Dissertação de Mestrado

Inatel
Instituto Nacional de Telecomunicações

MODELAGEM DE COEFICIENTES DE ADAPTAÇÃO PARA SISTEMAS DE RECONHECIMENTO AUTOMÁTICO DE FALA

TATIANE MELO VITAL

SETEMBRO/2013

Modelagem de Coeficientes de Adaptação para Sistemas de Reconhecimento Automático de Fala

Tatiane Melo Vital

Dissertação apresentada ao Instituto Nacional de Telecomunições, como parte dos requisitos para obtenção do título de Mestre em Telecomunicações.

Orientador: Prof. Dr. Carlos Alberto Ynoguti

Santa Rita do Sapucaí
2013

Vital, Tatiane Melo
V836m

Modelagem de Coeficientes de Adaptação para Sistemas de Reconhecimento Automático de Fala. / Tatiane Melo Vital. – Santa Rita do Sapucaí, 2013.

108 p.

Orientador: Prof. Dr. Carlos Alberto Ynoguti.

Dissertação de Mestrado – Engenharia de Telecomunicações – Instituto Nacional de Telecomunicações – INATEL.

Inclui bibliografia e anexo.

1. Reconhecimento de Fala Robusto 2. MAP 3. Coeficiente de Adapatação 4. Multi-Estilo 5. Ruído 6. Engenharia de Telecomunicações. I. Ynoguti, Carlos Alberto. II. Instituto Nacional de Telecomunicações – INATEL. III. Título.

CDU 621.39

Dissertação defendida e aprovada em 30/09/2013, pela comissão julgadora:

(Prof. Dr. Carlos Alberto Ynoguti / INATEL)

(Prof. Dr. Francisco José Fraga da Silva / UFABC)

(Prof. Dr. Alberto Yoshihiro Nakano / UTFPR)

Prof. Dr. José Marcos Câmara Brito

Coordenador do Curso de Mestrado / INATEL

Ao meu esposo e minha
linda princesa.

Agradecimentos

Agradeço, primeiramente, a Deus Pai que me escolheu antes mesmo de eu nascer, protegeu e cuidou de mim desde o ventre materno.

Em especial à meu esposo, Richard, mais que amigo e patrocinador deste sonho. Que esteve presente em todos os momentos, na alegria e na tristeza, no ânimo e no desânimo, nas lutas e nas conquistas. E, que me ajudou a transpor os obtáculos que surgiram no caminho mantendo sempre otimismo e palavras de incentivo quando o desânimo surgia em meio às dificuldades.

À minha pequenina princesa, Letícia, presente a mim entregue por Deus para me dar forças a continuar lutando pelos meus ideais. Que me deu fôlego e forças para não desanimar quando tudo parecia não ter mais solução. A quem prometi que concluiria este trabalho com êxito.

A meus pais Manoel e Adélia que sempre investiram na minha educação e incentivaram meu crescimento pessoal e profissional. Aos familiares que acreditaram no meu potencial e deram forças nos momentos difíceis.

Em especial à minha sogra Elita, uma mãe. Sempre atenciosa, amável e presente não só nos momentos alegres, mas principalmente nos momentos difícies.

À Vó Lúcia (*in memorian*), sei que este momento representaria muito para ela.

A meu orientador, Doutor Carlos Alberto Ynoguti, que direcionou os diversos temas de estudo e auxiliou o desenvolvimento deste trabaho.

Ao conselho do mestrado pela compreensão, apoio e por acreditarem na concretização deste trabalho.

Aos meus amigos, pelo incentivo e pelas palavras de ânimo.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ANN *Artificial Neural Networks* - Redes Neurais Artificiais
ASR *Automatic Speech Recognition* - Reconhecimento Automático de Fala
CMN *Cepstral Mean Normalization* - Normalização da Média Cepstral
DCT *Discrete Cossine Transform* - Transformada Discreta de Cosseno
DFT *Discrete Fourier Transform* - Transformada Discreta de Fourier
EMR *Electronic Medical Records* - Registros Médicos Eletrônicos
FWT *Fast Wavelet Transform* - Transformada Wavelet Rápida
GMM *Gaussian Mixture Models* - Modelo de Misturas Gaussianas
GPS *Global Positioning System* - Sistema de Posicionamento Global
HMM *Hidden Markov Models* - Modelos Ocultos de Markov
IMELDA *Integrated Mel-scale with Linear Discriminant Analysis* - Análise Discriminante Linear com Escala Mel Integrada
LCT *Local Cossine Transform* - Transformada do Cosseno Local
MAP *Maximum a Posteriori* - Máximo a Posteriori
MFCCs *Mel Frequency Cepstrum Coefficients* - Coeficientes Mel-Cepstrais
MLLR *Maximum Likelihood Linear Regression* - Regressão Linear de Máxima Verossimilhança
MMSE *Minimum Mean Square Error* - Mínimo Erro Médio Quadrático
NAT *Noise Adaptative Training* - Treinamento Adaptativo com ruído
PLP *Perceptual Linear Predictive* - Predição Linear Perceptual
PMC *Parallel Model Combination* - Modelo de combinação parelela
IP *Internet Protocol* - Protocolo de comunicação utilizado na rede mundial de computadores
RASTA-PLP *RelAtive SpecTrAl Perceptual Linear Predictive* - Predição Linear Perceptual Espectral Relativa
SCTK *Speech Recognition Scoring Toolkit* - Ferramenta de avaliação de Reconhecimento de Fala
SGMM *Subspace Gaussian Mixture Models* - Modelo de Mistura no subespaço Gaussiano
SNR *Signal-to-Noise Ratio* - Relação Sinal-Ruído
SVL *Sistema de Verificação de Locutor*

SVM *Support Vector Machines* - Máquinas de Vetor de Suporte
VHS *Video Home System* - Sistema de Vídeo Doméstico
VTS *Vector Taylor Series* - Séries de Taylor Vetorial
VOIP *Voice over Internet Protocol* - Voz sobre IP
WA *Word Accuracy* - Taxa de acerto
WER *Word Error Rate* - Taxa de palavras erradas

Lista de Símbolos

a Inclinação da curva logística
α Coeficiente de adaptação
α' Valor médio ponderado para o coeficiente de adaptação
b Deslocamento horizontal da curva logística
c Deslocamento vertical da curva logística
C Número de palavras corretas
D Número de deleções
ΔWA Ganho na taxa de acerto de palavras
dB Decibel
$E_i(x^2)$ Variância
$E_i(x)$ Média
$Energia_{sinaldafala}$ Energia do sinal de fala
$Energia_{ruído}$ Energia do ruído
$f(x)$ Função logística
$h(t)$ Resposta impulsiva do canal e microfone
$h_{canal}(t)$ Resposta impulsiva do canal
$h_{mic}(t)$ Resposta impulsiva do microfone
$H(\omega)$ Resposta em frequência do canal e microfone
Hz Unidade que representa a frequência de um sinal
I Número de inserções de palavras
M Número de densidades Gaussianas
$n(t)$ Componente do ruído aditivo no domínio do tempo
$n_1(t)$ Ruído de fundo no domínio do tempo
$n_2(t)$ Ruído aditivo do canal de transmissão no domínio do tempo
$n_3(t)$ Ruído no receptor no domínio do tempo
n_i Peso
$N(\omega)$ Resposta em frequência do ruído aditivo
N Número total de palavras na sentença de referência
p Função densidade de probabilidade
$Pr(i|x_t)$ Alinhamento probabilístico dos vetores de treinamento
$s(t)$ Sinal de fala limpo no domínio do tempo

$S(\omega)$ Resposta em frequência do sinal de fala limpo
S Número de substituições de palavras
$y(t)$ Sinal de fala corrompido no domínio do tempo
$Y(\omega)$ Resposta em frequência do sinal de fala corrompido
x Nível do ruído em dB
$WA(i)$ Taxa de acerto de palavras
ω Peso da mistura
$\hat{\omega}_i$ Peso após adaptação
$\hat{\mu}_i$ Média após processo de adaptação
$\hat{\sigma}_i^2$ Variância após proesso de adaptação
W Unidade de potência elétrica

Resumo

O descasamento entre as condições acústicas das locuções utilizadas no treinamento e aquelas vivenciadas pelos sistemas de reconhecimento automático de fala é um dos fatores responsáveis pela degradação de seu desempenho quando operam em ambientes ruidosos. Esta é uma questão relevante na realidade atual, com o aumento do uso destes sistemas em dispositivos móveis.

Dentre as várias técnicas propostas na literatura para minimizar este problema, destaca-se a adaptação baseada no critério do Máximo a Posteriori (MAP), onde os modelos acústicos gerados na etapa de treinamento podem ser adaptados para a condição de ruído (tipo e intensidade) experimentada pelo sistema. Nesta abordagem, amostras do ruído são utilizadas para modificar os parâmetros dos modelos acústicos de modo a maximizar a taxa de acertos. A intensidade desta modificação depende de um coeficiente de adaptação, que em geral é calculado de forma empírica, através um processo de varredura.

Nesta dissertação é realizado um modelamento de como os valores ótimos deste coeficiente se comportam com o tipo e a intensidade do ruído e, a partir deste resultado, propõe-se um algoritmo para determinar um valor adequado para o mesmo. Este baseia-se no ajuste paramétrico através da aplicação da curva logística minimizando tempo de processamento. Não se consegue com este algoritmo determinar o coeficiente de adaptação que retorne a máxima taxa de acertos em todos os casos, mas a um coeficiente que proporcione um aumento desta taxa. Nos testes realizados, obteve-se um ganho médio de 3% na taxa de acertos.

Palavras-chave: Reconhecimento de Fala Robusto, MAP, Coeficiente de Adaptação.

Abstract

The mismatch between the acoustic conditions of the training utterances and those experienced by automatic speech recognition systems is one of the responsible factors for its performance degradation when operating in noisy environments. This is a relevant issue in the current reality with increasing use of these systems on mobile devices.

Among the various techniques proposed in the literature to minimize this challenge, the adaptation based on Maximum a Posteriori criteria (MAP) stands out where the acoustic models from training stage can be adapted to the noise condition (type and level) experienced by the system. In this approach, noise samples are used to modify the parameters of the acoustics models to maximize the word accuracy. The intensity of this modification depends on an adaptation coefficient which is usually calculated empirically through a grid search.

In this dissertation, a modeling of how the great values of these coefficients behave according the type and level of noise is performed. From this result, an algorithm to determine an appropriate value for it is proposed. It is based on the parametric adjustment by application of logistic curve minimizing the processing time. The adaptation coefficient provided by this algorithm does not lead the maximum word accuracy for all cases, but it always provides gain. The experimental results show an gain of 3% on word accuracy.

Keywords: Robust speech recognition, MAP, adaptation coefficient.

Capítulo 1

Introdução

O intenso fluxo de mercadorias, capitais, produtos, serviços e tecnologias entre os países gerou um novo conceito, a Globalização. Esta integração mundial possibilitou não apenas um avanço, mas uma revolução tecnológica que trouxe uma nova concepção mercadológica, onde as exigências e as críticas, ao longo de décadas, contribuíram positivamente para o desenvolvimento de novos produtos e aplicações [1].

Uma nova cultura padronizada de consumo tem surgido e com ela o conceito de uniformização, a convergência digital. O avanço na área computacional e das telecomunicações tem possibilitado não apenas a integração das diversas tecnologias já existentes, mas tem propiciado o desenvolvimento de novas tecnologias.

Hoje, o dia a dia das pessoas de diferentes faixa etárias é regado de produtos e serviços em constante aperfeiçoamento. A inovação nos processos produtivos aumentou a capacidade de produção; A TV preta e branca deu lugar à TV de alta definição 3D; o antigo cartão perfurado, hoje, completamente extinto, foi substituído por circuitos integrados de altíssimo desempenho e as antigas fitas cassetes e VHS por blue-rays; no ramo das telecomunicações, a Internet tornou-se um dos meios mais importantes de comunicação; além da possibilidade da voz sobre IP (VOIP), os antigos aparelhos pesados de telefones fixos deram lugar à mobilidade proporcionada pelos novos aparelhos compactos e leves, que possibilitam não apenas executar uma simples chamada, mas tarefas mais complexas como o envio de mensagens de texto ou voz, agendas, busca de informações úteis, GPS, entre outros serviços via comandos de voz.

A constante busca por inovação, qualidade, praticidade e, principalmente, conforto tem contribuído para o fortalecimento do reconhecimento da fala. O reconhecimento da voz representa uma mudança de paradigma: alternativamente ao uso de chaves ou botões, as máquinas podem também responder a comandos de voz. Esta é uma forma humanizada de interação entre o homem e a máquina. A

aplicação do reconhecimento de voz como interface em vários serviços tem sua empregabilidade calcada na redução de custos operacionais, aumento de segurança, inovação, avanço tecnológico, aumento de receitas, acessibilidade a portadores de necessidades especiais, entre outros [2].

Os sistemas de reconhecimento automático de fala (ASR - *Automatic Speech Recognition*) têm se mostrado eficientes quando integrados a plataformas escaláveis e de funcionamento ininterrupto, permitindo o uso inteligente de árvores de atendimento e, proporcionando um meio simples e natural de comunicação dispensando interfaces adicionais entre homem e máquina. Devido a esta característica, esta tecnologia tem sido amplamente utilizada como interface de atendimento em *call centers*, por exemplo [3].

A automação no atendimento pode ser empregada em diversas aplicações, tais como na área de entretenimento e informação. Podem-se criar produtos que disponibilizam menus de fácil navegação e interatividade, como acesso a agendas, discagem ativada por voz, roteamento de chamadas, horóscopo, previsão do tempo, leitura de e-mails, controle de acessos, notícias, novelas, horários de vôos e partidas de transportes públicos (metrô, trem e ônibus), serviços bancários, *voice-commerce*, entre outros [4].

Em sistemas de controle de acesso a serviços, informações ou lugares, a aplicação da voz para identificação de usuário (SVL - Sistema de Verificação de Locutor) é atrativa e um instrumento valioso, pois proporciona maior confiabilidade e praticidade [5].

O reconhecimento de fala tem sido empregada em cursos e treinamentos, como exemplo, cursos de língua estrangeira e treinamento de controladores de tráfego aéreo. Esta tecnologia vem sendo explorada em alguns programas da área militar em aplicações como: comando do sistema de piloto automático, orientação de coordenadas e parâmetros de lançamento de armas, ajuste de frequência de comunicação, controle de monitores de navegação entre outros. Adicionalmente, a tecnologia ASR já está presente no mercado e tem transformado a realidade de deficientes físicos e visuais, contribuindo para o diagnóstico de patologias, acompanhamento de tratamentos e processos de documentação médica (EMR - *Electronic Medical Records*) [6].

O contínuo crescimento em aplicações na área de reconhecimento de voz e locutor, tem exigido sistemas cada vez mais robustos e imunes aos diferentes tipos de ruídos [7][8][9][10][11]. Diversas técnicas e métodos existentes têm sido alvo de estudo com intuito de identificar melhoria e novos procedimentos que possam contribuir no desempenho destes sistemas. A seção a seguir descreve as principais características e funcionamento de um sistema básico de reconhecimento de fala.

1.1 Sistema básico de reconhecimento

A Figura 1.1 apresenta o diagrama básico de um sistema de reconhecimento de fala.

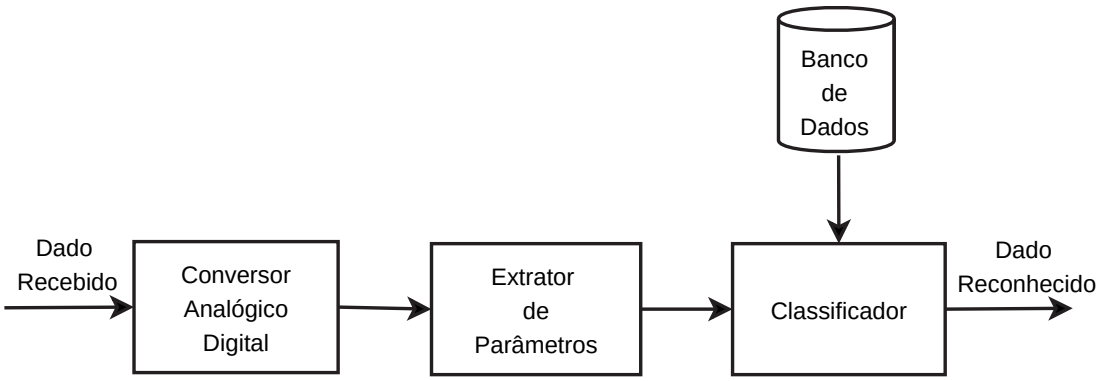

Figura 1.1: *Sistema básico de reconhecimento de locutor ou fala*

A função do conversor analógico/digital é converter o sinal acústico recebido em um sinal digital através da amostragem e quantização do sinal analógico [9].

O módulo de extração de parâmetros é responsável por extrair as informações relevantes da fala representando-a de forma a facilitar a etapa de reconhecimento.

O banco de dados é responsável por armazenar os padrões da fala (palavras inteiras para o caso de vocabulários de pequeno porte ou subunidades no caso de vocabulários de médio ou grande porte, podendo ainda conter informações lexicais e gramaticais além das semânticas).

O objetivo do classificador é mapear a locução a ser reconhecida para um dos padrões armazenados. Dentre os classificadores mais usuais estão: Modelos Ocultos de Markov (HMM - *Hidden Markov Models*), Redes Neurais Artificiais (ANN - *Artificial Neural Networks*), Máquinas de Vetor de Suporte (SVM - *Support Vector Machines*) e Modelos de Misturas de Gaussianas (GMM - *Gaussian Mixture Models*). O HMM é um dos modelos acústicos mais comumente empregado devido à sua simplicidade e flexibilidade [12][13].

1.2 Adversidades no reconhecimento da fala

Os sistemas ASRs são susceptíveis a diversos fatores, tais como: erros introduzidos entre codificadores e decodificadores de voz, saturação, ruído ambiente, alterações na voz, nasalidade, variabilidade em termos da pronúncia e de acentuação, semântica variável, efeito Lombard, velocidade de pronúncia, stress da voz, reverberação, atraso no canal de comunicação, ruído aditivo e convolucional,

latência, ruído musical proveniente de processamento para redução do ruído entre outros [14].

Sistemas automáticos de reconhecimento de fala ou locutor são altamente sensíveis a estes fatores e apresentam perda significativa de desempenho na presença de ruído tanto aditivo quanto convolucional. Portanto, faz-se necessária a aplicação de técnicas que proporcionem robustez a estes sistemas.

Existem diversas pesquisas nos campos de identificação de locutor bem como reconhecimento de fala onde diferentes técnicas e estratégias têm sido empregadas em cada um os blocos do sistema básico mostrado na Figura 1.2. O alvo dos estudos é oferecer soluções que proporcionem melhor desempenho do sistema, tanto na taxa de acertos quanto no tempo de processamento.

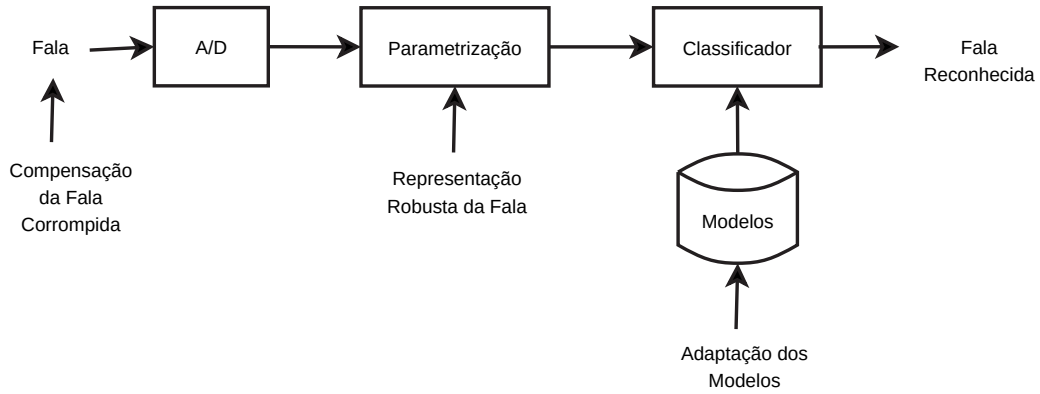

Figura 1.2: *Representação da interação dos métodos de reconhecimento robusto da fala em um ASR*

As principais técnicas propostas atuam na representação robusta da fala, na compensação dos parâmetros da fala corrompida ou na adaptação dos modelos de fala limpa [12].

O principal objetivo da parametrização robusta é escolher parâmetros da fala que minimizem as distorções introduzidas pelos ruídos aditivo e convolucional. Nesta abordagem, não é requerido processamento adicional. O foco das técnicas de compensação dos parâmetros da fala corrompida é processar o sinal da fala contaminada de forma a estimar o sinal de fala limpa. Para tal fim, os métodos compreendidos nesta categoria utilizam informações estatísticas do ruído. Na última categoria enquadram-se as técnicas que atuam nos modelos de fala limpa adaptando-os de forma a garantir robustez aos ASRs [15].

Dentre os métodos utilizados, pode-se citar:

1. Parametrização robusta

    - *PLP (Perceptual Linear Predictive) e suas variações RASTA-PLP (RelAtive SpecTrAl Perceptual Linear Predictive), J-RASTA-PLP e*

RASTA-MFCC (RelAtive SpecTrAl Mel-Frequency Cepstral Coefficient), as técnicas de predição linear e similares baseiam-se no modelo auditivo. O foco do PLP é realizar um modelamento tal que o mesmo seja insensível a variações muito lentas no domínio da frequência de forma similar ao comportamento do ouvido. O aprimoramento do PLP, RASTA-PLP é indicado para robustez ao ruído convolucional e, seu aprimoramento, J-RASTA, aplica-se tanto para ruído convolucional quanto aditivo. O RASTA-MFCC possui melhor desempenho comparado ao RASTA-PLP devido à eficiência proveniente da filtragem temporal do sinal da fala [9][12][16];

- *CMN (Cepstral Mean Normalization)*, esta técnica foi desenvolvida com o propósito de reduzir as distorções introduzidas pelo canal. A idéia básica é que o ruído convolucional pode ser analisado no domínio cepstral como uma distorção aditiva, o que facilita sua implementação. A fala limpa é parametrizada utilizando parâmetros cepstrais e um valor médio é subtraído de cada elemento Cepstral [9][12][17];
- *Análise Cepstral*, onde os parâmetros MFCCs (*Mel Frequency Cepstrum Coefficients*) são empregados para representar o vetor característico da fala. A vantagem desse tipo de parametrização é possibilitar uma representação equiparável a resposta em frequência do ouvido humano e permitem descrever características fonéticas importantes o sinal da fala. Modificações baseadas nestes coeficientes têm sido propostas devido ao desempenho crítico na representação da fala corrompida [5][12][16][18];
- *IMELDA (Integrated Mel-scale with Linear Discriminant Analysis)*, possibilita mais robustez que MFCCs, porém a mesma descreve com a redução da relação sinal-ruído [12].

2. Compensação da fala corrompida

- *Subtração Espectral*, esta técnica tem como foco estimar o sinal da fala a partir do sinal da fala corrompida e de uma estimativa do ruído obtida em períodos sem atividade vocal. Para tal fim, considera-se que o sinal da fala e o ruído são descorrelacionados. O principal desafio desta técnica é a inserção do ruído musical proveniente da introdução de não-linearidades no espectro [9][12][17][19][20][21];
- *MMSE (Minimum Mean Square Error) e variações*, diversas técnicas baseadas em MMSE-LSA (*Minimum Mean Square Error-log Spectral Amplitude*) são empregadas para estimar o espaço característico com

o intuito de melhorar o desempenho de sistemas de reconhecimento de fala em condições adversas [22][23];

- *MAP (Maximum a Posteriori)*, este método calcula os parâmetros do ruído maximizando a probabilidade *a-posteriori* da fala limpa a partir da fala corrompida e estatísticas da fala limpa [24][25][26];
- *Transformadas, DFT (Discrete Fourier Transform), DCT (Discrete Cossine Transform), LCT (Local Cossine Transform), FWT (Fast Wavelet Transform) e suas variações*, largamente empregadas nos algoritmos de reconhecimento de fala devido a facilidade que proporcionam na separação do sinal da fala do ruído no domínio da frequência [27];
- *VTS (Vector Taylor Series)*, a estimativa dos parâmetros da fala corrompida é obtida a partir de uma aproximação da série finita de Taylor que pode causar descasamento residual entre os dados observados e o modelo adaptado. O processamento desta técnica está diretamente relacionado aos parâmetros a serem compensados [28][29];
- *SGMM (Subspace Gaussian Mixture Models)*, emprega um modelo de subespaço globalmente compartilhado entre os estados de forma a capturar as maiores variações do modelo provendo uma representação compacta dos modelos acústicos resultando em uma estimação robusta dos parâmetros e melhora no desempenho de sistemas de reconhecimento de fala, principalmente, quando a base de treinamento é reduzida [30][31].
- *Filtro de Wiener*, solução proposta por Norbert Wiener com o propósito de minimizar a distância da média quadrática entre o sinal estimado e o sinal desejado. O objetivo é reduzir o efeito do ruído no sinal da fala corrompida através da comparação com uma estimativa do sinal da fala limpa [12][20][32][33];

3. Adaptação dos modelos de fala limpa

- *MLLR (Maximum Likelihood Linear Regression)*, técnica que se enquadra na categoria de adaptação de modelo cujo princípio básico é considerar que o descasamento entre as condições de teste e treinamento podem ser modeladas por transformações lineares dos modelos acústicos. Os parâmetros das transformações são estimados através do critério de máxima verossimilhança na adaptação linear dos dados [24];
- *PMC (Parallel Model Combination)*, o modelo da fala corrompida é derivado da combinação entre o modelo do ruído e da fala limpa. A

principal desvantagem do emprego deste método é necessidade de se estimar os parâmetros do modelo do ruído para compensar posteriormente os parâmetros da fala. Diferentes tipos de aproximações podem ser usados para obter-se os parâmetros, tais como: log-normal, log-add e integração numérica[9][15][22][28];

- *Combinação de modelo no domínio cepstral*, técnica similar ao PMC, porém os modelos são atualizados no domínio cepstral, o que reduz a complexidade computacional devido a não requerer conversões entre o domínio linear e cepstral [22];
- *NAT (Noise Adaptative Training)*, usa treinamento multi-estilo e, posteriormente transforma o modelo obtido em um modelo "pseudo-limpo" com objetivo de reduzir as distorções do ambiente. Esta técnica surgiu com objetivo de contornar os problemas encontrados na adaptação VTS [29].

Existem métodos baseados nas técnicas já anteriormente citadas que propõem melhorias e técnicas híbridas, formadas a partir da combinação de dois ou mais métodos [10][11][14][18][21][22][34][35][36][37][38][39][40][41][42][43][44].

Este trabalho emprega uma técnica de reconhecimento robusto, a adaptação Bayesiana baseada no MAP combinada ao treinamento multi-estilo, que serão abordadas no Capítulo 3.

Considerando a aplicabilidade de sistemas de reconhecimento em diferentes tipos de ambiente, o presente trabalho visa contribuir com análise de desempenho dos mesmos em face ao ruído aditivo. Para exemplificar o emprego destes sistemas, pode-se citar os smartphones, hoje comumente utilizados nos mais adversos meios.

1.3 Objetivo e contribuições

O presente trabalho analisa o desempenho de um sistema de reconhecimento de fala que emprega a técnica de treinamento multi-estilo seguida da adaptação Bayesiana dos modelos acústicos.

O procedimento de determinação do valor ótimo do coeficiente usado na adaptação MAP envolve uma busca por varredura e, portanto, é um procedimento de elevado custo computacional. A principal contribuição desta dissertação é permitir a escolha adequada de coeficientes de adaptação para diferentes relações sinal-ruído (SNR - *Signal-to-Noise Ratio*) de cada tipo de ruído utilizado através de um algoritmo baseado no ajuste paramétrico através da aplicação da curva logística sem aumentar o custo computacional.

Os testes realizados foram baseados em dois tipos de treinamentos: com locuções limpas e locuções corrompidas com diferentes tipos de ruído aditivo para relações sinal-ruído 15 dB e 20 dB.

A análise do desempenho do sistema de reconhecimento pode ser resumido em duas etapas:

- *Treinamento multi-estilo*, o HMM foi treinado com locuções corrompidas por diferentes tipos de ruído com SNR 15 dB e 20 dB. Esta abordagem proporcionou um ganho de 6,89 % na taxa de acerto de palavras comparado ao sistema treinado apenas com dados limpos;

- *Adaptação baseada no MAP de um sistema utilizando a técnica de multi-estilo.* O HMM proveniente do treinamento da etapa anterior foi adaptado para diferentes tipos de ruído considerados presentes no momento do reconhecimento. Um ganho adicional de 1,74 % foi alcançado. Portanto, as técnicas proporcionaram conjuntamente um ganho de 8,63 % com base no desempenho de referência do sistema.

Dentre as várias aplicações possíveis, este trabalho, está focado no reconhecimento robusto de fala contínua com independência de locutor e vocabulário de médio porte.

1.4 Estrutura da dissertação

Este trabalho está estruturado em 6 capítulos.

O segundo capítulo enfoca os efeitos do ruído nos sistemas de reconhecimento de fala. Nas seções deste são apresentados os modelos dos ruídos convolucional e aditivo, bem como, a influência destes ruídos em sistemas de reconhecimento automático de fala e o modelo real da fala corrompida em ambientes adversos. Posteriormente, é descrito o modelo simplificado após considerações dos efeitos desprezíveis em condições normais.

O terceiro capítulo aborda as principais técnicas empregadas para adaptação dos modelos acústicos da fala limpa para ambientes ruidosos.

No quarto capítulo é apresentado o sistema de reconhecimento utilizado, mostrando suas principais características e configurações necessárias para a execução dos treinamentos e testes de reconhecimento automático. Neste capítulo é caracterizada a base de dados utilizada nas etapas de treinamento e reconhecimento e são apresentadas as técnicas de avaliação de desempenho de ASRs.

O quinto capítulo apresenta os resultados experimentais.

Finalmente, no sexto capítulo são apresentadas as considerações finais e as sugestões para trabalhos futuros.

Capítulo 2

A influência de fontes ruidosas no desempenho de sistemas de reconhecimento de fala

O reconhecimento de fala em ambientes adversos tem sido alvo de diversas discussões e estudos [45]. Em geral, tais sistemas são projetados para operarem em condições de baixo ruído e interferência. O descasamento acústico entre as condições de treino e reconhecimento está entre os fatores que têm grande influência no desempenho destes, pois eles são altamente susceptíveis à presença de ruído e interferência do meio.

Estudos com objetivo de garantir robustez a estes sistemas têm levado a novos métodos ou mesmo combinação de técnicas já anteriormente propostas com intuito de vencer um dos maiores obstáculos à aplicabilidade da tecnologia de reconhecimento automático de fala: o ruído. O foco das pesquisas é encontrar uma relação de compromisso entre o ruído eliminado e as distorções introduzidas no sinal de fala devido ao procedimento empregado.

Neste Capítulo são abordados tópicos relativos ao modelo da fala e do reconhecimento robusto.

2.1 Representação real do modelo do sinal de fala

Em diversas aplicações verifica-se que o sinal de voz é contaminado com ruído aleatório proveniente do ambiente, o ruído aditivo e, com o ruído da resposta

em frequência do microfone e do canal, denominado ruído convolucional. Nas subseções a seguir serão abordados primeiramente a influência dos ruídos aditivo e convolucional separadamente de forma a facilitar a compreensão e, posteriormente, será feita uma análise completa da ação destes e de outros efeitos indesejados no sinal de fala.

2.1.1 Ruído aditivo

O ruído aditivo é uma distorção introduzida no sinal de fala no domínio do tempo proveniente de qualquer fonte ruidosa, tais como: conversações paralelas, automóveis, equipamentos elétricos ou eletrônicos, etc. Este ruído degrada o desempenho dos sistemas de reconhecimento de fala ao provocar a diminuição da SNR. O mesmo pode ser representado no domínio do tempo por:

$$y(t) = s(t) + n(t) \tag{2.1}$$

onde $y(t)$ representa o sinal de fala corrompido, $s(t)$ o sinal de fala limpo e $n(t)$ é o ruído aditivo.

No domínio da frequência, o sinal corrompido pode ser representado por:

$$Y(\omega) = S(\omega) + N(\omega) \tag{2.2}$$

onde $Y(\omega)$ representa o espectro do sinal de fala corrompido, $S(\omega)$ é o espectro do sinal de fala limpo e $N(\omega)$ é o espectro do ruído aditivo.

O ruído aditivo é variante no tempo e pode corromper as componentes do sinal de fala. Esta particularidade de aleatoriedade do ruído aditivo dificulta a definição do método mais adequado para redução dos efeitos do mesmo.

2.1.2 Ruído convolucional

O ruído convolucional é uma distorção introduzida no sinal de fala devido à influência do canal de transmissão e microfone que é representado pela convolução no domínio do tempo mostrado por:

$$y(t) = s(t) * h(t) \tag{2.3}$$

onde $y(t)$ representa o sinal de fala corrompido, $s(t)$ o sinal de fala limpo e $h(t)$ é a resposta impulsiva combinada dos efeitos do canal e do microfone.

No domínio da frequência tem-se:

$$Y(\omega) = S(\omega).H(\omega) \tag{2.4}$$

onde $Y(\omega)$ representa o espectro do sinal de fala corrompido, $S(\omega)$ é o espectro do sinal de fala limpo e $H(\omega)$ é a resposta em frequência do canal e microfone.

A resposta impulsiva $h(t)$ é considerada invariante no tempo e independente do sinal de fala [9]. Esta característica facilita a supressão do ruído convolucional.

2.1.3 Modelo completo do sinal de fala corrompido

A modelagem completa do sinal de fala num sistema de reconhecimento é complexa tendo em vista as diferentes distorções que o mesmo pode sofrer, sendo estas particulares de cada ambiente e de outras características acústicas. Estas distorções podem ser estacionárias ou não estacionárias, contínuas ou descontínuas, correlacionadas ou descorrelacionadas com o sinal de fala [12].

Considerando a abordagem apresentada em [12], na qual as fontes de ruído são consideradas independentes, pode-se representar o sinal de fala corrompida pela Equação (2.5):

$$y(t) = [\{([s(t)|^{Stress}_{Lombard}]_{n_1(t)} + n_1(t)) * h_{mic}(t) + n_2(t)\} * h_{canal}(t)] + n_3(t) \quad (2.5)$$

onde $y(t)$ representa o sinal de fala corrompido, $s(t)$ o sinal de fala limpo, $n_1(t)$ é o ruído de fundo, $h_{mic}(t)$ é a resposta impulsiva do microfone, $n_2(t)$ e $h_{canal}(t)$ são respectivamente o ruído aditivo e resposta impulsiva do canal de transmissão e $n_3(t)$ é o ruído no receptor. Esta simplificação é possível, pois $s(t)$, $n_1(t)$ e $n_2(t)$ são descorrelacionadas.

De acordo com Gales, o efeito Lombard (tendência natural das pessoas em aumentar o esforço vocal em ambientes ruidosos) e o stress fisiológico (causado por diversos fatores como execução de diversas atividades em paralelo ou até mesmo cansaço físico) sob condições normais podem ser desprezados possibilitando a simplificação da Equação (2.5) [12]:

$$y(t) = s(t) * h(t) + n(t) \quad (2.6)$$

onde $y(t)$ representa o sinal de fala corrompido, $s(t)$ o sinal de fala limpo, $h(t)$ é a resposta impulsiva do efeito combinado do canal e do microfone e $n(t)$ representa o ruído aditivo.

No domínio da frequência, a Equação (2.6) é representada por:

$$Y(\omega) = S(\omega).H(\omega) + N(\omega) \quad (2.7)$$

onde $Y(\omega)$ representa a resposta em frequência do sinal de fala corrompido, $S(\omega)$ é a resposta em frequência do sinal de fala limpo, $H(\omega)$ é a resposta em frequência

do canal e microfone e $N(\omega)$ é a resposta em frequência do ruído aditivo.

O modelo de fala corrompido simplificado é representado na Figura 2.1.

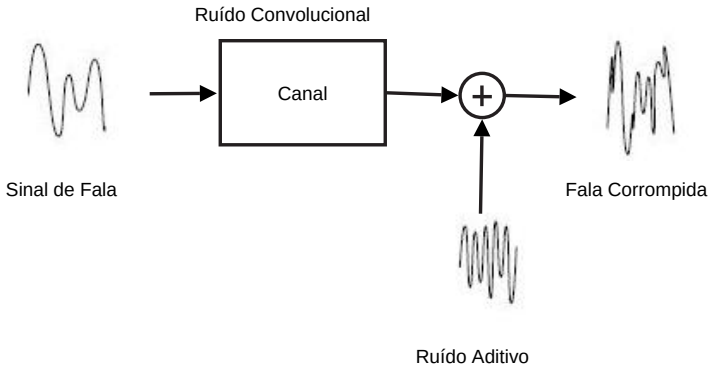

Figura 2.1: *Modelo simplificado das distorções no sistema de reconhecimento*

2.2 Reconhecimento de fala robusto

É comumente abordado na literatura relativa ao reconhecimento de fala que as condições ambientais adversas bem como as variabilidades decorrentes da fala, tais como stress, pronúncia, entre outras, degradam o desempenho dos sistemas ASRs. Além disso, são diversas as fontes ruidosas que podem interferir durante a operação destes sistemas.

Em virtude de tais distorções faz-se necessário a aplicação de técnicas e algoritmos que minimizem estes efeitos indesejados que alteram o sinal de voz de forma a garantir integibilidade aos sistemas de reconhecimento de fala.

Como apresentado no primeiro capítulo, para suprimir as distorções introduzidas no sinal da fala é possível utilizar-se da parametrização robusta do sinal limpo, da compensação dos parâmetros da fala corrompida ou de técnicas que atuem diretamente no módulo de reconhecimento.

Capítulo 3

Treinamento Multi-Estilo, Adaptação Bayesiana e Modelo Logístico

Neste capítulo serão descritas em detalhes as duas técnicas utilizadas neste trabalho: o treinamento multi-estilo e a adaptação Bayesiana. Além disso, propõe-se um algoritmo para a modelagem do coeficiente de adaptação segundo aproximação para curva logística.

3.1 Treinamento Multi-Estilo

O treinamento multi-estilo, também conhecido como multi-condição, emprega locuções corrompidas na etapa de treinamento, de forma a minimizar a queda de desempenho dos sistemas de reconhecimento de fala operando em ambientes ruidosos [46].

Esta técnica é bastante versátil podendo ser empregada de diferentes formas: o sistema pode ser treinado para um determinado tipo e nível de ruído, ou, com diferentes níveis de um tipo específico de distorção, ou ainda, com diferentes tipos e níveis de ruídos. Este trabalho emprega a última abordagem baseado nos resultados experimentais apresentados em [48].

Com o propósito de prover robustez aos HMMs às diversas variabilidades ambientais, distorção de canal, reverberação, além de outros efeitos indesejados, o treinamento multi-estilo depende da disponibilidade de base de dados adquirida em ambientes reais, de forma a capturar não apenas o sinal de voz, mas também, o sinal correspondente ao ruído. Porém, a construção de uma base de dados que re-

flita todas as possíveis situações reais de utilização do sistema ASR é impraticável dada a grande variabilidade de tipos e níveis de ruído.

Desta forma, em geral emprega-se uma base de dados artificial. Ou seja, as locuções ruidosas são geradas através da combinação entre sentenças gravadas em ambientes livres de ruído e diferentes tipos e níveis de ruído [46].

A Seção a seguir apresenta as principais características e fundamentação do MAP.

3.2 Adaptação baseada no MAP

A adaptação baseada no critério de Máximo a Posteriori, também denominada adaptação Bayesiana, mapeia os modelos acústicos do ambiente de treinamento para o modelo acústico do ambiente de reconhecimento. Geralmente, métodos calcados na utilização de modelos acústicos proporcionam melhor desempenho comparados às técnicas que baseiam-se no mapeamento do espaço característico do vetor de reconhecimento para o espaço característico de treinamento, pois possibilitam o modelamento da incerteza causada pelas estatísticas das condições ambientais adversas [49].

O HMM gerado a partir do treinamento utilizando locuções limpas ou corrompidas denomina-se modelo canônico. Estes modelos são posteriormente adaptados a partir de estatísticas do ruído (média, peso e variância) do meio no qual o sistema de reconhecimento está em operação. Portanto, o modelo de fala hipotético é derivado pela adaptação dos parâmetros do modelo canônico a partir dos dados do treinamento [47].

As equações relativas à adaptação são descritas a seguir. Dada uma amostra de ruído e os vetores do treinamento da fala hipotetizada (modelo canônico), $X = x_1, x_2, ..., x_T$, o alinhamento probabilístico dos vetores de treinamento dentro da amostra de ruído nas componentes do modelo, ou seja, para mistura i no modelo canônico, tem-se:

$$\Pr(i|x_t) = \frac{\omega_i p_i(x_t)}{\sum_{j=1}^{M} \omega_j p_j(x_t)} \tag{3.1}$$

onde M é o número de densidades Gaussianas, ω é o peso da mistura e p é a função densidade de probabilidade.

Então, $Pr(i|x_t)$ e x_t são usados para determinar os parâmetros estatísticos ruidosos: peso (n_i), média ($E_i(x)$) e variância ($E_i(x^2)$) conforme apresentado nas equações a seguir.

$$n_i = \sum_{t=1}^{T} Pr(i|x_t) \tag{3.2}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t) x_t \tag{3.3}$$

$$E_i(x^2) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t) x_t^2 \tag{3.4}$$

Finalmente, estas estatísitcas estimadas do ruído são usadas para adaptar os modelos canônicos gerando um novo modelo. As equações de adaptação relativas a estes parâmetros são:

$$\hat{\omega}_i = \left[\alpha_i^{\omega} n_i/T + (1 - \alpha_i^{\omega})\omega_i\right] \gamma \tag{3.5}$$

$$\hat{\mu}_i = \alpha_i^{m} E_i(x) + (1 - \alpha_i^{m})\mu_i \tag{3.6}$$

$$\hat{\sigma}_i^2 = \alpha_i^{\nu} E_i(x^2) + (1 - \alpha_i^{\nu})(\sigma_i^2 + \mu_i^2) - \mu_i^2 \tag{3.7}$$

onde:

- ω_i, μ_i and σ_i^2 são peso da mistura, média e variância do sistema treinado com dados limpos ou corrompidos;
- $\hat{\omega}_i$, $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ são peso da mistura, média e variância após o processo de adaptação;
- n_i, $E_i(x)$ and $E_i(x^2)$ são as estatísticas ruidosas e
- γ é o fator de escala usado para garantir que a soma de todos os pesos da mistura após a adaptação seja igual ao valor unitário.

O coeficiente de adaptação α controla o quanto da informação do ruído será introduzida no modelo. O intuito é empregar informações do ruído de forma a aproximar o modelo canônico que será utilizado como referência no classificador ao modelo de fala capturado no ambiente, minimizando o descasamento acústico entre as condições de treino e teste. Os coeficientes de adaptação α_i^{ω}, α_i^{m} e α_i^{ν} podem assumir valores no intervalo $[0, 1)$ e permitem realizar o balanço entre as velhas e novas estimativas para pesos, médias e variâncias, respectivamente.

Uma boa escolha destes parâmetros depende do tipo e nível de ruído, pois valores altos do coeficiente de adaptação enfatizam as estimativas do ruído enquanto baixos valores de α tendem a preservar o modelo original.

É possível empregar valores diferentes de coeficiente para adaptar pesos, médias e variâncias. Entretanto, o emprego de valores diferentes provê um ganho pequeno comparado ao uso de um único valor, ou seja, $\alpha_i^{\omega}=\alpha_i^{m}=\alpha_i^{\nu}=\alpha$, no processo de adaptação. Portanto, este trabalho usa um valor único do coeficiente de adaptação para todos os parâmetros, a exemplo do que foi feito em [47].

3.3 Modelo logístico ruidoso

Visto que a escolha de um fator de adaptação para um determinado tipo e nível de ruído é um fator impactante no desempenho do sistema ASR, o presente trabalho propõe um algoritmo que provê um valor de α baseado nas estatísticas dos distúrbios de fundo focando uma relação de compromisso entre o desempenho do sistema e a máxima taxa de acerto de palavras (WA - *Word Accuracy*) que será definida no Capítulo [**?**].

O método sugere uma nova aproximação empírica para a modelagem da relação entre a SNR e o coeficiente de adaptação para determinado tipo e nível de ruído. O objetivo é predizer o valor de α para uma larga faixa de diferentes intensidades do ruído de acordo com as diversas condições ambientais a partir de resultados de testes de reconhecimento para um número limitado de SNRs.

A técnica proposta tem como foco principal prover um coeficiente de adaptação que leve um valor para este parâmetro que proporcione ganho comparado a resposta do sistema usado como referência.

As etapas do algoritmo são:

- O passo inicial consiste em identificar coeficientes de adaptação no intervalo [0,1], que proporcionem ganho à resposta do sistema. No experimento foram empregados os seguintes níveis de SNR: 0, 5, 10, 15 e 20 dB;

- Os processos de reconhecimento realizados demonstraram que o passo inicial pode retornar um ou mais valores de α e estes não são necessariamente adjacentes. Neste sentido, verificou-se que o valor médio ponderado pela WA representa adequadamente o modelo, pois considera a influência de cada valor de α conforme Equação (3.8).

$$\alpha' = \frac{\sum_i WA(i) \times \alpha(i)}{\sum_i WA(i)} \tag{3.8}$$

  onde $WA(i)$ é a taxa de acerto de palavras para cada coeficiente de adaptação proveniente do passo inicial e α' é o valor médio ponderado que será

utilizado nos passos posteriores.

Portanto, após este passo, existe um conjunto de pares SNR x α.

- O passo final consiste em solucionar o sistema de equações proveniente do estágio anterior, seguido da aproximação para o modelo da curva logística. O modelo utilizado para ajuste paramétrico dos resultados experimentais possui 3 parâmetros livres como mostra a Equação (3.9).

$$f(x) = \frac{1}{1 + e^{b-ax}} - c \tag{3.9}$$

onde x é o nível do ruído em dB, $f(x)$ é o coeficiente de adaptação α'. Os parâmetros de configuração a, b e c controlam, respectivamente, inclinação, deslocamento horizontal e vertical da curva logística. Quanto menor o valor do parâmetro a, mais rápido é reduzido o valor do α frente à variação da intensidade do ruído. A Figura 3.1 mostra diferentes valores para o parâmetro a. Se o valor de b diminui, a curva é deslocada para direita e, vice-versa, conforme demonstrado na Figura 3.2. Quanto menor o valor de c, maior é o coeficiente de adaptação para uma dada SNR, visto que, a curva é deslocada para cima. Observe a Figura 3.3 com diferentes valores de c. Esta etapa é imprescendível na viabilização da escolha de coeficientes α para intensidades de ruído diferentes das usadas na geração das curvas que descrevem o comportamento característico de cada distorção.

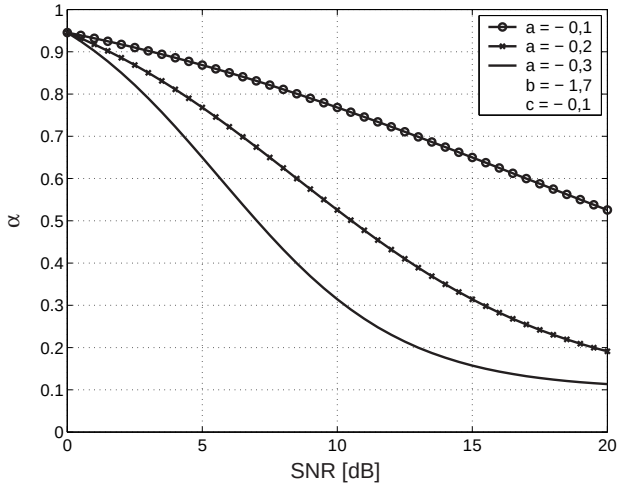

Figura 3.1: *Exemplo de diferentes valores para o parâmetro livre 'a'*

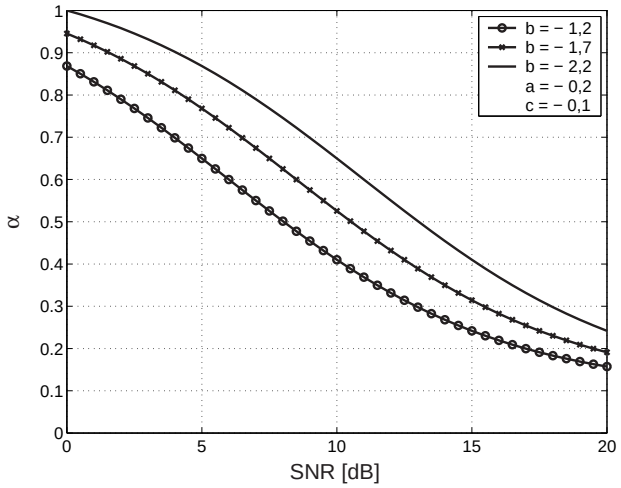

Figura 3.2: *Exemplo de diferentes valores para o parâmetro livre 'b'*

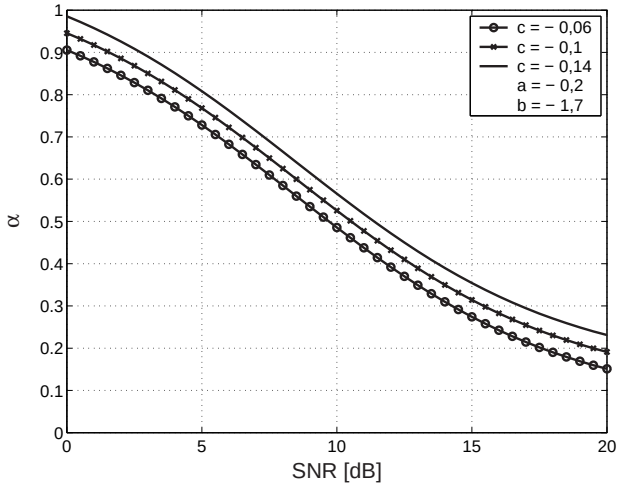

Figura 3.3: *Exemplo de diferentes valores para o parâmetro livre 'c'*

Capítulo 4

Materiais e Métodos

Este capítulo descreve a base de dados utilizada para avaliação do sistema ASR proposto por [50], o sistema de reconhecimento utilizado e a métrica para medição do desempenho.

4.1 Base de dados

Para avaliação dos métodos que são alvos de estudo deste trabalho optou-se por utilizar a base de dados gerada por Ynoguti [50]. As gravações das locuções que formam a base de dados foram efetuadas em ambiente silencioso com taxa de amostragem de 11.025 Hz e 16 bits.

Esta base de dados foi construída a partir da colaboração de 40 locutores, sendo 20 mulheres e 20 homens. Esta foi subdividida com objetivo de formar um banco de dados para o treinamento e outro para a verificação de desempenho do sistema ASR. O conjunto de treinamento conta com 30 locutores (15 mulheres e 15 homens), onde cada um gravou 40 locuções totalizando 1200 locuções. Para avaliação do sistema, utilizou-se as demais 400 locuções gravadas pelos outros 10 locutores (5 mulheres e 5 homens). As locuções que compõem a base foram distribuídas em 20 listas com 10 frases cada uma segundo [50]. O conjunto conta com 694 palavras do idioma Português, o que caracteriza um cenário de vocabulário de tamanho médio.

Com o objetivo de avaliar o desempenho do sistema ASR em diferentes situações, propôs-se dois tipos de treinamentos:

- *treinamento com dados limpos*;
- *treinamento multi-estilo.*

Como o propósito é a robustez de sistemas ASRs a ruídos, foram executados testes com locuções corrompidas com diferentes tipos de ruídos e relações sinal-ruído. Para atender a estes treinamentos e testes foi necessário produzir uma base de dados corrompida a partir da base de dados limpa. Isto foi feito com os sinais de ruído presentes na base Aurora.

4.2 Ruídos da base AURORA

Para gerar as locuções corrompidas, utilizou-se os oito diferentes tipos de ruído disponíveis no banco de dados da base AURORA [52]. Estes compreendem: ruído de aeroporto, balbúcio, carro, exposições, restaurante, rua, metrô e trem. Estes ruídos estão disponíveis no formato SHORT com taxa de amostragem de 8 kHz e resolução de 16 bits.

A escolha dos ruídos está calcada no emprego da tecnologia de reconhecimento de fala em diferentes tipos de ambientes, por exemplo, aplicativos de celulares, aparelho utilizado por usuários de diferentes faixas etárias nos mais diversos tipos de meios [3].

As Figuras 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7 e 4.8 descrevem o comportamento dos ruídos no domínio do tempo e da frequência.

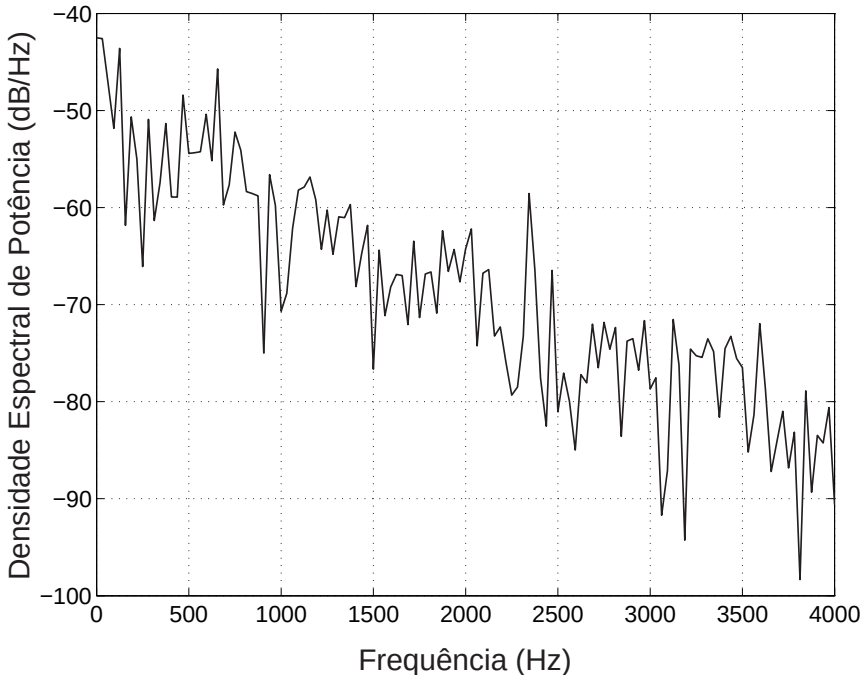

Figura 4.1: *Representação do ruído aeroporto da base AURORA nos domínios do tempo e da frequência*

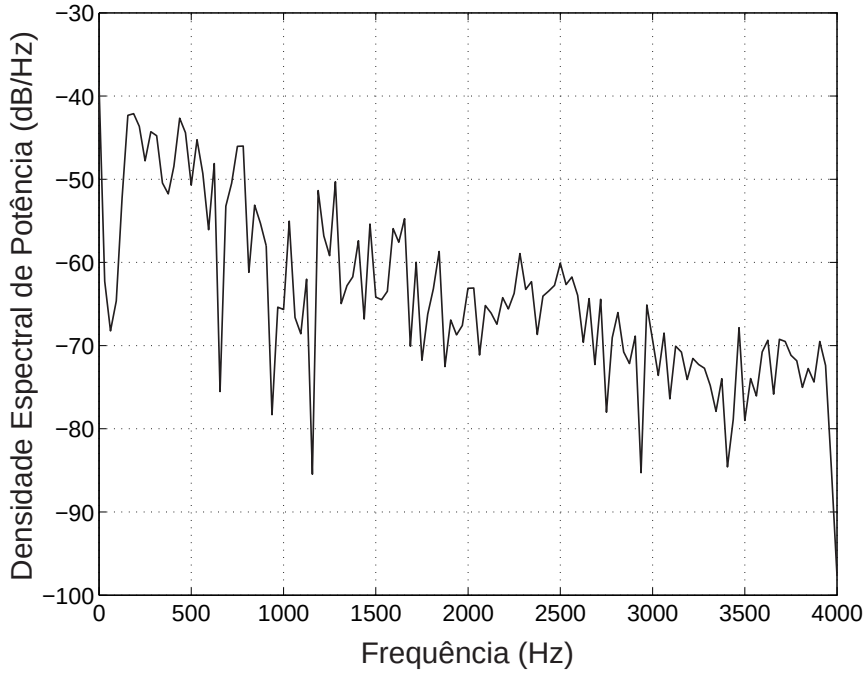

Figura 4.2: *Representação do ruído balbúcio da base AURORA nos domínios do tempo e da frequência*

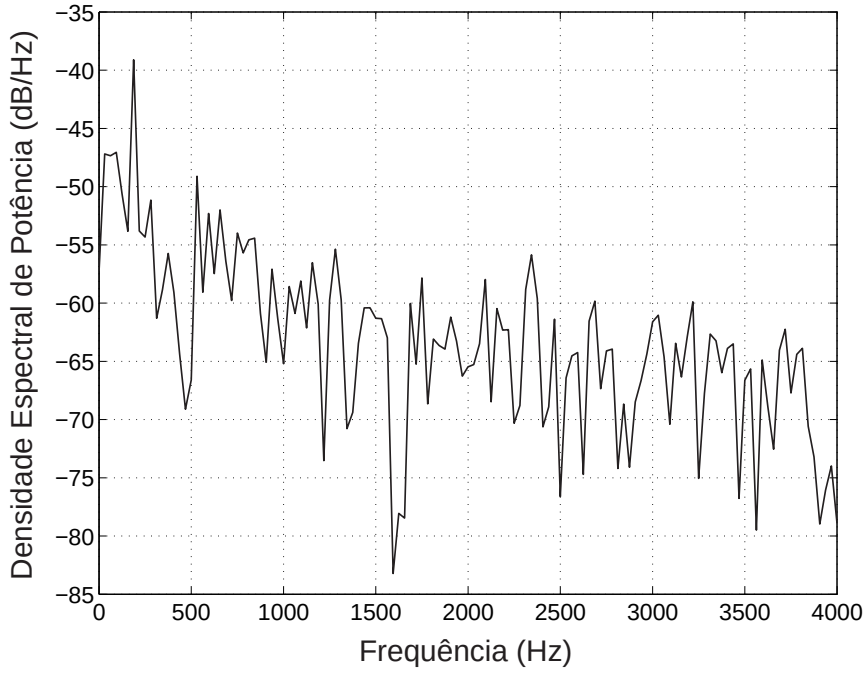

Figura 4.3: *Representação do ruído carro da base AURORA nos domínios do tempo e da frequência*

4.3 Base de dados corrompida artificialmente

Com o objetivo de avaliar o desempenho do sistema ASR em diferentes situações, propôs-se dois tipos de treinamentos:

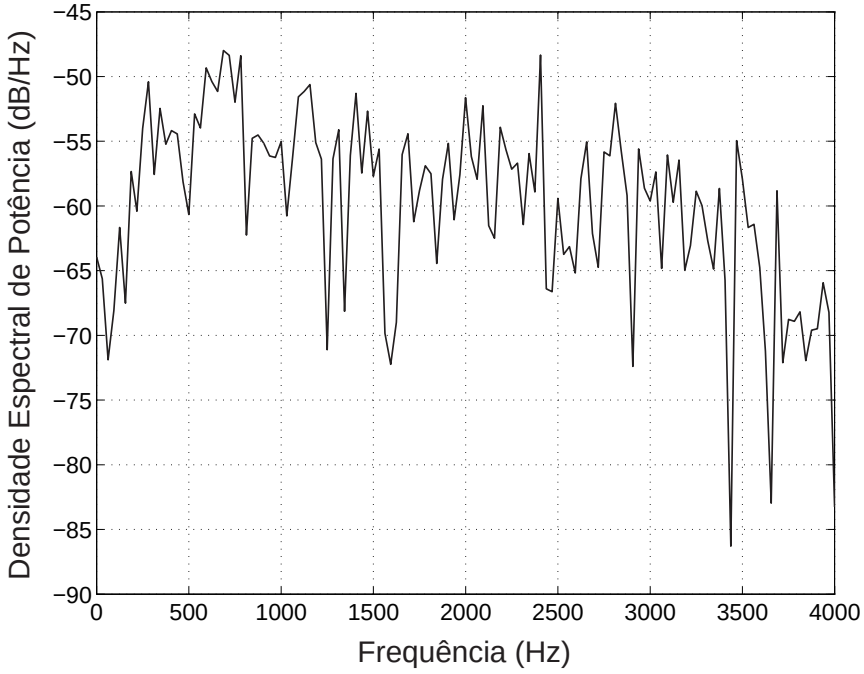

Figura 4.4: *Representação do ruído exposição da base AURORA nos domínios do tempo e da frequência*

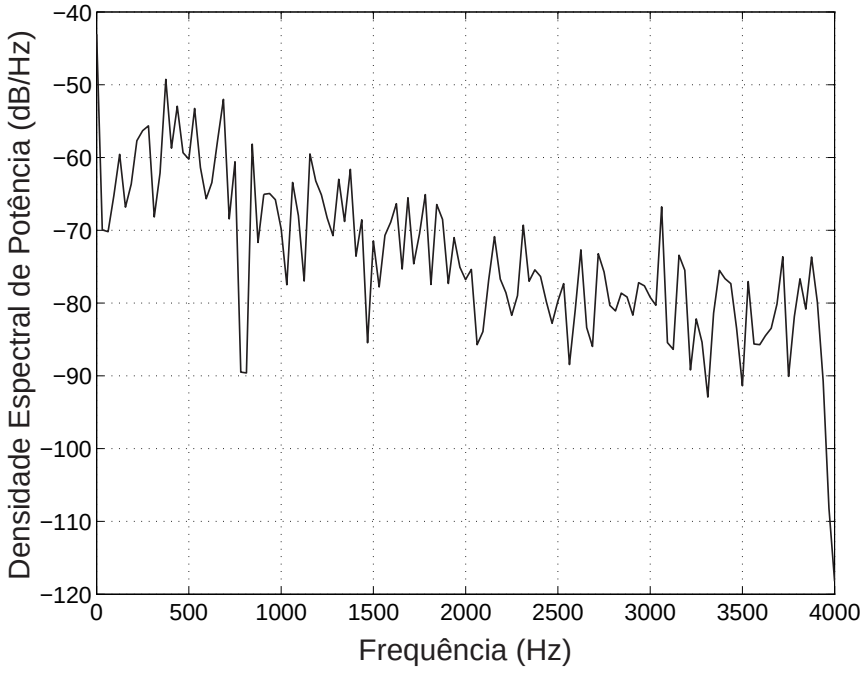

Figura 4.5: *Representação do ruído restaurante da base AURORA nos domínios do tempo e da frequência*

- *treinamento com dados limpos*;
- *treinamento multi-estilo.*

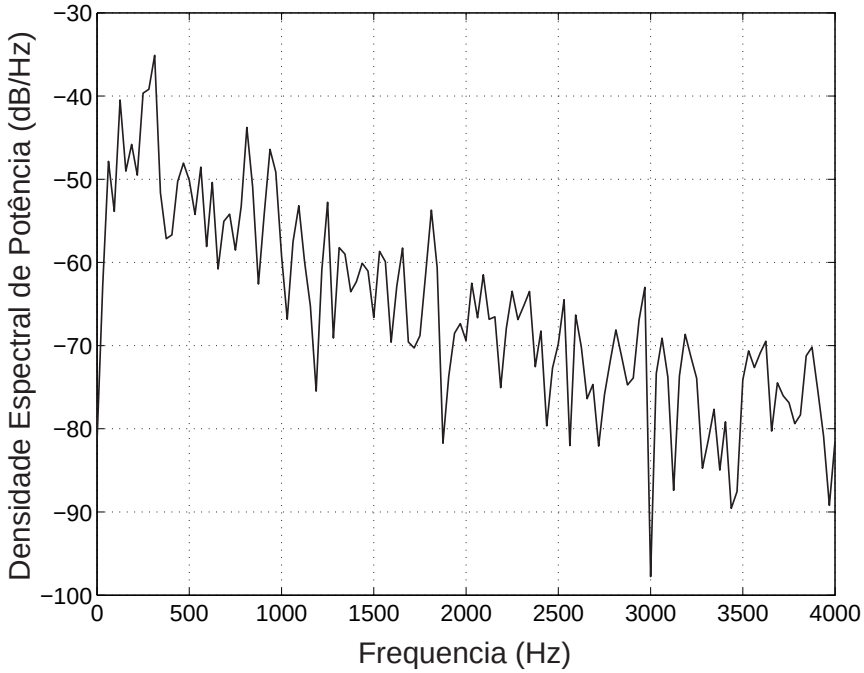

Figura 4.6: *Representação do ruído rua da base AURORA nos domínios do tempo e da frequência*

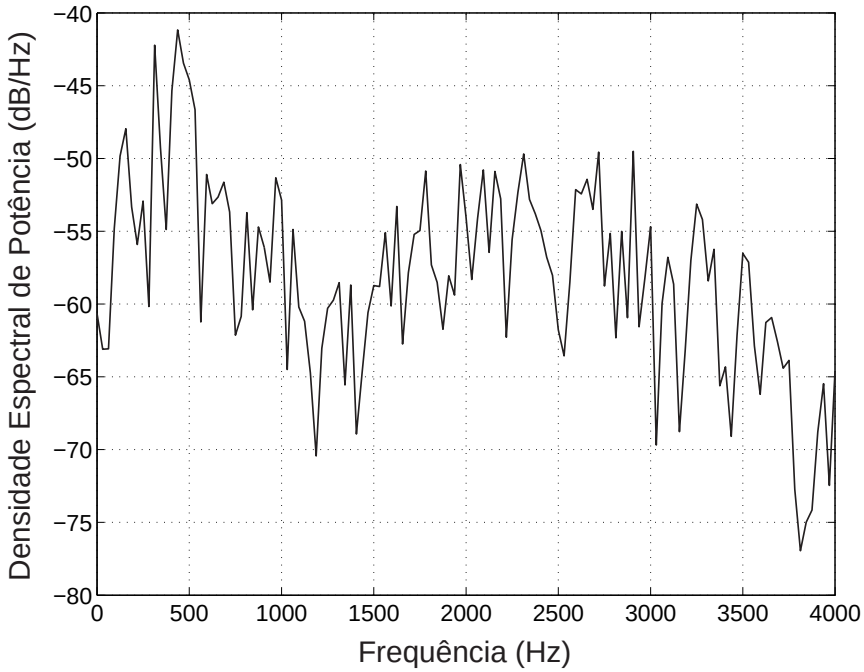

Figura 4.7: *Representação do ruído metrô da base AURORA nos domínios do tempo e da frequência*

Como o propósito é a robustez de sistemas ASRs a ruídos, foram executados testes com locuções corrompidas com diferentes tipos de ruídos e relações

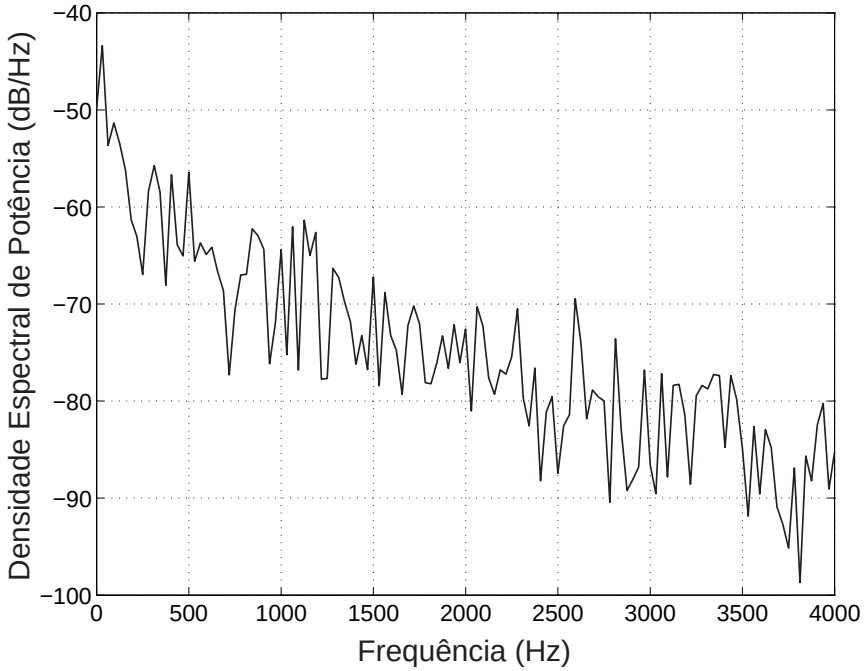

Figura 4.8: *Representação do ruído trem da base AURORA nos domínios do tempo e da frequência*

sinal-ruído. Para atender a requisitos foi necessário produzir uma base de dados corrompida a partir da base de dados limpa.

O primeiro passo foi realizar a reamostragem da base de dados original utilizando a frequência de 8 kHz tornando-a compatível com a frequência de amostragem da base AURORA.

Os ruídos disponíveis na base de dados AURORA foram artificialmente adicionadas à base de dados criada por Ynoguti através da ferramenta MATLAB. Adicionou-se a cada versão original das locuções da base limpa uma porção do ruído de mesmo tamanho. Para determinação da relação sinal-ruído desejada utilizou-se a definição matemática da SNR, razão entre a energia do sinal da fala pela energia do ruído, como descrito pela Equação (4.1):

$$SNR = 10 \times \log_{10} \frac{Energia_{sinaldafala}}{Energia_{ruído}} \tag{4.1}$$

onde SNR representa a relação sinal-ruído em dB, $Energia_{sinaldafala}$ a energia do sinal de fala limpo e $Energia_{ruído}$ é a energia do ruído.

No âmbito das locuções que compõem o banco de dados do treinamento, foi gerado um novo grupo com 2400 locuções corrompidas com SNR de 15 dB e 20 dB com cada um dos oito tipos de ruídos da base AURORA, totalizando 19200

locuções.

Para compor o banco de dados do teste, foram gerados 8 grupos, um para cada tipo de ruído da base AURORA, com 400 locuções corrompidas com SNR de -5 dB, 0 dB, 5 dB, 10 dB, 15 dB e 20 dB, totalizando 2400 locuções por grupo.

A Seção a seguir apresenta particularidades do sistema ASR utilizado nos testes de reconhecimento.

4.4 O sistema utilizado no reconhecimento de fala

Nesta seção é descrito detalhes sobre o sistema de reconhecimento utilizado neste trabalho.

4.4.1 O sistema ASR

O software do sistema ASR baseado em Modelos Ocultos de Markov contínuos, cujos diagramas em blocos dos módulos de treinamento e reconhecimento são representados nas Figuras 4.9 e 4.10, desenvolvido em [50], foi utilizado para execução dos testes experimentais.

Este sistema modela cada uma das subunidades fonéticas por um HMM de 3 estados, conforme arquitetura apresentada na Figura 4.11. E, em cada estado de um HMM foram utilizadas 10 gaussianas. Cada HMM foi inicializado a partir do algoritmo *Segmental K-Means* e treinado posteriormente via algoritmo *Baum Welch*. O sistema ASR emprega fones independentes de contexto e utiliza como algotitmo de busca o "*One Step*" [53]. Empregou-se os parâmetros acústicos mel-cepstrais de ordem 12 com suas respectivas primeira e segunda derivadas (parâmetros delta e delta-delta), portanto constituem vetores característicos de dimensão 36. Empregou-se modelo linguístico do tipo bigrama. A transcrição fonética da base de dados foi realizada para cada uma das locuções utilizando 36 subunidades fonéticas conforme mostrado no Anexo A.

A Seção a seguir apresenta algumas métricas utilizadas para medir o desempenho de sistemas ASRs.

4.5 Métricas de avaliação de desempenho

Existem diversas métricas que permitem avaliar e medir a precisão de sistemas ASR [54][55][56].

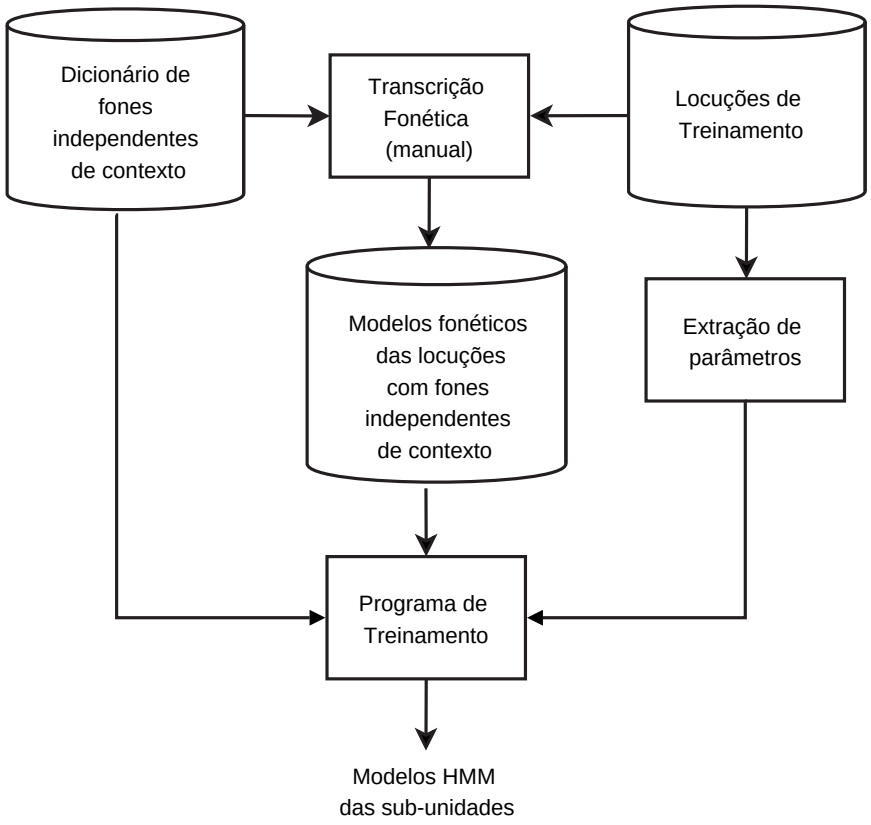

Figura 4.9: *Diagrama em Blocos da Etapa de Treinamento*

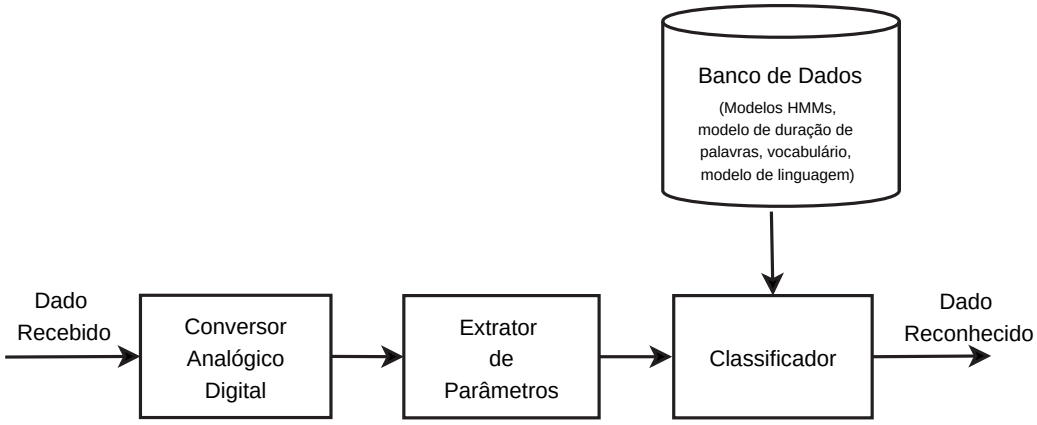

Figura 4.10: *Diagrama em Blocos da Etapa de Reconhecimento*

Para avaliar o desempenho do sistema ASR utiliza-se um arquivo de referência contendo as informações que correspondam as possíveis falas, palavras ou locuções que é posteriormente comparado ao dado reconhecido. Alguns softwares de medição de desempenho de sistemas permitem que esta comparação seja

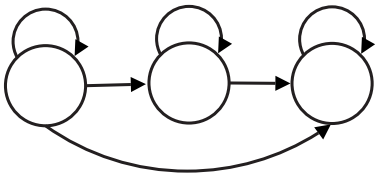

Figura 4.11: *Modelo HMM utilizado para cada subunidade fonética*

realizada tanto através do alinhamento temporal como do alinhamento textual. O principal desafio na avaliação de desempenho é a possibilidade da sentença reconhecida ter tamanho diferente da sentença de referência.

A métrica comumente empregada em sistemas ASRs é a taxa de erro de palavra (WER - *Word Error Rate*) [54][55][57][58]. Ela permite quantificar e identificar os tipos de erros cometidos pelo classificador.

Desta forma, WER pode ser definida como:

$$WER = \frac{S + D + I}{N} \tag{4.2}$$

onde WER taxa de erro de palavra, S é o número de substituições, D é o número de deleções, I é o número de inserções e N é o número total de palavras na sentença de referência.

Para avaliar o desempenho do sistema ASR utilizado neste trabalho, optou-se por utilizar a precisão de palavras que é definida como:

$$WA = 1 - WER \tag{4.3}$$

onde WER taxa de erro de palavra.

Como N é o número total de palavras na sentença de referência, é possível obter-se WER maior que 1 caso hajam mais inserções de palavras do que palavras corretas na sentença hipotética. Portanto, é possível obter valores negativos para descrever a WA.

Neste trabalho foi empregado o SCLITE [62], para cálculo do WER.

4.5.1 SCLITE

O SCLITE é uma ferramenta do pacote SCTK desenvolvido pelo NIST que possibilita contabilizar e avaliar os dados hipotéticos gerados pelos classificadores dos sistemas de reconhecimento de fala [62].

O princípio básico de funcionamento desta ferramenta baseia-se na comparação entre a saída hipotética do classificador e a referência e permite coletar várias informações estatísticas e realizar uma variedade de relatórios com informações relativas ao desempenho de sistemas ASRs.

Para este trabalho foi utilizado o seguinte comando para coleta das estatísticas e geração do relatório com detalhes da precisão do sistema:

sclite $-i$ *wsj* $-r$ *arquivo_ref [fmt]* $-h$ *arquivo_hip [fmt]* *arquivo_saida*

onde a opção $-i$ define como serão interpretados as identificações das locuções dos arquivos de entrada, wsj define que cada linha dos arquivos de entrada contém uma locução e que cada uma delas é uma sequência de palavras seguida pela sua *identificação* entre parênteses, $-r$ *arquivo_ref [fmt]* e $-h$ *arquivo_hip [fmt]* são argumentos de entrada obrigatórios e, indicam, respectivamente, o arquivo de referência e o arquivo que contém as informações hipotéticas geradas durante o processo de reconhecimento, *[fmt]* especifica o formato destes arquivos e *arquivo_saída* corresponde ao relatório que conterá todas as informações de desempenho do sistema.

O arquivo de saída do Sclite acima mostra detalhes da taxa reconhecimento de cada uma das 400 locuções de teste. As colunas $SPKR$, Snt, $\#Wrd$, $Corr$, Sub, Del, Ins, Err e $S.Err$ descrevem respectivamente: número correspondente à sentença em análise, número de frases na sentença, número de palavras, porcentagem de palavras corretas, taxa de substituições, taxa de palavras deletadas, taxa de inserção de novas palavras, taxa de palavras erradas e taxa de sentenças erradas. A linha Sum/Avg mostra a compilação do resultado para todas as locuções da base de dados de teste.

sclite: 2.3 TK Version 1.3
Begin alignment of Ref File:'base_reco.txt' and Hyp File:
'/home/tatiane/Documentos/Software_Reco/hmmreco/resultado/
resultado_10g_adaptado/reco_corrupted/
resultado_gmm10g8k_Tcorrupted_Aairport_alpha015_Rcorrupted_airport_snr10.txt'

Alignment# 1 for speaker 0)

Alignment# 1 for speaker 1)

.
.
.

Alignment# 1 for speaker 398)

Alignment# 1 for speaker 399)

SYSTEM SUMMARY PERCENTAGES by SPEAKER

/home/tatiane/Documentos/Software_Reco/hmmreco/resultado/resultado_10g_
adaptado/reco_corrupted/_gmm10g8k_Tcorrupted_Aairport_alpha015_
Rcorrupted_airport_snr10.txt

SPKR	# Snt	# Wrd	Corr	Sub	Del	Ins	Err	S.Err
0)	1	4	75.0	25.0	0.0	25.0	50.0	100.0
1)	1	6	100.0	0.0	0.0	0.0	0.0	0.0
2)	1	7	71.4	14.3	14.3	14.3	42.9	100.0
3)	1	6	100.0	0.0	0.0	0.0	0.0	0.0
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
397)	1	7	28.6	57.1	14.3	14.3	85.7	100.0
398)	1	6	100.0	0.0	0.0	0.0	0.0	0.0
399)	1	8	50.0	37.5	12.5	12.5	62.5	100.0
Sum/Avg	400	2620	75.5	21.1	3.4	9.4	33.8	75.5
Mean	1.0	6.5	75.6	21.3	3.0	9.9	34.3	75.5
S.D.	0.0	1.3	24.6	22.7	7.2	14.8	31.5	43.1
Median	1.0	7.0	83.3	15.5	0.0	0.0	28.6	100.0

Successful Completion

Capítulo 5

Resultados Experimentais

Para avaliação do comportamento isolado das técnicas de robustez ao ruído apresentadas no Capítulo 3, bem como, a combinação das mesmas, foram realizadas as seguintes simulações que serão detalhadas nas subseções seguintes:

1. na primeira etapa foi realizado o treinamento com locuções limpas seguido do reconhecimento de locuções limpas, para verificar o ganho máximo possível numa situação ideal, ou seja, avaliar o desempenho do sistema ASR na ausência de ruído;

2. na segunda etapa, o HMM treinado com locuções limpas foi utilizado para os testes de reconhecimento de locuções ruidosas com o intuito de verificar a redução da WA quando o sistema ASR está operando em condições adversas;

3. na terceira etapa, o HMM treinado com locuções corrompidas por ruído com SNR 15 dB e 20 dB foi utilizado nos testes de reconhecimento de locuções limpas com a finalidade de identificar a influência do treinamento multi-estilo no desempenho do sistema ASR em ambiente livre de ruído. A escolha destas SNRs foi baseada nos resultados experimentais apresentados em [48];

4. na quarta etapa, o HMM treinado com locuções corrompidas por ruído com SNR 15 dB e 20 dB foi utilizado nos testes de reconhecimento de locuções ruidosas, com a finalidade de identificar a influência do treinamento multi-estilo na resposta do sistema ASR;

5. na quinta etapa, o HMM treinado com locuções limpas e adaptado para um determinado tipo e nível de ruído é empregado nos testes com locuções livres de ruído, a fim de medir o ganho proporcionado pelo MAP;

6. na sexta etapa, o HMM treinado com locuções limpas e adaptado para um determinado tipo e nível de ruído é empregado nos testes com locuções corrompidas, a fim de medir o ganho proporcionado pelo MAP em condições reais;

7. na sétima etapa, o HMM treinado com locuções corrompidas por ruído com SNR 15 dB e 20 dB adaptado para um determinado tipo e nível de ruído foi utilizado nos testes com locuções limpas, para avaliar o desempenho do sistema proporcionado pela integração das técnicas propostas na ausência de ruído.

8. e, por fim, o HMM treinado com locuções corrompidas por ruído com SNR 15 dB e 20 dB adaptado para um determinado tipo e nível de ruído foi utilizado nos testes com locuções corrompidas, cujo principal alvo, foi determinar o ganho proporcionado pela combinação das técnicas propostas.

5.1 Treinamento e reconhecimento utilizando dados limpos

Esta seção apresenta os resultados experimentais para condição ideal de operação do sistema ASR, ou seja, na ausência de ruído. Para medir seu desempenho nestas condições, o sistema foi treinado a partir dos dados limpos e, posteriormente foram realizados testes de reconhecimento utilizando a base de teste com locuções limpas. A WA obtida foi de 75,6 %.

5.2 Treinamento com dados limpos e reconhecimento utilizando locuções corrompidas

A Tabela 5.1 apresenta os resultados para o sistema ASR treinado com locuções limpas, porém testado com locuções corrompidas artificialmente por diferentes tipos e níveis de ruído.

A partir da análise dos resultados apresentados é possível verificar que houve uma queda significativa de desempenho do sistema ASR comprovando a sensibilidade deste na presença de ruído.

Verifica-se que para alguns tipos de ruído com SNR 0 dB, obteve-se um valor de WA negativo. Isto ocorreu devido ao elevado número de erros por inserção.

Tabela 5.1: *WA, em %, para um sistema treinado com locuções limpas e testado com locuções corrompidas*

Tipo de Ruído	SNR −5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
Aeroporto	1,0	−3,8	15,6	44,7	66,3	66,3
Balbúcio	0,5	−2,1	21,6	54,9	69,6	69,6
Carro	0,3	4,8	23,3	56,4	70,5	70,5
Exposições	0,6	−3,2	8,3	45,3	68,5	68,5
Restaurante	0,7	4,7	0,7	20,3	52,3	52,3
Rua	2,9	9,3	42,2	66,4	74,6	74,6
Metrô	1,1	−1,5	11,9	43,9	64,6	64,6
Trem	3,8	6,2	26,8	60,5	73,7	73,7
Média	1,4	1,8	10,8	42,2	31,5	31,5

5.3 Treinamento Multi-Estilo e reconhecimento utilizando locuções limpas

Nesta etapa, o sistema foi treinado com base corrompida pelos modelos acústicos dos 8 tipos de ruído disponíveis na base AURORA com SNR de 15 dB e 20 dB conforme descrito na Seção 4.3. Posteriormente, foram realizados testes de reconhecimento para base de dados limpa e a WA obtida foi de 74,3 %. Houve uma queda de apenas 1,3 % comparado ao sistema treinado e testado com locuções limpas. Este resultado indica que o descasamento acústico entre as condições de teste e treino interfere no desempenho do sistema.

5.4 Treinamento Multi-Estilo e reconhecimento utilizando dados corrompidos

A fim de minimizar a influência dos distúrbios ambientais, empregou-se a técnica de multi-estilo utilizando locuções corrompidas com SNR 15 dB e 20 dB seguida de testes de reconhecimento de locuções corrompidas.

A Tabela 5.2 apresenta os resultados experimentais obtidos para diferentes tipo e níveis de ruído. Verifica-se que esta abordagem proporcionou um ganho médio de aproximadamente 6,83 % na resposta do sistema comparado ao desempenho do sistema treinado com locuções limpas e testado com dados corrompidos.

Os resultados indicam que para alguns tipos e níveis de ruído houve um ganho no desempenho do sistema comparado à resposta do sistema treinado e testado com locuções limpas, pois o sistema contém informações do ruído minimizando o

Tabela 5.2: *WA, em %, para um sistema treinado e testado com locuções corrompidas*

Tipo de Ruído	SNR −5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
Aeroporto	1,2	3,2	25,4	62,9	78,0	78,0
Balbúcio	2,4	4,6	32,2	66,3	77,1	77,1
Carro	0,3	5,3	31,7	65,8	76,5	76,5
Exposição	1,2	0,2	16,4	58,2	75,5	75,5
Restaurante	0,5	13,8	4,4	35,5	69,2	69,2
Rua	3,4	15,6	56,1	73,6	75,8	75,8
Metrô	0,8	−0,3	18,4	57,9	73,0	73,0
Trem	2,3	7,6	34,0	69,0	78,1	78,1
Média	1,5	6,3	27,3	61,2	65,8	65,8

descasamento acústico entre o treinamento e teste.

5.5 ASR treinado com dados limpos, adaptado com ruído e testado com locuções limpas

Uma outra proposta para agregar robustez ao sistema foi a técnica MAP. Nesta etapa, o modelo obtido a partir do treinamento com dados limpos foi adaptado para cada um dos 8 tipos de ruído (aeroporto, balbúcio, carro, exposição, restaurante, rua, metrô e trem) e, posteriomente, foram realizados testes de reconhecimento dos dados limpos.

Os testes foram realizados com fator de adaptacão α no intervalo de 0,01 a 0,09 com variação de 0,01 e, entre 0,1 e 0,95 variando em 0,05. O desempenho do sistema ASR é apresentado na Tabela 5.3.

Analisando os resultados, verifica-se que certos valores de α proporcionam aumento no desempenho do sistema ASR comparado à resposta do sistema treinado e testado com locuções limpas, enquanto outros valores provocam uma degradação no desempenho. Vale ressaltar que pequenos valores do coeficiente de adaptação enfatizam o sinal original enquanto valores maiores dão mais ênfase ao ruído. Assim, conclui-se que o procedimento de adaptação depende fortemente de uma escolha adequada para o parâmetro α.

Tabela 5.3: *WA baseado, em %, um sistema adaptado para determinado tipo de ruído, treinado e testado com locuções limpas*

α	Aeroporto	Balbúcio	Carro	Exposição	Restaurante	Rua	Metrô	Trem
0,01	75,4	**75,2**	75,5	75,4	75,3	75,5	75,4	**75,6**
0,02	75,6	75,1	75,4	**75,8**	**75,6**	75,1	75,3	75,4
0,03	75,5	75,0	75,1	75,3	75,1	**75,8**	**76,0**	75,5
0,04	**76,3**	75,1	75,4	74,9	75,2	**75,8**	75,8	75,3
0,05	75,8	**75,2**	**75,9**	74,1	75,3	75,2	75,7	75,3
0,06	75,4	**75,2**	75,6	74,1	74,8	75,3	75,0	75,0
0,07	74,8	74,8	74,9	73,9	75,0	74,8	75,2	74,8
0,08	74,8	74,1	75,1	73,7	75,0	75,0	74,9	74,9
0,09	74,4	74,2	75,3	73,7	75,0	75,0	74,8	74,3
0,10	74,4	74,0	75,0	72,9	74,8	74,8	74,0	74,7
0,15	74,4	73,2	74,2	72,1	73,9	74,1	73,2	73,9
0,20	74,3	73,1	73,3	72,0	74,0	73,9	72,5	73,6
0,25	74,2	73,2	72,0	70,3	73,3	73,1	71,3	73,5
0,30	73,8	72,5	72,9	70,1	72,6	71,8	70,8	73,2
0,35	72,9	72,6	71,5	68,9	72,1	71,1	69,5	71,2
0,40	71,4	72,2	69,4	67,7	72,2	70,8	67,3	71,7
0,45	70,5	70,3	67,7	65,4	71,0	68,5	64,4	68,4
0,50	70,2	69,1	66,0	62,5	68,9	67,2	63,6	66,7
0,55	69,1	66,5	62,1	57,1	66,8	64,8	60,6	65,5
0,60	65,8	62,0	58,7	50,9	65,5	62,8	56,7	61,3
0,65	62,8	58,3	54,0	45,4	62,6	59,5	50,0	58,3
0,70	58,6	53,6	45,5	35,5	58,7	54,8	43,2	52,7
0,75	52,0	46,0	35,7	21,6	52,6	49,9	33,8	43,8
0,80	45,5	38,2	25,2	10,1	42,9	42,4	19,0	36,8
0,85	33,5	30,0	9,9	−0,3	31,1	31,6	7,1	25,5
0,90	18,7	16,8	−4,1	−8,5	17,4	18,8	−5,6	10,6
0,95	−1,5	−0,2	−9,0	−9,9	3,9	3,9	−12,2	−7,6

5.6 ASR treinado com dados limpos, adaptado com ruído e testado com locuções corrompidas

A fim de avaliar-se o desempenho do sistema ASR adaptado operando em condições ambientais adversas, o modelo canônico resultante do treinamento com dados limpos foi adaptado para cada um dos 8 tipos de ruídos (aeroporto, balbúcio, carro, exposição, restaurante, rua, metrô e trem) e, então, o novo modelo gerado foi utilizado nos testes de reconhecimento de locuções corrompidas artificialmente.

Os testes foram realizados com fator de adaptacão no intervalo de 0,01 a 0,09 com variação de 0,01 e, entre 0,1 e 0,95 variando em 0,05. As WA do sistema para diferentes tipos e níveis de ruído são mostradas nas Figuras 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7 e 5.8.

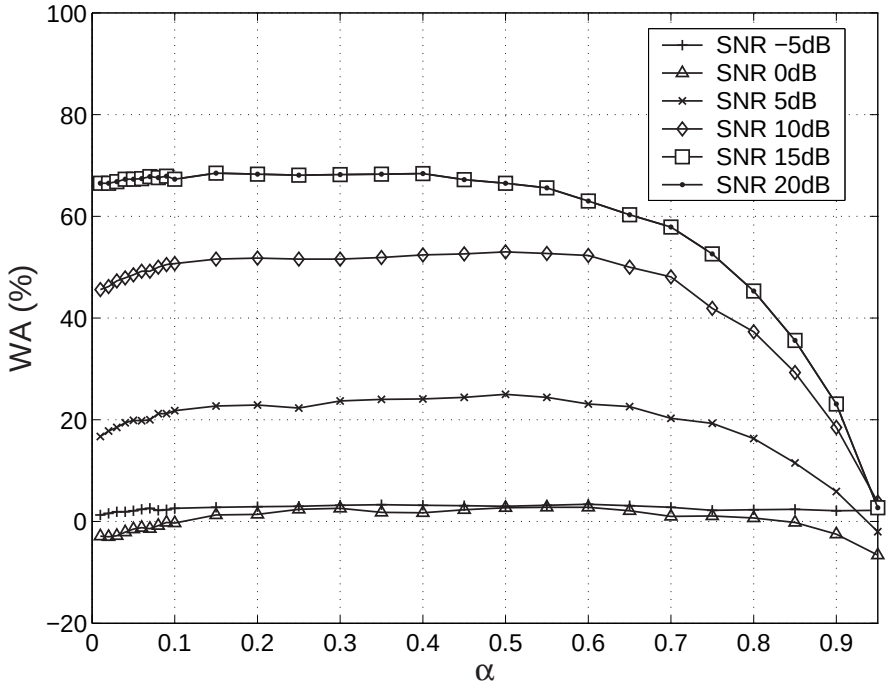

Figura 5.1: *WA para o sistema adaptado com ruído de aeroporto e treinamento com dados limpos*

Em geral, os resultados demonstram que quanto menor a SNR maior é a degradação do desempenho do sistema devido ao descasamento acústico entre as condições de treinamento e teste. Excepcionalmente, para os ruídos aeroporto, balbúcio, exposição e subway, obteve-se um desempenho superior para SNR -5 dB comparado à SNR 0 dB. Do ponto de vista estatístico, o nível de -5 dB é agressivo para o sistema fornecendo resultados não confiáveis.

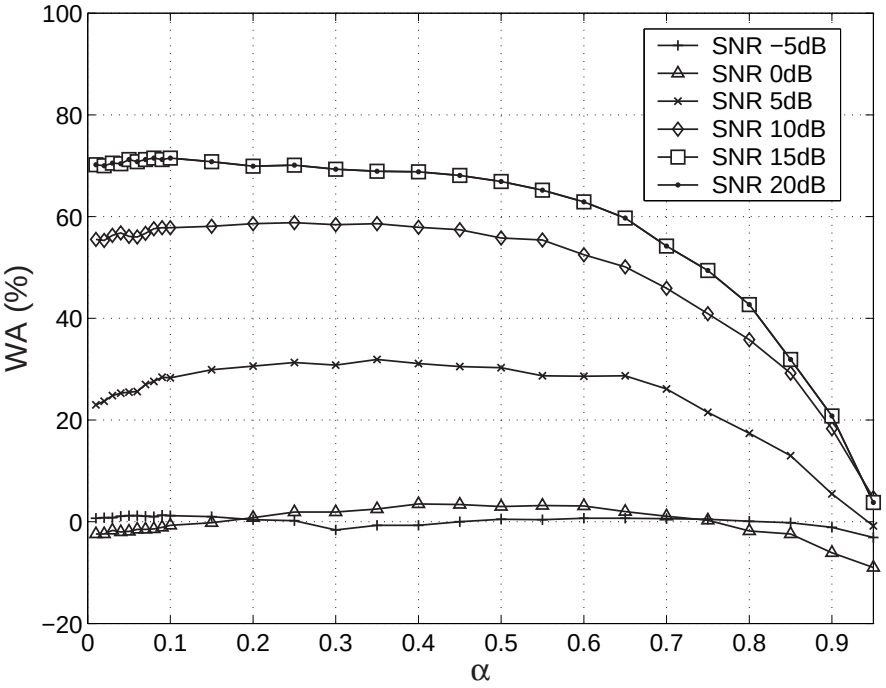

Figura 5.2: *WA para o sistema adaptado com ruído de balbúcio e treinamento com dados limpos*

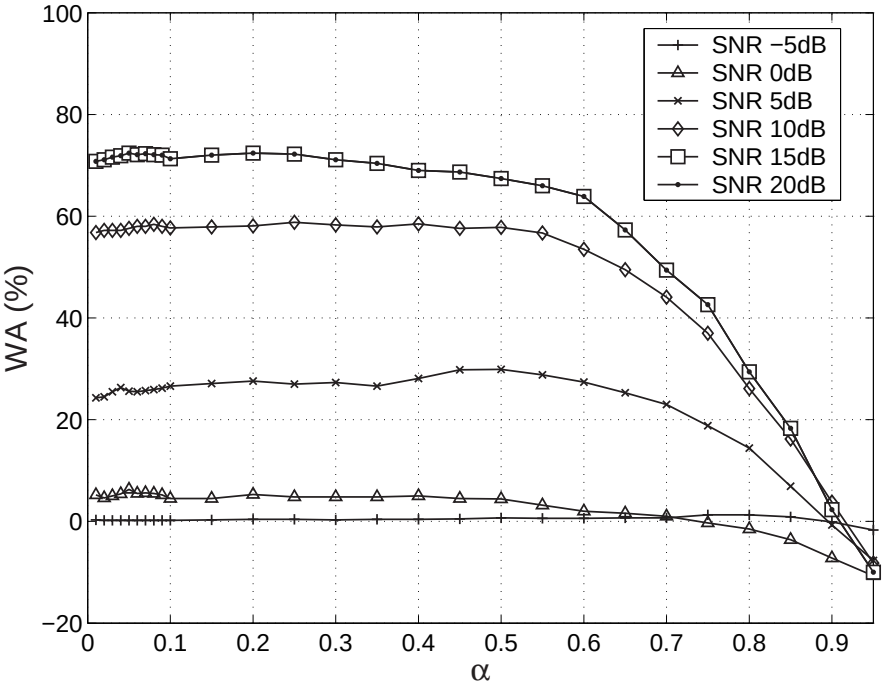

Figura 5.3: *WA para o sistema adaptado com ruído de carro e treinamento com dados limpos*

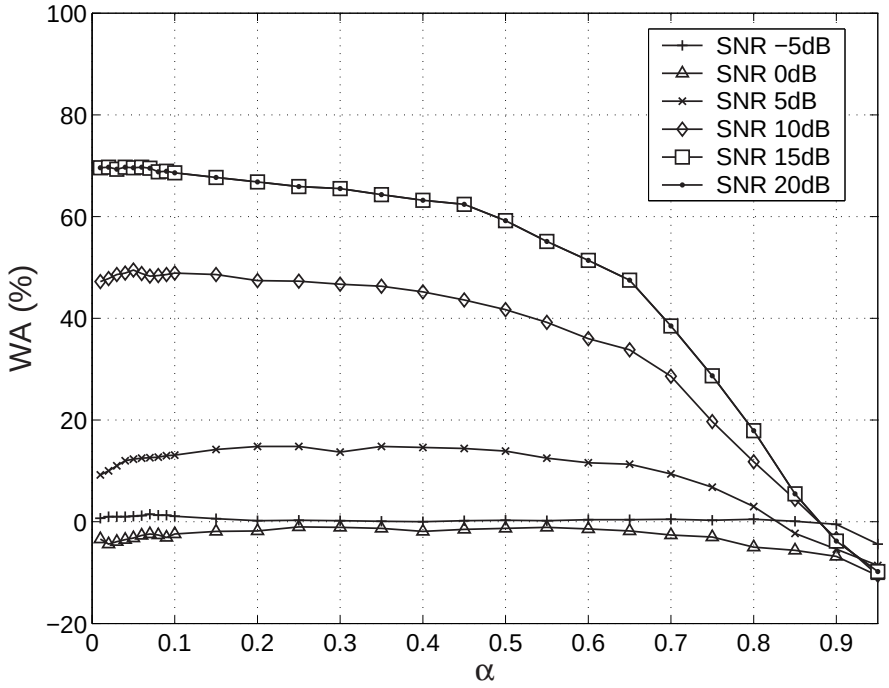

Figura 5.4: *WA para o sistema adaptado com ruído de exposição e treinamento com dados limpos*

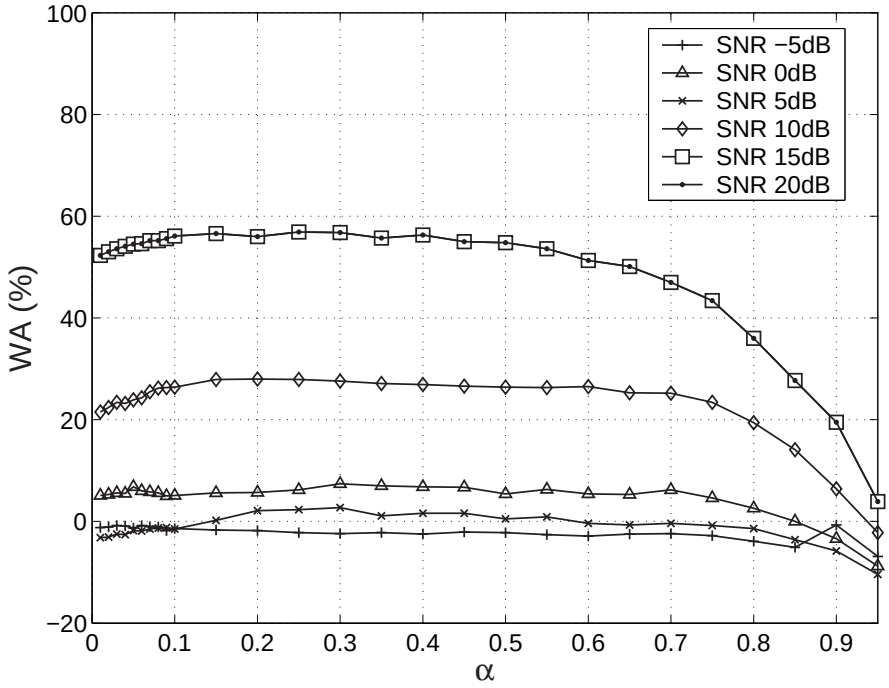

Figura 5.5: *WA para o sistema adaptado com ruído de restaurante e treinamento com dados limpos*

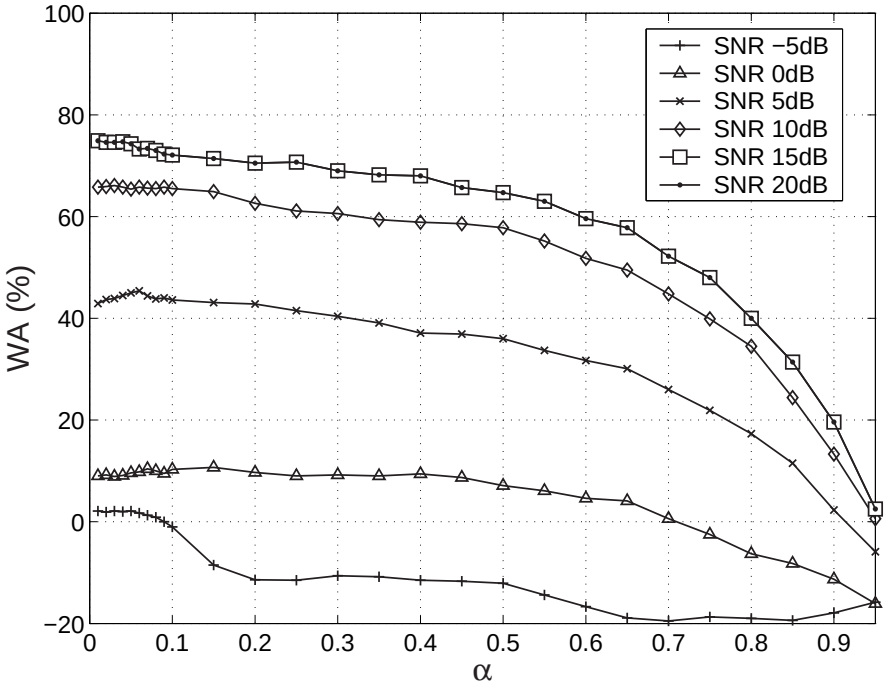

Figura 5.6: *WA para o sistema adaptado com ruído de rua e treinamento com dados limpos*

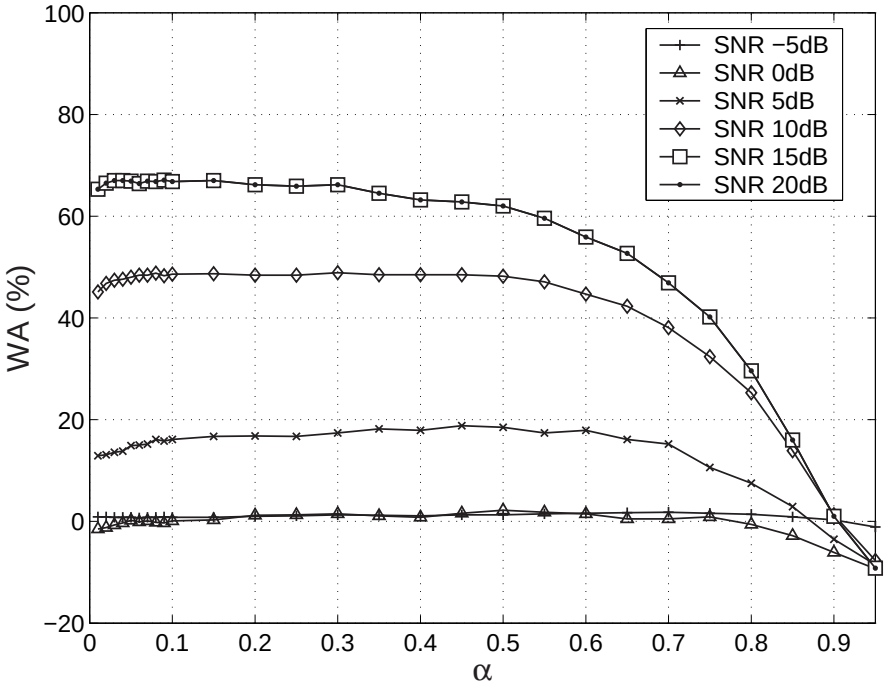

Figura 5.7: *WA para o sistema adaptado com ruído de metrô e treinamento com dados limpos*

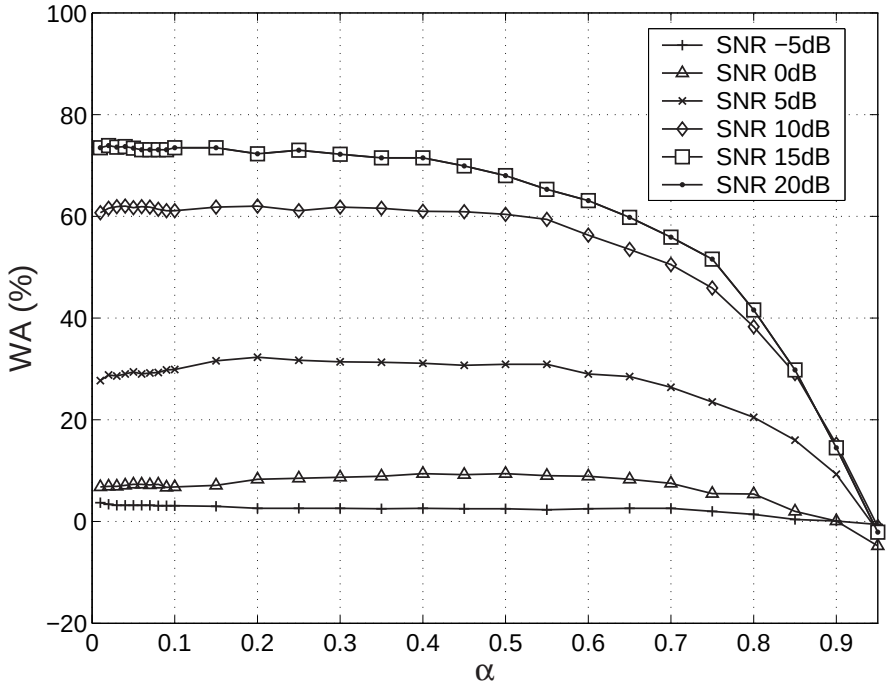

Figura 5.8: *WA para o sistema adaptado com ruído de trem e treinamento com dados limpos*

Para alguns tipos e níveis de ruído, obteve-se um valor de WA negativo. Isto ocorreu devido ao elevado número de erros por inserção comparado ao número total de palavras corretas.

Verifica-se ainda, que a adaptação com determinados valores de coeficiente de adaptação proporcionou um desempenho melhor nas mesmas condições de treino e teste.

5.7 ASR treinado com Multi-estilo, adaptado com ruído e testado com locuções limpas

Tendo como foco avaliar as vantagens propiciadas pela combinação das técnicas multi-estilo e MAP, o modelo canônico resultante do treinamento com locuções ruidosas foi adaptado com estimativas estatísticas de cada um dos tipos de ruído disponíveis na base AURORA, formando 8 novos modelos. A partir destes, foram realizados testes de reconhecimento de dados limpos cujos resultados foram compilados e estão apresentados na Tabela 5.4. Os testes foram realizados com coeficiente de adaptacão no intervalo de 0,01 a 0,09 com variação de 0,01 e, entre 0,1 e 0,95 variando em 0,05.

Verifica-se que os valores menores de α proporcionaram uma WA maior, con-

ferindo que o sinal original foi enfatizado. Em poucos casos, houve queda de desempenho comparado ao sistema treinado através da técnica multi-estilo e testado com dados limpos e, em outros o MAP proporcionou um ganho.

Tabela 5.4: *WA, em %, para um sistema adaptado, treinado com locuções ruidosas e testado com locuções limpas*

α	Aeroporto	Balbúcio	Carro	Exposição	Restaurante	Rua	Metrô	Trem
0,01	74,1	74,1	74,0	**73,8**	**74,1**	74,0	**74,0**	74,0
0,02	73,8	74,2	**74,1**	73,1	74,0	74,0	73,2	73,8
0,03	**74,4**	**74,3**	73,7	72,5	74,0	**74,5**	72,9	**74,2**
0,04	74,2	**74,3**	73,8	71,7	73,5	73,6	71,9	73,6
0,05	73,9	73,7	73,6	71,6	73,7	73,5	71,5	73,6
0,06	73,8	73,8	73,6	71,2	73,8	73,2	71,8	73,6
0,07	73,8	73,9	73,2	70,7	73,1	72,9	71,7	72,9
0,08	73,7	73,7	72,8	71,1	72,9	72,3	71,5	73,1
0,09	73,4	72,9	72,6	70,6	72,7	72,1	71,1	72,9
0,10	73,5	72,5	72,0	70,4	71,9	71,7	70,6	72,7
0,15	72,3	71,4	70,5	69,8	71,4	70,4	70,0	71,0
0,20	70,8	70,1	68,5	68,5	70,3	69,4	69,4	70,2
0,25	70,0	69,1	67,6	66,9	69,5	68,0	68,5	68,5
0,30	69,4	67,2	66,6	65,7	69,0	67,3	66,7	66,9
0,35	67,9	66,8	65,0	64,0	68,2	66,1	65,8	65,6
0,40	66,3	64,7	63,1	62,6	67,0	65,7	65,3	64,6
0,45	64,2	64,5	61,1	59,2	63,9	64,9	64,2	62,8
0,50	62,4	62,1	57,4	56,5	62,4	61,9	60,6	60,6
0,55	58,8	59,7	53,5	52,3	61,1	59,2	56,4	57,4
0,60	55,4	56,8	49,1	45,1	57,1	56,9	49,7	52,7
0,65	51,8	52,4	42,8	38,3	54,3	53,0	43,6	47,9
0,70	47,1	46,9	35,7	31,4	50,0	48,2	35,5	42,2
0,75	41,9	40,4	26,1	20,4	43,5	41,8	26,8	36,8
0,80	32,1	31,1	14,4	9,9	35,7	35,2	15,2	27,1
0,85	20,8	19,9	4,5	−2,9	26,5	25,0	4,5	13,9
0,90	5,2	3.8	−4.5	−8,9	9,1	9,7	−6,8	−1,1
0,95	−7,6	−8.7	−8.6	−9,8	−5,7	−4,5	−7,3	−10,5

5.8 ASR treinado com locuções corrompidas, adaptado com ruído e testado com locuções ruidosas

Esta Seção consiste da última etapa dos testes realizados na qual adaptou-se o modelo canônico resultante do treinamento com dados corrompidos artificialmente dando origem a 8 novos modelos. É importante ressaltar que o mesmo tipo de ruído foi utilizado nas fases de treinamento e adaptação. Portanto, o modelo resultante do treinamento cujos dados utilizados foram corrompidos por ruído de aeroporto foi adaptado para ruído de aeroporto e, assim, sussessivamente. Posteriormente, foram executados testes de reconhecimento de locuções ruidosas cujos resultados estão demonstrados nas Figuras 5.9, 5.10, 5.11, 5.12, 5.13, 5.14, 5.15 e 5.16. Os testes foram realizados com coeficiente de adaptacão no intervalo de 0,01 a 0,09 com variação de 0,01 e, entre 0,1 e 0,95 variando em 0,05.

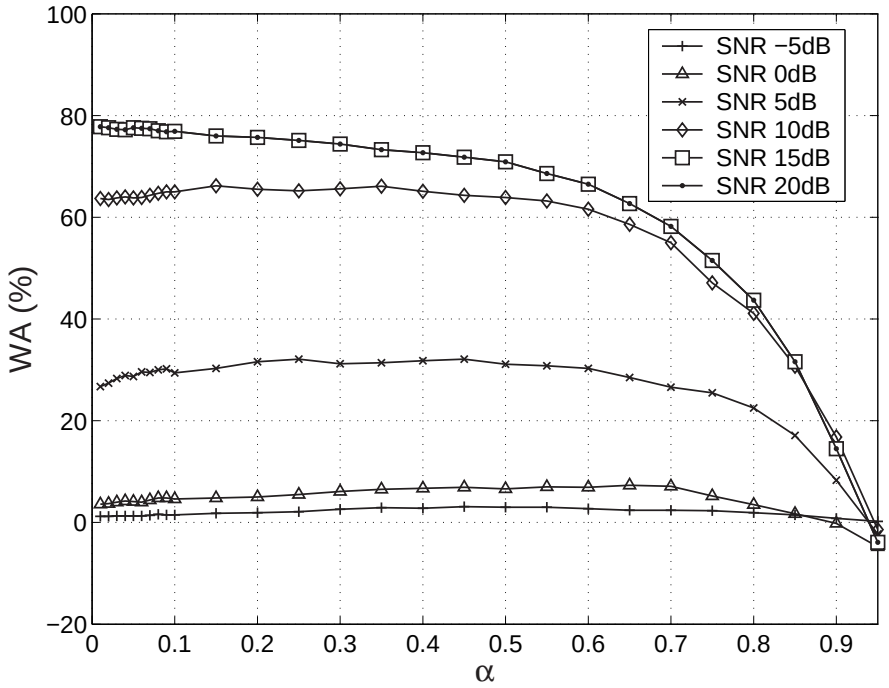

Figura 5.9: *WA para o sistema adaptado com ruído de aeroporto e treinamento multi-estilo*

A partir dos resultados, verifica-se que o emprego conjunto das técnicas proporcionou robustez ao sistema ASR aumentando a precisão de reconhecimento do sistema. Esta proposta proporcionou uma redução de aproximadamente 2,1% na taxa de erro de palavras.

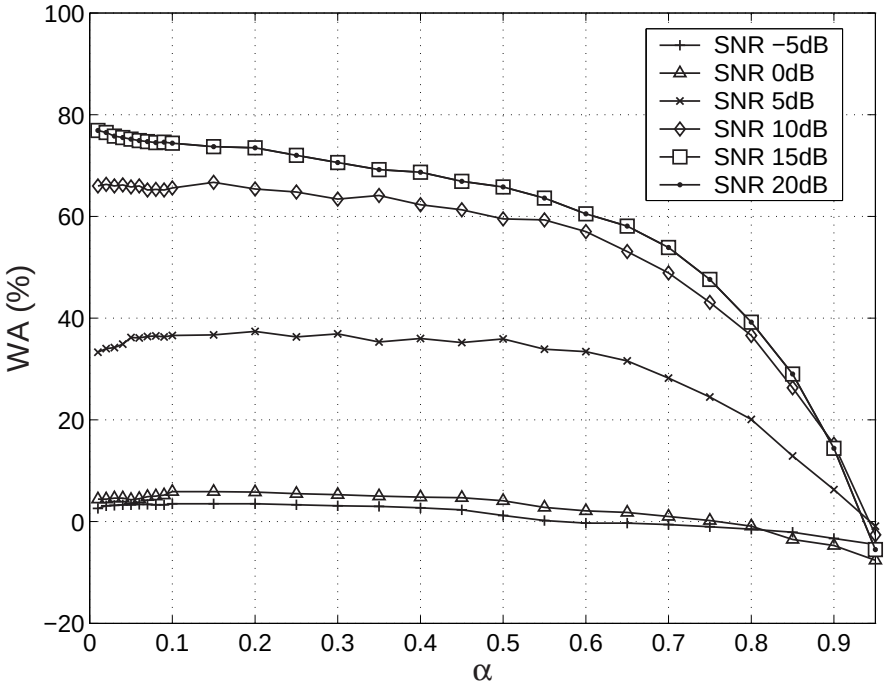

Figura 5.10: *WA para o sistema adaptado com ruído de balbúcio e treinamento multi-estilo*

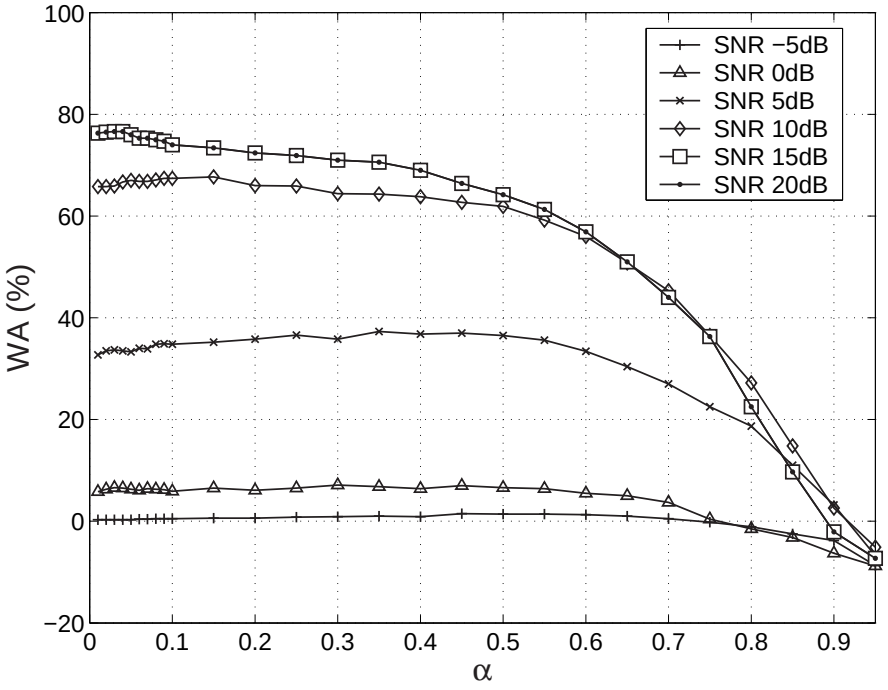

Figura 5.11: *WA para o sistema adaptado com ruído de carro e treinamento multi-estilo*

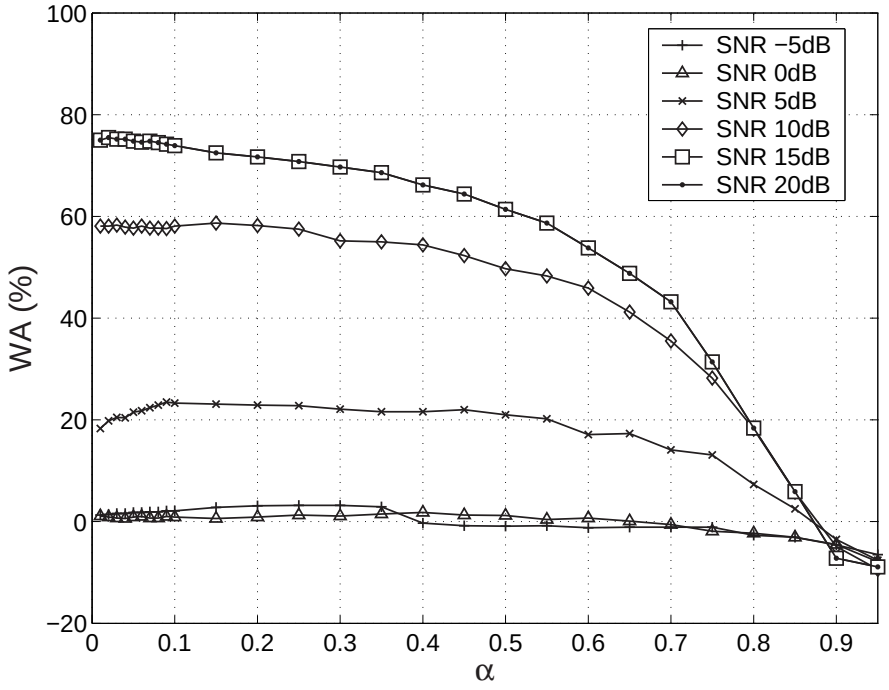

Figura 5.12: *WA para o sistema adaptado com ruído de exposição e treinamento multi-estilo*

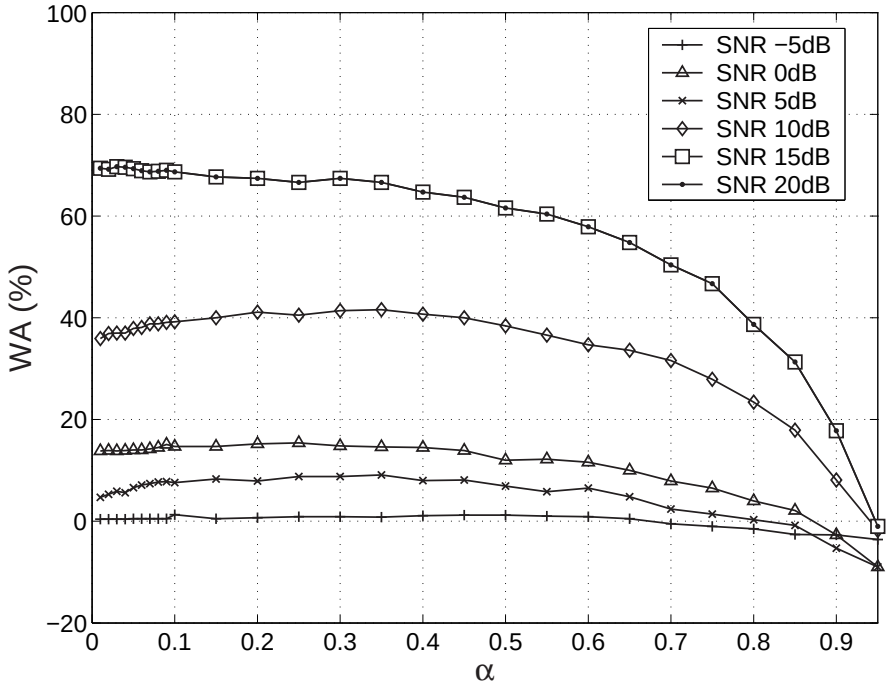

Figura 5.13: *WA para o sistema adaptado com ruído de restaurante e treinamento multi-estilo*

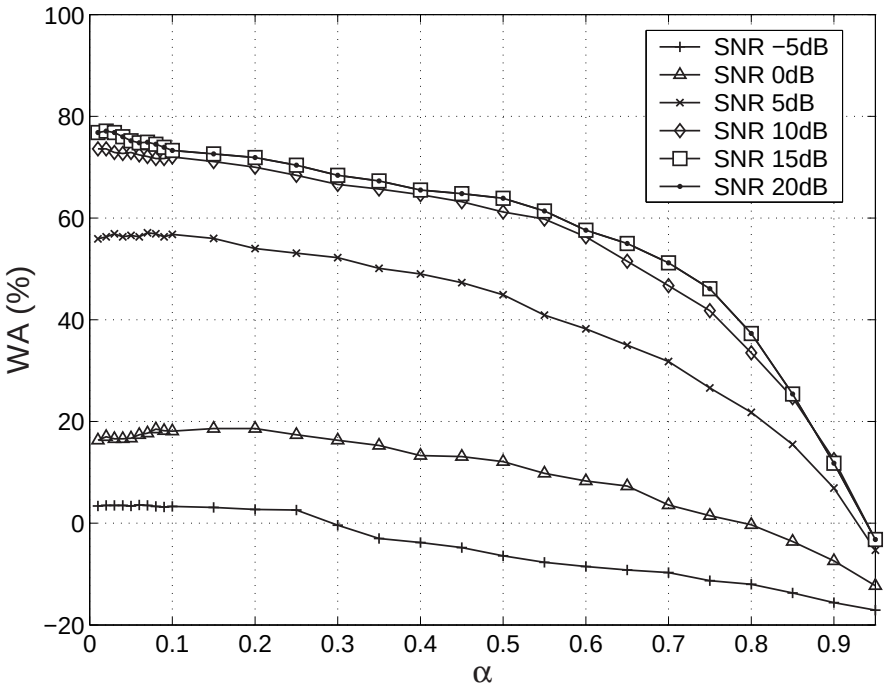

Figura 5.14: *WA para o sistema adaptado com ruído de rua e treinamento multi-estilo*

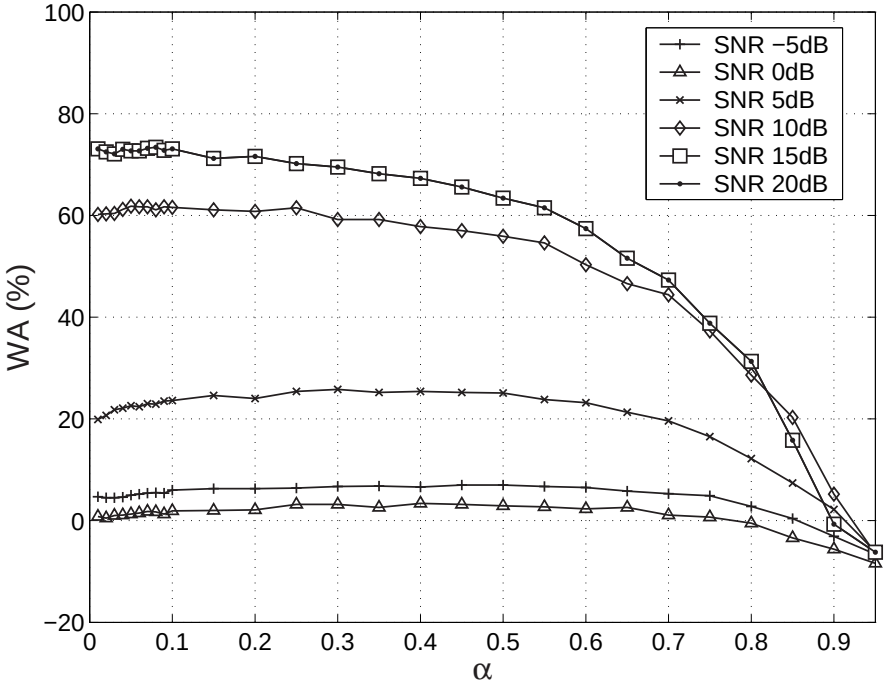

Figura 5.15: *WA para o sistema adaptado com ruído de metrô e treinamento multi-estilo*

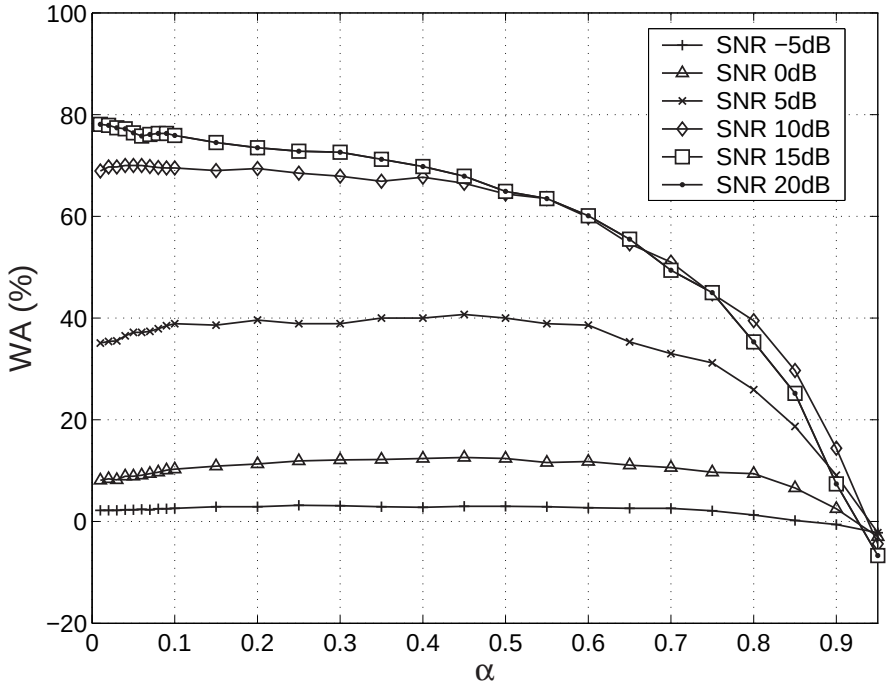

Figura 5.16: *WA para o sistema adaptado com ruído de trem e treinamento multi-estilo*

5.9 Modelagem do coeficiente de adaptação

Os resultados apresentados anteriormente mostram que a aplicação conjunta das técnicas multi-estilo e MAP propiciaram maior robustez ao ruído reduzindo a influência dos distúrbios ambientais no desempenho do sistema ASR.

Visando alcançar a resposta para motivação principal deste trabalho, que é identificar uma forma de predizer um valor adequado do coeficiente de adaptação para um dado tipo e nível de ruído, foram realizados estudos dos resultados obtidos para aplicação das técnicas multi-estilo e MAP.

O desafio deste trabalho é encontrar um meio de determinar um valor adequado do coeficiente de adaptação para um dado tipo e nível de ruído. Tendo em vista encontrar uma forma de apontar estes valores, o primeiro passo foi identificar os valores do coeficiente de adaptação que reduziram a WER comparado à referência, que neste caso é o sistema treinado com locuções ruidosas de 15 dB e 20 dB e testado com locuções ruidosas.

Para análise dos resultados dos testes, considerou-se que valores da taxa de acertos acima do valor de referência são valores aceitáveis, uma vez que introduziram ganho no processo de reconhecimento. Desta forma, foi determinada a faixa de valores do fator de adaptação para cada tipo e nível de ruído, conforme mostrado na Tabela 5.5.

As Figuras 5.17, 5.18, 5.19, 5.20, 5.21, 5.22, 5.23 e 5.24 mostram a relação entre

Tabela 5.5: *Faixa ótima de valores do coeficiente de adaptação para cada tipo e nível de ruído*

Tipo de Ruído	SNR −5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
Aeroporto	0,03-0,85	0,01-0,80	0,01-0,75	0,01-0,55	–	–
Balbúcio	0,01-0,40	0,03-0,04 0,07-0,45	0,01-0,60	0,15	–	–
Carro	0,06-0,70	0,01-0,60	0,01-0,60	0,03-0,25	–	–
Exposição	0,01-0,35	0,01-0,60	0,01-0,65	0,03 0,15	–	–
Restaurante	0,10 0,20-0,60	0,02 0,04-0,45	0,01-0,65	0,01-0,55	0,01 0,03-0,05	0,01 0,03-0,05
Rua	0,02-0,04 0,06	0,01-0,30	0,02-0,10	–	0,01-0,04	0,01-0,04
Metrô	0,01-0,80	0,01-0,75	0,01-0,70	0,01-0,35	0,01 0,07-0,08 0,1	0,01 0,07-0,08 0,1
Trem	0,06 0,08-0,70	0,01-0,80	0,01-0,65	0,01-0,20	–	–

SNR x α que fornece máxima WA, cujos valores estão no intervalo apresentados na Tabela 5.5.

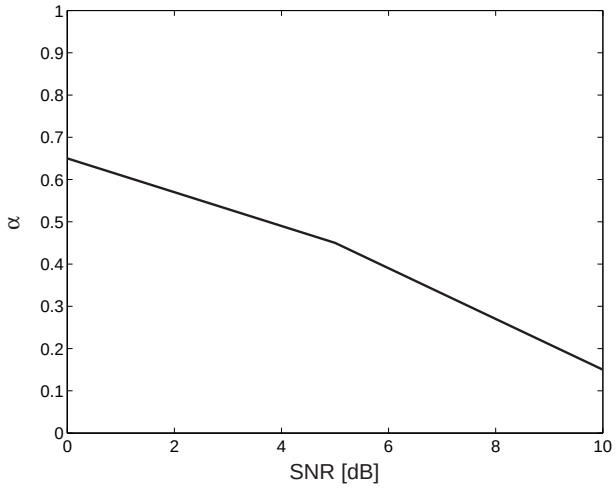

Figura 5.17: *Valores de α que fornecem máxima WA para ruído de aeroporto*

De forma a ilustrar a análise desenvolvida, a Figura 5.25 mostra os valores dos coeficientes de adaptação para diferentes níveis do ruído aeroporto. É importante salientar que existe uma nuvem de valores do coeficiente que proporcionam ganho ao sistema comparado à referência.

Uma vez determinada a região de valores, o passo seguinte foi identificar

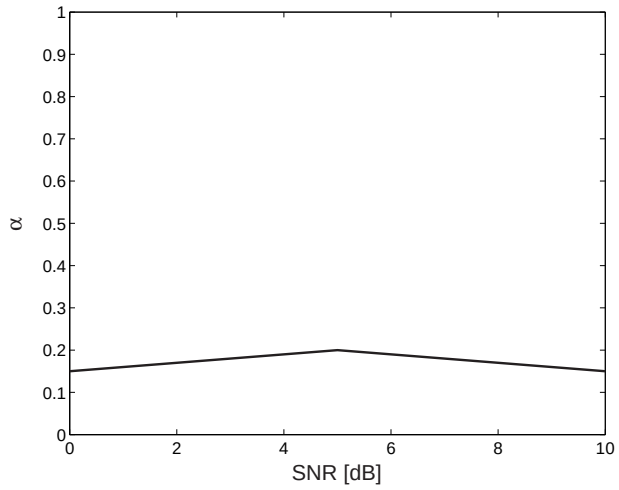

Figura 5.18: *Valores de α que fornecem máxima WA para ruído de balbúcio*

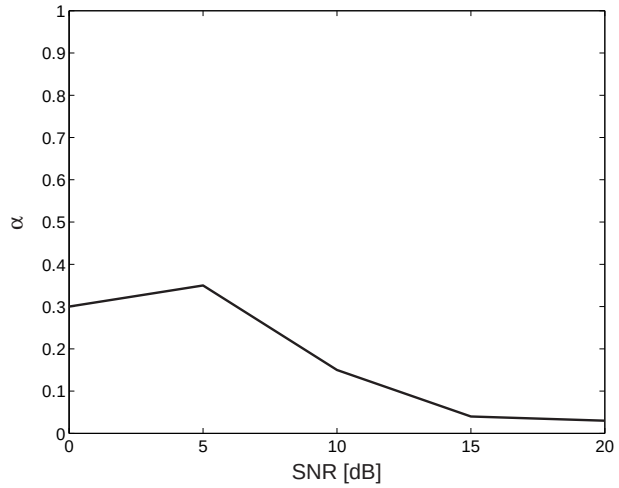

Figura 5.19: *Valores de α que fornecem máxima WA para ruído de carro*

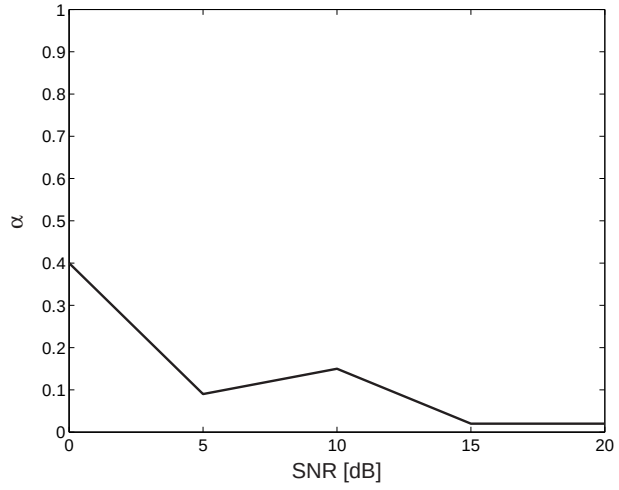

Figura 5.20: *Valores de α que fornecem máxima WA para ruído de exposição*

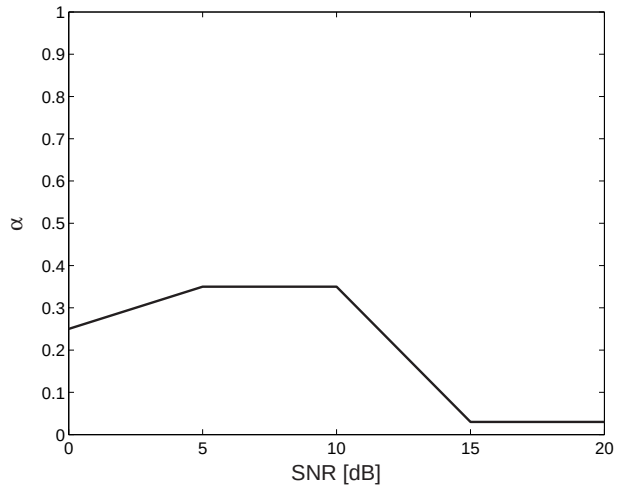

Figura 5.21: *Valores de α que fornecem máxima WA para ruído de restaurante*

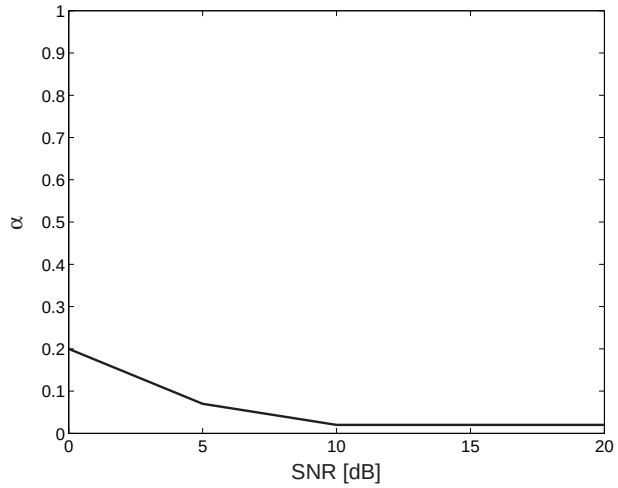

Figura 5.22: *Valores de α que fornecem máxima WA para ruído de rua*

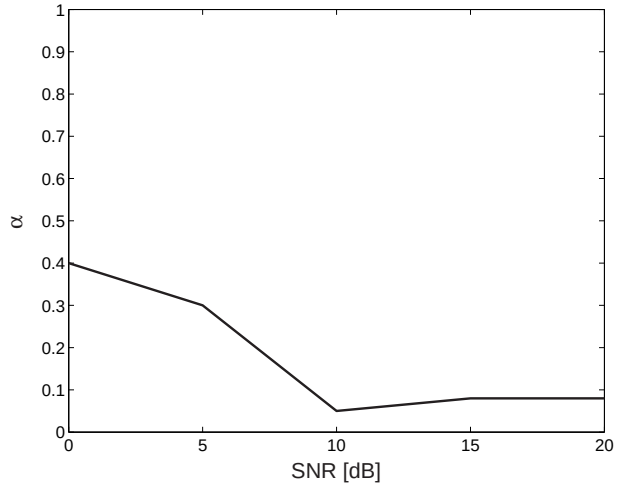

Figura 5.23: *Valores de α que fornecem máxima WA para ruído de metrô*

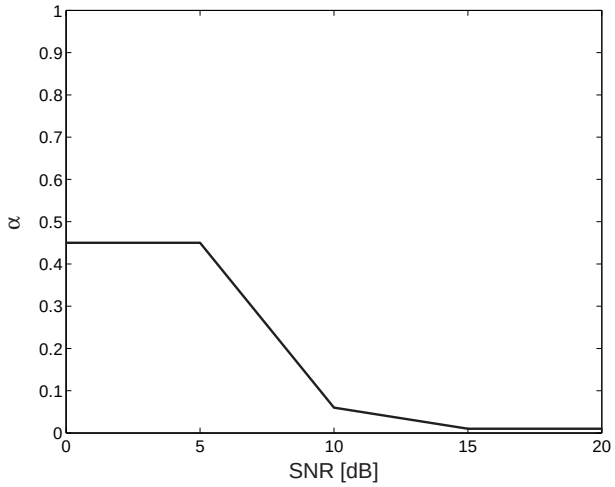

Figura 5.24: *Valores de α que fornecem máxima WA para ruído de trem*

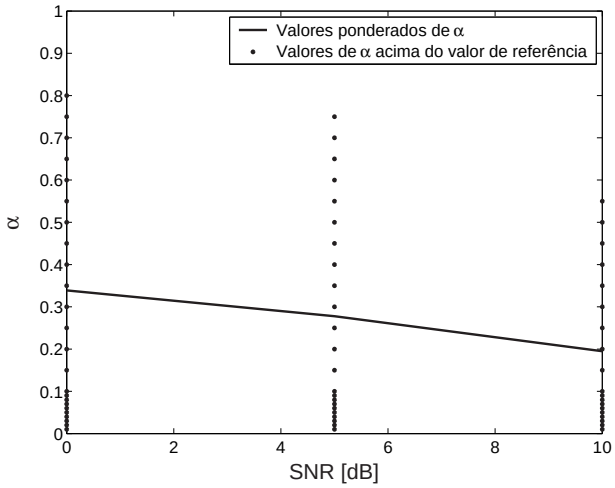

Figura 5.25: *Região de valores aceitáveis para coeficiente de adaptação para ruído de aeroporto*

como chegar a uma solução que permitisse descrever o comportamento do fator de adaptação para os diferentes tipos e intensidades de cada ruído.

Ao se analisar os dados provenientes da Figura 5.25, não foi possível determinar uma curva que descrevesse o valor adequado para cada nível do ruído. Porém, estudos seguintes mostraram que a média ponderada poderia modelar o comportamento da variação do coeficiente de adaptação α para cada SNR e, que este poderia ser aproximado à uma curva logística.

Análise similar foram realizadas para os demais ruídos: balbúcio, carro, exposição, restaurante, rua, metrô e trem, conforme pode ser verificado nas Figuras 5.26, 5.27, 5.28, 5.29, 5.30, 5.31 e 5.32.

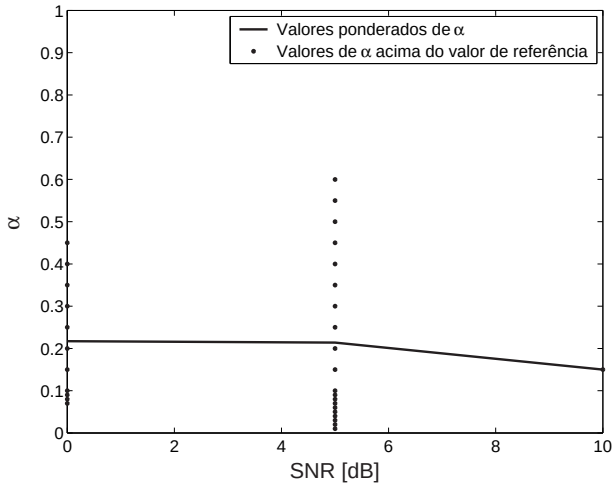

Figura 5.26: *Região de valores aceitáveis para coeficiente de adaptação para ruído de balbúcio*

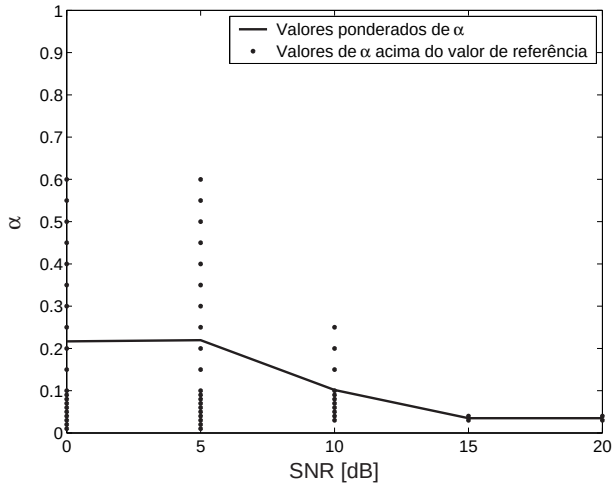

Figura 5.27: *Região de valores aceitáveis para coeficiente de adaptação para ruído de carro*

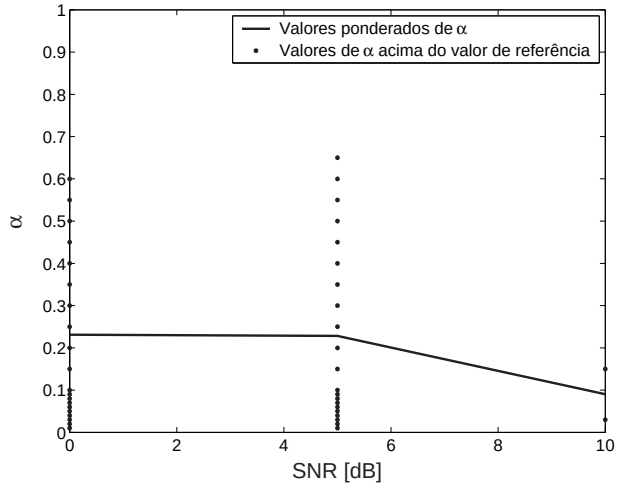

Figura 5.28: *Região de valores aceitáveis para coeficiente de adaptação para ruído de exposição*

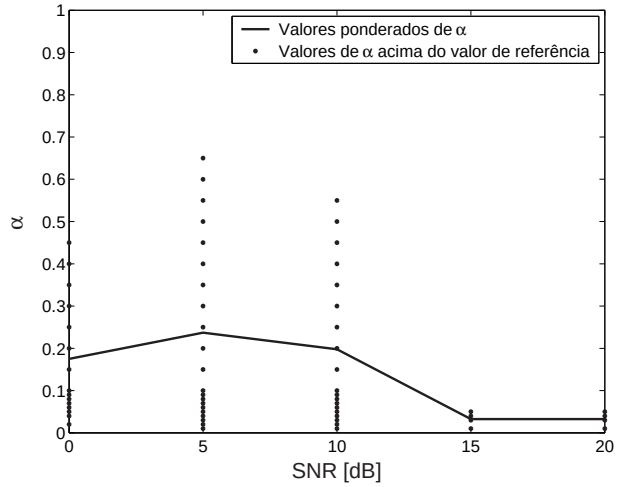

Figura 5.29: *Região de valores aceitáveis para coeficiente de adaptação para ruído de restaurante*

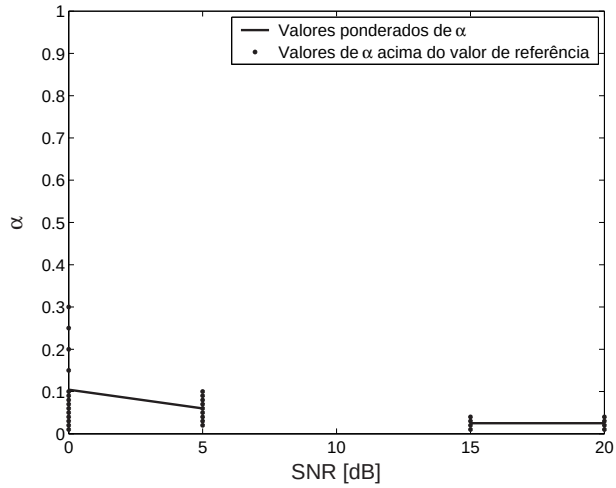

Figura 5.30: *Região de valores aceitáveis para coeficiente de adaptação para ruído de rua*

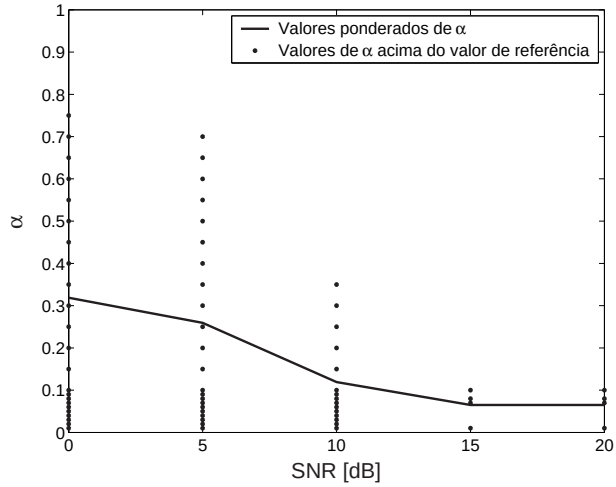

Figura 5.31: *Região de valores aceitáveis para coeficiente de adaptação para ruído de metrô*

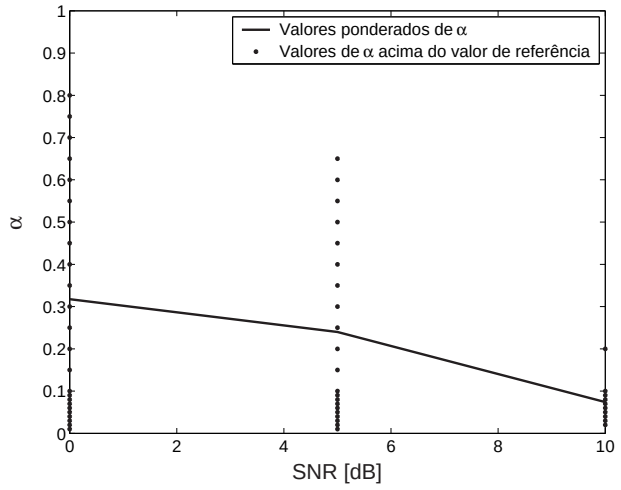

Figura 5.32: *Região de valores aceitáveis para coeficiente de adaptação para ruído de trem*

Seguindo o modelo proposto no Capítulo [**?**], uma vez identificados os valores ótimos do coeficiente α para cada tipo e nível de ruído, conforme mostrado na Tabela 5.5, foram calculados os respectivos α ponderados, cujos valores podem ser verificados nas Tabelas 5.6, 5.7, 5.8, 5.9, 5.10, 5.11, 5.12 e 5.13.

Tabela 5.6: *Comparação entre desempenhos obtidos para valores de α ponderado e máximo para ruído de aeroporto*

SNR (dB)	WA de referência (%)	α para WA máxima	WA máxima (%)	α'	WA para α' (%)	ΔWA (%)
0	$3,2$	$0,6500$	$7,3$	$0,3387$	$6,3$	$3,1$
5	$25,4$	$0,4500$	$32,1$	$0,2779$	$31,1$	$5,7$
10	$62,9$	$0,1500$	$66,2$	$0,1950$	$65,6$	$2,7$
15	$78,0$	$0,0100$	$77,8$	–	–	–
20	$78,0$	$0,0100$	$77,8$	–	–	–

Nas Tabelas 5.6, 5.7, 5.9, 5.11 e 5.13 é possível verificar que para alguns níveis de SNR não existe um valor para α ponderado, pois nenhum dos testes realizados retornou WA maior que a WA de referência. Para alguns casos, a adaptação introduziu uma pequena queda de desempenho ou, simplesmente, não proporcionou ganho. Entretanto, os resultados monstram que a combinação das técnicas resulta em boa relação custo benefício.

Baseado no algoritmo proposto e resultados experimentais obtidos durante a etapa de reconhecimento, através da ferramenta matemática, MATHCAD, obteve-se os valores para os parâmetros livres a, b e c que permitem a caracterização da relação entre SNR e coeficente de adaptação para cada um dos oito

Tabela 5.7: *Comparação entre desempenhos obtidos para valores de α ponderado e máximo para ruído de balbúcio*

SNR (dB)	WA de referência (%)	α para WA máxima	WA máxima (%)	α'	WA para α' (%)	ΔWA (%)
0	4,6	0,1500	5,9	0,2172	5,8	1,2
5	32,2	0,2000	37,4	0,2139	37,1	4,9
10	66,3	0,1500	66,7	0,1500	66,7	0,4
15	77,1	0,0100	76,9	–	–	–
20	77,1	0,0100	76,9	–	–	–

Tabela 5.8: *Comparação entre desempenhos obtidos para valores de α ponderado e máximo para ruído de carro*

SNR (dB)	WA de referência (%)	α para WA máxima	WA máxima (%)	α'	WA para α' (%)	ΔWA (%)
0	5,3	0,3000	7,1	0,2167	6,0	0,7
5	31,7	0,3500	37,3	0,2195	36,3	4,6
10	65,8	0,1500	67,7	0,1017	67,5	1,7
15	76,5	0,0400	76,6	0,0350	76,3	−0,2
20	76,5	0,0400	76,6	0,0350	76,3	−0,2

Tabela 5.9: *Comparação entre desempenhos obtidos para valores de α ponderado e máximo para ruído de exposição*

SNR (dB)	WA de referência (%)	α para WA máxima	WA máxima (%)	α'	WA para α' (%)	ΔWA (%)
0	0,2	0,4000	1,8	0,2310	1,0	0,8
5	16,4	0,0900	23,5	0,2283	23,1	6,7
10	58,2	0,1500	58,7	0,1267	58,0	−0,2
15	75,5	0,0200	75,5	–	–	–
20	75,5	0,0200	75,5	–	–	–

Tabela 5.10: *Comparação entre desempenhos obtidos para valores de α ponderado e máximo para ruído de restaurante*

SNR (dB)	WA de referência (%)	α para WA máxima	WA máxima (%)	α'	WA para α' (%)	ΔWA (%)
0	13,8	0,2500	15,4	0,1751	14,8	1,0
5	4,4	0,3500	9,1	0,2371	8,5	4,1
10	35,5	0,3500	41,6	0,1976	41,1	5,6
15	69,2	0,0300	69,7	0,0325	69,7	0,5
20	69,2	0,0300	69,7	0,0325	69,7	0,5

Tabela 5.11: *Comparação entre desempenhos obtidos para valores de α ponderado e máximo para ruído de rua*

SNR (dB)	WA de referência (%)	α para WA máxima	WA máxima (%)	α'	WA para α' (%)	ΔWA (%)
0	15,6	0,2000	18,6	0,1043	18,4	2,8
5	56,1	0,0700	57,1	0,0600	56,3	0,2
10	73,6	0,0200	73,6	–	–	–
15	75,8	0,0200	77,1	0,025	76,8	1,0
20	75,8	0,0200	77,1	0,025	76,8	1,0

Tabela 5.12: *Comparação entre desempenhos obtidos para valores de α ponderado e máximo para ruído de metrô*

SNR (dB)	WA de referência (%)	α para WA máxima	WA máxima (%)	α'	WA para α' (%)	ΔWA (%)
0	−0,3	0,4000	3,4	0,3186	2,7	3,0
5	18,4	0,3000	25,8	0,2594	25,6	7,2
10	57,9	0,0500	61,8	0,1193	61,7	3,8
15	73,0	0,0800	73,4	0,0650	72,9	−0,1
20	73,0	0,0800	73,4	0,0650	72,9	−0,1

Tabela 5.13: *Comparação entre desempenhos obtidos para valores de α ponderado e máximo para ruído de trem*

SNR (dB)	WA de referência (%)	α para WA máxima	WA máxima (%)	α'	WA para α' (%)	ΔWA (%)
0	7,6	0,4500	12,6	0,3174	12,1	4,5
5	34,0	0,4500	40,7	0,2398	38,8	4,8
10	69,0	0,0400	70,0	0,0749	69,9	0,9
15	78,1	0,0100	78,1	–	–	–
20	78,1	0,0100	78,1	–	–	–

modelos de distorção disponíveis na base AURORA. Na Tabela 5.14 verifica-se os valores para cada um destes parâmetros.

Tabela 5.14: *Parâmetros para modelagem paramétrica usando função logística*

Noise	a	b	c
Aeroporto	−0,104175	−0,810155	0,186922
Balbúcio	−0,319246	1,338508	−0,009345
Carro	−0,442008	1,445377	−0,025966
Exposição	−0,124371	1,049212	0,028005
Restaurante	−0,148612	1,701653	−0,020750
Rua	−1,264809	2,320370	−0,014900
Metrô	−0,131225	1,000045	−0,049691
Trem	−0,299248	0,882365	0,000910

A partir dos valores apresentados na Tabela 5.14 traçou-se os gráficos que permitem a determinação do valor de α para diferentes condições ruidosas. Nas Figuras 5.33, 5.34, 5.35, 5.36, 5.37, 5.38, 5.39 e 5.40 verifica-se o comportamento de cada distorção, bem como, relacioná-la com o valor que retorna o desempenho máximo do sistema.

5.10 Análise da Parametrização Logística

Nesta Seção são apresentados os resultados experimentais que validam o método proposto.

No intuito de validar as curvas logísticas apresentadas na Seção anterior, escolheu-se aleatoriamente três valores de diferentes níveis de SNR (2 dB, 7 dB e 12 dB). Uma vez obtido o valor do coeficiente α oriundo das curvas, para cada intensidade dos diferentes ruídos, executou-se a adaptação seguida do teste de reconhecimento. A WA obtida em cada teste foi comparada ao valor de referência

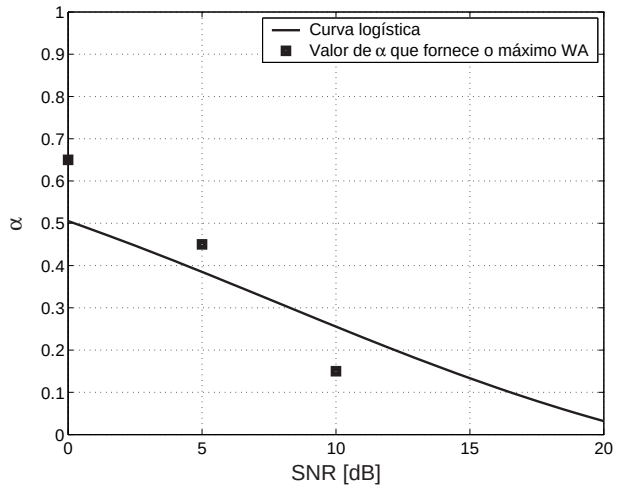

Figura 5.33: *Curva logística para ruído de aeroporto*

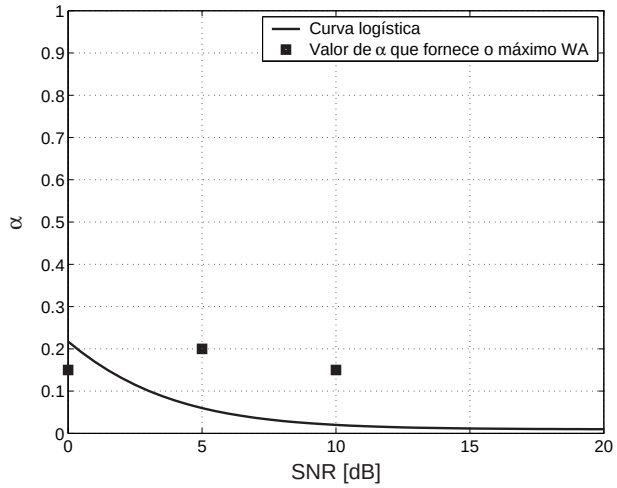

Figura 5.34: *Curva logística para ruído de balbúcio*

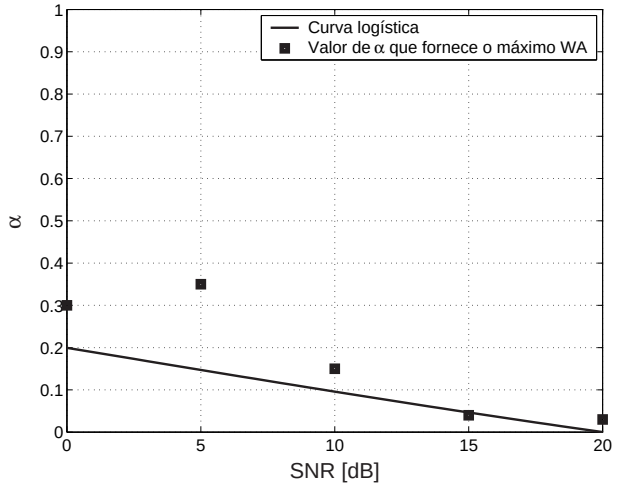

Figura 5.35: *Curva logística para ruído de carro*

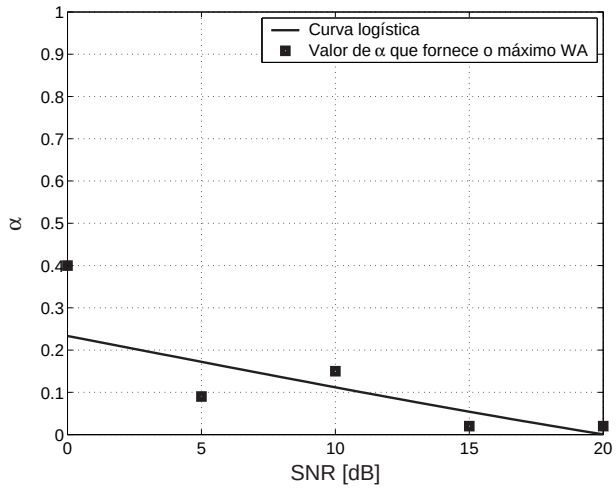

Figura 5.36: *Curva logística para ruído de exposição*

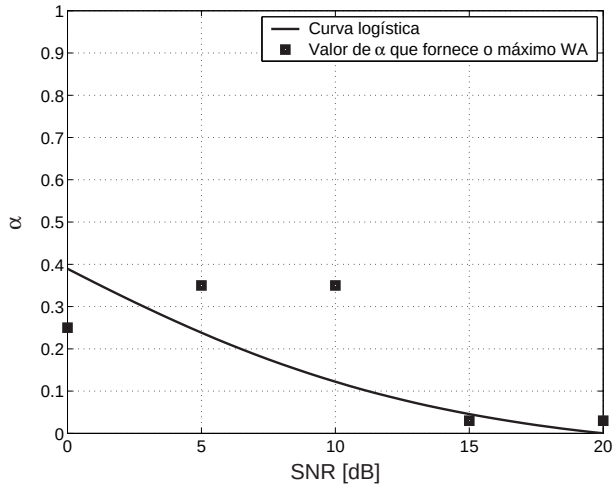

Figura 5.37: *Curva logística para ruído de restaurante*

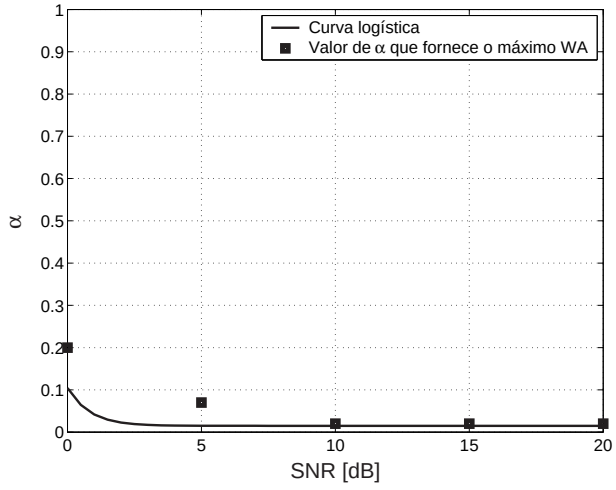

Figura 5.38: *Curva logística para ruído de rua*

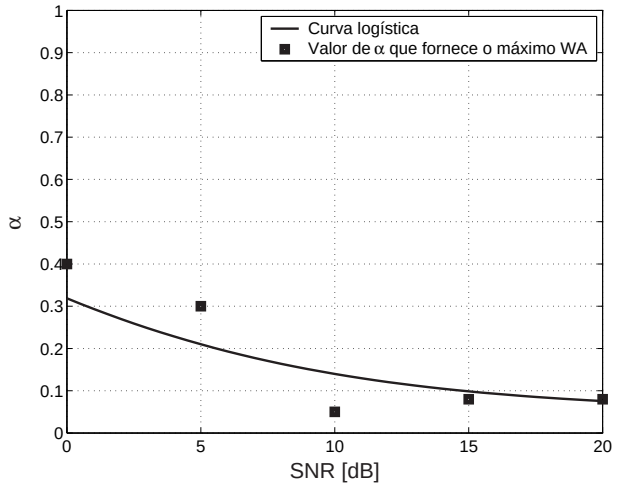

Figura 5.39: *Curva logística para ruído de metrô*

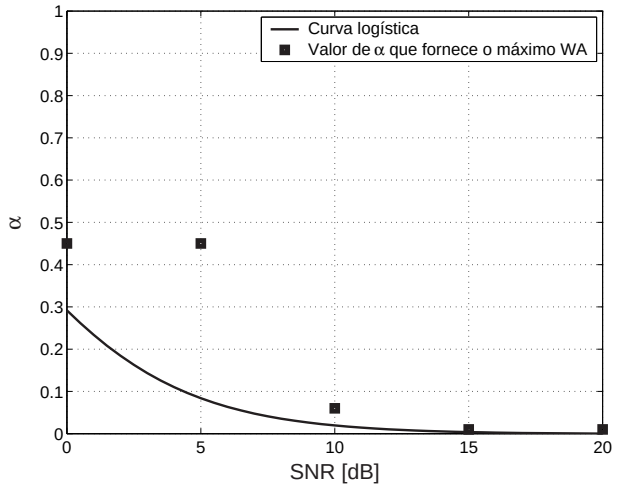

Figura 5.40: *Curva logística para ruído de trem*

(WA para o sistema treinado e testado com dados corrompidos). O processo de validação das curvas propostas consiste em verificar se a taxa de acertos obtida foi maior que a de referência.

As Tabelas 5.15, 5.16 e 5.17 apresentam os resultados experimentais encontrados nos processos de reconhecimento. Verifica-se, claramente, que no geral houve uma melhora no desempenho do sistema para todos os tipos de ruído utilizados. No pior dos casos, o processo não retornou ganho, porém a taxa obtida foi igual ao valor de referência. Desta forma, comprova-se que as curvas logísticas provenientes da modelagem proposta são válidas para o levantamento do valor adequado de α para um dado nível de ruído.

Tabela 5.15: *WA para SNR = 2 dB usando o valor de α proveniente da curva logística*

Tipo de ruído	α da Curva Logística	WA de Referência (%)	WA usando a Curva Logística (%)	ΔWA (%)
Aeroporto	$0,4591$	$7,9$	$12,8$	$4,9$
Balbúcio	$0,1310$	$10,5$	$15,4$	$4,9$
Carro	$0,1147$	$12,5$	$15,9$	$3,4$
Exposição	$0,1865$	$2,4$	$6,9$	$4,5$
Restaurante	$0,1401$	$6,1$	$7,5$	$1,4$
Rua	$0,0227$	$30,3$	$31,5$	$1,2$
Metrô	$0,2702$	$2,3$	$9,0$	$6,7$
Trem	$0,1844$	$12,3$	$15,6$	$3,3$

Tabela 5.16: *WA para SNR = 7 dB usando o valor de α proveniente da curva logística*

Tipo de ruído	α da Curva Logística	WA de Referência (%)	WA usando a Curva Logística (%)	ΔWA (%)
Aeroporto	$0,3333$	$39,8$	$47,3$	$7,5$
Balbúcio	$0,0366$	$48,7$	$50,3$	$1,6$
Carro	$0,0365$	$49,8$	$50,8$	$1,0$
Exposição	$0,0999$	$31,3$	$38,4$	$7,1$
Restaurante	$0,0813$	$11,4$	$15,5$	$4,1$
Rua	$0,0149$	$65,9$	$65,9$	$0,0$
Metrô	$0,1777$	$33,9$	$40,8$	$6,9$
Trem	$0,0476$	$51,6$	$53,4$	$1,8$

Tabela 5.17: *WA para SNR = 12 dB usando o valor de α proveniente da curva logística*

Tipo de ruído	α da Curva Logística	WA de Referência (%)	WA usando a Curva Logística (%)	ΔWA (%)
Aeroporto	$0,2048$	$71,7$	$72,4$	$0,7$
Balbúcio	$0,0150$	$73,7$	$74,2$	$0,5$
Carro	$0,0271$	$73,9$	$73,9$	$0,0$
Exposição	$0,0450$	$69,1$	$69,6$	$0,5$
Restaurante	$0,0505$	$53,2$	$54,9$	$1,7$
Rua	$0,0149$	$76,0$	$76,0$	$0,0$
Metrô	$0,1205$	$66,8$	$67,3$	$0,5$
Trem	$0,0104$	$74,6$	$74,6$	$0,0$

Capítulo 6

Conclusões

6.1 Considerações finais

Os estudos que nortearam o presente trabalho demonstram que o descasamento acústico entre as condições de treinamento e teste causam uma queda drástica de desempenho dos sistemas automáticos de reconhecimento de fala. Neste contexto, a aplicação conjunta das técnicas treinamento multi-estilo e adaptação Bayesiana agrega robustez contra as distorções ambientais indesejadas promovendo uma melhora significativa no desempenho dos sistemas de reconhecimento automático de fala contínua. Entretanto, verifica-se na literatura que um dos desafios da aplicação da adaptação baseada no critério Máximo a Posteriori é encontrar o valor adequado do coeficiente de adaptação para um dado tipo e nível de ruído a ser utilizado no processo de adaptação que antecede a fase de reconhecimento de fala. Este procedimento de busca em geral é feito em forma de varredura, o que representa um alto custo computacional, o que é ruim para um sistema que deve operar em tempo real. Portanto, este fator pode impactar a aplicação conjunta das técnicas em tarefas que demandam rápido tempo de processamento, tais como: controle de sistemas veiculares, aplicações ligadas à segurança física, controle de dispositivos dedicados à pessoas portadoras de deficiências, interfaces para controle de processos industriais, entre outras.

Desta forma, este estudo concentra-se na avaliação e validação de um algoritmo que modela o comportamento da variação do coeficiente de adaptação de acordo com a intensidade e tipo da distorção ambiental indesejada presente no meio durante a etapa de reconhecimento. Este algoritmo baseia-se no ajuste paramétrico que aproxima o comportamento do sistema a uma função decrescente do tipo logística. Os resultados experimentais obtidos para os valores do coeficiente de adaptação relacionados a uma determinada relação sinal-ruído, tomados das curvas propostas para cada um dos tipos de ruído (aeroporto, balbúcio, carro,

exposição, restaurante, rua, metrô e trem), comprovaram a eficácia do método apresentado. Portanto, quando comparado aos valores de referência (WA para o sistema treinado e testado com locuções corrompidas), verificou-se um ganho médio de aproximadamente 3 %.

Uma vantagem proporcionada pelo algoritmo proposto neste trabalho é que uma vez detectado e identificado o tipo e intensidade de ruído no presente momento do reconhecimento em aplicações reais, o valor adequado do coeficiente α a ser utilizado na etapa de adaptação pode ser facilmente verificado a partir dos dados das curvas traçadas reduzindo a complexidade computacional e tempo de processamento. Vale ressaltar que o método não necessariamente retorna o valor de α que proporciona o máximo ganho possível ao sistema. O foco principal do método é facilitar a escolha do coeficiente de adaptação que retorne melhora no desempenho do ASR para toda a faixa experimental (SNR entre -5 dB e 20 dB).

6.2 Sugestão para trabalhos futuros

Seguindo a linha de pesquisa na área de processamento de voz no que se diz respeito a minimizar os efeitos indesejados causados pelas condições ambientais adversas em tarefas de reconhecimento de fala, verifica-se a constante busca por técnicas de adaptação e redução da influência do ruído. Geralmente, verifica-se a aplicação de dados corrompidos artificialmente na análise e validação de métodos, conforme estudo realizado neste trabalho. Desta forma, uma interessante pesquisa a ser realizada é a verificação e validação do modelamento proposto utilizando uma base de dados gravada em ambientes reais, avaliando, portanto, não apenas a influência do ruído aditivo na resposta do sistema mas também o ruído convolucional.

Outro tema importante a ser explorado é o estudo estatístico dos ruídos. Uma vez detectado o ruído presente durante a fase de reconhecimento, é possível verificar a correlação do mesmo com os modelos de distorções previamente conhecidas. Desta forma, pode-se identificar qual curva logística proposta no presente trabalho é a mais aplicável no processo de adaptação numa aplicação real.

Anexo A

Subunidades Fonéticas

#
a
an
e
E
en
i
y
in
o
O
on
u
un
b
d
D
f
g
j
k
l
L
m
n
N
p

r
rr
R
s
t
T
v
x
z

Anexo B

Resultados do processo de reconhecimento para o sistema treinado com locuções limpas, adaptado com ruído e testado com locuções corrompidas

Tabela B.1: *Taxa de acertos de palavras, em %, para um sistema treinado com dados limpos, adaptado com ruído de aeroporto e testado com locuções ruidosas*

α	SNR -5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
$0,01$	$1,3$	$-2,9$	$16,7$	$45,6$	$66,5$	$66,5$
$0,02$	$1,6$	$-3,0$	$17,8$	$46,2$	$66,5$	$66,5$
$0,03$	$1,9$	$-2,8$	$18,5$	$47,3$	$66,8$	$66,8$
$0,04$	$1,9$	$-2,1$	$19,4$	$47,9$	$67,3$	$67,3$
$0,05$	$2,1$	$-1,5$	$19,9$	$48,5$	$67,3$	$67,3$
$0,06$	$2,4$	$-1,2$	$19,8$	$49,2$	$67,4$	$67,4$
$0,07$	$2,6$	$-1,4$	$20,0$	$49,2$	$67,8$	$67,8$
$0,08$	$2,2$	$-0,8$	$21,2$	$50,0$	$67,6$	$67,6$
$0,09$	$2,3$	$-0,2$	$21,2$	$50,5$	$67,9$	$67,9$
$0,10$	$2,6$	$-0,3$	$21,8$	$50,7$	$67,3$	$67,3$
$0,15$	$2,8$	$1,3$	$22,7$	$51,6$	**68,5**	**68,5**
$0,20$	$2,9$	$1,4$	$22,9$	$51,8$	$68,3$	$68,3$
$0,25$	$3,0$	$2,4$	$22,3$	$51,6$	$68,1$	$68,1$
$0,30$	$3,2$	$2,6$	$23,7$	$51,6$	$68,2$	$68,2$
$0,35$	$3,3$	$1,8$	$24,0$	$51,9$	$68,3$	$68,3$
$0,40$	$3,2$	$1,7$	$24,1$	$52,4$	$68,4$	$68,4$
$0,45$	$3,1$	$2,3$	$24,4$	$52,6$	$67,2$	$67,2$
$0,50$	$3,0$	$2,7$	**25,0**	**53,0**	$66,5$	$66,5$
$0,55$	$3,2$	**2,8**	$24,4$	$52,7$	$65,6$	$65,6$
$0,60$	**3,4**	**2,8**	$23,1$	$52,3$	$63,0$	$63,0$
$0,65$	$3,1$	$2,1$	$22,6$	$50,0$	$60,3$	$60,3$
$0,70$	$2,8$	$1,0$	$20,3$	$48,1$	$57,9$	$57,9$
$0,75$	$2,2$	$1,1$	$19,3$	$41,9$	$52,6$	$52,6$
$0,80$	$2,3$	$0,7$	$16,3$	$37,3$	$45,3$	$45,3$
$0,85$	$2,4$	$-0,2$	$11,5$	$29,3$	$35,6$	$35,6$
$0,90$	$2,1$	$-2,5$	$5,9$	$18,5$	$23,1$	$23,1$
$0,95$	$2,2$	$-6,6$	$-2,0$	$3,8$	$2,7$	$2,7$

Tabela B.2: *Taxa de acertos de palavras, em %, para um sistema treinado com dados limpos, adaptado com ruído de balbúcio e testado com locuções ruidosas*

α	SNR −5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
0,01	0,7	−2,4	23,0	55,5	70,2	70,2
0,02	0,8	−2,4	23,7	55,3	70,0	70,0
0,03	0,8	−1,7	24,8	56,3	70,5	70,5
0,04	1,1	−2,0	25,3	56,8	70,4	70,4
0,05	1,2	−1,9	25,5	56,1	71,2	71,2
0,06	1,2	−1,5	25,6	56,0	70,8	70,8
0,07	1,1	−1,5	27,0	56,7	71,2	71,2
0,08	1,0	−1,4	27,6	57,6	**71,5**	**71,5**
0,09	**1,3**	−1,1	28,4	57,8	71,2	71,2
0,10	1,2	−0,7	28,3	57,8	**71,5**	**71,5**
0,15	1,0	−0,2	29,9	58,1	70,8	70,8
0,20	0,4	0,8	30,6	58,6	69,9	69,9
0,25	0,2	1,9	31,3	**58,8**	70,1	70,1
0,30	−1,6	1,9	30,8	58,4	69,3	69,3
0,35	−0,7	2,5	**31,9**	58,6	68,9	68,9
0,40	−0,7	**3,5**	31,1	57,9	68,8	68,8
0,45	0,0	3,4	30,5	57,4	68,1	68,1
0,50	0,5	3,0	30,3	55,8	66,9	66,9
0,55	0,4	3,2	28,7	55,4	65,2	65,2
0,60	0,7	3,1	28,6	52,5	62,9	62,9
0,65	0,7	2,0	28,7	50,1	59,7	59,7
0,70	0,6	1,1	26,1	45,9	54,2	54,2
0,75	0,5	0,3	21,5	40,9	49,4	49,4
0,80	0,1	−1,8	17,4	35,8	42,7	42,7
0,85	−0,2	−2,4	13,0	29,2	31,9	31,9
0,90	−1,1	−6,1	5,5	18,3	20,8	20,8
0,95	−3,1	−9,0	−0,8	4,7	3,8	3,8

Tabela B.3: *Taxa de acertos de palavras, em %, para um sistema treinado com dados limpos, adaptado com ruído de carro e testado com locuções ruidosas*

α	SNR -5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
$0,01$	$0,3$	$5,2$	$24,3$	$56,8$	$70,8$	$70,8$
$0,02$	$0,2$	$4,6$	$24,5$	$57,2$	$71,1$	$71,1$
$0,03$	$0,2$	$5,0$	$25,5$	$57,2$	$71,6$	$71,6$
$0,04$	$0,2$	$5,4$	$26,3$	$57,2$	$71,9$	$71,9$
$0,05$	$0,2$	**6,3**	$25,6$	$57,6$	**72,4**	**72,4**
$0,06$	$0,2$	$5,5$	$25,5$	$58,0$	$72,1$	$72,1$
$0,07$	$0,2$	$5,6$	$25,7$	$58,0$	$72,3$	$72,3$
$0,08$	$0,2$	$5,5$	$25,9$	$58,4$	$72,1$	$72,1$
$0,09$	$0,2$	$5,2$	$26,2$	$58,0$	$72,0$	$72,0$
$0,10$	$0,2$	$4,5$	$26,6$	$57,7$	$71,3$	$71,3$
$0,15$	$0,3$	$4,5$	$27,1$	$57,9$	$72,0$	$72,0$
$0,20$	$0,4$	$5,3$	$27,6$	$58,1$	**72,4**	**72,4**
$0,25$	$0,4$	$4,8$	$27,0$	**58,8**	$72,2$	$72,2$
$0,30$	$0,3$	$4,8$	$27,3$	$58,3$	$71,1$	$71,1$
$0,35$	$0,4$	$4,8$	$26,6$	$57,9$	$70,4$	$70,4$
$0,40$	$0,4$	$5,0$	$28,1$	$58,5$	$69,0$	$69,0$
$0,45$	$0,5$	$4,5$	$29,8$	$57,6$	$68,7$	$68,7$
$0,50$	$0,7$	$4,4$	**29,9**	$57,8$	$67,4$	$67,4$
$0,55$	$0,6$	$3,2$	$28,8$	$56,7$	$66,0$	$66,0$
$0,60$	$0,6$	$2,0$	$27,4$	$53,5$	$63,9$	$63,9$
$0,65$	$0,7$	$1,6$	$25,3$	$49,5$	$57,3$	$57,3$
$0,70$	$0,7$	$1,0$	$23,0$	$44,1$	$49,4$	$49,4$
$0,75$	**1,3**	$-0,3$	$18,8$	$37,0$	$42,6$	$42,6$
$0,80$	**1,3**	$-1,5$	$14,4$	$26,1$	$29,4$	$29,4$
$0,85$	$0,9$	$-3,6$	$6,9$	$16,2$	$18,3$	$18,3$
$0,90$	$-0,1$	$-7,2$	-0.7	$3,8$	$2,3$	$2,3$
$0,95$	$-1,7$	$-10,7$	$-7,7$	$-8,2$	$-10,0$	$-10,0$

Tabela B.4: *Taxa de acertos de palavras, em %, para um sistema treinado com dados limpos, adaptado com ruído de exposição e testado com locuções ruidosas*

α	SNR −5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
0,01	0,7	−3,4	9,2	47,2	69,6	69,6
0,02	1,0	−4,4	10,0	47,8	**69,7**	**69,7**
0,03	1,0	−3,9	11,0	48,6	69,3	69,3
0,04	1,0	−3,5	12,0	48,9	**69,7**	**69,7**
0,05	1,1	−3,2	12,3	**49,5**	69,6	69,6
0,06	1,2	−2,7	12,5	48,8	**69,7**	**69,7**
0,07	**1,5**	−2,4	12,6	48,3	69,5	69,5
0,08	1,3	−2,6	12,7	48,4	68,8	68,8
0,09	1,3	−3,1	13,0	48,6	68,9	68,9
0,10	1,1	−2,4	13,1	48,9	68,6	68,6
0,15	0,6	−1,9	14,2	48,6	67,7	67,7
0,20	0,2	−1,8	**14,8**	47,4	66,8	66,8
0,25	0,3	**-1,0**	**14,8**	47,3	65,9	65,9
0,30	0,2	−1,1	13,7	46,7	65,5	65,5
0,35	0,1	−1,3	**14,8**	46,3	64,3	64,3
0,40	0,0	−1,9	14,6	45,2	63,2	63,2
0,45	0,2	−1,5	14,4	43,6	62,4	62,4
0,50	0,3	−1,3	13,9	41,7	59,2	59,2
0,55	0,2	−1,1	12,5	39,2	55,1	55,1
0,60	0,4	−1,4	11,6	36,0	51,4	51,4
0,65	0,4	−1,8	11,3	33,8	47,5	47,5
0,70	0,5	−2,6	9,4	28,6	38,5	38,5
0,75	0,3	−3,0	6,8	19,7	28,7	28,7
0,80	0,5	−5,0	3,0	11,8	17,9	17,9
0,85	0,1	−5,6	−2,3	4,4	5,5	5,5
0,90	−0,5	−6,8	−5,4	−3,3	−3,8	−3,8
0,95	−4,4	−10,6	−8,6	−10,5	−9,8	−9,8

Tabela B.5: *Taxa de acertos de palavras, em %, para um sistema treinado com dados limpos, adaptado com ruído de restaurante e testado com locuções ruidosas*

α	SNR −5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
0,01	−1,2	5,1	−3,2	21,5	52,3	52,3
0,02	−1,1	5,3	−3,1	22,4	53,0	53,0
0,03	**-0,8**	5,6	−2,5	23,4	53,6	53,6
0,04	−0,9	5,5	−2,6	23,2	54,1	54,1
0,05	−1,3	6,8	−1,8	23,9	54,5	54,5
0,06	**-0,8**	6,0	−1,9	24,3	54,6	54,6
0,07	−1,0	5,8	−1,5	25,5	55,2	55,2
0,08	−1,1	5,6	−1,3	26,2	55,2	55,2
0,09	−1,8	5,0	−1,2	26,3	55,6	55,6
0,10	−1,4	5,1	−1,5	26,4	56,1	56,1
0,15	−1,7	5,6	0,2	27,9	56,6	56,6
0,20	−1,8	5,7	2,1	**28,0**	56,0	56,0
0,25	−2,2	6,2	2,3	27,9	**56,9**	**56,9**
0,30	−2,4	**7,4**	**2,7**	27,6	56,8	56,8
0,35	−2,2	7,0	1,1	27,1	55,7	55,7
0,40	−2,5	6,8	1.6	26,9	56,3	56,3
0,45	−2,1	6,7	1,6	26,6	55,0	55,0
0,50	−2,2	5,4	0,5	26,4	54,8	54,8
0,55	−2,6	6,3	0,9	26,3	53,6	53,6
0,60	−2,9	5,4	−0,4	26,5	51,3	51,3
0,65	−2,5	5,3	−0,7	25,3	50,1	50,1
0,70	−2,4	6,2	−0,4	25,2	47,0	47,0
0,75	−2,8	4,6	−0,8	23,4	43,4	43,4
0,80	−3,9	2,6	−1,4	19,4	36,0	36,0
0,85	−5,1	0,0	−3,6	14,1	27.7	27,7
0,90	−0,7	−3,4	−5,8	6,4	19,5	19,5
0,95	−6,9	−8,8	−10.4	−2,2	3,9	3,9

Tabela B.6: *Taxa de acertos de palavras, em %, para um sistema treinado com dados limpos, adaptado com ruído de rua e testado com locuções ruidosas*

α	SNR −5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
0,01	**2,1**	9,0	42,9	65,8	**74,9**	**74,9**
0,02	1,9	9,2	43,7	65,9	74,6	74,6
0,03	**2,1**	8,9	43,9	**66,1**	74,6	74,6
0,04	2,0	9,1	44,5	65,8	74,7	74,7
0,05	**2,1**	9,6	45,0	65,4	74,3	74,3
0,06	1,7	9,8	**45,4**	65,8	73,3	73,3
0,07	1,3	10,4	44,4	65,6	73,4	73,4
0,08	0,9	10,0	43,8	65,5	73,0	73,0
0,09	0,0	9,5	44,0	65,8	72,3	72,3
0,10	−1,0	10,3	43,6	65,5	72,1	72,1
0,15	−8,5	**10,7**	43,1	64,9	71,4	71,4
0,20	−11,4	9,7	42,8	62,6	70,5	70,5
0,25	−11,5	9,0	41,5	61,1	70,7	70,7
0,30	−10,6	9,2	40,4	60,6	69,0	69,0
0,35	−10,8	9,0	39,1	59,4	68,2	68,2
0,40	−11,5	9,4	37,1	58,9	68,0	68,0
0,45	−11,7	8,7	36,9	58,6	65,7	65,7
0,50	−12,1	7,1	36,0	57,8	64,7	64,7
0,55	−14,4	6,1	33,7	55,2	63,0	63,0
0,60	−16,7	4,6	31,7	51,8	59,6	59,6
0,65	−18,9	4,1	30,1	49,5	57,8	57,8
0,70	−19,5	0,6	26,0	44,8	52,2	52,2
0,75	−18,7	−2,5	21,9	39,9	48,0	48,0
0,80	−19,0	−6,3	17,3	34,5	40,0	40,0
0,85	−19,4	−8,2	11,5	24,4	31,4	31,4
0,90	−17,9	−11,3	2,3	13,3	19,6	19,6
0,95	−15,8	−16,1	−5,9	0,6	2,5	2,5

Tabela B.7: *Taxa de acertos de palavras, em %, para um sistema treinado com dados limpos, adaptado com ruído de metrô e testado com locuções ruidosas*

α	SNR -5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
$0,01$	$0,9$	$-1,5$	$12,9$	$45,1$	$65,3$	$65,3$
$0,02$	$0,9$	$-1,2$	$13,1$	$46,8$	$66,5$	$66,5$
$0,03$	$0,8$	$-0,7$	$13,6$	$47,4$	$67,0$	$67,0$
$0,04$	$0,8$	$-0,3$	$13,8$	$47,6$	$67,0$	$67,0$
$0,05$	$0,8$	$0,2$	$14,9$	$48,0$	$66,9$	$66,9$
$0,06$	$0,7$	$-0,1$	$15,0$	$48,4$	$66,4$	$66,4$
$0,07$	$0,8$	$0,1$	$15,2$	$48,4$	$66,9$	$66,9$
$0,08$	$0,8$	$-0,2$	$16,1$	$48,8$	$66,8$	$66,8$
$0,09$	$0,9$	$-0,3$	$15,8$	$48,3$	**67,1**	**67,1**
$0,10$	$0,8$	$0,1$	$16,1$	$48,6$	$66,8$	$66,8$
$0,15$	$0,8$	$0,3$	$16,7$	$48,7$	$67,0$	$67,0$
$0,20$	$1,0$	$1,2$	$16,8$	$48,4$	$66,2$	$66,2$
$0,25$	$1,1$	$1,3$	$16,7$	$48,4$	$65,9$	$65,9$
$0,30$	$1,3$	$1,5$	$17,4$	**48,9**	$66,2$	$66,2$
$0,35$	$1,2$	$1,1$	$18,2$	$48,5$	$64,5$	$64,5$
$0,40$	$1,1$	$0,8$	$17,9$	$48,5$	$63,2$	$63,2$
$0,45$	$1,3$	$1,6$	**18,8**	$48,5$	$62,8$	$62,8$
$0,50$	$1,3$	**2,2**	$18,5$	$48,2$	$62,0$	$62,0$
$0,55$	$1,5$	$1,8$	$17,4$	$47,1$	$59,6$	$59,6$
$0,60$	$1,6$	$1,5$	$17,9$	$44,7$	$55,9$	$55,9$
$0,65$	$1,7$	$0,5$	$16,1$	$42,3$	$52,7$	$52,7$
$0,70$	**1,8**	$0,5$	$15,2$	$38,1$	$46,9$	$46,9$
$0,75$	$1,6$	$0,9$	$10,6$	$32,4$	$40,2$	$40,2$
$0,80$	$1,4$	$-0,6$	$7,5$	$25,3$	$29,6$	$29,6$
$0,85$	$0,9$	$-2,8$	$2,9$	$13,9$	$16,0$	$16,0$
$0,90$	$0,3$	$-6,1$	$-3,5$	$1,3$	$1,0$	$1,0$
$0,95$	$-1,1$	$-9,3$	$-8,5$	$-7,7$	$-9,2$	$-9,2$

Tabela B.8: *Taxa de acertos de palavras, em %, para um sistema treinado com dados limpos, adaptado com ruído de trem e testado com locuções ruidosas*

α	SNR −5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
0,01	**3,7**	6,8	27,7	60,7	73,5	73,5
0,02	3,4	6,9	28,8	61,6	**73,9**	**73,9**
0,03	3,2	6,9	28,6	61,9	73,6	73,6
0,04	3,2	7,1	29,0	**62,0**	73,7	73,7
0,05	3,2	7,3	29,4	61,7	73,4	73,4
0,06	3,2	7,3	29,0	61,9	73,1	73,1
0,07	3,2	7,2	29,2	61,8	73,1	73,1
0,08	3,1	7,3	29,3	61,4	73,1	73,1
0,09	3,1	6,7	29,8	61,1	73,1	73,1
0,10	3,1	6,8	29,9	61,1	73,5	73,5
0,15	3,0	7,1	31,6	61,8	73,5	73,5
0,20	2,6	8,3	**32,3**	**62,0**	72,3	72,3
0,25	2,6	8,5	31,7	61,1	73,0	73,0
0,30	2,6	8,7	31,4	61,8	72,2	72,2
0,35	2,5	8,9	31,3	61,6	71,5	71,5
0,40	2,6	**9,4**	31,1	61,0	71,5	71,5
0,45	2,5	9,2	30,7	60,9	69,9	69,9
0,50	2,5	**9,4**	30,9	60,4	68,0	68,0
0,55	2,3	9,0	30,9	59,4	65,3	65,3
0,60	2,5	8,9	29,0	56,3	63,1	63,1
0,65	2,6	8,3	28,5	53,5	59,8	59,8
0,70	2,6	7,5	26,4	50,5	55,9	55,9
0,75	2,0	5,5	23,5	45,9	51,6	51,6
0,80	1,4	5,4	20,5	38,3	41,6	41,6
0,85	0,4	2,0	16,0	29,1	29,8	29,8
0,90	0,1	0,1	9,3	15,2	14,5	14,5
0,95	−0,6	−4,8	−1,9	−1,0	−2,1	−2,1

Anexo C

Resultados do processo de reconhecimento para o sistema treinado com locuções ruidosas, adaptado com ruído e testado com locuções corrompidas

Tabela C.1: *Taxa de acertos de palavras, em %, para um sistema adaptado com ruído de aeroporto, treinado e testado com locuções ruidosas*

α	SNR −5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
0,01	1,2	3,6	26,7	63,7	**77,8**	**77,8**
0,02	1,2	3,7	27,4	63,5	77,6	77,6
0,03	1,3	4,0	28,3	63,8	77,3	77,3
0,04	1,3	4,2	28,9	64,0	77,2	77,2
0,05	1,3	4,1	28,7	63,8	77,6	77,6
0,06	1,3	4,0	29,6	63,9	77,5	77,5
0,07	1,4	4,4	29,5	64,3	77,4	77,4
0,08	1,6	4,8	30,0	64,7	77,0	77,0
0,09	1,5	4,8	30,2	65,0	76,8	76,8
0,10	1,5	4,6	29,4	65,0	76,9	76,9
0,15	1,8	4,8	30,3	**66,2**	76,0	76,0
0,20	1,9	5,0	31,6	65,5	75,7	75,7
0,25	2,1	5,5	**32,1**	65,2	75,1	75,1
0,30	2,6	6,1	31,2	65,6	74,4	74,4
0,35	2,9	6,5	31,4	66,1	73,3	73,3
0,40	2,8	6,7	31,8	65,1	72,7	72,7
0,45	**3,1**	6,9	**32,1**	64,3	71,8	71,8
0,50	3,0	6,6	31,1	63,9	70,9	70,9
0,55	3,0	7,0	30,8	63,2	68,6	68,6
0,60	2,7	6,9	30,3	61,6	66,5	66,5
0,65	2,4	**7,3**	28,5	58,6	62,7	62,7
0,70	2,4	7,1	26,6	55,0	58,2	58,2
0,75	2,3	5,2	25,5	47,1	51,5	51,5
0,80	1,9	3,5	22,5	41,1	43,7	43,7
0,85	1,5	1,7	17,1	30,7	31,6	31,6
0,90	0,8	−0,2	8,3	16,8	14,5	14,5
0,95	0,2	−4,8	−2,3	−1,4	−3,9	−3,9

Tabela C.2: *Taxa de acertos de palavras, em %, para um sistema adaptado com ruído de balbúcio, treinado e testado com locuções ruidosas*

α	SNR -5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
$0,01$	$2,6$	$4,4$	$33,3$	$66,0$	**76,9**	**76,9**
$0,02$	$3,1$	$4,4$	$34,0$	$66,3$	$76,5$	$76,5$
$0,03$	$3,2$	$4,6$	$34,2$	$66,0$	$75,8$	$75,8$
$0,04$	$3,3$	$4,6$	$34,9$	$66,2$	$75,5$	$75,5$
$0,05$	$3,3$	$4,3$	$36,2$	$65,8$	$75,2$	$75,2$
$0,06$	$3,4$	$4,5$	$36,1$	$66,0$	$74,9$	$74,9$
$0,07$	$3,4$	$4,9$	$36,4$	$65,2$	$74,7$	$74,7$
$0,08$	$3,3$	$5,0$	$36,5$	$65,3$	$74,5$	$74,5$
$0,09$	$3,3$	$5,2$	$36,3$	$65,2$	$74,6$	$74,6$
$0,10$	**3,5**	**5,9**	$36,6$	$65,6$	$74,4$	$74,4$
$0,15$	**3,5**	**5,9**	$36,7$	**66,7**	$73,7$	$73,7$
$0,20$	**3,5**	$5,8$	**37,4**	$65,4$	$73,5$	$73,5$
$0,25$	$3,3$	$5,5$	$36,3$	$64,8$	$72,0$	$72,0$
$0,30$	$3,1$	$5,3$	$36,9$	$63,4$	$70,6$	$70,6$
$0,35$	$3,0$	$5,0$	$35,3$	$64,1$	$69,2$	$69,2$
$0,40$	$2,7$	$4,8$	$36,0$	$62,3$	$68,7$	$68,7$
$0,45$	$2,3$	$4,7$	$35,2$	$61,3$	$66,9$	$66,9$
$0,50$	$1,2$	$4,1$	$35,9$	$59,5$	$65,8$	$65,8$
$0,55$	$0,2$	$2,8$	$33,9$	$59,3$	$63,6$	$63,6$
$0,60$	$-0,3$	$2,1$	$33,4$	$57,0$	$60,5$	$60,5$
$0,65$	$-0,3$	$1,8$	$31,6$	$53,1$	$58,1$	$58,1$
$0,70$	$-0,6$	$1,0$	$28,2$	$48,9$	$53,9$	$53,9$
$0,75$	$-1,0$	$0,2$	$24,5$	$43,1$	$47,6$	$47,6$
$0,80$	$-1,5$	$-0,9$	$20,1$	$36,6$	$39,2$	$39,2$
$0,85$	$-2,1$	$-3,5$	$12,9$	$26,3$	$29,0$	$29,0$
$0,90$	$-3,3$	$-4,7$	$6,3$	$15,2$	$14,4$	$14,4$
$0,95$	$-4,5$	$-7,6$	$-0,9$	$-2,6$	$-5,5$	$-5,5$

Tabela C.3: *Taxa de acertos de palavras, em %, para um sistema adaptado com ruído de carro, treinado e testado com locuções ruidosas*

α	SNR −5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
0,01	0,3	5,8	32,7	65,8	76,3	76,3
0,02	0,3	6,3	33,5	65,8	76,5	76,5
0,03	0,3	6,6	33,7	65,9	**76,6**	**76,6**
0,04	0,3	6,5	33,5	66,7	**76,6**	**76,6**
0,05	0,3	6,3	33,3	67,0	76,0	76,0
0,06	0,4	6,1	34,0	66,8	75,3	75,3
0,07	0,4	6,4	33,9	66,8	75,3	75,3
0,08	0,5	6,3	34,8	67,1	75,0	75,0
0,09	0,5	6,2	34,9	67,4	74,7	74,7
0,10	0,5	5,9	34,8	67,4	74,0	74,0
0,15	0,6	6,5	35,2	**67,7**	73,4	73,4
0,20	0,6	6,1	35,8	66,0	72,4	72,4
0,25	0,8	6,5	36,6	65,9	71,9	71,9
0,30	0,9	**7,1**	35,8	64,4	71,0	71,0
0,35	1,0	6,8	**37,3**	64,3	70,6	70,6
0,40	0,9	6,4	36,8	63,8	69,0	69,0
0,45	**1,5**	7,0	37,0	62,7	66,4	66,4
0,50	1,4	6,6	36,5	61,9	64,2	64,2
0,55	1,4	6,4	35,6	59,2	61,3	61,3
0,60	1,3	5,5	33,4	56,0	56,9	56,9
0,65	1,0	5,0	30,4	50,8	51,0	51,0
0,70	0,5	3,7	27,0	45,3	44,0	44,0
0,75	−0,2	0,4	22,5	36,5	36,3	36,3
0,80	−1,1	−1,5	18,7	27,2	22,5	22,5
0,85	−2,5	−3,2	11,0	14,8	9,7	9,7
0,90	−3,8	−6,3	3,2	2,6	−2,1	−2,1
0,95	−8,7	−8,8	−6,8	−5,1	−7,3	−7,3

Tabela C.4: *Taxa de acertos de palavras, em %, para um sistema adaptado com ruído de exposição, treinado e testado com locuções ruidosas*

α	SNR -5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
$0,01$	$1,3$	$1,1$	$18,3$	$58,1$	$75,0$	$75,0$
$0,02$	$1,4$	$0,9$	$19,8$	$58,1$	**75,5**	**75,5**
$0,03$	$1,6$	$0,7$	$20,5$	$58,3$	$75,2$	$75,2$
$0,04$	$1,6$	$0,6$	$20,4$	$57,9$	$75,2$	$75,2$
$0,05$	$1,8$	$0,9$	$21,5$	$57,7$	$74,8$	$74,8$
$0,06$	$1,8$	$1,0$	$21,8$	$58,1$	$74,6$	$74,6$
$0,07$	$1,9$	$0,7$	$22,4$	$57,7$	$74,8$	$74,8$
$0,08$	$1,9$	$0,7$	$22,9$	$57,7$	$74,5$	$74,5$
$0,09$	$2,1$	$1,0$	**23,5**	$57,6$	$74,2$	$74,2$
$0,10$	$2,1$	$0,9$	$23,3$	$58,1$	$73,9$	$73,9$
$0,15$	$2,8$	$0,6$	$23,1$	**58,7**	$72,5$	$72,5$
$0,20$	$3,1$	$0,9$	$22,9$	$58,2$	$71,7$	$71,7$
$0,25$	**3,2**	$1,3$	$22,8$	$57,5$	$70,8$	$70,8$
$0,30$	**3,2**	$1,1$	$22,1$	$55,2$	$69,7$	$69,7$
$0,35$	$2,9$	$1,5$	$21,6$	$55,0$	$68,6$	$68,6$
$0,40$	$-0,3$	**1,8**	$21,6$	$54,4$	$66,2$	$66,2$
$0,45$	$-0,8$	$1,3$	$22,0$	$52,3$	$64,4$	$64,4$
$0,50$	$-0,9$	$1,2$	$21,0$	$49,7$	$61,4$	$61,4$
$0,55$	$-0,8$	$0,4$	$20,2$	$48,3$	$58,7$	$58,7$
$0,60$	$-1,2$	$0,7$	$17,1$	$45,9$	$53,8$	$53,8$
$0,65$	$-1,1$	$0,1$	$17,3$	$41,2$	$48,8$	$48,8$
$0,70$	$-1,1$	$-0,6$	$14,1$	$35,5$	$43,2$	$43,2$
$0,75$	$-1,1$	$-1,9$	$13,1$	$28,2$	$31,4$	$31,4$
$0,80$	$-2,7$	$-2,3$	$7,3$	$18,2$	$18,4$	$18,4$
$0,85$	$-3,1$	$-3,1$	$2,5$	$5,8$	$5,9$	$5,9$
$0,90$	$-4,6$	$-4,5$	$-3,5$	$-4,9$	$-7,2$	$-7,2$
$0,95$	$-6,5$	$-7,9$	$-7,6$	$-9,4$	$-8,9$	$-8,9$

Tabela C.5: *Taxa de acertos de palavras, em %, para um sistema adaptado com ruído de restaurante, treinado e testado com locuções ruidosas*

α	SNR −5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
0,01	0,4	13,8	4,7	35,9	69,4	69,4
0,02	0,4	13,9	5,3	36,9	69,2	69,2
0,03	0,4	13,8	5,9	37,0	**69,7**	**69,7**
0,04	0,4	13,9	5,6	37,0	69,6	69,6
0,05	0,5	14,0	6,6	37,9	69,3	69,3
0,06	0,5	14,0	7,1	38,1	68,9	68,9
0,07	0,5	14,2	7,4	38,8	68,7	68,7
0,08	0,5	14,5	7,7	38,8	68,8	68,8
0,09	0,5	15,1	7,8	39,1	69,0	69,0
0,10	**1,3**	14,7	7,6	39,2	68,7	68,7
0,15	0,5	14,7	8,3	40,0	67,7	67,7
0,20	0,7	15,2	7,9	41,1	67,4	67,4
0,25	0,9	**15,4**	8,8	40,5	66,6	66,6
0,30	0,9	14,8	8,8	41,4	67,4	67,4
0,35	0,8	14,6	**9,1**	**41,6**	66,6	66,6
0,40	1,1	14,5	8,0	40,7	64,7	64,7
0,45	1,2	13,9	8,1	40,0	63,7	63,7
0,50	1,2	12,0	6,9	38,4	61,6	61,6
0,55	1,0	12,2	5,8	36,6	60,4	60,4
0,60	0,9	11,6	6,5	34,7	57,9	57,9
0,65	0,5	10,0	4,8	33,6	54,8	54,8
0,70	−0,5	7,9	2,4	31,6	50,4	50,4
0,75	−1,0	6,5	1,4	27,9	46,7	46,7
0,80	−1,5	4,0	0,3	23,4	38,7	38,7
0,85	−2,6	2,1	−0,8	17,9	31,3	31,3
0,90	−2,7	−2,7	−5,3	8,1	17,8	17,8
0,95	−3,6	−9,0	−9,0	−1,8	−1,0	−1,0

Tabela C.6: *Taxa de acertos de palavras, em %, para um sistema adaptado com ruído de rua, treinado e testado com locuções ruidosas*

α	SNR -5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
$0,01$	$3,4$	$16,3$	$55,9$	**73,6**	$76,8$	$76,8$
$0,02$	$3,5$	$17,0$	$56,3$	**73,6**	**77,1**	**77,1**
$0,03$	$3,5$	$16,6$	$56,9$	$72,9$	$76,8$	$76,8$
$0,04$	$3,5$	$16,6$	$56,3$	$72,7$	$76,0$	$76,0$
$0,05$	$3,4$	$16,7$	$56,6$	$72,9$	$75,2$	$75,2$
$0,06$	**3,6**	$17,4$	$56,3$	$72,4$	$74,8$	$74,8$
$0,07$	$3,5$	$17,7$	**57,1**	$72,1$	$74,9$	$74,9$
$0,08$	$3,3$	$18,5$	$56,9$	$71,7$	$74,5$	$74,5$
$0,09$	$3,2$	$18,2$	$56,3$	$71,7$	$73,9$	$73,9$
$0,10$	$3,3$	$18,1$	$56,8$	$72,0$	$73,3$	$73,3$
$0,15$	$3,1$	**18,6**	$56,0$	$71,1$	$72,6$	$72,6$
$0,20$	$2,7$	**18,6**	$54,0$	$70,0$	$71,9$	$71,9$
$0,25$	$2,6$	$17,4$	$53,1$	$68,4$	$70,4$	$70,4$
$0,30$	$-0,4$	$16,3$	$52,2$	$66,6$	$68,4$	$68,4$
$0,35$	$-3,0$	$15,3$	$50,1$	$65,7$	$67,3$	$67,3$
$0,40$	$-3,8$	$13,3$	$49,0$	$64,6$	$65,5$	$65,5$
$0,45$	$-4,8$	$13,1$	$47,3$	$63,2$	$64,8$	$64,8$
$0,50$	$-6,4$	$12,1$	$44,9$	$61,2$	$63,9$	$63,9$
$0,55$	$-7,7$	$9,8$	$40,9$	$59,8$	$61,4$	$61,4$
$0,60$	$-8,5$	$8,3$	$38,2$	$56,3$	$57,6$	$57,6$
$0,65$	$-9,2$	$7,3$	$35,0$	$51,5$	$55,0$	$55,0$
$0,70$	$-9,7$	$3,6$	$31,8$	$46,7$	$51,2$	$51,2$
$0,75$	$-11,3$	$1,5$	$26,6$	$41,8$	$46,1$	$46,1$
$0,80$	$-12,0$	$-0,3$	$21,8$	$33,5$	$37,3$	$37,3$
$0,85$	-13.7	$-3,6$	$15,5$	$24,7$	$25,4$	$25,4$
$0,90$	-15.6	$-7,4$	$6,9$	$12,5$	$11,8$	$11,8$
$0,95$	-17.1	$-12,3$	$-5,3$	$-3,2$	$-3,2$	$-3,2$

Tabela C.7: *Taxa de acertos de palavras, em %, para um sistema adaptado com ruído de metrô, treinado e testado com locuções ruidosas*

α	SNR −5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
0,01	4,7	0,8	19,9	60,2	73,1	73,1
0,02	4,5	0,5	20,7	60,3	72,5	72,5
0,03	4,5	1,0	21,8	60,4	72,1	72,1
0,04	4,6	1,1	22,1	61,2	73,0	73,0
0,05	5,0	1,3	22,6	**61,8**	72,7	72,7
0,06	5,2	1,5	22,4	61,7	72,7	72,7
0,07	5,4	1,8	23,0	61,7	73,2	73,2
0,08	5,5	1,7	22,9	61,1	**73,4**	**73,4**
0,09	5,4	1,3	23,5	61,7	72,8	72,8
0,10	6,0	1,9	23,6	61,6	73,1	73,1
0,15	6,3	2,0	24,6	61,1	71,2	71,2
0,20	6,3	2,1	24,0	60,8	71,6	71,6
0,25	6,4	3,2	25,4	61,5	70,2	70,2
0,30	6,7	3,2	**25,8**	59,2	69,5	69,5
0,35	6,8	2,6	25,2	59,2	68,2	68,2
0,40	6,6	**3,4**	25,4	57,8	67,3	67,3
0,45	**7,0**	3,2	25,2	57,0	65,6	65,6
0,50	**7,0**	2,9	25,1	55,9	63,4	63,4
0,55	6,7	2,7	23,8	54,6	61,5	61,5
0,60	6,5	2,3	23,2	50,3	57,4	57,4
0,65	5,8	2,6	21,3	46,6	51,6	51,6
0,70	5,3	1,1	19,6	44,4	47,3	47,3
0,75	4,9	0,7	16,5	37,3	38,8	38,8
0,80	2,8	−0,5	12,2	28,6	31,3	31,3
0,85	0,4	−3,4	7,4	20,3	15,8	15,8
0,90	−3,1	−5,6	2,2	5,2	−0,7	−0,7
0,95	−6,6	−8,4	−6,1	−6,5	−6,2	−6,2

Tabela C.8: *Taxa de acertos de palavras, em %, para um sistema adaptado com ruído de trem, treinado e testado com locuções ruidosas*

α	SNR -5 dB	SNR 0 dB	SNR 5 dB	SNR 10 dB	SNR 15 dB	SNR 20 dB
$0,01$	$2,2$	$8,1$	$35,1$	$68,9$	**78,1**	**78,1**
$0,02$	$2,2$	$8,4$	$35,4$	$69,7$	$77,9$	$77,9$
$0,03$	$2,2$	$8,2$	$35,5$	$69,7$	$77,4$	$77,4$
$0,04$	$2,3$	$8,9$	$36,5$	**70,0**	$77,2$	$77,2$
$0,05$	$2,3$	$8,9$	$37,2$	**70,0**	$76,4$	$76,4$
$0,06$	$2,4$	$9,1$	$37,2$	**70,0**	$75,8$	$75,8$
$0,07$	$2,3$	$9,4$	$37,4$	$69,8$	$76,1$	$76,1$
$0,08$	$2,5$	$9,7$	$37,9$	$69,6$	$76,3$	$76,3$
$0,09$	$2,5$	$10,1$	$38,5$	$69,5$	$76,3$	$76,3$
$0,10$	$2,6$	$10,3$	$38,9$	$69,5$	$75,9$	$75,9$
$0,15$	$2,9$	$10,9$	$38,6$	$69,0$	$74,5$	$74,5$
$0,20$	$2,9$	$11,3$	$39,6$	$69,4$	$73,5$	$73,5$
$0,25$	**3,2**	$11,9$	$38,9$	$68,5$	$72,8$	$72,8$
$0,30$	$3,1$	$12,1$	$38,9$	$67,9$	$72,6$	$72,6$
$0,35$	$2,9$	$12,2$	$40,0$	$66,9$	$71,2$	$71,2$
$0,40$	$2,8$	$12,4$	$40,0$	$67,7$	$69,8$	$69,8$
$0,45$	$3,0$	**12,6**	**40,7**	$66,5$	$67,9$	$67,9$
$0,50$	$3,0$	$12,4$	$40,0$	$64,4$	$64,9$	$64,9$
$0,55$	$2,9$	$11,6$	$38,9$	$63,5$	$63,5$	$63,5$
$0,60$	$2,7$	$11,8$	$38,6$	$59,8$	$60,1$	$60,1$
$0,65$	$2,6$	$11,1$	$35,3$	$54,6$	$55,5$	$55,5$
$0,70$	$2,6$	$10,6$	$33,0$	$51,0$	$49,4$	$49,4$
$0,75$	$2,1$	$9,7$	$31,2$	$44,8$	$45,0$	$45,0$
$0,80$	$1,3$	$9,4$	$25,9$	$39,5$	$35,3$	$35,3$
$0,85$	$0,2$	$6,6$	$18,7$	$29,7$	$25,2$	$25,2$
$0,90$	$-0,6$	$2,5$	$9,1$	$14,4$	$7,4$	$7,4$
$0,95$	$-2,2$	$-3,0$	$-2,2$	$-4,4$	$-6,7$	$-6,7$

Referências Bibliográficas

[1] FERREIRA, P. C.; SANTOS, M. R. **Globalization and the Industrial Revolution**. 34o. Encontro Brasileiro de Econometria, 2012.

[2] RABINER, L. R.; PARK, F. **Applications of Speech Recognition in the Area of Telecommunications**. Procceedings in IEEE Workshop on Automatic Speech Recognition and Understanding, 1997.

[3] WAIBEL, A.; LEE, K-F. **Readings in Speech Recognition**. Book. Editor: Michael B. Morgan. 1990.

[4] KUMAR, K.; AGGARWAL, R. K. **Hindi speech recognition system using HTK**. International Journal of Computing and Business Research, Vol.2, 2011.

[5] MISHRA, A. N.; CHANDRA, M.; BUSWAS, A.; SHARAN, S. N. **Robust features for connected Hindi digits recognition**. International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 4, No. 2, 2011.

[6] CROVATO, C. D. P. **Classificação de sinais de voz utilizando a Transformada Wavelet Packet e Redes Neurais Artificiais**. Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal do Rio Grande do Sul. Porto Alegre, 2004.

[7] LOIZOU, P. C. **Speech enhancement theory and pratice**. CRC Press. Nova Iorque. Estados Unidos. 2013.

[8] UMARANI, S. D.; WAHIDABANU, R. S. D.; RAVIRAM, P. **Speaker invariant and noise robust speech recognition using enhanced auditory and VTL based features**. International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), Vol. 199, p. 77-82, 2013.

[9] PEGORARO, T. F. **Algoritmos robustos de reconhecimento de voz aplicados a verificação de locutor**. Dissertação (Mestrado em Engenharia Elétrica) - Universidade Estadual de Campinas. São Paulo, 2000.

[10] ATTIAS,H.; DENG, L.; ACERO, A.; PLATT, J. **A new method for speech denoising and robust speech recognition using probabilistic models for clean speech and for noise**. Eurospeech, p. 1903-1906, 2001.

[11] PLAPOUS, C.; MARRO, C.; MAUUARY, L.; SCALART, P. **A two-step noise reduction technique**. IEEE International Conference on Acoustics Speech an Signal Processing (ICASSP), Vol. 1, p. 289-292, 2004.

[12] GALES, M. J. F. **Model-based techniques for noise robust speech recognition**. Dissertação (Doutorado) - Universidade de Cambridge, 1995.

[13] NIMJE, K.; SHANDILYA, M. **Automatic isolated digit recognition system: an approach using HMM**. Journal of Scientific and Industrial Research, Vol. 70, p. 270-272, 2011.

[14] GELBART, D.; MORGAN, N. **Double the Trouble: Handling Noise and Reverberation in Far-Field Automatic Speech Recognition**. 7th International Conference on Spoken Language Processing (ICSLP), 2002.

[15] GALES, M. J. F.; YOUNG, S. J. **Robust continuous speech recognition using parallel model combination**. IEEE Transactions on Audio, Speech, and Language Processing, Vol 4, p. 352-359, 1996.

[16] HACHKAR, Z.; MOUNIR, B.; FARCHI, A.; ABBADI, J. **Comparison of MFCC and PLP Parameterization in pattern recognition of Arabic Alphabet Speech**. Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition, Vol.2, No.3, 2011.

[17] KERMORVANT, C. **A comparison of noise reduction techniques for robust speeh recognition**. IDIAP Research Report, 1999.

[18] MARTIN, R.; NAGATHIL, A. **Cepstral modulation ratio regression (CMRARE) parameters for audio signal analysis and classification**. IEEE International Conference on Acoustics Speech an Signal Processing (ICASSP), p. 321-324, 2009.

[19] BOLL, S. F. **Suppression of Acoustic Noise in Speech Using Spectral Subtraction**. IEEE Transactions on Accoustics, Speech and Signal Processing, Vol. 27, No. 2, p. 113-120, 1979.

[20] MAYER, A. G.; KING, R. W.; RATHMELL, J. G. **A Comparison of Noise Reduction Techniques for Speech Recognition in Telecommunications Environments**. The Institution of Engineers Australia Communications Conference, 1992.

[21] KAMATH, S. D.; LOIZOU, P. C. **A multi-band spectral subtraction method for enhancing speech corrupted by colored noise**. IEEE International Conference on Acoustics Speech an Signal Processing (ICASSP), Vol. 4, 2002.

[22] KIM, H. K.; ROSE, C. R. **Cepstrum Domain Model Combination Based on Decomposition of Speech and Noise Using MMSE-LSA for ASR in Noisy Environments**. IEEE Transactions on Audio, Speech, and Language Processing, Vol 17, pp. 704-713, 2009.

[23] GERKMANN, T.; HENDRIKS, R. **Improved MMSE-based noise PSD tracking using temporal cepstrum smoothing**. IEEE International Conference on Acoustics Speech an Signal Processing (ICASSP), p. 105-108, 2012.

[24] ILLINA, I.; MOSTEFA, D. **Structural Maximum a Posteriori Adaptation for Mixture Stochastic Trajectory Framework**. ITEW on Adaptation Methods for Speech Recognition, p. 147-150, 2001.

[25] MATROUF, D.; BELLOT, O.; NOCERA, P.; LINARES, G.; BONASTRE, J. **A Posteriori and a Priori Transformations for Speaker Adaptation in Large Vocabulary Speech Recognition Systems**. Eurospeech, 2001.

[26] GRIMM, M.; KROSCHEL, K. **Discrete-Mixture HMMs-based Approach for Noisy Speech Recognition**. Robust Speech Recognition and Understanding, I-Tech Education and Publishing. Viena. Áustria. 2007.

[27] GHANBARI, Y.; KARAMI, M. R. **Spectral subtraction in the Wavelet domain for speech enhancement**. International Conference on Intelligent Knowledge Systems, Vol. 1, p. 26-29, 2004.

[28] GALES, M. J. F. **Predictive model-based compensation schemes for robust speech recognition**. Journal Speech Communication, Vol. 25, Issue 1-3, p. 49-74, 1998.

[29] KALINLI, O.; SELTZER, M. L.; DROPPO, J.; ACERO, A. **Noise Adaptative Training for robust automatic speech recognition**. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 8, p. 1889-1901, 2010.

[30] LU, L.; GHOSHAL, A.; RENALS, S. **Regularized Subspace Gaussian Mixture Models for speech recognition**. IEEE Signal Processing Letters, Vol. 18, No. 7, p. 419-422, 2011.

[31] BURGET, L.; SCHAWARZ, P.; AGARWAL, M.; AKYAZI, P.; FENG, K.; GHOSHAL, A.; GLEMBEK, O.; GOEL, N.; KARAFIÁ, M.; POVEY, D.; RASTRO, A.; ROSE, R. C.; THOMAS, S. **Multilingual acoustic modeling for speech recognition based on subspace Gaussian Mixture Models**. IEEE International Conference on Acoustics Speech an Signal Processing (ICASSP), p. 4334-4337, 2011.

[32] SALAVEDRA, J. M.; HERNANDO, J. **Third-order cumulant-based Wiener filtering algorithm applied to robust Speech Recognition**. 8th European Signal Processing Conference, p. 1603-1606, 1996.

[33] QI, Z. **Real-Time Adaptive Noise Cancellation for Automatic Speech Recognition in a car environment**. Dissertação (Doutorado em Engenharia da Computação) - Massey University School of Engineering an Advanced Technology. Auckland, New Zealand, 2008.

[34] RAJ, B.; GOUVÊA, E. B.; MORENO, P. J.; STERN, R. S. **Cepstral compensation by polynomial approximation for environment-independent Speech Recognition**. 4th International Conference on Spoken Language Processing (ICSLP), Vol.4, p.2340-2343, 1996.

[35] CLAES, T.; XIE, F.; COMPERNOLLE, D. V. **Spectral estimation and normalisation for robust speech recognition**. 4th International Conference on Spoken Language Processing (ICSLP), Vol. 4, p.1997-2000, 1996.

[36] LEBART, K.; BOUCHER, J. M. **A New Method Based on Spectral Subtraction for Speech Dereverberation**. Acta Acustica. Vol. 87, p. 359-366, 2001.

[37] KOTNIK, B.; VLAJ, D.; KACIC, Z.; HORVAT, B. **Robust MFCC feature extraction algorithm using efficient additive and convolutional noise reduction procedures**. 7th International Conference on Spoken Language Processing (ICSLP), p.445-448, 2002.

[38] SHAFRAN, I.; ROSE, R. **Robust speech detection and segmentationfor real-time ASR applications**. IEEE International Conference on Acoustics Speech an Signal Processing (ICASSP), Vol. 1, p. 432-435, 2003.

[39] YEUNG, S. A.; SIU, M. **Improved Performance of Aurora 4 using HTK and Unsupervised MLLR Adaptation**. Interspeech, 8th Inter-

national Conference on pokn Language Processing, Jeju Island, Korea, p. 161-164, 2004.

[40] COHEN, I. **Relaxed Statistical Model for Speech Enhancement and A Priori SNR Estimation**. IEEE Transactions on Audio, Speech, and Language Processing, Vol 13, p. 870-881, 2005.

[41] SHINGH, R.; RAO, P. **Spectral Subtraction Speech Enhancement with RASTA Filtering**. National Conference on Communications,2007.

[42] REN, Y.; JOHNSON, M. T. **An improved SNR estimator for speech enhancement**. IEEE International Conference on Acoustics Speech an Signal Processing (ICASSP), p. 4901-4904, 2008.

[43] RAO C. V. R.; MURTHY, M. B. R.; RAO, K. S. **Noise reduction using mel-scale spectral subtraction with perceptually defined subtraction parameters-A new scheme**. Signal and Image Processing International Journal, Vol. 2, No. 1, 2011.

[44] CHATTERJEE, S.; KLEIJIN, W. B. **Auditory Model Based Design and Optimization Of Feature Vectors for Automatic Speech Recognition**. IEEE Transactions on Audio, Speech, and Language Processing, Vol 19, p. 1813-1825, 2011.

[45] JUANG, B. H. **Speech recognition in adverse environments**. Computer Speech and Language, Vol. 5, p. 275-294, 1991.

[46] LIPPMAN, R.; MARTIN, E.; PAUL, D **Multi-style training for robust isolated-word speech recognition**. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, vol. 12, pp. 705- 708, 1987.

[47] REYNOLDS, D. A.; QUARTIERI, T. F.; DUNN, R. B. **Speaker verification using adapted Gaussian mixture models**. Digital Signal Processing, 10, pp.19-41, 2000.

[48] VALÉRIO, T. A. F. **Treinamento multi-estilo e adaptação de modelos via MAP para reconhecimento de fala em ambientes ruidosos**. Dissertação de Mestrado. Inatel. Santa Rita do Sapucaí, 2011.

[49] BUERA, L.; LlEIDA, E.; MIGUEL, A.; ORTEGA, A.; SAZ, O. **Cepstral Vector Normalization Based on Stereo Data for Robust Speech Recognition**. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, pp. 1098-1113, 2007.

[50] YNOGUTI, C. A. **Reconhecimento de Fala Contínua usando Modelos Ocultos de Markov**. Tese (Doutorado em Engenharia Elétrica) - Universidade Federal de Campinas. Campinas, 1999.

[51] ALCAIM, A.; SOLEWICZ, J. A.; MORAES, J. A. **Freqüência de ocorrência dos fones e lista de frases foneticamente balanceadas no português falado no Rio de Janeiro**. Revista da Sociedade Brasileira de Telecomunicações, Vol 7(1), pp. 23-41, 1992.

[52] HIRSCH, H. G.; PEARCE, D. **The Aurora experimental framework for the evaluation of speech recognition systems under noisy conditions**. In Proc. ASR-2000, pp 181-188,September 2000.

[53] NEY, H. **The use of a one-stage dynamic programming algorithm for conected word recognition**. IEEE Transactions on Acoustics, Speech and Signal Processing, ASSSP-32(2), 1984.

[54] EVERMANN, G. **Minimum Word Error Rate Decoding**. Dissertação. University of Cambrigde, 1999.

[55] COWAN, Mc. I.; MOORE, D.; DINES, J.; GATICA-PEREZ, D.; FLYNN, M.; WELLNER, P. BOURLARD, H. **On the use of information retrieval measures for speech recognition evaluation**. IDIAP Research Report, 2005.

[56] GHAI, W.; SINGH, N. **Literature review on Automatic Speech Recognition**. International Journal of Computer Applications, Vol. 41, No. 8, 2012.

[57] MANGU, L.; BRILL, E., STOLCKE, A. **Finding consensus in speech recognition: word error minimization and other applications of confusion networks**. Computer Speech and Language, Vol. 14, No. 4, p. 373-400, 2000.

[58] STANLEY, F. C.; BEEFERMAN, D.; ROSENFIELD, R. **Evaluation metrics for language models**. In Proccedings of the DARPA Broadcast News Transcription an Ubderstanding Workshop, p. 275-280, 1998.

[59] ACHNER, K.; WAIBEL, A. **Minimizing word error rate in textual summaries of spoken language**. In proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics(NAACL), Seattle, WA, p. 186-193, 2000.

[60] WANG, Y.; ACERO, A.; CHELBA, C. **Is word error rate a good indicator for spoken language understanding accuracy**. IEEE Workshop

on Automatic Speech Recognition and Understanding (ASRU), p. 577-582, 2003.

[61] ANANTHAKRISHNAN, S.; NARAYANAN, S. **Unsupervised adaptation of categorical prosody models for prosody labeling and speech recognition**. IEEE Transactions on Audio, Speech and Language Processing, Vol. 17, No. 1, 2009.

[62] **sctk-1.3 - Speech Recognition Scoring Toolkit SCTK Version 2.4 (Includes the SCLITE Scoring program)**. ftp://jaguar.ncsl.nist.gov/pub/sctk-1.3.tgz (Janeiro/2012).

[63] CHIOVATO, A. G. **Avaliação da relação entre qualidade perceptual da fala e taxa de acerto de sistemas de reconhecimento de fala em ambientes ruidosos**. Dissertação de Mestrado. Inatel. Santa Rita do Sapucaí, 2005.