INSTITUTO MILITAR DE ENGENHARIA

CAP DIRCEU GONZAGA DA SILVA

ESTUDO DE COMPENSAÇÃO DE CANAL E ANÁLISE FRACTAL APLICADA AO RECONHECIMENTO DE LOCUTOR

Dissertação de Mestrado apresentada ao Curso de Mestrado em Engenharia Elétrica do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Mestre em Ciências em Engenharia Elétrica.

Orientador: Prof. José Antonio Apolinário Jr. - D.Sc.
Co-orientador: Prof. Rosângela F. Coelho - Dr. ENST

Rio de Janeiro
2002

A Deus, fonte de toda a inspiração; a minha amada esposa Andréia e a minha querida filha Rebeca.

AGRADECIMENTOS

Ao Instituto Militar de Engenharia, pela oportunidade de realizar este curso de Mestrado.

Aos professores José Antônio Apolinário e Rosângela Fernandes Coelho pela orientação, apoio e confiança, mesmo nos momentos difíceis do começo da realização deste trabalho.

Ao professor Roberto Miscow Filho pelos ensinamentos transmitidos e colaboração durante o desenvovimento deste trabalho.

Ao amigo Charles Borges de Lima, pelas sugetões no desenvolvimento da dissertação.

Aos companheiros de mestrado pelo incentivo durante a realização deste trabalho e o pelo apoio direto ou indireto para a conclusão da dissertação.

Ao professores e funcionários do DE-3 pela colaboração com a qual sempre pude contar.

Aos meus familiares que, mesmo distantes, mantiveram um incentivo contante durante a execução desta dissertação.

À Congregação Batista Ladeira dos Tabajaras por suas constantes orações à Deus por meu sucesso.

A minha filha Rebeca, por superar todos os momentos de ausência aos quais a submeti devido ao trabalho.

A minha esposa Andréia por estar sempre ao meu lado, muitas vezes se sacrificando para que eu pudesse estudar.

Finalmente, mas em primeiro lugar, a DEUS, pela saúde, inspiração, e pelas vidas de todas estas pessoas acima citadas que, recrutadas por Ele, formaram um exército a meu favor. A DEUS, pois, toda honra e toda a glória, bem consciente de que estou de que sem Ele, nada poderia ser feito.

"... porque sem mim, nada podeis fazer."
(palavras de JESUS CRISTO em João 15:5)

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

A/D	Conversor Analógico-Digital
AV	*Abry-Veich*
AR-vetorial	Modelo Autorregressivo Vetorial
CMS	*Cepstral Mean Subtraction*
CMN	*Cepstral Mean Normalization*
DCT	*Discrete Cosine Transform*
DFT	*Discrete Fourier Transform*
DTW	*Dinamic Time Warping*
EMQN	*Erro Médio Quadrático Normalizado*
FBI	*Federal Bureau of Investigations*
fdp	função densidade de probabilidade
FFT	*Fast Fourier Transform*
FIR	*Finite Impulse Response*
FT	*Fourier Transform*
GMM	*Gaussian Mixture Model*
HMM	*Hidden Markov Model*
IDFT	*Inverse Discrete Fourier Transform*
IFFT	*Inverse Fast Fourier Transform*
LAR	*Log Area-Ratio*
LBG	*Linde Buzo Gray*
LPC	*Linear Prediction Coefficients*
LFCC	*Linear Frequency Cepstrum Coeficients*
LPCC	*Linear Prediction Cepstrum Coeficients*
MCC	*Mel-Cepstrum Coefficients*
NIST	*National Institute of Standards and Thecnology*
PLP	*Perceptual Linear Predictive*
QV	Quantização Vetorial
RAL	Reconhecimento Automático de Locutor
RNA	Rede Neural Atificial
R/S	*Rescaled Adjusted Range*
STFT	*Short Time Fourier Transform*

RESUMO

Esta dissertação trata da identificação de locutor independente do texto utilizando sinais de voz distorcidos por um canal linear invariante no tempo e da aplicação da análise fractal no reconhecimento de locutor dependente e independente do texto.

Para introduzir a robustez na identificação, foi estudada a compensação de canal através da técnica de subtração da média *cepstral*(conhecida como CMS). Desta técnica foram estudadas as suas vantagens e desvantagens, sendo propostas duas modificações para minimizar tais problemas.

Foi estudado também, o efeito do tamanho da janela de voz no processamento de tempo curto, na identificação cega de canal utilizada pela técnica CMS.

Como características de voz foram utilizados os coeficientes *cepstrum* extraídos do LPC (LPCC), da FFT (LFCC) e do banco de filtros espaçados segundo a escala MEL (MCC). Para identificação de locutor foram utilizadas as técnicas de quantização vetorial e a distância Bhattacharyya. A primeira para avaliar as compensações quadro a quadro e a segunda para avaliar a robustez da evolução dos coeficientes.

Foi verificado que as modificações propostas para o CMS melhoram o reconhecimento com QV, quando se utiliza, na extração de características, a superposição de 50% entre janelas adjacentes. E foi verificado também que o LFCC, em geral, possui a melhor compensação de canal.

Com a distância Bhattacharyya foi verificado que a evolução das características sofre pouca alteração mesmo em situações de distorção devido ao canal.

Para a análise fractal aplicada ao reconhecimento de locutor dependente e independente do texto foi proposto o parâmetro de Hurst como uma nova característica de voz.

Para a identificação dependente do texto foi utilizada uma rede neural linear, obtendo-se taxas de acertos semelhantes às obtidas com os coeficientes MCC. Para a identificação independente do texto foi utilizada a distância Bhattacharyya com o Hurst e os coeficientes MCC. Foi mostrado que o parâmetros de Hurst melhorou as taxas de reconhecimento.Desta forma confirmamos que o parâmetro de Hurst pode ser considerado como uma nova característica de voz, podendo ser utilizado no reconhecimento de locutor.

ABSTRACT

This dissertation addresses the text independent speaker identification with speech signals corrupted by linear and time invariant channels. Moreover, the application of fractal analysis in—text dependent and text independent—speaker recognition was examined.

Aiming the robustness of the identification process, channel compensation using *cepstral* mean subtraction (CMS) was studied. The CMS advantages and drawbacks were studied and two modifications were proposed, to minimize such problems. The effect of the window size in short time speech processing was also studied for a blind channel identification.

The *cepstrum* coefficients extracted from the LPC (LPCC), from the FFT (LFCC), and from the MEL scale spaced filter bank (MCC) were used as speech features. The Bhattacharyya distance and the vector quantization (VQ) techniques were used to form the recognition system. The former was used to evaluate the compensation in a frame to frame basis while the latter was used to evaluate the robustness of the *cepstrum* coefficients evolving in time.

It was observed that the proposed CMS modifications improved the recognition with VQ for speech features extraction using a 50% overlapping between adjacent windows. It was also observed that the LFCC, in general, presented the best channel compensation results.

With the Bhattacharyya distance, it was noted that features time evolving do not have strong changings even in the presence of channel spectral distortion.

The fractal analysis was applied to text dependent and independent speaker recognition and the Hurst parameter was proposed as a new speech feature.

For the text dependent identification, a linear neural network was used and the error rates obtained were similar to those obtained with the Mel *cepstrum* coefficients (MCC). It was shown that the Hurst parameter improved the recognition rates.

Thus, it was demonstrated that the Hurst parameter can be considered as a new speech feature that should be used in speaker recognition.

A voz é um dos mais complexos meios de comunicação do homem. Ela envolve muitos estágios entre a codificação de um pensamento e a decodificação no receptor. Neste meio de comunicação, o sinal acústico resultante do sistema de produção da fala é o portador da informação. O sinal produzido carrega não só a mensagem resultante de um pensamento mas também a informação da fonte que a produziu, ou seja, do locutor. As informações extraídas do locutor incluem a identidade, o sexo, o idioma ou dialeto e, possivelmente, a condição física e emocional do locutor. Com esta riqueza de informações, não é de surpreender que, com o advento dos microcomputadores, a voz tenha achado uma ampla gama de aplicações na interação homem máquina.

A extração da informação do sinal de voz pode ser classificada em (CAMPBELL, 1997):

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

Dentre as aplicações envolvendo a voz, destacam-se: os comando por telefone para controlar operações financeiras com verificação de locutor, ditados contínuos e reconhecimento de locutor para fins forenses. A aplicação determina o tipo de informação a ser extraída do sinal de voz. Por exemplo: para o propósito de RAV, a presença de vários locutores pode levar à confusão, degradando o desempenho do sistema. Da mesma forma no RAL, a variabilidade fonética pode levar a um menor desempenho no reconhecimento de locutores. Além das degradações impostas pelo tipo de aplicação, o sinal acústico produzido é freqüentemente corrompido por diversos agentes do meio durante a sua transmissão via ambiente natural ou através de um canal de comunicações. Tais agentes poderiam ser fontes de ruído, ondas refletidas (eco, reverberação), distorções lineares e não-lineares introduzidas pelo meio de transmissão, entre outros tipos de perturbações.

Esses fatores, em geral, degradam o sinal de voz e, portanto, inviabilizam algum tipo de análise ou aplicação. Desta forma é de extrema importância o estudo de técnicas que minimizem os efeitos provocados por essas degradações.

Neste trabalho, são estudados os efeitos das degradações provocadas pelo canal de comunicações no RAL. A aplicação de RAL para fins forenses, foi motivada por um recente convênio firmado entre o Instituto Militar de Engenharia e a Secretaria de Segurança Pública do Rio de Janeiro.

A seguir serão descritos alguns conceitos sobre o RAL, sua classificação e posteriormente apresentadas as principais fontes de distorções do sinal de voz. Logo após será apresentado o estado da arte para compensação de canal. Finalmente, os objetivos desta dissertação serão descritos na seção 1.4.

1.1 RECONHECIMENTO DE LOCUTOR

Reconhecimento Automático de Locutor é um termo genérico que se refere à tarefa de discriminar pessoas, baseando-se apenas nas características de sua voz. O RAL pode ser classificado segundo a tarefa a ser executada, segundo o texto pronunciado ou segundo ao grau de cooperação na fala dos locutores. O esquema abaixo mostra essas três divisões.

- Quanto a Tarefa
  - Identificação
    - grupo-aberto
    - grupo-fechado
  - Verificação
- Quanto ao Texto
  - Dependente do Texto
  - Independente do Texto
- Quanto à Cooperação
  - Cooperativo
  - Não-cooperativo

Com relação à "tarefa" existem dois[1] campos que têm sido estudados intensivamente: a Identificação de Locutor e a Verificação de Locutor (este último, algumas vezes referido como autenticação ou detecção de locutor).

Identificação de locutor é a classificação de uma locução, pronunciada por um locutor qualquer, como pertencente a um locutor de um conjunto de N locutores de referência (N

[1]Mais recentemente o *National Institute of Standards and Technology* (NIST) (MARTIN; NIST) apresentou uma nova tarefa no RAL que é a segmentação e agrupamento de Locutores. Esta tarefa visa determinar os instantes de fala de um determinado locutor ou de vários locutores durante uma conversação. O agrupamento é feito ao juntar-se os segmentos de uma mesmo locutor.

possíveis saídas). O sistema de reconhecimento neste caso deverá identificar qual dos N locutores de referência pronunciou a locução. A tarefa de identificação requer comparação com as locuções de referência de todos os N locutores, o que se torna impraticável se o número de locutores for muito grande. Como a locução é comparada com cada um dos N padrões de referência, existe a probabilidade de ocorrer uma decisão incorreta a cada comparação; logo, a probabilidade de uma decisão incorreta cresce em função de N. A FIG. 1.1 mostra um exemplo ilustrativo da identificação de locutor.

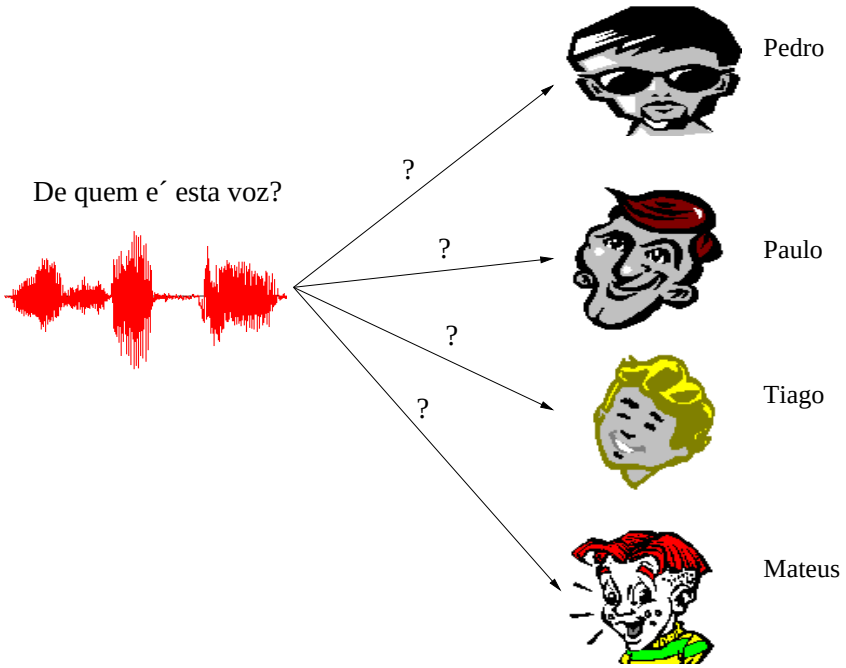


FIG. 1.1: Representação de uma Identificação de Locutores

Verificação de Locutor é um processo em que é decidido se uma determinada locução (pronunciada por um pretenso locutor) pertence ou não a uma pessoa conhecida. A verificação requer uma decisão binária, isto é, aceita ou rejeita o pretenso locutor, dependendo de um limiar de similaridade entre as locuções. A FIG. 1.2 mostra um exemplo ilustrativo da verificação de locutor. Na verificação, a probabilidade de erro tende a um valor limite constante quando o número de locutores tende a infinito. No caso da identificação, a probabilidade de erro tende para um (ROSEMBERG, 1976). Na tarefa de identificação, existem os conceitos de grupo fechado e aberto. O primeiro envolve as tarefas onde o grupo de possíveis locutores é conhecido, já no grupo aberto existem alguns locutores na população que são desconhecidos e portanto devem ser rejeitados pelo sistema. Desta forma, o grupo aberto é uma combinação de identificação de grupo fechado com verificação de locutor.

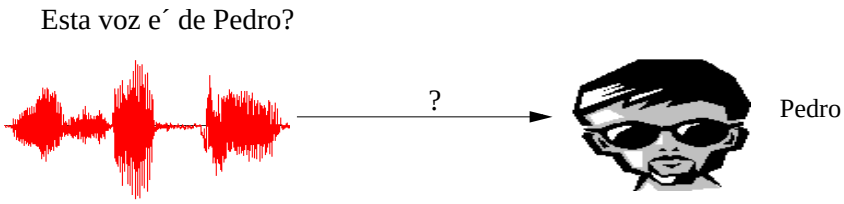


FIG. 1.2: Representação de uma Verificação de Locutor

O grau de controle sobre a geração das locuções é outro importante fator a ser levado em consideração no reconhecimento de locutores. Locuções de texto fixo (chamadas dependentes do texto) são utilizadas em aplicações nas quais o pretenso locutor deseja ser reconhecido e é conseqüentemente cooperativo, ou seja, não introduz propositadamente grandes variações na voz. Locuções de texto livre (chamadas independente do texto) são necessárias naquelas aplicações onde tal controle não pode ser mantido, ou porque o locutor é não-cooperativo ou porque o reconhecimento precisa ser realizado de maneira reservada (ex: escuta telefônica). Nos sistemas dependentes do texto, após adequado alinhamento temporal da locução, pode-se realizar uma comparação com padrões armazenados de uma forma mais precisa, pois há uma similaridade fonética. Isto faz com que o desempenho de tais sistemas seja alto. Sistemas dependentes do texto normalmente utilizam apenas 2-3 segundos para treinamento e teste, enquanto sistemas independentes do texto requerem 10-30 segundos para treinamento e 5-10 segundos para teste em situações de alta relação sinal-ruído (JAYANT, 1990).

O diagrama em blocos ilustrado na FIG. 1.3 apresenta os principais estágios do RAL. O primeiro estágio é composto pelo sistema de aquisição, onde a voz produzida pelo locutor é convertida de pressão sonora em sinal elétrico através de um transdutor. Esse sinal acústico é digitalizado e amostrado segundo uma taxa de Nyquist (OPPENHEIM, 1989). O segundo estágio é composto pelo processamento do sinal e extração de característica,

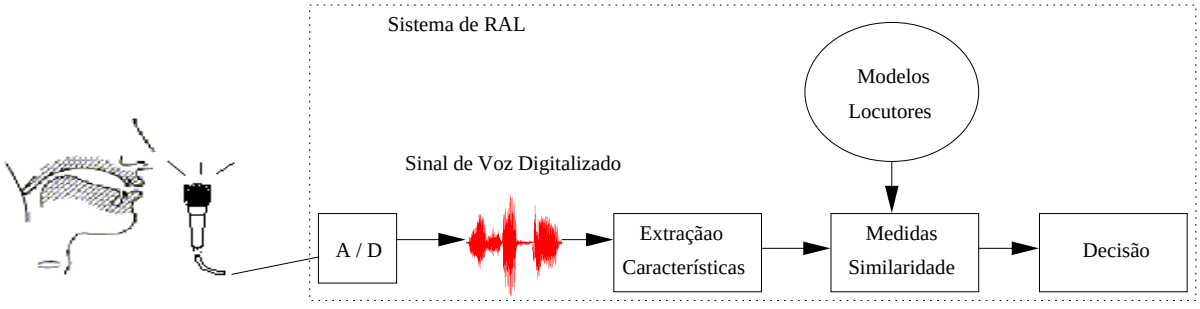


FIG. 1.3: Sistema Genérico de Reconhecimento

onde os parâmetros relevantes da voz que identificam o locutor são extraídos do sinal acústico. A extração de características está baseada no conhecimento do processamento da voz, tais como: modelos do sistema articulatório e sistema auditivo (HERMANSKY, 1994, 1990; O'SHAUGHNESSY, 1986), teoria de fonética e lingüística (DELLER, 1993), processos de transmissão (RABINER, 1978) ou qualquer outra característica envolvida em alguma aplicação específica. O terceiro estágio envolve o cálculo das similaridades (RABINER, 1993) entre as características extraídas do sinal de voz do pretenso locutor com os modelos dos locutores de referência, estes armazenados previamente durante uma fase de treinamento.

No quarto estágio é realizada a comparação entre o valor de similaridade obtido do estágio anterior com limiares previamente estabelecidos no caso de verificação de locutor ou identificação com grupo aberto. Para a identificação de locutor de grupo fechado é feita a seleção pela menor distância ou maior similaridade.

A seguir serão abordadas as principais distorções introduzidas no sinal de voz que reduzem o desempenho do sistema de RAL.

1.2 PROBLEMAS NO RECONHECIMENTO DE LOCUTOR

O que mais se busca em sistemas de RAL é a robustez a ruído ambiente ou a distorções introduzidas nos sinais de voz pelos canais de comunicações. Os sinais podem ser fortemente distorcidos pelo ambiente ruidoso de gravação ou pelas más condições do meio de transmissão (linha telefônica, rádio). Podem aparecer juntamente com o sinal de interesse, linhas cruzadas, interferências estáticas ou zumbidos.

A FIG. 1.4 apresenta um detalhamento da parte inicial do esquema geral de RAL, mostrado na FIG. 1.3, onde são introduzidas uma variedade de fontes de erro. As fontes de distorções podem estar todas presentes ou não. Para as distorções apresentadas podemos destacar:

- *Stress* - tem sido provado que o *stress* e condições emocionais afetam o desempenho no reconhecimento de locutor (HANSEN, 1987);

- *Efeito Lombard* - ocorre devido ao esforço do locutor em compensar o ruído ambiente para poder se comunicar, alterando, assim, as suas características de fala. O nível do efeito *Lombard* dependerá do tipo e nível do ruído ambiente (JUNQUA, 1993; HANSON, 1993; HANSEN, 1994);

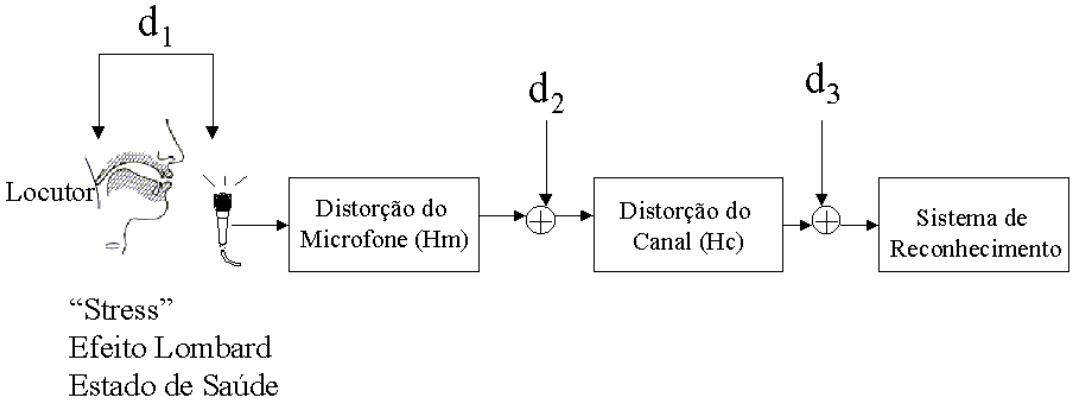


FIG. 1.4: Fontes de distorção introduzidas no sinal de voz

- *Ruído aditivo* - representado na FIG 1.4 por d_1, d_2, d_3, os quais representam o ruído inserido pelo ambiente onde o locutor está falando, o ruído devido ao meio de gravação e o ruído devido ao canal de comunicações, respectivamente;

- *Descasamento acústico* - o descasamento acústico é obtido quando as locuções de treinamento e teste passam, cada uma, por diferentes microfones ou diferentes canais de comunicações ou mídias de gravação. Esse descasamento provoca diferenças no espectro da voz que degradam o reconhecimento.

- *Idade* - o ser humano altera as características do sistema de produção da fala com o envelhecimento do corpo (FURUI, 1981);

- *Período da gravação* - tem-se verificado que o tempo decorrido entre a gravação das locuções de treinamento e teste (FURUI, 1981; OLIVEIRA, 2001), degradam o reconhecimento;

- *Codificação* - o tipo de codificação utilizada para telefonia celular ou armazenamento de voz pode causar um empobrecimento das características dos locutores degradando o reconhecimento.

Para cada tipo de degradação citada, existe pelo menos uma técnica de processamento de sinal de voz para tentar minimizar o seu efeito. Muitos estudos tem sido realizados no intuito de tornar as características de voz menos sensíveis a ruídos, filtragem linear ou distorções por convolução. Por exemplo, alguns estudos procuram melhores características de voz (HUNT, 1989; ROSEMBERG, 1988), outros procuram compensar as

distorções provocadas pelo ruído através de subtração espectral (BOLL, 1979) ou filtragem de Wiener (OPPENHEIM, 1978). Em (HERMANSKY, 1993) é mostrada uma técnica de normalização para ruído aditivo e distorções por convolução.

Neste trabalho serão estudadas as distorções devidas ao canal puramente convolucional. A seguir serão mostrados algumas técnicas utilizadas para compensar tais distorções.

1.3 ESTADO DA ARTE

A robustez no reconhecimento de locutor pode ser vista de uma forma diferente para cada problema. Um reconhecedor de locutor pode ser robusto para uma determinada distorção e não apropriado para outra. O desempenho dos sistemas de reconhecimento de voz atuais, assumem um ambiente livre de ruído e assim eles degradam rapidamente na presença de ruído, distorção ou *stress*.

Seguindo o diagrama mostrado na FIG. 1.3, a compensação pode ser feita nos estágios de extração de características, das medidas de similaridade e de decisão. Sendo assim, existem, atualmente, três campos de atuação para minimizar os efeitos das distorções devidas ao canal. O diagrama abaixo apresenta um resumo dos métodos de compensação mais encontrados na literatura.

1. Sinal
   - Filtragens Lineares e não-lineares
   - Características
     - Intantâneo
     - Longo Termo
2. Sistema de Classificação
   - GMM
   - HMM
   - Híbridos
3. Sistema de Decisão
   - HNorm,TNorm,ZNorm
   - Background

O primeiro método normalmente atua sobre o sinal no tempo ou nas características de voz antes delas serem entregues ao classificador, da seguinte forma:

- através de filtragens do sinal no tempo: compensação através de uma filtragem linear inversa (STOCKHAM, 1975), após a estimação do canal, e não-lineares (KISHORE, 2000), com a utilização de redes neurais;

- atuando diretamente nas características:

- instantâneo: técnicas que compensam as características dentro de cada quadro sem levar em consideração as relações entre quadros, baseiam-se na atribuição de pesos a cada coeficiente do vetor de características (MAMMONE, 1996).
- longo termo: são técnicas que levam em consideração as relações entre quadros; a mais conhecida é a técnica chamada de CMS (ATAL, 1976; MAMMONE, 1996) do inglês *Cepstral Mean Subtraction* ou subtração da média *cepstral*.

Atuar nas características implica, também, na pesquisa de novas características de voz, que sejam naturalmente robustas às distorções introduzidas pelo canal. Considerando isto, será investigada nesta dissertação a utilização da análise fractal no RAL.

O segundo método visa tornar o classificador mais robusto a fim de compensar os efeitos das distorções no estágio de classificação. A utilização de técnicas estatísticas tais como *Gaussian Mixture Models* (REYNOLDS, 1995b) (GMM) e *Hidden Markov Models* (RABINER, 1993) (HMM) fazem parte dessa abordagem. Além da introdução de robustez nos classificadores, procura-se também medidas de similaridade que tenham a capacidade de minimizar os efeitos das distorções.

No sistema de decisão, uma das técnicas mais recentes procura atuar na compensação das verossimilhanças obtidas com o GMM através de um *background*[2] (REYNOLDS, 1994b) obtido com sinais de voz distorcidos através de um canal. Também utilizam-se as normalizações de verossimilhanças chamadas de Tnorm, ZNorm e Hnorm (AUCKENTHALER, 2000), que procuram retirar a polarização introduzida nas verossimilhanças resultantes dos canais de comunicações.

Na última década, o reconhecimento de locutor, principalmente a tarefa de verificação, tem experimentado um crescente número de técnicas de extração de características, de técnicas de modelagem de locutor e de métodos de avaliação. Estas técnicas tem sido abordadas em vários tutoriais (ATAL, 1976; ROSEMBERG, 1976; DODDINGTON, 1985; CAMPBELL, 1997; JAYANT, 1990). A TAB. 1.1 mostra um levantamento recente feito em (CAMPBELL, 1997), onde é apresentada uma coletânea de sistemas de reconhecimento de locutor comerciais.

Desta tabela, pode-se observar que:

1. as aplicações pressupõem locutores cooperativos;

[2]Modelo de GMM obtido com a utilização de vários locutores para treinamento. Este modelo simula o espaço de impostores e retira, com a subtração das verossimilhanças, aquilo que é comum em todos os modelos treinados.

TAB. 1.1: Cronologia selecionada no reconhecimento de locutor.

Fonte	Organiz.	Característ.	Método	Qualid. da voz	Texto	Loc.	Erro(%) temp. teste(s) i:identif. v:verific.
Atal, 1974	AT& T	Cepstrum	Associação de padrões	Laborat.	Depend.	10	i:2% 0,5s v:2% 1s
Markel and Davis, 1979	STI	LP	Estatist. de longo termo	Laborat.	Independ.	17	i:2% 39s
Furui, 1981	AT& T	Cepstrum normalizado	Associação de padrões	Telefone	Depend.	10	v:0,2% 3s
Schwartz, et al. 1982	BBN	LAR	Pdf não paramétrica	Telefone	Independ.	21	i:2,5% 2s
Li and Wrench, 1983	ITT	LP Cepstrum	Associação de padrões	Laborat.	Independ.	11	i:21% 3s v:4%, 10s
Doddington, 1985	TI	Banco de Filtros	DTW	Laborat.	Depend.	200	v:0,8% 6s
Soong, et al. 1985	AT& T	LP	VQ (64)	Telefone	10 dígitos isolados	100	i:5% 1,5s i:1,5% 3,5s
Higgins and Wohlford, 1986	ITT	Cepstrum	DTW	Laborat.	Independ.	11	v:10% 2,5s v:4,5% 10s
Attili, et al. 1988	RPI	Cepstrum LP Autocorr.	Projeção estatística de longo termo	Laborat.	Depend.	90	v:1%, 3s
Higgins, et al. 1991	ITT	LAR LP-cepstrum	DTW	Escritório	Depend.	186	v:1,7% 10s
Tishby, 1991	AT& T	LP	HMM (AR mix)	Telefone	10 dígitos isolados	100	v:2,8% 1,5s v:0,8% 3,5s
Reynolds and Carlson, 1995	MIT-LL	Mel-cepstrum	GMM	Escritório	Depend.	138	i: 0,8%, 10s v: 0,12% 10s
Che and Li, 1995	Rutgers	Cepstrum	HMM	Escritório	Depend.	138	i:0,56% 2,5s i:0,14% 10s v:0,62% 2,5s
Colombi, et al. 1996	AFIT	Cepstrum energia Dcepstr. DDcepstr.	HMM Monofone	Escritório	Depend.	138	i:0,22% 10s v:0,28% 10s
Reynolds, 1996	MIT-LL	Mel-cepstrum D-Mel-cepstr.	GMM	Telefone	Independ.	416	v:11%/16% 3s v:6%/8% 10s v:3%/5% 30s mesmo telef./ outro telef.

2. os sistemas dependentes do texto obtém melhores desempenhos;

3. o tempo de teste influi no desempenho;

4. os sistemas dependentes do meio de transmissão tem um melhor desempenho;

5. em termos de características de voz, há uma predominância das características *cepstrum*.

Em projetos de RAL, é desejável que as características extraídas do sinal de voz atendam aos seguintes requisitos (JAYANT, 1990):

- possuam uma grande variabilidade entre locutores e uma reduzida variabilidade intra-locutor;

- sejam de fácil extração no sinal de voz;

- sejam estáveis no tempo;

- não sejam suscetíveis à mímica.

Alguns desses requisitos têm seu emprego determinado pela aplicação do sistema de RAL. Um exemplo disso é que em aplicações do tipo forense, a premissa de tempo não é tão fundamental, pois dispõe-se de um tempo razoavelmente longo para o processamento do sinal de voz.

Uma das características que atendem a esses requisitos e que tem sido extensivamente utilizada atualmente, como apresentado na (TAB. 1.1), são os coeficientes *cepstrum*. Esses coeficientes podem ser extraídos dos coeficientes LPC, chamado de LPC-*cepstrum* (LPCC) (DELLER, 1993), da Transformada Discreta de Fourier (chamado de LFCC - *Linear Frequency Cepstral Coefficients*) (DELLER, 1993), do banco de filtros espaçados segundo uma escala mel, chamado de mel-*cepstrum* (MCC) (MERMELSTEIN) ou dos coeficientes PLP (*Perceptual Linear Prediction*), chamados de PLP-*cepstrum* (HERMANSKY, 1990).

Em (ATAL, 1974; BEZERRA, 1994; SOUZA, 1996), foi mostrado que os coeficientes LPCC, MCC e LFCC apresentam discriminação e resistência a mímicos superior às características tipo *pitch* ou formantes. Em (REYNOLDS, 1994a) foram analisados os resultados de uma identificação de locutor com os quatro coeficientes *cepstrum* citados no parágrafo anterior. Os resultados mostraram que o PLPC apresentou a pior capacidade de discriminação entre locutores e que os coeficientes MCC, LPCC, LFCC apresentam

taxas de acertos muitos próximos, mesmo em situações de descasamento de canal. Já em (OPENSHAW, 1978), os autores mostraram que o PLPC, juntamente com a filtragem RASTA (HERMANSKY, 1994), apresentou uma taxa de erro menor que o MCC quando o sinal está misturado com ruído e que uma combinação entre MCC e PLPC apresenta uma menor taxa de erro.

Uma das técnicas de normalização de canal mais utilizadas atualmente, para o *cepstrum*, é a CMS (ATAL, 1974; FURUI, 1981), também conhecida como *cepstral mean normalization* (CMN). Esta técnica visa normalizar as locuções de treinamento e teste, retirando o nível DC obtido da evolução temporal das características *cepstrum*. Essa média temporal é uma estimativa grosseira do canal de transmissão ou da resposta do microfone. Ela é aplicada tanto no reconhecimento de voz como de locutor (ATAL, 1974; FURUI, 1981).

Em (NAIK, 1995), foi proposto uma modificação no CMS, chamado de *Pole-Filter*. Esta técnica tem por finalidade compensar as distorções introduzidas pelos pólos do modelo auto-regressivo utilizado na identificação do canal. O objetivo principal desta técnica era de melhorar a estimativa do canal através de manipulações no círculo unitário utilizado no cálculo dos coeficientes LPC. Com esta técnica conseguiu-se melhorar a taxa de acertos em cerca de 5% com a base de dados TIMIT.

Para que a média dos *cepstrum* seja uma estimativa do canal, o CMS parte do pressuposto que a média dos *cepstrum* do sinal de voz limpo tende para zero (MAMMONE, 1996). Será visto no próximo capítulo que esse pressuposto em geral não ocorre, e será proposta uma técnica para minimizar essa distorção levando-se em conta o efeito da polarização introduzida pela língua na estimação de canal utilizada no CMS.

Em recentes estudos sobre verificação de locutor utilizando critérios perceptuais (NIELSEN, 1998), verificou-se que o reconhecimento de locutores realizado por pessoas é igual ou melhor que o obtido com os melhores sistemas de verificação automática de locutor existentes. Foi observado também que o ser humano supera em cerca de 10% o resultado dos sistemas automáticos quando existe descasamento devido ao canal de comunicações. Este resultado era de se esperar visto que o ser humano não se detém apenas nas características fisiológicas ou instantâneas, mas ele extrai as informações da evolução das características, forma de falar e entonação.

As características temporais (ROSEMBERG, 1988) carregam informação de locutor e são robustas às distorções provocada pelo descasamento de canal entre as locuções de treinamento e de teste. Sistemas que casam (FLOCH, 1996) as características instan-

tâneas e as características temporais tais como as derivadas de primeira e segunda ordem da evolução dos coeficientes *cepstrum* ou o AR-vetorial proposto em (BIMBOT, 1992), mostram um aumento na taxa de acertos na identificação de locutores, mesmo em situações de descasamento acústico. Em (CAMPBELL, 1997), a distância Bhattacharyya (dB) (FUKUNAGA, 1990) foi utilizada com sucesso no RAL. Esta distância fornece uma medida da diferença entre distribuições gaussianas obtidas da evolução das características sendo portanto, uma medida temporal.

Dentre as pesquisas sobre novas características de voz, tem-se buscado na análise fractal uma melhor modelagem do sinal de voz, através de sistemas não-lineares e dinâmicos (BOSHOFF, 1991; MANN, 1999). Na área de reconhecimento de voz, em (BOHEZ, 1992) é feita uma análise fractal para a classificação de fonemas onde foram utilizadas a dimensão fractal em conjunto com os parâmetros do *iterated function system* (IFS)(BARNSLEY, 1988). Em (PETRY, 2001) é feita uma aplicação da dimensão fractal juntamente com os coeficientes LPC no RAL. Esta combinação apresentou um considerável ganho na taxa de acerto do reconhecimento quando comparado ao LPC isolado, mostrando que a dimensão fractal possui importantes características do locutor.

Neste trabalho é realizada uma investigação sobre a utilização da análise fractal, através do parâmetro de Hurst, no reconhecimento de locutores.

1.4 OBJETIVOS

Esta dissertação tem dois objetivos principais: estudar a eficácia da técnica de compensação CMS sobre as características *cepstrum*, extraídas do LPC, da FFT e do banco de filtros espaçados segundo a escala Mel, na aplicação de identificação de locutor independente do texto com grupo fechado; e avaliar o parâmetro de Hurst como uma nova característica de voz e sua aplicação no reconhecimento de locutor.

Para atingir esses objetivos, o estudo envolveu:

- a análise da influência das distorções devidas ao canal puramente convolucional no sinal de voz;
- a análise da influência das distorções devidas ao canal nas características de voz LPCC, MCC e LFCC;
- um estudo sobre a eficácia da técnica de subtração da média cepstral na normalização dos coeficientes *cepstrum* na identificação de locutor independente do texto, utilizando a quantização vetorial e a distancia Bhattacharyya como classificadores;

- uma investigação sobre a aplicação da análise fractal, através da dependência temporal, no reconhecimento de locutor;

1.5 ORGANIZAÇÃO DA DISSERTAÇÃO

Esta dissertação está dividida em duas partes. A primeira trata do estudo da compensação de canal nos coeficientes *cepstrum* e da identificação de locutor independente do texto com a técnica CMS. A segunda parte apresenta um estudo sobre a aplicação da análise fractal no reconhecimento de locutor.

Apresentamos a seguir a organização da dissertação através de uma descrição sumária do conteúdo dos capítulos.

Parte I

- **Capítulo 2:** neste capítulo são apresentados os principais conceitos do processamento de voz que permitem uma análise da influência do canal no sinal de voz e nas características *cepstrum*.

- **Capítulo 3:** este capítulo aborda a técnica CMS, suas vantagens e problemas. São propostos dois algoritmos para compensar os problemas apresentados no CMS e é feita uma comparação entre os valores de compensação dos três algoritmos.

- **Capítulo 4:** neste capítulo são aplicadas as três técnicas de compensação estudadas no capítulo 3, na identificação de locutor independente, utilizando a quantização vetorial e a distância Bhattacharyya.

Parte II

- **Capítulo 5:** este capítulo apresenta os conceitos sobre dependência temporal. São apresentados os estimadores do parâmetro Hurst, e sua utilização no processamento de voz.

- **Capítulo 6:** este capítulo apresenta os resultados da aplicação do parâmetro de Hurst no reconhecimento de locutor dependente do texto por fonemas e frases, utilizando redes neurais. São apresentados também alguns resultados com independêcia do texto, utilizando distância Bhattacharyya.

Conclusão: este capítulo apresenta as principais conclusões obtidas dos resultados desta dissertação e aponta propostas para trabalhos futuros.

PARTE I

Estudo da Compensação de Canal

Este capítulo aborda os conceitos básicos de processamento de sinal de voz na análise das perturbações provocadas pelo canal[3]. Apresentamos inicialmente uma revisão sobre o sistema de produção da fala, envolvendo aspectos fonéticos e de informação de locutor no aparelho fonador. Em seguida são apresentados conceitos sobre a análise em tempo curto, conceito este importante para determinar o efeito do canal. Também é apresentada a análise dos efeitos do canal no sinal de voz e nas características *cepstrum*. Encerrando o capítulo, é apresentado um resumo das principais observações referentes ao efeitos do canal no sinal e nas características de voz.

2.1 PRODUÇÃO DA FALA

O sistema de produção da fala está mostrado na FIG. 2.1 (CAMPBELL, 1997). A voz é produzida por uma fonte de excitação, que gera um fluxo de ar dos pulmões através da traquéia e das cordas vocais. A fonação ocorre quando as cordas vocais modulam o fluxo de ar por um movimento de abertura e fechamento sob tensão e pressão do ar. O resultado é um fluxo periódico de ar que excita o trato vocal, causando uma ressonância nas suas freqüências características. Este processo ocorre na produção dos sons ditos sonoros tais como as vogais. As freqüências características são conhecidas como *freqüências formantes*. As freqüências formantes podem ser modificadas por mudanças na configuração do trato vocal, por um processo chamado articulação. Isto acontece quando qualquer dos articuladores tais como língua, úvula (conhecida como campainha) ou os lábios são movimentados. Caso a passagem de ar seja feita de forma forçada produzindo uma turbulência no trato vocal e sem vibração das cordas vocais, tem-se os chamados sons surdos (DELLER, 1993).

Os sons fricativos ocorrem quando existe uma turbulência de ar em alguma região do trato vocal. Eles podem ser sonoros (ex. /v/ e /z/) ou surdos (ex. /f/ e /ch/), dependendo do movimento das cordas vocais. Os sons conhecidos como oclusivos são formados quando a corrente de ar encontra na boca obstáculo total e ocorre uma abertura repentina (ex. /p/, /t/ e /k/). Quando a úvula abaixa, o ar é desviado para o trato nasal produzindo os sons nasais.

[3] O Apêndice 9.1 mostra a classificação adotada neste trabalho para os canais de comunicação.

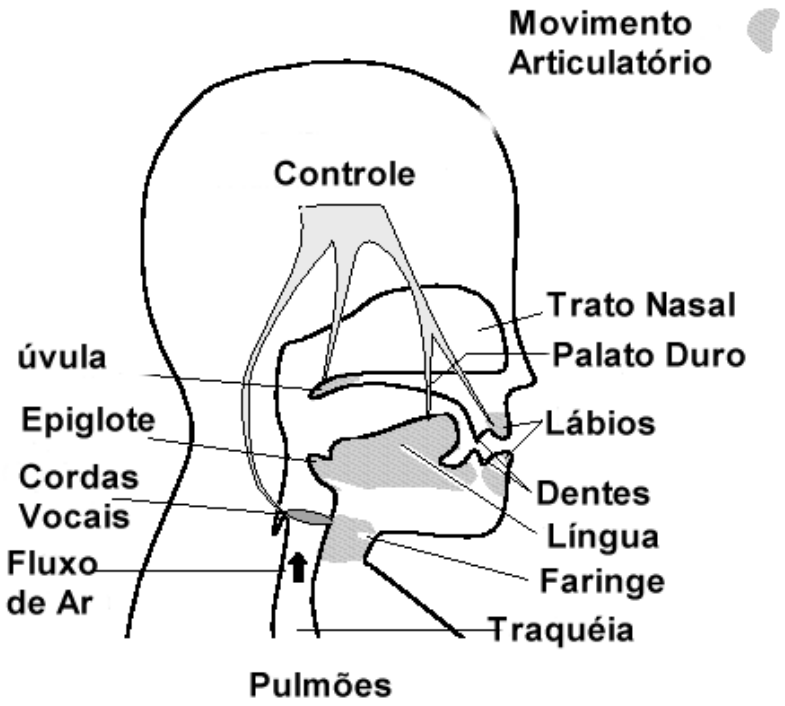


FIG. 2.1: Sistema de produção da Fala

No reconhecimento de locutor deseja-se descobrir os atributos fisiológicos derivados do sinal de voz que tenham correlação com a identidade do locutor. Durante décadas de pesquisa, foram encontradas algumas características, tais como:

- variações no tamanho do trato vocal produzem diferenças no espectro do sinal de voz. O comprimento do trato vocal afeta o espectro como um todo (ATAL, 1974).
- variações no tamanho das cordas vocais estão associadas com mudanças na *pitch* média ou freqüência fundamental da voz (ATAL, 1972).
- variações na úvula e tamanho da cavidade nasal produzem diferenças espectrais nos sons nasalados (ROSEMBERG, 1992).
- a configuração dos dentes e do palato afetam os fricativos (LADEFOGED, 1993).

As informações de locutor, na fala, podem ser classificadas em dois níveis (JAYANT, 1990):

- características de alto nível; são baseadas na lingüística, semântica, ou seja, as características de prosódia tais como o uso das palavras, a pronúncia, os hábitos de linguagem e a maneira de falar.

- características de baixo nível; são as de natureza acústica e descrevem, por exemplo, a sensibilidade auditiva, a taxa de fala ou velocidade, a nasalidade e a sonoridade.

As características de alto nível estão associadas ao comportamento de fala do locutor e são difíceis de quantificar. As de baixo nível estão associadas com as características físicas e estruturais do locutor e são mais fáceis de quantificar; por isso, os principais estudos no reconhecimento de locutor concentram-se nas características de baixo nível.

A *pitch*, apesar de ser um bom discriminador de locutor (ATAL, 1972), é suscetível à mímica (ROSEMBERG, 1976). Além disso, o estado emocional e o efeito *Lombard* também mudam os valores de *pitch* significativamente (JAYANT, 1990). As características espectrais associadas com o trato vocal e nasal tem tido mais sucesso no RAL (CAMPBELL, 1997; SANTOS, 1989; BEZERRA, 1994; PARANAGUÁ, 1997).

As freqüências formantes contém informações específicas do locutor, pois elas fornecem as freqüências de ressonância do trato vocal e podem ser utilizadas para a tarefa de RAL. Porém os algoritmos utilizados no cálculo dos formantes são, ainda, pouco precisos (JAYANT, 1990).

Quando os fonemas e sílabas são articulados, o trato vocal muda sua forma lentamente com o tempo de forma que ele pode ser modelado por um filtro variante no tempo que acompanha as propriedades da resposta em freqüência do trato vocal. Pode-se assumir que o espectro de voz é estacionário para um tempo entre 10 e 40 ms (ATAL, 1974; RABINER, 1978). A análise espectral do sinal de voz sobre intervalos curtos pode prover características que apontem para a forma do trato vocal do indivíduo.

A seguir serão apresentados os conceitos básicos da análise em tempo curto. Estes conceitos serão importantes para o entendimento da análise do efeito do canal no sinal de voz e nas características.

2.2 ANÁLISE DE TEMPO CURTO

A Transformada de Fourier de Tempo Curto (STFT - *Short Time Fourier Transform*) tem sido freqüentemente utilizada na análise do sinal de voz (ATAL, 1974; OPPENHEIM, 1989; RABINER, 1978). A principal idéia do processamento é a de tratar o sinal de voz como quase-estacionário em curtos intervalos de voz (10-40 ms), extraindo desses segmentos a análise espectral necessária para a aplicação.

Seja o sinal de voz $s(m)$ e a seqüência $w(m)$, a STFT é definida como:

$$S(n,\omega) = TF[w(n-m)s(m)] = \sum_{m=-\infty}^{\infty} w(n-m)s(m)e^{-j\omega m} \tag{2.1}$$

onde TF é a abreviatura de Transformada de Fourier.

A janela $w(m)$ é normalmente referida como *janela de análise*. Ela estabelece a seqüência real que determina a porção do sinal de voz a ser analisada para um determinado tempo n. O segmento do sinal "janelado" é normalmente referido como *quadro* e $S(n,\omega)$ como espectro de tempo curto. Na prática usamos a DFT (*Discrete Fourier Transform*) ou seu algoritmo rápido FFT (*Fast Fourier Transform*), ao invés da Transformada de Fourier (TF). Para existir a FFT, o sinal $w(n-m)s(m)$ deve ser absolutamente somável (OPPENHEIM, 1989) - um requisito que é atendido quando a janela é finita.

A taxa de amostragem em n determina o número de quadros a serem extraídos do sinal como um todo, sendo normalmente referida como *taxa de quadros*. Os componentes do vetor espectral, ou alguma transformação do espectro para cada quadro, são normalmente chamadas de características de voz.

A EQ. 2.1 pode ser reescrita em termos da TF do sinal $s(m)$ e da janela de análise $w(m)$, como:

$$S(\omega) = \sum_{m=-\infty}^{\infty} s(m)e^{-j\omega m}, \qquad W(\omega) = \sum_{m=-\infty}^{\infty} w(m)e^{-j\omega m}. \tag{2.2}$$

Para isso considera-se que o segmento $s(m)$ é um segmento de tempo curto e, que fora do cumprimento da janela de análise o sinal é zero ou periódico. Do teorema da modulação (OPPENHEIM, 1989), temos:

$$\begin{aligned} S(n,\omega) &= TF[w(n-m)s(m)] & (2.3)\\ &= TF[w(n-m)] *_\omega TF[s(m)] & (2.4) \end{aligned}$$

ou seja,

$$S(n,\omega) = \frac{1}{\pi}\int_{-\pi}^{\pi} W(\theta)e^{j\theta n}S(\omega+\theta)\mathrm{d}\theta, \tag{2.5}$$

onde o operador $*_\omega$ representa a convolução periódica com respeito à freqüência de análise ω. A EQ. 2.5 mostra a interpretação no domínio da freqüência da análise em tempo curto (RABINER, 1978). A convolução entre a TF do sinal de voz e a TF da janela de análise resultam na Transformada de Fourier de Tempo Curto. O espectro estimado dessa forma, provê uma estimativa do espectro da voz, com a resolução em freqüência limitada pela largura de faixa da janela de análise. Dentre as várias janelas existentes (OPPENHEIM,

1989; RABINER, 1978), a janela de Hamming tem se popularizado no processamento de voz. Ela é definida como:

$$w(m) = \begin{cases} 0.54 - 0.46\cos(2\pi m/(L-1)) & ,0 \leq m \leq L-1 \\ 0 & ,\text{outro} \end{cases}$$

A faixa de passagem em Hz de uma janela de Hamming de L pontos e com freqüência de amostragem do sinal de voz F_s, para o primeiro zero, é dada por:

$$B = 2F_s/L. \tag{2.6}$$

Este cálculo é importante para se poder estabelecer a freqüência de amostragem (ou taxa de quadros θ_s em Hz) das janelas consecutivas a fim de evitar o *aliasing*, ou seja, uma perda de informação na evolução das características (RABINER, 1978). Logo, para reproduzir o sinal sem o *aliasing*, deve-se respeitar o teorema da amostragem (OPPENHEIM, 1989) e, portanto, o sinal deverá ser amostrado a uma taxa maior ou no mínimo igual a $\theta_s = 2B$ quadros por segundo a cada intervalo de janelamento.

A TAB. 2.1 apresenta alguns valores de tamanho de janela e taxa de quadros para evitar o *aliasing* com a janela de Hamming. Esses dados são importantes para determinar a superposição entre janelas adjacentes, pois para a janela de Hamming, obedecendo-se θ_s, a superposição deve ser de, no mínimo, 75%. Para a janela retangular a superposição deve ser de 50%. As demais propriedades da janela de Hamming estão detalhadas em (OPPENHEIM, 1989; RABINER, 1978). A escolha do comprimento da janela poderá influenciar diretamente na compensação de canal e nos resultados de reconhecimento de locutor, como será mostrado posteriormente.

TAB. 2.1: Taxa de amostragem mínima para evitar aliasing para vários comprimentos da janela de Hamming para um sinal de voz amostrado a 8 kHz.

T (ms)	10	20	32	40
L (amostras)	80	160	256	320
B (Hz)	200	100	64	50
θ_s (Hz)	400	200	128	100

A interpretação de filtragem linear da Transformada de Fourier de Tempo Curto (RABINER, 1993) está apresentada na FIG. 2.2. O espectro de tempo curto $S(n,\omega)$ pode ser obtido através da demodulação (multiplicação por uma exponencial complexa) do sinal de voz seguida por uma convolução com a janela de análise. Isto poder ser visto através da EQ. 2.7

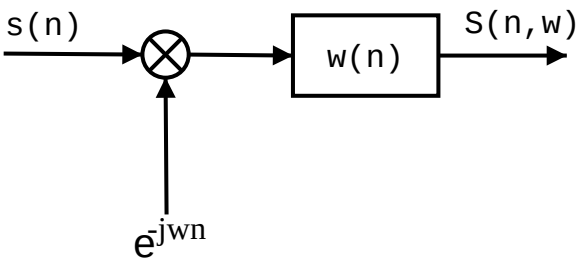


FIG. 2.2: Interpretação da STFT por Filtro Linear

$$S(n,\omega) = \sum_{m=-\infty}^{\infty} w(n-m)s(m)e^{-j\omega m} = s(n)e^{-j\omega n} *_n w(n) \tag{2.7}$$

onde $*_n$ representa uma convolução linear com respeito ao tempo n.

A interpretação é que seqüência de $S(n,\omega_k)$, variando-se n e fixando-se uma banda k, pode ser obtida através de uma demodulação do sinal original $s(m)$ com a aplicação de um filtro passa baixas $w(m)$ para atenuar os componentes de *aliasing* (OPPENHEIM, 1989). Este é um passo básico no processo de demodulação em muitos sistemas de comunicações.

No domínio da freqüência, é mostrado em (RABINER, 1993), que aplicada a Transformada de Fourier na EQ. 2.7 chega-se a:

$$TF(S(n,\omega)) = S(\omega+\theta)W(\theta) \tag{2.8}$$

Ou seja, $W(\theta)$ é um filtro passa-baixas com uma faixa de passagem estreita (isto é, $w(m)$ é quase constante), fazendo com que o espectro de tempo curto $S(n,\omega)$ venha a ser uma boa aproximação da Transformada de Fourier do espectro $S(\omega)$.

Quando o sinal de voz é filtrado por um canal linear invariante no tempo, acredita-se que este canal afete apenas o nível DC da seqüência do logarítmico da energia de uma determinada banda do espectro (ATAL, 1974; FURUI, 1981). A FIG. 2.3 apresenta uma interpretação da STFT considerando um canal com resposta impulsiva $h(n)$. O sinal

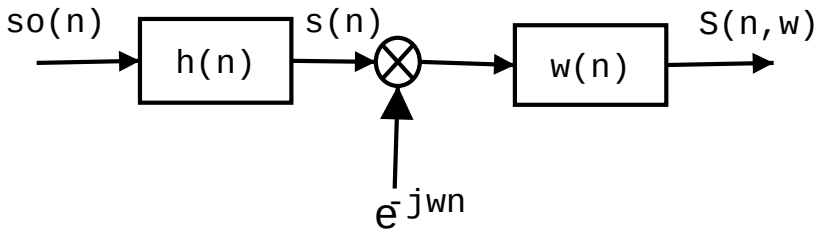


FIG. 2.3: Interpretação da STFT por Filtro Linear com canal de comunicações

degradado é referido como $s(n)$ e o original por $s_o(n)$. Da interpretação da STFT como

filtro linear, pode-se chegar a:

$$S(n,\omega) = ITF_\theta[S_o(\omega+\theta)H(\omega+\theta)W(\theta)]. \tag{2.9}$$

onde ITF_θ representa a transformada de Fourier inversa com respeito à freqüência θ.

A interpretação é semelhante a anterior, isto é, $W(\theta)$ comporta-se como um filtro passa baixas de forma que $W(\theta) \rightarrow A\delta(\theta)$ sendo A constante, isto é, $w(m)$ é quase constante. Com isso, o espectro de tempo curto $S(n,w)$ aproxima-se do espectro de $S_o(\omega)H(\omega)$ a menos de uma constante. Para que essa aproximação seja real, é necessário que a resposta ao impulso do canal $h(n)$ seja mais curta que o comprimento da janela de análise $w(m)$. Em (AVENDAÑO, 1997), foi verificado que o comprimento de uma janela de cerca de 4 vezes a resposta ao impulso do canal fornece uma boa aproximação. Aplicando o logarítmico e obedecendo os requisitos de tamanho de janelas descritos, chega-se à seguinte equação:

$$\log|S(n,\omega)| \approx \log|S_o(n,\omega)| + \log|H(\omega)|. \tag{2.10}$$

Esta equação é o princípio da deconvolução homomórfica, que será estudada no Capítulo 3.

2.3 EFEITO DO TAMANHO DA JANELA NA ESTIMAÇÃO DO CANAL

Nesta dissertação, será estudado o canal linear invariante no tempo sem a influência do ruído aditivo. A voz será distorcida unicamente pela influência de uma filtragem. Serão utilizados dois modelos de dois canais telefônicos: um deles segue as recomendações G.151 da ITU, o qual será chamado de Canal A e o outro é uma simulação do canal continental pobre de voz, chamado de Canal B, ambos podem ser vistos na FIG. 2.4. Os valores dos polinômios que definem ambos os filtros estão no Apêndice 9.2.

Os canais escolhidos para este trabalho possuíam respostas ao impulso bem diferentes. O canal A com uma resposta ao impulso de cerca de 25 amostras e o canal B com cerca de 150 amostras. A fim de verificar o efeito do tamanho da resposta ao impulso do canal na aproximação de STFT, foi feito um teste conforme a configuração mostrada na FIG. 2.5.

O teste consistiu em avaliar o erro médio quadrático entre o logaritmo do espectro do sinal obtido com a STFT após a filtragem pelo canal $h(n)$, com o logaritmo do espectro do STFT do sinal $s(n)$ somado com o logaritmo espectro do canal.

Foram utilizados 3 valores para o comprimento da janela de Hamming: 20 ms, 40 ms e 60 ms, o que corresponde a 160, 320 e 480 amostras, respectivamente. Foi utilizado

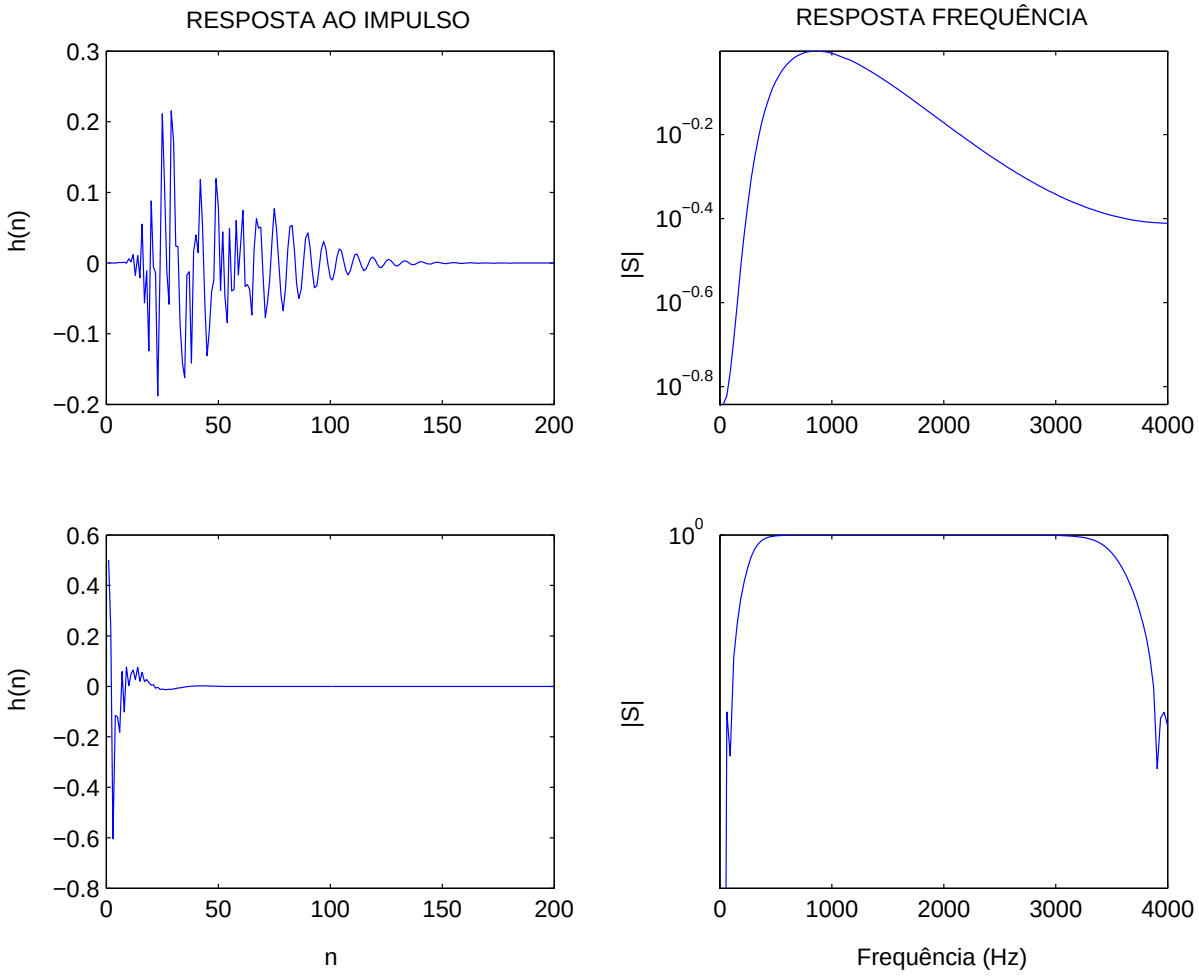


FIG. 2.4: Resposta ao impulso e em freqüência dos Canais A e B. Na parte superior é mostrado a resposta ao impulso e em freqüência do canal B; na parte de baixo a resposta ao impulso e em freqüência do canal A

um sinal de voz de 2 minutos amostrado a 8 kHz. Os resultados estão apresentados na FIG. 2.6. A figura (a) mostra o erro quando utiliza-se o canal A, a (b) para o canal B.

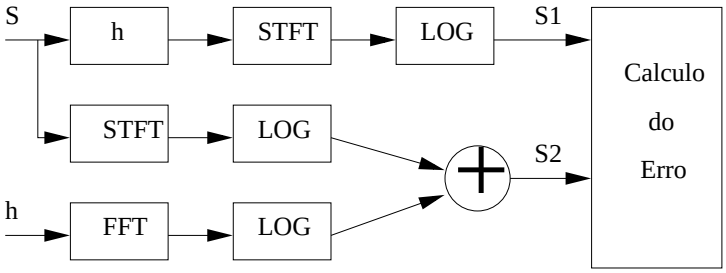


FIG. 2.5: Configuração para medida do erro devido ao tamanho da janela e da resposta ao impulso do canal

A figura (c) mostra uma comparação entre os canais A e B para o caso de uma janela de 20 ms. Pode-se ver que o erro utilizando o canal A é menor que o do canal B visto que a resposta ao impulso dele é da mesma ordem da janela de 20 ms. Aumentando-se o tamanho da janela, percebe-se claramente que o erro diminui melhorando a aproximação dada pela EQ. 2.10. Na letra (c), pode-se ver que a principal diferença entre os canais A e B está na altas freqüências. Este resultado era esperado uma vez que a janela tem o efeito de truncar a resposta ao impulso do canal B, fornecendo desta forma uma baixa resolução em freqüência para o canal.

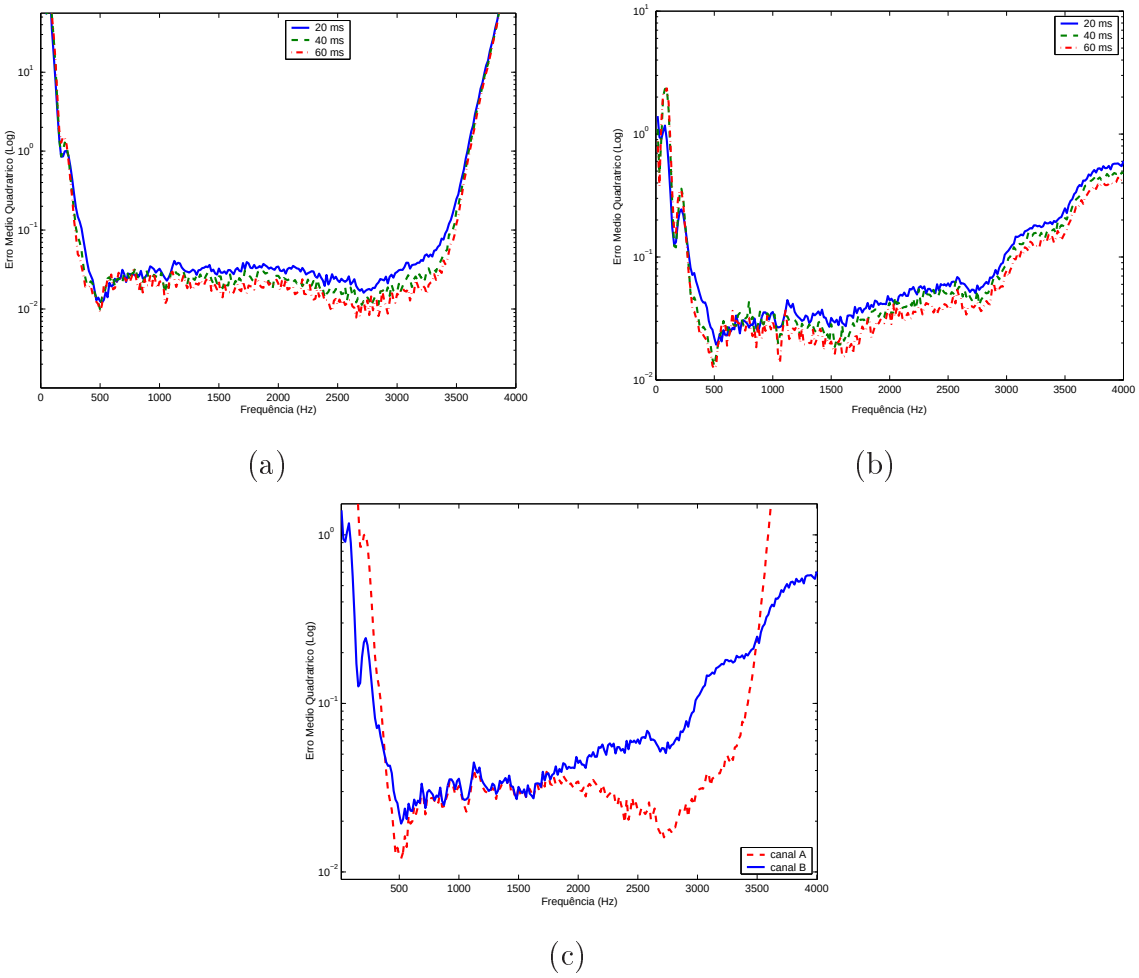


FIG. 2.6: Erro médio quadrático da comparação dos comprimentos da janela e a resposta ao impulso do canal. Foram utilizadas janelas de 20, 40 e 60 ms. (a) Comparação com canal A, (b) com o Canal B,(c) comparação entre os canais A e B para a janela de 20 ms

As características de voz mais utilizadas no RAL, conforme visto na TAB 1.1, são os coeficientes *cepstrum*. Estes coeficientes podem ser extraídos diretamente dos coeficientes de predição linear (LPC - *Linear Prediction Coefficients*) (ATAL, 1974; RABINER, 1978, 1993) ou da FFT (OPPENHEIM, 1989). Além dessas duas formas, há o *cepstrum* extraído da energia de banco de filtros como no caso Mel-*Cepstrum* (MCC)(DODDINGTON, 1985; MERMELSTEIN; RABINER, 1993). Os coeficientes *cepstrum* pertencem ao domínio chamado *quefrency* (CHILDERS, 1977). Vários aspectos gerais tem sido levantados na aplicação das características de voz no RAL. Entre esses aspectos destacam-se:

- em (GRAY, 1976) foi mostrado que uma distância RMS do logaritmo da energia do espectro é significante no processamento de voz. Uma representação logarítmica pode ser adequada, além de poder ser utilizada em modelamentos estatísticos (CAMPBELL, 1997) e num modelo para tratamento de distorções convolucionais (HERMANSKY, 1994; RABINER, 1978).

- baseado em considerações do sistema auditivo, deve-se procurar a utilização de um banco de filtros com freqüências espaçadas de forma logarítmica tais como Mel e Bark (MERMELSTEIN; O'SHAUGHNESSY, 1987; HERMANSKY, 1990).

- baseado na teoria de mascaramento auditivo (O'SHAUGHNESSY, 1987; HERMANSKY, 1990), deve-se procurar utilizar o conceito de bandas críticas, a fim de se ter uma resolução em freqüência ajustada no sistema auditivo humano.

- como um canal telefônico (DODDINGTON, 1985) possui uma faixa de passagem de cerca de 300-3400 Hz, a energia espectral fora dessa faixa tende a ser inconsistente além de introduzir ruído ao conteúdo do sinal de voz. Ao se ignorar estas energias, pode-se aumentar a robustez para o descasamento acústico (DODDINGTON, 1985; REYNOLDS, 1992).

A análise de voz através do espectro de tempo curto apresenta alguns problemas quando consideradas sem a informação de contexto. Há estudos (HERMANSKY, 1998) que indicam que a informação de voz espalha-se sobre o comprimento de uma sílaba, cerca de 200 ms, devido aos efeitos da co-articulação. Além disso, o espectro de tempo curto é sensível às distorções de canal (DODDINGTON, 1985) e requerem a informação de contexto para aumentarem a sua robustez (AVENDAÑO, 1997; ROSEMBERG, 1988).

A seguir, serão feitas algumas considerações sobre as características adotadas, naquilo que é relevante para a compreensão do problema do efeito do canal nas características. Uma descrição detalhada poderá ser obtidas nas referências citadas.

2.4.1 *CEPSTRUM* LPC

LPC são os coeficientes de um filtro só de pólos que modelam a função de transferência do trato vocal. O número de coeficientes é determinado aproximadamente pelo número de pólos do trato vocal dentro da faixa de freqüências do sinal de voz (ATAL, 1976). Os LPC são bastante utilizados porque sua obtenção é simples. O espectro da voz é modelado segundo a equação:

$$P(\omega) = |H(e^{j\omega})| = \frac{G^2}{|A(e^{j\omega})|^2} \tag{2.11}$$

onde $A(e^{j\omega})$ é o polinômio de p coeficientes e é dados por:

$$A(e^{j\omega}) = 1 + \sum_{k=1}^{p} a_k e^{-jk\omega}. \tag{2.12}$$

Existem vários algoritmos para o cálculo dos LPC (RABINER, 1978; DELLER, 1993). O mais utilizado na prática (MARKEL, 1976) é o método da autocorrelação que possui as seguintes vantagens (MAKHOUL, 1975; PICONE, 1993): a) disponibilidade de um algoritmo eficiente para a sua implementação, o algoritmo de Levinson-Durbin (MARKEL, 1976; PICONE, 1993); b) os coeficientes gerados por este método resultam em filtros garantidamente estáveis.

O algoritmo de Levinson-Durbin é baseado no cálculo da autocorrelação de uma seqüência. Este cálculo normalmente é feito nas amostras do sinal de voz, fazendo dessa forma uma estimação para todo o espectro. Quando o sinal de voz passa por um filtro tipo telefônico, é necessário restringir a banda de estimação por motivos explicados no item 2.2. Para restringir o espectro pode-se utilizar a predição linear seletiva descrita em (MARKEL, 1976; MAKHOUL, 1975).

Suponha que se deseje modelar o espectro $S(\omega)$ de uma dado quadro janelado do sinal de voz, somente na região $\omega_a \leq \omega \leq \omega_b$ por um modelo só de pólos dados pelas EQ. 2.11 e 2.12. A fim de calcular os novos coeficientes para o espectro $P(\omega)$, primeiro deve-se realizar um mapeamento linear da região do espectro selecionado para metade do círculo unitário, ou seja,

$$\omega' = \pi \frac{\omega - \omega_a}{\omega_b - \omega_a} \tag{2.13}$$

de forma que a região selecionada é mapeada em $0 \leq \omega' \leq \pi$.

Em seguida, faz-se $P(-\omega') = P(\omega')$, definindo a potência espectral sobre todo o círculo unitário e calcula-se a IFFT sobre $P(\omega)$, obtendo-se a autocorrelação e desta, através do algoritmo de Levinson-Durbin, os coeficientes de predição linear para a faixa do espectro desejada.

Depois de obtidos os coeficientes de predição linear, procede-se a transformação para os coeficientes *cepstrum* através da equação recursiva que pode ser encontrada em (PICONE, 1993). O programa utilizado nesta dissertação para calcular a predição linear seletiva encontra-se descrito no Apêndice 9.3.

A análise do efeito do canal nos coeficientes de predição linear não é obtida de forma fácil; por isso, optamos por obter resultados práticos na avaliação do canal nos coeficientes *cepstrum* obtidos dos LPC.

2.4.2 *CEPSTRUM* FFT

Os coeficientes *cepstrum* podem ser obtidos da FFT através da seguinte equação (OPPENHEIM, 1989):

$$LFCC = IFFT\{\log|FFT(s_n)|\}, \tag{2.14}$$

ou seja, através da transformada inversa de Fourier do logaritmo do módulo da transformada de Fourier da n-ésima janela do sinal de voz.

A análise do efeito do canal sobre os *cepstrum* é obtida diretamente através da EQ. 2.10, onde o sinal resultante aproxima-se da soma do espectro do sinal de entrada com o espectro do canal. Tendo-se uma boa estimativa do canal, é possível minimizar seu efeito, reconstituindo o sinal de entrada.

Para um sinal passado através de um canal telefônico, procede-se como no item anterior, ou seja, calcula-se a transformada inversa somente na faixa de freqüência desejada. O programa utilizado para o LFCC, encontra-se descrito no Apêndice 9.3.

2.4.3 *CEPSTRUM* MEL

Uma das formas de simular a resposta auditiva do homem é utilizando o conceito de bandas críticas[4] e aplicá-lo ao espectro de tempo curto calculado através da STFT. Em

[4]Banda Crítica - função de freqüência que quantifica a faixa de passagem do filtro coclear. Ou seja, é um critério subjetivo do conteúdo de freqüência de um sinal que se refere a faixa de passagem para a qual respostas subjetivas, tais como *loudness* (ou intensidade percebida), tornam-se significativamente diferentes. O *loudness* de uma banda de ruído para uma pressão sonora constante, permanece constante

seguida, realiza-se uma combinação linear de suas componentes de freqüência dentro das faixas de banda crítica da resposta auditiva (MERMELSTEIN). Os pontos do espectro auditivo são calculados através de:

$$A_j(n) = \sum_{f=f_{jl}}^{f_{jh}} W_j(f) S(n,f) \tag{2.15}$$

onde $W_j(f)$ são os pesos atribuídos ao espectro na faixa de banda crítica (f_{jl}, f_{jh}), e $S(n,f)$ é a potência espectral de tempo curto do sinal de voz no tempo n e freqüência f. Normalmente utiliza-se o banco de filtros triangular definido em (MERMELSTEIN) para o cálculo das energias das bandas críticas. O banco de filtros está mostrado na FIG. 2.7.

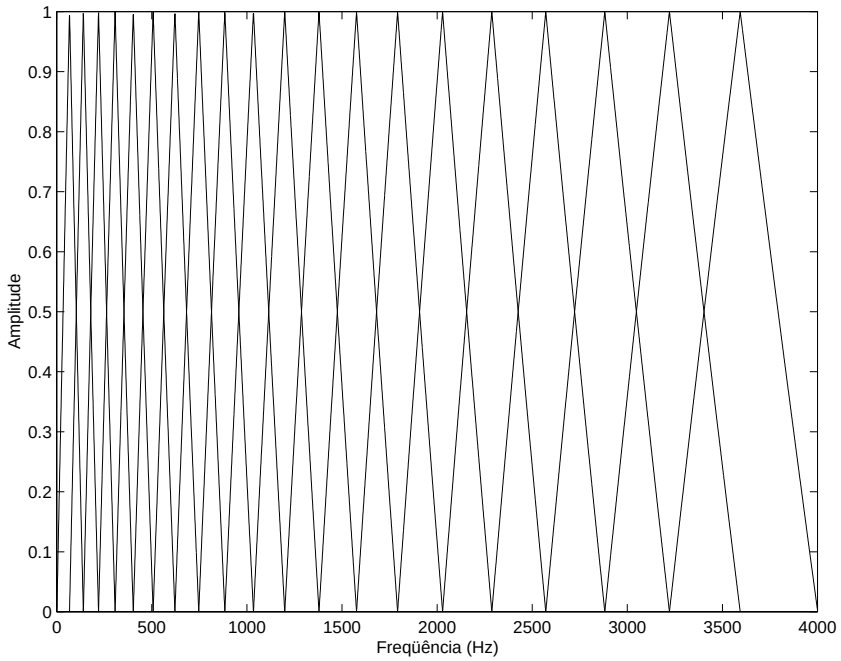


FIG. 2.7: Banco de filtros triangulares espaçados segundo a escala Mel

Cada filtro tem uma freqüência central dada pela escala Mel $f' = 2595 \log_{10}(1 + f/700)$. Normalmente utilizam-se de 20 a 24 filtros (MERMELSTEIN) para toda o espectro de voz (0 - 4000 Hz). Quando se leva em consideração um canal telefônico, utilizam-se os filtros que estiverem dentro da faixa útil do canal.

Para o cálculo do *cepstrum* processa-se a Transformada Discreta Cosseno (DCT) (MERMELSTEIN) dada pela equação:

quando a faixa de ruído aumenta até a largura da banda crítica; após ultrapassar o limite percebe-se mudança no *loudness* (RABINER, 1993).

$$MCC_i = \sum_{j=1}^{N} m_j \cos(\frac{\pi i}{N}(j - 0.5)) \tag{2.16}$$

onde N é o número de filtros utilizado e m_j é a energia do filtro de número j.

O efeito do canal é uma convolução do sinal de voz com a resposta ao impulso do canal; desta forma, é introduzido um fator multiplicativo na EQ. 2.15:

$$A'_j(n) = \sum_{f=f_{jl}}^{f_{jh}} W_j(f)S(n,f)H(f), \tag{2.17}$$

onde $H(f)$ representa a potência espectral do canal.

Considerando-se que a janela de voz é maior do que a resposta ao impulso do canal, pelos motivos já explicados, e considerando-se que o canal varia mais lentamente que a resposta em freqüência da voz, ou seja, a resposta em freqüência do canal não muda dentro da faixa (f_{jl}, f_{jh}), pode-se aplicar o logaritmo na equação acima passando a ter:

$$\log\{A'_j(n)\} = \log\{\sum_{f=f_{jl}}^{f_{jh}} W_j(f)S(n,f)H(f)\} \tag{2.18}$$

ou,

$$\log\{A'_j(n)\} = \log\{A_j(n)\} + C_j, \tag{2.19}$$

onde

$$C_j = \log H(f) = \text{constante}, \qquad f_{jl} \leq f \leq f_{jh} \tag{2.20}$$

Respeitando-se as condições descritas acima, pode-se ver que o canal representa uma constante somada ao espectro e após a aplicação da DCT, ao *cepstrum*, sendo possível a sua retirada ou ter a seu efeito no sinal minimizado através de uma subtração.

2.5 RESUMO

Neste capítulo apresentamos os principais conceitos envolvidos no processamento de voz que possuem alguma influência no efeito do canal do sinal de voz e nas suas características. Assim, foi visto que:

- o tamanho da janela é importante para se ter uma eficácia maior na minimização dos efeitos do canal. Esta janela deve ter no mínimo, 4 vezes o número de pontos da resposta impulsiva do canal;
- além disso, foi visto que a banda de freqüência do sinal voz deve ser limitada pela faixa útil do canal;

- o equacionamento do Mel-cepstrum mostra que ele pode ser bastante afetado pelo canal, apesar de ser, atualmente, muito utilizado no RAL.

Neste capítulo será estudada a técnica de compensação de canal denominada CMS. Iniciamos o capítulo apresentando a história da deconvolução cega aplicada ao processamento de sinais. Em seguida, a deconvolução homomórfica é revista assim como sua utilização com o CMS. Serão abordados, também, os principais problemas envolvidos com esta técnica e mostradas duas modificações propostas nesta dissertação. A primeira visa compensar as polarizações introduzidas na estimação cega do canal feita durante o CMS e a segunda proposta procura normalizar as locuções por um sinal de referência. Serão apresentados ainda os dois modelos de canais telefônicos utilizados nesta dissertação, sendo mostradas as distorções provocadas por ambos no espectro do sinal de voz e o efeito do tamanho da janela, discutido no Capítulo 2. Por último será realizada uma comparação entre o CMS e as duas técnicas propostas.

3.1 REVISÃO

A idéia da normalização de canal está baseada na teoria de filtragem homomórfica desenvolvida por Oppenheim (OPPENHEIM, 1968). Para os sinais convoluídos, uma separação por subtração é possível num domínio no qual os sinais possam ser somados, isto é, o logaritmo do espectro ou o *cepstrum* (OPPENHEIM, 1989). Esta deconvolução é possível quando se conhece algum dos sinais; quando não há informação dos sinais, os métodos de deconvolução cega podem ser utilizados. Nestes métodos, quaisquer informações sobre as características de um dos sinais poderá melhorar os resultados.

O trabalho pioneiro em deconvolução cega foi apresentado em (STOCKHAM, 1975), inspirando algumas das mais populares técnicas de normalização de canal utilizadas atualmente. Naquele trabalho, o objetivo era restaurar uma gravação musical removendo as distorções convolucionais introduzidas por um aparelho de gravação antigo (ano de 1907). Deste modo nenhum conhecimento explícito existia do sinal ou do aparelho de gravação. Para obter algum conhecimento do sinal original, foi realizada uma gravação mais recente da música que eles desejavam restaurar (*Vesti la Guibba* por Enrico Caruso). A gravação recente foi realizada com equipamentos modernos e foi cantada por um tenor com a voz similar a de Enrico Caruso. O espectro médio desta nova versão foi utilizado para estimar as características originais da música de Caruso. Dividindo o espectro médio da antiga

gravação por esta nova versão, eles obtiveram uma estimativa da função de transferência do equipamento original. Um dos pressupostos foi que a resposta em freqüência do equipamento utilizado recentemente possuísse uma resposta em freqüência plana. Após obter a estimativa da resposta do gravador, foi projetado um filtro inverso para compensar as suas distorções, reconstituindo-se, assim, a gravação original do Caruso.

No contexto de RAL, o canal envolvido pode ser um microfone, uma cápsula de telefone, um canal telefônico, enfim, qualquer um dos canais típicos mostrados no APÊNDICE 9.1. Estes canais normalmente possuem respostas ao impulso curtas (AVENDAÑO, 1997) e não prejudicam a conversação entre pessoas (WATKINS, 1991). No entanto, para sistemas de RAL, há uma degradação significativa (para as técnicas atuais) quando há um descasamento entre os dados de treinamento e os de teste, conforme será apresentado nesta dissertação.

3.2 DECONVOLUÇÃO HOMOMÓRFICA

A normalização de canal é aplicada sobre alguma característica extraída do sinal de voz. As características *cepstrum* são as mais utilizadas para esse fim, pois são baseadas no conceito de deconvolução homomórfica (OPPENHEIM, 1989).

Seja $y(t)$ o resultado da convolução entre o sinal de voz $s(t)$ e a resposta ao impulso do canal $h(t)$, então:

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

Após a digitalização de $y(t)$ e aplicação da DFT para cada quadro de voz, a operação acima passa para uma multiplicação no domínio da freqüência.

$$|Y_{k,i}| = |S_{k,i}||H_k| \tag{3.2}$$

onde k é o índice da DFT e i é o índice do quadro. Para simplificar a notação serão utilizados apenas $\mathbf{Y}_i$, $\mathbf{S}_i$ e $\mathbf{H}$ para indicar os vetores de pontos da DFT para cada quadro.

Para a recuperação de um dos sinais, aplica-se o logaritmo a ambos os lados da equação, fazendo com que a multiplicação transforme-se numa soma.

$$\log|\mathbf{Y}_i| = \log|\mathbf{S}_i| + \log|\mathbf{H}| \tag{3.3}$$

Dessa forma, é possível recuperar um dos dois sinais, s ou h, conhecendo-se um deles. Aplicando-se a transformada inversa na EQ. 3.3, passamos para o domínio denominado *quefrency* (CHILDERS, 1977) ou,

$$\hat{\mathbf{y}}_i = \hat{\mathbf{s}}_i + \hat{\mathbf{h}} \tag{3.4}$$

onde $\hat{\mathbf{y}}_i$, $\hat{\mathbf{s}}_i$ e $\hat{\mathbf{h}}$ são os vetores de coeficientes *cepstrum* para cada quadro do sinal de voz distorcido do sinal limpo e do canal, respectivamente.

Para que a EQ. 3.4 seja verdadeira é necessário que o tamanho da janela seja maior que a resposta ao impulso do canal, como discutido no capítulo anterior.

3.3 SUBTRAÇÃO DA MÉDIA CEPSTRAL

A técnica de normalização CMS tem como base a identificação cega de canais. Portanto, qualquer conhecimento prévio de alguma das características dos sinais envolvidos na convolução pode trazer alguma vantagem para o algoritmo. Tomando a média do sinal no domínio *quefrency*, temos:

$$\bar{\hat{\mathbf{y}}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\mathbf{s}}_i + \hat{\mathbf{h}}). \tag{3.5}$$

onde N é o número total de quadros do sinal.

Considerando que o canal linear invariante no tempo tenha o seu efeito somente no nível DC do espectro da voz e desconsiderando quaisquer efeitos de não-linearidade que possam ser introduzidos pelos circuitos ou pelo canal de transmissão, podemos expressar a EQ. 3.5 como:

$$\bar{\hat{\mathbf{y}}} = \hat{\mathbf{h}} + \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{s}}_i \tag{3.6}$$

ou,

$$\bar{\hat{\mathbf{y}}} = \hat{\mathbf{h}} + \bar{\hat{\mathbf{s}}} \tag{3.7}$$

De acordo com (MAMMONE, 1996), caso o sinal de voz seja balanceado[5], a média *cepstral* do sinal de voz tende para zero ou $\bar{\hat{s}} \to 0$ e então $\bar{\hat{\mathbf{y}}} \approx \hat{\mathbf{h}}$.

Logo, tendo obtido a estimativa do canal podemos obter um sinal compensado realizando a subtração:

$$\hat{\mathbf{s}}_i \approx \hat{\mathbf{y}}_i - \bar{\hat{\mathbf{y}}}. \tag{3.8}$$

Esta técnica apresenta duas deficiências: a primeira diz respeito à polarização da estimação de canal realizada durante o CMS e a segunda, à redução de informação de locutor, quando se retira o nível DC da evolução dos coeficientes.

Em (MAMMONE, 1996) é mostrado que para a média *cepstral* do sinal de voz tender para zero é necessário que o sinal de voz seja balanceado entre fricativos, oclusivos e sonoros. A FIG. 3.1 apresenta um gráfico da ocorrência dos fonemas sonoros e surdos da

[5] Equilíbrio entre a ocorrência de sons sonoros, fricativos e oclusivos

língua inglesa e portuguesa extraídos de (ALCAIM, 1992; DENES, 1963). Pode-se ver que as duas línguas apresentam características distintas no balanceamento entre os fonemas surdos e sonoros. Para a língua inglesa há uma maior proximidade da ocorrência entre os surdos e sonoros, para os fonemas de maior ocorrência; já para a língua portuguesa, há uma diferença inicial de cerca de 10% entre os surdos e sonoros. Quando se leva em consideração todos os fonemas, as duas línguas apresentam um volume de sonoros bem maior que o de surdos, fazendo com que o balanceamento falado acima não ocorra, prejudicando a estimação do canal. A relação completa dos fonemas utilizados para esse gráfico encontra-se no APÊNDICE 9.6.

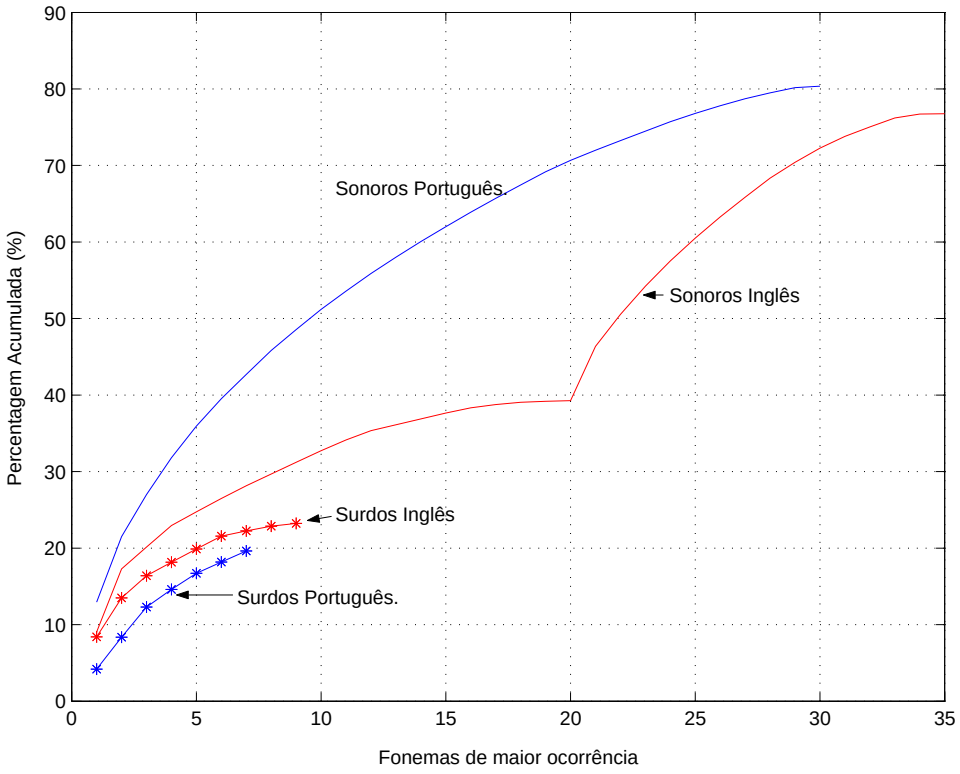
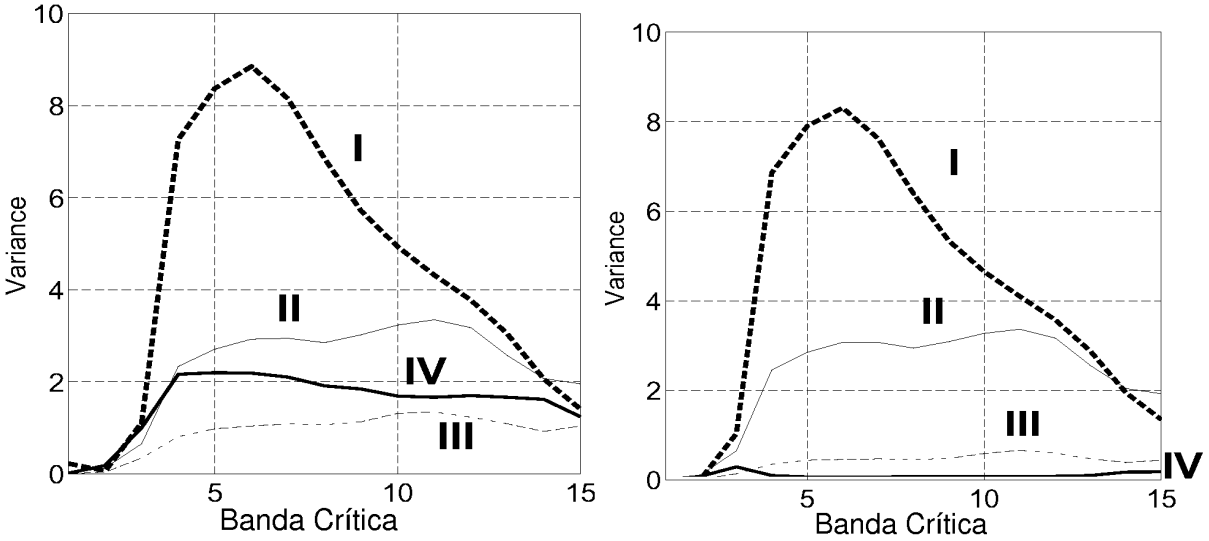


FIG. 3.1: Gráfico de ocorrência acumulada dos fonemas surdos e sonoros da língua inglesa e portuguesa.

A outra limitação do CMS é que ele não remove somente o nível DC atribuído ao canal, mas também tudo o que for constante e comum a todos os quadros do sinal de voz. Isto é mostrado em (KAJAREKAR, 1999), onde foi analisada a variância do sinal[6] antes e depois do CMS, mostrando que a variância dos locutores é reduzida após o

[6] A análise foi feita utilizando a ANOVA (*Analysis of Variance*) fator 3, com a base de dados HTIMIT (base derivada do TIMIT após passagem por oito canais telefônicos). Foram utilizados 133 locutores falando cerca de 30 seg cada. A análise foi realizada no espectro de 15 bandas críticas.

CMS. A FIG. 3.2 ilustra os resultados obtidos no citado artigo, onde foram analisadas as variâncias do sinal de voz com respeito a variabilidade fonética, de contexto, de locutor e a introduzida pelo canal. É possível ver no gráfico que a variabilidade fonética e de contexto não são alteradas com o CMS. A variância devida ao canal teve, no entanto, uma redução mais acentuada que a dos locutores. Isto pode reduzir a taxa de acertos no reconhecimento de locutor, visto que reduz a variabilidade de locutor. O ideal é que a técnica reduzisse o efeito do canal e fonética sem alterar a variabilidade de locutor.



(a) Sem CMS (b) Com CMS

FIG. 3.2: Gráfico de análise de variância da voz por banda crítica extraído de (KAJAREKAR, 1999). I - variância fonética; II - variância de contexto; III - variância de locutor e IV - variância de canal.

A presença de ruído é comum nos ambientes onde são gravadas as locuções. Assumindo que o ruído $v(n)$ seja descorrelacionado com o sinal de voz original $s(n)$, seja $y(n)$ o resultado de uma convolução entre $s(n)$ e $h(n)$ somado ao sinal $v(n)$, ou seja:

$$y(n) = s(n) * h(n) + v(n) \tag{3.9}$$

Passando para o domínio da freqüência (em módulo), temos:

$$|Y(f)| = |S(f)||H(f)| + |V(f)| \tag{3.10}$$

colocando $SH = |S(f)||H(f)|$ em evidência, temos:

$$Y = SH(1 + VS^{-1}H^{-1}) \tag{3.11}$$

aplicando o logarítmo a ambos os lados da equação acima temos:

$$\log Y = \log S + \log H + \log(1 + \frac{V}{SH}) \tag{3.12}$$

A EQ. 3.12 possui duas importantes implicações. A primeira é que, com a presença de ruído, a distorção devido a convolução não pode ser totalmente removida. Da mesma forma, o ruído também não pode ser retirado na presença de distorção convolucional. A segunda implicação é que, apesar do canal considerado ser invariante no tempo e linear, o ruído pode ser variante no tempo o que ocorre na prática, dificultando ainda mais a análise de um método de normalização para compensar a distorção devido ao canal e ao ruído.

3.4 CMS E A MÉDIA DA LÍNGUA

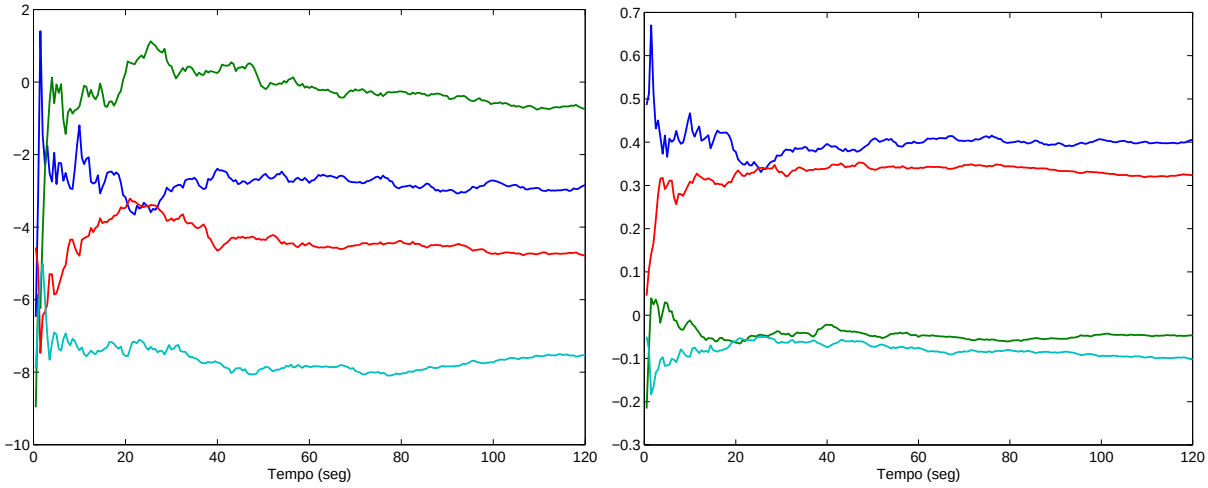
A primeira modificação proposta nesta dissertação (Proposto I ou PI) visa compensar a polarização na estimação cega do canal, introduzida pelo termo $\bar{\hat{\mathbf{s}}}$ da EQ. 3.7. Para isso foi verificado que a média da evolução dos coeficientes *cepstrum* tende para uma constante e que ela pode refletir a tendência do peso fonético do idioma.

A FIG. 3.3 apresenta a evolução da média para os 4 primeiros coeficientes *cepstrum* extraídos do MCC, LPCC e LFCC. Os coeficientes foram extraídos de um sinal de voz da língua portuguesa de 2 minutos, gravados por um único locutor. A estimativa da média foi obtida através de

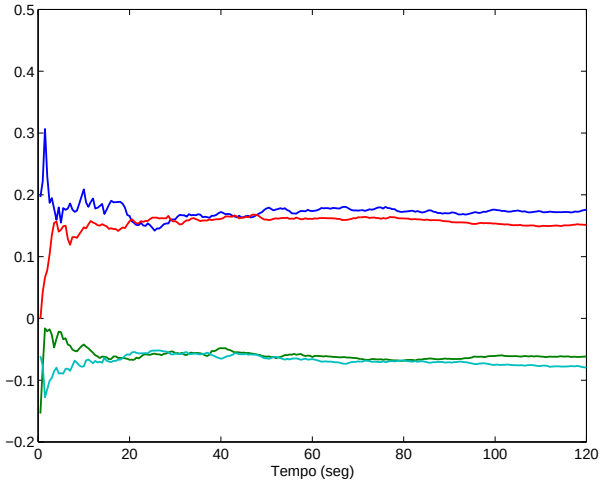
$$m_T = \frac{1}{T}\sum_{i=1}^{T} \hat{\mathbf{s}}_i \qquad \text{, onde} \qquad T = 2...N. \tag{3.13}$$

Teoricamente a média é obtida quando $N \to \infty$. Contudo, podemos observar nos gráficos apresentados que, a partir de aproximadamente 40 segundos para o MCC, 30 segundos para o LPCC e 20 segundos para o LFCC, a estimação da média dos coeficientes não se altera mais. Vários testes foram realizados com outros textos falados por outros locutores e os resultados foram semelhantes. Esses resultados diferem dos obtidos em (NEUMEYER, 1994), para a língua inglesa. No citado artigo, utilizando medidas indiretas, o autor chegou a conclusão de que o valor de $\bar{\hat{s}}$ apresenta pouca variação na estimação do canal a partir de aproximadamente 8 segundos.

Assumindo que o termo $\bar{\hat{\mathbf{s}}}$ seja aproximadamente igual a média $\mathbf{m}$ dos coeficientes *cepstrum* obtidos do sinal limpo, pode-se obter uma estimativa mais fiel do canal, conforme



(a) MCC (b) LPCC



(c) LFCC

FIG. 3.3: Evolução temporal da estimativa da média dos 4 primeiros *cepstrum*.

a equação abaixo.

$$\bar{\hat{\mathbf{y}}} - \mathbf{m} = \hat{\mathbf{h}} + \bar{\hat{\mathbf{s}}} - \mathbf{m} \tag{3.14}$$

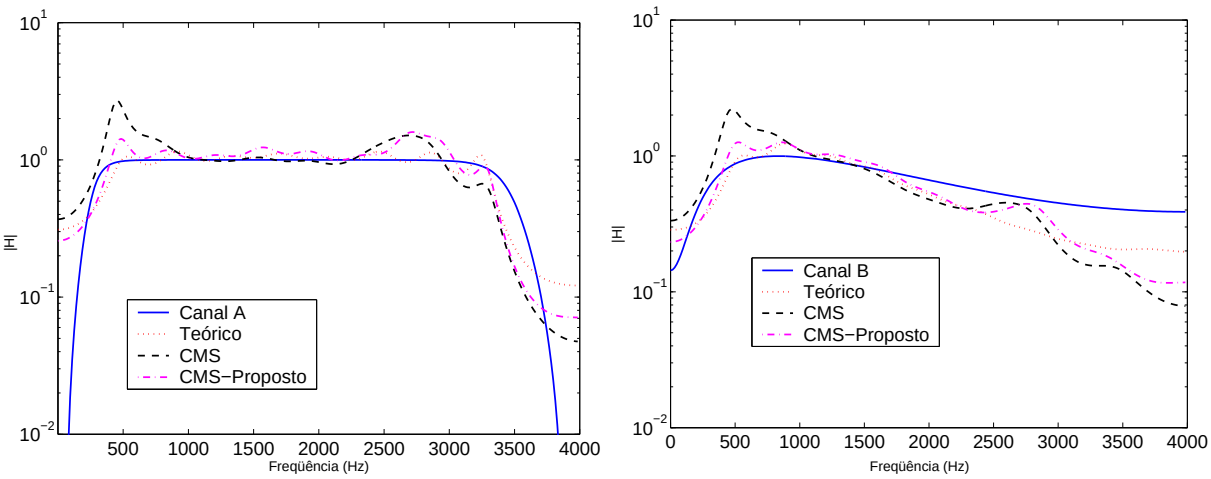
Admitindo que $\bar{\hat{\mathbf{s}}} \approx \mathbf{m}$, temos:

$$\bar{\hat{\mathbf{y}}} - \mathbf{m} \approx \hat{\mathbf{h}}. \tag{3.15}$$

Para a análise de identificação cega de canal foram utilizados dois minutos de fala extraída de uma gravação das frases descritas em (ALCAIM, 1992). O sinal foi gravado originalmente em 22 kHz e reduzido para 8 kHz com 16 bits por amostra. O silêncio foi retirado de todas as frases e os trechos de sinal de voz foram concatenados para formar os dois minutos. Utilizamos janelas de Hamming de 20 ms e superposição de 50% em todos os testes. A média foi estimada utilizando-se 10 outros locutores, cada um falando

cerca de 20 frases distintas descritas em (ALCAIM, 1992). O tempo total foi de cerca de 6 minutos de voz, já retirado o silêncio. O texto utilizado para o sinal de 2 minutos era diferente para o de 6 minutos. Para esta análise foram extraídos 15 coeficientes *cepstrum* das três formas descritas na seção 2.4.

Além da identificação utilizada para o CMS convencional e da estimação pelo método proposto, foi obtida uma estimação teórica dos canais (estimação do canal através da subtração do *cepstrum* do sinal corrompido do *cepstrum* do sinal limpo, conforme EQ. 3.4). A estimativa teórica não pode ser obtida num caso prático e foi mostrada somente como uma referência para o melhor resultado que poderíamos obter. Esses resultados de estimação de canal podem ser vistos na FIG. 3.4.



(a) Canal A[7] (b) Canal B[7]

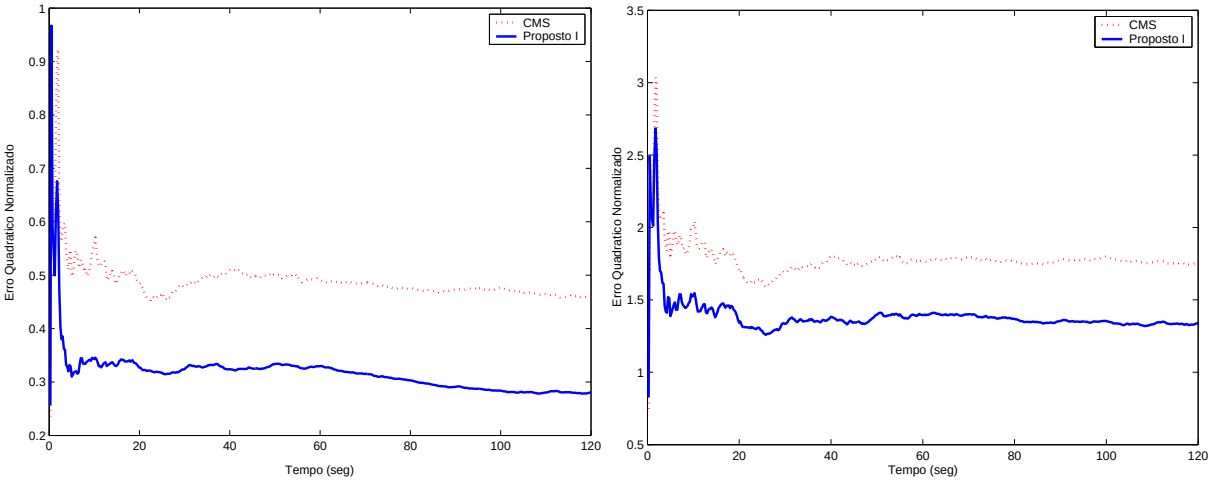
FIG. 3.4: Estimação do canal através do CMS convencional do CMS proposto e da melhor estimativa. A linha cheia é o canal simulado

Pode-se notar que a estimação feita pelo método proposto é superior à do CMS convencional, para os dois canais, aproximando-se mais da estimativa teórica. Observa-se que a polarização introduzida pela energia nas baixas freqüências devida aos sons sonoros, foi reduzida pela subtração da média da língua.

A FIG. 3.5 apresenta o erro quadrático normalizado (EQN) entre a resposta em freqüência do canal e a estimação do canal feita pela característica LPCC, pelo tempo. O EQN é dado pela equação:

$$EQN = \frac{\|H - \hat{H}\|^2}{\|H\|\|\hat{H}\|} \tag{3.16}$$

[7]Estes canais estão descritos na seção 2.3.

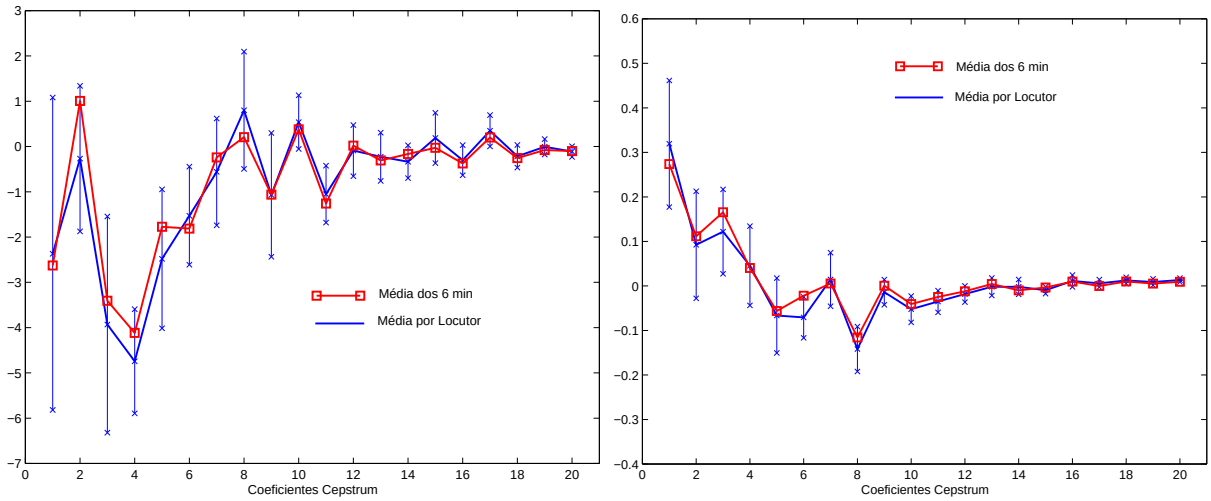


(a) Canal A (b) Canal B

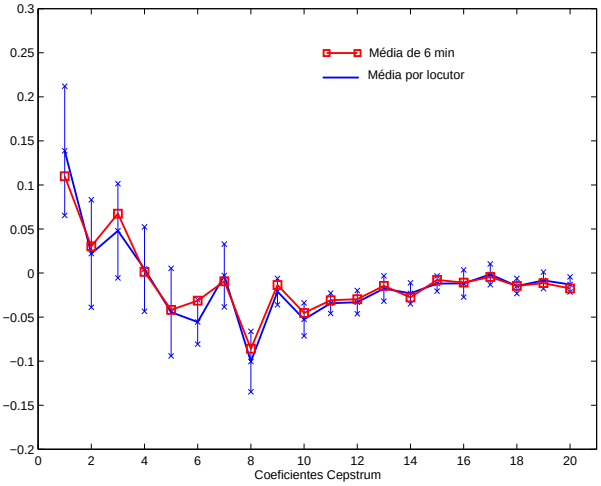
FIG. 3.5: Comparação entre a evolução do erro da estimação do canal com CMS e Proposto I

Pode-se notar pela figura que a técnica proposta melhora considerável sobre a estimação obtida com média simples utilizada no CMS. Outro ponto interessante é que a partir de aproximadamente 10 segundos o erro apresenta um decaimento muito lento, podendo ser considerado constante. Há uma diferença menor para a estimação do canal B, visto que a resposta ao impulso desse canal é da ordem do tamanho da janela de análise. Isto prejudica a estimação do canal conforme discutido no Capítulo 2.

Para que esta técnica proposta possa ser aplicada, é necessário que a média $\mathbf{m}$ obtida seja representativa para qualquer locutor do mesmo sexo. Para isso, a Figura 3.6 apresenta o resultado da média *cepstral* de 15 locutores masculinos, cada um falando cerca de 30 segundos em comparação com a média *cepstral* obtida do sinal de cerca de 6 minutos citado anteriormente. Os gráficos apresentam a média com um desvio padrão para cima e para baixo, mostrado pelas barras verticais na figura. Pode-se perceber que a média *cepstral* dos 6 minutos está muito próxima da média obtida pelos 15 locutores. O MCC é a única que apresenta a maior variação, isto é devido à lenta convergência da média mostrada na FIG. 3.3. Para o LPCC e o LFCC, os resultados dos experimentos realizados sugerem que a média $\mathbf{m}$ é uma constante representativa para o sinal de voz, podendo ser utilizada como uma aproximação $\bar{\hat{\mathbf{s}}}$ em sinais mais curtos.



(a) MCC (b) LPCC



(c) LFCC

FIG. 3.6: Comparação dos coeficientes cepstrum extraídos para cada locutor e da média dos *cepstrum* obtidos a partir de um sinal de 6 min.

3.5 NORMALIZAÇÃO POR SINAL DE REFERÊNCIA

A segunda técnica proposta (Proposto II ou PII) visa compensar a queda de informação de locutor provocada pelo CMS. Ela parte do pressuposto que existe uma outra locução gravada pelo mesmo locutor, em condições que podem ser consideradas como referência. Tendo essa locução, é possível, utilizando o CMS, minimizar o efeito do canal na locução em análise e substituí-lo pelo canais estimados da locução de referência.

Esta técnica pode ser vista como uma técnica de normalização pois procura igualar os canais, fazendo com que as locuções estejam sob as mesmas condições de distorção. A FIG. 3.7 mostra a idéia desta normalização. Sejam s_u e s_v sinais de voz limpos, que

são filtrados pelos canais A e B, respectivamente. A idéia é, após extraído o *cepstrum* de ambos os sinais, aplicar o CMS no sinal filtrado pelo canal B e estimar o canal A através do sinal s_u. Em seguida realizar a soma desta estimação com o sinal filtrado s_v, após o CMS, fazendo com que os dois sinais de saída estejam sob as mesmas condições de distorção.

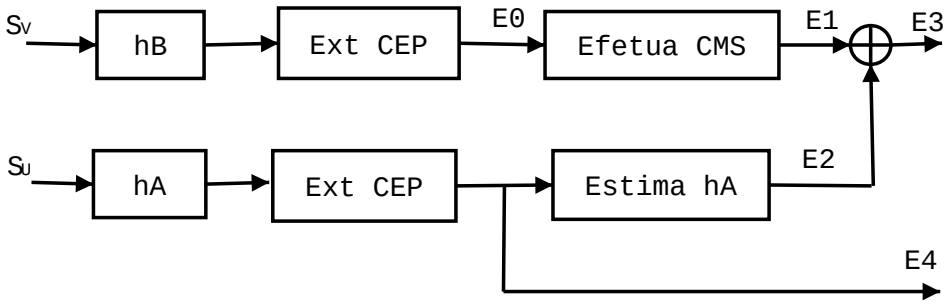


FIG. 3.7: Esquema da compensação de canal através do sinal de treinamento

Os pontos E0, E1, E2, E3 e E4 mostram o desenvolvimento matemático para essa técnica. Sejam $\hat{\mathbf{s}}_u$ e $\hat{\mathbf{s}}_v$ os coeficientes *cepstrum* obtidos dos respectivos sinais. E seja h_A e h_B as respostas ao impulso dos canais A e B.

- no ponto E0 temos o sinal s_v após a extração do *cepstrum*, então:

$$Eo = \hat{\mathbf{s}}_v + \hat{\mathbf{h}}_B$$

- no ponto E1 após a aplicação do CMS em $\hat{\mathbf{s}}_v$, temos:

$$E1 = \hat{\mathbf{s}}_v + \hat{\mathbf{h}}_B - (\bar{\hat{\mathbf{s}}}_v + \hat{\mathbf{h}}_B)$$

  ou

$$E1 = \hat{\mathbf{s}}_v - \bar{\hat{\mathbf{s}}}_B$$

  onde $\bar{\hat{\mathbf{s}}}_v$ é a média cepstral de $\hat{\mathbf{s}}_v$ computada pelo ramo correspondente ao canal B.

- no ponto E2 obtém-se a média *cepstral* do sinal $\hat{\mathbf{s}}_u = \hat{\mathbf{s}}_u + \hat{\mathbf{h}}_A$:

$$E2 = \bar{\hat{\mathbf{s}}}_u + \hat{\mathbf{h}}_A$$

  com $\bar{\hat{\mathbf{s}}}_u$ sendo a média cepstral de $\hat{\mathbf{s}}_u$ computada pelo ramo correspondente ao canal A.

- no ponto E3 é realizada a soma de E1 com E2 chegando a :

$$E3 = (\hat{\mathbf{s}}_v - \bar{\hat{\mathbf{s}}}_v) + (\bar{\hat{\mathbf{s}}}_u + \hat{\mathbf{h}}_A)$$

  supondo que $\bar{\hat{\mathbf{s}}}_v \cong \bar{\hat{\mathbf{s}}}_u$ por serem o mesmo locutor, chegamos a:

$$E3 = \hat{\mathbf{s}}_v + \hat{\mathbf{h}}_A$$

- no ponto E4 temos:

$$E4 = \hat{\mathbf{s}}_u + \hat{\mathbf{h}}_A$$

Esta técnica procura realizar uma troca de canal, ou seja, retirar o efeito do canal B e introduzir o efeito do canal A no sinal de teste, fazendo com que os sinais E3 e E4 estejam sob as mesmas distorções. No caso de uma tarefa de RAL o sinal $E4$ poderia ser utilizado como locutor de referência e o resultante E3 como o pretenso locutor. O algotimo em Matlab utilizado para a tarefa de identificação de locutor, encontra-se descrito no APÊNDICE 9.3.

3.6 COMPARAÇÃO ENTRE O CMS E AS TÉCNICAS PROPOSTAS

A fim de avaliar o desempenho da técnica de normalização sobre diferentes canais, foram gravados dois minutos de um sinal limpo. O sinal foi filtrado através dos canais A e B, obtendo-se dessa forma dois sinais de voz distorcidos pelos mesmos. Em seguida foram extraídos 15 coeficientes MCC, LPCC e LFCC utilizando janelas de Hammming de 20 ms com superposição de 50%.

A comparação foi realizada através de três métodos:

- comparação pelo erro médio quadrático normalizado (EMQN): esta comparação verifica o erro entre os vetores antes e depois das técnicas de compensação;

- comparação através da Quantização Vetorial (QV): esta comparação visa obter uma percentagem de erro entre os vetores de pertinência obtidos com os dados agrupados antes e depois das técnicas de compensação;

- comparação através da distância Bhattacharyya: visa obter a distância entre a seqüência de vetores cepstrum antes e depois das compensações. Esta medida foi escolhida visto que a distância Bhattacharyya avalia a forma da evolução dos coeficientes.

Para a QV e a distância Bhattacharyya foram testadas várias das configurações de extração de características descritas na literatura e comentadas no Capítulo 2. Foram realizados testes com 6 configurações diferentes, visando avaliar a influência dos seguintes parâmetros: tamanho de janela, superposição entre janelas e faixa de freqüência utilizada. As configurações encontram-se descritas na TAB. 3.1.

TAB. 3.1: Configurações utilizadas nos testes de compensação

	a	b	c	d	e	f
Tam Janela (ms)	20	20	20	20	40	40
Superposição (%)	50	50	75	75	75	50
Banda (Hz)	0-4000	300-3400	0-4000	300-3400	300-3400	300-3400
θ_s (Hz)	100	100	200	200	100	50

3.6.1 COMPARAÇÃO PELO ERRO MÉDIO QUADRÁTICO NORMALIZADO (EMQN)

Após as compensações, foi calculado o EMQN entre os vetores de coeficientes *cepstrum* dos dois sinais de teste para cada quadro e depois foi tomada a média de todas as N janelas do sinal, conforme equação abaixo.

$$Erro = \frac{1}{N}\sum_{i=1}^{N}\frac{\|\mathbf{a}_1^i - \mathbf{a}_2^i\|^2}{\|\mathbf{a}_1^i\|\|\mathbf{a}_2^i\|} \tag{3.17}$$

onde $\mathbf{a}_1^i$ e $\mathbf{a}_2^i$ são vetores de *cesptrum* de dimensão K para a janela i dos dois sinais.

A FIG. 3.8 apresenta a taxa de compensação por característica e por método de compensação em relação ao erro obtido com o sinal sem compensação. Os valores de erro foram obtidos pela EQ. 3.17 e a taxa de compensação (M) foi calculada pela equação:

$$M = \frac{SC - T}{SC} \tag{3.18}$$

onde SC representa os valores do EMQN obtidos com os sinais sem compensação e T, o valores obtidos apos o CMS, PI e PII. Na FIG. 3.8, pode-se observar que a normalização realizada pelas técnicas propostas apresentam uma taxa de compensação maior que as obtidas pelo CMS convencional, para as três características estudadas. Há uma compensação sobre o CMS de 2,41% para o MCC, cerca de 1,47% para o LPCC, 1 ,28% para o LFCC, no método Proposto I e de 33,43%, 17,72% e 31,33% para o Proposto II. O método proposto II reduziu a diferença entre as taxas de compensação obtidas por cada características, fazendo com que as diferenças entre as normalizações observadas no caso do CMS e do Proposto I fossem minimizadas. Além disso, pode-se notar que o LPCC

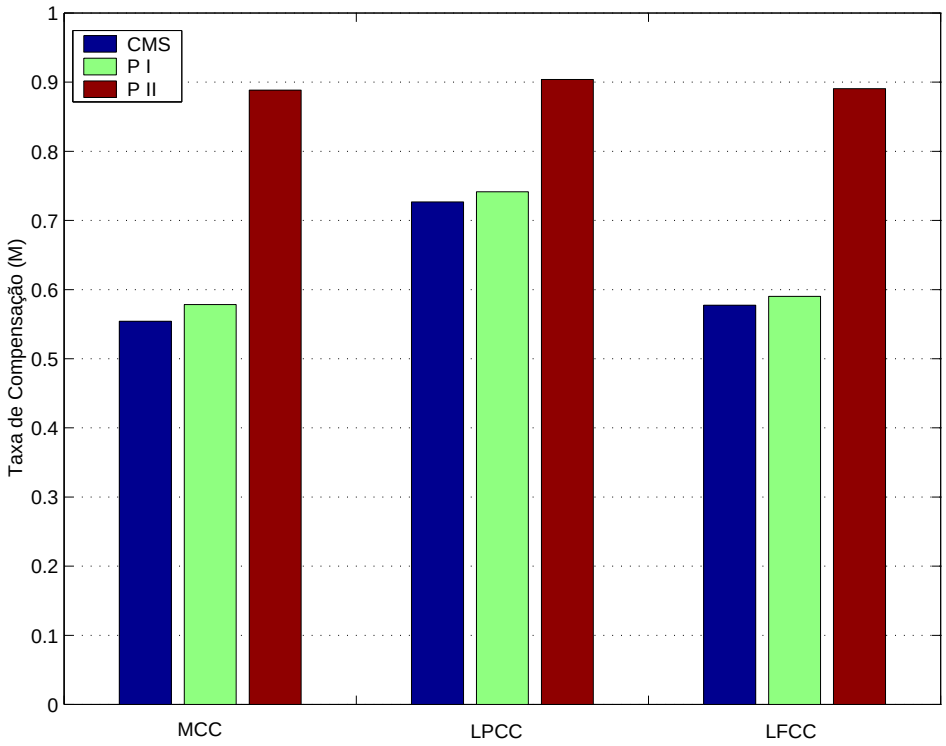


FIG. 3.8: Comparação entre a taxa de compensação obtida pelo CMS, Proposto I e Proposto II, em relação ao sinal sem compensação, para as três características estudadas.

apresentou a melhor compensação. Isto pode ser devido ao amaciamento do espectro provocado pelo LPC, o que vem a facilitar a compensação.

Esses resultados mostram que, levando-se em consideração o conhecimento da língua pode-se melhorar a estimativa do canal e, por conseguinte, a normalização das características de voz.

3.6.2 COMPARAÇÃO ATRAVÉS QUANTIZAÇÃO VETORIAL

Com a quantização vetorial, pode-se agrupar os vetores de características de voz segundo um número de grupos especificado e atribuir um índice para cada um desses grupos. Com isso é possível obter um vetor que atribui a cada quadro de voz um índice do grupo ao qual ele pertence. Este vetor é chamado de *vetor de pertinência*.

A comparação realizada nesta seção calcula o percentagem de erro entre os vetores de pertinência obtidos com a vetores de *cepstrum* do sinal corrompido pelo canal A e os *cepstrum* obtidos com o sinal corrompido pelo canal B.

Esses testes foram realizados com todas as técnicas de compensação já descritas. Uma melhor descrição da quantização vetorial poderá ser vista no Capítulo 4. Para

este teste foi utilizado o algoritmo LBG (LINDE, 1980) para treinamento da QV com distância euclidiana. A TAB. 3.2 apresenta os resultados obtidos neste experimento.

TAB. 3.2: Erro em % obtido comparando os vetores de pertinência do sinal corrompido com o Canal A e B, obtido através quantização vetorial

	S/C			CMS		
	MCC	LPCC	LFCC	MCC	LPCC	LFCC
a	73,18	64,80	78,31	19,18	19,21	30,73
b	74,15	82,84	80,01	2,09	3,50	2,71
c	71,92	65,97	75,05	17,29	18,44	31,47
d	74,85	81,55	78,91	2,21	4,04	2,90
e	74,05	83,03	83,64	1,77	2,95	2,47
f	75,43	85,28	64,43	1,17	4,05	2,60
	P I			P II		
	MCC	LPCC	LFCC	MCC	LPCC	LFCC
a	19,18	19,21	30,73	19,18	19,21	30,73
b	2,09	3,50	2,71	2,09	3,50	2,71
c	17,29	18,44	31,47	17,29	19,44	31,47
d	2,21	4,04	2,90	2,21	4,04	2,90
e	1,77	2,95	2,47	1,77	2,95	2,47
f	1,17	3,83	2,03	1,17	3,55	2,15

As principais observações sobre os resultados da QV são:

- em todas as características as configurações (a) e (c) obtiveram o pior resultado após as compensações, isto é devido à inconsistência discutida no Capítulo 2.

- as configurações (e) e (f) tiveram a menor taxa de erro, isto pode ser devido ao tamanho da janela conseguir estimar de uma melhor forma a resposta do canal, uma vez que a janela de 40 ms é cerca de duas vezes maior que a resposta ao impulso do canal B. A única característica em que isso não foi verdade foi o LPCC, isto pode ser devido ao processo de aproximação realizado pelo LPC, tornando-se mais sensível às mudanças no espectro.

- outro ponto importante a observar é que para o MCC as técnicas modificadas não produziram nenhum efeito tendo seus valores idênticos aos obtidos pelo CMS convencional.

- para o LPCC e o LFCC somente a configuração (f) apresentou diferenças o que pode ser devido ao baixo número de quadros obtidos com a taxa de quadros de 50

Hz. Isto produz uma quantidade de quadros insuficiente para a estimação do canal, sendo necessário um sinal de maior tamanho.

3.6.3 COMPARAÇÃO ATRAVÉS DA DISTÂNCIA BHATTACHARYYA

A distância Bhattacharyya (d_B), descrita no Capítulo 4, serve para comparar duas distribuições através do cálculo das médias e covariâncias. Ela é dada pela equação:

$$d_B = \frac{1}{2}\ln\frac{\frac{|C_i+C_j|}{2}}{|C_i|^{1/2}|C_j|^{1/2}} + \frac{1}{8}(\mu_i-\mu_j)^T(\frac{C_i+C_j}{2})^{-1}(\mu_i-\mu_j) \tag{3.19}$$

as primeira e segunda parte do lado direito da EQ. 3.19 foram definidas como: a primeira (d_C) é responsável pela forma da evolução das características e é extraída da matriz covariância. A segunda parte (d_M) é responsável pela média da evolução das características, ou nível DC. Podemos reescrever a equação acima como:

$$d_B = d_C + d_M \tag{3.20}$$

As TAB. 3.3, 3.4 e 3.5 apresentam as distâncias d_B, d_C e d_M. Pode-se ressaltar que:

- sem compensação de canal a distância d_B é dominada pela distância devida à média (d_M), visto que a distorção do canal prejudica principalmente o nível DC do espectro da voz;

- a configuração (a) e (c) apresentam as maiores distância em todos os casos, tanto para d_C como para d_M. Isto pode ser devido a introdução de ruído nas faixas de freqüência abaixo de 300 Hz e acima de 3400 Hz; conforme discutido no Capítulo 2, há uma inconsistência nessas faixas provocada pela filtragem do canal;

- após a compensação, a distância d_C passa a dominar e pode-se ver que ela não se altera mesmo sem compensação. Trata-se, pois, de uma medida robusta à distorção provocada pelo canal;

- as distâncias das configurações (b) e (d), (e) e (f) mostram que a superposição pouco influi no valor da distância, o que mais altera é a faixa de passagem do sinal de voz;

- a característica que mais sofre com a distorção do canal é o LPCC, isto pode ser devido ao processo de predição linear, visto que o LPCC é obtido de uma aproximação do espectro de voz; dessa forma os erros nessa aproximação refletem-se de forma mais acentuada nos valores obtidos.

- após as compensações, a média torna-se zero pois todas as técnicas apresentadas tem como base o CMS, que retira a média, ou o nível DC, da evolução das características;

- pode-se ver que as técnicas de compensação não produzem efeito sobre d_C, pois as técnicas apresentadas não interferem na evolução das características, elas alteram unicamente o nível DC.

TAB. 3.3: Distância Bhatacharyya para o MCC

	S/C			CMS			P I			P II		
	$\mathbf{d_B}$	d_C	d_M	$\mathbf{d_B}$	d_C	d_M	$\mathbf{d_B}$	d_C	d_M	$\mathbf{d_B}$	d_C	d_M
a	**1,14**	0,1	1,04	**0,10**	0,1	0	**0,10**	0,1	0	**0,10**	0,1	0
b	**0,27**	0,0	0,27	**0,0**	0,0	0	**0,0**	0,0	0	**0,0**	0,0	0
c	**1,15**	0,0	1,04	**0,10**	0,1	0	**0,10**	0,1	0	**0,10**	0,1	0
d	**0,27**	0,1	0,27	**0**	0,0	0	**0,0**	0,0	0	**0,0**	0,0	0
e	**0,29**	0,0	0,29	**0**	0,0	0	**0,0**	0,0	0	**0,0**	0,0	0
f	**0,29**	0,0	0,29	**0**	0,0	0	**0,0**	0,0	0	**0,0**	0,0	0

TAB. 3.4: Distância Bhatacharyya para o LPCC

	S/C			CMS			P I			P II		
	$\mathbf{d_B}$	d_C	d_M	$\mathbf{d_B}$	d_C	d_M	$\mathbf{d_B}$	d_C	d_M	$\mathbf{d_B}$	d_C	d_M
a	**2,18**	0,93	1,25	**0,93**	0,93	0	**0,93**	0,93	0	**0,93**	0,93	0
b	**0,29**	0,02	0,27	**0,02**	0,02	0	**0,02**	0,02	0	**0,02**	0,02	0
c	**2,18**	0,93	1,26	**0,93**	0,93	0	**0,93**	0,93	0	**0,93**	0,93	0
d	**0,29**	0,02	0,27	**0,02**	0,02	0	**0,02**	0,02	0	**0,02**	0,02	0
e	**0,31**	0,02	0,29	**0,02**	0,02	0	**0,02**	0,02	0	**0,02**	0,02	0
f	**0,31**	0,02	0,29	**0,02**	0,02	0	**0,02**	0,02	0	**0,02**	0,02	0

TAB. 3.5: Distância Bhatacharyya para o LFCC

	S/C			CMS			P I			P II		
	$\mathbf{d_B}$	d_C	d_M	$\mathbf{d_B}$	d_C	d_M	$\mathbf{d_B}$	d_C	d_M	$\mathbf{d_B}$	d_C	d_M
a	**1,14**	0,14	1,26	**0,15**	0,15	0	**0,15**	0,15	0	**0,15**	0,15	0
b	**0,27**	0,0	0,27	**0,0**	0,0	0	**0,0**	0,0	0	**0,0**	0,0	0
c	**1,41**	0,15	1,26	**0,15**	0,15	0	**0,15**	0,15	0	**0,15**	0,15	0
d	**0,27**	0,0	0,27	**0**	0,0	0	**0,0**	0,0	0	**0,0**	0,0	0
e	**0,31**	0,0	0,31	**0**	0,0	0	**0,0**	0,0	0	**0,0**	0,0	0
f	**0,31**	0,0	0,31	**0**	0,0	0	**0,0**	0,0	0	**0,0**	0,0	0

Este capitulo abordou a técnica CMS na compensação dos coeficientes *cepstrum*. Foi visto que esta técnica possui duas deficiências, quais sejam: ela utiliza uma estimação de canal polarizada nas freqüências mais baixas, devido ao peso de fonemas sonoros na língua; e, como ela retira o nível DC da evolução dos coeficientes, é retirado, também, informações do locutor, podendo desta forma reduzir a variabilidade ente locutores numa aplicação de RAL.

Para compensar a primeira deficiência foi feita uma técnica que leva em consideração a polarização do idioma. Com isto foi possível obter uma estimação de canal mais eficientes. A segunda deficiência foi compensada através de outra técnica que realiza uma normalização por um sinal de referência.

A avaliação foi feita através de três algoritmos: uma comparação pelo erro médio quadrático normalizado, uma comparação utilizando QV e última através da distância Bhattacharyya.

Na comparação com o EMQN e com a QV as técnicas propostas superaram a compensação realizada pelo CMS. Com a distância Bhattacharyya não houve melhoras em relação ao resultados obtidos com o CMS, isto pode ser explicado pelo fato de que a evolução dos coeficientes sofre pouca alteração com a introdução do canal.

No Capítulo 4 serão apresentados os resultados de uma identificação de locutor com as técnicas descritas neste capítulo. Será permitido observar se as melhoras na normalização produzem de fato melhoras na taxa de reconhecimento.

4 RESULTADOS DA IDENTIFICAÇÃO DE LOCUTOR COM AS TÉCNICAS DE COMPENSAÇÃO

Neste capítulo são apresentados os resultados da identificação de locutor para as técnicas CMS, Proposto I e Proposto II. Os testes foram realizados utilizando a técnica de QV e a distância Bhattacharyya. Para a análise não se está interessado na taxa de erro em valor absoluto, mas em avaliar a robustez e o comportamento dos coeficientes segundo a compensação de canal realizada pelas diferentes técnicas descritas no Capítulo 3.

Inicialmente, serão apresentados os sistemas de decisão utilizados e a base de dados escolhida para a identificação de locutor. Depois serão apresentados os principais resultados com a QV e com a distância Bhattacharyya, sem e com as técnicas de compensação. Outros resultados com a QV, encontram-se no APÊNDICE 9.4.

Os resultados serão apresentados segundo duas considerações. Ou seja, a taxa de melhora no reconhecimento proporcionada pelas técnicas de compensação e da compensação feita pelo CMS da diferença de tempo entre as seções de gravação das locuções de treinamento e teste.

4.1 SISTEMAS DE DECISÃO ADOTADOS

4.1.1 QUANTIZAÇÃO VETORIAL

A Quantização Vetorial (QV), aplicada ao RAL, foi proposta em (SOONG, 1987). Este sistema de decisão escolhido para a avaliação das técnicas de compensação, pois ele permite avaliar os locutores levando-se em consideração os vetores de voz de forma instantânea. Um outro objetivo ao escolher a QV é a sua proximidade com GMM, citado no Capítulo 1, que é umas das técnicas que tem tido maior emprego nos últimos anos.

O livro-código (*codebook* como é conhecido em inglês) de cada locutor foi obtido através de 32 agrupamentos das características de voz extraídas da elocução falada por cada locutor. Para o treinamento foi utilizado o algoritmo LBG descrito em (LINDE, 1980). O esquema utilizado para o sistema de identificação é o descrito na FIG. 4.1, onde D_k indica a distância total obtida para cada locutor, e é dada por:

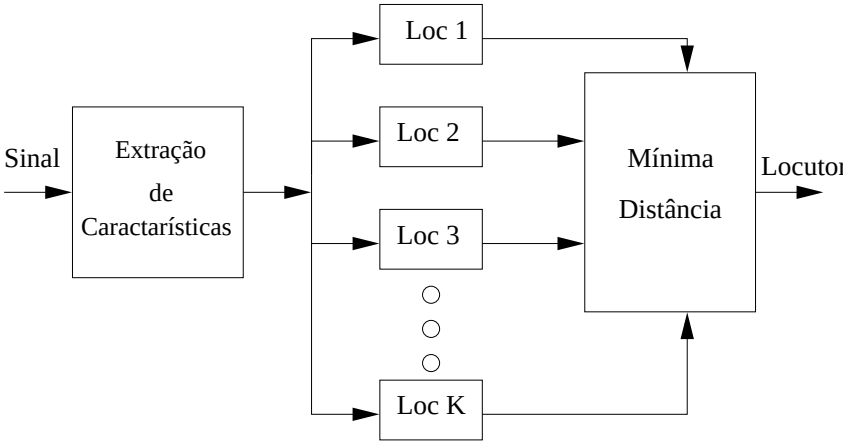


FIG. 4.1: Sistema de Identificação por Quantização Vetorial.

$$D_k = \frac{1}{N}\sum_{i=1}^{N} \min_{1\leq j\leq M} d(a_i, b_j) \tag{4.1}$$

sendo k o índice do locutor, M o número de centróides e N o número de janelas do sinal de teste. Para o cálculo de $d(a_i, b_j)$ foi utilizada a distância euclidiana. Esta distância foi escolhida pois o principal objetivo da análise era verificar a eficácia das técnicas de compensação de canal e não melhorar em valor absoluto a taxa de reconhecimento. Para um melhor desempenho do sistema de reconhecimento seria adequado utilizar a distância de Mahalanobis (CAMPBELL, 1997).

4.1.2 DISTÂNCIA BHATTACHARYYA

A distância Bhattacharyya, já mencionada anteriormente, está descrita em muitos textos de reconhecimentos de padrões como por exemplo em (FUKUNAGA, 1990). A seguir, apresenta-se uma descrição sucinta. Considera:

- ω_i: classe i, i = 1,2
- μ_i: vetor média da classe ω_i
- C_i: matriz covariância da classe i

A distância Bhattacharyya, d_B, mede a separabilidade entre duas distribuições gaussianas e é definida por (FUKUNAGA, 1990):

$$d_B = \frac{1}{2}\ln\frac{\frac{|C_i+C_j|}{2}}{|C_i|^{1/2}|C_j|^{1/2}} + \frac{1}{8}(\mu_i-\mu_j)^T(\frac{C_i+C_j}{2})^{-1}(\mu_i-\mu_j) \tag{4.2}$$

onde T indica transposto. De forma resumida temos:

$$d_B = d_C + d_M \tag{4.3}$$

onde d_C fornece a distância devido a diferença entre as matrizes covariância das classes e d_M fornece a separação entre as médias das classes.

Aplicando ao RAL, a distância d_C fornece uma medida da forma da evolução das características de voz e a distância d_M é a avaliação do nível DC de cada coeficiente (CAMPBELL, 1997).

Para o caso de RAL, cada classe seria composta pelos dados de um locutor e a comparação seria realizada diretamente entre o padrão de referência e o pretenso locutor. O esquema representativo da identificação é o mesmo mostrado na FIG. 4.1, mudando-se a distância para d_B, d_C ou d_M.

4.2 BASE DE DADOS UTILIZADA

Para realizar os testes de identificação de locutor com as compensações, foi utilizada a base de dados gravada no Laboratório de Voz do Instituto Militar de Engenharia. A base é composta de gravações de 50 locutores masculinos, cada um falando 20 grupos de 10 frases extraídas de (ALCAIM, 1992). As frases foram concatenadas e, após extraído o silêncio, foram divididas conforme segue:

- treinamento e teste: utilizando os primeiros 18 grupos
  - número de locutores de treinamento: foram utilizados 40 locutores de treinamento falando 1 min cada.
  - número de locuções de teste: foram utilizados 474 locuções de teste, gravadas pelos 40 locutores. Cada locução possui 30 segundos de fala. Dos 474 testes, 450 foram montados com as gravações do mesmo período da gravação das locuções de treinamento (Grupo I) e 24 gravados por quatro locutores com uma diferença de aproximadamente 6 meses (Grupo II). O texto das 24 gravações é formado pelos primeiros 10 grupos de 20 frases. Isto implica que este grupo de teste possui frases iguais às de treinamento.
- gravações para média da língua: foram utilizados os últimos 10 locutores e os dois últimos grupos de textos lidos, para que as locuções fossem diferentes das de treinamento e teste.

As locuções de treinamento foram filtradas pelo canal A e as de teste pelo canal B.

Diante dos resultados mostrados no capítulo anterior, TAB. 3.2 a 3.5, escolhemos analisar os tipos de extração de características (b), (d), (e) e (f), as quais passaremos a chamar de (20ms50%), (20ms75%), (40ms75%) e (40ms50%), respectivamente. Os valores 20 e 40 fazem referência ao tamanho da janela e os valores 50 e 75, a superposição.

4.3 RESULTADOS DA IDENTIFICAÇÃO DE LOCUTOR UTILIZANDO QV

A TAB. 4.1 apresenta a taxa de erros para a identificação de locutor sem compensação. Pode-se ver que a taxa de erros[8] é elevada devido ao descasamento entre os dados de treinamento e teste. Estes resultados são apresentados para se ter uma idéia da taxa de erros alcançada pelo descasamento de canal e para posterior comparação com as compensações.

Da tabela, pode-se perceber que não há uma característica que se destaque pela robustez ao descasamento de canal. Se tivéssemos que escolher uma característica, o MCC é a que apresenta as menores taxas de erros.

TAB. 4.1: Erro em % da identificação de locutor através quantização vetorial sem compensação

	MCC	LPCC	LFCC
20ms50%	79,54	81,86	80,80
20ms75%	79,96	80,80	78,90
40ms75%	79,11	83,12	84,60
40ms50%	81,01	82,70	84,18

A seguir serão apresentados os resultados e os comentários considerando-se as técnicas de compensação descritas no Capítulo 3. A TAB. 4.2 apresenta os resultados da taxas de erros obtidas com as três técnicas de compensação, para os quatro tipos de extração de características.

As comparações serão apresentadas através de gráficos que descrevem a taxa de melhora sobre a taxa de erros na identificação, obtida com o sinal sem compensação (VSC) e foi calculada pela equação:

$$M = \frac{VSC - VC}{VSC} \tag{4.4}$$

onde VC representa os valores da taxas de erros após a compensação.

[8] Razão entre o número de identificações erradas e o número total de teste

TAB. 4.2: Taxa de erro em % da identificação de locutor através quantização vetorial com o CMS, Proposto I e Proposto II

	CMS			Proposto I			Proposto II		
	MCC	LPCC	LFCC	MCC	LPCC	LFCC	MCC	LPCC	LFCC
20ms50%	1,05	0,42	0,21	0,42	0,21	0,42	0,63	0,42	0,21
20ms75%	0,42	0,63	0,21	0,42	0,63	0,00	0,42	0,63	0,21
40ms75%	0,84	0,42	0,21	0,84	0,42	0,21	0,84	0,42	0,21
40ms50%	1,05	0,21	1,27	0,84	1,05	0,42	0,84	1,05	0,42

4.3.1 COMPARAÇÃO ENTRE AS TÉCNICAS DE COMPENSAÇÃO

A FIG. 4.2 apresenta os valores de M para os experimentos com a QV, pode-se destacar:

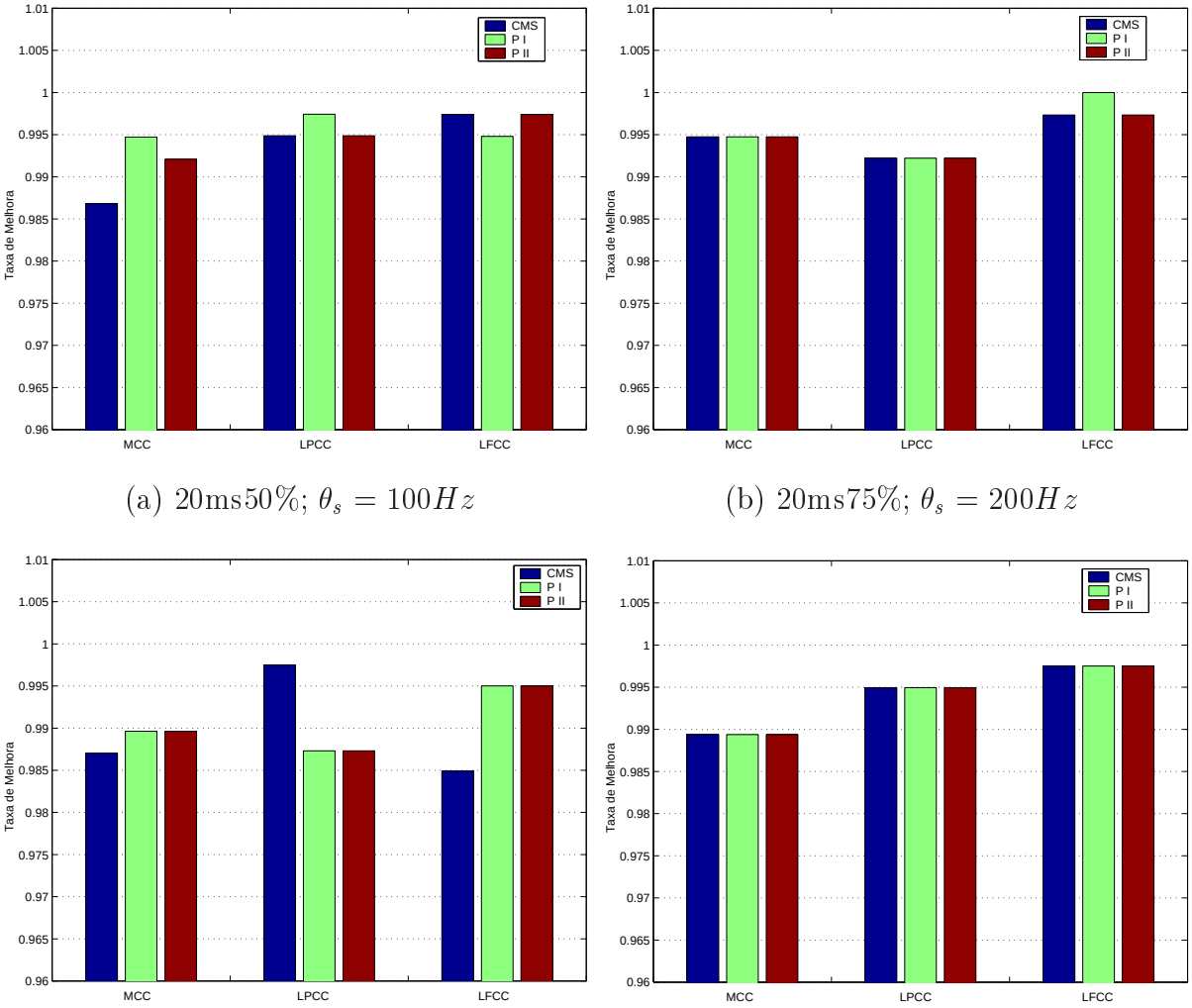
- As FIG. 4.2 (b) e (d) mostram que para a superposição de 75% não houve diferença entre as técnicas de compensação. Nas figuras (a) e (c) é mostrado que o reconhecimento utilizando sinais com superposição de 50 % torna-se mais sensível às técnicas de compensação. Isto pode ser devido a alguma compensação feita pelas técnicas Proposto I e II da perda de informação provocada pelo *aliasing* discutido no Capítulo 2.

- Para a superposição de 50% o Proposto I, em geral, obteve as melhores taxas de compensação. Já o Proposto II obteve valores que ficaram entre o CMS e o Proposto I.

- Pode ser observado que a taxa de quadros (θ_s), discutidas no Capítulo 2, influiu na taxa de compensação. Esta conclusão foi retirada observando que:

  - os quadros (a) e (d) que possuíam $\theta_s = 100Hz$ apresentaram taxas de compensação, na média, iguais.
  - as maiores taxas de compensação foram obtidas com $\theta_s = 200Hz$, mostrada na figura (b).
  - as menores taxas foram obtidas com $\theta_s = 50Hz$, figura (c).

  Isto mostra que o número de quadros no CMS e consequentemente nas técnicas Proposto I e II, junto com a modelagem do locutor feita na QV, são sensíveis à taxa de quadros.



(a) 20ms50%; $\theta_s = 100Hz$

(b) 20ms75%; $\theta_s = 200Hz$

(c) 40ms50%; $\theta_s = 50Hz$

(d) 40ms75%; $\theta_s = 100Hz$

FIG. 4.2: Comparação entre as técnicas CMS, P I e P II, por característica e por tipo de extração.

- A característica MCC obteve, em geral, as menores taxas de compensação. Para as extrações (20ms50%) e (40ms50%), as técnicas Proposto I e II, suplantaram o CMS em 1,5% (acerto de 3 locutores) e 1% (acerto de 2 locutores), respectivamente. Este desempenho pode ser devido às aproximações realizadas pelo banco de filtros, também discutida no Capítulo 2.

- O LFCC foi a característica que obteve as melhores taxas de compensação, com especial destaque para a técnica Proposto I com janela de 20 ms e superposição de 75%, onde ocorreu a compensação de todos os erros.

4.3.2 OBSERVAÇÕES SOBRE A DIFERENÇA DE TEMPO ENTRE AS SEÇÕES DE GRAVAÇÃO

Esta análise foi incluída pela sua importância para aplicações de reconhecimento que envolvam diferença de tempo entre os dados de treinamento e os dados de teste, como pode ser o caso de uma aplicação forense.

Nestes testes, inicialmente esperávamos obter uma degradação considerável com as locuções do Grupo II, visto que, em trabalhos anteriores e na literatura, foi observado que esses tipos de gravações prejudicam a taxa de acertos nos sistemas de RAL.

No entanto, ao avaliar os resultados utilizando o CMS e as técnicas propostas, ficamos surpresos, pois todas as locuções do Grupo II haviam sido compensadas. A seguir são descritos os resultados encontrados para cada técnica de compensação.

- **Observações com o CMS**

  Para podermos apresentar os resultados e tirarmos algumas conclusões, foi verificado o erro, em locutores, para cada grupo de teste Grupo I e Grupo II. Foi realizado um reconhecimento com o sinal limpo, porém com a limitação de banda entre 300 e 3400 Hz, a fim de termos um parâmetro de análise dos resultados com o CMS.

  As Tabelas 4.3 e 4.4 mostram que, com o sinal limpo, os erros concentram-se nas locuções do Grupo II, mostrando que o tempo influi no reconhecimento. Já para o reconhecimento com o sinal corrompido, o CMS produziu um efeito contrário, reduzindo os erros do segundo grupo a zero. De alguma forma o CMS conseguiu compensar o efeito do tempo, retirando o fator que introduzia os erros.

  Pode-se ver pelas tabelas que o LFCC, para as extrações (20ms50%), (20ms75%) e (40ms75%) manteve o número de erros tanto com o sinal limpo como para o corrompido para o Grupo I. O LFCC mostrou ser uma boa característica para se trabalhar em ambiente de descasamento acústico.

  Além disso pode ser observado que o MCC introduziu mais erros no primeiro grupo após o CMS, isso pode ser devido ao que já foi discutido anteriormente sobre o efeito do banco de filtros na estimação de canal e na compensação.

- **Observações com o Proposto I**

  Neste tópico, a tendência de compensar os dados do grupo II manteve-se igual aos resultados obtidos com o CMS, ou seja, não houve nenhum erro neste grupo. As

TAB. 4.3: Erro em nr de locutores da identificação de locutor através quantização vetorial com o sinal limpo e banda de 300 - 3400 kHz para os dois grupos de teste. A barra vertical dupla na tabela separa os erros por cada grupo, o primeiro com 450 e o segundo com 24 locuções

	Grupo I			Grupo II		
	MCC	LPCC	LFCC	MCC	LPCC	LFCC
20ms50%	0	1	1	4	6	7
20ms75%	0	1	1	3	6	8
40ms75%	0	1	1	2	7	9
40ms50%	0	1	1	3	6	10

TAB. 4.4: Erro em nr de locutores da identificação de locutor através quantização vetorial com o CMS para os dois grupos de teste. A barra na tabela separa os erros por cada grupo, o primeiro com 450 e o segundo com 24 locuções

	Grupo I			Grupo II		
	MCC	LPCC	LFCC	MCC	LPCC	LFCC
20ms50%	5	2	1	0	0	0
20ms75%	2	3	1	0	0	0
40ms75%	4	2	1	0	0	0
40ms50%	5	1	6	0	0	0

diferenças das percentagens mostrados na TAB. 4.2, devem-se totalmente aos dados do grupo I. Isto se deve ao fato da técnica proposta ser uma variação do CMS.

- **Observações com o Proposto II**

  Os resultados para este tópico são apresentados na TAB. 4.5. As diferenças nas taxas de erro foram devidas unicamente ao dados do grupo I. Isto pode ser devido ao fato de que no sinal de teste foi aplicado o CMS, sendo mostrado desta forma, que a compensação de tempo deve ocorrer basicamente no sinal de teste, não havendo necessidade de aplicá-lo nos dados de treinamento.

  Com a finalidade de verificar a hipótese da troca de canais, a TAB. 4.6 mostra os erros em número de locutores para um treinamento e teste realizado com o canal ITU e o sinal limitado pela banda de 300 a 3400 Hz.

  Podemos ver que os erros concentram-se principalmente no Grupo II de teste. Já no caso onde os filtros dos sinais de treinamento e teste são diferentes, após o CMS nos dados de teste os erros ficam restritos ao dados do grupo I.

  Para o LFCC as três primeiras formas de extração apresentaram os mesmos resul-

TAB. 4.5: Erro em nr de locutores da identificação de locutor através quantização vetorial com o sinal de compensado pelo método Proposto II. A barra vertical dupla na tabela separa os erros por cada grupo, o primeiro com 450 e o segundo com 24 locuções

	Grupo I			Grupo II		
	MCC	LPCC	LFCC	MCC	LPCC	LFCC
20ms50%	3	2	1	0	0	0
20ms75%	2	3	1	0	0	0
40ms75%	4	2	1	0	0	0
40ms50%	4	5	2	0	0	0

TAB. 4.6: Erro em nr de locutores da identificação de locutor através quantização vetorial com o sinal de treinamento e teste filtrados pelo canal ITU. A barra vertical dupla na tabela separa os erros por cada grupo, o primeiro com 450 e o segundo com 24 locuções

	Grupo I			Grupo II		
	MCC	LPCC	LFCC	MCC	LPCC	LFCC
20ms50%	0	1	1	5	6	6
20ms75%	0	1	1	1	6	8
40ms75%	0	1	1	1	6	8
40ms50%	0	1	1	4	6	7

tados com apenas um locutor errado para o Grupo I de teste. Isso mostra que para essa característica a hipótese de troca de canais e de cancelamento da perda de informação de locutor introduzida pelo CMS foi verificada. Uma hipótese para que o MCC e LPCC não tenham tido o mesmo desempenho e que elas possuem uma convergência mais lenta na estimação cega do canal, prejudicando o método.

4.4 IDENTIFICAÇÃO DE LOCUTOR COM A DISTÂNCIA BHATTACHARYYA

Para a distância Bhattacharyya os resultados serão apresentados segundo o tipo de extração de características, tipo de característica e por últimos serão mostrados algumas observações sobre a diferença de tempo entre as seções de gravação de treinamento de teste.

A TAB. 4.7 apresenta os resultados da identificação de locutor com a distância Bhattacharyya, sem compensação de canal. Dos dados apresentados pode-se ver que:

- a distância d_M possui um erro muito alto visto que o canal interfere principalmente no nível DC da evolução das características;
- a distância d_C mostra que a evolução das coeficientes *cepstrum* pouco se altera,

sendo desta forma uma característica robusta ao erro introduzido pelo canal. Esta distância será alvo de considerações a seguir.

TAB. 4.7: Erro em % da identificação de locutor através da Distância Bhatacharyya.

	MCC			LPCC			LFCC		
	d_B	d_C	d_M	d_B	d_C	d_M	d_B	d_C	d_M
20ms50%	7,17	**2,74**	97,26	12,87	**2,95**	98,52	2,32	**2,74**	97,05
20ms75%	7,38	**2,74**	97,05	12,66	**2,74**	98,52	2,32	**2,74**	97,05
40ms75%	4,43	**1,69**	97,26	9,07	**2,95**	98,95	1,9	**1,9**	97,47
40ms50%	9,49	**0,63**	79,32	21,94	**2,32**	83,54	18,78	**1,48**	82,91

Os resultados para as técnicas Proposto I e II, igualaram-se completamente ao caso do CMS; por isso, só será analisado o CMS. Isto é devido ao fato que eles tem como base o CMS. A TAB. 4.8 mostra os resultados da identificação com o CMS. Só estão apresentados os resultados de d_C uma vez que a média dos coeficientes sendo zero, não há sentido em apresentar d_M.

TAB. 4.8: Erro em % da identificação de locutor através da Distância Bhatacharyya aplicando CMS (valores idênticos a P I e P II). Valor da distância d_C.

	MCC	LPCC	LFCC
20ms50%	1,05	2,32	2,11
20ms75%	1,05	2,32	2,11
40ms75%	0,63	2,32	1,05
40ms50%	0,63	2,32	1,48

- **Tipo de Extração de Característica**

  Pode-se perceber que os tipos (40ms75%) e (40ms50%) apresentaram, com e sem compensação, os melhores resultados para todas as características. Isto pode ser devido ao maior tamanho da janela produzir uma melhor definição do espectro, pois quanto maior a janela, melhor a resolução do espectro.

  Após aplicar o CMS, o tamanho da janela foi o que mais interferiu; a janela de 40 ms teve os melhores resultados. Pelas duas tabelas, pode-se ver que o tipo (40ms50%) mostrou ser o mais robusto, visto que, com ou sem compensação, obteve os mesmos resultados, ou seja, não sofreu alteração na taxa de erros.

  Outro ponto interessante é que, para as três características, após o CMS, as taxas de erros para os mesmos tamanhos de janelas tenderam a se igualar. O CMS, de

alguma forma, realizou uma filtragem na evolução dos coeficientes, retirando o que estava causando a diferença entre os tipos de mesmo tamanho de janela.

- **Tipo de Característica de Voz**

  A característica que obteve o melhor desempenho foi o MCC. Isto pode ser devido ao fato dele realizar um amaciamento diretamente do espectro.

  Após o CMS o LFCC foi sensível a superposição, tendo a superposição de 75% obtido uma melhor resultado.

- **Diferença de Tempo Entre Treinamento e Teste**

  A TAB. 4.10 apresenta os resultados por grupo de teste com o CMS para a identificação com a distância Bhattacharyya. Pode-se observar que os maiores erros concentram-se no grupo I. Diferentemente do que ocorreu na QV, mesmo após o CMS, para o MCC e o LFCC houveram erros no grupo II. O LPCC consegui compensar as diferenças para o grupo II, porém aumentou bastante os erros do grupo I.

TAB. 4.9: Erro em nr de locutores da identificação de locutor através da Distância Bhattacharyya com o CMS para os dois grupos de teste.

	Grupo I			Grupo II		
	MCC	LPCC	LFCC	MCC	LPCC	LFCC
20ms50%	10	11	7	3	3	6
20ms75%	10	10	7	3	3	6
40ms75%	2	11	6	6	3	3
40ms50%	1	11	5	2	0	2

TAB. 4.10: Erro em nr de locutores da identificação de locutor através da Distância Bhattacharyya com o CMS para os dois grupos de teste.

	Grupo I			Grupo II		
	MCC	LPCC	LFCC	MCC	LPCC	LFCC
20ms50%	3	11	9	2	0	1
20ms75%	3	11	9	2	0	1
40ms75%	1	11	4	2	0	1
40ms50%	1	11	5	2	0	2

4.5 RESUMO

4.5.1 QUANTIZAÇÃO VETORIAL

Para a quantização vetorial pode-se concluir que:

- as técnicas Proposto I e II apresentaram, na maiorias dos casos, melhora quando os sinais de voz utilizaram superposição de 50%;
- os melhores resultados de compensação, para as três características, foram obtidas com janelas de 20 ms e superposição de 75%;
- o LFCC foi a característica que obteve as melhores taxas na compensação para a superposição de 75%;
- todas as três técnicas conseguiram compensar as diferenças de tempo entre as seções de gravação dos dados de treinamento e teste.
- a técnica Proposto II, conseguiu atingir, com o LFCC, as mesmas taxas de erros obtidas com os sinais de treinamento e teste filtrados pelo canal A.

4.5.2 DISTÂNCIA BHATTACHARYYA

Para a Distância Bhattacharyya pode-se concluir que:

- a medida da forma da evolução dos coeficientes é uma medida robusta à distorção de canal, visto que o canal afeta principalmente o nível DC da evolução dos coeficientes;
- o MCC mostrou-se ter o comportamento mais estável para a obtenção da forma da evolução das características;
- o CMS, aparentemente realiza alguma filtragem na evolução das características, levando a uma melhor resultado para todas as características;
- as técnicas propostas não alteram os resultados obtidos com CMS.
- a extração (40ms50%), 40 ms e superposição 50 ms, obteve o melhor desempenho, sendo um tipo robusto à distorção de canal, visto que com ou sem compensação obteve os mesmo resultado na taxa de erro;
- a distância Bhattacharyya não consegue retirar totalmente o efeito da diferença de tempo entre as gravações de treinamento e teste.

PARTE II

Análise Fractal Aplicada ao Reconhecimento de Locutor