FACULDADE DE TECNOLOGIA DE SÃO PAULO

ESTEVAM LEAL MORAES DA SILVA

UTILIZANDO O CSLU TOOLKIT PARA O RECONHECIMENTO AUTOMÁTICO DE FALA POR COMPUTADOR

São Paulo

2011

FACULDADE DE TECNOLOGIA DE SÃO PAULO

ESTEVAM LEAL MORAES DA SILVA

UTILIZANDO O CSLU TOOLKIT PARA O RECONHECIMENTO

AUTOMÁTICO DE FALA POR COMPUTADOR

Monografia submetida como exigência
parcial para a obtenção do Grau de
Tecnólogo em Processamento de Dados
Orientador: Prof. Gabriel Shammas

São Paulo

2011

Aos meus pais que nunca deixaram de me incentivar a chegar até aqui

AGRADECIMENTOS

Agradeço ao professor Gabriel Shammas pelo incentivo, paciência e compreensão para comigo.

A todos os amigos que acompanharam a minha tragetória e me incentivaram durante todos estes anos: Alex de Moraes, Daniel Bertuqui, Dimas Alves, Marco Aurélio de Salles e Rafael Araújo e a Ellen Chen pela colaboração.

RESUMO

Dotar máquinas da capacidade de reconhecer a fala humana é um problema desafiador estudado já a muito tempo e que saiu dos livros de ficção científica para a vida real nas últimas décadas, ajudando humanos desde a se comunicarem com equipamentos eletrônicos através da fala a até aumentar a produtividade e a qualidade de *call centers* ao fornecer dados valiosos por meio da transcrição automática de chamadas telefônicas, graças à constante evolução dos sistemas de reconhecimento automático de fala que têm paulatinamente reduzido suas taxas de erro na realização desta tarefa.

O objetivo principal deste trabalho é consolidar as diretrizes básicas que permeiam o assunto de reconhecimento automático de fala a fim de se obter uma visão geral dos desafios e soluções modernas comuns propostas para o tema. Optou-se aqui por não se aprofundar demais nos fundamentos matemáticos, estatísticos e computacionais inerentes ao problema, mas sim abordá-los de forma objetiva para que sejam úteis na compreensão do problema e possam servir de base para um estudo mais pormenorizado. Além disto, será apresentado neste trabalho o *CSLU Toolkit,* uma ferramenta poderosa para construção de sistemas de reconhecimento automático de fala que dispensa a necessidade de grandes conhecimentos de informática para o seu manuseio.

Palavras chaves: reconhecimento automático de fala; RAF; CSLU Toolkit;

ABSTRACT

Enabling machines of speech recognition capabilities is a challenging problem that has been studied for a long time and that has come out from sci-fi books to real life in the recent decades, helping humans from communicate with electronic devices through speech up to increase the productivity and the quality of call centers by mining valuable data out of phone calls with the automated speech recognition that has seen its error rates constantly decrease.

The main aim of this work is to consolidate the basic directives that surround the problem of the automated speech recognition and thus have an overview of the challenges and modern solutions commonly proposed to the subject. The option here is not get too deep in the mathematics, statistics and computational foundations, but to mention them objectively in a form that it may be useful in the comprehension of the matter and may serve as a basis to a more detailed investigation. Further, will be presented here the CSLU Toolkit, a powerful tool to authoring automated speech recognition systems that dismisses the need of a great computational knowledge to handle it.

Keywords: automatic speech recognition; ASR; CSLU Toolkit;

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

O reconhecimento automático de fala (RAF) é a tecnologia que confere aos autômatos a capacidade de interpretarem a fala humana. A pesquisa sobre o problema já é antiga e há relatos de um equipamento capaz de reconhecer dígitos falados já em 1952[1].

Porém, devido à grande capacidade computacional demandada por tais sistemas, a pesquisa não pode se desenvolver de forma eficaz até a década de 80, quando computadores mais potentes se tornaram comercialmente viáveis e, a partir de então, as principais pesquisas passaram a girar em torno de aproximações estatísticas baseadas nos Modelos Ocultos de Markov em conjunto com a Teoria das Redes Neurais[2].

A complexidade do assunto deriva do fato do problema exigir o estudo de diferentes modalidades de ciências, entre elas, reconhecimento de padrões, processamento de sinais, fonética, processamento de linguagem natural, ciências da computação, teoria da inxformação e inteligência artificial[3].

Existem hoje em dia inúmeras aplicações para o RAF: acionamento de funcionalidades de equipamentos através de comando de voz, como a discagem por

---

[1] JUANG, B. H.; RABINER, L. R. *Automatic Speech Recognition: A Brief History of the Technology Development*. Santa Barbara, EUA: Rutgers University and the University of California. p. 6.
[2] *Ibid*, p. 10.
[3] SEWARD, A. *Efficient Methods for Automatic Speech Recognition*. Estocolmo, Suécia: Royal Institute of Technology. p. 5.

voz em celulares ou acionamento de luzes em cômodos que contam com automação, interação com URAs (unidades de resposta audível), sinalização da ocorrência de determinados vocábulos em arquivos de áudio, transcrição de fala de gravações, etc..

Com o intuito de fomentar a pesquisa relacionada à interação homem-máquina, o *Center for Spoken Language Understanding* (CSLU), órgão ligado à *Oregon Health & Science University*, dos Estados Unidos, iniciou em 1992 o desenvolvimento de um conjunto de aplicativos de código aberto para este propósito, o *CSLU Toolkit*, que permite que até mesmo pessoas que não têm conhecimento em desenvolvimento de sistemas possam utilizá-lo. Uma das possibilidades de utilização do CSLU Toolkit é o auxílio no desenvolvimento de aplicativos capazes de transcrever o conteúdo falado e sintetizar a voz humana a partir de um texto escrito.

1.1 – OBJETIVO DESTE TRABALHO

Neste trabalho, procura-se consolidar os fundamentos básicos que permeiam o assunto de RAF para dar uma visão geral dos desafios e soluções comuns propostas para o tema. Optou-se aqui por não se aprofundar demais nos fundamentos matemáticos, estatísticos e computacionais inerentes ao problema, justamente por estes serem demasiadamente extensos e estão bem expostos na literatura especializada e aqui referenciada.

Espera-se que esta monografia sirva de base para o aprofundamento na pesquisa sobre o problema de RAF, com o apoio prático do *CSLU Toolkit* para a avaliação da aplicabilidade de sistemas de RAF para o mundo real e, quem sabe, para o desenvolvimento de uma ferramenta proprietária voltada para a solução de problemas relacionados.

1.2 – ESTRUTURA E CONTEÚDO DO TRABALHO

O capítulo 2 trata das características do som e da fala humana. Compreender como a fala é produzida e percebida pelos humanos é fundamental para que se possa desenvolver sistemas de RAF com taxas de acertos que viabilizem o uso prático destes.

O capítulo 3 descreve o problema do RAF e os desafios e dificuldades que o assunto carrega.

O capítulo 4 apresenta a estrutura comum dos sistemas de RAF modernos e aborda as soluções comuns adotadas por estes.

O capítulo 5 apresenta o *CSLU Toolkit* e mostra como utilizá-lo para a tarefa do RAF.

O capítulo 6 apresenta as considerações finais sobre o desenvolvimento deste trabalho.

2 CARACTERÍSTICAS DO SOM E DA FALA HUMANA

2.1 PROCESSO DA FALA

Uma seqüência de sons compõe os sinais da fala que, por sua vez, são orientados pelos critérios de linguagem e pelas peculiaridades do orador. Para a captação, a compreensão, a sintetização, o reconhecimento, enfim, o processamento dos sinais da fala imprescindível se faz o entendimento do mecanismo de sua produção.[4]

A produção da fala inicia-se com uma mensagem formulada a partir da formação de uma idéia a ser exteriorizada, ou seja, expressada. Com isso, o sistema de expressão lingüística é provocado, convertendo essa idéia em um conjunto de palavras para transmissão.[5]

A partir da definição das palavras e dos fonemas, segue-se para o mapeamento neuromuscular que, por sua vez, dá início aos trabalhos do trato vocal para que este possa emitir corretamente os sons associados à mensagem original.

Assim, depois do locutor executar a sua fala, há a propagação do som produzido pelo ar e que, por conseguinte, alcança o ouvinte. Cumprida essa etapa, inicia-se o processo de reconhecimento da fala. O destinatário da mensagem, ou seja, o

---

[4] Produção de Fala Humana. In: *DEETC – Departamento de Engenharia e Eletrônica e Telecomunicações e de Computadores*. Disponível em: <http://www.deetc.isel.ipl.pt/comunicacoesep/disciplinas/pdf/sebenta/pdf/producao_2.pdf>. Acesso em 25 nov. 2011.

[5] Ibid.

ouvinte, é estimulado por meio da membrana de seu ouvido que, por sua vez, realiza a análise espectral do sinal que, depois, transmuda-se para um sinal elétrico no nervo auditivo (transdução neural).

A etapa seguinte consiste na codificação dos sinais em expressão lingüística, ao longo do nervo auditivo, por meio de elementos como o vocabulário e a gramática. Após a codificação, passa-se para o processo de reconhecimento e compreensão da mensagem pelo ouvinte.[6] Esse mecanismo de produção, reconhecimento e compreensão da fala é melhor demonstrado pelo fluxograma a seguir:

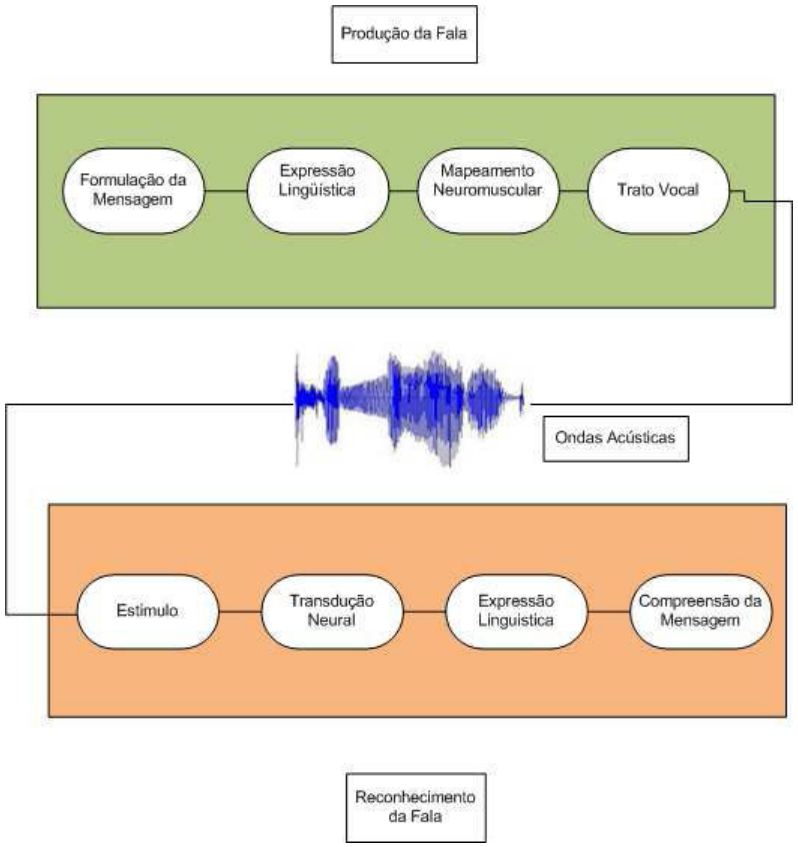


Figura 2.1 – Processo de Produção e Reconhecimento da fala humana.[7]

[6] Produção de Fala Humana. In: *DEETC – Departamento de Engenharia e Eletrônica e Telecomunicações e de Computadores*. Disponível em: <http://www.deetc.isel.ipl.pt/comunicacoesep/disciplinas/pdf/sebenta/pdf/producao_2.pdf>. Acesso em 25 nov. 2011.

[7] SILVA, Patrick. *Sistemas de reconhecimento de voz para o português brasileiro utilizando os Corpora Spoltech e OGI-22*. Trabalho de conclusão de curso, Universidade Federal do Pará, Instituto de Tecnologia, 2008.

2.2 ELEMENTOS FÍSICOS DO SOM

Os elementos constitutivos do som são altura, intensidade e timbre. São eles que conferem diversificação nas produções sonoras na medida das várias formas de combinação.

Altura está associada à qualidade do som. Dependendo da qualidade, o som pode ter a seguinte classificação: grave, apresentando baixa freqüência; e agudo, alta freqüência. O som de freqüência inferior a 16 Hz é denominado “infrassom”, e, o de freqüência superior a 17.000 Hz, por sua vez, “ultrassom”.[8] As freqüências audíveis pelo ouvido humano situam-se entre 20 e 4.000 Hz.

Intensidade é propriedade expressiva da força e do alcance do som, no sentido de possibilitar a sua percepção em maior ou menor distância da fonte sonora. Quanto à intensidade, o som pode ser classificado em forte ou fraco. No mais, a intensidade do som:

- É proporcional ao quadrado da amplitude da onda sonora;
- É mais intensa quanto maior for à superfície de vibração da fonte sonora;
- Aumenta com a densidade do meio em que ele se propaga;
- Diminui com o quadrado da distância entre o observador e fonte sonora, quando o som se propaga em meio homogêneo e infinito;
- Depende da proximidade de ressonadores, pois eles reforçam a intensidade do som;
- É alterada pelos ventos. Estes interferem na intensidade do som quando a distância entre a fonte e o observador é maior do que 6m.[9]

Por fim, o timbre é propriedade que difere sons que se encontram em altura e intensidade iguais, mas que são oriundos de fontes sonoras diversas, a partir de um conjunto de sons secundários (sons harmônicos) associados ao som principal.[10]

---

[8] Produção de Fala Humana. In: *DEETC – Departamento de Engenharia e Eletrônica e Telecomunicações e de Computadores*. Disponível em: <http://www.deetc.isel.ipl.pt/comunicacoesep/disciplinas/pdf/sebenta/pdf/producao_2.pdf>. Acesso em 25 nov. 2011.

[9] LOUZADA, Jailton Alkimin. *Reconhecimento automático de fala por computador*. Trabalho de conclusão de curso, Pontifícia Universidade Católica de Goiás, Ciência da Computação, 2010. p. 7-8.

[10] Ibid. p. 8.

2.2.1 REPRESENTAÇÃO DIGITAL DO SOM

Para que um computador possa processar o som, é necessária primeiramente a captação das ondas sonoras por meio de um transdutor – um microfone ou um telefone, filtragem do sinal e a sua conversão analógico-digital, conforme mostra a figura 2.2[11].

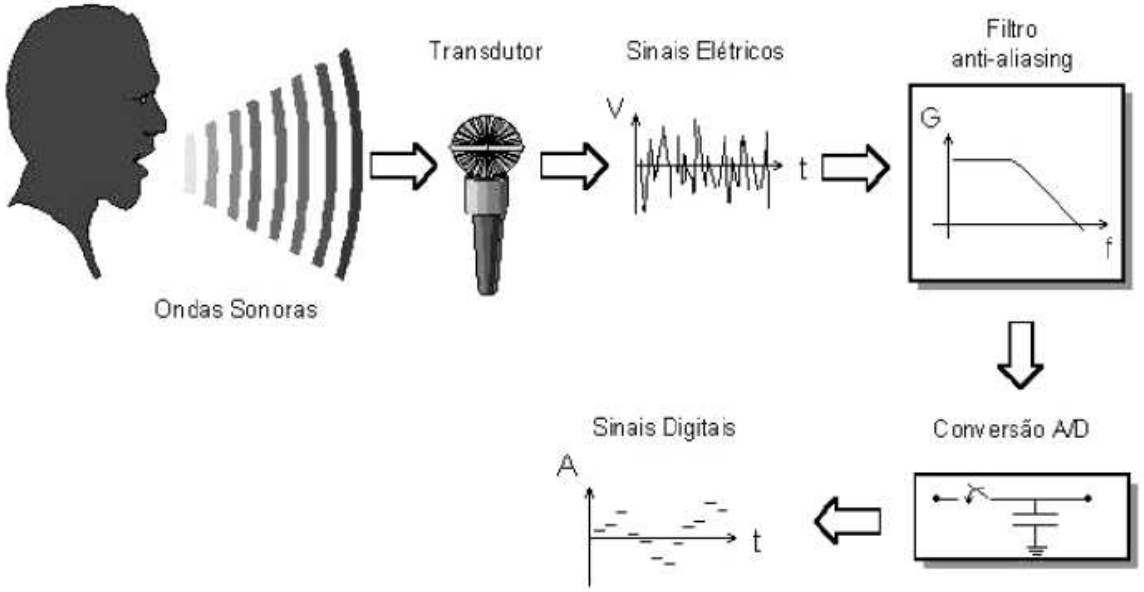


Figura 2.2: Processo de aquisição do sinal de fala

Enquanto o som natural é uma onda contínua, a sua representação digital é feita na forma de valores discretos através da obtenção da amplitude de sua onda com uma freqüência pré-estabelecida, processo conhecido como amostragem (ou *sampling*).

Por conta na natureza discreta da representação digital, é esperada a perda de informações durante a conversão do som para o formato digital. A figura 2.3 mostra a diferença entre a onda natural e a digital.

---

[11] DA SILVA, Anderson Gomes. *Reconhecimento de voz para palavras isoladas*. Recife, PE: [s.n.], 2009. p. 7.

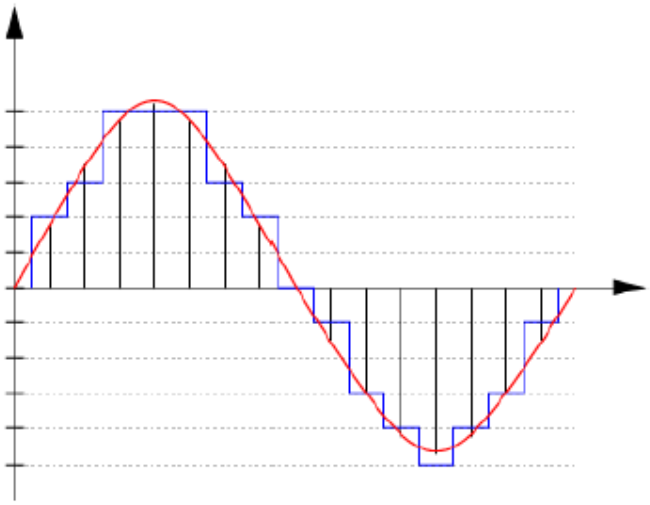


Figura 2.3: Representação analógica (em vermelho) e representação digital (em azul) do áudio[12].

Porém, quanto maior for a resolução, ou tamanho em bits do valor de cada amostra, menor será essa perda. Um inteiro de 8 *bits* pode representar 256 níveis da amplitude da onda enquanto um inteiro de 16 *bits* é capaz de representar 65.536 níveis diferentes.

Segundo o teorema de Nyquist, para se representar um sinal contínuo em formato digital, a freqüência de amostragem do mesmo deve ser maior ou igual ao dobro da freqüência que compõe a onda. Sabe-se que a fala humana ocorre dentro da faixa de freqüência de 500hz a 4.000hz, portanto, a taxa de amostragem de 8khz é a amostragem mínima para se registrar uma conversa no formato digital. Em telecomunicações, o áudio é transmitido com uma taxa de amostragem de 8khz e resolução de 8 *bits*[13].

---

[12] Pedrosa, Diogo Pinheiro Fernandes, *Conceitos básicos de áudio digital*, Universidade Federal do Rio Grande do Norte, pág 5.
[13] Ibid.

O formato mais comum de representação digital do áudio é o *Pulse Code Modulation*, ou PCM, utilizado em telecomunicações e CDs de música em que não há compressão dos dados.

2.3 FISIOLOGIA DO APARELHO FONADOR HUMANO

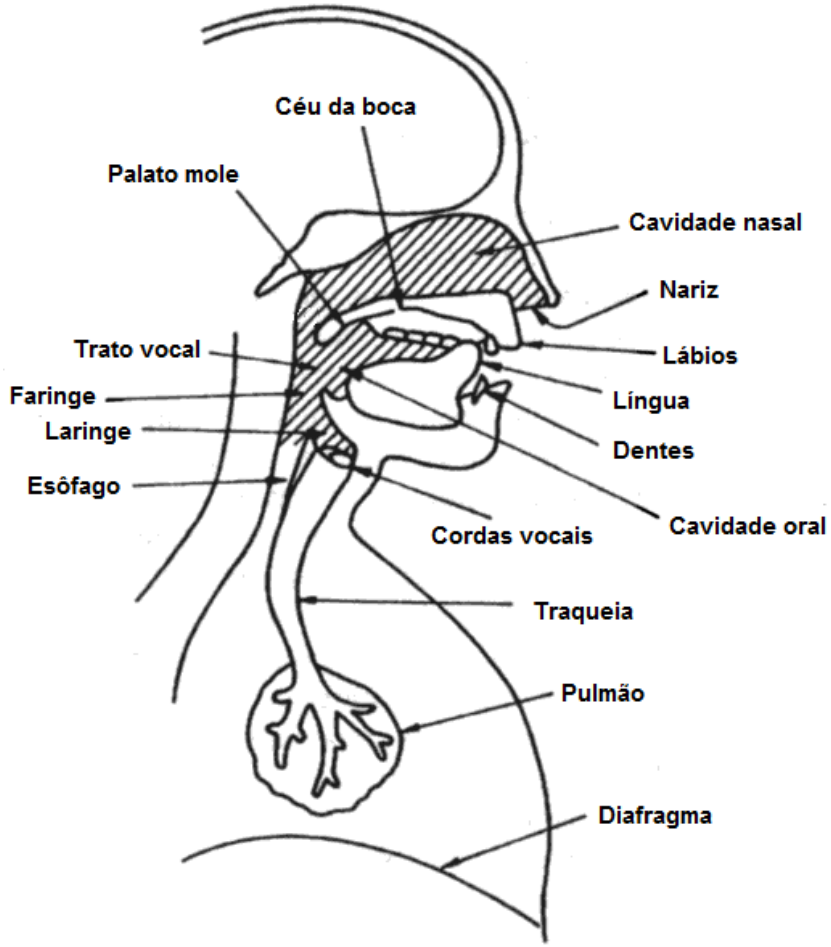


Figura 2.2 – Aparelho fonador humano.[14]

O aparelho fonador é um complexo sistema de produção de voz e abrange várias estruturas do corpo humano. Constituem o aparelho fonador: aparelho

---

[14] Biometria: impressão vocal. In: *Grupo de Teleinformática e Automação da Universidade Federal do Rio de Janeiro – GTA/UFRJ*. Disponível em < http://www.gta.ufrj.br/grad/09_1/versao-final/impvocal/propdosinal.html>. Acesso em 25 nov. 2011.

broncopulmonar; laringe; cavidades de ressonância; articuladores de voz; audição; e sistema nervoso.[15]

É da conjugação de todas essas estruturas que se obtém uma boa emissão de voz. Segue breve descrição de cada uma delas e as respectivas funções e participações na produção de voz:

> APARELHO BRONCOPULMONAR: Os pulmões são os propulsores da voz. Eles são os fornecedores do sopro aéreo, combustível essencial à vibração das pregas vocais.
>
> LARINGE: Na laringe situam-se as pregas vocais. Elas adotam posição ligeiramente aproximada uma da outra, realizando movimentos sucessivos em sentido látero-medial. É este movimento que produz as ondas sonoras. O movimento lateral das pregas vocais ocorre passivamente, devido à pressão do ar expelido. O movimento medial é obtido por contrações musculares, que ajustam tensão, forma e massa das pregas vocais.
>
> CAVIDADES DE RESSONÂNCIA: As cavidades de ressonância são compostas pelas estruturas do trato aerodigestivo localizadas acima da glote. Desse modo, compõem as cavidades de ressonância o ventrículo de Morgagni e demais componentes supraglóticos laríngeos, a faringe, as cavidades bucal e nasal e os seios paranasais. A faringe, principal componente ressoador, realiza movimentos musculares complexos, adaptando-se a todos os fonemas. Ela é capaz de ampla variação tonal, instintiva ou voluntária. As cavidades de ressonância conferem timbre à voz, a partir do movimento vibratório de seus constituintes.
>
> ARTICULADORES DA VOZ: Este grupo é compreendido pelo palato, língua, mandíbula e pelos lábios. O palato mole tem a função de ocluir as fossas nasais na emissão de fonemas vocálicos puros, de modo a diminuir ou eliminar a participação nasal nestes fonemas. A pressão do palato mole sobre a parede faríngea é variável, de acordo com a vogal pronunciada, sendo máxima em "i" e mínima em "a". A língua, formada por extensa rede muscular, executa os mais variados movimentos, tendo grande participação na articulação de fonemas, sejam eles vocálicos ou consonantais. Os movimentos labiais influenciam amplamente na produção do som articulado, devido aos seus diferentes graus de abertura disponíveis. Os lábios participam na emissão de fonemas consonantais bilabiais e labiodentais e dão timbre a todos os fonemas vocálicos. Os articuladores da voz alteram o fluxo de ar, convertendo o som em fala.
>
> AUDIÇÃO: A audição é fundamental para o bom controle da voz e para o desenvolvimento da fala. Através do feedback auditivo regulamos constantemente a intensidade do som que emitimos, bem como ajustamos a correta articulação de um fonema. Pela audição decodificamos e interpretamos o que nos é dito. Isto permite posterior uso de um fonema memorizado, quando em situação semelhante à vivida anteriormente. Entende-se, portanto, o quão essencial é a função auditiva para a fala.

[15] MOUSSALLE, Sérgio (Org.); et. al. *Guia prático de otorrinolaringologia*: anatomia, fisiologia e semiologia. Porto Alegre: EDIPUCRS, 1997. p. 112.

> SISTEMA NERVOSO: A via nervosa é composta por seis pares cranianos, pelo cerebelo e pelo sistema extrapiramidal. Os impulsos nervosos para o movimento laríngeo partem do córtex cerebral, seguindo pelas vias córtico-talâmicas ao tálamo. Neste ponto, as informações são processadas e coordenadas, antes do estímulo partir em direção aos núcleos bulbares dos pares cranianos. Os pares cranianos (V, VII, IX, X, XI e XII), com núcleos motores no bulbo, inervam todos os músculos que participam da fonação. Estes núcleos recebem impulsos corticais e mantêm relação com o cerebelo e o sistema motor extrapiramidal. O cerebelo atua na coordenação e o sistema extrapiramidal atua sobre o ritmo e o automatismo da fala.[16]

2.4 CLASSIFICAÇÃO DOS CONTÓIDES E VOCÓIDES

Como anteriormente explanado, são os articuladores de voz que controlam o fluxo de ar oriundos do aparelho broncopulmonar. O modo como a corrente de ar passa pelas pregas vocais determina a natureza do som que, nesse sentido, pode ser classificada em: livre ou comprimida/detida. Na primeira modalidade, o ar passa pelos articuladores de voz e sai livremente pela boca, sobre a língua, produzindo sons vocálicos, denominados vocóides. A posição do lábio e da língua e, o movimento do maxilar inferior são fatores determinantes da qualidade dos sons vocálicos, porque controlam a saída do fluxo de ar pela cavidade bucal. Na segunda modalidade, como o próprio nome sugere, o fluxo de ar não passa livremente, sofrendo uma pausa momentânea imposta pelos articuladores de voz conjugados com os movimentos da língua, produzindo, então, sons consonânticos, denominados contóides.

2.4.1 Classificação dos contóides

Os contóides, também chamados de ruídos, caracterizam a "fonética das consoantes"[17], e "classificam-se pelo **modo** e **ponto** de articulação, pela **sonoridade** e pelas **articulações secundárias**".[18]

---

[16] MOUSSALLE, Sérgio (Org.); et. al. *Guia prático de otorrinolaringologia*: anatomia, fisiologia e semiologia. Porto Alegre: EDIPUCRS, 1997. p. 113-114.
[17] VEGINI, Valdir. *Linguística aplicada à estrutura da língua materna*: fonética e fonologia (Módulo 1). Fundação Universidade Federal de Rondônia. p. 32.

2.4.1.1 Modo de articulação

a) **As Oclusivas** – São os contoides (ruídos) que resultam de uma oclusão (impedimento) momentâneo da passagem de ar, seguida de uma abertura brusca ou explosão. Em **português**, ela é realizada nos seguintes pontos:
Oclusão bilabial – Um lábio contra o outro: [p] e [b] (***p***á e ***b***oi).
Oclusão ápico-dental – A ponta da língua contra os dentes ou gengivas: [t] e [d] (***t***eu e ***d***eu).
Oclusão dorso-palatal – O dorso da língua contra o palato duro: [k] e [g] (***q***uilo) e ***g***uia).
Oclusão dorso-velar – O dorso da língua contra o palato mole ou véu palatino: [k] e [g] (***c***ulpa e ***g***ula). **Oclusão glotal** – As cordas ou pregas vocais se tocam, fechando momentaneamente a passagem do ar: [ʔ] **Obs**. Como em "kanãy´wet" [kaˈnãrˈʔwet] em tupy ramarama.

b) **As Fricativas** – São as caracterizadas por um estreitamento da passagem do ar, que produz um ruído de fricção ao passar entre dois articuladores.
Fricativa labiodental – [f] e [v] [ˈfaka] - [ˈvaca] (***f***aca e ***v***aca)
Fricativas alveolares – [s] e [z] [ˈsɛla] - [ˈzɛla] (***s***ela e ***z***ela)

[18] Ibid.

Fricativas álveo-palatais – [ʃ] e [ʒ] [ˈʃato - [ˈʒato] (***ch***ato e ***j***ato)
Fricativa velar – [x] [ˈkaxo] (ca***rr***o) [
Fricativa glotal – [h] [ˈkaho] (ca***rr***o; [ˈhəv] ***h***ave (ter em inglês), ***h***abu [haˈbu] (homem em karajá)

d) **<u>As laterais</u>** – As consoantes laterais são produzidas por um contato da língua com o centro do canal bucal, deixando sair o ar pelos lados.
Lateral ápico-dental – [l] - [ˈleʲte] (***l***eite)
Lateral palatal – [ʎ] [ˈpaʎa] (pa***lh***a)
Lateral retroflexo – [ɭ] [ˈkaɭdo] (ca***l***do)

c) **<u>As Nasais</u>** – É um tipo de oclusiva pronunciada com o palato mole ou véu palatino em posição abaixada, permitindo a passagem do ar pelo nariz. Embora haja uma oclusão na boca, o ar não sai como uma explosão, que é o caso das oclusivas, porque a cavidade nasal fica permanentemente aberta. Um nasal, por conseguinte, é uma oclusão em relação à articulação bucal, mas livre, se for levando em conta a passagem do ar pelo nariz. O contoide [m] seria um [b] não fosse o abaixamento da úvula.
Nasal bilabial – [m] - [ˈmala] (***m***ala)
Nasal ápico-dental [n] - [ˈnada] (***n***ada)
Nasal palatal – [ɲ] - [ˈmaɲa] (ma***nh***a)
Nasal velar – [ŋ] - [mãŋga] (ma***n***ga) (manga)
<u>Obs.</u> As nasais são normalmente sonoras, mas podem, em algumas línguas, serem surdas.

f) **As Africadas** – Combinação entre uma articulação oclusiva e outra fricativa:

[tʃ] e – [dʒ] - ['tʃia] ['dʒia] (*t*ia e *d*ia)

e) **As Vibrantes** – As consoantes vibrantes são articuladas com o ápice (a ponta) da língua ou a úvula ocluindo (tapando) rápida e repetidamente um ponto da boca.

Vibrante anterior ou apical ou simples (*flap*) – ['kaɾo] (ca**r**o)

Vibrante anterior ou apical ou rolada ou múltipla (*trill*) – [r] - ['karo] (ca**rr**o)

Vibrante uvular – [ʀ] - ['kaʀo] (ca*rr*o)

g) **As Aproximantes** – Impedimento da passagem da corrente de ar em menor escala que no caso das fricativas fica numa posição intermediária entre uma fricativa e uma vogal.

[j] e [ɯ] – ['paj] e ['paɯ] (pa*i* e pa*u*)

Figura 2.1 – Modo de articulação.19

2.4.1.2 Ponto de articulação

O ponto de articulação nada mais é que o articulador superior. Junto com o inferior, configuram os dois articuladores necessários para a pronúncia de um contóide. Seguem as formações possíveis envolvendo o ponto de articulação:

[19] VEGINI, Valdir. *Linguística aplicada à estrutura da língua materna*: fonética e fonologia (Módulo 1). Fundação Universidade Federal de Rondônia. p. 32-34.

a) **Bilabial** – O lábio inferior articula (toca no) com o lábio superior.
b) **Lábio-dental** – O lábio inferior articula (toca no) com os dentes superiores.
c) **Dental** – A ponta da língua articula (toca no) com os dentes.
d) **Alveolar** – A ponta da língua articula (toca na) com a arcada alveolar.
e) **Palatal** – A lâmina da língua articula (toca no) com o palato duro.
f) **Velar** – O dorso da língua articula (toca no) com o palato mole ou véu palatino.
g) **Uvular** – O dorso da língua articula (toca na) com a úvula.
h) **Faríngeo** – A raiz da língua articula (toca na) com a parede posterior da faringe.
i) **Glotal** – As duas cordas ou pregas vocais articulam-se (tocam-se) entre si.

Tabela 2.2 – Ponto de articulação.[20]

2.4.1.3 Sonoridade

Via de regra, "todos os contóides podem ser sonoros (vozeados) ou surdos (desvozeados), conforme haja vibração ou não das cordas ou pregas vocais".[21]

2.4.1.4 Articulações secundárias

Como anteriormente exposto, os sons secundários determinam o timbre e diferem sons que se encontram em altura e intensidade iguais, porém, oriundos de fontes sonoras diversas, quando associados ao som principal. Seguem breves descrições acerca das variações das articulações secundárias:

---

[20] VEGINI, Valdir. *Linguística aplicada à estrutura da língua materna*: fonética e fonologia (Módulo 1). Fundação Universidade Federal de Rondônia. p. 34-35.
[21] VEGINI, Valdir. *Linguística aplicada à estrutura da língua materna*: fonética e fonologia (Módulo 1). Fundação Universidade Federal de Rondônia. p. 35.

a) **Labialização** – Um contoide pode se tornar arredondado, quando sua articulação sofre influência (antecipação da labialização) do vocoide que lhe segue. Exemplo: [kwatɾo] (quatro) → [ˈkʷatɾo] (quatro).
b) **Palatalização** – A língua pode-se elevar na direção do palato duro por influência (antecipação da palatalização) do vocoide seguinte. Exemplo: [kiˈabo] (quiabo) → [ˈkʲabo] (quiabo).
c) **Faringalização** – A língua pode ser retraída na direção da parede posterior da faringe, fenômeno articulatório que não ocorre no português. .
d) **Retroflexão** – Um som apical pode ser emitido com a ponta da língua retraída para cima e para trás. Exemplo: [verˈdade] (verdade) → [veɽˈdade] (verdade).

Tabela 2.3 – Articulações secundárias.[22]

2.5 FONES E FONEMAS

Os conceitos de fone e fonema são frequentemente confundidos entre si e, via de conseqüência, os campos da fonética e fonologia também.

Nesse sentido, cabe esclarecer, de início, que o fone é objeto de estudo da Fonética, e, o fonema, da Fonologia. Vê-se, portanto, que, apesar da confusão que existe entre elas, os focos de estudo dessas duas searas são diferentes. Isto, porque, "enquanto a Fonética estuda a natureza física da produção e da percepção dos sons da fala (...), a Fonologia preocupa-se com a maneira como eles se organizam dentro de uma língua, classificando-os em unidades capazes de distinguir significados".[23]

---

[22] Ibid. p. 35-36.

[23] Fonologia. In: *Wikipédia*: a enciclopédia livre. Disponível em < http://pt.wikipedia.org/wiki/Fonologia>. Acesso em 28 nov. 2011.

Assim, a “Fonologia (do Grego *phonos* = voz/som e *logos* = palavra/estudo) é o ramo da Linguística que estuda o sistema sonoro de um idioma, do ponto de vista de sua função no sistema de comunicação lingüística”,[24] mas “também estuda outros tópicos, como a estrutura silábica, o acento e a entonação”.[25] A Fonética, por sua vez, “preocupa-se com a parte significante do signo linguístico e não com o seu conteúdo”, segundo Francisco S. Borba.[26]

2.6 FONÉTICA DO PORTUGUÊS BRASILEIRO

O português praticado no Brasil apresenta 39 fones. Carlos Alberto Ynoguti sugere que, dentre eles, alguns sejam aglutinados por aproximação, por uma questão sistemática que acabaria por reduzir o número de subunidades fonéticas. Todavia, esse agrupamento deve ser elaborado com o devido cuidado, considerando as características dos fones, a fim de se evitar inconsistências.[27] As fusões submetidas a testes foram as seguintes:

- [i] e [j];
- [u] e [w];
- [a] e [ɑ];
- [e] e [ə] (esta fusão foi efetivada já na transcrição fonética original, sem realização de teste).[28]

Na lista a seguir, seguem os fones constantes do português brasileiro:

---

[24] Ibid.
[25] Ibid.
[26] BORBA, Francisco S.. *Introdução aos estudos linguísticos*. São Paulo: Companhia Editora Nacional, 1975. p. 251.
[27] YNOGUTI, Carlos Alberto. *Reconhecimento de fala contínua usando modelos ocultos de Markov*. Campinas, SP: [s.n.], 1999. p. 83.
[28] YNOGUTI, Carlos Alberto. *Reconhecimento de fala contínua usando modelos ocultos de Markov*. Campinas, SP: [s.n.], 1999. p. 83-84.

Vogais orais		Consoantes orais	
[i]	l*i*vro	[p]	*p*á
[e]	P*e*dro	[b]	*b*ata
[ɛ]	t*e*rra	[t]	*t*arde
[a]	p*a*to	[d]	*d*ado
[ɑ]	m*a*no	[k]	*c*ão
[ɔ]	g*o*la	[g]	*g*ato
[o]	p*o*ço	[s]	*s*ábado
[u]	p*u*lar	[z]	ca*s*a
[ə]	s*e*car	[ʃ]	*ch*ão
		[ʒ]	*j*ardim
		[f]	*f*ado
Vogais nasais		[v]	*v*aca
		[l]	*l*ado
[ĩ]	p*in*to	[ʎ]	fi*lh*o
[ẽ]	d*en*te	[r]	po*r*ta
[ɑ̃]	c*an*to	[r̄]	ca*rr*o
[õ]	p*on*te	[R]	po*r*ta (velar)
[ũ]	f*un*do	[R̄]	ca*rr*o (velar)
[õ]	p*on*te	[R]	po*r*ta (velar)
[ũ]	f*un*do	[R̄]	ca*rr*o (velar)
		[tʃ]	*t*ia
		[dʒ]	*D*ia
Semivogais		**Consoantes nasais**	
[j]	pa*i*	[m]	*m*ãe
[w]	pa*u*	[n]	*n*ada
		[ʎ]	pi*nh*o

Tabela 2.4: Lista dos fones presentes no português falado no Brasil.[29]

A metodologia adotada para verificar a manutenção ou não de uma fusão foi a seguinte:

- Inicialmente foram gerados e treinados os modelos HMM de todos os 39 fones listados [...].
- Com esses modelos calculou-se a probabilidade média dos modelos HMM das locuções de treinamento gerarem as sequências de

[29] YNOGUTI, Carlos Alberto. *Reconhecimento de fala contínua usando modelos ocultos de Markov*. Campinas, SP: [s.n.], 1999. p. 84.

observação correspondentes. Esta probabilidade é tomada então como referência.

- Para cada uma das fusões propostas acima, foram criados e testados os modelos HMM correspondentes e calculada novamente a probabilidade de os modelos gerarem as sequências de observação. Se esta probabilidade fosse maior que a de referência, a fusão era adotada.[30]

Seguem os resultados dos referidos testes na tabela a seguir:

Testes	log (P($O\|\lambda$))
todos os fones independentes de contexto (referência)	-1693.13
a) juntando [i] e [j]	-1693.65
b) juntando [u] e [w]	-1692.96
c) juntando [$\overline{\text{R}}$] e [R]	-1693.21
d) juntando [a] e [α]	-1693.05

Tabela 2.5: Resultados dos testes realizados para fusão de fones independentes de contexto.[31]

A partir desses testes, com exceção da fusão entre [i] e [j], todas as outras foram consolidadas, culminando nas unidades listadas na tabela a seguir:

[30] Ibid. p. 85.
[31] YNOGUTI, Carlos Alberto. *Reconhecimento de fala contínua usando modelos ocultos de Markov*. Campinas, SP: [s.n.], 1999. p. 85.

Fone	Símbolo utilizado	Exemplo	Frequência Relativa (%)		Número de ocorrências
			Alcaim et. al.	Observada	
a	a	***a*** çafrão	12,94	13,91	6031
e	e	***e*** levador	4,82	2,15	933
ɛ	E	p ***e*** le	1,91	6,35	2785
i	i	s ***i*** no	8,57	1,90	821
j	y	fu ***i***	3,13	0,95	410
o	o	b ***o*** lo	2,71	4,14	1798
ɔ	O	b ***o*** la	1,00	6,23	2691
u	u	l ***u*** a	8,69	2,57	1124
α̃	an	maç ***ã***	2,12	4,04	1773
ẽ	en	s ***en*** ta	2,30	1,16	501
ĩ	in	p ***in*** to	3,23	0,69	296
õ	on	s ***om*** bra	0,75	8,41	3648
ũ	un	um	2,50	1,98	860
b	b	***b*** ela	1,09	1,18	511
d	d	***d*** ádiva	2,64	3,14	1346
dʒ	D	***d*** iferente	1,92	1,49	665
f	f	***f*** eira	1,46	1,44	625
g	g	***g*** orila	0,93	0,87	378
ʒ	j	***j*** iló	1,32	0,75	325
k	k	***c*** achoeira	4,19	3,63	1575
l	l	***l*** eão	1,72	1,91	830
ʎ	L	***lh*** ama	0,21	0,35	152
m	m	***m*** ontanha	4,12	3,77	1637
n	n	***n*** évoa	2,40	2,26	982
ɲ	N	i ***nh*** ame	0,68	0,42	185
p	p	***p*** oente	2,29	2,49	1081
r	r	ce ***r*** a	3,58	4,05	1759
r̄	rr	ce ***rr*** ado	2,06	0,89	363
R	R	ca ***r*** ta	-	1,32	598
s	s	***s*** apo	4,18	6,52	2832
t	t	***t*** empes ***t*** ade	3,94	4,02	1737
tʃ	T	***t*** igela	1,44	1,20	531
v	v	***v*** erão	1,23	1,51	656
ʃ	x	***ch*** ave	2,12	0,32	132
z	z	***z*** abumba	1,81	1,96	859

Figura 2.6: subunidades acústicas utilizadas na transcrição fonética das locuções, com exemplos e frequências relativas de ocorrência, e aquelas encontradas na transcrição fonética da base de dados coletada. Também são listados os números de ocorrências observados para cada subunidade.[32]

[32] YNOGUTI, Carlos Alberto. *Reconhecimento de fala contínua usando modelos ocultos de Markov*. Campinas, SP: [s.n.], 1999. p. 30.

3 DESCRIÇÃO DO PROBLEMA DO RAF

3.1 O QUE É O RAF

Um sistema moderno de RAF são, basicamente, arquiteturas de software capazes de gerar uma seqüência de palavras hipotéticas a partir do processamento de um sinal acústico através de algoritmos populares baseados em sua maioria em métodos estatísticos[33].

As palavras “hipotética” e “estatísticos” indicam que o resultado obtidos com tais sistemas não garantem a identificação exata das palavras contidas em tais sinais.

Este é um fato perfeitamente aceitável, uma vez que a comunicação oral entre os seres humanos também é passível de erros, seja por falha de pronúncia do emissor, seja por dificuldades encontradas pelo receptor da mensagem, como ruído excessivo, má qualidade do sinal acústico, etc..

Portanto, o produto de um sistema de RAF será sempre uma palavra ou seqüência de palavras associados a um indicador de confiança daquele resultado.

---

[33] Varile, Giovanni Battista; Zampoli, Antonio. *Survey of the State of The Art in Human Language Technology*. Cambridge University Press. p. 21.

3.2 POR QUE O RAF É DIFÍCIL

A compreensão da comunicação oral entre humanos é uma atividade complexa e vai muito além da simples associação entre sinais acústicos e palavras.

Forsberg (2003) lista uma série de fatores que interferem na tarefa de RAF, algumas das quais serão enumeradas a seguir.

3.2.1 - A compreensão humana da fala comparada com o RAF

Os humanos utilizam mais do que somente os ouvidos para compreender a fala, eles fazem uso do conhecimento que têm do locutor e do assunto. A estrutura gramatical das línguas demanda uma ordenação das palavras numa sentença e permitem que o ouvinte possa prever as palavras que serão ditas.

É possível modelar a estrutura gramatical de uma língua e, à partir daí, utilizar modelos estatísticos para melhorar a predição das palavras. Porém, o desafio é modelar o conhecimento sobre as inúmeras matérias que podem ser abordadas em uma conversa e até que ponto elas são essenciais para viabilizar o RAF em todos os níveis.

3.2.2 – Linguagem corporal

As pessoas também usam o corpo para se comunicar através do movimento das mãos e dos olhos, da postura, etc..

3.2.3 – Ruído

O ruído – informação indesejada no sinal sonoro – produzido por outro elemento presente no mesmo ambiente do locutor precisa ser identificado e filtrado pelos

sistemas de RAF. Por exemplo, o ruído produzido por carros, aparelhos sonoros e o próprio eco.

3.2.4 – Diferença entre linguagem falada e linguagem escrita

Enquanto a comunicação escrita é, geralmente, unidirecional, a linguagem falada é orientada ao diálogo.

Em um diálogo, há a resposta ao sinal recebido, negociação sobre o significado das palavras, adaptação mútua entre os interlocutores, etc..

Outro ponto importante são as disfluências presentes na fala: hesitações, repetições, mudança de assunto, erros de pronúncia, etc..

Ainda, a linguagem falada é gramaticalmente diferente da linguagem escrita.

3.2.5 – Fala contínua

Na fala contínua as palavras são pronunciadas foneticamente emendadas umas nas outras, o que pode gerar ambigüidade dentro de frases devido à dificuldade de identificação do limite entre palavras dentro das mesmas. Um exemplo bem conhecido deste efeito pode ser percebido no Hino Nacional Brasileiro, onde tem-se a frase “De um povo **heróico o brado** retumbante”, o trecho grifado pode ser ouvido como “**herói cobrado**”.

3.2.6 – Variabilidade do canal

Alterações nos níveis de ruído no decorrer do tempo, o equipamento que capta o áudio e qualquer outro fator que altere o conteúdo da onda acústica entre o seu emissor até a sua representação do na forma digital.

3.2.7 – Dialetos regionais e sociais

Dialetos são as variações de uma mesma língua relacionadas a grupos. Há os dialetos regionais – ligados a uma determinada área geográfica – e os dialetos sociais – ligados a um determinado grupo social.

Em ambos os casos, notam-se variações de pronúncia, vocabulário e gramática. Por exemplo, a pronúncia do s em biscoito na região metropolitana da cidade de São Paulo é diferente daquela realizada na região metropolitana do Rio de Janeiro.

3.2.8 – Anatomia do trato vocal

A anatomia do trato vocal altera a forma da realização da fala. As suas características variam em função de características genéticas, da idade e do sexo, portanto, todos esses fatores em conjunto influenciam os sinais vocais.

3.3 CARACTERIZANDO AS CAPACIDADES DE UM SISTEMA DE RAF

De acordo com Varile (1997), um sistema de RAF pode ser classificado de acordo com as suas capacidades conforme demonstrado na figura 3.1.

3.3.1 – Modo de pronúncia

De acordo com o problema descrito no item 3.2.5, quanto à pronúncia, um sistema de RAF pode ser capaz de reconhecer a fala de palavras isoladas, isto é, quando há pausa entre elas, e capaz de reconhecer fala contínua.

Sistemas de RAF que reconhecem palavras isoladas podem ser aplicados a

equipamentos que obedecem a comandos voz, como a discagem por voz em celulares ou a automação de residências.

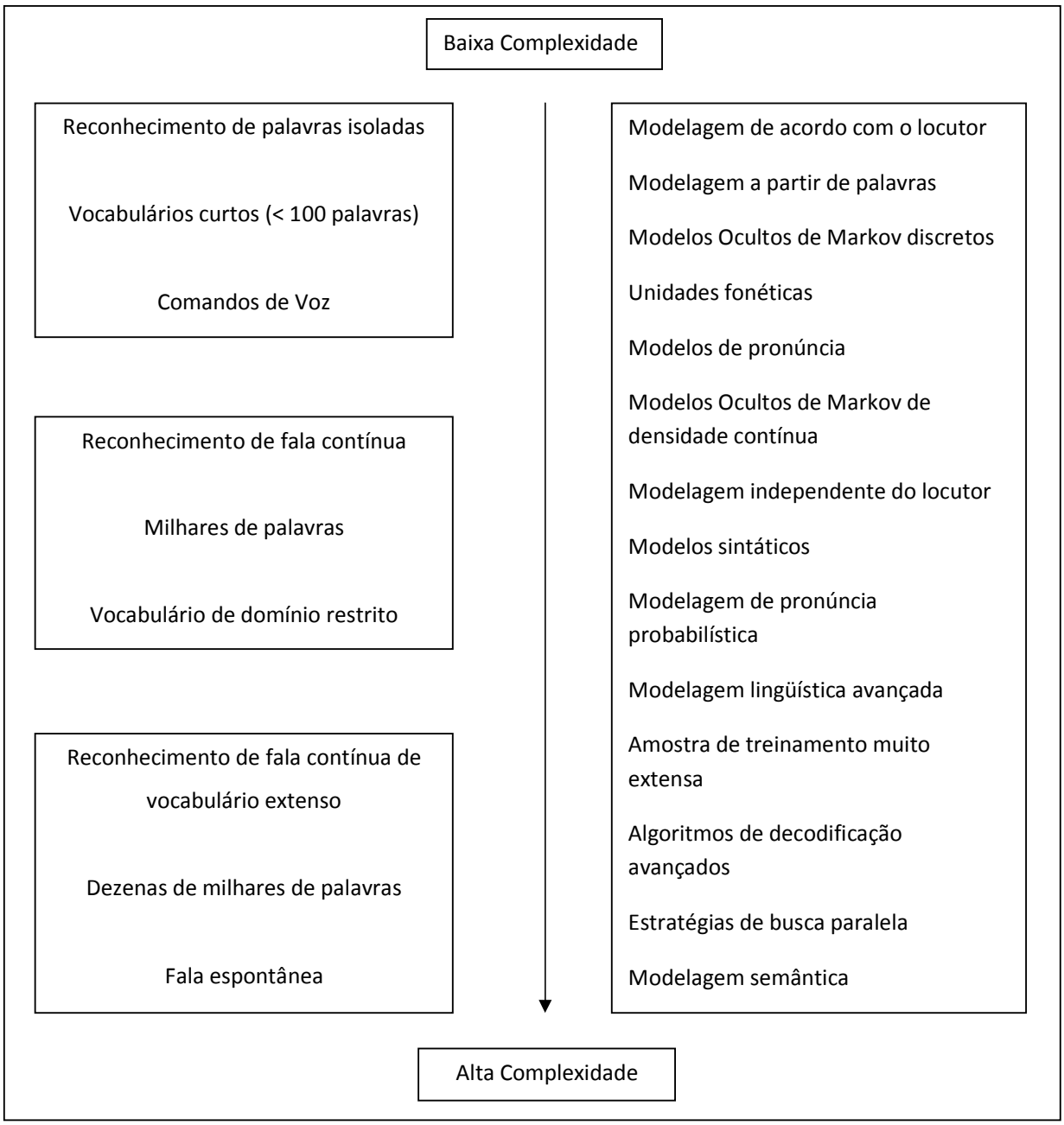


Figura 3.1: RAF, das tarefas mais simples às mais avançadas[34].

[34] Varile, Giovanni Battista; Zampoli, Antonio. *Survey of the State of The Art in Human Language Technology*. Cambridge University Press. p. 9

3.3.2 – Estilo de pronúncia

Um sistema de RAF pode ser mais eficiente em reconhecer fala produzida a partir da leitura de um texto do que a fala espontânea, já que esta contém disfluências que dificultam o processamento da fala, conforme descrito no item 3.2.4.

3.3.3 – Treinamento

Para lidar com os problemas descritos nos itens 3.2.6 a 3.2.8, alguns sistemas de RAF são dependentes do locutor, ou seja, precisam ser treinados com exemplos da voz dos locutores que o utilizarão para aprenderem sobre o seu timbre, estilo de pronúncia, etc.. e assim aumentar significativamente a precisão do sistema.

Tarefa bem mais complexa é realizada pelos sistemas de RAF independentes de locutor, que identificam a fala sem treinamento prévio de um locutor específico.

3.3.4 – Vocabulário

Quanto mais extenso e mais palavras que soem semelhantes contiver o vocabulário, mais difícil se torna a tarefa de RAF.

3.3.5 – Modelo de Linguagem

Para processar a fala de uma seqüência de palavras, os sistemas de RAF se valem de modelos de linguagem para prever e restringir as palavras subseqüentes.

Há dois tipos comuns de modelos de linguagem: os modelos de estados finitos, quando as palavras que podem seguir outra são definidas de modo explícito e os

modelos sensíveis ao contexto, que analisam o contexto das palavras e assim se aproximam mais da linguagem natural.

3.3.6 – Perplexidade

Perplexidade é uma forma de medir a quantidade de palavras que podem seguir outra, que será influenciada diretamente pelo tamanho do vocabulário e pelo domínio específico ao qual o sistema de RAF possa estar direcionado, como meio jurídico, médico, etc..,

O modelo de linguagem do sistema de RAF irá tentar restringir a quantidade de palavras que possam suceder a última para assim melhorar o despenho do sistema e a dificuldade dessa tarefa será proporcional à perplexidade do sistema[35].

3.3.7 – Qualidade do sinal e nível de ruído

A taxa de amostragem, a qualidade da captação e a relação ruído / sinal do áudio podem ser um fatores limitantes para os sistemas de RAF.

[35] LOUZADA, Jailton Alkimin. *Reconhecimento automático de fala por computador*. Trabalho de conclusão de curso, Pontifícia Universidade Católica de Goiás, Ciência da Computação, 2010.

4 ESTRUTURA DE UM SISTEMA MODERNO DE RAF

Apesar de existirem diversas técnicas diferentes para se realizar o RAF, algumas etapas são comuns e essenciais para o funcionamento destes, conforme demonstrado na figura abaixo:

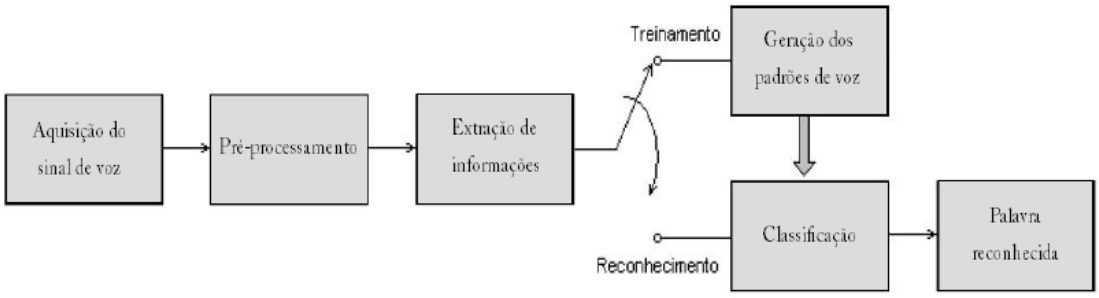


Figura 4.1: Diagrama de blocos de um sistema de reconhecimento de voz.[36]

As primeiras tentativas de se desenvolver sistemas de RAF consistiam em se comparar os sinais vocais a modelos de referência. O sistema mantinha um modelo de cada palavra que fosse capaz de identificar de forma isolada e foi estendido a sistemas de RAF capazes de processar a fala contínua, porém com um número limitado de palavras.

Porém, este tipo de arquitetura se mostra limitada, já que é incapaz de processar fala independente do locutor com vocabulário mais amplo. Por isso, os sistemas passaram a trabalhar com modelos de unidades fonéticas, que podem ser

[36] SILVA, Anderson Gomes da. *Reconhecimento de voz para palavras isoladas*. Trabalho de graduação, Universidade Federal de Pernambuco, Engenharia da Computação, 2009. p. 8.

concatenados aos modelos de palavras. Desta forma, é possível criar um dicionário fonético que torna bastante flexível a inclusão de novas palavras ao vocabulário do sistema.

A seguir, serão discutidos os detalhes da estrutura de um sistema moderno de RAF.

4.1 CAPTAÇÃO DO ÁUDIO

A captação de áudio, bem como a sua forma de representação digital pode ser encontrada no capítulo 2.2.1.

4.2 PRÉ-PROCESSAMENTO

O pré-processamento do áudio é etapa de preparação do áudio para a extração das informações relevantes para a seqüência do processo de reconhecimento.

Devido a fatores físicos dos elementos de captação do áudio, pode surgir um tipo de ruído no mesmo conhecido como nível DC, que causa o deslocamento da amplitude da onda. Para a sua correção, são utilizados algoritmos para a remoção do nível DC, forçando o valor médio da amplitude da onda a se fixar em zero.[37 e 38]

Outra etapa do pré-processamento é a normalização, que consiste em padronizar o volume do som, fazendo com que a amplitude das ondas esteja dentro de uma mesma faixa de valores.[39] Desta forma garante-se que a diferença de volume entre os áudios não influencie o processo de RAF.

---

[37] DC Offset: the case of the missing headroom. In: *Harmony Central*. Disponível em < http://www.harmonycentral.com/docs/DOC-1082>. Acesso em 14 dez. 2011.
[38] DC bias. In: *Wikipédia*: a enciclopédia livre. Disponível em < http://en.wikipedia.org/wiki/DC_bias>. Acesso em 14 dez. 2011.
[39] Audio normalization. In: *Wikipédia*: a enciclopédia livre. Disponível em < http://en.wikipedia.org/wiki/Audio_normalization>. Acesso em 14 dez. 2011.

Por último, é feita a remoção do silêncio do início e do fim da amostra de áudio de forma a isolar os trechos que contém fala.

4.3 EXTRAÇÃO DE CARACTERÍSTICAS

A extração de características é a etapa do sistema de RAF responsável por criar uma representação parametrizada dos sinais de voz contidos no áudio analisado na forma de vetores.

Essa parametrização visa minimizar a quantidade de informações contida no sinal digital do áudio para assim otimizar o processamento dos dados ali contidos. A parametrização visa também criar um padrão de representação para as ondas sonoras para facilitar a correlação entre o áudio analisado e o modelo já obtido.

A seguir serão listadas as técnicas mais comuns de extração de características do áudio.

4.3.1 – Análise do espectro de energia (FFT)

O espectro de energia de um áudio descreve a freqüência do sinal em função tempo. Essa função é obtida através da transformada discreta de Fourier.

4.3.2 – Análise preditiva linear (LPC)

A análise preditiva linear é um poderoso método tanto para estimar as características da fala quanto para se gerar um modelo de fala computacionalmente eficiente.

A idéia básica por trás deste método é a de que uma amostra específica de áudio num dado momento pode ser aproximada por combinação linear de amostras anteriores de áudio. Através da minimização da soma das diferenças dos quadrados

em um intervalo finito entre a amostra de áudio de fato e os valores preditos linearmente, um único conjunto de coeficientes pode ser determinado.

Estes coeficientes são então transformados em um conjunto mais robusto de parâmetros conhecidos como coeficientes cepstrais.

4.3.3 – Predição linear perceptual (PLP)

Esta técnica é similar à análise linear preditiva, porém, efetua modificações no espectro do áudio para simular algumas características psicofisiológicas da percepção da fala pelos humanos.

Mas, assim como outras técnicas de análise de espectro de intervalos curtos, este método é vulnerável às alterações espectrais em intervalos curtos causadas pela resposta à freqüência do canal de comunicação. Para contornar o problema, o filtro RASTA (Relative Spectral – espectro relativo) é empregado.

4.3.4 – Análise Cepstral da Escala Mel (MEL)

Técnica similar à PLP, com a diferença que a MEL realiza as modificações dos espectros de intervalo curto de acordo com a Escala Mel, fazendo com que as saídas das duas análises sejam diferentes.

4.4 TREINAMENTO DO SISTEMA

A fase de treinamento de um sistema de RAF consiste da montagem de classificadores probabilísticos gerados a partir dos áudios coletados.[40] Desta forma,

---

[40] SEWARD, Alexander. *Efficient methods for automatic speech recognition*. Dissertação de doutorado, Royal Institute of Technology, Stockholm, 2003. p. 5-6.

é possível criar modelagem de palavras com diferenciação de pronúncia bem como modelos de gramática.[41]

4.4.1 – Os Modelos Ocultos de Markov

Desde a metade da década de 70, os Modelos Ocultos de Markov (ou HHM – *Hidden Markov Models*, em inglês) são utilizados como classificadores probabilísticos nos sistemas de RAF e são considerados mais adequados para a modelagem de unidades acústico-fonéticas concatenadas, comum para o processamento de fala contínua. Os HHMs podem lidar com as variações da velocidade da fala, de pronúncia e na identificação de fones dependentes de contexto.[42]

4.4.1.1 – Definição

Um modelo de Markov, também chamado de cadeia de Markov, consiste de um conjunto finito de estados ligados entre si por transições associadas a um processo estocástico[43], formando máquinas de estados.

Espíndola (2009) diz:

> Talvez não seja possível observar diretamente a dinâmica estocástica que rege um dado processo do mundo real, mas muito provavelmente esse processo produz observáveis, também chamados "sinais", a partir dos quais o sistema pode ser modelado. Esses sinais podem ou não ser de fonte estacionária (sistema em equilíbrio), ser de natureza discreta ou contínua, tratar-se de sinais limpos ou ruidosos, dentre outras características imagináveis. [44]

---

[41] ELLIS, Dan. *ASR*: training and systems. Columbia University, Electrical Engineering, 2003. p. 20.
[42] JUANG, B. H.; RABINER, Lawrence R. *Automatic speech recognition*: a brief history of the technology development. p. 11.
[43] Um processo estocástico é uma família de variáveis aleatórias de um algum espaço de probabilidade dentro de um espaço de estados (SEWELL, Martin. *Stochastic processes*. 2006.).
[44] ESPINDOLA, Luciana da Silveira. *Um estudo sobre modelos ocultos de Markov* (HMM – Hidden Markov Model). Trabalho de pós-graduação, Pontifícia Universidade Católica do Rio Grande do Sul, Informática, 2009. p. 9.

Nestes casos, um modelo estocástico baseado em sinais pode ser utilizado para descrever tais processos, como os HHM. Portanto, os HHM são utilizados em modelos quando a evolução da cadeia de Markov está escondida do observador, ou seja, os observáveis não podem ser obtidos de forma direta. Considere como exemplo um torcedor de um determinado time de futebol que comparece aos seus jogos no estádio em função do resultado da última partida deste mesmo time. Geralmente, ele comparece ao estádio se o seu time preferido venceu a última partida. É possível inferir que há uma probabilidade maior deste time ter vencido seu último jogo se for observado que o torcedor compareceu ao estádio, mas não pode ser descartada a possibilidade de o time ter sido derrotado recentemente[45].

4.4.1.2 – Elementos de um HHM

Os elementos básicos de um HHM são[46]:

- N, o número de estados do modelo. Os estados individuais são rotulados como $S = \{S_1, S_2, S_3, ..., S_N\}$, e o estado em t como q_t.
- M, o número de símbolos de observação distintos por estado. Os símbolos individuais são denotados como $V = \{v_1, v_2, v_3, ..., v_M\}$.
- A distribuição de probabilidade de transição do estado $A = \{a_{ij}\}$, onde

$$a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i),\ 1 \leq i, j \leq N$$

- A distribuição de probabilidade de símbolos de observações no estado j, $B = \{b_j(k)\}$, onde

$$b_j(k) = P(O_t = v_k \mid q_t = S_j),\ 1 \leq j \leq N,\ 1 \leq k \leq M$$

---

[45] Ibid.
[46] SILVA, Anderson Gomes da. *Reconhecimento de voz para palavras isoladas*. Trabalho de graduação, Universidade Federal de Pernambuco, Engenharia da Computação, 2009. p. 21.

- A distribuição do estado inicial $\pi = \{ \pi_i \}$, onde

$$\pi_i = P(q_1 = S_i),\ 1 \leq i \leq N$$

Para um definição completa de um HHM, faz-se necessário especificar os parâmetros N e M, a seqüência de observações ($O = O_1, O_2, ..., O_T$), onde T é o número de observações na seqüência) e a especificação de três conjuntos de medidas de probabilidade A, B e π. Para indicar o conjunto de parâmetros completos do modelo, utiliza-se a seguinte notação compacta:

$$\lambda = (A, B, \pi)$$

4.4.1.3 - Topologias de HHM

Existem duas topologias principais de HHM[47].

A primeira é o modelo ergótico, na qual todos os estados podem ser alcançados a partir de qualquer outro estado, estando assim todos os estados conectados, conforme mostra a figura 4.2.

[47] Ibid., p. 22.

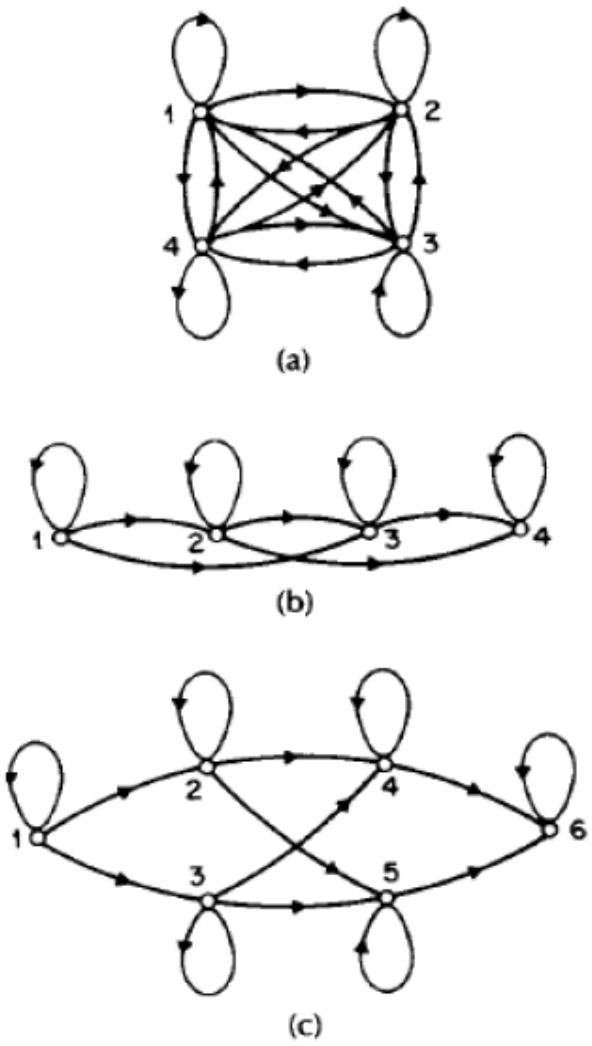


Figura 4.2: Ilustração de 3 topologias de HHM distintas. a) Modelo ergótico. b) Modelo esquerda-direita. c) Modelo esquerda-direita paralelo.[48]

A segunda topologia é conhecida como modelo esquerda-direita, que é assim conhecida graças à sua propriedade de que, à medida que o tempo aumenta, o índice do estado aumenta ou permanece o mesmo, conforme ilustrado na figura 4.2.

Esta topologia possui ainda a variante esquerda-direita paralelo, conforme ilustra a figura 4.2.

[48] SILVA, Anderson Gomes da. *Reconhecimento de voz para palavras isoladas*. Trabalho de graduação, Universidade Federal de Pernambuco, Engenharia da Computação, 2009. p. 23.

Esta propriedade pode ser descrita da seguinte forma:

$$a_{ij} = 0, j < i$$

ou seja, não existem conexões entre o estado atual e os estados com índices menores que estes.

A topologia de HHM esquerda-direita é a mais comumente utilizada nos sistemas de RAF.

4.4.1.4 – Os três problemas canônicos do HHM [49] e [50]

A modelagem de um sistema ou de uma realidade é apenas uma versão simplificada destes. Assim, não há modelo absoluto, e sim, alguns modelos que são mais adequados do que outros para um dado sistema. Portanto, o processo de modelagem de um sistema é realizado em duas fases: a definição dos parâmetros do modelo e o ajuste deste através da resolução de "problemas-controle" [51].

No caso dos HHM, há três problemas canônicos (ou fundamentais) a serem resolvidos que são responsáveis pelo ajuste de um modelo que são descritos a seguir[52].

4.4.1.4.1 – Problema da avaliação

Dado um modelo $\lambda = (A, B, \pi)$ e uma seqüência de observações $O = O_1, O_2, ..., O_T$, como calcular eficientemente a probabilidade da seqüência de observações ter sido

---

[49] Ibid., p. 24.
[50] ESPINDOLA, Luciana da Silveira. Um *estudo sobre modelos ocultos de Markov* (HMM – Hidden Markov Model). Trabalho de pós-graduação, Pontifícia Universidade Católica do Rio Grande do Sul, Informática, 2009. p. 13.
[51] Ibid.
[52] RABINER, Lawrence R. *A tutorial on hidden Markov models and selected applications in speech recognition*.

gerada por um determinado modelo, ou seja, $P(O \mid \lambda)$?
Pensando em um sistema de RAF que tenha um modelo de HHM para cada palavra que este seja capaz de identificar, ao se obter uma seqüência de observações através do processamento do áudio, a determinação da palavra é feita calculando-se a probabilidade de cada modelo ter gerado a seqüência de observáveis obtida para assim encontrar o modelo mais adequado.

A solução direta para o problema seria identificar cada seqüência de estados que possa gerar as observações obtidas, o que resultaria num algoritmo de ordem exponencial, o que pode exigir uma grande capacidade computacional dependendo dos parâmetros envolvidos, inviabilizando o uso dos HHMs[53].

Para reduzir a complexidade dos cálculos, utiliza-se o algoritmo *forward*, que é de ordem polinomial.

4.4.1.4.2 – Problema da decodificação

Dado um modelo $\lambda = (A, B, \pi)$ e uma seqüência de observações $O = O_1, O_2, ..., O_T$, como encontrar a seqüência de estados mais provável que gerou as observações O?

Não existe uma solução ótima para este problema, logo, existem várias formas de resolvê-lo. Em sistemas de RAF, o algoritmo de Viterbi é muito empregado para se identificar palavras quando estas são modeladas a partir de subunidades fonéticas. Este algoritmo de ordem polinomial é visto como uma aplicação de programação dinâmica para encontrar o caminho de máxima verossimilhança em um grafo.

---

[53] ESPINDOLA, Luciana da Silveira. Um *estudo sobre modelos ocultos de Markov* (HMM – Hidden Markov Model). Trabalho de pós-graduação, Pontifícia Universidade Católica do Rio Grande do Sul, Informática, 2009.

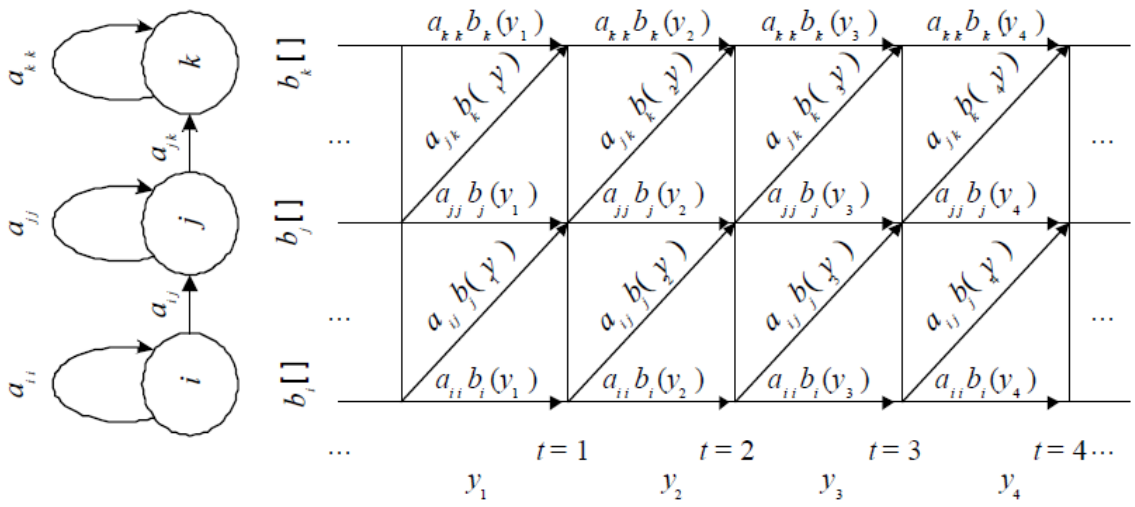


Figura 4.3: exemplo de funcionamento do algoritmo de Viterbi.[54]

A figura 4.3 mostra uma estrutura conhecida como treliça utilizada pelo algoritmo de Viterbi, com as linhas representando os estados[55].

Cada coluna armazena os valores das verossimilhanças acumuladas em cada estado do HHM para todos os instantes de tempo e todo intervalo entre duas colunas consecutivas corresponde a uma observação ou quadro de áudio analisado em um instante de tempo em um sistema de RAF.

As setas na treliça representam transições no modelo que correspondem a possíveis caminhos no modelo do instante inicial até o final. O cálculo é realizado por colunas, atualizando as probabilidades dos nós a cada quadro, utilizando fórmulas de recursão as quais envolvem os valores de uma coluna adjacente, as probabilidades de transição dos modelos, e os valores das densidades de saída para o quadro correspondente.

---

[54] YNOGUTI, Carlos Alberto. *Reconhecimento de fala contínua usando modelos ocultos de Markov*. Tese de doutorado, Universidade Estadual de Campinas, Engenharia Elétrica e Computação, 1999. p. 39.
[55] Ibid.

4.4.1.4.3 – Problema do treinamento

Como ajustar os parâmetros $\lambda = (A, B, \pi)$ do modelo para maximizar $P(O \mid \lambda)$?

Ou seja, como a partir de uma coleção de observações criar um modelo que melhor as represente para uma determinada aplicação? O objetivo é, a cada seqüência de observações processada, aumentar a probabilidade do modelo de gerar estas mesmas observações.

Como no caso do problema de decodificação, não existe solução ótima para o problema do treinamento. A técnica mais utilizada é o algoritmo *forward-backward*, também conhecido como algoritmo Baum-Welch.

Através desse algoritmo, é feita a reavaliação dos parâmetros do modelo a cada sessão de treinamento ou iteração e, ao final desta, é avaliado o grau de convergência pelo cálculo da distância (verossimilhança ou *Maximum Likelihood*) entre o novo modelo e o anterior. O treinamento é repetido até que a diferença relativa entre a verossimilhança da época atual e da época anterior atingir um valor menor que 0,0001, conforme ilustrado pela figura 4.4.

Conforme demonstrado por Ynoguti (1999), o procedimento de treino depende dos seguintes pré-requisitos:

- Determinar as unidades fundamentais que serão treinadas (subunidades fonéticas ou palavras);

- Criar o modelo de HHM para cada unidade fundamental[56];

- Alimentar o sistema com transcrições de locuções contendo as unidades

[56] Ibid.

fundamentais a serem treinadas;

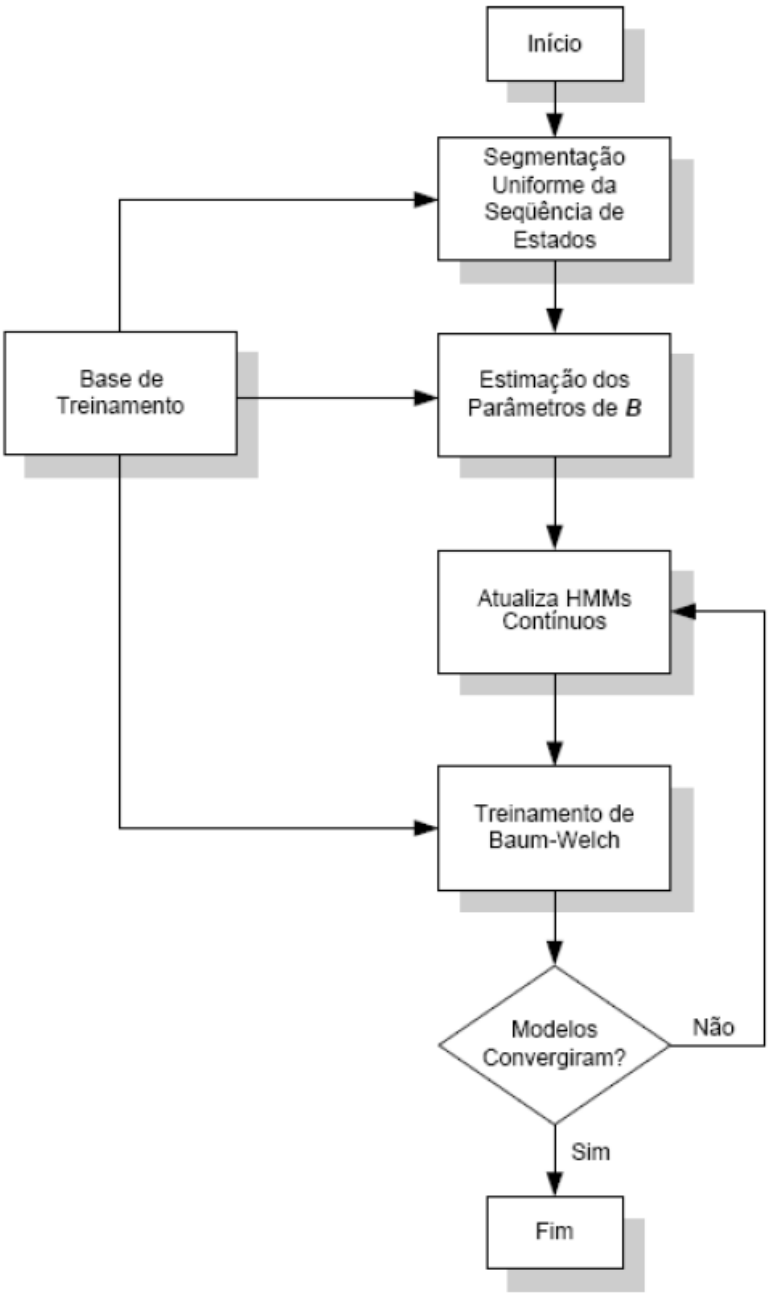


Figura 4.4: Exemplo de procedimento de treinamento.[57]

[57] SILVA, Anderson Gomes da. *Reconhecimento de voz para palavras isoladas*. Trabalho de graduação, Universidade Federal de Pernambuco, Engenharia da Computação, 2009. p. 39.

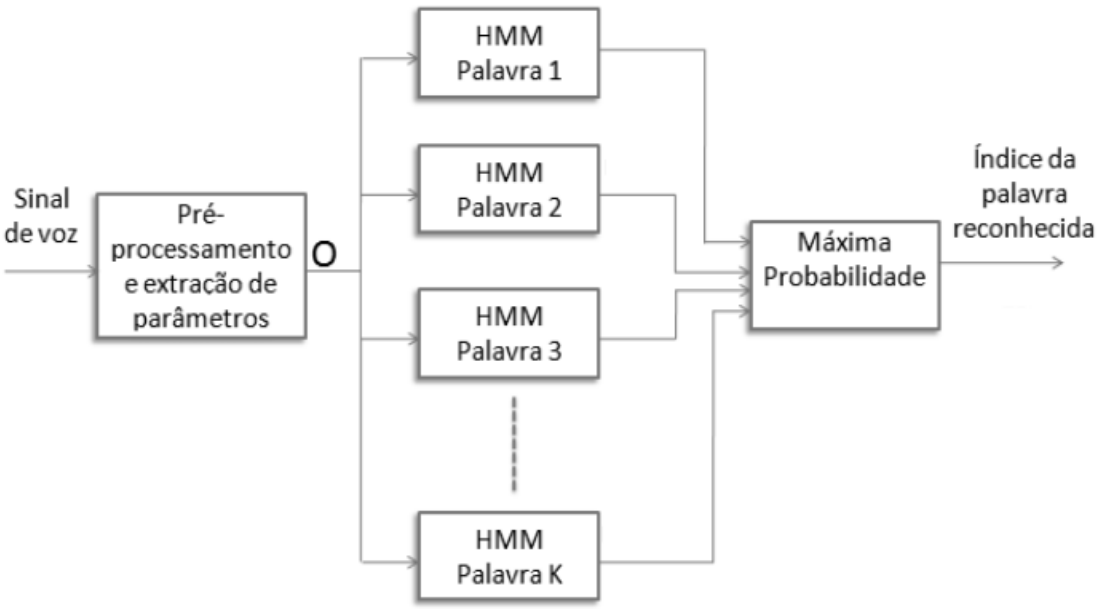
4.4.2 – Tipos de modelagem acústica

Um ponto de suma importância no desenvolvimento de um sistema de RAF é a forma como será feita a modelagem acústica do sistema. A modelagem acústica adequada para a finalidade do sistema é essencial para a eficiência do mesmo.

4.4.2.1 – Modelagem de palavras

Nesta abordagem, é criado um modelo de HHM para cada palavra que se deseja que o sistema seja capaz de reconhecer. A figura 4.5[58] ilustra o processo de reconhecimento neste caso [59].

Deve-se estabelecer o número de estados por modelo de HHM. Não existe uma quantidade ideal, esse valor deve ser descoberto através da experiência ou de testes com o HHM.



[58] Ibid.
[59] Ibid., p. 40.

Figura 4.5: Procedimento de reconhecimento para palavras isoladas.[60]

4.4.2.1 – Modelagem de subunidades fonéticas

No caso de sistemas de RAF com vocabulário extenso, torna-se inviável a manutenção de um modelo para cada palavra, além da inclusão de novos vocábulos ser altamente onerosa, já que cada uma deverá ser treinada individualmente novamente.

Como as palavras são compostas de um número finito de subunidades fonéticas, em sistemas de RAF com vocabulários extensos é mais vantajoso realizar a identificação das palavras a partir destas subunidades, já que estas são capazes de representar quaisquer outras palavras existentes. Uma abordagem pormenorizada sobre fonética e como ela é descrita para o português brasileiro pode ser vista no capítulo 2.

Há várias formas de se trabalhar com subunidades fonéticas: a partir dos fones, sílabas ou qualquer outra divisão que seja conveniente. Conforme pôde ser visto no capítulo 2.7, existem 39 fones no português brasileiro e o número de sílabas é muito maior.

Porém, assim como no caso da identificação de palavras conectadas, as subunidades fonéticas são dependentes de contexto. Por exemplo, a frase “as artes”, quando pronunciada de forma fluente, soa como “*azartes*”. Note que entre as duas palavras aparecerá o som do “z”, o que não ocorre se ambas forem pronunciadas com uma pausa entre elas. Mesmo que seja feita a divisão das palavras em sílabas, as formas de suas pronúncias serão alteradas tanto pelas sílabas que as antecedem quanto pelas que as sucedem, portanto, as subunidades fonéticas devem ser representadas, modeladas e treinadas levando em conta o

[60] SILVA, Anderson Gomes da. *Reconhecimento de voz para palavras isoladas*. Trabalho de graduação, Universidade Federal de Pernambuco, Engenharia da Computação, 2009. p. 40.

contexto em que elas serão identificadas.

De forma a contornar esse problema, pode ser utilizada a modelagem baseada em bifones ou trifones, onde é feita a modelagem das combinações de dois ou três fones, conforme demostrado abaixo na decomposição da palavra “casa”:[61]

Casa:	k	a	z	a	Unidades independentes de contexto
Casa:	<sil>-k-a	k-a-z	a-z-a	z-a-<sil>	Trifones

4.4.3 – Modelos de linguagem

É lógico inferir que o custo computacional para um sistema de RAF processar o áudio aumenta de forma proporcional à extensão de seu vocabulário de reconhecimento, já que a busca de um modelo dentre vários que possam ter gerado as observações obtidas também aumenta. Uma conta simplificada (e, de certa forma, exagerada), um vocabulário de tamanho *V*, o reconhecimento de *N* palavras resulta em V^N possibilidades, o que dá uma idéia da dimensão do problema.[62]

A forma encontrada para restringir esta busca é através do uso de modelos de linguagem, que tem como função estimar da forma mais confiável possível a probabilidade da ocorrência de uma determinada seqüência de palavras.

É comum em sistemas de ASR a utilização do modelo de linguagem n-gramas. Um *n*-grama uma seqüência de *n* símbolos utilizado para predizer a ocorrência de cada símbolo a partir de seus predecessores. A construção da coleção de *n*-gramas é feita através do processo de treinamento, quando se conta a quantidade dos mesmos para o cálculo da probabilidade de sua ocorrência em um texto desconhecido.

---

[61] JUANG, Biing-Hwang; RABINER, Lawrence. *Fundamentals of speech recognition*. New Jersey: Prentice Hall International, Inc., 1993. p. 459.

[62] LOUZADA, Jailton Alkimin. *Reconhecimento automático de fala por computador*. Trabalho de conclusão de curso, Pontifícia Universidade Católica de Goiás, Ciência da Computação, 2010. p. 21.

4.5 – RECONHECIMENTO DE FALA

O módulo de reconhecimento de fala de um sistema de RAF será aquele que irá utilizar todos os outros módulos para realizar a tarefa de transcrição de áudio propriamente dita.

A visão de geral de um modelo de sistema de RAF pode ser vista na figura 4.6.

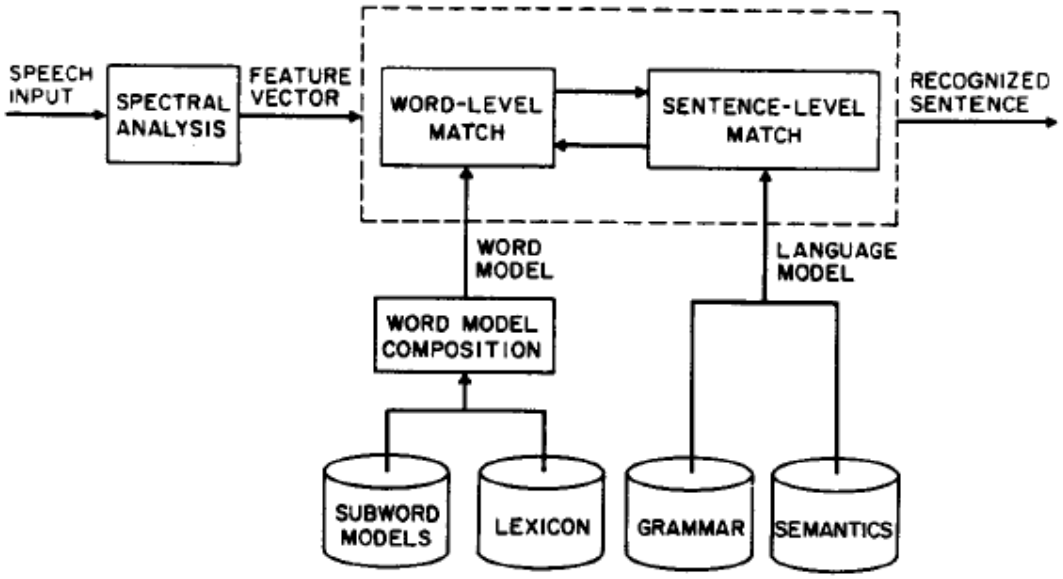


Figura 4.6: Diagrama de blocos de um modelo de sistema de RAF.[63]

[63] JUANG, Biing-Hwang; RABINER, Lawrence. *Fundamentals of speech recognition*. New Jersey: Prentice Hall International, Inc., 1993. p. 451.

5 CONSTRUINDO UM SISTEMA DE RAF COM O CSLU TOOLKIT

5.1 INTRODUÇÃO À FERRAMENTA

O *Center for Spoken Language Understanding (CSLU)* é um instituto ligado ao Departamento de Engenharia Biomédica da *Oregon Health & Science University* dos Estados Unidos que se dedica a pesquisas relacionadas à tecnologia biomédica na área da fala e visão[64].

Com o intuito de facilitar e incentivar a pesquisa da interação homem-máquina, o instituto desenvolveu o *CSLU Toolkit*, um *framework* com ferramentas básicas para este fim. A ferramenta inclui interfaces gráficas que permitem que pessoas sem um conhecimento vasto de informática e computação possam explorá-la sem grandes dificuldades e assim criar de forma intuitiva aplicações como uma Unidade de Resposta Audível (URA) capaz de interagir com usuário exclusivamente através de comandos de voz.

Para tarefas relacionadas ao RAF, o *CSLU Toolkit* disponibiliza *APIs* (*Application Programming Interfaces*) e *scripts TCL (Tool Command Language)* para o desenvolvimento de aplicações, sendo que estas últimas serão abordadas neste capítulo.

---

[64] Oregon Health & Science University. Disponível em <http://ogi.edu/bme/cslu/>. Acesso em 28 nov. 2011.

O instituto disponibiliza a ferramenta ao público livremente para avaliação ou para fins educacionais, porém, veda o seu uso comercial sem o prévio licenciamento [65].

O instituto disponibiliza a ferramenta também para uso comercial, porém sob um licenciamento diferenciado, oferecendo neste caso todo o suporte e treinamento necessário para a sua plena utilização[66].

5.2 – CONFIGURAÇÕES INICIAIS

Uma vez que o sistema é baseado na linha de comando do *Windows*, é necessário adicionar à variável de sistema *path* os diretórios "*bin*" e "*script\hmm_2.0*" para que os comandos "*.tcl"* possam ser reconhecidos e interpretados a partir de qualquer diretório do sistema operacional.

É possível a criação de arquivos com configurações pré-definidas. Ao se utilizar com um comando, pode-se utilizar o opção "*-config <arquivo de configuração>*" e especificar que as opções serão obtidas a partir do arquivo explicitado.

Exemplo de arquivo de configurações:

```
experiment.cfg
--------------
set config(tool,option1) x
set config(tool,option2) y
set param(tool,param1) param1
set param(tool,param2) param2
```

Exemplo de comando utilizando o arquivo de configurações:

```
hhmtool.cfg - - -config experiment.cfg
```

---

[65] License Terms. Disponivel em <http://www.cslu.ogi.edu/toolkit/download/license.html>. Acesso em 28 nov. 2011.

[66] Commercial Information. Disponível em <http://www.cslu.ogi.edu/toolkit/commercial.html>. Acesso em 28 nov. 2011.

Note que cada hífen seguido de espaços em branco representa uma opção que será lida do arquivo de configuração “experiment.cfg”.

5.3 – DESENVOLVIMENTO DA GRAMÁTICA

A gramática do sistema deve ser criada utilizando expressões regulares, conforme o exemplo abaixo:

```
$digit = one | two | three | four | five | six | seven |
eight | nine | zero | oh;
$grammar = [sil] <$digit [sil]>;
```

5.4 – DESENVOLVIMENTO DO MODELO DE PALAVRAS

Os modelos de pronúncia de cada palavra contida na gramática devem ser definidos em um arquivo seguindo as definições do alfabeto fonético *WorldBet*. A partir daí, são definidos os modelos de HMM para cada palavra.

No exemplo a seguir, foi criado um modelo de HHM com cinco estados para cada modelo de fone utilizado. Também foi definida a matriz de probabilidade de transição de estados para o HHM.

```
#!hscript
#
outputmodel "digit.0";
vecsize 39;
prototype mono numstate 5 mixtures 4 transp
0.000 1.000 0.000 0.000 0.000
0.000 0.600 0.400 0.000 0.000
0.000 0.000 0.500 0.500 0.000
0.000 0.000 0.000 0.600 0.400
0.000 0.000 0.000 0.000 0.000;
define mono <z> <I> <\9r> <oU>
<w> <^> <n>
<th> <u>
```

```
<T> <i:>
<f> <\>r>
<aI> <v>
<s> <ks>
<E> <&>
<ei>
<.pau> <.garbage>;
```

5.5.1 - Divisão do corpus

É sugerido dividir os arquivos de áudio em 3 categorias: os que serão utilizados para o treinamento do sistema (três quintos), os que serão usados para o desenvolvimento do sistema (um quinto) e os que serão usados para o teste do sistema (um quinto).

5.5.2 - Transcrições fonéticas das amostras de áudio

Para que o sistema possa ser treinado, é necessário realizar a transcrição fonética das amostras de áudio. Para facilitar esta tarefa, o *CSLU Toolkit* oferece o *Label GUI*, uma interface gráfica que facilita esta tarefa.

Ali, o usuário irá abrir a amostra de áudio do qual deseja fazer a transcrição fonética e visualizar o espectograma de diversos tipos do mesmo, como ilustra a figura 5.1.

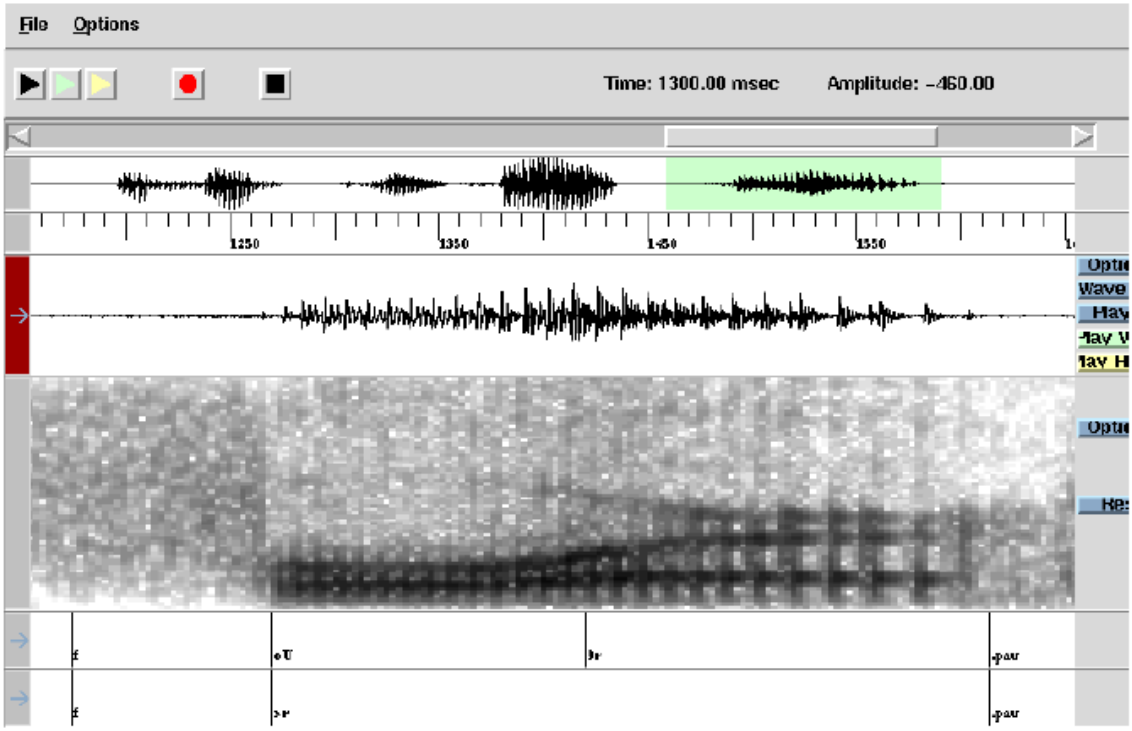


Figura 5.1: Ferramenta *Label GUI*.

5.5.3 - Extração das características do áudio

Nesta etapa, o sistema irá processar as amostras de áudio para extrair as características de áudio. O parâmetro sugerido é o cálculo de 13 coeficientes cepstrais da escala MEL para cada 10 milisegundos de áudio da amostra.

O sistema irá gerar os arquivos com no diretório especificado no arquivo de configuração:

```
set config(feature, basedir)
set config(feature, corpusdir)
```

Exemplo de uso do script:

```
genfeature.tcl -config digit.cfg
```

5.6 - TREINO DO SISTEMA

5.6.1 - Inicialização do modelo

Com as características das amostras de áudio já calculadas, o sistema irá calcular os parâmetros iniciais dos modelos HHM através da quantização vetorial e do realinhamento com o algoritmo de Viterbi.

5.6.2 - Treino individual dos modelos

Nesta etapa, os parâmetros dos modelos são refinados através do algoritmo de Baum-Welch. O arquivo de configuração deverá especificar quantidade de iterações, bem como as especificações de gramática descritas no item 5.5.

5.6.3 - Avaliação do modelo

A partir deste ponto, é possível fazer a avaliação do desempenho do reconhecedor desenvolvido. O script *buildsearch.tcl* irá criar uma gramática de estado finito a partir da gramática básica definida na configuração.

O script *hmmsearch.tcl* irá realizar a avaliação do modelo processando as amostras de áudio em busca das palavras e as repostas mais prováveis serão inseridas no arquivo definido na configuração através do parâmetro:

```
set config(search, output)
```

O comando abaixo irá listar os resultados da avaliação:

```
hmmscore.tcl -config digit.cfg
```

Saída:

```
# words : 425
# insertions : 10 (2.35294117647)
# deletions : 5 (1.17647058824)
# substitutions: 17 (4.0)
Word Correct : 94.8235294118
Sentence Correct: 77.0833333333
Accuracy : 92.4705882353
```

6 CONSIDERAÇÕES FINAIS

Esta monografia atesta o quão amplo é o problema de RAF. Negligenciar qualquer uma das áreas de conhecimento envolvidas no tema pode condenar todo o esforço empregado no desenvolvimento de um sistema de RAF. Definitivamente, somente com uma equipe multidisciplinar de pesquisa é possível chegar a resultados satisfatórios.

É impossível não notar os avanços alcançados ao longo dos anos e, com a constante evolução da capacidade computacional comercialmente disponível, é uma mera questão de tempo termos sistemas de RAF com taxas de erro comparáveis à humana.

É importante também notar que a complexidade e as tecnologias que envolvem um sistema de RAF variam e dependem muito do seu objetivo. Um sistema de reconhecimento de palavras isoladas, um sistema de reconhecimento de fala contínua e de fala espontânea, a dependência do locutor entre outras características influenciam na forma como o sistema deverá ser desenvolvido e treinado.

Como sugestão de outros trabalhos que podem se desdobrar a partir deste aqui apresentado, pode-se relacionar:

- Desenvolvimento de um sistema de reconhecimento de comandos para o português brasileiro utilizando o *CSLU Toolkit*;

- Desenvolvimento de um sistema de reconhecimento de fala contínua para o português brasileiro utilizando o *CSLU Toolkit*;

- Desenvolvimento de um sistema de reconhecimento de fala para o português brasileiro utilizando as técnicas aqui descritas e comparação de seu desempenho com o *CSLU Toolkit*;

REFERÊNCIAS BIBLIOGRÁFICAS

BORBA, Francisco S.. *Introdução aos estudos linguísticos*. São Paulo: Companhia Editora Nacional, 1975.

DA SILVA, Anderson Gomes. *Reconhecimento de voz para palavras isoladas*. Recife, PE: [s.n.], 2009.

ELLIS, Dan. *ASR*: training and systems. Columbia University, Electrical Engineering, 2003.

ESPINDOLA, Luciana da Silveira. *Um estudo sobre modelos ocultos de Markov* (HMM – Hidden Markov Model). Trabalho de pós-graduação, Pontifícia Universidade Católica do Rio Grande do Sul, Informática, 2009.

JUANG, B. H.; RABINER, L. R. *Automatic Speech Recognition: A Brief History of the Technology Development*. Santa Barbara, EUA: Rutgers University and the University of California.

JUANG, Biing-Hwang; RABINER, Lawrence. *Fundamentals of speech recognition*. New Jersey: Prentice Hall International, Inc., 1993.

LOUZADA, Jailton Alkimin. *Reconhecimento automático de fala por computador*. Trabalho de conclusão de curso, Pontifícia Universidade Católica de Goiás, Ciência da Computação, 2010.

MOUSSALLE, Sérgio (Org.); et. al. *Guia prático de otorrinolaringologia*: anatomia, fisiologia e semiologia. Porto Alegre: EDIPUCRS, 1997.

PEDROSA, Diogo Pinheiro Fernandes, *Conceitos básicos de áudio digital*, Universidade Federal do Rio Grande do Norte.

RABINER, Lawrence R. *A tutorial on hidden Markov models and selected applications in speech recognition*.

SCHALKWYK, Johan; Hosom, Paul; Kaise Ed et al. *CSLU-HMM: The CSLU Hidden Markov Modeling Environment*, Oregon Graduate Institute of Science & Technology, 2000.

SEWARD, Alexander. *Efficient methods for automatic speech recognition*. Dissertação de doutorado, Royal Institute of Technology, Stockholm, 2003. p. 5-6.

SILVA, Anderson Gomes da. *Reconhecimento de voz para palavras isoladas*. Trabalho de graduação, Universidade Federal de Pernambuco, Engenharia da Computação, 2009. p. 8.

SILVA, Patrick. *Sistemas de reconhecimento de voz para o português brasileiro utilizando os Corpora Spoltech e OGI-22*. Trabalho de conclusão de curso, Universidade Federal do Pará, Instituto de Tecnologia, 2008.

Varile, Giovanni Battista; Zampoli, Antonio. *Survey of the State of The Art in Human Language Technology*. Cambridge University Press. p. 21.

YNOGUTI, Carlos Alberto. *Reconhecimento de fala contínua usando modelos ocultos de Markov*. Campinas, SP: [s.n.], 1999.

Documentos disponíveis na *internet*

Audio normalization. In: *Wikipédia*: a enciclopédia livre. Disponível em < http://en.wikipedia.org/wiki/Audio_normalization>. Acesso em 14 dez. 2011.

Biometria: impressão vocal. In: *Grupo de Teleinformática e Automação da Universidade Federal do Rio de Janeiro – GTA/UFRJ*. Disponível em < http://www.gta.ufrj.br/grad/09_1/versao-final/impvocal/propdosinal.html>. Acesso em 25 nov. 2011.

Commercial Information. Disponível em <http://www.cslu.ogi.edu/toolkit/commercial.html>. Acesso em 28 nov. 2011.

DC bias. In: *Wikipédia*: a enciclopédia livre. Disponível em < http://en.wikipedia.org/wiki/DC_bias>. Acesso em 14 dez. 2011.

DC Offset: the case of the missing headroom. In: *Harmony Central*. Disponível em < http://www.harmonycentral.com/docs/DOC-1082>. Acesso em 14 dez. 2011.

Fonologia. In: *Wikipédia*: a enciclopédia livre. Disponível em < http://pt.wikipedia.org/wiki/Fonologia>. Acesso em 28 nov. 2011.

License Terms. Disponivel em <http://www.cslu.ogi.edu/toolkit/download/license.html>. Acesso em 28 nov. 2011.

Oregon Health & Science University. Disponível em <http://ogi.edu/bme/cslu/>. Acesso em 28 nov. 2011.

Produção de Fala Humana. In: *DEETC – Departamento de Engenharia e Eletrônica e Telecomunicações e de Computadores*. Disponível em: <http://www.deetc.isel.ipl.pt/comunicacoesep/disciplinas/pdf/sebenta/pdf/producao_2.pdf>. Acesso em 25 nov. 2011.