UNIVERSIDADE FEDERAL DE SANTA CATARINA

Programa de Pós-Graduação em Ciências da Computação

RECONHECIMENTO DE VOZ:

UMA ABORDAGEM UTILIZANDO LÓGICA DIFUSA

Dissertação submetida à Universidade Federal de Santa Catarina para a obtenção do grau de Mestre em Ciências da Computação

Norberto de Castro Peil

Florianópolis, Fevereiro de 1998

RECONHECIMENTO DE VOZ:
UMA ABORDAGEM UTILIZANDO LÓGICA DIFUSA

NORBERTO DE CASTRO PEIL

ESTA DISSERTAÇÃO FOI JULGADA ADEQUADA PARA OBTENÇÃO DO TÍTULO DE

MESTRE EM CIÊNCIAS DA COMPUTAÇÃO

ESPECIALIDADE EM SISTEMAS DE CONHECIMENTO E APROVADA EM SUA FORMA FINAL PELO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO.

Prof. Jorge Muniz Barreto
Coordenador do Curso

BANCA EXAMINADORA:

Prof. Édis Mafra Lapolli, Dra.
Orientadora

Prof. Lia Caetano Bastos, Dra.

Prof. Ana Maria Franzoni, M.Sc.

Com gratidão e amor,
à minha mãe Diana,
ao meu filho Andrei e
à minha esposa Luciana.

AGRADECIMENTOS

Agradeço em primeiro lugar a Deus que me deu bem mais do que pedi.

Ao meu pai Norberto, pela vida e pela confiança que sempre demonstrou.

Especialmente à minha mãe Diana, pelo amor sem medida, pela educação e por ter feito tudo quanto possível para que este momento chegasse.

À minha esposa Luciana, pelo incentivo, confiança e paciência em muitos momentos.

Ao meu filho Andrei, que nasceu durante esta caminhada, pela alegria que presenteou-me tantas vezes e, que em alguns momentos teve que ouvir um triste "agora não posso".

À minha orientadora Édis, por aceitar-me no meio do caminho, nunca negando atenção e apoio, dignificando o significado da palavra professor.

À secretária Vera Lúcia Sodré Teixeira, pela atenção e simpatia que sempre demonstrou.

Postumamente ao meu primeiro orientador Hermann Adolf Harry Lucke, por sua dedicação e contribuição para o desenvolvimento tecnológico de um país que não era o seu, mas que sem dúvida amava.

Aos demais familiares, irmãos e amigos , pela amizade e apoio que nunca faltaram.

RESUMO

O desenvolvimento de interfaces naturais tem ocupado uma fatia importante dos projetos de pesquisa nos centros tecnológicos pelo mundo inteiro. Esta busca por meios mais naturais de comunicação homem-máquina tem sido possível graças ao desenvolvimento tecnológico do hardware e a utilização de novas técnicas de inteligência artificial. A fala, sem dúvida, é uma forma natural para o ser humano comunicar-se, e os computadores do futuro deverão ser capazes de entende-la e inclusive manter um diálogo com o locutor. Este trabalho faz uma análise dos conceitos que envolvem um Sistema de Reconhecimento de Voz (SRV) e suas potencialidades. A abordagem utilizada na identificação das palavras faladas é a lógica difusa, que também é conceituada e observada na prática. Um SRV capaz de comandar uma calculadora por voz é desenvolvido, alcançando uma taxa de reconhecimento de 89% em ambiente favorável. O sistema também tem uma função didática permitindo alterações nos formatos dos conjuntos difusos e em outros parâmetros que poderão aumentar ou diminuir a taxa de reconhecimento.

ABSTRACT

The development of the natural interfaces has got a great important part of the research projects in the technological centers all over the world. This search for more natural ways of man-machine communication has been possible thanks to the technological development of the hardware and the utilization of new technics of artificial intelligence. The speech, no doubt, is a natural form of human being to communication, and computers of the future will be able to understand it and inclusive keep a dialog with the speaker. This work analyses the concepts that belongs to the voice recognition system (VRS) and yours potentialities. The approach utilized in the indentification of speak words is the fuzzy logical. An VRS able to command a calculation by voice is developed, reaching 89% of recognition in favorable ambient. The system also have an didatic character which is permitting alterations in the formats of the fuzzy sets and others parameters that increase or decrease the recognition tax.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

CAPÍTULO I

INTRODUÇÃO

1.1 Origem do Trabalho

O mundo que nos cerca está repleto de sons, ou seja, ondas sonoras produzidas em uma ampla faixa de freqüências, por nós ou por qualquer outra fonte. Alguns destes sons não conseguimos escutar por estarem fora da faixa de audição do ser humano.

Quando falamos de som em computadores, certamente imaginamos um usuário diante dele apenas escutando, numa comunicação sonora unidirecional. Esta realidade aos poucos vai se modificando pois a necessidade por interfaces sonoras de mão dupla, ou a simples possibilidade de uma alternativa ao teclado e ao mouse , tem sido atendida por alguns fabricantes de software. Claro que estamos ainda longe daquele sonho: "Abra as portas, HAL" da década de 60, porém os passos estão sendo dados. Os sistemas atuais nos permitem vislumbrar pessoas com microfones presos a cabeça, em salas ou escritórios com baixo nível de ruído.

Os sistemas de inteligência artificial tem procurado reproduzir através de software e com uso de hardware, o intrincado funcionamento do cérebro humano. Os desafios são grandes, a pesquisa não para e o futuro é dependente cada vez mais do hardware.

Com a idéia básica de propor uma associação do reconhecimento de voz com a lógica difusa, de forma a aumentar a eficiência do reconhecimento de voz, teve origem este trabalho.

1.2 Objetivos

Associar lógica difusa com reconhecimento de voz é unir duas linhas de pesquisa. Enquanto a lógica difusa tem uma bem desenvolvida teoria baseada em conceitos e definições, o reconhecimento de voz sofre com a complexidade do som da fala, portanto entre os objetivos principais deste trabalho estão:

- Apresentar os conceitos fundamentais envolvidos com o som, o reconhecimento de voz e a lógica difusa.

- Relatar as técnicas mais utilizadas em reconhecimento de voz e resultados alcançados, tendo em conta que diversas são as abordagens e grande o número de variáveis envolvidas no "reconhecer uma palavra falada".
- Utilizar os conceitos da teoria de conjuntos difusos em um sistema de reconhecimento de voz, dependente do locutor e com pronúncia não contínua, ou seja, palavras pronunciadas espaçadamente.

- Permitir ao usuário alterar parâmetros que possam melhorar o percentual de reconhecimento de voz.

- Criar uma aplicação em ambiente DOS que faça uso do sistema.

1.3 Importância do trabalho

A lógica difusa esquecida por algum tempo, e redescoberta nos últimos anos, tem sido cada vez mais utilizada na produção intelectual. Suas aplicações aumentam enormemente a cada dia em todo o mundo. A associação com sistemas de reconhecimento de voz não é substancial. Publicações científicas mais detalhadas sobre implementações de sistemas de reconhecimento de voz são raras. Com utilização de lógica difusa mais raras ainda, o que torna o desafio bem maior.

A possibilidade, em futuro bem próximo, do uso de interfaces homem-máquina cada vez mais naturais, como alternativas ao teclado e ao mouse, será imprescindível para os softwares que quiserem competir no mercado mundial. A utilização da fala permite a liberação das mãos e dos olhos, da posição de sentado em uma cadeira, e até mesmo das preocupações com lesões de esforço repetitivo que tem atacado profissionais da informática em todo o mundo.

Baseado no que foi exposto, este trabalho oferece uma aplicação prática da lógica difusa relacionada a uma área em que a pesquisa está em franca expansão, e cujos conceitos precisam ser dominados por aqueles que tem interesse em acompanhar a evolução tecnológica nesta área. Esta evolução leva sem dúvida ao diálogo natural homem-máquina, onde o reconhecimento e a síntese da voz estarão perfeitamente integradas ao sistema computacional.

1.4 Estrutura do Trabalho

Este trabalho é composto de 6 capítulos que procuram apresentar os tópicos principais envolvidos na elaboração de um sistema de reconhecimento de voz e no uso da lógica difusa.

O capítulo 1 - Introdução - fornece uma visão geral do quadro onde se insere o reconhecimento de voz e a lógica difusa destacando sua importância e objetivos do trabalho.

O capítulo 2 - Som - destaca os conceitos que descrevem o som, tanto o som natural quanto o som digital.

O capítulo 3 - Sistemas de Reconhecimento de Voz - trata inicialmente das categorias e classificações dos SRV, mostrando sua importância e discutindo possibilidades. A seguir é feita uma análise dos sistemas de reconhecimento de voz, trazendo um histórico das pesquisas desenvolvidas, abordagens realizadas, bem como resultados alcançados.

No capítulo 4 - Lógica Difusa - é apresentado um histórico e os conceitos fundamentais que envolvem a teoria de conjuntos difusos e a lógica difusa. Sua utilização e aplicação na produção intelectual e na indústria também estão presentes.

O capítulo 5 - Lógica Difusa em Reconhecimento de Voz - apresenta o processo de desenvolvimento de um sistema de reconhecimento de voz, a utilização da lógica difusa e sua implementação em um sistema real com análise de resultados.

O capítulo 6 - Conclusão - apresenta as conclusões do trabalho, bem como traz sugestões para futuros trabalhos.

Finalmente, uma extensa bibliografia sobre o assunto é apresentada.

CAPÍTULO II

SOM

2.1 Introdução

Como escreve Ridge [RID94] "O som é uma vibração que se propaga através do ar, por cortesia das moléculas de ar, que passam a vibração para a frente, até nossos ouvidos." Situação parecida com aquela experimentada quando se joga uma pedra num lago: a pedra ao chocar-se com a água forma uma série de ondas concêntricas, que se propagam em todas as direções até que a amplitude (neste caso, comprimento) das ondas seja tão pequena que elas não possam mais ser vistas.

2.2 Formação do Som

O ser humano possui dentro da cabeça uma estrutura física que lhe permite escutar os sons que o rodeiam. A figura 2.1 ilustra este sistema:

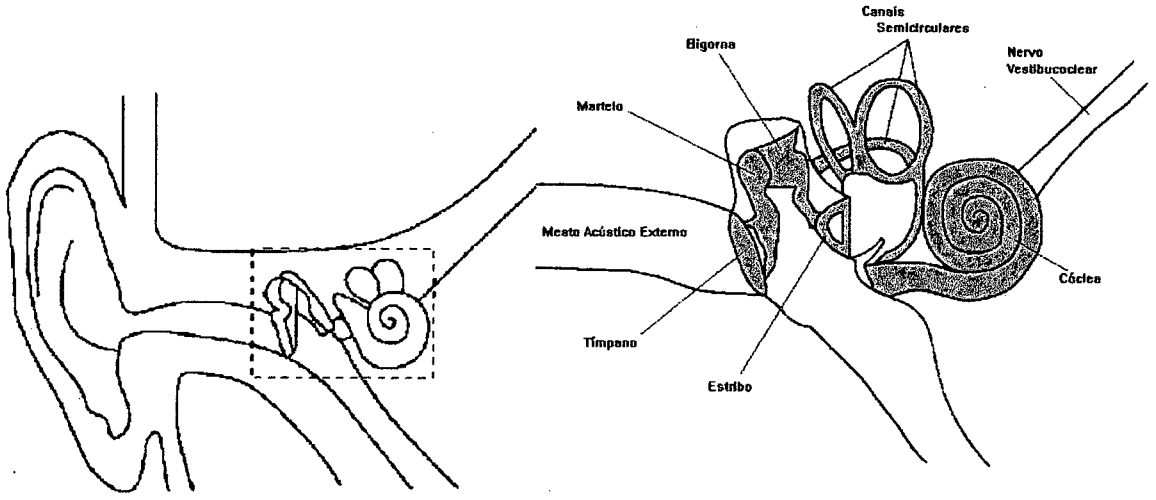

Figura 2.1 - O ouvido humano.
Fonte: [MOO94]

De acordo com [MOO94], para o ouvido o som existe na forma de rápidas variações na pressão do ar. Essas variações entram no canal auditivo em ondas. Conforme [DAN95], a pressão do ar empurra e puxa o tímpano, que move o martelo, que move a bigorna, que move o estribo - até que a onda seja finalmente transferida para a cóclea. Na cóclea estas ondas mecânicas são transformadas em ondas líquidas, que estimulam os receptores do nervo vestibulococlear, permitindo que o cérebro perceba o som. A intensidade do som vai fazer com que as ondas se propaguem com mais ou menos intensidade.

Isto nos dá idéia de como escutar é complexo, e do que é capaz esse pedaço de hardware orgânico que possuímos entre nossas orelhas.

2.3 Descrição de um Som

As variações de pressão do ar que chegam aos nossos ouvidos são originárias de um número incalculável de fontes sonoras, seja na forma de música, voz ou simplesmente ruído. As variações de pressão do ar que compõe o som tem características importantes. Na figura 2.2, onde uma onda senoidal está representando um som puro, podemos percebe-las.

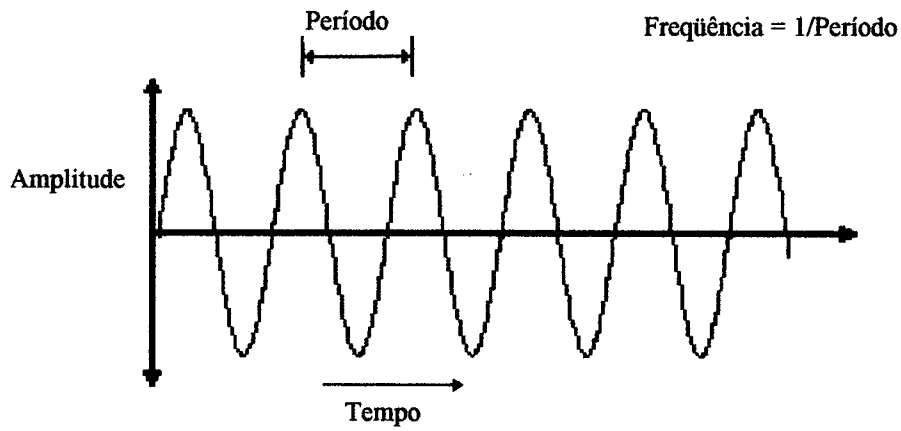

Figura 2.2 - Formato de uma onda senoidal produzida por um oscilador eletrônico. Fonte:[MOO94]

2.3.1 Freqüência

A freqüência do som é a quantidade de vezes com que ele vibra por segundo. A unidade de medida da Freqüência é o hertz (Hz). Um som que vibra uma vez por segundo mede 1 Hz. As freqüências são descritas normalmente em quilohertz (KHz), a unidade que representa 1.000 Hz. O ser humano saudável pode perceber sons na faixa de aproximadamente 20 a 20.000 Hz (20 KHz).

2.3.2 Amplitude

A amplitude de uma onda representa sua intensidade de energia. Como diz Ridge em [RID94], um balanço mostra bem a importância da amplitude. À medida que alguém se embala mais, maior é a amplitude do balanço. Observa-se ainda que para subir só um pouco, muita energia é exigida.

A unidade de medida da intensidade do som é o decibel, abreviado como dB. A sensibilidade do ouvido humano é extraordinária, podendo perceber sons de intensidade muito alta e logo em seguida perceber um som muito baixo.

Como o ouvido precisa de grandes variações na intensidade do som antes que seja sentida uma alteração, conclui-se que a sensibilidade do ouvido à intensidade do som é logarítmica. O ouvido funciona como um dispositivo logarítmico; portanto o decibel, uma unidade de medida logarítmica, é uma escolha apropriada para medir a intensidade do som. Portanto, um som de 46dB tem duas vezes a intensidade do som a 43dB e quatro vezes a intensidade do som a 40dB. É preciso um aumento de 10 dB para fazer com que o som pareça duas vezes mais agudo aos nossos ouvidos.

Dadas essas características, pode-se traçar um gráfico que represente o som.

Figura 2.3 Forma de onda do som da palavra "divide"
Fonte: software "Wave Studio"

O que se vê na Figura 2.3 é a representação gráfica de como um som de voz se parece, no caso, o som da palavra "divide" O tempo corre da esquerda para a direita. A amplitude, ou volume do som, é indicada pela massa escura que se move para cima e para baixo da linha horizontal central. Quanto maior a distância da linha horizontal central, mais intenso é o som.

Ampliando-se a forma de onda (figura 2.4) percebe-se melhor os detalhes do formato do som e as ondas individuais. A freqüência ou tom (pitch) do som, como já foi dito, é medida pelo número de ciclos por segundo.

A altura dessas ondas, a distância entre os máximos de cada onda e o formato das ondas é que fazem o som único.

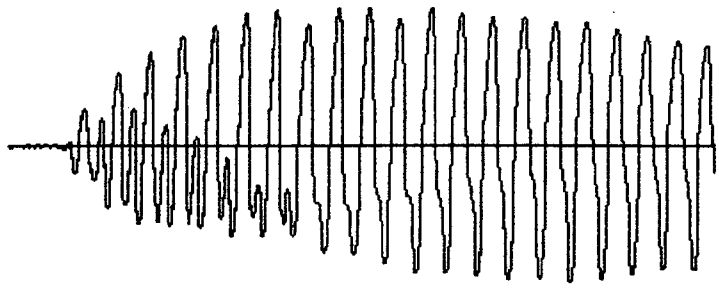

Figura 2.4 Forma de onda ampliada.
Fonte:[MOO94]

2.3.3 Envoltória

A Figura 2.5 ilustra um outro conceito importante para descrever o som, que é a envoltória do som. A forma da envoltória é uma das maneiras de se descrever um som em função do tempo, da amplitude e da freqüência.

Figura 2.5 A envoltória da palavra "divide".

2.3.4 Timbre

Diferentemente de outras formas de onda, a forma da onda senoidal é tal que apenas uma única nota pode ser ouvida, sem outras notas mais graves ou agudas. Um diapasão pode produzir uma onda senoidal, assim como um oscilador eletrônico. Entretanto, a maioria dos sons orgânicos é muito mais complicada do que uma onda senoidal. Essas complicações é que são responsáveis pelo timbre do som.

Para [MOO94], o timbre é a característica da qualidade tonal de um som. O timbre é responsável pela diferença entre o som de um violino e o de um saxofone. O timbre é composto por diversos elementos, incluindo o formato da envoltória e a complexidade das freqüências dentro da envoltória.

2.3.5 Freqüência Fundamental e Harmônica

O termo harmônica descreve o relacionamento entre ondas, onde uma onda tem Freqüência que é múltipla da assim chamada Freqüência Fundamental, a onda dominante (Freqüência mais forte). A segunda onda harmônica tem o dobro da Freqüência da onda fundamental, a terceira o triplo e assim por diante.

Por exemplo, se a freqüência fundamental é 440 Hz a segunda harmônica é 880 Hz, a terceira harmônica é 1.320 Hz e assim sucessivamente.

2.4 Digitalização do Som

Segundo [SPA94], o ouvido e um microfone conectado à uma placa de som qualquer possuem a mesma função: converter pequenas variações na pressão do ar em um sinal elétrico que pode ser entendido e armazenado pelo cérebro humano ou pela CPU no computador. Para [MOO94], sinal é o termo usado para designar informações como o som, quando ele está sendo transformado de sua forma original - moléculas de ar colidindo uma com as outras - em uma versão elétrica que pode ser salva, manipulada e tocada.

O processo para digitalizar uma onda senoidal ou um som exige que se determine inicialmente de quanto em quanto tempo se quer medir sua amplitude. Supondo que uma onda senoidal dure exatamente um segundo e se queira medir sua amplitude a cada 1/10 segundos, significa que se terá 11 amostras durante o segundo. Como mostrado na Figura 2.6 faz-se as amostras no início do tempo (T0), 1/10 segundos depois do início do tempo (T1),1/10 segundos depois (T2), e assim sucessivamente.

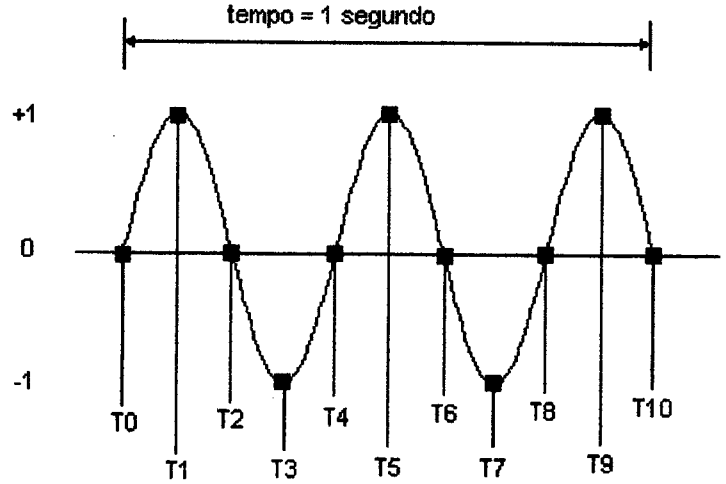

Figura 2.6 Onda senoidal amostrada a cada 1/10 de segundo.
Fonte:[MOO94]

Assim é possível elaborar uma tabela com os números que representam a onda senoidal:

Nr da amostra	0	1	2	3	4	5	6	7	8	9	10
Amplitude	0	+1	0	-1	0	+1	0	-1	0	+1	0

No tempo T0, a amplitude é 0; após 1/10 segundos, em T1, a amplitude é +1; e assim sucessivamente. Neste exemplo realizamos uma amostragem a uma Freqüência de 10 Hz (ou seja, a cada 1/10 segundos). Desta forma digitalizamos a onda senoidal tornando um sinal analógico em um sinal digital que pode ser armazenado e processado em um computador.

Trata-se de um processo parecido com o utilizado por produtores de discos compactos (CDs) para gravação de música em um CD.

2.4.1 Taxas de Amostragem

Se utilizarmos uma Freqüência de 10 Hz, como a do exemplo anterior, como taxa de amostragem, muita informação será perdida sobre o som que se pretende digitalizar. Deste modo, taxas bem mais altas são necessárias. Segundo [RID94], placas de som como a Sound Blaster utilizam freqüências de até 44.000 Hz como taxa de amostragem.

Quanto maior for a taxa de amostragem mais fiel será sua reprodução pelo equipamento. Se fizermos o processo inverso, ou seja, tentarmos reproduzir a onda senoidal digitalizada criada no exemplo anterior, obteremos uma forma de onda como a da figura 2.7.

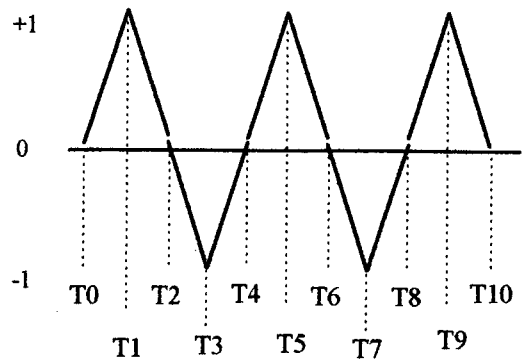

Figura 2.7 Resultado de uma taxa de amostragem baixa.
Fonte: [MOO94]

A baixa taxa de amostragem da onda senoidal só consegue alcançar os valores da onda nos seus máximos, mínimos e quando passa pelo ponto de

amplitude zero. Taxas baixas de amostragem proporcionam baixa fidelidade na reprodução do som original.

Resumindo, é assim que a amostragem digital funciona. Quando se fala no microfone conectado a uma placa de som tipo Sound Blaster, ela realiza uma medida da voz de até 44.000 vezes por segundo e armazena essa informação na forma de números na memória ou no disco do computador.

CAPÍTULO III

SISTEMAS DE RECONHECIMENTO DE VOZ

3.1 Introdução

Desde o fim dos anos 60, nossas expectativas por tecnologias de reconhecimento de voz poderiam ser caracterizadas pelas palavras: "Abra as portas, HAL". Embora estejamos ainda longe de manter um diálogo inteligente com um computador, finalmente o reconhecimento de voz no PC está avançado o suficiente para mudar a maneira como pessoas e máquinas se relacionam.

Os pacotes de reconhecimento de voz para PCs estão disponíveis há aproximadamente seis anos, mas até recentemente eram pouco mais que curiosidades que requeriam hardware demais. Nos últimos anos aconteceu uma explosão de produtos disponíveis comercialmente que trabalham em plataformas padrão, sem a necessidade de muito hardware adicional.

Hoje, os produtos de reconhecimento de voz mais simples permitem controlar aplicativos para Windows e OS/2 com comandos verbais em vez de mouse ou teclado. Sistemas mais sofisticados, com grande vocabulário, são capazes de tomar ditados ou servir como base para seus próprios aplicativos de reconhecimento de voz.

3.2 Categorias Básicas e Classificações

3.2.1 Categorias Básicas

Os produtos para reconhecimento de voz são classificados segundo RASH [RAS95] em três categorias básicas: *navegação, ditado e desenvolvimento*. Dos três, os produtos de navegação são os mais numerosos comercialmente.

3.2.1.1 **Sistemas de Ditado** - Capturam o texto à medida que o usuário fala. Trabalham com processadores de texto para criar documentos a partir da fala; alguns trabalham também com funções de navegação. Pode-se também permitir correção de erros durante o funcionamento, baseados no contexto. O aprendizado dos particulares padrões e inflexões da voz permitem uma aperfeiçoada precisão no reconhecimento.

3.2.1.2 **Navegadores de Voz** - Permitem controlar verbalmente programas como Windows ou OS/2 substituindo o mouse e o teclado na maioria das funções. Os navegadores analisam os comandos falados para manobrar através de menus e janelas, abrir e fechar aplicativos e controlar os movimentos do cursor. Alguns produtos vêm prontos para operar aplicativos comuns, outros precisam ser ensinados por repetição.

3.2.1.3 **Plataformas de Desenvolvimento** - Destinadas a programadores em C ou linguagens visuais, permitem construir aplicativos ativados por voz e acrescentar capacidades de reconhecimento de voz, para tarefas específicas, a aplicativos existentes. As ferramentas de desenvolvimento incluem código de amostra, APIs documentadas e as bibliotecas C necessárias ou controles personalizados.

3.2.2 Classificações

Os sistemas de reconhecimento estão classificados ao longo de um número de dimensões padrão. As capacidades de um sistema particular dependem fortemente de onde um sistema cai neste espaço, desta forma temos segundo [RAS95] e [WIL98]:

3.2.2.1 **Dependente do Locutor (SD) x Independente (SI):** se o sistema é treinado para reconhecer a voz de um indivíduo ou se pode reconhecer a voz de qualquer pessoa. Os sistemas adaptativos ao locutor também existem, o qual funciona inicialmente como independente do locutor mas tornam-se ajustados para a voz de usuários individuais.

3.2.2.2 **Discreto (IWR) x Contínuo (CSR):** se o usuário precisa pronunciar as palavras com pequenas pausas entre elas ou não. Os sistemas de reconhecimento de palavras isoladas são fáceis de implementar desde que o sistema conheça a extensão exata de cada palavra. Quando a fala é contínua problemas surgem principalmente na separação de palavras.

3.2.2.3 **Tamanho do Vocabulário**: dependente da tarefa, e é claro que quanto menor o vocabulário mais simples precisa ser o sistema.

3.2.2.4 **Tempo Real x Processamento Offline**: se a frase precisa ser traduzida para texto à medida que é pronunciada, ou se é permitido gastar algum tempo para isto.

3.3 Vantagens e Desafios

Os fabricantes de computadores estão agindo na suposição de que a fala se tornará um componente importante da interface de computador. Entre os motivos fundamentais para tal conclusão segundo [RUD94] estão situações como a seguir:

.**Fala como um atalho**. Ao invés de abrir um arquivo percorrendo muitos níveis de hierarquia, um usuário diz apenas "ABRA ORÇAMENTO".

.**Mãos ocupadas/Olhos ocupados**. Alterar o estilo da fonte enquanto um usuário está digitando, ou alterar a ferramenta de desenho enquanto o usuário está desenhando.

.**Recuperação da Informação**. Interfaces gráficas com o usuário são inconvenientes para especificação da recuperação baseada em restrições ("encontrar todos os documentos de João recebidos após Março").

.**Aplicações Portáveis**. Como computadores encolhem em tamanho do desktop ao notebook e subnotebooks, os teclados serão mais difíceis de usar ou mesmo sumirão, tornando assim a fala uma alternativa competitiva.

O principal desafio ao reconhecimento de voz em microcomputadores é a existência de uma madura e eficiente alternativa - o **teclado** (e apontador). É improvável que a fala possa substituir completamente estes dispositivos. Antes, a interface do futuro provavelmente combinará várias interfaces e permitirá ao usuário selecionar o modo de entrada ou combinação de modos que lhe será mais útil. Outros desafios a utilização da fala na interface com PC incluem **eficiente gerenciamento do erro, realimentação apropriada** ao usuário, e também **integração** dentro do ambiente computacional.

3.4 Problemas Existentes

Conforme colocado em [TER94] a fonte dos sons de vogais são as cordas vocais. O formato do trato vocal é modificado para alterar a ressonância, e a informação da voz é acrescentada as ondas sonoras. Nas consoantes, a fonte é uma parte do trato vocal, ao invés das cordas vocais. O reconhecimento de voz é a extração da informação lingüística da saída de voz que é formada nesta forma. O método principal atualmente para reconhecimento de voz é a combinação de padrões.

Os principais problemas apresentados por esta técnica são:

(1) A flutuação temporal dos padrões de voz característicos. O tamanho das palavras não é uniforme, ou seja, a velocidade da pronúncia de uma palavra pode variar. Naturalmente que, com pessoas diferentes esta velocidade varia mais.

(2) Os efeitos de diferenças no tamanho dos órgãos de fala de cada pessoa. Isto causa uma diferença na freqüência de ressonância, mesmo se o som é gerado com órgãos de fala com o mesmo formato.

Há também os sons que são influenciados pelos sons pronunciados antes ou depois deles, ou seja , problemas com co-articulações e diferenças no ambiente do locutor tais como dialetos.

3.5 Produtos Disponíveis no Mercado

Para ter sucesso, o reconhecimento de voz incorpora diversos tipos de tarefas. Quando falam, as pessoas tendem a juntar palavras e não dizem as mesmas palavras duas vezes da mesma forma. Assim, o programa deve dividir o fluxo da fala em palavras separadas e, para grafar corretamente cada palavra, os sistemas de ditado precisam também entender o contexto, a fim de distinguir, por exemplo, concerto de conserto.

Como este é um processo bastante complexo, alguns sistemas requerem que o usuário insira pausas entre as palavras. Esta fala discreta, torna o produto mais difícil de usar. Dizer. Cada. Palavra. Separadamente... é com certeza desanimador em muitas situações.

De forma compreensível, a tendência atual no mercado é na direção da fala contínua, em que os sistemas de reconhecimento trabalham da forma como as pessoas realmente falam. De modo geral, a fala contínua é usada pelos produtos de navegação e desenvolvimento. Os sistemas de fala contínua dependem de fonemas audíveis, os subcomponentes das palavras faladas. O programa de reconhecimento de voz deve seguir o rastro desses fonemas e discernir quando eles criam uma palavra.

Treinar sistemas dependentes do locutor requer que se pronunciem as palavras ou frases até que a máquina aprenda a fala do interlocutor suficientemente bem para reconhecer as palavras. Segundo Rash [RAS95] alguns produtos ditos como independentes do locutor na verdade precisam de um pouco de treinamento para assim o serem. Um dos produtos de ditado, o Kurzweil

Voice for Windows, requer relativamente pouco treinamento e está perto da independência.

Alguns sistemas como o Dragon NaturallySpeaking incorporam construtores de tópicos, construtores de vocabulários, facilidades do uso de macros. O construtor de tópicos permite observar os arquivos do usuário e em cima deles criar vocabulários voltados para as palavras destes arquivos. Usando múltiplos tópicos pode-se passar de uma taxa entre 90% e 97% para algo em torno de 95% a 99% de precisão no reconhecimento. Segundo o fabricante é possível atingir com o produto da Dragon Systems, Inc., uma velocidade de reconhecimento de 160 palavras por minuto.

A IBM lançou em setembro de 1996 seu primeiro programa de reconhecimento de fala contínua, o IBM Medspeak/Radiology, um sistema voltado para a área médica de grande desempenho e alto custo.

Abaixo são listados alguns dos principais produtos disponíveis no mercado segundo análise feita em [RAS95] e também através de dados dos próprios fabricantes obtidos em suas "home pages" em Janeiro de 1998.

.**Dragon NaturallySpeaking Deluxe** (Dragon Systems, Inc)- independente do locutor - fala contínua - múltiplos tópicos - vocabulário com 60.000 palavras - velocidade de 160 pal./min

.**DragonDictate for Windows** (Dragon Systems, Inc)- dependente do locutor - fala discreta - combina ditado com um navegador Windows - o treinamento exige que se falem cerca de 750 palavras individuais (20 min) - vocabulário com 230.000 palavras.

.**IBM ViaVoice** (IBM Corporation) - independente do locutor - ditado - fala contínua - vocabulário com até 64.000 palavras - velocidade de 125 pal./min.

.**Kurzweil Voice for Windows** (Kurzweil Applied Intelligence, Inc.) - independente - ditado -fala discreta - pouco treinamento - 85% sem treinamento - vocabulário com 60.000 palavras.

.**Listen for Windows** (Verbex Voice Systems, Inc.) - navegador - independente - algum treinamento.

.**Phonetic Engine** (Speech Systems, Inc) - independente - fala contínua - kit de desenvolvimento incluindo API - vocabulário com 60.000 palavras.

.**Microsoft Dictation** (Microsoft Corporation) - ditado - independente - vocabulário com 60.000 palavras - vocabulário pode ser aumentado.

.**SpeechMagic versão 2.0** (Philips Electronics) - independente - fala contínua - múltiplos contextos - vocabulário adaptativo - algum treinamento - vocabulário com 64.000 palavras.

3.6 Tendências Atuais

As atividades de pesquisa atuais em reconhecimento de voz são mais fortes na indústria, mas com presença acadêmica significante. De acordo com [WIL98] o principal patrocinador da pesquisa em reconhecimento de voz nos EUA tem sido o DARPA (Defense Advanced Research Projects Agency) através de seu programa de Tecnologia de Reconhecimento de Voz. Este programa tem se concentrado sobre dois domínios, reconhecimento de fala contínua baseado em textos do Wall Street Journal e no serviço de linguagem falada do Air Travel Information Service (ATIS). O primeiro está pretendendo estender a tecnologia de reconhecimento básica no que diz respeito a aumento do tamanho do vocabulário e operação em ambientes de dificuldade acústica. O serviço ATIS promove a integração de fala e processamento de linguagem natural e enfoca questões de diálogo e interface.

Segundo [RUD94], o programa DARPA produz uma avaliação competitiva anual de sistemas sob um corpo de testes comum, que tem provado ser altamente útil na promoção de rápidas melhorias em algoritmos. Esta competição anual começou a atrair a participação internacional. O DARPA tem encorajado o desenvolvimento de técnicas de processamento de linguagem natural que determinam a interpretação da fala espontânea e foca sua atenção em projetos de interface de voz.

As pesquisas de reconhecimento baseadas em redes neurais está também ativa, com ênfase atual posta sobre a transição da tecnologia para domínios mais amplos estudados pela comunidade de "Hidden Markov Model" (HMM). Sucesso tem sido alcançado por sistemas que combinam tanto a tecnologia HMM quanto redes neurais, usando a última para classificação do sinal e o primeiro para modelação da fala no tempo. Atualmente, estas comunidades de pesquisa tem andado mais ou menos independentemente.

Fora da comunidade DARPA a pesquisa tem se concentrado em diferentes frentes. Por exemplo a AT&T Bells Labs continua a fazer trabalhos significativos

de reconhecimento em telefone de largura de banda, enquanto a IBM continua a trabalhar com sistemas de dicção de grandes vocabulários.

A tabela I de [RUD94] mostra o aumento das capacidades de reconhecimento nos últimos anos. Como regra, se a taxa de erro estiver abaixo de 5% o sistema está pronto para ser comercializado, porém isto dependerá da aplicação específica.

Tarefa	Fim dos anos 70	Meio dos anos 80	Início dos anos 90
DL/DIS/Alfabeto	30%	10%	4%
IL/CON/Dígitos	10%	6%	0.4%
DL/CON/Rápido,1000 palavras(perplex.4)	2%	0.1%	
IL/CON/Rápido,1000 palavras(perplex.60)	-	60%	3%
DL/DIS/Dicção, 5000 palavras	-	10%	2%
IL/CON/Dicção, 5000 palavras	-	-	5%
IL/CON/Dicção, 20000 palavras	-	-	13%

[DL:Dep.Loc., IL:Indep.Loc., CON: Fala Contínua, DIS: Fala Discreta]
Tabela I. Progresso em reconhecimento de voz, expressado pela taxa de erro de palavra. Fonte:[RUD94]

Não apenas a taxa de erro de palavra deve ser adotada como método de avaliação, mas também o erro de expressão (se uma sentença inteira foi corretamente transcrita) ou erro de semântica (se a intenção do usuário foi identificada e a ação correta tomada). Em [OLI96] os erros em sistemas de reconhecimento de voz são definidos como sendo de 4 tipos:

- Rejeição: quando uma palavra falada não é retornada.
- Substituição: quando uma palavra é substituída por uma outra.
- Inserção: quando uma palavra não dita é retornada.
- Deleção: quando uma palavra é ignorada.

A disponibilidade da tecnologia de reconhecimento de voz tem encorajado o desenvolvimento de protótipos que aliam capacidades de fala com as interfaces de computador de propósito geral.

De forma geral, como diz [WIL98], os fabricantes tem procurado desenvolver uma máquina ou programa de software que possa:

- Reconhecer qualquer fala sem treinamento anterior.
- Reconhecer qualquer palavra.
- Permitir ao usuário falar de qualquer lugar.

- Discernir entre comandos, textos, e conversa do usuário.
- Trabalhar sob qualquer condição (ruído de fundo por exemplo).

Estes sistemas para se tornarem eficientes deverão além da tarefa de reconhecimento, também realizar a síntese da fala, permitindo ao computador falar. O uso da inteligência artificial é que fará com que o sistema possa tomar decisões e aprender a partir da experiência, e assim como diz Allen em [ALL96], "o esforço a ser realizado para comunicar-se com uma máquina será o mesmo que para comunicar-se com uma outra pessoa".

CAPÍTULO IV

LÓGICA DIFUSA

4.1 Introdução

A lógica difusa é uma poderosa técnica de resolução de problemas com uma vasta gama de aplicabilidades, especialmente nas áreas de controle e tomada de decisão. Em geral, é mais utilizada nos problemas ou situações difíceis de definir por meio de modelos matemáticos precisos.

Com a lógica difusa é possível tomar decisões e gerar respostas com base na informação vaga, ambígua, incompleta, ou imprecisa. Neste aspecto, os sistemas baseados em lógica difusa tem uma capacidade de raciocínio similar a dos humanos. A implementação de um sistema difuso requer muito menos memória e esforço computacional que os métodos convencionais, e além disto sua construção é mais compreensível, de fácil manutenção e mais robusta. Isto tudo gerará sistemas menores e menos dispendiosos.

Lotfi Zadeh, professor da Universidade da Califórnia em Berkeley, é a pessoa mais facilmente associada com a lógica difusa. De acordo com [JAN95], em 1965 ele apresentou o "paper" original definindo formalmente a teoria de conjuntos difusos, a partir do qual a lógica difusa emergiu. Tradicionalmente, uma premissa lógica tem dois extremos: ou completamente verdade ou completamente falso. Contudo, no mundo difuso, uma premissa pode ser parcialmente verdadeira e/ou parcialmente falsa segundo um grau de certeza de 0 a 100 por cento.

Pela incorporação deste conceito de "grau de certeza", a lógica difusa estende a lógica tradicional sobre dois caminhos. Primeiro, os conjuntos são rotulados qualitativamente (usando termos lingüísticos tais como "alto", "quente", "forte", "próximo" e assim por diante), e aos elementos destes conjuntos são atribuídos graus de pertinência. Por exemplo, homens de 1,80 m e de 1,95 m podem ambos serem membros de um conjunto de homens "altos", sendo que o homem de 1,95 m tem um grau de pertinência maior. Em segundo lugar, qualquer ação ou saída resultante de uma premissa existente verdadeira executa uma força refletindo o grau pelo qual a premissa é verdadeira.

Por acompanharem continuamente as entradas, as saídas podem evitar alterações abruptas, mesmo quando as entradas transcendem os limites do conjunto.

4.2 Origem

A referência a conjuntos difusos foi introduzida por Lotfi A. Zadeh, Universidade da Califórnia em Berkeley, 1965. Conforme [SCH94], neste trabalho Zadeh defendeu a tese de que um dos motivos pelos quais os homens são melhores ao controle, do que as máquinas existentes, é que eles são mais hábeis na tomada de decisões com base em informação lingüística imprecisa. Portanto, seria possível melhorar a performance dos controladores eletromecânicos, modelando a forma como os homens trabalham intelectualmente com este tipo de informação.

A teoria desenvolveu-se de forma lenta no início, mas no começo dos anos 70 atraiu um pequeno segmento internacional. Entre eles estavam um certo número de ocidentais, na maioria matemáticos, e um pequeno número de engenheiros japoneses. O interesse foi impulsionado pela curiosidade científica, mesmo sem muita fé na aplicabilidade final da teoria. Durante este tempo, as investigações focaram, principalmente, as propriedades matemáticas de conjuntos difusos.

Em 1971 Zadeh introduziu uma teoria de equações de estado para descrição do comportamento dos sistemas difusos (sistemas cujos parâmetros descritivos são valores difusos). Um marco importante deste desenvolvimento foi um paper de 1973, que introduziu a noção básica de variável lingüística, isto é, uma variável cujos valores são termos lingüísticos ao invés de números. O conceito de variável lingüística em combinação com a noção básica de regra difusa SE-ENTÃO, como por exemplo, "SE pressão é muito grande ENTÃO volume é muito pequeno", desempenhou um grande papel no que diz respeito a aplicação da teoria em problemas de tempo real.

Segundo [SCH94] ao final dos anos 70, o interesse em sistemas difusos cresceu explosivamente, atraindo pesquisadores de todo o mundo, e gerando milhares de publicações em sua maioria sobre trabalhos teóricos. A primeira aplicação comercial de porte de FLC foi um controlador de temperatura para um forno de cimento, desenvolvido pela Smith and Co. da Dinamarca.

O primeiro chip de inferência difuso foi desenvolvido no começo dos anos 70 nos laboratórios Bell da AT&T por Togai e Watanabe, implementando as operações Min e Max, que são usadas para representar a intersecção e união de conjuntos difusos.

Ao lado de redes neurais e algoritmos genéticos, conforme [ZAD94], a lógica difusa sustenta-se como um dos pontos de partida para a era do controle inteligente.

4.3 Teoria de Conjuntos Difusos

Em um conjunto clássico segundo [ZIM90], um elemento do universo, pertence ou não pertence a um conjunto. Isto é, a pertinência de um elemento é crisp - ela é sim ou não. Um conjunto difuso admite um grau de pertinência (número real entre 0 e 1) para cada elemento. Se o grau é zero, o elemento não pertence ao conjunto, e se é 1 o elemento pertence 100% ao conjunto. Qualquer outro valor entre 0 e 1 faz com que o elemento pertença parcialmente ao conjunto.

Supondo que os alunos de uma sala de aula são o universo e considerando um conjunto de pessoas "jovens". Seria correto considerar como "jovem" uma pessoa com 30 anos e "não jovem" uma com 31 ? O natural seria associar um grau de juventude a cada elemento, como por exemplo, {Ana/0.8, José/0.1, Cátia/1}. Talvez Ana tenha 28 anos, José 40, e Cátia 23. É claro que esses graus de pertinência poderiam ser diferentes de acordo com a situação e a intenção da pessoa que vai defini-los.

Tendo cada elemento do conjunto associado com um grau de pertinência como no exemplo anterior tem-se de fato a fundamentação de conjuntos difusos, bem como de sistemas difusos. A partir daí pode-se definir ou derivar várias propriedades e operações, muitas das quais tendo suas correspondentes em conjuntos ordinários, relações e lógica, enquanto outras pertenceriam apenas aos sistemas difusos.

Para [MUN94], um conjunto difuso A pode ser representado como um conjunto de pares ordenados de um elemento genérico x pertencente ao universo U, e seu grau de pertinência $\mu_A(x)$ como a seguir:

$$A = \{(x,\mu_A(x)) \mid x \in U\}$$

As operações difusas que possuem correspondência incluem a união, intersecção, complemento, relações binárias e composição de relações. Operações tais como fusificação são unicamente de conjuntos difusos. Existem várias formas diferentes de definição da união, intersecção e complemento. Segundo Mendel em [MEN95], as definições mais comuns são:

$$\mu_{A\cup B} = max(\mu_A(x), \mu_B(x))$$
$$\mu_{A\cap B} = min(\mu_A(x), \mu_B(x))$$
$$\mu_{\bar{A}}(x) = 1 - \mu_A(x)$$

Baseado nestas definições, podemos derivar as versões difusas de propriedades de conjuntos ordinários, tais como leis comutativas, leis de "de Morgan", etc. (ex. a lei comutativa para a definição de união precedente mantém-se $A \cup B = B \cup A$).

Como diz [ZIM90], os conjuntos difusos são extensões de conjuntos ordinários, a lógica difusa é uma extensão da lógica tradicional. Como existe correspondências entre os conjuntos ordinários e a lógica tradicional, também existe entre a teoria de conjuntos difusos a e a lógica difusa; por exemplo, união $\longleftrightarrow$ OU, intersecção $\longleftrightarrow$ E, e complemento $\longleftrightarrow$ NÃO. O grau de um elemento em um conjunto difuso pode corresponder ao valor verdade de uma proposição em lógica difusa. A lógica difusa pode representar implicações difusas tais como $A \Rightarrow B$, isto é, "*Se* A *então* B", onde A e B são conjuntos difusos. Por exemplo, "*Se* x é jovem *então* y é pequeno", ou simplesmente "*Se* jovem *então* pequeno" é uma implicação difusa.

O controle difuso é uma técnica de controle baseada na inferência difusa, discutida anteriormente. Segundo [GOM93] a lógica difusa constitui a base para o desenvolvimento de métodos e algoritmos de modelagem e controle de processos, permitindo a redução da complexidade de projeto e implementação, tornando-se uma solução para controle até então intratáveis por técnicas clássicas. Dadas entradas que são medições típicas do sistema a ser controlado, determinamos a saída para controlar o sistema. Em controle difuso, a entrada, saída, e ou regras podem envolver efusividade , levando ao uso de inferência difusa. Por exemplo, uma regra pode estabelecer que "se a temperatura é moderadamente alta então a velocidade é alta".

4.4 Sistemas Difusos

A figura 4.1 de [VIO93] ilustra o fluxo dos dados através de um sistema difuso.As entradas do sistema passam por três transformações para se tornarem saídas do sistema. Primeiro, é realizado um processo de fusificação que usa funções de pertinência pré definidas para mapear cada entrada do sistema dentro de um ou mais graus de pertinência. Então, as regras na base de regras (também pré definidas) são avaliadas pela combinação de graus de pertinência para formar

a potência da saída. E ao final, no processo de defusificação são calculadas as saídas do sistema baseado na potência e funções de pertinência.

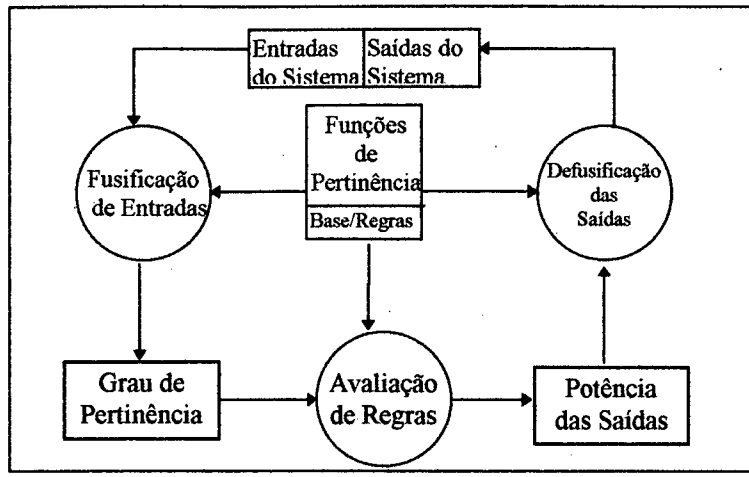

Figura 4.1: Fluxo de dados do Sistema Difuso.
Fonte: [VIO93]

4.4.1 Fusificação das Entradas.

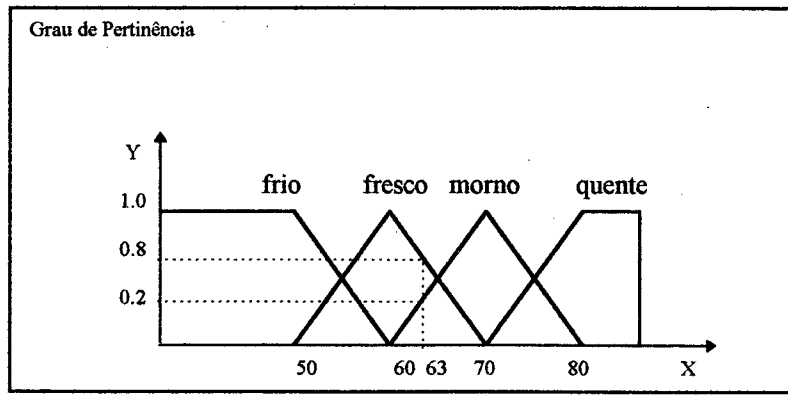

Figura 4.2 Conjuntos Difusos para temperatura.
Fonte: [VIO93]

Para [MEN95] a fusificação é o processo de atribuição ou cálculo de um valor para representar um grau de pertinência de uma entrada em um ou mais agrupamentos qualitativos, chamados "conjuntos difusos". A figura 4.2 mostra uma entrada do sistema, temperatura, com conjuntos difusos frio, fresco, morno e quente. Cada valor de temperatura tem um grau de pertinência em cada um destes conjuntos. O grau de pertinência é determinado pela função de pertinência, que é definida baseada na experiência ou intuição.As funções de pertinência podem ser

alteradas até que o sistema esteja ajustado para apresentar as respostas desejadas a determinadas entradas. Normalmente, no momento em que o sistema está em operação, as funções de pertinência não são mais modificadas. Formas simples como trapezóides e triângulos são freqüentemente usadas para definir pertinências em conjuntos difusos, mas outras funções podem ser usadas.

O número de funções de pertinência dos conjuntos difusos e o formato a escolher depende de coisas tais como precisão, responsividade e estabilidade do sistema, facilidade de implementação, manipulação e manutenção.

A sobreposição entre os limites dos conjuntos é desejável e chave para uma operação suave do sistema. Na figura 4.2, 63 graus podem ser tanto fresco quanto morno, mas é fresco com um grau maior.

Com a fusificação um projetista pode definir ou modificar o comportamento de um sistema usando uma linguagem natural, e assim facilitar descrições mais claras e concisas de tarefas complexas.

4.4.2 Avaliação de Regras.

Conforme [JAN95] para controlar o comportamento do sistema, o projetista desenvolve um conjunto de regras que tem a forma de declarações SE-ENTÃO. O lado SE de uma regra contém uma ou mais condições, chamadas "antecedentes"; o lado ENTÃO contém uma ou mais ações, chamadas "conseqüências". Os antecedentes das regras correspondem diretamente a graus de pertinência calculados durante o processo de fusificação.Por exemplo, considere uma regra potencial de um sistema de mercado de ações: SE o preço da ação é *decrescente* E o volume de negócios é *forte*, ENTÃO a ordem é *vender*. As duas condições "preço da ação é decrescente" e "volume de negócios é forte" são os antecedentes das regras.Cada antecedente tem um valor de grau-de-verdade (pertinência) atribuído a ele como resultado da fusificação. A ação da regra (ou "saída difusa") é vender ações. Durante a avaliação das regras, as potências são calculadas em cima dos valores antecedentes e então atribuídos as saídas difusas das regras. Geralmente, a função mínimo é usada de forma que para a potência de uma regra é atribuído o valor de seus antecedentes mais fracos ou menos verdadeiros.

Outros métodos, como multiplicação dos valores antecedentes, para calcular a potência de uma regra podem ser usados. Deste modo a quantidade de ações a serem vendidas está baseada no grau pelo qual o preço da ação é decrescente e o volume de negócios é forte.

Regra 1: Se A & B então Z
Regra 2: Se C & D então Z
Potência da Regra 1 = min(A,B)
Potência da Regra 2 = min(C,D)
X = Potência da Regra 1
Y = Potência da Regra 2
Z = max (X, Y)
= max (min(A,B),min(C,D))

Figura 4.3: Cálculo da avaliação de regras.
Fonte:[VIO93]

Freqüentemente, mais que uma regra aplica-se a mesma ação específica, e neste caso a prática comum é usar a mais forte ou mais verdadeira das regras, veja a figura 4.3.

4.4.3 Defusificação das Saídas.

A defusificação deve ser feita por duas razões principais segundo [ZIM90]. A primeira é decodificar o sentido de ações vagas (difusas), tais como "ordem é vender" usando funções de pertinência. A segunda é resolver conflitos entre ações diversas tais como "ordem é vender" e "ordem é segurar", que podem ter sido disparadas por certas condições durante a avaliação das regras.

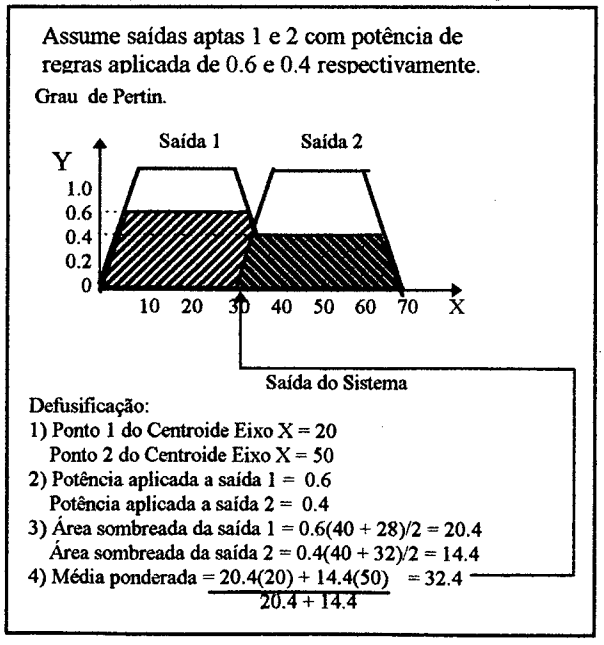

Figura 4.4.Defusificação das saídas pelo método do centro de gravidade ou centróide.
Fonte: [VIO93]

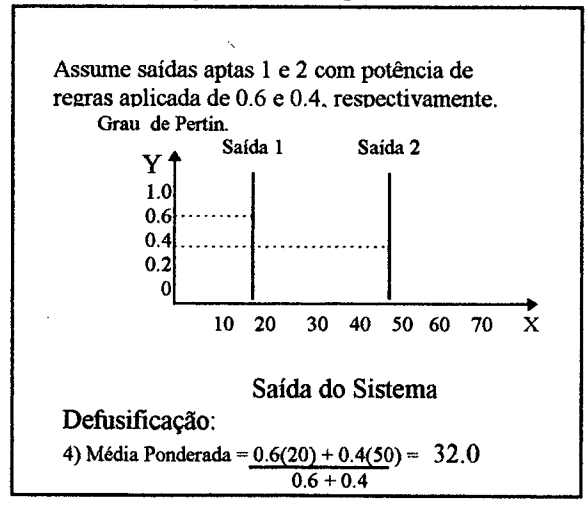

Figura 4.5: Defusificação das saídas pelo método de pulsos (ou singleton).
Fonte: [VIO93].

Conforme [VIO93] uma técnica de defusificação comum, o "método do centro de gravidade ou centróide", consiste de vários passos. Inicialmente, um ponto centróide do eixo X é determinado para cada função de pertinência de saída. Então, as funções de pertinência são limitadas na altura pela potência da regra aplicada, e as áreas das funções de pertinência são computadas. Finalmente, a saída defusificada é obtida por uma média ponderada do ponto centróide do eixo X e das áreas computadas, com as áreas servindo como pesos. O método do centro de gravidade é ilustrado na Figura 4.4.

Algumas vezes, pulsos são usados para simplificar o processo de defusificação; veja Figura 4.5. Um pulso é uma função de pertinência de saída representada por uma única linha vertical. Já que um pulso intercepta o eixo X em um único ponto, o cálculo do centro de gravidade reduz-se exatamente a um cálculo da média ponderada de pontos do eixo X e potência de regras, com a potência das regras usadas como pesos.

4.5 Limitações dos Sistemas Difusos

Em [SCH94] , [MUN94] e [LI89] encontramos algumas limitações ao uso de sistemas difusos, quais sejam:

(1) **Estabilidade**: Ao contrário da teoria clássica de controle, que já tem estabelecidos vários critérios de estabilidade que podem ser aplicados ao sistema, para os sistemas difusos isto não existe.

(2) **Capacidade de Aprendizagem:** Aos sistemas difusos faltam capacidades de aprendizagem e memorização, e este é o motivo por estarem surgindo cada vez mais sistemas híbridos neurodifusos.

(3) **Determinação ou Ajuste das funções de Pertinência e regras difusas** não são muito fáceis. Mesmo após muitos testes é difícil dizer quantas regras ou quantas funções de pertinência são realmente necessárias.

(4) Existe uma **concepção geral do termo difuso** como significando impreciso ou imperfeito. Muitos pensam que a lógica difusa representa alguma mágica sem fundamentação matemática firmada.

4.6 Aplicações da Lógica Difusa

A aplicação da tecnologia de conjuntos difusos nos produtos de consumo japoneses iniciou em 1990. Estes produtos são agora também comercializadas no resto do mundo. A maioria destas aplicações não faz uso chips de inferência, mas

ao invés disto usam simulações através de tabelas de busca em GIs digitais padrão. Tal abordagem é útil onde não se requer a velocidade de um chip de inferência difuso especialmente dedicado. Como exemplo podemos citar informações de [SCH94]. A Canon introduziu um controlador lógico difuso no mecanismo de autofoco de uma câmera de vídeo de 8mm O 'Palmcord'da Matsushita (Panasonic) usa lógica difusa para estabilização de imagem. Matsushita Eletric, Hitachi, Sanyo e Sharp agora tem suas próprias "máquinas de lavagem difusa", que automaticamente ajustam ciclos de lavagem em resposta as várias combinações de tamanho de carga, tipo de sujeira (terra versus gordura), quantidade de sujeira, e tipo de fabricação. Na máquina da Matsushita, o tipo e quantidade de sujeira são detectados por meio de sensores óticos, que também usam controles difusos.

Schwartz em [SCH94] classifica as aplicações de FLC com relação a "tempo de atraso"no ciclo de controle, da seguinte forma:

Com tempo de atraso pequeno em controle operacional estariam:

- sistemas de condicionamento de ar
- freios anti-derrapantes
- robôs de soldagem
- diafragma de auto-íris para VTR
- foco automático para câmera compacta
- transmissão automática
- robôs autônomos
- ajuste de cor da TV
- controle para grupo de elevadores
- máquina de dragagem
- fornecedor de água quente para chuveiro
- controle de velocidade para automóveis
- escavador de túneis

Sob Controle de Processos (grande tempo de atraso) podemos colocar

- controle de combustão de fábrica de incineração de lixo
- fornos de cimento
- fornalha de fundição de vidro
- bomba de água
- moinho
- ventilação de túneis
- processo de purificação de água

O gráfico da figura 4.6 baseado em dados de [MUN94] mostra que as aplicações de sistemas difusos tem crescido significativamente nos últimos anos, o que leva crer que num futuro próximo ocupará espaço importante nestas áreas.

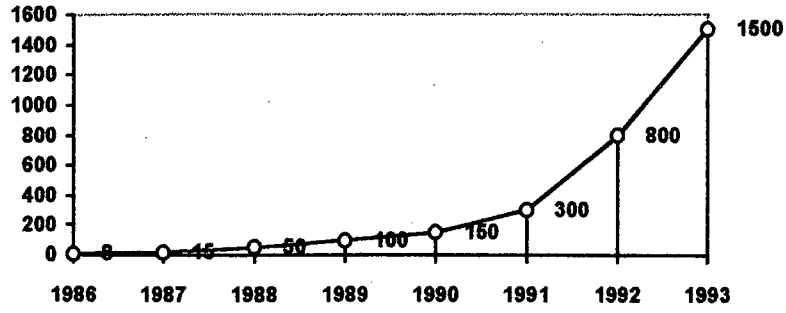

Figura 4.6. Número aproximado de aplicações industriais e comerciais de sistemas difusos. Fonte: baseado em dados de [MUN94]

4.7 Sistemas Híbridos

A tendência atual é utilizar várias formas de sistemas híbridos de lógica difusa com outras áreas tais como redes neurais e algoritmos genéticos.

.Sistemas Híbridos Difuso-Rede Neural - Existem diversas pesquisas sugerindo várias formas do uso combinado de lógica difusa e redes neurais [JAN95], [KOM93], colocando-as como técnicas complementares. Por exemplo, não existem capacidades de aprendizagem ou memória em sistemas difusos. Os sistemas híbridos poderiam acrescentar tais capacidades.

.Sistemas Híbridos Difuso-Algoritmo genético - Aplicações de algoritmos genéticos combinado com controle difuso estão sendo investigados não apenas a nível acadêmico mas também a nível comercial. Segundo [GOL94] os algoritmos genéticos são particularmente apropriados para ajuste das funções de pertinência.

4.8 Considerações Finais

Os conceitos sobre conjuntos e lógica difusa permitem aproximar um pouco mais a programação da forma como o homem raciocina. Qualquer programador pode sem problemas maiores escrever código para implementar uma máquina de inferência difusa. No entanto, existem muitas ferramentas de desenvolvimento difuso que permitem ao projetista se preocupar mais com aplicação e o comportamento do sistema

A implementação de máquinas difusas manualmente, no entanto, oferece a possibilidade de entender e adaptar o sistema à aplicação em questão.

CAPÍTULO V

LÓGICA DIFUSA EM RECONHECIMENTO DE VOZ

5.1 Introdução

Os sistemas de reconhecimento de voz já estão disponíveis há algum tempo, mas suas limitações tem impedido a difusão de seu uso. Nenhum sistema de fala é 100% preciso, o que leva a crer que muito ainda precisa ser feito, haja visto que para aumentar o reconhecimento em 0.1 % é necessário muito esforço científico.

Os sistemas de reconhecimento de voz existentes e citados na bibliografia fazem uso de técnicas muito variadas. A razão disto tem muito haver com a também diversidade de características do sistema, onde estão incluídos: performance necessária, dimensão do sistema, idioma e ambiente do usuário.

A utilização da lógica difusa como base de um sistema de reconhecimento de voz, para operar uma calculadora especificamente, e não qualquer outro tipo de sistema, deve-se a simplicidade de sua implementação, teste e também a sua grande utilidade prática. O sistema proposto vai estar limitado as seguintes características:

.Vocabulário limitado a 18 palavras.
.Fala Discreta (não contínua).
.Dependência do locutor
.Ambiente com baixo nível de ruído

Aplicar lógica difusa no reconhecimento de voz não prima pelo ineditismo, mas também não faz parte da lista de técnicas mais utilizadas em reconhecimento de voz. São poucas as citações bibliográficas encontradas, o que dificulta sua implementação. Neste capítulo é proposta uma abordagem, baseada na lógica difusa para reconhecimento de voz, bem como a verificação da sua utilidade dentro dos limites propostos.

5.2 Etapas do Reconhecimento de Voz

Para construir um sistema de reconhecimento de voz deve-se levar em conta as três etapas mostradas na figura 5.1.

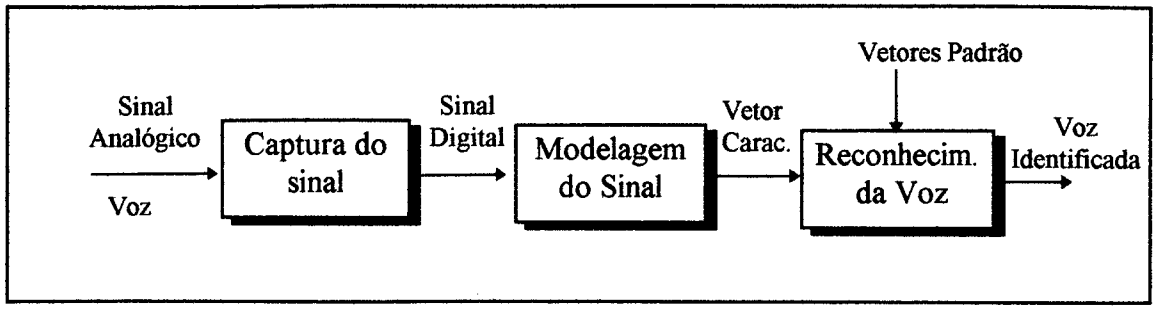

Figura 5.1 Etapas de um sistema de reconhecimento de voz.
Fonte:[PIC93]

5.2.1 Captura do sinal

Para captura do sinal de voz são necessários um microfone e uma placa de som que transformará o sinal analógico proveniente do microfone em um sinal digital compreendido pelo computador.

Ridge [RID94] afirma que certamente a escolha do microfone e da placa de som são importantes, uma vez que isto poderá aumentar ou diminuir o desempenho do sistema. Microfones de baixa qualidade poderão atenuar ou até mesmo eliminar freqüências significativas do sinal de voz, além de introduzir ruído no sistema. Em [FLE98] é feita uma análise da influência do microfone no desempenho dos softwares de reconhecimento de voz "Via Voice" da IBM e "Dragon NaturallySpeaking" da Dragon Systems.

A placa de som também pode ser responsável por problemas nas etapas seguintes do processamento. A placa deve possuir uma largura de faixa que cubra toda a faixa de freqüências da voz humana. Se isto não ocorrer informações importantes poderão ser perdidas complicando a tarefa de reconhecimento. Conforme [RID94], a largura de faixa global do sistema de som depende do elo mais fraco da cadeia, que pode ser o microfone ou a placa de som. O ruído e a distorção introduzidos pelos componentes eletrônicos e pela fiação também devem estar a um nível mínimo possível.

O conversor A/D da placa de som converte sinais analógicos contínuos a partir do microfone em uma série de valores digitais discretos por meio de amostragem, tomando medições de amplitude instantânea do sinal a uma taxa constante. Se essa medição for feita a uma taxa suficientemente alta, se obterá uma forma de onda mais parecida com a forma de onda do sinal analógico. Isto pode ser observado na figura 5.2.

A fidelidade de som digital depende da taxa de amostragem e do tamanho da amostra, a quantidade de bytes usados para armazenar cada amostra. A taxa de amostragem deve ser o mais alta possível, o limite está no espaço em disco necessário para armazenar o som.

Para um som de voz é adequada uma taxa de amostragem em torno de 8.000 Hz. Placas de som como a Sound Blaster que utilizou-se nesse trabalho possuem taxas que vão de 8.000 a 44.100 Hz. Para economizar também espaço em disco e reduzir o tamanho do arquivo a ser tratado usaremos amostras de 8 bits que são suficientes e o mais indicado. Com estes valores para armazenar um som que dure 2 segundos serão ocupados 16 Kbytes do disco.

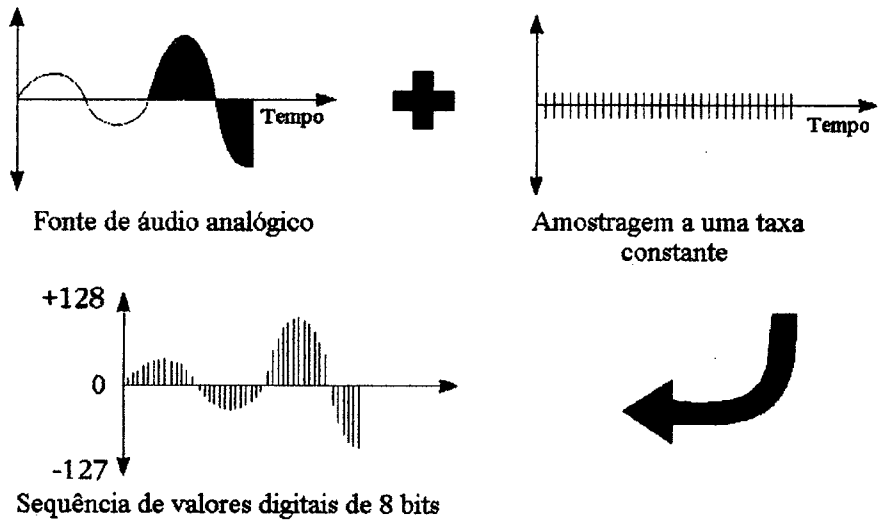

Figura 5.2 Conversão Analógico-Digital.
Fonte: [RID94]

5.2.2 Modelagem do sinal

A modelagem do sinal é uma etapa importante da tarefa do reconhecimento de voz porque nela são definidos e obtidos os parâmetros que facilitarão a identificação da palavra falada. Existem três itens principais que devem ser observados no projeto da modelagem do sinal, quais sejam:

(1) As parametrizações devem procurar aquilo que represente um aspecto saliente do sinal de voz;
(2) As parametrizações precisam ser robustas para superar variações no canal, locutor e transdutor;
(3) Os parâmetros que capturam a dinâmica espectral, ou variações do espectro no tempo, referido como *correlação temporal*;

Segundo [PIC93] a modelagem do sinal requer agora menos de 10% do tempo de processamento requerido em uma aplicação de reconhecimento de voz baseada em grandes vocabulários.

A modelagem do sinal que é bom para um tipo de aplicação pode não ser para uma outra. Nas aplicações de identificação do locutor ou dependentes do locutor, características singulares de aprendizagem do usuário e seu ambiente acústico são importantes.

Nesta etapa são utilizadas técnicas que fazem o pré-processamento do sinal. Entre estas pode-se citar:

- filtragem de bandas de freqüência - eliminar do sinal informações que não fazem parte da faixa de voz. Normalmente as freqüências ficam limitadas entre 100 Hz e 3.200 Hz.
- normalização do sinal - compensar diferenças no volume ou no timbre de voz por exemplo.
- filtragem pré-ênfase - enfatizar a faixa de freqüência em que o ouvido humano possui maior sensibilidade (normalmente acima de 1.000 Hz).

Os parâmetros mais importantes na modelagem do sinal estão relacionados com a amplitude e freqüência do sinal

5.2.2.1 Amplitude

A amplitude do sinal é normalmente calculada em intervalos regulares de tempo, com o objetivo de permitir uma análise de sua dinâmica. Estes intervalos são definidos como *tempo de duração de janela.*

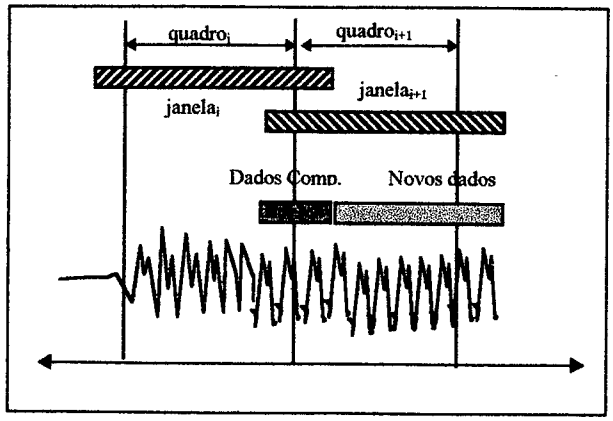

Figura 5.3 Análise de sobreposição baseada em quadros.
Fonte: [PIC93].

A figura 5.3 mostra a análise baseada em *quadros*,que será utilizada no sistema. Neste caso, uma sobreposição de 33% é mostrada.Um terço dos dados usados em cada quadro de análise é compartilhado com o quadro anterior. Note que apenas um terço dos dados são particulares ao quadro atual -- os restantes dois terços são compartilhados entre quadros adjacentes onde o sinal é dividido em intervalos de tempo ou quadros. Esta sobreposição permite controlar o quão rapidamente os parâmetros poderão variar de quadro para quadro. Quantidades grandes de sobreposição podem esconder variações significativas no sinal.A duração ou período do quadro é definido como o comprimento do tempo (em segundos) sob o qual um conjunto de parâmetros é válido. A taxa de quadro, também um outro termo comum, é o número de quadros computados por segundo (hertz).

A duração do quadro normalmente varia entre 20 e 10 ms, nos sistemas práticos. A duração do quadro está intimamente dependente da velocidade dos articuladores no sistema de produção de voz (taxa de alteração do molde do traço vocal).Enquanto alguns sons de voz (tais como consoantes ou ditongos) exibem transições espectrais agudas que podem resultar em picos espectrais tais como 80 Hz/ms, durações de quadros menores que aproximadamente 8 ms não são normalmente usados. Para o sistema proposto é verificado qual a melhor opção.

5.2.2.2 Freqüência

A freqüência também é uma informação importante do sinal de voz, pois nela estão contidas informações que vão permitir diferenciar sons graves de sons agudos. Palavras como "um" possuem uma freqüência média menor que a da palavra "seis" por exemplo. Cada pessoa também possui um tom de voz mais agudo ou mais grave. Esta informação vai complicar o reconhecimento em situações onde o treinamento do sistema foi feito para uma determinada voz e foi usado por outra.

Como uma palavra é composta de diversos fonemas, e cada fonema possui uma freqüência média, também a freqüência precisa levar em conta o tempo de ocorrência do fonema na palavra. Isto fará com que haja diferenças significativas para o sistema, entre pronunciar as palavras "possa" e "sapo" por exemplo.

Estes parâmetros, que ainda poderão incluir o tempo de pronúncia da palavra, vão compor um vetor característico da palavra falada. Este vetor deverá ser comparado com um vetor padrão de cada palavra que comporá o vocabulário do sistema. Esta etapa de comparação e identificação de palavras é de responsabilidade do reconhecedor de voz propriamente.

Uma abordagem muito utilizada e muito citada nesta etapa é a que faz uso *de Transformada Rápida de Fourier* (FFT - Fast Fourier Transform). Segundo [ALL90] a FFT traduz uma função no domínio do tempo para uma função no domínio da freqüência. A FFT permite decompor uma onda complexa em diversas ondas que somadas irão gerar a onda original. Com a FFT consegue-se reduzir consideravelmente a quantidade de sinais tratados, porém perde-se uma informação significativa para o reconhecimento de voz, que é a relação temporal entre a pronúncia dos fonemas das palavras. Voltamos ao problema do "sapo"e do "possa".

5.2.3 Reconhecimento da Voz

Uma vez vencidas as etapas anteriores cabe ao projetista definir qual será a abordagem utilizada para realizar a comparação entre os vetores característicos gerados a cada palavra pronunciada e aqueles que compõem o banco de padrões obtidos através de alguma forma de treinamento pré-realizado.As técnicas utilizadas incluem:

- Modelagem Oculta de Markov (HMM - hidden Markov Model) - utilizado em sistemas de reconhecimento de fala contínua como o sistema SPHINX é um método de aprendizagem estatística composto de estados e transições [RAB89].
- Redes Neuronais - baseadas no funcionamento do cérebro humano, possuem elementos que simulam os neurônios biológicos permitindo aprendizagem a partir de treinamento [MOR95].
- Lógica Difusa - utiliza lógica difusa associada com alguma outra técnica [KOM93], [JAN95], ou isoladamente [TER94]e [PAL77].

Também em muitos dos sistemas de fala contínua são utilizadas gramáticas como no SPHINX e no TANGORA. O objetivo é reduzir ou ampliar a possibilidade de que uma palavra possa ocorrer em função das pronunciadas anteriormente.

5.3 Desenvolvimento do modelo computacional

O ambiente escolhido para desenvolvimento do sistema foi o MS-DOS, que permitiu uma resposta em tempo real adequada aos propósitos do trabalho. O compilador Borland C++ versão 3.1 foi utilizado como ferramenta para sua elaboração.

O sistema passou por diversas fases e testes que permitiram regular parâmetros e escolher métodos que alcançassem uma taxa de reconhecimento maior.

Na fase inicial foram feitas as escolhas que determinaram limites, características e métodos a serem adotados na construção do sistema e na sua utilização ao final.

Na segunda fase foram implementados os algoritmos de pré-processamento do sinal que permitem, após gerar o vetor característico da palavra, o reconhecimento da mesma na fase seguinte.

Na terceira fase foi utilizada a lógica difusa para realizar a tarefa de reconhecimento. Foram cumpridas etapas de definições e escolhas relacionadas ao formato dos conjuntos difusos, regras e métodos de fusificação e defusificação.

Na última etapa foi feito uma realimentação, ou seja, foram reajustados parâmetros de diversas partes do sistema, modificados formatos dos conjuntos difusos, e revistos alguns métodos. A busca do aumento da taxa de reconhecimento foi o objetivo a ser alcançado.

A seguir são descritas estas fases e relatados os resultados alcançados.

5.3.1 Fase 1: Escolha das características da aplicação

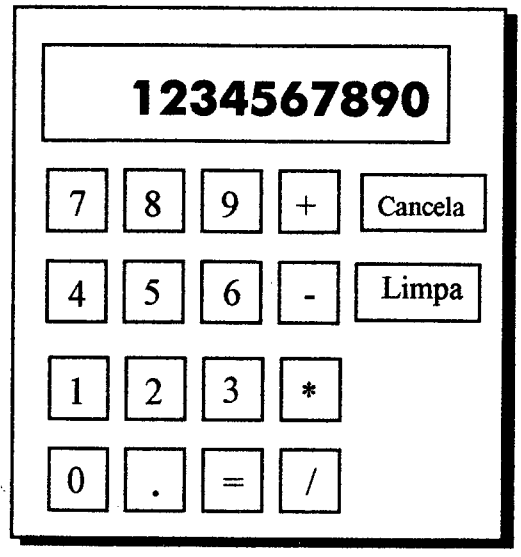

Figura 5.4 Calculadora controlada por voz.

Para operar uma calculadora "não científica" com capacidade de visualização de nove dígitos, como aquela mostrada na figura 5.4 foram escolhidas dezoito palavras que irão compor o vocabulário do sistema.

- 10 números: *Zero, Um, Dois, Três, Quatro, Cinco, Seis, Sete, Oito* e *Nove*.
- 4 operadores: *Mais, Menos, Vezes* e *Divide*.
- 1 para execução: *Igual*.
- 1 para vírgula decimal: *Ponto* (sistema americano: mais comum).
- 2 para controle: *Cancela* e *Limpa*.

O sistema não faz uso de nenhum método que trate uma maior probabilidade de uma palavra ser pronunciada ou não. Por exemplo: se já foi digitado(ou em nosso caso "falado") o ponto decimal no cálculo atual e a última palavra foi identificada também como *ponto*, porém com uma pequena diferença no índice em relação a palavra "*oito*". Poderíamos usar um algoritmo que levasse em conta a maior probabilidade de ter sido pronunciada a palavra *oito* ao invés de *ponto*, não só neste caso mas também em outros onde a situação é semelhante, como com:

- operadores em seqüência.
- número pronunciado com visor já cheio.
- operadores antes de qualquer número.
- etc.

5.3.2 Fase 2: O Pré-processamento do sinal

Nesta fase são pré-gravadas as diversas palavras do vocabulário e baseado no que foi colocado no capítulo 2 feitas as escolhas das entradas do sistema. Foram escolhidas três entradas: freqüência, amplitude e envoltória. O método para obte-las a partir do sinal de voz foi o seguinte:

Freqüência: Conforme mostra a figura 5.5 foi medido o período de um ciclo do sinal digital e a partir daí calculada a freqüência. Cada quadro posteriormente será composto de vários ciclos com uma freqüência média correspondente. As freqüências com valores fora da faixa de audição humana são desprezadas.

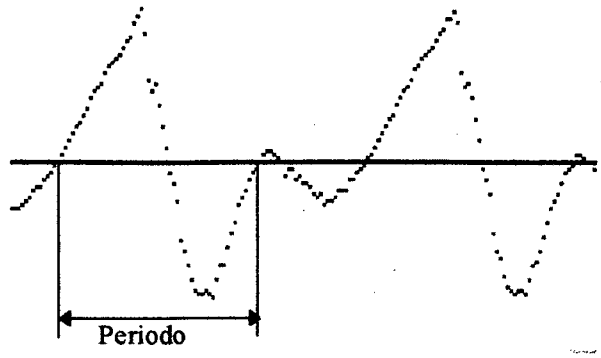

Figura 5.5 A freqüência do sinal de voz.

Amplitude: É calculada em função da soma dos valores instantâneos do ciclo positivo do sinal como mostra a figura 5.6. Como os valores dependerão da intensidade do sinal de voz, precisam ser normalizados pelo valor máximo. Deste modo o volume do sinal de voz não será um problema crítico na operação do sistema.

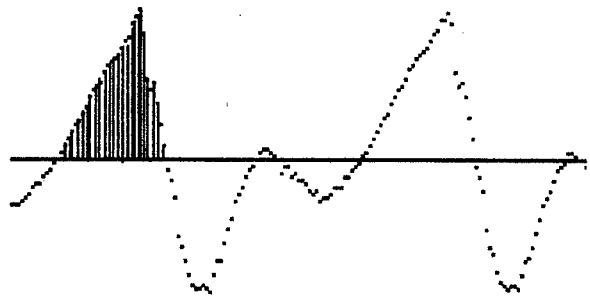

Figura 5.6 A amplitude do sinal de voz.

Envoltória: No caso da envoltória,como mostra a figura 5.7, são tomados apenas os valores de pico de cada ciclo. Assim como acontece com a amplitude, os valores precisam passar por um processo de normalização.

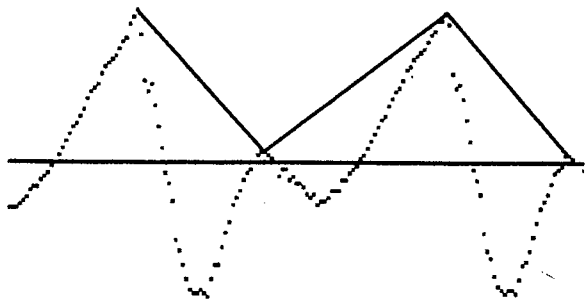

Figura 5.7 A envoltória do sinal de voz.

A partir destas informações o sistema divide o sinal de voz em vinte quadros como mostrado na figura 5.8. Cada quadro é composto por n ciclos. Este número de ciclos será determinado em função do tamanho do sinal de voz, e portanto variará de palavra para palavra.

De acordo com Picone [PIC93] uma taxa de sobreposição de quadros é utilizada para que haja uma suavização na mudança das informações de um quadro para outro. Para introduzir o conceito de sobreposição são criadas janelas que possuem 100% dos dados referentes a um quadro específico e um percentual dos dados referentes aos quadros adjacentes. Deste modo o sistema trabalhará não com vinte quadros mas com vinte janelas. Esta divisão em janelas permite captar a dinâmica espectral do som, ou seja, a evolução no tempo da palavra falada, extraindo daí informações que ajudem no reconhecimento.

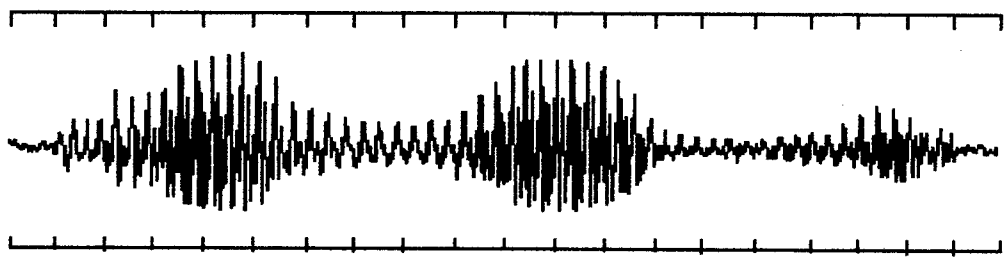

Figura 5.8 Sinal de voz dividido em vinte quadros

Os dados referentes as três entradas (freqüência, amplitude e envoltória) são processados gerando um vetor característico do sinal de voz. Para chegar até este vetor o sinal passou por etapas de filtragem de bandas de freqüência, normalização do sinal, e eliminação de informações insignificantes.

5.3.3 Fase 3: Aplicação da Lógica Difusa

A lógica difusa tem se mostrado extremamente útil em aplicações de controle e tomada de decisão, baseadas no modelamento da forma como funciona o raciocínio humano. No caso do reconhecimento de voz esta situação não se caracteriza tão claramente, pois os conjuntos difusos não são termos lingüísticos como "alto", "quente" ou "forte". O que são utilizados são faixas de valores difusos.

A implementação do sistema difuso tem como base o sistema sugerido por [VIO93]. A implementação é feita em linguagem C e fornece as rotinas básicas para fusificação das entradas, defusificação das saídas, avaliação de regras, etc.

5.3.3.1 Fusificação das Entradas

A escolha do número de conjuntos difusos de cada entrada e sua função de pertinência (trapezoidal ou triangular) é sempre uma tarefa difícil e demorada em um sistema difuso. Em nosso caso também foi assim. Foram escolhidos 10 conjuntos difusos para cada entrada significando cada um deles uma faixa de valores(de F0 a F9), ou seja, uma faixa F0 teria um valor médio de freqüências(ou amplitude, ou envoltória) menor que o da faixa F1, e assim por diante, como mostrado na figura 5.9. Deste modo um determinado valor de freqüência como entrada, vai possuir um grau de pertinência em um ou mais conjuntos difusos.

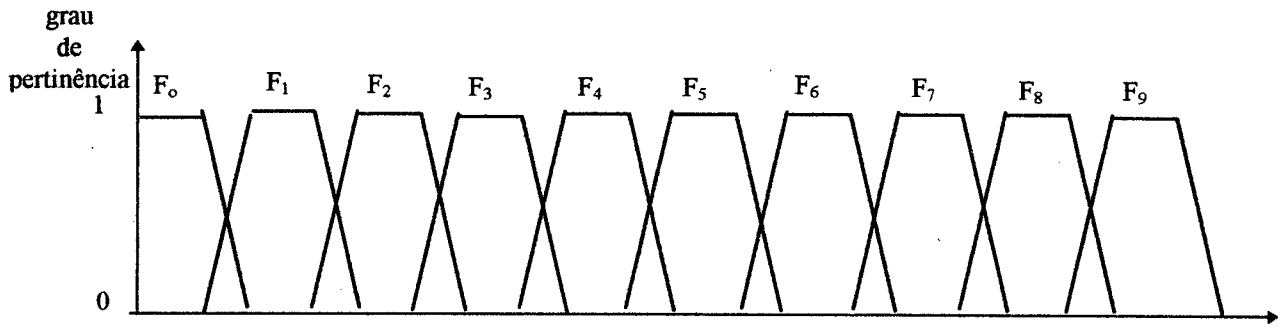

Figura 5.9. Conjuntos difusos de entrada

A escolha definitiva das funções de pertinência foi determinada na fase final de ajustes, através da utilização de um grupo de palavras pré-gravadas e da observação da taxa de reconhecimento alcançada com cada variação. O programa permite alterar com facilidade o formato das funções de pertinência de entrada o que facilita a observação dos resultados após alterações. Cada função tem um valor de largura para a subida e outro para a descida e também um valor de topo como mostra a figura 5.10.

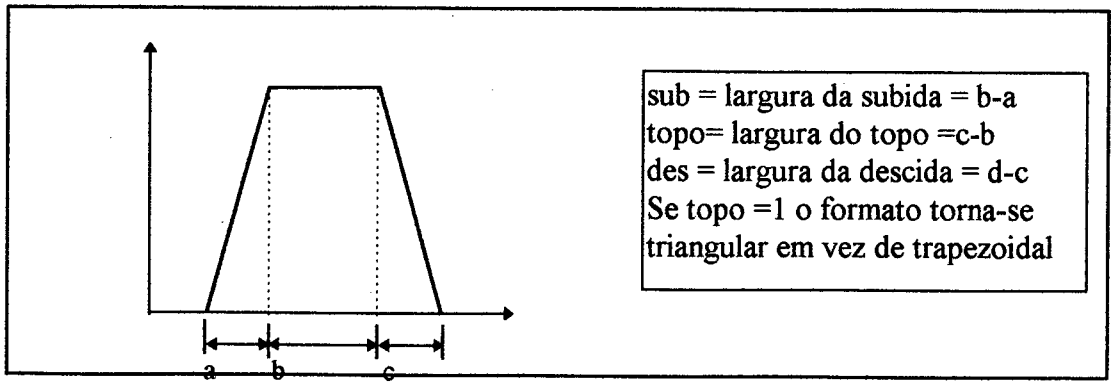

Figura 5.10 Parâmetros que permitem determinar o formato da função de pertinência de cada conjunto difuso.
Fonte:[VIO93]

Estas informações sobre as funções são gravadas em disco em um arquivo como exemplificado na figura 5.11. Na primeira linha o nome do conjunto. Na última linha os valores de sub, topo e desc. Nas linhas intermediárias os valores do número de faixa a, b, c e d , conforme a figura 5.10.

5.3.3.2 Regras

Para realizar o processo de inferência o sistema dispõe de 18 regras, uma para cada palavra do vocabulário. Estas regras tem a forma de declarações se-então e determinarão a saída do sistema. O lado "se" da regra é composto de informações sobre as faixas de freqüência, amplitude e envoltória de cada janela do sinal, e construída de acordo com os dados do treinamento. Desta forma as regras serão alteradas no momento em que um outro usuário utilizar o sistema e fizer um novo treinamento.

```
FREQÜÊNCIA
0 0 1 10 40
1 0 10 20 50
2 0 20 30 60
3 0 30 40 70
4 10 40 50 80
5 20 50 60 90
6 30 60 70 100
7 40 70 80 110
8 50 80 90 120
9 60 90 100 130
30 10 30
```

Figura 5.11 Formato de um dos arquivos do conjunto difuso de entrada do sistema.

A figura 5.12 mostra uma parte das regras, onde Fj_n é a freqüência média da janela n, Aj_n a amplitude média da janela n, Ej_n o valor da envoltória média da janela n, F_n o conjunto difuso da freqüência n, A_n o conjunto difuso da amplitude n, E_n o conjunto difuso da envoltória n, Pal a saída, PA_n o conjunto difuso de saída.

```
Se   (Fj0  = F0 e Aj0  = A2 e Ej0  = E3)então Pal = PA0
Se   (Fj0  = F4 e Aj0  = A2 e Ej0  = E3)então Pal = PA1
Se   (Fj0  = F6 e Aj0  = A0 e Ej0  = E4)então Pal = PA2
     ...   ...   ...
Se   (Fj0  = F8 e Aj0  = A2 e Ej0  = E1)então Pal = PA17
     ...   ...   ...
Se   (Fj1  = F4 e Aj1  = A5 e Ej1  = E9)então Pal = PA0
     ...   ...   ...
     ...   ...   ...
Se   (Fj19 = F6 e Aj19 = A0 e Ej19 = E0)então Pal = PA17
```

Figura 5.12 Conjunto de regras do sistema.
Fonte:[VIO93]

Os operadores de intersecção e união serão definidos durante o período de teste do sistema, com intuito de analisar qual se adapta melhor a situação em questão.

5.3.3.3 Defusificação das saídas

A saída do sistema possui 18 conjuntos difusos (um para cada palavra do vocabulário) de acordo com a figura 5.13.

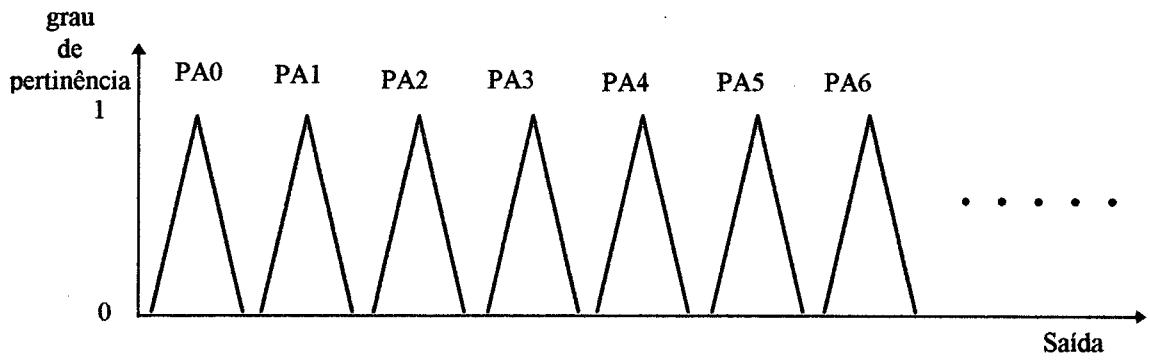

Figura 5.13 Conjuntos difusos de saída.

Não há sobreposição dos conjuntos pelo fato de ser impossível colocar uma palavra entre duas outras e dizer que ela tem um som intermediário entre estas elas.

Para defusificação das saídas não será utilizado o método do centróide ou centro de gravidade. O sistema tomará como saída o valor médio do conjunto com maior grau de pertinência. Este valor médio identificará uma das palavras que compõem o vocabulário.

5.3.3.4 Estrutura dos dados difusos

A estrutura dos dados difusos é feita segundo o modelo sugerido por Viot [VIO93]. As entradas do sistema são arranjados em uma lista lincada e nodos função de pertinência como mostra a figura 5.14. Esta estrutura está mais detalhada na figura 5.15.

O nodo de entrada do sistema é alocado na memória e contém um nome de entrada, um ponteiro da função de pertinência, e um ponteiro da próxima entrada. Mais interessante é a estrutura da função de pertinência, que contém dois pontos no eixo X e dois valores de inclinação que descrevem a função de pertinência na

forma trapezoidal. Estas informações são usadas para calcular os valores antecedentes(graus de pertinência), como mostrado na figura 5.16.

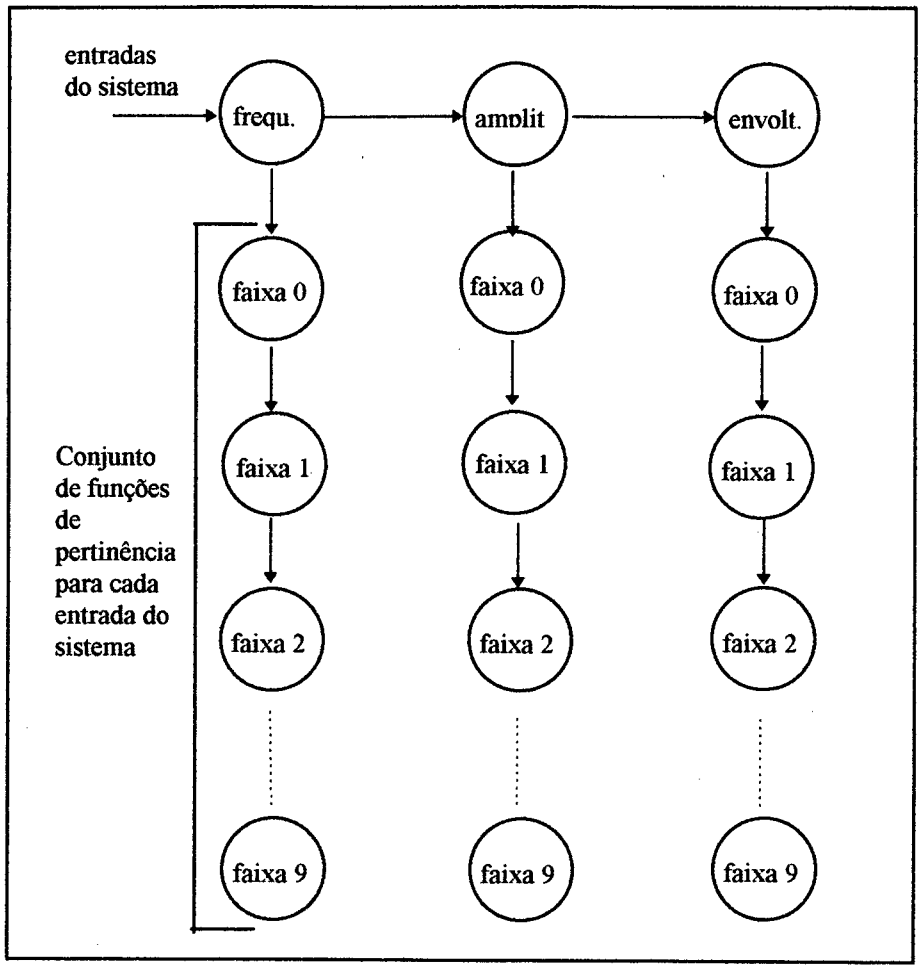

Figura 5.14. Arranjo da entrada de dados.
Fonte:Adaptado de [VIO93]

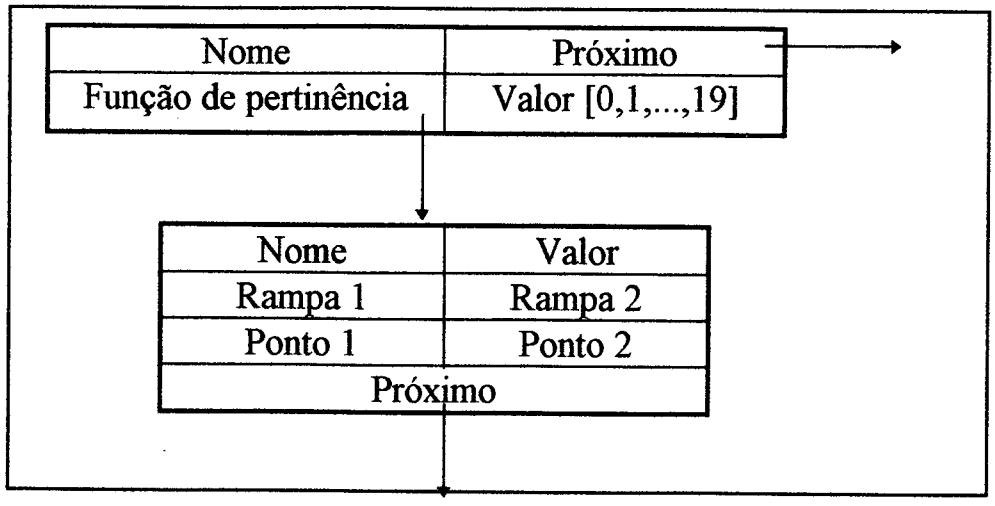

Figura 5.15. Estrutura de dados mais detalhada.
Fonte:[VIO93]

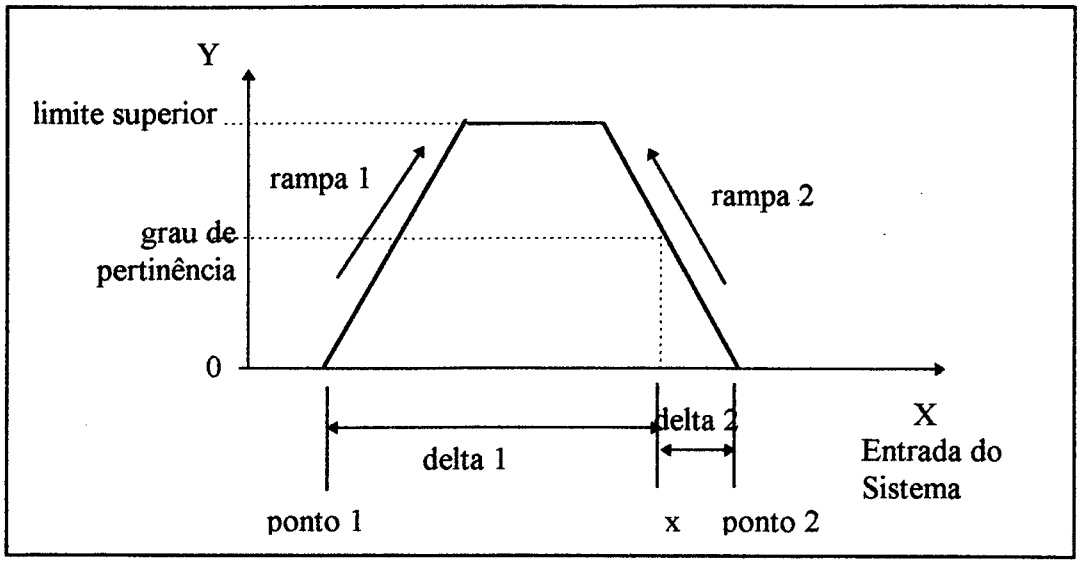

Figura 5.16. Cálculo dos graus de pertinência.
Fonte:[VIO93]

Para o cálculo do grau de pertinência são seguidos os seguintes passos:

1- cálculo dos termos delta:

delta1 = x - ponto 1
delta2 = ponto 2 - x

2. se (delta1 <= 0) ou (delta2 <=0) então
grau de pertinência = 0
senão
grau de pertinência = min(delta1*rampa1,delta2*rampa2,limite superior)

O valor antecedente resultante é armazenado no campo "valor" da estrutura função de pertinência.

As regras são representadas por dois conjuntos de ponteiros como mostra a figura 5.17. O primeiro conjunto indica que os valores antecedentes são usados para determinar a potência das regras, e o segundo conjunto aponta para locações de saída onde a potência deve ser aplicada.

A saída possui uma estrutura de dados semelhante a da entrada, representada na figura 5.14, diferindo apenas no fato de ser apenas uma saída e não três como acontece com a entrada.

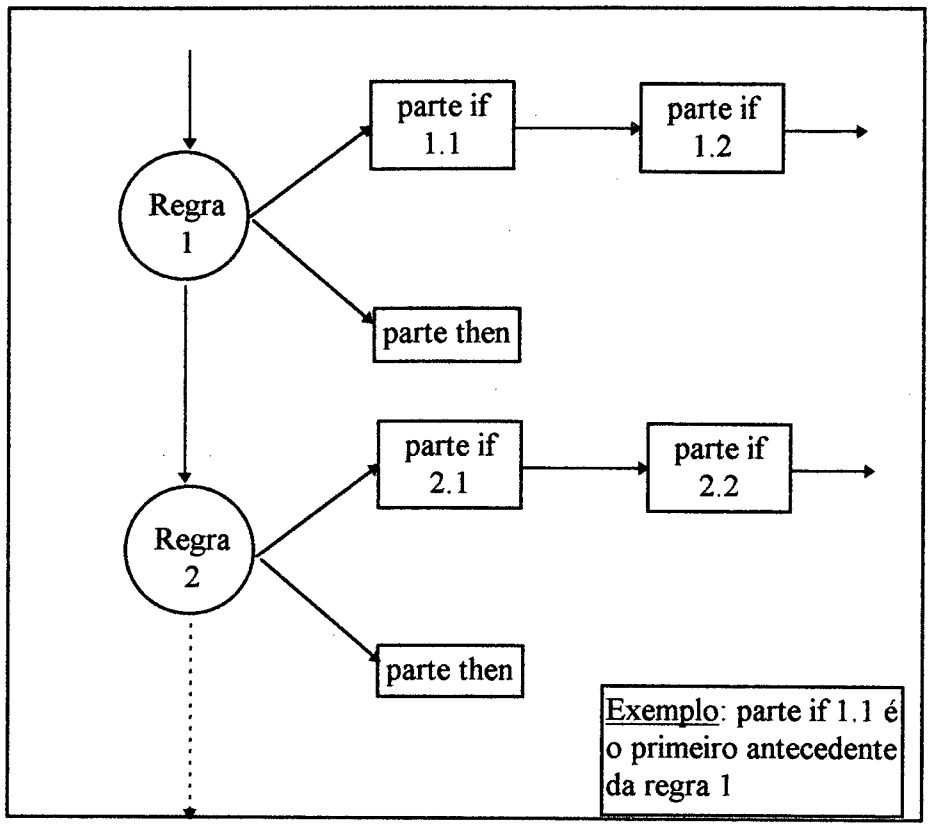

Figura 5.17. Estrutura da base de regras.
Fonte:[VIO93]

5.3.4 Fase 4: Testes, Ajustes e Resultados alcançados

Indispensável em qualquer sistema pois irá determinar o quanto o programa atingiu ou não seus objetivos. Nesta fase parâmetros maus ajustados poderão comprometer todo a performance do sistema. Portanto, descreveremos cada uma das variáveis do sistema que foram reguladas sempre olhando para a taxa de reconhecimento.

5.3.4.1 Formato das funções de pertinência dos conjuntos difusos de entrada

Segundo [ZIM90] uma das etapas mais demoradas de um sistema que utilize a lógica difusa, e que irá determinar sua performance, é a escolha dos formatos das funções de pertinência. Para chegar aos formatos ideais, fez-se: aplicou-se apenas uma entrada ao sistema, ou seja, só freqüência, só amplitude ou só envoltória. Em cima disso regulou-se cada um dos três conjuntos para aqueles valores com maior taxa de reconhecimento. A tabela II mostra as taxas alcançadas nestes testes.

Formato	Freqüência	Amplitude	Envoltória
0-10-0	59%	47%	51%
20-1-20	78%	59%	66%
30-1-30	82%	64%	70%
40-1-40	82%	66%	72%
30-10-30	82%	66%	71%
40-10-40	80%	68%	71%
30-20-30	80%	66%	70%
40-20-40	78%	67%	70%
20-30-20	80%	64%	69%
30-30-30	78%	66%	69%
40-30-40	77%	66%	68%
20-40-20	75%	65%	68%
30-40-30	73%	65%	68%
40-40-40	75%	65%	68%

Tabela II - Formato dos conjuntos difusos de entrada e taxas alcançadas.

Algumas palavras obtiveram índices de reconhecimento de 100% enquanto que outras tiveram índices de até 50 % nos formatos de melhor performance.

Baseado nestes testes, as funções de pertinência ficaram com os formatos mostrados na figura 5.18. Observamos que a freqüência tem um formato triangular enquanto que a amplitude e a envoltória possuem um formato trapezoidal. Outro fato que chama a atenção é o tamanho da subida e da descida em todos os conjuntos, um valor de 40 dentro de uma faixa que iria até 120 no máximo.

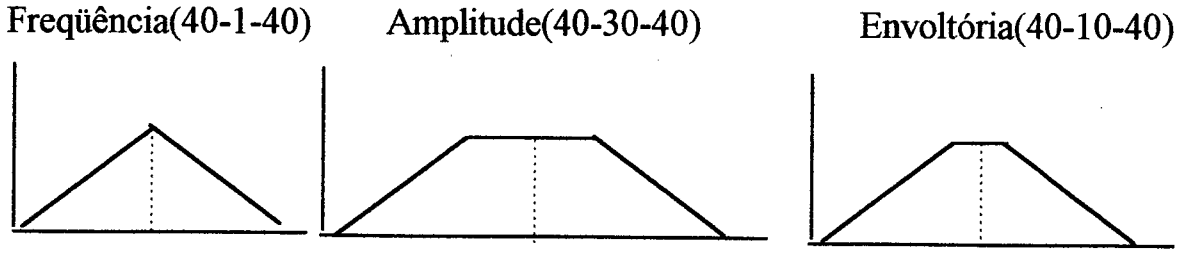

Figura 5.18 Formato das funções de pertinência dos conjuntos difusos de entrada.

Com estas três entradas e respectivos conjuntos difusos foi obtida uma taxa de reconhecimento de 89% e cada palavra ficou com um índice conforme a tabela III:

palavra	taxa
zero	100%
um	88%
dois	88%
tres	94%
quatro	61%
cinco	83%

palavra	taxa
seis	100%
sete	77%
oito	83%
nove	100%
mais	72%
menos	100%

palavra	taxa
vezes	83%
divide	94%
igual	100%
ponto	77%
cancela	100%
limpa	100%

Tabela III - Taxa de Reconhecimento de cada palavra do vocabulário.

Observa-se que a taxa de reconhecimento neste caso ficou entre 61% e 100%.

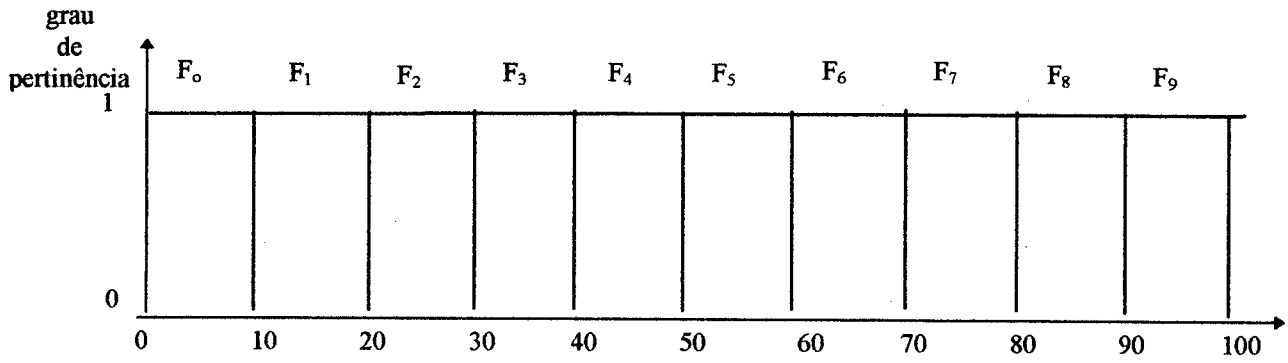

Figura 5.19. Conjuntos de entrada sem difusão

Outra observação que foi feita foi considerando que os conjuntos de entrada não tivessem sobreposição (conforme ilustra a figura 5.19), ou seja, não aplicar a força da lógica difusa que é o fato de uma entrada possuir um grau de pertinência em mais de um conjunto

Assim uma entrada de valor 21 estaria dentro da faixa F2 apenas e de forma alguma pertenceria aos valores do conjunto faixa F1 que iriam apenas até 20. Com estes conjuntos observou-se que a taxa de reconhecimento cai abruptamente para 43%.

5.3.4.2 Taxa de sobreposição de janelas

Outro parâmetro que foi regulado nesta etapa foi o índice de sobreposição de janelas, que também afetará o rendimento do sistema. A tabela IV mostra como ficou a taxa de reconhecimento para valores maiores ou menores de sobreposição, ou seja, o quanto os dados de uma janela farão parte das janelas adjacentes.

índice de sobrep.	taxa	índice de sobrep.	taxa
1.00	84%	1.32	89%
1.10	86%	1.34	87%
1.20	88%	1.40	87%
1.30	88%	1.50	84%

Tabela IV - Taxa de Reconhecimento para diversos valores de sobreposição de janelas.

Ao observar a tabela percebemos que o valor ideal do índice de sobreposição é de 1.32, ou seja, 32% dos dados de uma janela também fazem parte das janelas adjacentes. Pouca sobreposição ou muita reduzem a performance do reconhecimento.

5.3.4.3 Operadores de intersecção na avaliação das regras

Para Zimmermann [ZIM90] a escolha dos operadores para agregação dos conjuntos difusos pode ser confusa e difícil em um modelo específico ou situação. Para tanto ele sugere alguns critérios, entre eles:

a)comportamento apropriado em sistemas reais;
b)adaptabilidade ao contexto;
c)esforço computacional;
d)capacidade e faixa de compensação;
e)comportamento na agregação.

Seguindo estes critérios utilizamos alguns dos operadores citados em [ZIM90]:

a) mínimo:

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = min\left\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)\right\}$$

b) produto algébrico:

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = \mu_{\tilde{A}}(x).\mu_{\tilde{B}}(x)$$

c) diferença limitada:

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = max\left\{0, \mu_{\tilde{A}}(x) + \mu_{\tilde{B}}(x) - 1\right\}$$

d) produto de Hamacher:

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = \frac{\mu_{\tilde{A}}(x).\mu_{\tilde{B}}(x)}{\mu_{\tilde{A}}(x) + \mu_{\tilde{B}}(x) - \mu_{\tilde{A}}(x).\mu_{\tilde{B}}(x)}$$

e) produto de Einstein:

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = \frac{\mu_{\tilde{A}}(x).\mu_{\tilde{B}}(x)}{2-\left(\mu_{\tilde{A}}(x)+\mu_{\tilde{B}}(x)-\mu_{\tilde{A}}(x).\mu_{\tilde{B}}(x)\right)}$$

f) produto drástico:

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = \begin{cases} min\left\{\mu_{\tilde{A}}(x),\mu_{\tilde{B}}(x)\right\} se\ max\left\{\mu_{\tilde{A}}(x),\mu_{\tilde{B}}(x)\right\}=1 \\ 0 \quad \text{caso contrario} \end{cases}$$

g) intersecção de Hamacher:

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = \frac{\mu_{\tilde{A}}(x).\mu_{\tilde{B}}(x)}{\gamma+(1-\gamma)\left(\mu_{\tilde{A}}(x)+\mu_{\tilde{B}}(x)-\mu_{\tilde{A}}(x).\mu_{\tilde{B}}(x)\right)}, \gamma \geq 0$$

h) intersecção de Dubois e Prade:

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = \frac{\mu_{\tilde{A}}(x).\mu_{\tilde{B}}(x)}{max\left\{\mu_{\tilde{A}}(x),\mu_{\tilde{B}}(x),\alpha\right\}}, \alpha \in [0,1]$$

i) intersecção de Yager:

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = 1 - min\left\{1,\left(\left(1-\mu_{\tilde{A}}(x)\right)^{p}+\left(1-\mu_{\tilde{B}}(x)\right)^{p}\right)^{1/p}\right\}, p \geq 1$$

j) média aritmética:

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = \frac{\mu_{\tilde{A}}(x)+\mu_{\tilde{B}}(x)}{2}$$

k) média aritmética ponderada:

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = \frac{\mu_{\tilde{A}}(x).p1+\mu_{\tilde{B}}(x).p2}{p1+p2}$$

l) média geométrica:

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = \sqrt[2]{\mu_{\tilde{A}}(x).\mu_{\tilde{B}}(x)}$$

Para cada um deles obteve-se uma taxa conforme a tabela V.

Operador	**Taxa**	**Operador**	**Taxa**
Média geométrica	89%	Produto de Einstein	82%
Mínimo	87%	Diferença limitada	81%
Média aritmética ponderada	86%	Média aritmética	81%
Intersecção de Yager	85%	Produto de Hamacher	78%
Intersecção de Dubois e Prade	83%	Produto drástico	78%
Produto algébrico	82%	Intersecção de Hamacher	75%

Tabela V - Taxa de Reconhecimento para diversos operadores de intersecção.

Como se vê pelos resultados, a média geométrica registrou o melhor índice de reconhecimento de palavras: 89%, enquanto a intersecção de Hamacher chegou apenas a 75%.

5.3.4.4 Método de defusificação

Para obter o valor de saída que identificará uma das palavras do vocabulário como sendo a pronunciada, o programa toma como saída o valor médio do conjunto difuso com maior grau de pertinência.

Este grau de pertinência é alcançado tomando-se a média dos diversos graus de pertinência de cada janela, ou seja, de cada um dos vinte intervalos de tempo que são observados.

A escolha do tipo de média a ser usada foi feita também através de experimentação. A tabela VI mostra as médias testadas e os resultados:

Média	**Taxa**
Média Geométrica	89%
Média Aritmética	87%
Média Harmônica	87%

Tabela VI - Taxa de Reconhecimento para cada tipo de média na defusificação.

Mais uma vez a média geométrica adaptou-se melhor a esta finalidade, mesmo que por uma diferença de apenas dois pontos percentuais.

5.4 Operação do sistema

O sistema tem um menu principal com as seguintes opções:

1. Usuário
2. Treinamento
3. Modifica Conjuntos
4. Vê Conjuntos
5. Atualiza Regras
6. Mostra médias
7. Calculadora
8. Vê som
9. Identifica sons

Em que:

1. Usuário: Permite a identificação do Usuário.
2. Treinamento: Permite fazer o treinamento do vocabulário.
3. Modifica Conjuntos: Permite alterar o formato dos conjuntos difusos de entrada.
4. Vê Conjuntos: Permite visualizar o formato dos conjuntos difusos de entrada.
5. Atualiza Regras: Permite atualizar o arquivo de regras após alterações nos conjuntos.
6. Mostra médias: Mostra gráfico de médias dos valores de entrada de cada janela a partir dos dados de treinamento.
7. Calculadora: Permite utilizar a calculadora com interface por voz.
8. Vê som: Permite visualizar arquivos .WAV graficamente.
9. Identifica sons: Identifica sons .WAV pré-gravados.

O usuário após identificar-se com um código de três letras, necessita primeiramente passar por uma etapa de treinamento das palavras a serem pronunciadas. Este treinamento, que deve ser realizado em um ambiente com baixo índice de ruído, consiste da pronúncia em seqüência de todo o vocabulário, ou seja, das dezoito palavras. Cada palavra deve ser pronunciada três vezes. O programa fará o tratamento destes dados extraindo características que permitirão o reconhecimento das palavras ao final.

O sistema como já foi dito anteriormente é dependente do locutor e a taxa de reconhecimento cairá muito se o usuário utilizar dados de treinamento de outra pessoa.

As palavras pronunciadas ficam gravadas em disco com a extensão ".WAV" e podem ser reproduzidas em softwares que reconhecem este tipo de arquivo ou visualizadas neste próprio programa na opção "Vê Som".

Terminado o treinamento, o programa exige que as regras sejam atualizadas para que o arquivo de regras esteja adaptado aos novos padrões de voz, ou novos parâmetros. Esta opção na verdade poderia ser retirada do menu e ser realizada automaticamente pelo programa sempre que necessário, porém ficou presente para fins didáticos. Este é o mesmo motivo porque ficaram presentes as opções "Modifica conjuntos", "Vê conjuntos" e "Mostra médias".

Atualizadas as regras pode-se finalmente fazer uso da calculadora com interface de voz. Para tanto basta pronunciar os dígitos, operadores, ou controles como se estiverem sendo digitados via teclado.

CAPÍTULO VI

CONCLUSÃO

6.1 Conclusões sobre o trabalho

De acordo com palavras de Bill Gates da Microsoft citadas em [WIL98] "A fala será parte do sistema operacional" , ou seja, será possível comandar o Windows 95 por voz. Isto está mais perto graças a possibilidade de contar com máquinas de 600 Mhz, ou mais rápidas ainda, com até 64 K de RAM e discos rígidos de mais de 3 Gigabytes. Somente com a ajuda do desenvolvimento tecnológico do hardware será possível chegar em uma taxa de 100% (ou 99,9..%), no caso das interfaces de reconhecimento de voz. As barreiras vão sendo vencidas, o tratamento do erro aperfeiçoado, e a independência do locutor e da linguagem sendo atingidas gradualmente. A concretização desta possibilidade trará um grande incentivo ao uso de sistemas computacionais por pessoas que ainda tem receios ou dificuldades com o uso dos meios tradicionais de interfaces de entrada (mouse e teclado especialmente).

Durante as diversas etapas de desenvolvimento do trabalho foram comentados os principais tópicos relacionados a lógica difusa e ao reconhecimento de voz, buscando-se uma taxa de reconhecimento o mais alta possível, e atingiu-se 89% em condições favoráveis. Comparando-se com sistemas já existentes e citados no trabalho, esta taxa poderia ser melhor, porém o objetivo foi atingido, ou seja, utilizar e explorar a lógica difusa na tarefa de reconhecimento de voz. A busca por um aumento da taxa de reconhecimento poderia continuar, seja mudando alguns procedimentos, seja aplicando outros conceitos envolvidos com o som de voz. O trabalho não para aqui, espera-se continua-lo e aperfeiçoa-lo.

O conceito de *lógica difusa* foi aplicado e sem dúvida pelos resultados alcançados é possível utilizá-lo também nesta área de pesquisa, como já vem sendo feito. A dificuldade aqui ficou no sentido da comparação com sistemas de mesmo porte, com as mesmas limitações tecnológicas e que utilizam técnicas diferentes. Como o objetivo deste trabalho não foi implementar outros métodos, que por si só já seriam motivo de nova dissertação, as metas foram alcançadas dentro dos limites propostos.

6.2 Sugestões para trabalhos futuros

A utilização de técnicas de pré-processamento do sinal que permitam identificar e diferenciar as imagens acústicas não de uma palavra inteira, mas de um fonema desta palavra, seria uma das alternativas que pode-se citar como sugestão para trabalhos futuros. Os sistemas comerciais estão baseados nesta análise de fonemas, que na maioria das linguagens alcança centenas. Este fato gera um grande esforço computacional, o que, no caso de sistemas com vocabulário limitado a algumas dezenas de palavras, torna-se improdutivo. A sugestão neste caso seria trabalhar, a título de experimentação, apenas com os fonemas incluídos no vocabulário do sistema.

O estudo da probabilidade de uma palavra vir após ter sido pronunciada uma outra, aumenta a performance do reconhecimento. Em um vocabulário com vinte palavras, tem-se em principio a probabilidade de 1/20 de uma palavra ter sido pronunciada. Com a utilização de técnicas que levem em conta um nível de perplexidade diferente para cada palavra, esta probabilidade é reduzida consideravelmente.

O grande número de idiomas no planeta tem feito com que os fabricantes, na busca por uma fatia maior do mercado, forneçam versões de seus softwares em vários destes idiomas. As pesquisas atuais em sistemas de reconhecimento de voz tem se dirigido no sentido da independência do idioma. Isto ainda está longe de ser alcançado por sua complexidade e necessidade por velocidade de processamento e capacidade de armazenamento. Como sugestão, poderia-se implementar este mesmo sistema (calculadora) de forma a reconhecer as palavras em diversos idiomas. Na prática isto significaria apenas o aumento do vocabulário.

Técnicas de tratamento do erro poderiam ser implementadas de forma a aumentar a precisão do sistema. A repetição por parte do computador da palavra identificada e o tratamento de uma identificação incorreta de forma eficiente são sugestões neste sentido. Se o sistema quiser alcançar pessoas que não tem a visão, por exemplo, a repetição de forma audível das palavras pronunciadas e dos resultados dos cálculos é imprescindível.

REFERÊNCIAS BIBLIOGRÁFICAS

[ALL96] ALLEN, James. **The Next Step Beyond Speech Recognition**. Speech Technology Magazine. CI Publishing. Out/Nov.1996.

[ALL90] ALLEN, James. **Speech Recognition.** Encyclopedia of Artificial Intelligence, v. 2, p 1065-1069, New York, Wiley,1990.

[DAN95] DÂNGELO, José Geraldo, FATTINI, Carlo Américo. **Anatomia Humana Sistêmica e Segmentar - para o estudante de medicina.** Ed. Atheneu, 1995.

[FLE98] FLEMING, Peter, ANDERSEN, Robert. **Continuous Speech: Better Over Time**. Speech Technology Magazine. CI Publishing. Jan/Fev. 1998.

[FU82] FU, King Sun. **Syntatic Pattern Recognition and Applications.** Prentice-Hall, 1982.

[GOL94] GOLDBERG, D.E. **Genetic And Evolutionary Algorithms Come Of Age.** Communications of the ACM v.37, n.3 ,mar. 1994.

[GOM93] GOMIDE, F., GUDWIN, R., NETTO, M.L. Andrade, **Controle de Processos por Lógica Difusa**, Unicamp,1993.

[HUG95] HUGO, Marcel. **Uma Interface de Reconhecimento de Voz para o Sistema de Gerenciamento de Central de Informação de Fretes**. Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção. Professor orientador: Ricardo Miranda Barcia. Florianópolis, UFSC, 1995.

[JAN95] JANG, Jyh-Shing Roger. **Neuro Fuzzy Modeling and Control**. Proceedings of the IEEE, v.83, n.3, mar. 1995.

[KOM93] KOMORI, Yasuhiro, SHIGEKI, Sagayama, WAIBEL, Alexander H., **A Neural Fuzzy Training Approach for Improving Speech Recognition.** Systems and Computers in Japan, v.24, n.8, 1993.

[LI89] LI, Y.F., LAU, C.C.. **Development of Fuzzy Algorithms for Servo Systems.** IEEE Control Systems Magazine, p 65-72, jun.1989.

[MEN95] MENDEL, Jerry M. **Fuzzy Logic Systems for Enginneering: A Tutorial.** Proceedings of the IEEE, v.83, n.3, p.345-377, mar.1995.

[MOO94] MOORE, Martin L. **Sound Blaster: o livro definitivo**, tradução de Flávio Pareschi, Rio de Janeiro, Ed Campus, 1994.

[MOR95] MORGAN, Nelson, BOURLARD, Hervè A.,**Neural Networks for Statistical Recognition of Continuos Speech,** Proceedings of the IEEE, Vol. 83, No 5, Mai 1995.

[MUN94] MUNAKATA, Toshinori, JANI, Yashvant, **Fuzzy Systems: An Overview**, Communications of the ACM, Vol 37, no 3,Mar 1994.

[OLI96] OLIVER, Gaugarin. **Measuring ASR Success**. Speech Technology Magazine. CI Publishing. out/nov.96.

[PAL77] PAL, Sankar K. , MAJUNDER D.D. **Fuzzy Sets an Decionmaking Approaches in Vowel and Speaker Recognition.** IEEE Trans., Syst., Man, Cybern., ago. 1977.

[PIC93] PICONE, Joseph W. **Signal Modeling Techniques in Speech Recognition.** Proceedings of the IEEE, v.81, n.9, p.1214-1247, set. 1993.

[RAB81] RABINER, Lawrence R., WILPON, J.G. e ACKENHUSEN, J.G. **On the Effects of Varying Analysis Parameters on an LPC-Based Isolated Word Recognizer.**The Bell System Journal, v.60, n.6, jul.-ago. 1981.

[RAB89] RABINER, Lawrence R. **A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.** Proceedings of the IEEE, v.77, n.2, fev. 1989.

[RAO93] RAO, Valluru B., RAO, HAYAGRIVA, V. **C++ Neural Networks and Fuzzy Logic**, ed MIS, 1993.

[RAS95] RASH, Wayne Jr., HOLZER Elisabeth H. **Talk Show.** PC Magazine, p.71-84, fev. 1995.

[RIC93] RICH, Elaine, KNIGHT, Kevin. **Inteligência Artificial.** tradução de Maria Claudia S. R. Ratto, São Paulo, Makron Books, 1993.

[RID94] RIDGE, Peter M. et al. **O livro oficial de Sound Blaster**, tradução de Flávio D. Steffen, São Paulo, Makron Books, 1994.

[RUD94] RUDNICKY, Alexander I., HAUPTMANN, Alexander G., LEE, Kai-Fu. **Survey of Current Speech Technology.** Communications of the ACM , v.37, n.3, mar. 1994.

[SCH94] SCHWARTZ, Daniel G. **Applications of Fuzzy Sets and Approximate Reasoning**, Proceedings of the IEEE, vol 82, no.4, p. 482-498, abr. 1994.

[SPA94] SPANIAS, Andreas S. **Speech Coding: A Tutorial Review**, Proceedings of the IEEE, v.82, n.10, out.1994.

[TER94] TERANO, Toshiro, ASAI, Kiyoji, SUGENO, Michio. **Applied Fuzzy Systems**, Academic Press, 1994.

[VIO93] VIOT, Greg. **Fuzzy Logic in C.** Dr Dobb's Journal, p.40-49,fev 1993.

[WAN92] WANG, Li-Xin, Jerry M. Mendel. **Generating Fuzzy Rules by Learning from Examples.** IEEE Transactions on systems, man and Cybernetics, v. 22, n.6, p.1414-1427, nov-dez 1992

[WIL98] WILLIS, William. **Speech Recognition: Instead de Typing and Clicking, Talk and Command.** T.H.E. Journal, jan 1998.

[XUA93] XUAN ZHONG-, Yuan, CHONG-ZHI, Yu, YUAN, Fang. **Text Independent Speaker Identification Using Fuzzy Mathematical Algorithm.** IEEE, 1993.

[ZAD94] ZADEH, Lotfi A. **Fuzzy logic, Neural Networks, and Soft Computing.** Communications of the ACM, v.37, n.3,mar 1994.

[ZIM90] ZIMMERMANN, H.-J. **Fuzzy Set Theory - and its Applications.** Kluwer Academic Publishers, 1990.