MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
SECRETARIA DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA
CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMUNICAÇÕES

AI ALBINO ADRIANO ALVES CORDEIRO JÚNIOR
AI CAIMI FRANCO REIS

SOBRE A COMPENSAÇÃO DE MICROFONE E DE CANAL TELEFÔNICO EM VERIFICAÇÃO AUTOMÁTICA DE LOCUTOR

Rio de Janeiro
2003

INSTITUTO MILITAR DE ENGENHARIA

AI AIBINO ADRIANO ALVES CORDEIRO JÚNIOR
AI CAIMI FRANCO REIS

SOBRE A COMPENSAÇÃO DE MICROFONE E DE CANAL TELEFÔNICO EM VERIFICAÇÃO AUTOMÁTICA DE LOCUTOR

Dissertação de Projeto de Fim de Curso apresentada ao Curso de Graduação em Engenharia de Comunicações do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de graduação em Engenharia de Comunicações.

Orientador: TC José Antonio Apolinário Júnior

Co-orientador: Cel R/1 Roberto Miscow Filho

Rio de Janeiro
2003

INSTITUTO MILITAR DE ENGENHARIA

AI ALBINO ADRIANO ALVES CORDEIRO JÚNIOR
AI CAIMI FRANCO REIS

SOBRE A COMPENSAÇÃO DE MICROFONE E DE CANAL TELEFÔNICO EM VERIFICAÇÃO AUTOMÁTICA DE LOCUTOR

Dissertação de Projeto de Fim de Curso apresentada ao Curso de Graduação em Engenharia de Comunicações do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de graduação em Engenharia de Comunicações.

Orientador: TC José Antonio Apolinário Júnior – D.Sc.
Co-orientador : Cel R/1 Roberto Miscow Filho – M.C

Aprovada em XV de Novembro de 2003 pela seguinte Banca Examinadora:

_____
TC José Antonio Apolinário Júnior

_____
Cel R/1 Roberto Miscow Filho

Rio de Janeiro
2003

AGRADECIMENTOS

Agradecemos aos nossos pais Albino Cordeiro e Rosa Maria, Felisberto Reis e Maria Lúcia, que tanto nos apoiaram para vencermos mais uma importante etapa de nossas vidas.

Aos nossos orientadores TC Apolinário e Cel Miscow pela atenção com que nos orientaram, bem como ao ex-aluno César Medina e ao servidor civil Bomfim pelo apoio técnico que nos prestaram.

RESUMO

O bom desempenho de um sistema de Verificação Automática de Locutor depende essencialmente da compensação de três fatores: ruído, canal e microfone. Este trabalho visa fundamentalmente a estudar com mais detalhes as técnicas utilizadas para a compensação de microfones e de um canal telefônico fixo, quando aplicadas a um sistema de Verificação Automática de Locutor. Foi apresentado um Modelo Geral de Compensação aplicado à língua portuguesa. Para o teste de algumas técnicas de compensação foi desenvolvida uma base de dados de sinais de voz (corpus) chamada IME2003, que foi levantada em ambiente acusticamente isolado utilizando equipamentos de alta qualidade. As possíveis aplicações para o modelo apresentado seriam os acessos autenticados por voz a bancos, casas inteligentes, auxílios em investigações policiais e intranet's entre outros.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

1 INTRODUÇÃO

1.1 INTRODUÇÃO

Um dos sistemas mais complexos do ser humano é o sistema vocal. Constituído pelos pulmões, pelo trato vocal, pela boca e pelo nariz, seu funcionamento depende fundamentalmente de um estímulo do sistema nervoso, e qualquer alteração em alguma de suas partes modifica diretamente a voz humana.

A importância do estudo do sistema vocal está no fato de a voz carregar consigo informações sobre várias características de uma pessoa, tais como identidade, sexo, idioma e, possivelmente, condições físicas e emocionais. No entanto, quando o sinal de voz passa por um determinado canal, algumas perturbações são inseridas pelo mesmo no sinal de voz, impedindo a extração imediata das informações pessoais do locutor. Portanto, de um modo geral, o conteúdo de informação da voz de um locutor, após passar por algum tipo de canal, como por exemplo o canal telefônico, é composto pelas características da pessoa, pela sua frase falada, pelas suas emoções, pelo ruído, e pelas transformações do canal, que influenciam sobremaneira a extração das características do locutor.

Na prática, é fundamental uma adequada extração dessas informações para diversas aplicações, tais como síntese de voz, tradução automática e reconhecimento de locutor. E, dentro dessa perspectiva, o processamento digital de sinais de voz ganha importância ao trazer soluções mais eficientes para a extração e para a modelagem das características de uma pessoa dentro do processo de reconhecimento.

Geralmente, a extração das informações de um sinal de voz para o reconhecimento automático pode ser classificada em (SILVA, 2001) :

- Reconhecimento Automático de Voz (RAV) é o processo de extrair a mensagem contida no sinal de voz;
- Reconhecimento Automático de Locutor (RAL) é o nome atribuído ao processo de extrair a identidade do locutor; e
- Reconhecimento Automático de Idioma (RAI) que é a identificação do idioma ou do dialeto falado pelo locutor.

Para o caso de reconhecimento automático de locutor, que foi a maior ênfase deste trabalho, a principal tarefa consiste em extrair e separar as características dependentes do locutor de todas as outras inseridas pelo canal.

O reconhecimento de locutor pode ser dividido em três etapas. A primeira etapa consiste em extrair, a partir de janelas temporais do sinal de voz, as características dependentes do locutor. Para isso, são utilizados três métodos principais, que melhor refletem a identidade do locutor. Esses métodos estão listados abaixo:

- Coeficientes mel-cepstrum;
- Coeficientes cepstrum de predição liner (LPC);
- Coeficientes cepstrum da Transformada de Fourier Discreta (DFT);

Nesta primeira etapa, o que se procura é tentar minimizar a influência de qualquer meio que atrapalhe a correta extração das informações das pessoas. São os canais os principais óbices da extração, porque podem inserir não-linearidades e, como isso, impedir o uso de técnicas para a compensação. Atualmente, a técnica mais utilizada para compensação do canal é o *Cepstral Mean Subtraction* (CMS), que só pode ser aplicada em canais convolutivos, sendo o canal telefônico o que melhor atende a estas condições. Por outro lado, os microfones utilizados para amostrar ou transmitir a voz de uma pessoa não são padronizados, e alguns tipos existentes também podem inserir não-linearidades que prejudicam enormemente o reconhecimento.

O segundo passo é usar um modelo estatístico que represente a forma do conjunto de características escolhidas. Os modelos estatísticos mais utilizados para construir este conjunto são:

- Quantização Vetorial;
- Gaussian Mixture Models (GMM);
- Hidden Markov Models (HMM);
- Arquiteturas de Redes Neurais;

A terceira etapa, que é a etapa da decisão, é onde se compara a voz de entrada com o modelo de locutor especificado e toma-se a decisão sobre a identidade do locutor.

Os sistemas de reconhecimento de locutor são agrupados dentro de duas categorias, conhecidas como verificação e identificação de locutor. Verificação de locutor é o mesmo que determinar, através de uma amostra de voz, se uma pessoa é realmente quem ela diz ser. Por outro lado, a identificação de locutor determina, dentro de um grupo de vozes conhecidas, qual a que mais se aproxima da amostra de entrada.

As aplicações do reconhecimento de locutor têm aumentado significativamente nos últimos anos (CAMPBELL, 1997). Dentre tais aplicações, podemos destacar (LIMA, 2001):

- Controle de acesso: a dispositivos, redes de trabalho e dados;
- Autenticação de transações comerciais como ferramenta para prevenção de fraudes: compras por telefone com cartão de crédito, transações na internet e operações bancárias;
- Segurança pública: monitoração em penitenciárias, aplicações forenses;
- Auxílio a deficientes físicos;
- Uso militar: informações que requeiram identificação de locutor.

1.2 OBJETIVO DO PROJETO

O desempenho de um sistema de reconhecimento é afetado por diversos fatores tais como o canal, o microfone utilizado pelo locutor, o ruído e as diferentes características emocionais (alegre, deprimido, etc) e físicas (cansado, doente, etc) do locutor. Sendo assim, é bastante importante que o sistema seja indiferente às influências provocadas aos sinais de voz pelos meios externos ao locutor. Com base nessa condição necessária, o objetivo deste Projeto Final de Curso foi apresentar um modelo modificado do CMS para a língua portuguesa e também um modelo para a compensação das cápsulas de telefone mais comuns, de modo a minimizar a influência do microfone e do canal telefônico na verificação automática de locutor.

1.3 CONTRIBUIÇÃO DESTE PROJETO

Este projeto apresentou como principal contribuição a proposta de junção de dois tipos de compensações em um único sistema de reconhecimento dependente da língua: a compensação do canal através do CMS modificado para a língua portuguesa e a compensação dos microfones através de duas técnicas: HM (Handset Mapping) e HNORM (Handset Normalization).

Vale ressaltar também que houve um avanço na aplicação das técnicas descritas acima no sentido de que, para algumas delas, foi utilizado um banco de vozes limpas, obtido em uma câmara acusticamente isolada.

1.4 ORGANIZAÇÃO DO PROJETO

O projeto foi divido em 6 capítulos. O Capítulo 2 apresenta os conceitos básicos de reconhecimento automático de locutor e enfatiza a Verificação Automática de Locutor - o tipo de reconhecimento mais utilizado neste trabalho. Em seguida, ele aborda os principais conceitos relacionados ao GMM e aos cepstra mais utilizados. O Capítulo 3 trata da compensação de canal telefônico, dando uma ênfase no CMS modificado para o português. Já o Capítulo 4 faz uma abordagem sobre os métodos mais recentes para a compensação de microfones. O Capítulo 5 reúne as técnicas mais importantes citadas nos capítulos anteriores e apresenta um modelo integrado para compensação de microfone e para compensação de canal telefônico aplicado à língua portuguesa. Finalmente, o último capítulo engloba as principais conclusões obtidas com o trabalho e sugere as possíveis pesquisas que podem ser desenvolvidas com base neste trabalho.

2 O RECONHECIMENTO AUTOMÁTICO DE LOCUTOR

2.1 INTRODUÇÃO

O reconhecimento de locutor nada mais é do que a capacidade de reconhecer alguma pessoa por meio de sua voz. Quando o reconhecimento é feito por um computador, ele é conhecido como reconhecimento automático de locutor (RAL). Segundo (ATAL, 1976), existem evidências sugerindo que o reconhecimento realizado pelo dispositivo eletrônico convencional pode chegar a superar o reconhecimento feito por uma pessoa. Daí o interesse pelo estudo do RAL.

De um modo geral, a extração das informações de um sinal de voz para reconhecimento automático pode ser classificada em (RABINER, 1978) :

- reconhecimento automático de voz (RAV): é o processo de se extrair a mensagem contida no sinal de voz;
- reconhecimento automático de locutor (RAL): é o processo de se extrair a identidade do locutor; e
- reconhecimento automático de idioma (RAI): é a identificação do idioma ou do dialeto falado pelo locutor.

Cada um dos tipos de reconhecimentos pode ser esquematizado conforme aparece na FIG. 2.1. Neste trabalho foi dada uma maior ênfase ao Reconhecimento automático de locutor.

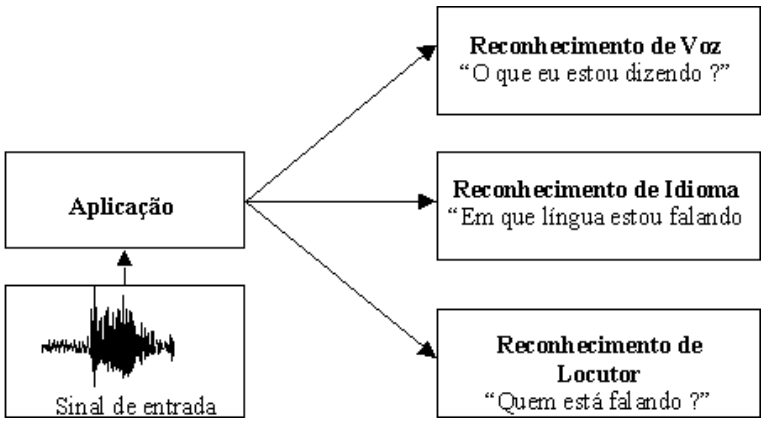

Figura 2.1 – Os diversos tipos de reconhecimentos automáticos

2.2 O RECONHECIMENTO AUTOMÁTICO DE LOCUTOR

O RAL é a tarefa de discriminar pessoas pelas características do seu sinal de voz. Ele pode ser classificado segundo a tarefa a ser executada, segundo o texto pronunciado ou segundo o grau de cooperação na fala dos locutores, conforme pode ser explicado abaixo:

a.Quanto à Tarefa

- Identificação;
- Verificação.

A identificação é a tarefa de identificar uma pessoa por meio de sua voz em um conjunto conhecido de N locutores, denominado fechado quando envolver N decisões ou aberto quando houver (N+1) decisões (decide-se também, se a voz pertence a algum dos componentes do conjunto ou a nenhum deles) (REYNOLDS, 1992). O desempenho de um sistema de identificação é degradado pelo aumento do número de locutores, pois é aumentado o número possível de decisões (JAYANT, 1990). Resumindo: a função deste processo é responder à seguinte pergunta: “De quem é essa determinada voz?”. A FIG. 2.2 ilustra a identificação de locutor.

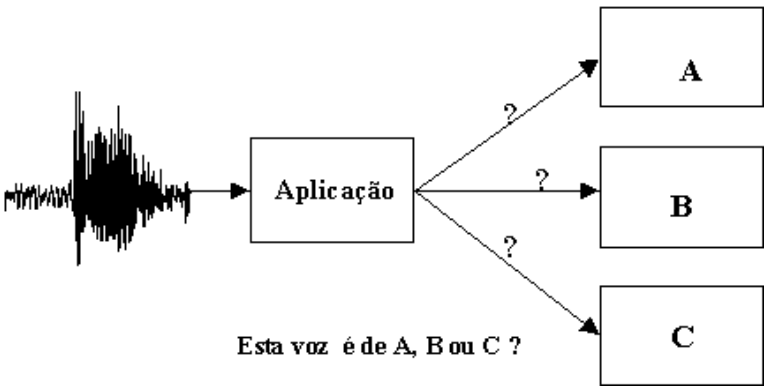

Figura 2.2 – Exemplo de identificação de locutor

A verificação é a tarefa em que se procura constatar se uma dada voz (locução) pertence ou não a uma determinada pessoa. Envolve, portanto, uma decisão binária. A decisão é feita no denominado conjunto aberto de locutores (REYNOLDS, 1995), porque o reconhecimento é efetuado em um grupo de locutores desconhecidos (possíveis impostores). Considerando-se um bom projeto do sistema, com estatística suficiente, o desempenho é independente do número de locutores (JAYANT, 1990). Resumindo, a função deste processo é responder à seguinte pergunta: “Esta voz é de fulano?”. Um esquema deste tipo de reconhecimento pode ser visto na FIG. 2.3.

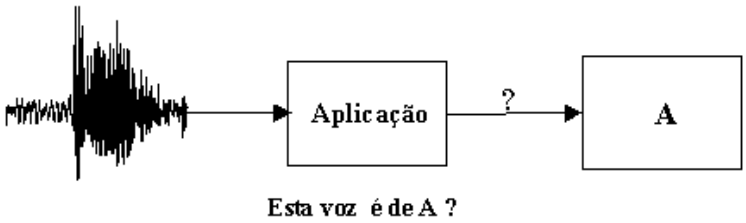

Figura 2.3 – Exemplo de verificação de locutor

b.Quanto ao Texto

- Dependente do Texto;
- Independente do Texto.

Quanto à dependência do texto, o reconhecimento pode ser dependente ou independente. Os sistemas que exigem uma fala pré-determinada são dependentes do texto. Esses sistemas podem fazer comparações precisas e confiáveis entre duas locuções com o mesmo texto, em ambientes foneticamente similares e exigindo de 2 a 3 segundos de voz para treinamento e teste (JAYANT, 1990). Em sistemas independentes do texto, tais comparações não são fáceis de serem obtidas, o que torna o desempenho de tais sistemas inferiores aos sistemas dependentes do texto. Para que se obtenha uma estatística razoável do sinal, geralmente são exigidos de 10 a 30 segundos de voz para treinamento e teste (JAYANT, 1990). Trabalhos recentes têm utilizado de 1 a 2 minutos para treinamento e de 3 a 30 segundos para teste, com qualidade variável do sinal de voz (MARTIN, 2000).

c.Quanto à Cooperação

Quanto à cooperação (SOUSA, 1996), os locutores podem ser:

- Cooperativos: sabem que estão sendo reconhecidos; é o caso quando pronunciam palavras específicas ou não, mas falam de forma clara para ajudar o sistema de reconhecimento;
- Não-cooperativos: não sabem que estão sendo reconhecidos; por isso sua fala não é condicionada para ajudar o sistema de reconhecimento.

2.3 UM SISTEMA BÁSICO DE RAL

Uma aproximação geral de um sistema de reconhecimento de locutor, consiste basicamente em três etapas: aquisição do sinal de voz, extração das características do

sinal de voz pertinentes ao reconhecimento, modelagem estatística e, por último, sistema classificador. A FIG 2.4 esquematiza um sistema básico de reconhecimento de locutor.

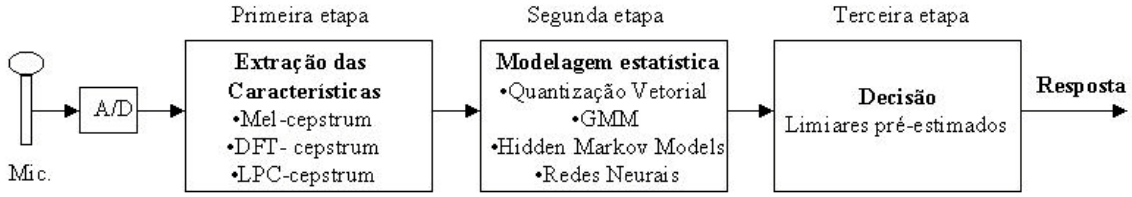

Figura 2.4 - Esquema com as etapas de um sistema de reconhecimento de locutor

Na aquisição do sinal, a voz (pressão acústica) é transformada em um sinal elétrico (analógico) pelo microfone, e, em seguida, esse é convertido para um sinal digital por um conversor analógico/digital. Esse sinal é então pré-processado, para supressão de informações desnecessárias e/ou ênfase nas importantes (RABINER, 1978). Finalmente, as características desejadas são extraídas constituindo vetores de características, que gerarão os padrões para o classificador. Este, após treinamento, terá como objetivo separar as classes distintas. Na identificação de locutor, todos os modelos treinados são avaliados com uma locução de teste. O modelo que apresentar o melhor resultado é aceito como *verdadeiro*, ou seja, a locução de teste é tida como pertencente ao locutor cujo modelo venceu. Na verificação, o modelo de um pretenso locutor determinará se a locução de teste pertence ou não ao locutor. Tal modelo pode ser comparado com outro, que corresponde ao universo de locutores falsos e/ou ruído. A decisão será dada com base em limiares pré-estimados, de acordo com os valores apresentados pelo classificador.

2.4 OS PRINCIPAIS TIPOS DE CARACTERÍSTICAS CEPSTRAIS

2.4.1 COEFICIENTES DFT-CEPSTRUM

O cepstrum DFT é obtido pela DFT inversa (*Discrete Fourier Transform*) do logaritmo do módulo -desprezando-se a fase - da DFT de um sinal. Quando aplicado a estudos de RAL é importante aplicar as transformações anteriores às janelas do sinal. A FIG. 2.5 mostra a seqüência de passos necessários para a obtenção do cepstrum DFT.

O janelamento mostrado na FIG. 2.5 consiste de um processo de extração de um certo número de amostras do sinal e a multiplicação dessas por uma função janela, para depois serem analisadas. O janelamento é realizado devido ao fato de que a voz é um processo não-estacionário, mas em intervalos pequenos, de 20*ms* ou até 40*ms,* pode ser tratada como um sinal localmente estacionário. Isto ocorre pois o trato vocal muda de forma lentamente com o passar do tempo (OPPENHEIM, 1989). Vale ressaltar que ele é importante para a maioria dos processos de extração de características de voz.

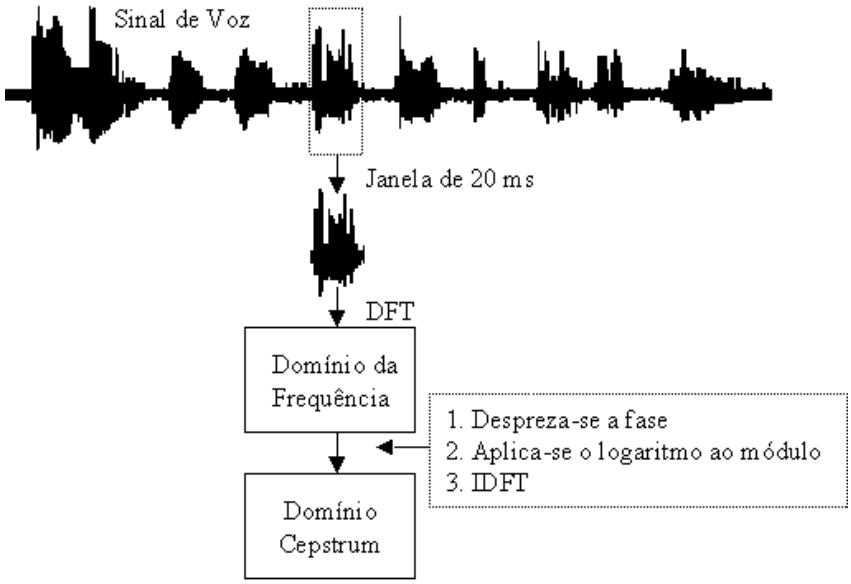

Figura 2.5 – Os passos para obter o cepstrum DFT

2.4.2 COEFICIENTES MEL

É uma das características mais utilizadas no reconhecimento de locutor (REYNOLDS, 1994). Os coeficientes mel-cepstra (mel-cepstrum coefficients - MCC), são obtidos de um sistema que aproxima a resposta em freqüência do ouvido humano, do qual são extraídos os coeficientes cepstrais (PICONE, 1991). Presume-se que o sucesso no reconhecimento de locutor é devido à capacidade que o MCC tem em capturar as diferenças interlocutor, pois carrega informações sobre o trato vocal em conjunto com características da percepção auditiva (SHARMA, 1999). O processo para obtenção desses coeficientes pode ser visto na FIG. 2.6, em que DCT é a transformada cosseno discreta, e o banco de filtro pode ser mais bem observado na

FIG. 2.7, que mostra os filtros de banda crítica para o cálculo de um sinal x(n) amostrado a 8kHz.

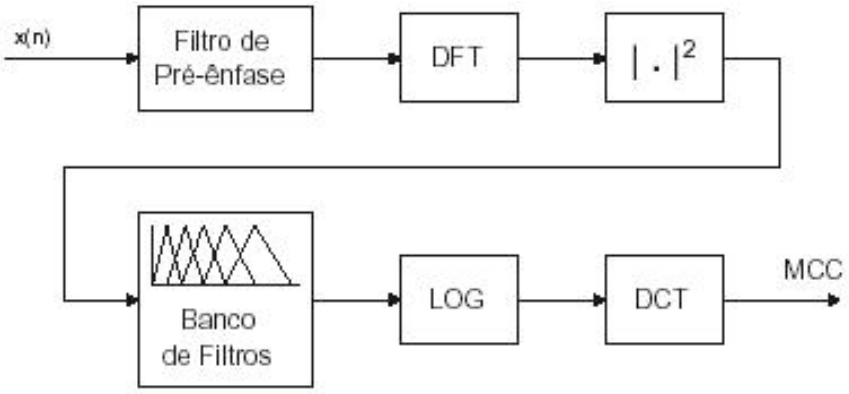

Figura 2.6 – Diagrama em blocos para a extração dos coeficientes mel-cepstrum (LIMA, 2001)

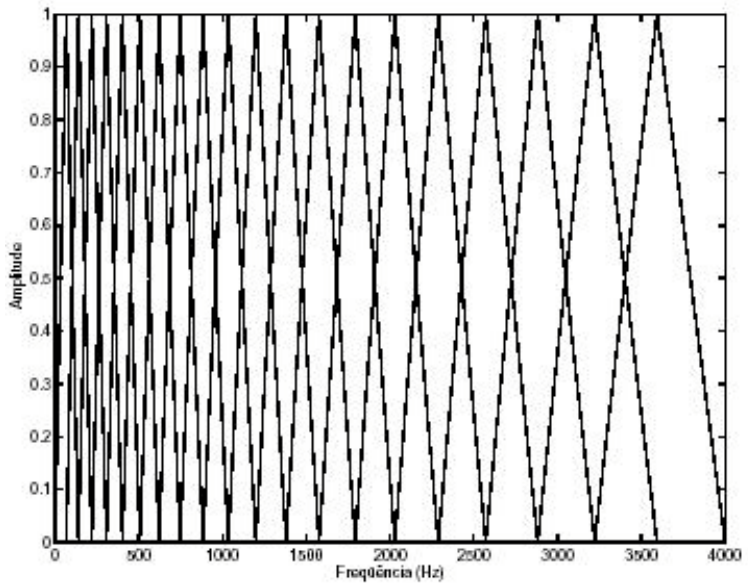

Figura 2.7 – Magnitude do espectro dos filtros de banda larga utilizados na produção dos coeficientes mel-cepstrais (LIMA, 2001)

2.4.3 COEFICIENTES LPC

Coeficientes de predição linear (*Linear Prediction Coefficientes* - LPC): a idéia principal da predição linear é a de que uma amostra de voz pode ser aproximada por

uma combinação linear de amostras passadas. O modelo é estimado por um filtro digital (pólos e/ou zeros) que simula o trato vocal e nasal. A ordem do filtro é escolhida de tal modo que, se ele for excitado por impulsos, sua saída apresente um espectro muito próximo do espectro do sinal que está sendo modulado (RABINER, 1978).

2.5 VERIFICAÇÃO AUTOMÁTICA DE LOCUTOR COM GMM

Dentre as possíveis formas de modelagens estatísticas presentes na segunda etapa do esquema de verificação de locutor está a abordagem por GMM (Gaussian Mixture Models), conforme o esquema da FIG. 2.4. Como vantagens deste método, podem-se destacar: o fato de que há um baixo custo computacional, é baseado em um modelo estatístico bem estudado e é insensível aos aspectos temporais da fala, modelando apenas a distribuição das observações acústicas de um locutor (REYNOLDS, 2000). A eficácia dos sistemas baseados em GMM vem sendo comprovada, particularmente em verificação independente do texto, no evento anual promovido pelo NIST (*National Institute of Standards and Technology,* americano) denominado SER (*Speaker Recognition Evaluation*), no qual os referidos sistemas obtêm os melhores resultados desde 1996.

A seguir, será feita uma explanação resumida dos princípios de um sistema de verificação automático de locutor fundamentado em GMM.

2.5.1 Razão de Verossimilhanças Logarítmica

Em um problema de verificação, partindo de um trecho gravado de fala Y, desejamos verificar se Y foi pronunciado por um locutor hipotético de quem possuímos gravado um trecho de fala S. Nesse caso, teremos as seguintes possibilidades:

- H_o: Y foi falado pelo locutor hipotético da fala S;
- H_1: Y não foi falado pelo locutor hipotético da fala S.

Primeiramente, dados dois eventos A e B, definimos verossimilhança como o valor da função densidade de probabilidade p(A|B) quando B é a variável independente da função.

Assim, uma excelente forma de se decidir por uma das hipóteses (H_o ou H_1) é através da razão de verossimilhança:

$$L = \frac{p(Y \mid H_0)}{p(Y \mid H_1)} \tag{2.1}$$

onde, aceita-se H_o caso $L \geq \theta$ e rejeita-se H_o se $L < \theta$. Onde θ é o limiar (*threshold*) de decisão e é estimado em laboratório, empiricamente e sob condições controladas.

Na prática, para o cálculo de $p(Y|H_o)$ e $p(Y|H_1)$, o segmento de voz Y é substituído pela seqüência $X = \{x_1,..., x_T\}$ de vetores cepstrais x_t, indexados no tempo discretizado t = 1,..., T, adquiridos na primeira etapa como ilustrado na FIG. 2.4, e a hipótese H_o é substituída por um modelo matemático λ_{hipo} que representa a locução S no espaço cepstral de x. A hipótese H_1 é representada pelo modelo $\overline{\lambda}_{hipo}$.

A razão de verossimilhanças fica então $p(X \mid \lambda_{hipo}) \big/ p(X \mid \overline{\lambda}_{hipo})$; no entanto, usualmente utiliza-se a razão de verossimilhanças logarítmica, ou seja :

$$\Lambda(X) = \log p(X \mid \lambda_{hipo}) - \log p(X \mid \overline{\lambda}_{hipo}) \tag{2.2}$$

O modelo λ_{hipo} é estimado utilizando-se a locução de treinamento S. O modelo $\overline{\lambda}_{hipo}$ é estimado a partir de um background como treinamento, ou seja, reúne-se uma grande quantidade de locuções de locutores distintos e treina-se um único modelo, o UBM (*Universal Background Model*) λ_{bkg} (REYNOLDS, 2000).

2.5.2 Modelos de Mistura Gaussiana – GMM (Gaussian Mixture Models)

GMM's são na verdade as funções escolhidas para estimar $p(X|\lambda)$. Esta escolha foi feita pela primeira vez em (REYNOLDS, 1992). Normalmente, pode-se supor que os vetores de X são independentes, logo:

$$\log p(X \mid \lambda) = \sum_{t=1}^{T} \log p(x_t, \lambda) \tag{2.3}$$

Considerando-se um vetor cepstral x com dimensão D, onde D é o número de janelas utilizado no processo de extração das características (vide Seção. 2.3), tem-se:

$$p(x \mid \lambda) = \sum_{i=1}^{M} \omega_i p_i(x) \tag{2.4}$$

ou seja, a densidade é uma combinação linear, ou uma soma ponderada, de M densidades Gaussianas unimodais $p_i(x)$, onde cada uma destas é parametrizada por um vetor de médias, μ_i,e uma matriz de covariâncias, Σ_i de acordo com:

$$p_i(x) = \frac{1}{(2\pi)^{D/2}\left|\sum_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'(\sum_i)^{-1}(x-\mu_i)\right\} \tag{2.5}$$

Para os pesos ω_i existe ainda a condição $\sum_{i=1}^{M} \omega_i = 0$. Desse modo, define-se, então, o modelo $\lambda = \{\omega_i, \mu_i, \sum_i\}$, onde i=1,...,M.

Dado um conjunto de vetores de treinamento, os parâmetros do modelo são estimados usando o algoritmo iterativo de maximização da esperança (ME). Este método é descrito em (DEMPSTER, 1977). Na verdade, os parâmetros de um modelo GMM vão sendo refinados iterativamente de modo que, para k interações e para k+1, $p(X|\lambda^{(k+1)}) > p(X|\lambda^{(k)})$. Geralmente, apenas cinco iterações são necessárias para a convergência dos parâmetros do modelo (REYNOLDS, 2000).

3 COMPENSAÇÃO DE CANAL

3.1. INTRODUÇÃO

O desempenho de sistemas de reconhecimento automático de locutor é significativamente degradado por descasamentos acústicos entre as condições de treinamento e teste. Tais descasamentos são comumente encontrados em sistemas que aplicam a voz sobre redes telefônicas, onde os diferentes microfones e as diferentes rotas impõem distorções convolutivas variadas no sinal de fala (GARCIA, 1999).

Além dos microfones diferentes, efeitos do canal englobam outros fatores tais como o ambiente acústico (isto é, escritório, carro, etc.) e meios de transmissão (linhas de telefones fixos, linhas celulares, VoIP). Uma vez que as informações do canal e do locutor estão concentradas no espectro, qualquer coisa que modifique o espectro pode acrescentar dificuldades.

Técnicas de compensação para os efeitos do canal são aplicadas em geral em três domínios: domínio cepstral, domínio da razão de verossimilhança, e domínio do modelo estatístico. Na entrada, compensações no domínio *cepstral* são as mais indicadas para remover os efeitos do canal nos vetores cepstrais, antes de se modelar o treinamento e a verificação. Essas compensações incluem técnicas bem conhecidas e amplamente utilizadas tais como *cepstral mean subtraction* (CMS), filtragem RASTA e subtração espectral. Sendo que o *Cepstral Mean Subtraction* (CMS) é o método mais recente e mais popular empregado para melhorar os efeitos de variabilidade de canal em sistemas de reconhecimento de fala e locutor (GARCIA, 1999). Já na saída, compensações no domínio da razão de verossimilhança são necessárias devido às mudanças causadas pelas variações nas condições do canal. Exemplos de compensações nesse domínio são as técnicas *Hnorm* e Mapeamento Não-linear. Em compensações no domínio do modelo estatístico, o objetivo é modificar os modelos de verificação para minimizar os efeitos da variabilidade dos canais (REYNOLDS, 2003).

Neste trabalho foram estudas apenas as técnicas Hnorm, Mapeamento Não-linear e CMS.

3.2 CMS e CMS modificado

Sabendo-se que um canal telefônico é convolutivo pode-se dizer que o sinal de saída é a convolução entre o sinal de entrada e a resposta do canal ao impulso. Em seguida, pode-se aplicar o logaritmo no módulo da DFT e, depois, fazendo a IDFT e obtendo os coeficientes cepstrais, pode-se obter a sua média ao longo do tempo (janelas). De acordo com (MAMMONE, 1996), se o sinal de voz é balanceado em termos de sons vozeados, não-vozeados e explosivos, a média cepstral tende para zero, para o caso dos trabalhos científicos feitos para a língua inglesa. Assim, basta conhecer as características do canal para se compensar o seu efeito e obter o sinal limpo.

Para o idioma português, no entanto, a média cepstral não é zero (SILVA, 2001); surge assim, a necessidade de se conhecer a média cepstral para a língua de modo a fazer uma compensação mais adequada. A média cepstral é obtida através da base CORPUS IME2003, que está detalhada no APÊNDICE A.

Os passos anteriores podem ser descritos matematicamente da seguinte maneira:

Supondo um sinal de entrada s(t) passando por um canal convolutivo h(t) e com uma saída y(t), conforme a FIG. 3.1, tem-se:

$$y(t) = s(t) + h(t) \tag{3.1}$$

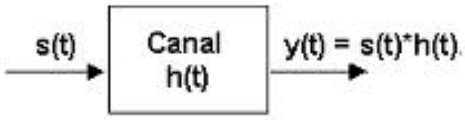

Figura 3.1 – Canal convolutivo apresentado como exemplo hipotético

Aplicando a Transformada de Fourier sobre o módulo, com o desprezo da fase, a equação poderá ser reescrita como:

$$|Y(f)| = |S(f)| + |H(f)| \tag{3.2}$$

Tomando o logaritmo neperiano da EQ. (3.2), tem-se a EQ.(3.3):

$$\ln|Y(f)| = \ln|S(f)| + \ln|H(f)| \tag{3.3}$$

Aplicando a função Inversa da Transformada de Fourier têm-se as três funções no domínio cepstral, conforme pode ser observado na EQ.(3.4):

$$\overset{\Lambda}{y}(t) = \overset{\Lambda}{s}(t) + \overset{\Lambda}{h}(t) \tag{3.4}$$

Tomando-se o valor médio para o vetor, pode-se escrever

$$\overset{\Lambda}{y}_m(t) = \overset{\Lambda}{s}_m(t) + \overset{\Lambda}{h}_m(t) \tag{3.5}$$

Na EQ. (3.5) vale observar que, como o canal é considerado invariante no tempo por ser um canal de telefone fixo, $\overset{\Lambda}{h}(t)$ é considerado igual ao $\overset{\Lambda}{h}_m(t)$. Assim, de acordo com o que foi explicado no início desta seção, ao assumir que a média cepstral tende a zero para a língua inglesa, isto é, $s_m(t) \approx 0$, a EQ. (3.5) pode ser reescrita conforme a equação abaixo:

$$\overset{\Lambda}{y}_m(t) \approx \overset{\Lambda}{h}(t) \tag{3.6}$$

Agora basta subtrair o canal da EQ. (3.4):

$$\overset{\Lambda}{y}(t) - \overset{\Lambda}{h}(t) \approx \overset{\Lambda}{s}(t) \tag{3.7}$$

Portanto, o sinal está compensado.

Já para o caso da língua portuguesa não se podem garantir as mesmas condições e, neste caso, para uma melhor compensação é necessário calcular, através da CORPUS IME2003, a média cepstral para a língua portuguesa $\overset{\Lambda}{s}_m(t)$. Assim, não basta efetuar as mesmas operações anteriores na EQ. (3.5), pois $\overset{\Lambda}{s}_m(t) \neq 0$. Mas com posse do valor da média cepstral para a língua, pode-se fazer:

$$\overset{\Lambda}{y}_m(t) - \overset{\Lambda}{s}_m(t) = \overset{\Lambda}{h}(t) \tag{3.8}$$

Logo, substituindo a EQ.(3.8) na EQ.(3.4), vem:

$$\overset{\Lambda}{y}(t) - \overset{\Lambda}{h}(t) \approx \overset{\Lambda}{s}(t) \tag{3.9}$$

Portanto, o sinal de entrada poderá, assim, ser compensado, pois o canal é conhecido e a média cepstral pode ser calculada.

4 COMPENSAÇÃO DE MICROFONES

4.1 INTRODUÇÃO

A degradação das respostas dos sistemas de verificação automática de locutor (VAL), associada às não-linearidades de certos microfones, deve-se, de acordo com os experimentos descritos em (REYNOLDS, 1997), às situações em que o sinal de voz de teste tenha sido adquirido com um microfone de resposta não-linear e o sinal de treinamento tenha sido obtido com um microfone de resposta linear ou com resposta não-linear, mas, estranha à resposta do microfone de teste e vice-versa.

Tanto nos experimentos com base de dados em inglês –descritos no referido trabalho do Reynolds– quanto em experimento realizado com a base de dados de voz em português do laboratório de voz do IME, observou-se total ineficácia do CMS (*Cepstral Mean Subtraction*) na compensação dos efeitos dos microfones. Entende-se que isso se deu pois o CMS assume que o canal seja convolutivo, portanto linear; logo, torna-se uma ferramenta inútil neste problema posto que a resposta não-linear é uma característica de certos microfones.

De fato, no experimento com o IME2002 (vide APÊNDICE A) a utilização do CMS chega a deteriorar ainda mais –em relação à não utilização do CMS- os percentuais de acerto do sistema como se pode ver na curva DET da FIG. 4.1, que mostra o resultado de um experimento realizado no laboratório de voz do IME com o intuito de estudar a performance do CMS Modificado. Considerando o fato de que um experimento semelhante (SILVA, 2001) foi realizado antes, no qual, a diferença principal foi que neste a base de dados havia sido gravada utilizando-se apenas um único microfone de eletreto, e o CMS, assim como o CMS Modificado, obteve resultado positivo. Assim, o presente resultado explica-se pela variabilidade de microfones entre gravações.

Dado que as principais motivações atuais para as pesquisas em tecnologias de VAL são as aplicações imersas em sistemas telefônicos, tais como tele-marketing, tele-atendimento, escuta telefônica para combate ao crime, dentre outros, fica evidente a importância do estudo dos efeitos e possíveis paliativos para as respostas não-lineares dos microfones.

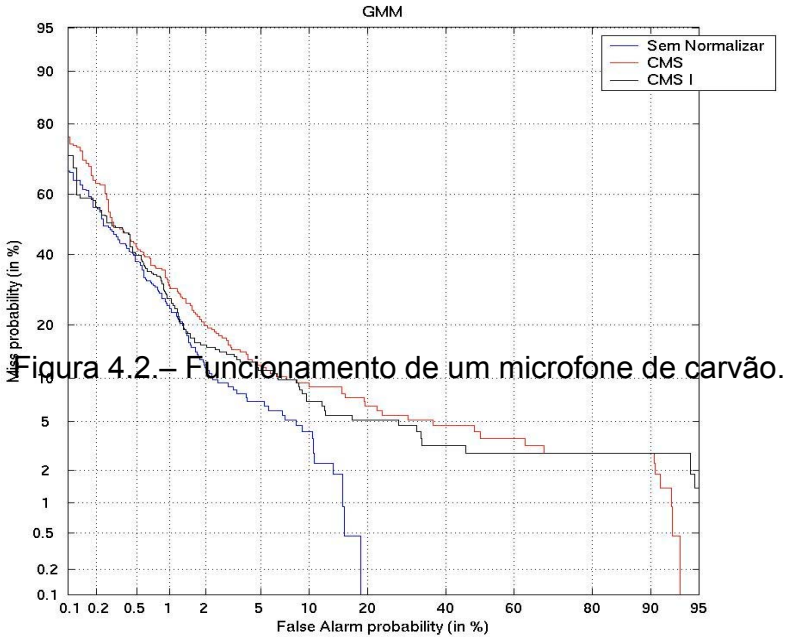

Figura 4.2.– Funcionamento de um microfone de carvão.

Figura 4.1 – Curva DET com resultado de experimento com a IME2002

Nesta seção, foi descrita uma metodologia e as principais técnicas levantadas para o tratamento desse problema peculiarmente importante.

4.2 PRINCIPAIS TIPOS DE MICROFONES TELEFÔNICOS

Existem vários tipos de microfones. O microfone mais antigo mas ainda bastante utilizado é o **microfone de carvão**. A FIG. 4.2 mostra o funcionamento de um microfone de carvão.

A voz do usuário provoca variações na pressão do ar que atua sobre uma membrana de alumínio. Essa pressão variável modifica a resistência ôhmica entre os pontos de contacto da cápsula. A eficiência do microfone depende muito da aplicação da tensão correta na cápsula. Uma tensão baixa pode acarretar uma transmissão ruim, e uma tensão alta pode provocar a queima dos grânulos de carvão.

Outro tipo de microfone largamente utilizado em aparelhos telefônicos é o **microfone de eletreto** (vide FIG. 4.3).

O eletreto é um material dielétrico utilizado para armazenar carga elétrica quase que indefinidamente. Quando o eletreto é colocado como dielétrico entre as duas placas de metais, forma um tipo especial de capacitor. A relação ente a tensão (V), a carga (Q) e a capacitância (C) é dada por

$$V = \frac{Q}{C} \tag{4.1}$$

A carga Q armazenada no dielétrico é mantida praticamente constante por causa do eletreto. O pequeno movimento do diafragma de metal devido à ação do sinal sonoro, acarreta pequenas variações na capacitância do capacitor, fazendo com que haja variações na tensão V. Essas variações de tensão são pequenas e devem ser amplificadas. O sinal de tensão captado, portanto, é proporcional ao sinal sonoro que excita o diafragma.

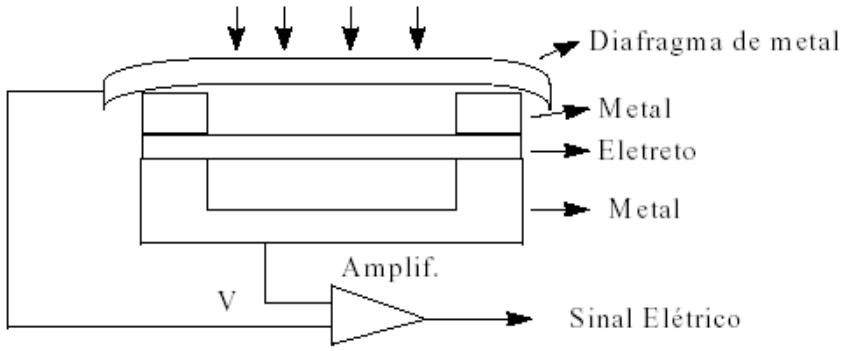

Figura 4.3. Esquema de funcionamento do microfone de eletreto.

Os efeitos não-lineares do microfone de eletreto são reconhecidamente menores quando comparados com os da não-linearidade do microfone de carvão. E isso pode ser facilmente comprovado na prática, utilizando-se um banco de vozes como o HTIMIT (REYNOLDS, 1997) que é gravação de uma parte das vozes do banco TIMIT através de diferentes tipos de microfones de carvão e eletreto.

Um outro tipo de microfone utilizado em telefone, ainda que escassamente, é o **microfone eletromagnético**. O diagrama da FIG. 4.4 mostra o esquema de funcionamento de um microfone eletromagnético.

A pressão acústica ocasiona o movimento da bobina. O movimento da bobina imersa no campo magnético induz uma corrente proporcional a esse movimento. Essa corrente é de pouca intensidade e necessita ser amplificada.

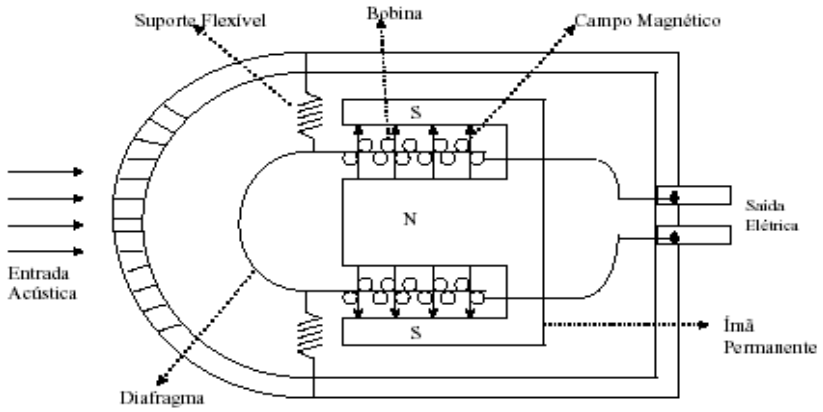

Figura 4.4: Esquema de um microfone eletromagnético.

Os microfones eletromagnéticos têm sido utilizados por muitos anos em equipamentos de comunicações de campanhas militares pelo Exército Brasileiro.

Vale ressaltar que, de acordo com o documento do ITU-T *Recomendation P.62*, enquanto a distorção não linear dos receptores telefônicos é em geral desprezível, microfones (e particularmente o microfone de carbono do tipo usualmente utilizado em alguns aparelhos telefônicos comerciais) mostram uma considerável não-linearidade, ou seja, a relação entre a variação da resistência do microfone e pressão acústica no diafragma é não linear.

A não-linearidade torna-se mais acentuada quanto mais sensível for o microfone. Ainda no referido texto as seguintes observações são feitas:

- o microfone é menos sensível à pressão acústica menor do que certo valor (limiar de excitação);
- como conseqüências da inércia mecânica dos grânulos de carvão (demora em estabelecer contato elétrico entre os grânulos), os vários estados de agitação do carbono sobre a influência das ondas acústicas não são os mesmos para todas as freqüências (por exemplo, batimentos lentos de dois sons são usualmente amplificados).

A informação existente no efeito final da distorção harmônica na qualidade de uma locução por telefone indica que o efeito da distorção de segunda ordem é consideravelmente menor do que a distorção de terceira ordem.

4.3 UM IDENTIFICADOR AUTOMÁTICO DO TIPO DE MICROFONE

Os métodos para solucionar o problema da variação do transdutor em sistemas RAL atualmente conhecidos têm um passo em comum anterior à compensação: a identificação automática do tipo de microfone ou, mais simplesmente, seleção de microfone. Neste passo, dado um sinal de voz de um locutor qualquer, procura-se identificar qual, mais provavelmente, foi o microfone utilizado para gravá-lo.

Será descrita agora uma técnica de identificação automática de microfone baseada em GMM. Esta técnica pode ser encontrada com alguns melhoramentos em (Mak, 2002).

Basicamente, um trecho de sinal de fala é fornecido a um certo número H de modelos GMM's denominados $\{\Gamma_k\}_{k=1}^{H}$ (um para cada tipo de microfone possível). O mais provável tipo de microfone é selecionado de acordo com

$$k^* = \arg\max_{k=1}^{H} \sum_{t=1}^{T} \log p(y_t \mid \Gamma_k) \tag{4.2}$$

onde $p(y_t \mid \Gamma_k)$ é a densidade de probabilidade para o k-ésimo microfone, y_t é o vetor com o sinal de voz distorcido e T é a ordem usada no treinamento dos GMM's de cada tipo de microfone (MAK, 2002).

Em outras palavras, Γ é o modelo matemático associado a um tipo de microfone e, portanto, sendo H o número de tipos de microfones telefônicos utilizados para o treinamento dos diferentes modelos Γ (GMM) utiliza-se as mesmas falas dos mesmos locutores, entretanto, estes devem ter sido gravados através dos H tipos de microfones.

Como um exemplo de experimento prático, após a compensação linear padrão (CMS), poderia treinar-se um GMM de ordem 1024 com os discursos do corpus gravados através de microfone de eletreto e treinar-se um outro GMM também de ordem 1024 com o mesmo corpus anterior sendo que agora gravado através de microfone de carvão, ou seja, H igual a dois. Números semelhantes foram utilizados na realização dos experimentos em (REYNOLDS, 2003)

Considerando-se que o corpus utilizado tenha uma quantidade razoável de locutores –por volta de 50 indivíduos do mesmo sexo (MEDINA, 2003)– e que os efeitos lineares foram atenuados com o CMS, entende-se que os modelos assim construídos caracterizam os dois tipos de microfones telefônicos principais.

Nesse contexto prático, a implementação da EQ. (4.2) tem obtido acertos na ordem de 98~99% (MAK, 2002), portanto, satisfatórios para maioria das aplicações de verificação atualmente desenvolvidas.

4.4 MODELAGEM DO PROBLEMA DA DISTORÇÃO NÃO-LINEAR

Será exposta, a seguir, uma modelagem do problema. Baseando-se nesta modelagem será, mais adiante, descrita uma primeira forma de compensação da distorção não-linear do microfone: o *mapeamento não-linear*.

Deste ponto em diante partir-se-á da premissa que todo sinal de treinamento é isento de distorções não-lineares, pois, dessa forma pode-se simplificar a organização dos experimentos e, simultaneamente, diminuir o custo computacional dos mesmos.

4.4.1 TEORIA DA DISTORÇÃO NÃO-LINEAR POLINOMIAL EM VOZ

Observa-se, quando se tratando de microfones utilizados em aparelhos telefônicos, que os mais populares são os microfones de carvão e os baseados em eletretos. Será investigado primeiramente o efeito da não-linearidade polinomial na resposta ao impulso de um único trato vocal. A motivação é mostrar que polinômios de ordem finita podem gerar **formantes**[1] **fantasmas**, ou seja, formantes "artificiais" produtos do *batimento* entre os formantes originais os quais ocorrem em múltiplos, somas e diferenças dos formantes originais de maneira consistente com as medidas feitas com transdutores não-lineares.

Considere-se que uma única ressonância $x[n] = r^n \cos(\omega n)$, a resposta ao impulso de um único trato vocal, é passada através de uma não-linearidade polinomial de ordem 3. A saída dessa distorção consiste de dois termos dados por

$$y[n] = 2r^{3n} \cos(\omega n) + 2r^{3n} \cos(3\omega n) \tag{4.3}$$

Observe-se que as ressonâncias foram alargadas, correspondendo ao decaimento mais rápido, e a ressonância do segundo termo, isto é, o formante fantasma, está localizado a três vezes a freqüência da ressonância original.

[1] Freqüências de ressonância do trato vocal.

Este desenvolvimento, baseado na resposta a um único trato vocal, não é exatamente igual ao desenvolvimento para respostas periódicas, – que são mais próximas da fala humana real-; no entanto, é útil, pois a diferença entre as mesmas é mínima para uma grande gama de espécies de não-linearidade (REYNOLDS, 1998).

Para ilustração do efeito da não-linearidade polinomial, fez-se a distorção polinomial de ordem três de uma janela de um sinal de fala real. Neste experimento numérico, realizou-se a gravação da vogal *a* durante alguns segundos utilizando microfone de computador comercial, a digitalização foi feita a uma taxa de amostragem de 22050Hz e resolução de 16 bits. A partir deste sinal gerou-se um segundo elevando-se ao cubo cada amostra do primeiro.

Na FIG. 4.5 pode-se ver o espectro LPC de uma janela do sinal original e então após a distorção, esta mesma janela, com a aparição do formante fantasma logo após o formante principal indicado no gráfico com uma seta.

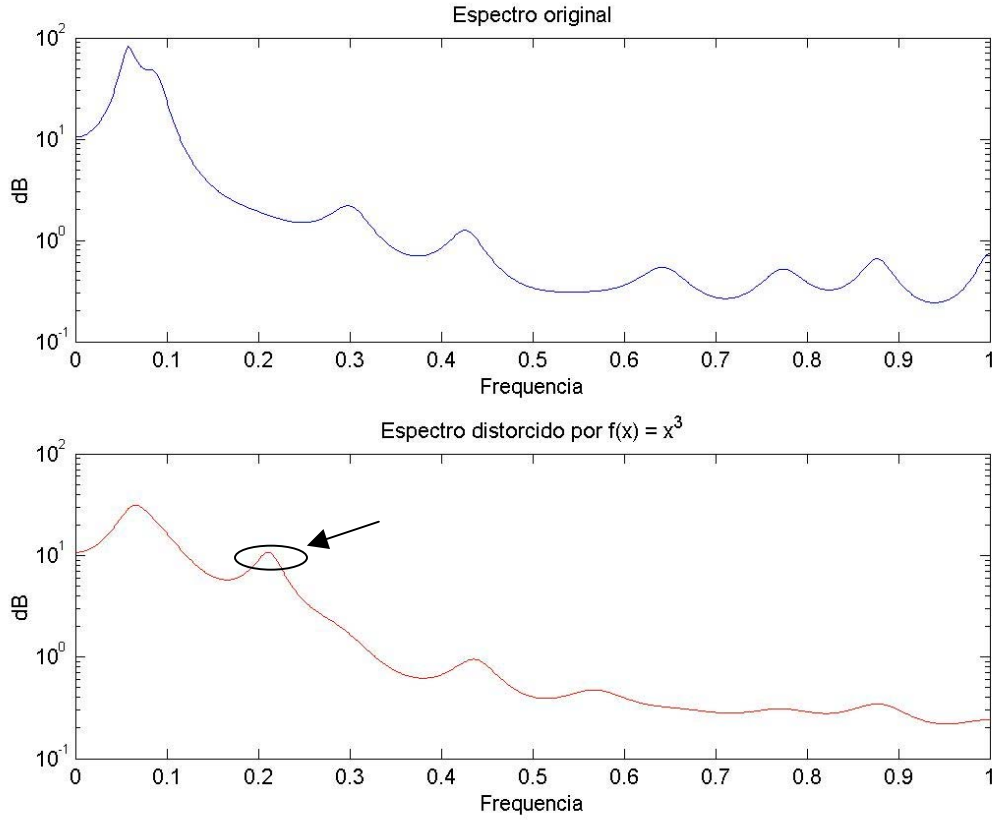

Figura 4.5: Distorção não-linear polinomial.

4.4.2 Modelo de Um Microfone Não-linear

Na prática, microfones telefônicos adicionam distorção espectral linear de magnitude e fase (dispersão). Por esta razão, a não-linearidade fica “prensada” entre um pré-filtro linear e um pós-filtro, linear os quais supõe-se que ambos sejam FIR (REYNOLDS, 1998). Denominamos esta conexão em série de filtros linear, não-linear, linear de *modelo L-N-L* de microfone. A FIG. 4.6 mostra um diagrama de blocos do esquema de um L-N-L, onde x[n] é o sinal puro e y[n] é o mesmo sinal depois de distorcido pelo microfone.

Figura 4.6.– Modelo L-N-L para Microfones**.**

O pós-filtro proporciona uma deformação adicional. A dispersão do pré-filtro proporciona memória para a não-linearidade. Em microfones reais, pode-se notar o aparecimento de efeitos mais complexos de não-linearidade, no entanto, descrever-se-á neste trabalho uma não-linearidade concordante com o modelo polinomial proposto na seção anterior.

Assim, à não-linearidade referente ao bloco central da FIG. 4.6 associaremos um polinômio de ordem finita Q, ao pré-filtro associaremos uma transformação convolucional g[n] e ao pós-filtro, de forma análoga à anterior, teremos h[n]. Portanto, na saída do L-N-L tem-se:

$$y[n] = Q\big(g[n] * x[n]\big) * h[n] \tag{4.4}$$

onde Q é um operador não-linear polinomial de ordem *p* para o qual uma entrada *u* retorna

$$Q(u) = q_o + q_1 u + q_2 u^2 + \ldots + q_p u^p \tag{4.5}$$

onde g[n] é um pré-filtro FIR de ordem M e h[n] um filtro FIR de ordem N.

4.5 COMPENSAÇÃO POR MAPEAMENTO NÃO-LINEAR

Baseando-se no modelo não linear proposto na seção anterior e considerando-se que, para todos os tipos de microfones utilizados para gravação da base de testes, já se dispõe dos polinômios Q_k e das funções h_k e g_k que estimam o comportamento não-linear de cada k-ésimo tipo de microfone, desenvolve-se a técnica de compensação por mapeamento não-linear.

Lembre-se: assumimos o canal de gravação do sinal de treinamento como livre de não-linearidades, logo, há descasamento entre as características do canal de gravação do sinal de teste e as características do canal de gravação do sinal de treinamento quando o sinal de **teste** estiver distorcido não-linearmente.

Em suma, o princípio da compensação por mapeamento não-linear é o de que para melhorar os resultados finais do sistema RAL se atenue o efeito do descasamento por meio de uma distorção não-linear **deliberada** do sinal de treinamento, utilizando-se do modelo L-N-L definido pelos parâmetros $[Q_k, h_k, g_k]$ associados ao microfone mais provavelmente utilizado para gravar o sinal de teste, antes de se realizar o processo de verificação em si. Este método foi primeiramente desenvolvido em (REYNOLDS, 1998).

O diagrama em blocos da FIG. 4.7 ilustra esse processo levando-se em conta que se utilizou o GMM como método de verificação de locutor, onde *y[n]* é o vetor

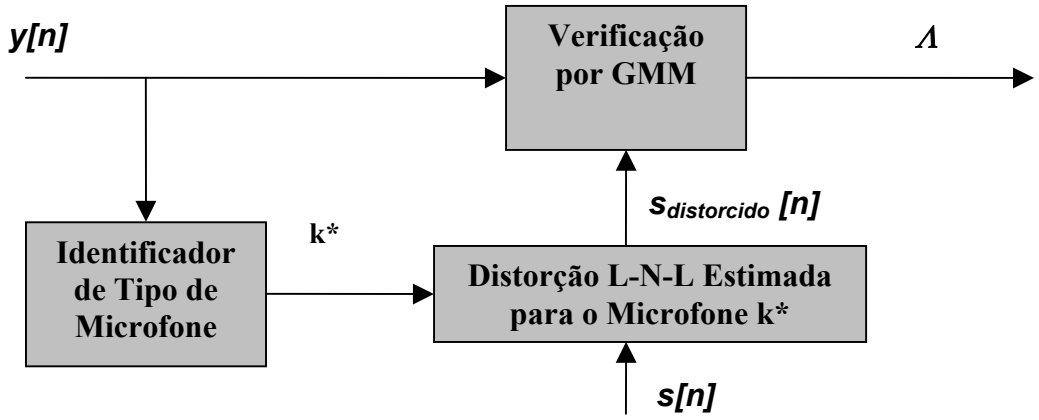

Figura 4.7. Esquema do processo de Compensação por Mapeamento Não-Linear.

característica do sinal de teste distorcido não-linearmente, *k** é a etiqueta do tipo de microfone que distorceu *y, s[n]* é o sinal de treinamento(*), $s_{distorcido}[n]$ é o sinal de treinamento distorcido pelo L-N-L associado ao microfone *k** já no domínio das características e, por fim, Λ é a razão de verossimilhança logarítmica entre *y* e $s_{distorcido}$. Este último valor, a razão de verossimilhança, é usado diretamente na decisão de aceitação ou rejeição por parte do sistema de verificação.

4.6 COMPENSAÇÃO POR NORMALIZAÇÃO DA RAZÃO DE VEROSSIMILHANÇA LOGARÍTMICA (HNORM)

O método de normalização de Λ –vide SEÇÃO 2.4 sobre GMM– o Hnorm (Handset Normalization) foi motivado pela observação de que diferentes modelos alvos de Λ tem diferentes polarizações e escalas para gravações com diferentes tipos de microfones (REYNOLDS, 2000). A idéia principal do Hnorm é a remoção dessas polarizações e escalas da Λ.

Na aplicação do Hnorm, considerando-se que a maior parte dos aparelhos telefônicos comerciais utiliza ou o microfone de carvão ou o de eletreto, identifica-se primeiramente qual o microfone utilizado –etapa realizada pelo identificador descrito na SEÇÃO 5.3- então, de posse da identificação ou etiqueta HS(X) do microfone mais provável para o sinal de entrada, passa-se ao cálculo dos parâmetros de normalização, que são na verdade a média e o desvio padrão dos Λ's possíveis utilizando-se o UBM de gravações por apenas um tipo de microfone.

Para exemplificar: em um eventual experimento prático poder-se-ia, no treinamento do modelo do UBM (*Universal background model)* no qual um grande número de locutores é preciso –vide SEÇÃO 2.3 sobre GMM-, fazer-se uso de um GMM de ordem 1024 usando apenas discursos gravados através de microfones de carvão e outro GMM também de ordem 1024 usando apenas discursos gravados microfones de eletreto.

Daí, supondo-se que estes tivessem distribuição gaussiana, calcular-se-ía as médias μ(carv) e μ(elet) para o primeiro e segundo modelo respectivamente e, da mesma forma, os desvios padrões σ(elet) e σ(carv). De posse desses parâmetros normaliza-se o Λ da seguinte forma:

(*) *s[n]* será *mapeado* por uma não-linearidade Q_{k*}, sendo esta a origem da denominação deste método de

$$\Lambda^{Hnorm}(X) = \frac{\Lambda(X) - \mu(HS(X))}{\sigma\big(HS\big(X\big)\big)} \tag{4.6}$$

compensação.

5 MODELO GERAL PARA COMPENSAÇÃO

Para facilitar a explicação do Modelo Geral para Compensação apresentado neste trabalho, ele foi dividido em duas partes, que podem ser observadas na FIG 5.1 e na FIG 5.2. No entanto, para o caso de implementação basta desenvolver duas rotinas e implementa-las em um único programa. A parte inicial das duas figuras citadas anteriormente mostra o bloco CMS (*Cepstral Mean Subtraction*) que executa a compensação do canal. No caso deste trabalho, esse bloco equivale ao CMS modificado, em que o vetor média cepstral idioma é utilizado para fazer a compensação para a língua portuguesa.

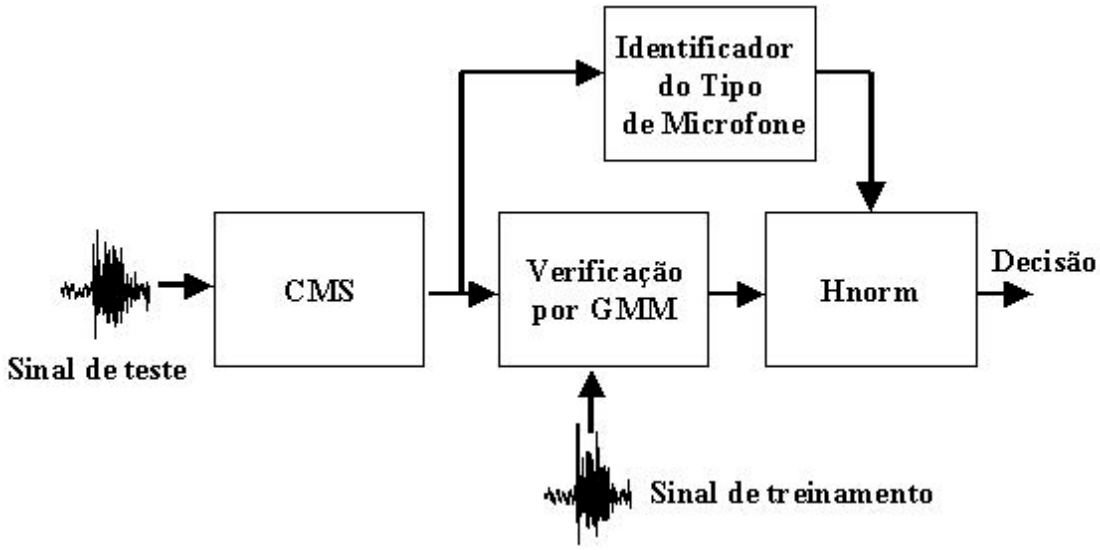

Figura 5.1 - Modelo Geral de Compensação com Normalização da Verossimilhança

O processo ilustrado na FIG 5.1 pode ser descrito em seis partes. A primeira parte é a extração das características cepstrais dos sinais de teste e de treinamento descrita na SEÇÃO 2.3. A segunda parte consiste em realizar a compensação do canal telefônico (CMS) no domínio cepstral. A matriz contendo o resultado da etapa anterior deve ser fornecida ao algoritmo de identificação de microfones já apresentado na SEÇÃO 4.3, esta mesma matriz será utilizada pelo algoritmo de cálculo de razão de verossimilhanças baseados em GMM (vide SEÇÃO 2.4). Por fim, antes de se aplicarem os critérios de decisão (última etapa), deve haver uma compensação dos

efeitos inseridos pela variabilidade dos microfones utilizados para aquisição dos sinais de voz sendo que esta última etapa é realizada por normalização de razão de verossimilhanças, conforme o bloco Hnorm da figura citada anteriormente.

A diferença entre os sistemas ilustrados nas FIG 5.1 e 5.2 encontra-se na maneira como estes realizam a compensação do microfone. Na FIG 5.1, como já foi dito, utiliza-se a normalização da razão de verossimilhanças (vide SEÇÃO 4.6), enquanto no sistema da FIG 5.2 faz-se uso do método de *mapeamento não-linear* descrito na SEÇÃO 4.5.

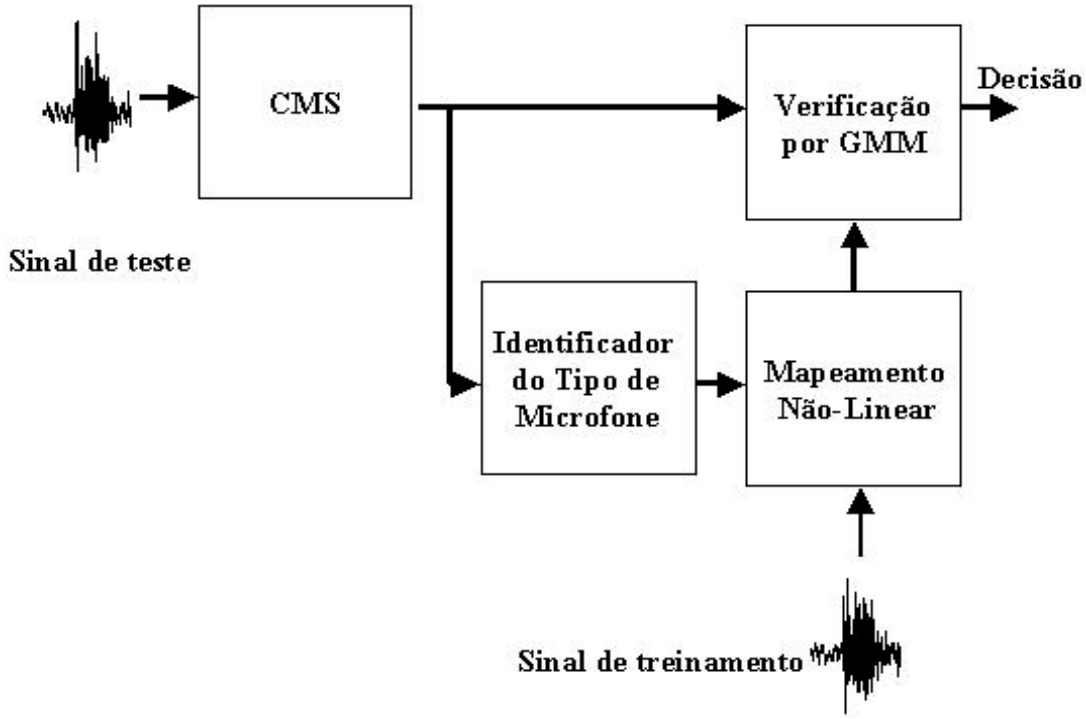

Figura 5.2 - Modelo Geral de Compensação com Mapeamento Não-Linear

É interessante frisar que a compensação dos efeitos dos microfones baseada em Hnorm é mais facilmente implementável do ponto de vista prático, haja visto que este método utiliza dados que devem necessariamente ser calculados independentemente da compensação dos microfones.

O segundo método, que também tem o propósito de realizar a compensação do microfone (mapeamento não-linear), necessita da estimação polinomial da resposta dos microfones (vide SEÇÃO 4.5), porém, as técnicas para cálculo dessas estimações não estão ainda bem desenvolvidas e testadas.

,6 CONCLUSÕES

Os principais objetivos alcançados com este trabalho foram a elaboração dos bancos IME2003 primário e secundário, e a apresentação de um modelo geral para a compensação de microfones e de canais telefônicos. Sendo que para tal foi importante a pesquisa realizada sobre compensação de microfones, na qual foi constatado que estes são os grandes responsáveis pelo alta taxa de erro em sistemas de verificação de locutor.

De acordo com os experimentos realizados, um estudo mais profundo sobre a eficácia do CMS modificado em relação ao CMS, fará sentido somente se for realizada a compensação dos microfones utilizados.

Para a validação dos métodos apresentados aqui, é necessário que haja uma base limpa de sinais de voz com um maior números de locutores, o que não existe até o momento.

Como sugestões para o futuro podem ser destacadas as seguintes:

- desenvolver novos bancos de sinais de voz, que explorem os seguintes aspectos: maior número de pessoas na base IME2003 com períodos de gravação por pessoa menores, gravações efetuadas diretamente em um microfone de carvão, gravações das letras do alfabeto da língua portuguesa realizadas por diferentes pessoas;
- elaborar um software que implemente a identificação do tipo de microfone, dentro do modelo geral para compensação descrito no CAP. 5;
- elaborar um software que implemente o Hnorm no CAP. 4;
- Montar um sistema que implemente o modelo geral descrito no CAP. 5 utilizando as duas implementações anteriores;
- Testar o modelo anterior com a base IME2002, gerando uma curva DET para comparação com experimentos anteriores.

7 Referências Bibliográficas

ALCAIM, A., SOLEWICZ, J. A. e MORAES, J. A. **Freqüência de ocorrência dos fonemas e listas de frases foneticamente balanceadas no português falado no Rio de Janeiro**. Revista da Sociedade Brasileira de Telecomunicações, 7(1), dezembro 1992.

ATAL, B.S. **Automatic recognition of speaker from their voices**. Proceedings of the IEEE, volume 64, pg 460-475, abril 1976.

DEMPSTER, A., Laird, N., and Rubin, D., **Maximum likelihood from incomplete data via the EM algorithm**, J. Roy. Stat. Soc. 39 (1977), p. 1–38, 1977.

GARCIA, Alvin A e Mammone, Richard J. **Channel-Robust speaker identification using modified-mean cepstral mean normalization with frequency warping**. Proceedings of the ICASSP'99, 1999.

JAYANT, M. Naik. **Speaker Verification: A Tutorial**. IEEE Communications Magazine, p. 42-47, Jan. 1990.

LIMA, Charles Borges de. **Sistemas de Verificação de Locutor Independente do Texto Baseados em GMM e Ar-Vetorial Utilizando PCA**. 2001. Tese (Mestrado em Ciências) - Instituto Militar de Engenharia, 2001.

MAMMONE, R. J, Zhang, X. and Ramachandran, R. P. **Robust Speaker Recognition-A Feature-based Approach**. IEEE Signal Processing Magazine, pp.58-71, Sep. 1996.

MARTIN, Alvin. **The DET Curve in Assessment of Detection Task Performance**. Proceedings of EuroSpeech 97, v. 4, p.1895-1898, 1997.

MARTIN, Alvin, and Mark Przybocki. **The NIST 1999 Speaker Recognition Evaluation - An Overview**. Digital Signal Processing, v. 10, p. 1-18, 2000.

OPPENHEIM, Alan V., Schafer, R. **Discrete-Time Signal Processing**. Prentice-Hall Signal Processing Series, 1989.

PICONE, Joseph W. **Signal Modeling Techniques in Speech Recognition**. Proceedings of IEEE, v. 81, n. 9, p. 1215-1247, Sept. 1991.

RABINER, Lawrence, and Ronald Shafer. **Digital Signal Processing of Speech Signals**. USA: Prentice Hall Inc., 1978.

REYNOLDS, Douglas A. **A Gaussian Mixture Modeling Approach to Text Indepent Speaker Identification**. 1992. Tese (Doctor of Philosophy) - Georgia Institute of Technology, 1992.

REYNOLDS, Douglas A. **Experimental Evaluation of Features for Robust Speaker Identification**. IEEE Transactions on Speech and Audio Processing, v. 2, n. 4, p. 639-643, Oct. 1994.

REYNOLDS, Douglas A. **Robust Text-Idenpendent Speaker Identification Using Gaussian Mixture Speaker Model**. IEEE Transactions on Speech and Audio Processing, v. 3, n. 1, p. 72-83, Jan. 1995.

REYNOLDS, Douglas A. **Speaker Identification and Verification Using Gaussian Mixture Speaker Models**. Speech Communication, v. 17, p. 91-108, 1995.

REYNOLDS, Douglas A. HTIMIT and LLHDB: **Speech Corpora for study the of handset transducer effects**. Proceedings of the ICASSP'97, pp.1535-538, Maio 1997.

REYNOLDS, Douglas A. Thomas F. Quatieri, and Robert B. Dunn.
Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing, v. 10, p. 19-41, 2000.

REYNOLDS, Douglas A. **Channel Robust Speaker Verification Via Mapping.** Proceedings of the ICASSP'99, 1999.

SHARMA, Sridevi Vedula. **A Segment-Based Speaker Verification System Using SUMMIT**. Dissertation (Master of Science) - Massachusetts Institute of Technology, 1999.

SILVA, Dirceu Gonzaga da, Apolinário, José A. Jr., and Lima, Charles B. de. **On the Effect of the Language in CMS Channel Normalization.** Proceedings of International Telecommunications Symposium – ITS2002, 2002.

SOUSA, Ricardo Honório Guedes de. **Estudo de Características Relevantes do Sinal de Voz para o Reconhecimento Automático do Locutor Desprevenido, Independente ao Texto**. 1996. Dissertação (Mestrado em Ciências) - Instituto Militar de Engenharia, 1996.

APÊNDICE A

BANCO IME2003

Para realização deste trabalho foi de grande relevância o desenvolvimento de uma base de dados de sinais de voz (corpus) acusticamente limpa, conforme foi citado nos capítulos anteriores. O nome dado a esse banco é IME2003, dando seqüência aos nomes dos bancos passados: IME2001 e IME2002. A base IME2001 foi obtida através da gravação de frases de diversas pessoas utilizando-se apenas um tipo de microfone, que no caso foi eletreto. Já a base IME2002 foi feita através da gravação de ligações telefônicas externas ao IME em que eram registrados alguns dados da pessoa e o tipo de telefone de cada ligação. Com relação a base IME2003 é importante ressaltar neste anexo algumas características peculiares à mesma, no que se refere a três aspectos: 1) Local da gravação (Câmara Acústica), 2)Características da gravação e 3) Organização do banco de vozes.

1 CÂMARA ACÚSTICA

Em relação aos dados da Câmara Acústica, seguem abaixo alguns detalhes:
"Em 2002 foi construída no Laboratório de Processamento de Sinais de Voz do IME uma Câmara Acústica destinada à gravação de sinais de voz com alta qualidade. Esta câmara ou sala acústica foi projetada pelo Escritório de Arquitetura Thompson Motta (especialista em projetos acústicos), teve seu sistema de condicionamento de ar projetado pela VETOR- Consultoria e Projetos S/C ltda. e foi construída pela Decibel Thermo-Acústica Ltda. O resultado em termos de isolamento acústico foi medido em 15/10/2002 pelo Sr. Valdir Garcia da ACUSTERMO-Tratamento Termo-Acústico Ltda., fabricante das portas da sala acústica (www.acustermo.com.br), e estão abaixo resumidos:

1.1 Tom de 1000Hz colocado em frente à porta de entrada:

- NR (nível de ruído) na parte externa = 102 dB (A)
- NR antecâmara (sala do operador) = 50 dB (A)
- NR câmara (aquário) = 42 dB (A)

1.2 Teste de Ruído de fundo (RF):

- RF na parte externa = 56 dB (A).
- RF na antecâmara = MIN 31,2 e LAV5(média) = 34,3 dB (A).
- RF na câmara = 28 dB (A).”

2 ESPECIFICAÇÕES DA GRAVAÇÃO

2.1 Banco de vozes limpas original

O Corpus IME2003 teve como principal característica a excelente qualidade dos sinais de voz, devido ao fato de as gravações terem sido feitas na câmara acústica acima citada, utilizando um microfone de qualidade (AKG C 3000B) e gravado por um equipamento de aquisição de dados (A/D) profissional (Delta). As principais características do Corpus IME2003 podem ser vistas na TAB A1. O esquema de gravação utilizado para a construção do banco limpo primário pode ser observado com mais detalhes na Figura A1.

Número de pessoas	15
Sexo	Masculino
Quantização	16 bits
Amostragem	22050 Hz
Tempo por pessoa	~10 minutos
Microfone	C 3000 B (AKG)
Placa de Aquisição	Delta 44/ M-audio
Lista de Vozes	(ALCAIM, 1992)

Tabela A1. Características da gravação primária do banco IME2003

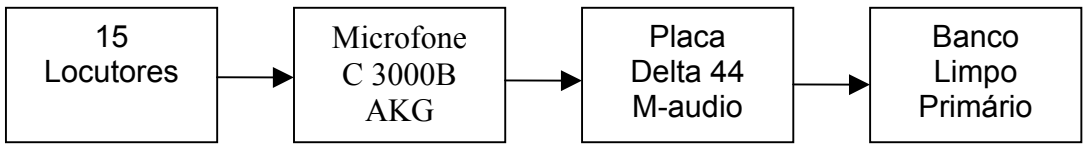

FIG. A1 - Modelo esquemático utilizado para a gravação do banco primário de vozes

2.2 Banco secundário de vozes

A criação desse banco de vozes secundário foi motivada pelo fato de que, numa Verificação Automática de Locutor, quando se faz um treinamento em um determinado tipo de microfone e o teste em um modelo diferente (“handset mismatch”), o desempenho do sistema degrada fortemente inclusive com possibilidade de inserção de não-linearidades ocasionadas pelo uso de microfones de carvão. Assim, este banco torna-se muito útil para o estudo mais detalhado deste problema.

O método empregado para a gravação deste banco foi baseado no método utilizado por (REYNOLDS,1997) para a construção da base HTIMIT e consistiu em passar o banco primário com vozes limpas, obtidas anteriormente, pelos dois principais tipos de cápsulas telefônicas existentes atualmente: os microfones de carvão e eletreto. Sendo que, para isto, foi utilizado um alto-falante que foi conectado aos dois microfones. Deste modo, foram obtidas mais 30 gravações a partir das 15 originais. Um diagrama com o método utilizado para gravação pode ser visto na figura A2, abaixo.

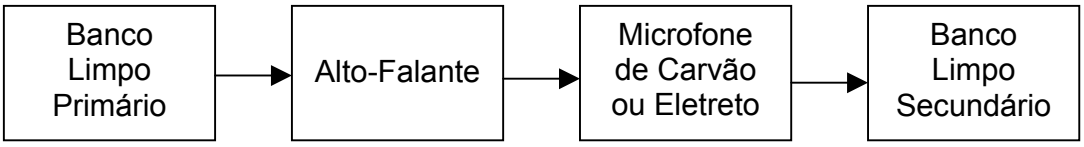

FIG. A2 - Modelo esquemático utilizado para a gravação do banco de vozes secundário

3 ORGANIZAÇÃO DO BANCO IME2003

O banco foi gravado em dois CD’s sendo que as divisões foram feitas por locutores, isto é, cada locutor apresenta uma pasta com três gravações: uma original, outra passando por um microfone de eletreto e uma última passando por um microfone de carvão. A Tabela A2 ilustra a organização do banco.

	CD 1/2 (Locutores)	CD 2/2 (Locutores)	Total
Original	8	7	15
Eletreto	8	7	15
Carvão	8	7	15

Tabela A2 – Organização do banco IME2003