UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Neilza Andréa de Oliveira

Reconhecimento de Fala utilizando Modelos Matemáticos e Redes Neurais

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação

João Bosco da Mota Alves

Florianópolis, Abril de 2002

Reconhecimento de Fala utilizando Modelos Matemáticos e Redes Neurais

Neilza Andréa de Oliveira

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação Área de Concentração Sistemas de Conhecimentos e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____

Prof. Fernando A. O. Gauthier , Dr. – Coord. do Curso

Banca Examinadora

_____

Prof. João Bosco da Mota Alves , Dr.

_____

Prof. Luiz Fernando Jacintho Maia, Dr.

_____

Prof. Marcelo Stemmer, Dr.

"Você sabe", admitiu o Diabo, "nem mesmo os melhores matemáticos de outros planetas, todos muito mais avançados que o seu, conseguiram resolvê-lo. Tem um sujeito em Saturno, ele parece um cogumelo sobre pernas de pau, que resolve mentalmente equações diferenciais, e mesmo ele desistiu."
(Artur Poges em "O Diabo e Simon Flagg", retirado do livro: O Último Teorema de Fermat)

Ofereço este trabalho aos pesquisadores da área de reconhecimento de Fala, pela nossa dedicação na procura de soluções, nossos projetos de pesquisa são uma pequena semente para um futuro melhor.

Gostaria de agradecer a todos que de uma forma ou de outra vem auxiliando em meus estudos e desbravamentos pessoais que estou fazendo para seguir dentro da pesquisa. Ao meu marido, obrigado pela sua compreensão pelas noites que venho dedicando a esses estudos e pelo apoio concedido. A minha filha Marjury, é também por você que estou fazendo isso.
Ao professor João Bosco Alves, obrigado pelo seu empenho e confiança depositados sobre a minha pessoa, espero continuar sempre retribuindo. Agradeço também a Universidade Federal de Santa Catarina pela oportunidade de ter realizado esta Pós-graduação.

SUMÁRIO

ÍNDICE DE FIGURAS

RESUMO

O reconhecimento de fala tem várias áreas de aplicação: tradução de textos, ditados, interfaces de computadores, serviços automáticos por telefone e aplicações industriais de propósito gerais. A principal razão para o sucesso dos sistemas de reconhecimento tem sido demonstrada pelo aumento na produtividade propiciada por estes, que assistem ou substituem operadores humanos. Esta dissertação tem como objetivo o desenvolvimento de um sistema de reconhecimento de fala.

As redes neurais artificiais surgem como o principal paradigma para o desenvolvimento destes sistemas, já que estas têm como principais características seu paralelismo, capacidade de treinamento, generalização, não linearidade e robustez. Essas vantagens são confirmadas através dos experimentos realizados neste trabalho, no qual comprova-se a importância das redes neurais artificiais para tais aplicações.

ABSTRACT

There are several application areas for speech recognition: translation of texts, dictation, computer interfaces, automatic telephone services and general purpose industrial applications. The success of the speech recognition system has been demostration by the increasing productivity provide by them, attending or substituting the human users or operators. This work is focused on the development of a speech recognition system.

The artificial neural network emerge as the main paradigm to the development of those system, since they have parallelism, trainability, generatization, nonlinearity and robustness characteristics. These advantages are confirmed through the experiments performed in this work, where they attest the importance of the artificial neural networks for applications of that kind.

1. INTRODUÇÃO

1.1 Visão Geral do Trabalho

Para podermos identificar um fonema com a utilização de modelos matemáticos, se faz necessário termos o universo que compõem a abrangência deste trabalho.

Fonologia é à parte da gramática que estuda e classifica os fonemas de uma língua.

Fonema é a menor unidade sonora capaz de estabelecer uma diferença de significado entre palavras da língua: ***p***ato(diferente)***m***ato(diferente)***b***ato

No caso do exemplo acima, /p/ /m/ e /b/ são fonemas.

Letra é sinal gráfico utilizado para representar os fonemas da língua. Em alguns casos são necessárias duas letras para representar um único fonema; diz-se que ocorreu um dígrafo quando isso acontece.

Não podemos pensar no processo de reconhecimento de Fonemas, sem antes entender a composição dos sons, para isto se faz necessário o entendimento de Processamento Digital de Sinais (DSP – Digital Signal Processing).

Processamento Digital de Sinais é uma das maravilhas tecnológicas que a engenharia produziu neste século vinte. Revoluções tecnológicas aconteceram graças a este campo: comunicação, imagens medicas, radar&sonar, alta fidelidade na reprodução de músicas, e outras. Cada uma dessas áreas tem sido desenvolvida com a utilização de algoritmos, matemática, processamento digital de sinais e técnicas especializadas. Os primeiros estudos na área de processamento digital de sinais ocorreram entre 1960 e 1970 [16].

Os sinais, de uma forma ou de outra, constituem um ingrediente básico de nossa vida diária. Por exemplo, uma forma de comunicação se desenvolve através do uso de sinais da fala, seja na conversação frente a frente ou por um canal telefônico. Outra forma comum de comunicação humana é visual por natureza, com sinais assumindo a forma de imagens de pessoas ou objetos que nos cercam. Sinal é formalmente definido como uma função de uma ou mais variáveis, a qual vincula informações sobre a natureza de um fenômeno físico [16].

O reconhecimento de fala é uma da área que se utiliza o processamento digital de sinais e é considerada uma das mais complexas áreas de pesquisa no estudo de sinais. O reconhecimento de palavras isoladas não é muito útil na pratica, sendo necessário

técnicas que possibilitem a identificação do início e término de cada palavra dentro do sinal da fala. Através do DSP temos que extrair as informações necessárias para podermos trabalhar com a identificação desta informação.

Mas a fala é um fenômeno muito mais complexo do que se normalmente se imagina. Cada locutor apresenta um conjunto de características que o diferencia dos outros, como: freqüência (determinada fundamentalmente pelo sexo), entonação (por razões regionais), o tom com que se pronunciam as palavras (evidenciando o estado emocional), etc. [32]

Para a extração das informações existentes dentro de um sinal, temos que trabalhar com: técnicas de filtragem de ruídos, Fourier, Wavelet, Laplace, Transformada Z, LPC, Cepstral, Segmentação, Janelas e outras. A cada dia surgem novas técnicas para resolução de problemas específicos [54].

Após a extração da informação pelos modelos matemáticos, existe a necessidade de identificarmos a natureza da informação, para isto se faz necessário o uso de redes neurais (Backpropagation, Kohonen, Hopfield e outras) ou modelos probabilísticos (Cadeias Ocultas de Markov) [54].

1.2 Objetivos

1.2.1 Objetivos Gerais

Reconhecimento de fonemas utilizando modelo matemático e um modelo de rede neural.

1.2.2 Objetivos Específicos

Para a realização deste projeto, se faz necessário o cumprimento de etapas seqüenciais, sendo cada etapa um obstáculo para a próxima, mas o caminho para a realização do mesmo.

[obj.01] Captura de sons analógicos para seu devido armazenamento em dispositivos de mídia magnética, transformando este sinal em digital. A captura será realizada através de microfones e armazenados em formato de arquivos WAVE.

[obj.02] Representar graficamente os sinais analógicos e digitais, para o seu devido entendimento. Nesta representação, tem que existir opções de configurações do esboço do sinal, para a visualização de uma parte específica.

[obj.03] A eliminação de ruídos através do uso de algoritmos de filtragem.

[obj.04] Aplicar algoritmos de modelos matemáticos para o trabalho com as ondas.

[obj.05] Aplicar a onda tratada em uma rede neural, e verificar o percentual de acertos do reconhecimento de fonemas.

1.3 Estrutura do Trabalho

Este trabalho esta dividido em: “O Reconhecimento da Fala”, “Conceitos Básicos de Sons”, “Redes Neurais”, “Materiais e Métodos”, “Resultados e Discussão”, “Conclusões” e “Anexos”.

O capitulo intitulado “Reconhecimento da Fala” apresenta uma visão do estado da arte na área de reconhecimento de fala e as dificuldades existentes. Este item traz uma visão do processo de produção da fala e do processo de audição da fala pelos seres humanos.

O capitulo “Conceitos Básicos de Sons”, apresenta uma visão de alguns dos principais modelos matemáticos existentes para o tratamento de sinais. Vários dos modelos apresentados foram usados neste trabalho.

O capitulo “Redes Neurais”, é mostrado algumas das redes neurais existentes, identificando suas vantagens e desvantagens.

O capitulo “Materiais e Métodos”, descrevem o processo de criação deste trabalho para que possa ser repetido por outros pesquisadores.

O capitulo “Resultados e Discussão”, nos traz os resultados obtidos e apresenta alguns itens que podem ser debatidos para a melhoria dos resultados encontrados neste trabalho.

O capitulo “Conclusões”, traz o fechamento do trabalho, mostrando de maneira sucinta informações pertinentes ao mesmo.

Nos anexos, são apresentados dois artigos gerados a partir deste trabalho e encaminhados a congresso.

2. O RECONHECIMENTO DA FALA

2.1. Breve Resumo Histórico [41]

Os estudos sobre reconhecimento da fala iniciaram-se faz aproximadamente cinco décadas. A partir de 1950 iniciaram-se os estudos de pesquisadores que procuravam explorar as idéias básicas da fonética acústica sem resultados satisfatórios.

Dentre os anos 1950 até 1959, os estudos foram liderados, basicamente por pesquisadores americanos e ingleses. Dentre as investigações mais relevantes destas décadas podemos mencionar os trabalhos de Olson e Belar, dos Laboratórios RCA e os trabalhos de Fry e Denes da Universidade de College, Inglaterra.

A partir de 1960, se inicia uma verdadeira explosão nas pesquisas de reconhecimento da fala e pesquisadores de outros países, principalmente Japão, se incorporam às pesquisas nesta área. Sem dúvida a técnica mais expressiva desenvolvida naquela época foi a análise de Cruzamento por Zero (Zero-Crossing Analysis) que conseguia distinguir entre várias regiões do estímulo de entrada, facilitando o reconhecimento.

Uma outra técnica, não menos importante, desenvolvida na década de 60, é a Programação dinâmica para alimentação temporal. Esta técnica foi desenvolvida pelo cientista russo Vintsvnc. A pesar que a base do conceito foi considerada rudimentar, as versões mais avançadas do seu algoritmo foram muito usadas na década de 80.

Em 1970, já existia um grande número de pesquisadores trabalhando na área. Os pesquisadores japoneses apresentaram a idéia da Codificação Preditiva Linear (LPC-Linear Predictive Coding) que tem sido amplamente usada. Por outro lado, nos Laboratórios Bell se iniciavam serie de experimentos com o objetivo de criar um sistema de reconhecimento de fala independente da locução.

Foi em meados da década dos 80 que surgiu a idéia de aplicar as Redes Neurais ao reconhecimento de padrões da fala. As RNA tinham sido introduzidas na década dos 50, no entanto seus resultados tinham sido considerados insatisfatórios devido a um conjunto de problemas operacionais, os quais foram sendo solucionados na medida que o conhecimento avançou.

A década de 80 foi à década onde a pesquisa no reconhecimento da fala apresenta melhores resultados. Como exemplo podemos mencionar o sistema desenvolvido pela

Agência de Projetos de Pesquisa Avançados para a defesa (DARPA – USA) que reconhece com precisão mil palavras em comunicação contínua.

Durante os últimos 10 anos tem se observado um progresso significativo nas tecnologias de reconhecimento da fala. As taxas de erro de palavras caem por um fator de 2 a cada dois anos. Os sistemas são cada vez mais independentes do locutor, a fala mais contínua e o vocabulário maior.

Existem vários fatores que tem contribuído para isto:

- Alcançou-se uma maturidade em ferramentas como modelos ocultos de Markov e as redes neurais artificiais.
- Estabeleceu-se padrões para a avaliação de desempenho. Anteriormente, os pesquisadores só podiam usar dados obtidos localmente e não se preocupavam com a escolha dos conjuntos de treinamento, o que provoca dificuldade na comparação de sistemas, e a degradação dos mesmos quando lhe eram apresentados dados nunca vistos. Hoje se tem grande quantidade de dados no domínio público, junto com especificações de padrões de avaliação.
- Os progressos nas tecnologias computacionais também contribuíram. A disponibilidade de computadores rápidos e mais baratos diminuiu grandemente o tempo de implementação das idéias.

Existem já em vários países sistemas de reconhecimento de fala em redes telefônicas e celulares. Outras realidades são os sistemas de ditados para a geração de documentos.

2.2. Os problemas do reconhecimento da Fala [86]

Para os seres humanos, reconhecer a fala é uma tarefa simples e bastante natural. Não se pode dizer o mesmo dos computadores. Fazer um computador responder a um comando falado é uma tarefa extremamente difícil e complexa. Muitos são os fatores variantes no reconhecimento da voz humana. Um dos grandes problemas é o tamanho do vocabulário, existe ainda a questão da palavra isolada e da palavra concatenada. Essa questão se refere ao fato de como exatamente as palavras estão sendo adquiridas pelo sistema. As palavras faladas isoladamente apresentam um espectro, quando pronunciadas de forma concatenadas apresentam outro. Esse processo, da palavra

concatenada, envolve um sistema de procura da palavra dentro de uma seqüência inteira de sons.

Outro problema é que as palavras raramente são pronunciadas da mesma maneira duas vezes. Essa mudança, por menor que seja, certamente afetará o circuito de decisão estabelecido pelo sistema. Ainda sobre fatores de lingüística, existe o problema *speaker-dependent* (depende-do-locutor) versus *speaker-independent* (independe-do-locutor). Esse fator é importante e precisa ser considerado. Um sistema tipo *speaker-dependent*, para ter uma boa performance necessita de um novo treinamento toda vez que mudar o orador do sistema. Os sistemas tipo *speaker-independent* possuem a vantagem de serem treinados para diversos tipos de oradores, entretanto, sua performance é menor que os sistemas *speaker-dependent*.

Todos esses fatores fazem com que as pesquisas de reconhecimento de fala permaneçam em contínua mudança. São esses os fatores que, por ora, impedem que se produza um sistema comercialmente perfeito e 100 % confiável.

Lembramos, porém, que os problemas apontados acima não incluem, em momento algum, processamento ou análise de sinal, seja analógico ou digital. Muitos outros problemas podem ser encontrados dependendo da técnica utilizada no processamento do sinal. O sinal, na maioria das vezes, precisa e deve sofrer algum tipo de tratamento (modificação), de modo a facilitar o seu uso e análise. Para realizar esses tratamentos, é necessário um conhecimento mais apurado sobre o som e a voz humana de uma maneira geral. A fase de aquisição de fala e processamento do sinal, impreterivelmente, antecede a fase de análise da voz.

2.3. A produção dos sons na linguagem humana [87]

Falar é tão natural para os seres humanos, como é o olfato, a visão e o paladar, que só nós possuímos para examinar seu funcionamento nos casos de deficiência ou de privação. No entanto, é essa capacidade de falar do modo como o fazemos que singulariza o homem de todos os outros animais.

É comum, ao falarmos sobre a linguagem, ter como ponto de referência à língua escrita. E muitas vezes, o estudo dessa faculdade distintiva da espécie humana fica reduzido ao estabelecimento das regras do bem escrever das quais se derivam às regras do bem falar. A linguagem é, porém, uma atividade primordialmente oral. A importância atribuída à língua escrita, importância essa que ocasiona até mesmo uma

inversão dos fatos, advém do papel capital que a escrita desempenha nas sociedades e de massa para a coesão política e social e para a comunicação a longa distância. A história dessas sociedades revela, contudo, que o uso difundido e sistemático da escrita é relativamente recente em comparação aos milênios de anos em que era privilégio de uns poucos ou aos vários milênios durante os quais nem mesmo existia. Ainda hoje há povos que nunca desenvolveram um sistema de escrita e as línguas por eles faladas em nada diferem em essência, das línguas faladas pelas populações letradas.

A linguagem humana se distingue dos demais sistemas simbólicos por ser segmentável em unidades menores, unidades essas em número finito para cada língua e que têm a possibilidade de se recombinarem para expressar idéias diferentes. O contínuo sonoro pode, pois, ser escandido em segmentos linearmente dispostos cuja presença ou ausência, assim como a sua ordem, tem uma função distintiva, isto é, ocasiona mudança no significado de uma palavra. Assim distinguimos ‘parte’ de ‘arte’ porque na primeira há um segmento p inexistente na segunda. Já em ‘Roma’ e ‘amor’ é a ordem dos segmentos que diferencia os dois vocábulos.

As unidades constitutivas do contínuo sonoro são produzidas por um mecanismo fisiológico específico a que se convencionou chamar aparelho fonador, e do qual fazem parte os pulmões, a laringe, a faringe, as cavidades oral e nasal. Observa-se que as partes constituem do aparelho fonador têm funcionamentos outros, distintos dos usados para a produção de sons. Assim os pulmões e a cavidade nasal têm um desempenho específico no processo de respiração, mas para a produção do som servem de câmara da corrente de ar, e a cavidade nasal funciona como câmara de ressonância para a produção dos sons nasais ou nasalizados. A diferença no funcionamento dos pulmões e das fossas nasais para as duas atividades – a respiração em repouso e a respiração para a fala – se evidência pelo fato de que a respiração em repouso há uma perfeita sincrônia entre a atividade dos músculos inspiratórios e o aumento do volume da cavidade torácica, atividade essa que cessa quando se inicia o movimento expiratório e conseqüente diminuição do volume torácico. Para a fala a atividade dos músculos inspiratórios contínuos na fase expiratória. Na respiração vital o ar sai pelo nariz e na respiração para a fala o ar sai pela boca. Os dentes e a língua são órgãos cruciais para a trituração dos alimentos mas na produção dos sons passam a articuladores que modificam a corrente de ar egressa dos pulmões.

Costumava-se, por isso, dizer que a linguagem é uma função secundária ou sobreposta, desempenhada por vários órgãos cujas funções biológicas primárias são de outra ordem. Essa perspectiva é, hoje em dia, ao menos polêmica por estar subjacente à teoria psicológica que considera a linguagem uma capacidade adquirida e não uma faculdade inata da espécie humana. Os argumentos em que se ancora a posição de que a linguagem é uma faculdade inata derivam dos mecanismos do tipo que vimos no parágrafo anterior: qualquer atividade que requeira uma sustentação do movimento inspiratório é penosa e arduamente aprendida, como por exemplo, nadar por debaixo d'água, tocar flauta, etc. Porém, uma criança começa a falar sem que jamais faça um treinamento específico para controlar esse mecanismo.

A finalidade última da linguagem é a comunicação. Um meio de representar esquematicamente o mecanismo da comunicação é imaginar uma fonte (o falante), um transmissor (o aparelho fonador), um canal (o ar atmosférico), um receptor (o aparelho auditivo) e um alvo (o ouvinte). Um ser humano tem algo a exprimir a outrem e para tal entra em funcionamento o seu sistema nervoso, impulsionando o aparelho fonador que opera sobre a informação a ser transmitida e a codifica em determinados padrões de ondas sonoras (a linguagem, o código, a mensagem). Essa operação é denominada codificação. As ondas sonoras, emitidas pelo falante, são conduzidas pelo ar atmosférico circundante indo atingir o aparelho auditivo do ouvinte (figura 1), que captura os sons convertendo as ondas sonoras em atividade nervosa que é levada ao cérebro. Essa operação é denominada decodificação. Está fechado o circuito e o processo pode repetir-se passando o ouvinte a falante. No estudo da faculdade de linguagem costuma-se imaginar uma mesma pessoa como fonte e receptora de um falante-ouvinte.

A produção dos sons é assim estudada de três ângulos diversos: 1) partindo-se do falante (da fonte) e examinando-se o que se passa no aparelho fonador; 2) focalizando-se os efeitos acústicos da onda sonora produzida pela corrente de ar em sua passagem pelo aparelho fonador ou, então, 3) examinando-se a percepção da onda sonora pelo ouvinte, isto é, o estudo das impressões acústicas e de suas interpretações no processo de decodificação.

A técnica mais difundida é a do exame da produção do som pelo aparelho fonador e registro de ouvido. Tal disciplina é denominada fonética articulatória ou fonética

fisiológica. Embora os dados proporcionados pela análise acústica sejam mais objetivos, a maior utilização da fonética articulatória se deve à relativa simplicidade com que pode ser aplicada, em contraposição à fonética acústica, a qual exige um aparelhamento mais dispendioso, pouco acessível em países em desenvolvimento, ao lado de um conhecimento de física, fato pouco comum aos estudiosos da área de letras e lingüística. Ademais, mesmo nos estudos em que se focalizam as propriedades físicas da onda sonora, que na sua produção, quer na sua percepção, os princípios de segmentação e as unidades depreendidas pela fonética articulatória estão presentes, tornando-se indispensável, portanto, o seu conhecimento.

O ser humano é capaz de produzir uma gama variadíssima de sons vocais. Porém nem todos eles são utilizados para fins lingüísticos de gerar, num enunciado, uma diferença de sentido por substituição ou por rearranjo. Por exemplo, o arroto, que é um som produzido com ar proveniente do esôfago, pode, em algumas culturas, exprimir plenitude após uma refeição. Mas em língua alguma funciona como um segmento na composição das palavras, formando com outros sons pares distintivos, como acontece, em português, na substituição do p de 'pata' por m em 'mata'. E mais, dentro do inventário de possibilidades usadas com fins fonológicos, cada língua seleciona apenas um subconjunto que utiliza com propósitos distintivos.

Assim a designação fonética articulatória tem dois sentidos. No seu mais amplo propósito é descrever qualquer som produzido pelos seres humanos; no mais restrito trata de esmiuçar os mecanismos existentes nas línguas humanas para comporem a enunciação. Beneficia-se da fonética experimental, isto é, de estudos que utilizam aparelhos como o oscilógrafo, o espectrógrafo, o sintetizador de fala, para um exame mais apurado da fisiologia acústica da produção dos sons.

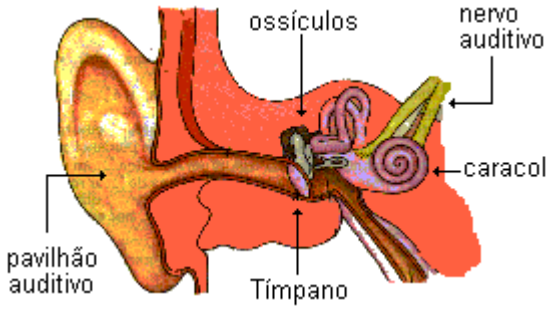


Figura 1- Ouvido humano

2.4. Sumário

Neste capitulo abordou o histórico do reconhecimento da fala, os problemas existentes nesta área de pesquisa, como funciona a produção da fala humana e como funciona a recepção deste som pelo ouvido humano.

No próximo capitulo iremos ver como podemos trabalhar com um sinal de fala utilizando modelos matemáticos, sendo os modelos apresentados usados neste projeto de pesquisa.

3. CONCEITOS BÁSICOS DO SOM [50]

O som é uma forma de energia caracterizada pela mudança de pressão, de natureza oscilatória, que se propaga em um meio físico, podendo ele ser gasoso, líquido, ou sólido.

O som tem dois tipos de características básicas: “intensidade” e o “timbre”, que podemos relacionar a um instrumento musical que pode emitir a mesma nota musical levemente ou com mais força. Pode emitir também duas notas diferentes com a mesma “força”.

O timbre está diretamente ligado ao número de vezes que a onda se repete a cada segundo.

Do ponto de vista do universo matemático existe uma grande semelhança entre o som e a imagem: o som é um sinal unidimensional enquanto a imagem é um sinal bidimensional.

A freqüência de amostragem é o número de vezes que o processo de amostragem é realizada em cada unidade de tempo. A unidade mais comum de medida de freqüência em engenharia e ciências aplicadas é o Hertz (Hz), que mede o número de ocorrências, oscilações, de um fenômeno, por segundo. Desta maneira podemos dizer que um som possui uma taxa de amostragem com 10.000 Hz, significa dizer que para cada segundo foram realizadas 10.000 operações de amostragem.

Um sinal de som digital é representado perfeitamente por um vetor.

Quando um som é descrito simplesmente através de uma função de atributos f definida sobre um intervalo [a,b] descrevendo a “pressão do ar” a cada instante de tempo (isto é chamado de representação temporal).

Uma interpretação interessante para a utilização de representações tempo X freqüência pode ser emprestada da álgebra linear: mudança de base.

Existem diversas transformadas que realizam uma transformação tempo X freqüências, dentre elas podemos citar a transformada de Fourier com janelas e a transformada de Wavelets.

3.1. Fourier

Análise de Fourier esta em uma família de técnicas matemáticas, todas baseadas na decomposição de um sinal em senóides. A transformada discreta de Fourier (DFT) é um membro da família usada com sinais digitalizados.

O estudo de sinais e sistemas, usando representações senoidais, é denominado análise de Fourier em homenagem a Joseph Fourier (1768-1830) por suas contribuições à teoria de representação de funções como superposições ponderadas de senóides. Os métodos de Fourier têm aplicações difundidas indo além dos sinais e sistemas; eles são usados em todos os ramos da engenharia e da ciência [16].

O que a transformada de Fourier faz?

A série de Fourier mostra que toda função periódica pode ser decomposta como uma soma (infinita) de funções periódicas (senos e co-senos). Essa decomposição torna fácil uma análise das freqüências de f: existe uma freqüência fundamental (w), e todas as outras freqüências são múltiplos inteiros (wj) dessa freqüência.

Em resumo, a transformada de Fourier transforma domínio temporal para o domínio de freqüência.

A localização de freqüência com a transformada de Fourier é impossível em geral. Isso se deve ao fato de que a função moduladora (exponencial) utilizada para medir a densidade de freqüência não possui suporte compacto: a integral que define a transformada se estende a toda a reta real. Portanto f(s) informa apenas que a freqüência s, ou freqüências próximas a s, estão presentes na função f.

A dificuldade em localizar freqüências no domínio do tempo, torna a transformada de Fourier uma ferramenta ineficiente para analisar funções [20].

A análise de Fourier é portanto mais eficiente no estudo de sinais que não sofrem variações muito bruscas ao longo do tempo. Esses sinais são chamados de estacionários.

O princípio da incerteza afirma: “Não podemos obter localizações precisas simultaneamente no domínio do tempo e da freqüência” [20].

3.2. Fourier com Janelas

Uma forma de tentar se obter a localização temporal de certos eventos espectrais é através do uso de uma função de janelas. Uma função de janelas tem como objetivo manter apenas uma região do domínio válida e eliminar tudo o que fica fora dela. Diversas funções podem ser usadas. Alguns exemplos clássicos usados em

processamento de sinais para o projeto de filtros com resposta finita são as janelas retangulares, hamming, hanning, Blackmann, Kaiser [20].

Após a multiplicação da função pela janela transformada, preserva-se apenas a região desejada e é então possível calcular a transformada de Fourier. Esse tipo de análise se chama Transformada de Fourier com Janelas.

Desse modo, a transformada de Fourier com janela fornece uma análise da função f no domínio tempo-frequência, no sentido de que temos informações localizadas tanto no domínio do tempo quanto no domínio da freqüência. Esse é um resultado na direção do problema que discutimos anteriormente: detectar freqüências em f e determinar sua localização no domínio temporal.

3.3. Janela Hamming

Funções de janela são sinais que estão concentrados no tempo, com freqüência de duração limitada. Enquanto funções de janela tal uma triangular, kaiser, barlett e prolate spheroidal ocasionalmente aparecem em sistemas de processamento digital de fala, as janelas retangulares, hanning e hamming são as mais usadas. Funções de janelas são também concentradas em freqüências baixas [2].

A função de janela hamming é definida como:

$$h_h[n] = \begin{cases} (1-\alpha) - \alpha\cos(2\pi n / N) & 0 \le n < N \\ 0 & em\ outros\ aspectos \end{cases}$$

Neste trabalho usamos a janela hamming para realizar um enquadramento do sinal da fala, para seu processamento por outros modelos matemáticos.

3.4. Transformada Z

O comportamento dos filtros no domínio da freqüência esclarece uma série de características que não são facilmente discerníveis apenas pelos seus coeficientes no tempo. Vimos que a transformada de Fourier é um recurso utilizado para estudar o sinal no domínio da freqüência. Nesta parte vamos introduzir uma transformada mais adequada, que na realidade contem a transformada de Fourier.

A Transformada-Z é uma generalização da Transformada de Fourier. A Transformada-Z de um sinal digital h[n] que é definida:

$$H(z) = \sum_{n=-\infty}^{n=+\infty} h[n]\, z^{-n}$$

Onde z é uma variável complexa. De fato, a Transformada de Fourier h[n] é igual à Transformada-Z $z=e^{jw}$ no cálculo. Normalmente usamos a transformada de Fourier para plotar a resposta da freqüência de um filtro e a transformada-Z para analisar as características gerais do filtro, dando uma função polinominal. Podemos usar a transformada-Z para entender os filtros, isto não é possível com a transformada de Fourier.

Neste trabalho a transformada-Z é usada para analisar um sinal e identificar os valores existentes num plano Z, estes valores são os fonemas existentes no sinal.

3.5. LPC – Predição Linear [2]

A predição linear tornou-se num dos métodos dominantes na estimação de parâmetros do sinal da fala, numa trama em que se considera o sinal estacionário. A idéia básica por detrás da predição linear é a de que o valor de uma amostra pode ser aproximado (predito), por combinação linear dos valores das amostras anteriores, tirando partido da correlação entre estas. Os coeficientes de predição linear ou coeficientes LPC (Linear Predictive Coding) são estimados por minimização do erro quadrático entre a amostra atual e a sua predição.

A analise de LPC é utilizada para encontrar os coeficientes da função de transferência do filtro que modela o sistema. Se o modelo é capaz de uma predição no sinal com um erro muito baixo, se tem que o LPC foi capaz de alcançar a informação necessária. Em analogia com um instrumento musical, LPC seria um instrumento de vento que soprado emite um som com um timbre particular a voz que representa.

O principio do LPC é que um valor de uma amostra de sinal de voz, s(n), pode se predito a partir de um numero finito de amostras anteriores: s(n-1),...,s(n-p), com um erro associado e(n) utilizando um filtro linear nos pólos:

$$s(n) = e(n) + \sum_{k=1}^{p} a_k \, s(n-k)$$

O erro da predição (também conhecido como seno residual), e(n), é simplesmente a diferença entre o valor atual do seno, s(n), e o valor predito, ?(n):

$$e(n) = s(n) \qquad ?(n)a_k$$

Os fatores que definem o peso a_k, sons encontrados e minimizados no erro quadrático, encontrados em N amostras (E):

$$E = (S \quad e^{2}_{i=0}^{n-1} (i)) / 2$$

A utilização da predição linear não se limita à codificação, sendo exemplos de outras aplicações no processamento de fala o reconhecimento e síntese e a identificação e verificação do orador. Exemplos de aplicações noutros campos são: a prospecção de petróleo através da análise da vibração da terra causada pela explosão de cargas de dinamite; o diagnóstico do cérebro através da análise de sinais do electroncefalograma; e a codificação digital de imagens.

Neste trabalho utilizamos LPC para extrair informações do sinal para o processamento utilizando a transformada Z.

3.6. Sumário

Neste capítulo foram apresentados conceitos fundamentais sobre reconhecimento de fala. Os métodos matemáticos utilizados neste projeto foram descritos, sendo na sua definição já apresentado o objetivo no escopo deste trabalho.

No capitulo seguinte serão apresentados alguns fundamentos de redes neurais artificiais, citando-se comparações entre os vários modelos existentes e um trabalho de reconhecimento de fala.

4. REDES NEURAIS

4.1. Introdução

Há alguns anos, a comunidade científica mundial tornou a reconhecer a importância da aplicação de técnicas de Inteligência Artificial (IA) na área de redes neurais, na resolução daqueles problemas que não seriam facilmente solucionados pela utilização de técnicas e ferramentas convencionais. As técnicas de IA recebem esta denominação pelo fato de tentarem "imitar" a maneira pela qual o homem soluciona problemas, especialmente problemas que a máquina, utilizando técnicas convencionais, tem dificuldade em processar, como *reconhecimento de padrões*, manipulação de variáveis qualitativas, etc.

Existem três abordagens básicas para a análise por Inteligência Artificial: os Paradigmas *Simbólico, Conexionista* e *Evolucionário*.[7]

A abordagem *Simbólica Baseada em Regras* (a qual é o tipo mais conhecido de abordagem simbólica, dentre os muitos existentes para este paradigma) procura definir um determinado sistema através de *regras de produção* simples (do tipo "SE-ENTÃO") elicitadas a partir do conhecimento de um especialista ou grupo de especialistas, para permitir a descrição das variáveis qualitativas e a tomada de decisões (com inspiração humana) sobre estas variáveis [7]. Por exemplo, para um sistema procedural e convencional baseado em computador seria muito difícil decidir-se por um ou por outro diagnóstico, da maneira como um médico (o *especialista* em questão) faria rotineiramente. Porém, se as regras para se chegar a este diagnóstico (exemplo simplificado: *'SE há febre E há dor-de-cabeça ENTÃO diagnóstico = gripe*") puderem ser definidas, o programa de computador poderá compilar todas estas regras, criando um *Sistema Especialista* e, em tese, chegar a, pelo menos, uma hipótese de diagnóstico para cada caso específico.

O Paradigma Conexionista, por sua vez, também procura imitar o discernimento humano, mas sem o estabelecimento de nenhuma regra verbal conhecida de condução da tomada de decisão. A abordagem conexionista procura reproduzir a estrutura do cérebro humano, que não precisa consultar regras de conduta, ou de produção (pelo menos, conscientes) a todo instante, para seguir um procedimento [7].

Apenas ocorre o "processamento" de informações de entrada, resultando em ações de saída intuitivas, na maior parte das vezes. Tanto isto ocorre que, por exemplo, é muito difícil para um médico experiente explicar em palavras *de que maneira* opta por um diagnóstico específico, frente a uma determinada condição clínica. Assim, a informação chega ao cérebro por meio das células nervosas dos olhos, por exemplo, e segue na forma de impulsos elétricos através de uma extensa e complexa rede de neurônios do córtex cerebral, até que se obtenha uma saída definitiva para os dados de entrada. O que acreditam os cientistas da área Neurológica, é que cada experiência vivida pelo ser humano resulta no estabelecimento de *conexões sinápticas* entre neurônios. Se as experiências forem se repetindo, as conexões são fortalecidas. Caso contrário, as conexões são enfraquecidas com o tempo. Este processo é chamado de *Aprendizagem* ou *Treinamento*. Assim, quando algum evento ocorre, um cérebro *treinado* age enviando "mensagens" ao longo de determinados "caminhos" pré-estabelecidos de neurônios interconectados, segundo a relação entre estes caminhos de conexão da rede e o próprio evento [7].

As Redes Neurais Artificiais, portanto, consistem de um ou mais conjuntos de "neurônios" artificiais interconectados, geralmente dispostos em "camadas" formando redes, que, uma vez devidamente *treinadas*, são capazes de processar matematicamente dados de entrada ao longo das camadas até que estes atinjam a camada de saída da rede, onde será disponibilizada a "resposta" da rede. Ou nas palavras de De Azevedo: "Redes neurais são sistemas complexos, constituídos por elementos representando algumas das características dos neurônios que constituem o sistema nervoso dos seres vivos e permitindo sua interação com o ambiente que os cerca" [7].

A Computação Evolutiva, ou Evolucionária, baseia-se em simulação, via programas de computador, do processo Evolutivo[1] que deu origem às espécies que hoje habitam o planeta Terra [7]. Simplificando, cria-se uma "população virtual" de "indivíduos" com características aleatórias distintas entre si (que lhe atribuem uma determinada "aptidão"). Estas entidades nascem, sofrem mutações, reproduzem e morrem; tal como qualquer animal; só que virtualmente. E após os "cruzamentos" entre

[1] Esta teoria foi desenvolvida por Charles Darwin e estabeleceu as bases para o que hoje chamamos de *Genética*. Darwin estudou o mecanismo natural que selecionou, ao longo dos tempos, as espécies mais "aptas" ou "adaptáveis" dentre todas as que existiram, e como estas foram capazes de transmitir seus genes às gerações posteriores, ao contrário de espécies mais "fracas" ou "menos adaptáveis", que se extinguiram.

estas entidades, o programa deve identificar quais dentre os novos indivíduos gerados seriam mais “resistentes”, ou “aptos”, a um determinado “meio-ambiente” e quais aqueles que por não possuírem as características necessárias à sobrevivência tenderão extinguir-se [7]. Desta forma, após várias “gerações”, ou iterações do programa, pode-se identificar as entidades “otimizadas”, ou que “evoluíram”. Este paradigma tem sido bastante aplicado, justamente, à otimização de processos. Atualmente, os algoritmos evolutivos também têm sido aplicados na otimização do treinamento de Redes Neurais Artificiais, o que permite um maior controle sobre o desempenho que a rede vai alcançar [7].

Uma quarta abordagem, que vem sendo muito utilizada nos últimos anos, é a Híbrida, ou seja, que utiliza dois ou mais paradigmas de Inteligência Artificial simultaneamente. Neste trabalho a abordagem Híbrida empregará os Paradigmas Conexionista e Simbólico, tal como vários trabalhos científicos têm feito, nos últimos anos. Esta abordagem oferece a vantagem de aproveitar as características mais fortes de cada paradigma, onde estas forem mais necessárias. Por exemplo, onde existirem variáveis relevantes e regras explícitas acerca destas variáveis, para uma tomada de decisão, pode-se utilizar uma abordagem simbólica. Já, por exemplo, onde não sejam conhecidas as regras de produção ou estas sejam muito complexas, pode-se aproveitar à abordagem conexionista. A interação entre ambas as abordagens deve ser feita segundo as características particulares de cada sistema e universo de análise.

4.2. Fundamentos Biológicos [23]

A criação de Redes Neurais Artificiais foi baseada nos neurônios biológicos (Figura 2) e nos sistemas nervosos. Camilo Golgi, em 1875, deu um dos primeiros passos na neuroanatomia com a descoberta de um método que possibilita isolar e observar de forma individual os neurônios através do uso de corantes. A partir do método de Golgi, Cajal apresentou dois resultados muito importantes: o primeiro foi à adoção da idéia inicial de sistema nervoso, que postulava que a comunicação entre as células era realizada através da sinapse (ponto de conexão entre neurônios). O segundo foi de que a interconexão entre neurônios não se dava ao acaso, mas sim de maneira altamente específica e estruturada.

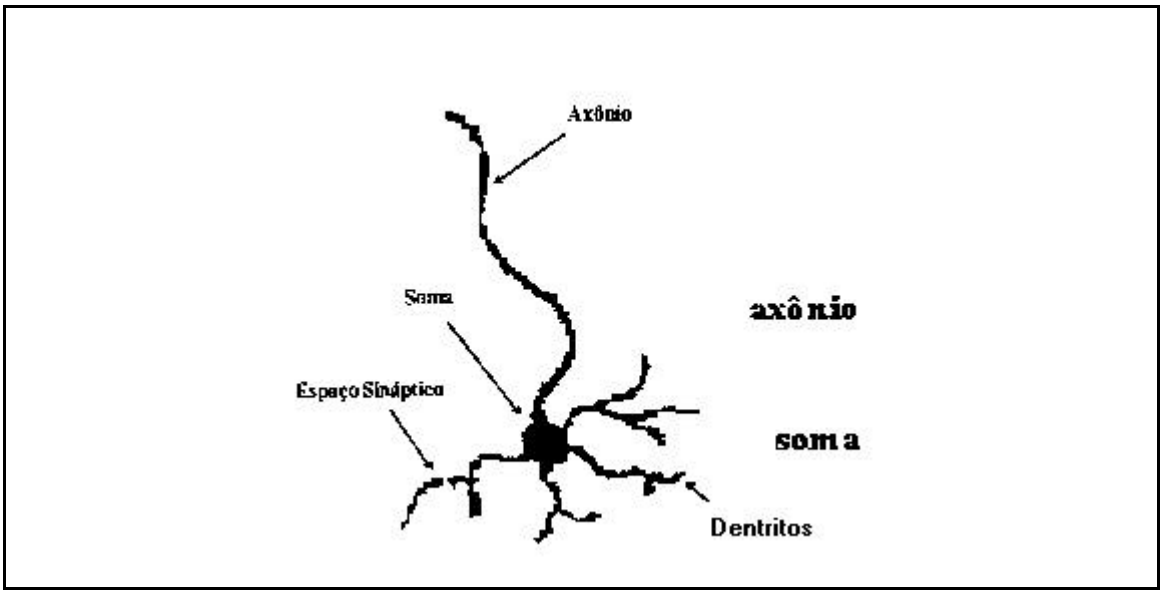


Figura 2- Neurônio Biológico

Atualmente, sabe-se que o neurônio tem um corpo celular denominado soma e variáveis, ramificações denominadas dendritos que são responsáveis por conduzir os sinais elétricos das extremidades para o corpo celular dos neurônios. As ramificações que tem por função transmitir o sinal do corpo celular às extremidades são denominadas axônios e geralmente, são uma só ramificação. O axônio é conectado aos dendritos de outros neurônios ou a outros axônios.

De acordo com James Weston, descobridor do DNA, o cérebro é a peça mais complexa do maquinário biológico existente na Terra. Com intuito de entender as funções do cérebro, neurobiologistas estudaram curvas características do estimulo-resposta de neurônios e, juntamente, redes de neurônios; paralelamente, psicólogos estudavam tais funções a nível comportamental e cognitivo. Ao longo dos anos, observou-se que estas duas abordagens vêm convergindo em seus estudos, ainda serão muito anos de pesquisa para o completo entendimento destes órgãos de aproximadamente 10^{11} neurônios, que recebe e envia sinais através de até 10 mil sinapses por neurônio [7].

4.3. Propriedades das Redes Neurais Artificiais [23]

A seguir serão citadas as principais propriedades que motivam a utilização das redes neurais artificiais na aplicação de sistemas de reconhecimento de fala.

- Paralelismo: as redes são altamente paralelas por natureza, assim, é recomendada sua implementação sobre computadores paralelos, permitindo maior rapidez no processamento dos dados.
- Capacidade de treinamento: a rede pode ser treinada para se adequar a qualquer padrão de entrada e saída. Isto pode ser usado, por exemplo, para uma rede aprender a classificar padrões de fala.
- Generalização: as redes não memorizam somente os dados treinados. Elas podem, a partir de dados treinados, generalizar seu processamento para novos padrões de entrada. Isto é essencial para o caso de sistemas de reconhecimento de fala, porque os padrões acústicos nunca são exatamente os mesmos.
- Não linearidade: as redes podem representar não linearidade, funções paramétricas de suas entradas, habilitando-as a desenhar arbitrariamente transformações complexas de dados. Isto é útil, já que a voz é um processo altamente não linear.
- Robustez: as redes são tolerantes a ambos os problemas: dados físicos e dados ruidosos; no caso de dados ruidosos as redes podem ajudar a formar melhores generalizações. Esta é uma valiosa característica, porque o padrão de voz é notavelmente ruidoso.

4.4. Aplicações, Vantagens e Desvantagens por Modelo [8]

Nome da Rede:	ADALINE/MADALINE
Ano de Publicação:	1960
Desenvolvedor(es):	B. Widrow
Aplicações básicas:	Filtragem de sinais adaptativos; equalização adaptativa
Vantagens:	Rápida, fácil de implementar, tanto em circuito analógico como VLSI
Desvantagens:	Assume relação linear entre entrada e saída desejada. Somente é possível classificar espaços linearmente separáveis.

Nome da Rede:	Adaptative Resonance Theory (ART)
Ano de Publicação:	1983
Desenvolvedor(es):	G. Carpenter & S. Grossberg
Aplicações básicas:	Reconhecimento de Padrões
Vantagens:	Capaz de aprender novos padrões, de novas categorias de padrões, e reter as categorias já aprendidas.
Desvantagens:	Natureza dos exemplos categóricos podem mudar com o aprendizado.

Nome da Rede:	Backpropagation Perceptron
Ano de Publicação:	1974-1986
Desenvolvedor(es):	P.J. Werbos, D. Parker, D. Rumelhart
Aplicações básicas:	Reconhecimento de padrões, filtragem de sinal, remoção de ruído, segmentação de sinal/imagem, classificação, mapeamento, controle robótico adaptativo, compressão de dados.
Vantagens:	Operação rápida. Boa em formar representações internas das características dos dados de entrada ou classificação e outras tarefas. Bem compreendida, com muitas aplicações de sucesso.
Desvantagens:	Tempo de treinamento longo.

Nome da Rede:	Recurrent
Ano de Publicação:	1987
Desenvolvedor(es):	Almeida, Pineda
Aplicações básicas:	Controle robótico, reconhecimento da fala, previsão do elemento seqüencial.
Vantagens:	Melhor tempo até agora para classificação, mapeamento de informações variando no tempo.
Desvantagens:	Rede complexa, pode ser difícil treinar e otimizar.

Nome da Rede:	Time-Delay
Ano de Publicação:	1987
Desenvolvedor(es):	D.W. Tank & J.J. Hopfield
Aplicações básicas:	Reconhecimento da fala.
Vantagens:	Desempenho equivalente aos melhores métodos convencionais, rápida operação.
Desvantagens:	Janela fixada para a atividade temporal representada, responde desastradamente para diferenças em escala na entrada.

Nome da Rede:	Rede de ligações funcionais
Ano de Publicação:	1988
Desenvolvedor(es):	Y.H. Pao
Aplicações básicas:	Classificação, mapeamento
Vantagens:	Somente duas camadas (entrada e saída) são necessárias, rápida para treinar.
Desvantagens:	Não é claro o modo de identificar funções adotadas para "ligações funcionais"

Nome da Rede:	Redes de funções de base radial
Ano de Publicação:	1987-1988
Desenvolvedor(es):	Múltiplos pesquisadores
Aplicações básicas:	Classificação, mapeamentos
Vantagens:	Uma rede com uma única camada oculta de neurônios é equivalente a rede Perceptron multicamadas básica com duas camadas ocultas.
Desvantagens:	Não é bem conhecida ainda

Nome da Rede:	Backpropagation de função utilidade no tempo
Ano de Publicação:	1974
Desenvolvedor(es):	P.J. Werbos
Aplicações básicas:	Maximiza o índice de desempenho ou a função utilidade no tempo; neurocontrole (robótica).
Vantagens:	Abordagem neural mais compreensiva para modelos de controle e/ou previsão.
Desvantagens:	Pode ser usado somente depois de identificado o modelo diferenciável, adaptação off-line se o modelo é dinâmico, e assume que o modelo é exato.

Nome da Rede:	BAM – Memória Associativa Bidirecional
Ano de Publicação:	1987
Desenvolvedor(es):	B. Kosko
Aplicações básicas:	Heteroassociativa (memória endereçada por conteúdo)
Vantagens:	Simples, regra de aprendizado, arquitetura e dinâmica clara. Prova clara da estabilidade dinâmica.
Desvantagens:	Capacidade de armazenamento e precisão de recuperações pobres.

Nome da Rede:	Boltzmann Machine, Cauchy Machine
Ano de Publicação:	1984, 1986
Desenvolvedor(es):	G. Hinton, T. Sejnowski, D. Ackley, H. Szu
Aplicações básicas:	Reconhecimento de padrões (imagens, sonar, radar), otimização.
Vantagens:	Capaz de formar representação ótima das características dos padrões. Segue superfície de energia para obter otimização no ponto mínimo.
Desvantagens:	A máquina de Boltzmann possue tempo de aprendizado longo, enquanto a máquina de Cauchy oferece aprendizado rápido.

Nome da Rede:	Brain-State-in-a-Box(BSB)
Ano de Publicação:	1977
Desenvolvedor(es):	J. Anderson
Aplicações básicas:	Revocação autoassociativa
Vantagens:	Possivelmente melhor desempenho que a rede hopfield
Desvantagens:	Incompleta exploração em termos de desempenho e aplicações em potêncial.

Nome da Rede:	Hopfield
Ano de Publicação:	1982
Desenvolvedor(es):	J. Hopfield
Aplicações básicas:	Evocação autoassociativa, otimização
Vantagens:	Conceitualmente simples, possui estabilidade dinâmica, de fácil implementação em circuitos VLSI.
Desvantagens:	Incapaz de aprender novos estados (pesos fixados para Hopfield discreta), armazenamento de memória pobre, pode estabilizar em muitos estados espúrios.

Nome da Rede:	Quantização do Vetor de Aprendizagem
Ano de Publicação:	1981
Desenvolvedor(es):	T. Kohonen
Aplicações básicas:	Revocação autoassociativa (complementação do padrão a um padrão parcial apresentado), compreensão de dados.
Vantagens:	Capaz de auto-organizar representações vetoriais de distribuições aleatórias em dados apresentados. Execução após treinamento completo.
Desvantagens:	Características não resolvidas na seleção do número de vetores usados e tempo de treinamento apropriado. Treinamento lento.

Nome da Rede:	Neocognitron
Ano de Publicação:	1975-1982
Desenvolvedor(es):	K. Fukushima
Aplicações básicas:	Reconhecimento de caracteres manuscritos e outras figuras.
Vantagens:	Capaz de reconhecer padrões independentes da escala, translação e rotação.
Desvantagens:	Requer muitos Eps e camadas, estrutura complexa, medida de escala para palavras é um problema ainda não resolvido.

Nome da Rede:	Mapas de preservação da topologia de auto-organização
Ano de Publicação:	1981
Desenvolvedor(es):	T. Kohonen
Aplicações básicas:	Mapeamentos complexos (envolvendo relações de vizinhança), compressão de dados, otimização.
Vantagens:	Capaz de auto-organizar representações vetoriais de dados com uma ordenação significativa entre as representações.
Desvantagens:	Características não resolvidas na seleção do número de vetores usados e tempo de treinamento apropriado. Treinamento lento.

4.5. Reconhecimento da Fala [8]

Um dos primeiros trabalhos de reconhecimento de fala foi produzido por Kohonen, onde foi implementado um reconhecedor de fonemas na língua finlandesa, este trabalho é apresentado abaixo.

Kohonem implementou uma “máquina de escrever fonética” para entradas de fala usando mapas de auto-organização para localizar e reconhecer em fala contínua para o idioma finlandês e o japonês. A rede sofreu um ajuste fino para precisão decisória ótima através de quantização de vetor de aprendizagem. Depois, num estágio de pós-processamento a gramática auto-aprendida foi aplicada para corrigir a maioria dos erros de co-articulação e deriva suas numerosas regras de transformação automaticamente a partir dos exemplos dados.

É conhecido que organismos sensoriais, como os encontrados no interior do ouvido, são usualmente hábeis em adaptar-se a sinais transientes de modo rápido e em caminho não-linear. Nesta aplicação, optou-se por aplicar análise de freqüências convencionais para o processamento da fala. A razão é que os fundamentos da filtragem digital são bem conhecidos, e a análise de Fourier digital é correta e rápida de ser efetuada.

Os detalhes do pré-processamento acústico estão abaixo relacionados:

- Uso de um filtro de passagem de banda baixa de 5,3 KHZ.
- Conversor analógico/digital de 12 bits com taxa de amostragem de 13,02 KHZ.
- Transformada de Fourier de 256 pontos formados a cada 9,83 ms usando janela Hamming.
- Logaritmação e alisamento da potencia espectral.
- Combinação de canais espectrais entre as freqüências de 200 KHZ a 5 KHZ em vetores padrões de 15 componentes.
- Subtração da média das componentes.
- Normalização dos vetores padrões.

O tipo mais simples de mapas de fala formados por auto-organização é o mapa de fonemas estáticos. Existem 21 fonemas no finlandês. Como mencionado nos detalhes técnicos, para a representação fonética foram usados espectros de tempo de 9,83 ms, cobrindo 256 pontos computados pela transformada de Fourier, de cuja potência espectral obtém-se vetores de 15 componentes, formados por argumentos de canais. Estes vetores, após os processamentos numéricos de subtração da média e normalizações constituíram as entradas da rede. No estudo presente todas as amostras espectrais, mesmo as de regiões transitórias, foram empregadas e apresentadas ao algoritmo na ordem natural de sua elocução. Após a adaptação, o mapa foi calibrado por ajuste fino usando fonemas estacionários.

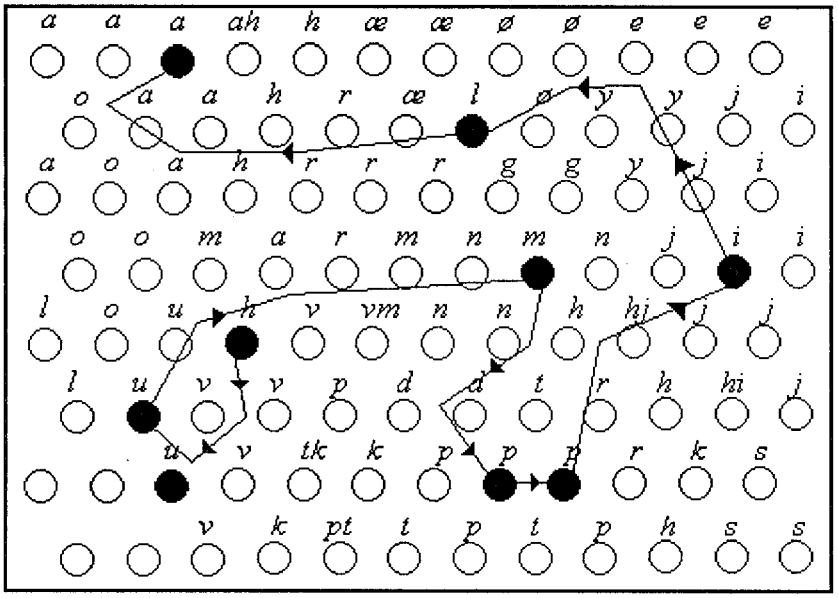


Figura 3 - O mapa ordenado de fonemas e a trajetória para a palavra *humppila* (unidades ativadas em preto)

Na prática, para locutores novos, 200 a 300 palavras são analisadas pelo método de segmentação automática. Na fase de treinamento supervisionado um conjunto de 2000 casos deve ser repetido ciclicamente ou randomicamente. O mapa para um locutor padrão pode ser modificado para outro locutor, usando apenas umas 100 palavras seguidas de ajuste fino.

4.6. Sumário

Este capítulo apresentou as redes neurais artificiais a partir de um resumo de seu histórico. Procurou-se esboçar uma idéia geral a respeito das propriedades das redes neurais artificiais e seu fundamento biológico. Alem disso foram apresentadas as aplicações, vantagens e desvantagens de vários modelos de redes neurais. Por fim, foi apresentado um projeto de reconhecimento de fala produzido por Kohonen, um dos primeiros projetos de pesquisa nesta área.

O capítulo seguinte apresentará o desenvolvimento do sistema de reconhecimento de fala, descrevendo todos os procedimentos adotados em sua implementação. Além disso, será descrita a aplicação desenvolvida.

5. IMPLEMENTAÇÃO DO SISTEMA DE RECONHECIMENTO DE FALA

Este projeto foi realizado através do estudo de livros, publicações científicas e informações disponibilizadas na internet, bem como o auxilio científico de professores e colegas da universidade, nos mais diversos departamentos.

Para a realização dos objetivos específicos deste trabalho foram realizados estudos em conformidade com os itens dos Objetivos Específicos.

- *Estudo 01:* Estudo de sinais para a sua captura e sistemas para o processamento deste sinal. O sinal é armazenado em arquivo WAVE, sendo necessário o estudo deste formato de arquivo. Através da utilização de software já disponíveis no mercado, poderemos ter a comprovação da validade das informações, verificando se as mesmas estão corretas.
- *Estudo 02:* Um sinal analógico pode ser representado por ondas senoidais, devendo a informação de entrada estar neste formato. O sinal dentro do sistema é tratado na forma digital, devendo o mesmo também ser representado. Faz-se necessário à possibilidade de marcar uma determinada região do histograma para o seu estudo mais detalhado, podemos verificar a validade da informação através de software existentes no mercado.
- *Estudo 03:* Com a configuração dos arquivos WAVE no formato de trabalho 8000 Hz, mono e 16 Bits, podem identificar soluções para realizarmos Filtragem e desta maneira eliminarmos os ruídos. Existem várias literaturas no mercado mostrando como utilizar e identificar os resultados esperados.
- *Estudo 04:* Com o uso de algoritmos buscamos o fonema e a sua representação gráfica através de histograma, através de softwares existentes no mercado verificaremos a validade da informação identificada.
- *Estudo 05:* Na realização de aplicação do modelo matemático, buscamos identificar o início e o término das palavras. O modelo matemático foi estudado e comparado seus resultados através de software e livros.
- *Estudo 06:* Nesta etapa identificamos o fonema através da utilização de redes neurais. Para a realização destes estudos utilizamos livros e orientação a profissionais da área fonoaudiologia.

Após o estudo dos itens relacionados e uma revisão no estado da arte existente em publicações de internet, publicações científicas , livros, dissertações de mestrado e teses de doutorado a visão do caminho para a realização deste projeto de pesquisa se tornou claro.

A primeira etapa a ser realizada foi à criação de arquivos de sons no formato desejado, sendo este formato de 8000 Hz, 16 bits mono. Foram gravados 300 arquivos no formato wave. Estes arquivos foram divididos em grupos de 10, contendo cada grupo as palavras: zero, um, dois, três, quatro, cinco, seis, sete, oito e nove. Este banco de dados de palavras foi produzido por três locutores, totalizando 100 arquivos waves por locutor.

A segunda etapa do projeto foi à escolha da ferramenta que se possibilita o desenvolvimento do projeto, nesta etapa, após testes utilizando linguagens de programação como C, C++ e Java, foi escolhido a ferramenta de simulações e desenvolvimento MATLAB.

A utilização do Matlab trouxe a facilidade da prototipação, permitindo criar e testar ferramentas de maneira muito rápido e amigável.

Com a utilização do Matlab, o primeiro modelo desenvolvido foi o de carga dos arquivos wave e as suas respectivas plotagem (visualização gráfica do sinal da fala), conforme figura 4. O próximo passo foi colocar o sinal da fala dentro de uma janela Hamming, permitindo um enquadramento do sinal, conforme a figura 5.

Uma vez o sinal enquadrado em uma janela Hamming, ele sofreu o processamento do modelo matemático LPC de tamanho 14, que permitiu a extração de informações necessárias do sinal (figura 6). Com os dados extraídos do sinal através da técnica de LPC, foi aplicado o modelo matemático denominado Transformada Z, que num plano Z identifica os pontos de convergência do sinal da fala, possibilitando uma visualização dos picos existentes no sinal, estes picos representam a informação buscada neste trabalho, que é o fonema. O Sinal após o processamento matemático da Transformada Z é reduzido para 250 pontos e cada ponto é calculado seu Log e multiplicado por 20 (figura 7). Esta redução para 250 pontos ocorre para que se possa treinar a rede neural com uma quantidade pequena de dados de entrada.

Os sinais após o processamento matemático apresentaram um índice de padronização que possibilita o treinamento da rede neural, os dez números uns

gravados, após o Janelamento, LPC, Transfromada-Z e Log, apresentaram graficamente as mesmas características, o mesmo ocorreu com os demais números, sendo humanamente possível identificar graficamente o sinal de cada palavra, desta maneira o trabalho chegou a ponto de utilização dos recursos de redes neurais.

Foi adotada a rede neural Backpropagation, pela sua facilidade de utilização e compreensão, sendo a sua estrutura criada contendo 250 neurônios de entrada, 100 neurônios ocultos e 10 neurônios de saída.

Os 250 neurônios de entrada foram definidos por serem exatamente o mesmo numero de pontos existentes no sinal da fala após o seu processamento matemático e 10 neurônios de saída por termos no nosso universo de palavras os numero de zero a nove, totalizando 10 tipos de palavras.

O treinamento foi utilizado a taxa de aprendizado 0.1.

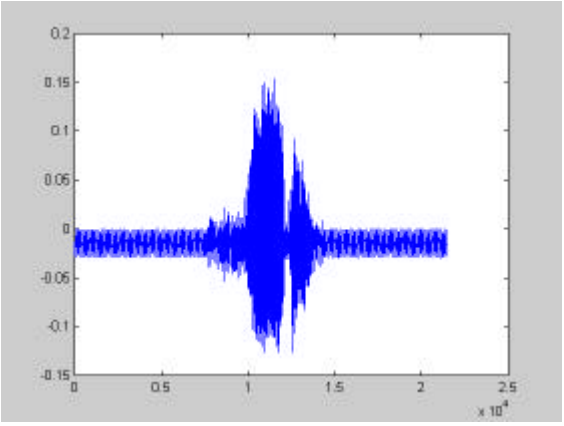


Figura 4 - Plotagem da sinal da fala no seus estado bruto

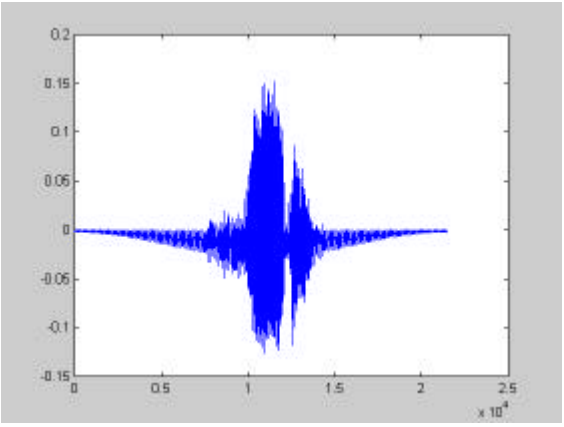


Figura 5 - Sinal plotado após o processamento de uma janela Hamming

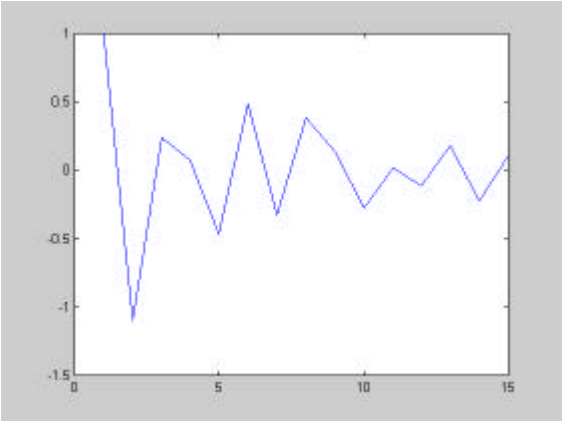


Figura 6 - Sinal da fala representado pelo processamento matemático LPC de tamanho 14

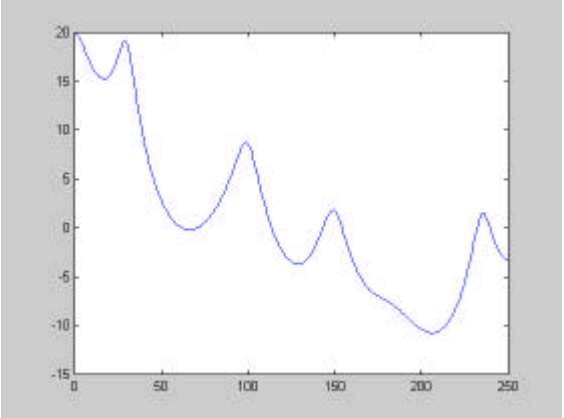


Figura 7 - Sinal após processamento da Transformada Z e calculado o Log

5.1. Sumário

Neste capítulo foram abordadas as questões referentes ao sistema reconhecimento de fala implementado, propriamente dito, onde apresentamos o ciclo da entrada do sinal, seu processamento matemático para extração de informações e o processo de treinamento da rede neural com as informações extraídas do sinal da fala.

No próximo capítulo irá apresentar alguns dos resultados obtidos neste trabalho através do modelo adotado.

6. RESULTADOS E DISCUSSÃO

Para o treinamento da rede neural, foram utilizados cinqüenta porcento (50%) dos arquivos de sons gerados e os outros cinqüenta porcento foram utilizados para a validação da técnica utilizada mediante aos resultados obtidos.

A rede neural Backpropagation obteve uma media de acerto equivalente a oitenta porcento (80%) dos arquivos apresentados a ela. Sendo o vinte porcento (20%) de erro em sinais que não podem ser distinguidos pelo ser humano na sua representação gráfica, mas podendo ser identificado pelo ser humano na sua execução no formato wave para a identificação do som contido nele.

O percentual de acerto não pode ser comparado a outros sistemas de reconhecimento de fala por ter utilizado um banco de dados de palavras próprio. Para uma comparação de acertos deste sistema a outros existentes se faz necessário a utilização de banco de dados de palavras como TIMIT, que é utilizado como referencia para mensuração de eficiência de sistemas de reconhecimento de fala.

A não utilização do banco de dados TIMIT neste projeto de pesquisa foi devido a falta de palavras em português no mesmo, sendo hoje o TIMIT composto apenas de palavras na língua inglesa.

Para testar o índice de acerto na rede neural, é introduzida na rede uma seqüência de dez sons já processados pelos modelos matemáticos apresentados, sendo estes sons referentes aos números de um, dois, três, quatro, cinco, seis, sete, oito, nove e zero e introduzidos na rede neural nesta seqüência.

Para a identificação dos acertos dos números apresentados a rede neural foi utilizada uma visualização gráfica dos resultados obtidos, conforme as figuras 8 e 9.

Cada linha das figuras 8 e 9 refere-se a um numero testado pela rede neural, mostrando graficamente numa escala de um a dez na sua base, onde estes números de um a dez representam os neurônios de saída da rede neural.

Os índices de acertos podem ser melhorados fazendo alguns acertos no processo de gravação das palavras, pela seqüência grande de palavras capturadas fica impossível à gravação das mesmas num único dia e sem interrupção no processo, como os jogos de palavras foram geradas em dias diferentes e em ambiente nem sempre adequado, o índice de ruído existente nos arquivos é grande, bem como a falta de padrão nos mecanismos de pronuncia das palavras e a posição do microfone.

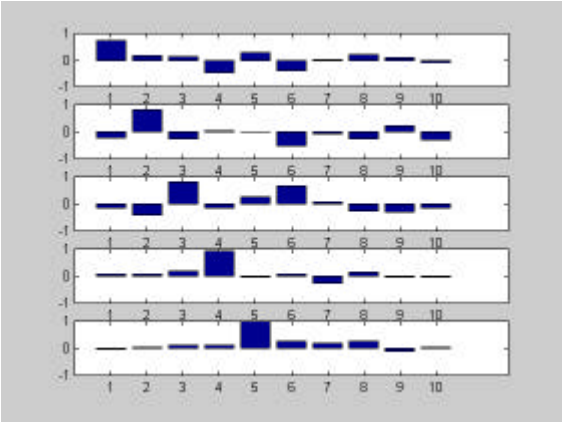


Figura 8 - Apresentação dos resultados dos números um ate cinco

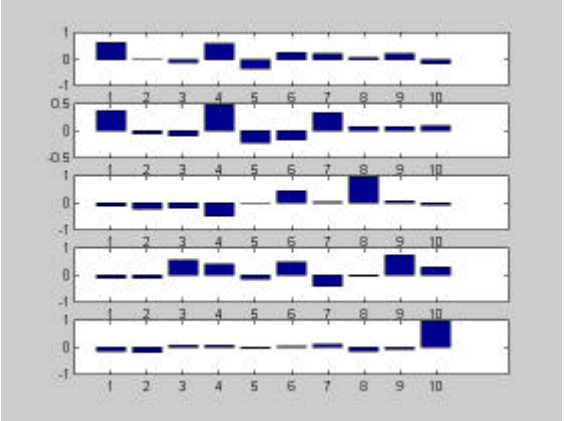


Figura 9 - Apresentação dos resultados dos números seis ate nome e o numero dez representa o numero zero

6.1. Sumário

Neste capitulo foram apresentados os resultados obtidos no presente trabalho, através da utilização da rede neural backpropagation e um banco de dados na língua portuguesa de palavras, este capitulo faz referencia ao banco de dados TIMIT que é da língua inglesa e que é utilizado como parâmetro de mensuração de acerto de sistemas de reconhecimento de fala.

O próximo capitulo apresentará as conclusões obtidas ao longo deste trabalho, bem como as perspectivas para trabalhos futuros.

7. CONCLUSÕES

Este trabalho teve como objetivo o desenvolvimento de um sistema de reconhecimento de palavras, utilizando, para isso, modelos matemáticos para o tratamento do sinal da fala e as redes neurais artificiais. Para a realização deste trabalho foi implementado um software que possibilitou a visualização passo a passo dos tratamentos dos sinais através dos modelos matemáticos, o treinamento da rede neural e os seus testes para a visualização gráfica dos resultados obtidos através das técnicas adotadas.

O sistema funciona apenas com sinais já gravados, mas as técnicas desenvolvidas através do software MATLAB podem ser facilmente implementadas em outra linguagem de programação para a sua utilização no momento que o sinal esta sendo gerado pelo locutor.

Este trabalho traz a sua contribuição para os pesquisadores da área de reconhecimento de fala, dando um passo pequeno, mas com um acréscimo no auxilio dos pesquisadores desta área, que não é nova, mas que tem muito a ser feito ainda.

7.1. Perspectivas de Trabalhos Futuros

Em relação às perspectivas de futuros trabalhos nesta área, ou como continuação do trabalho apresentado, podem ser citadas:

- O aumento do número de palavras identificadas pelo sistema;
- Variação da técnica de analise do sinal de fala, variação da arquitetura de rede, entre outros, a fim de avaliar e comparar o desempenho deste sistema;
- Testar outros algoritmos de treinamento de redes neurais;
- Testar outros modelos de redes neurais;
- Testar Modelos Ocultos de Markov;
- Testar técnicas de reconhecimento de fala continua.

8. ANEXOS

Anexo 1: Desbravando a Área de Reconhecimento de Fala.

Anexo 2: Reconhecimento de Fonemas utilizando Modelos Matemáticos e Redes Neurais.

Desbravando a Área de Reconhecimento de Fala

Neilza Andréa de Oliveira, UFSC, neilza@inf.ufsc.br,
Msc. Nelson Abu Samra Rahal Junior, UFSC, nelson_abu@das.ufsc.br,
Dr. Marcelo Stemmer, UFSC, marcelo@das.ufsc.br,
Dr. João Bosco Alves, UFSC, jbosco@inf.ufsc.br

Abstract. This work presents a global vision on necessary steps to speech recognition, presenting from state-of-art study to signal treatment details and hidden model Markov, showing some peculiarities among the approached items. On this presentation is shown the HTK software, an Open Source tool created by scientific community that allows the entrance of new research professionals on this area. The objective of this article is to explain each item trying to bring new research professionals to this area, which is not new, but there is still much to be done.

Resumo

Este trabalho apresenta um giro de 360 graus nas etapas necessárias para o reconhecimento de fala, apresentado desde o estudo da arte até detalhes de tratamentos de sinais e Cadeias Ocultas de Markov, mostrando algumas peculiaridades entre os itens abordados. Nesta apresentação é mostrado o software HTK, uma ferramenta Open Source criada pela comunidade científica e que permite o ingresso de novos pesquisados nesta área. O objetivo deste artigo é explanar sobre cada item tentando trazer novos pesquisadores para esta área que não é nova, mas que tem muita coisa a ser feita ainda.

1. Introdução

As interfaces de linguagem falada para computadores tem atraído e fascinada tanto aos engenheiros quanto aos cientistas da fala durante décadas. Além de ser um assunto polêmico por causa do desafio que implica. Interfaces de fala são cada dia mais uma necessidade, elas facilitam ao cidadão a comunicação com as novas tecnologias de redes, que ainda hoje, estão limitadas a uma pequena parte da população, inclusive nos países mais desenvolvidos.

Mas a fala é um fenômeno muito mais complexo do que normalmente se imagina. Cada locutor apresenta um conjunto de características que o diferencia dos outros, como: freqüência (determinada fundamentalmente pelo sexo), entonação (por razões regionais), o tom com que se pronunciam as palavras (evidentemente o estado emocional), etc.

Isto, assim como o alto grau de não linearidade da produção da fala e de variação na sua percepção, tem provocado que mesmo tendo havido um enorme progresso nos últimos dez anos, as máquinas ainda estão longe de poder reconhecer a fala convencional, sem limite no vocabulário e sem dependência do locutor.

2. Definição do problema

Reconhecimento da fala é o processo mediante o qual se converte o sinal acústico produzido pelo homem e capturado por um microfone ou telefone, em um conjunto de palavras. As palavras reconhecidas podem ser o resultado final do sistema. Este é o caso

das aplicações de comandos de controle, entradas de dados ou preparação de documentos.

Elas também podem servir de entrada a outros processamentos lingüísticos para conseguir a compreensão da fala.

Os sistemas de reconhecimento da fala, podem ser caracterizados e avaliados por vários parâmetros, como:

Parâmetros	*Faixa de Valores*
Vocabulário	Pequeno (< 20 palavras) a Grande (>20000 palavras)
Enrollment	Independente do locutor a Dependente do locutor
Modo de fala	Palavras isoladas e Fala contínua
Perplexity	Pequena (<10) a Grande (>100)
SNR	Alto (>30dB) a Baixo (<10dB)

Tabela 1 – Definição de parâmetros para um software de reconhecimento de fala

Vocabulário: Não existe uma definição estabelecida em relação a seu tamanho, mas geralmente se segue a faixa apresentada.

Enrollment: Diz-se que um sistema precisa de enrollment com o locutor quando este lhe deve fornecer amostras da sua fala antes de usá-lo. Se este não for o caso o sistema é dito independente do locutor.

Modo de fala: Em um sistema de reconhecimento de fala que trabalha com palavras isoladas é necessário que estas estejam separadas por pausas ou silêncios artificiais para que sejam reconhecidas. Isto não é exigido nos sistemas que trabalham com fala contínua. Estes são mais difíceis de implementar já que, neles, a pronúncia de uma palavra pode interferir na palavra subseqüente. Este efeito é chamado de coarticulação.

Perplexity: Esta é uma medida muito usada para medir a complexidade da tarefa,a qual constitui a média geométrica do número de palavras que podem seguir uma palavra em uma dada gramática. A gramática se encarrega de estabelecer regras e limites para a formação de uma sentença a partir de um vocabulário determinado. Sem a gramática, o número de possibilidades pode ser extremamente grande, mesmo com um vocabulário pequeno.

Condições adversas: Um sistema pode ser degradado por uma série de fatores, tais como: ruído ambiente, distorções acústicas, diferentes microfones, largura de banda, e maneiras alteradas de falar.

3. Estado da arte

3.1 Breve resumo histórico

Os estudos sobre reconhecimento da fala iniciaram-se aproximadamente em 1950, onde pesquisadores que procuravam explorar as idéias básicas da fonética-acústica sem resultados satisfatórios. Entre os anos de 1950 até 1959, os estudos foram liderados, basicamente por pesquisadores americanos e ingleses. Dentre as investigações mais relevantes desta década podemos mencionar os trabalhos de Olson e Belar, dos Laboratórios RCA e os trabalhos de fry e Denes da Universidade de College, Inglaterra.

A partir de 1960, se inicia uma verdadeira explosão nas pesquisas de reconhecimento da fala e pesquisadores de outros países, principalmente do Japão, se incorporam às pesquisas nesta área. Sem dúvida a técnica mais expressiva desenvolvida naquela época foi a Análise de Cruzamento por Zero (Zero-Crossing Analysis) que conseguia

distinguir entre várias regiões do estímulo de entrada, facilitando o reconhecimento. Uma outra técnica, não menos importante, desenvolvida na década de 60 é a Programação Dinâmica para alinhamento temporal. Esta técnica foi desenvolvida pelo cientista russo Vintsync. Apesar de que a base do conceito foi considerada rudimentar, as versões mais avançadas do seu algoritmo foram muito usadas na década dos anos 80.

Em 1970, já existia um grande número de pesquisadores trabalhando na área. Os pesquisadores japoneses apresentaram a idéia da Codificação Preditiva Linear (LPC - Linear Predictive Coding) que tem sido amplamente usada. Por outro lado, nos Laboratórios Bell se iniciavam uma serie de experimentos com o objetivo de criar um sistema de reconhecimento de fala independente do locutor.

Foi em meados da década dos anos 80 que surgiu a idéia de aplicar Redes Neurais ao reconhecimento de padrões da fala. As RNA tinham sido introduzidas na década dos anos 50, no entanto seus resultados tinham sido considerados insatisfatórios devido a um conjunto de problemas operacionais, os quais foram sendo solucionados na medida que o reconhecimento avançou.

A década dos anos 80, foi à década onde a pesquisa no reconhecimento da fala apresenta melhores resultados. Como exemplo podemos mencionar o sistema desenvolvido pela Agência de Projetos de Pesquisas Avançadas para a Defesa (DARPA - USA) que reconhece com precisão mil palavras em comunicação contínua.

3.2 Atualidade

Durante os últimos 10 anos tem se observado um progresso significativo nas tecnologias de reconhecimento de fala. As taxas de erro de palavras caem por um fator de 2 a cada dois anos. Os sistemas são cada vez mais independentes do locutor, as falas mais contínuas e o vocabulário maior.

Existem vários fatores que tem contribuído para isto:

- Alcançou-se uma maturidade em ferramentas como modelos ocultos de Markov e as redes neurais artificiais.
- Estabeleceram-se padrões para avaliação de desempenho. Anteriormente, os pesquisadores só podiam usar dados obtidos localmente e não se preocupavam com a escolha dos conjuntos de treinamento, o que provoca dificuldades na comparação entre sistemas, e a degradação dos mesmos quando lhe eram apresentados dados nunca vistos. Hoje se tem grande quantidade de dados no domínio publico, junto com especificações de padrões de avaliação.
- Os progressos na tecnologia computacionais também contribuíram. A disponibilidade de computadores rápida e mais barata diminuiu grandemente o tempo de implementação das idéias.

Existem já em vários países sistemas de reconhecimento de fala em redes telefônicas e celulares. Outras realidades são os sistemas de ditado para geração de documentos.

3.3 Modelos acústicos mais usados hoje

Durante a análise acústica a fim de analisar os quadros da fala, se necessita de alguns modelos acústicos. Estes diferem entre si por suas propriedades, além do modo como são apresentados. Entre as representações mais conhecidas temos:

Templates: é a forma de representação mais simples, na qual apenas uma amostra de unidade da fala a ser modelada é armazenada. Uma palavra ou fonema pode ser identificado através de sua simples comparação com templates conhecidas. Neste tipo

de apresentação, o principal problema é que como a variação acústica é muito grande, torna-se difícil modelar as unidades de fala. já que seriam necessárias múltiplas templates para cada unidade.

Estados: É uma representação mais usada do que o template por ser mais flexível. Está baseada no treinamento do modelo acústico, através da utilização de estados. Neste tipo de representação cada palavra é modelada por uma seqüência de estados treinados. Cada estado indica o som que provavelmente pode ser ouvido em aquele segmento de palavra (fonema), usando uma distribuição probabilística sobre o espaço acústico. Para a representação deste tipo de modelo acústico podemos utilizar modelos ocultos de Markov - HMM (Hidden Markov Models) e/ou redes neurais.

A maioria dos modelos acústicos hoje em dia estão baseados em Modelos Ocultos de Markov, uma estrutura estática capaz de suportar o modelamento acústico como temporal do sinal da fala. No entanto, as redes neurais artificiais também têm sido usadas por terem a capacidade de classificar.

3.4 Sistemas Híbridos

Com o objetivo de extrair parâmetros acústicos adicionais ou de explorar as vantagens associadas ao aprendizado de algoritmos como o backpropagation, existem os sistemas híbridos. Aqui, as redes neurais (ANN) são normalmente utilizadas para obter estes parâmetros e depois se integram na parte HMM do sistema.

No trabalho[2]: "Modelo Oculto de Markov Parametrização Através de uma ANN e sua Aplicação no Reconhecimento de Fala contínua", apresenta um sistema híbrido ANN-HMM ("Artificial Neural Network - Hiddem Markov Model") para reconhecimento de fala contínua, onde o modelamento acústico é confiado a ANN e o HMM responsabiliza-se pelo modelamento temporal, explorando, portanto, as principais vantagens de cada uma das estruturas. Este sistema foi treinado e avaliado a partir de uma base de dados dependente de locutor, consistindo de um total de 100 frases (totalizando 5,21 minutos de locução). Estas frases foram cedidas pelo laboratório de Fonética Acústica e Psicolingüística Experimental (LAFAPE) do Instituto de Estudos da Linguagem (IEL) da UNICAMP. O treinamento do sistema foi realizado em nível de fonemas (36 fonemas) e o teste foi realizado em nível de frases, apresentando uma probabilidade de acerto de 99%.

4. Processamento digital da fala

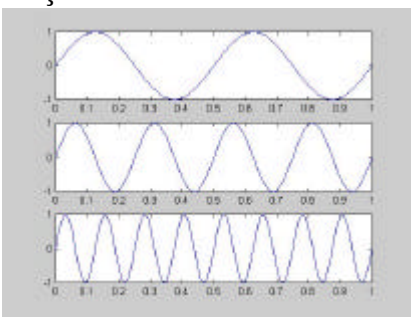
Os sinais se caracterizam pela sua freqüência, então podemos concluir que um bom começo para analisar uma função é iniciar pelo estudo de sua freqüência. freqüência de uma função é uma função periódica $f(t) = A\,\mathrm{sen}(2\prod \boldsymbol{w} t), A > 0$. O parâmetro A mede a amplitude da função (os valores máximo e mínimo assumido pela função). O parâmetro $\boldsymbol{w}$ é um número que indica quantos ciclos completos de período existe no intervalo [0,1]. Esse número está portanto diretamente relacionado com o número de oscilações da função na unidade de tempo, que é a freqüência da função. A figura 4.1 mostra o gráfico de f para diferentes valores de $\boldsymbol{w}$.

Se uma função é periódica de período L > 0, isto é, f(t + L) = f(t), a teoria de séries de Fourier nos ensina que ela pode ser escrita na forma $f(s) = \sum_{-\infty}^{\infty} a_j e^{i2\prod wjs}, a_j \in R$ onde $\boldsymbol{w}$

é uma constante. Essa decomposição de uma função periódica é chamada de série de Fourier.

A série de Fourier mostra que toda função periódica pode ser decomposta como uma soma (infinita) de funções periódicas (senos e cossenos). Essa decomposição torna fácil uma analise das freqüências de f: existe uma freqüência fundamental (w), e todas as outras freqüências são múltiplos inteiros (wj) dessa freqüência.

O coeficiente a_j na equação da série de Fourier mede a amplitude do componente de freqüência wj na função f. Em particular, se $a_j=0$, essa freqüência não se encontra presente na função. A amplitude a_j é dada pela equação $a_j \int_0^L f(u)e^{i2\Pi wju}du$, onde L é o período da função.



```
t = 0:1/100:1;
y1 = sin(pi*t*4);
y2 = sin(pi*t*8);
y3 = sin(pi*t*16);
subplot(3,1,1);
plot(t,y1);
subplot(3,1,2);
plot(t,y2);
subplot(3,1,3);
plot(t,y3);
```

Tabela 2 - Diferentes freqüências de uma senóide: item a) sen 4 Π t b) sen 8 Π t c) sen 16 Π t

O objetivo do processamento digital de sinais da fala é a representação digital dos sinais da fala, análise e extração de características e o desenvolvimento de modelos de síntese. Todas estas ferramentas são cruciais na implementação de sistemas de comunicação falada, seja esta comunicação a distancia ou comunicação homem-máquina. Tradicionalmente, estes sistemas estão divididos em sistemas de codificação, síntese, reconhecimento de sinais da fala e verificação e identificação do orador.

Podemos reproduzir graficamente o sinal da fala, utilizando a seguinte fórmula matemática:

- A variável *t* representara um vetor com valores no intervalo de 0 a 1, sendo incrementado por valores 0.001.
- A variável *y* representa a composição de ondas senoidais referente a variável *t*.
- Desta maneira temos y = sin(2*pi*50*t), onde 50 é o valor em Hz.

Sobre o valor de *y* que iremos utilizar as técnicas de processamento digital, técnicas como: Filtros Digitais, MFCC, LPC, Fourier, e outras. O resultado destas técnicas matemáticas de processamento digital de sinais, possibilita extrairmos informações para serem utilizadas nas redes neurais ou cadeias ocultas de Markov, para o reconhecimento da fala.

5. Modelos de Markov não observáveis [7]

A metodologia de reconhecimento da fala baseada nos modelos de Markov não-observáveis (HMM - Hiddem Markov Models) é uma das mais utilizadas. A teoria dos HMM foi publicada por Baum em 1966 [5], sendo a primeira aplicação em reconhecimento de fala proposta por Jelinek logo em 1969 [6]. No entanto, só a partir de 1975 estas aplicações começaram a ser reportadas com regularidade e só na década

de 80 apareceram na literatura um conjunto de artigos explicitando a teoria básica que permitiu que esta metodologia se tornasse tão popular.

5.1 Processo discreto de Markov

Considere-se um sistema que num determinado instante de tempo se encontra na estado i de entre N possível $S_1, S_2, \ldots, S_N$. A intervalos de tempo regulares o sistema evolui para outro estado ou eventualmente permanece no mesmo, em função de uma probabilidade de transição entre estados. Designaremos os diversos instantes de tempo por t=1,2,... e o estado no instante t por q. A descrição probabilística deste processo estocástico requer o conhecimento dos estados ocupados nos instantes passados, ou seja $P(q = S \mid q = S, q = S, \ldots) 1 \le i, j, k \le N$.

Num processo de Markov de primeira ordem a descrição probabilística é condicionada apenas ao estado no instante anterior, podendo ser representado através de uma matriz de transição entre estados A={a}, independente do instante de tempo, em que cada elemento é definido por: $a = P(q = S \mid q = S) 1 \le i, j \le N$.

Esta matriz verifica as restrições estocásticas de definição de probabilidades,nomeadamente: $a \ge 0, \sum a = 1$.

Como exemplo, considere o modelo de Markov com 3 estados, para descrever de um modo simplificado o estado do tempo. Neste modelo, cada estado correspondente à observação, uma vez por dia, das seguintes condições atmosféricas: estado 1: Dia chuvoso; Estado 2: Dia nublado; Estado 3: Dia com sol;

Assumindo que o estado do tempo num dia apenas depende do estado do tempo no dia anterior e que a matriz de transição de estados é dado por:

$$A = \{a\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}.$$

Obtém-se a seguinte cadeia de Markov.

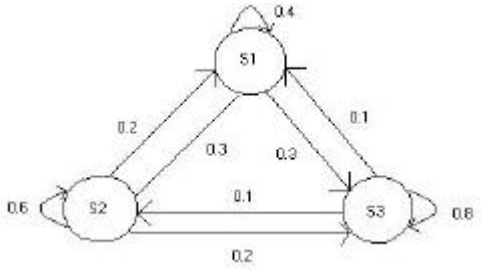


Figura 1 – Exemplo de Cadeia de Markov com 3 estados

Admitindo que o tempo num determinado dia é sol (estado 3), pode perguntar-se, por exemplo, qual a probabilidade de os 7 dias seguintes serem dias de sol-sol-chuva-chuva-sol-nublado-sol. Se definirmos a seqüência de observações Os correspondentes à seqüência de estados {S,S,S,S,S,S,S}, a probabilidade da seqüência O dado o modelo é dado por:

P(O|Modelo) = P(S|S) P(S|S) P(S|S) P(S|S) P(S|S) P(S|S) P(S|S)
= a a a a a a a
= 0.8x0.8x0.1x0.4x0.3x0.1x0.2
=1.536x10

O processo descrito é denominado de modelo de Markov observável, uma vez que a cada observação corresponde um estado. Este modelo é no entanto bastante restritivo e incapaz de ser utilizando em muitos problemas reais. Para tornar o modelo mais flexível, associa-se a cada estado uma função de distribuição de probabilidades e observações. Assim, cada estado pode gerar uma observação, de entre um conjunto, de acordo com esta distribuição. A mesma seqüência de observações pode assim ser gerada, com probabilidades diferentes, através de seqüências diferentes de estados. A seqüência de estados que gera uma seqüência de observações não é conhecida, pelo que este modelo se denomina de não-observável. Estes modelos encontram aplicações na solução de uma grande variedade de problemas.

Para ilustrar melhor este duplo processo estocástico, considere-se um conjunto de N urnas, cada uma com M bolas de cores diferentes. Estas urnas estão colocadas por detrás de uma cortina, apenas visíveis a um indivíduo que as manuseia. Estes indivíduos, de acordo com determinado processo aleatório, escolhem uma urna inicial e retira dela uma bola mostrando-a através da cortina. A cor é a única observação para quem está na sala. Seguidamente, a bola é colocada na urna de onde foi retirada e é escolhida outra urna através de um processo aleatório que depende apenas da última urna escolhida. Para este caso, os estados correspondentes às urnas escolhidas e as cores das bolas às observações, sendo a probabilidade de cada cor definida diferentemente para cada urna.

5.2 Aplicação dos HMM em reconhecimento de fala

No reconhecimento baseado em HMM, existem modelos probabilísticos das entidades do vocabulário a reconhecer. O reconhecimento é efetuado determinando a probabilidade da entidade a reconhecer, ter sido gerada por cada um dos modelos.

Para a construção de um reconhecedor de sinais de fala utilizando HMM, deve-se inicialmente construir um conjunto de modelos, um para cada classe de sons (fonemas, palavras, etc.) a reconhecer, através dos seguintes passos que constituem a fase de treino:

1. Definir o conjunto de classes de sons a reconhecer que corresponderá ao número L de modelos a treinar;
2. Escolher a topologia (o tipo de modelo, o número de estados e o número de observações por estado);
3. Obter, para cada classe, um conjunto com dimensão razoável de dados de treino;
4. Treinar os modelos utilizados, por exemplo, a reestimação de Baum-Welch.

Para o reconhecimento de um som, começa-se por extrair a seqüência de observações correspondente ao sinal da fala. Seguidamente é calculada a probabilidade da seqüência de observações, dada cada um dos modelos. Atribuí-se à seqüência de observações a som (classe) associado ao modelo que obteve a máxima probabilidade.

$$P(O \mid l) = \max P(O \mid l) 1 \le i \le L$$

O esquema de blocos do reconhecedor utilizando estes modelos é apresentado na figura 2.

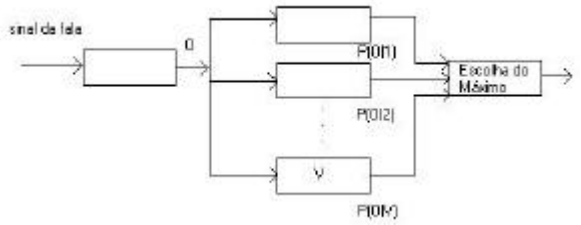


Figura 2 – Esquema de blocos de um reconhecedor de entidades isoladas

As características do sinal de entrada que servem como observações, obtidos trama-a-trama, são normalmente parâmetros espectrais derivados de LPC tais como os cepstrum, a energia, e as respectivas variações em relação à trama anterior (delta cepstrum e delta energia). sendo estes valores contínuos, é necessário proceder à sua quantificação vetorial, tornando-se um conjunto finito de símbolos, resultando numa degradação da percentagem de reconhecimento, a menos que se utilize um livro de código bastante grande. Outra solução para este problema é a utilização de modelos contínuos, onde as distribuições associadas às observações são caracterizadas por uma mistura de funções, densidade de probabilidades, normalmente com distribuição gaussiana: $b(O) = \sum c \aleph(O, m, U) 1 \le j \le N$ em que O é o vetor a ser modelado, c é o peso ou coeficiente da m-ésima mistura no estado S e $\aleph(O, m, U)$, com média m_{jm} e covariância U. A função densidade de probabilidade da equação $P(O \mid l) = \max P(O \mid l) 1 \le j \le N$ pode ser usada, desde que com o número suficiente de misturas, para aproximar qualquer função contínua. A equação da reestimação de bj(k) dada pela equação $b(O) = \sum c \aleph(O, m, U) 1 \le j \le N$ desdobra-se nas equações para reestimar c, m e U.

Nas aplicações dos HMM para o reconhecimento de fala, não se usam normalmente modelos completamente ligados mas sim modelos esquema-direto, ou seja, modelos em que de um estado Si só é possível transitar para o estado S, ou permanecer no mesmo estado, como mostra na figura 3.

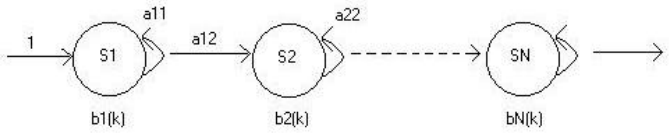


Figura 3 Modelo Esquerda – Direita normalmente utilizado nas aplicações de reconhecimento

Esta topologia traz algumas simplificações na estrutura dos parâmetros do modelo, que se explicitam seguidamente.

1. A distribuição de estados inicial Π tem apenas um valor não nulo (igual a 1), correspondendo ao estado S.
2. A matriz de distribuição das probabilidades de transição entre estados, tem, por cada linha, apenas dois valores não nulos. Os valores correspondentes a a_{ii}

e $a_{i(i+1)}$ $\begin{bmatrix} a & a & 0 & 0 \\ 0 & a & a & 0 \\ 0 & 0 & a & a \\ 0 & 0 & 0 & 1 \end{bmatrix}$ sendo o último estado um estado absorvente, só podendo transitar para si mesmo.

6. Softwares de reconhecimento da fala

	HTK	***Sphinx III***	***CSLU***	***ISIP***	***Spock***
Instituição	Microsoft & Cambridge Univ.	CMU	OGI	Mississippi State University	UCSD
URL	Htk.eng.cam.ac.uk	www.speech.cs.cmu.edu	cslu.cse.ogi.edu	www.isip.msstate.edu	speech.ucsd.edu/spock
Linguagem	C	C, Perl	C, Tcl/TK	C	Java
Plataforma	Unix e DOS	Linux FreeBSD , SunOS, HP/UX, Digital Unix, Windows NT/2000	Unix, DOS, Windows	Unix	Unix, Windows
Tecnologia	HMM	HMM	ANN & HMM	HMM	HMM
Recursos	Muito flexível, ultima versão suporta novos front end baseados em PLP e normalização vocal		Aplicação de rápido desenvolvimento (RAD) para criação de diálogos.	Open Source, fornece um extenso vocabulário	Fácil de usar
Objetivo	Grandes vocabulários				Restrito a palavras isoladas

Tabela 3 - O software HTK apresentou o menor índice de erro em comparação a seus concorrentes no workshop Broadcast News em 1998, sendo os resultados publicados no DARPA.

7. HTK

HTK é um kit de ferramentas para construção de um Modelo Oculto de Markov (HMM). HTK é primariamente construído para ser utilizado como uma ferramenta de reconhecimento de fala. O desenvolvimento desta ferramenta é realizada por: Speech Vision Robotics Group, no Departamento de Engenharia da Universidade de Cambridge (http://svr-www.eng.cam.ac.uk).

HTK é um software livre (free) que pode ser descarregado (download) pela internet, mas não pode ser redistribuído, pela necessidade de se registrar no site para realizar o processo de baixa do software (download).

8. IBM ViaVoice

O software IBM ViaVoice é um produto que tem como seus consumidores usuários domésticos e empresas. Este produto é de fácil instalação e configuração, e demostra como a tecnologia de reconhecimento de fala continua pode ser utilizada para fornecer benefícios e melhorias na qualidade de vida das pessoas. O IBM ViaVoice permite sua customização ao Sistema Operacional em uso, permitindo executar operações através dos comandos verbais, esta ferramenta também de adapta aos aplicativos de uso geral como editor de texto e navegador de web, permitindo o controle desses aplicativos através da pronuncia de palavras.

Para os desenvolvedores de software a IBM disponibiliza um Kit de ferramentas para customização dos aplicativos desenvolvidos com o IBM ViaVoice, permitindo que os desenvolvedores adaptem a tecnologia de reconhecimento aos seus softwares.

9. Trabalhos futuros

O HTK permite o uso exclusivo do Modelo Oculto de Markov, a ferramenta não traz em seu conjunto modelos de redes neurais (ANN) para testes ANN x ANN e ANN x HMM, para identificação do modelo de trabalho mais eficiente e eficaz. O HTK traz toda a estrutura necessária para o trabalho com arquivos de sons e seus códigos fontes estão disponíveis, o que permite adicionarmos os recursos para testes mais aprofundados.

Os resultados em publicações científicas demonstram o uso de sistemas híbridos com resultados extremamente satisfatórios, mas qual o melhor modelo híbrido para ser utilizado, no geral os pesquisados estão focalizando energias no modelo backpropagation e HMM, só os resultados testando os outros modelos de redes responderá esta incógnita.

10. Conclusões

O reconhecimento de fala com um vocabulário acima de 60 mil palavras e independente do locutor já é uma realidade, mas com custos computacionais elevados, como podemos manter a qualidade e diminuir o custo computacional? Hoje os computadores estão evoluindo cada vez mais rápidos, devemos deixar para a evolução do Hardware a solução dos problemas existentes hoje? Ficarmos satisfeitos com o modelo backpropagation x HMM pelos seus bons resultados não é uma estagnação na evolução das nossas próprias capacidades?

Essas são perguntas a serem respondidas, com certeza temos uma única resposta coerente, existe muito trabalho a ser feito nesta área de pesquisa, não só no processo de criação e testes de novas técnicas bem como na aplicabilidade das técnicas existentes, para a proliferação deste recurso que com certeza trará uma qualidade de vida superior ao ser humano.

11. Bibliografia

[1] Vitor Zue, Ron Cole. Survey of the State of the Art in Human Language Technology. http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html, 1997.

[2] Edmilson S. Morais, Fábio Violaro, Márcio L. Andrade. Modelos Ocultos de Markov Parametrização Através de uma ANN e sua aplicação no Reconhecimento de Fala Contínua. III Congresso Brasileiro de Redes Neurais, Florianópolis, Julho de 1997.

[3] Young, Steve; Evermann, Gunnar; Moore, Gareth. The HTK Books. Cambrigde University Engineering Department, December 2001.

[4] Z.L.Kovács. Redes Neurais Artificiais: Fundamentos e Aplicações. Edição Acadêmica, São Paulo, 1996.

[5] L.E.Baum, T.Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chain., Ann. Math. Stat. Vol. 37, pp. 1554-1563, 1996.

[6] F.Jelinek. A Fast Sequencial Decoding Algorithm using a Stack. IBM, J. Res. Develop., vol 13, pp. 675-685, 1969.

[7] Ribeiro, Carlos E. de Meneses. Processamento Digital de fala. ISEL-DEEC-SCPS, Abril de 2001.

[8] Gomes, Jonas. Wavelets Teoria Software e Aplicações. 21º Coloquio Brasileiro de Matemática, 21-25 Julho, 1997.

[9] RABINER, Lawrence R. A Tutorial on Hidden Markov Models and selected Applications in Speech Recognition. Proceeding of the IEEE, Vol. 77, No. 2, Bebruary 1989.

Reconhecimento de Fonemas utilizando Modelos Matemáticos e Redes Neurais

Neilza Andréa de Oliveira, UFSC, neilza@inf.ufsc.br,
Msc. Nelson Abu Samra Rahal Junior, UFSC, nelson_abu@das.ufsc.br,
Dr. Marcelo Stemmer, UFSC, marcelo@das.ufsc.br,
Dr. João Bosco Alves, UFSC, jbosco@inf.ufsc.br

Abstract. This work presents the necessary mechanisms to phonemes identification using mathematical models and phoneme recognition using neural nets. It will be presented an algorithm used to speech recognition research and obtained results.

Resumo

Este trabalho apresenta os mecanismos necessários para a identificação de fonemas utilizando modelos matemáticos e o reconhecimento do fonema utilizando redes neurais.

Será apresentado um algoritmo utilizado para a pesquisa na área de reconhecimento da fala e os resultados obtidos.

Introdução

Os problemas existentes nos projetos de pesquisas na área de reconhecimento da fala englobam desde a definição do numero de palavras (sistemas pequenos ate 100 palavras e sistemas complexos acima de 1000 palavras), dependência e independência do locutor (sistemas personalizados a um usuário ou a vários usuários), problemas regionais (entonação), estado emocional do locutor (tom) e sexo do locutor (freqüência). Estas peculiaridades do sinal da voz, tem que produzir um modelo de representação para que se possa ser utilizado em uma rede neural ou num modelo probabilístico como as Cadeias Ocultas de Markov.

Um modelo representativo das ondas acústicas baseadas em Fourier que traz a representação da freqüência não permite um treinamento de uma rede neural com um volume grande de palavras, pois a rede não consegue identificar um padrão nos dados apresentados para o treinamento, e quando consegue a identificação das palavras apresentadas a rede neural não traz um índice de acerto adequado, isto graças ao modelo representativo das freqüências.

Através da utilização do modelo matemático LPC (coeficiente da predição linear) e da FreqZ (Transformada-Z de uma resposta de freqüência de um filtro digital) conseguimos identificar em um sinal acústico o ponto aproximado da existência de um fonema, possibilitando a criação de um vetor de dados que representa o sinal da voz, treinando a rede neural com este vetor e permitindo a convergência da rede no seu treinamento.

Configurações adotadas

Este projeto foi executado com as seguintes características:

Banco de dados de arquivos de sons no formato wave gravados em 8000Hz/16Bits/Mono.

Cada arquivo wave possui um palavra, sendo a abrangência do conjunto de palavras de números de zero a nove (zero, um, dois, três, quatro, cinco, seis, sete, oito e nove).

Cada palavra foi gravada 10 vezes, totalizando um banco de dados de 100 arquivos de sons.

Rede neural Backpropagation com 250 neurônios de entrada, 100 neurônios ocultos e 10 neurônios de saída, ficando 250x100x10.

50% dos arquivos de som foram utilizados para o treinamento e os outros 50% para os testes de reconhecimentos.

Software utilizado Matlab 5.3.0 (R11).

Tratamento matemático para os arquivos de som

Inicialmente o sinal é processado com a janela hamming de tamanho igual ao do sinal de entrada e o resultado é calculo pelo modelo matemático Lpc com tamanho da ordem igual a 14. O resultado do Lpc é processado pela transformada-Z (freqZ). O resultado é reduzido para 250 pontos, que são introduzidos na rede neural.

```
1Carga do arquivo wave
[n00_01 f00_01] = wavread(['8000Hz16bits\zero01']);
2Tamanho do arquivo
t00_01 = length(n00_01);
3Janelamento do sinal
j00_01 = hamming(t00_01).*n00_01;
4Fourier
A00_01 = fft(j00_01, f00_01);
5Lpc
[a00_01 ,g00_01] = lpc(j00_01,14);
6FreqZ
H00_01 = freqz(g00_01,a00_01,f00_01/2);
7Reduz o sinal para 250 pontos
H00_01r = diminuiWave(H00_01,16);
8Função de redução do sinal
function retorno = diminuiWave(vetor, tamanho)
echo off;
clear y;
ncount = 1;
for x = 1:length(vetor)
  [r] = mod(x,tamanho)
  if (r == 0.00)
    y(ncount,1) = vetor(x);
    ncount = ncount + 1;
  end
end
retorno=y;
```

Tabela 4 - Algoritmo para identificação dos fonemas

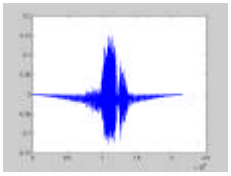
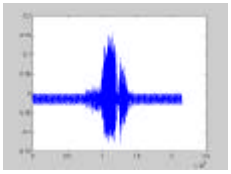
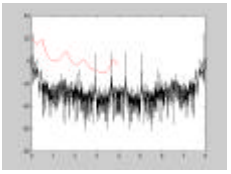
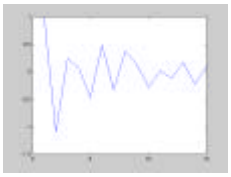
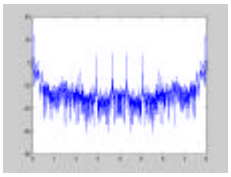
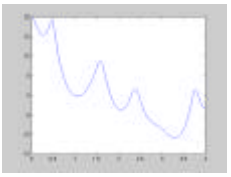
Janelamento do sinal	Carga do arquivo wave	Fourier x FreqZ
		
Lpc	**Fourier**	**FreqZ**
		

Tabela 5 - Visualização gráfica dos tratamentos do sinal

Backpropagation

O algoritmo apresentado demonstra a carga dos arquivos wave já com o processamento do sinal e a sua redução para 250 pontos, tendo este vetor de dados alguns itens selecionados para a sua introdução na rede neural para o treinamento.

Cada variável (numero01, numero02, ...) possui um conjunto de palavras gravadas, sendo este conjunto composto de palavras de zero a nove, somando 10 variáveis e cada variável com 10 palavras gravadas, totalizando 100 informações para manipulação com a rede neural.

```
[numero01, numero02, numero03, numero04, numero05,...
 numero06, numero07, numero08, numero09, numero10, tamanho] = cargaWave;

P = [numero01, numero02, numero03, numero04, numero05];
T = [tamanho tamanho tamanho tamanho tamanho];

alphabet     = P;
targets  = tamanho;

[R,Q] = size(alphabet);
[S2,Q] = size(targets);

S1 = 100;
net = newff(minmax(alphabet),[S1 S2],{'tansig' 'tansig'},'traingdx');
%net.LW{2,1} = net.LW{2,1}*0.01;
%net.b{2} = net.b{2}*0.01;

net.performFcn = 'sse';        % Sum-Squared Error performance function
net.trainParam.goal = 0.5;     % Sum-squared error goal.
net.trainParam.show = 20;      % Frequency of progress displays (in epochs).
net.trainParam.epochs = 25000; % Maximum number of epochs to train.
net.trainParam.mc = 0.95;      % Momentum constant.

[net,tr] = train(net,P,T);
```

Tabela 6 - Algoritmo de rede neural (backpropagation)

Para a realização dos testes após o treinamento, deve-se selecionar uma variável para ser apresentada a rede neural, para que se possa processar o reconhecimento. A variável apresentada no algoritmo abaixo é "numero10" que possui as palavras zero, um, dois, três, quatro, cinco, seis, sete, oito e nove dentro dela.

```
A = sim(net,numero10);
graficoBackpropagation(A);
```

Tabela 7 - Entrada de dados na rede neural para a sua identificação

A função gráfica Backpropagation apresenta um gráfico de barras com a representação dos números de 0 a 9, mostrando graficamente o índice do maior valor identificado pela rede neural para cada palavra.

```
function graficoBackpropagation(vetor)
figure(1)
    subplot(5,1, 1), bar(vetor( 1:10))
    subplot(5,1, 2), bar(vetor(11:20))
    subplot(5,1, 3), bar(vetor(21:30))
    subplot(5,1, 4), bar(vetor(31:40))
    subplot(5,1, 5), bar(vetor(41:50))
figure(2)
    subplot(5,1, 1), bar(vetor(51:60))
    subplot(5,1, 2), bar(vetor(61:70))
    subplot(5,1, 3), bar(vetor(71:80))
    subplot(5,1, 4), bar(vetor(81:90))
    subplot(5,1, 5), bar(vetor(91:100))
```

Tabela 8 - Algoritmo de apresentação gráfica dos resultados da rede neural

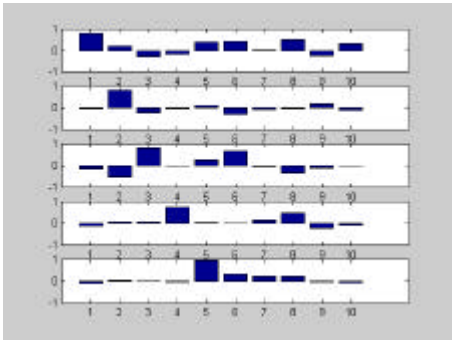


Figura 10 - Apresentação dos gráficos referentes aos números um, dois, três, quatro e cinco, juntamente com o valor identificado pela rede neural

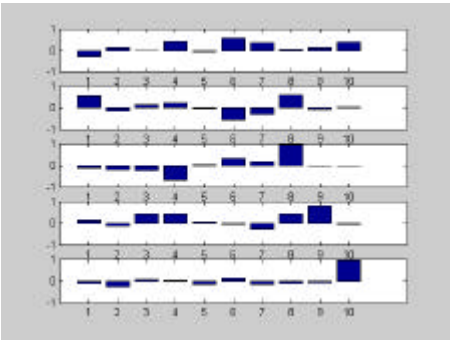


Figura 11 - Apresentação dos gráficos referentes aos números seis, sete, oito, nove e zero (representado por 10), juntamente com o valor identificado pela rede neural

Resultados

O reconhecimento de fonema utilizando os modelos matemáticos de janela (hamming), Lpc e Transformada-Z geraram um sinal de saída que obteve um percentual de acerto equivalente a 80% com a rede neural backpropagation.

Conclusões

As técnicas matemáticas utilizadas demonstraram ser eficiente para a identificação dos fonemas, sendo esta identificação comparada com a utilização de uma rede neural. A rede aprendeu e identificou as palavras testadas.

Os próximos trabalhos necessários têm como por objetivo testar o modelo matemático com outras redes neurais para mensurar o índice de resultados, para a identificação da melhor rede neural.

Trabalhos de identificação de fonemas utilizando o modelo matemático MFCC tem sido aplicado e com resultados satisfatórios.

Bibliografia

[1] S. HAYKIN, Redes neurais: princípios e prática. Porto Alegre: Bookman, 2001, 2º.ed. 900p.

[2] X HUANG et al., Spoken Language Processing: A Guide to Thery, Algorithm, and System Development. New Jersey: PH PTR, 2001. 980p.

[3] M. TAFNER et al., Redes Neurais Artificiais: introdução e princípios de neurocomputação. Blumenau: FURB, 1996. 199p.

[4] R. KOMPE, Prosody in speech understanding systems. New York: Springer, 1997. 357p.

[5] J. M. BARRETO, Inteligência Artificial no limiar do século XXI: abordagem hídrica simbólica conexionista e evolutiva. Florianópolis: Duplic, 2001, 3º.ed. 392p.

[6] F. M. AZEVEDO et al., Redes Neurais: com aplicações em controle e em sistemas especialistas. Florianópolis: Visual Books, 2000. 401p.

[7] C. LOESCH and S. SARI, Redes Neurais Artificiais: Fundamentos e Modelos. Blumenau: FURB, 1996. 166p.

[8] D. HANSELMAN and B. LITTLEFIELD, Versão do estudante Matlab 5: Guia do Usuário. São Paulo: Makron Books, 1999. 413p.

[9] É. MATSUMOTO, Matlab 6: Fundamentos de Programação. São Paulo: Érica, 2001. 310p.

[10] ENTENDENDO O COMPUTADOR, Comunicação entre Computadores. São Paulo: Nova Cultura, ISBN 85-13-00097-3, 1986. 56p.

[11] S. HAYKIN et. al, Sinais e Sistemas. Porto Alegre: Bookman, 2001. 668p.

[12] D. HANSELMAN and B. LITTLEFIELD, Matlab: Versão do Estudante: Guia do Usuário. São Paulo: Makron Books, 1997. 305p.

[13] J. M. GOMES and L. VELHO, Wavelets: Teoria, Software e Aplicações. Rio de Janeiro: IMPA, 1997. 216p.

[14] P. M. ANGEL, Inteligência Artificial II - Redes Neurais. Porto Alegre: UFRGS, 1999. 51p.

[15] E. P. RAPOSO, Desenvolvimento de um Sistema de Reconhecimento de Comandos Verbais para Robôs Baseados na Técnica de Redes Neurais Artificiais. Florianópolis: Mestrado em Engenharia Elétrica, 1997. 74p.

[16] K. GURNEY, Multilayer nets and backpropagation. Dept. Human Sciences, Brunel University, Uxbridge, Middx. UK.

[17] D. R. SÁNCHES, Redes Neurais e outras Tecnologias para o Reconhecimento de Fala. Florianópolis: LCMI - Departamento de Automação e Sistemas - UFSC.

[18] S. C. B. do SANTOS and A. Abraham, Sílabas como Unidades Fonéticas para o Reconhecimento Automático de Voz Contínua em Português. SBA Controle & Automação Vol. 12 no. 01/Jan., Fev., Mar., Abril de 2001.

[19] D. R. SÁNCHEZ, Introdução à Análise de Sinais. Florianópolis: LCMI - Departamento de Automação e Sistemas - UFSC.

9. GLOSSÁRIO

LPC – Modelo matemático para tratamento de sinais.

HMM – Camadas Ocultas de Markov.

ANN – Redes Neurais.

BACKPROPAGATION – Tipo de Rede Neural.

TRANSFORMADA-Z – Modelo matemático para tratamento de sinais.

FOURIER – Modelo matemático para tratamento de sinais.

HAMMING – Modelo matemático para tratamento de sinais

WAVE – Formato de arquivo de sons.

10.REFERÊNCIAS BIBLIOGRÁFICAS

[1] HAYKIN, Simon. Redes neurais: princípios e prática. Porto Alegre: Bookman, 2001, 2º.ed. 900p.

[2] HUANG, Xuedong; ACERO, Alex; HON, Hsiao-Wuen. Spoken Language Processing: A Guide to Thery, Algorithm, and System Development. New Jersey: PH PTR, 2001. 980p.

[3] TAFNER, Malcon A; XEREZ, Marcos de; FILHO, Ilson W. Rodriges. Redes Neurais Artificiais: introdução e princípios de neurocomputação. Blumenau: FURB, 1996. 199p.

[4] KOMPE, Ralf. Prosody in speech understanding systems. New York: Springer, 1997. 357p.

[5] BARRETO, Jorge Muniz. Inteligência Artificial no limiar do século XXI: abordagem hídrica simbólica conexionista e evolutiva. Florianópolis: Duplic, 2001, 3º.ed. 392p.

[6] FILHO, Ogê Marques; NETO, Hugo Vieira. Processamento Digital de Imagens. Rio de Janeiro: Brasport, 1999. 406p.

[7] AZEVEDO, Fernando Mendes; BRASIL, Lourdes Mattos; OLIVEIRA, Roberto Célio Limão de. Redes Neurais: com aplicações em controle e em sistemas especialistas. Florianópolis: Visual Books, 2000. 401p.

[8] LOESCH, Claudio; SARI, Solange T. Redes Neurais Artificiais: Fundamentos e Modelos. Blumenau: FURB, 1996. 166p.

[9] SCHILDT, Herbert. Inteligência Artificial Utilizando Linguagem C. São Paulo, McGraw-Hill, 1989. 349p.

[10] HANSELMAN, Duane; LITTLEFIELD, Bruce. Versão do estudante Matlab 5: Guia do Usuário. São Paulo: Makron Books, 1999. 413p.

[11] MATSUMOTO, Élia Yathie. Matlab 6: Fundamentos de Programação. São Paulo: Érica, 2001. 310p.

[12] RIMMER, Steve. Multi-Mídia: Programação for Windows. São Paulo: McGraw-Hill, 1994. 410p.

[13] KUELKAMP, Nilo. Cálculo 1. Florianópolis: Ed. UFSC, 1999. 470p.

[14] BOLDRINI, José Luiz; COSTA, Sueli I. Rodrigues; FIGUEIREDO, Vera Lúcia; "et al". Álgebra Linear. Campinas: Harbra, 1980, 3º.ed. 411p.

[15] ENTENDENDO O COMPUTADOR. Comunicação entre Computadores. São Paulo: Nova Cultura, ISBN 85-13-00097-3, 1986. 56p.

[16] HAYKIN, Simon; VEEN, Barry Van. Sinais e Sistemas. Porto Alegre: Bookman, 2001. 668p.

[17] GIOVANNI, José Ruy; BONJORNO, José Roberto. Matemática. São Paulo: FTD, 1992. 296p.

[18] HANSELMAN, Duane; LITTLEFIELD Bruce. Matlab: Versão do Estudante: Guia do Usuário. São Paulo: Makron Books, 1997. 305p.

[19] MELO, Jair Candido de. Princípios de Telecomunicações. São Paulo: McGraw-Hill, 1976. 220p.

[20] GOMES, Jonas de Miranda; VELHO, Luiz; GOLDENSTEIN, Siome. Wavelets: Teoria, Software e Aplicações. Rio de Janeiro: IMPA, 1997. 216p.

[21] ALVES, João Bosco da Mota. Sinais de Dados em Telecomunicações. Florianópolis: UFSC, 1999. 95p.

[22] ANGEL, Paulo Martins. Inteligência Artificial II – Redes Neurais. Porto Alegre: UFRGS, 1999. 51p.

[23] RAPOSO, Emerson Pereira. Desenvolvimento de um Sistema de Reconhecimento de Comandos Verbais para Robôs Baseados na Técnica de Redes Neurais Artificiais. Florianópolis: Mestrado em Engenharia Elétrica, 1997. 74p.

[24] **Redes Neurais Artificiais:** <http://www.icmsc.sc.usp.br/~prico/neural1.html>

[25] **Modelo de Redes FeedForward:** <**http://www.digitus.com.br/~mauryjr/Rna_Ffow.html**>

[26] GURNEY, Kevin. Multilayer nets and backpropagation. Dept. Human Sciences, Brunel University, Uxbridge, Middx. UK.

[27] **BPN.class:**< **http://www.patol.com/java/NN/index.html**>

[28] **Backpropagation:** <**http://www.timestocome.com/javaai/nnbackpropagation.html**>

[29] **The Backpropagation Network:** <**http://www.geocities.com/CapeCanaveral/1624/bpn.html**>

[30] OSÓRIO, Fernando; BITTENCOURT, João Ricardo. **Sistemas Inteligentes baseados em Redes Neurais Artificiais aplicados ao Processamento de Imagens.** I Workshop de Inteligência Artificial UNISC – Universidade de Santa Cruz do Sul – Departamento de Informática – Junho 2000. <http://www.inf.unisinos.br/~osorio>

[31] **Tutorial of Supervised Artificial Neural Networks:** <**http://lslwww.epfl.ch/~aperez/**>

[32] SÁNCHES, Damián Rodriguez. Redes Neurais e outras Tecnologias para o Reconhecimento de Fala. Florianópolis: LCMI – Departamento de Automação e Sistemas – UFSC.

[33] HOSOM, John-Paul; COLE, Ron; FANTY, Mark. **Speech Recognition Using Neural Networks at the Center for Spoken Languagem Undestanting.** Oregon Graduate Institute of Science and Technology, July 6, 1999. <**http://cslu.cse.ogi.edu/tutordemos/nnet_recog/recog.html**>

[34] SANTOS, Sidney Cerqueira Bispo do; ALCAIM, Abraham. Sílabas como Unidades Fonéticas para o Reconhecimento Automático de Voz Contínua em Português. SBA Controle & Automação Vol. 12 no. 01/Jan., Fev., Mar., Abril de 2001.

[35] SÁNCHEZ, Damián Rodríguez. Introdução à Análise de Sinais. Florianópolis: LCMI – Departamento de Automação e Sistemas – UFSC.

[36] OLIVEIRA, Neilza de Andrea. **Filtragem Espacial.** <**http://www.inf.ufsc.br/~visao/1999/neilza/c_filtro.htm**>

[37] ARGOUD, Fernanda I. Marques. Introdução à Transformada Wavelet. Florianópolis: Departamento de Engenharia Elétrica de Pesquisas em Engenharia Bio-Médica, 1988. 10p.

[38] ARGOUD, Fernanda I. Marques. Seminário: A Transformada Wavelet. Florianópolis: Departamento de Engenharia Elétrica de Pesquisas em Engenharia Bio-Médica, 1988. 21p.

[39] GRAPS, Amara. An Introduction to Wavelets. IEEE Computational Science and Engineering, Summer 1995, vol. 2, num. 2. 18p.

[41] CISTER, Angelo Maia; JACCOULD, Carlos Felipe de Brito. Reconhecimento del Habla Mediante una Red Neural con Pre-processamento Wavelet. Florianópolis: III Congresso Brasileiro de Redes Neurais, IV Escola de Redes Neurais, 1997. 330-336p.

[42] SÁNCHEZ, Damián Rodríguez. Os conceitos do Modelo Oculto de Markov no reconhecimento de fala. Florianópolis: LCMI – Departamento de Automação e Sistemas – UFSC. 19p.

[43] BENGIO, Yoshua. Markovian Models for Sequential Data. Montreal: Dept. Informatique et Recherche Opérationnellee, Université de Montréal. 129-163p.

[44] **A graphical Viterbi decoder:** <**http://lcmwww.epfl.ch/APPLET/VITERBI/Viterbi.html**>

[45] SUHM, B.; LEVIN, L.; COCCARO, N.; "et al.". Speech-Language Integration in a Multi-Lingual Speech Translation System. Carnegie Mellon University (USA). 8p.

[46] HONKELA, Timo. Self-Organizing Maps in Natural Language Processing. Finland: Helsinki University of Technology Neural Networks Research Centre, 1997. 63p.

[47] WU, Duanpei; GOWDY, John N. K-subspeces and Time-Delay Autoassociators for Phoneme Recognition. Clemson: Departament of Electrial and Computer Engineering Clemson University. 1871-1876p.

[48] DUARTE, M. Heveline V.; SOUZA, Márcio N. de. Técnicas Bio-inspiradas para Reconhecimento Automático de Voz. Rio de Janeiro: Programa de Engenharia Biomédica - COPPE/UFRJ, Departamento de Engenharia Eletrônica – Engenharia Elétrica/UFRJ. 443-444p.

[49] SCHUCK, A. Jr.; ZARO, M. A.; VILHENA, M.T.M.B. Comparison of the use of discrete and continuous Wavelet Transform in the Assessment of Dysphonic Voice. Proceedings of the EMBEC'99. 526-527p.

[50] GOLDENSTEIN, Siome Klein. Metamorfose de Sons e Aplicações. 1997. 66p.

[51] Álgebra Linear com Aplicações. Porto Alegre: Bookman, 2001. 305-308p.

[52] **Signal Processing Toolbox 5: for algorithm development, signal and linear system analysis, and time-series modeling**:< http://www.mathworks.com>

[53] KANUNGO, Tapas. **Hidden Markov Models – II.** Language and Media Processing Lab Center for Automation Research, University of Maryland. 25p. <**http://www.cfar.umd.edu/~kanungo**>

[54] DUGAD, Rakesh. A Tutorial on Hidden Markov Models. Signal Processing and Artificial Neural Networks Laboratory, Departament of Eletrical Engineering, Indian Institute of Technology – BombayPowai, Mumbai 400 076, India, 1996. 16p.

[55] SHEN, Jia-lin; HUNG, Jeih-weih; LEE, Lin-shan. Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments. Institute of Information Science, Academia Sinica, Taipei, Taiwan, Republic of Chine. 4p.

[56] COSI, Peiro. **D,DD,DDD,DDDD ..... Evidence Against Frame-Based Analysis Techniques.** Istituto di Fonética e Dialettologia – C.N.R. Italy. 5p. <**http://www.csrf.pd.cnr.it**>

[57] COSI, Piero. **SLAM: Segmentation and Labelling Automatic Module.** stituto di Fonética e Dialettologia – C.N.R. Italy. 4p.< **http://www.csrf.pd.cnr.it**>

[58] SHYU, Ruey-Ching; WANG, Jhing-Fa; LEE, Jau-Yien. Improvement in Connected Mandarin Digit Recognition by Explicitly Modeling Coarticulatory Information. Departament of Eletrical Engineering, National Cheng Kung University, Tainan, Taiwan, 2000. 649-660p.

[59] LLORENTE, Juan I. Goldinho; NAVARRO, Santiago Aguilera; ESPINOSA, Carlos Hernández; "et al.". On the Selection of Meaningful Speech Parameters Used by a Pathologic/Non Pathologic Voice Register Classifier. Madrid, Spain, Ciudad Universitaria. 4p.

[60] KOMPE, R.; DENZLER, J.; NIEMANN, H.; "et al". Going Back to the Source: Inverse Filtering of the Speech Signal with ANNs. EUROSPEECH, 1993, Berlin, Vol. 1, pp. 111-114.

[61] KOB, Malte; ALHAUSER, Nils; REITER, Ulrich. Time-Domain Model of the Singing Voice. Procceding of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99), NTNU, Trondhein, December 9-11, 1999. pp. w99-1, w99-4.

[62] WAN, Eric A.; NELSON, Alex T. Networks for Speech Enhancement. Oregon Graduate Institute of Science and Technology, 1998. 29p.

[63] WANG, Xihong; ZAHORIAN, Stephen A.; AUBERG, Stefan. Analysis of Speech Segments Using Variable Spectral/Temporal Resolution. Departament of Eletrical and Computer Engineering, Old Dominion University Norfolk. 4p.

[64] ALI, Ahmed M. Abdelatty; SPIEGEL, Dr. Jan Van der; MUELLER, Dr. Paul. Stop Consonant Classification Using Recurrant Neural Networks. NSF Summer Undergraduate Fellowship in Sensor Technologies David Auerbach (physics), Swarthmore College. 14p.

[65] BARROS, Allan Kardec; RUTKOWSKI, Tomaz; CICHOCKI, Andrzej. Speech Extraction from Interferences in real Environment using bank filters and blind source separation. Depto. de Engenharia Elétrica, Universidade Federal do Maranhão, São Luiz – Ma – Brasil.4p.

[66] CASTRO, Maria J.; PEREZ, Juan C. Comparison of Geometric, Connectionist and Structural Techniques on a Difficult Isolated Word Recognition Task. Dpto. de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Valencia, Spain. 5p.

[67] **Enhanced Silence Detection in Variable Rate Coding System using Voice Extraction.** IC Tech, Inc., Okemos, USA,< **ictech@ic-tech.com**.> 2p.

[68] WEIJER, Joost van de. Language Input to a Prelingual Infant. Max Plank Institute for Psycholinguistics, Nijmegen, The Netherlands. 4p.

[69] ROY, Deb. A Computational Model of Word Learning from Multimodal Sensory Input. Media Laboratory, Massachusetts Institute of Technology, Cambridge, USA. 8p.

[70] MOLZ, Rolf F.; ENGEL, Paulo M.; MORAES, Fernando G. Uso de um Ambiente Codesign para a Implementação de Redes Neurais. Proceeding of the IV Brazilian Conference on Neural Networks, pp. 013-018, July 20-22, 1999.

[71] ERGIN, Rivarol; SHAUGHNESSY, Douglas O', FARHAT, Azarshid. Generalized Mel frequency Cepstral Coefficients for Large-Vocabulary Speaker-Independent Continuous-Speech Recognition. IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 5, Setember 1999. pp. 525-532.

[73] GAUVAIN, Jean-Luc; LAMEL, Lori. Large-Vocabulary Continuous Speech Recognition: Advances and Applications. Proceeding of the IEEE, Vol. 88, No. 8, August 2000. pp. 1181-1200.

[74] ZAZULA, Damjan; GYERGYEK, Ludvik. Complexity in Signal Processing Using Cepstral Approach. 9p.

[75] DONG, Minghui; LUA, Kim-Teng; Automatic Prosidic Break Labeling for MandarinChinese Speech Data. Departament of Computer Science, School of Computing, National University of Singapore. 4p.

[76] COSTA, M.; GAY P.; PALMISANO, D.; "el al.". A Neural Ensemble for Speech Recognition. Dipartimento di Elettronica – Politecnico di Torino, Torino, Italy. 4p.

[77] **Digital Signal Processing Tutorial:** <http://www.dsptutor.freeuk.com>

[78] **FACENS – Projeto Vocallis:** <http://www.facens.br/site/ensino/projetos/andre/resumo.html>

[79] **Processos de Reconhecimento de Voz:** <http://www.inf.ufsm.br/~roberto/desenvolvimento.html>

[80] GANAPATHIRAJU, A.; WEBSTER, L.; TRIMBLE, J.; "el al.". **Comparison of Energy-Based EndPoint Detectors for Speech Signal Processing.** Departament of Eletrical and Computer Enginnering Mississippi State University. <http://www.isip.msstate.edu/publications/1996/secon'96/endpt.fm >

[81] **Cepstral Analysis:** <http://svr-www.eng.cam.ac.uk/~ajr/SA95/node33.html>

[82] **Fonema X Letra:** <http://www.agoraeusei.hpg.com.br/fonema.html>

[83] **Fonética Articulatória:** <http://www.peb.ufrj.br/~lib/TeseMscTujal/fonetica.html>

[84] **Português Gramática:**< http://cpsico.hypermart.net/oriestportugues.htm>

[85] **Fonemas e Letras:** <http://interaula.com/Portugues_fonemas_e_letras_01.html>

[86] TAFNER, Malcon Anderson. **Reconhecimento de Palavras Faladas Isoladas Usando Redes Neurais Artificiais.** Departamento de Engenharia de Produção – UFSC, Tese de Mestrado, Janeiro de 1996.

[87] CUNHA, Celso Ferreira da. **Gramática da Língua Portuguesa.** Fename, Rio de Janeiro, 1980.