ESCOLA POLITÉCNICA
DE PERNAMBUCO

Departamento de
Sistemas
Computacionais

Reconhecimento de voz dependente de locutor utilizando Redes Neurais Artificiais

Trabalho de Conclusão de Curso

Engenharia da Computação

Nome do Aluno: Petrônio de Luna Braga
Orientador: Prof. MSc. Renato Fernandes Corrêa

Recife, maio de 2006

UNIVERSIDADE
DE PERNAMBUCO

ESCOLA POLITÉCNICA
DE PERNAMBUCO

Departamento de Sistemas Computacionais

Reconhecimento de voz dependente de locutor utilizando Redes Neurais Artificiais

Trabalho de Conclusão de Curso

Engenharia da Computação

Este Projeto é apresentado como requisito parcial para obtenção do diploma de Bacharel em Engenharia da Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

Nome do Aluno: Petrônio de Luna Braga
Orientador: Prof. MSc. Renato Fernandes Corrêa

Recife, maio de 2006

UNIVERSIDADE
DE PERNAMBUCO

Petrônio de Luna Braga

Reconhecimento de voz dependente de locutor utilizando Redes Neurais Artificiais

Resumo

O problema de reconhecimento de fala é de difícil tratabilidade. A maior dificuldade é a sua natureza interdisciplinar. Além dessa, variabilidades acústicas, do transdutor, intra-locutor, entre-locutores estão relacionadas com o problema. Mas, em contra partida, várias áreas podem ser beneficiadas com o uso desta técnica, tornando um campo desafiador para ser pesquisado. Este trabalho apresenta o uso de uma rede neural artificial do tipo *Self-Organizing Map* (SOM), utilizado no reconhecimento de fala, através do modelo de subdivisão fonética. Utilizou-se técnicas de pré-processamento e extração de características do sinal da fala através do modelo cepstrum [41]. Relatou-se desde técnicas de pré-ênfase até os métodos de extração de coeficientes mel-cepstrais e energia. Foi criada e definida uma base de dados contendo os padrões fonéticos contidos na tese de Ynoguti [41].

Foi desenvolvido um sistema que teve como fim, facilitar o entendimento e ajudar pesquisadores que trabalham na área de reconhecimento de fala. Foram desenvolvidos três experimentos: reconhecimento de vogais, fonemas e frases. Os resultados obtidos foram razoáveis, quando levado em consideração a complexidade do problema. Com isso, mostra-se que as redes SOM são bastante adequadas a esse tipo de problema. Por fim, foi citado possíveis melhoramentos, de onde podem culminar novas pesquisas. Esses melhoramentos poderão ser incorporados de modo a se conseguir melhores resultados.

Abstract

The problem of speech recognition is of difficult resolution. The biggest difficulty is its interdisciplinary nature. Moreover, acoustics variabilities, of the transducer, intra speaker, beetween speakers are also related to this problem. On the other hand, some areas can profit from the use of this technique, making it a challenging field to be researched. This work presents the use of an artificial neural network. Cepstrum model's pre-processing techniques and characteristics of speak signal extraction were used [41]. Techniques of pre-emphasis and methods of extration of mel-cepstrals and energy coefficients had been related [41]. A database containing the phonetic standards was created and defined.

A system was developed to facilitate the agreement and to help the researchers that work in the area of recognition of speak. Three experiments had been developed: recognition of vowels, phonemes and phrases. The obtained results were reasonable which reveals that the artificial neural networks of type SOM are sufficiently adjusted to this type of problem. Finally, it was cited possible improvements, of where new research can culminate. These improvements could be incorporated in order to obtain better results.

Sumário

Índice de Figuras

Índice de Tabelas

Tabela de Símbolos e Siglas

RAF - automatic speech recognition (reconhecimento automático de fala)
LER - lesão por esforço repetitivo
HMM - hidden Markov models - modelos ocultos de Markov
GMM - gaussian mixture models - modelos de mistura gaussiana
RNA - redes neurais artificiais
SOM - self-organizing maps (mapa auto-organizáveis)
FFT - fast Fourier transform (transformada rápida de Fourier)
MCP - McCulloch-Pitts
MLP - multi-layer perceptron
LVQ - learning vector quantization
A/D - analógico/digital
LPC - linear predictive coding (codificação por predição linear)
DCT - discrete cosine transform (transformada discreta inversa do co-seno)
PCM - pulse-code modulation
WAV - wave
DC - corrente contínua
IDE - integrated development environment (ambiente de desenvolvimento integrado)
API - application programming interface

Agradecimentos

Gostaria de agradecer a todos que contribuíram de alguma forma para que este trabalho se tornasse realidade:

À Deus, a Ele devo tudo que tenho, sou e espero ser.

Aos meus pais, Petrônio Braga e Lúcia Braga, que me permitiram estar aqui hoje e pelo apoio inconteste.

Aos meus irmãos, Rodolfo Braga e Jorge Braga, pela paciência, companheirismo e grande ajuda em ler e escrever a minha monografia.

À minha namorada, Ana Régia Macedo Uchôa, companheira, ombro amigo e que me ajudou bastante na etapa de separação dos fonemas.

Ao meu orientador, Renato Fernandes Corrêa, pelo aprendizado, apoio, pela forma que me entendeu e pelas cobranças.

Ao grande amigo e irmão, César Augusto Lins Oliveira, pela disponibilidade, amizade e ajuda.

Aos amigos Cleyton Mário, Diogo Pacheco, Filipe Regueira, Laura Moraes, Marcelo Nunes, Milena Rodrigues, Nívia Quental que conheci e convivi na universidade, pela amizade e companheirismo nestes 5 anos. Sentirei saudades!

Aos meus professores e colegas da POLI, pelo aprendizado e caminhada durante o curso.

Aos meus familiares e amigos.

Capítulo 1

Introdução

As interfaces homem-máquina vêm sendo aperfeiçoadas para as mais diversas aplicações e fins [37]. Interfaces antes só imaginadas no campo da ficção científica, hoje, estão se tornando uma realidade comum [17]. Dentre essas, a tecnologia de reconhecimento de fala está ganhando, cada vez mais, visibilidade na computação empresarial e pessoal [17, 37, 41]. A grande aceitação dessa tecnologia é, em geral, devida ao aumento na interação do usuário, produtividade e diminuição de custos [17].

O reconhecimento automático de fala (RAF) - *Automatic Speech Recognition* - tem como objetivo primordial, dada uma entrada em forma de um sinal (onda acústica), produzir uma saída em forma de seqüência de fonemas, palavras ou sentenças correspondentes ao sinal de entrada, ou seja, um sistema RAF transcreve a fala em texto [24, 41].

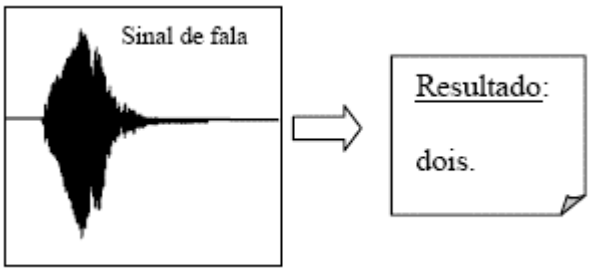


Figura 1. Processo de reconhecimento de fala.

A figura 1 ilustra um sinal de fala captado por um microfone, o qual passou por um sistema de RAF produzindo como saída o texto *dois*.

De forma simplificada, o RAF consiste no processo de extrair a informação lingüística no sinal da fala. Esse processo normalmente acontece em três passos [1]:

1. Aquisição do sinal de fala
2. Extração de Parâmetros
3. Reconhecimento do Padrão

O primeiro passo consiste em realizar a captação do sinal da fala através de um transdutor, geralmente, um microfone ou telefone, e repassar o sinal a uma interface analógica/digital (uma placa de som, por exemplo), que consiste na entrada da informação (sinal de voz) em forma analógica e recolhermos na saída essa mesma informação de forma digital.

O segundo passo é o pré-processamento. A meta é extrair do sinal capturado as características que descrevem adequadamente o sinal de voz.

O terceiro passo consiste em comparar os dados extraídos na fase de pré-processamento com os padrões armazenados anteriormente medindo a similaridade entre eles e escolhendo o padrão que melhor representa o sinal da fala.

A principal meta das pesquisas na área de reconhecimento de fala é desenvolver um modelo que seja capaz de decodificar a fala com uma alta taxa de acerto sem a dependência do usuário e que se adeque a todos os ambientes, possibilitando assim, uma comunicação homem-máquina mais amigável e natural, como a utilizada entre os seres humanos.

Embora esse objetivo esteja distante de ser tangenciado, muitos avanços foram conseguidos nos últimos anos, fazendo com que a tecnologia de reconhecimento de fala passasse do limiar da aceitabilidade [34, 38], tornando-a aplicável ao dia-a-dia dos seres humanos. Isso é decorrente, graças ao avanço dos algoritmos disponíveis para modelar os problemas de reconhecimento de fala e do barateamento dos sistemas de alto desempenho. Hoje, é possível ver aparelhos com sistemas de reconhecimento de fala, principalmente as embarcadas em celulares e *handhelds* [23].

A fala como meio de comunicação entre o computador e o usuário traz diversos benefícios. Dentre eles, podem ser citados [24, 25, 26, 27, 41]:

- Rapidez: torna a interação mais rápida e assim acelera a realização de tarefas.
- Evita problemas médicos: reduz o risco de lesão por esforço repetitivo (LER).
- Maior mobilidade: o usuário pode utilizar o sistema enquanto está se movendo ou fazendo uma outra atividade que requer o uso das mãos.
- A rede telefônica pode ser usada para a passagem da informação, possibilitando dessa forma o acesso remoto ao sistema.
- Aplicações Médicas: utilização da fala na manipulação e interação com os equipamentos cirúrgicos.
- Possibilita o acesso de deficientes visuais e físicos serem incluídos no contexto social e tecnológico.

Porém alguns pormenores devem ser levados em consideração. A utilização da fala como um meio de comunicação entre o homem e a máquina apresenta seus percalços e complicações. Em um sistema que incorpore este tipo de funcionalidade, aspectos como: características do microfone, conversão analógico-digital do sinal, estado físico, emocional e cultural do locutor, assim como características intrínsecas do ambiente precisam ser levadas em consideração [12, 27, 28, 36]. Porém, o locutor, é o aspecto que introduz a maior variabilidade ao sistema [27].

Os sistemas de reconhecimento de fala têm a sua capacidade baseada na associação entre o locutor e o modo e estilo de pronúncia, treinamento, vocabulário, modelo de linguagem, perplexidade, relação sinal ruído e transdutor [41].

Para efetuar o reconhecimento em si, os principais métodos usados são os baseados em Modelos Ocultos de Markov (HMMs) [23, 38], Modelos de Mistura Gaussiana (GMMs) [6, 31, 32] e Redes Neurais Artificiais (RNAs) [23, 29, 41]. A utilização de cada método é dependente principalmente da modalidade de texto associada ao problema. As HMMs têm demonstrado melhores resultados em aplicações dependentes de texto, enquanto que os GMMs e as RNAs têm melhores resultados em aplicações independentes de texto [23].

Outros trabalhos relacionados na área de reconhecimento de fala para o Português-Brasil podem ser encontrados [1, 3, 23, 24, 25, 30, 34, 35, 41].

O presente trabalho objetiva a construção de um sistema que trabalhe com reconhecimento de fala contínua baseado em unidades fonéticas, com dependência de locutor e vocabulário médio, onde o usuário através da sua fala possa reconhecer textos em forma de padrões fonéticos, ou seja, dada uma frase, dizer os fonemas contidos na mesma.

Além desse, outro objetivo definido foi o de se criar um sistema de ajuda, incorporado ao anterior, que pudesse ser utilizado por outros pesquisadores, tendo uma interface visual bastante intuitiva, com a intenção de diminuir o tempo entendimento de projetos da área de reconhecimento da fala.

O segundo capítulo tem por finalidade discorrer sobre o histórico, características da produção da fala humana, os problemas e dificuldades e as características mais comuns dos sistemas de reconhecimento de fala. Além disso, apresentar uma visão geral do estado da arte, técnicas atualmente mais utilizadas, para os sistemas de reconhecimento de fala.

O terceiro capítulo faz uma breve revisão de literatura, além disso, introduz as redes neurais artificiais, citando as principais arquiteturas usadas no reconhecimento de fala e aprofundando-se na rede SOM (*Self-Organizing Maps*). Nesse capítulo, também é mostrado e explicado o experimento que Teuvo Kohonen realizou, “datilógrafo fonético” (The “neural” Phonetic Typewriter) [21], experimento que converte a linguagem falada, nos idiomas finlandês e japonês, em um texto escrito, através do reconhecimento de fonemas em fala contínua usando uma rede SOM [21].

O quarto capítulo descreve desde a etapa de aquisição do sinal acústico até as técnicas de pré-processamento e extração de características mais utilizadas atualmente na área. Dentre as técnicas de pré-processamento, descritas neste trabalho, podemos citar: filtro de pré-ênfase, técnica de divisão do sinal em quadros, técnica de janelamento e endpoints. Também, citou-se as algumas técnicas de extração dos parâmetros para o reconhecimento de fala, aprofundando-se nos coeficientes cepstrais baseados na escala de freqüências mel, técnica de extração de características mais utilizada atualmente nos sistemas de reconhecimento de fala, devido a sua boa adequação as características do trato vocal humano.

O quinto capítulo descreve a implementação e todas as funcionalidades do sistema proposto. Todas as funcionalidades são explicadas e imagens do sistema são mostradas para uma melhor idéia do trabalho desenvolvido, assim como, são apresentadas a arquitetura básica do sistema e a definição dos módulos construídos.

No sexto capítulo, são descritos a base de dados e a sua criação, o treinamento da rede SOM aliado as explicações do uso dos parâmetros escolhidos, as técnicas de pré-processamento utilizadas, a técnica de extração de característica escolhida e como foram realizados os experimentos propostos.

O sétimo capítulo apresenta e analisa os resultados obtidos nos experimentos para as dimensões de redes propostas.

No sexto capítulo, são discutidas as conclusões, principais contribuições, limitações e propostas para futuros trabalhos.

Capítulo 2

Sistemas de reconhecimento de fala

O RAF é um problema de difícil resolução. Segundo Rabiner e Juang [27], a maior dificuldade ao pleno desenvolvimento de um sistema de RAF é a natureza interdisciplinar do problema. Para se obter sucesso no desenvolvimento e implementação de um sistema completo de RAF, são necessários, entre outros, conhecimentos especializados das áreas de processamento de sinais, fonética articulatória e acústica, reconhecimento de padrões, teoria das comunicações, lingüística estrutural, neurofisiologia, inteligência artificial, ciência da computação, psicologia cognitiva [33]. Devido a isso, não existe, hoje, um sistema capaz de reconhecer a fala de qualquer pessoa, em um ambiente indeterminado, pronunciada de qualquer maneira e abrangendo um vocabulário ilimitado em qualquer idioma.

Atualmente, os sistemas de RAF que atingem uma alta taxa de corretude, conseguem-no, restringindo o domínio das variáveis acima citadas, isto é, limitam-se a um objetivo específico [24, 33].

2.1 Histórico

A literatura reporta que a primeira pesquisa na área de reconhecimento de fala é datada de 1952 [9]. Foi proposta por Davis, Biddulph e Balashek nos laboratórios da Bell [12]. O objetivo desse trabalho foi criar um sistema que reconhecesse dígitos isolados falados por um único locutor. Desde então, muitos pesquisadores, têm trabalhado e se dedicado nessa área.

Em meados das décadas de 50 e 60, vários sistemas reconhecedores de dígitos e fonemas foram implementados e bons resultados foram alcançados [13]. Em decorrência desses bons resultados alcançados, no final dos anos 60 e início dos anos 70, um grande impulso nas pesquisas em reconhecimento de fala foi obtido [13].

Existem vários exemplos de sistemas de reconhecimento de fala, dentre os quais podem ser citados [24]:

- Dragon - 1975 - Carnegie-Mellon University: reconhecimento de fala contínua dependente do locutor com vocabulário de 194 palavras com taxa de acerto de 84%.

- Hearsay - 1975 - Carnegie-Mellon University: reconhecimento de fala contínua dependente do locutor com vocabulário de 1011 palavras com taxa de acerto de 87%.
- Harpy - 1976 - Carnegie-Mellon University: reconhecimento de fala contínua dependente do locutor com vocabulário de 1011 palavras com taxa de acerto de 97%.
- Bell Labs – 1982 - reconhecimento de palavras isoladas independente do locutor com vocabulário de 129 palavras com taxa de acerto de 91%.
- Prina - 1982 - Ericson Business Systems: reconhecimento de palavras isoladas dependente do locutor para vocabulário pequeno (menos que 25 palavras).
- Feature - 1983 - Carnegie-Mellon University: reconhecimento de palavras isoladas independente do locutor, com vocabulário constituído pelas letras do alfabeto com taxa de acerto de 90%.
- Tangora - 1985 - IBM: reconhecimento de palavras isoladas dependente do locutor com vocabulário de 5000 palavras com taxa de acerto de 97%.
- Bell Labs - 1988: reconhecimento de dígitos conectados independente do locutor com taxa de acerto de 97.1%.
- Byblos - 1988 - BBN: reconhecimento de fala contínua dependente do locutor com vocabulário de 997 palavras com taxa de acerto de 93%.
- Sphinx - 1988 - Carnegie-Mellon University: reconhecimento de fala contínua independente do locutor com vocabulário de 997 palavras com taxa de acerto de 96.2%.
- Teleton - 1988 - Deutsche Bundespost Telekom: reconhecimento de palavras isoladas independente do locutor com vocabulário de 12 palavras.
- Babsy - 1990 - Deutsche Bundespost Telekom: reconhecimento de palavras isoladas independente do locutor com vocabulário de 18 palavras com taxa de acerto de 95%.
- Mairievox - 1990 - France Telecom: reconhecimento de palavras isoladas e conectadas independente do locutor com vocabulário de 21 palavras com taxa de acerto de 88%.
- Citruf - Deutsche Bundespost Telekom: reconhecimento de palavras isoladas independente do locutor para vocabulário pequeno (menos que 25 palavras).
- Teledialogue - 1992 - Jydsk Telefon: reconhecimento de palavras isoladas independente do locutor para vocabulário pequeno (menos que 25 palavras).
- Audiotex - 1992 - Telefonica I. D. de Espanha: reconhecimento de palavras isoladas independente do locutor com vocabulário de 12 palavras com taxa de acerto de 96%.
- World Window - 1992 - Global Communications Ltd.: reconhecimento de palavras conectadas dependente do locutor com vocabulário de 200 palavras.
- Les Balandins - 1992 - France Telecom: reconhecimento de palavras isoladas e conectadas independente do locutor com vocabulário de 26 palavras com taxa de acerto de 95%.
- IBM - 1993: lança primeiro software comercial para reconhecimento de fala.
- Dragon Systems - 1994 - Dragon Dictate: sistemas de reconhecimento de ditados.
- Philips Dictation Systems - 1996: reconhecimento de palavras isoladas dependente do locutor com vocabulário de 64000 palavras.
- IBM - 1996 - MedSpeak/Radiology: primeiro produto para reconhecimento da fala em tempo real.
- Dragon Systems - 1997 - disponibiliza reconhecimento de fala contínua em inglês.
- IBM -1997 – ViaVoice: sistema de reconhecimento de fala.
- IBM -1998 - ViaVoice: sistema de reconhecimento de fala em português.
- MicroPower - 1998 – DeltaTalk: sintetizador de voz em português.
- Philips - 1999 - FreeSpeech 2000: sistema com reconhecimento de português.
- Telemar – 2001 - Vocall: serviço de voz aberto ao público, com síntese e reconhecimento de fala, para e-mails e agenda.

- Microsoft - 2001 - Microsoft Speech Application Software Development Kit: suporta as linguagens: visual basic e c++.
- Microsoft - 2001 - Office X: apresenta recursos de voz (para ditados e voz).
- Microsoft - 2005 - Microsoft Speech Application Software Development Kit 1.1: que adiciona suporte a funcionalidades de voz para aplicações WEB com a linguagem ASP.NET.

2.2 Sistema de produção de fala humana

Para se obter um bom resultado no reconhecimento de fala é preciso um conhecimento prévio acerca dos sons da fala, suas classificações e características. É também necessário o entendimento do sistema de produção da fala humana, devido ao fato dos módulos responsáveis pela extração dos parâmetros da fala levarem em consideração essas características.

2.2.1 Sons da fala

A forma mais usual de um ser humano se comunicar é através da linguagem falada. A linguagem falada contém um número de elementos básicos de sons, denominados fonemas. Para um melhor entendimento sobre o assunto indicamos a leitura da dissertação de mestrado de Ishi [16].

Conforme Alencar [1], os sons da fala podem ser classificados em: vogais, nasais, fricativas e oclusivas, cujas características são descritas abaixo:

- Vogais – são originárias de vibrações das cordas vocais, tendo duração dependente da vizinhança fonética que a acompanha.
- Nasais – são caracterizados pela saída pelo narinas.
- Fricativas – são caracterizadas por um chiado contínuo criado a partir de constrições estreitas no aparelho vocal.
- Oclusivas – são caracterizados por um excesso de pressão criado em um ponto do aparelho vocal, seguido de um desprendimento repentino de ar.

Para a criação de um sistema de RAF é imprescindível que se conheça as unidades fundamentais, pois são elas que definem o arcabouço do reconhecedor e qual a unidade básica que ele trabalhará [1, 41].

2.3 O sistema de reconhecimento de fala

O desenvolvimento de interfaces controladas pelo uso da fala, tem como meta principal a substituição, em alguns casos, das interfaces tradicionais como mouses, teclados e outros dispositivos [17]. O RAF vislumbra a criação de uma interface mais amigável e natural entre a máquina e o homem.

2.3.1 Unidades fundamentais

Geralmente, em sistemas onde o tamanho do vocabulário é pequeno (de 20 a 50 palavras) [16, 35, 41], é comum utilizar-se às próprias palavras que pretendemos reconhecer como unidades

fundamentais. Para um treinamento adequado desses sistemas, é necessário ter um grande número de exemplos de cada palavra. No entanto, para sistemas com tamanho de vocabulários maiores, torna-se inviável e impraticável ter uma grande quantidade de exemplos de cada palavra.

Com isso, o uso de unidades fundamentais, tais como fonemas, sílabas, etc., é uma alternativa bastante plausível, pois com isso reduz-se drasticamente a quantidade de exemplos necessários para uma maior generalização. Contudo, o uso dessas unidades fundamentais pode aumentar significativamente a complexidade do sistema.

Dois parâmetros relevantes para uma boa escolha de sub-unidades são [16, 35, 41]:

- consistência: exemplos diferentes de uma unidade devem ter características similares.
- treinabilidade: devem existir exemplos de treinamento suficientes de cada subunidade para criar um modelo robusto.

Existem diversos tipos de sub-unidades, das quais podem ser citadas: sílabas, fones, difones, trifones, dentre outras. Entretanto, sílabas, difones e trifones são bastante consistentes, mas são difíceis de treinar, enquanto que unidades menores, tais como os fones, são simples de treinar, mas não apresentam uma consistência desejável.

Uma forma de minimizar o problema de consistência encontrado nos fones, é usar fones dependentes do contexto [16, 35, 41]. Estas unidades têm uma consistência muito boa, isso ocorre pelo fato de se levar em consideração o efeito de coarticulação dos fones vizinhos.

2.3.2 O reconhecedor de fala

Um reconhecedor de fala tem como entrada um sinal acústico, capturado por meio de um transdutor. A partir desse sinal, a saída é mapeada em forma de seqüências de fonemas ou palavras correspondentes ao sinal de entrada, ou seja, um reconhecedor de fala transcreve a fala em texto [1, 23, 36].

A figura 2 ilustra os componentes que um sistema de RAF precisa ter.

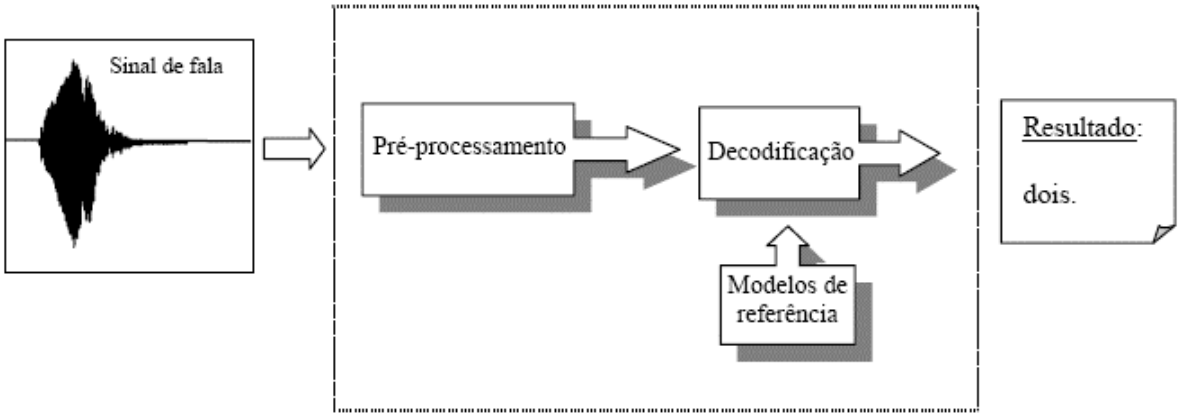


Figura 2. Sistema de Reconhecimento de Fala.

Uma forma comum encontrada na literatura [1, 23] é dividir um sistema de RAF em componentes, conforme a figura 2 nos mostra, deixando o seu entendimento mais fácil. Cada componente tem sua meta bem definida. Dos quais podem ser citados:

- Pré-processamento – responsável desde a captação do sinal acústico até a extração das características espectrais e temporais do sinal de fala capturado.
- Decodificação – processo que tenta simular, através de modelos estatísticos, a mente humana. Uma decisão precisa ser tomada, um conhecimento anterior é preciso, e esse é representado pelos modelos de referência obtidos a partir da base dos sinais.
- Modelos de Referência – é à base do conhecimento. Contém a representação dos fonemas, palavras ou sentenças. Os modelos de referência são obtidos a partir de exemplos das unidades a serem reconhecidas.
- Pós-processamento – nesta fase, as probabilidades obtidas na comparação com os modelos de referência são usadas para definir o padrão que melhor corresponde ao padrão que se deseja reconhecer. Para ajudar, podem-se usar restrições sintáticas e semânticas (uma gramática, por exemplo). Com isso podemos diminuir a dimensionalidade das possibilidades dos padrões possíveis.

Embora, o entendimento da construção de um sistema de RAF seja fácil, o processo de desenvolvimento e implementação do mesmo é extremamente difícil. A isso se relaciona a natureza interdisciplinar do problema. Entretanto, outras dificuldades estão relacionadas com esse problema, dentre as quais [10, 25, 26, 41]:

- Variabilidades acústicas - ruído, temperatura e umidade do ambiente;
- Variabilidades do transdutor – tipo de transdutor usado na captação e características do transdutor.
- Variabilidades intra locutor – podem resultar de mudanças do estado físico/emocional dos locutores, velocidade de pronúncia ou qualidade de voz.
- Variabilidades entre locutores – podem proceder das diferenças na condição sócio-culturais, geográficas, forma e tamanho do trato vocal para cada uma das pessoas.

2.3.2.1 Características

Podem-se caracterizar os sistemas de RAF de várias maneiras. Algumas das mais importantes características, quanto à capacidade de sistemas de RAF, encontram-se condensadas abaixo [23, 24, 35, 37].

Quanto ao modo de pronúncia:

- Reconhecedor de palavras isoladas – são os sistemas mais simples de serem desenvolvidos, onde devem existir pequenas pausas entre as palavras de uma locução.
- Reconhecedor de palavras conectadas – são sistemas mais complexos que os anteriores, utilizam palavras como unidade fonética padrão e reconhecem sentenças pronunciadas de forma natural. Entretanto, essas sentenças devem ser bem pronunciadas.
- Reconhecedor de voz contínua – são os mais complexos e difíceis de serem implementados, pois devem ser capazes de lidar com todas as características e vícios da forma natural de falar, dentre os quais podem ser citados: durações de palavras desconhecidas, efeitos de coarticulação e pronúncia descuidada.

Quanto ao estilo de pronúncia:

A quantidade de coarticulações aumenta à medida que se vai de uma pronúncia em modo de leitura para uma pronúncia espontânea. Desse modo, o grau de dificuldade de se reconhecer palavras aumenta da mesma forma.

Quanto ao treinamento:

- Sistemas dependentes de locutor: necessitam de uma fase de treinamento para cada usuário antes de serem utilizados;
- Sistemas independentes do locutor: não precisam da fase de treinamento, já que, foram previamente treinados com vários locutores.

Quanto ao tamanho do vocabulário:

A dificuldade de um reconhecedor de fala é diretamente proporcional ao aumento do vocabulário utilizado ou de palavras parecidas. É comum classificar sistemas que reconhecem menos de 20 a 50 palavras como pequenos, enquanto sistema que reconhecem mais de 20000 palavras como grandes [41].

Quanto ao modelo de linguagem:

Uma das maneiras mais comuns encontradas no reconhecimento é a partir da fala produzir seqüências de fonemas ou palavras, onde são utilizados modelos de linguagem para diminuir a dimensionalidade das possibilidades de seqüências fonemas ou palavras. Os modelos podem sofrer bastante simples, que são os fundamentados numa máquina de estados finita onde há um mapeamento para as possíveis palavras que poderão vir após uma determinada palavra a modelos de linguagem mais completos e gerais, onde são definidas gramáticas sensíveis ao contexto.

Quanto à dificuldade de reconhecer uma palavra:

Um conceito bastante aplicado para calcular a dificuldade de se reconhecer uma palavra é a perplexidade. Ela se baseia na combinação do tamanho do vocabulário e do modelo de linguagem aplicado. Pode-se defini-la como a média do número de palavras que pode seguir uma palavra depois que o modelo de linguagem foi aplicado.

2.4 Base de dados

A linguagem falada é o modo mais usado e natural de comunicação entre humanos. Seu modelo é regido por estruturas fonológicas, sintáticas e semânticas da língua [31]. A fala é produzida de forma diferente de pessoa para pessoa. Aspectos como: idioma, dialeto, forma e tamanho do trato vocal, velocidade da pronúncia, fatores culturais, geográficos, étnicos e sexuais entre outros fatores. Ainda, os padrões de fala são modificados pelo ambiente físico, contexto social, estado físico e emocional das pessoas, dentre outros [12].

As implicações causadas por variáveis não modeladas ou mal modeladas no desempenho dos sistemas de RAF são desastrosas. Dessa forma, para fornecer exemplos em número suficiente para que os métodos estatísticos funcionem de forma mais adequada, a base de dados precisa ser de tamanho adequado para o problema que se quer resolver.

Tanto o fator financeiro quanto o tempo despendido para se criar uma base de dados dessa é alta.

Uma forma para resolver esse inconveniente seria a de conseguir parcerias com empresas, instituições de pesquisa e agências financiadoras, de maneira a diluir custos, evitar duplicação de esforços e distribuir as tarefas. Mas, para se conseguir envolver um maior número de agentes neste processo, é necessário que esta base de dados não seja direcionado a um sistema ou serviço

específico, mas tentar atender as necessidades de vários grupos, linhas de pesquisa e desenvolvimento, em diversas áreas do conhecimento.

Parcerias com esse intuito já existem, principalmente em países desenvolvidos. Na Europa, um conjunto de 8 países integrou o projeto EUROM_1. Os países foram: Itália, Inglaterra, Alemanha, Holanda, Dinamarca, Suécia, Noruega e França, com a adesão posterior de Grécia, Espanha e Portugal. A construção da base de dados foi criada com o número de locutores (30 homens e 30 mulheres), escolhidos através dos mesmos critérios e gravados em condições acústicas semelhantes, e no mesmo formato. Ainda, em Portugal, foi criada uma base de dados chamada BD-PUBLICO. Da mesma forma, os EUA também fizeram grandes esforços neste sentido, e já existem disponíveis bases de dados (TIMIT, TI-DIGITS e SWITCHBOARD) para o seu idioma.

Aqui no Brasil, um trabalho pioneiro foi feito pelo pesquisador Carlos Alberto Ynoguti na sua tese de doutorado [41]. No seu trabalho foi definida e criada uma base de dados para o Português-Brasil utilizando 40 locutores (20 homens e 20 mulheres), na qual foi está disponibilizada para ser utilizada. No apêndice A se encontra todas as frases e suas transcrições fonéticas contidas nessa base.

As gravações da base de dados criada por Ynoguti foram realizadas em ambiente relativamente silencioso, com um microfone direcional de boa qualidade, utilizando uma placa de som SoundBlaster AWE64. A taxa de amostragem utilizada foi de 11.025kHz, e resolução de 16 bits. Os dados foram armazenados no formato Windows PCM (WAV).

Com a existência e a disponibilização destas bases, a área da tecnologia da fala tem evoluído bastante, não só porque poupa-se um trabalho penoso, caro e demorado, como também se abriu a possibilidade de se comparar os resultados de cada novo método proposto.

2.5 Estado da arte

De um modo geral, os reconhecedores de fala podem ser divididos em três classes principais. Essa divisão é feita levando-se em consideração a técnica utilizada para o reconhecimento [23, 24]. Existem os baseados em Modelos Ocultos de Markov (*HMMs – Hidden Markov Models*), em Modelos de Mistura Gaussiana (*GMMs - Gaussian Mixture Models*) e em Redes Neurais Artificiais (*RNAs - Artificial Neural Networks*).

Atualmente, o método mais eficaz para o reconhecimento depende principalmente da modalidade de texto associada ao problema [23]. As modalidades variam de dependentes a independentes do texto falado. A diferença fundamental é que as dependentes do texto, as palavras que serão reconhecidas são previamente definidas.

Os HMMs têm demonstrado os melhores resultados em aplicações que têm dependência do texto. Os HMMs são modelos estatísticos, com grande capacidade de modelagem das dependências temporais associadas aos sinais de fala. Para um melhor entendimento sobre a aplicação e os resultados do uso de HMMs para reconhecimento automático de locutor veja as referências citadas [7, 10, 24, 27, 34, 35, 37, 41].

Os GMMs são também modelos estatísticos, em que as probabilidades de ocorrência dos vetores de atributos para cada locutor são modeladas como combinações ponderadas de variáveis aleatórias vetoriais com Gaussianas. Apresentam resultados excelentes em aplicações independentes de texto [6, 23, 31].

As RNAs são modelos conexionistas não lineares, com grande capacidade de reconhecimento e classificação de padrões. Muitas arquiteturas de RNAs foram experimentadas em RAF, sendo que os melhores resultados são conseguidos pelo uso de arquiteturas baseadas em

quantização vetorial para aplicações independentes de texto. Seu desempenho é comparável ao dos GMMs [23].

Um caso de bastante sucesso da aplicação de técnicas de processamento de sinal, no caso FFT, junto a RNAs na resolução do problema de RAF foi o "datilógrafo fonético" (*The "neural" Phonetic Typewriter*) [21] implementado pelo pesquisador da universidade de Helsinki, Teuvo Kohonen, no final da década de 80. Esse sistema converte a linguagem falada, nos idiomas finlandês e japonês, em um texto escrito, através do reconhecimento de fonemas em fala contínua [21]. Kohonen utilizou para a resolução do problema uma RNA que continha uma topologia de auto-organização, que depois ficou conhecida como: rede SOM – *Self-Organizing Maps*. O experimento chegou a uma taxa de acerto de 92 a 97% nas conversões.

Capítulo 3

Redes neurais artificiais

Neste capítulo são apresentados conceitos referentes a redes neurais artificiais, desde sua fundamentação biológica até seu mapeamento matemático. Redes SOM (*Self-Organizing Maps*) também têm seus conceitos apresentados e explicados. Além disso, é descrito o experimento que Teuvo Kohonen realizou, "datilógrafo fonético" (The "neural" Phonetic Typewriter) [21], experimento que converte a linguagem falada, nos idiomas finlandês e japonês.

3.1 Redes Neurais Artificiais

3.1.1 Neurônio biológico

A capacidade do cérebro humano de processar informações de forma não-linear e paralela permite a resolução de tarefas complexas. Esse poder de processamento é creditado aos neurônios e suas conexões. A partir da "experiência" acumulada com o tempo, são reforçadas ou inibidas as conexões entre eles. O potencial do cérebro humano, então, não provém da simplicidade de cada neurônio em si, mas da complexidade e da grande quantidade das interconexões formadas ao longo do tempo [14]. Conforme a lista abaixo, podemos imaginar o grau de complexidade do cérebro humano.

- Natureza assíncrona;
- Quantidade de neurônios: 10^{11} neurônios aproximadamente;
- Quantidade de sinapses: 10^{3} aproximadamente;
- Tipos de neurônios: 10^{2} aproximadamente.

O entendimento do funcionamento do cérebro se tornou menos desconhecido devido aos pesquisadores Ramón y Cajál [14], os quais introduziram o conceito de que neurônios, na realidade, são os constituintes do cérebro. O processo de aprendizado, então, está fortemente relacionado aos próprios neurônios. Para que esse processo ocorra, uma ligação entre os neurônios é necessária. Essa ligação é denominada *sinapse*, cuja finalidade é impor ao neurônio

receptivo um grau de excitação ou inibição. Contudo, esses sinais inibitórios e excitatórios trafegam através dos *axônios*, que nada mais são que as linhas de transmissão, e dos *dendritos*, que são as zonas receptivas. A figura 3 ilustra um exemplo de um neurônio biológico.

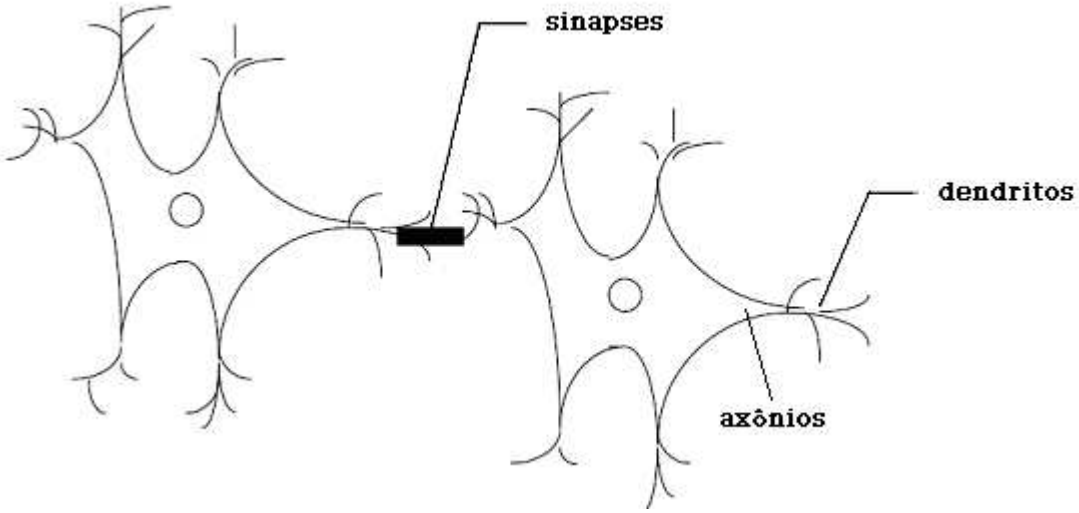


Figura 3. Neurônio biológico indicando onde se localizam os axônios e dendritos e onde ocorrem as sinapses.

3.1.2 Conceitos gerais

Segundo Haykin [14], uma rede neural artificial é um processador paralelo e distribuído, constituído de unidades de processamento que processam funções matemáticas quaisquer a fim de armazenar conhecimento e utilizá-lo. A estrutura desses sistemas são dispostas em camadas e interligadas através de conexões. Associadas às conexões, geralmente, encontram-se os pesos, os quais são os responsáveis pelo conhecimento armazenado na rede.

Os benefícios das redes neurais se evidenciam em sua habilidade para executar computação distribuída e na sua generalização. Segundo Haykin [14], a generalização se refere à capacidade de apresentar saídas coerentes para entradas que não estavam presentes durante o treinamento (aprendizagem). Além disso, RNAs possuem outras potencialidades, tais como: capacidade de trabalhar com problema não-lineares, adaptabilidade (habilidade de se ajustar a novas informações) e tolerância a falhas (capacidade de oferecer boas respostas mesmo com falta de informação, confusão ou dados ruidosos).

Quanto ao processo de aprendizagem, mais conhecido como Algoritmo de aprendizagem, os pesos sinápticos (“força” das conexões) da rede são modificados de forma a alcançar um objetivo pré-definido.

3.1.3 Modelo neural

A unidade de uma rede neural é o neurônio, que é o principal responsável pelo processamento das informações apresentadas na entrada da rede. Um dos modelos mais conhecido de neurônio artificial é o modelo *McCulloch-Pitts* (MCP) [5, 14], no qual a saída de um neurônio assume o valor 1, se um limar for ultrapassado, e 0 caso contrário. A Figura 4 abaixo retrata o modelo neural MCP.

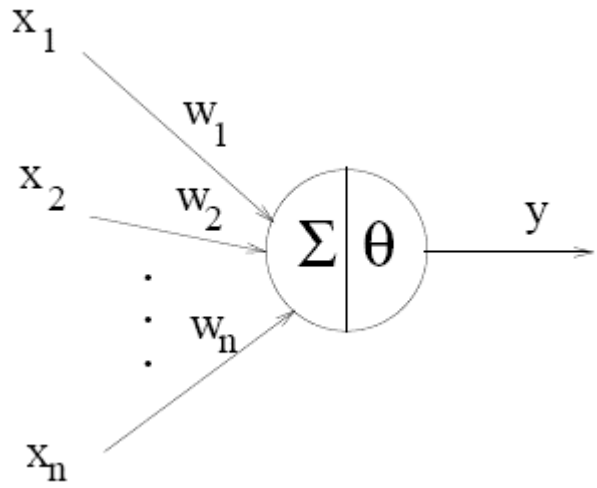


Figura 4. Neurônio MCP.

Onde o vetor X (x_1, x_2, ..., x_i) é apresentado como entrada da rede, o vetor W (w_1, w_2, ..., w_i) corresponde ao vetor de pesos, um terminal de saída Y e Θ (*Threshold*) é um limiar de ativação.

O neurônio atua quando a soma dos impulsos que ele recebe ultrapassa o seu limiar de ativação (*threshold*). O funcionamento é baseado num mecanismo simples que faz a soma dos valores de $x_i w_i$ recebidos pelo neurônio e decide se o neurônio deve ou não disparar comparando a soma obtida ao limiar do neurônio. No MCP, a ativação do neurônio é obtida através da aplicação de uma função de ativação, que ativa ou não a saída, dependendo do valor da soma das suas entradas. O neurônio MCP terá então sua saída ativa quando:

$$\sum_{i=1}^{n} x_i w_i \geq \Theta$$

As redes Perceptron com uma camada são o tipo mais antigo de redes neurais[5, 14]. Elas são formadas por uma camada única de neurônios de saída, os quais estão conectados por pesos às entradas.

As redes perceptron são geralmente treinadas por uma regra-delta (descida do gradiente) [5, 14]. Esse algoritmo de treinamento calcula os erros a partir da diferença da saída dos dados calculados e da saída desejada, e utiliza isso para ajustar os pesos. Elas têm como principal limitação a incapacidade de lidar com problemas não linearmente separáveis. Problemas linearmente separáveis são aqueles cuja solução pode ser delimitada por meio de uma reta ou hiperplano (para problemas n-dimensionais). A reta ou hiperplano é responsável por dividir o espaço das possíveis soluções em classes. A Figura 5 ilustra um problema linearmente separável, onde é possível dividir a solução em duas classes através de uma reta.

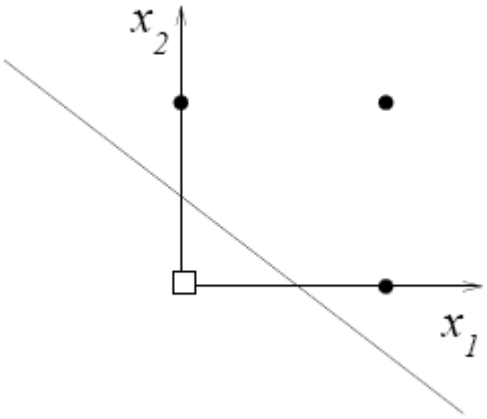


Figura 5. Um problema linearmente separável.

Entretanto, os casos mais comuns, casos reais, não são possíveis de serem solucionados por meio de uma única reta, daí a necessidade de um conjunto de retas ou regiões espaciais mais complexas para a resolução desse tipo de problema. A figura 6 exemplifica um caso.

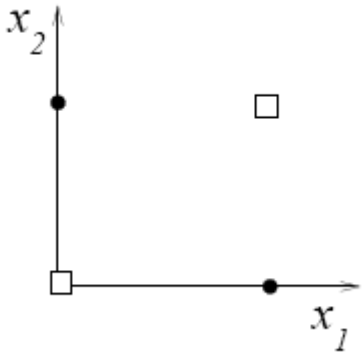


Figura 6. Um problema não linearmente separável.

3.1.4 Funções de ativação

As funções de ativação definem, segundo o modelo exposto, o valor da saída em função do campo local. As mais usadas são expostas [5, 14] a seguir, podendo ser encontradas com variações nos limites de definição.

1. Função de limiar: sua saída é 1 quando a entrada é positiva e 0 em caso contrário. Essa é a função utilizada no neurônio MCP, conforme visto na seção anterior. A figura 7 ilustra esse tipo de função.

$$f(x) = 1, \quad x \geq 0$$
$$f(x) = 0, \quad x \prec 0$$

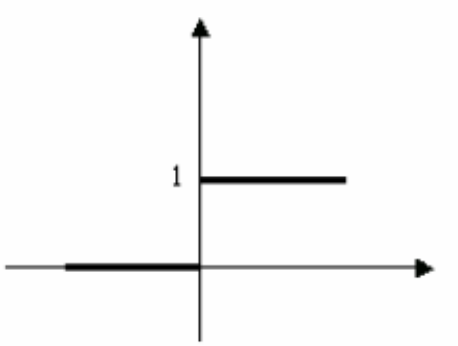


Figura 7. Função de ativação degrau.

2. Função linear por partes: produz valores constantes (*v*) em um determinado intervalo . Um exemplo de uma função linear por partes é a função abaixo, também representada na figura 8.

$$f(x) = 1, \quad x \geq 1/2$$
$$f(x) = x, \quad -1/2 \prec x \prec 1/2$$
$$f(x) = 0, \quad x \prec 1/2$$

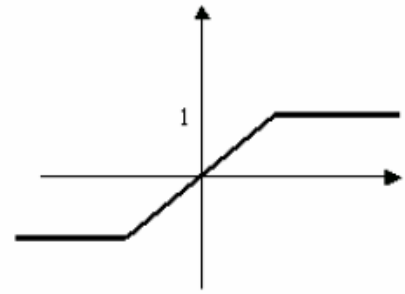


Figura 8. Gráfico da função linear.

3. Função sigmóide: o gráfico desse tipo de função apresenta forma de S. É o tipo de função mais utilizada na elaboração de RNAs devido ao fato de ser diferenciável. Um exemplo de função sigmóide é a função logística, que é dada pela equação abaixo. A Figura 9 ilustra graficamente a função sigmóide.

$$f(x) = \frac{1}{1 + \exp(-\alpha x)}$$

onde α determina a inclinação da curva.

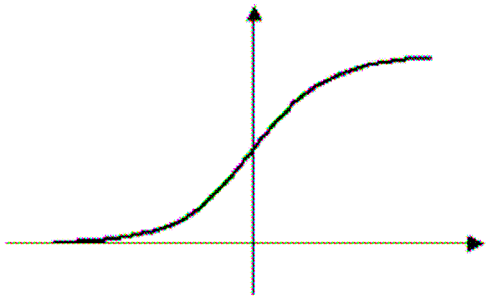


Figura 9. Gráfico da função sigmóide.

3.1.5 Arquiteturas de redes neurais

O algoritmo de aprendizagem a ser utilizado para a aquisição do conhecimento está intimamente ligado a arquitetura de uma rede neural. Essa é uma decisão que deve ser tomada nas etapas do projeto de uma RNA. Contudo, em geral, podemos identificar três tipos básicos de estrutura de redes neurais [14].

1. Redes Perceptron: rede composta de uma camada de entrada e de uma camada de saída (note que a camada de entrada não é contabilizada como camada de fato). Os nós (neurônios ou nodos) da camada de entrada são conectados aos nós da camada de saída. Essa conexão pode ser parcial, na qual nem todos os neurônios de entrada estão conectados com os neurônios de saída, ou total, no qual existe conexões entre todos neurônios entrada-saída. A figura 10 ilustra um exemplo desse tipo arquitetural.

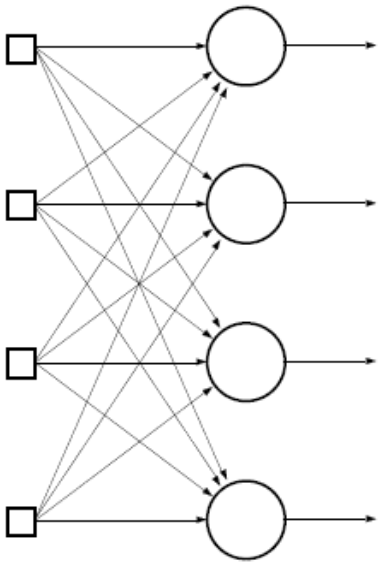


Figura 10. Rede Perceptron.

2. Redes Perceptron Multi-Camadas (*MLP – Multi-layer Perceptron*): esse tipo de arquitetura se distingue do tipo anterior apenas pelo acréscimo de camadas ocultas (ou camada intermediária), com seus respectivos neurônios ocultos. Sabe-se que o número de camadas determina, além de outras coisas, a eficiência da rede. Redes de uma única camada não conseguem resolver problemas não linearmente separáveis. Redes de camadas múltiplas conseguem. Contudo, a determinação da quantidade de camadas ocultas é realizado na base de teste-erro. Alguns autores [5] apontam que apenas uma única camada oculta é suficiente para a obtenção de uma solução satisfatória, enquanto outros estudos indicam que duas camadas, ou mesmo três delas atinjam melhores resultados. Apesar disso, o problema em questão é quem de fato determina o número de camadas a ser utilizado. Complexidade, tempo de treinamento, tempo de classificação, e possíveis outros parâmetros devem ser levados em consideração quando da escolha da quantidade de camadas ocultas. As conexões entre os neurônios das camadas ocultas e das camadas de saída, podem ser totalmente ou parcialmente conectados. A figura 11 ilustra um exemplo desse tipo de topologia.

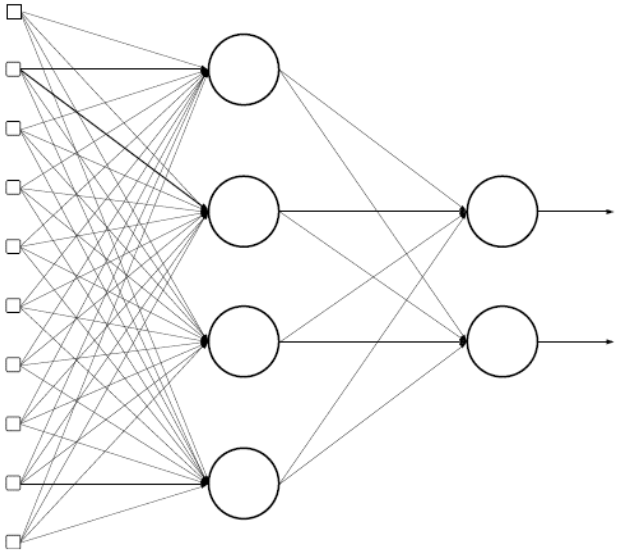


Figura 11. Rede MLP.

3. Redes recorrentes: são RNAs que utilizam um ou mais laços de realimentação. Esse tipo de rede permite a presença de memória, o que, por sua vez, introduz um comportamento dinâmico ao sistema [5, 14]. Ainda, a complexidade das conexões da rede tende a ser aumentada, se comparada as redes não-recorrentes. A realimentação pode ser local, na qual a saída de um neurônio serve de entrada para outro neurônio, ou global, na qual a saída de uma camada serve de entrada para outra. A figura 12 ilustra um exemplo dessa arquitetura de rede, onde Z^{-1} representa elementos de atraso unitário o que resulta em um comportamento dinâmico não-linear.

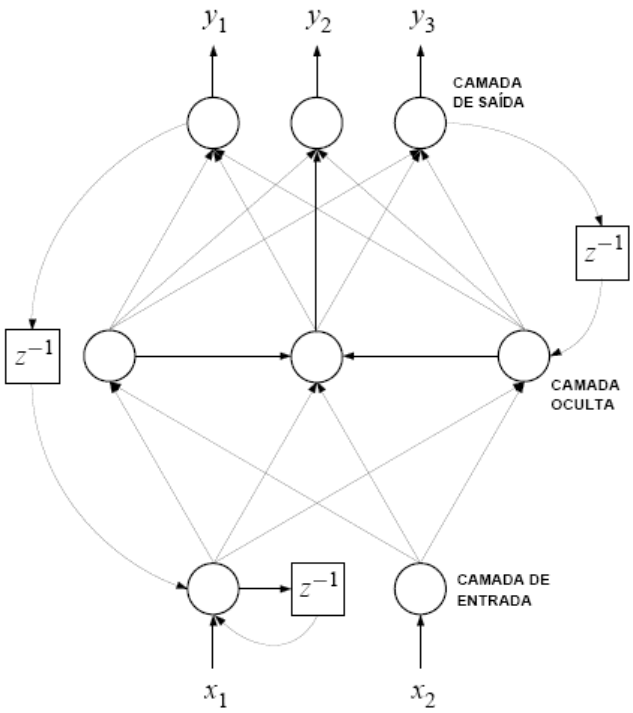


Figura 12. Rede recorrente.

3.1.6 Paradigmas de aprendizado

É sabido que RNAs extraem características de um conjunto de exemplos fornecidos à camada de entrada e processadas pela rede durante a fase de treinamento. Essa extração de características, que ao final definirá o grau de generalização da rede, é obtida pela aplicação de um algoritmo de aprendizado, que consiste em um conjunto pré-estabelecido de regras bem definidas. A aprendizagem, então, consiste na adaptação dos parâmetros livres (por exemplo, pesos e limiar de ativação) de uma rede neural. Como já dissemos, essa adaptação ocorre durante a fase de treinamento. A maneira pela qual uma RNA se relaciona com seu ambiente é, basicamente, dada de três maneiras: aprendizagem com um professor, aprendizagem sem um professor e aprendizagem por reforço [14].

1. Aprendizado supervisionado: também conhecida como aprendizagem supervisionada, nesse tipo de paradigma um conjunto de exemplos entrada-saída é fornecido à rede por meio de um professor que conhece o ambiente (domínio do problema). No entanto, a rede neural não conhece esse ambiente. Nesse processo de aprendizagem, uma medida utilizada para incorporar conhecimento à rede, é o erro, o qual consiste na diferença entre a resposta apresentada *y(t)* pela rede e a resposta desejada *d(t)*. O erro *e(t)* é usado para o ajuste dos parâmetros da rede, através de um processo iterativo.

$$e(t) = d(t) - y(t)$$

A Figura 13 mostra um diagrama que ilustra o aprendizado supervisionado.

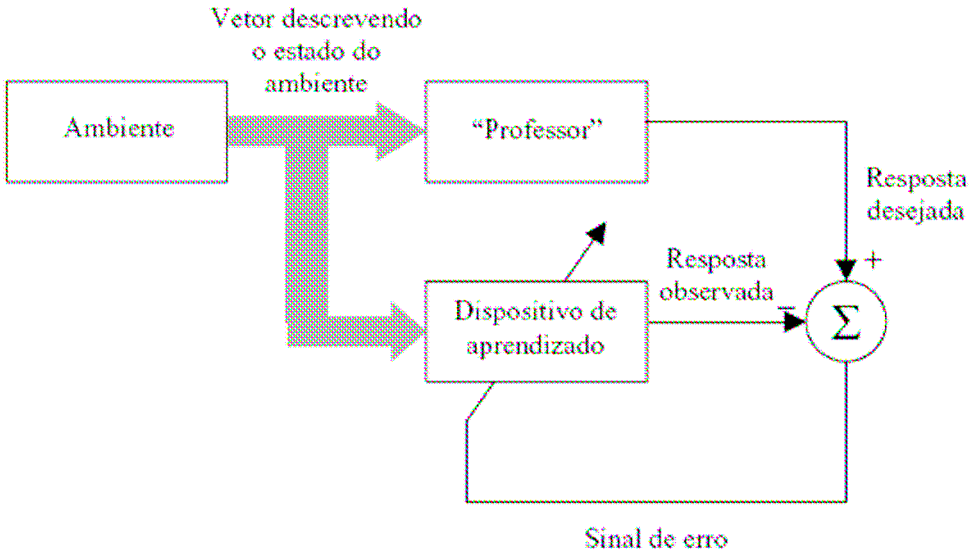


Figura 13. Aprendizado supervisionado.

2. Aprendizado não-supervisionado (auto-organização): não utiliza um professor externo que supervisiona o processo de aprendizagem. Devido a essa restrição, padrões redundantes, ou seja, com regularidade estatística, são utilizados para favorecer o processo de aquisição do conhecimento. Se essa redundância inexistir, será impossível a rede aprender. A figura 14 apresenta um diagrama da aprendizagem não-supervisionada.

Figura 14. Aprendizagem não-supervisionada.

3. Aprendizado por reforço: este tipo de aprendizado, que pode ser visto como um caso particular de aprendizagem supervisionada baseia-se em qualquer medida que possa ser dada ao sistema onde a informação de realimentação fornecida é se uma determinada saída está correta ou não, diferentemente do aprendizado supervisionado, cuja medida é baseada no critério do erro. A figura 15 ilustra a representação do diagrama de aprendizagem por reforço.

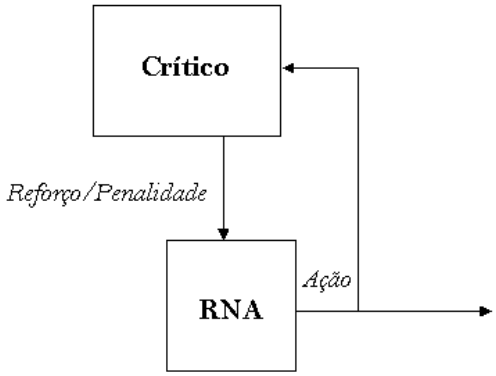


Figura 15. Aprendizagem por reforço.

3.2 Redes neurais no reconhecimento de fala

As pesquisas científicas mostram que as RNAs podem ser uma alternativa viável aos modelos estatísticos tradicionais em aplicações de RAF [21, 23, 38], se mostrando adequadas na extração de características.

Vários modelos têm sido apresentados e os mais utilizados para reconhecimento de fala são [24]:

- Redes Recorrentes;
- Multi-layer Perceptron (MLP);
- Kohonen ou redes auto-organizadas.

As redes recorrentes são sistemas em que as entradas de cada elemento consistem de uma combinação das entradas da rede com as saídas de outros elementos da rede. Essas redes são apropriadas para sistemas com entradas que podem ser representadas por valores binários.

As redes MLP são o tipo mais comum usado em RAF [24].

As redes SOM, também conhecidas como redes de Kohonen, são redes que se auto-organizam a partir da apresentação seqüencial dos vetores de entrada. Essa rede pode ser usada em reconhecimento de fala como um quantizador vetorial [21].

A rede utilizada neste trabalho foi a SOM (*Self-Organizing Maps*). A motivação dessa escolha foi o de querer refazer, com algumas pequenas diferenças, o experimento de Kohonen "datilógrafo fonético" (The "neural" Phonetic Typewriter) [21].

3.2.1 Redes SOM (Self-Organizing Maps)

As redes SOM, são mapas topográficos dos conjuntos de padrões de entrada nos quais as localizações espaciais dos neurônios indicam algum grau de semelhança. Na realidade, o desenvolvimento de redes SOM teve base na auto-organização do cérebro humano. Sabe-se que a cada área do córtex do cérebro humano encontram-se associadas várias funções. A figura 16 ilustra o particionamento do cérebro.

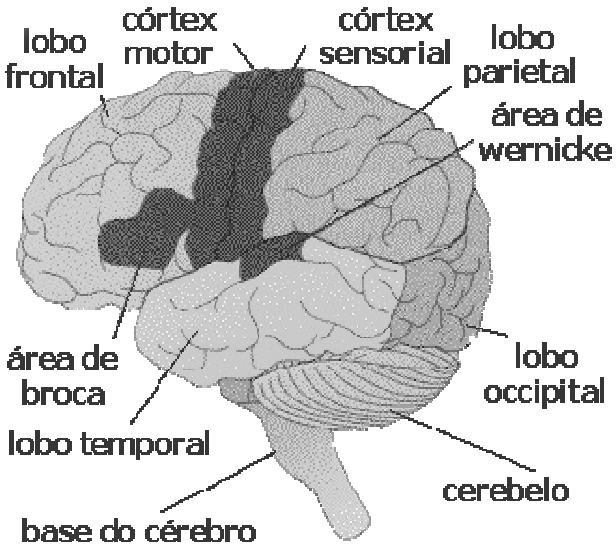


Figura 16. Regiões do cérebro humano.

Cada região do cérebro humano tem sua função bem definida. Por exemplo, o lobo occipital está ligado ao sentido de visão, assim como o lobo temporal é responsável pela audição e memória.

Organismos humanos são uma fonte de motivação para o desenvolvimento de modelos que proporcionam um arcabouço para o desenvolvimento de algoritmos de aprendizado e adaptação.

Por inspirações biológicas, as redes SOM podem ser aplicadas na resolução de diversos tipos de problemas, principalmente em problemas de reconhecimento de padrões e categorização de dados em que as classes não são conhecidas antecipadamente. A idéia de utilizarmos a rede SOM para reconhecimento de padrões é buscar agrupar padrões que compartilham características comuns apenas em uma classe. Para realizar este agrupamento (*clusters*), um algoritmo SOM necessita encontrar características significativas nos conjuntos de dados de entrada, sem o auxilio de um professor. A utilização deste algoritmo só é possível em casos em que se houver redundância nos conjuntos de dados de entrada. A redundância dos dados de entrada fornece conhecimentos à rede sobre semelhanças e diferenças entre estes dados, enquanto que a ausência da redundância torna impossível encontrar similaridade nas características dos padrões.

3.2.2 Arquitetura Self-Organizing Maps (SOM)

As redes SOM foram idealizadas por Kohonen [20]. Elas são baseadas no aprendizado competitivo, onde os neurônios competem entre sim para ver quem é a unidade vencedora e conseqüentemente terá o direito de atualizar seus pesos. O principal objetivo é mapear o conjunto de entradas em um mapa topográfico, geralmente unidimensional ou bidimensional. Uma ilustração de uma possível arquitetura de uma rede SOM é ilustrada na figura 17.

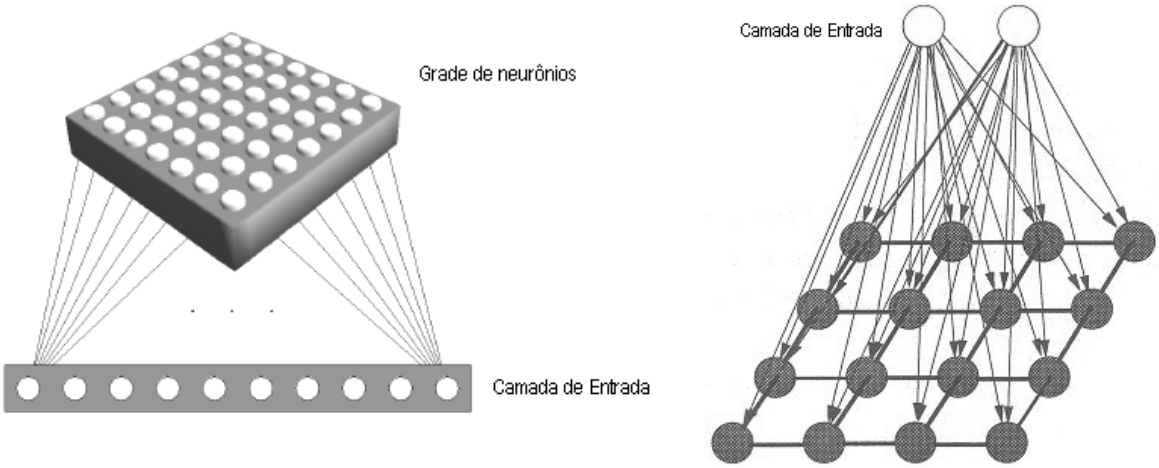


Figura 17. Exemplo da arquitetura de uma rede SOM.

Os nodos que compõem a rede SOM se arranjam em forma de grade. Podemos encontrar grades unidimensionais, bidimensionais e com três ou mais dimensões, mas as mais comuns são as bidimensionais. Os nodos de saída de uma grade bidimensional são organizados em forma de linhas e colunas, as quais cada nodo recebe todas as entradas e funciona como um extrator de características.

Os neurônios que constituem o mapa topográfico são ativados seletivamente de acordo com os vários padrões de entrada durante o processo de aprendizado. Esta ativação seletiva implementa a quantização vetorial dos vetores que contém os padrões de entrada. Desse modo, uma rede SOM é caracterizada por formar um mapa topográfico dos padrões de entrada, no qual as coordenadas das unidades são indicativos de atributos estatísticos intrínsecos contidos nestes padrões [5, 14, 20]. De forma que, o estado de ativação de um neurônio é determinado pela distância entre seu peso e o vetor de entrada. A função de ativação, mais usada, da rede SOM é baseada na medida de distância euclidiana, conforme citada abaixo:

$$d_j = \sum \left(x_i(t) - w_{ij}(t) \right)^2$$

Onde:

$xi(t) = valor\ do\ neurônio\ de\ entrada\ i\ no\ tempo\ t.$

$wij(t) = peso\ sináptico\ entre\ o\ neurônio\ de\ entrada\ i\ e\ o\ nerurônio\ de\ saída\ \ j\ no\ tempo\ t.$

3.2.3 Algoritmo de treinamento

A inicialização dos pesos é através da atribuição de valores pseudo-aleatórios pequenos, de forma a impedir a imposição de uma pré-ordenação qualquer. O algoritmo pode-se ser dividido em 3 estágios: competição, cooperação e adaptação, descritos a seguir [14].

3.2.3.1 Competição

Neurônios competem entre si onde: para cada padrão de entrada apresentado, é calculado seu valor de saída através do cálculo de uma função matemática. O neurônio vencedor é aquele que apresentou a menor distância do padrão de entrada.

3.2.3.2 Cooperação

O neurônio vencedor determina a localização de determinado padrão no mapa topográfico, dando origem a um cluster. Isso só é possível, escolhendo inicialmente uma dimensão para o mapa, o sistema de coordenadas nesta dimensão e uma métrica de distância entre as unidades. Podemos usar para os casos de mapas unidimensionais e bidimensionais, as seguintes métricas:

$$d(X_{i1}, X_{i2}) = | X_{i1} - X_{i2} |, \text{ para mapas unidimensionais}$$
$$d(X_{i1}, X_{i2}) = || r_{Xi1} - r_{Xi2} ||^{1/2}, \text{ para mapas bidimensionais}$$

Conseqüentemente, dado um neurônio vencedor X_{i*}, calcula-se inicialmente a distância entre ele e cada um dos demais $d(X_{i*}, X_i)$, onde $i \neq i^*$. A intensidade da interação entre a unidade vencedora e suas vizinhas é dada por uma função de vizinhança h (d, t).

A função Gaussiana é a mais usada, devido à amplitude ser atenuada com o aumento da distância lateral. A função gaussiana é ilustrada na figura 18.

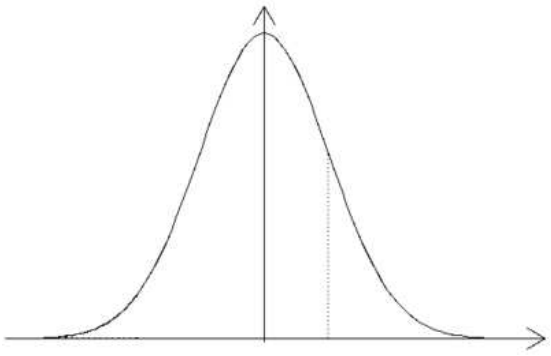


Figura 18. Função gaussiana.

$$h(d,t) = \exp\left(-\frac{d^2}{2\rho^2(t)}\right)$$

Onde $\rho(t)$ é a função largura efetiva da vizinhança topológica, que é diminuída com o tempo, reduzindo a intensidade da interação cooperativa com os neurônios vizinhos:

$$\rho(t) = \rho_0 \exp\left(-\frac{t}{\tau_1}\right)$$

Onde ρ_0 sendo a largura efetiva inicial (em t = 0) e τ_1 a constante de tempo de diminuição.

Dessa forma, quando é terminado o cálculo das distâncias entre o neurônio vencedor e todos os outros neurônios, é calculado o valor da função de vizinhança para cada um deles, determinando assim as intensidades de cooperação entre eles. A figura 19 ilustra a cooperação

entre neurônios em duas diferentes topologias de vizinhança, (a) hexagonal e (b) retangular. Os números 0, 1 e 2 que se encontram na figura são delimitações da atuação de diferentes raios de vizinhança.

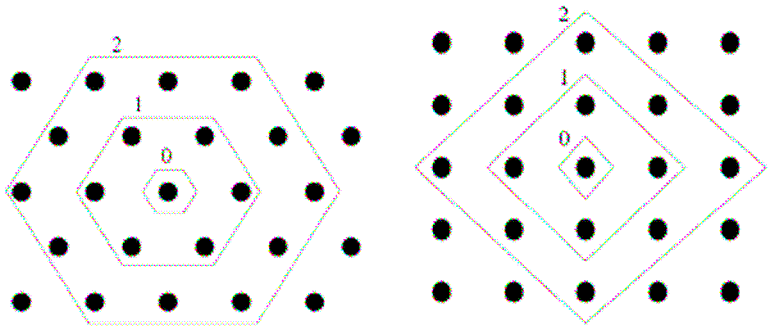


Figura 19. Cooperação entre os neurônios em duas diferentes vizinhanças (a) hexagonal (b) retangular.

3.2.3.3 Adaptação

Neste estágio, a adaptação dos pesos é feita de forma que o peso do neurônio vencedor é incrementado, assim como os de sua vizinhança. Essa atualização leva em consideração parâmetros como taxa de aprendizado e um raio de vizinhança. A equação de atualização do peso da conexão entre o neurônio *j* e o neurônio *i* é dada por:

$$W_j(t+1) = W_j(t) + \eta(t) * (X(t) - W_j(t))$$

Onde $W_j(t)$ é o vetor de peso do neurônio *j* no instante *t*. $\eta(t)$ é a função taxa de aprendizado, que começa com um valor inicial alto η_0 e diminui exponencialmente com o tempo:

$$\eta(t) = \eta_0 \exp(-\frac{t}{\tau_2})$$

O processo de adaptação da rede SOM ocorre em duas fases: fase de ordenação e fase de convergência [14]:

Na fase de ordenação, ocorre a ordenação topológica dos vetores de pesos, que são inicialmente orientados de forma aleatória. Nesta etapa, que dura tipicamente cerca de 1000 iterações, deve-se usar uma taxa de aprendizado alta e uma largura de vizinhança alta. Para a taxa de aprendizado, pode-se assumir uma taxa inicial η_0 = 0.1, decaindo para 0.01 após 1000 iterações ($\tau 2$ = 1000). Quanto à largura da vizinhança, deve-se incluir inicialmente quase todas as unidades da rede (ρ_0 é o "raio" do mapa), decaindo para somente algumas unidades vizinhas ou somente para a unidade vencedora ao final. Para isso, pode-se usar a constante de tempo τ_1 = 1000 / log ρ_0.

Na fase de convergência, o processo adaptativo faz a sintonia fina do mapa de atributos, fazendo uma quantização estatística do conjunto de entradas. A duração desta etapa é de aproximadamente 500 vezes o número de neurônios que compõem o mapa. A taxa de

aprendizado deve ser mantida fixa, com um valor baixo (η_0 = 0.01). A vizinhança deve ser mantida constante, em aproximadamente uma ou nenhuma unidade vizinha.

O algoritmo padrão da rede SOM é basicamente o seguinte:

1. Inicialização dos pesos:

Antes de qualquer operação na rede, inicializar todos os pesos das sinapses com valores pseudo-aleatórios.

2. Apresentar uma nova entrada.

3. Calcular as distâncias entre os neurônios da camada competitiva e os neurônios da camada de entrada:

Para cada neurônio da camada competitiva, calcular a distância d_j entre o neurônio j da camada competitiva e cada neurônio i da camada de entrada.

$$d_j = \sum \left(x_i(t) - w_{ij}(t) \right)^2$$

Onde:
$xi(t)$ = *valor do neurônio de entrada i no tempo t.*
$wij(t)$ = *peso sináptico entre o neurônio de entrada i e o nerurônio de saída j no tempo t.*

4. Selecionar o neurônio vencedor

Seleção do neurônio j cuja distância d_j seja a menor possível.

5. Atualização de pesos da vizinhança do neurônio vencedor

A Atualização de pesos é realizada no próprio neurônio vencedor bem como numa área que abrange

$$W_j(t+1) = W_j(t) + \eta(t) * (X(t) - W_j(t))$$

Onde:
$\eta(t)$ = *taxa de aprendizagem que deve ser decrementada no decorrer do treinamento.*

6. Voltar ao passo 2 para novo treinamento.

Pode ocorrer, ainda, após o término da fase do aprendizado, alguns padrões estarem com limites muito próximos ou sobrepostos. Para tentar resolver isso, Kohonen apresentou três outros métodos de quantização que chamou de LVQ1, LVQ2 e LVQ3 [20].

Os métodos de quantização do vetor de aprendizado (*Learning Vector Quantization*) permitem que, através de processos iterativos, sejam reforçados, ou inibidos, os pesos de determinados neurônios da rede, caso eles proporcionem uma resposta correta, ou incorreta, respectivamente, a um padrão de entrada.

3.2.4 Aplicação da rede SOM no reconhecimento de fala

Um caso de bastante sucesso da aplicação de uma RNA na resolução do problema de RAF foi a "datilógrafo fonético" (The "neural" Phonetic Typewriter) [21] implementado pelo pesquisador da universidade de Helsinki, Teuvo Kohonen. Esse sistema converte a linguagem falada, nos idiomas finlandês e japonês, em um texto escrito, através do reconhecimento de fonemas em fala contínua [21].

Nesse experimento, Kohonen definiu uma etapa de pré-processamento na qual englobou desde a captação do sinal acústico até a extração dos vetores acústicos que contêm os parâmetros retirados da locução. As características mais importantes desta etapa foram:

- Cancelamento do ruído de ambiente (Microfone);
- Utilizou um filtro passa baixa centrado em 5.3Khz.
- Conversor de 12 bits (analógico para digital) e taxa de amostragem de 13KHz;
- FFT com 256 pontos, computados a cada 9.83ms usando uma janela de Hamming com 256 pontos;
- Normalização do vetor resultante;

Os valores do espectro retirados com a FFT foram guardados em um vetor de 15 dimensões reais. Kohonen imaginou que os espectros de diferentes fonemas da fala ocupam diferentes regiões do espaço, então eles podem ser detectados por alguns tipos de métodos de discriminação multi-dimensional. As distribuições do espectro de diferentes classes de fonemas sofrem sobreposição, então não é possível distinguir fonemas com 100% de certeza. Daí a necessidade da sobreposição dos quadros. Esse assunto será discorrido melhor no próximo capítulo.

Kohonen percebeu por meio do experimento que existiam alguns fonemas que eram mais difíceis de se classificar, devido à resposta de transientes ser muito rápida. Baseado nesse fato, ele definiu mapas auxiliares para tratar desses fonemas. De cada fonema foram retiradas 50 amostras e utilizou um mapa bidimensional de dimensões 8x12.

O experimento chegou a uma taxa de acerto de 92 a 97% de acerto nas conversões.

A figura 20 mostra como ficou o mapa fonético, depois que a rede foi treinada com os conjuntos dados de entrada. Cada célula, neurônio, é marcada pelo rótulo do fonema gerado Células respondendo ao mesmo fonema formam domínios, os quais são agrupados de acordo com a similaridade entre os animais.

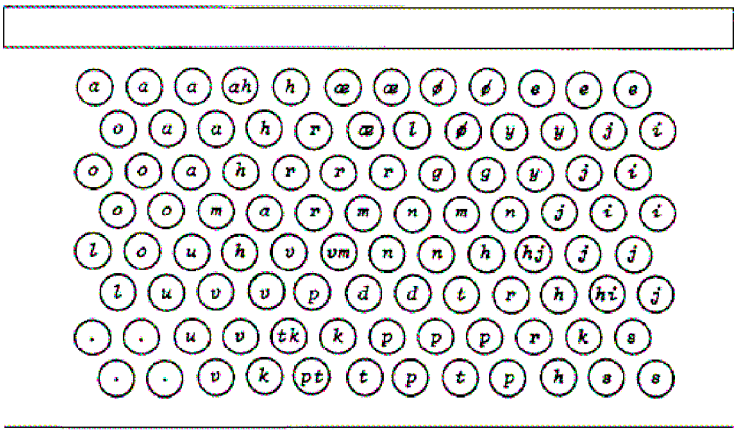


Figura 20. Neurônios, mostrados como círculos, que foram classificados como fonemas na melhor resposta que a rede proporcionou [21].

Capítulo 4

Pré-processamento da fala

Para se obter um bom desempenho na tarefa de reconhecimento de fala, é de extrema importância que a etapa de pré-processamento do sinal seja bem modelada. A etapa de pré-processamento engloba desde a captação do sinal acústico até a extração dos vetores acústicos que contêm as características retiradas da locução [28]. Para efeitos didáticos, dividiu-se a etapa de pré-processamento em sub-etapas, as quais serão comentadas abaixo.

4.1 Aquisição da fala

A primeira etapa de um reconhecedor consiste em realizar a aquisição do sinal da fala através de um transdutor. Nessa etapa, é realizada a digitalização do sinal da fala, que engloba as seguintes operações, conforme é mostrado na figura 21 [26].

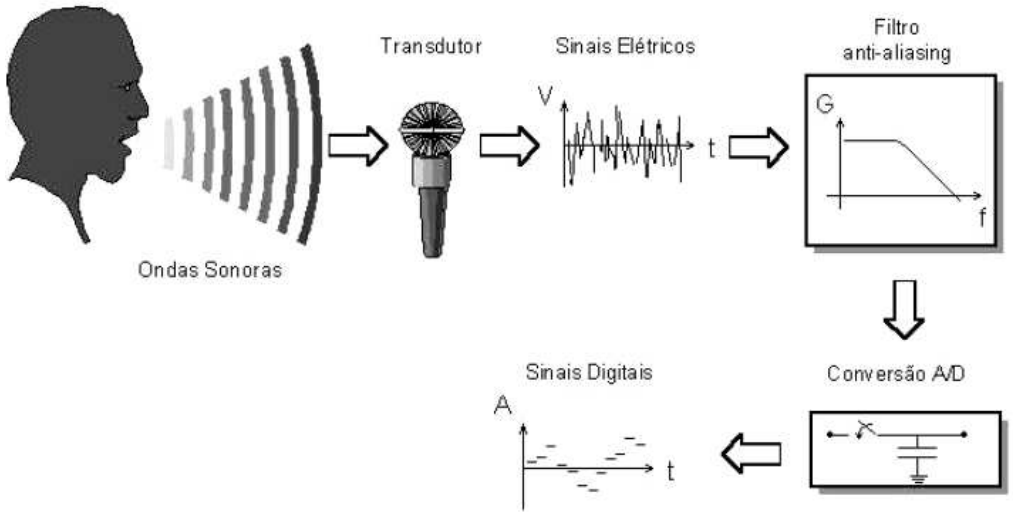


Figura 21. Processo de aquisição do sinal de fala.

4.1.1 Transdução do sinal da fala

A transdução do sinal da fala em modo acústico em sinal elétrico é necessária, pois o microcomputador é um dispositivo eletrônico, ou seja, funciona mediante a presença de sinais elétricos. Na fase de transdução os equipamentos mais freqüentemente utilizados são: microfones ou telefones, cujo objetivo consiste em transformar um sinal acústico em um sinal elétrico.

4.1.2 Filtragem do sinal da fala

A filtragem é realizada com o intuito de estreitar a largura de banda do sinal da fala de modo a tentar eliminar possíveis ruídos, que possam estar presentes no sinal da fala.

4.1.3 Conversão A/D

Realiza a conversão A/D do sinal da fala, ou seja, transforma o sinal da fala analógico em sinal da fala digital, para possibilitar o processamento do mesmo através do microcomputador.

Nesta etapa, são escolhidos o ganho, a taxa de amostragem que assegure o não aparecimento do efeito aliasing [7, 28] e a precisão usados para a gravação do sinal da fala.

4.2 Pré-processamento

Depois da sub-etapa de aquisição do sinal, vem a fase de pré-processamento propriamente dita, conforme é mostrado na figura 22. Algumas vezes, pode-se ainda utilizar outros filtros antes da fase de pré-ênfase. Um exemplo seria a remoção do nível DC, que serve para remover possíveis sinais espúrios advindos de problemas com o transdutor utilizado.

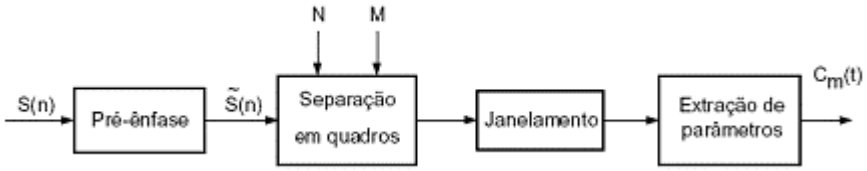


Figura 22. Pré-processamento e extração de parâmetros.

4.2.1 Pré-ênfase

A finalidade do filtro utilizado na pré-ênfase é compensar a atenuação nas altas freqüências do sinal de fala gerado pelo processo de produção da fala na glote, tornando o seu espectro de freqüência mais plano [8, 16, 21, 23, 41]. A resposta em freqüência de um filtro de pré-ênfase é mostrada na figura 23. A função de transferência do filtro de pré-ênfase é dada pela equação abaixo:

$$H(z) = 1 - \alpha z^{-1}, \quad 0.9 \le \alpha \le 1.0$$

Neste caso, a saída da pré-ênfase, ~s(n), está relacionada à entrada, s(n), pela equação:

$$\sim s(n) = s(n) - \alpha s(n-1)$$

Onde:
H(z) – função de transferência do filtro
z - freqüência

s(n) – sinal
α – coeficiente de pré-ênfase

Neste trabalho adotou-se α igual a 0.95.

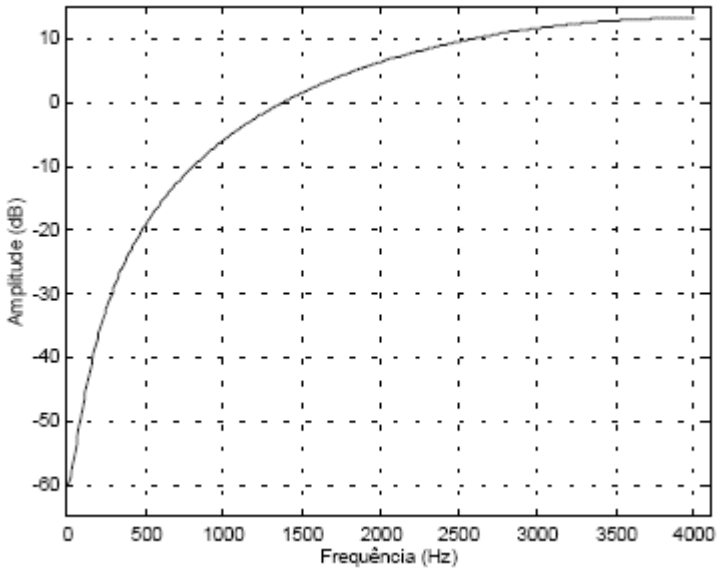


Figura 23. Resposta em freqüência do filtro de pré-ênfase para α = 0.95.

A figura 24 ilustra o espectro de um sinal de fala (a) sem o uso do filtro de pré-ênfase e (b) com o uso de um filtro de pré-ênfase.

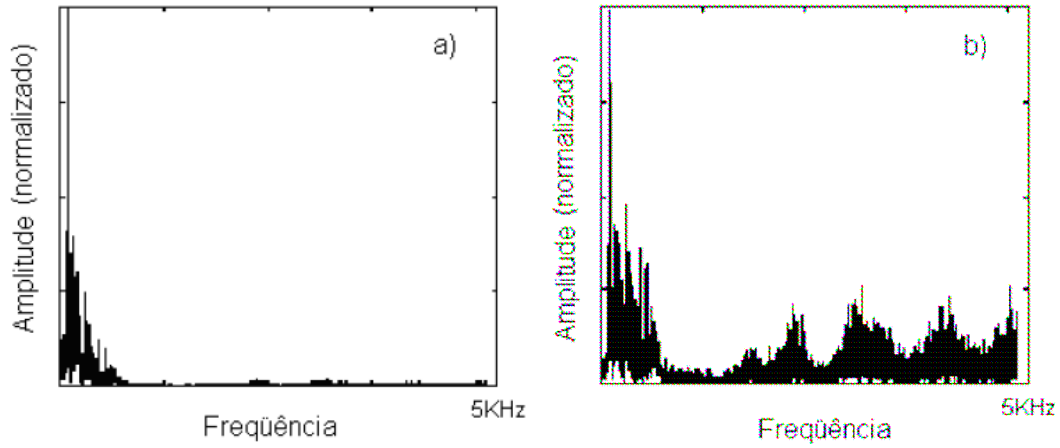


Figura 24. Espectro de freqüências para um sinal de fala a) sem pré-ênfase e b) com pré-ênfase.

4.2.2 Divisão do sinal em quadros e janelamento

Logo após as fases da aquisição e de pré-ênfase do sinal da fala passa-se à etapa da divisão do sinal em quadros e janelamento [8, 21, 27]. Nesta etapa são extraídos quadros de N amostras a partir do sinal s(n), sendo os quadros adjacentes separados por M amostras. Tal divisão é extremamente importante devido ao fato de um sinal de fala ser estatisticamente variante no tempo. As divisões em pequenos segmentos variam de 10 a 25 ms. Essa divisão é possível, devido a se assumir que o sinal de fala é invariante no tempo sobre um intervalo menor que 25

ms [35]. Fisicamente, isto significa que o formato do trato vocal permanece constante sobre este intervalo.

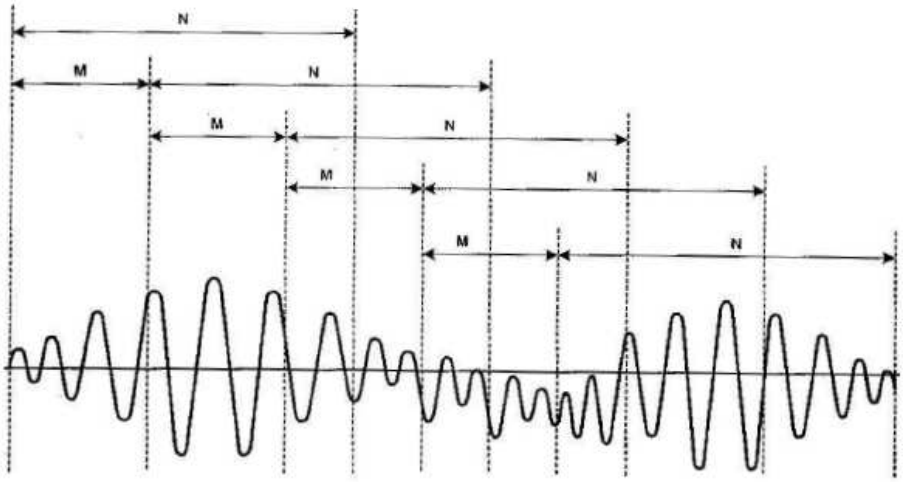


Figura 25. Divisão do sinal em quadros.

Ponderações podem ser feitas entre a relação M/N. Se M < N, então os quadros adjacentes sobrepõem-se, como na figura 25, e os valores estimados serão correlatados de quadro para quadro; se M << N, então as estimativas de um quadro para o outro serão bem suaves. Por outro lado, se M ≥ N, não há sobreposição entre quadros adjacentes.

A divisão em quadros é feita através do janelamento do sinal de fala. Aplicar uma janela a um sinal no domínio do tempo é de forma simples fazer uma operação de multiplicação do sinal pela função que representa a janela. A multiplicação no domínio do tempo é equivalente a convolução no domínio da freqüência, de forma que espectro de um sinal janelado é a convolução do espectro do sinal original com o espectro da janela [15]. Dessa maneira, a operação de janelamento modifica a forma do sinal tanto no domínio do tempo quanto no da freqüência.

A utilização do janelamento é uma forma se conseguir aumentar as informações espectrais de um sinal amostrado [2]. Esse "aumento" das informações é decorrente da minimização das margens de transição em forma de ondas truncadas e de uma melhor separação do sinal de pequena amplitude de um sinal de grande amplitude com freqüências muito próximas uma da outra. Muitos tipos diferentes de janelas podem ser utilizados. A tabela 1 mostra os tipos mais conhecidos de janelas e suas respectivas equações matemáticas, onde N é o número de pontos da janela e n o índice avaliado. A figura 26 ilustra o formato que essas janelas assumem.

Tabela 1. Equação matemática para tipos mais conhecidos de janelas.

Janela	Equação Matemática
Retangular	$\begin{cases} 1 & 0 \le n \le N-1 \\ 0 & n > N-1 \end{cases}$
Bartlett	$\begin{cases} 1 - \left\lvert \dfrac{n - \frac{1}{2}N}{\frac{1}{2}N} \right\rvert & 0 \le n \le N-1 \\ 0 & n > N-1 \end{cases}$
Blackman	$\begin{cases} 0{,}42 - 0{,}5\cos\left(\dfrac{2\pi n}{N-1}\right) + 0{,}08\cos\left(\dfrac{4\pi n}{N-1}\right) & 0 \le n \le N-1 \\ 0 & n > N-1 \end{cases}$

Hamming	$\begin{cases} 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N-1}\right) & 0 \le n \le N-1 \\ 0 & n > N-1 \end{cases}$
Hanning	$\begin{cases} 0{,}5 - 0{,}5\cos\left(\frac{2\pi n}{N-1}\right) & 0 \le n \le N-1 \\ 0 & n > N-1 \end{cases}$
Welch	$\begin{cases} 1 - \left(\frac{n - \frac{1}{2}N}{\frac{1}{2}N}\right)^2 & 0 \le n \le N-1 \\ 0 & n > N-1 \end{cases}$

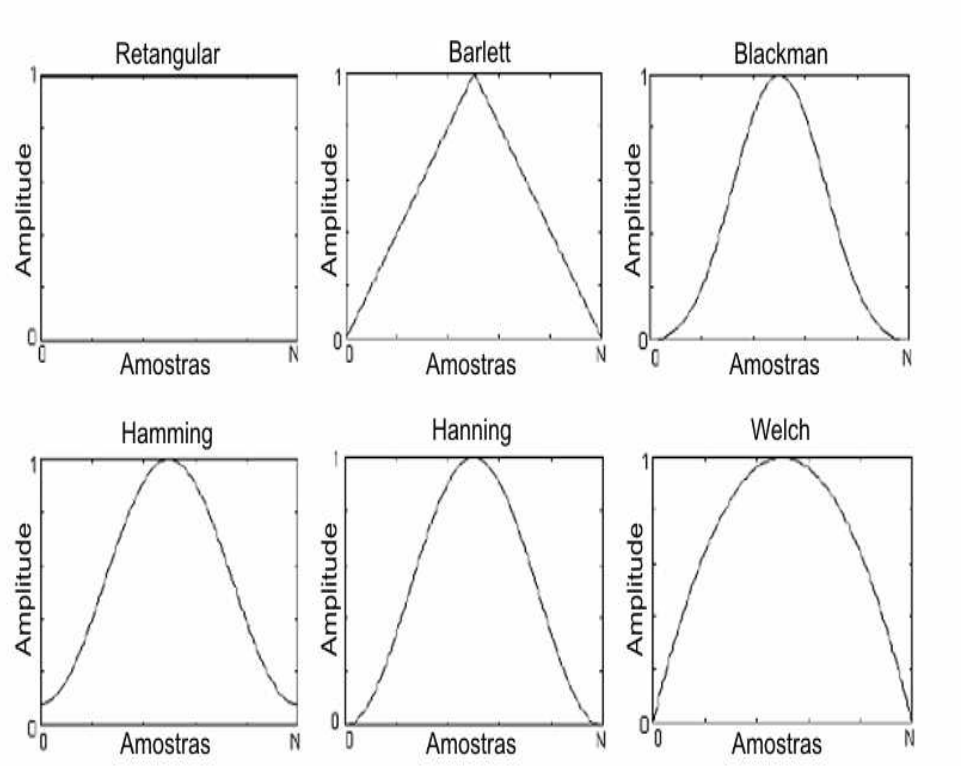


Figura 26. Formato dos tipos mais conhecidos de janelas.

Cada tipo de janela ilustrada acima tem um efeito final diferente quando aplicada ao sinal. Por exemplo, se aplicarmos uma janela retangular a um sinal é igual à não utilizar qualquer janela. Ela pode ser utilizada para a análise de transientes que possuem uma duração menor do que a da janela em análise. Já a janela de Hanning é útil para a análise de transientes maiores que o tempo de duração da janela e também para aplicações de objetivos gerais. Por sua vez, a janela de Hamming é bastante parecida com a de Hanning, mas com uma pequena diferença no domínio do tempo, a janela de Hamming não se aproxima do zero como a janela de Hanning.

Normalmente, a janela de *Hamming* é mais utilizada nos sistemas de RAF, por apresentar características espectrais interessantes e suavidade nas bordas [1, 21]. As sucessivas janelas usualmente possuem uma região de sobreposição, para que a variação dos parâmetros entre janelas adjacentes seja mais gradual, e para que a avaliação dos elementos localizados nos extremos de alguma janela não seja prejudicada. A figura 27 ilustra uma janela de *Hamming* de 20ms com superposição de 50%.

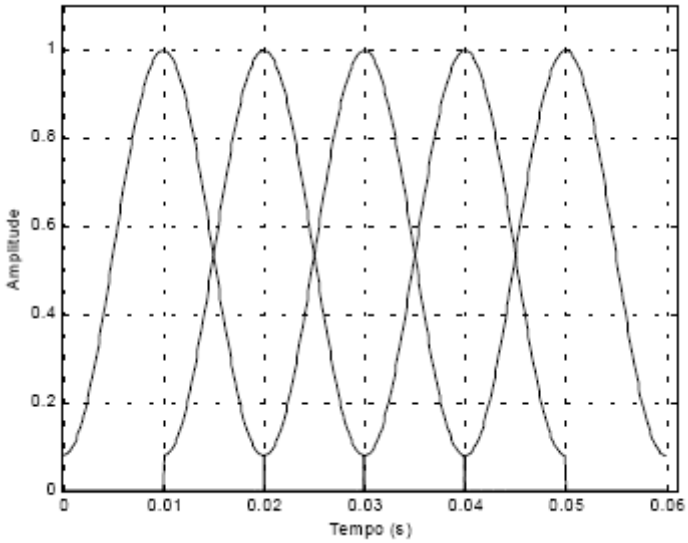


Figura 27. Janelas de *Hamming* de 20ms com superposição de 50%.

A quantidade de superposição é dada pela equação:

$$\text{sup}\,erposicão(\%) = \left(\frac{T_w - T_f}{T_w}\right) * 100$$

Onde T_W é a duração da janela, e T_F é a duração do bloco. Na prática, estas medidas são ajustadas aos pares. Por exemplo, para blocos de 20ms, costuma-se usar janelas de 30ms; enquanto para blocos de 10ms, é comum o uso de janelas com duração de 20ms.

4.2.3 Endpoints

Outro fator de grande importância em um reconhecedor de fala é a necessidade de se determinar de forma eficiente e precisa, o início e o final de uma locução, com a finalidade de excluir os silêncios que não trazem nenhuma informação adicional sobre a locução a ser reconhecida, evitando carga computacional e economizando tempo, além de servir como marco de início e fim de um segmento de fala [1, 3, 8]. Estes fatores são de grande importância, pois minimizam a carga do reconhecedor, visto que o mesmo não terá que processar atributos de reconhecimento de trechos sem informação de fala.

O processo de determinação dos limites de uma palavra pode representar um aspecto crucial na performance de um sistema de RAF. De fato, trata-se de um problema extremamente complicado, particularmente para palavras que começam ou acabam em fonemas de baixa energia, como fricativas ou nasais ou, ainda, palavras que possuem oclusivas, pois o silêncio que precede a oclusiva pode ser confundido com o fim da palavra.

Os endpoints são determinados pelo primeiro quadro onde o sinal de fala realmente se inicia e pelo último quadro do sinal de fala. A determinação dos endpoints deve ser feita de forma cuidadosa, pois os mínimos erros nesta estimação podem degradar o processo de reconhecimento.

Através de um classificador de fala pode-se diferenciar entre sons sonoros, surdos ou silêncio. Neste trabalho foi utilizado um classificador baseado nas características temporais do

sinal. O algoritmo é baseado na estimação da amplitude média do sinal. Os 100ms iniciais e 30ms finais da locução são considerados como ruído de fundo. Com este algoritmo, determinou-se o final e início da locução a ser processada pelo sistema de reconhecimento, ou seja, os endpoints.

4.3 Extração de características do sinal de fala

A seleção da melhor representação paramétrica dos dados acústicos é uma tarefa importante no projeto de qualquer sistema de reconhecimento de fala. Os principais objetivos na seleção de uma representação paramétrica são a eliminação de informações irrelevantes com respeito à análise fonética dos dados e a enfatização dos aspectos do sinal de fala que contribuem significativamente para a detecção das diferenças fonéticas. Além disso, quando uma quantidade considerável de informações de referência deve ser armazenada, a armazenagem compacta da informação torna-se uma consideração prática importante.

Dentre as técnicas mais comuns de análise espectral encontram-se os métodos de banco de filtros (*Filter Bank*), transformada rápida de Fourier (FFT - *Fast Fourier Transform*), análise homomórfica (*cepstrum*) e os métodos de codificação por predição linear (LPC - *Linear Predictive Coding*) [27, 28].

Os métodos de banco de filtros, transformada rápida de Fourier e a de codificação por predição linear foram largamente usados para a extração de informação espectral da fala. Entretanto, elas apresentam restrições. A mais pronunciável é a de não resolver as características do trato vocal. O método cepstrum trata essa restrição. A idéia por trás do cepstrum é a obtenção de uma relação linear entre a excitação da energia do sinal e(n) com o filtro utilizado v(n). A literatura reporta que os melhores resultados, na maioria dos casos, são os obtidos pelo método cepstrum [16, 35].

Os coeficientes mel-cepstrais, advindos do método espectral cepstrum, são obtidos pela representação em freqüência na escala mel [16, 35, 41]. Um mel é uma unidade de medida da freqüência percebida de um tom. Não corresponde linearmente às freqüências físicas do tom, à medida que o sistema auditivo humano aparentemente não percebe a freqüência de maneira linear. Assim, a escala mel foi construída de tal maneira que os incrementos iguais na escala mel correspondem a incrementos subjetivos iguais em freqüência. Para obtermos o valor de uma freqüência em mel é só fazer o logaritmo na base 10 da freqüência.

Os coeficientes mel cepstrais são obtidos a partir de cada janela do sinal, depois de realizados os seguintes processamentos:

- Aplicação do banco de filtros triangulares em escala mel e cálculo do logaritmo da energia de saída de cada filtro. A aplicação do logaritmo é necessária para a obtenção do cepstro. São utilizados geralmente 20 filtros de formato triangular, como mostrado na Figura 28. O espaçamento e a largura de faixa dos filtros usados são os tabelados em [12];
- Cálculo da transformada discreta inversa do co-seno (*DCT*) do vetor do logaritmo da energia de saída do banco de filtros através da equação:

$$c(n) = \sum_{k=1}^{M} (\log_{10} X(k)) \cos\left(\frac{n(k-0.5)\Pi}{M}\right), \quad 1 \le n \le N$$

Onde:
n – índice dos coeficientes mel cepstrais;
N – número total de coeficientes mel cepstrais;

k - índice do filtro
M – número total de filtros
X(k) – energia de saída do filtro *k*.

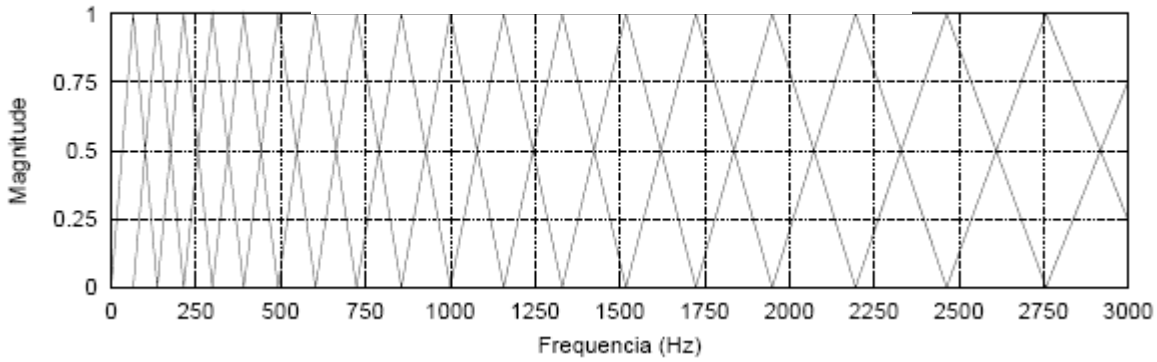


Figura 28. Banco de filtros triangulares na escala mel, incremento de 100 Hz.

Quanto mais detalhada for a extração das características do sinal de fala, melhor será o resultado do sistema de reconhecimento. Resultados publicados na literatura mostram que o emprego de coeficientes mel-cepstrais e energia e da primeira e segunda derivada desses parâmetros (mel-cepstrais e energia) melhoram sobremaneira a taxa de acerto em reconhecimento de fala [1, 16, 35]. A primeira e a segunda derivada dos coeficientes cepstrais são obtidas pelas seguintes equações:

$$Dc_i^1(n) = \sum_{k=-K}^{K} \left(\frac{kc_{i-k}(n)}{2K+1} \right)$$

$$Dc_i^2(n) = \sum_{k=-K}^{K} \left(\frac{kDc_{i-k}^1(n)}{2K+1} \right)$$

Onde:
i – índice do quadro do sinal;
n – índice do coeficiente mel cepstral;
K – número de quadros utilizados no cálculo das derivadas;

O cálculo da energia de cada quadro da amostra apresentado, definiu-se como sendo a média dos valores das amostras contidas nos quadros, obtida segundo a equação:

$$E(x) = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Onde:
x – índice do quadro do sinal
N – número de amostras retiradas do quadro
i – índice do número de amostras

O cálculo da primeira e segunda derivada da energia é feito pelas próprias equações dos cálculos da primeira e segunda derivada dos coeficientes mel-cepstrais, substituindo o vetor de coeficientes *c*(*n*) pela energia do quadro.

4.4 Quantização vetorial

Sistemas discretos requerem a representação no domínio discreto do conjunto de parâmetros espectrais. Uma forma eficiente de se discretizar o conjunto de parâmetros é através da quantização vetorial.

O procedimento mais habitual é, a partir da base de dados, definir um dicionário de códigos seguindo um critério de otimização. O dicionário de códigos é acessado a cada tarefa de quantização de um conjunto de parâmetros. A determinação do vetor mais adequado é resultado de uma busca exaustiva da menor distância entre o vetor de parâmetros e o vetor do dicionário de códigos. Várias medidas de distorção podem ser utilizadas, porém, a mais comum é a medida de distorção euclidiana [3]. Muitas vezes ainda é preciso utilizar um processo de otimização, o qual faz o levantamento dos vetores que melhor representam todo um conjunto de parâmetros.

Neste trabalho, a solução aplicada foi baseada na que Kohonen utilizou no "datilógrafo fonético" [21]. A principal diferença foi a técnica usada na extração dos parâmetros que compõem o vetor de características do sinal da fala. Kohonen empregou uma RNA intitulada SOM com a função de ser o quantizador vetorial, no qual o conjunto de parâmetros foi extraído com o auxílio da FFT de 256 pontos aplicado a uma janela de Hamming, enquanto que neste trabalho optou-se pelos coeficientes cepstrais e de energia, além das suas primeira e segunda derivadas. A Figura 29 ilustra melhor o processo de quantização.

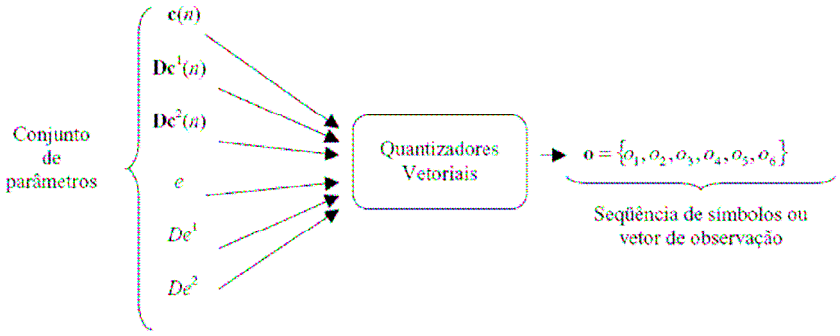


Figura 29. Processo de quantização vetorial.

Capítulo 5

O sistema desenvolvido

A necessidade de um sistema que tenha a capacidade de lidar com as dificuldades encontradas no reconhecimento da fala é uma motivação forte para a sua criação. O projeto consiste em um sistema de fácil uso e interface amigável com capacidade de fornecer aos usuários um modo de entender melhor o problema. A finalidade do sistema vai desde o pré-processamento do sinal da fala até o reconhecimento dos fonemas contidos em uma frase. Para isso, gráficos e relatórios contendo as médias e desvios padrões dos 39 coeficientes, além de arquivos contendo os padrões fonéticos de uma determinada frase podem ser gerados.

Todos os resultados de pré-processamentos e extração das características são salvos em um arquivo .mce de maneira que se possa servir para o próprio treinamento do sistema em questão. Assim como, as redes treinadas são salvas num arquivo .cod.

As figuras 30 e 31 mostram a janela principal do sistema quando o mesmo é inicializado.

Figura 30. Tela inicial.

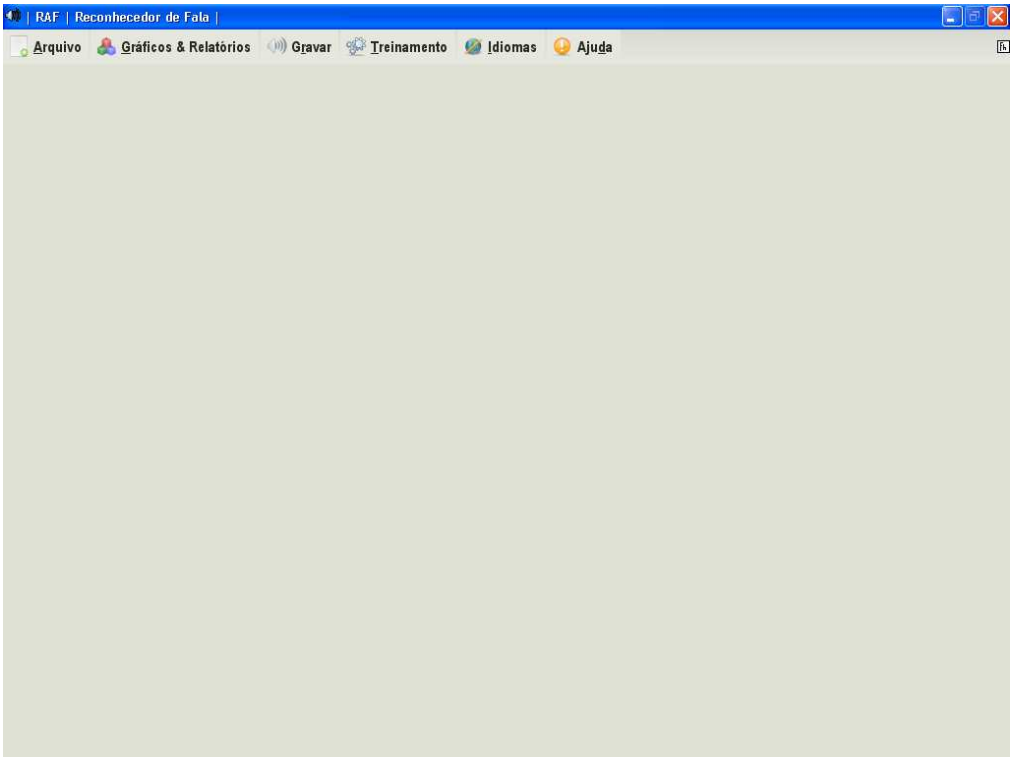


Figura 31. Tela principal do sistema.

5.1 Características técnicas

Os programas foram implementados na linguagem Java 5 [18], sendo utilizado o ambiente de desenvolvimento integrado (Integrated Development Environment - IDE) eclipse versão 3.1 [11], em conjunto com as APIs (*application programming interface*): Java Sound, Java Swing e JMusic [19], JFreeChart e IReport, que facilitam o trabalho com componentes visuais, geração de gráficos e relatórios, com o tratamento de arquivos de áudio e com a comunicação com outras linguagens.

Na implementação, teve-se o cuidado de criar uma interface visual bastante amigável e intuitiva, um código estruturado, de forma que outros pesquisadores possam desenvolver seus testes.

O sistema está dividido em pacotes. O uso de pacotes é essencial por questões de engenharia de software. Com o uso de pacotes a leitura do código fica expressivamente mais simples. A figura 27 mostra o digrama de pacotes do sistema. Desses pacotes, três merecem um melhor detalhamento.

Os pacotes de pré-processamento e extração de características, é responsável por fazer toda a parte de captação do sinal de fala, fazer a remoção do espectro DC, utilizar um filtro de pré-ênfase no sinal, dividir o sinal em quadros, utilizar a técnica de janelamento, cálculo da FFT e cálculo dos coeficientes mel-cepstrais. Todos essas funcionalidades foram implementadas utilizando a linguagem JAVA 5 [18]. É importante salientar que foram implementados todos os filtros para janelamento citados nesta monografia cuja finalidade é servir para a utilização em outras pesquisas.

O pacote da rede SOM é o responsável pela implementação da rede SOM (*self-organizing maps*). Esse módulo se utiliza do pacote SOM_PAK [22], que é implementado na linguagem C. Devido a isso, foi utilizado uma comunicação com este código nativo. A escolha foi por questões

de desempenho, já que nos testes preliminares feitos sentimos a necessidade de melhoria no desempenho geral da rede SOM. O treinamento da rede também está incorporado a esse pacote do sistema. Para o treinamento da rede é necessária a passagem de um arquivo com extensão .mce que contém os padrões de entrada. Vale salientar que, quando a rede está treinada, é guardada a rede treinada em um arquivo local com extensão .cod.

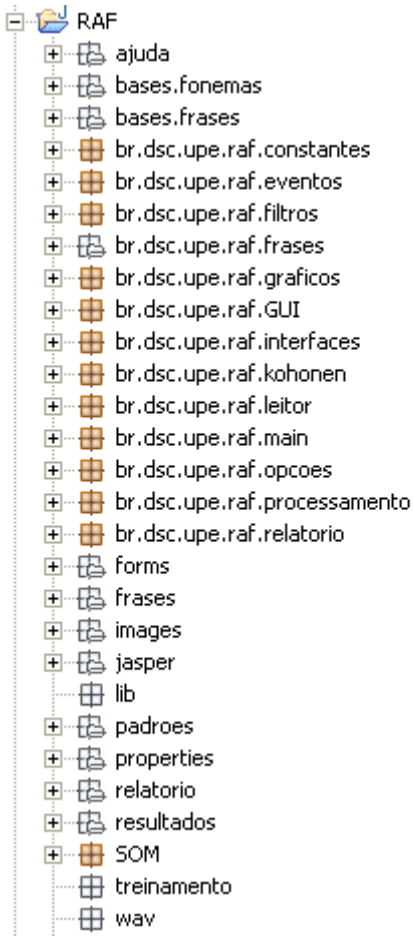


Figura 32. Visualização dos pacotes do sistema.

5.2 Funcionalidades

As funcionalidades do sistema podem ser acessadas através de 4 menus:

- Menu arquivo;
- Menu Gráfico e Relatório;
- Menu Gravar;
- Menu Treinamento;

5.2.1 Menu Arquivo

O menu arquivo têm duas funcionalidades básicas:

- Abrir arquivos .wav e gerar os padrões fonéticos constituintes;
- Sair do sistema;

5.2.2 Menu Gráficos e Relatório

Ao selecionar o menu “Gráficos e relatórios” é aberta um janela de contendo todas as opções possíveis para a geração de gráficos e relatórios. A figura 33 apresenta a janela de opções. Conforme podemos ver, existem algumas opções para a geração dos gráficos e relatórios. É possível escolher um título, uma descrição, o nome que aparecerá no eixo das abcissas e coordenadas do gráfico a ser gerado. Assim como, quais características serão mostradas. A figura 34 ilustra uma saída de gráfico gerado pelo sistema. É possível, também, gerar relatórios contendo as médias e desvios padrões das características do sinal. A figura 35 mostra a saída parcial de um relatório do sistema.

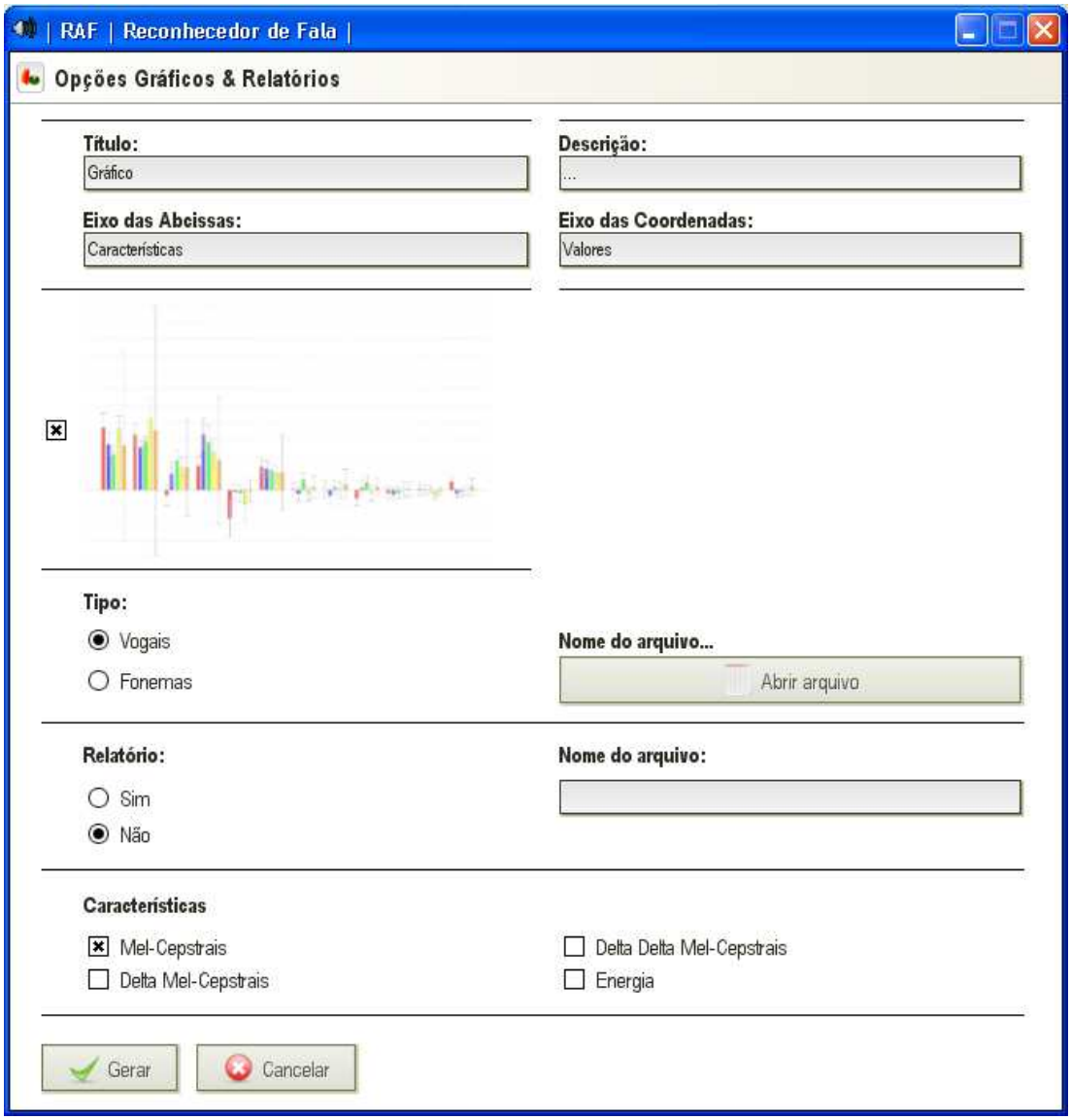


Figura 33. Tela de opções para a geração dos gráficos e relatórios.

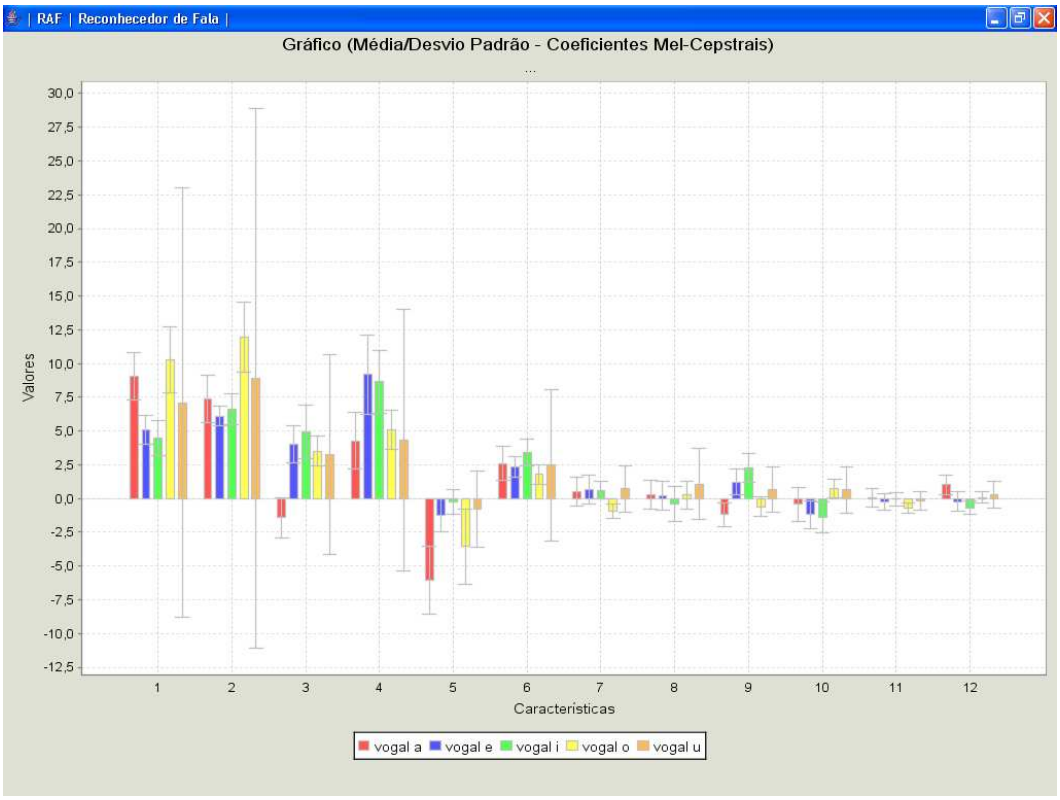


Figura 34. Visualização do gráfico gerado.

RAF Reconhecedor Automático de Fala

Fonema #

Coeficientes Mel-Cepstrais			Coeficientes Delta Mel-Cepstrais			Coeficientes Delta-Delta Mel-Cepstrais			Energia		
#	Média	Desvio Padrão	#	Média	Desvio Padrão	#	Média	Desvio Padrão	#	Média	Desvio Padrão
1	-3.54383372041	5.490826880564	13	-0.03094642717	0.648375215644	25	0.01775110378	0.191117741499054	37	0.10911010183	0.05658983007533
2	2.09852404669	2.695528626910	14	0.00984333555	0.330526301972	26	-0.00640008766	0.112145619531913	38	0.75547822145	0.05127307876033
3	1.18002784466	1.096889176479	15	0.00901702228	0.274400248814	27	-0.00467760229	0.112929790499516	39	0.76051474832	0.02220943604674
4	0.99026551048	1.955848967520	16	0.00408149918	0.306454810493	28	-0.00240584235	0.125102915645957			
5	0.01401299282	0.835319391462	17	-0.00280818183	0.245248710273	29	-1.37568212618	0.110607983107041			
6	0.63744403853	1.177505420669	18	0.00161222673	0.217152994093	30	-0.00185243218	0.090712402382292			
7	0.51626186665	0.657120808603	19	0.00477603840	0.184273572857	31	-0.00235945722	0.078102292955887			
8	0.32437501242	0.803439662876	20	0.00252211881	0.198241323237	32	-0.00111120773	0.089733218941939			
9	0.17649689909	0.606968656333	21	0.00273943089	0.183971060711	33	-0.00104390868	0.080841743739865			
10	-0.26131889935	0.605037649612	22	-0.00200075165	0.188636676415	34	7.18701592392	0.084703560847532			
11	0.07152654866	0.552286701911	23	0.00248967735	0.173850487232	35	-6.32707327043	0.078528045752044			
12	0.14698121460	0.567767754555	24	0.00260363176	0.160067743973	36	-6.08488970893	0.071713114572342			

Quantidade de Padrões	974

Relatório de Coeficientes para os Fonemas Página 1 de 36

Figura 35. Visualização do relatório gerado.

5.2.3 Menu Gravar

O menu “Gravar” tem a responsabilidade de fazer as gravações do sinal de fala do usuário. Esta funcionalidade ainda não está completa. A essa funcionalidade será adicionada um banco de dados simples que terá por objetivo guardar quais frases foram treinadas por uma determinada pessoa. Com isso, o sistema poderá fazer treinamentos personalizados para cada usuário em particular.

Figura 36. Tela de gravação do sinal da fala do usuário.

5.2.4 Menu Treinamento

O menu de treinamento do sistema é responsável, por diversas funcionalidades do sistema. A partir dela podem ser gerados:

- Gerar os arquivos contendo as características mel-cepstrais e de energia que serão utilizadas para o treinamento da rede SOM;
- Setar algumas opções que serão utilizadas na etapa de pré-processamento e extração de características do sinal. A figura 37 apresenta a tela que contém as opções possíveis;
- Fazer o treinamento da rede SOM;
- Setar algumas opções que serão utilizadas para o treinamento da rede SOM. A figura 38 apresenta a tela que contém as opções possíveis;

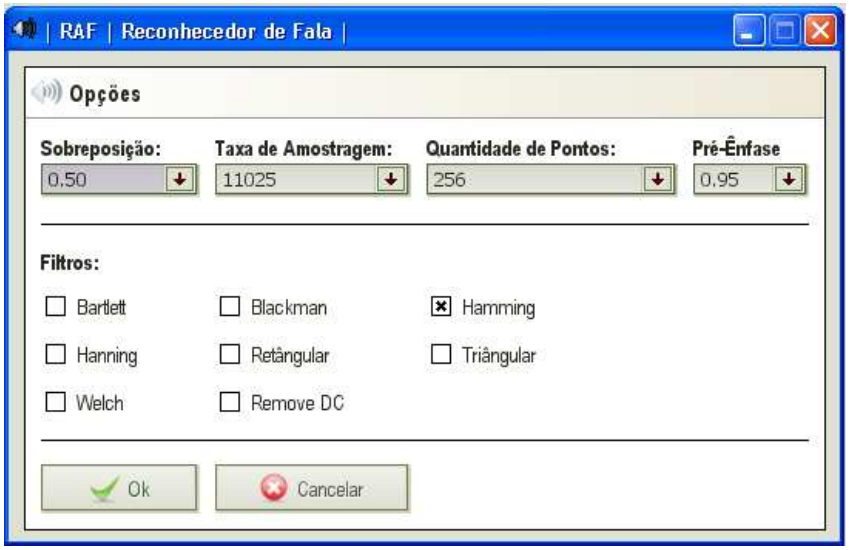


Figura 37. Tela de opções para o pré-processamento e extração de características.

Figura 38. Tela de opções dos parâmetros da rede SOM.

Capítulo 6

Experimentos

As tecnologias mais usadas, atualmente, na área de reconhecimento de fala (RNAs e HMMs) utilizam métodos de modelagem estatística que aprendem através de amostras de entrada [23, 38]. Devido a isso, é necessário um conjunto de dados de treinamento que tente cobrir estas variações.

6.1 Base de dados

No presente trabalho, utilizou-se a base de dados confeccionada por Ynoguti, como ponto de partida para os experimentos.

A partir dessa base de dados, foi confeccionada uma outra base contendo apenas as sub-unidades fonéticas mostradas na tabela 2. Para cada sub-unidade fonética, foram criados 25 arquivos no formato Windows PCM (WAV), o qual contém o espectro do sinal do fonema. A taxa de amostragem foi de 11.025kHz, e resolução de 16 bits.

Tabela 2. Sub-unidades acústicas utilizadas na transcrição fonética das locuções com exemplos[41].

Símbolo utilizado	***Exemplo***
a	**a** çafrão
e	**e** levador
E	p **e** lê
i	s **i** no
y	fu **i**
o	b **o** lo
O	b **o** la
u	l **u** a
an	maç **ã**
en	s **en** ta
in	p **in** to

on	S **om** bra
un	**um**
b	**b** ela
d	**d** ávida
D	**d** iferente
f	**f** eira
g	**g** orila
j	**j** iló
k	**c** achoeira
l	**l** eão
L	**Lh** ama
m	**m** ontanha
n	**n** évoa
N	i **nh** ame
p	**p** oente
r	ce **r** a
rr	ce **rr** ado
R	ca **r** ta
s	**s** apo
t	**t** empes **t** ade
T	**t** igela
v	**v** erão
x	**ch** ave
Z	**z** abumba

6.2 Transcrição fonética

A transcrição fonética é uma técnica usada com o intuito de separar as palavras em formas de fonemas. A transcrição fonética utilizada nessa monografia foi feita a partir da base de dados confeccionada por Ynoguti [41].

Para a execução desta tarefa foi utilizado um programa de visualização gráfica do espectrograma e forma de onda do sinal chamado *audacity* [4].

O audacity é um editor de áudio gratuito, que pode ser usado para gravar sons, tocar músicas, importar e exportar arquivos WAV, AIFF e MP3. Ele ainda suporta a funcionalidade de editar sons, cortá-los, copiar e colar pedaços de som. A figura 39 mostra o espectro da frase “A justiça é a única vencedora” junto com a análise de freqüências da mesma.

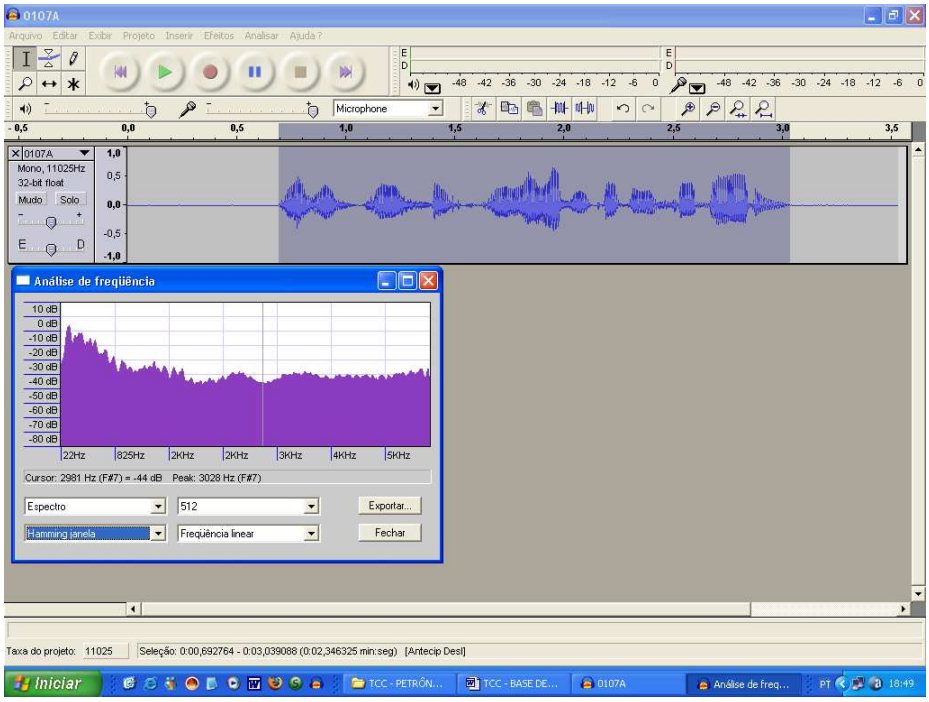


Figura 39. O audacity em execução, mostrando o espectro da frase “A justiça é a única vencedora”.

Na figura 40 é mostrada outra funcionalidade, que é de cortar e copiar pedaços de uma faixa de áudio, mostrando o mapeamento de um fonema.

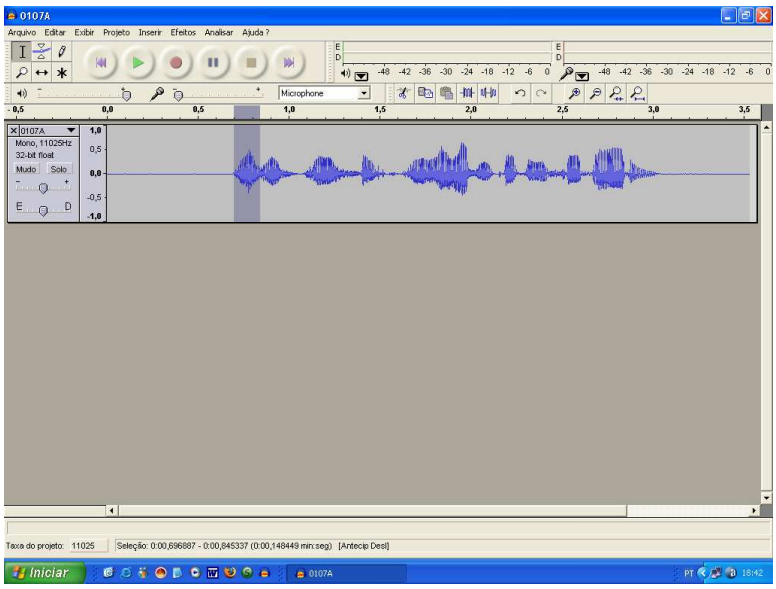


Figura 40. Mapeamento do fonema /a/ da frase “A justiça é a única vencedora”.

O procedimento da geração da base de dados contendo as sub-unidades fonéticas foi manual, conforme foi mostrado na figura 40. As sub-unidades utilizadas nesta tarefa são mostradas na tabela 2. É importante frisar que o conjunto dos fonemas utilizados nesta monografia foi o que Ynoguti especificou na sua tese [41].

6.3 Pré-processamento e extração de parâmetros

Nesta etapa do reconhecedor, cuja entrada é um sinal de fala no formato WAV, são calculados os parâmetros da locução. São extraídos 12 coeficientes mel-cepstrais, 12 coeficientes delta e 12 coeficientes delta-delta mel-cepstrais e 1 log-energia normalizada, 1 delta e 1 delta-delta do log-energia normalizada, formando um total de 39 parâmetros extraídos de cada quadro, na freqüência de amostragem 11.025kHz, com 16 bits de resolução.

Os parâmetros são calculados utilizando-se janelas de 23.22ms, atualizadas a cada 11.61ms. Antes da extração, o sinal é submetido a alguns pré-processamentos: retirada do nível DC que pode aparecer devido a problemas com o microfone, pré-ênfase com um filtro passa altas ($1-0{,}95\ z^{-1}$), e janelamento através de uma janela de Hamming, conforme é mostrada na figura 41.

Figura 41. Diagrama de blocos do processo de pré-processamento e extração dos parâmetros mel-cepstrais e de energia.

6.4 Arquiteturas SOMs usadas

Para o melhor entendimento do problema a ser estudado, resolveu-se criar dois protótipos de um sistema de RAF. Entende-se por protótipo de um sistema como uma implementação de parte das funcionalidades do sistema requerido. Esses protótipos ajudaram na concepção e modelagem do sistema de RAF final.

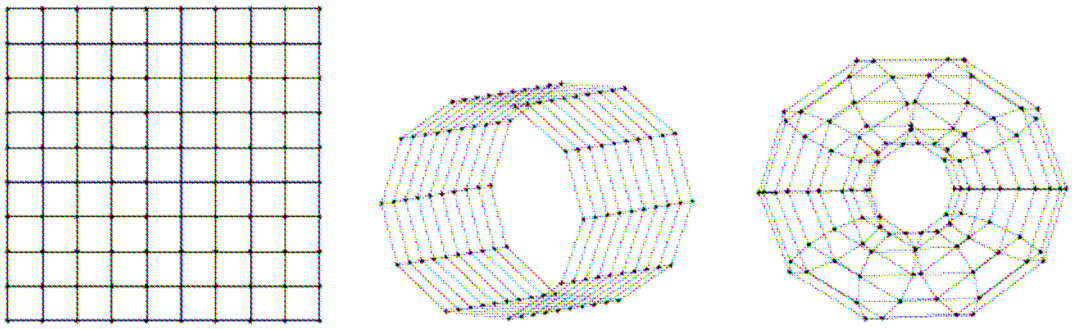
Foram testadas diversas arquiteturas de SOMs, que se distinguiram principalmente pelas dimensões do mapa. Em todas elas, foram usados mapas bidimensionais hexagonais.

Para os protótipos foram avaliadas diversas características que poderiam influenciar no resultado do problema, dentre as quais podemos citar: topologia da rede, forma da rede e função para cálculo da vizinhança.

A figura 19 ilustra uma rede retangular e uma hexagonal, enquanto a figura 42 ilustra possíveis formas que uma rede SOM pode assumir. As implicações das diferentes topologias e formas que uma rede SOM pode adquirir são: a política de atualização dos pesos dos vizinhos do neurônio vencedor e de tratamento dos efeitos de borda. Quanto à topologia, optou-se pela hexagonal. Quanto à forma da rede, optou-se pela estrutura plana (*sheet*). Quanto à função utilizada para o cálculo da vizinhança, usamos a função cutgaussian, mostrada na equação abaixo.

$$h_{ci}(t) = e^{-\frac{d_{ci}^2}{2\sigma_t^2}} 1(\sigma_t - d_{ci})$$

Onde σ é o raio da vizinhança no instante de tempo t, $d_{ci} = \| r_c - r_i \|$ é a distância entre as unidades de mapa c e i no mapa topográfico e $1(x)$ é a função degrau.



Estrutura plana | Estrutura cilíndrica | Estrutura toroidal

Figura 42. Exemplos de possíveis arquiteturas de rede SOM.

O primeiro protótipo foi idealizado com o intuito de reconhecer as cinco vogais do alfabeto Português-Brasil. Para esse experimento foram definidos, além das já citadas características, os seguintes parâmetros mostrados na tabela 3.

Tabela 3. Parâmetros utilizados no experimento de reconhecimento de vogais.

Dimensão	ρ_0	ρ_i	Épocas de ordenação	Épocas de refinamento
5x6	4	0.85	10	20
10x12	7	0.85	10	20
20x24	13	0.85	10	20

O segundo protótipo teve por objetivo reconhecer os 36 fonemas, propostos por Ynoguti [40], para o Português-Brasil. Nesse experimento, além das já citadas características, os seguintes parâmetros encontrados na tabela 4.

Tabela 4. Parâmetros utilizados no experimento de reconhecimento de fonemas.

Dimensão	ρ_0	ρ_i	Épocas de ordenação	Épocas de refinamento
20x24	13	0.85	10	20
40x48	25	0.85	10	20
60x72	37	0.85	10	20

Três esclarecimentos precisam ser feitos para o melhor entendimento do experimento.

- A primeira vista parece ser pequena a quantidade de épocas utilizadas, mas para o treinamento da rede foi utilizado o modo batch.
- As dimensões foram estabelecidas a partir da quantidade de classes do experimento em questão.
- O raio inicial (ρ_0) e final (ρ_i) foram definidos a partir da dimensão da rede.

Os protótipos, como já descrito, além de ajudarem no entendimento do problema, tiveram um papel decisivo na escolha de qual dimensão da rede utilizar no sistema de reconhecimento automático de fala.

6.4.1 Treinamento e teste

Em cada experimento, as SOMs foram treinadas com uma das arquiteturas citadas, sobre o conjunto de vogais e fonemas. Os vetores contendo os 39 coeficientes extraídos do sinal da fala foram apresentados à rede SOM. A quantidade de amostras utilizadas nos protótipos de vogais e de fonemas podem ser encontradas, respectivamente, nas tabelas 5 e 6.

Tabela 5. Quantidade de amostras e padrões de entrada contidas na base de dados de vogais.

Fonemas	Quantidades de amostras no conjunto de treinamento	Quantidade de amostras no conjunto de teste
a	116	30
e	86	30
i	118	30
o	97	30
u	126	30

Tabela 6. Quantidade de amostras e padrões de entrada contidas na base de dados de fonemas.

Símbolo utilizado	Quantidades de amostras no conjunto de treinamento	Quantidade de amostras no conjunto de teste
#	153	30
a	117	30
an	168	30
b	116	30
d	90	30
D	116	30
e	86	30
E	86	30
en	170	30
f	156	30
g	109	30
i	119	30
in	169	30
j	105	30
k	136	30
l	93	30
L	105	30
m	120	30
n	98	30
N	131	30
o	98	30
O	146	30
on	186	30
p	143	30
r	69	30
R	177	30
rr	107	30
s	193	30

t	141	30
T	116	30
u	127	30
un	141	30
v	93	30
x	184	30
y	101	30
z	128	30

Para testar o reconhecedor de fala, foram utilizadas 10 frases da base de Ynoguti, mostradas abaixo, com seus respectivos padrões fonéticos.

1. Nosso telefone quebrou.
   # n O s u t e l e f o n y k e b r o u #
2. Desculpe se magoei o velho.
   # D y s k u p y s y m a g u e y u v E L u #
3. Queremos discutir o orçamento.
   # k e r e m u z D i s k u T i r u o R s a m e in t u #
4. Ela tem muita fome.
   # E l a t e in m u y t a f O m y #
5. Uma índia andava na mata.
   # u m a in D y a an d a v a n a m a t a #
6. Zé, vá mais rápido!
   # z E # v a m a y s rr a p i d u #
7. Hoje dormirei bem.
   # o j y d o R m i r e y b e in #
8. João deu pouco dinheiro.
   # j u an un d e u p o u k u D in N e y r u #
9. Ainda são seis horas.
   # a in d a s an un s e y z O r a s #
10. Ela saía discretamente.
    # E l a s a i a D i s k r E t a m e in T y #

Capítulo 7

Resultados

Todos os experimentos criados, foram implementados e testados, usando a ferramenta Matlab [39] junto com toolbox SOMToolbox[1] [40] onde se encontra a implementação da rede SOM.

No experimento de reconhecimento de vogais foram feitos levantamentos de alguns tamanhos de mapas. Utilizou-se a técnica de normalização para testar sua influência sobre os vetores dos padrões de entrada. Aplicou-se, também, o treinamento sem aplicar a normalização.

A tabela 7 mostra os resultados do treinamento e teste conseguidos com uma rede SOM de dimensão 5x6, enquanto que a figura 43 ilustra a comparação entre os resultados. Pode-se notar que não houve melhoria significativa com a utilização da normalização. A figura 44 mostra a visualização da localização das vogais no mapa topográfico gerado ao longo do fase de teste.

Tabela 7. Comparativo entre os resultados do treinamento e teste de vogais usando um mapa topográfico de dimensões 5x6.

Tipos de normalização	Treinamento (% de acerto)	Teste (% de acerto)
Sem normalização	76,501	62,134
Normalização com vetores entre 0 e 1	76,08	65,47
Normalização com vetor unitário	76,501	62,134
Normalização com vetor unitário e vetor entre 0 e 1	75,732	65,998

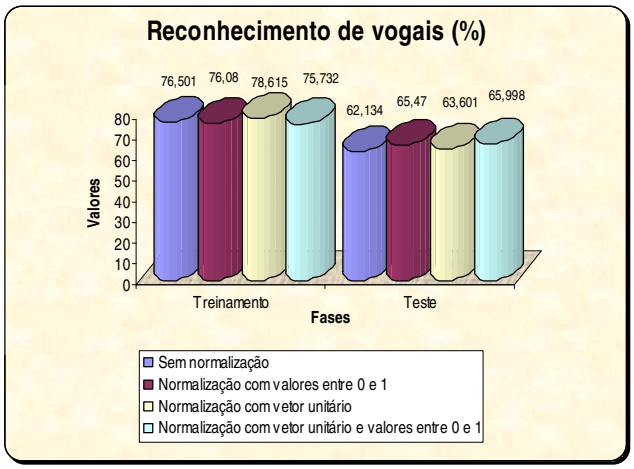


Figura 43. Visualização da taxa de acerto (%) no treinamento e teste utilizando um mapa 5x6 com e sem a normalização dos dados.

[1] *A utilização desse pacote se deu pelo fato de que o módulo da rede SOM do sistema proposto não estar devidamente testado e validado na época em que foram feitos os testes.*

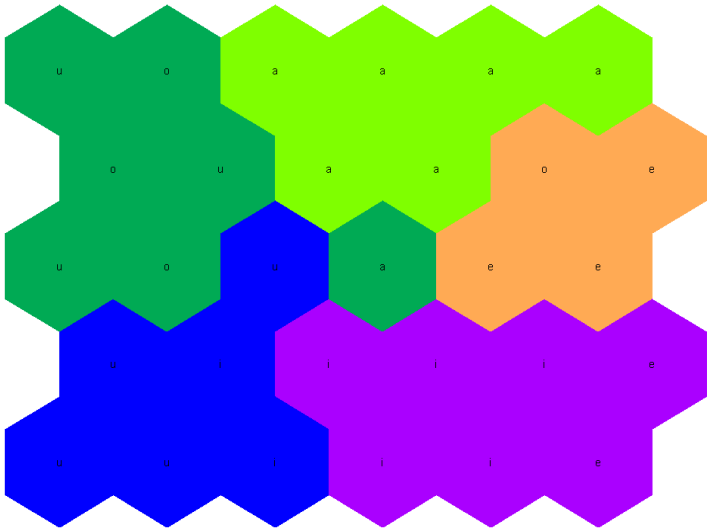


Figura 44. Visualização da localização das vogais no mapa 5x6.

Os resultados utilizando uma rede 10x12 foram melhores que os conseguidos com uma rede 5x6. Credita-se essa melhora ao fato do problema investigado ser reconhecidamente complexo, necessitando, dessa forma, de um aumento de unidades processadoras (neurônios). A tabela 8 mostra e a figura 45 ilustra os resultados do treinamento e teste conseguidos com tal rede. De maneira semelhante, notou-se que a taxa de acerto não se alterou substancialmente com a aplicação da normalização. A figura 46 mostra a visualização da localização das vogais na rede de dimensão 10x12. Subentende-se, com isso, que, com o aumento da dimensão das unidades processadoras, as regiões fronteiriças entre as vogais, encontram-se mais bem delimitadas.

Tabela 8. Comparativo entre os resultados do treinamento e teste de vogais usando um mapa topográfico de dimensões 10x12.

Tipos de normalização	**Treinamento (% de acerto)**	**Teste (% de acerto)**
Sem normalização	87,362	74,667
Normalização com vetores entre 0 e 1	86,85	74,468
Normalização com vetor unitário	86,996	72,665
Normalização com vetor unitário e vetor entre 0 e 1	87,015	76,266

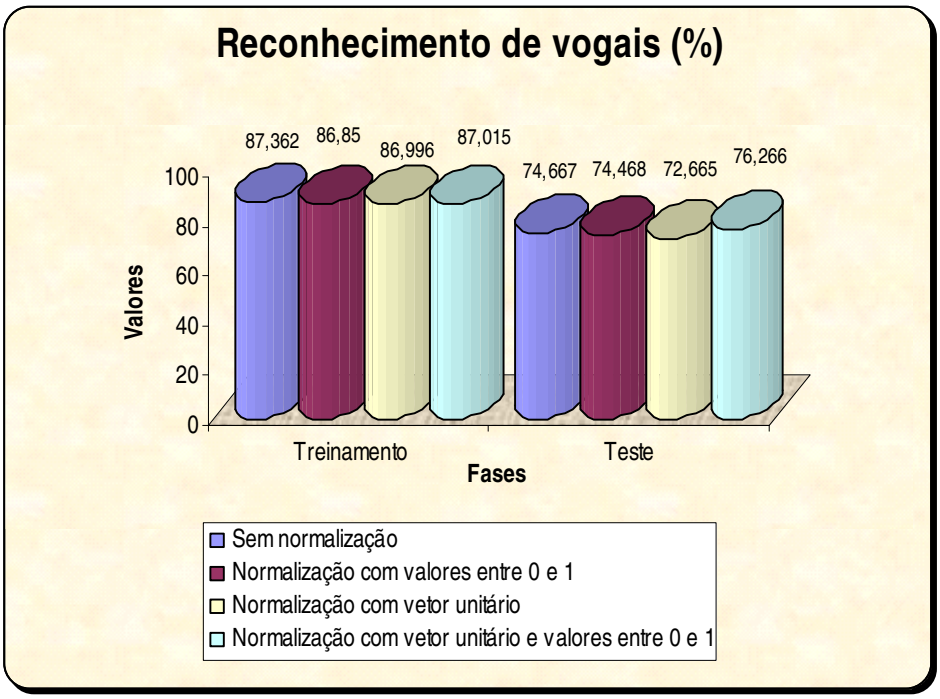


Figura 45. Visualização da taxa de acerto (%) no treinamento e teste utilizando um mapa 10x12 com e sem a normalização dos dados.

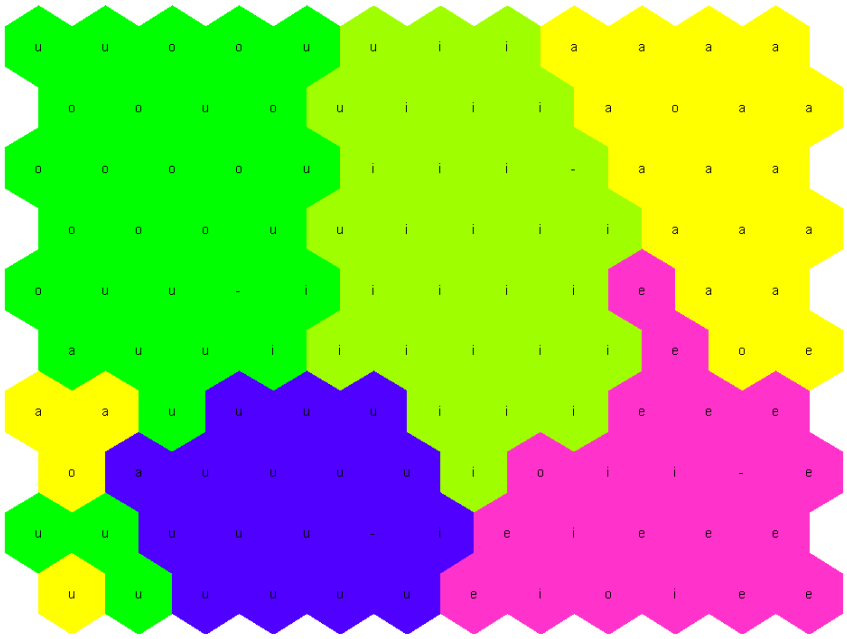


Figura 46. Visualização da localização das vogais no mapa 10x12.

Novamente, com o aumento da complexidade da rede SOM, observou-se um aumento na taxa de acerto dos padrões de entrada. Não somente a complexidade da rede cresceu, mas, também, o tempo necessário para a confecção do mapa. Enquanto o tempo de treinamento e teste, da rede de dimensão 5x6, aproximou-se dos 5 segundos, a rede de dimensão 10x12 obteve 10

segundos. A rede 20x24 atingiu um tempo total de 16 segundos. A normalização do conjunto de entrada não influenciou significativamente no desempenho da rede. As figuras 47 e 48 ilustram os resultados do treinamento e teste conseguidos com tal rede, e o mapa topográfico gerado pelo algoritmo de aprendizado, respectivamente. Os resultados são mostrados abaixo, na tabela 9.

Tabela 9. Comparativo entre os resultados do treinamento e teste de vogais usando um mapa topográfico de dimensões 20x24.

Tipos de normalização	Treinamento (% de acerto)	Teste (% de acerto)
Sem normalização	95,879	76,602
Normalização com vetores entre 0 e 1	96,282	76,334
Normalização com vetor unitário	96,246	76,6
Normalização com vetor unitário e vetor entre 0 e 1	96,116	76,867

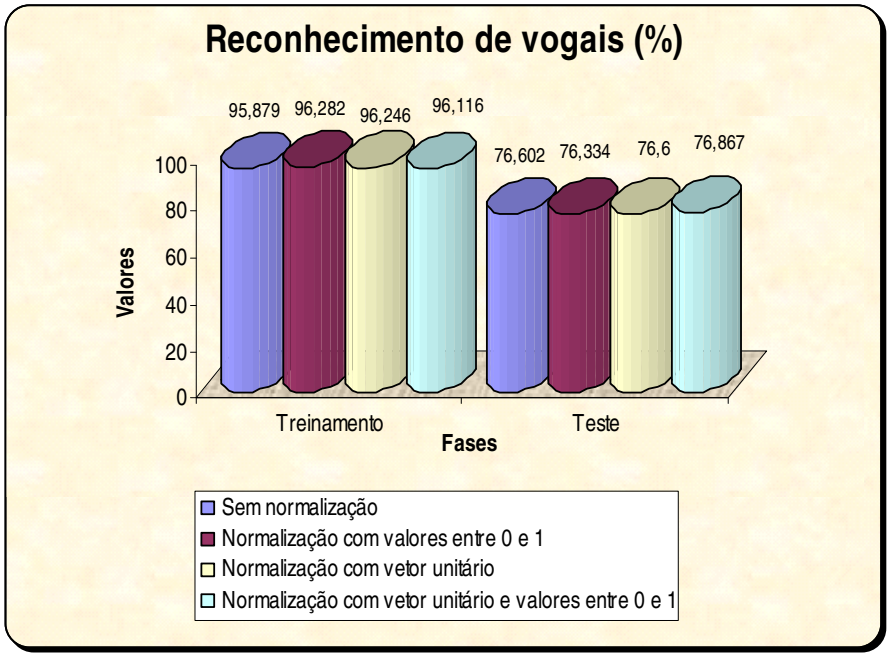


Figura 47. Visualização da taxa de acerto (%) no treinamento e teste utilizando um mapa 20x24 com e sem a normalização dos dados.

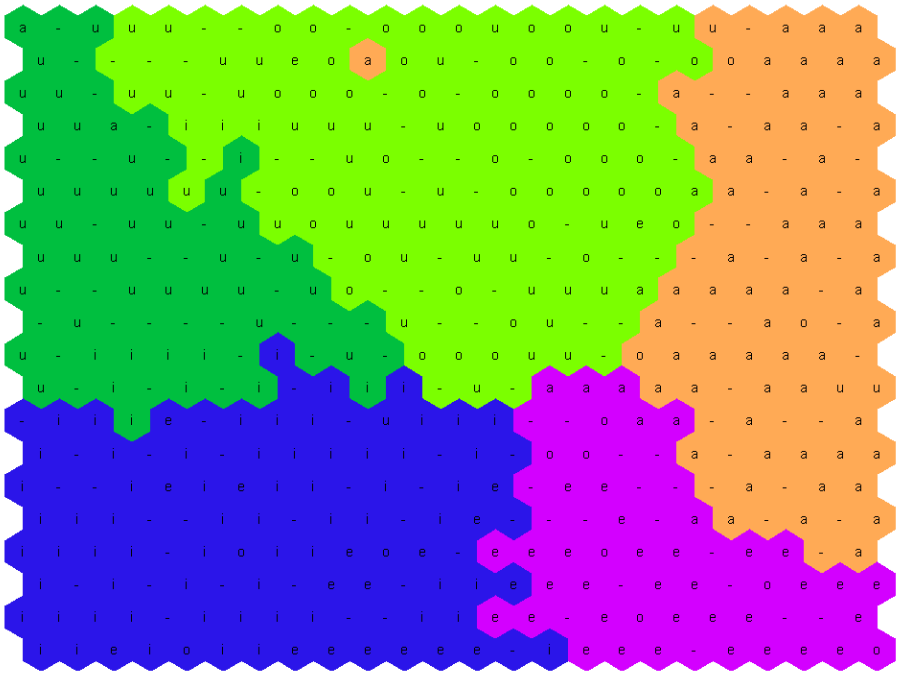


Figura 48. Visualização da localização das vogais no mapa 20x24.

Com relação aos fonemas, como já citado, há 36 tipos de fonemas, ao todo, foram realizados 30 treinamentos com redes de dimensões 20x24, 40x48 e 60x72. Para cada rede, 10 treinamentos foram feitos, contabilizando-se a média aritmética e desvio padrão das taxas de treinamento e teste obtidas. O processo de normalização foi aplicado, mas, apesar disso, não foi constatada melhora significativa nos resultados obtidos. As tabelas 10 e 11 comparam a taxa de acerto, média e desvio padrão, nas fases de treinamento e teste com e sem normalização.

Tabela 10. Comparativo entre os resultados da média do treinamento e teste de fonemas.

Dimensão	**20x24 (%)**		**40x48 (%)**		**60x72 (%)**	
Fase	Treinamento	Teste	Treinamento	Teste	Treinamento	Teste
Normalização com vetor unitário	37,66	32,06	67,56	57,28	81,47	66,98
Normalização com valores entre 0 e 1	38,83	32,28	68,18	57,58	82,1	67,2
Normalização com vetor unitário e com valores entre 0 e 1	37,35	32,95	67,69	56,24	82,57	67,98
Sem normalização	38,73	33,414	68,1	57,336	82,892	67,754

Tabela 11. Comparativo entre os resultados do desvio padrão do treinamento e teste de fonemas.

Dimensão	**20x24 (%)**		**40x48 (%)**		**60x72 (%)**	
Fase	Treinamento	Teste	Treinamento	Teste	Treinamento	Teste
Normalização com vetor unitário	2,35	1,95	3,24	3,56	2,97	3,28
Normalização com valores entre 0 e 1	2,08	1,75	2,97	3,02	2,68	2,95
Normalização com vetor unitário e com valores entre 0 e 1	2,06	1,9	2,86	3,05	2,48	2,69
Sem normalização	1,56	1,68	3,14	3,34	3,05	3,48

A rede de dimensão 20x24, conforme se observa na tabela 11, obteve uma taxa de acerto (média) de aproximadamente 33,414%. A porcentagem baixa mostra a insuficiência da quantidade de neurônios para representar os fonemas corretamente.

Quanto à rede de dimensão 40x48, a taxa de acerto aumentou significativamente, em comparação ao resultado obtido pela rede anterior. Esse índice, de aproximadamente 57,336%, implicou no aumento da complexidade, que, por sua vez, encontra-se refletida no acréscimo de unidades processadoras. Não somente a complexidade, mas, também, o tempo necessário para a associação padrão-fonema se elevou. Na rede anterior, o tempo foi de 21 segundos, enquanto nessa obtivemos 140 segundos. Uma ilustração da localização dos fonemas pode ser visto na figura 49.

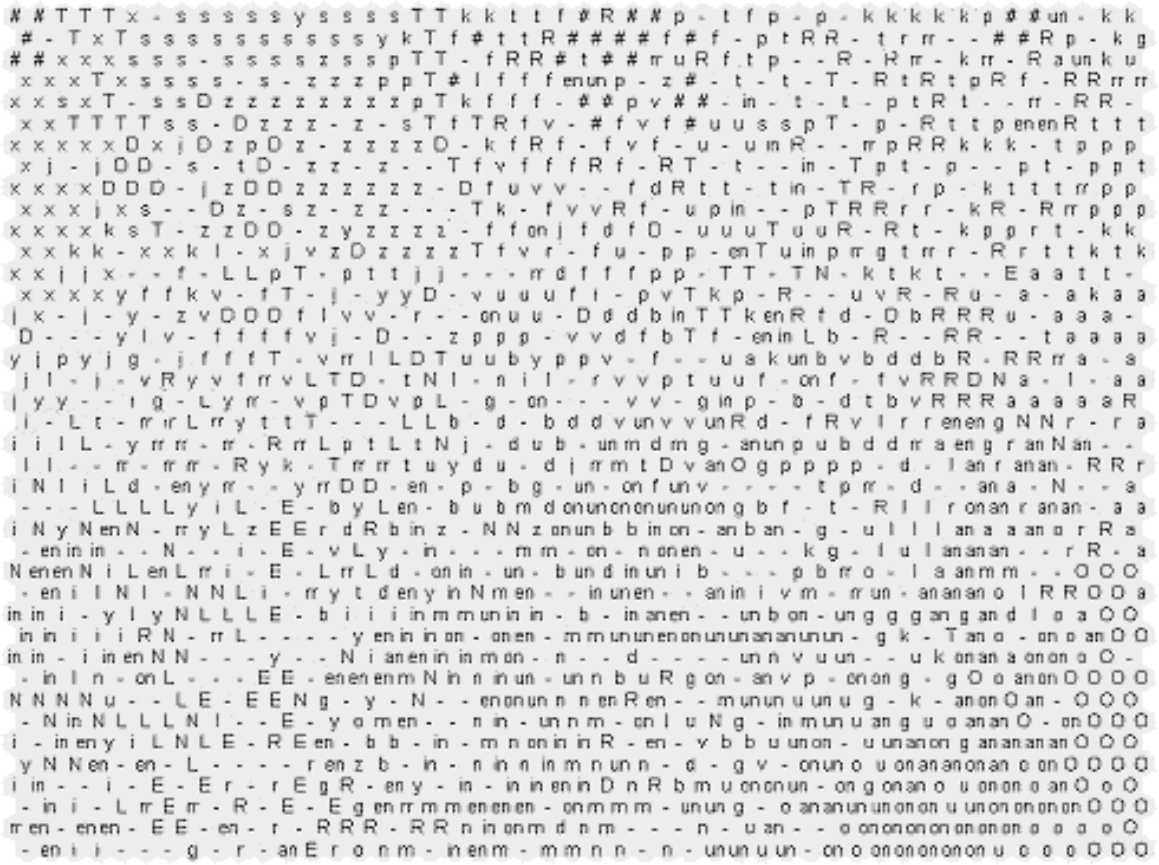


Figura 49. Visualização da localização dos fonemas no mapa 40x48.

O mapa 60x72 obteve uma precisão (taxa de acerto) de aproximadamente 67,754% através de um altíssimo custo computacional. Além da enorme quantidade de neurônios, o tempo de processamento, isto é, treinamento e teste, foram superior a 30 minutos. Observou-se que vários neurônios contidos no mapa topográfico não foram ativados, ou seja, não houve nenhuma associação padrão-neurônio. Isso é decorrente da dimensão do mapa obtido exceder a dimensão realmente necessária.

Para o experimento de reconhecimento das frases, foi escolhida a rede de dimensão 40x48 por questões de desempenho.

Logo após os experimentos já citados, prosseguiu-se para o reconhecimento dos fonemas nas frases contidas no tópico 6.4.1 desta monografia (página 58). Foi escolhida a rede de

dimensão 40x48 tendo em vista que, entre as três arquiteturas utilizadas nos experimentos, ela que apresentou melhor relação tempo de processamento e taxa de acerto no conjunto de teste.

Os resultados obtidos no reconhecimento das referidas frases se encontram no Apêndice B deste trabalho. Notou-se que o reconhecimento obtido não atingiu um desempenho satisfatório. Vários fatores podem ter ocasionado isso, a saber:

- Modelamento manual da base de dados fonética utilizada;
- Quantidade de padrões insuficientes tanto para o treinamento quanto para o teste;
- A bem conhecida dificuldade de associar padrões a transição de fonemas.

Para um melhor entendimento, o resultado do reconhecimento dos fonemas contidos em uma frase é mostrado, logo abaixo:

Nosso telefone quebrou.

n O s u t e l e f o n y k e b r o u

'-' '#' '#' '#' '#' '-' 'rr' '-' '-' '#' '#' '#' '-' '-' '#' '#' '#' '-' '-' '#' '#' '#' '-' '-' '-' '-' '-'
' '#' '#' '#' '#' '#' '-' '-' '-' '#' '-' '-' 'm' '-' '-' '-' 'rr' '-' '-' '-' 'O' 'O' '-' 'an' '-' 'O'
'-' '-' '-' '-' 's' '-' '-' '-' '-' '-' '-' '-' 'y' 'u' '-' '-' 'rr' '-' 'rr' '-' '-' '-' '-' 'en' '-' 'd' '-'
'-' '-' 'l' '-' 'E' 'E' '-' '-' 'E' 'E' '-' '-' '#' '#' '-' '#' 'f' 'f' '-' '-' '-' 'o' 'u' '-' 'un' 'on'
'o' 'an' 'o' 'o' '-' 'on' '-' '-' '-' 'D' 'en' 'en' 'N' 'N' 'e' '-' '-' '-' '-' '-' '-' '-' '-' '-' 'e' 'i'
'-' '-' '-' 'r' '-' '-' '-' '-' '-' '-' 'v' 'g' '-' '-' 'l' '-' 'un' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-'
'-' '-'
'-' '#' '#' '-' '-' '#' '#' '#' '#' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-'

Conforme o resultado nos mostra, poucos padrões foram classificados. Além disso, os fonemas que foram classificados não tiveram um resultado perceptivelmente bom.

A notação utilizada no resultado foi: frase, fonemas formadores da frase e fonemas reconhecidos sem o pós-processamento.

Como os resultados do experimento não foram satisfatórios, alterou-se a forma de solucionar o problema. Ao invés de treinar a rede SOM com os fonemas extraídos da base de dados de Ynoguti, treinou-se com as próprias frases da referida base. Como já foi citado, a base contém 200 frases. Nesse experimento, a rede foi treinada com 50, 100 e 200 frases, utilizando duas dimensões de mapas, 20x24 e 40x48. Para o rotulamento da rede, utilizaram-se os fonemas extraídos, base de dados fonética confeccionada neste trabalho, da base de dados de Ynoguti. Os resultados são mostrados na tabela 12 e ilustrados na figura 50.

Tabela 12. Comparativo entre as taxas de acerto (%) do teste de fonemas usando os mapas topográficos de dimensões 20x24 e 40x48.

Quantidade de frases no treinamento	**20x24 (Taxa de Acerto (%))**	**40x48 (Taxa de Acerto (%))**
50 frases	39,31	58,14
100 frases	39,33	59,98
200 frases	33,23	47,36

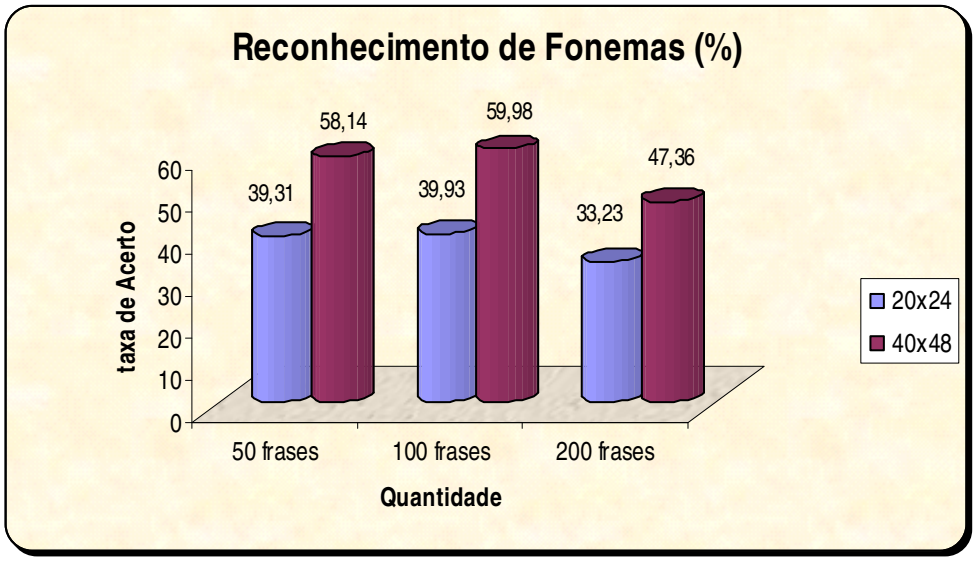


Figura 50. Visualização da taxa de acerto (%) no teste utilizando mapas de 20x24 e 40x48.

Analisando os resultados obtidos no reconhecimento dos fonemas, conseguidos nos dois experimentos, nota-se uma semelhança nos resultados encontrados.

A partir dos resultados, escolheu-se trabalhar com uma rede de dimensão 40x48 e utilizando 50 frases. Por questões de desempenho e taxa de acertos, essa escolha foi realizada. Os resultados obtidos podem ser encontrados no Apêndice C.

Conforme o resultado nos mostra, os fonemas que foram classificados tiveram um resultado aceitável como mostrado abaixo:

Nosso telefone quebrou.

n O s u t e l e f o n y k e b r o u

'-' '-' '#' '#' '-' '-' '-' '#' '#' '-' '-' '-' '-' '-' '-' '-' '#' '-' '-' '-' '-' '-' '#' '#' '#' '#' '#' '#'
'#' '#' '#' '#' '#' 'v' '-' '-' '#' '-' '-' '-' '-' '-' '-' 'L' 'n' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O'
'-' '-' '-' '-' 's' 's' 's' 's' 'T' '-' '-' 'l' '-' 'u' 'u' 'u' '-' '-' '-' '-' '-' '#' '-' 'en' '-' 'e' 'e'
'e' '-' 'l' 'l' '-' '-' '-' 'en' 'e' '-' 'D' 'f' 'f' '-' '-' '-' '-' 'f' 'f' '-' '-' 'on' 'on' 'on' 'b' 'on'
'O' 'O' '-' 'o' 'on' 'on' 'on' 'n' '-' '-' '-' '-' '-' 'in' 'N' 'y' 'y' '-' 'k' 'k' '#' '#' '#' '-' '-'
'-' 'en' 'en' 'b' 'R' 'n' '-' '-' '-' 'i' 'en' '-' '-' 'r' 'r' 'R' 'o' '-' 'o' 'o' 'o' 'o' 'o' 'o' 'o'
'o' '-' 'o' 'o' 'o' 'on' '-' '-' '-' 'O' 'u' 'p' 'p' 'u' 'u' '-' '-' 'f' 'u' '-' '-' '#' '-' 'f' 'f' '-'
'#' '-' '-' '-' '#' '#' '-' '#' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '#' '#'

A explicação para que este experimento tenha apresentado melhor resultado, se deve ao fato do treinamento com frases apresentar melhor representação dos fonemas e das variações dos mesmos, enquanto que no treinamento, com fonemas, apenas, essas variações não se encontram bem representadas e modeladas. Isso é patente nos resultados encontrados através da não-associação entre padrão-fonema, ilustrada por meio do símbolo “-“.

Capítulo 8

Conclusões e Trabalhos Futuros

O presente trabalho apresentou, primeiramente, uma introdução à área de reconhecimento de fala, relatando o histórico da área, características de um reconhecedor precisa ter e trabalhos relacionados.

Técnicas de aprendizagem de máquina foram também revisadas, focando nas redes neurais artificiais do tipo *Self-Organizing Map* (SOM), a qual foi utilizada na resolução do problema.

Outro assunto descrito foi o pré-processamento dos sinais de fala. Foram relatadas técnicas de pré-ênfase, de janelamento do sinal em quadros, de detecção do começo e fim da fala, de extração de características representativas da fala do locutor utilizando os coeficientes mel-cepstrais e de energia e quantização vetorial.

Foi criada e descrita uma base de dados contendo todos os padrões fonéticos citados na tese de Ynoguti. Além disso, buscou-se arquitetar um arcabouço para a construção de sistemas que facilitassem e ajudassem o desenvolvimento de sistemas de reconhecimento de fala. Foi criado um sistema que busca aplicar todo ferramental criado.

A implementação do sistema foi desenvolvida com a concepção de pacotes, cujo objetivo é facilitar a reusabilidade dos módulos e códigos.

Foram desenvolvidos dois protótipos, utilizando o MATLAB, antes de implementar o reconhecedor fonético, com o intuito de entender e refinar o problema proposto. O primeiro protótipo foi o reconhecedor de vogais, o qual obteve no melhor caso, uma taxa de acerto de aproximadamente 77%. O segundo protótipo desenvolvido foi um reconhecedor fonético que conseguiu uma taxa de acerto de aproximadamente 68%. Os resultados dos testes indicaram que as arquiteturas com mais unidades processadoras possuem maior poder discriminativo, respondendo com menores erros de quantização durante o treinamento e maiores taxas de acerto durante os testes, porém requerem maior tempo para o processamento.

Para testar a solução, foi proposto um sistema de reconhecimento de fala. Utilizou-se a modelagem da fala através de uma rede SOM, treinada com as características extraídas da fala do locutor. A técnica de reconhecimento de fala apresentada consiste na comparação fonética. O fonema escolhido é o que apresenta o menor erro de quantização vetorial da fala do locutor em relação à rede SOM treinada.

Nesse teste de reconhecimento de uma frase, não se conseguiu uma taxa de acerto desejável. Através do detalhamento deste erro, observou-se que as frases de teste mais curtas

foram responsáveis pelos erros. Isso pode ser devido ao não modelamento da transição entre fonemas, fazendo com que o reconhecedor não consiga associar um fonema ao padrão passado. Esse problema pode estar relacionada com a base de dados criada. Tamanho da base de dados fonética, assim como possíveis erros na confecção da base podem ter influenciado no resultado final.

Outro experimento foi proposto, que se baseia no treinamento da rede SOM com as próprias frases da base, utilizando a base de dados fonética para rotulação da rede. Esse experimento mostrou uma sensível melhora. A possível explicação para tal melhora é que com esse treinamento, foi possível obter uma melhor representação tanto dos fonemas quando de suas variações, isso repercutindo no resultado final do reconhecimento.

A partir das dificuldades encontradas no trabalho atual, propomos, para trabalhos futuros, o uso do modelo oculto de Markov para a extração dos fonemas das frases, de modo que a saída seja a entrada para uma rede SOM. Com isso, reduzimos significativamente a possibilidade de erros gerados na criação e modelagem da base de dados, além da redução do tempo despendido.

Outras propostas de trabalho futuro consistem em:

- Utilizar outra rede neural artificial, de forma a comparar os resultados conseguidos;
- Pesquisar outras técnicas de pré-processamento e extração de características que possam melhorar o reconhecimento, tais como: técnicas de endpoints e wavelets.
- Outro trabalho seria a da criação de um pós-processador que modele a saída obtida neste trabalho para uma saída que contenha apenas os fonemas constituintes da frase. Além disso, criar um pós-processador que contenha as funções de análise sintática e semântica, com isso pode-se conseguir um acréscimo significativo no resultado do reconhecedor de fonemas.
- Notou-se ainda que dentre os 36 fonemas, alguns tem um grau de dificuldade de reconhecimento maior. Dentre eles podemos citar: /p/, /b/ e /t/. Então, a partir dessa constatação pode-se estudar o porque disso.
- Outras bases de dados também podem e devem ser utilizadas, para reforçar a capacidade que a rede SOM tem para reconhecer padrões de fala. Também pode se tentar aumentar a quantidade de padrões da base de dados confeccionada neste trabalho para que um treinamento mais completo seja realizado.

Bibliografia

[1] ALENCAR, V. F. S. *Atributos e domínios de interpolação eficientes em reconhecimento de voz distribuído*. Dissertação de mestrado. PUC-Rio, 2005.
[2] ANDRADE, A. O., SOARES, A. B. *Técnicas de janelamento de sinais*. Universidade Federal de Uberlândia, 2002.
[3] ANDREÃO, R.V. *Implementação em tempo real de um sistema de reconhecimento de dígitos conectados*. Dissertação de mestrado. UNICAMP, 2001.
[4] Audacity. Software disponível em: <http://sourceforge.net/projects/audacity/>. Acesso em 2 de fevereiro de 2006.
[5] BRAGA, A. P., CARVALHO, A. P. L. F., LUDERMIR, T. B. *Redes Neurais Artificiais: Teoria e Aplicações*, Rio de Janeiro: LTC, 2000.
[6] CARICATI, A. M., WEIGANG, L. *Reconhecimento de locutores em língua portuguesa com modelos de redes neurais e gaussianos*. UNB, 2001.
[7] CHOU, W., JUANG, B.H. *Pattern recognition in speech and language processing*. CRC Press, 2003.
[8] CHU, W. C. *Speech coding algorithms*. Wiley-Interscience, 2003.
[9] Davis, H., BIDDULPH, R., BALASHEK, S. *Automatic recognition of spoken digits*. The Journal of the Acoustical Society of America, 1952.
[10] DIAS, R. S. F. *Normalização de locutor em Sistemas de Reconhecimento de Fala*. Tese de mestrado. UNICAMP, SP, 2000.
[11] Eclipse Platform Technical Overview. Object Technology International Inc., Software disponível em: <http://www.eclipse.org>. Acesso em 20 de outubro de 2005.
[12] FURUI, S. *Digital speech processing, synthesis and recognition*. Marcel Dekker, Inc., 1989.
[13] FURUI, S. *Speech Recognition - Past, Present and Future*. NTT Review, 1995.
[14] HAYKIN, S. *Redes Neurais: Princípios e Práticas*, ed. Trad. Paulo Martins Engel. Porto Alegre: Bookman, 2001.
[15] HAYKIN, S., VEEN, B. V. *Signals and systems*. Wiley, 2002.
[16] ISHI, C. T. *Análise de um sistema de reconhecimento de voz baseado em fonemas.* Dissertação de mestrado. Centro Técnico Aeroespacial, 1998.
[17] *Java Speech API Programmer's Guide v 1.0*. Sun Microsystems, Inc., 1998.
[18] Java Technology. Sun Microsystems, Inc., <http://java.sun.com>. Acesso em 20 de outubro de 2005.
[19] JMusic. Music Composition in Java. <http://jmusic.ci.qut.edu.au/index.html>. Acesso em 2 de fevereiro de 2006.
[20] KOHONEN, T. *Self-Organizing Maps*. Springer, 2001.
[21] KOHONEN, T. *The "Neural" Phonetic Typewriter*, Helsinki University of Technology, March 1988.

[22] KOHONEN, Teuvo et al. The Self-Organizing Map Program Package. Helsink, Finlândia: Helsink University of Technology, 2996.

[23] MAFRA, A. T. *Reconhecimento automático de locutor em modo independente de texto por Self-Organizing Maps*. Dissertação de mestrado. USP, SP, 2002.

[24] MARTINS, J. A. *Avaliação de diferentes técnicas para reconhecimento de fala*. Tese de doutorado. UNICAMP, SP, 1997.

[25] MORAES, E. S. *Reconhecimento automático de fala contínua empregando Modelo Híbridos ANN + HMM*. Dissertação de mestrado. UNICAMP, SP, 1995.

[26] PETRY, A. *Reconhecimento Automático de Locutor Utilizando medidas de invariantes dinâmicas não-lineares*. Tese de doutorado. UFRS, 2002.

[27] RABINER, L. R., JUANG, B. H. *Fundamentals of speech recognition*. Prentice Hall, 1993.

[28] RABINER, L. R., SCHAFER, R.W. *Digital processing of speech signals*. Prentice Hall, 1978.

[29] RABINER, L. R. *A tutorial on hidden markov models and selected applications in speech recognition*. Proceedings of the IEEE, VOL. 77, Nº. 2, 1989.

[30] PAULA, M. B. *Reconhecimento de palavras faladas utilizando redes neurais artificiais*. Monografia de final de curso. UFPEL, 2000.

[31] REYNOLDS, D. A., QUATIERI, T. F., DUNN, R. B. *Speaker verification using adapted gaussian mixture models*. Digital signal processing review journal, 2000.

[32] REYNOLDS, D. A. *Speaker identification and verification using gaussian mixture models.* Speech communications, v.17, p.91-108, 1995.

[33] RUSSEL, S., NORVIG, P. *Inteligência Artificial*, São Paulo: Campus, 2º Edição, 2004.

[34] SANTOS, S. C. B., ALCAIM, A. *Sílabas como unidades fonéticas para o reconhecimento automático de voz em português*. IME e PUC, 2001.

[35] SILVA, F. J. F. *Conversão fala-texto em português do Brasil integrando segmentação sub-silábica e vocabulário ilimitado*. Tese de doutorado. Centro Técnico Aeroespacial, 1998.

[36] SIMÕES, F. O. *Implementação de um Sistema de Conversão Texto-Fala para o Português do Brasil*. Tese de Mestrado. UNICAMP, SP, 1999.

[37] SOUSA, L. C. *Adaptação de Locutor de Sistemas de Reconhecimento de Fala Contínua Empregando "Eigenvoices"*. Dissertação de Mestrado. UNICAMP, SP, 2004.

[38] TEBELSKIS, J. *Speech recognition using neural networks*. Pittsburg¸ Pensilvania. PhD Thesis. School of Computer Science. Carnegie Mellon University. 1995.

[39] THE MATHWORKS. Matlab 7 (R) - The Language of Technical Computing. The Mathworks Inc. 2004. Disponível em: <https://tagteamdbserver.mathworks.com/ttserverroot/Download/18842_ML_91199v00.pdf> Acesso em: 1 maio 2006.

[40] VERSANTO, Juha et al. SOM Toolbox for Matlab 5. Helsinki, Finlândia: Helsinki University of Technology, 2000.

[41] YNOGUTI, C. A. *Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov*. Tese de Doutorado. UNICAMP, SP, 1999.

Apêndice A

Frases com sua descrição fonética

1. A questão foi retomada no congresso.
   # a k e s t an un f o y rr e t o m a d a n o k on g r E s u #)
2. Leila tem um lindo jardim.
   # l e y l a t e in un l in d u j a R D in #
3. O analfabetismo é a vergonha do país.
   # u a n a u f a b e T i z m u E a v e R g on N a d u p a i s #
4. A casa foi vendida sem pressa.
   # a k a z a f o y v e in D i d a s e in p r E s a #
5. Trabalhando com união rende muito mais.
   # t r a b a L an d u k on u n i an un rr e in D y m u y t u m a y s #
6. Recebi nosso amigo para almoçar.
   # rr e s e b i n O s u a m i g u p a r a u m o s a R #
7. A justiça é a única vencedora.
   # a j u s T i s a E a u n i k a v e in s e d o r a #
8. Isso se resolverá de forma tranqüila.
   # i s u s y rr e z o u v e r a D y f O R m a t r an k u y l a #
9. Os pesquisadores acreditam nessa teoria.
   # u s p e s k i z a d o r y z a k r e D i t an un n E s a t e o r i a #
10. Sei que atingiremos o objetivo.
    # s e y k y a T in j i r e m u z u o b y j e T i v u #
11. Nosso telefone quebrou.
    # n O s u t e l e f o n y k e b r o u #
12. Desculpe se magoei o velho.
    # D y s k u p y s y m a g u e y u v E L u #
13. Queremos discutir o orçamento.
    # k e r e m u z D i s k u T i r u o R s a m e in t u #
14. Ela tem muita fome.
    # E l a t e in m u y t a f O m y #
15. Uma índia andava na mata.
    # u m a in D y a an d a v a n a m a t a #

16. Zé, vá mais rápido!
    # z E # v a m a y s rr a p i d u #
17. Hoje dormirei bem.
    # o j y d o R m i r e y b e in #
18. João deu pouco dinheiro.
    # j u an un d e u p o u k u D in N e y r u #
19. Ainda são seis horas.
    # a in d a s an un s e y z O r a s #
20. Ela saía discretamente.
    # E l a s a i a D i s k r E t a m e in T y #
21. Eu vi logo a Iôiô e o Léo.
    # e u v i l O g u a y o y o y u l E u #
22. Um homem não caminha sem um fim.
    # un O m e in n an un k a m in N a s e in un f in #
23. Vi Zé fazer essas viagens seis vezes.
    # v i z E f a z e r E s a z v i a j e in s e y z v e z y s #
24. O atabaque do Tito é coberto com pele de gato.
    # u a t a b a k y d u T i t u E k u b E R t u k un p E l y D y g a t u #
25. Ele lê no leito de palha.
    # e l y l e n u l e y t u D y p a L a #
26. Paira um ar de arara rara no Rio Real.
    # p a y r a un a R D y a r a r a rr a r a n u rr y u rr e a u #
27. Foi muito difícil entender a canção.
    # f o y m u y t u D i f i s y u in t en d e r a k an s an un #
28. Depois do almoço te encontro.
    # d e p o y z d u a u m o s u T in k on t r u #
29. Esses são nossos times.
    # e s y s an un n O s u s T i m y s #
30. Procurei Maria na copa.
    # p r o k u r e y m a r i a n a k O p a #
31. A pesca é proibida nesse lago.
    # a p E s k a E p r o i b i d a n e s y l a g u #
32. Espero te achar bem quando voltar.
    # y s p E r u T y a x a R b e in k u an d u v o u t a R #
33. Temos muito orgulho da nossa gente.
    # t e m u z m u y t u o R g u L u d a n O s a j e in T y #
34. O inspetor fez a vistoria completa.
    # u in s p e t o R f e y z a v i s t o r i a k on p l E t a #
35. Ainda não se sabe o dia da maratona.
    # a in d a n an un s y s a b y u D i a d a m a r a t o n a #
36. Será muito difícil conseguir que eu venha.
    # s e r a m u y t u D i f i s y u k on s e g i R k y e u v en N a #
37. A paixão dele é a natureza.
    # a p a y x an un d e l y E a n a t u r e z a #
38. Você quer me dizer a data?
    # v o s e k E R m y D i z e r a d a t a #
39. Desculpe, mas me atrasei no casamento.
    # D y s k u p y # m a z m y a t r a z e y n u k a z a m e in t u #

40. Faz um desvio em direção ao mar!
    # f a z un d e z v i u en D i r e s an un a u m a R #
41. A velha leoa ainda aceita combater.
    # a v E L a l e o u a a in d a s e y t a k on b a t e R #
42. É hora do homem se humanizar mais.
    # E O r a d u O m e in s y u m an n i z a R m a y s #
43. Ela ficou na fazenda por uma hora.
    # E l a f i k o u n a f a z en d a p u r un m a O r a #
44. Seu crime foi totalmente encoberto.
    # s e u k r i m y f o y t o t a u m e in T y en k o b E R t u #
45. A escuridão da garagem assustou a criança
    # a y s k u r i d an un d a g a r a j e in a s u s t o u a k r i an s a #
46. Ontem não pude fazer minha ginástica.
    # on t e in n an un p u D y f a z e R m in N a j i n a s T i k a #
47. Comer quindim é sempre uma boa pedida.
    # k o m e R k in D in E s e in p r y un m a b o u a p e D i d a #
48. Hoje eu irei precisar de você.
    # o j i e u i r e y p r e s i z a R D y v o s e #
49. Sem ele o tempo flui num ritmo suave.
    # s e in e l y u t en p u f l u y n un rr i T y m u s u a v y #
50. A sujeira lançada no rio contamina os peixes.
    # a s u j e y r a l an s a d a n o rr y u k on t a m i n a u s p e y x y s #
51. O jogo será transmitido bem tarde.
    # u j o g u s e r a t r an z m i T i d u b e in t a R D y #
52. É possível que ele já esteja fora de perigo.
    # E p o s i v e u k y e l y j a y s t e j a f O r a D y p e r i g u #
53. A explicação pode ser encontrada na tese.
    # a e s p l i k a s an un p O D y s e r in k on t r a d a n a t E z y #
54. Meu vôo tinha sido marcado para as cinco.
    # m e u v o u T in N a s i d u m a R k a d u p a R a s in k u #
55. Daqui a pouco a gente irá pousar.
    # d a k i a p o u k u a j en T i r a p o u z a R #
56. Estou certo que mereço a atenção dela.
    # y s t o u s E R t u k y m e r e s u a t en s an un d E l a #
57. Era um belo enfeite todo de palha.
    # E r a un b E l u in f e y T y t o d u D y p a L a #
58. O comércio daqui tem funcionado bem.
    # u k o m E R s y u d a k y t e in f un s y o n a d u b e in #
59. É a minha chance de esclarecer a notícia.
    # E a m in N a x an s y D y s k l a r e s e r a n o T i s y a #
60. A visita transformou-se numa reunião íntima.
    # a v i z i t a t r an s f o R m o u s y n un m a rr e u n i an un in T i m a #
61. O cenário da história é um subúrbio do Rio.
    # u s e n a r y u d a i s t O r y a E un s u b u R b y u d u rr y u #
62. Eu tenho ótima razão para festejar.
    # e u t e N u O T i m a rr a z an un p a r a f e s t e j a R #
63. A pequena nave medirá o campo magnético.
    # a p y k e n a n a v y m e D i r a u k an p u m a g y n E T i k u #

64. O prêmio será entregue sem sessão solene.
    # u p r e m y u s e r a in t r E g y s e in s e s an un s o l e n y #
65. A ação se passa numa cidade calma.
    # a s an un s y p a s a n u m a s i d a D y k a u m a #
66. Ela e o namorado vão a Portugal de navio.
    # E l a y u n a m o r a d u v an un a p o R t u g a u D y n a v i u #
67. O adiamento surpreendeu a mim e a todos.
    # u a D i a m e in t u s u R p r i en d e u a m in y a t o d u s #
68. A gente sempre colhe o que plantou.
    # a j e in T y s e in p r y k O L y u k y p l an t o u #
69. Aqui é onde existem as flores mais interessantes.
    # a k i E on D y e z i s t e in a s f l o r y z m a y z in t e r e s an T y s #
70. A corrida de inverno aconteceu com vibração.
    # a k o rr i d a D in v E R n u a k on t e s e u k o un v i b r a s an un #
71. Esse empreendimento será de enorme sucesso.
    # e s in p r e en D i m en t u s e r a D y e n O R m y s u s E s u #
72. As feiras livres não funcionam amanhã.
    # a s f e y r a z l i v r y z n an un f un s i o n an un a m an N an #
73. Fumar é muito prejudicial à saúde.
    # f u m a r E m u y t u p r e j u D i s i a u a s a u D y #
74. Entre com seu código e o número da conta.
    # e in t r y k o un s e u k O D i g u y u n u m e r u d a k on t a #
75. Reflita antes e discuta depois.
    # rr e f l i t a an T y z y D i s k u t a d e p o y s #
76. As aulas dele são bastante agradáveis.
    # a z a u l a z d e l y s an un b a s t an T y a g r a d a v e y s #
77. Usar aditivos pode ser desastroso.
    # u s a r a D i T i v u s p O D y s e R d e z a s t r o z u #
78. O clima não é mau em Calcutá.
    # u k l i m a n an un E m a u e in k a u k u t a #
79. A locomotiva vem sem muita carga.
    # a l o k o m o T i v a v e in s e in m u y t a k a R g a #
80. Ainda é uma boa temporada para o cinema.
    # a in d a E u m a b o u a t en p o r a d a p a r a u s i n e m a #
81. Os maiores picos da Terra ficam debaixo d'água.
    # u z m a y O r y s p i k u z d a t E rr a f i k an un d e b a y x u d a g u a #
82. A inauguração da vila é quarta-feira.
    # a i n a u g u r a s an un d a v i l a E k u a R t a f e y r a #
83. Só vota quem tiver o título de eleitor.
    # s O v O t a k e in T i v E r u T i t u l u D y e l e y t o R #
84. É fundamental buscar a razão da existência.
    # E f un d a m en t a u b u s k a r a rr a z an un d a e z i s t e in s y a #
85. A temperatura só é boa mais cedo.
    # a t e in p e r a t u r a s O E b o u a m a i s e d u #
86. Em muitas regiões a população está diminuindo.
    # e in m u y t a s rr e j i o in z a p o p u l a s an un y s t a D i m i n u in d u #
87. Nunca se pode ficar em cima do muro.
    # n un k a s y p O D y f i k a r in s i m a d u m u r u #

88. Pra quem vê de fora o panorama é desolador.
p r a k e in v e D y f O r a u p an n o r an m a E d e z o l a d o R
89. É bom te ver colhendo flores.
E b o un T y v e R k o L en d u f l o r y s
90. Eu me banho no lago ao amanhecer.
e u m y b an N u n u l a g u a u a m an N e s e R
91. É fundamental chegar a uma solução comum.
E f un d a m en t a u x e g a r a u m a s o l u s an un k o m un
92. Há previsão de muito nevoeiro no Rio.
a p r e v i z an un D y m u y t u n e v o e y r u n u rr y u
93. Muitos móveis virão às cinco da tarde.
m u y t u z m O v e y z v i r an un a s in k u d a t a R D y
94. A casa pode desabar em algumas horas.
a k a z a p O D y d e z a b a r e in a u g u m a z O r a s
95. O candidato falou como se estivesse eleito.
u k an D i d a t u f a l o u k o m u s y s T i v E s y e l e y t u
96. A idéia é falha, mas interessa.
a i D E y a E f a L a # m a z in t e r E s a
97. O dia está bom para passear no quintal.
u D i a y s t a b o un p a r a p a s y a R n u k in t a u
98. Minhas correspondências não estão em casa.
m in N a s k o rr e s p on d e in s y a z n an un y s t an un in k a z a
99. A saída para a crise dele é o diálogo.
a s a i d a p a r a k r i z y d e l y E u D i a l u g u
100. Finalmente o mau tempo deixou o continente.
f i n a u m e in T y u m a u t e in p u d e i x o u k on T i n e in T y
101. Um casal de gatos come no telhado.
un k a z a u D y g a t u s k O m y n u t e L a d u
102. A cantora foi apresentar seu último sucesso.
a k an t o r a f o y a p r e z en t a R s e u T i m u s u s E s u
103. Lá é um lugar ótimo para tomar uns chopinhos.
l a E un l u g a r O T i m u p a r a t o m a r un s x o p in N u s
104. O musical consumiu sete meses de ensaio.
u m u z i k a u k o un s u m y u s E T y m e z y z D in s a y u
105. Nosso baile inicia após as nove.
n O s u b a y l i n i s i a p O z a z n O v y
106. Apesar desses resultados, tomarei uma decisão.
a p e z a R d e s y s rr e z u t a d u s # t o m a r e y un m a d e s i z an un
107. A verdade não poupa nem as celebridades.
a v e R d a D y n an un p o u p a n e in a s e l e b r i d a D y s
108. As queimadas devem diminuir este ano.
a s k e y m a d a z d E v e in D i m i n u i R e s T y an n u
109. O vão entre o trem e a plataforma é muito grande.
u v an un en t r y u t r e in y a p l a t a f O R m a E m u y t u g r an D y
110. Infelizmente não compareci ao encontro.
in f e l i z m en T y n an un k on p a r e s i a u en k on t r u
111. As crianças conheceram o filhote de ema.
a s k r i an s a s k on N e s e r an un o f i L O T y D y e m a

112. A bolsa de valores ficou em baixa.
a b o u s a D y v a l o r y s f i k o u e in b a y x a
113. O congresso volta atrás em sua palavra.
u k on g r E s u v O u t a t r a z e in s u a p a l a v r a
114. A médica receitou que eles mudassem de clima.
a m E D i k a rr e s e y t o u k y e l y z m u d a s e in D y k l i m a
115. Não é permitido fumar no interior do ônibus.
n an un E p e R m i T i d u f u m a R n u in t e r i o R d u o n i b u s
116. A apresentação foi cancelada por causa do som.
a p r e z en t a s an un f o y k an s e l a d a p u R k a u z a d u s o un
117. Uma garota foi presa ontem à noite.
u m a g a r o t a f o y p r e z a on t e in a n o y T y
118. O prato do dia é couve com atum.
u p r a t u d u D y a E k o u v y k o un a t un
119. Eu viajarei ao Canadá amanhã.
e u v i a j a r e y a u k a n a d a a m an N an
120. A balsa é o meio de transporte daqui.
a b a u s a E u m e y u D y t r an s p O R T y d a k i
121. O grêmio ganhou a quadra de esportes.
u g r e m y u g an N o u a k u a d r a D y s p O R T y s
122. Hoje irei à vila sem meu filho.
o j i r e y a v i l a s e in m e u f i L u
123. Essa magia não acontece todo dia.
E s a m a j i a n an un a k on t E s y t o d u D i a
124. Será bom que você estude esse assunto.
s e r a b o un k y v o s e y s t u D y e s y a s un t u
125. O menu incluía pratos bem saborosos.
u m e n u in k l u i a p r a t u z b e in s a b o r O z u s
126. Podia dizer as horas, por favor?
p u D i a D i z e r a z O r a s # p u R f a v o R
127. A casa é ornamentada com flores do campo.
a k a z a E o R n a m en t a d a k on f l o r y z d u k an p u
128. A Terra é farta, mas não infinita.
a t E rr a E f a R t a # m a z n an un in f i n i t a
129. O sinal emitido é captado por receptores.
u s i n a u e m i T i d u E k a p y t a d u p u R rr e s e p y t o r y s
130. A mensalidade aumentou mais que a inflação.
a m e in s a l i d a D y a u m e in t o u m a i s k y a in f l a s an un
131. O tele-jornal termina às sete da noite.
u t E l e j o R n a u t e R m i n a s E T y d a n o y T y
132. A cabine telefônica fica na próxima rua.
a k a b i n y t e l e f o n i k a f i k a n a p r O s i m a rr u a
133. Defender a ecologia é manter a vida.
d e f en d e r a e k o l o j i a E m an t e r a v i d a
134. Nesse verão o calor está insuportável.
n e s y v e r an un u k a l o r y s t a in s u p o R t a v e u
135. Um jardim exige muito trabalho.
un j a R D in e z i j y m u y t u t r a b a L u

136. O mamão que eu comprei estava ótimo.
u m an m an un k y e u k on p r e y y s t a v a O T i m u
137. Meu primo falará com a gerência amanhã.
m e u p r i m u f a l a r a k o un a j e r e in s y a m an N an
138. De dia apague a luz sempre.
D y D i a a p a g i a l u s en p r y
139. A sociedade uruguaia tem que se mobilizar.
a s o s i e d a D y u r u g u a y a t e in k y s y m o b i l i z a R
140. Suas atitudes são bem calmas.
s u a z a T i t u D y s an un b e in k a u m a s
141. Dezenas de cabos eleitorais buscavam apoio.
d e z e n a z D y k a b u z e l e y t o r a y z b u s k a v an un a p o y u
142. A vitória foi paga com muito sangue.
a v i t O r y a f o y p a g a k o un m u y t u s an g y
143. Nossa filha tem amor por animais.
n O s a f i L a t e in a m o R p u r an n i m a y s
144. Esse peixe é mais fatal que certas cobras.
e s y p e y x y E m a y s f a t a u k y s E R t a s k O b r a s
145. O time continua lutando pelo sucesso.
u T i m y k on T i n u a l u t an d u p e l u s u s E s u
146. Essa medida foi devidamente alterada.
E s a m e D i d a f o y d e v i d a m e in T y a u t e r a d a
147. O estilete é uma arma perigosa.
u y s T i l e T y E u m a a R m a p e r i g O z a
148. Aguarde, quinta eu venho jantar em casa.
a g u a R D y # k in t a e u v en N u j an t a r e in k a z a
149. A mudança é lenta, porém duradoura.
a m u d an s E l e in t a p o r e in d u r a d o u r a
150. O clima não é mais seco no interior.
u k l i m a n an un E m a y s e k u n u in t e r i o R
151. A sensibilidade indicará a escolha.
a s e in s i b i l i d a D in D i k a r a e s k o L a
152. A Amazônia é a reserva ecológica do globo.
a m a z o n y a E a rr e z E R v a e k o l O j i k a d u g l o b u
153. O ministério mudou demais com a eleição.
u m i n i s t E r y u m u d o u D y m a y s k un a e l e y s an un
154. Novos rumos se abrem para a informática.
n O v u s rr u m u s y a b r in p a r a in f o R m a T y k a
155. O capital de uma empresa depende da produção.
u k a p i t a u D y u m a in p r e z a d e p e in D y d a p r o d u s an un
156. Se não fosse ela, tudo seria contido.
s y n an un f o s y E l a # t u d u s e r i a k on T i d u
157. A principal personagem no filme é uma gueixa.
a p r in s i p a u p e R s o n a j e in n u f y u m y E u m a g e y x a
158. Receba seu jornal em sua casa.
rr e s e b a s e u j o R n a u e in s u a k a z a
159. A juventude tinha que revolucionar a escola.
a j u v en t u D y T in N a k y rr e v o l u s y o n a r a y s k O l a

160. A atriz terá quatro meses para ensaiar seu canto.
a t r i s t e r a k u a t r u m e z y s p a r a in s a y a R s e u k an t u
161. Muito prazer em conhecê-lo.
m u y t u p r a z e r in k on N e s e l u
162. Eles estavam sem um bom equipamento.
e l y z y s t a v an un s e in un b o un e k i p a m e in t u
163. O sol ilumina a fachada de tarde.
u s O u i l u m i n a f a x a d a D y t a R D y
164. A correção do exame está coerente.
a k o rr e s an un d u y z an m y s t a k o e r e in T y
165. As portas são antigas.
a s p O R t a s an un an T i g a s
166. Sobrevoamos Natal acima das nuvens.
s o b r e v o an m u z n a t a u a s i m a d a z n u v e in s
167. Trabalhei mais do que podia.
t r a b a L e y m a y z d o k y p u D i a
168. Hoje eu acordei muito calmo.
o j y e u a k o R d e y m u y t u k a u m u
169. Esse canal é pouco informativo.
e s y k a n a u E p o u k u in f o R m a T i v u
170. Parece que nascemos ontem.
p a r E s y k y n a s e m u z on t e in
171. Receba meus parabéns pela apresentação.
rr e s e b a m e u s p a r a b e in s p e l a p r e z en t a s an un
172. Eu planejo uma viagem no feriado.
e u p l an n e j un m a v i a j e in n u f e r i a d u
173. No lado de cá do rio há uma boa sombra.
n u l a d u D y k a d u rr y u a u m a b o u a s on b r a
174. A maioria dos visitantes gosta deste monumento.
a m a y o r i a d u z v i z i t an T y s g O s t a d e s T y m o n u m e in t u
175. Minha filha é especialista em música sacra.
m in N a f i L a E y s p e s y a l i s t a e in m u z i k a s a k r a
176. A casa só tem um quarto.
a k a z a s O t e in un k u a R t u
177. A duração do simpósio é de cinco dias.
a d u r a s an un d u s in p O z y u E D y s in k u D i a s
178. Ao contrário de nossa expectativa, correu tranqüilo.
a u k on t r a r y u D y n O s y s p e k y t a T i v a # k o rr e u t r an k u i l u
179. A intenção é obter apoio do governante.
a in t en s an un E o b y t e r a p o y u d u g o v e R n an T y
180. A fila aumentou ao longo do dia.
a f i l a u m en t o u a u l on g u d u D i a
181. À noite a temperatura deve ir a zero.
a n o y T y a t en p e r a t u r a d E v i r a z E r u
182. A proposta foi inspecionada pela gerência.
a p r o p O s t a f o in s p e s y o n a d a p e l a j e r e in s y a
183. O quadro mostra uma face do cotidiano.
u k u a d r u m O s t r a u m a f a s y d o k o T i D i an n u

184. Já era bem tarde quando ele me abordou.
j a E r a b e in t a R D y k u an d u e l y m y a b o R d o u
185. O canário canta ao amanhecer.
u k a n a r y u k an t a a u a m an N e s e R
186. A lojinha fica bem na esquina de casa.
a l O j in N a f i k a b e in n a y s k i n a D y k a z a
187. Meu time se consagrou como o melhor.
m e u T i m y s y k on s a g r o u k o m u o m e L O R
188. Um instituto deve servir a sua meta.
un in s T i t u t u d E v y s e R v i r a s u a m E t a
189. Ele entende quando se fala pausadamente.
e l in t e in D y k u an d u s y f a l a p a u z a d a m e in T y
190. Seu saldo bancário está baixo.
s e u s a u d u b an k a r y u y s t a b a y x u
191. O termômetro marcava um grau.
u t e R m o m e t r u m a R k a v a un g r a u
192. O discurso de abertura é bem longo.
u D i s k u R s u D y a b e R t u r a E b e in l on g u
193. Eu precisei de microfone na conferência.
e u p r e s i z e y D y m i k r o f o n y n a k on f e r e in s y a
194. Joyce esticou sua temporada até quinta.
j O y s y s T i k o u s u a t en p o r a d a a t E k in t a
195. Nada como um almoço ao ar livre.
n a d a k o m u un a u m o s u a u a R l i v r y
196. Nossa filha é a primeira aluna da classe.
n O s a f i L a E a p r i m e y r a l u n a d a k l a s y
197. Gostaria de deitar um pouco.
g o s t a r i a D y d e y t a r un p o u k u
198. Não fizemos uma viagem muito cansativa.
n an un f i z e m u z un m a v i a j e in m u y t u k an s a T i v a
199. Ainda tenho cinco telefonemas para dar.
a in d a t en N u s in k u t e l e f o n e m a s p a r a d a R
200. O hotéis do sudoeste são fantásticos.
u z o t E y z d u s u d o E s T y s an un f a n t a s T i k u s

Apêndice B

Resultados do reconhecimento das frases – Experimento 1

1. Nosso telefone quebrou.

n O s u t e l e f o n y k e b r o u

'-' '#' '#' '#' '#' '-' 'rr' '-' '-' '#' '#' '#' '-' '-' '#' '#' '#' '-' '-' '#' '#' '#' '-' '-' '-' '-' '-'
' '#' '#' '#' '#' '#' '-' '-' '-' '#' '-' '-' 'm' '-' '-' '-' 'rr' '-' '-' '-' 'O' 'O' '-' 'an' '-' 'O'
'-' '-' '-' '-' 's' '-' '-' '-' '-' '-' '-' '-' 'y' 'u' '-' '-' 'rr' '-' 'rr' '-' '-' '-' '-' 'en' '-' 'd' '-'
'-' '-' 'l' '-' 'E' 'E' '-' '-' 'E' 'E' '-' '-' '#' '#' '-' '#' 'f' 'f' '-' '-' '-' 'o' 'u' '-' 'un' 'on'
'o' 'an' 'o' 'o' '-' 'on' '-' '-' '-' 'D' 'en' 'en' 'N' 'N' 'e' '-' '-' '-' '-' '-' '-' '-' '-' '-' 'e' 'i'
'-' '-' '-' 'r' '-' '-' '-' '-' '-' '-' 'v' 'g' '-' '-' 'l' '-' 'un' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-'
'-' '-'
'-' '#' '#' '-' '-' '#' '#' '#' '#' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-'

2. Desculpe se magoei o velho.

D y s k u p y s y m a g u e y u v E L u

'#' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '-' '-' '-' 'v' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-'
'i' '-' '-' '-' 'z' '-' '-' 's' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' 'on' '-' '-' '-' '-'
'-' '-' '-' 'v' '-' '-' 'v' '-' 'rr' '-' '-' 'L' '-' 'i' 'L' '-' '-' '-' '-' '-' '-' '-' '-' 's' 's' '-' '-'
'en' 'en' '-' '-' '-' '-' '-' '-' 'an' 'an' '-' '-' '-' 'an' '-' '-' 'n' 'g' 'v' '-' '-' '-' '-' '-' '-' '-'
'r' '-' 'E' '-' 'e' 'e' 'i' '-' '-' '-' '-' 'in' 'd' 'g' 'u' '-' 'un' 'un' '-' 'R' '-' '-' '-' '-' 'rr' '-'
'-' '-' 'E' '-' '-' '-' 'E' '-' '-' '-' 'd' '-' '-' 'L' 'N' 'en' 'en' 'y' '-' '-' '-' 'g' 'g' '-' 'on' '-'
'-' '-' '-' '-' '-' '-' '-' '-' '#' '-' '-' '-' '-' '-' 'p' '-' '-' '-' '#' '#' '#' '-' '-' '-' '#' '-' '-' '#'
'#' '#' '-' '#' '#' '-' '-' '-' '-' '-' '#' '-' '-' '-' '-' '#' '-' '#' '-' '-'

3. Queremos discutir o orçamento.

k e r e m u z D i s k u T i r u o R s a m e i n t u

'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '#' '#' '#' '#' '-' '-' '#' '#' '#' '#'
'#' '#' '#' '#' '#' '#' '#' '-' '#' '#' '#' '#' '#' '-' '#' '#' '-' '-' '#' '#' '#' '#' '-' '#' '#' '-'

'-' '-' '-' '-' '#' '-' '#' '#' '-' '-' '-' '-' '-' 'T' '-' 'e' '-' 'i' 'i' 'e' '-' '-' '-' 'g' '-' '-' 'i' '-'
'i' 'e' 'i' 'e' 'e' 'N' '-' '-' '-' '-' '-' '-' '-' 'm' '-' '-' 'o' '-' '-' '-' '-' '-' '-' 's' '-' '-' '-'
'-' '-' '-' '-' '-' '-' '-' '-' '-' '-' 's' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' 'd' 'R' '-' 'D' '-'
'v' '-' '-' '-' 'x' '-' 'x' '-' '-' 'e' 'e' 'e' 'e' '-' '-' '-' 'e' '-' '-' '-' '-' '-' 'o' 'on' 'un' 'un'
'u' 'o' 'un' 'u' 'u' '-' '-' '-' 'R' '-' '-' '-' '-' 'z' 's' '-' '-' 's' 's' 's' 's' 's' '-' 'an' '-' '-' 'r'
'a' 'n' '-' 'on' '-' '-' '-' 'E' 'D' '-' 'e' 'e' 'e' '-' '-' '-' 'in' 'n' '-' '-' '-' '-' '-' '-' '-' '-' '-'
'#' '-' '-' '-' 'g' '-' 'g' '-' '-' 'n' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '#' '-' '-' '#' '-' '#' '-' '-'
'-' '-' '-' '#' '#' '#' '-' '#' '-' '-' '-' '-' 'f' 'f' '#' 'f' '#' '-' '#' '-' '#' '-' '#' '-' '#' '#' '#'
'#' '#' '#' '#' '#' '#' '-' '#' '-' '#' '#' '-' '#' '-' '-' '#' '#' '-' '-' '-' '-'

4. Ela tem muita fome.

E l a t e in m u y t a f O m y

'#' '-' '#' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '-' '#' '#' '-' '#' '-' '-' '#' '#' '#' '#' '-' '-' '-' '-' '#' '#' '-'
'-' '-' '-' '-' '-' '-' '-' '-' '-' '-' 'E' '-' '-' 'E' 'E' '-' 'r' 'r' 'r' 'r' 'r' '-' 'r' '-' '-' 'on' '-'
'rr' 'rr' '-' '-' '-' 'v' 'E' 'e' '-' 'e' '-' 'v' 'in' 'N' '-' '-' '-' 'en' 'en' 'in' '-' 'in' '-' '-' '-'
'un' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' 'en' 'an' '-' '-' '-' '-' 'g' '-' 'v' '-' '-'
'f' 'f' '#' 'f' '-' '-' '-' '-' 'o' 'O' 'o' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'o' 'o' '-' '-' '-' 'in'
'-' '-' 'e' 'in' 'l' 'l' 'l' '-' '-' '-' 'N' 'N' 'y' '-' '-' '-' '-' '-' '-' '-' '-' '#' '-' '-' '#' '#' '#'
'#' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '#' '-' '-' 'f' '#' '-' '-' '-' '#'
'-' '-' '#' '-' '-' '-' '-' '-' '-' '-'

5. Uma índia andava na mata.

u m a in D y a an d a v a n a m a t a

'#' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '-' '-' '#' '#' '#' '-' '-' '-' '#' '-' '-' '-
' '#' '#' '-' '-' '#' '#' '-' '-' '#' '#' '-' '-' '#' '-' '-' '-' '-' 'un' '-' '-' 'r' '-' '-' 'm' 'l' 'l'
'on' '-' '-' '-' '-' 'a' '-' 'E' '-' 'e' '-' '-' 'en' 'en' 'en' 'in' 'en' '-' '-' 'en' '-' '-' 'in' '-' 'D'
'-' '-' '-' 'T' 'j' '-' '-' '-' 'd' 'd' 'an' '-' '-' 'an' '-' '-' '-' 'r' 'a' 'r' 'r' 'p' 'r' '-' '-' 'r' 'r'
'r' '-' '-' 'on' '-' '-' '-' '-' '-' '-' 'r' '-' 'r' '-' 'a' '-' 'a' '-' '-' '-' 'on' 'm' '-' '-' 'on' '-' '-'
'O' '-' '-' '-' '-' '-' 'r' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' 'u' '-' '-' '-' '-' '-' 'a' '-' '-' '-'
'-' '-' '-' '-' '-' '-' '-' '-' 'L' '-' 'a' '-' 'r' '-' '-' '-' '-' '-' '-' a su '-' '-' '-' '-'
'#' '#' '#' '-' '#' '#' '#' '#' '-' '-' '-' '-' '-' '#' '#' '-' '#' '#' '#' '#' '#' '#' 'f' '-' '-' '-' '-'
'#' '#' '#' '#' '-' '#' '-' '-' '#' '#' '#' '-' '-' '#' '-' '#' '#' '#' '-' '#' '-' '-' '#' '#' '-' '#' '-
' '#' '#' '-' '-'

6. Zé, vá mais rápido!

z E # v a m a y s rr a p i d u

'#' '#'
'#' '#' '#' '#' '-' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-'
'-' '#' '#' 'f' 'f' '-' '#' '#' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' 'z' '-' 'an' '-' 'E' 'E' 'E'
'-' '-' '-' '-' '-' '-' '-' '-' '-' '-' 'E' 'p' 'rr' 'rr' 'rr' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-'
'-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '#' '-' '-' '-' '#' '-' '#' '#' '#' '#' '#' '-' '-' '#' '#'
'#' '#' '#' '-' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '#' '#' '#' '#' '#' '#' '#'
'#' '#' '#' '-' '-' '#' '#' '#' '#' '#' '#' '#' '#' '-' '-' '-' '-' 'l' '-' '-' '-' '-' 'a' '-' '-' '-' '-'
'on' '-' '-' 'D' 'on' '-' '-' 'a' 'a' '-' '-' '-' '-' '-' '-' '-' 's' 's' 's' '-' '#' '-' '-' '-' '-' 'a'
'a' 'a' 'a' 'a' 'a' 'a' '-' '-' '-' '-' '-' '-' 'rr' '-' '-' 'v' '-' 'e' '-' 'n' '-' 'D' 'D' '-' '-' '-' '-'
'-' '-' 'g' '-' 'g' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '#' '#' '#' '#' '#' '#' '#' '#' '#'
'#' '#' '#' '-' '#' '-' '#' '#' '#' '#' '#' '#' '#' '-' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '-'
'#' '#' '-' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '#' '#' '-' '-'

7. Hoje dormirei bem.

o j y d o R m i r e y b e in

'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '#' '-' '-' '#' '#' '#' '-' '#' '-' '-'
'#' '#' '#' '#' '#' '-' '-' '#' '-' '#' '#' '#' '#' '#' '-' '-' '-' '#' '-' '-' '#' '-' '-' 'D' 'on' '-'
'-' '-' '-' '-' '-' '-' '-' '-' 'j' 'j' 'j' 'j' '-' '-' 'L' '-' '-' 'v' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-'
'R' 'R' '-' '-' '-' '-' '-' '-' '-' 'N' 'in' 'in' '-' '-' '-' 'on' '-' 'rr' '-' 'L' '-' '-' '-' 'e' 'L' 'en'
'i' '-' '-' '-' '-' '-' '-' '-' '-' '-' 'rr' 'L' '-' 'e' 'e' 'L' 'e' 'e' 'en' 'N' 'N' 'in' '-' '-' 'in' '-'
'-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' 'f' 'f' '-' 'f' 'f' '-' '-' '-' 'f' '#' '-' '-' 'f' 'f'
'-' '-' '-' '#' '#' '#' '-' '#' '#' '#' '#' '#' '-' '#' '#' '-' '#' '#' '#' '#' '-' '-'

8. João deu pouco dinheiro.

j u an un d e u p o u k u D in N e y r u

'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '#' '-' '#' '#' '#' '#' '#' '#' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '#' '#' '-' '#' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-'
'-' '-' '-' '-' '-' '-' '-' '-' '-' '-' 'z' 'z' 'f' 'D' '-' '-' 'on' 'on' '-' '-' 'o' 'o' '-' '-' '-' 'O'
'O' 'u' 'O' 'o' 'o' 'o' '-' 'on' '-' '-' 'un' '-' 'g' 'r' '-' '-' '-' '-' '-' 'd' 'd' 'rr' 'O' '-' 'on'
'-' '-' '-' 'R' '-' '-' '-' '-' '-' '-' '-' 'on' '-' '-' '-' '-' '-' 'on' '-' '-' 'on' 'on' 'v' '-' '-' '-'
'-' '-' 'en' '-' 'd' '-' '-' 'on' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' 'in' 'in' 'in' 'in' 'in' 'in' 'in'
'-' '-' '-' '-' 'y' '-' '-' 'e' 'e' '-' '-' 'in' '-' '-' '-' '-' '-' '-' 'y' 'rr' '-' 'in' '-' '-' 'n' 'r' 'r'
'r' '-' 'g' 'u' '-'
'-' 'f' '-' '-' '-' 'g' '-' '-' '-' '-' '-' '#' '#' '-' '-' '-' '-' '-' '#' '#' '-' '#' '#' '-' '-' '-' '-'
'#' '-' '-' '-' '-' '-' '-'

9. Ainda são seis horas.

a in d a s an un s e y z O r a s

'#' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '-' '#' '#' '#' '#' '-' '#' '#'
'-' '-' '-' '-' '-' '#' '#' '-' '-' '-' '-' '-' 'a' 'a' 'a' '-' '-' '-' 'p' '-' '-' '-' '-' 'en' 'N' 'N'
'N' 'in' 'in' 'en' '-' '-' 'in' '-' 'E' 'E' 'an' '-' '-' '-' '-' '-' 'z' '-' 's' 's' 's' '-' '-' '-' 's' 'z'
'-' '-' 'an' '-' '-' 'on' 'on' '-' '-' '-' '-' 'z' 's' 's' 's' '-' 's' 's' 's' '-' 'en' '-' 'L' '-' '-' '-'
'-' '-' 'e' '-' 'y' '-' '-' 'z' '-' '-' '-' 's' '-' '-' 'u' 'd' '-' 'on' '-' '-' '-' '-' '-' '-' '-' '-' 'r'
'in' '-' '-' '-' 'p' '-' '-' '-' '-' '-' '-' 's' '-' 's' 's' 's' '-' 's' 's' 's' 's' 's' 's' 's' '-' '-' '-'
's' 's' 's' 's' 's' '-' '-' '-' 'R' '-' '-' 'v' '-' '-' '-' '-' 'v' '-' 'v' 'v' '-' '-' '-' '#' '#' '#' '#'
'-' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '#' '#' '#' '-' '-' '#' '#' '-' '#' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '-'

10. Ela saía discretamente.

E l a s a i a D i s k r E t a m e in T y

'#' '#'
'#' '-' '#' '-' '-' '#' '-'
'-' '-' '-' '-' '-' '-' '-' '-' '-' '#' '-' '-' '-' '-' '#' '#' '-' '-' '-' '-' '#' '-' '-' 'E' '-' '-' '-' '-'
'E' '-' '-' '-' '-' '-' '-' 'd' 'd' '-' '-' '-' '-' '-' 'z' '-' '-' '-' '-' '-' '-' '-' 'z' 'z' '-' '-' 'a'
'-' 'a' '-' '-' 'i' 'e' 'y' '-' '-' '-' '-' '-' '-' 'e' 'e' 'i' 'i' 'i' 'i' '-' 'L' '-' '-' '-' '-' 'D' 'D'
'-' '-' 'z' '-' 's' 's' '-' '-' '-' '-' 'f' '-' '-' '-' '-' 'y' '-' '-' '-' '-' 'E' 'E' '-' 'E' '-' 'D' '-'
'-' '-' '-' '-' 'D' '-' '-' 'a' '-' 'a' '-' 'n' 'n' '-' '-' '-' '-' '-' 'e' '-' '-' '-' 'in' 'en' 'N' '-' '-'
' '-' 'in' '-' '-' '-' '-' '-' '-' '-' '-' 'x' '-' 'x' 'T' '-' 'k' 'y' 'y' 'in' '-' '-' '-' '-' '-' '-' '-'
'-' '#' '#' '-' '#' '#' '#' '#' '#' '-' '-' '-' '-' 'g' 'f' '-' '#' '-' '-' '-' '-' '#' '#' '-' '#' '-' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '#' '-' '#' '#' '#' '-' '-' '#' '-' '-'

Apêndice C

Resultados do reconhecimento das frases – Experimento 2

1. Nosso telefone quebrou.

n O s u t e l e f o n y k e b r o u

'-' '-' '#' '#' '-' '-' '-' '#' '#' '-' '-' '-' '-' '-' '-' '-' '#' '-' '-' '-' '-' '-' '#' '#' '#' '#' '#' '#'
'#' '#' '#' '#' '#' 'v' '-' '-' '#' '-' '-' '-' '-' '-' '-' 'L' 'n' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O'
'-' '-' '-' '-' 's' 's' 's' 's' 'T' '-' '-' 'l' '-' 'u' 'u' 'u' '-' '-' '-' '-' '-' '#' '-' 'en' '-' 'e' 'e'
'e' '-' 'l' 'l' '-' '-' '-' 'en' 'e' '-' 'D' 'f' 'f' '-' '-' '-' '-' 'f' 'f' '-' '-' 'on' 'on' 'on' 'b' 'on'
'O' 'O' '-' 'o' 'on' 'on' 'on' 'n' '-' '-' '-' '-' '-' 'in' 'N' 'y' 'y' '-' 'k' 'k' '#' '#' '#' '-' '-'
'-' 'en' 'en' 'b' 'R' 'n' '-' '-' '-' 'i' 'en' '-' '-' 'r' 'r' 'R' 'o' '-' 'o' 'o' 'o' 'o' 'o' 'o' 'o'
'o' '-' 'o' 'o' 'o' 'on' '-' '-' '-' 'O' 'u' 'p' 'p' 'u' 'u' '-' '-' 'f' 'u' '-' '-' '#' '-' 'f' 'f' '-'
'#' '-' '-' '-' '#' '#' '-' '#' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '#' '#'

2. Desculpe se magoei o velho.

D y s k u p y s y m a g u e y u v E L u

'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '#' '#' '#' '-' '#' '#' '#' '#' '#' '#' '#' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' 'f' '-' 'f' 'D' 'f' 'D' 'f' 'D' 'D' 'j' 'j' 'z' '-' '-' '-'
'L' '-' 'T' 'T' 's' 's' 's' 's' 's' 's' 's' 'k' 'k' 'k' '#' '#' 'k' 'k' 'k' '-' '-' '-' 'o' 'o' '-' '-'
'u' 'u' '-' 'on' 'D' 'v' 'L' 'u' '-' '-' '-' 'rr' '-' 'i' 'i' 'i' '-' '-' 'in' 'y' 's' 's' 's' 's' 's' 's'
's' 's' '-' '-' 'R' 'R' 'n' 'on' '-' '-' 'O' 'O' 'O' 'O' '-' '-' 'on' 'on' 'an' '-' '-' '-' 'g' 'g'
'on' 'on' 'on' 'o' 'o' '-' 'r' 'R' '-' 'i' 'i' 'y' '-' '-' 'e' 'e' 'en' 'l' '-' '-' '-' 'an' '-' 'un' '-'
'-' 'd' 'f' 'rr' '-' 'r' 'r' 'E' 'E' 'E' 'E' 'E' '-' 'E' 'E' 'E' 'e' 'L' 'L' 'L' 'L' 'N' '-' 'en' 'i'
'y' '-' 'in' '-' '-' '-' 'on' '-' '-' '-' '-' '-' '-' 'u' '-' '-' '-' '-' '-' '-' '-' '-' '#' '-' '-' '#' '-'
'-' '-' '#' '-' '-' '#' 'T' '-' '-' '-' '-' '-' '#' '-' '#' '-' '-' '-' '-' '-' '-' '-' '#' '#' '-' '-' '#'

3. Queremos discutir o orçamento.

k e r e m u z D i s k u T i r u o R s a m e in t u

'#' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '-' '#' '#' '#' '#' '-' '#' '#' '-' '-' 'k'
'#' '#' '#' '#' '-' '-' '-' '-' '#' '-' 'x' 'x' 'x' '-' 'x' '-' '-' 'en' '-' '-' '-' '-' 'in' 'd' '-' 'L'

'e' '-' 'en' 'e' '-' 'o' 'en' 'N' 'N' 'd' '-' 'N' '-' '-' '-' '-' 'un' '-' '-' '-' '-' '-' '-' 'z' 's' 's'
's' 's' 'z' 'x' 'x' 'x' 'k' 'j' '-' 'y' '-' '-' '-' 'z' 's' 's' 's' 's' '-' '-' '-' '-' '-' 'k' '-' 'on'
'u' 'on' 'm' 'D' 'v' 'y' 'T' 'T' '#' 'x' 'x' 'x' 'D' '-' '-' 'i' 'i' 'i' 'i' 'i' 'i' '-' '-' 'en' 'un'
'-' '-' '-' '-' '-' '-' '-' 'o' '-' 'o' 'o' 'o' 'R' 'R' 'R' 'R' 'R' '-' 'D' 's' 's' 's' 's' 's' 's' 's'
's' 'y' '-' 'an' 'a' 'a' 'a' 'O' 'on' 'm' 'm' 'm' '-' '-' '-' 'e' '-' '-' 'en' '-' 'en' 'in' 'en' 'in'
'-' '-' 'en' 'en' 'en' '-' 't' '-' 'p' 'p' '-' 'r' 'r' '-' '-' 'u' 'un' 'u' 'u' '-' 'v' 'm' 'u' 'g' '-'
'u' 'u' 'u' '-' '-' '-' 'f' 'v' 'un' '#' 'k' '#' 'f' 'f' 'u' '#' '#' '#' '-' '#' '-' 'f' '-' '-' 'u' '-'
'-' 'f' '#' '-' '-' '#' '#' '#' '-' '-' '-' '#' '#' '#' 'f' '#' '#' '#' 'R' '#' '-' '-' '-' '#' '-' '#'
'#' '-' '-' '#' '#' '#' '#' '#' '#'

4. Ela tem muita fome.

E l a t e in m u y t a f O m y

'#' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '-' '#' '#' '#' '#' '-' '-' '-' '#' '-' '#'
'-' '-' '-' '#' 'l' 'R' 'R' '-' '-' 'E' 'a' 'E' 'E' 'E' 'E' 'E' '-' 'r' 'r' 'an' 'l' 'l' 'l' 'r' 'r' 'd'
'R' '-' '-' '-' 'k' 't' 'd' 'e' 'e' '-' 'en' 'y' 'R' 'in' 'in' 'in' 'in' '-' '-' '-' 'N' 'in' '-' '-' '-' '-
' un' '-' '-' 'on' 'an' 'y' 'y' 'in' '-' 'n' '-' 'y' '-' '-' 'en' 't' 'an' '-' 'E' 'a' 'a' 'R' 'O' 'L'
'-' 'L' 'f' 'f' 'f' '#' 'f' 'f' '-' '-' 'on' 'on' 'on' 'o' 'O' 'O' 'O' 'O' 'O' 'O' 'an' 'O' '-' '-' '-'
'-' '-' '-' '-' 'm' '-' '-' '-' '-' '-' 'in' 'in' '-' 'y' 'y' 'N' 'N' 'e' 'e' 'u' '-' '-' 'u' 'u' '-' '-'
'u' '#' '-' '#' '#' '-' '#' '#' '#' '-' '-' 'f' '-' '-' '#' '#' '-' '#' '-' '#' '-' '#' '-' '#' '-' '#' '#'
'-' '#' '-' '-' '-' '#' '-' '#' '#' '-' '#' '-' '-' 'f' 'f'

5. Uma índia andava na mata.

u m a in D y a an d a v a n a m a t a

'#' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '#' '#' '-' '#' '#' '-' '-' '-' '#' '#' '#' 'f'
'#' '-' '-' '-' '-' '-' '-' '-' '#' '#' '#' '-' '#' '-' '-' '-' un' un' un' on' '-' 'm' 'm' 'N' 'N'
'u' 'O' '-' 'O' 'a' 'a' 'l' 'l' 'en' 'rr' '-' '-' '-' 'en' 'in' 'en' 'N' 'in' 'N' 'in' 'in' 'in' 'in' '-'
'-' 'D' 'D' 'D' 'D' '-' 'j' '-' '-' 'E' 'E' 'E' 'r' 'E' 'R' 'a' 'a' 'a' 'a' '-' 'N' 'N' '-' 'R' 'r'
'an' 'R' 'an' 'an' 'r' 'R' '-' 'un' '-' 'in' 'in' 'in' '-' 'R' '-' 'O' 'l' '-' '-' '-' '-' 'O' 'a' 'on'
'on' '-' 'v' '-' 'O' 'O' 'O' 'O' 'a' 'a' 'n' 'm' 'n' 'r' 'a' 'a' 'a' '-' 'O' '-' '-' '-' '-' 'm' 'm'
'O' 'O' '-' 'O' '-' '-' '-' '-' '-' 'O' 'an' 'd' 'R' '-' '-' '-' '-' '-' '-' 'l' '-' 'an' 'a' 'R' 'R' 'R'
'R' '-' '-' 'an' '-' '-' '-' '#' '#' '#' '#' '-' '-' '-' '-' '#' '#' 'T' 'T' '-' '#' '#' '#' '#' '#' '#'
't' '#' 'f' 'f' '#' '-' '#' '#' '-' '#' 'f' '-' '-' '#' '-' '#' '-' '-' '#' '#' '#' '#' '-' '-' 'f' '#' '-'
'#' '#' 'R' '-' '-' '-' '-' '#' '-' '-' 'f' 'f'

6. Zé, vá mais rápido!

z E # v a m a y s rr a p i d u

'#' '#'
'#' '-' '#' 't' '-' '#' '-'
'#' '#' '#' un' 'f' '#' '#' '#' 'f' 'z' 'z' 'z' 'z' 'z' 'z' 'z' 'z' 'z' 'z' 'z' 'z' '-' 'E' 'E' 'en' 'E'
'E' 'E' 'E' '-' 'E' 'E' 'E' 'E' 'E' '-' 'E' 'en' 'en' 'en' 'f' '-' '-' 'k' 'k' 'k' 'k' 'R' 'R' 'R'
'R' 'R' '-' '-' '-' '-' '-' '#' 'rr' 'rr' '-' '-' '#' '#' '-' '-' '#' '#' '-' '#' '-' '-' '#' '#' '-' '#'
'-' 'R' 'f' '#' '#' '#' '#' '#' '-' '#' '#' '#' '#' '-' '#' '#' 'R' '#' '#' '#' '#' '#' '#' '#' 'z' '#'
'#' '#' '#' '#' '-' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' 'f' '#' '#' '-' '-' 'O' 'O' 'O'
'O' 'O' '-' '-' '-' 'O' '-' '-' 'm' 'm' '-' 'O' '-' '-' 'a' 'a' '-' '-' '-' '-' 'z' 'z' 's' 's' 's' 's'
'-' '-' 'k' 'k' 'k' '#' 'rr' 'rr' 'a' '-' 'a' 'a' 'a' '-' '-' 'O' 'a' '-' '-' 'k' '-' '-' 'L' 'i' 'i' 'i' '-'
'un' 'm' 'm' '-' 'T' 't' 'T' '-' 'un' '-' 'm' '-' '-' '-' '-' '-' 'v' 'v' 'R' '-' 'T' 'T' 'T' '#' '#'
'#' '#' '#' '#' '#' '#' 'f' 'f' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '#' 'z' '#' '-' '-' '#'
'f' '#' '#' '#' '#' '#' '#' '-' '-' 'f' 'f' '-' 'f' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '-' '#'
'#' '#'

7. Hoje dormirei bem.

o j y d o R m i r e y b e in

'#' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '-' '-' '-' '#' '#' 'f' '-' 'T' 'on' '-'
'o' 'o' '-' '-' 'r' 'r' 'y' '-' 'j' 'D' '-' 'j' '-' '-' 'L' 'v' 'v' '-' '-' 'v' '-' 'd' '-' '-' 'an' 'n'
'R' '-' 'R' 'u' 'on' '-' 'm' 'm' 'in' '-' '-' 'N' '-' 'N' '-' 'i' 'i' '-' '-' 'e' 'o' '-' '-' '-' '-' '-'
'-' 'i' 'y' 'i' 'i' 'i' 'en' 'L' 'L' 'L' '-' '-' '-' 'e' 'e' 'e' 'e' 'e' '-' 'e' 'y' 'in' 'in' 'in' 'in' '-'
'-' '-' 'in' 'in' 'in' 'in' 'in' 'in' '-' '-' '-' '-' '-' 'in' 'in' 'f' '-' '-' 'v' '-' '-' 'v' 'v' '-' 'v' 'f'
'f' 'f' 'u' '-' '-' '-' 'v' '#' 'f' '-' '-' '#' '#' '-' '#' '-' '#' '#' 'v' '#' '#' '#' '#' '-' '#' '#'

8. João deu pouco dinheiro.

j u an un d e u p o u k u D in N e y r u

'#' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' 'f' '-' '#' '-' '-' '-' '-' '#' '#' '#' 't' '#' '-' '-
' 'j' '-' 'j' 'j' 'T' 'D' 'k' 'j' 'y' '-' 'z' '-' '-' '-' 'un' '-' '-' '-' 'on' 'o' 'o' 'o' 'O' 'O' 'O'
'O' 'an' 'u' 'O' 'on' 'an' 'on' 'on' 'on' 'on' '-' '-' '-' '-' '-' 'D' 'l' 'l' '-' '-' 'o' 'r' '-' 'on'
'on' 'o' 'on' 'g' '-' '-' '-' 'k' '-' 'un' 'R' 'R' 'un' 'on' 'on' 'o' 'o' 'o' '-' '-' 'o' 'on' '-' '-'
'-' 'e' '-' '-' 'k' 'g' '-' '-' 'u' 'l' 'r' '-' '-' 'D' 'D' 'D' 'j' 'D' 'x' '-' 'y' 'N' 'N' 'N' 'in'
'en' 'in' '-' '-' '-' '-' 'in' 'y' '-' 'i' 'i' 'i' 'i' 'i' 'i' '-' 'i' 'i' 'i' 'i' 'i' 'i' 'N' 'N' '-' '-' '-'
'-' 'n' 'N' 'n' 'm' 'u' 'u' '-' '-' '-' '-' 'in' 'u' 'v' 'R' 'R' '-' '-' '-' '-' 'f' '#' 'v' 'v' 'f' 'v'
'v' '-' '-' 'f' 'v' '-' '-' '-' '-' 'f' '-' '-' '-' '-' 'v' 'v' '-' 'R' '#' '-' '#' '#' '#' '-' '-' '-' '-'
'-' 'f' '#' '-' '-' '-' '-' '-' '-' '-' '#' '#' '#'

9. Ainda são seis horas.

a in d a s an un s e y z O r a s

'#' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '#' '#' '#' '#' '#' '#' '#' '#' '-' '-' '#'
'-' '#' '#' '#' 'f' '#' '#' '-' '#' '#' 'O' 'a' 'a' 'a' 'a' '-' '-' 'a' 'D' 'N' '-' '-' '-' 'in' 'in'
'in' 'in' 'in' 'd' '-' '-' 'on' 'in' 'en' 'en' '-' '-' '-' 'a' 'E' 'a' '-' '-' 'z' 's' 's' 's' 's' 's' 's'
's' '-' '-' 'O' 'an' 'an' 'an' '-' '-' '-' '-' '-' '-' '-' 's' 's' 's' 's' 's' 's' 's' 's' '-' '-' '-' '-'
'en' 'en' '-' '-' 'i' '-' '-' 'i' 'z' 'z' 'z' 'z' 'z'
'z' 'z' '-' 'in' 'o' 'R' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'g' 'an' '-' '-' 'r' 'r' '-' '-' 'a' '-'
'-' '-' 's' 's' 's' 's' 't' 's' 's' 's' 't' '-' '-' 's' 's' 's' 's' 's' 's' 's' 's' 's' 's' '-' '-' '-'
'-' '-' '-' '-' '-' '-' '-' '-' '-' '-' '-' 'D' '-' 'f' '-' '#' '#' '-' '-' '#' '#' '#' '#' '#' '-' '#' '#'
'-' '-' '-' '#' '#' '-' '#' '-' '#' '-' '-' '-' '#' '#' '#' '#' '-' '#' '-' '#' '#' '#' '-' '-' '#' '-' '-'
'-' '#' '#' '-' '#' '#' '#'

10. Ela saía discretamente.

E l a s a i a D i s k r E t a m e in T y

'#' '#'
'#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '#' '-' '#' '#' '#' '#' '#' '#' 'v' '#' '-' '-' 'f' '-'
'#' '-' '-' '-' '-' 'R' '-' '-' '-' '#' '#' '#' '#' '-' '-' '#' '#' '#' '#' '-' '-' '-' '-' 'E' '-' 'E'
'E' 'E' 'E' 'E' 'E' '-' 'an' 'an' 'an' 'l' 'an' 'an' '-' 'an' '-' 'z' 'z' 'z' 's' 'z' 'z' 's' 's'
'-' '-' '-' 'a' 'a' '-' '-' '-' 'E' 'L' '-' '-' 'l' 'in' 'N' 'l' 'i' 'i' 'y' 'y' '-' '-' 'L' 'L' '-' '-' '-'
'-' 'D' '-' 'D' '-' 'y' 'y' 'z' 'z' 'z' 's' '-' '-' '-' '-' 'f' 'k' '-' 'k' 'r' 'E' 'E' 'rr' 'E' 'E' 'E'
'E' 'E' '-' 'R' 'E' 'g' 'rr' 'r' 'k' '-' '-' 'r' 'r' 'r' 'O' 'O' '-' 'm' 'm' 'm' 'en' 'en' '-' 'e' '-'
'en' 'en' 'in' 'in' '-' 'in' 'in' 'in' 'en' 'in' 'en' 'en' '-' '-' '-' 'T' 'T' 'T' 'x' 'x' 'x' '-' 'x' 'x'
'-' '-' 'en' '-' 'en' 'rr' 'rr' '-' '-' '-' '-' 'f' '#' '#' '#' '#' '#' '#' 'D' 'D' '-' '-' '-' '-' 'f' 'u'
'-' 'v' 'v' '-' '-' '#' '#' '-' '#' '-' '#' '-' '#' '#' '#' '#' 'f' 'R' '-' '#' '-' '-' '#' '#' '-' '-' '-
' 'T' '#' '-' '-' '#' 'R' '-' 'R'