelo ndivíduo 4. A única ocorrência não assertiva para esse caso foi registrado para a palavra '*Evolucionismo*', treinada pelo indivíuo 6, cujo resultado apurado fora a palavra '*Cruzeiro*', treinada pelo Indivíduo 1.

Além disso, outra ponderação pode ser realizada analisando os resultados das tabelas. Quando um indivíduo que treinou o sistema enuncia um comando no banco de dados que não é de sua autoria, o sistema errôneamente aufere o resultado à algum comando que esse próprio locutor tenha treinado. Portanto, nesse caso, o sistema identifica corretamente o locutor, mas assimila um comando incorreto ao comando testado. Esse fato está registrado para os casos da palavra '*Multipotencialidade*', testada pelo Indivíduo 2, treinada pelo Indivíduo 3, e cujo comando assimilado foi a palavra '*Elo*' treinada pelo Indivíduo 2; e da palavra '*Evolucionismo*',

testada pelo Indivíduo 3, treinada pelo Indivíduo 4, e cujo comando assimilado foi a palavra '*Elo*', treinada pelo Indivíduo 4.

Os resultados apresentados pelos 2 comandos de testes escolhidos para cada indivíduo apresentados nas tabelas e gráficos anteriores são extrapolados para os demais comandos que foram testados por todos os indivíduos participantes do experimento.

A análise conduzida acerca da capacidade do sistema em identificar corretamente o locutor foi verificada assertivamente para 91,19% dos casos. Os 80 sinais de fala de comandos de testes reproduziram 1600 amostras de distorções VQ. Para 1459, os resultados foram assertivos, isso é, ou os locutores que treinaram o sistema foram corretamente identificados, ou os indivíduos que apenas testaram o sistema fora corretamente não identificados.

Em alguns casos, embora a identificação do locutor fora correta, ao avaliar resultados relacionandos à outros locutores, esses se apresentavam dentro da região de identificação do locutor. Situações como estas foram integradas ao cálculo de resultados incorretos do sistema.

O índice de acertos para capacidade do sistema em identificar o locutor foi considerado satisfatório para a quantidade de sinais de fala coletados e número de pessoas que participaram do experimento. Os resultados obtidos são concordantes com as pesquisas conduzidas por Nijhawan e Soni (2014), Bharti e Bansal (2015).

Com relação à capacidade do sistema em identificar comandos, verificou-se para a situação em que o locutor que participou da fase de treinamento do sistema enuncia algum comando treinado por ele, obteve-se 90% de reconhecimento assertivo. Para os 20 comandos treinados, 18 foram corretamente identificados quando enunciados pelos próprios autores. Portanto, em 90% dos casos, o sistema foi capaz de identificar assertivamente o locutor e o comando por ele enunciado. No entanto, quando esses indivíduos enunciavam comandos distintos daqueles que usaram para treinar o sistema, os resultados apontaram para algum comando treinado pelo próprio autor. Esse fato foi registrado para todas as 20 ocorrências.

Ao investigar amostras nas quais os locutores treinados enunciam comandos do banco de dados de outros autores, e analisando apenas as distâncias Euclidianas com relação aos comandos treinados por esses mesmos indivíduos, isto é, desconsiderando as distâncias Euclidianas associadas aos comandos treinados pelo indivíduo locutor, em 4 das 8 ocorrências desse fato, o comando foi assertivamente associado.

Quando um locutor que apenas testou o sistema enuncia algum comando do banco de dados, em 9 das 20 amostras para esta situação, a distância mínima reportada correspondeu ao comando presente no banco de dados (45% dos casos).

Portanto, apurou-se que o sistema apresentou uma sensibilidade para reconhecimento de comandos enunciados em 31 das 48 amostras em que se enunciou um comando do banco de dados, correspondendo a 64,58%.

Outra importante ponderação pode ser realizada ao avaliar os dados coletados. Nota-se uma clara distinção entre as distorções VQ quando um indivíduo enuncia um comando e este é comparado com locuções de um banco de dados associadas à um indivíduo do sexo oposto. Pode-se perceber este fato, por exemplo, ao analisar os resultados do Indivíduo 1 (indivíduo do gênero masculino) no gráfico apresentado na Figura 4.4. As distâncias comparadas com comandos treinados pelos indivíduos 2 e 3, que são do gênero feminino, são superiores com relação aos comandos treinados pelo Indivíduo 4, do gênero masculino. O oposto também é notável. Assim o sistema mostra-se suscetível ao sexo do locutor. Estes resultados também foram observados pelos pesquisadores Patel e Prasad (2013).

A seguir estão apresentados os resultados referentes às frases de treinamentos através das Tabelas 4.8 a 4.10.

Tabela 4.9 Resultados do Experimento 01 – Frases de Testes (Ind. 1-2-3)

		Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)			
		Ind. 1	Ind. 2	Ind. 3	
Locutor (Ind.)	Frases de Treinamento ↓	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	"Felicidade só é real quando compartilhada" (H. David Thoreau)	"Ser feliz sem motivo é mais autêntica forma de felicidade" (Carlos Drummond de Andrade)	Acertos
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	4,206	8,912	9,463	✓
2	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	8,874	4,542	8,602	✓
3	"Felicidade só é real quando compartilhada" (H. David Thoreau)	9,256	7,537	4,721	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	7,459	10,027	10,395	✓

Tabela 4.10 Resultados do Experimento 01 – Frases de Testes (Ind. 4-5-6)

		Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)			
		Ind. 4	Ind. 5	Ind. 6	
Locutor (Ind.)	Frases de Treinamento ↓	"Liberdade, Igualdade e Fraternidade". (Lema da Revolução Francesa)	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	"Saber como pensar torna a pessoa muito mais capaz do que aquele que apenas sabe o que deve pensar" (Neil deGrasse Tyson)	Acertos
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	7,516	8,761	9,663	✓
2	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	8,632	9,855	11,362	✓
3	"Felicidade só é real quando compartilhada" (H. David Thoreau)	9,471	10,279	10,857	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	4,637	9,305	9,225	✓

Tabela 4.11 Resultados do Experimento 01 – Frases de Testes (Ind. 7-8)

		Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)		
		Ind. 7	Ind. 8	
Locutor (Ind.)	Frases de Treinamento ↓	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	Acertos
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	9,347	8,631	✓
2	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	7,524	8,910	✓
3	"Felicidade só é real quando compartilhada" (H. David Thoreau)	8,771	9,424	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	9,805	6,892	✓

As distâncias Euclidianas apresentados nas tabelas com relação às frases de testes são superiores, quando comparados com comandos de testes. Esse fato se deve pelas frases terem maiores variabilidades fonéticas, portanto, são mais propícias a diferenciações durante as gravações de treinamento e teste.

Os resultados obtidos com relação às frases de testes são passíveis das mesmas conclusões generalizadas para os comandos de testes. A identificação dos locutores treinados e não reconhecimento daqueles que não estão registrados no banco de dados gerou um total de acertos de 92,50% com relação às frases de testes, isto é, em 148 das 160 amostras produzidas.

Com relação ao reconhecimento da frase de testes, para o caso em que a amostra foi treinada e testada pelo próprio locutor, houve 100% de acerto, fato que ocorreu para 4 amostras. Para as 12 amostras nas quais foram enunciadas frases presentes no banco de dados por outros autores, 10 amostras foram corretamente identificadas. Para esse espaço amostral, foi considerado a condição de indivíduos que treinaram o sistema quando enunciam frases do banco de dados que não são de sua autoria, e analisa-se as distâncias Euclidianas com excessão daquelas referentes às frases treinadas pelo próprio autor.

Assim, a sensibilidade para reconhecimento das frases de testes é de 87,5%, fato registrado para 14 das 16 amostras. Em 100% dos casos, o locutor e a frase por ele enunciada foram identificados corretamente, fato registrado para 4 amostras.

O gráfico apresentado na Figura 4.8 avalia a eficiência do sistema de reconhecimento automático de voz desenvolvido em sua configuração elementar para comandos e frases.

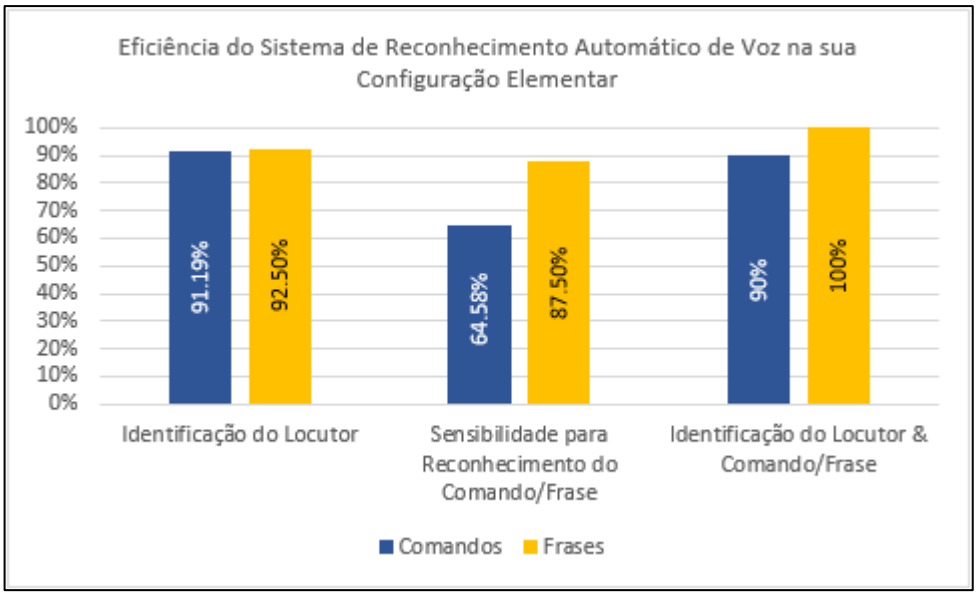

Figura 4.8 - Gráfico de eficiência do sistema de reconhecimento de voz para Comandos e Frases

4.4.2 Experimento 02 – Ensaio de Avaliação do Parâmetro VAD

O segundo experimento dessa pesquisa objetiva a avaliação do parâmetro de detecção de voz ativa (VAD) como um atributo que possa ser empregado no sistema de reconhecimento automático de voz para aumentar sua eficiência. Esta técnica tem como objetivo reduzir a quantidade de dados analisados distinguindo regiões silenciadas e com reproduções de ruído de regiões vocalizadas que contém informações úteis para o sistema de reconhecimento de voz. As Tabelas 4.11 a 4.13 esboçam os resultados para esse experimento.

Tabela 4.12 Resultados do Experimento 02 – Comandos de Testes (Ind. 1-2-3)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 1		Ind. 2		Ind. 3		
Locutor (Ind.) ↓	Comandos de Treinamentos ↓	Sinestesia	Vendemiário	Descanso	Multipoten-cialidade	Poesia	Evolucionismo	Acertos
1	Elo	2,943	2,648	5,825	6,352	7,498	6,553	✓
	Sinestesia	2,042	3,765	6,440	6,990	6,459	6,225	✓
	Terapêutico	3,254	4,309	6,195	5,978	7,200	5,448	✓
	Cosmologia	3,027	4,050	6,477	6,606	6,383	6,353	✓
	Cruzeiro	2,953	3,939	6,945	6,412	6,751	6,908	✓
2	Elo	8,345	7,002	2,446	3,289	5,950	7,269	✓
	Descanso	6,709	5,864	2,320	3,200	5,632	5,480	✓
	Natureza	6,750	6,549	2,860	3,418	5,587	5,761	✓
	Endorfina	7,082	7,348	3,274	4,055	5,392	5,807	✓
	Musicalidade	6,785	7,330	3,081	3,185	5,527	5,425	✓
3	Elo	6,755	6,158	5,443	6,045	2,590	3,004	✓
	Insólito	7,585	7,219	5,488	6,392	2,437	3,463	✓
	Multipotencialidade	7,747	6,688	6,119	5,660	3,449	4,100	✓
	Poesia	7,053	7,070	5,760	6,528	2,176	3,120	✓
	Rudimentar	7,455	6,482	6,472	6,006	3,290	2,985	✓
4	Elo	5,576	6,046	6,209	7,700	6,055	5,895	✓
	Genealogia	5,341	5,520	6,795	6,765	6,500	6,430	✓
	Orgânico	5,728	6,339	6,343	6,685	6,351	5,986	✓
	Evolucionismo	5,879	5,362	7,098	6,733	6,356	6,739	✓
	Herbáceo	5,872	5,526	6,397	6,576	6,966	6,068	✓

Tabela 4.13 Resultados do Experimento 02 – Comandos de Testes (Ind. 4-5-6)

		Comandos de Testes- Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 4		Ind. 5		Ind. 6		
Locutor (Ind.) ↓	Comandos de Treinamentos ↓	Alquimia	Elo	Cosmologia	Sinapses	Evolucionismo	Destreza	Acertos
1	Elo	6,392	5,889	5,844	5,881	5,957	6,719	✓
	Sinestesia	7,563	6,442	6,138	5,680	6,659	6,120	✓
	Terapêutico	6,348	5,933	5,854	6,482	6,918	6,346	✓
	Cosmologia	5,836	6,775	5,422	5,904	6,222	6,273	✓
	Cruzeiro	6,122	6,569	6,001	6,178	6,311	6,612	✓
2	Elo	7,025	5,789	6,778	6,627	6,543	6,290	✓
	Descanso	7,457	5,959	7,290	6,446	6,404	6,518	✓
	Natureza	7,266	5,952	6,690	6,427	7,439	6,536	✓
	Endorfina	7,352	7,030	7,156	6,773	6,845	6,859	✓
	Musicalidade	8,235	7,343	7,706	6,653	7,275	7,021	✓
3	Elo	6,927	6,259	6,873	5,795	6,443	6,090	✓
	Insólito	7,040	6,606	7,220	6,912	8,353	6,330	✓
	Multipotencialidade	7,211	7,261	7,112	7,442	6,988	7,401	✓
	Poesia	6,919	6,361	6,697	6,547	6,910	7,885	✓
	Rudimentar	8,514	6,721	6,954	6,746	7,320	6,725	✓
4	Elo	3,249	2,209	6,086	6,194	6,756	5,331	✓
	Genealogia	3,112	2,336	5,538	6,484	7,666	6,950	✓
	Orgânico	3,651	2,790	6,138	6,180	6,834	6,633	✓
	Evolucionismo	4,302	2,902	6,784	6,258	5,649	7,313	✓
	Herbáceo	3,128	2,674	5,872	6,051	6,334	6,724	✓

Tabela 4.14 Resultados do Experimento 02 – Comandos de Testes (Ind. 7-8)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)				
		Ind. 7		Ind. 8		
Locutor (Ind.) ↓	Comandos de Treinamentos ↓	Natureza	Poesia	Cruzeiro	Orgânico	Acertos
1	Elo	6,243	6,004	5,794	6,536	✓
	Sinestesia	7,031	5,483	5,827	6,597	✓
	Terapêutico	7,251	6,959	7,840	5,881	✓
	Cosmologia	6,693	6,415	6,608	6,384	✓
	Cruzeiro	6,202	7,086	5,389	6,608	✓
2	Elo	6,272	5,547	7,130	6,807	✓
	Descanso	5,812	6,448	7,060	6,249	✓
	Natureza	5,823	5,933	6,407	6,515	✓
	Endorfina	5,576	5,832	6,890	6,936	✓
	Musicalidade	6,108	6,123	7,483	6,935	✓
3	Elo	6,228	5,661	6,392	6,682	✓
	Insólito	6,574	6,476	7,717	6,287	✓
	Multipotencialidade	5,831	7,058	7,581	6,992	✓
	Poesia	5,640	5,266	6,573	6,801	✓
	Rudimentar	5,748	5,647	7,139	6,204	✓
4	Elo	6,886	6,675	6,504	6,413	✓
	Genealogia	6,913	6,343	5,795	6,382	✓
	Orgânico	6,347	7,479	6,344	6,273	✓
	Evolucionismo	6,758	7,249	6,913	6,703	✓
	Herbáceo	6,634	7,137	6,660	6,863	✓

A apuração dos resultados das distorções VQ para os comandos de testes utilizando o atributo de detecção de voz ativa (VAD) mostrou uma melhora no sistema de reconhecimento automático de voz desenvolvido.

Ao implementar o atributo VAD, o sistema foi capaz de reconhecer o locutor em 92,37% dos casos, resultado superior aos 91,19% obtidos no primeiro experimento. Em termos de reconhecimento do comando, para a condição de identificação dos comandos treinados e

testados pelo próprio indivíduo, o sistema teve a mesma performance, detectando 18 dos 20 comandos treinados (90% das vezes). Nesses casos, o locutor também foi corretamente identificado.

Com relação aos resultados associados aos locutores que treinaram o sistema e que enunciam comandos presentes no banco de dados, mas que não são de sua prórpia autoria, houve um aumento com relação à assertividade do reconhecimento desses comandos de 4 para 5 das 8 amostras possíveis, aumento de 12,5%. A sensibilidade do sistema para locutores não treinados que enunciaram comandos presentes no banco de dados permaneceu igual a 9 das 20 amostras possíveis.

Para a configuração utilizando VAD como atributo adicional ao sistema, houve um aumento na sensibilidade para reconhecimento de comandos enunciados presentes no banco de dados com relação aos resultados obtidos no Experimento 1. Foram registradas amostras assertivas em 32 dos 48 casos possíveis, um total de 66,67% de acertos.

As Figuras 4.9 e 4.10 mostram uma comparação dos resultados do sistema com a configuração elementar (apresentada no Experimento 1) e a configuração utilizando VAD para a palavra “Sinestesia”, quando esta é treinada e enunciada pelo Indivíduo 1, e para a palavra “Orgânico”, testada pelo Indivíduo 8 e treinada pelo Indivíduo 4, respectivamente.

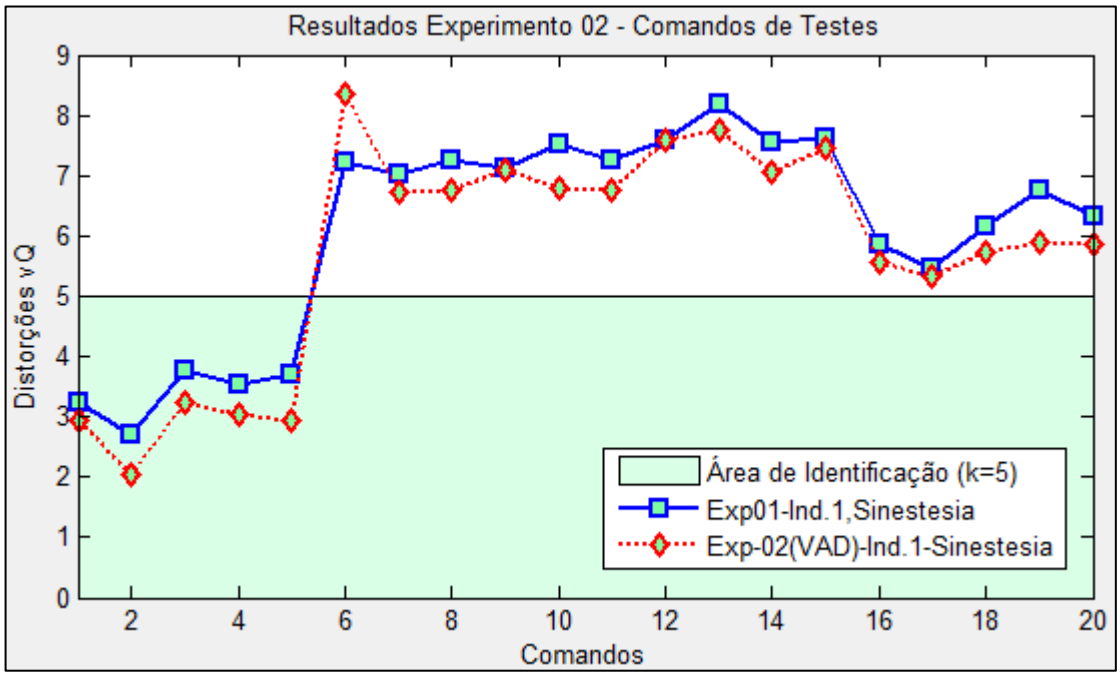

Figura 4.9 - Comparação de resultados do experimento 1 e 2 para o comando “Sinestesia”.

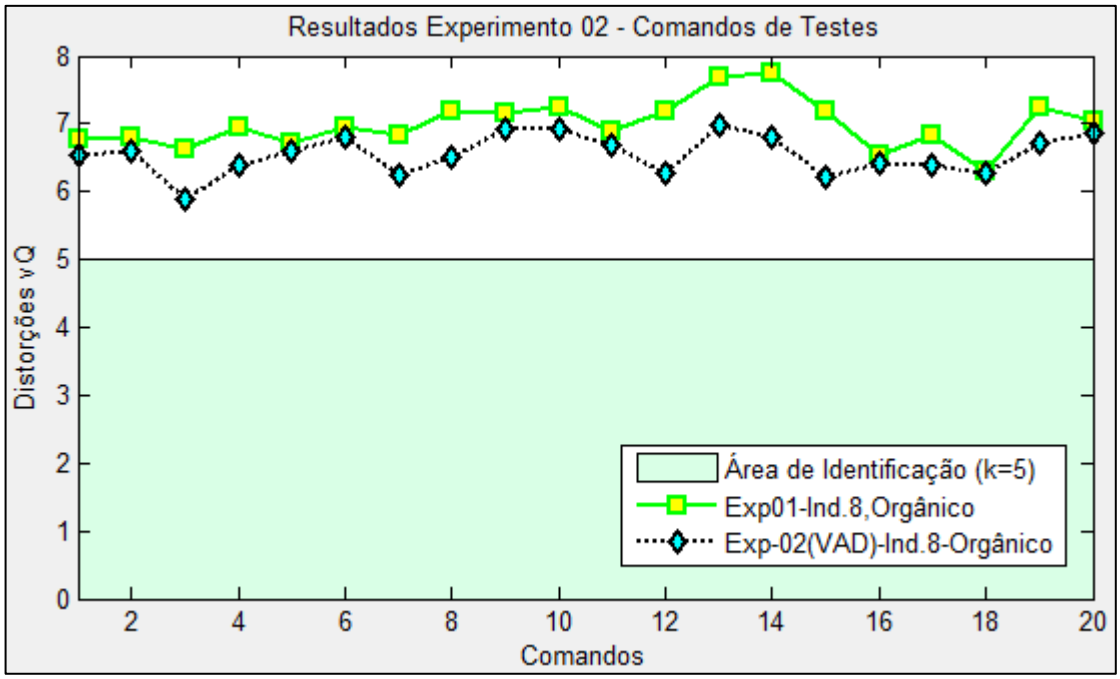

Figura 4.10 - Comparação de resultados do experimento 1 e 2 para o comando "Orgânico".

Os resultados para as frases de testes utilizando-se o atributo VAD estão apresentados a seguir através das Tabelas 4.14 a 4.16.

Tabela 4.15 Resultados do Experimento 02 – Frases de Testes (Ind. 1-3)

		Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)			
		Ind. 1	Ind. 2	Ind. 3	
Locutor (Ind.)	Frases de Treinamento ↓	Seu sonho é tão palpável quão grande sua vontade de realiza-lo.	"Felicidade só é real quando compartilhada" (H. David Thoreau)	"Ser feliz sem motivo é mais autêntica forma de felicidade". (Carlos Drummond de Andrade)	Acertos
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	2,707	8,744	8,377	✓
2	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	6,973	3,880	7,234	✓
3	"Felicidade só é real quando compartilhada" (H. David Thoreau)	7,078	7,260	2,442	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	6,622	8,477	9,455	✓

Tabela 4.16 Resultados do Experimento 02 – Frases de Testes (Ind. 4-5-6)

		Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)			
		Ind. 4	Ind. 5	Ind. 6	
Locutor (Ind.)	Frases de Treinamento ↓	"Liberdade, Igualdade e Fraternidade". (Lema da Revolução Francesa)	Seu sonho é tão palpável quão grande sua vontade de realiza-lo.	"Saber como pensar torna a pessoa muito mais capaz do que aquele que apenas sabe o que deve pensar". (Neil deGrasse Tyson)	Acertos
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo.	6,980	7,372	8,165	✓
2	"Tudo o que você precisa é de amor". (John Lennon & Paul McCartney)	7,140	9,430	11,300	✓
3	"Felicidade só é real quando compartilhada". (H. David Thoreau)	9,272	8,285	9,168	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo". (Robin Williams, *Death Poets Society*)	3,709	8,008	7,142	✓

Tabela 4.17 Resultados do Experimento 02 – Frases de Testes (Ind. 7-8)

		Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)		
		Ind. 7	Ind. 8	
Locutor	Frases de Treinamento ↓	"Tudo o que você precisa é de amor". (John Lennon & Paul McCartney)	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo". (Robin Williams, *Death Poets Society*)	Acertos
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo.	7,516	7,202	✓
2	"Tudo o que você precisa é de amor". (John Lennon & Paul McCartney)	6,857	8,378	✓
3	"Felicidade só é real quando compartilhada". (H. David Thoreau)	8,431	8,442	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo". (Robin Williams, *Death Poets Society*)	7,475	4,939	✓

Os resultados obtidos para as frases de testes nesse experimento apontaram que a inclusão do atributo de detecção de voz ativa usando o algoritmo de Qiang He aumentou a performance do sistema. Foram corretamente identificados os locutores em 149 das 160 amostras, um índice de 93,12%.

O sistema de reconhecimento de voz com VAD reconheceu as 4 frases treinadas e os indivíduos que as enunciaram corretamente, índice de 100% de acertos. Além disso, com essa configuração o sistema obteve o mesmo resultado encontrado para o primeiro experimento quanto á sensibilidade de identificação das frases presentes no banco de dados, quando essas eram enunciadas por indivíduos diferentes daqueles que as treinaram.

A Figura 4.11 compara os resultados para comandos e frases de testes obtidos nesse experimento com aqueles obtidos no primeiro experimento, o qual não utilizou a ferramenta VAD.

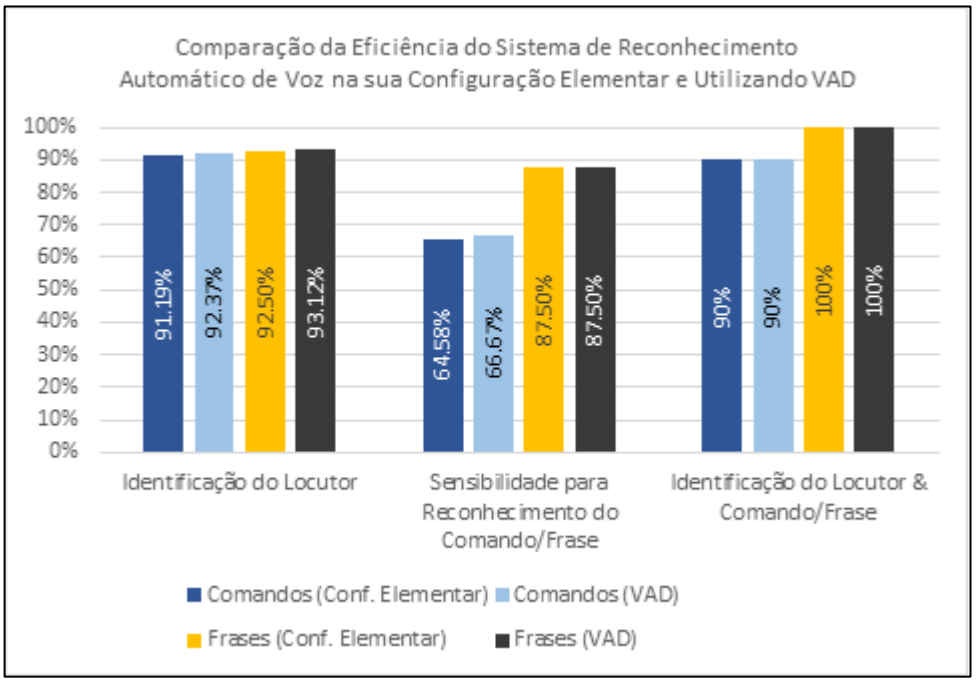

Figura 4.11 - Gráfico comparativo da eficiência do sistema usando VAD para comandos e frases.

Embora as gravações tenham sido realizadas em ambientes acusticamente controlados, a utilização da técnica VAD mostrou-se benéfica para a configuração do sistema. Esse resultado é compartilhado pelos pesquisadores Ling et al. (2009), Górriz et al. (2005).

4.4.3 Experimento 03 – Ensaio de Avaliação dos Parâmetros Dinâmicos DDC e SDC

O terceiro experimento dessa pesquisa estuda a implementação de coeficientes dinâmicos ao sistema de reconhecimento de voz. Serão avaliadas duas técnicas, *Delta-Delta Coefficients* (DDC) e *Shifted-Delta Coefficients* (SDC). O estudo contempla se o emprego de informações dinâmicas dos sinais de voz aos coeficientes estáticos calculados através de MFCC aperfeiçoará o sistema para reconhecimento do locutor e identificação de comandos e frases.

Os novos coeficientes são adicionados ao vetor acústico juntamente com os *Mel Frequency Cepstral Coefficients*, dessa forma, o vetor acústico contém informações estáticas (MFCC) e dinâmicas (DDC ou SDC).

Para esse experimento, o limiar das distâncias Euclidianas adotado foi igual a 3,2. Esse valor foi encontrado experimentalmente, e usado para maximizar o índice de reconhecimento correto dos comandos e frases enunciados, uma vez que os valores das distâncias VQ decaíram em relação aos experimentos anteriores.

As Tabelas 4.17 a 4.19 mostram resultados desse experimento. Nessas tabelas estão apresentados os valores das distorções VQ obtidos para cada atributo estudado (DDC e SDC) para um comando por indivíduo daqueles selecionados para apresentação nas tabelas dos experimentos anteriores.

Tabela 4.18 Resultados do Experimento 03 – Comandos de Testes (Ind. 1-2-3) (continua)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 1		Ind. 2		Ind. 3		
		DDC	SDC	DDC	SDC	DDC	SDC	
Locutor (Ind.)	Comandos de Treinamentos ↓	Sinestesia	Sinestesia	Multipotencialidade	Multipotencialidade	Poesia	Poesia	Acertos
1	Elo	2,329	1,363	5,109	3,879	4,693	3,877	✓
	Sinestesia	1,796	0,983	4,735	3,973	4,613	3,948	✓
	Terapêutico	2,312	1,497	4,313	4,265	4,915	4,014	✓
	Cosmologia	1,936	1,509	4,400	4,072	4,428	4,297	✓
	Cruzeiro	2,220	1,655	4,512	4,193	4,700	3,806	✓

Tabela 4.19 Resultados do Experimento 03 – Comandos de Testes (Ind. 1-2-3) (conclusão)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 1		Ind. 2		Ind. 3		
		DDC	SDC	DDC	SDC	DDC	SDC	
Locutor (Ind.)	Comandos de Treinamentos ↓	Sinestesia	Sinestesia	Multipotencialidade	Multipotencialidade	Poesia	Poesia	Acertos
2	Elo	4,296	4,115	1,928	2,086	3,817	3,347	✓
	Descanso	4,566	4,309	2,279	2,141	3,610	3,732	✓
	Natureza	4,469	3,836	2,027	1,983	3,761	3,829	✓
	Endorfina	4,606	4,141	2,536	2,107	4,007	3,846	✓
	Musicalidade	4,735	4,067	2,122	1,867	3,838	4,199	✓
3	Elo	4,652	3,819	4,185	3,626	1,923	1,351	✓
	Insólito	4,881	3,921	4,441	4,118	2,612	1,982	✓
	Multipotencialidade	4,521	4,281	3,809	3,556	2,188	1,897	✓
	Poesia	4,670	4,034	4,028	3,914	1,776	0,961	✓
	Rudimentar	4,916	4,469	4,150	4,087	2,178	1,405	✓
4	Elo	3,815	3,716	4,338	4,123	5,038	4,621	✓
	Genealogia	3,757	3,668	4,211	4,549	4,460	4,107	✓
	Orgânico	4,127	4,116	4,412	4,332	4,502	4,432	✓
	Evolucionismo	4,315	3,884	4,296	4,203	4,506	4,767	✓
	Herbáceo	4,058	3,805	4,239	4,925	4,577	4,636	✓

Tabela 4.20 Resultados do Experimento 03 – Comandos de Testes (Ind. 4-5-6)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 4		Ind. 5		Ind. 6		
		DDC	SDC	DDC	SDC	DDC	SDC	
Locutor (Ind.)	Comandos de Treinamentos ↓	Elo	Elo	Cosmologia	Cosmologia	Evolucionismo	Evolucionismo	Acertos
1	Elo	3,724	3,342	4,109	3,880	4,517	3,752	✓
	Sinestesia	4,091	3,528	4,330	4,579	4,406	4,007	✓
	Terapêutico	4,248	3,639	4,214	3,619	4,580	3,663	✓
	Cosmologia	4,223	3,444	4,043	3,297	4,386	3,678	✓
	Cruzeiro	4,422	3,931	4,986	3,802	4,957	3,971	✓
2	Elo	4,213	3,981	3,967	3,627	4,672	3,969	✓
	Descanso	4,635	4,112	4,233	3,994	4,490	4,091	✓
	Natureza	4,519	4,190	4,111	4,041	4,725	3,960	✓
	Endorfina	4,795	4,268	4,270	4,345	4,989	4,258	✓
	Musicalidade	5,068	4,037	4,521	4,290	5,202	4,344	✓
3	Elo	4,325	3,946	4,768	3,816	5,131	3,810	✓
	Insólito	4,454	4,190	4,591	4,179	4,972	4,118	✓
	Multipotencialidade	5,165	4,266	4,630	4,547	4,954	4,330	✓
	Poesia	4,412	4,028	4,479	4,154	5,088	3,952	✓
	Rudimentar	4,714	4,488	5,004	4,837	4,638	4,232	✓
4	Elo	1,567	1,158	4,323	3,537	4,736	3,984	✓
	Genealogia	2,059	1,297	4,299	3,766	5,330	3,808	✓
	Orgânico	2,164	1,992	4,146	3,963	4,749	3,870	✓
	Evolucionismo	2,457	2,225	4,391	4,169	4,234	3,295	✓
	Herbáceo	1,892	1,970	4,270	4,368	4,832	3,619	✓

Tabela 4.21 Resultados do Experimento 03 – Comandos de Testes (Ind. 7-8)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)				
		Ind. 7		Ind. 8		
		DDC	SDC	DDC	SDC	
Locutor (Ind.)	Comandos de Treinamentos ↓	Poesia	Poesia	Orgânico	Orgânico	Acertos
1	Elo	5,073	3,861	4,071	4,024	✓
	Sinestesia	4,629	4,082	4,174	3,652	✓
	Terapêutico	4,789	4,176	4,183	3,841	✓
	Cosmologia	4,931	3,763	4,103	3,424	✓
	Cruzeiro	4,406	4,188	4,226	3,669	✓
2	Elo	4,166	3,478	4,691	4,031	✓
	Descanso	4,223	3,765	4,803	4,105	✓
	Natureza	4,188	3,726	5,015	4,226	✓
	Endorfina	4,468	4,215	4,465	4,127	✓
	Musicalidade	4,407	4,302	4,793	4,271	✓
3	Elo	3,929	3,619	4,566	3,945	✓
	Insólito	4,174	3,965	4,642	4,156	✓
	Multipotencialidade	4,249	3,754	4,577	4,184	✓
	Poesia	3,826	3,513	4,605	4,323	✓
	Rudimentar	4,152	3,950	4,585	4,267	✓
4	Elo	5,033	4,159	4,105	3,931	✓
	Genealogia	5,007	4,262	4,044	3,859	✓
	Orgânico	5,084	4,372	3,882	3,649	✓
	Evolucionismo	4,533	4,492	4,188	3,698	✓
	Herbáceo	5,145	4,336	4,265	4,011	✓

A análise dos resultados dos comandos de teste para esse experimento mostrou que a utilização de atributos dinâmicos incorporados aos coeficientes cepstrais aumentam a eficiência do sistema para o cumprimento das tarefas propostas. Concluiu-se ainda que os *Shifted-Delta Coefficients* (SDC) apresentaram resultados superiores aos *Delta-Delta Coefficients* (DDC).

A identificação do locutor foi assertiva para 1513 amostras usando-se DDC, índice de acertos igual a 94,56% e 1528 amostras usando-se SDC, índice de acertos igual a 95,5%. Em

ambos os casos, todos os 20 comandos enunciados pelos próprios autores foram reconhecidos, assim como seus autores.

Com relação às amostras nas quais os locutores treinados enunciaram comandos do banco de dados treinados por outros indivíduos, os resultados foram semelhantes para ambos os casos, com reconhecimento assertivo para 5 das 8 amostras dessa ocorrência (resultado similar ao obtido pelo Experimento 2).

Para a configuração do sistema de reconhecimento de voz usando DDC, em 9 das 20 amostras, nas quais os locutores que apenas testaram o sistema enunciam comandos do banco de dados, houveram a correta identificação do comando enunciado. Para a configuração do sistema usando o atributo SDC, esse resultado subiu para 11 amostras, aumento de 10% no índice de acertos.

Assim, quanto à sensibilidade para reconhecimento de comandos do banco de dados, para a configuração utilizando o atributo DDC os resultados assertivos ocorreram para 34 das 48 amostras com essa ocorrência, índice de 70,83%. E, para o sistema utilizando o atributo SDC, os resultados foram superiores, sendo estes assertivos para 36 das 48 amostras, índice de 75% de assertividade.

As Figuras 4.12 e 4.13 comparam graficamente os resultados coletados para os atributos dinâmicos DDC e SDC com aqueles obtidos no primeiro experimento para a palavra "Sinestesia", treinada pelo Indivíduo 1 para a palavra "Orgânico", treinada pélo indivíduo 4 e testada pelo indivíduo 8, respectivamente.

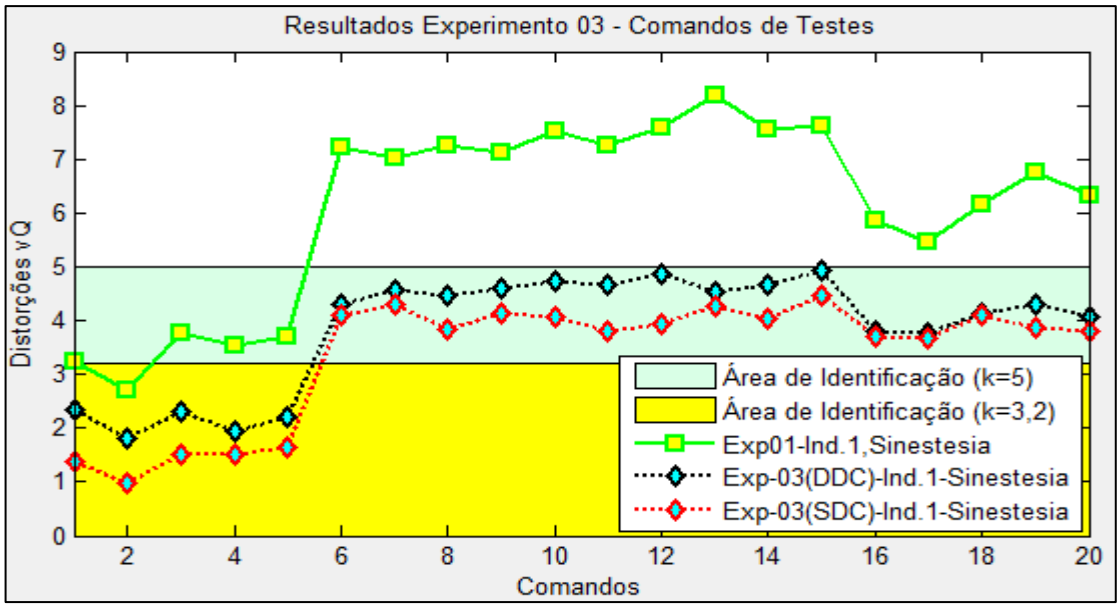

Figura 4.12 - Comparação de resultados para a palavra "Sinestesia" entre o Experimento 1 (com coeficientes estáticos) e o Experimento 3 (com coeficientes dinâmicos: DDC ou SDC).

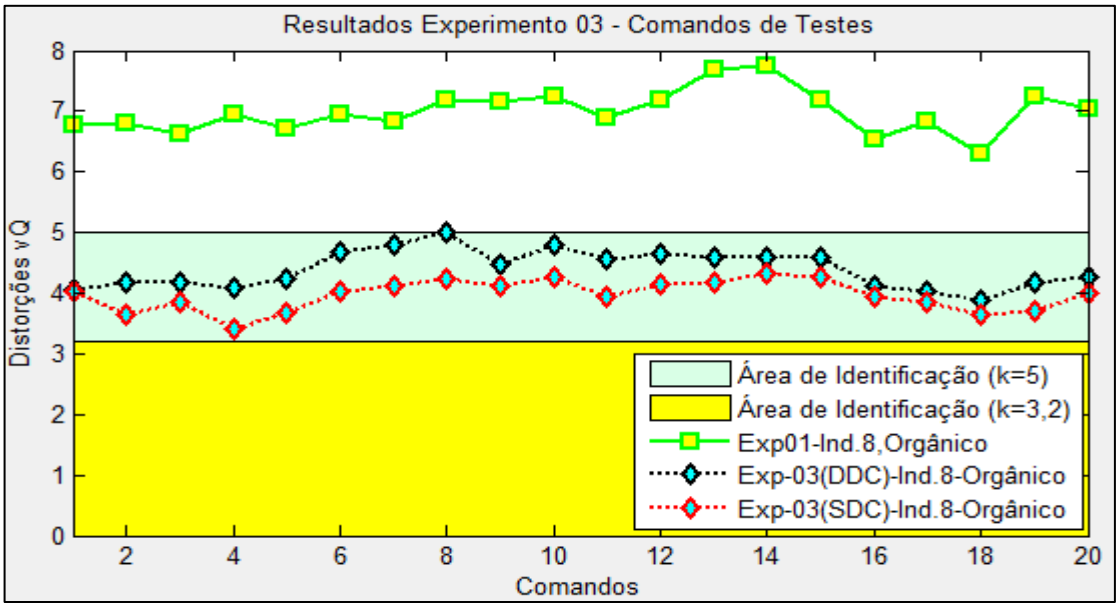

Figura 4.13 - Comparação de resultados para a palavra "Orgânico" entre o Experimento 1 (com coeficientes estáticos) e o Experimento 3 (com coeficientes dinâmicos: DDC ou SDC).

Os resultados para as frases de testes são apresentados nas Tabelas 4.20, 4.21 e 4.22.

Tabela 4.22 Resultados do Experimento 03 – Frases de Testes (Ind. 1-2-3)

Locutor (Ind.)	Frases de Treinamento ↓	Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						Acertos
		Ind. 1		Ind. 2		Ind. 3		
		Seu sonho é tão palpável quão grande sua vontade de realiza-lo.		"Felicidade só é real quando compartilhada". (H. David Thoreau)		"Ser feliz sem motivo é mais autêntica forma de felicidade". (Carlos Drummond de Andrade)		
		DDC	SDC	DDC	SDC	DDC	SDC	
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	2,122	1,862	5,943	6,096	5,975	5,055	✓
2	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	5,969	3,983	2,494	1,775	4,892	3,838	✓
3	"Felicidade só é real quando compartilhada" (H. David Thoreau)	4,755	5,267	4,236	4,469	2,960	1,842	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	4,105	3,447	6,236	4,604	5,272	4,544	✓

Tabela 4.23 Resultados do Experimento 03 – Frases de Testes (Ind. 4-5-6)

		Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 4		Ind. 5		Ind. 6		
Locutor (Ind.)	Frases de Treinamento ↓	"Liberdade, Igualdade e Fraternidade". (Lema da Revolução Francesa)		Seu sonho é tão palpável quão grande sua vontade de realiza-lo.		"Saber como pensar torna a pessoa muito mais capaz do que aquele que apenas sabe o que deve pensar". (Neil deGrasse Tyson)		Acertos
		DDC	SDC	DDC	SDC	DDC	SDC	
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	3,569	4,109	4,876	3,464	4,367	4,802	✓
2	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	5,306	4,861	6,794	6,077	7,816	6,138	✓
3	"Felicidade só é real quando compartilhada" (H. David Thoreau)	4,788	4,553	7,200	4,052	7,721	6,204	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	2,118	1,916	4,921	5,352	6,790	5,729	✓

Tabela 4.24 Resultados do Experimento 03 – Frases de Testes (Ind. 7-8) (continua)

		Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)				
		Ind. 7		Ind. 8		
Locutor	Frases de Treinamento ↓	"Tudo o que você precisa é de amor". (John Lennon & Paul McCartney)		"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo". (Robin Williams, *Death Poets Society*)		Acertos
		DDC	SDC	DDC	SDC	
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	6,487	4,482	4,128	4,815	✓
2	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	3,866	3,441	6,314	4,590	✓
3	"Felicidade só é real quando compartilhada" (H. David Thoreau)	4,471	3,939	4,506	5,935	✓

Tabela 4.25 Resultados do Experimento 03 – Frases de Testes (Ind. 7-8) (conclusão)

Locutor	Frases de Treinamento ↓	Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)				Acertos
		Ind. 7		Ind. 8		
		"Tudo o que você precisa é de amor". (John Lennon & Paul McCartney)		"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo". (Robin Williams, *Death Poets Society*)		
		DDC	SDC	DDC	SDC	
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	5,019	5,688	3,552	4,087	✓

Os resultados para as frases de testes para esse experimento, assim como aqueles obtidos para os comandos de testes, mostraram que houve um aperfeiçoamento no sistema de reconhecimento de voz quando se utiliza atributos dinâmicos nos vetores acústicos. Esse fato foi registrado para ambas as técnicas em análise nesse experimento, DDC e SDC.

O sistema de reconhecimento de voz incrementado com a técnica DDC foi capaz de identificar corretamente o locutor em 151 das 160 amostras para esse caso, índice de 94,37%. Usando a técnica SDC os resultados foram ligeiramente superiores, com assertividade para 153 amostras, índice de 95,62%.

Como nos experimentos anteriores, tanto para o sistema implementado com a técnica DDC quanto com SDC, houve o correto reconhecimento da frase do banco de dados e identificação do locutor para todas as 4 amostras para essa ocorrência. Quanto à sensibilidade para o reconhecimento das frases presentes no banco de dados, quando essas eram enunciadas por indivíduos que não as treinaram, para ambos os sistemas em estudo nesse experimento, em 11 das 12 amostras observou-se a distância mínima referente à frase enunciada, apurando-se, portanto, uma sensibilidade para reconhecimento de frases do sistema em 93,75% dos casos.

A Figura 4.14 compara a eficiência dos sistemas estudados nesse experimento, utilizando coeficientes dinâmicos com a configuração elementar, quando usou-se apenas coeficientes estáticos, e a Figura 4.15 compara os dados referentes às frases de testes para as mesmas condições.

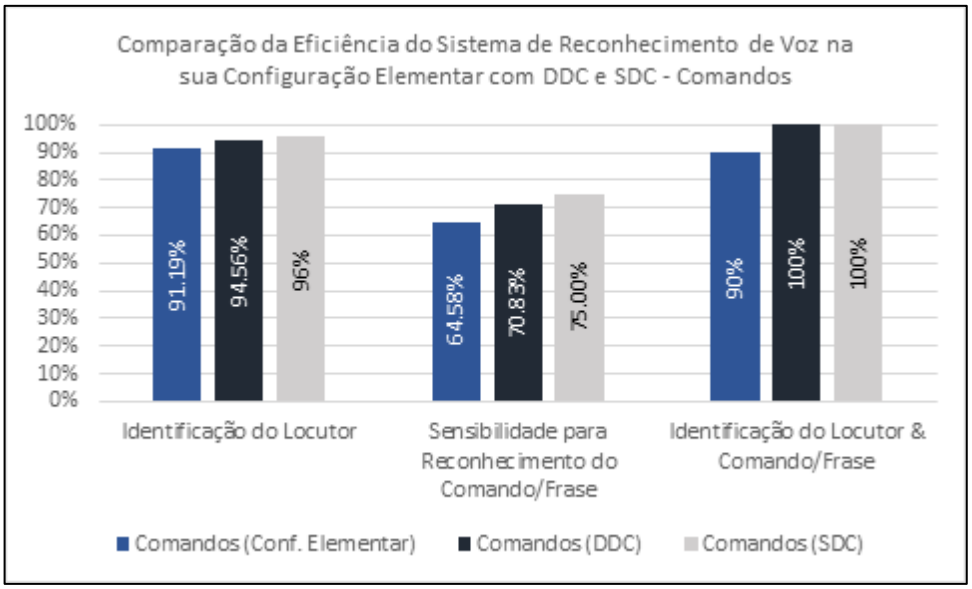

Figura 4.14 - Gráfico da eficiência do sistema usando DDC e SDC para comandos de testes.

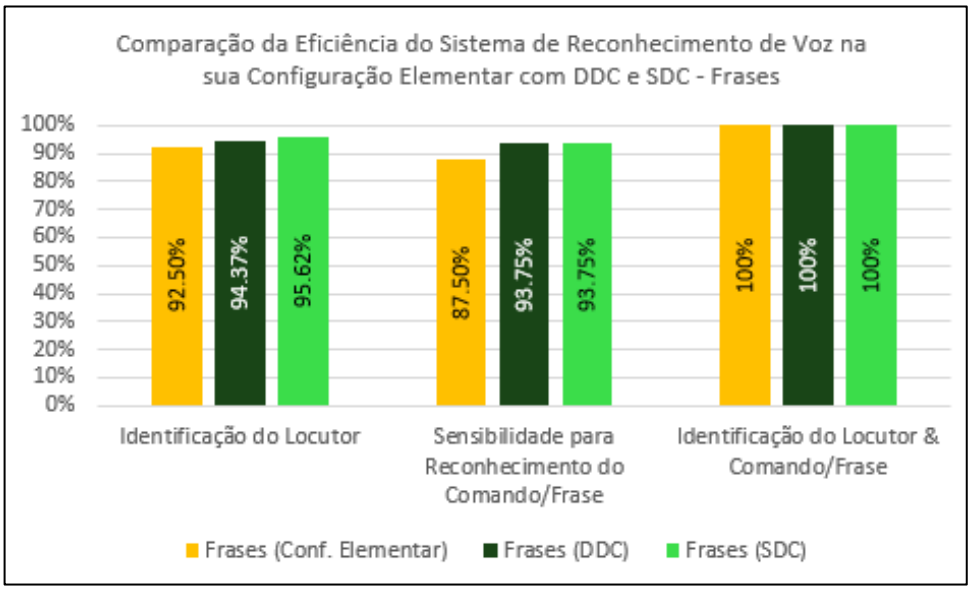

Figura 4.15 - Gráfico da eficiência do sistema usando DDC e SDC para frases de testes.

Os resultados apurados foram satisfatórios e correspondentes às pesquisas realizadas por Allen et al. (2005), Torres Carrasquillo et al. (2002) e Kumar et al. (2011).

4.4.4 Experimento 04 – Ensaio de Avaliação dos Parâmetros de Normalização

O quarto experimento dessa pesquisa estuda a aplicação de atributos de normalização dos coeficientes cepstrais. Foram analisados 3 diferentes técnicas de normalização: *Cepstral Mean and Variance Normalization* (CMVN); *Windowed Cepstral Mean and Variance Normalization* (WCMVN) e *Short-Time Gaussianization* (STG).

As técnicas de normalização são aplicadas para compensar efeitos de ruídos externos e provenientes do canal de entrada, além de diferenças entre os ambientes de treinamentos e testes captados por microfones, e de reduzir efeitos de ressonâncias geradas no momento de gravação. As técnicas de normalização foram aplicadas sobre os coeficientes cepstrais na etapa de pós-processamento do sinal.

O limiar das distâncias Euclidianas adotado nesse experimento foi igual a 5 para o sistema de reconhecimento de voz implementado com a técnica CMVN e igual a 3,2, para os sistemas iomplementados com WCMVN e STG, mesmo valor utilizado no Experimento 3.

Os resultados para os três diferentes modelos de normalização em estudo neste experimento estão apresentados nas Tabelas 4.23 a 4.26 para comparação. Como para o Experimento 3, serão apresentadas as distorções VQ de apenas um dos comandos por indivíduo (os mesmos utilizados no Experimento 03).

Tabela 4.26 Resultados do Experimento 04 – Comandos de Testes (Ind. 1-2) (continua)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 1			Ind. 2			
		CMVN	WCMVN	STG	CMVN	WCMVN	STG	
Locutor (Ind.)	Comandos de Treinamentos ↓	Sinestesia	Sinestesia	Sinestesia	Multipotencialidade	Multipotencialidade	Multipotencialidade	Acertos
1	Elo	2,321	1,917	1,654	6,057	5,589	4,458	✓
	Sinestesia	1,995	1,422	1,257	5,993	5,329	4,897	✓
	Terapêutico	2,665	2,312	1,818	5,948	4,795	4,888	✓
	Cosmologia	2,444	1,888	1,962	7,565	5,033	4,732	✓
	Cruzeiro	2,548	2,236	2,020	5,979	5,229	4,863	✓

Tabela 4.27 Resultados do Experimento 04 – Comandos de Testes (Ind. 1-2) (conclusão)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 1			Ind. 2			
		CMVN	WCMVN	STG	CMVN	WCMVN	STG	
Locutor (Ind.)	Comandos de Treinamentos ↓	Sinestesia	Sinestesia	Sinestesia	Multipotencialidade	Multipotencialidade	Multipotencialidade	Acertos
2	Elo	7,859	5,030	4,877	2,982	2,122	1,897	✓
	Descanso	6,494	5,309	5,061	2,544	2,174	2,074	✓
	Natureza	6,295	4,690	4,433	3,696	2,261	2,053	✓
	Endorfina	6,349	4,488	4,354	2,283	2,337	2,169	✓
	Musicalidade	6,372	4,713	4,599	2,013	2,033	1,815	✓
3	Elo	6,468	4,722	4,627	6,144	4,870	4,538	✓
	Insólito	6,466	4,775	4,992	5,714	4,636	4,533	✓
	Multipotencialidade	6,640	4,458	4,753	5,260	4,118	4,181	✓
	Poesia	7,480	5,086	4,626	5,896	4,502	4,253	✓
	Rudimentar	6,207	5,295	5,147	5,458	4,753	4,430	✓
4	Elo	6,027	3,923	3,792	6,614	5,077	4,845	✓
	Genealogia	5,856	3,860	3,641	5,987	4,945	4,982	✓
	Orgânico	6,066	4,324	4,180	6,166	5,401	5,084	✓
	Evolucionismo	5,515	4,018	3,876	6,171	5,095	4,735	✓
	Herbáceo	5,308	4,273	4,035	5,762	5,274	5,289	✓

Tabela 4.28 Resultados do Experimento 04 – Comandos de Testes (Ind. 3-4)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 3			Ind. 4			
		CMVN	WCMVN	STG	CMVN	WCMVN	STG	
Locutor (Ind.)	Comandos de Treinamentos ↓	Poesia	Poesia	Poesia	Elo	Elo	Elo	Acertos
1	Elo	7,825	4,803	4,924	5,646	4,117	4,081	✓
	Sinestesia	6,088	5,315	4,951	6,282	4,305	4,457	✓
	Terapêutico	6,174	5,060	4,846	5,758	4,636	4,428	✓
	Cosmologia	6,369	4,658	4,202	6,282	4,275	4,506	✓
	Cruzeiro	6,789	4,987	4,682	5,810	4,861	4,388	✓
2	Elo	5,838	4,397	3,837	6,124	4,865	4,866	✓
	Descanso	5,531	4,622	4,285	6,874	5,074	5,017	✓
	Natureza	5,609	4,719	4,119	6,754	4,964	4,993	✓
	Endorfina	6,042	4,162	4,369	6,545	5,418	4,916	✓
	Musicalidade	5,890	4,944	4,228	6,393	5,325	5,020	✓
3	Elo	2,048	2,071	1,795	6,531	4,530	4,767	✓
	Insólito	2,189	2,064	1,864	7,004	5,200	4,924	✓
	Multipotencialidade	2,423	1,506	2,245	6,962	5,171	4,954	✓
	Poesia	1,729	1,829	1,279	6,715	4,931	5,103	✓
	Rudimentar	3,335	2,010	2,090	6,966	5,226	5,352	✓
4	Elo	6,370	5,167	5,149	1,561	1,876	1,711	✓
	Genealogia	6,162	5,273	4,767	1,909	2,044	1,459	✓
	Orgânico	6,746	5,288	4,337	2,255	2,477	2,074	✓
	Evolucionismo	7,053	5,112	4,793	2,284	2,350	1,927	✓
	Herbáceo	6,338	5,283	5,262	2,288	2,195	2,196	✓

Tabela 4.29 Resultados do Experimento 04 – Comandos de Testes (Ind. 5-6)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 5			Ind. 6			
		CMVN	WCMVN	STG	CMVN	WCMVN	STG	
Locutor (Ind.)	Comandos de Treinamentos ↓	Cosmologia	Cosmologia	Cosmologia	Evolucionismo	Evolucionismo	Evolucionismo	Acertos
1	Elo	6,060	4,198	4,230	6,036	4,323	3,959	✓
	Sinestesia	5,989	4,236	4,218	5,887	4,248	4,185	✓
	Terapêutico	5,773	4,407	4,568	6,142	4,573	4,260	✓
	Cosmologia	5,584	4,008	4,092	6,761	4,436	4,039	✓
	Cruzeiro	6,288	4,572	4,737	6,381	4,502	4,344	✓
2	Elo	5,929	4,630	4,802	7,494	4,422	4,385	✓
	Descanso	6,069	4,228	4,373	6,737	4,909	4,485	✓
	Natureza	6,072	4,931	4,390	7,343	5,391	4,250	✓
	Endorfina	6,317	5,151	4,268	7,260	5,449	4,746	✓
	Musicalidade	6,564	5,244	4,497	6,826	5,400	4,901	✓
3	Elo	5,912	5,426	4,544	7,062	5,370	4,727	✓
	Insólito	6,143	5,338	4,696	6,643	4,927	4,142	✓
	Multipotencialidade	6,394	4,928	4,579	6,920	5,070	4,802	✓
	Poesia	5,963	4,855	4,656	6,510	5,291	4,815	✓
	Rudimentar	6,330	5,557	5,050	6,191	4,820	4,499	✓
4	Elo	5,996	4,147	4,647	6,253	5,102	4,133	✓
	Genealogia	5,843	4,194	4,747	6,504	5,412	4,009	✓
	Orgânico	5,853	4,340	4,219	6,147	5,060	4,290	✓
	Evolucionismo	5,784	4,209	4,214	5,702	4,167	3,873	✓
	Herbáceo	5,988	4,579	4,381	5,980	5,446	4,341	✓

Tabela 4.30 Resultados do Experimento 04 – Comandos de Testes (Ind. 7-8)

		Comandos de Testes Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 7			Ind. 8			
		CMVN	WCMVN	STG	CMVN	WCMVN	STG	
Locutor (Ind.)	Comandos de Treinamentos ↓	Poesia	Poesia	Poesia	Orgânico	Orgânico	Orgânico	Acertos
1	Elo	6,255	5,162	4,468	6,194	4,274	4,126	✓
	Sinestesia	6,350	5,119	4,560	6,302	4,344	4,231	✓
	Terapêutico	6,452	5,185	5,011	6,064	4,480	4,649	✓
	Cosmologia	6,335	5,223	4,767	6,066	4,356	3,999	✓
	Cruzeiro	6,660	4,863	4,847	6,419	4,796	4,136	✓
2	Elo	6,337	4,791	4,268	7,531	4,782	4,509	✓
	Descanso	5,732	4,424	4,181	7,156	5,303	4,859	✓
	Natureza	5,665	4,590	4,150	7,250	5,509	4,486	✓
	Endorfina	6,117	4,631	4,450	7,538	5,151	5,074	✓
	Musicalidade	6,112	4,324	4,553	6,911	5,270	4,432	✓
3	Elo	5,884	4,419	4,096	6,481	5,115	4,705	✓
	Insólito	6,372	4,630	4,323	7,284	5,330	5,206	✓
	Multipotencialidade	5,799	4,565	4,251	7,104	5,198	4,921	✓
	Poesia	5,549	4,116	3,775	6,787	5,272	4,684	✓
	Rudimentar	6,173	5,041	4,197	7,084	4,964	5,481	✓
4	Elo	6,077	5,218	4,652	6,327	4,779	4,876	✓
	Genealogia	7,336	5,260	4,460	6,732	4,534	4,204	✓
	Orgânico	6,282	5,259	4,972	6,031	4,227	4,003	✓
	Evolucionismo	6,162	5,199	4,944	6,415	4,696	4,296	✓
	Herbáceo	6,387	4,686	5,343	6,255	4,756	4,263	✓

As Figuras 4.16 e 4.17 comparam graficamente os resultados.

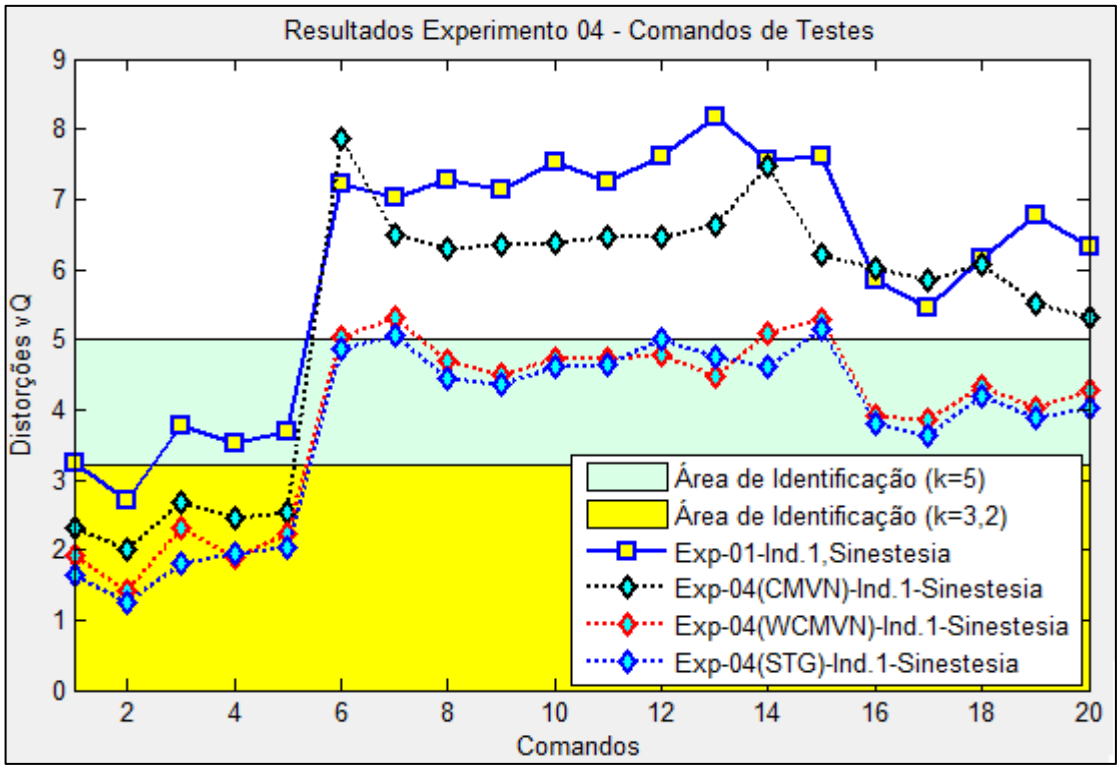

Figura 4.16 - Comparação de resultados para a palavra "Sinestesia" entre o Experimento 1 e o Experimento 4 para as técnicas: CMVN, WCMVN, e STG.

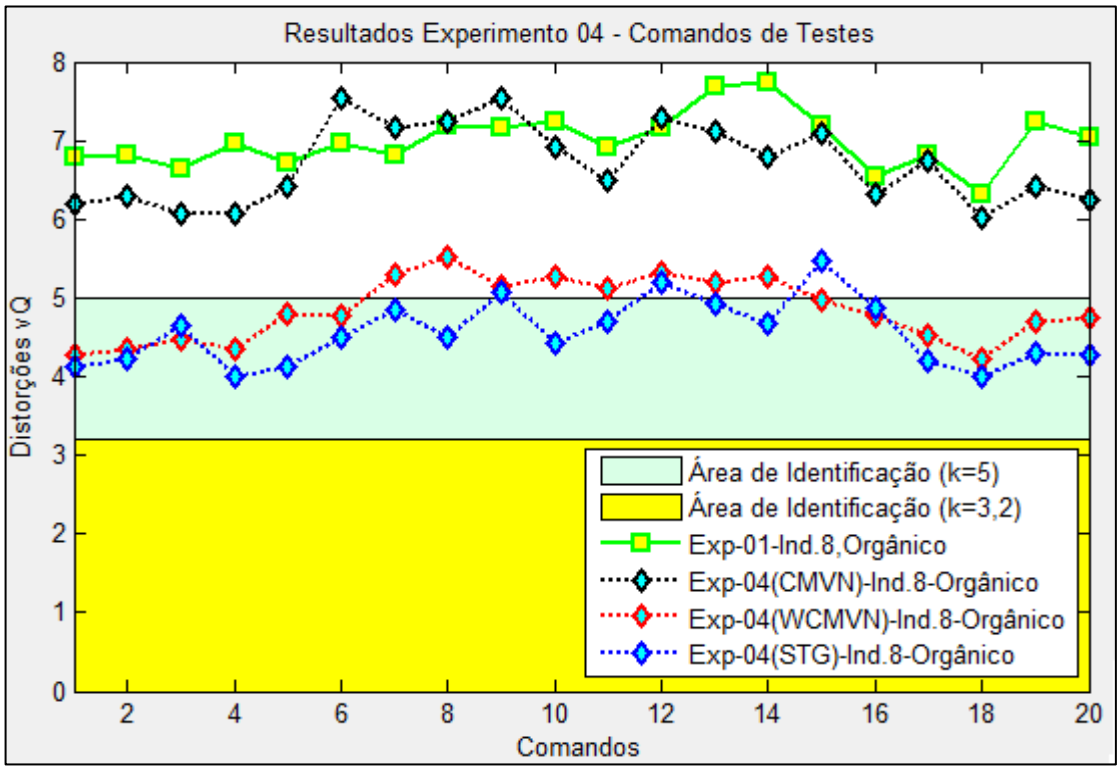

Figura 4.17 - Comparação de resultados para a palavra "Orgânico" entre o Experimento 1 e o Experimento 4 para as técnicas: CMVN, WCMVN, e STG.

Os resultados coletados para os sistemas de reconhecimento de voz implementados com atributos de normalização apontaram uma melhora na capacidade de identificação dos comandos do banco de dados e reconhecimento do locutor, em relação ao Experimento 1.

Os sistemas configurados com atributos de normalização apresentaram os seguintes índices de reconhecimento do locutor: 1489 amostras usando CMVN, índice de 93,06% de acertos; 1509 amostras usando WCMVN, índice de 94,31% de acertos; e 1511 amostras usando STG, índice de 94,43% de assertividade.

Para todas as configurações do sitema com ferramentas de normalização, houve 100% de identificação dos comandos treinados e testados pelos prórpios autores. Além disso, em todos os casos, verificou-se a menor distância Euclidiana para comandos presentes no banco de dados e testados por outros indivíduos que treinaram o sistema em 6 das 8 amostras com essa ocorrência.

Com relação às amostras em que um locutor que apenas testou o sistema enuncia algum comando do banco de dados, para os sistemas implementados com CMVN e WCMVN, em 9 dessas 20 ocorrências houveram a detecção assertiva dos comandos. Este número aumentou para 11 amostras quando o sistema foi implementado com a técnica STG. Assim, os índices de reconhecimento dos comandos presentes no banco de dados foram iguais a 72,92% para CMVN e WCMVN, e 77,08% para STG.

As Tabelas 4.27 a 4.30 apresentam os resultados do sistema utilizando atributos de normalização para as frases de testes.

Tabela 4.31 Resultados do Experimento 04 – Frases de Testes (Ind. 1-2) (continua)

		Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 1			Ind. 2			
Locutor	Frases de Treinamento ↓	Seu sonho é tão palpável quão grande sua vontade de realiza-lo			"Felicidade só é real quando compartilhada" (H. David Thoreau)			Acertos
		CMVN	WCMVN	STG	CMVN	WCMVN	STG	
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	2,078	1,686	0,903	9,027	5,039	5,068	✓
2	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	6,774	6,322	3,912	2,650	1,921	1,530	✓

Tabela 4.32 Resultados do Experimento 04 – Frases de Testes (Ind. 1-2) (conclusão)

Locutor	Frases de Treinamento ↓	Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						Acertos
		Ind. 1			Ind. 2			
		Seu sonho é tão palpável quão grande sua vontade de realiza-lo			"Felicidade só é real quando compartilhada" (H. David Thoreau)			
		CMVN	WCMVN	STG	CMVN	WCMVN	STG	
3	"Felicidade só é real quando compartilhada" (H. David Thoreau)	6,530	4,334	4,379	6,187	4,035	4,213	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	5,754	4,088	4,162	8,025	5,252	4,777	✓

Tabela 4.33 Resultados do Experimento 04 – Frases de Testes (Ind. 3-4)

Locutor	Frases de Treinamento ↓	Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						Acertos
		Ind. 3			Ind. 4			
		"Ser feliz sem motivo é mais autêntica forma de felicidade" (Carlos Drummond de Andrade)			"Liberdade, Igualdade e Fraternidade" (Lema da Revolução Francesa)			
		CMVN	WCMVN	STG	CMVN	WCMVN	STG	
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	7,013	6,342	4,405	5,864	3,801	3,701	✓
2	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	6,278	4,492	4,026	6,019	4,928	4,118	✓
3	"Felicidade só é real quando compartilhada" (H. David Thoreau)	3,766	2,295	1,602	7,078	4,357	4,066	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	7,226	4,596	5,150	2,417	1,815	1,333	✓

Tabela 4.34 Resultados do Experimento 04 – Frases de Testes (Ind. 5-6)

Locutor	Frases de Treinamento ↓	Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						Acertos
		Ind. 5			Ind. 6			
		Seu sonho é tão palpável quão grande sua vontade de realiza-lo			"Saber como pensar torna a pessoa muito mais capaz do que aquele que apenas sabe o que deve pensar" (Neil deGrasse Tyson)			
		CMVN	WCMVN	STG	CMVN	WCMVN	STG	
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	6,378	4,433	3,824	6,510	4,474	4,240	✓
2	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	9,188	5,630	6,082	7,221	6,985	4,750	✓
2	"Felicidade só é real quando compartilhada" (H. David Thoreau)	6,944	6,804	4,783	9,079	5,949	4,978	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	7,149	4,699	4,679	6,917	5,298	4,594	✓

Tabela 4.35 Resultados do Experimento 04 – Frases de Testes (Ind. 7-8) (continua)

Locutor	Frases de Treinamento ↓	Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						Acertos
		Ind. 7			Ind. 8			
		"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)			"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)			
		CMVN	WCMVN	STG	CMVN	WCMVN	STG	
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	7,242	5,567	3,995	6,734	4,058	4,114	✓
2	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	5,252	4,009	3,329	7,415	6,208	4,424	✓

Tabela 4.36 Resultados do Experimento 04 – Frases de Testes (Ind. 7-8) (conclusão)

Locutor	Frases de Treinamento ↓	Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						Acertos
		Ind. 7			Ind. 8			
		"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)			"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)			
		CMVN	WCMVN	STG	CMVN	WCMVN	STG	
3	"Felicidade só é real quando compartilhada" (H. David Thoreau)	6,186	4,343	3,652	7,909	4,348	4,863	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	7,838	5,806	4,573	5,164	3,901	3,532	✓

As frases de testes analisadas para as configurações com atributos de normalização corroboraram com os resultados obtidos para os comandos de testes. Isso é, houve uma melhora no sistema tanto para a identificação das frases quanto para o reconhecimento dos locutores, comparando esses resultados com o primeiro experimento.

A identificação do locutor para as frases de testes ocorreu de maneira assertiva para 156 das 160 amostras em todos os três casos (sistema de reconhecimento de voz implementado com a técnica de normalização dos coeficientes cepstrais CMVN, ou WCMVN, ou STG), índice de 97,5% de acertos.

Todas as 4 amostras de frases treinadas e enunciadas pelos próprios autores foram corretamente identificadas nos três sistemas. Além disso, para todos os casos constatou-se a sensibilidade para o reconhecimento de frases pertencentes ao banco de dados que não foram enunciadas pelos próprios autores em 100% das ocorrências.

As Figuras 4.18 e 4.19 apresentam gráficos comparativos dos resultados obtidos nesse experimento usando as três técnicas de normalização com o primeiro experimento para os comandos e frases de testes, respectivamente.

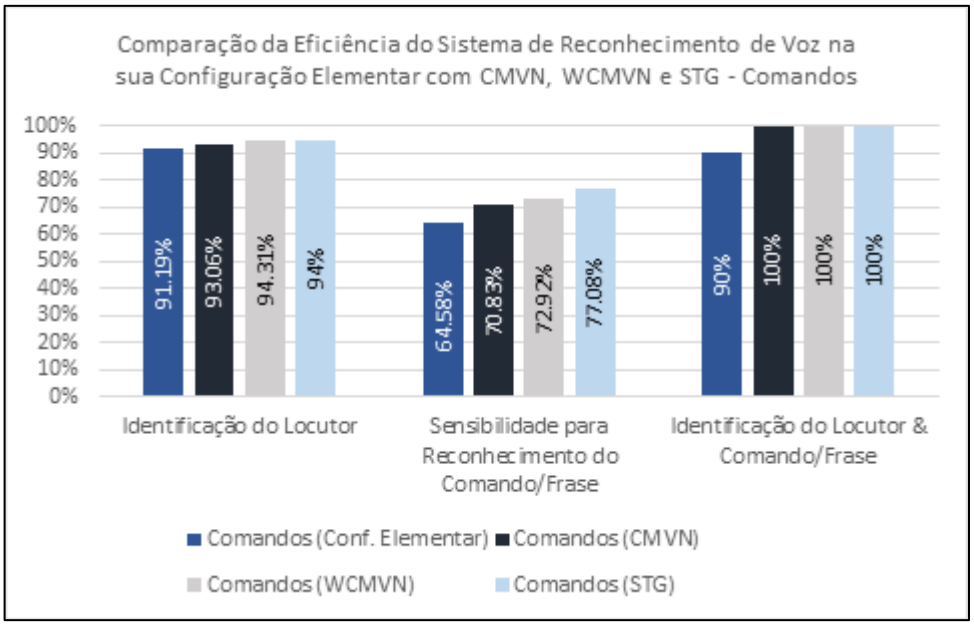

Figura 4.18 - Gráfico de eficiência do sistema para CMVN, WCMVN e STG - comandos de testes.

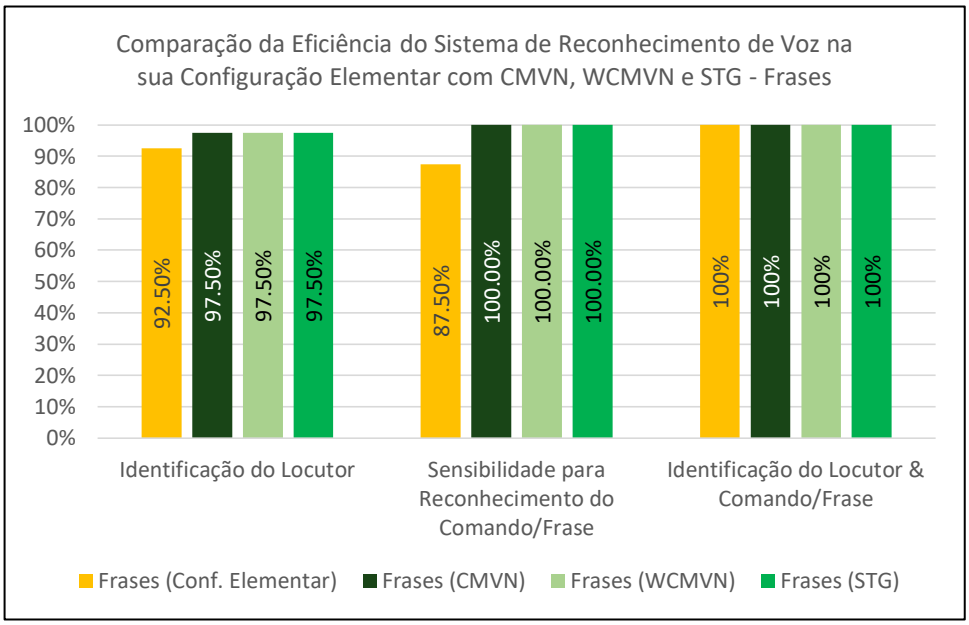

Figura 4.19 - Gráfico de eficiência do sistema para CMVN, WCMVN e STG - frases de testes.

A investigação conduzida neste experimento acerca da melhor técnica de normalização utilizada no sistema concluiu que os melhores resultados foram apresentados usando-se STG. Os resultados obtidos para as técnicas CMVN e WCMVN foram satisfatórios e concordantes com as pesquisas conduzidas por Alam et al. (2011) e Zheng et al. (2006).

4.4.5 Experimento 05 – Ensaio do Sistema Composto dos Melhores Atributos

Os quatro primeiros experimentos realizados nessa pesquisa permitiram identificar se houve aperfeiçoamento do sistema de reconhecimento de voz com o uso de ferramentas como: detecção de voz ativa (VAD); coeficientes dinâmicos (DDC e SDC); e técnicas de normalização dos coeficientes cepstrais (CMVN, WCMVN e STG). Além disso, os resultados apontaram quais dentre as técnicas utilizadas para incrementar propriedades dinâmicas aos coeficientes cepstrais e técnicas de normalização apresentaram melhor rendimento para as tarefas de identificação dos comandos e frases de testes e de reconhecimento do locutor.

Este experimento estuda o sistema de reconhecimento de voz que reúne as melhores técnicas e ferramentas utilizadas nos experimentos anteriores para avaliar se usando essas técnicas agrupadas o sistema apresentará melhor performance que aqueles já estudados nessa pesquisa.

Para esse experimento, o sistema foi configurado utilizando-se: detecção de voz ativa (VAD); a técnica MFCC para extração das propriedades acústicas dos sinais de fala e obtenção dos coeficientes cepstrais estáticos; *Shifted-Delta Coefficients* (SDC) que incorpora as propriedades dinâmicas dos sinais de fala aos vetores acústicos; STG como técnica de normalização dos coeficientes cepstrais estáticos e dinâmicos; e a técnica Quantização Vetorial para classificação e reconhecimento de padrões.

O limiar das distâncias Euclidianas que define o reconhecimetno do locutor foi usado nesse experimento com valor igual a 3,2. Este valor foi encontrado experimentalmente nessa pesquisa após analisar os experimentos 3 e 4 cujos sistemas continham atributos dinâmicos e de normalização respectivamente. Para comparação de resultados, este valor foi mantido em ambos os experimentos, e será novamente usado nesse estudo, em particular.

As Tabelas 4.31 a 4.33 apresentam os resultados do sistema de reconhecimento de voz em estudo obtidos para 2 comandos de testes por indivíduo. Foram analisados os mesmos comandos apresnetados nas tabelas dos Experimentos 1 e 2.

Tabela 4.37 Resultados do Experimento 05 – Comandos de Testes (Ind. 1-2-3)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 1		Ind. 2		Ind. 3		
Locutor (Ind.)	Comandos de Treinamentos ↓	Sinestesia	Vendemiário	Descanso	Multipotencialidade	Poesia	Evolucionismo	Acertos
1	Elo	0,960	1,396	4,069	4,056	4,161	3,821	✓
	Sinestesia	0,837	1,277	4,133	4,208	4,014	4,569	✓
	Terapêutico	1,411	1,873	3,917	4,036	4,194	4,209	✓
	Cosmologia	1,309	1,419	3,945	3,969	3,801	4,118	✓
	Cruzeiro	1,284	1,515	4,215	4,168	4,082	4,148	✓
2	Elo	4,273	3,950	1,401	1,189	3,712	4,023	✓
	Descanso	4,681	3,926	0,824	1,705	3,878	3,953	✓
	Natureza	4,303	4,076	1,149	1,957	3,921	3,841	✓
	Endorfina	3,982	3,995	1,735	1,184	3,915	4,145	✓
	Musicalidade	4,437	4,406	1,282	1,199	4,036	3,994	✓
3	Elo	4,895	3,975	3,637	3,700	1,273	1,251	✓
	Insólito	4,457	4,124	3,844	3,935	1,617	1,122	✓
	Multipotencialidade	4,494	4,040	3,653	3,684	1,390	1,375	✓
	Poesia	3,940	4,056	3,819	3,744	0,862	1,014	✓
	Rudimentar	4,175	4,827	4,070	3,806	1,404	0,848	✓
4	Elo	3,500	3,952	3,941	4,744	3,847	4,097	✓
	Genealogia	3,434	3,688	4,279	4,179	3,922	4,868	✓
	Orgânico	3,689	3,889	4,356	4,210	4,299	4,306	✓
	Evolucionismo	3,719	3,713	4,253	3,985	3,846	4,033	✓
	Herbáceo	4,102	4,074	4,234	4,201	4,263	4,724	✓

Tabela 4.38 Resultados do Experimento 05 – Comandos de Testes (Ind. 4-5-6)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)						
		Ind. 4		Ind. 5		Ind. 6		
Locutor (Ind.)	Comandos de Treinamentos ↓	Alquimia	Elo	Cosmologia	Sinapses	Evolucionismo	Destreza	Acertos
1	Elo	4,086	3,674	4,031	3,692	3,975	4,098	✓
	Sinestesia	3,931	3,739	3,623	3,851	3,839	4,271	✓
	Terapêutico	3,826	3,933	4,025	4,453	4,120	4,169	✓
	Cosmologia	3,695	4,047	3,562	3,928	3,900	4,205	✓
	Cruzeiro	3,825	3,986	3,943	3,858	4,204	3,978	✓
2	Elo	3,942	4,021	4,659	3,980	3,742	4,347	✓
	Descanso	4,202	3,934	4,326	4,256	4,227	4,045	✓
	Natureza	4,154	3,991	4,128	4,137	4,396	3,911	✓
	Endorfina	4,000	4,296	4,056	4,275	4,148	4,297	✓
	Musicalidade	4,616	4,148	4,962	4,235	4,169	4,165	✓
3	Elo	4,014	4,112	4,341	4,027	4,047	4,088	✓
	Insólito	4,286	4,363	4,425	4,259	4,251	4,373	✓
	Multipotencialidade	4,255	4,422	3,962	4,432	3,949	4,145	✓
	Poesia	4,400	3,940	4,303	3,864	4,106	4,128	✓
	Rudimentar	4,267	4,064	4,216	4,867	4,350	4,238	✓
4	Elo	1,413	0,711	4,189	3,983	3,859	3,907	✓
	Genealogia	2,195	1,721	3,650	3,720	3,686	4,261	✓
	Orgânico	1,637	1,090	4,173	3,798	3,925	4,110	✓
	Evolucionismo	1,497	1,269	4,099	3,846	3,367	4,052	✓
	Herbáceo	1,856	1,184	3,716	3,738	4,075	3,958	✓

Tabela 4.39 Resultados do Experimento 05 – Comandos de Testes (Ind. 7-8)

		Comandos de Testes - Distância do sinal de voz às *codewords* (distorção VQ)				
		Ind. 7		Ind. 8		
Locutor (Ind.)	Comandos de Treinamentos ↓	Natureza	Poesia	Cruzeiro	Orgânico	Acertos
1	Elo	4,196	4,230	3,965	4,148	✓
	Sinestesia	4,099	3,959	4,193	3,872	✓
	Terapêutico	4,205	4,381	4,084	3,917	✓
	Cosmologia	4,155	4,121	3,949	3,974	✓
	Cruzeiro	4,197	4,309	3,705	3,902	✓
2	Elo	3,602	3,893	4,053	4,330	✓
	Descanso	3,934	4,118	4,354	3,948	✓
	Natureza	3,507	3,719	4,166	3,883	✓
	Endorfina	3,757	3,771	4,191	4,326	✓
	Musicalidade	3,909	4,107	4,180	4,039	✓
3	Elo	3,957	4,085	4,156	4,181	✓
	Insólito	4,210	3,963	4,114	4,179	✓
	Multipotencialidade	3,995	3,843	4,499	4,502	✓
	Poesia	3,928	3,677	4,780	4,000	✓
	Rudimentar	4,185	4,056	4,083	4,043	✓
4	Elo	4,390	4,184	3,713	3,810	✓
	Genealogia	4,174	4,063	3,789	4,080	✓
	Orgânico	4,422	4,108	3,948	3,798	✓
	Evolucionismo	4,281	4,264	4,140	3,941	✓
	Herbáceo	4,542	4,379	3,848	3,976	✓

A configuração do sistema de reconhecimento automático de voz utilizando os atributos detecção de voz ativa (VAD), *Shifted-Delta Coefficients* (SDC), e STG (*Short-Time Gaussianization*) além das técnicas MFCC e VQ obtiveram resultados superiores a qualquer outra configuração testada neste trabalho.

Os resultados dos comandos de testes apontam para uma melhora com relação à configuração elementar do sistema em 5,18% para identificação assertiva de locutores, em

20,83% com relação à capacidade de reconhecer um comando do banco de dados e em 10% para identificação simultânea do locutor e do comando por ele enunciado.

As Figuras 4.20 e 4.21 ilustram a comparação dos resultados da configuração elementar com o sistema composto dos atributos VAD, SDC e STG para as palavras "Sinestesia", treinada e testada pelo Indivíduo 1 e "Orgânico", treinada pelo Indivíduo 4 e testada pelo Indivíduo 8.

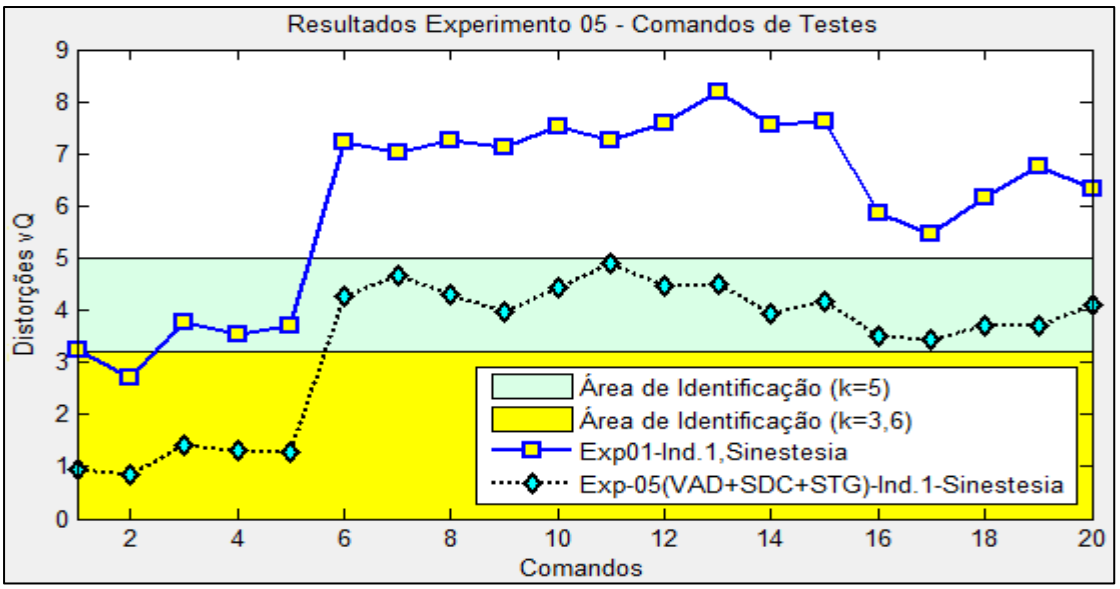

Figura 4.20 - Comparação de resultados do Experimento 01 com o Experimento 05 para a palavra "Sinestesia".

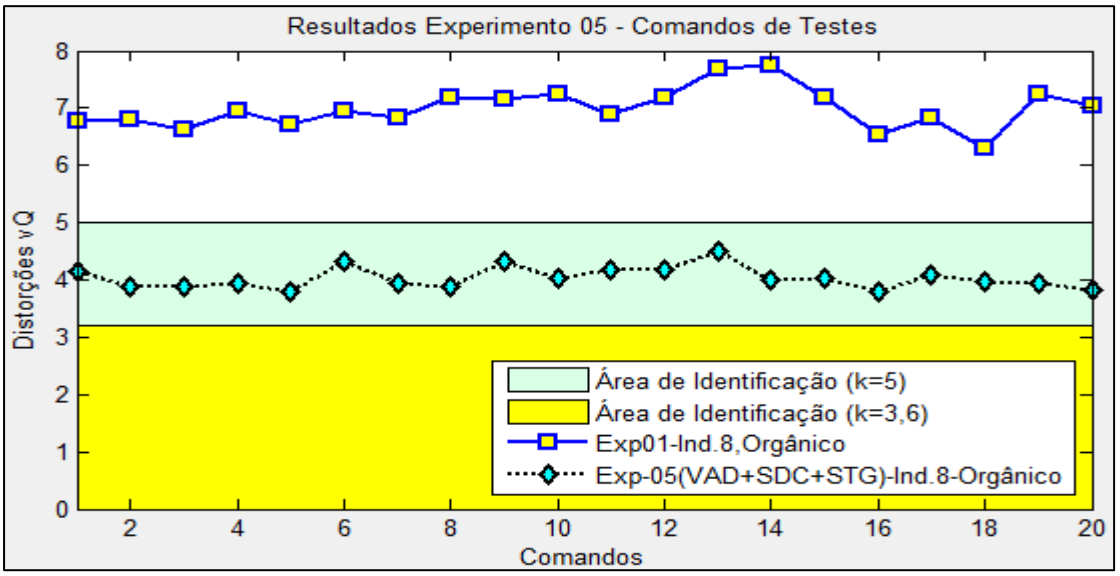

Figura 4.21 - Comparação de resultados do Experimento 01 com o Experimento 05 para a palavra "Orgânico".

Os resultados do sistema de reconhecimento de voz desse experimento para as frases de estão apresentados nas Tabelas 4.34 a 4.36.

Tabela 4.40 Resultados do Experimento 05 – Frases de Testes (Ind. 1-2-3)

		Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)			
		Ind. 1	Ind. 2	Ind. 3	
Locutor (Ind.)	Frases de Treinamento ↓	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	"Felicidade só é real quando compartilhada" (H. David Thoreau)	"Ser feliz sem motivo é mais autêntica forma de felicidade" (Carlos Drummond de Andrade)	Acertos
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	1,280	4,655	4,234	✓
2	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	3,816	1,761	3,799	✓
3	"Felicidade só é real quando compartilhada" (H. David Thoreau)	4,298	4,364	2,092	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	3,513	5,579	4,712	✓

Tabela 4.41 Resultados do Experimento 05 – Frases de Testes (Ind. 4-5-6) (continua)

		Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)			
		Ind. 4	Ind. 5	Ind. 6	
Locutor (Ind.)	Frases de Treinamento ↓	"Liberdade, Igualdade e Fraternidade" (Lema da Revolução Francesa)	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	"Saber como pensar torna a pessoa muito mais capaz do que aquele que apenas sabe o que deve pensar" (Neil deGrasse Tyson)	Acertos
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	3,767	3,629	3,819	✓
2	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	4,563	5,952	5,305	✓

Tabela 4.42 Resultados do Experimento 05 – Frases de Testes (Ind. 4-5-6) (conclusão)

		Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)			
		Ind. 4	Ind. 5	Ind. 6	
Locutor (Ind.)	Frases de Treinamento ↓	"Liberdade, Igualdade e Fraternidade" (Lema da Revolução Francesa)	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	"Saber como pensar torna a pessoa muito mais capaz do que aquele que apenas sabe o que deve pensar" (Neil deGrasse Tyson)	Acertos
3	"Felicidade só é real quando compartilhada" (H. David Thoreau)	4,255	4,969	4,475	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	2,149	4,419	3,948	✓

Tabela 4.43 Resultados do Experimento 05 – Frases de Testes (Ind. 7-8)

		Frases de Testes - Distância do sinal de voz às *codewords* (distorção VQ)		
		Ind. 7	Ind. 8	
Locutor (Ind.)	Frases de Treinamento ↓	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	Acertos
1	Seu sonho é tão palpável quão grande sua vontade de realiza-lo	4,352	4,173	✓
2	"Tudo o que você precisa é de amor" (John Lennon & Paul McCartney)	3,347	4,521	✓
3	"Felicidade só é real quando compartilhada" (H. David Thoreau)	3,860	4,790	✓
4	"Poesia, beleza, romance e amor. Por estas razões continuamos vivendo" (Robin Williams, *Death Poets Society*)	4,210	3,506	✓

Os resultados para as frases de testes analisadas para o sistema de reconhecimento de voz implementado com as técnicas VAD, SDC, STG, MFCC e VQ apresentaram resultados

similares aos analisados para as técnicas de normalização, com 97,5% de índice de assertividade para identificação dos locutores, 100% de reconhecimento das frases treinadas e enunciadas pelos próprios autores durante a fase de testes e 100% de reconhecimento das frases do banco de dados quando essas eram enunciadas por outros indivíduos que não as treinaram.

As Figuras 4.22 e 4.23 apresentam os gráficos comparativos entre a configuração elementar do sistema e a configuração em estudo nesse experimento, para comandos e frases de testes, respectivamente.

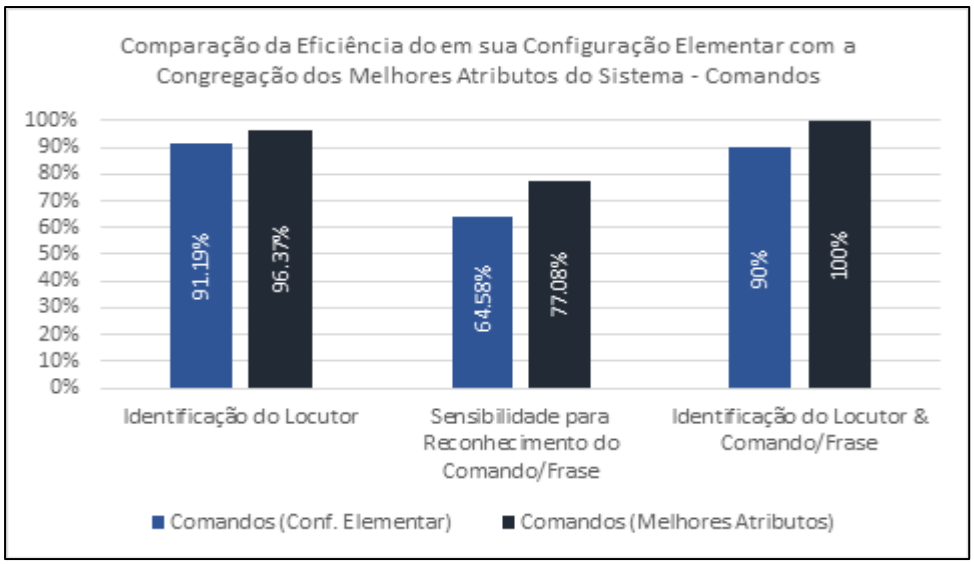

Figura 4.22 - Gráficos de eficiência da configuração robusta do sistema - comandos de testes.

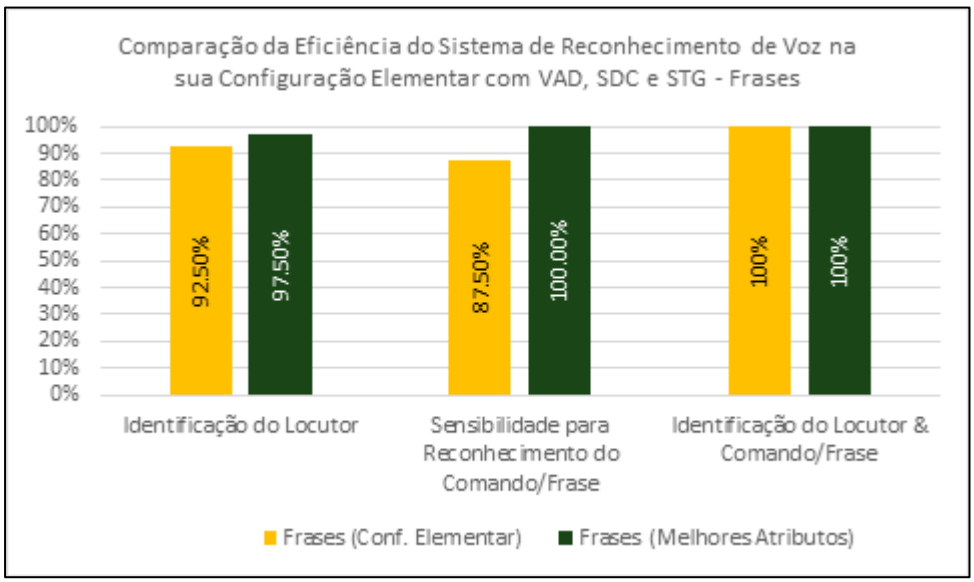

Figura 4.23 - Gráficos de eficiência da configuração robusta do sistema - frases de testes.

5 CONCLUSÕES E TRABALHOS FUTUROS

O desenvolvimento da presente pesquisa abrangeu uma revisão do estado da arte acerca das ferramentas usuais em sistemas de reconhecimento automático de voz, especificamente daquelas que simulam o mecanismo de percepção da fala humana sob o enfoque de sintetizar um programa computacional para captar, digitalizar, processar sinais de fala, extrair suas propriedades acústicas e os classificar em um contexto linguístico, com objetivo de identificar locutores e as suas locuções.

A modelagem da plataforma de reconhecimento de voz mostrou-se útil por sua versatilidade em termos operacionais e de velocidade de processamento, diligenciando os experimentos conduzidos neste estudo.

A escolha das técnicas M*el Frequency Cepstral Coeficcients* (MFCC) e Quantização Vetorial para integrarem o sistema de reconhecimento de voz em estudo se deu pela robustez, eficiência e fácil implementação das mesmas. Além disso, essas técnicas têm alta repercussão no meio científico, sendo vastamente abordadas em pesquisas atuais correlacionadas a este tema e apresentando resultados com relevância significativa para o âmbito de reconhecimento de voz.

Particularmente, MFCC apresentou grande capacidade de extração das propriedades acústicas contidas nos sinais de fala em função da utilização de uma escala que tenta reproduzir o aparelho auditivo humano (escala Mel). Quantização Vetorial mostrou-se uma técnica versátil com capacidade de comprimir dados dos vetores acústicos para criação de bancos de dados, e o emprego do algoritmo LBG proporcionou ótimo desempenho em termos de tempo de processamento do sistema. A adesão de novos vetores acústicos ao banco de dados levou em média cerca de 2 segundos, ao passo que a análise de similitude entre os vetores de teste com aqueles presentes no banco de dados levaram em média, aproximadamente 7 segundos.

Os resultados obtidos referentes à eficiência do sistema desenvolvido, quanto a sua capacidade de identificação do locutor e de suas enunciações, corroboraram para a excelente avaliação das técnicas MFCC e Quantização Vetorial para extração das propriedades acústicas e classificação e reconhecimento de padrões, respectivamente. Estas técnicas operando de maneira isolada, isto é, sem a utilização de atributos de detecção de voz ativa, de normalização ou dinâmicos, obtiveram um índice de reconhecimento do locutor de 91,19% para comandos de testes e de 92,5% para as frases de testes. Ainda nesta configuração, o sistema identificou de

maneira assertiva em 90% das vezes o comando enunciado pelo próprio autor, e em 100%, as frases treinadas e testadas pelos próprios autores.

A incorporação da ferramenta de detecção de voz ativa com intuito de segregar partes vocalizadas do sinal de áudio de partes ressoantes e silenciadas apresentou resultados sutilmente superiores aos obtidos para a configuração elementar do sistema (usando apenas as técnicas MFCC e VQ), ainda que esta técnica seja usual para ambientes ruidosos. Este sistema, que fora confeccionado com as metodologias STE (*Short-Time Energy*) e ZCR (*Zero Crossing Rate*) apresentou uma capacidade de identificação de locutores de 92,37% para comandos de testes e 93,12% para frases de testes. E nesta configuração, os índices de reconhecimento do comando ou frase mantiveram-se os mesmos para a configuração elementar do sistema.

O uso dos *Delta-delta Coefficients* (DDC) e dos *Shifted-Delta Coefficients* (SDC) também aprimoraram o sistema agregando coeficientes dinâmicos aos coeficientes cepstrais estáticos. Houve uma melhora tanto para identificação dos locutores quanto para reconhecimento de suas enunciações. Usando-se DDC, o sistema fora capaz de identificar assertivamente o locutor em 94,56% para os comandos de testes e em 94,37% para as frases de testes. Utilizando-se SDC, os resultados foram ainda mais expressivos, com o índice de identificação de locutores para 95,5% dos casos dos comandos de testes e 95,62% das frases de testes. Para ambos, o índice de reconhecimento tanto do comando quanto das frases de testes enunciadas pelos próprios autores foi de 100%. Associa-se este ganho nos resultados como reflexo das características cepstrais dinâmicas combinadas ao comportamento prosódico conseguido usando-se a técnica SDC.

Da mesma maneira, constatou-se que o uso de ferramentas de normalização incrementou de maneira favorável o sistema de reconhecimento de voz, melhorando seus resultados. A investigação constatou que dentre os três métodos em análise, CMVN, WCMVN e STG, este último apresentou os melhores resultados, apresentando um índice superior de acertos, comparando-se com os outros atributos estudados. Este método apresentou 94,43% de identificação do locutor para os comandos de testes, e 97,50% para as frases de testes. Além disto, o sistema implementado com STG foi capaz de reconhecer, em todos os casos, comandos e frases gravados e enunciados por seus próprios autores. Os ótimos resultados apurados quanto ao uso de STG são atribuídos ao fato desta técnica modificar a distribuição dos vetores acústicos em um curto período de tempo, para seguir uma distribuição referencial ou padrão.

A combinação dos atributos que apresentaram os melhores índices de reconhecimento com as técnicas MFCC e VQ aprimoraram ainda mais o sistema. Tal configuração congregou as ferramentas VAD, SDC e STG à configuração elementar. Os resultados obtidos

demonstraram que o uso de diferentes atributos para detecção de voz ativa, adição de propriedades dinâmicas aos coeficientes cepstrais e normalização de vetores acústicos aumentaram a eficiência do sistema. Em comparação com o uso apenas das técnicas MFCC e VQ, a identificação de locutores aumentou 5,18% para comandos de testes e 5% para frases de testes. E em termos de reconhecimento dos comandos e frases enunciados pelos próprios autores, este resultado aumentou 10% em ambos os casos.

Em uma perspectiva geral, as seguintes conclusões foram auferidas quanto aos resultados obtidos:

1. O sistema apresentou resultados sólidos quanto a capacidade de identificação correta do locutor, para todas as configurações testadas, mostrando-se altamente capaz de realizar esta tarefa;
2. Comandos e frases treinados e enunciados por seus próprios autores apresentaram elevado índice de acertos para a configuração elementar e com detecção de voz ativa, mas apenas os sistemas implementados com atributos dinâmicos, de normalização e para a condição utilizando o conjunto dos melhores atributos, o índice de assertividade foi de 100%;
3. Notou-se que o sistema apresenta maior eficiência de reconhecimento para as frases (locuções de maiores durações) do que para os comandos, ainda que esta diferença seja sutil. Tal fato pode ser explicado em função das sentenças apresentarem maior quantidade de pronunciações, particularizando para o sistema a voz e a locução de cada indivíduo, ainda que, por serem compostas de maiores variabilidades, estas são mais propícias a diferenciações durante os períodos de gravações de treinamentos e testes;
4. A utilização da técnica VAD, ainda que atribuída ao sistema operando em ambientes de ruído controlado, mostrou-se vantajosa, eliminando regiões ruidosas e ressoantes;
5. O uso de atributos dinâmicos conjugados aos coeficientes cepstrais melhorou consideravelmente o sistema. Com o uso destas ferramentas, não só os índices de reconhecimento foram aprimorados, mas também os de reconhecimento das locuções. O modelo SDC sobressaiu-se sobre o modelo DDC;
6. Da mesma maneira, os atributos de normalização usados sobre os vetores acústicos apresentaram ótimos resultados. Conclui-se que o uso de uma janela deslizante sobre os cepstros com a média e variância normalizados apresentou melhores resultados quando comparado com a normalização CMVN. No

entanto, a transformação da distribuição dos vetores acústicos para uma distribuição padrão através da ferramenta STG melhorou ainda mais o sistema, alcançando inclusive, os melhores índices dentre todos os atributos experimentados;

7. Observou-se através de uma grande quantidade de avaliações que se faz necessário a flexibilização do limiar adotado para identificação dos locutores, a depender do tipo de atributo utilizado no sistema. Para os sitemas com a configuração elementar ou aqueles implementados com as técnicas VAD e CMVN, o valor do limiar das distâncias Euclidianas adotado para reconhecimento do locutor foi igual a 5. Para os demais sistemas, utilizou-se o valor para este limiar igual a 3,2, valor encontrado experimentalmente nessa pesquisa;
8. Além da solidez quanto a identificação do locutor, o sistema mostrou-se sensível à altura das vozes daqueles que pronunciavam as frases e comandos, mostrando-se capaz de discernir o gênero do locutor. Os resultados foram robustos para esta característica;
9. Os resultados apurados permitiram a conclusão que em geral, os sistemas de reconhecimento de voz estudados apresentaram sensibilidade para a identificação de comandos e frases do banco de dados, quando estes eram enunciados por indivíduos que apenas testaram o sistema. Tal fato foi constatado, por exemplo, para 77,08% dos casos na condição do conjugado dos melhores atributos em estudo;
10. E, finalmente, observou-se que para as situações em que indivíduos que treinaram o sistema enunciavam comandos ou frases do banco de dados, mas que não de sua autoria, os sistemas apresentaram-se mais sensíveis à identificação do locutor, uma vez que as menores distâncias Euclidianas encontradas nessas situações são referentes aos comandos ou frases encunciados por aquele locutor durante a fase de treinamentos. Muito embora, para uma notada quantidade de amostras com essa ocorrência, quando se desconsiderava as distâncias para os comandos ou frases que o próprio autor treinou, o sistema identificou corretamente o comando ou frase que fora testado.

5.1 Trabalhos Futuros

Sugere-se para pesquisas futuras e complementares à esta desenvolvida, a experimentação do sistema robusto construído em ambientes ruidosos, determinando a efetiva contribuição da técnica VAD, examinando assim, se esta possa ter um efeito superior aos demais atributos. Além disto, recomenda-se o uso de maiores corpora, e de uma maior quantidade de indivíduos voluntários para a realização das fases de treinamento e de teste, verificando assim, se com um maior número de locuções e vetores acústicos ocupando o espaço vetorial gerado para cada experimento, o sistema apresentará os mesmos índices de reconhecimento. É interessante ainda, o estudo da comparação entre diferentes técnicas de extração das propriedades acústicas, como LPC e PLP com MFCC, assim como de GMM e HMM com Quantização Vetorial.

6 REFERÊNCIAS

AKILA, A.; CHANDRA, E. Comparative Study of Endpoint Detection Algorithms SuiTabela for Isolated Word Recognition. In: **BIJIT-BVICAM International Journal of Information Technology**. New Delhi, India. September 2014.

ALAM, M. J.; OUELLET, P.; KENNY, P.; O'SHAUGHNESSY, D. Comparative Evaluation of Feature Normalization Techniques for Speaker Verification. In: **Proceedings 5th International Conference on Nonlinear Speech Processing, NOLISP**. Las Palmas de Gran Canaria, Spain. November 2011. ISSN: 0302-9743.

ALLEN, F.; AMBIKAIRAJAH, E.; EPPS, J. Language Identification Using Warping and Shifted Delta Cepstrum. In: **IEEE 7th Workshop on Multimedia Signal Processing**. November, 2005.

ANUSUYA, M. A.; KATTI, S. K. Speech Recognition by Machine: A Review. In: **(IJCSIS) International Journal of Computer Science and Information Security**, Vol. 6, No. 3, 2009.

ARYA, S.; MOUNT, D. M. Algorithms for Fast Vector Quantization. In: **Proc. Data Compression Conference**, J. A. Storer and M. Cohn, eds. Snowbird, Utah, 1993, IEEE Computer Society Press.

ATKINSON, Q. D. Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa. In: **Science Magazine 332 (6027)**. March 2011.

BHARTI, R.; BANSAL, P. Real Time Speaker Recognition System using MFCC and Vector Quantization Technique. In: **International Journal of Computer Applications (0975 – 8887).** Vol. 117 – No.1. May 2015.

CHOU, W.; JUANG, B. H. Minimum Classification Error (MCE) Approach in Pattern Recognition. In: **Pattern Recognition in Speech and Language Processing**. The Electrical Engineering and Applied Signal Processing Series. 2003. Editado por Alexander Poularikas. ISBN 0-8493-1232-9.

DAVE, N. Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition. In: **International Journal for Advance Research in Engineering and Technology**. Vol. 1. July 2013.

DHINGRA, S. D.; NIJHAWAN, G.; PANDIT, P. Isolated Speech Recognition Using MFCC and DTW. In: **International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering**. Vol.2, Issue 8, 2013.

DRAKE, N. Human Evolution 101. **National Geographic**. Published September 11, 2015. Disponível em: <http://news.nationalgeographic.com/2015/09/human-evolution-101/>. Acesso em 18 de novembro. 2015.

DUHAMEL, P.; VETTERLI, M. Fast Fourier Transforms: A Tutorial Review and State of the Art. In: **The Digital Signal Processing Handbook, Digital Signal Processing**

Fundamentals. 2ª Ed. Boca Raton, United States of America, 2010. Taylor & Francis Group. Editor-in-chief Vijay K. Madisetti. ISBN 978-1-4200-4606-9.

ENQING, D.; GUIZHONG, L; YATONG, Z.; YU, C. *Voice Activity Detection Based on Short-Time Energy and Noise Spectrum Adaptation*. In: **Proceedings**

GÂTA, M.; TODEREAN, G. System of Speaker Identification Independent of Text for Romanian Language based on Gaussian Mixture Models extract from the MFCC Vectors. In: **International Conference on Computer Systems and Technologies – CompSysTech'06.** 2006.

GERSHO, A.; GRAY, R. M. **Vector Quantization and Signal Compression**. Springer Science & Business Media. New York, 1992. ISBN 978-1-4613-6612-6.

GILL, M. K.; KAUR, R.; KAUR, J. Vector Quantization based Speaker Identification. In: **International Journal of Computer Applications (0975 – 8887)**. Vol. 4 – No.2. July 2010.

GOLD, B.; MORGAN, N.; ELLIS, D. **Speech and Audio Signal Processing: Processing and Perception of Speech and Music.** 2ª Ed. John Wiley & Sons. New Jersey, 2011. ISBN 978-0-470-19536-9.

GOPI, E. S. **Digital Speech Processing Using Matlab**. Springer India, India, 2014. ISBN 978-81-322-1676-6.

GUPTA, S.; JAAFAR, J.; AHMAD, W. F.; BANSAL, *A.* Feature Extraction Using MFCC. In: **Signal & Image Processing: An International Journal (SIPIJ)**.Vol.4, NO.4, August 2013.

HUANG, X.; ACERO, A.; HON, H.-W. Spoken Language Processing*:* A Guide to Theory, Algorithm, and System Development. **Prentice Hall PTR**, New Jersey, 2001.

JAIN, A.; SHARMA, O. P. A Vector Quantization Approach for Voice Recognition Using Mel Frequency Cepstral Coefficient (MFCC): A Review. In: **International Journal of Electronics & Communication Technology (IJECT)** Vol. 4, Issue Spl - 4, April - June 2013. ISSN: 2230-7109.

JENKINS, W. K. Fourier Methods for Signal Analysis and Processing. In: **The Digital Signal Processing Handbook, Digital Signal Processing Fundamentals**. 2ª Ed. Boca Raton, United States of America, 2010. Taylor & Francis Group. Editor-in-chief Vijay K. Madisetti. ISBN 978-1-4200-4606-9.

JUANG, B. H.; RABINER, L.R. Hidden Markov Models for Speech Recognition. In: **Technometrics**. Vol. 33, No. 3. August. 1991.

KABIR, A.; AHSAN, S. M. M. Vector Quantization in Text Dependent Automatic Speaker Recognition Using Mel-frequency Cepstrum Coefficient. In: **6th WSEAS International Conference on Circuits, Systems, Electronics, Control & Signal Processing**. Cairo, Egypt, Dec 29-31, 2007.

KUMAR, K.; KIM, C.; STERN, R. M. Delta-spectral Cepstral Coefficients for Robust Speech Recognition. In: **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. Prague, Czech Republic. 2011.

LANDELL, B. P; NAYLOR, J. A.; WOHLFORD, R. P. Effect of Vector Quantization on a Continuous Speech Recognition System. In: **ITT Defense Communications Division, San Diego, CA**. March 1984. DOI: 10.1109/ICASSP.1984.1172586.

LINDE, Y.; BUZO, A.; GRAY, R. Algorithm for Vector Quantizer *Design*. In: **Transactions on Communication IEEE**, Vol. COM-28, NO. 1. January, 1980. ISSN: 0090-6778.

LING, J.; SUN, S.; ZHU, J.; LIU, X. Speaker Recognition with VAD. In: **Second Pacific-Asia Web Mining and Web-based Application, WMWA'09**. Junho 2009.

LOGAN, B. Mel Frequency Cepstral Coefficients for Music Modeling. In**: International Symposium of Music Information Retrieval ISMIR**, 2000.

LOKHANDE, N. N.; NEHE, N. S.; VIKHE, P. S. Voice Activity Detection Algorithm for Speech Recognition Applications. In: **International Conference in Computational Intelligence (ICCIA), Proceedings published in International journal of Computer Applications (IJCA)**. Março 2012.

MAFRA, A. T. **Reconhecimento Automático do Locutor em Modo Independente de Texto por Self-organizing Maps**. 2002. Dissertação (Mestrado em Engenharia Mecatrônica e de Sistemas Mecânicos) Escola Politécnica da Universidade de São Paulo, Departamento de Engenharia Mecatrônica e de Sistemas Mecânicos.

MAKHOUL, J.; ROUCOS, S.; GISH, H. *Vector* Quantization in Speech Coding. In: **Proceedings of the IEEE**, Vol. 73, NO 11. November 1985.

MCLOUGHLIN, I. Applied Speech and Audio Processing With Matlab® Examples. **Cambridge University Press**. New York, 2009. ISBN-13 978-0-521-51954-0.

MOATTAR, M. H.; HOMAYOUNPOUR, M. M. A Simple but Efficient Real-Time Activity Detection Algorithm. In: **17th European Signal Processing Conference**. Agosto 2009.

MOHAN, B. J.; BABU, R. Speech Recognition Using MFCC and DTW. In: **International Conference on Advances in Electrical Engineering (ICAEE)**, 2014.

MOLAU, S.; PITZ, M.; SCHLÜTER, R.; NEW, H. Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum. In: **IEEE International Conference on Acoustics Speech and Signal Processing**. Germany, 2001.

MOLLA, M. K. I.; HIROSE, K. On the Effectiveness of MFCC's and Their Statistical Distribution Properties in Speaker Identification. In: **VECIMS 2004 – IEEE International Conference on Virtual Environments, Human-Computer Interfaces, and Measurement Systems.** Boston, MA, USA. July 2004.

MUDA, L.; BEGAM, M.; ELAMVAZUTHI, I. Voice Recognition Algorithms Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. In: **Journal of Computing**, Volume 2, Issue 3, March 2010, ISSN: 2151-9617.

NGOC, V. V.; WHITTINGTON, J.; DEVLIN, J. Real-time Hardware Feature Extraction with Embedded Signal Enhancement for Automatic Speech Recognition. In: **Speech Technologies**. 2011. Edited by Ivo Ipšić. ISBN 978-953-307-996-7.

NIJHAWAN, G.; SONI, M. K. Speaker Recognition Using MFCC and Vector Quantization. In: **Int. J. on Recent Trends in Engineering and Technology**. Vol. 11 – No. 1. July 2014.

O'SHAUGHNESSY, D. Speech Communication: Human and Machine. Addison-Wesley, 2ª ed. 2000, New York. ISBN 987-0-201-16520-3.

OPPENHEIM, A. V.; SCHAFER, R. W. **Discrete-time Signal Processing**. 2ª Ed. Prentice Hall, Inc. New Jersey, 1999. INSB 0-13-754920-2.

PATEL, K.; PRASAD, R. K. Speech Recognition and Verification Using MFCC & VQ. In: **International Journal of Emerging Science and Engineering (IJESE)**. Volume-1, Issue-7, May, 2013. ISSN:2319-6378,

PATRA, S. **Robust Speaker Identification System**. 2007. Dissertação (Master of Science in Engineering). Faculty of Engineering, Super Computer Education and Research Centre, Indian Institute of Science. Bangalore, India.

PICONE, W. J. Signal Modeling Techniques in Speech Recognition. In: **Proceedings of the IEEE**. Vol. 81, No. 9. September 1993.

PRASAD, N. V.; UMESH, S. Improved Cepstral Mean and Variance Normalization Using Bayesian Framework. In: **IEEE 2013 Workshop on Automatic Speech Recognition and Understanding (ASRU)**. Dezembro 2013.

PROAKIS, J. G.; MANOLAKIS, D. G. **Digital Signal Processing, Principles, Algorithms, and Applications**. 3ª Ed. Prentice Hall, Inc. New Jersey, 1996. ISBN 0-13-394338-9.

QIANG, H.; YOUWEI, Z. On Prefiltering and Endpoint Detection of Speech Signal. In: **Proceedings of ICSP**. 1998.

RABINER, L.; JUANG, B.-H. Fundamentals of Speech Recognition. **Prentice Hall PTR**, New Jersey, 1993. ISBN 0-13-285826-6.

RABINER, L.; SCHAFER, R. W. **Digital Processing of Speech Signals**. Prentice Hall, Inc., Signal Processing Series. Bell Laboratories Incorporated. New Jersey 1978. ISBN 0-13-213601-1.

RABINER, L.; SCHAFER, R. W. **Introduction to Digital Speech Processing**. Foundations and Trends® in Signal Processing, vol. 1, no 1–2. Boston, 2007. ISBN: 978-1-60198-070-0.

RAJSEKHAR, A. **Real Time Speaker Recognition Using MFCC and VQ**. 2008. Dissertação (Master of Technology in Telematics and Signal Processing). Department of Electronics & Communication Engineering, National Institute of Technology. Rourkela, India.

RAMACHANDRAN, R. P. Quantization of Discrete Time Signals. In: **The Digital Signal Processing Handbook, Digital Signal Processing Fundamentals**. 2ª Ed. Boca Raton, United States of America, 2010. Taylor & Francis Group. Editor-in-chief Vijay K. Madisetti. ISBN 978-1-4200-4606-9.

RAMIREZ, J.; GORRIZ, J. M.; SEGURA, J. C. Voice Activity Detection. Fundamentals and Speech Recognition Systems Robustness. In: **Robust Speech Recognition and Understanding**. Vienna, Austria, 2007. Edited by Michael Grimm and Kristian Kroschell.

RONG, T. **Automatic Speaker and Language Identification**. 2006. Tese de Doutorado (Doctor of Philosophy Thesis in Computer Engineering). Nanyang Technological University. Western Water Catchment, Singapore.

ROSENGREN, K. E. **Communication: An Introduction**. Sage Publications Ltd. 2000. ISBN-10-8039-7836-7.

RUSSEL, S. J.; NORVIG, P. **Artificial Intelligence, a Modern Approach**. Prentice-Hal, Inc. A Simon & Schuster Company. New Jersey, 1995. ISBN D-IH-IQBSOS-E.

SINGER, E.; TORRES-CARRASQUILLO, P. A.; GLEASON, T. P.; CAMPBELL, W. M.; REYNOLDS, D. A. Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification. In: **Proc. Eurospeech'03**. Geneva, 2003.

SONG, F. K.; ROSENBERG, A. E.; JUANG, B. H. A Vector Quantization Approach to Speaker Recognition. In: **AT&T Technical Journal**. Vol. 66-2, Março, 1987.

STEVENS, S. S.; VOLKMAN, A. D. J.; NEWMAN, E. B. A Scale for the Measurement of the Psychological Magnitude Pitch. In: **Journal of the Acoustical Society of America**. Vol. 8. January 1937. DOI: 10.1121/1.1915893.

TEVAH, R. T. **Implementação de um Sistema de Reconhecimento de Fala Contínua com Amplo Vocabulário para o Português Brasileiro**. 2006. Dissertação (Mestrado em Engenharia Elétrica). Universidade Federal do Rio de Janeiro, Brasil.

TEXAS INSTRUMENTS INC. **Understanding Data Converters**. [S1], 1995 (Application Report, SLAA013).

TIWARI, V. MFCC and its Applications in Speaker Recognition. In: **International Journal on Emerging Technologies**. 2010. ISSN: 0975-8364.

TORRES-CARRASQUILLO, P. A.; SINGER, E.; KOHLER, M. A.; GREENE, R. J.; REYNOLDS, D. A.; DELLER, J. R. Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features. In: **Proc. ICSLP**. 2002.

VIRTANEN, T.; SINGH, R.; RAJ, B. **Techniques for Noise Robustness in Automatic Speech Recognition**. John Wiley & Sons, Ltd. Chichester, England, 2013. ISBN: 978-0-470-97409-4.

XIANG, B.; CHAUDHARI, U. V.; NAVRATIL, J.; RAMASWAMY, G. N.; GOPINATH, R. A. Short-time Gaussianization for Robust Speaker Verification. In: **IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP)**. Vol. 1. Orlando, Florida, USA. Maio 2002.

YOUNG, S.; EVERMAN, G.; GALES, M.; HAIN, T.; KERSHAW, D.; LIU, X.; MOORE, G.; ODELL, J.; OLLASON, D.; POVEY, D.; VALTCHEV, V.; WOODLAND, P. **The HTK Book Version** *3.4*. Cambridge University. Cambridge, 2006.

YU, D.; DENG, L. **Automatic Speech Recognition, A Deep Learning Approach**. Springer-Verlag. London, 2015.ISBN 978-1-4471-5778-6.

ZHENG, R.; ZHANG, S.; XU, B. A Comparative Study of Feature and Score Normalization for Speaker Verification. In: **International Conference, ICB 2006, Proceedings**. Hong Kong, China. Janeiro 2006.