UNIVERSIDADE FEDERAL DE VIÇOSA
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

RECONHECIMENTO DE FALA DE LOCUTOR RESTRITO PARA ACIONAMENTO DE DISPOSITIVOS USANDO MODELOS OCULTOS DE MARKOV.

TIAGO ZANOTELLI

VIÇOSA
MINAS GERAIS - BRASIL
DEZEMBRO/2008

TIAGO ZANOTELLI

RECONHECIMENTO DE FALA DE LOCUTOR RESTRITO PARA ACIONAMENTO DE DISPOSITIVOS USANDO MODELOS OCULTOS DE MARKOV.

Parte manuscrita do Projeto de Graduação do aluno Tiago Zanotelli, apresentado ao Departamento de Engenharia Elétrica, do Centro de Ciências Exatas e Tecnológicas da Universidade Federal de Viçosa, para obtenção do título de Engenheiro Eletricista.

VIÇOSA

MINAS GERAIS - BRASIL

NOVEMBRO/2008

Agradecimentos

Gostaria de agradecer a todos que me apoiaram de forma direta ou indireta durante a minha caminhada acadêmica. Agradeço a Deus, minha família, meus pais, Marinado Zanotelli e Izabel Zanotelli, aos meus professores, em especial ao Leonardo Bonato, pela orientação e ajuda.

Agradeço aos frangos do Bonde Dos Patolas pelos momentos de alegria, a República Raça Danada que suportou-me durante 6 anos de Viçosa, a galera do Munta No Porco por ser muito doida e não poderia esquecer da mulherada fácil de Viçosa.

Obrigados a todos !

"Se o conhecimento pode criar problemas, não
é através da ignorância que podemos
solucioná-los."

Isaac Asimov

“Frango que acompanha pato, morre afogado !”

Sabedoria popular

Índice

Índice de Figuras

Índice de Tabelas

RESUMO

RECONHECIMENTO DE FALA DE LOCUTOR RESTRITO PARA ACIONAMENTO DE DISPOSITIVOS USANDO MODELOS OCULTOS DE MARKOV

As pesquisas no campo de reconhecimento de fala iniciaram-se na década de 50. Sua evolução só foi possível com os avanços dos sistemas digitais e técnicas de processamento de sinais. Atualmente, sistemas reconhecedores de fala são aplicados no acionamento de dispositivos, e.g cadeiras de rodas. Esse trabalho tem como objetivo implementar um sistema de reconhecimento de um locutor restrito, sendo o vocabulário pré-definido, constituído por cincos comandos. O sistema deve classificar o locutor como autorizado ou não, caso seja autorizado classificar o comando. Os dados de voz foram inicialmente adquiridos em um ambiente com a relação sinal/ruído de 25 dB e passaram por um pré-processamento (filtragem, normalização e corte). Para extração das características do sinal utilizou-se os coeficientes Mel-Cepstrais (CMCs) e Delta-Mel-Cepstrais (Delta-CMCs), que leva em conta o sistema de percepção auditiva humana, e também possibilita a diminuição dos dados sem a perda significativa da informação útil do sinal de entrada. Os CMCs foram usados como entradas para Modelos Ocultos de Markov, que são classificadores estocásticos que caracterizam muito bem um sinal variante no tempo, como a fala. O sistema obteve uma taxa de acerto de 99,53% para classificação do locutor e de 100,00% para classificação do comando. Para um ambiente com a relação sinal/ruído de 25 dB, conclui-se que o sistema reconhecedor mostrou eficiente na classificação do locutor e do comando.

Palavras-chave: *Reconhecimento de fala; processamento de sinais; Modelos Ocultos de Markov (HMM).*

1 Introdução

As crescentes inovações tecnológicas fizeram com que o reconhecimento de fala seja um campo de estudo fascinante. A partir da década de 50 iniciaram-se pesquisas nesse campo de estudo, destacando-se as atividades do Instituto de Tecnologia de Massachusets (MIT), nos Estados Unidos, e da Universidade de Kyoto, no Japão (PETRY e BARONE, 2000).

Em 1997 a empresa IBM lança o *ViaVoice* (ver **Figura 1-1**), software de uso doméstico que reconhece fala contínua.

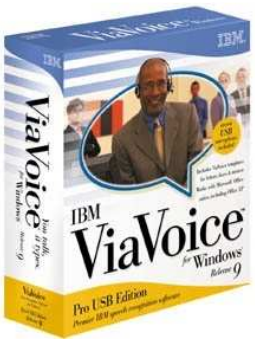


Figura 1-1 *Software* ViaVoice da empresa IBM

Aplicação do reconhecimento de fala é muito ampla, sendo a engenharia de reabilitação uma área que pode se beneficiar deste tipo de técnica, visando à melhoria da qualidade de vida de deficientes físicos.

Segundo o IBGE (IBGE, 2005), em 1991, 1,1% da população brasileira tinha algum tipo de deficiência. Desse percentual, 12,1% corresponde à pacientes paraplégicos e 2,8% a pacientes tetraplégicos. Esses deficientes se beneficiariam com a implementação de protótipos acionados por comandos de voz, que ajudaria na realização das atividades cotidianas dessas pessoas.

O reconhecimento de fala, também é aplicado em interface de atendimento de linhas telefônicas, sistemas de seguranças, etc. E hoje já existe no mercado dispositivos que reconhecem o comando dito pelo locutor e o traduz para outra língua.

A evolução dessa área só foi possível com os avanços em termos computacionais, os quais possibilitaram a conversão do som em dados digitais. Outra área que contribuiu para isso foi a de processamento de sinais, através de métodos de extração de parâmetros que caracterizam a voz, como os Coeficientes Cepstrais e Mel-Cepstrais.

O reconhecimento de padrões pode ser realizado através de metodologias estatísticas. Estas técnicas fazem comparações de padrões através da medida da função de verossimilhança, ou probabilidade condicional, a partir da observação de um modelo. As principais são: Modelos de Misturas Gaussianas (GMM, Gaussians Mixtures Models) e Modelos Ocultos de Markov (HMM, Hidden Markov Models).

Nos GMMs, as probabilidades de ocorrência dos vetores de atributos para cada locutor são modeladas com combinações ponderadas de variáveis aleatórias vetoriais com funções densidade de probabilidades (PDF) gaussianas. São usados com excelentes resultados em aplicações independentes de texto (MAFRA, 2002).

Os HMMs, são modelos com grandes capacidades de modelagem das dependências temporais associadas aos sinais de voz, apresentando os melhores resultados em aplicações dependentes de texto (MAFRA, 2002). Os HMM são largamente utilizados em sistemas corrompidos por ruídos, devido à utilização de modelos probabilísticos para representação dos sistemas (OLIVEIRA e MORITA, 2005), sendo hoje o que se pode chamar de “Estado da Arte” em reconhecimento de fala.

1.1 Tipos de Sistemas de Reconhecimento de Fala

O sistema de reconhecimento de fala pode ser classificado quanto a dependência do locutor, ao estilo da fala e pelo tamanho do vocabulário.

Quanto ao locutor podem ser dependentes ou independentes. Sistemas dependentes são capazes de reconhecer apenas o conjunto de locutores para o qual foi treinado. Podendo serem aplicados na Verificação Automática de Locutores (ASV, *Automatic Speaker Verification*), que identifica se a pessoa é autorizada ou não, ou na Identificação Automática de Locutores (ASR, *Automatic Speaker Recognition*), que reconhece diferentes locutores dentro de um conjunto de pessoas (PARREIRA, 2005).

Já os sistemas independentes são capazes de reconhecer a fala de qualquer pessoa, sendo aplicado em sistema de telefonia. Entretanto devido à grande variedade dos aspectos lingüístico (idade, sexo, nível sócio-cultural, regionalidade) necessita de um banco de dados maior do que um sistema dependente.

Quanto ao estilo da fala pode ser contínuo ou isolado. No contínuo o locutor fala normalmente, ocorrendo a coarticulação do final de uma palavra com o inicio da próxima. No isolado cada palavra é falada isoladamente ou com uma pausa entre elas, para que o sistema possa separar cada palavra.

E por último, sistemas reconhecedores de fala podem ser classificados quanto ao tamanho do vocabulário que pode ser grande, médio, pequeno (**Tabela 1-1**).

Tabela 1-1 Classificação quanto ao tamanho do vocabulário (DIAS, 2000)

Classificação	Número de palavras
Pequeno	1 a 99
Médio	100 a 999
Grande	Mais que 999

Em sistemas com vocabulário grande é inviável gravar um banco de dados com muitas palavras, para contornar esse problema o sistema modela unidades fonéticas menores (exemplo:fonemas), contudo ao fazer isso é inserido o erro devido à coarticulação de cada palavra.

1.2 Objetivo do trabalho

Este trabalho teve como objetivo implementar um sistema de Reconhecimento de Palavra com Locutor Restrito, para acionamento automático de dispositivos, e.g. cadeira de rodas, equipamentos industriais. O sistema inicialmente classifica o tipo do locutor em autorizado ou não-autorizado, em seguida, caso seja autorizado decodifica o comando proferido pelo locutor. E os comandos utilizados são: Frente, Trás, Esquerda, Direita, Para.

Para extração das características do sinal utilizou-se os Coeficientes Mel-Cepstrais (CMCs), pois possibilitam a redução do volume de dados sem perda de informação. Esses coeficientes foram utilizados no reconhecedor de padrões de voz baseado em HMM (Modelos Ocultos de Markov).

2 Metodologia

Neste capítulo são abordados os métodos utilizados no sistema de reconhecimento de fala. O sistema implementado pode ser visto na **Figura 2-1**.

Na aquisição dos dados, o sinal de fala é amostrado. Logo depois o sinal amostrado passa por um pré-processamento, que elimina as informações inúteis do sinal. A extração de parâmetros reduz o tamanho dos dados, extraindo coeficientes que preservam apenas as informações mais significativas do sinal. E esses coeficientes são as entradas do reconhecedor, parte do sistema que identifica o locutor como restrito ou não restrito, e sendo restrito classifica a palavra dentro de um conjunto de palavras pré-definidas. Após o trabalho de reconhecimento, a saída desse é utilizada para o acionamento de um dispositivo.

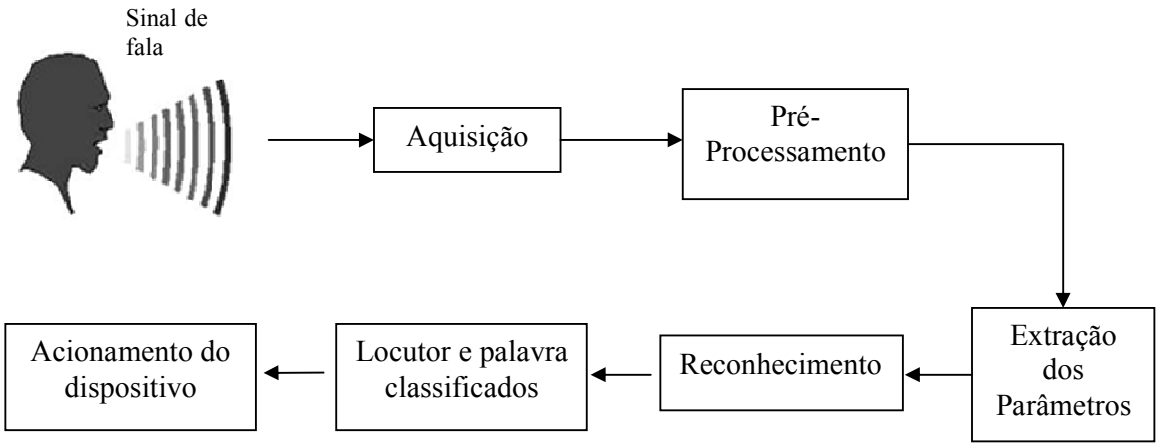


Figura 2-1 - Sistema de Reconhecimento de Fala

2.1 Aquisição de dados

Os dados foram gravados com freqüência de amostragem de 11025HZ, em um único canal com codificação PCM de16 bits, sendo gravados no formato *wave* (*.wav*).

Um dos problemas na gravação do sinal é presença de ruídos, que podem ser do tipo: aditivo, eco ou distorção do canal.

Os ruídos aditivos correspondem ao som produzido por outras fontes, que não seja o locutor. E são classificados em dois tipos: estacionário e não estacionário. O estacionário tem a potência espectral constante ao longo do tempo, e.g ventilador, e o não estacionário a potência espectral é diferente em cada instante, e.g televisão, quando fecha a porta, rádio (HUANG, ACERO e HON, 2001). E o eco corresponde ao som que é refletido por algum obstáculo e depende da acústica da ambiente.

Os dados foram gravados numa sala fechada e pequena (4 x 3 m). No instante da gravação o interior da sala estava em silêncio, evitando assim ruídos sonoros não estacionários. No entanto, não foi possível eliminar o som produzido pela ventoinha do computador, que é um ruído estacionário.

Os ruídos de distorção do canal são gerados pela resposta freqüência em do microfone e conversor A/D. Nas gravações utilizou um único microfone, mantendo-o sempre na mesma posição em relação à boca do locutor para todas as gravações. Almejando, assim, que os ruídos devido à distorção do canal sejam os mesmos para todas as gravações.

Antes de armazenar os dados, o sinal é dividido em janelas com duração de 10ms e calculado a energia de cada janela. Considere a energia do ruído (E) como a média da energia nos primeiros 100ms (equivale as 10 primeiras janelas) e a energia do sinal de voz (S) como a máxima energia do conjunto das janelas, e pela **equação 2-1** calcula a relação sinal/ruído (S/R), caso essa seja menor que 25 dB a amostra é descartada. Contudo, mesmo após o sinal passar pelo processo de seleção citado anteriormente não garante que ele seja isento de ruídos.

$$Sinal/Ruído(S/R) = 10 \times \log_{10}\left(\frac{S}{R}\right) \qquad \textbf{(eq. 2-1)}$$

Onde:

R = Energia do ruído

S = Energia do sinal de Voz

O banco de dados é constituído por gravações das palavras: Frente, Trás, Esquerda, Direita, Para. Foram gravadas 96 locuções de cada palavra por um mesmo locutor (l1, locutor restrito), nas condições de ambiente descrita

anteriormente, resultando no total de 480 gravações. Também foram gravadas mais 10 locuções de cada palavra pelo locutor (l2), 3 de cada palavra pelo l3 e 2 de cada palavra pelos locutores l4 e l5 resultando em 85 gravações de locutores não restritos.

2.2 Pré-Processamento

Os sinais gravados passaram por pré-processamento com objetivo de reduzir as perturbações e ressaltar as informações úteis, pois até mesmo os melhores sistemas de reconhecimento sofrem substancial degradação de seu desempenho quando trabalham com sinais de fala corrompidos por ruídos (CHIOVATO, 2001).

O pré-processamento é composto pelas etapas (ver **Figura 2-2**) de filtragem, corte, normalização e divisão em janelas do sinal, preparando o sinal digital para a etapa de extração de característica.

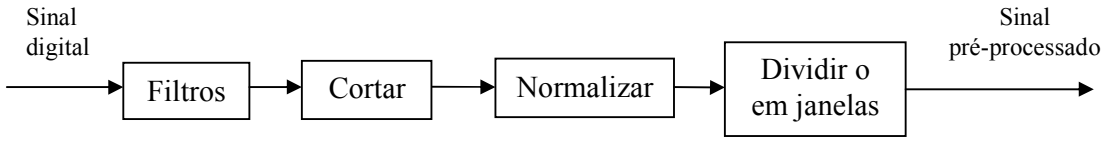


Figura 2-2 - Etapa do pré-processamento dos sinais

2.2.1 Filtros

O sinal gravado passa por dois filtros, ver **Figura 2-3**.

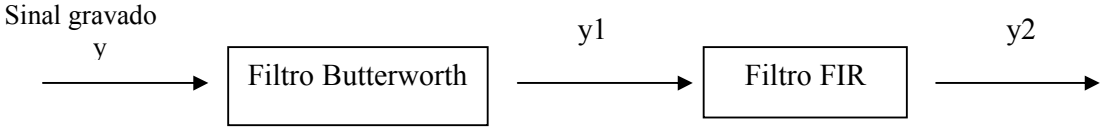


Figura 2-3- Filtros

Segundo o Teorema da Amostragem de Nyquist, para que o sinal original possa ser recuperado a maior freqüência do sinal deve ser a metade da freqüência de amostragem, sendo assim a banda passando do sinal é de 0Hz a 5512,5Hz.

O primeiro filtro tem como objetivo reduzir essa largura de banda. Para a escolha da freqüência de corte inferior levou em conta que as pessoas possuem uma freqüência fundamental de ressonância que depende das características do trato vocal, sendo freqüências entre 80 a 150 Hz para homens e 150 a 250 Hz para mulheres (Brandão, 2006), sendo assim, a freqüência de corte inferior é 80 Hz. E a freqüência de corte superior utilizada foi de 3.400 Hz, pois a freqüência máxima da voz humana é em torno desse valor.

O filtro utilizado é um Butterworth passa-faixa, de ordem 10 e com freqüência de corte de 80 a 3400 Hz (**Figura 2-4**).

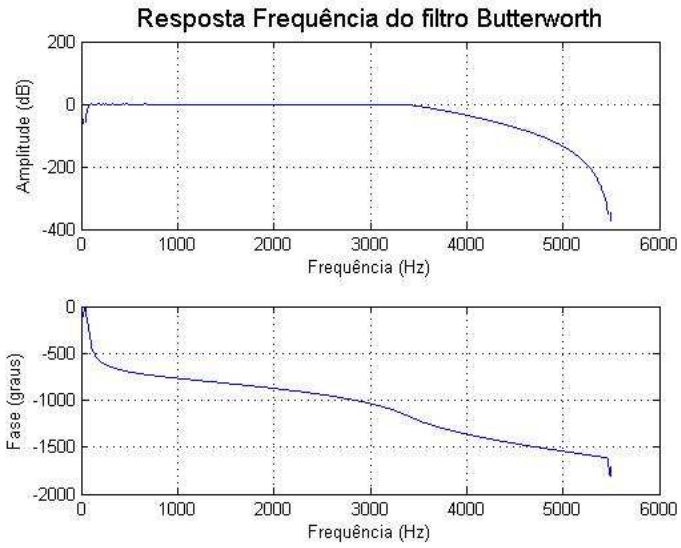


Figura 2-4- Diagrama de Bode do Filtro Butterworth

Como o Butterworth apresenta fase não-linear, para corrigir essa distorção o sinal passado pelo processo de filtragem do esquema da **Figura 2-5** (THE MATHWORKS INC., 2002).

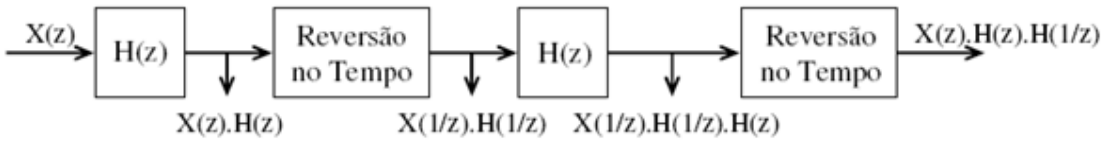


Figura 2-5 -Diagrama de Blocos para eliminação de fase, figura retirada de NIQUINI, 2007.

Sendo $X(Z)$ e $H(Z)$ as Transformada-z do sinal do tempo $x(n)$ e da função de transferência do filtro Butterworth, respectivamente, e $X(1/Z)$ a Transformada-z da função revertida no tempo $x(-n)$. A resposta em freqüência da saída se reduz a: $X(e^{jw}) \cdot \left| H\left(e^{jw}\right) \right|^2$. Portanto, a distorção de fase da saída ocasionada pelo processo de filtragem, é nula (NIQUINI, 2007).

Na **Figura 2-6** pode ser visto o sinal gravado (y) e depois de ser filtrado (y1) .

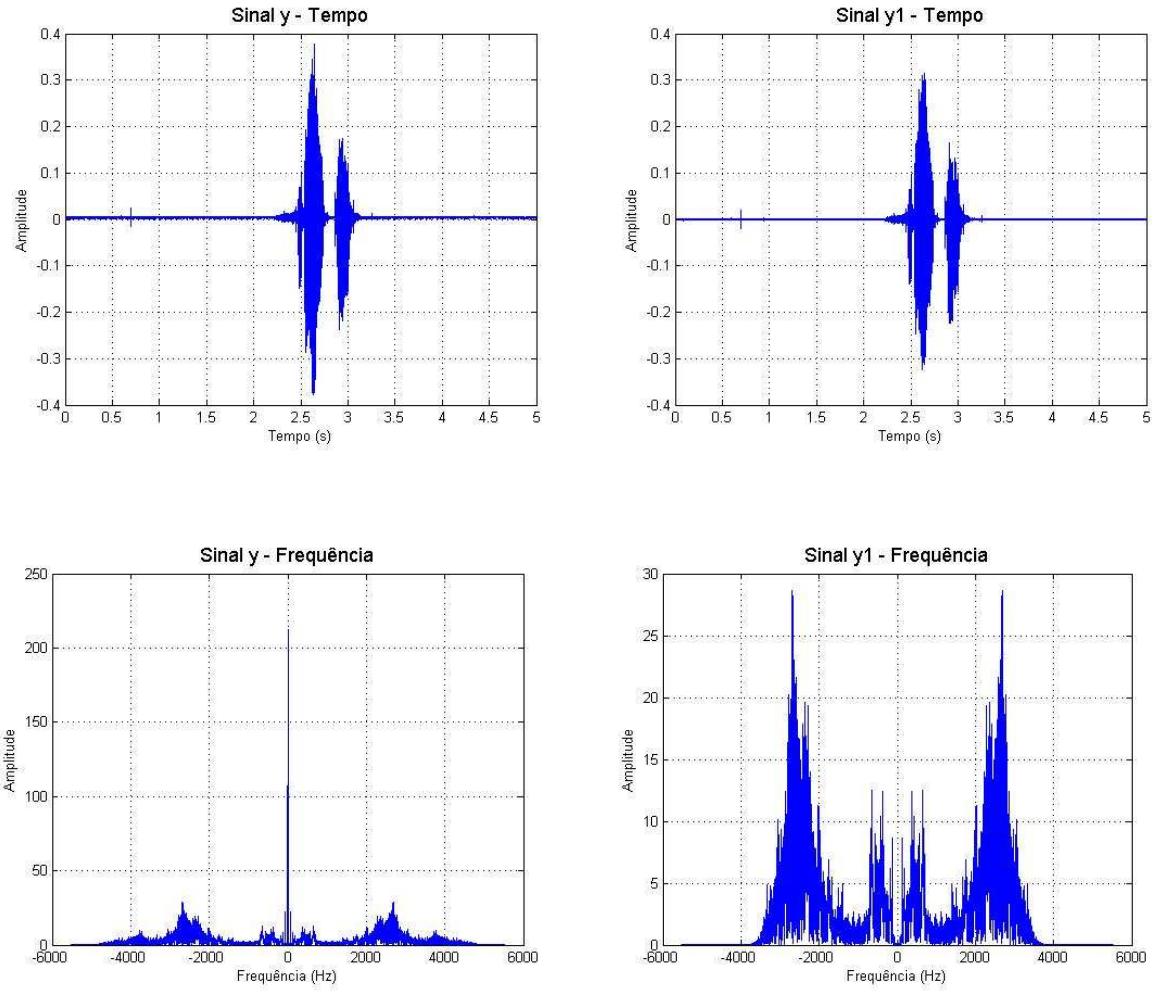


Figura 2-6 - Gráfico dos sinais y e y1 no domínio do tempo e freqüência, para o comando Frente

O sinal $y1$ por um segundo filtro, que é um filtro FIR (*finite –impulse response*) passa altas ($1 - 0{,}95z^{-1}$), com o objetivo de compensar a atenuação de 6db/oitava nas altas freqüências. Esta atenuação é ocasionada pelo efeito combinado do espectro decrescente dos pulsos glotais (-12db/oitava) e pelo efeito de radiação dos lábios (6db/oitava) (MARTINS,1997). A resposta em freqüência do filtro FIR pode ser visto na **Figura 2-7**, e o sinal de entrada (y1) e saída (y2)do filtro FIR , na **Figura 2-8**.

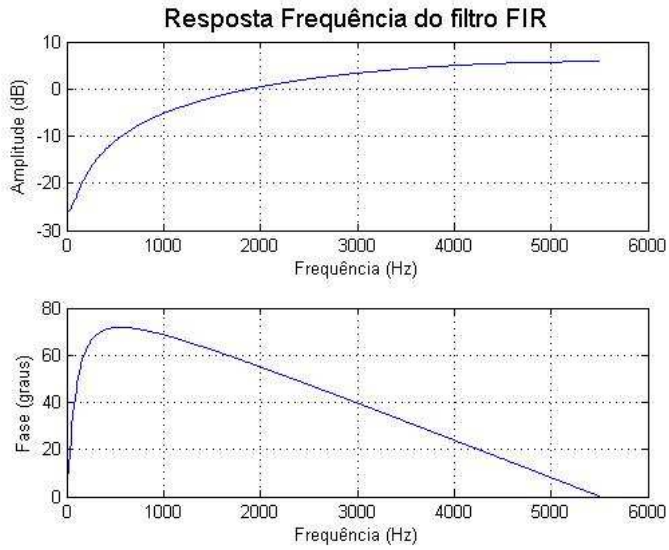


Figura 2-7- Resposta Freqüência do filtro FIR

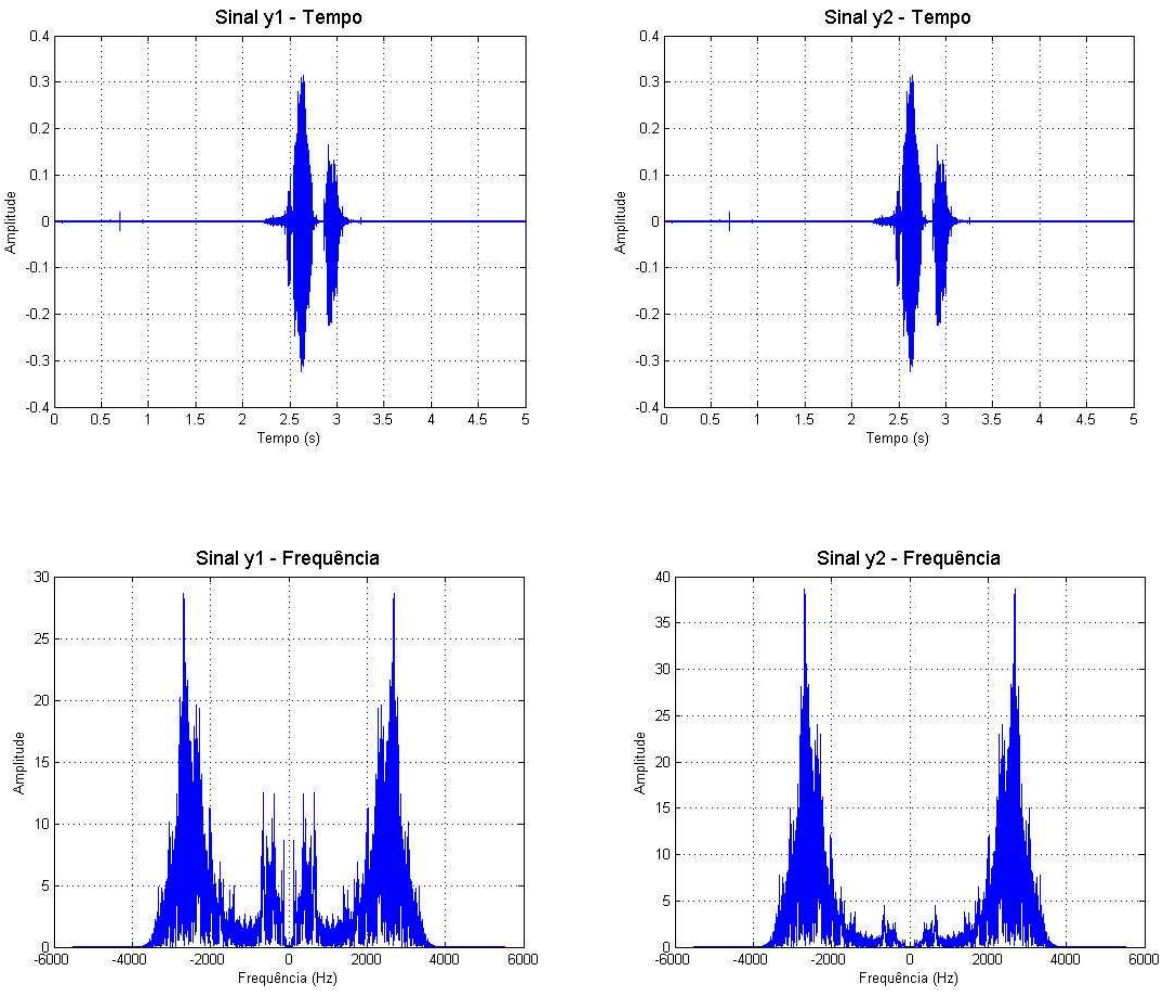


Figura 2-8 - Gráfico no domínio do tempo e freqüência do sinal y1 e y2

2.2.2 Corte

O processo de corte detecta o inicio e final da palavra, reduzindo assim o número de dados do sinal.

O algoritmo proposto para o corte, basea-se no algoritmo "*Endpoint"* (RABINER E SAMBUR,1975), com algumas modificações. Inicialmente o sinal é dividido em janelas e calculado a energia. É considerado início da palavra quando a energia do sinal passa um valor limiar de energia, E, *e* mantêm esse valor por Ni janelas, e o final quando a energia é menor que E e mantém esse valor por Nf janelas.

O sinal gravado pode apresentar níveis diferentes de ruído. Assim, para tornar o sistema mais robusto, calcula-se a relação sinal/ruído (S/R) (ver **eq. 2-1**). A partir do valor de S/R classifica o sinal em nível 1, 2 ou 3. Nível 1 corresponde a altos valores de S/R, como o nível de ruído é baixo no sinal, utiliza um valor de E baixo . No nível 2 ou médio, utiliza um valor de E um pouco maior que o nível 1. E no nível 3, ou baixo, corresponde uma baixa relação S/R, nessa condição o valor de E é maior, para evitar que o algoritmo considera a parte ruidosa como inicio. O valor de N_i e N_f permanece o mesmo para os diferentes níveis.

Sendo S o valor máximo de energia do sinal (ver **eq. 2-1**), os valores de E para os níveis 1, 2 e 3 correspondem, respectivamente, a 0,1%, 5% e 8% do valor de S.

Algoritmo: Para encontrar o início e final da palavra

Passo 1. Divide o sinal em janelas de 10ms, e calcula a energia de cada janela (**eq. 2-2**).

$$Energia\ da\ janela\ n = yy(n) = \sum_{i=1}^{M} y_n^2(i) \qquad \textbf{(eq. 2-2)}$$

Sendo:

$M = fix(fs \times 0{,}01) = 110\ pontos$

Passo 2. Calcular o S/R.

Passo 3. Classificar o nível do sinal.

Passo 4. Inicializar os argumentos: Energia (E), número de quadros para o início (Ni), número de quadros para o final (Nf), número de atraso (Nt), número de adiantamento (Nd).

Passo 5. Verificar que janela possui energia maior e menor que os valores limitares pré-estabelecido.

$$a(n) = \begin{cases} 1, & se\ yy(n) \geq Ei \\ 0, & se\ yy(n) < 0 \end{cases} \quad \textbf{(eq. 2-3)}$$

$$b(n) = \begin{cases} 1, & se\ yy(n) \leq Ef \\ 0, & se\ yy(n) > Ef \end{cases} \quad \textbf{(eq. 2-4)}$$

Passo 6. Encontrar o início:

a. Inicializa a variável $i = N_i + 1$
b. Calcula a variável aux:

$$aux = \sum_{i}^{i+Ni-1} a(i) \quad \textbf{(eq. 2-5)}$$

c. Se $aux = Ni$, $início = i - Nt$, e passa para o **passo 7**.
d. Caso contrário, se $aux \leq Ni$, $i = i + 1$, volte para o **passo b**.
e. Caso contrário, não encontrou o início da palavra, terminar o algoritmo.

Passo 7. Encontrar o final:

a. Inicializa a variável $i = início$
b. Calcula a variável aux:

$$aux = \sum_{i}^{i+Ni-1} b(i) \quad \textbf{(eq. 2-6)}$$

c. Se $aux = Nf$, $final = i + Nd$, e passa para o **passo 8**.
d. Caso contrário, se $aux \leq Nf$, $i = i + 1$, volte para o **passo b**.
e. Caso contrário, não encontro final, $final = número\ de\ janelas$.

Passo 8. Corte:

$$\boldsymbol{y_{cortado}} = y((início - 1) \times M + 1 : (final - 1) \times M) \quad \textbf{(eq. 2-7)}$$

2.2.3 Normalização

Depois de filtrado e cortado, os dados são normalizados, sendo enquadrados na faixa de -1 a 1, de acordo com a **eq. 2-8**.

$$y(n) = \frac{x(n)}{max|\boldsymbol{x}|} \quad \textbf{(eq. 2-8)}$$

2.2.4 Dividir em janelas

O sinal de voz é não-estacionário, suas características intrínsecas variam como o tempo. Porém, o trato vocal muda de forma muito lentamente na voz contínua, muitas partes da onda acústica podem ser supostas estacionária, assim para tamanhos de janelas com duração de 10 a 40ms, o processo pode ser considerado estacionário (RABINER e SCHAFER, 1978).

No trabalho utilizou janelas (*frames)* com duração de 23,21ms, pois janelas com essa duração correspondem a 256 pontos, que é uma potência de dois (2^8). Os algoritmos utilizados na extração de características são mais rápidos quando utilizado um número de pontos múltiplo de dois (2^n).

Um *frame* de voz é definido matematicamente como o produto de uma janela discreta $w(n)$ de tamanho L e determinada no tempo "l" com o sinal de voz (**eq. 2-9**).

$$frame(k) = y(n) \cdot w(l - n) \quad \textbf{(eq. 2-9)}$$

Onde:

$frame(k)$ = sinal janelado, $1 < k < número\ de\ janelas$.

$y(n)$ = sinal de voz

A janela utilizada no trabalho é a de Hamming (**eq. 2-10**), por apresentar baixa relação entre os lóbulos laterais e principal, aproximadamente 40dB. E amortecer o efeito do “Fenômeno de Gibbs”, fenômeno que ocorre no janelamento retangular devido à descontinuidade da janela (HAYKIN e VEEN, 2001).

$$w(n) = \begin{cases} 0 & n < 0 \\ 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{L-1}\right) & 0 \le n < L \\ 0 & n \ge L \end{cases} \qquad \textbf{(eq. 2-10)}$$

No trabalho o janelamento é realizado utilizando superposição de 50% entre sinais consecutivos. A superposição aumenta a correlação entre as janelas consecutivas, evitando variações bruscas entre as características extraídas das janelas, no entanto os números de dados serão maiores.

2.3 Extração das Características

O sinal, mesmo que esteja janelado, apresenta muita informação. Isso torna o processamento do computador lento, não sendo interessante para sistemas de reconhecimento de fala, pois esse às vezes necessita do reconhecimento quase instantâneo em relação ao pronunciamento das palavras. Portanto, torna-se necessário extrair parâmetros que possam caracterizar cada locução, reduzindo a quantidade de dados, mas sem a perda significativa de informações.

No trabalho para extração das características de cada janela (*frame*) é utilizado os Coeficientes Mel-Cepstrais (CMCs), que baseiam no princípio da audição humana (HUANG, ACERO e HON, 2001).

2.3.1 Coeficientes Mel-Cepstrais

Os Coeficientes Mel-Cepstrias (CMCs) são utilizados por muitos sistemas reconhecedores de fala (PICONNE, 2003; DELLER, PROAKIS e HANSEN, 1987). Os coeficientes são obtidos conforme a **Figura 2-9**.

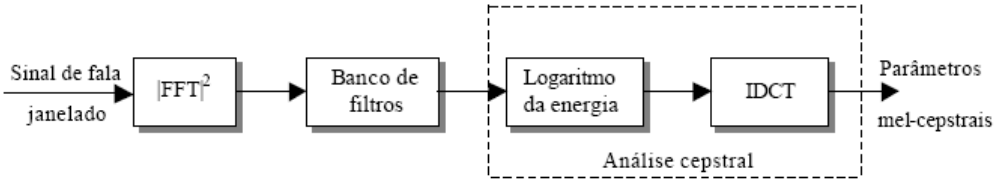


Figura 2-9 -Extração dos CMCs, figura retirada de Dias,2000

2.3.1.1 |FFT| ²

A FFT (Fast Fourier Transform) é um algoritmo computacional que realiza a DTFT (Transformada de Fourier de Tempo Discreto), mapeando o sinal para o domínio da freqüência (Haykin e Veen, 2001). As formantes (harmônicas da freqüência fundamental), são melhores caracterizados no domínio da freqüência (TIMOSZCZUK, 2004).

Ao sinal no domínio da freqüência aplica-se o operador modulo (| |), assim descartando a informação da fase, desprezada para trabalhos de reconhecimento de fala. A fase é relevante em trabalho, como por exemplo: codificadores de voz tipo Vocoders (HUANG, ACERO e HON, 2001). E por fim, aplica-se o operador potência de 2 (x^2), que equivale a potência do espectro de freqüência do sinal.

2.3.1.2 Banco de Filtros

Na audição humana o sistema responsável pelo sensoriamento dos estímulos sonoros recebidos pelo ouvido chama-se membrana basiliar, que é um filamento de aproximadamente 30mm, onde cada ponto da membrana basiliar é sensível a uma determinada freqüência. Dividindo espacialmente a membrana em partes iguais, cada uma dessa equivale a um filtro passa-faixa com largura de faixa crescente com a freqüência.

De acordo com esse modelo, aplica-se ao sinal um banco de filtros triangulares, que corresponde aos filtros passa-faixa ao longo da membrana basiliar. E esses filtros são espaçados linearmente na escala mel. Sendo MEL uma unidade de medida de freqüência, cuja sensibilidade aos sinais de voz se processa em uma

escala não linear e se semelhante ao do sistema da audição humana (PICKLES, 1988; ALLEN, 1985).

O mapeamento da freqüência acústica (f em Hz) para as freqüências percebidas em Mel (em mels), é feito pela **equação 2-11** (ver **Figura 2-10**).

$$B(f) = 1125 \cdot ln\left(1 + \frac{f}{700}\right) \quad (mel) \qquad \textbf{(eq. 2-11)}$$

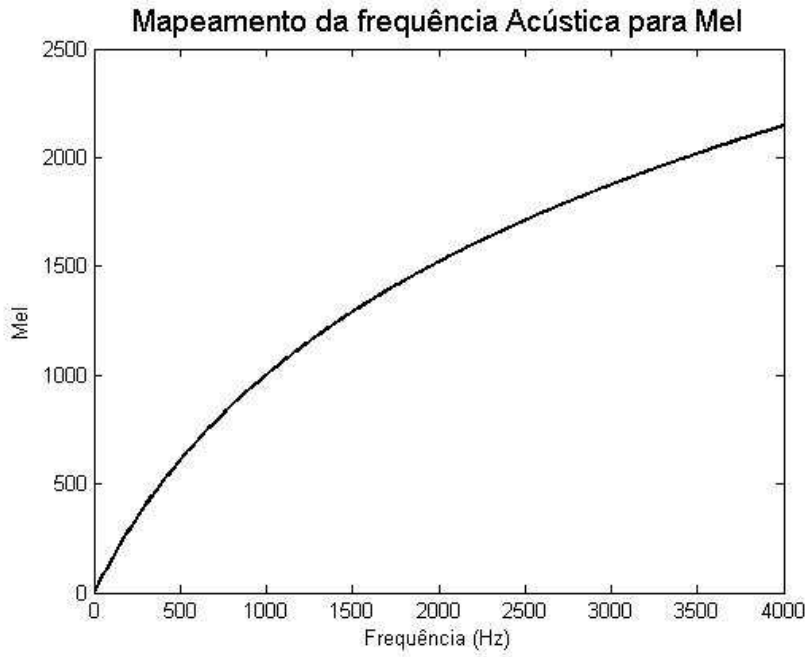


Figura 2-10 - Mapeamento da freqüência acústica de Hz para Mel

O banco de filtro triangular de tamanho *M* é definido pela **equação 2-12**.

$$H_m(k)=\begin{cases}0 & K<f(m-1)\\ \dfrac{\left(k-f(m-1)\right)}{\left(f(m)-f(m-1)\right)} & f(m-1)\le K\le f(m)\\ \dfrac{\left(f(m-1)-k\right)}{\left(f(m+1)-f(m)\right)} & f(m)\le K\le f(m+1)\\ 0 & k>f(m+1)\end{cases} \quad \textbf{(eq. 2-12)}$$

Onde:

M= números de filtros

m= número do filtro, $1 \le m \le M$

$f(m)$ = *f*reqüência central de cada filtro

Cada filtro triangular apresenta uma freqüência central $f(m)$, e sua base entre $f(m-1)$ a $f(m+1)$, **Figura 2-11**. E $f(m)$ é calculado pela **equação 2-13**.

$$f(m)=\left(\frac{N}{F_s}\right)B^{-1}\left(B(f_l)+m\frac{B(f_h)-B(f_l)}{M+1}\right) \quad \textbf{(eq. 2-13)}$$

Onde:

fl= freqüência mínima

fh=freqüência máxima

N= número de amostras

F_s= freqüência de amostragem

$B^{-1}(f)$ = inverso de B :

$$B^{-1}(f)=700\times\left(e^{\left(\frac{b}{1125}\right)}-1\right) \quad \textbf{(eq. 2-14)}$$

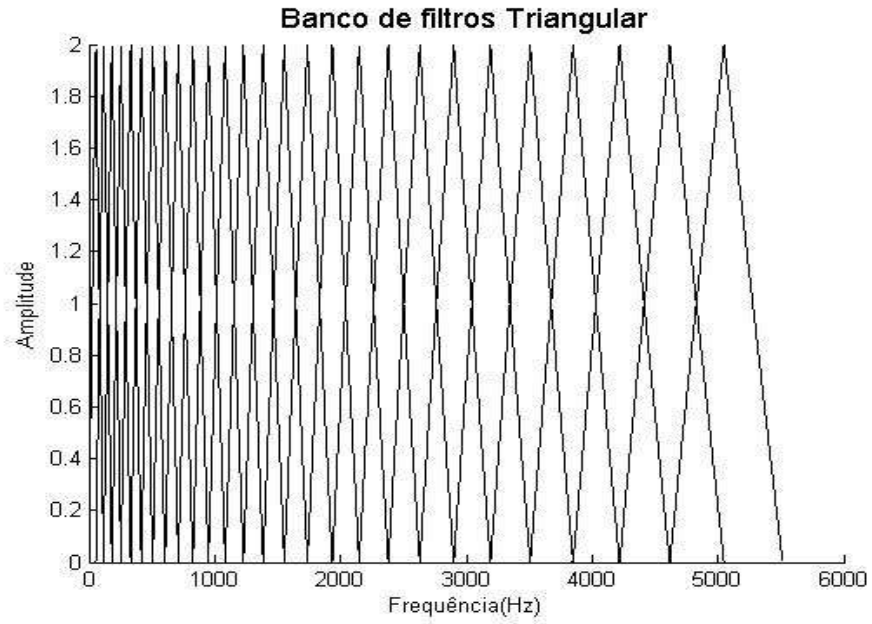


Figura 2-11 – Banco de Filtros Triangulares

2.3.1.3 Análise Cepstral

A energia do sinal é multiplicada separadamente por cada um dos filtros do banco de filtros, e calcula a log-energia da saída de cada filtro (**eq. 2-15**). Uma das vantagens de somar a saída da energia de cada filtro, é que desse modo os CMCs serão mais robustos a ruídos (HUANG, ACERO e HON, 2001).

$$S(m) = \ln\left[\sum_{k=1}^{N} \left|Y_a(k)\right|^2 H_m(k)\right] \qquad \textbf{(eq. 2-15)}$$

$$1 < m \leq M$$

Onde:

$Y_a(k)$= sinal no domínio da freqüência

$H_m(k)$= filtro m do banco de filtro na escala Mel

Os coeficientes mel-cepstrais (*mel frequency cepstral coefficients*) –são então calculados, aplicando a transformada inversa do coseno (IDT) , ver **eq. 2-16**.

$$c(n) = \sum_{m=1}^{M} S(m) \cos\left(\frac{\pi n \left(m - \frac{1}{2} \right)}{M} \right) \qquad \textbf{(eq. 2-16)}$$

$$1 \leq n < M$$

2.3.1.4 Coeficientes derivados dos Coeficientes Mel-Cepstrais

Os coeficientes derivados dos Coeficientes Mel-Cepstrais tem como objetivo melhorar a taxa de reconhecimento do sistema.

- Coeficientes Delta

Os coeficientes deltas melhor caracterizam a variação temporal do sinal da fala ao longo do tempo. Os coeficientes deltas são calculados pela **eq. 2-17 (**PICONE,1993). No trabalho utilizou k igual a 4.

$$delta(n) = \sum_{k=-K}^{K} \left(k \times c_{i-k}(n) \right) \, /(2K+1) \qquad \textbf{(eq. 2-17)}$$

2.4 Reconhecimento

O reconhecimento consiste em: dado um vetor de parâmetros, classificar este se veio de um locutor restrito ou não-restrito, e quando restrito classificar também a que classe de palavra a entrada pertence, dentro de um conjunto de palavras pré-definida. As etapas do reconhecimento estão na **Figura 2-12**. No trabalho foi utilizados Modelos Ocultos de Markov (HMM) discreto e contínuo. O HMM discreto trabalha com valores discretos, assim torna necessária a etapa da quantificação vetorial, que mapeia um vetor no espaço continuo (vetor de parâmetros) para um valor discreto. O HMM contínuo utiliza valores contínuos, assim não faz necessidade da quantização vetorial.

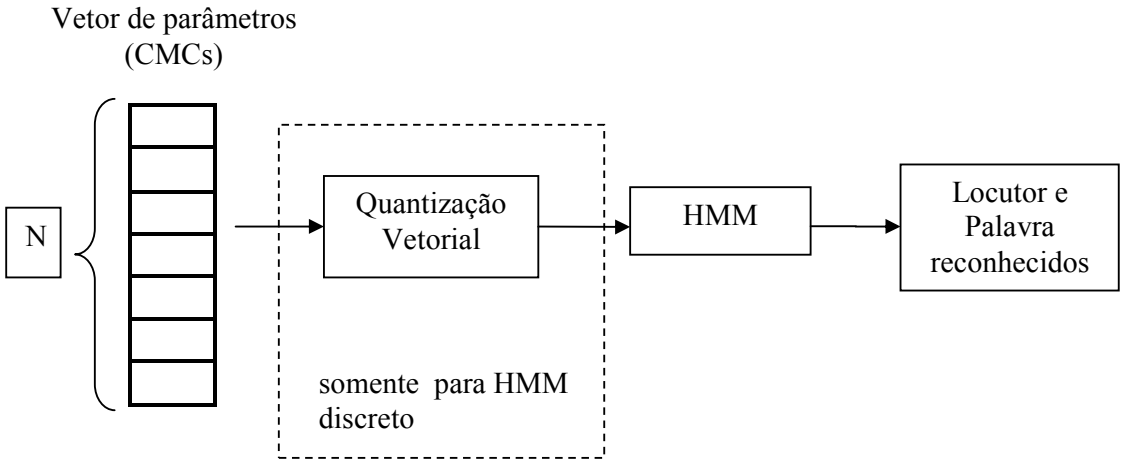


Figura 2-12 – Processo de Reconhecimento

2.4.1 Quantização Vetorial (somente para HMM discreto)

Os parâmetros que caracterizam o sinal de fala (mel-cepstrais) apresentam natureza contínua. No entanto, o HMM discreto (como o próprio nome define) emite observações de natureza discreta. Daí a necessidade do processo de quantização, que consiste em mapear um sinal de natureza contínua para valores discretos.

A quantização vetorial é realizada em duas fases: geração do codebook e quantização.

Na geração do dicionário (*codebook*), divide o espaço d-dimensional em M vetores, sendo cada um desses vetores associados a um índice ($1 \leq índice \leq M$). O método para geração do *codebook* foi o Lindo-Buzo-Gray (LGB) (RABINER e JUANG, 1993; GERSHO, 1989),que é descrito abaixo:

Algoritmo: LBG com *centroid splitting*

Passo 1. Inicialização: O primeiro *codebook* é formado por um único vetor ($M = 1$). Este é a média (centróide) de todo conjunto de treinamento.

Passo 2. Splitting: Divide cada vetor do *codebook* em outros dois, de acordo com a regra:

$$\mathbf{y_{n1}} = \mathbf{y_n} \cdot (1 + \varepsilon) \text{ e } \mathbf{y_{n2}} = \mathbf{y_n} \cdot (1 - \varepsilon) \qquad \textbf{(eq. 2-18)}$$

$\mathbf{y_n}$ = centróide n

$\mathbf{y_{n1}}$ e $\mathbf{y_{n2}}$= dois novos centróides

ε = parâmetros splitting

No final dessa etapa, tem um novo *codebook com* $2M$ vetores

Passo 3. K-means: Utiliza *K-means* para encontrar o melhor centróide do novo *codebook*. Algoritmo K-means:

- **a.** Divide todos os vetores do conjunto de treinamento em grupos, utilizando o *codebook* atual.
- **b.** Calcule a média de cada grupo.
- **c.** Atualize os valores dos centróides pelos encontrado em **b.**
- **d.** Calcule a função de custo :

$$\mathrm{D} = \frac{1}{\mathrm{M}} \sum_{\mathrm{j}=1}^{\mathrm{M}} \frac{1}{\mathrm{L_j}} \sum_{\mathrm{i}=1}^{\mathrm{L_i}} \mathrm{dist}(\mathbf{x}_\mathrm{i}(\mathrm{l}), \mathbf{y}_\mathrm{i}) \qquad \textbf{(eq. 2-19)}$$

M = número de partições

L_i= número de vetores para partição i

$\boldsymbol{x}_i(l)$ = vetor número l da partição i

$\boldsymbol{y}_i$= centróide da partição i

$dist(\boldsymbol{x}_i, \boldsymbol{y}_i)$ = distância Euclidiana (ver **eq. 2-20**)

e. Caso $D \leq D_{limiar}$ ou $\Delta D \leq \Delta D_{limiar}$ vá para **passo 4**, caso contrário volte para **a.**

Passo 4. Condição de término: Caso M atingi o valor desejado, termino do algoritmo. Caso contrário volte para o **passo 2**.

O processo de quantização visa encontrar o índice do vetor pertencente ao *codebook* que possui a menor distorção em relação ao vetor dos parâmetros característicos (vetor de entrada), **Figura 2-13**. A medida de distorção utilizada é a distância Euclidiana, **eq. 2-20**.

$$\mathrm{dist}(\mathbf{x}, \mathbf{c_i}) = \sum_{\mathrm{j}=1}^{\mathrm{K}} (\mathrm{x(j)} - \mathrm{c_i(j)})^2 \qquad \textbf{(eq. 2-20)}$$

Onde:

K = número de coeficientes característicos

i = índice do centróide, $1 \leq i \leq M$

x_j = coeficientes característicos j

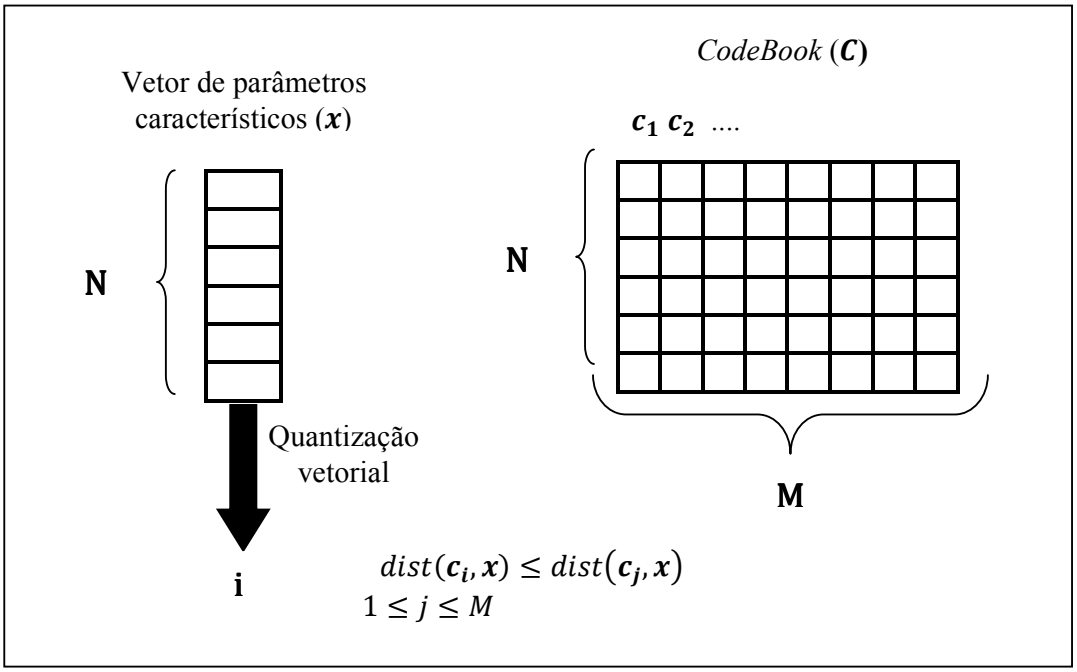


Figura 2-13- Processo de quantização, o vetor de parâmetros depois de quantizado é representado por um valor natural

O tamanho de um *codebook* (M) deve ser escolhido com o cuidado de não escolher pequeno demais, pois aumenta o erro devido à quantização, e nem grande demais, pois exige maior esforço computacional. No trabalho utilizou M igual a 64, 128, 256.

2.4.2 Modelos Ocultos de Markov

A teoria de Modelos Ocultos de Markov (*Hidden Markov Models (HMM))* começou a ser formulada no final da década 60 e inicio dos anos 70. (BAUM e PETRIE, 1966; BAUM e EAGON, 1967; BAUM, 1972).

A aplicação na área de reconhecimento de fala começou nos anos 70 (BAKER,1975), e hoje o HMM é uma das principais ferramentas no reconhecimento de fala.

O sucesso do HMM no reconhecimento de fala deve-se ao fato de ser um modelo estocástico, que caracteriza muito bem sinais que variam no tempo. Outras vantagens é a sua natureza probabilística, que é apropriada para sinais corrompidos por ruídos, como a fala, e a sua fundação teórica devido à existência de algoritmos poderosos para ajuste automático dos parâmetros do modelo através de procedimentos interativos (YACOUBI, SABOURIN, GILLOUX e SUEN, 1999).

2.4.2.1 Definição

O HMM é um processo duplamente estocástico, tendo um processo escondido (oculto) e outro observável. O escondido consiste de um conjunto de estados conectados por transições com probabilidades, e o processo observável consiste em um conjunto de saídas ou observações.

Os elementos que caracterizam o HMM (ver **Figura 2-14**) são:

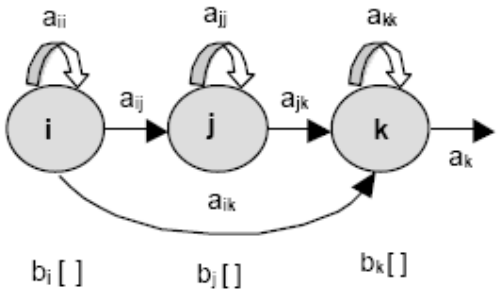


Figura 2-14 – Modelo Oculto de Markov, figura retirada de Dias, 2000.

- $\Omega = \{1,2,..,N\}$, conjunto de estados. Sendo N, número de estados do modelo. Os estados individuais no tempo t são rotulados como q_t.
  Na **figura 2-14** : $\Omega = \{i, j, k\}$.
- $\boldsymbol{O^t} = \{\boldsymbol{o_1}, \boldsymbol{o_2}, \dots, \boldsymbol{o_T}\}^t$, conjunto de observações (alfabeto). Sendo T, número de observações distintas. As observações individuais são denotadas como $\boldsymbol{o_t}$ $(1 \leq i \leq T)$. O índice t representa o tempo, como as variações do sinal ao longo do tempo são caracterizadas pela seqüência das janelas, assim, t=1 corresponde à janela 1, t=2 a janela 2.
  No HMM continuo $\boldsymbol{o_t}$ corresponde ao vetor de parâmetros característicos. No HMM discreto é o valor quantizado, assim neste caso: $\dim \boldsymbol{o_t} = \{1,1\}$ e $o_t \in \mathbb{N} = \{1,2,3, \dots, tamanho\ do\ codebook\}$.
- $\boldsymbol{A^t} = \{a_{ij}\}^t$, A matriz de transição de estados, onde a_{ij} é a probabilidade de transição do estado i para o estado j.
  A **eq. 2-21** significa que a seqüência no tempo $t-1$ estava no estado i e no tempo t está no estado j, e a probabilidade dessa transição foi de a_{ij}.

$$a_{ij} = P(q_t = j | q_{t-1} = i) \qquad \textbf{(eq. 2-21)}$$

  E a_{ij} deve obedecer às seguintes regras :

$$a_{ij} \geq 0 \quad 1 \leq i, j \leq N \qquad \textbf{(eq. 2-22)}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \ , \ 1 \leq i \leq N \qquad \textbf{(eq. 2-23)}$$

- $\boldsymbol{B^t}$, matriz de probabilidade de saída. Os HMM são classificados quanto à natureza da Função de Probabilidade ($\boldsymbol{B^t}$), que pode ser discreta ou continua.

  HMM Discreto: O número possível de valores que cada símbolo pode assumir é finito, e a matriz de probabilidade de saída é representada por $\boldsymbol{B}^t = \{b_j(k)\}$, onde $b_j(k)$, **eq. 2-24**, define a probabilidade de emissão do símbolo k, no estado j.

$$b_j(k) = P(o = k|q_t = j) \quad \textbf{(eq. 2-24)}$$

E b_j deve obedecer às seguintes regras:

$$b_j(k) \geq 0 \qquad 1 \leq j \leq N, 1 \leq k \leq M \quad \textbf{(eq. 2-25)}$$

$$\sum_{k=1}^{M} b_j(k) = 1 \quad \textbf{(eq. 2-26)}$$

HMM Contínuo: O número possível de valores que cada símbolo pode assumir é infinito. A função de densidade de probabilidade é contínua, e matriz de probabilidade de saída é representada por $\boldsymbol{B^t} = \{b_j(\boldsymbol{o_t})\}^t$, onde $b_j(\boldsymbol{o_t})$, (ver **equação 2-27**), define a probabilidade de emissão do símbolo $\boldsymbol{o_t}$, no estado j.

$$b_j(\boldsymbol{o_t}) = P(\boldsymbol{o} = \boldsymbol{o_t}|q_t = j) \quad \textbf{(eq. 2-27)}$$

A função de densidade usada no trabalho é uma mistura gaussiana da forma (RABINER e JUANG, 1993) :

$$b_j(\boldsymbol{o_t}) = \sum_{m=1}^{M} c_{jm} G(\boldsymbol{o_t}, \boldsymbol{\mu_{jm}}, \boldsymbol{U_{jm}}), \qquad 1 \leq j \leq N \quad \textbf{(eq. 2-28)}$$

Onde :

$\boldsymbol{o_t}$= vetor de entrada

M= número de misturas

c_{jm}= coeficientes da m-ésima mistura nos estados q_j.

G = função densidade de probabilidade multidimensional como vetor de médias $\boldsymbol{\mu}$ e matriz de covariância $\boldsymbol{U}$.

A função de densidade de probabilidade Gaussiana multidimensional é dada por (Miller e Freund, 1985) :

$$G(\boldsymbol{o_t}, \boldsymbol{\mu_{jm}}, \boldsymbol{U_{jm}}) = \frac{1}{(2\pi)^{\frac{dim}{2}} \left|\boldsymbol{U_{jm}}\right|^{\frac{1}{2}}} \exp\left\{-\left(\boldsymbol{o_t} - \boldsymbol{\mu_{jm}}\right)\boldsymbol{U_{jm}^{-1}}\left(\boldsymbol{o_t} - \boldsymbol{\mu_{jm}}\right)'/2\right\} \quad \textbf{(eq. 2-29)}$$

Onde:

dim = dimensão de $\boldsymbol{o_t}$

$\left|\boldsymbol{U_{jm}}\right|$= determinante da matriz de covariância $\boldsymbol{U_{jm}}$

$\boldsymbol{U_{jm}^{-1}}$ = matriz covariância inversa

E os coeficientes de misturas devem obedecer às seguintes regras :

$$\sum_{m=1}^{M} c_{jm} = 1, \qquad 1 \leq j \leq N \quad \textbf{(eq. 2-30)}$$

$$c_{jm} \geq 0\,, \qquad 1 \leq j \leq N \; e \; 1 \leq m \leq M \quad \textbf{(eq. 2-31)}$$

$$\int_{-\infty}^{+\infty} b_j(x)dx \; = 1, \qquad 1 \leq j \leq N \quad \textbf{(eq. 2-32)}$$

- $\boldsymbol{\pi} = \{\boldsymbol{\pi_i}\}$, vetor de probabilidade inicial.

O modelo HMM pode ser escrito na sua forma compacta:

$$\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}) \quad \textbf{(eq. 2-33)}$$

2.4.2.2 Topologia

Existem diversos tipos de arranjos para os estados em um HMM, no reconhecimento de fala utiliza o modelo simplificado de HMM conhecido como *left-right* ou Bakis (RABINER e JUANG, 1993), **Figura 2-15**.

Nesse modelo não pode haver dado um estado i uma transição para estado j, sendo esse ultimo inferior ao primeiro. Assim:

$$a_{ij} = 0, \;\; i > j \qquad \textbf{(eq. 2-34)}$$

São permitidas apenas transições para o próprio estado ou até dois estados à frente.

$$a_{ij} = 0, \;\; j > i + 2 \qquad \textbf{(eq. 2-35)}$$

E a função de probabilidade inicial é dada por:

$$\boldsymbol{\pi_i} = \begin{cases} 1, & i = 1 \\ 0, & i \neq 1 \end{cases} \qquad \textbf{(eq. 2-36)}$$

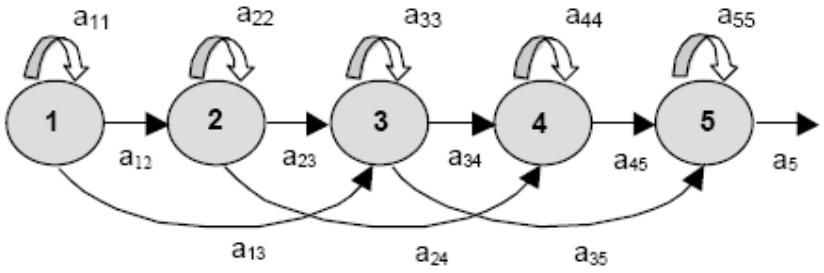


Figura 2-15 – HMM de Bakis, figura retirada de Dias (2000).

2.4.2.3 HMM em Reconhecimento de Fala

Os HMM podem ser usados para modelar qualquer unidade fonética, tais como palavras, fones, ditongos, trifones. Unidades maiores representam melhores os efeitos da coarticulação, entretanto, à medida que o tamanho da unidade básica aumenta, o número de dados também cresce. Como o vocabulário utilizado nesse trabalho é pequeno, constituído por cinco palavras, escolheu-se como unidade básica a palavra.

Para a escolha do número de estados não existe regras, normalmente tem-se o número de estados igual ao número de fonemas da palavra. Nesse trabalho procurou variar o número de estados, com o objetivo de comparar a influência que eles exercem nos resultados do reconhecimento de fala.

Os modelos de reconhecimento de fala são considerados modelos Markov de primeira ordem, e são feitas duas hipóteses (DIAS, 2000) :

1) **Hipótese de Markov:** A probabilidade de uma cadeia estar em um dado instante t, depende apenas de seu estado no instante $t-1$.
2) **Hipótese de Independência de Emissão de Símbolos**: A probabilidade de emissão de um símbolo de saída, no instante t, depende apenas da transição realizada neste instante de tempo.

Dado uma seqüência de observação ($\boldsymbol{O}^t = \{\boldsymbol{o_1}, \boldsymbol{o_2}, \dots, \boldsymbol{o_T}\}^t$) e um HMM para cada palavra (λ_i, HMM da palavra i), o reconhecimento consiste em determinar qual modelo λ_i melhor representa a seqüência de observação desejada ($\max_i[P(\lambda_i|\mathbf{O})]$).

Pela probabilidade condicional tem-se:

$$\max_i[P(\lambda_i|\mathbf{O})] = \max_i[P(\boldsymbol{O}|\lambda_i) \cdot P(\lambda_i)] \qquad \textbf{(eq. 2-37)}$$

Assumindo λ_i são equiprováveis, então :

$$\max_i[P(\lambda_i|\mathbf{O})] = \max_i[P(\boldsymbol{O}|\lambda_i], \qquad \textbf{(eq. 2-38)}$$

A **equação 2-38** classifica a seqüência de observação ($\boldsymbol{O}$) em alguma categoria i e a sua solução corresponde na identificação da palavra, sendo que o modelo λ_i encontrado é o que apresenta a maior probabilidade de reproduzir a seqüência de observação desejada ($\boldsymbol{O}$).

2.4.2.4 Dois problemas básicos do HMM

Existem dois problemas básicos do HMM que devem ser resolvidos para que possa ser feita a aplicação em reconhecimento de palavras (HUANG, ACERO e HON, 2001).

1. **Avaliação** – Dado um modelo $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ e a seqüência de observação $\boldsymbol{O} = \{\boldsymbol{o_1}, \boldsymbol{o_2}, \dots, \boldsymbol{o_T}\}$, qual probabilidade de $P(\boldsymbol{O}|\lambda)$, ou seja, qual a probabilidade do sistema tenha gerado a seqüência de observações ?

2. **Treinamento** – Dado uma seqüência de observação $\boldsymbol{O} = \{\boldsymbol{o_1}, \boldsymbol{o_2}, \dots, \boldsymbol{o_T}\}$, como ajustar os parâmetros do modelo $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$, que maximiza $P(\boldsymbol{O}|\lambda)$?

2.4.2.5 Soluções dos problemas básicos

2.4.2.5.1 Problema de Avaliação

Dado um modelo $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ com N estados e uma seqüência de observação $\boldsymbol{O} = \{\boldsymbol{o_1}, \boldsymbol{o_2}, \dots, \boldsymbol{o_T}\}$ de duração T, a probabilidade de $P(\boldsymbol{O}|\lambda)$ pode ser calculada pelo procedimento *forward-backward.*

A variável *forward* $(\alpha_t(i))$ é a probabilidade da seqüência de observações $\boldsymbol{o_1}, \boldsymbol{o_2}, \dots, \boldsymbol{o_T}$ até o tempo T, assim:

$$\alpha_t(i) = P\left(\boldsymbol{o_1}, \boldsymbol{o_2}, \dots, \boldsymbol{o_t}, q_T \middle| \lambda \right) \qquad \textbf{(eq. 2-39)}$$

E $\alpha_t(i)$ pode ser resolvido pelo algoritmo *forward* (HUANG, ACERO e HON, 2001).

Algoritmo *forward*

Passo 1. **Inicialização:**

$$\alpha_t(i) = \pi_i b_i(\boldsymbol{o_1}), \qquad 1 \leq i \leq N \qquad \textbf{(eq. 2-40)}$$

Passo 2. **Indução**

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_i(i) a_{ij} \right] \cdot b_j(\boldsymbol{o_{t+1}}) \qquad \textbf{(eq. 2-41)}$$

Onde:

$1 \leq t \leq T - 1 \;\; e \;\; 1 \leq j \leq N$

j= estado

$t =$ Tempo (como unidade discreta, número da observação)

Passo 3. **Terminação**

$$P(\boldsymbol{O}|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \qquad \textbf{(eq. 2-42)}$$

A variável *backward* ($\beta_t(i)$) é a probabilidade da seqüência de observações $\boldsymbol{o_{t+1}}, \boldsymbol{o_{t+2}}, \ldots, \boldsymbol{o_t}$, assim:

$$\beta_t(i) = P\left(\boldsymbol{o_{t+1}}, \boldsymbol{o_{t+2}}, \ldots, \boldsymbol{o_T} \middle| \lambda \right) \qquad \textbf{(eq. 2-43)}$$

E $\beta_t(i)$ pode ser resolvido pelo algoritmo *backward* (HUANG, ACERO e HON, 2001).

Algoritmo *backward*

Passo 1. **Inicialização:**

$$\beta_t(i) = 1, \qquad 1 \leq i \leq N \qquad \textbf{(eq. 2-44)}$$

Passo 2. **Indução**

$$\alpha_{t+1}(j) = \sum_{j=1}^{N} a_{ij} b_j(\boldsymbol{o_{t+1}}) \beta_{t+1}(j) \qquad \textbf{(eq. 2-45)}$$

$t = T-1, T-2, \ldots, 1 \;\; e \; 1 \leq i \leq N$

$j =$ estado

$t =$ Tempo (como unidade discreta, número da observação).

Passo 3. **Terminação**

$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i) \qquad \textbf{(eq. 2-46)}$$

Através de qualquer um dos algoritmos *forward* ou *backward*, chega à solução de $P(\boldsymbol{O}|\lambda)$. A ordem do custo computacional para cada um desses dois métodos é de $N^2 \times T$ (HUANG, ACERO e HON, 2001).

2.4.2.5.2 Problema de treinamento

O problema de treinamento procura ajustar os parâmetros do modelo $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$, que maximiza $P(\boldsymbol{O}|\lambda)$. Para isso utiliza o procedimento interativo de *Baum-Welch* (também conhecido como método EM, *expectation-maximization*),ou técnicas de gradiente (RABINER,1989). Esse método é supervisionado, necessitando de um conjunto de treinamento .

O método *Baum-Welch* ajusta os parâmetros de λ, maximizando $P(\boldsymbol{O}|\lambda)$ para um máximo local. Às vezes, o máximo local não coincide com o máximo global, podendo o modelo encontrado não ser o que apresenta a maior $P(\boldsymbol{O}|\lambda)$. Uma forma de contornar esse problema é realizar diversos treinamentos, com a inicialização dos parâmetros diferentes, e escolher o modelo que obteve os melhores resultados.

Algoritmo de *Baum-Welch*

Passo 1. **Inicialização**- Inicializa um conjunto inicial de parâmetros $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$. E o número de interação $(int) = 1$, máximo de interação $(Mint) = L$.

Passo 2. **Estimação** – Calcule $\bar{A}$, $\bar{B}$ de acordo com as fórmulas:

Probabilidade de Transição $\overline{a_{ij}}$:

$$\overline{a_{ij}} = \frac{número\ esperado\ de\ transição\ do\ estado\ i\ para\ o\ estado j}{número\ esperado\ de\ transição\ do\ estado\ i}$$

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i)\, a_{ij} b_j\left(\mathbf{o_{t+1}}\right) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)} \qquad \textbf{(eq. 2-47)}$$

Probabilidade de emisão b_j :

Discreto :

$$\overline{b_i(k)} = \frac{\textit{número esperado de ocorrer o simbolo k no estado i}}{\textit{número esperado de ocorrer o estado i}}$$

$$\overline{b}_i(k) = \frac{\sum_{\substack{t=1 \\ o_t=k}}^{T} \alpha_t(i)\beta_t(i)}{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)} \qquad \textbf{(eq. 2-48)}$$

Sendo: $P(\boldsymbol{O}/\lambda) = \sum_{i=1}^{N} \alpha_t(i) \cdot \beta_t(i)$

Continuo :

Defina-se a variável de probabilidade *posteriori* (ver **eq. 2-49** e **eq. 2-50**), como sendo a probabilidade de estar no estado j e pertencer à gaussiana m no tempo t .

$$\gamma_t(j,m) = P(q_t = j, m/\boldsymbol{O}, \lambda) \qquad \textbf{(eq. 2-49)}$$

$$\gamma(j,m)=\left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{k=1}^{N}\alpha_t(k)\beta_t(k)}\right]\left[\frac{c_{jm}G\left(\mathbf{o_t},\boldsymbol{\mu}_{\mathbf{jm}},\mathbf{U_{jm}}\right)}{b_j(\mathbf{o_t})}\right] \quad \textbf{(eq. 2-50)}$$

As equações de reestimações para $\overline{c_{jm}}$, $\overline{\boldsymbol{\mu}_{\boldsymbol{jm}}}$, $\overline{\boldsymbol{U}_{\boldsymbol{jm}}}$, são:

$$\bar{c}_{jm}=\frac{\textit{número esperado de ocorrer o gaussiana } m \textit{ no estado } j}{\textit{número esperado de ocorrer o estado } j}$$

$$\overline{c}_{jm}=\frac{\sum_{t=1}^{T}\gamma_t(j,m)}{\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_t(j,m)} \quad \textbf{(eq. 2-51)}$$

$$\bar{\boldsymbol{\mu}}_{\boldsymbol{jm}}=\frac{\textit{num. esperado de ocorrer a gaussiana } m \textit{ no estado } j \textit{ ponderada pela } o_t}{\textit{número esperado de ocorrer o estado } j \textit{ na mistura } m}$$

$$\overline{\boldsymbol{\mu}}_{\mathbf{jm}}=\frac{\sum_{t=1}^{T}\gamma_t(j,m)\mathbf{o_t}}{\sum_{t=1}^{T}\gamma_t(j,m)} \quad \textbf{(eq. 2-52)}$$

$\bar{\boldsymbol{U}}_{\boldsymbol{jm}}$

$$=\frac{\textit{núm. esperado de ocorrer a gaussiana } m \textit{ no estado } j \textit{ ponderada pela matriz de covariância}}{\textit{número esperado de ocorrer o estado } j \textit{ na mistura } m}$$

$$\overline{\mathbf{U}}_{\mathbf{jm}}=\frac{\sum_{t=1}^{T}\gamma_t(m)(\mathbf{o_t}-\boldsymbol{\mu}_{\mathbf{jm}})(\mathbf{o_t}-\boldsymbol{\mu}_{\mathbf{jm}})'}{\sum_{t=1}^{T}\gamma_t(j,m)} \quad \textbf{(eq. 2-53)}$$

Passo 3. **Terminação –** fazer $\boldsymbol{A} = \overline{\boldsymbol{A}}$, e calcular $P(\boldsymbol{O}|\lambda)$, se $P(\boldsymbol{O}|\lambda) \geq limiar$ ou $int > Mint$, termino do algoritmo, senão $int = int + 1$ e volte para o **passo 2**.

2.4.2.5.3 Inicialização do Modelo:

Como já foi dito anteriormente, a função de probabilidade inicial é dada por:

$$\pi_i = \begin{cases} 1, & i = 1 \\ 0, & i \neq 1 \end{cases}$$

Utilizaram-se métodos de inicialização diferentes para cada tipo de HMM. O HMM discreto apresenta menor sensibilidade ao valor inicial das matrizes $\boldsymbol{A}$ e $\boldsymbol{B}$, o HMM contínuo apresenta uma sensibilidade maior em relação à inicialização da matriz $\boldsymbol{B}$ (DIAS,2000).

Inicialização HMM discreto

As probabilidades de transmissão são consideradas equiprováveis. Para um modelo de N estados e de acordo com o modelo de Bakis (ver capítulo **2.4.2.2**), a matriz de transição de estados $\boldsymbol{A} = \{a_{ij}\}$ tem-se as seguintes restrições:

$$a_{ij} = 0, \;\; i > j \text{ ou } \;\; j > i + 2 \qquad \textbf{(eq. 2-54)}$$

Para os outros valores considera-se que a transição de um estado para outro seja equiprovável, assim temos:

$$\mathrm{a_{ij}} = \frac{1}{3}, 1 \leq \mathrm{i} \leq \mathrm{N} - 2 \text{ e } 1 \leq \mathrm{j} \leq \mathrm{N} \qquad \textbf{(eq. 2-55)}$$

$$\mathrm{a_{ij}} = \frac{1}{2}, \mathrm{i} = \mathrm{N} - 1 \text{ e } \mathrm{N} - 1 \leq \mathrm{j} \leq \mathrm{N} \qquad \textbf{(eq. 2-56)}$$

$$a_{ij} = 1, i = j = N \qquad \textbf{(eq. 2-57)}$$

A para a inicialização da matriz de probabilidade de saída $\boldsymbol{B} = \{b_j(k)\}$, considera, também, que cada observação é equiprovável em cada estado, como o número de observações igual a M, os valores de $\boldsymbol{B}$ são :

$$b_j(k) = \frac{1}{M}, 1 \leq K \leq M \; e \; 1 \leq j \leq N \qquad \textbf{(eq. 2-58)}$$

Inicialização HMM contínuo

O HMM contínuo não apresenta muita sensibilidade quanto à inicialização da transição de estados $\boldsymbol{A} = \{a_{ij}\}$, sendo assim utilizou o mesmo método de inicialização do HMM discreto.

Na inicialização da probabilidade de saída $\boldsymbol{B} = \{b_j(\boldsymbol{o_t})\}$, utilizou o seguinte algoritmo (EVANDRO,1997):

Passo 1. Segmentar o conjunto de observação ($\boldsymbol{O}$) de cada amostra em N partes de mesmo tamanho. N corresponde ao número de estados.

Passo 2. Caso na divisão do **Passo 1** resta alguma observação. Somar o resto na primeira ou última parte, a escolha da parte é feita de modo alternativa.

Passo 3. Agrupar os segmentos de cada amostras. O primeiro segmento da amostra k deve ser agrupado ao primeiro segmento da amostra $k + 1$, esse mesmo raciocínio segue para as outras amostras.

Passo 4. Dividir as observações correspondentes a cada grupo em M *clusters*. M é o número de misturas gaussianas. Para esse passo utilizou o método de K-means.

Passo 5. Calcular os parâmetros de cada *clusters*:

Cada grupo corresponde a um estado:

c_{jm}= número de observações no cluster m no estado q_j dividido pelo número de observações no estado q_j.

$\boldsymbol{\mu_{jm}}$= média das observações no cluster m no estado q_j.

$\boldsymbol{U_{jm}}$= matriz de covariância das observações no cluster m no estado q_j.

A matriz de covariância ($\boldsymbol{U_{jm}}$) de inicialização, utilizado nesse trabalho é a diagonal, despreza os elementos fora da diagonal por apresentar os valores mais baixo que os elementos da diagonal. Além disso a matriz diagonal converge o sistema mais rápido que a matriz cheia[1] (MARTINS,1997).

2.4.2.6 Normalização dos coeficientes forward e backward (RABINER,1989;LEVINSON,1983;JUANG, 1986).

Os valores a_{ij} e b_j são menores que 1, na medida que aumenta a seqüência de observações, o tamanho de $\alpha_t(i)$ diminui, podendo atingir valores menores que a faixa de precisão do computador, causando *underflow*. Usa-se um fator de escala ($\hat{c}_t$) independente do número de estados, dependente somente de t (tempo, que no domínio discreto corresponde ao número da observação). O método para obtenção do fator de normalização ($\hat{c}_t$) pode ser visto abaixo:

Algoritmos para obtenção dos fatores de normalização

Passo 1. **Inicializando**:

$$\tilde{\alpha}_1(i) = \alpha_1(i) \qquad \textbf{(eq. 2-59)}$$

$$\hat{c}_1 = \frac{1}{\sum_{i=1}^{N} \alpha_1(i)} \qquad \textbf{(eq. 2-60)}$$

$$\hat{\alpha}_1(i) = \tilde{\alpha}_1 \cdot \hat{c}_i \qquad \textbf{(eq. 2-61)}$$

[1] Matriz cheia: considera uma matriz cheia quando os elementos fora da diagonal são não nulos.

Onde:

$\hat{\alpha}_t$= variável *forward* normalizada no tempo t

$\tilde{\alpha}_t$=variável *forward* calculada no com os valores normalizados em t

$\hat{c}_t$ =fator de escala no tempo t

Passo 2. **Interação**

a. Calcular:

$$\tilde{a}_{t+1}(j) = \left[\sum_{i=1}^{N} \hat{a}_t(i) a_{ij} \right] \cdot b_j(O_{t+1}) \qquad \textbf{(eq. 2-62)}$$

N = número de estados e $1 \leq j \leq N$

$$\hat{c}_t = \frac{1}{\sum_{i=1}^{N} \alpha_t(i)} \qquad \textbf{(eq. 2-63)}$$

$$\hat{\alpha}_t(i) = \tilde{\alpha}_t(i) \cdot \hat{c}_t \qquad \textbf{(eq. 2-64)}$$

b. Termino: caso $t \leq T - 1$ (sendo T número de amostras), fazer: $t = t + 1$ e volte para o passo **a**, caso contrário termino do algoritmo.

Depois de encontrado os fatores de normalização ($\hat{c}_t$), calcula-se o valor de $\log(P(\boldsymbol{O}|\lambda))$ e pode-se chegar ao seu valor pelo seguinte manipulação matemática:

Pela definição, temos que:

$$P[\mathbf{O} \mid \lambda] = \sum_{i=1}^{N} \alpha_T(i) \qquad \textbf{(eq. 2-65)}$$

Aplicando $\log$:

$$\log\left(P\left[\mathbf{O} \mid \lambda\right]\right) = \log\left(\sum_{i=1}^{N} \alpha_T(i)\right) = \log(\alpha_1(i) + \alpha_2(i) + ... + \alpha_T(i)) \quad \textbf{(eq. 2-66)}$$

Como:

$$\hat{\alpha}_t(i) = \alpha_t(i) \cdot \hat{c}_t \rightarrow \alpha_t(i) = \frac{\hat{\alpha}_t(i)}{\hat{c}_t} \quad \textbf{(eq. 2-67)}$$

Substituindo a **equação 2-67** em **2-66**, e isolando o $\hat{c}_t$

$$\log\left(\frac{\sum_{t=1}^{N} \hat{\alpha}_t(i)}{\hat{c}_t}\right) = \log\left(\sum_{t=1}^{N} \hat{\alpha}_t(i)\right) - \log(\hat{c}_t) \quad \textbf{(eq. 2-68)}$$

$$como = \sum_{t=1}^{N} \hat{\alpha}_t(i) = 1$$

Então:

$$\log\left(P\left[\mathbf{O} \mid \lambda\right]\right) = -\sum_{t=1}^{T} \log \hat{c}_t(i) \quad \textbf{(eq. 2-69)}$$

2.4.2.7 Treinamento para múltiplas seqüências de treinamento

Para tornar o sistema mais robusto, ao treinar o modelo utiliza-se um banco de treinamento com múltiplas seqüências de treinamento, sendo D o número de amostras. Utilizou-se as **equações 2-70** a **2-75** para o treinamento (MARTINS, 1997).

Probabilidade de Transição $\overline{a_{ij}}$:

$$\overline{a}_{ij} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d-1} \hat{\alpha}_t^d(i)\, a_{ij} b_j\left(\mathbf{O}^d{}_{t+1}\right) \hat{\beta}_{t+1}^d(j)}{\sum_{d=1}^{D}\sum_{t=1}^{T_d-1} \frac{\hat{\alpha}_t^d(i)\hat{\beta}_t^d(i)}{\hat{c}_t^{\,d}}} \qquad \textbf{(eq. 2-70)}$$

$\widehat{\alpha_t}$ = *forward* normalizado

$\hat{\beta}_t(i)$ = *backward* normalizado ($\hat{\beta}_t(i) = \beta_{(t)}(i) \cdot c_t$))

Probabilidade de emisão $\overline{b_j}$:

Discreto:

$$\overline{b}_i(k) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \frac{\hat{\alpha}_t^d(i)\hat{\beta}_t^d(i)}{\hat{c}_t^d}}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \frac{\hat{\alpha}_t^d(i)\hat{\beta}_t^d(i)}{\hat{c}_t^d}} \qquad \textbf{(eq. 2-71)}$$

$o_k = q_k$

Continuo:

Sendo $N_t(j,m)$ igual a:

$$N_t(j,m) = \frac{c_{jm} G\left(\mathbf{o_t}, \boldsymbol{\mu}_{\mathbf{jm}}, \mathbf{U}_{\mathbf{jm}}\right)}{\sum_{k=1}^{M} c_{jk} G\left(\mathbf{o_t}, \boldsymbol{\mu}_{\mathbf{jm}}, \mathbf{U}_{\mathbf{jm}}\right)} \qquad \textbf{(eq. 2-72)}$$

Então $\overline{c_{jm}}$, $\overline{\boldsymbol{\mu_{jm}}}$ e $\overline{\boldsymbol{U_{jm}}}$ são :

$$\overline{c}_{jm} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^{\,d}(j)\hat{\beta}_t^{\,d}(j)N_t^{\,d}(j,m)/\hat{c}_t^{\,d}}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^{\,d}(j)\hat{\beta}_t^{\,d}(j)/\hat{c}_t^{\,d}} \qquad \textbf{(eq. 2-73)}$$

$$\overline{\boldsymbol{\mu}}_{\mathbf{jm}} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^{\,d}(j)\hat{\beta}_t^{\,d}(j)N_t^{\,d}(j,m)\mathbf{o}_\mathbf{t}^{\mathbf{d}}/\hat{c}_t^{\,d}}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^{\,d}(j)\hat{\beta}_t^{\,d}(j)/\hat{c}_t^{\,d}} \qquad \textbf{(eq. 2-74)}$$

$$\overline{\mathbf{U}}_{\mathbf{jm}} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^{\,d}(j)\hat{\beta}_t^{\,d}(j)N_t^{\,d}(j,m)(\mathbf{o}_\mathbf{t}^{\mathbf{d}}-\boldsymbol{\mu}_{\mathbf{jm}})(\mathbf{o}_\mathbf{t}^{\mathbf{d}}-\boldsymbol{\mu}_{\mathbf{jm}})'/\hat{c}_t^{\,d}}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^{\,d}(j)\hat{\beta}_t^{\,d}(j)N_t^{\,d}(j,m)/\hat{c}_t^{\,d}} \qquad \textbf{(eq. 2-75)}$$

2.4.2.8 Sistema de Reconhecimento e Valor Limiar para $log(P(O|\lambda))$

A medida de probabilidades usada no trabalho com a finalidade de comparar os modelos é a log-probabilidade, que compreende valores entre 0 a $-inf$. Quanto maior o valor mais próximo estará de 0, e quanto menor mais próximo do $-inf$.

O sinal de fala depois de extraído suas características, compreende a um conjunto de observações $\boldsymbol{O} = \{\boldsymbol{o_1}, \boldsymbol{o_2}, \dots, \boldsymbol{o_T}\}$. Esses conjuntos de observações são os parâmetros de entrada do sistema de reconhecimento.

O sistema de reconhecimento é constituído por cinco Modelos de Markov (HMM – λ), cada um representando uma palavra. Dado um vetor de entrada é calculada a probabilidade de cada modelo emitir tal entrada (ver **Figura 2-16**).

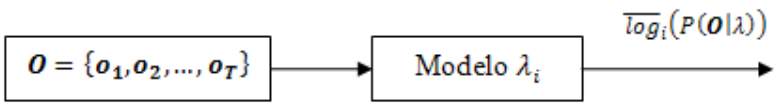


Figura 2-16 - Probabilidade do HMM i emitir a seqüência de observação O

A saída de cada modelo, passa pela condição de se o valor da probabilidade é maior que um valor $limiar$, caso seja verdadeira a condição o locutor é válido (autorizado) e a saída é a própria probabilidade, caso contrário o locutor é falso (não-autorizado) e a saída é um conjunto vazio (ver **figura 2-17**).

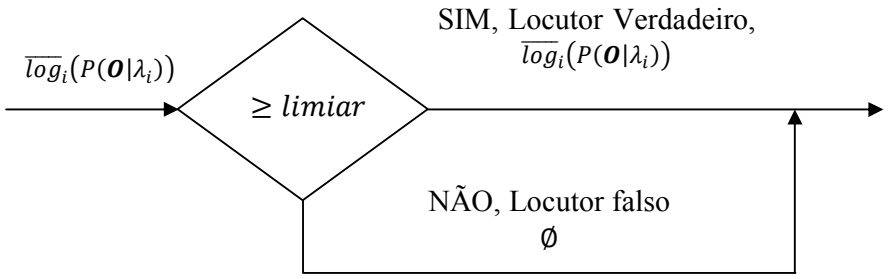


Figura 2-17 – Condição para o reconhecimento do locutor

A saída de todas as condições são entradas de uma função $Max[\,]$ e se o valor de saída dessa função for conjunto vazio (∅) a locutor é dito como não autorizado, se não é dito como autorizado e o comando será o modelo que apresenta a maior probabilidade. Na **figura 2-18** mostra o processo de reconhecimento completo.

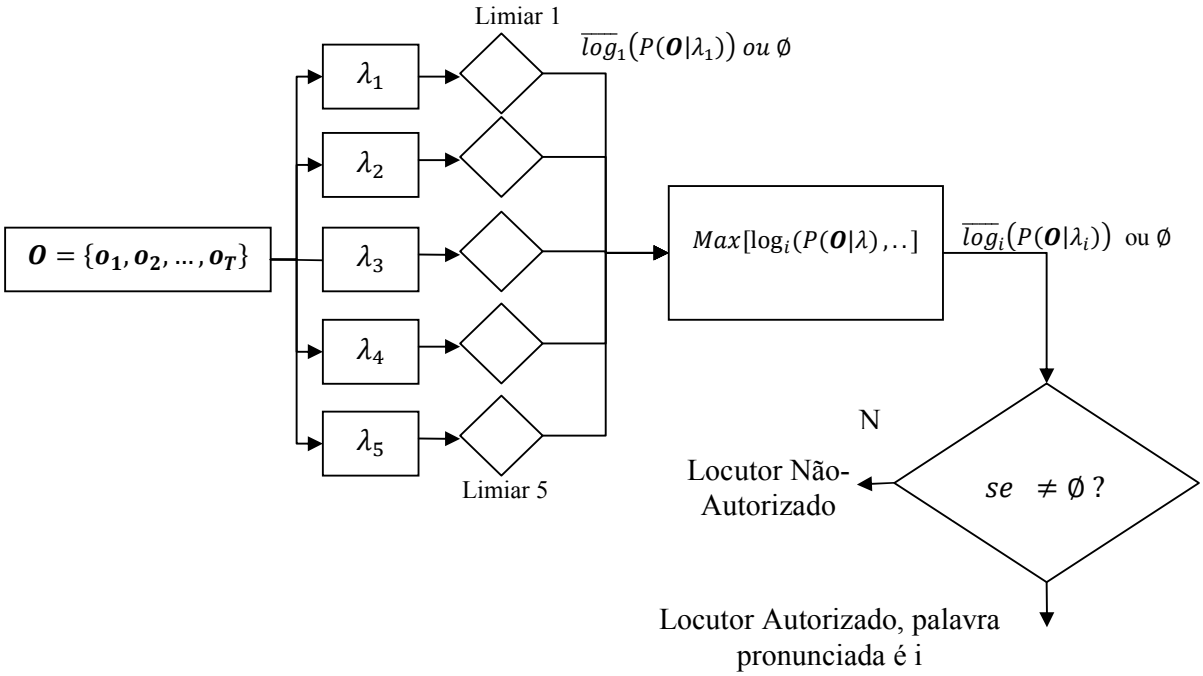


Figura 2-18 - Sistema de Reconhecimento

Ao calcular a probabilidade de cada modelo faz-se o uso da normalização no reconhecimento. Isso deve-se ao fato de que a probabilidade é proporcional ao número de observações, assim caso locutores não identificados pronunciem uma palavra rapidamente podem ser classificados como verdadeiro, para evitar esse problema, divide a probabilidade pelo número de observação, **eq. 2-76** (EVANDRO,1997).

$$\overline{log}\big(P(\boldsymbol{O}|\lambda)\big) = \frac{log(P(\boldsymbol{O}|\lambda))}{T} \qquad \textbf{(eq. 2-76)}$$

O algoritmo utilizado no calculo da probabilidade nesse trabalho foi o *forward*, por apresentar um desempenho superior ao algoritmo *viterbi* (MARTINS, 1997). E quando os parâmetros de entradas apresentarem os coeficientes CMCs e Delta-CMCs, foram considerados os modelos de cada um desses coeficientes independentes, assim têm-se para cada comando dois HMM, um para os CMCs e outro para os Delta-CMCs, e a probabilidade final emitida por cada comando é a soma da probabilidade dos dois HMM (**Figura 2-19**).

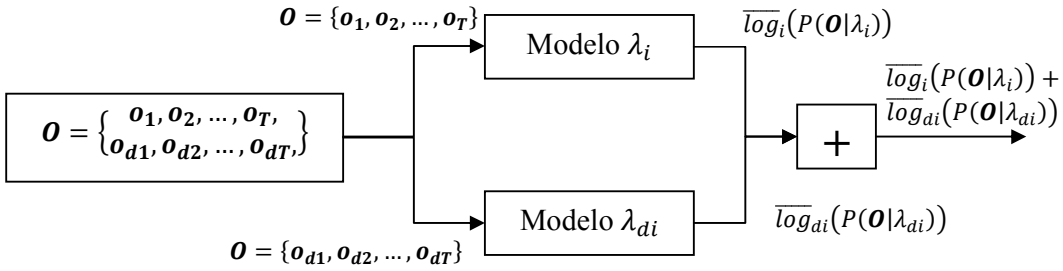


Figura 2-19 – Cálculo da probabilidade para sistemas como Delta-MFCC

2.5 Avaliação do Reconhecedor

Nessa seção é descrito os parâmetros utilizados na avaliação dos reconhecedores.

2.5.1 Banco de dados

O banco de dados é constituído por 5 palavras pré-definidas: FRENTE, TRÁS, ESQUERDA, DIREITA, PARA, sendo o locutor L1 o locutor autorizado e os outros locutores não autorizados (L2, L3, L4, L5). O banco foi dividido em três conjuntos de dados:

- C1 (Conjunto 1): constituído por 70 gravações de cada uma das 5 palavras pré-definidas, resultando no total de 350 amostras (72,92 %dos dados de L1). E esse conjunto é utilizado no treinamento.
- C2 (Conjunto 2): constituído por 26 gravações de cada uma das 5 palavras pré-definidas, resultando no total de 130 amostras (27,08% dos dados de L1). E esse conjunto é utilizado na avaliação do modelo.
- C3 (Conjunto 3): constituído por 17 gravações de cada uma das 5 palavras pré-definidas, resultando no total de 85 amostras. A gravação foi realizada pelos locutores não autorizados e será utilizado na verificação do modelo.

2.5.2 Taxas utilizadas para validação e escolha do modelo

A validação do modelo é feita utilizando os conjuntos C2, C3. O conjunto C2, refere a locutor restrito, L1, assim a taxa de acerto é calculada pela fórmula:

$$Taxa\ de\ aceto = \frac{Numero\ certo}{Número\ total\ de\ amostras} \qquad \textbf{(eq. 2-77)}$$

$Numero\ certo$= número de palavras que foram classificadas corretamente

Nota-se, quando classifica a palavra em qualquer categoria, já está subtendido que o modelo classificou o locutor como verdadeiro, assim no número de acerto (**eq. 2-77**), subtende-se que modela classificou o locutor e a palavra corretamente.

O conjunto C3, refere a locutores não restritos, o sistema deve classificar as amostras desse grupo como locutor falso (**eq. 2-78**)

$$Taxa\ de\ aceto = \frac{Numero\ de\ Locutores\ falsos}{Número\ total\ de\ amostras} \quad \textbf{(eq. 2-78)}$$

2.6 Acionamento do dispositivo

O acionamento do dispositivo ocorre pela troca de informação do computador, onde é realizado o trabalho de reconhecimento, para um dispositivo externo.

Nesse trabalho o dispositivo acionado é um conjunto de leds (**Figura 2-20**), e cada um desses representa uma das possíveis saídas do sistema reconhecedor. A forma de transmissão utilizada foi paralela e USB.

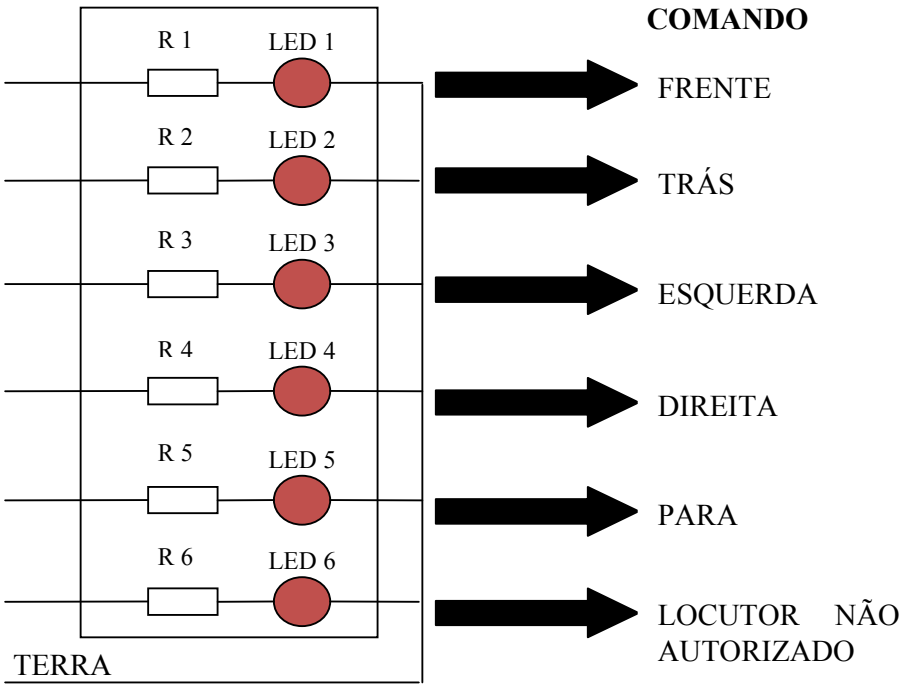


Figura 2-20-Dispositivo acionado

2.6.1 Paralela

Na transmissão paralela cada led é conectado a uma saída diferente da porta paralela do computador (LPT1), ver **Figura 2-21**. E cada bit de informação é transmitido pelas saídas no mesmo instante de tempo.

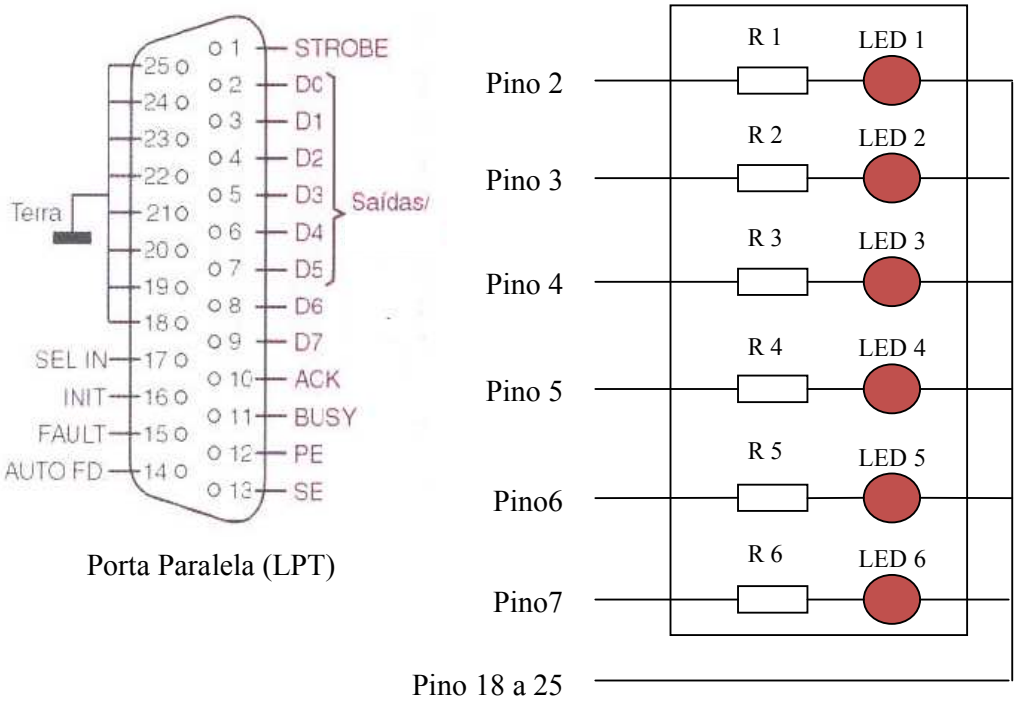


Figura 2-21 - Conexão entre a porta paralela e o dispositivo

2.6.2 USB

O padrão Universal Serial Bus (USB) foi idealizado em 1995 por um grupo de empresas de tecnologia: Compaq, Hewlett-Packard, Intel, Lucent, Microsoft, NEC, Philips. Esse padrão pode transmitir os dados de forma serial até a 60 Mbytes (última versão o USB 2.0) e existe a possibilidade de acoplar no máximo 127 dispositivos em um único barramento, mediante a utilização de *hub*. O USB na sua extremidade, pode apresentar conectores do tipo A ou B (ver **Figura 2-22**), e esse padrão utiliza quatro fios: alimentação (V+), terra e um par de fios para transmissão de informação (D+ e D-)

Figura 2-22 - Conector USB, o da direita é o tipo A e da esquerda é o tipo B

Nesse trabalho foi utilizado o circuito (**Figura 2-23**). O microcontrolador PIC18F4550 recebe os dados pela porta USB e os processa, mediante o resultado desse processo o PIC envia dados para o PORTA, onde estão conectados os leds (dispositivo **figura 2-20**).

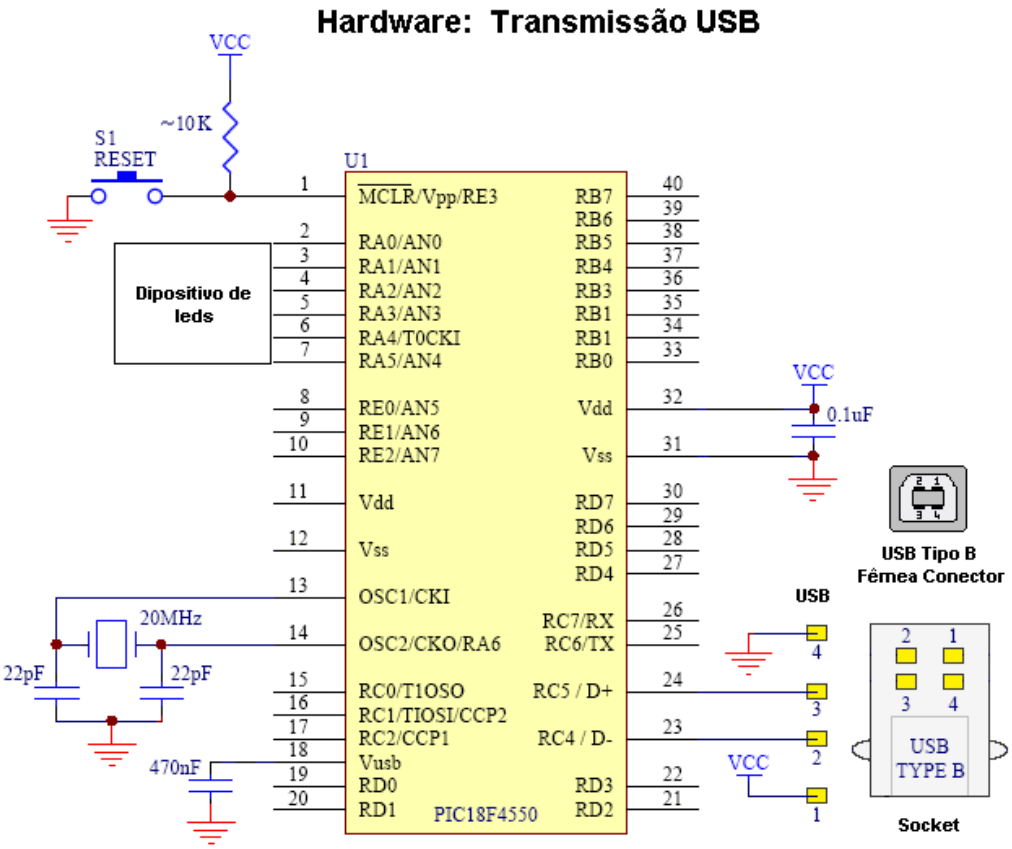


Figura 2-23 - Circuito para acionamento dos leds com comunicação USB

Para a comunicação entre PIC e o computador (PC) utilizou o *driver* “mchpcdc.inf” e “mpusbapi.dll”, fornecidas pela Microchip para PICs com interface USB, através desses *drivers* o PC reconhece o PIC no instante em que é conectado no PC. Além disso, esses geram uma porta virtual RS-232 (COM) no PC, assim qualquer *software* com interface serial (porta serial COM) pode comunicar com o dispositivo ligado no terminal do USB através dessa porta virtual, **figura 2-24** (ROJVANIT,2004). E a velocidade de transmissão pode atingir taxas de até 80 Kbytes.

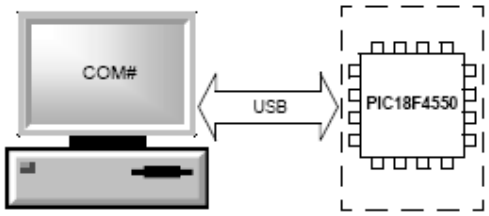


Figura 2-24 - O PC utiliza a comunicação USB através da porta RS-232 emulada pelo driver

3 Resultados e discussão

Os Modelos Ocultos de Markov apresentam diversos parâmetros que podem interferir no resultado final. Esse tópico tem como objetivo, de forma sistemática, discutir e avaliar alguns desses parâmetros e mostrar os resultados mais significativos encontrados.

3.1 Função de corte

O correto funcionamento da função de corte é de grande importância no desempenho final do sistema. Quando o corte é feito de forma errada, perdem-se informações úteis do sinal de voz.

Para testar a eficiência da função de corte, conferiu visualmente a saída da função para cada uma amostra. Como pode ser observado na **Tabela 3-1**, a função apresenta grande eficiência, porém, deve ressaltar que a relação sinal/ruídos (S/R) da amostras apresenta valores maiores que 25 dB.

Tabela 3-1 – Rendimento da função de corte

Função de corte (*Endpoint* modificada)	Acerto (%)
	100,00

Devido ao seu alto desempenho e para manter um padrão para o inicio e final de cada locução, utilizou-se a função de corte na avaliação dos sistemas de reconhecimento.

3.2 HMM discreto

O HMM discreto possui diversos parâmetros que influenciam no resultado final, como o número de Coeficientes Mel-Cepstrais, tamanho do *codebook*, número de estados e valor limiar. Procurou-se variar cada um desses modelos com o objetivo de comparar os parâmetros. Vale ressaltar novamente que para avaliar o

sistema quanto à classificação de locutor, devem-se observar as taxas de acerto de C2 e C3, porque a primeira corresponde ao locutor autorizado e a segunda ao locutor não-autorizado. E quanto à classificação do comando, observa-se apenas a taxa de acerto de C2.

Nas **Tabelas A-1** a **A-12** (ver **ANEXO A**), são mostrados os melhores resultados cara cada modelo. O parâmetro que apresentou a maior influência sobre o desempenho dos reconhecedores foi o valor limiar, isso pode ser visto comparando os resultados das **Tabelas A-1** a **A-3**, **A-4** a **A-6**, **A-7** a **A-9** e **A-10** a **A-12**.

Outra análise importante é quanto ao tamanho do *codebook*, o aumento do tamanho do *codebook* produz uma melhora no desempenho do sistema quanto às taxas de acerto de C3 (ver **Figura 3-1** e comparar os resultados nas Tabelas no **ANEXO A**).

Para o número de estados, na maioria dos casos, o incremento no número de estados em um número pequeno desse possui uma maior melhora no sistema do que o incremento em um número grande de estados, isso pode ser observado na variação da taxa de acerto entre os estados 10 e 5, e 20 e 15 (ver **Figuras 3-1** e **3-2**). Deve-se notar que para cada modelo existe um valor máximo de estado, para valores de estados maiores que esse máximo, o rendimento do sistema pode-se: diminuir (ver **Figura 3-2**, **modelo 3**) ou permanecer o mesmo, esse último ocorre para a maioria dos casos (ver **Figura 3-1**, **modelo 2**).

Tabela 3-2 - Descrição dos Modelos utilizados nas Figuras 3-1 e 3-2

Modelos	Tamanho do *Codebook*	Valor limiar	Números de CMCs
1	64	-30	64
2	128	-30	128
3	256	-30	256

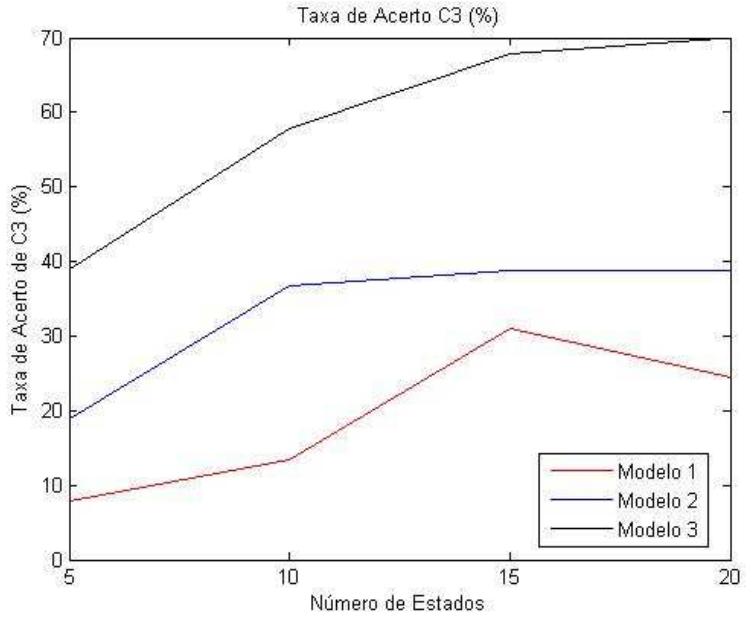


Figura 3-1 - Taxa de Acerto de C3 (%), para informações dos modelos ver Tabela 3-2

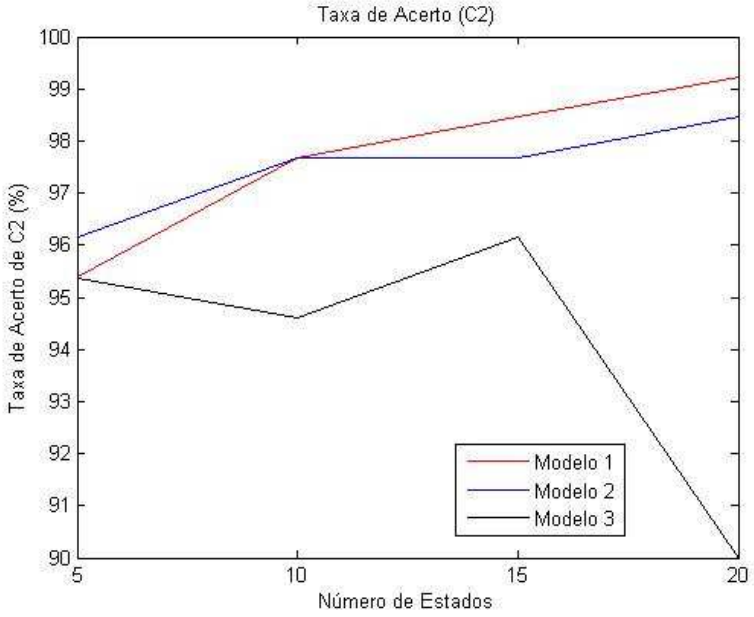


Figura 3-2 - Taxa de Acerto de C2 (%), para informações dos modelos ver Tabela 3-2

E quanto ao número de CMCs não interferiu significamente no desempenho do sistema, com uma pequena melhora para os modelos com 16 coeficientes.

Nas **Tabelas A-7** a **A-12**, foram utilizados junto com os CMCs os coeficientes Delta-CMCs. O uso dos coeficientes Delta-CMCs resultou numa melhoria no rendimento do sistema, principalmente na taxa de reconhecimento de locutor.

Resumindo, na **Tabela 3-3** têm-se os resultados dos melhores modelos.

Tabela 3-3 – Os melhores resultados encontrados pelo HMM discreto

Tamanho do *codebook*	Número de estados	Número de Coeficientes (Mel/Delta-Mel)	Valor Limiar	Taxa de Acerto % (C1)	Taxa de Acerto %(C2)	Taxa de Acerto %(C3)
256	20	13 / 0	-30,0	100,00	90,00	70,00
256	20	16/0	-30,0	100,00	92,30	70,58
256	20	13/13	-30,0	89,71	83,84	96,47
256	10	13/13	-50,0	96,28	95,38	88,23
256	10	16/16	-30,0	96,57	90,00	92,94

Para um sistema com o objetivo de classificar o locutor, como esse trabalho, não é recomendável a utilização do HMM discreto, pois o reconhecedor não consegue classificar corretamente o locutor. Contudo para a classificação do comando o desempenho do modelo é alto, como pode ser visto na **Tabela 3-4** para essa análise considerou apenas o banco de dados do locutor autorizado e retirou do sistema a classificação de locutor.

Tabela 3-4 – Rendimento do HMM discreto para reconhecimento de locutor

Tamanho do *codebook*	Número de estados	Número de Coeficientes (Mel/Delta-Mel)	Taxa de Acerto % (C1)	Taxa de Acerto %(C2)
64	10	16/ 0	100,00	99,23
64	10	16/16	100,00	100,00
128	10	16/0	100,00	100,00
128	10	16/16	100,00	100,00
256	10	16/16	100,00	100,00
256	10	16/16	100,00	100,00

3.3 HMM contínuo

O HMM contínuo possui como parâmetros variáveis o número de estados, número de CMCs e Delta-CMCs, número de misturas gaussianas e o valor limiar. Nos testes manteve-se o número de estados fixo e variou os outros parâmetros.

Uma boa estimativa do número de estados, segundo MARTINS (1997) é fazer o número de fonemas da palavra mais dois. Cada estado corresponderia a um fonema e os outros dois corresponde à região de silêncio do início e final da palavra. Sendo assim o número de estado dos comandos utilizado no trabalho são:

Tabela 3-5 - Número de estados de cada comando

Palavras	Número de estados
Frente	8
Trás	6
Esquerda	10
Direita	9
Para	6

Nas **Tabelas 3-6 a 3-9** são mostrados os melhores resultados para cada modelo. Comparando as tabelas e fazendo uma análise global observa-se que o aumento do número de gaussianas e coeficientes contribuiu com uma pequena melhora no rendimento do sistema.

Analisando as **Tabelas 3-6** e **3-7**, **3-8** e **3-9** ,observa-se a melhora do sistema com o uso dos coeficientes delta.

Tabela 3-6 – HMM contínuo, número de CMCs: 13

Número de coeficientes Mel:13				
Número de Gaussianas	Valor limiar	Taxa de Acerto % (C1)	Taxa de Acerto % (C2)	Taxa de Acerto %(C3)
1	-40	64,25	33,84	100,00
	-45	100,00	97,69	98,82
	-50	100,00	99,23	62,35
	-55	100,00	100,00	32,94
3	-40	52,00	31,53	98,82
	-45	100,00	94,61	98,82
	-50	100,00	100,00	62,35
	-55	100,00	99,23	31,76
4	-40	64,28	33,84	100,00
	-45	100,00	97,69	98,82
	-50	100,00	99,23	58,82
	-55	100,00	100,00	35,29
5	-40	71,42	33,07	100,00
	-45	100,00	98,46	98,82
	-50	100,00	100,00	64,70
	-55	100,00	100,00	36,47
7	-40	77,14	26,92	100,00
	-45	100,00	96,92	100,00
	-50	100,00	100,00	68,23
	-55	100,00	100,00	43,52

Tabela 3-7- HMM contínuo, número de CMCs e Delta-CMCs: 13

Número de coeficientes Mel e Delta-Mel:13				
Número de Gaussianas	Valor limiar	Taxa de Acerto % (C1)	Taxa de Acerto % (C2)	Taxa de Acerto %(C3)
1	-70	100,00	99,23	96,47
	-75	100,00	99,23	74,11
	-80	100,00	100,00	48,23
	-85	100,00	98,46	35,29
3	-70	100,00	99,23	96,47
	-75	100,00	100,00	70,58
	-80	100,00	100,00	48,23
	-85	100,00	100,00	32,94
4	-70	100,00	99,23	96,47
	-75	100,00	99,23	74,11
	-80	100,00	100,00	48,23
	-85	100,00	100,00	35,29
5	-70	100,00	100,00	97,64
	-75	100,00	100,00	75,29
	-80	100,00	100,00	49,41
	-85	100,00	100,00	37,64
7	-70	100,00	99,23	98,82
	-75	100,00	100,00	83,52
	-80	100,00	100,00	63,53
	-85	100,00	100,00	45,88

Tabela 3-8- HMM contínuo, número de CMCs: 16

Número de coeficientes Mel :16				
Número de Gaussianas	Valor limiar	Taxa de Acerto % (C1)	Taxa de Acerto % (C2)	Taxa de Acerto %(C3)
1	-40	0,00	0,00	100,00
	-45	0,00	0,00	100,00
	-50	24,57	16,15	98,82
	-55	90,28	85,38	97,64
3	-40	1,42	0,00	98,82
	-45	34,28	17,69	98,82
	-50	96,85	79,23	98,82
	-55	100,00	97,69	97,64
4	-40	0,85	0,00	98,82
	-45	29,71	11,53	98,82
	-50	97,14	72,30	98,82
	-55	100,00	99,23	94,11
5	-40	1,42	0,00	98,82
	-45	34,25	17,69	98,82
	-50	96,85	79,23	98,82
	-55	100,00	97,69	97,64
7	-40	2,85	0,00	100,00
	-45	51,71	14,61	98,82
	-50	100,00	68,46	98,82
	-55	100,00	95,38	95,29

Tabela 3-9- HMM contínuo, número de CMCs e Delta-CMCs: 16

Número de coeficientes Mel e Delta-Mel:16				
Número de Gaussianas	Valor limiar	Taxa de Acerto % (C1)	Taxa de Acerto % (C2)	Taxa de Acerto %(C3)
1	--70	15,14	5,38	98,82
	--75	38,28	32,30	98,82
	-80	90,57	84,61	98,82
	--85	99,71	96,92	94,11
3	--70	97,71	53,07	100,00
	--75	100,00	86,92	100,00
	-80	100,00	96,92	100,00
	--85	100,00	100,00	92,94
4	--70	89,71	46,23	98,82
	--75	99,14	83,07	98,82
	-80	100,00	96,92	98,82
	--85	100,00	100,00	87,05
5	--70	97,71	53,07	100,00
	--75	100,00	86,92	100,00
	-80	100,00	96,92	100,00
	--85	100,00	100,00	92,94
7	--70	99,71	49,23	98,82
	--75	100,00	79,23	98,82
	-80	100,00	95,38	98,82
	--85	100,00	96,92	95,29

Assim como HMM discreto o parâmetro que mais interfere no rendimento final do sistema com HMM contínuo é o valor limiar.

Deve-se tomar o cuidado ao escolher o valor limiar, para baixos valores o sistema torna mais rigoroso, aumenta a taxa de acerto do locutor não autorizado (C3), no entanto o sistema não reconhece o locutor autorizado (C2), resultando em baixas taxas de acerto (ver **Figuras 3-3** e **3-4**).

Para valores altos do limiar, o sistema torna-se menos rigoroso, aceitando locutores não autorizados como autorizado, resultando em baixas taxas de acerto (ver **Figuras 3-3** e **3-4**). Então, uma solução para encontrar o valor ideal limiar, foi

definir de forma heurística um valor limiar para cada um dos cincos HMM, com o objetivo de otimizar sistema.

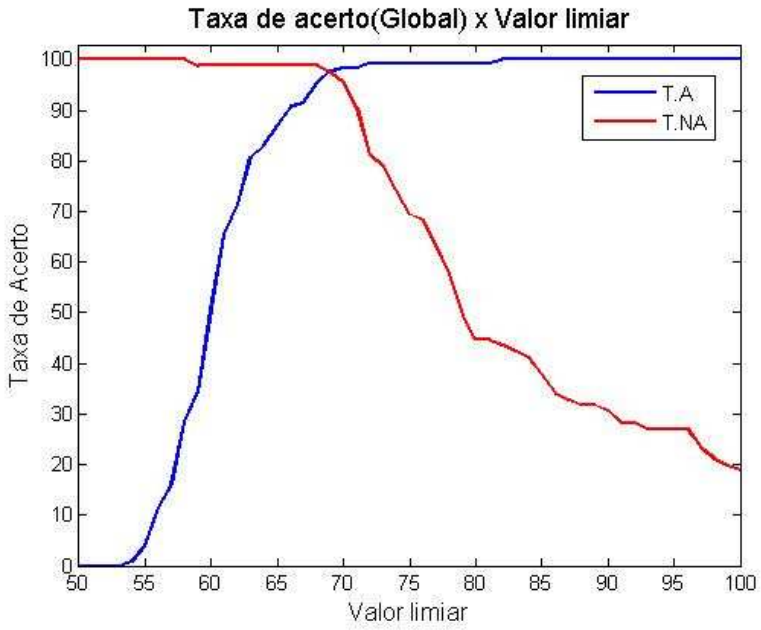


Figura 3-3 - Taxa de Acerto Global x Valor Limiar, T.A = Taxa de Acerto do Locutor Autorizado e T.AN = Taxa de Acerto do locutor não-Autorizado, modelo utilizado foi de 16 CMCs e Delta-CMCs e 4 gaussianas por estados.

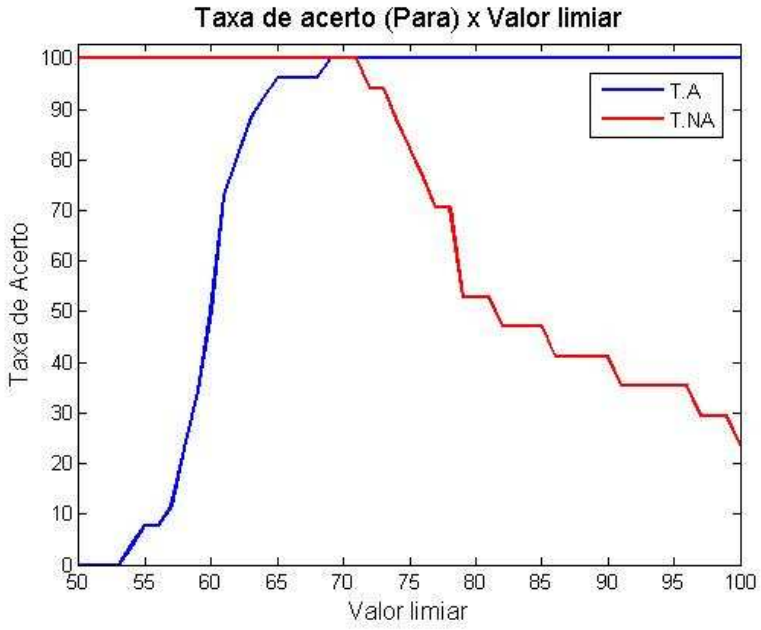


Figura 3-4- Taxa de Acerto (Para) x Valor Limiar, T .A = Taxa de Acerto do Locutor Autorizado e T.AN = Taxa de Acerto do locutor não-Autorizado, modelo utilizado foi de 16 CMCs e Delta-CMCs e 4 gaussianas por estados.

Nas **Tabelas 3-10** e **3-11**, podem ser observados os resultados do sistema para valores limiares diferentes para cada HMM do sistema. Obtendo resultados melhores do que os modelos anteriores (ver **Tabela 3-6** a **3-9**).

Tabela 3-10 – HMM contínuo, melhores resultados para 13 CMCs e Delta-CMCs e diferentes valores limiares

Número de coeficientes: 13 Mel e Mel-Delta								
Número de Gaussiana	Valor limiar					Taxa de Acerto %		
	Frente	Trás	Direita	Esquerda	Para	C1	C2	C3
1	-71	-68,5	-70	-67,5	-70	100,0	99,2	97,64
4	-71	-68	-70	-67,5	-70	100,0	98,46	98,82
5	-70	-70	-70	-67,5	-70	100,0	100,0	98,82

Tabela 3-11- HMM contínuo, melhores resultados para 16 CMCs e Delta-CMCs e diferentes valores limiares

Número de coeficientes: 16 Mel e Mel-Delta								
Número de Gaussiana	Valor limiar					Taxa de Acerto %		
	Frente	Trás	Direita	Esquerda	Para	C1	C2	C3
1	-85,5	-85,0	-86,5	-87,5	-80,0	99,71	99,23	95,29
4	-81,0	-80,0	-81,0	-82,5	-80,0	100,0	99,23	97,64
4	-82	-80	-84	-83	-80	100,0	100,0	98,82
7	-80	-80	-81	-80	-80	100,0	96,15	98,82

4 Conclusão

Os resultados no contexto geral alcançaram-se taxas de acerto acima de 90%. Destacando-se os modelos das **Tabelas 3-10** e **3-11.** O modelo que apresentou os melhore resultado, obteve uma taxa de acerto de 99,53% para classificação do locutor (média ponderada da taxa de acerto de C2 e C3) e de 100,00% para classificação do comando (taxa acerto de C2).

Deve-se ressaltar que o sistema implementado executou duas tarefas: reconhecimento da fala e a verificação do locutor. Para a tarefa de reconhecimento de fala, o HMM é chamado de “Estado da Arte” pelo seu alto desempenho em sistemas práticos (OLIVEIRA e MORITA, 2005). Porém, sistemas práticos utilizando HMM para a tarefa de verificação do locutor não apresentam altas taxas de acerto.

Nota-se que os resultados apresentados nesse trabalho foram obtidos em um ambiente de alta relação sinal/ruído, sendo superior a 25 dB, e as gravações obedeceram a certo padrão, quanto à distância do microfone e a utilização dos mesmos instrumentos para todas as gravações. Em outras condições de gravação não é garantido os mesmos resultados.

Uma importante análise dos resultados é quanto ao tamanho do *codebook* utilizado no HMM discreto, o aumento do tamanho do *codebook* produz uma melhora nas taxas de classificação do locutor, isso pode ser justificado pelo fato que o aumento do *codebook* diminui o erro da quantização.

Outra análise é quanto ao número de estado, o aumento desse na maioria das vezes, aumentou o desempenho do sistema. No entanto, a partir de um número máximo de estados o desempenho do sistema manteve-se o mesmo, isso ocorre porque a partir desse número o próprio treinamento exclui o excesso de estados, atribuindo a matriz de transição de estado ($\mathbf{A}$) valores baixos.

Com utilização dos Coeficientes Deltas, melhorou-se a taxa de acerto de locutor do sistema, isso porque a voz de cada locutor é caracterizada também pela forma que esse pronuncia, essas características são melhores representadas pelos coeficientes delta, que detectam as variações bruscas do espectro da voz.

Quanto aos tipos de Modelos Ocultos de Markov, o HMM discreto apresentou taxas elevadas na classificação do comando e um menor esforço computacional, porém apresentou taxas baixas na classificação de locutor, sendo assim recomendável em sistemas com o objetivo de reconhecimento de comando.

Os sistemas com HMM contínuo alcançaram um desempenho alto na classificação de locutor e de comando, no entanto seu custo computacional é grande.

Para aplicação do reconhecedor, algumas implementações deveriam ser feitas com o objetivo de torna o sistema mais robusto, tais como a utilização dos CMCs normalizados e o escalonamento do banco de filtros utilizados no calculo dos CMCs, melhorando assim o desempenho do sistema no reconhecimento do locutor.

5 Bibliografia

PETRY, A., BARONE, D. A. C. **Sistema para Controle de Elevadores por Voz.**Instituto de Informática, Universidade Federal do Rio Grande do Sul – UFRGS – 2000.

PARREIRA, W. D. **Reconhecimento de Locutor pela Voz usando o Classificador Polinomial e Quantização Vetorial.** Faculdade de Engenharia Elétrica, Universidade Federal de Uberlândia – UFU,2005.

IBGE(2005) Web site do Instituto Brasileiro de Geografia e Estatística, disponível em http://www.ibge.gov.br, [Acesso:21/11/2007]

MAFRA, A. T. **Reconhecimento Automático de Locutor em Modo Independente de Texto por Self-Organizing Maps.** Dissertação de Mestrado. Escola Politécnica da Universidade de São Paulo – 2002.

OLIVEIRA, L. E. S., MORITA, M. E. **Introdução aos Modelos Escondidos de Markov (HMM)**. Pontifícia Universidade Católica do Paraná – PUC-Pr – 2005.

CHOU, W., JUANG, B.H. **Pattern recognition in speech and language processing**, CRC Press, 2003.

MARTINS, J. A. **Avaliação de Diferentes Técnicas para Reconhecimento de Fala**. Tese de doutorado, FEEC/UNICAMP, Campinas, Dezembro 1997.

THE MATHWORKS INC. **Signal Processing Toolbox, User's Guide for use with Matlab** – 2002.

BRANDÃO, A .S., et al. **Redes Neurais Artificiais Aplicadas ao Reconhecimento de Comando de Voz.** VII Internacional Conference On Industrial Applications, 2006.

NIQUINI, F. M. M. **Reconhecimento de Comandos de Voz com Verificação de Locutores e Vocabulários Restrito Utilizando Redes Neurais Artificiais**. UFV, Viçosa, 2007.

RABINER, L. R, SCHAFER, R. W. **Digital Processing of Speech Signal,** Prentice-Hall, Englawood Clifts, New Jersey, 1978.

HAYKIN, S., VEEN, B. V. **Sinais e Sistemas**, trad. José Carlos Barbosa dos Santos. Bookman, Porto Alegre,2001.

PICONE, J. W. **Signal Modeling Techniques in Speech Recognition**, Proceedings of the IEEE, vol. 81, no. 9, Setembro 1993, 1215—1247.

DELLER,J., PROAKIS, J. G., HANSEN, J. H. L. **Discrete-time Processing of Speech Coding**, Prentice Hall, 1987.

DIAS, R. S. **Normalização de locutor em sistema de reconhecimento de fala**, Campinas, SP, 2000.

RABINER, L. R, JUANG, B. H. **Fundamentals of Speech Recognition**. Prentice Hall. 1993.

GERSHO, A. R. **Vector quantization and Signal Compression**. Kluwer. 1993

BAUM, L. E., PETRIE, T. **Statistical Inference for Probabilistic Functions of Finite State Markov Chains**, Ann Math, pp 1554-1563, 1966.

BAUM, L. E., EAGON, J. A., **An Inequality with Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to a Model for Ecology**, Bull. Amer. Math. Soc. 73, pp. 360-363, 1967.

BAKER, J. K., **Stochastic Modeling as a Means of Automatic Speech Recognition**, PhD. Dissertação, Carriegie-Mellon University, 1975.

BAUM, L. E., **An Inequality and Associated Maximisation Technique in Statistical Estimation for Probabilistic Functions of a Markov Process**, Inequalities III, pp.1-8, 1972.

YACOUBI , A., SABOURIN, R., GILLOUX, M., SUEN, C. Y. **Off-line Handwritten Word Recognition using Hidden Markov Modesl**, Knowledge Techniques in Character Recognition, CRC Press LLC, 1999.

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

LOU, H.L. **Implementing the Viterbi Algorithm**, IEEE Signal Processing Magazine, pp. 42-52, 1995.

RABINER, L. R., **A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition**, Procedings of the IEEE, Vol. 77, n. 2, Feb. 1989.

LEVINSON, S. E. et alii. **An Introduction to the Application Theory of Probabilistic Functions of a Markov Process to Automatic** Speech Recognition, Bell system Tech. J., 1983, 62(4), pp. 1035-1074.

JUANG, B. H. et alii. **Maximum Likelihood Estimation for Multivariate Mixture Observations of Markov Chains**, IEEE Trans. Informations Theory, Março 1986, vol. IT-32, nº 2, pp. 307-309

EVANDRO, D. S. P. **Reconhecimento de Locutores Utilizando Modelos de Markov Escondidos Contínuos**, Instituto Militar de Engenharia –IME, Rio de Janeiro, 1997.

MILLER, I., FREUND, J. E. **Probability and Statistics for Engineers.** Englewood Clis, 1985.

CHIVIATO, A. G., el al. **Sistema de Reconhecimento de Fala com Ruídos para Automóveis**, 20ª Feira Tecnológica do INATEL –FETIN, 2001.

RABINER, L. R., SAMBUR, M. R. **An Algorithm for Determining the Endpoints of Isolated Utterances**. The Bell System Technical Journal, pag. 297-315, Fevereiro 1975.

ROJVANIT, R. **Migrating Applications to USB from RS-232 UART with Minimal Impact on PC Software.** Microchip Technology Inc., 2004.

ANEXO A. Tabelas do HMM discreto

Tabelas em anexo referentes aos resultados para HMM discreto (**capítulo 3.2**). Nas **Tabelas A-1** a **A-6** foram utilizado apenas o CMCs.

Tabela A- 1 HMM Discreto, valor limiar: -10, números CMCs: 13

Limiar = -10/ números de CMCs = 13				
Tamanho do *codebook*	Número de estados	Taxa de Acerto % (C1)	Taxa de Acerto %(C2)	Taxa de Acerto %(C3)
64	5	99,71	93,83	11,11
	10	100,00	91,53	43,33
	15	100,00	90,00	34,44
	20	100,00	93,86	41,11
128	5	100,00	84,61	36,67
	10	100,00	79,23	64,44
	15	100,00	81,53	66,67
	20	100,00	78,38	74,44
256	5	100,00	71,53	73,33
	10	100,00	66,92	75,55
	15	100,00	73,84	78,88
	20	100,00	63,07	90,00

Tabela A- 2 HMM Discreto, valor limiar: -30, números CMCs: 13

Limiar = -30/ números de CMCs = 13				
Tamanho do codebook	Número de estados	Taxa de Acerto % (C1)	Taxa de Acerto %(C2)	Taxa de Acerto %(C3)
64	5	99,71	95,38	7,77
	10	100,00	97,69	13,33
	15	100,00	98,46	31,11
	20	100,00	99,23	24,44
128	5	100,00	96,15	18,88
	10	100,00	97,69	36,67
	15	100,00	97,69	38,88
	20	100,00	98,46	38,88
256	5	100,00	95,38	38,88
	10	100,00	94,61	57,77
	15	100,00	96,15	67,77
	20	100,00	90,00	70,00

Tabela A- 3 HMM Discreto, valor limiar: -50, números CMCs: 13

Limiar = -50/números de CMCs = 13				
Tamanho do *codebook*	Número de estados	Taxa de Acerto % (C1)	Taxa de Acerto %(C2)	Taxa de Acerto %(C3)
64	5	99,71	99,38	3,33
	10	100,00	97,69	5,55
	15	100,00	100,00	8,88
	20	100,00	100,00	13,33
128	5	100,00	96,82	15,15
	10	100,00	99,23	17,77
	15	100,00	100,00	37,77
	20	100,00	100,00	31,1
256	5	100,00	98,46	23,33
	10	100,00	100,00	33,33
	15	100,00	99,23	32,22
	20	100,00	97,69	52,22

Tabela A- 4 HMM Discreto, valor limiar: -10, números CMCs: 16

Limiar = -10/ números de CMCs = 16				
Tamanho do *codebook*	Número de estados	Taxa de Acerto % (C1)	Taxa de Acerto %(C2)	Taxa de Acerto %(C3)
64	5	100,00	94,61	11,76
	10	100,00	93,07	29,41
	15	100,00	95,38	38,82
	20	100,0	93,84	45,88
128	5	100,00	96,92	34,11
	10	100,00	86,15	57,64
	15	100,00	83,84	63,52
	20	100,00	86,15	71,76
256	5	100,00	78,38	74,11
	10	100,00	70,76	81,17
	15	100,00	62,30	85,88
	20	100,00	55,38	89,41

Tabela A- 5 HMM Discreto, valor limiar: -30, números CMCs: 16

Limiar = -30/ números de CMCs = 16				
Tamanho do *codebook*	Número de estados	Taxa de Acerto % (C1)	Taxa de Acerto %(C2)	Taxa de Acerto %(C3)
64	5	100,00	97,69	7,05
	10	100,00	96,92	14,11
	15	100,00	99,23	21,17
	20	100,00	99,23	21,17
128	5	100,00	98,46	17,64
	10	100,00	96,92	23,52
	15	100,00	98,46	37,64
	20	100,00	98,46	41,17
256	5	100,00	94,61	51,76
	10	100,00	97,30	55,29
	15	100,00	94,61	61,17
	20	100,00	92,30	70,58

Tabela A- 6 HMM Discreto, valor limiar: -50, números CMCs: 16

Limiar = -50/ números de CMCs = 16				
Tamanho do *codebook*	Número de estados	Taxa de Acerto % (C1)	Taxa de Acerto %(C2)	Taxa de Acerto %(C3)
64	5	100,00	97,69	4,70
	10	100,00	99,23	7,05
	15	100,00	99,23	10,58
	20	100,00	99,23	12,94
128	5	100,00	98,14	14,11
	10	100,00	98,46	20,00
	15	100,00	100,00	30,58
	20	100,00	100,00	25,88
256	5	100,00	99,23	32,94
	10	100,00	99,23	41,17
	15	100,00	99,23	48,23
	20	100,00	99,23	50,58

Nas **Tabelas de A-7** a **A-12** foram utilizados junto com os CMCs os coeficientes Delta-CMCs.

Tabela A- 7 HMM Discreto, valor limiar: -10, números CMCs e Delta-CMCs: 13

Limiar = -10/ números de CMCs e Delta-CMCs= 13				
Tamanho do *codebook*	Número de estados	Taxa de Acerto % (C1)	Taxa de Acerto %(C2)	Taxa de Acerto %(C3)
64	5	96,00	93,07	65,88
	10	93,71	91,53	72,94
	15	93,85	90,76	72,94
	20	90,57	84,61	76,47
128	5	91,14	76,92	88,23
	10	90,85	73,84	89,41
	15	89,42	74,61	90,58
	20	88,00	76,15	92,94
256	5	84,25	59,23	96,47
	10	83,42	54,61	100,00
	15	83,42	60,76	98,82
	20	81,71	56,92	100,00

Tabela A- 8 HMM Discreto, valor limiar: -30, números CMCs e Delta-CMCs: 13

Limiar = -30/ números de CMCs e Delta-CMCs= 13				
Tamanho do *codebook*	Número de estados	Taxa de Acerto % (C1)	Taxa de Acerto %(C2)	Taxa de Acerto %(C3)
64	5	98,57	97,69	36,47
	10	97,71	97,69	48,23
	15	98,00	95,39	74,05
	20	96,85	96,92	54,11
128	5	96,28	93,07	64,41
	10	96,00	93,07	72,94
	15	96,57	91,53	76,47
	20	95,14	93,07	87,05
256	5	92,00	90,00	91,76
	10	93,14	81,53	94,11
	15	91,14	84,61	95,29
	20	89,71	83,84	96,47

Tabela A- 9 HMM Discreto, valor limiar: -50, números CMCs e Delta-CMCs: 13

Limiar = -50/ números de CMCs e Delta-CMCs= 13				
Tamanho do *codebook*	Número de estados	Taxa de Acerto % (C1)	Taxa de Acerto %(C2)	Taxa de Acerto %(C3)
64	5	99,14	96,92	20,00
	10	99,42	98,46	35,29
	15	99,71	98,46	35,29
	20	99,42	100,00	37,64
128	5	98,00	97,69	50,58
	10	98,28	93,07	54,47
	15	97,14	96,15	61,17
	20	96,57	96,82	67,05
256	5	96,85	96,15	78,82
	10	96,28	95,38	88,23
	15	95,42	93,07	87,05
	20	95,14	93,84	88,23

Tabela A- 10 HMM Discreto, valor limiar: -10, números CMCs e Delta-CMCs: 16

Limiar = -10/ números de CMCs e Delta-CMCs= 16				
Tamanho do *codebook*	Número de estados	Taxa de Acerto % (C1)	Taxa de Acerto %(C2)	Taxa de Acerto %(C3)
64	5	95,14	91,53	63,52
	10	93,42	90,00	70,58
	15	93,71	86,15	71,76
	20	92,28	81,53	74,11
128	5	91,71	76,92	89,41
	10	88,85	82,30	91,76
	15	90,00	75,38	91,76
	20	88,00	73,84	95,29
256	5	81,71	50	96,47
	10	78,00	62,30	97,64
	15	78,57	57,69	100,00
	20	77,45	56,15	100,00

Tabela A- 11 HMM Discreto, valor limiar: -30, números CMCs e Delta-CMCs: 16

Limiar = -30/ números de CMCs e Delta-CMCs= 16				
Tamanho do *codebook*	Número de estados	Taxa de Acerto % (C1)	Taxa de Acerto %(C2)	Taxa de Acerto %(C3)
64	5	98,28	97,69	36,47
	10	99,14	98,46	47,05
	15	98,57	96,15	50,58
	20	97,42	99,23	56,47
128	5	95,42	95,38	70,58
	10	95,15	95,38	75,29
	15	94,57	96,15	80,00
	20	95,42	92,30	80,00
256	5	88,57	87,69	88,23
	10	96,57	90,00	92,94
	15	88,57	81,53	92,94
	20	87,71	80,79	94,11

Tabela A- 12 HMM Discreto, valor limiar: -50, números CMCs e Delta-CMCs: 16

Limiar = -50/ números de CMCs e Delta-CMCs= 16				
Tamanho do *codebook*	Número de estados	Taxa de Acerto % (C1)	Taxa de Acerto %(C2)	Taxa de Acerto %(C3)
64	5	99,14	98,46	20,00
	10	100,00	99,46	32,94
	15	99,71	99,23	36,47
	20	99,14	99,23	42,35
128	5	98,28	96,15	56,47
	10	98,00	96,15	57,64
	15	96,85	98,46	63,52
	20	97,71	96,92	65,88
256	5	94,85	94,61	81,17
	10	94,00	95,38	85,88
	15	94,00	91,53	84,70
	20	93,71	90,00	85,88

ANEXO B. Software para aplicação

Desenvolveu-se o seguinte *software* (ver **Figura B-1**)

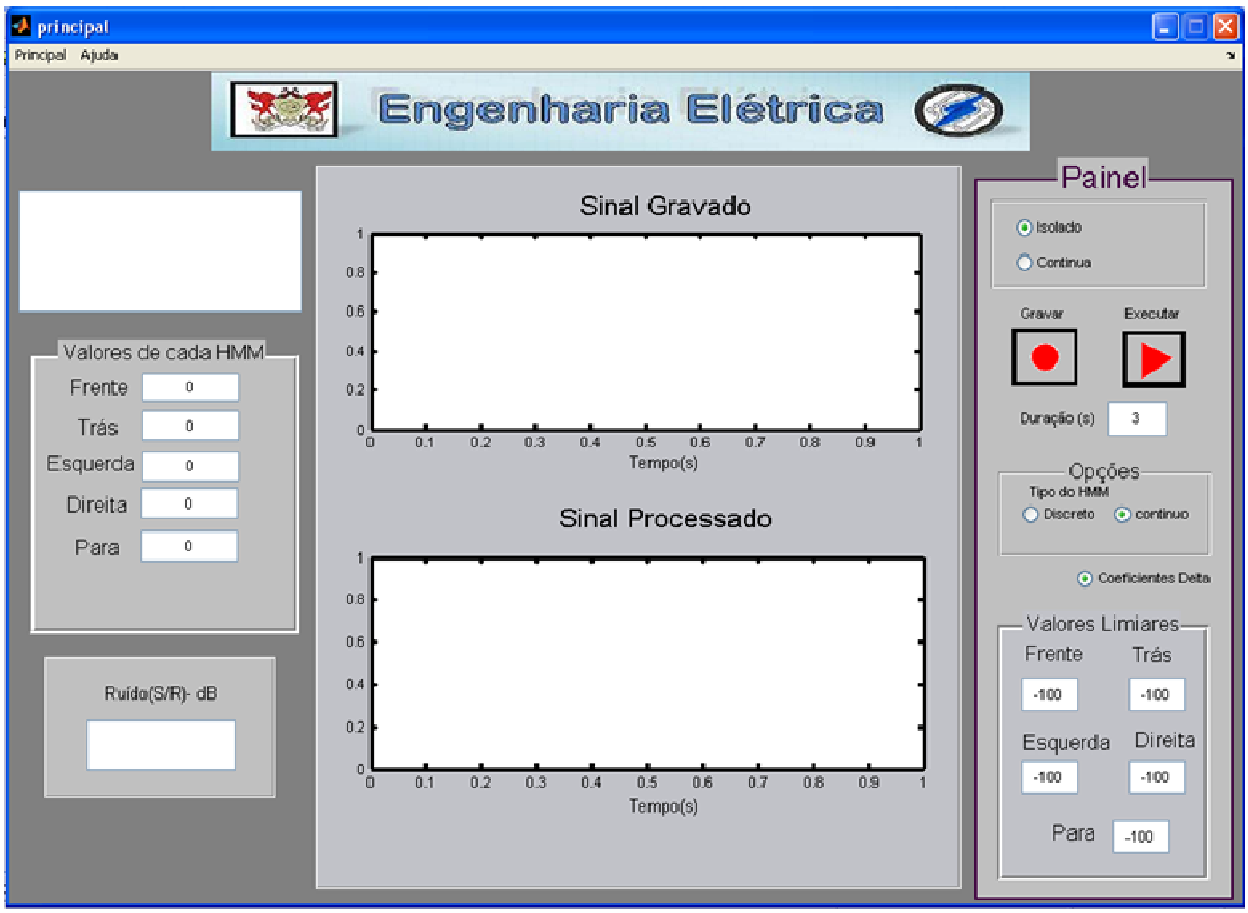
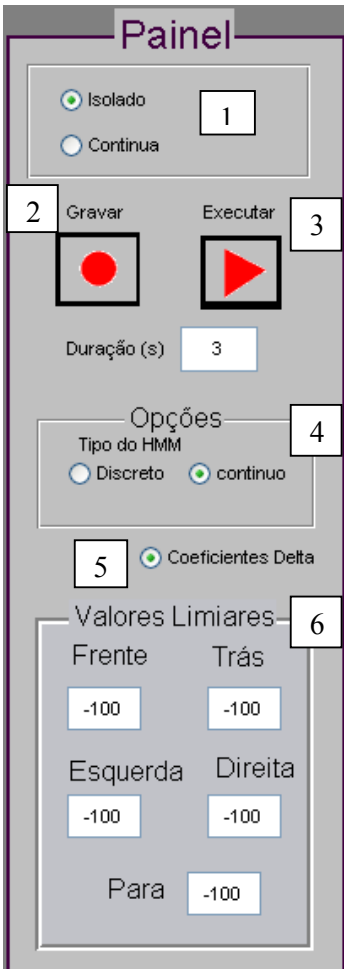


Figura B- 1 *Sofware* desenvolvido

O aplicativo apresenta na sua parte direita um painel, onde pode-se escolher o tipo de HMM, os valores limiares, iniciar o processo de reconhecimento. Esse painel será explicado a seguir.

Painel



1) Opção do tipo de gravação:pode-se escolher dois tipos de gravação:
Isolado: A gravação possui um tempo de duração.
Continua: A gravação é realizada em de forma contínua.

2) Gravar: inicia o processo de gravação. Após o termino realiza o reconhecimento automático do locutor e do comando.

3) Executar: Escuta a gravação (somente para gravação Isolada)

4) Opção do HMM: Pode escolher a opção do HMM discreto ou contínuo

5) Coeficientes Delta-CMCs: escolher se deseja utilizar os coeficietes Delta-CMCs

6) Valor Limiar: variar o valor limiar para cada HMM

Figura B- 2 Painel do *Software*

Após inicializar o processo de gravação (GRAVAR - **2**) o programa imprime o sinal amostrado em **7** (ver **Figura B-3**) e o sinal após o pré-processamento em **8**. Na janela **9** é mostrado o comando proferido pelo locutor, em **10** a log-probabilidade de cada um dos HMM e em **11** a relação .

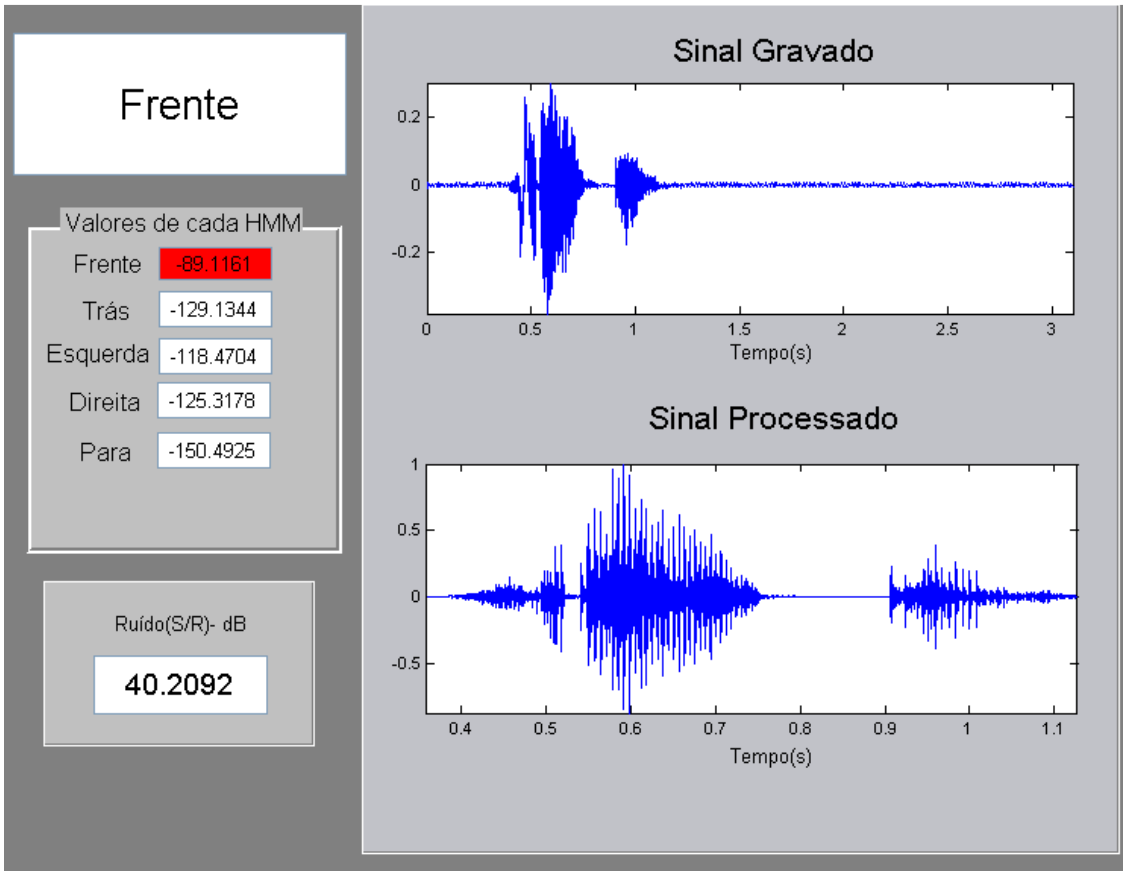


Figura B- 3 Software após a gravação da palavra "Frente"

O programa apresenta interface com a porta paralela (LPT), porta serial (COM) e com a webcam.

Realizaram-se alguns testes com o programa, onde nota-se que os resultados apresentaram altas taxas de reconhecimento de fala, por volta de 90%, e taxas mais baixas para a verificação do locutor, por volta de 60%. Esses valores correspondem quando o programa é executado no computador onde o banco de dados foi gravado.

Em testes realizados em outros computadores, o desempenho do programa diminuiu, no entanto a taxa de acerto do reconhecimento da fala permaneceu alta. Outro fator que influência também o desempenho desse *software* é o ambiente onde realiza as gravações.

Pode-se ser feitas as seguintes melhorias no programa:

- Tornar o sistema de reconhecimento mais robusto.
- A opção Contínua apresenta algumas *bug*, às vezes o *buffer* fica armazenado, causando erros no programa.