UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

EXTRAÇÃO DE CARACTERÍSTICAS DO SINAL DE VOZ UTILIZANDO ANÁLISE FATORIAL VERDADEIRA

ADRIANO NOGUEIRA MATOS

MANAUS
2008

FEDERAL UNIVERSITY OF AMAZONAS
EXACT SCIENCES INSTITUTE
POSTGRADUATE PROGRAM IN INFORMATICS

SPEECH SIGNAL FEATURE EXTRACTION USING TRUE FACTORIAL ANALYSIS

ADRIANO NOGUEIRA MATOS

MANAUS
2008

UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

ADRIANO NOGUEIRA MATOS

EXTRAÇÃO DE CARACTERÍSTICAS DO SINAL DE VOZ UTILIZANDO ANÁLISE FATORIAL VERDADEIRA

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal do Amazonas, como requisito parcial para a obtenção do título de Mestre em Informática, área de concentração Engenharia da Computação.

Orientador: Prof. Dr. José Luiz de Souza Pio

MANAUS
2008

Às minhas amadas esposa e filha, Ana Flávia e Clara.

Aos meus pais, Eríco e Aliane. Aos meus irmãos, Leonardo, Davi e Eduardo. Ao tio Eduardo e aos amigos de quem estive afastado enquanto realizava este trabalho.

Agradecimentos

Agradeço à Jesus, pois foi Seu Nome que eu clamei pedindo ajuda e, agora, estou aqui, contando essa história;

Ao meu irmão Leonardo, pelo exemplo e incentivo, que me levaram a buscar a realização deste objetivo;

Ao meu orientador, José Pio, pela indicação do caminho a seguir e pela liberdade com que me deixou trilhá-lo;

Ao Prof. Dr. Ynoguti, por permitir a utilização de sua base de locuções, imprescindível para realização deste trabalho;

Ao Prof. Dr. Fernando Gil, pelo apoio e por me colocar em contato com outros pesquisadores da área de reconhecimento de voz;

Ao Prof. Dr. Rogério Caetano, por me proporcionar a oportunidade de trabalhar em um projeto de pesquisa em reconhecimento de voz;

À Fundação Des. Paulo Feitoza e ao Instituto Nokia de Tecnologia, pelo apoio no desenvolvimento deste trabalho;

Ao pesquisador Dr. Paulo Esquef, pelas instigantes conversas sobre processamento de sinais;

Ao pesquisador Monik Jan, da lista de usuários HTK, que me ensinou como debugar os aplicativos HTK;

Ao colega Gilberto Martins, pelo apoio e por me ter disponibilizado seus modelos LaTeX, que utilizei na escrita deste trabalho;

À Edina, meu braço-direito em casa, que me permitiu ter tempo para dedicar ao mestrado.

Resumo

O processamento digital do sinal de voz é empregado em diversas aplicações computacionais, das quais as principais são: Reconhecimento, síntese e codificação da fala. Todas estas aplicações requerem que ocorra redução da quantidade de informações da onda acústica, de maneira a permitir o processamento por um computador. O processo de extração de características do sinal de voz, objeto de estudo deste trabalho, realiza esta tarefa. As características extraídas devem caracterizar o sinal de voz e não conter redundância, de forma a maximizar o desempenho dos sistemas que as utilizem. O método MFCC (*Mel Frequency Cepstral Coefficients*) de extração de características cumpre parcialmente esses requisitos, mas é seriamente degradado sob a incidência de ruído. A aplicação do método estatístico de Análise Fatorial objetiva filtrar o sinal de ruído das locuções. Os resultados obtidos dos experimentos realizados indicam a competitividade deste método, especialmente quando usado na geração dos modelos acústicos robustos em condições de ruído severo.

Palavras-Chave: Extração de características do sinal de voz, Filtragem de ruído, MFCC, Análise Fatorial.

Abstract

Digital processing of speech signal is applied in several computer applications, which the major ones are the following: Recognition, synthesis and coding of speech. All these applications require the amount of data in the acoustic signal to be reduced, in order to allow processing by a computer device. The feature extraction of speech signal, that is the goal of this study, performs this action. The features extracted should well depict the speech signal and should have no redundancy, in order to increase the performance of the systems using them. The feature extraction Mel Frequency Cepstral Coefficients (MFCC) method partially fulfills these requirements, but it is seriously damaged when noise signal is acting. The appliance of the statistical method of Factorial Analysis is intended to filter the noise components from the speech. The results of the experiments performed in this work shows that this is a competitive method, especially when used to generate acoustic models in severe noise conditions.

Key-words: Speech signal feature extraction, Noise filtering, MFCC, Factorial Analysis.

Lista de Figuras

Lista de Tabelas

Lista de Siglas

Sigla	**Significado**
ASR	*Automatic Speech Recognition*
HMM	*Hidden Markov Model*
LP	*Linear Prediction*
MFCC	*Mel Frequency Cepstral Coeficients*
WRR	*Word Recognition Rate*
RASTA	*Relative Spectra*
PLP	*Perceptual Linear Predictive*
FA	*Factorial Analysis*
PCA	*Principal Component Analysis*
HTK	*Hidden Markov Toolkit*
DCT	*Discrete Cosine Transform*
SNR	*Signal Noise Ratio*

Tabela 1: Lista de siglas

Sumário

Capítulo 1

Introdução

Esta dissertação trata do problema de extração de características do sinal de voz para fins de reconhecimento automático de fala (*ASR - Automatic Speech Recognition*). O reconhecimento de fala consiste no processo de conversão do sinal acústico em um conjunto de palavras. As palavras reconhecidas podem ser usadas em aplicações de controle, entrada de dados, preparação de documentos e processamento linguístico [V. Zue and Ward, 1996].

As características extraídas do sinal de voz devem ser suficientes para permitir a correta classificação do fonema, mesmo em presença de variáveis dependentes de locutor e do ambiente, tais como: tímbre de voz, modo de pronúncia, entonação, ruído e fala concorrente. Ainda, devem ser em quantidade tal que apenas os dados relevantes para o propósito de reconhecimento sejam apresentados ao sistema reconhecedor, de maneira a não aumentar o custo computacional resultante [Bozzeto, 2004].

O problema de extração de características do sinal de voz insere-se no escopo da área de pesquisa de *Processamento Digital de Sinais*. O propósito das pesquisas nessa área é a realização de operações sobre sinais representados por sequências numéricas. Esse formato é requerido para possibilitar o processamento dos sinais por um computador. Os sinais são obtidos, normalmente, por amostragem, quantização e codificação de sinais contínuos, como a fala [Oppenheim et al., 1999].

O processamento do sinal de voz objetiva, principalmente, a utilização em aplicações de reconhecimento, síntese e compressão de fala. Todas estas aplicações demandam redução da quantidade original de dados para uma escala menor, contendo apenas os parâmetros mais representativos do sinal de voz. O processo de extração de características, objeto de estudo deste trabalho, realiza essa tarefa.

O problema de extração de características do sinal de voz visando aplicação em reconhecimento automático de fala é objeto de estudo desde muitas décadas atrás. Nessa trajetória, diversos algoritmos, técnicas e abordagens foram propostas e avaliadas à medida que o entendimento dos processos envolvidos foram sendo desvendados.

Os primeiros trabalhos na área relacionavam as características do sinal de voz à posição dos formantes e à medidas de energia, obtidas da análise espectral de um segmento de curta duração do sinal de voz [Olson and Belar, 1956]. O processo de reconhecimento de fonemas, nesses sistemas, dava-se por verificação dessas características em um sistema de classificação, baseado em árvore de decisão binária [John R. Deller et al., 1993]. Devido à variabilidade acústica do sinal de voz, porém, o desenvolvimento de um sistema ASR baseado nessas características apresenta pouca robustez às diversas fontes de interferência a que é submetido o sinal de voz. As fontes de variabilidade são devidas principalmente ao ambiente (ruído e fala concorrente), ao locutor (tímbre de voz) e ao fenômento de *coarticulação*, que se caracteriza como a modificação da configuração acústica de um fone devido aos fones que o precedem e o sucedem em uma palavra. No estado-da-arte atual, a abordagem de classificação fonética baseada na posição dos formantes e medidas de energia não é mais utilizada.

Apartir do início dos anos setenta, a técnica de predição linear (*LP - Linear Prediction*) foi introduzida como método de análise do sinal de voz [Makhoul, 1973]. O conjunto de características resultantes da aplicação desta técnica corresponde aos coeficientes de um filtro *só-polos* que modela a forma de onda do segmento de sinal de voz. Essa técnica de extração de características foi, durante muito tempo, a mais utilizada em aplicações de reconhecimento de fala [Rabiner, 1989].

A análise do sinal de voz por meio de banco de filtros é o fundamento da técnica MFCC de extração de características *(Mel-Frequency Cepstral Coeficients)* [Davis and Mermelstein, 1990]. Essa técnica deriva de experimentos da percepção acústica por seres humanos [Schroeder, 1977] que indicam que a energia do sinal de voz em bandas de frequência distintas podem ser usadas para derivar elementos característicos do sinal de voz. A utilização dessa técnica em aplicações de reconhecimento automático de fala resultou no desenvolvimento de sistemas ASR estado-da-arte como BYBLOS [Schwartz et al., 1989] e SPHINX [Lee et al., 1990].

Deve-se destacar que o expressivo aumento de performance de sistemas ASR verificados nas últimas décadas é devido, principalmente, ao uso de sistemas de classificação baseados em modelos escondidos de Markov *(HMM - Hidden Markov Model)*[Rabiner, 1989]. Nos nossos dias, podem ser encontradas aplicações comerciais de sistemas ASR baseados em HMM orientados ao reconhecimento de palavras isoladas e vocabulário pequeno. O foco da pesquisa científica, porém, aponta para a aplicação de sistemas ASR em sistemas independentes de locutor, fala contínua e vocabulário muito grande.

Diversas outras técnicas de extração de características apresentadas na literatura são derivadas das técnicas LP e MFCC, citadas nos parágrafos anteriores. Dentre estas, podem ser citadas as técnicas PLP (*Perceptual Linear Predictive*) [Hynek Hermansky and Kohn, 1990] e RASTA (*Relative Spectra*) [Hermansky and Morgan, 1994]. Os experimentos conduzidos neste trabalho utilizam o método MFCC. Os detalhes relativos à esta técnica de extração de características são apresentados no próximo capítulo.

1.1 Motivação

A motivação principal deste trabalho é a busca pela inclusão dos portadores de deficiências físicas na sociedade da informação, por meio do desenvolvimento de interface de usuário por comando de voz.

Os portadores de deficiências físicas e de visão experimentam dificuldades na uso das interfaces *mouse* e teclado, que requerem alguma destreza motora para sua manipulação. Esse segmento da sociedade encontra-se, portanto, restringido para usufruir dos benefícios disponíveis na era da informação. Ainda, deve-se observar que a familiaridade na utilização de recursos computacionais tornou-se requisito de empregabilidade em muitos segmentos do mercado de trabalho.

Por outro lado, o aumento da capacidade de processamento e armazenamento dos computadores, aliado à crescente miniaturização dos dispositivos micro-processados, criou um segmento do mercado de eletrônicos bastante lucrativo. O desenvolvimento de uma interface de usuário por voz promoveria, consequentemente, o atendimento de uma demanda reprimida da indústria eletrônica para o consumo desses bens pela comunidade dos portadores de necessidades especiais. Percebe-se, portanto, que a inserção destes cidadãos na sociedade da informação atende à critérios não apenas humanitários, mas também econômicos.

1.2 Justificativa

Neste trabalho é apresentada uma técnica de extração de características do sinal de voz orientado à filtragem de ruído ambiente. Essa técnica é baseada na remoção das comunalidades, resultantes da aplicação do método de Análise Fatorial Verdadeira [Reyment and Jöreskog, 1996]. Essa técnica é validadada em tarefas de reconhecimento de fala, utilizando locuções corrompidas por ruído de naturezas diversas, como: ruído branco, ruído de máquina, ruído de conversação e outros. Os resultados obtidos indicam a aplicabilidade dessa técnica no contexto de ambientes ruidosos.

Sistemas ASR tratam do problema de presença de ruído, de duas formas, basicamente: Pela aplicação de métodos para remoção do sinal de ruído, em fase anterior à de criação de modelos acústicos e por construção de modelos acústicos robustos, através de treinamento utilizando bases de dados ruidosas. Neste trabalho, é utilizada esta última abordagem.

Dentre as várias aplicações potenciais de uso de sistemas ASR, há aquelas ocorrendo em am-

bientes não-controlados, sujeitas à ruído de diferentes fontes. A presença de ruído degrada o sinal de voz e faz decrescer a taxa de reconhecimento de palavras. A técnica desenvolvida nesse trabalho justifica-se, portanto, por buscar uma solução para o problema de reconhecimento automático de fala que desloca a aplicação ASR para o dia-a-dia das pessoas, removendo-a do ambiente de laboratório.

Este trabalho justifica-se também pela investigação do potencial de uso do método estatístico de Análise Fatorial Verdadeira, em sistemas ASR. Outros métodos de análise multivariada, como PCA, são usados costumeiramente em processos de extração de características de sinais. A investigação presente neste trabalho apresenta uma nova área de pesquisa a ser desenvolvida em trabalhos futuros.

1.3 Objetivos

Os objetivos almejados neste trabalho estão divididos em geral e específicos, como apresentados a seguir:

1.3.1 Objetivo Geral

Este trabalho objetiva apresentar o desenvolvimento de um método de extração de características do sinal de voz baseado na filtragem das comunalidades, resultantes da aplicação do método de Análise Fatorial Verdadeira.

1.3.2 Objetivos Específicos

Os objetivos específicos são apresentados a seguir:

- Criação de um sistema ASR para a língua portuguesa, com vocabulário de tamanho médio, independente de locutor e modo de pronúncia contínua, utilizando a base de locuções

desenvolvida por Ynoguti [Ynoguti, 1999] e o pacote de softwares HTK [Odell et al., 1995];

- Criação de bases de locuções ruidosas utilizando os arquivos de ruído da base de dados RSG-10 [Steeneken and Geurtsen, 1988];

- Avaliação da robustez do método de extração de características desenvolvido frente à incidência de ruído de diversas fontes.

1.4 Organização do Trabalho

O texto está organizado como segue: O Capítulo 2 apresenta a fundamentação teórica das técnicas de processamento de sinais e análise multivariada, empregadas neste trabalho. No Capítulo 3 são apresentados os trabalhos relacionados com este desenvolvido nesta dissertação. O Capítulo 4 apresenta a metodologia empregada no desenvolvimento deste trabalho de pesquisa. Experimentos e resultados dos testes realizados são apresentados no Capítulo 5. Finalmente, no Capítulo 6, são apresentadas as conclusões e as propostas para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta os fundamentos das técnicas de transformação Cepstral e análise multivariada de dados utilizando Análise Fatorial Verdadeira. Esses métodos formam a base teórica sobre os quais este trabalho é fundamentado.

2.1 A Transformação Cepstral

A transformação Cepstral pertence à categoria de métodos agrupados sob a classificação de *processamento homomórfico de sinais* [John R. Deller et al., 1993]. Esses métodos são utilizados para analisar sinais formados por relacionamentos descritos por uma generalização do *princípio da superposição*. O objetivo da aplicação da transformação Cepstral, neste trabalho, é selecionar do sinal de voz, o sinal sistema-trato vocal, para dele gerar o vetor de características do *frame* sinal de voz. Os detalhes que fundamentam essa técnica são apresentados nas próximas seções.

2.1.1 Análise Espectral do Sinal de Voz

A voz é uma onda de pressão acústica produzida como reação à passagem de ar, proveniente dos pulmões, pelas cavidades do trato vocal/nasal. A fala é modelada por movimentos voluntá-

rios dos orgãos articuladores do sistema de produção da fala, que constituem-se da mandíbula, lingua, dentes, lábios e velum. O posicionamento instantâneo destes orgãos configuram o trato vocal/nasal em suas características de comprimento e formato. A configuração física do trato vocal/nasal combinado à onda de excitação, originária dos pulmões, produz todos os fonemas [John R. Deller et al., 1993].

A onda de excitação pode apresentar características de periodicidade ou aleatoriedade, decorrente de o ar passar ou não pelas cordas vocais. No primeiro caso, as cordas vocais vibram interrompendo a passagem de ar para o trato vocal em intervalos regulares de tempo, denominado *período fundamental*. Os fones produzidos sob esta forma de excitação são chamados *vocalizados*, e dentre estes destacam-se as vogais. Os sons *não-vocalizados* ocorrem pela passagem de ar turbulento por estrangulações do trato vocal.

A modelagem do sistema trato vocal objetiva descrever matematicamente o sinal de voz. Pode-se demonstrar [John R. Deller et al., 1993] que, considerando não ocorrerem perdas pela passagem de ar pelo trato vocal, o sistema de produção da fala pode ser modelado por um *sistema linear e invariante no tempo* submetido à duas fontes de excitação: periódica e aleatória. A Figura 2.1 apresenta uma representação do modelo simplificado do sistema de produção de fala.

A modelagem do trato vocal por meio de um sistema linear e invariante no tempo é o fundamento da técnica MFCC de extração de características.

Um sistema linear e invariante no tempo transforma o sinal de entrada pela convolução deste com o sinal de resposta ao impulso do sistema. No domínio da frequência, esta transformação é equivalente à multiplicação das representações espectrais dos sinais de entrada e de resposta em frequência do sistema, obtidos com o emprego da transformação de Fourier. A análise do sinal de voz é realizado, normalmente, no domínio da frequência. Isso se deve à maior robustez às fontes de variabilidade que as características do sinal de fala apresentam nesse domínio [Rivarol Vergin and Farhat, 1999], o que favorece a utilização destas em aplicações de reconhecimento de fala.

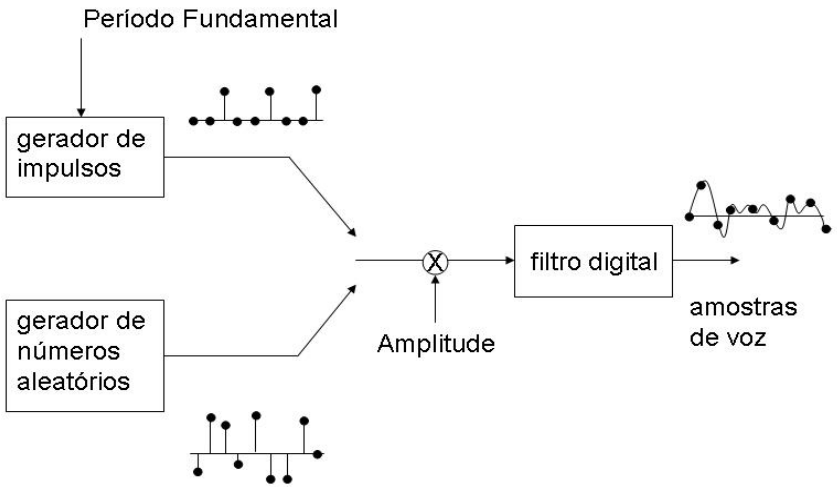

Figura 2.1: Modelo simplificado do sistema de produção de fala. Representação baseada em [Schafer and Rabiner, 1990].

A transformação Cepstral objetiva modelar o sistema trato vocal e utilizar os parâmetros derivados desse modelo como características representativas do sinal de voz. No entanto, devido à operação de convolução, os sinais do sistema trato vocal e de excitação encontram-se espalhados por todo espectro de frequência.

A observação do espectro de um sinal de fala vocalizado, Figura 2.2, permite observar o sinal de excitação modulado sobre o componente do sinal resposta em frequência do sistema trato vocal. É distinguivel também, que aquele primeiro sinal varia muito mais rapidamente que este último. Isto pode ser comprovado pela observação do sinal de *envelope*, Figura 2.3, que corresponde ao componente devido ao trato vocal, obtido aumentando a resolução espectral da transformação de Fourier.

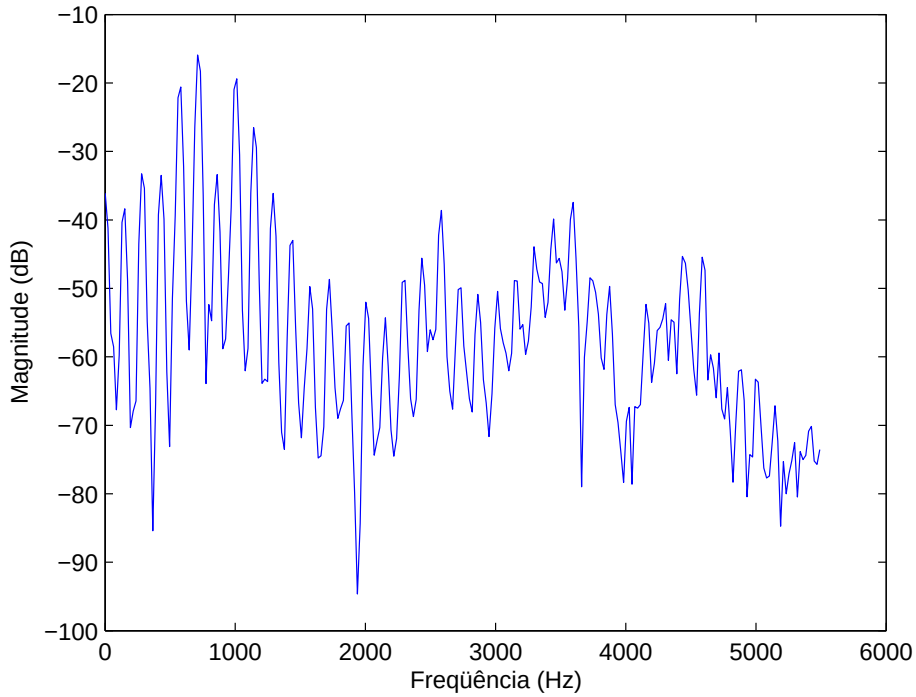

Figura 2.2: Espectro de potência de um segmento de sinal de voz correspondente à pronúncia do fonema /a/.

2.1.2 Os Coeficientes MFCC

A separação dos componentes do sinal de voz é obtida aplicando a função de logarítimo sobre a magnitude do espectro do frame de voz. Esta operação transforma a relação multiplicativa dos espectros dos sinais componentes em uma relação aditiva, conforme apresentado abaixo:

$$s[n] = e[n] * \theta[n] \rightleftarrows S(\omega) = E(\omega)\Theta(\omega), \tag{2.1}$$

$$C_s(\omega) = \log|S(\omega)| = \log|E(\omega)\Theta(\omega)|, \tag{2.2}$$

$$C_s(\omega) = \log|E(\omega)| + log|\Theta(\omega)|, \tag{2.3}$$

$$C_s(\omega) = C_e(\omega) + C_\theta(\omega). \tag{2.4}$$

Aplicando a transformação de Fourier sobre o sinal obtido da operação acima, obtem-se sinais

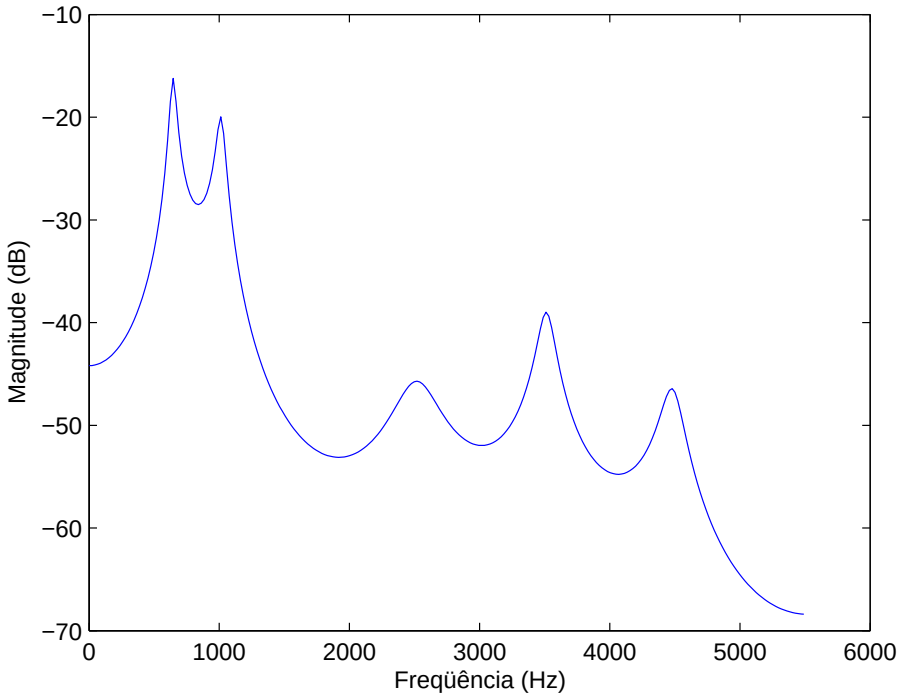

Figura 2.3: Sinal-Envelope de um segmento de sinal de voz correspondente à pronúncia do fonema /a/.

relacionados linearmente no novo domínio denominado *Cepstrum*. As amostras correspondentes ao trato vocal compreendem os componentes de baixa ordem da sequência cepstral. Esta característica deriva da observação de que os sinais componentes do sinal de voz apresentam componentes harmônicas distintas no domínio espectral, então, esses sinais devem ocupar bandas distintas no domínio cepstral [John R. Deller et al., 1993].

$$c_s[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} C_s(\omega)e^{\jmath\omega n}, \tag{2.5}$$

$$c_s[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} (C_e(\omega) + C_\theta(\omega))e^{\jmath\omega n}, \tag{2.6}$$

$$c_s[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} C_e(\omega)e^{\jmath\omega n} + \frac{1}{2\pi}\int_{-\pi}^{\pi} C_\theta(\omega)e^{\jmath\omega n}, \tag{2.7}$$

$$c_s[n] = c_e[n] + c_\theta[n]. \tag{2.8}$$

A implementação do método que deriva os coeficientes MFCC utiliza, no entanto, além dos conceitos de Transformação Cepstral apresentado acima, de resultados de experimentos de percepção acústica realizados em seres humanos por Stevens [Stevens and Newmann, 1937] e Schroeder [Schroeder, 1977].

Dois resultados desses experimentos são utilizados na derivação do método MFCC: Primeiro, verificou-se que a percepção de frequências acústicas não ocorre de forma linear em todo o domínio das frequências. A frequência real é percebida pelos seres humanos segundo um padrão que é aproximado pela expressão:

$$f_{mel} = 2595 \log_{10} \left(1 + \frac{f}{700} \right). \tag{2.9}$$

A frequência percebida define o que se convencionou denominar de *escala mel*. O segundo resultado desses experimentos é a constatação de que a energia em uma *banda crítica* em torno de uma frequência de interesse influi na percepção acústica desta frequência.

A implementação MFCC original [Davis and Mermelstein, 1990] usa um banco de filtros triangulares, Figura 2.4, espaçados linearmente de 100 *mels* no *range de Nyquist*, para acumular as medidas de energia nas bandas críticas.

A energia acumulada em cada filtro MFCC é dado por:

$$Y_i = \sum_{k=0}^{N/2} |S[k]| H_i \left(\frac{k2\pi}{N} \right), \quad i \in [1, N_{cb}], \tag{2.10}$$

onde N é o tamanho do segmento de sinal, $|S[k]|$ é a magnitude do espectro do segmento de voz, H_i corresponde à resposta em frequência do i-ésimo filtro e N_{cb} é a quantidade de bandas críticas no range de Nyquist.

A função logarítimo é aplicada então sobre a sequência de energia.

Os *coeficientes MFCC* são obtidos pela aplicação da transformada inversa de Fourier, ou, uma vez que a sequência de energias é simétrica, pela aplicação da transformada dos Cossenos, como

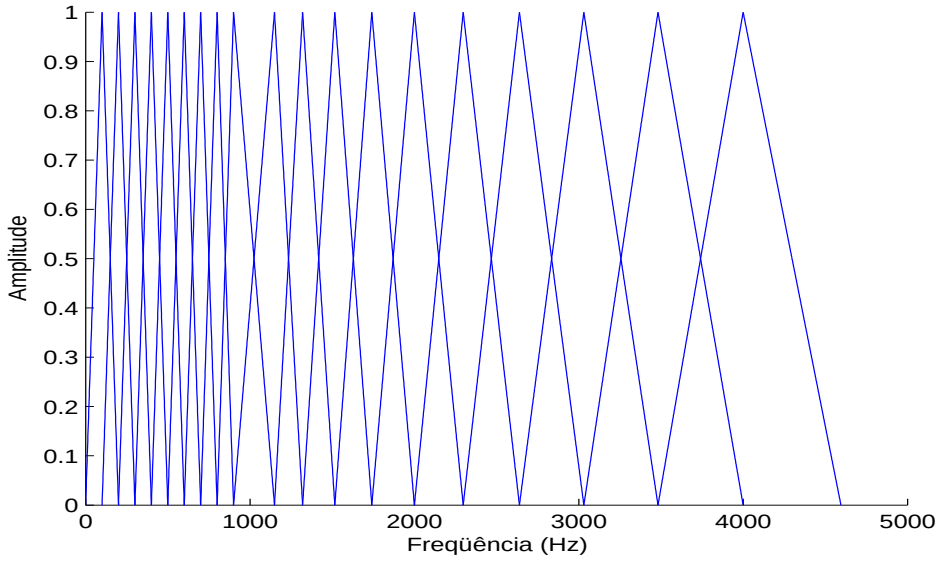

Figura 2.4: Banco de filtros triangulares espaçados linearmente em 100 *mels*.

apresentado abaixo:

$$c_s(m) = \frac{1}{N} \sum_{k=1}^{N_{cb}} \log(Y_k) \cos\left(k \frac{\pi}{N_{cb}} (m - 0.5k) \right). \tag{2.11}$$

2.2 Análise Fatorial

A análise estatística de um processo aleatório é realizada utilizando métricas que descrevem as probabilidades de ocorrências de eventos gerados por esse processo. As informações contidas nestas métricas permitem conhecer características do fenômeno aleatório subjacente e serem usadas em aplicações de predição e classificação de eventos.

Processos aleatórios unidimensionais são descritos, primordialmente, pelas métricas *média* e *variância* que representam, respectivamente, o valor esperado e o desvio do valor da média esperado de uma realização do processo aleatório. Processos aleatórios multivariados são descritos também por medidas extendidas dos valores unidimensionais, como o vetor de médias e o vetor de variâncias. Comumente, as variáveis de processos aleatórios multidimensionais são inter-

relacionadas. O grau de conexão entre estas variáveis é expresso pelas métricas *covariância* e *correlação*.

A grande quantidade de métricas necessárias para modelagem de um processo multidimensional constitui-se em um empecilho para a aplicação da análise estatística. A técnica de *Análise Fatorial* é empregada, nesse contexto, para encontrar uma representação simplificada do processo aleatório, enquanto preservando a informação essencial do conjunto de dados. Normalmente, uma quantidade significativamente menor de fatores que a quantidade de variáveis originais são suficientes para representar com precisão a estrutura de variabilidade dos dados [Mingoti, 2005]. A técnica de Análise Fatorial apresenta, portanto, característica de redução da dimensionalidade do conjunto de dados.

Para a realização da análise estatística de dados é necessário se dispôr de múltiplas realizações do processo investigado. Estas ocorrências são usadas, inicialmente, para criação de métricas, por meio de contagem de eventos. Um processo multidimensional gera realizações compostas de muitas variáveis, que conjuntamente constituem o *vetor de observações*. A reunião desses vetores de observações compõe a *matriz de dados*. A técnica de Análise Fatorial visa transformar o vetor de observações do domínio das variáveis originais para o domínio dos *fatores*, que equivalem à novas variáveis geradas pela transformação fatorial.

O modelo fatorial é representado matematicamente pela equação [Reyment and Jöreskog, 1996]:

$$Y_{(Nxp)} = F_{(Nxk)} A'_{(kxp)} + E_{(Nxp)}. \tag{2.12}$$

A variável $Y_{(Nxp)}$ é a matriz de dados, composta de *N* observações de *p* variáveis. A variável $F_{(Nxk)}$ é a matriz de escores e corresponde à quantização das observações no espaço dos fatores. $A_{(kxp)}$ é a matriz de pesos que quantifica o relacionamento das variáveis originais com o modelo fatorial. A matriz $E_{(Nxp)}$ contém os resíduos não explicados pelos fatores, e são assumidos, por definição, não relacionados à esses.

Re-ordenando a equação fatorial, obtem-se:

$$FA' = Y - E. \tag{2.13}$$

Supondo conhecida a matriz de resíduos, a solução do método fatorial corresponderia à solução do problema de decomposição matricial da matriz $Y - E$. Diversos pares de matrizes satisfazem este problema. No entanto, dado o objetivo de simplificação do modelo de dados, espera-se que a matriz de pesos **A** tenha *rank* menor (**k**), ($k \ll p$). Pode-se demonstrar, [Reyment and Jöreskog, 1996], que a matriz formada pela combinação linear dos *k* maiores autovalores, e autovetores correspondentes, não-nulos, de $[Y - E]'[Y - E]$, conforme apresentado abaixo, é solução ótima de *rank* **k**, do problema de aproximação da matriz **Y-E**.

$$Y - E \approx \gamma_1 v_1 u_1' + \gamma_2 v_2 u_2' + \ldots + \gamma_k v_k u_k'. \tag{2.14}$$

Os termos apresentados na expressão anterior advêm da decomposição de **Y-E** em valores singulares, utilizando o *teorema de Eckart-Young*. Esse teorema diz respeito à estrutura em que se decompõe uma matriz retangular qualquer em valores singulares, conforme abaixo:

$$Y - E = V\Gamma U'. \tag{2.15}$$

A matriz Γ é diagonal e seus elementos são as raízes quadradas dos autovalores positivos, não-nulos de $[Y - E]'[Y - E]$, dispostos em ordem decrescente. Os vetores linha da matriz *U'* são os autovetores correspondentes aos elementos de Γ e a matriz *V* resulta da operação linear:

$$V = [Y - E]U\Gamma^{-1}. \tag{2.16}$$

A melhor aproximação de $[Y - E]_{(Nxp)}$ de *rank* **k**, $(k < p)$, é, portanto, dada por:

$$Y - E \approx V_k \Gamma_k U_k', \tag{2.17}$$

onde:

- A matriz Γ_k é diagonal e seus elementos são as raízes quadradas dos *k* maiores autovalores positivos, não-nulos de $[Y - E]'[Y - E]$, dispostos em ordem decrescente;
- Os vetores linha de U_k' são autovetores de $[Y - E]'[Y - E]$ correspondentes aos elementos de Γ_k;
- Os vetores coluna da matriz V_k provêm da matriz V e são relacionados aos valores dos autovalores e autovetores conforme a Equação (2.16).

Pode-se deduzir que os termos relacionados aos autovalores de $[Y - E]'[Y - E]$ aproximam tanto mais a matriz **Y-E** ou, explicam a estrutura de variâncias e covariâncias dessa matriz, quanto maiores seus valores numéricos.

A proporção da variância explicada pelo *j-ésimo* fator equivale à:

$$\%[VAR]_j = \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}. \tag{2.18}$$

Os termos λ correspondem aos autovalores de $[Y - E]'[Y - E]$.

A análise gráfica da transformação fatorial corresponde à mudança do espaço original, cartesiano, dos dados multivariados para um espaço cartesiano de dimensão menor, tal que as direções de maior variabilidade dos dados coincidem com os eixos coordenados do novo espaço amostral.

A Figura 2.5 ilustra este conceito para o caso bidimensional.

Variantes da técnica de Análise Fatorial diferem no que determinam para o conteúdo da matriz de resíduos. Na seção seguinte são apresentadas as particularidades da técnica *Análise Fatorial Verdadeira*, utilizada neste trabalho.

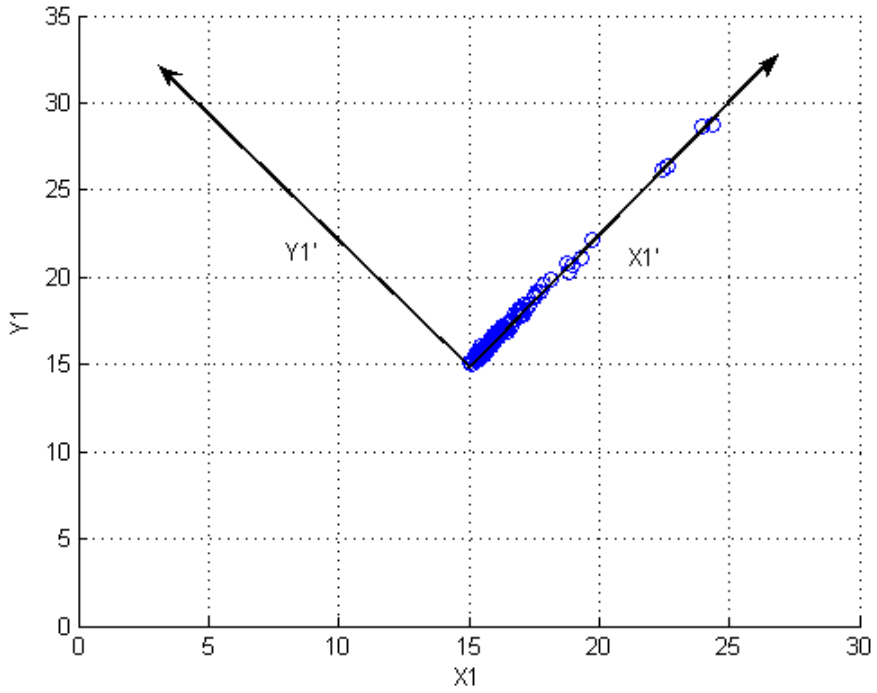

Figura 2.5: Transformação resultante da aplicação do método de Análise Fatorial.

2.2.1 Análise Fatorial Verdadeira

O modelo derivado da aplicação do método de *Análise Fatorial Verdadeira* é caracterizado por capturar todo relacionamento dos dados no domínio dos fatores. A matriz de resíduos resta conter apenas dados relativos às variâncias próprias das variáveis originais, não relacionadas entre-si. Decorre que a matriz de covariâncias residual ($E'E$) é diagonal.

A matriz de covariâncias da matriz de dados é expressa em notação fatorial formada pelos relacionamentos:

$$\frac{1}{N}Y'Y = A\left[\frac{1}{N}F'F\right]A' + A\left[\frac{1}{N}F'E\right] + \left[\frac{1}{N}E'F\right]A' + \frac{1}{N}E'E. \tag{2.19}$$

Por definição, são não-relacionados os fatores e os resíduos e os termos centrais no lado direito da equação anterior são anulados. A representação das estatísticas populacionais do fenômeno

aleatório, obtidas aumentando a quantidade de amostras **N**, é dada por:

$$\Sigma = A\Phi A' + \Psi. \tag{2.20}$$

Onde:

$$\begin{aligned}
\Sigma &= \frac{1}{N}Y'Y, \\
\Phi &= \frac{1}{N}F'F, \\
0 &= F'E = E'F, \\
\Psi &= \frac{1}{N}E'E.
\end{aligned}$$

Σ é a matriz de covariâncias populacionais das variáveis originais, Φ é a matriz de covariâncias dos escores fatoriais e Ψ é a matriz de covariâncias dos resíduos.

Assumindo os escores não-correlacionados e apresentando variâncias unitárias, o que pode ser obtido por multiplicação por matriz de rotação apropriada [Reyment and Jöreskog, 1996], a equação fatorial da matriz de covariâncias populacionais é simplificada para a forma:

$$\Sigma = AA' + \Psi. \tag{2.21}$$

A matriz de pesos é calculada supondo inicialmente conhecida a matriz de resíduos Ψ. Pré-multiplicando e pós-multiplicando a equação anterior por $\Psi^{-1/2}$, obtém-se a relação:

$$\begin{aligned}
AA' &= \Sigma - \Psi, \\
\Psi^{-1/2}AA'\Psi^{-1/2} &= \Psi^{-1/2}\Sigma\Psi^{-1/2} - \Psi^{-1/2}\Psi\Psi^{-1/2}, \\
A^*A^{*'} &= \Psi^{-1/2}\Sigma\Psi^{-1/2} - I.
\end{aligned} \tag{2.22}$$

onde, $A^* = \Psi^{-1/2}\,A$.

De acordo com Seber e Joreskog [Reyment and Jöreskog, 1996], a melhor estimativa para o valor de Ψ é obtido da aplicação do método da *Correlação Quadrática Múltipla*, e vale:

$$\hat{\Psi} = \theta(diag\,\Sigma^{-1})^{-1}, \tag{2.23}$$

onde:

$$\theta = \frac{1}{p-k}\sum_{m=k+1}^{p}\lambda_m, \tag{2.24}$$

é a média dos menores autovalores de:

$$S^* = (diag\,\Sigma^{-1})^{1/2}\,\Sigma\,(diag\,\Sigma^{-1})^{1/2}. \tag{2.25}$$

As matrizes A^* e $A^{*'}$ são obtidas por decomposição em valores singulares da matriz $\theta^2\,S^*$. A matriz de pesos é dada por:

$$A = \Psi^{1/2}A^*. \tag{2.26}$$

Os elementos da matriz de pesos são as medidas de relacionamento das variáveis originais e os escores, conforme demonstrado abaixo:

$$\begin{aligned}
\frac{1}{N}Y'F &= \frac{1}{N}(AF' + E')F, \\
\frac{1}{N}Y'F &= A\frac{1}{N}F'F + \frac{1}{N}E'F, \\
\frac{1}{N}Y'F &= AI + 0, \\
\frac{1}{N}Y'F &= A.
\end{aligned} \tag{2.27}$$

Os escores **F** são obtidos da projeção, por meio da matriz **Q**, dos vetores de observações-padronizadas no espaço dos fatores. O vetor padronizado é obtido, do vetor de observações, subtraindo deste os valores de média e dividindo cada elemento do vetor de observações pelo desvio

padrão da variável original correspondente. A operação de projeção, é expressa pela relação:

$$F = ZQ. \tag{2.28}$$

Pre-multiplicando a equação anterior por Z' e dividindo por N, o número de amostras, obtém-se:

$$\frac{1}{N}Z'F = \frac{1}{N}Z'ZQ. \tag{2.29}$$

O termo à esquerda corresponde à medida do relacionamento das variáveis originais e os escores, que foi demonstrado ser igual à matriz de pesos. O termo $\frac{1}{N}Z'Z$ é a matriz de correlação do conjunto original de dados, representado pela matriz R. O cálculo dos escores é realizado conforme apresentado a seguir:

$$\begin{aligned} A &= RQ, \\ Q &= R^{-1}A, \qquad (2.30) \\ F &= ZR^{-1}A. \qquad (2.31) \end{aligned}$$

Capítulo 3

Trabalhos Relacionados

Nesse capítulo são apresentados alguns trabalhos relacionados com a abordagem aqui desenvolvida. Esses trabalhos são todos orientados à aplicação de reconhecimento de fala e utilizam o método MFCC de extração de características ou métodos de análise de dados multivariados.

A técnica MFCC foi introduzida para a aplicação em reconhecimento de fala em trabalho de Davis e Mermelsteim [Davis and Mermelstein, 1990]. Este trabalho consiste em um estudo comparativo de diversos métodos de representação paramétrica, avaliados em tarefa de reconhecimento de palavras monosilábicas, equiprováveis, vocalizadas em ambiente controlado, por somente um locutor. O processo de classificação do reconhecedor constituia-se de medidas de distâncias para *templates* construídos das próprias locuções, utilizando algoritimo baseado em *Dinamic Time Warping*. Os resultados obtidos desses experimentos permitiram demonstrar a superioridade da técnica MFCC quando comparada às técnicas baseadas em predição linear. A explicação para esses resultados é dado pela maior robustez ao ruído aditivo, devido ao amortecimento do sinal de voz pelo banco de filtros MFCC, acarretando em menor variância dos vetores de energia. As técnicas baseadas em predição linear, ao contrário, não capturam as informações de ruído que se apresentam como zeros no espectro do sinal, devido à estratégia de modelagem LP realizada utilizando filtros *só-polos*.

O trabalho de Rivarol e O'Shaughnessy [Rivarol Vergin and Farhat, 1999] trata também da técnica MFCC em aplicação de reconhecimento de voz. Segundo o entendimento dos autores, o processo de filtragem MFCC desperdiça informações relevantes para o processo de discriminação fonética, como as posições dos formantes, por exemplo. Isso se deve à baixa resolução de frequências dos filtros MFCC, como consequência da largura de banda mínima de $100Hz$ da implementação MFCC original. O compromisso de resolução de frequências *versus* amortecimento para fins de supressão de ruído pode ser observado pela comparação desta abordagem com o trabalho apresentado no parágrafo anterior. Nesse trabalho, optou-se por remover os filtros MFCC do processo de geração dos coeficientes cepstrais. Os dados cepstrais são obtidos por projeção da sequência de log-energia em uma base extendida de séries de cossenos, semelhante à utilizada na transformação MFCC original. Portanto, a resolução de frequências obtida da tranformação de Fourier é preservada com a aplicação dessa técnica. A avaliação desse método ocorreu em experimento de reconhecimento de voz contínua, com vocabulário extenso e independência de locutor. Os resultados obtidos apresentaram valores ligeiramente superiores àqueles do método original.

O artigo de Skowronski e Harris [Skowronski and Harris, 2003] trata exatamente do compromisso resolução de frequências *versus* efeito de amortecimento dos filtros MFCC à incidência de ruído, discutido no parágrafo anterior. Este compromisso é materializado nos parâmetros de largura de banda dos filtros MFCC. Na implementação MFCC original, os filtros têm formato triangular e frequências centrais distribuidas linearmente na escala mel. A base de cada filtro, no entanto, é definida arbitrariamente pelas frequências centrais dos filtros adjacentes. A técnica proposta por Skowronski e Harris utiliza o conceito biológico ERB (*Equivalent Rectangular Bandwidth*) para determinação das larguras de banda dos filtros MFCC. Esta medida é uma aproximação da largura de banda do sistema auditivo humano sensível a um estímulo acústico em uma frequência específica. Os filtros construídos com a aplicação desse conceito têm larguras de banda menores que aquelas da implementação original e, portanto, maior resolução de frequências. Esta abordagem foi validada em tarefa de reconhecimento de dígitos, vocalizados por diversos locutores, em

ambiente ruidoso. Dos resultados obtidos, pôde-se verificar a performance superior desta técnica, quando comparada à implementação MFCC original, para adição de ruído branco em diferentes níveis SNR. No contexto de fala limpa ou corrompida por ruído de conversação, no entanto, os métodos são comparáveis em performance.

A abordagem apresentada no trabalho de Shang-Ming Lee [Lee et al., 2001] é também orientada à extração de características robustas à incidência de ruído. A proposta nesse trabalho é utilizar a técnica PCA para projetar as formas dos filtros MFCC. Pode ser verificado que a forma triangular comum dos filtros MFCC acarreta na ocorrência de níveis SNR distintos na sequência de log-energia, sob a incidência de ruído branco. Isso é devido aos componentes do sinal de voz apresentarem níveis de energia variáveis no *range* de frequências, enquanto esses valores são uniformes para o sinal de ruído branco. É desejável, portanto, a determinação de um procedimento que maximize os níveis SNR em todas as bandas de frequências. A abordagem adotada nesse trabalho consiste em modelar a forma dos filtros MFCC utilizando os coeficientes dos autovetores mais representativos, obtidos por decomposição da matriz de sequências espectrais. Esse trabalho demonstra que o mencionado procedimento maximiza a variância de saída dos filtros, bem como da relação sinal-ruído. Essa abordagem foi validada em tarefa de reconhecimento de dígitos em língua chinesa, submetidos à ruído branco com diferentes níveis SNR. Os resultados obtidos apresentaram melhoria significativa da taxa de reconhecimento de palavras para o sinal de voz corrompido por ruído branco e performance comparável à técnica original quando utilizando fala limpa.

O trabalho de Takiguchi e Ariki [Takiguchi and Ariki, 1990] é orientado ao desenvolvimento de um processo de extração de características com maior insensibilidade à incidência de ruído de reverberação. Esse tipo de ruído ocorre quando do posicionamento do microfone distante do locutor, na gravação de locuções em ambientes amplos. O efeito sobre o sinal de voz, nesse caso, é não-aditivo e os métodos empregados comumente para remoção de ruído não obtêm boa performance. A estratégia utilizada nesse trabalho é substituir a Transformada Discreta dos Cossenos

(DCT), do método de extração de características MFCC, pelo método de análise multivariada *Kernel PCA*. A aplicação do método Kernel PCA nesse trabalho é realizado mapeando a sequência de log-energia para um domínio de dimensão superior, através da aplicação da função kernel polinomial. Emprega-se, então, o método PCA para destacar os componentes principais da nova sequência de dados. Esse procedimento é representado graficamente na Figura 3.1. A validação dessa técnica ocorreu em experimentos de reconhecimento de fala, utilizando modelos HMM. Os resultados obtidos apresentaram melhoria significativa da taxa de reconhecimento de fala para o sinal de voz corrompido por ruído reverberante e performance comparável à técnica original quando utilizando fala limpa.

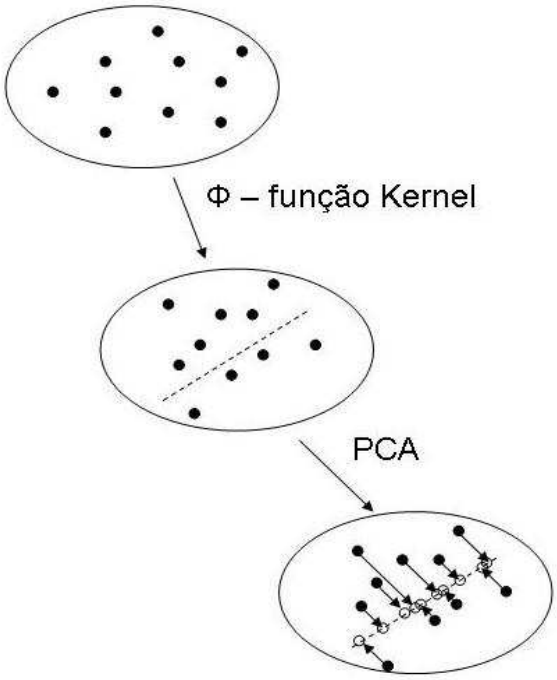

Figura 3.1: Representação gráfica da transformada Kernel PCA. Adaptado de [Lima et al., 2005].

Capítulo 4

Metodologia

Neste capítulo é apresentada a metodologia para extração de características do sinal de voz desenvolvida neste trabalho. Tal metodologia pode ser compreendida por meio dos seguintes passos:

1. Pré-processamento do sinal de voz;
2. Transformação MFCC;
3. Extração de características utilizando técnica baseada no método de Análise Fatorial Verdadeira;

Esses passos são descritos esquematicamente na Figura 4.1 onde se observa o sinal de voz transformado através dos procedimentos representados pelos retângulos. O sentido das transformações é o indicado pelas setas. O conjunto de características do sinal de voz é o resultado dessas transformações, os quais são validados quanto à robustez ao ruído em um sistema ASR, desenvolvido também neste trabalho.

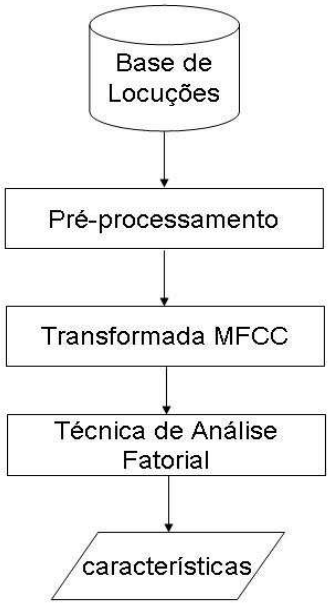

Figura 4.1: Desenho esquemático da metodologia desenvolvida.

4.1 Pré-processamento do Sinal de Voz

A etapa de pré-processamento compreende as operações realizadas no sinal de voz anteriores à aplicação dos métodos MFCC e Análise Fatorial. Uma vez que a base de locuções utilizada se encontra já no formato digital, não são executadas neste trabalho as operações de gravação, condicionamento, amostragem, quantização e codificação das amostras. A seguir, são descritas as operações da fase de pré-processamento:

- *Segmentação*: Corresponde à seleção de um segmento de sinal de voz de tamanho N, correspondente ao tempo de vocalização de um fonema. Dos segmentos de sinal de voz são extraídas as características representativas para fins de discriminação fonética. Os efeitos de articulação entre-fonemas são também capturados devido ao tamanho do deslocamento (d), da janela de segmentação, ser inferior ao tamanho do segmento. A resolução temporal do

processo de extração de características é dado pelo tamanho do deslocamento.

A operação de segmentação do sinal de voz, **S**, é representada pela expressão:

$$s_i = S[i * d; i * d + N]. \tag{4.1}$$

- *Remoção da Média*: Operação destinada à remoção do nível DC médio do segmento de sinal de voz. Essa operação tem efeito somente sobre a magnitude da amostra correspondente à frequência Zero *Hz*, na representação espectral.

  Assim, seja **m** o valor de amplitude médio do segmento de sinal $\mathbf{s_i}$. A sequência obtida da subtração deste valor de cada amostra do segmento de voz é dada por:

$$m = \frac{1}{N}\sum_{j=1}^{N} s_i[j], \tag{4.2}$$

$$s'_i[n] = s'_i[n] - m. \tag{4.3}$$

- *Pré-ênfase*: Essa operação objetiva compensar a atenuação dos componentes de alta-frequência do sinal de voz, resultado das características fisiológicas naturais do sistema humano de produção da fala. De acordo com Rivarol [Rivarol Vergin and Farhat, 1999], os segmentos vocalizados têm característica de decaimento de, aproximadamente, 20*dB* por década. A aplicação de um filtro de pré-ênfase tem o efeito de incrementar a magnitude dos componentes de alta frequência, favorecendo assim a relação sinal-ruído nesse *range*.

  Seja **p** o coeficiente de pré-enfase definido. A sequência obtida da aplicação dessa operação é expressa por:

$$ps_i[n] = s'_i[n] - p * s'_i[n-1], \tag{4.4}$$

$$ps_i[1] = s'_i[1](1-p). \tag{4.5}$$

- *Janelamento*: Essa operação tem por objetivo reduzir a incidência de ruído de alta frequência, causado pelo processo de segmentação. A operação de janelamento é realizada multiplicando o segmento de sinal por uma sinal-janela com característica de decaimento de amplitude nas extremidades. A aplicação desta técnica tem efeito de suavização do espectro, semelhante à aplicação de um filtro de médias.

  Seja **w[n]** um sinal-janela. O sinal obtido da operação de janelamento é expresso por:

$$f_i[n] = ps_i[n]\, w[n]. \tag{4.6}$$

4.2 A Transformação MFCC

Conforme apresentado no Capítulo 2, a transformação MFCC é realizada segundo os procedimentos listados a seguir:

- *Transformação de Fourier*: Realiza a transformação do sinal do domínio temporal para o domínio espectral, conforme a expressão:

$$F_i[k] = \sum_{n=1}^{N} f_i[n] e^{-\jmath(\frac{2\pi}{N})kn}. \tag{4.7}$$

- *Cálculo da Potência das Amostras*: A sequência espectral é transformada para uma sequência de valores de potência, obtida pela aplicação da operação:

$$P_i[k] = |F_i[k]|^2. \tag{4.8}$$

- *Particionamento da Sequência de Potência*: Dado a característica de simetria do espectro de potências de sinais reais, a sequência de potências é segmentada em duas partes de igual tamanho e descartada a metade superior.

Este procedimento é expresso em notação matemática pela expressão:

$$X_i[k] = P_i[k] \qquad 1 \leq k \leq N/2. \tag{4.9}$$

- *Redução por Banco de Filtros*: Essa operação realiza a transformação da sequência de potências para uma sequência acumulada de valores de potência, em bandas de frequência distintas, utilizando os filtros MFCC.

  Considera-se sejam distribuidos **K** filtros no *range* de Nyquist do sinal de voz (**B**). As frequências centrais ($\mathbf{fc_k}$) desses filtros, distribuidas linearmente na escala *mel* são calculadas segundo o procedimento apresentado a seguir:

  1. Cálculo da largura de banda de *Nyquist* na escala *mel*:

  $$B_{mel} = 2595 \log_{10} \left(1 + \frac{B}{700}\right). \tag{4.10}$$

  2. Cálculo das frequências-centrais dos filtros MFCC:

  $$fc_k = 700 * \left(10^{\frac{B_{mel} * k}{2595}} - 1\right), \quad k \in 0, \ldots, K. \tag{4.11}$$

  3. Cálculo do valor acumulado de potência no *k-ésimo* filtro:

     Dada a função de ganho ($\mathbf{H_k}$) do *k-ésimo* filtro no domínio espectral, o valor de potência acumulada nesse filtro é dada por:

  $$Y_k = \sum_{j=1}^{N/2} H_k(j) X_i(j). \tag{4.12}$$

- *Aplicação do Logarítimo*: A obtenção do relacionamento linear entre os sinais trato-vocal e excitação é realizado pela aplicação da função logarítimo à sequência obtida no passo

anterior, conforme descrito a seguir:

$$L_k = \log(Y_k). \tag{4.13}$$

- *Transformação Discreta dos Cossenos*: Essa operação realiza a separação dos componentes harmônicos da sequência de *log-potência* no domínio das *quefrências*.

  A expressão seguinte descreve essa operação:

$$c[n] = \sum_{k=1}^{K} L_k \cos\left(k\frac{\pi}{K}(n-0.5k)\right) \quad , \quad 1 \leq n \leq K. \tag{4.14}$$

- *Segmentação da Sequência Cepstral*: O sinal correspondente ao trato vocal é projetado na parte baixa do eixo das *quefrências*. A parte alta corresponde ao sinal de excitação e é descartado. As primeiras **M** amostras da sequência cepstral são mantidas, conforme expresso a seguir:

$$c_\theta[n] = c[n] \qquad 1 \leq n \leq M. \tag{4.15}$$

4.3 Técnica de Análise Fatorial Verdadeira

O método de extração de características do sinal de voz proposto nesta dissertação baseia-se na remoção das comunalidades, obtidas da aplicação da técnica de Análise Fatorial Verdadeira. Os procedimentos listados a seguir realizam essa etapa metodológica:

- *Criação da Matriz de Dados*: Operação que corresponde à reunião de sequências cepstrais, de comprimento **M**, representando a variabilidade acústica ocorrendo no universo das locuções. As locuções selecionadas são caracterizadas por buscar maximizar a variabilidade de frases e locutores.

  Seja $\mathbf{c}_\theta$, o vetor de características de comprimento **M**, obtido conforme apresentado na seção

anterior. A matriz de dados, de comprimento **N** ($\mathbf{N} >> \mathbf{M}$), é representada por:

$$\Theta = \begin{bmatrix} c_{\theta_{11}} & c_{\theta_{12}} & \dots & c_{\theta_{1M}} \\ c_{\theta_{21}} & c_{\theta_{22}} & \dots & c_{\theta_{2M}} \\ \vdots & \vdots & \vdots & \vdots \\ c_{\theta_{N1}} & c_{\theta_{N2}} & \dots & c_{\theta_{NM}} \end{bmatrix}$$

- *Cálculo da Matriz de Correlações*: A matriz de correlações $\mathbf{R_{(MxM)}}$ é calculada utilizando a matriz de dados **Θ**, conforme os procedimentos apresentados a seguir:

  1. Cálculo do vetor de médias:

$$\overline{C}_\theta = (\overline{c}_{\theta_1}, \overline{c}_{\theta_2}, \dots \overline{c}_{\theta_M}), \quad \text{onde} \quad \overline{c}_{\theta_i} = \frac{\sum_{j=1}^{N} c_{\theta_{ji}}}{N}. \tag{4.16}$$

  2. Cálculo da matriz de dados, removida a média:

$$Y = [y_{ij}], \quad \text{onde} \quad y_{ij} = c_{\theta_{ij}} - \overline{c}_{\theta_j}. \tag{4.17}$$

  3. Cálculo da matriz de covariâncias:

$$S = Y'Y/N, \tag{4.18}$$

     onde **Y**' é a matriz transposta de **Y**.

  4. Cálculo das correlações:

$$r_{ij} = s_{ij}/s_{ii}s_{jj}. \tag{4.19}$$

- *Decomposição Matricial*: Cálculo dos autovalores e autovetores da matriz definida na Equação (2.25), tomando Σ = **R**.

Os autovalores são calculados resolvendo a equação característica:

$$|R - \lambda I| = 0, \tag{4.20}$$

onde $\lambda = (\lambda_1, \lambda_2, \ldots \lambda_M)$ são os autovalores de **R** e **I** é a matriz identidade.

Determinados os autovalores, a equação seguinte é aplicada para obter cada autovetor (**u**):

$$Ru = u\lambda. \tag{4.21}$$

- *Determinação da quantidade de Fatores*: Conforme será visto no próximo capítulo, a quantidade de fatores mantidos no modelo fatorial é derivado de experimentação.

- *Cálculo do Fator de Compensação*: O fator θ de compensação da estimativa da matriz de resíduos é calculado aplicando a operação definida pela Equação (2.24).

- *Cálculo da Matriz de Pesos*: A matriz de pesos é calculada utilizando a equação a seguir, obtida da combinação das Equações (2.22) e (2.26).

$$A = (diagS^{-1})^{-1/2} U_k (\Lambda_k - \theta I)^{1/2}. \tag{4.22}$$

  Λ é matriz diagonal dos autovalores mantidos, **I** é a matriz identidade e $\mathbf{U_k}$ é matriz-coluna de dimensão **k** dos autovetores correspondentes.

- *Cálculo da Matriz de Projeção*: A matriz de projeção fatorial é obtida executando a operação expressa pela Equação (2.30).

- *Cálculo das Comunalidades do Vetor de Observações*: Os vetores de observações são projetados no espaço fatorial aplicando a matriz de projeção. O vetor de comunalidades é obtido da multiplicação do vetor de escores pela matriz de pesos, conforme o modelo fatorial. Uma

vez que a matriz de projeção fatorial é obtida da matriz de correlação amostral, é necessário, para a realização dessa atividade, padronizar as amostras cepstrais pelos vetores de média **m[n]** e desvio padrão **std[n]**. As equações abaixo descrevem esse procedimento:

$$c'_\theta[n] = \frac{c_\theta[n] - m[n]}{std[n]}, \tag{4.23}$$

$$F = \vec{c_\theta}' Q, \tag{4.24}$$

$$\overrightarrow{comun} = FA'. \tag{4.25}$$

$$\tag{4.26}$$

- *Cálculo dos Resíduos*: Os resíduos do vetor de observações, subtraidas as comunalidades explicadas pelo modelo fatorial, são tomadas como características do sinal de voz e utilizadas na tarefa de reconhecimento de fala. Esses dados são obtidos da aplicação da operação abaixo:

$$o[n] = c_\theta[n] - m[n] - comun[n]. \tag{4.27}$$

Capítulo 5

Resultados Experimentais

Neste capítulo são apresentados os recursos e atividades relacionados à parte experimental deste trabalho. Os experimentos realizados constituem-se de tarefas de reconhecimento automático de fala, por meio dos quais é avaliado o processo de extração de características desenvolvido.

A seção seguinte apresenta os recursos utilizados na implementação do método de extração de características e na criação de um sistema ASR completo baseado na implementação HTK [Odell et al., 1995] de redes de Modelos Escondidos de Markov (HMM). Os procedimentos empregados para realização dos processos de treinamento de modelos HMM e testes de reconhecimento de sentenças são baseados nos documentos HTKBook [Odell et al., 1995] e HTKTutorial [Moreau, 2002].

Para organização do texto, os recursos que compõem o sistema ASR são agrupados nas seguintes categorias:

- Implementação de redes HMM.
- Base de dados de locuções;
- Base de locuções ruidosas;
- Base de dados de texto;

5.1 Montagem do Arcabouço Experimental

Os experimentos realizados neste trabalho foram executados em computador PC com processador *Intel Core Duo2* à frequência de *clock* de Dois *GHz*, com Dois *GB* de memória RAM e executando sistema operacional *Windows XP Professional SP2*. Utilizou-se aproximadamente Um e meio *GB* de espaço em disco para instalação de aplicativos e arquivos necessários à realização dos experimentos.

O método de extração de características foi desenvolvido em *scripts* Matlab, versão 6.5.

A configuração do sistema ASR foi implementada em *scripts Python*, versão 2.5.1.

As seções seguintes apresentam os recursos constituintes do sistema ASR.

5.1.1 Implementação de Redes HMM

A estratégia de classificação do sistema decodificador de fala empregado neste trabalho é baseado em redes de modelos HMM.

O HMM é um modelo matemático utilizado para descrever processos aleatórios. Este modelo é representado por uma máquina de estados finitos duplamente encadeada que, no contexto de reconhecimento automático de fala, descreve um modelo acústico, como o fone. As transições entre os estados obedecem a um processo estacionário de Markov de primeira ordem. Os estados não são percebidos diretamente, mas geram eventos observáveis, independentes, modelados também por um processo probabilístico estacionário. As características extraídas do sinal de voz correspondem às observações emitidas quando da transição entre os estados.

A função de distribuição de probabilidades Gaussiana multivariada é utilizada, comumente, para modelar as probabilidades de emissão de observações. O *teorema do limite central* da estatística corrobora com essa escolha de função de distribuição, uma vez que são diversos os fatores envolvidos na geração do sinal de fala.

O objetivo de um sistema ASR é encontrar a sequência de palavras mais provável, dado a

sequência de vetores de observações. Esse problema pode ser decomposto em termos de modelos probabilísticos acústicos e linguísticos, o produto dos quais se busca maximizar. A implementação HTK, adotada neste trabalho, realiza este cálculo empregando a estratégia *bottom-up* que postula os encadeamentos HMM mais prováveis devido ao modelo linguístico e calcula as probabilidades de geração dos vetores de observações pelos modelos HMM correspondentes. Em seguida, o contexto linguístico é expandido e o processo repetido até esgotar a sequência de observações. A técnica de poda *beam-search* é utilizada, nesse contexto, para restringir os modelos avaliados que não atingem probabilidade mínima de geração da sequência acústica.

Os modelos acústico e linguístico são relacionados, respectivamente, à probabilidade de geração de uma sequência de observações por um modelo HMM e à probabilidade de postular, *a priori*, a ocorrência de uma sequência de unidades linguísticas.

Os algoritimos de *Viterbi* e *Baum-Welch* são consagrados na lituratura como solução dos problemas de decodificação e treinamento dos modelos HMM, respectivamente. O trabalho de Rabiner e Juang [Rabiner and Juang, 1993] dá detalhes de funcionamento desses algoritimos.

A conjunto de softwares HTK (*Hidden Markov Toolkit*) constitui-se de uma coleção de aplicativos que provêem as funcionalidades necessárias à manipulação dos modelos acústicos e de linguagem, na realização da tarefa de reconhecimento de fala.

Nas seções seguintes, os aplicativos HTK utilizados neste trabalho e as fases do processo de reconhecimento de fala em que esses atuam, são descritos resumidamente.

Fase de Preparação de Dados

Nessa fase são criados os arquivos de léxico e os arquivos de transcrições, denominados, na terminologia HTK, de arquivos *Label*.

- *HDMan*: Aplicativo utilizado para construção e edição do léxico que contém a transcrição fonética de cada palavra do vocabulário do sistema.

- *HLEd*: Aplicativo utilizado para edição dos arquivos de transcrições utilizando as entradas do léxico construído no passo anterior.

- *HCopy*: Aplicativo utilizado para manipulação dos arquivos de aúdio. Neste trabalho, esse aplicativo é empregado para calcular os parâmetros *delta* de primeira e segunda ordem, derivados do conjunto de características gerados pelo método fatorial desenvolvido.

Fase de Treinamento de Modelos Acústicos

Nessa fase são criados os modelos acústicos HMM.

Os parâmetros dos modelos HMM de cada fone ocorrendo na base de locuções são calculados, na estratégia HTK, em dois passos distintos: Inicialização e refinamento. Em ambas fases, os arquivos de transcrição são empregados na criação de modelos HMM encadeados, representando sentenças inteiras. São aplicados, então, algoritimos de treinamento tomando como entradas o modelo HMM composto e a sequência de características extraídas do sinal de fala. O resultado desse processo é um conjunto de parâmetros refinado, para cada modelo HMM. Os modelos HMM são, então, transformados em modelos HMM trifones, afim de lidar com a variabilidade acústica causada pelo fenômeno de *coarticulação*. Esse modelo captura as informações do fone central e dos fones à direita e à esquerda deste.

- *HCompV*: Aplicativo destinado à inicialização dos modelos HMM.

- *HERest*: Aplicativo utilizado para refinar os parâmetros dos modelos HMM, no processo de treinamento.

- *HHEd*: Aplicativo utilizado para edição dos modelos acústicos HMM, na realização da estratégia de compartilhamento de parâmetros pelos modelos trifones, afim de fazer frente ao problema de escassez de dados.

Fase de Criação de Modelo de Linguagem

O modelo de linguagem é parte constituinte da modelagem HMM e constitui-se de métricas que descrevem a probabilidade de ocorrência de sequências de palavras. A criação do modelo de linguagem utiliza uma base de dados de textos.

Os principais aplicativos HTK utilizados nessa fase são:

- *LNewMap*: Aplicativo utilizado na contagem de palavras da base de texto.
- *LGPrep*: Este aplicativo cria as tabelas *n-gram* que contêm as probabilidades de ocorrência de palavras e de sequência de palavras da base de texto.
- *LGCopy*: Aplicativo utilizado para filtrar das tabelas *n-gram* as palavras não cadastradas no léxico do sistema.
- *LBuild*: Aplicativo destinado à criação do modelo de linguagem.
- *HBuild*: O propósito deste aplicativo é transformar o modelo de linguagem criado no ítem anterior para o formato requerido pelo aplicativo decodificador *HVite*.

Fase de Teste

Os modelos acústicos e lingúistico criados nas fases anteriores são combinados nessa fase, para realização da tarefa de reconhecimento de fala.

- *HVite*: Esse aplicativo realiza o processo de decodificação das sequências acústicas em hipóteses de sequências de palavras.

Fase de Análise dos Resultados

Nesta fase são comparadas as sequências de palavras hipotetizadas com a correta transcrição das locuções para determinação da taxa de acerto de palavras.

- *HResults*: Esse aplicativo calcula a taxa de acerto de palavras da tarefa de reconhecimento, alinhando dinâmicamente as sentenças hipotetizadas e padronizadas e calculando as quantidades de palavras corretas e incorretas.

5.1.2 Base de Dados de Locuções

A base de locuções desenvolvida por Ynoguti [Ynoguti, 1999] foi utilizada neste trabalho. Essa base de locuções é constituída de vinte listas foneticamente balanceadas, contendo dez sentenças cada. O vocabulário dessa base é de 694 palavras, sendo considerada de tamanho médio. As locuções estão gravadas em formato WAVE à taxa de $11025Hz$ e são articuladas por 40 locutores distintos, sendo 20 do sexo masculino e 20 do sexo feminino. Foi adotada a classificação fonética proposta por Ynoguti e, portanto, considera-se ocorrerem nas locuções dessa base de dados, 35 fones distintos, mais 2 fones representando pausa entre palavras e silêncio longo.

A Tabela 5.1 apresenta o conjunto de fones proposto por Ynoguti, com as informações: Símbolo utilizado nas transcrições fonéticas, exemplo de utilização, frequência de ocorrências e número de ocorrências na base de locuções.

Devido ao fenômeno de coarticulação, a consistência espectral de qualquer fone apresenta larga variação, devido ao contexto. A utilização de modelos acústicos baseados em *trifones* é uma proposta de solução para esse problema. No entanto, o número de diferentes unidades trifones, admitindo-se qualquer combinação dos fones da tabela anterior, ultrapassa 45000. Em consequência desse grande número, utiliza-se a estratégia de compartilhamento de parâmetros entre os modelos HMM. Neste trabalho, foi empregada uma estratégia de agrupamento de contexto fonético baseado no local de articulação do som no trato vocal e na forma de excitação. Esta abordagem visa mitigar o problema de escassez de dados, que poderia acarretar em ineficiente processo de treinamento dos modelos acústicos.

As classes fonéticas listadas a seguir foram usadas, neste trabalho, para classificação dos con-

Tabela 5.1: Lista de fones.

Símbolo utilizado	Exemplo	Frequência Relativa(%)	Número de Ocorrências
a	*a* çafrão	13,91	6031
e	*e* levador	2,15	933
E	p *e* le	6,35	2785
i	s *i* no	1,90	821
y	fu *i*	0,95	410
o	b *o* lo	4,14	1798
O	b *o* la	6,23	2691
u	l *u* a	2,57	1124
an	maç *ã*	4,04	1773
en	s *en* ta	1,16	510
in	p *in* to	0,69	296
on	s *om* bra	8,41	3648
un	*um*	1,98	860
b	*b* ela	1,18	511
d	*d* ádiva	3,14	1346
D	*d* iferente	1,49	665
f	*f* eira	1,44	625
g	*g* orila	0,87	378
j	*j* iló	0,75	325
k	*c* achoeira	3,63	1575
l	*l* eão	1,91	830
L	*lh* ama	0,35	152
m	*m* ontanha	3,77	1637
n	*n* évoa	2,26	982
N	i *nh* ame	0,42	185
p	*p* oente	2,49	1081
r	ce *r* a	4,05	1759
rr	ce *rr* ado	0,89	363
R	ca *r* ta	1,32	598
s	*s* apo	6,52	2832
t	*t* empes *t* ade	4,02	1737
T	*t* igela	1,20	531
v	*v* erão	1,51	656
x	*ch* ave	0,32	132
z	*z* abumba	1,96	859

textos esquerdo e direito dos fones centrais:

1. vogais anteriores;
2. vogais médias;
3. vogais posteriores;
4. consoantes labiais plosivas;
5. consoantes labiais nasais;
6. consoantes labiais fricativas;
7. consoantes médias plosivas;
8. consoantes médias nasais;
9. consoantes médias fricativas;
10. consoantes posteriores plosivas;
11. consoantes posteriores nasais;
12. consoantes posteriores fricativas;
13. consoantes laterais;
14. consoantes vibrantes.

5.1.3 Base de Locuções Ruidosas

O processo de extração de características desenvolvido neste trabalho é validado em ambiente corrompido por ruído. Arquivos de locuções ruidosas são construídas utilizando a base de locuções de Ynoguti [Ynoguti, 1999] e a base de ruído RSG-10 [Steeneken and Geurtsen, 1988], segundo os procedimentos apresentados a seguir:

- *Segmentação*: A operação de segmentação seleciona amostras do sinal de voz e do sinal de ruído, de mesmo tamanho **N**, correspondente ao tempo de vocalização de um fonema.

Os segmentos de sinal de voz e de ruído são obtidos pela aplicação da operação:

$$s_i = S[i * N, (i+1) * N], \tag{5.1}$$

$$r_i = R[i * N, (i+1) * N]. \tag{5.2}$$

- *Cálculo da Energia dos Segmentos*: Os valores médios de energia dos segmentos de voz e de ruído são calculados conforme as expressões:

$$ES_i = \frac{1}{N} \sum_{j=1}^{N} |s_i[j]|^2, \tag{5.3}$$

$$ER_i = \frac{1}{N} \sum_{j=1}^{N} |r_i[j]|^2. \tag{5.4}$$

- *Cálculo do multiplicador SNR*: Esta operação objetiva calcular o valor multiplicador do segmento de ruído, necessário à obtenção de uma relação SNR definida. Dados os valores médios de energia dos segmentos de sinal e de ruído, o valor multiplicador SNR é dado por:

$$M_i = \frac{ES_i}{ER_i \, 10^{SNR/10}}. \tag{5.5}$$

- *Mistura dos segmentos de Voz e Ruído*: As locuções ruidosas são construídas misturando aditivamente os segmentos de voz e de ruído, ponderado pelo valor multiplicador SNR, conforme a expressão:

$$s'_i[n] = s_i[n] + M_i r_i[n]. \tag{5.6}$$

5.1.4 Base de Dados de Texto

O banco de sentenças textuais é um recurso destinado à construção do modelo de linguagem.

O banco de dados *CETENFolha* [Pinheiro and Aluísio, 2003] foi utilizado para criação do mo-

delo de linguagem. Esse recurso consiste de extratos de reportagens veiculadas no jornal *A Folha de São Paulo* no ano de 1994. O formato em que se apresenta esta base de dados contém marcações, no estilo XML, destinadas à identificação de elementos de estilo literário. Afim de prover compatibilidade dessa base textual com os algoritimos da implementação HTK, foram removidas a maior parte dessas marcações, bem como símbolos de pontuação e sentenças contendo números. A realização desse procedimento resultou em uma base de dados contendo ainda mais de 187000 palavras e 1000000 de segmentos.

O modelo de linguagem é construído calculando as probabilidades de ocorrência de palavras e de sequências de palavras na base de texto. A implementação HTK realiza esse atividade por meio de operações de contagem desses eventos e ponderação das métricas calculadas. Detalhes da operação desse procedimento pode ser encontrado no trabalho de Moore [Moore, 2001].

O banco de dados CETENFolha é caracterizado pelo discurso do domínio jornalístico, sendo mais apropriado para suportar atividades de reconhecimento orientado a este contexto específico. No entanto, a utilização, neste trabalho, desta base de dados provou-se bastante satisfatória.

5.2 Experimentos

A validação da metodologia desenvolvida neste trabalho é realizada pela execução dos experimentos apresentados nas próximas seções. Esses experimentos constituem-se de tarefas de reconhecimento automático de fala e são mensurados quanto à sua precisão utilizando a métrica *Taxa de Reconhecimento de Palavras*.

5.2.1 Taxa de Reconhecimento de Palavras

A avaliação da qualidade de sistemas ASR é realizada, comumente, através da métrica taxa de reconhecimento de palavras (*WRR - Word Recognition Rate*). Essa métrica exprime a relação da quantidade de palavras corretamente decodificadas relativo à quantidade total de palavras em

Tabela 5.2: Comparação das taxas WRR: Experimento *baseline versus* Ynoguti.

WRR ASR *baseline*	80,56
WRR ASR Ynoguti	81,01

uma sentença. Para obtenção dessa medida, é necessário alinhar a sentença contendo a correta transcrição da locução com a sentença hipotetizada pelo reconhecedor. Realizada essa operação, identificam-se as palavras hipotetizadas corretamente e as palavras removidas, inseridas e substituídas, relativo à sentença padrão.

A métrica WRR exprime-se pela relação:

$$WRR = \frac{N - S - I - D}{N}, \tag{5.7}$$

onde:

- N é a quantidade de palavras na sentença padrão;
- S é a quantidade de palavras substituídas;
- I é a quantidade de palavras inseridas;
- D é a quantidade de palavras removidas;

5.2.2 Experimento Baseline

O experimento rotulado *baseline* cumpre o objetivo de estabelecer uma referência para as realizações experimentais conduzidas neste trabalho.

Adotou-se, sem alteração, a proposta de tarefa de reconhecimento de fala apresentada no trabalho de Ynoguti [Ynoguti, 1999]. Com isso, pôde-se realizar a validação do sistema ASR desenvolvido, pela comparação das taxas de reconhecimento de palavras obtidas do sistema ASR desenvolvido por Ynoguti e desta *baseline*. Esses valores são apresentados na Tabela 5.2:

A convergência dos valores WRR observados indica a realização eficaz do sistema ASR desenvolvido.

A *baseline* é realizada segundo o procedimento metodológico apresentado no capítulo anterior, excluído a aplicação do método fatorial e utilizando fala livre de ruído. Os demais parâmetros são fixados e mantidos constantes nos demais experimentos realizados neste trabalho e são listados a seguir:

- Tamanho do segmento: 256 amostras;
- Tamanho do deslocamento: 110 amostras;
- Coeficiente de pré-ênfase: 0,97;
- Função-janela de Hamming;
- Quantidade de filtros MFCC: 26;
- Forma dos filtros MFCC: Triangular;
- Largura de banda dos filtros MFCC: Distância das frequências centrais dos filtros adjacentes;
- Quantidade de coeficientes cepstrais mantidos: 12;
- Emprego da energia total filtrada como característica adicional;
- Emprego dos coeficientes derivados de primeira e segunda ordem como características adicionais;
- Modelo HMM composto de três estados emissores;
- Modelo de linguagem;
- Tarefa de reconhecimento;

5.2.3 Experimento 1

O primeiro conjunto de experimentos foi realizado variando a quantidade de fatores mantidos na construção do modelo fatorial.

Nos experimentos realizados neste trabalho, a construção das matrizes de transformação fatorial é realizada empregando o mesmo conjunto de locuções selecionadas pelo critério apresentado na Seção 4.3. Estas locuções se diferenciam pela combinação com sinais de ruído de naturezas diversas. Neste experimento, em particular, o sinal de ruído é tornado nulo em todo intervalo de tempo.

As taxas da reconhecimento de palavras obtidas são apresentadas na Tabela 5.3 e na Figura 5.1.

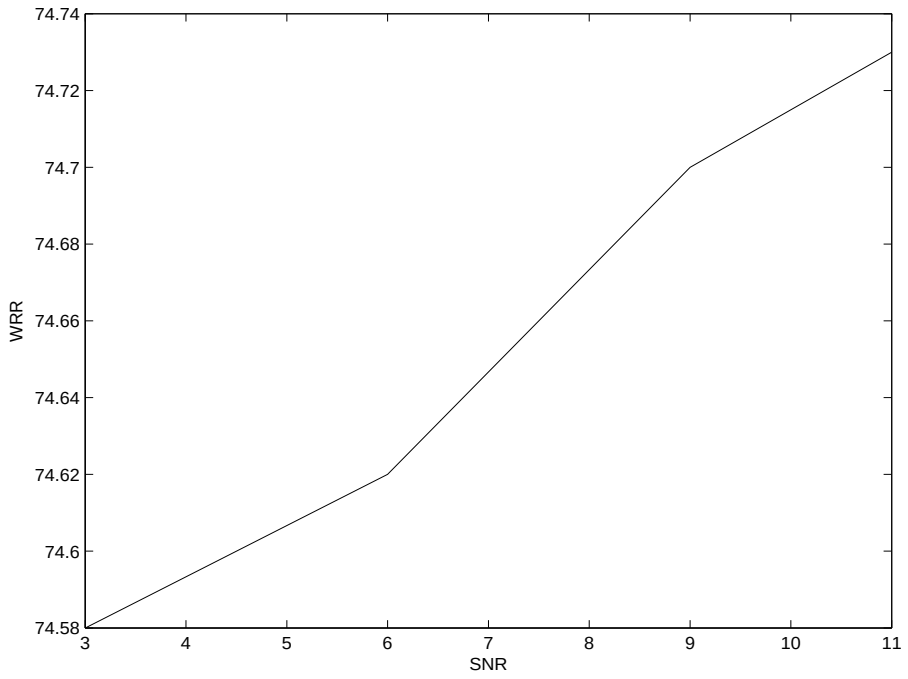

Figura 5.1: Experimento 1: Taxa WRR MFCC-FA *versus* quantidade de fatores. Experimento realizado com fala livre de ruído.

Neste experimento busca-se verificar a hipótese que a remoção das comunalidades do conjunto de características do sinal de voz influencia positivamente a taxa de reconhecimento de palavras.

Tabela 5.3: Experimento 1: Taxa WRR MFCC-FA *versus* quantidade de fatores. Experimento realizado com fala livre de ruído.

Quantidade de Fatores	3	6	9	11
WRR MFCC-FA	74,58	74,62	74,70	74,73

Pode-se observar que ocorre uma melhoria contínua da taxa WRR com o aumento da quantidade de fatores mantidos. Conforme apresentado no Capítulo 2, o modelo de covariâncias entre as amostras cepstrais é aproximada no modelo fatorial com o aumento da quantidade de fatores. O conjunto de características obtidas da subtração do vetor cepstral de sua projeção fatorial tem removida a informação de comunalidade. Assim, o resultado desse experimento é um indicativo da razoabilidade da hipótese declarada.

Nos experimentos seguintes, neste trabalho, o modelo fatorial terá 11 fatores.

5.2.4 Experimento 2

O segundo conjunto de experimentos objetiva investigar o desempenho do método fatorial de extração de características frente à variação da intensidade do sinal de ruído embutido nas locuções.

Bases de locuções corrompidas com *ruído branco*, com diferentes níveis SNR, foram criadas para realização deste experimento. Para cada condição SNR foram executadas tarefas de reconhecimento de fala empregando, inicialmente, o método MFCC original. Em seguida, foi avaliado o desempenho do método fatorial desenvolvido, doravante denominado MFCC-FA.

Os modelos acústicos foram criados com fala livre de ruído.

O desempenho do sistema ASR foi registrado para cada condição experimental e os resultados são apresentados na Tabela 5.4 e na Figura 5.2.

Este conjunto de experimentos investiga o desempenho do modelo fatorial em ambiente contaminado com ruído branco. Esse tipo de ruído se caracteriza por apresentar distribuição de potência uniforme em todo espectro. Os resultados obtidos revelam o declínio da taxa WRR com o aumento da potência do sinal de ruído. No entanto, o desempenho do sistema equipado com o

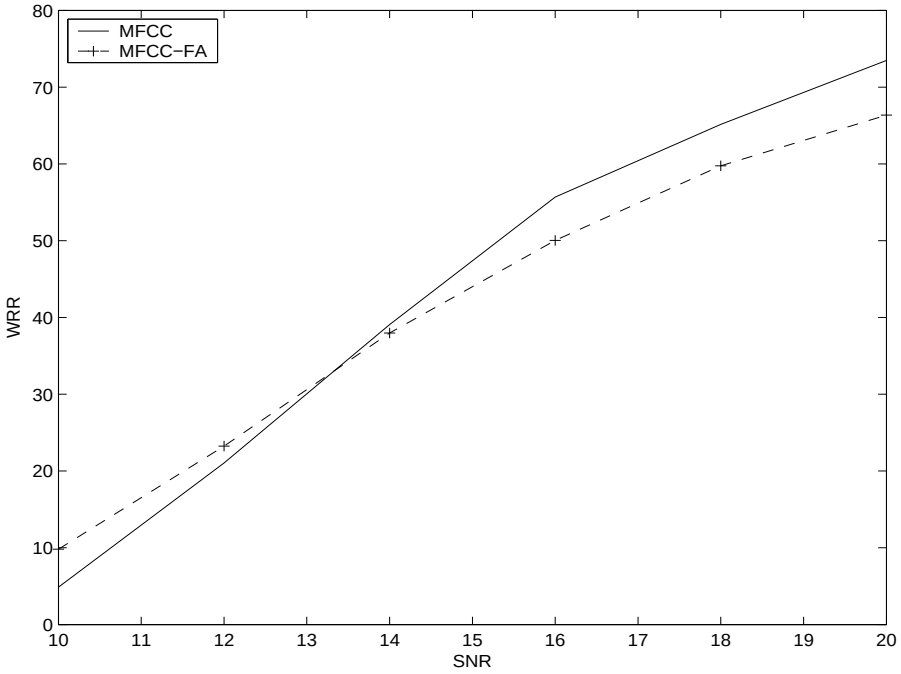

Figura 5.2: Experimento 2: Taxa WRR métodos MFCC e MFCC-FA *versus* razão SNR. Experimento realizado utilizando sinal de ruído branco.

método MFCC-FA decai mais lentamente que o desempenho do sistema ASR implementando o método MFCC original, e o ultrapassa para valores SNR inferiores à $13\,dB$. Este comportamento pode ser explicado pela característica de filtragem das comunalidades do método MFCC-FA, uma vez que o sinal de ruído branco incidente em todo espectro faz crescer os valores de covariância entre as amostras.

Tabela 5.4: Experimento 2: Taxa WRR métodos MFCC e MFCC-FA *versus* razão SNR. Experimento realizado utilizando sinal de ruído branco.

SNR(dB)	10	12	14	16	18	20
WRR MFCC	4,87	21,04	39,08	55,67	65,14	73,48
WRR MFCC-FA	9,82	23,25	37,98	50,04	59,74	66,36

5.2.5 Experimento 3

Os experimentos apresentados nesta seção investigam o desempenho do método fatorial de extração de características frente à variação do tipo de ruído incidente nas locuções.

De maneira inversa ao procedimento adotado na seção anterior, neste conjunto de experimentos o nível SNR das locuções ruidosas é mantido constante em 15dB. O sinal de ruído adicionado às locuções de teste foi variado entre os tipos: Branco, rosa, de conversação e ambiente de fábrica. Para cada tipo de sinal de ruído, o desempenho do sistema ASR foi avalidado empregando as técnicas de extração de características MFCC e MFCC-FA.

Os modelos acústicos foram criados com fala livre de ruído.

Os resultados desses testes são apresentados na Tabela 5.5 e na Figura 5.3.

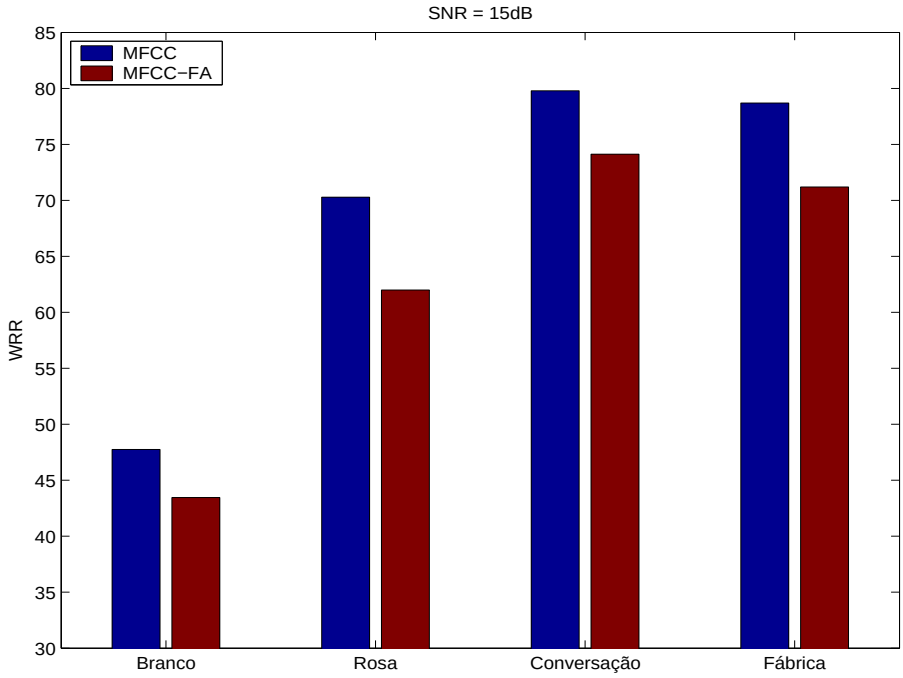

Figura 5.3: Experimento 3: Taxa WRR métodos MFCC e MFCC-FA *versus* tipo de sinal de ruído. A relação SNR em cada experimento é mantida constante e igual à 15dB.

Este conjunto de experimentos valida o método MFCC-FA sob a incidência de diversos tipos de sinais de ruído. Com exceção do ruído branco, os outros tipos de ruído investigados apresentam

Tabela 5.5: Experimento 3: Taxa WRR métodos MFCC e MFCC-FA *versus* tipo de sinal de ruído. A relação SNR em cada experimento é mantida constante e igual à 15*dB*.

Ruído	Branco	Rosa	Conversa	Fábrica
WRR MFCC	47,75	70,28	79,79	78,69
WRR MFCC-FA	43,46	61,99	74,12	71,19

concentração de potência nas baixas frequências, semelhante à voz humana. O ruído rosa é caracterizado pelo decaimento linear da potência com o aumento da frequência. Os sinais de ruído de conversação e ambiente de fábrica são não-estacionários. Observa-se, dos resultados obtidos que, para a taxa SNR de $15\,dB$, o acrescimo de energia nas baixas frequências tem pouco efeito sobre a taxa de reconhecimento de palavras. Somente os experimentos com ruído branco e rosa mostram alguma degradação da taxa WRR, devido a esses sinais apresentarem componentes de potência em frequências superiores. O desempenho do sistema ASR equipado com o método MFCC-FA é inferior, em cada caso, ao desempenho do sistema ASR implementando o método MFCC original. As taxas WRR, no entanto, são aproximadas. Esse comportamento pode ser explicado por serem os modelos acústicos gerados com fala limpa, pelo método MFCC tradicional, e serem as características obtidas do método MFCC-FA derivadas dos vetores cepstrais, por meio de transformação linear.

5.2.6 Experimento 4

Os experimentos apresentados nesta seção investigam a influência do método fatorial no processo de geração de modelos acústicos adaptados às condições ambiente.

Neste conjunto de experimentos, os modelos acústicos são criados por treinamento utilizando bases de locuções ruidosas. O desempenho do sistema ASR foi medido em experimentos com ruído dos tipos: Branco, rosa, de conversação e ambiente de fábrica, mantida constante a taxa SNR em 15*dB*. Foi avaliada também a influência do método fatorial no processo de treinamento dos modelos acústicos em condição livre de ruído.

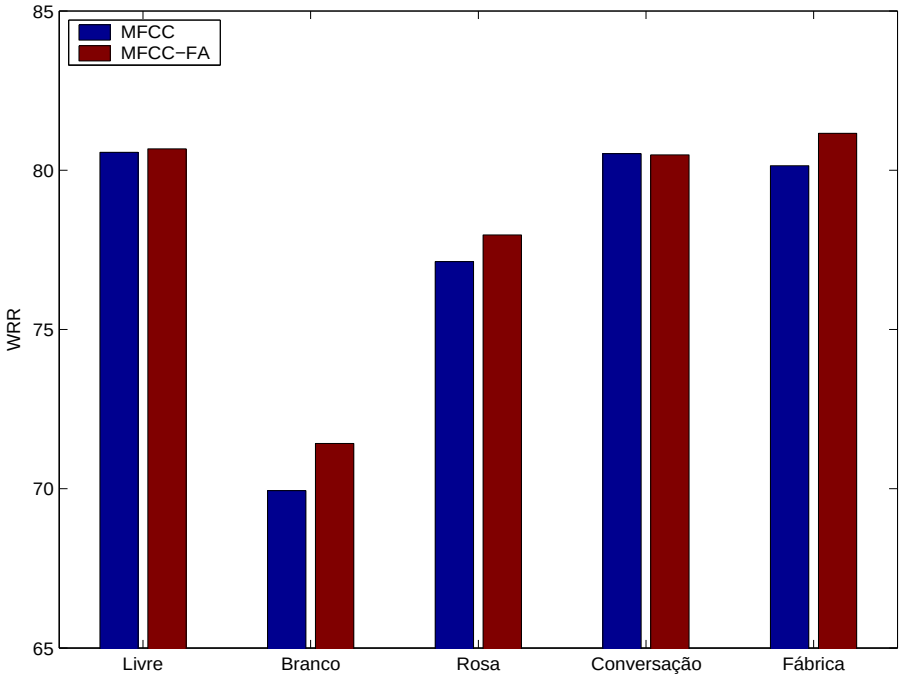

Figura 5.4: Experimento 4: Taxa WRR métodos MFCC e MFCC-FA em treinamento modelos HMM. Os experimentos realizados com locuções ruidosas mantêm a taxa SNR constante e igual à 15dB.

Conforme o procedimento adotado nos outros experimentos, foram realizados testes com os métodos MFCC e MFCC-FA.

Os resultados obtidos são apresentados na Tabela 5.6 e na Figura 5.4.

Tabela 5.6: Experimento 4: Taxa WRR métodos MFCC e MFCC-FA em treinamento modelos HMM. Os experimentos realizados com locuções ruidosas mantêm a taxa SNR constante e igual à 15dB.

Ruído	Livre	Branco	Rosa	Conversa	Fábrica
WRR MFCC	80,56	69,94	77,13	80,52	80,14
WRR MFCC-FA	80,67	71,42	77,97	80,48	81,16

Este conjunto de experimentos revela o que pode ser considerada a aplicação preferencial do método fatorial desenvolvido neste trabalho: a geração de modelos acústicos adaptados à condição ambiente. Neste experimento, o processo de treinamento é realizado com fala ruidosa, com observações constituídas dos resíduos fatoriais. A avaliação desta aplicação com fala livre de ruído,

bem como, com locuções corrompidas com ruído de naturezas diversas revelam o desempenho ligeiramente superior da técnica MFCC-FA comparada à técnica MFCC tradicional, à exceção do teste com ruído de conversação. No entanto, pode-se esperar, dos resultados e análise realizados do primeiro conjunto de experimentos que, com a diminuição da taxa SNR, o predomínio do método fatorial se torne mais evidente, inclusive para o caso de ruído de conversação.

5.3 Discussão

Nesta seção são discutidos os resultados obtidos dos experimentos descritos nas seções anteriores.

O Experimento um demonstrou ser factível a hipótese que a remoção das comunalidades do modelo fatorial influencia positivamente a taxa de reconhecimento de palavras.

O método MFCC-FA, desenvolvido neste trabalho, demonstrou ser competitivo em relação ao método MFCC original sendo derivado deste por transformação linear, como demonstram os relacionamentos das medidas de reconhecimento de palavras nos Experimentos dois e três.

A característica de filtragem de ruído do método MFCC-FA é mais facilmente observada em condições de ruído severo quando a taxa WRR da tarefa de reconhecimento utilizando MFCC-FA supera aquela do método MFCC original, mesmo quando os modelos acústicos são criados dos vetores cepstrais não-transformados. Esta conclusão é baseada na análise dos resultados do Experimento dois.

Finalmente, o Experimento quatro demonstra a utilização do método MFCC-FA na construção dos modelos acústicos que é, possivelmente, a aplicação preferencial desse método na aplicação de reconhecimento de voz. Os experimentos realizados com fala limpa e corrompida com ruído revelam desempenho ligeiramente superior do método MFCC-FA comparado ao método MFCC original, à exceção do teste com ruído de conversação.

Capítulo 6

Conclusões

O objeto de estudo deste trabalho é a investigação da aplicabilidade do método de Análise Fatorial Verdadeira no processo de extração de características do sinal de voz.

Elegeu-se a aplicação de reconhecimento automático de fala para validação da metodologia desenvolvida nesta pesquisa, por apresentar as seguintes características: 1) Por não haver implementada ainda uma solução robusta para esse problema, possível de ser utilizada em ambientes diversos, submetidos à interferência de ruído de naturezas variadas; 2) Por haver, desde já, demanda reprimida dos portadores de necessidades especiais para o desenvolvimento de produtos industriais economicamente viáveis e socialmente inclusivos; e 3) Por apresentar potencial de ruptura do modo de interação homem-máquina.

A investigação realizada partiu do estudo de soluções consagradas como estado-da-arte da área da pesquisa em sistemas ASR, quais sejam: O método de análise do sinal de voz utilizando transformação MFCC e o método de classificação de padrões baseado em redes de modelos escondidos de Markov (HMM). Utilizando implementações desses dois métodos, foi desenvolvido um sistema ASR orientado à língua portuguesa, fala contínua, vocabulário médio e multilocutor. Os recursos computacionais envolvidos na criação do sistema ASR são as implementações HMM do pacote de aplicativos HTK, versão 3.3, e a base de locuções desenvolvida por Ynoguti. Os resultados obtidos

dessa implementação ASR foram bastante satisfatórios, o que permitiu o avanço deste trabalho.

A análise do interrelacionamento dos dados das amostras cepstrais provocou a declaração da hipótese investigada neste trabalho, qual seja: A remoção das comunalidades, modelada pelo método fatorial, do vetor de características do sinal de voz resulta em um novo conjunto de características, relacionado linearmente com aquele primeiro, e mais robusto à incidência de ruído aditivo.

A necessidade de validação do método MFCC-FA, derivado da implementação do modelo hipotetizado, nas condições de ruído em que se vislumbra seu melhor caso de aplicação, levou à criação de bases de locuções ruidosas. Utilizou-se, para este fim, dos arquivos de ruído disponibilizados por Steeneken, em conjunto com as locuções devidas à Ynoguti.

Os resultados dos experimentos realizados neste trabalho sugerem a confirmação da hipótese investigada, e a potencialidade de uso desta abordagem como método robusto de extração de características do sinal de fala.

6.1 Propostas de Trabalhos Futuros

O prosseguimento da linha de investigação trilhada neste trabalho pode ser realizada por alguma abordagem derivada das idéias apresentadas a seguir:

- A utilização do método fatorial na remoção das comunalidades das amostras espectrais, em fase anterior à filtragem MFCC;
- A investigação da aplicabilidade do método PCA em estratégia de subtração de vetores de características, semelhante ao apresentado neste trabalho;
- A investigação do potencial de aplicação de funções de saturação, como a função *sigmoide*, em fase posterior à subtração dos vetores de características, na geração de novos conjuntos de características do sinal de voz.

Referências Bibliográficas

[Bozzeto, 2004] Bozzeto, C. (2004). Estratégias de modelamento acústico aplicadas a reconhecedores automáticos de fala baseados em hmm. Master's thesis, Departamento de Ciência da Computação - UFAM.

[Davis and Mermelstein, 1990] Davis, S. B. and Mermelstein, P. (1990). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Readings in speech recognition*, pages 65–74.

[Hermansky and Morgan, 1994] Hermansky, H. and Morgan, N. (1994). RASTA processing of speech. In *IEEE Transactions on Speech and Acoustics*, volume 2, pages 587–589.

[Hynek Hermansky and Kohn, 1990] Hynek Hermansky, Nelson Morgan, A. B. and Kohn, P. (1990). Perceptual linear predictive (plp) analysis for speech. *J. Acoust. Soc. Am.*

[John R. Deller et al., 1993] John R. Deller, J., Proakis, J. G., and Hansen, J. H. (1993). *Discrete Time Processing of Speech Signals*. Prentice Hall PTR, Upper Saddle River, NJ, USA.

[Lee et al., 1990] Lee, K.-F., Hon, H.-W., and Reddy, R. (1990). *An overview of the SPHINX speech recognition system*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Lee et al., 2001] Lee, S.-M., Fang, S.-H., weih Hung, J., and Lee, L.-S. (2001). Improved mfcc feature extraction by pca-optimized filter-bank for speech recognition. *Automatic Speech Recognition and Understanding*.

[Lima et al., 2005] Lima, A., Zen, H., Nankaku, Y., Tokuda, K., Kitamura, T., and Resende, F. G. (2005). Applying sparse kpca for feature extraction in speech recognition. *IEICE Transactions on Information and Systems*.

[Makhoul, 1973] Makhoul, J. (1973). Spectral analysis of speech by linear prediction. *IEEE Transactions on Audio and Electroacoustics*.

[Mingoti, 2005] Mingoti, S. A. (2005). *Análise de dados através de métodos de estatística multivariada: Uma abordagem aplicada*. Editora UFMG.

[Moore, 2001] Moore, G. L. (2001). *Adaptive Statistical Class-based Language Modelling*. PhD thesis, University of Cambridge.

[Moreau, 2002] Moreau, N. (2002). Htk (v.3.1): Basic tutorial. http://www.nue.tu-berlin.de/wer/moreau/.

[Odell et al., 1995] Odell, J., Ollason, D., Woodland, P., Young, S., and Jansen, J. (1995). *The HTK Book for HTK V2.0*. Cambridge University Press, Cambridge, UK.

[Olson and Belar, 1956] Olson, H. F. and Belar, H. (1956). Phonetic typewriter. *J. Acoust. Soc. Am.*, 28(6):1072–1081.

[Oppenheim et al., 1999] Oppenheim, A. V., Schafer, R. W., and Buck, J. R. (1999). *Discrete-time signal processing (2nd ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[Pinheiro and Aluísio, 2003] Pinheiro, G. M. and Aluísio, S. M. (2003). Córpus nilc: descrição e análise crítica com vistas ao projeto lacio-web. Technical report, Núcleo Interinstitucional de Linguística Computacional.

[Rabiner, 1989] Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286.

[Rabiner and Juang, 1993] Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[Reyment and Jöreskog, 1996] Reyment, R. A. and Jöreskog, K. G. (1996). *Applied Factor Analysis in the Natural Sciences*. Cambridge University Press, N. Y.

[Rivarol Vergin and Farhat, 1999] Rivarol Vergin, D. O. and Farhat, A. (1999). Generalized mel-frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. *IEEE Transactions on Speech and Audio Processing*, 7(5):525–532.

[Schafer and Rabiner, 1990] Schafer, R. W. and Rabiner, L. R. (1990). Digital representations of speech signals. *Readings in speech recognition*, pages 49–64.

[Schroeder, 1977] Schroeder, M. R. (1977). Recognition of complex acoustic signals. *Life Science Research Reports*.

[Schwartz et al., 1989] Schwartz, R., Barry, C., Chow, Y.-L., Derr, A., Feng, M.-W., Kimball, O., Kubala, F., Makhoul, J., and Vandegrift, J. (1989). The bbn byblos continuous speech recognition system. In *HLT '89: Proceedings of the workshop on Speech and Natural Language*, pages 94–99, Morristown, NJ, USA. Association for Computational Linguistics.

[Skowronski and Harris, 2003] Skowronski, M. D. and Harris, J. G. (2003). Improving the filter bank of a classic speech feature extraction algorithm. In *ISCAS (4)*, pages 281–284.

[Steeneken and Geurtsen, 1988] Steeneken, H. and Geurtsen, F. (1988). Description of the rsg-10 noise database. Technical report, TNO Institute for Perception, The Netherlands.

[Stevens and Newmann, 1937] Stevens, V. and Newmann (1937). A scale for the measurement of the psychological magnitude of pitch. *Journal of the Acoustical Society of America*.

[Takiguchi and Ariki, 1990] Takiguchi, T. and Ariki, Y. (1990). Robust feature extraction using kernel pca. *Readings in speech recognition*, pages 65–74.

[V. Zue and Ward, 1996] V. Zue, R. C. and Ward, W. (1996). Survey of the state of the art in human language technology. http://cslu.cse.ogi.edu/HLTsurvey/ch1node4.html.

[Ynoguti, 1999] Ynoguti, C. A. (1999). *Reconhecimento de Fala Contínua Utilizando Modelos Ocultos de Markov*. PhD thesis, Faculdade de Engenharia Elétrica - UNICAMP.