DAVID DANIEL E SILVA

CONTRIBUIÇÕES AO RECONHECIMENTO AUTOMÁTICO DE FALA ROBUSTO

FLORIANÓPOLIS
Agosto de 2010

Catalogação na fonte pela Biblioteca Universitária da
Universidade Federal de Santa Catarina

S586c Silva, David Daniel e
Contribuições ao reconhecimento automático de fala robusto [tese] / David Daniel e Silva ; orientador, Marcelo Ricardo Stemmer. - Florianópolis, SC, 2010.
270 p.: il., grafs., tabs., equações.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia de Automação e Sistemas.

Inclui referências

1. Engenharia de sistemas. 2. Automação. 3. Base de dados. Reconhecimento automático de fala.
I. Stemmer, Marcelo Ricardo. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia de Automação e Sistemas. III. Título.

CDU 621.3-231.2(021)

UNIVERSIDADE FEDERAL DE SANTA CATARINA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE AUTOMAÇÃO E SISTEMAS

CONTRIBUIÇÕES AO RECONHECIMENTO AUTOMÁTICO DE FALA ROBUSTO

Tese de doutorado submetida à
Universidade Federal de Santa Catarina
como requisito para a obtenção do título de
Doutor em Engenharia de Automação e Sistemas

DAVID DANIEL E SILVA

Florianópolis, 09 de agosto de 2010

CONTRIBUIÇÕES AO RECONHECIMENTO AUTOMÁTICO DE FALA ROBUSTO

David Daniel e Silva

Esta tese foi julgada adequada e dentro dos requisitos para Doutorado em Engenharia de Automação e Sistemas, Área de Concentração *Automação e Sistemas Mecatrônicos*, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia de Automação e Sistemas da Universidade Federal de Santa Catarina.

Prof. Marcelo Ricardo Stemmer, Dr.
Orientador

Prof. José Eduardo Ribeiro Cury, Dr.
Coordenador do Programa de Pós-Graduação em Engenharia de Automação e Sistemas

Banca Examinadora:

Prof. Marcelo Ricardo Stemmer, Dr., DAS/UFSC

Prof. Carlos Alberto Ynoguti, Dr., DEE/INATEL

Prof. Carlos Barros Montez, Dr., DAS/UFSC

Prof. Eduardo Camponogara, Dr., DAS/UFSC

Prof. Fábio Violaro, Dr., FEEC/DECOM/UNICAMP

Agradecimentos

Aos professores, Marcelo Ricardo Stemmer, Dr., e Carlos Alberto Ynoguti, Dr., não tenho palavras para agradecer o que a mim proporcionaram. Aqui deixo relatado, no entanto, o meu sincero agradecimento, pelo seu apoio e atenção. Sabemos todos que, pelo pioneirismo e falta de um grupo formado, o trabalho foi quase que hercúleo, mas acima de tudo estava a paciência e compreensão de vocês. Assim sendo, muito obrigado!!

Ao Professor Fábio Violaro, Dr., Unicamp, obrigado pelas ricas discussões feitas sobre o tema reconhecimento automático de fala.

Ao Programa de Pós-Graduação em Engenharia de Automação e Sistemas (PPGEAS), pelo bom nível do programa, pela estrutura física e humana, e, ao Prof. Eugênio Castelan, por sempre estar disposto a colaborar – muito obrigado!

É preciso deixar registrado que este trabalho jamais teria sido realizado sem qualquer uma das partes importantes que o compõem (aluno, orientadores, instituições de apoio, etc). Por isso, é preciso fazer um profundo agradecimento para a UDESC (Universidade do Estado de Santa Catarina), que custeou os estudos e para a qual só posso dizer – muito obrigado! Sem dúvida nenhuma sem o referido recurso eu não teria chegado aqui, sequer começado.

Ao INATEL (Instituto Nacional de Telecomunicações), que ajudou muito por meio do professor Carlos Alberto Ynoguti, com a sua experiência, recursos técnicos e a paciência. Esse apoio foi, sem dúvida, decisivo para o trabalho tomar o bom rumo. Muito obrigado!!!

À UFSC (Universidade Federal de Santa Catarina), instituição que acabou sendo a produtora da minha formação, desde a graduação e mestrado e agora o almejado doutorado, muito obrigado!

Agradeço às pessoas que ajudaram a abrandar os vários dias de sofrimento para a realização deste trabalho. Em especial à minha mãe Lindaura que ficou, sempre que pôde, cuidando de minha filha Bianca, que tinha apenas 3 anos quando comecei o doutorado, em agosto de 2005 e, desta forma, com o apoio desta amável mãe, ficou um pouco mais tranquilo estudar e seguir em frente.

À Jussara, que entrou na minha vida e na de minha filha em maio de 2006, e que com o nosso encontro, quando possível, suavizou momentos de angústia, com apoio amigo e de papel de mãe para a Bianca. Tenho certeza de que, se pudesse, ajudaria ainda mais, portanto, muito obrigado, minha querida, principalmente pelo apoio dado nesses últimos meses de, ainda, tanto trabalho. Obrigado também por suportar e compreender os vários dias e noites, inclusive da grande maioria dos feriados, em que fiquei verdadeiramente plantado em estudo nessa longa jornada.

Ao meu sobrinho Emanuel, que me auxiliou em vários momentos com o Linux, e com o HTK, muito obrigado! Ao Anderson, outro meu sobrinho, que acabou me auxiliando na edição da base de dados Lombard, também meus agradecimentos!

Ao Saulo, doutorando do DAS, que me ajudou nos últimos meses em alguns problemas importantes com o Linux – muito obrigado!

Ao senhor Nickolay V. Shmyrev, da lista de discussão sobre o HTK, muito obrigado pelas valiosas sugestões que acabaram determinando o funcionamento do HTK e finalmente o entendimento do complicado livro “HTKBook”.

Ao Carlos Montez, amigo de sempre, por ter ajudado com alguns conhecimentos de programação em C/C++, e livros sobre Markov/HMM, muito obrigado, amigo!!

Ao Rodrigo e Andre Iwersen, obrigado pelas vezes que participaram de discussões técnicas importantes!

Aos meus sobrinhos João Luiz, Íris, Lilian, Marcelinho, Fábio, Josué, Filipe, Emanuel e Anderson, obrigado pelas horas de alegria juntos. Sigam em frente sempre, pois só no final, depois de enfrentar as diversas tempestades do tempo, é que poderemos aprender – sem caminho não vamos a lugar algum.

Ao Paulo Conejo, amigo de tempo, por se dispor a avaliar as soluções matemáticas que foram desenvolvidas – obrigado!

Ao Jones Corso, também amigo de tempo, obrigado pela indicação da profissional para revisão do texto da tese, professora Cleusa Schneider, de Ijuí/RS.

Ao meu primo Cristiano Reitz e ao Evandro Heeck, da Dominik Metal Center, muito obrigado pela colaboração nas gravações do ruído de corte de metal.

Muito obrigado ao engenheiro José Roberto Rossi Filho e Fábio Pedro Dorigon (eletrotécnico), que permitiram e ajudaram nas

gravações de dentro e fora do túnel que liga o centro de Florianópolis à beira-mar sul.

Ao meu amigo Mário Gonzáles, doutorando da Engenharia Mecânica, pelas várias ajudas e pelos vários momentos em que acabamos nos encontrando e que sempre resultaram em boas reflexões.

Muito obrigado ao pessoal que colaborou para as gravações para a construção da base Lombard: *Alexandre Montenegro; Ana Paula Bastos; Bruna Vieira de Paula; Bruna Vieira de Paula; Bruno George Moraes; Camila Valerim; Cecilia Estela Giuffra Palomino; Christian Emanuel Mapurunga Silvano; Claudia Sampaio Ferreira; Daniel Habib; Dayana Spagnuelo; Eder Augusto Penharbel; Edileusa (DAS); Eduardo Dantas; Elisabete (DAS); Emanuel Carlos de Souza; Emerson Pereira Raposo;Fernanda Araújo Barrichello; Fernando Costa Bertoldi; Grazielle Susan Xavier; Heloisa Simon; Igor W. Khairalla; Íris Carolini; Izabel Gomes; Joice Andrea Balboa; Kayron Campos Beviláqua; Leide Sayuri Ogasawara; Marcelo Ricardo Stemmer; Maria de Fátima; Mauricio Edgar Stivanello; Mônica Tremarin; Polyana Liz; Patricia L. Barrufi Pinheiro; Priscilla Monique Silva Castanharo; Rafael Vasconcelos; Ricardo Hahn Barbosa de Souza; Rodrigo Lange; Saulo Popov Zambiasi; Shana Boff; Sigmar de Lima; Tanisia Foletto; Thiago de Souza; Underlea (DAS); Vanessa Dilda; Vilson Heck Junior; Vinicius da Rosa; Willian Fonseca*. Ah! Eu também gravei!

Peço desculpas se esqueci de alguém, mas a todos que ajudaram de alguma maneira para a realização deste trabalho, muito obrigado!!

Dedicatória

Querida Bianca:

Filha, você foi a grande motivação para eu começar, recomeçar e seguir em frente; fazer o doutorado e olhar para o futuro como um guerreiro; superar os problemas (que não foram poucos) e fazer com que os momentos bons fossem aproveitados e lembrados para sempre.

Nunca vou esquecer que você, em um dia que eu estava escrevendo um artigo, veio me pedir três folhas em branco e o grampeador... Na sequência, após algum tempo, veio me mostrar o que escreveu. Achei lindo e fiquei feliz por ter presenciado você, com apenas seis anos e meio, escrever seu primeiro livrinho, com início, meio e fim. A promessa de editá-lo vai ser cumprida! Muito obrigado, filha, eu te amo!!!

Obrigado tia Luizita, que soube cuidar dos últimos momentos de seu irmão, meu pai, buscando forças junto comigo, desde meados de 2005 até fevereiro de 2006, na difícil tarefa de ajudá-lo antes da sua partida para o outro plano. Que Deus o ilumine no plano divino, onde quer que esteja...

Minha mãe Lindaura, irmãos e sobrinhos:

A luz nos apresenta os cenários dos diversos caminhos, à noite todos os cenários somem, porém, independentemente disso, os personagens ainda continuam. Por isso, quando o sol se põe e a noite chega, é preciso buscar a luz interior - para passar melhor a noite - e nos dirigir melhor quando houver dia. A sabedoria é um bom ingrediente para conseguir ver com mais luz os dias anoitecerem e voltarem a ficar iluminados.

Obrigado por tudo, amo vocês!

David, em 18/01/2010

RESUMO

Resumo da tese apresentada à UFSC como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Automação e Sistemas.

CONTRIBUIÇÕES AO RECONHECIMENTO AUTOMÁTICO DE FALA ROBUSTO

David Daniel e Silva

Agosto/2010

Orientador: Professor Marcelo Ricardo Stemmer, Dr. Ing.
Área de Concentração: Controle, Automação e Sistemas.
Palavras-chave: base de dados Lombard, base de dados robusta, função logística, Pesq, reconhecimento automático de fala.

Número de páginas: 270

Reconhecimento Automático de Fala (RAF) é uma área fascinante e complexa. Durante décadas a demanda de pesquisas baseava-se em RAF para vocabulário não muito extenso, com técnicas que precisavam de alto desempenho computacional para processar dados produzidos em ambientes silenciosos de laboratórios. Dos meados da década de 80 para a frente, a tecnologia de processamento de voz avançou, com a utilização dos modelos ocultos de Markov (HMMs) e com o alto avanço de técnicas de programação e de processamento computacionais, conseguindo taxas de acerto, em ambientes silenciosos, próximas de 100%. Com a finalidade de colocar sistemas de RAF para funcionar na vida real, há alguns anos pesquisas intensas foram e continuam sendo feitas sobre reconhecimento de fala robusto. Por isso, aplicações como DSR (*Distributed Speech Recognition*), entre outras, surgiram no mercado. Para obter uma performance similar ao do ouvido humano em ambientes ruidosos, no entanto, sistemas desse tipo ainda são o foco de muitas pesquisas. Assim, este trabalho faz um estudo sobre sistemas de reconhecimento automático de fala robusto, objetivando a análise e comportamento de quatro tipos de ruídos (corte de metal, automóveis

em frente a um túnel, automóveis dentro do túnel e multidão de crianças), gravados em ambientes diferentes, para a avaliação e construção de bases de dados ruidosas. Desta forma, são desenvolvidas duas bases de dados, deixando como contribuição principal a metodologia para sua construção e o processo de análise e avaliação dos dados envolvidos na sua construção. Além disso, é apresentado um desenvolvimento matemático de um algoritmo que é a solução numérica para uma função logística de três parâmetros de difícil solução, empregada para modelar o comportamento dos sistemas WI007 e WI008 usados aqui. Um método de ajuste inicial logístico (Mail) das curvas Pesq vs. TA para a avaliação do comportamento do sistema de RAF adotado, também é uma das contribuições deste trabalho. Como um dos resultados da aplicação da metodologia proposta, obteve-se uma melhora significativa na taxa de acerto do WI007 para o ruído corte de metal que, em média, foi igual a 3,69%.

ABSTRACT

CONTRIBUTIONS TO ROBUST AUTOMATIC SPEECH RECOGNITION

David Daniel e Silva

August/2010

Advisor: Professor Marcelo Ricardo Stemmer, Dr. Ing.
Area of Concentration: Control, Automation and Systems.
Keywords: automatic speech recognition, Lombard database, logistic function, Pesq, robust database.

Page count: 270

Automatic Speech Recognition (ASR) is a fascinating and complex area. For decades the demand for research was based at ASR for not very extensive vocabulary, using techniques that need high performance computing to process the data produced in quiet laboratory environments. From the mid-80 forward, the speech processing technology has advanced, with the use of Hidden Markov Models (HMM) and the high advancement of programming techniques and computer processing, achieving recognition rates in quiet environments close to 100%. In order to put ASR systems to work in real life, several years of intensive research have been and are being made on robust speech recognition. Therefore, applications such as DSR (Distributed Speech Recognition), among others, appeared on the market. In order to achieve a performance similar to the human ear in noisy environments, however, such systems are still the focus of much research. This work makes a study on robust automatic speech recognition systems, aiming at the analysis and behavior of four types of noises (metal cutting, cars in front of a tunnel, cars inside the tunnel and a crowd of children), recorded in different environments for the evaluation and construction of noisy databases. Thus, two databases were developed, having as major contributions the methodology for their construction and the process of analysis and evaluation of data involved in its construction. Furthermore, we present a mathematical development of an algorithm which is the numerical solution to a logistic function of three parameters

of difficult solution, used to model the behavior of WI007 and WI008 systems employed here. A method for initial logistic adjustment (Mail) for Pesq vs. TA curves to evaluate the behavior of the adopted ASR system is also one of the contributions of this work. As one result of the proposed methodology, we obtained a significant improvement in the recognition rate for WI007 for the metal cutting noise which, on average, was equal to 3.69%.

Lista de figuras

Lista de tabelas

Lista de equações

Lista de links úteis

http://portal.acm.org

http://books.google.com

http://libdigi.unicamp.br

http://ieeexplore.ieee.org/Xplore

http://www.computer.org/portal

http://www.decom.fee.unicamp.br/lpdf/

http://www.bu.ufsc.br/

http://htk.eng.cam.ac.uk

http://www.voxforge.org/home/dev/acousticmodels/linux/create/htkjulius/tutorial

http://www.cs.cmu.edu/~robust

http://portal.etsi.org/stq/kta/DSR/dsr.asp

http://www.etsi.org/aurora

http://www.elda.org/article52.html

http://portal.etsi.org/stq/Summary.asp

http://www.opticom.de/technology/pesq.html

http://www.coolest-gadgets.com/20090323/sanwa-throat-microphone

http://www.minitabbrasil.com.br

Lista de siglas

ACR – *Absolute Category Rating*

AF-E – *Advanced Front-End*

Casa – *Computational Auditory Scene*

DES/STQ – *Document ETSI Standard/STQ*

DSR – *Distributed Speech Recognition*

EL – Efeito Lombard

EMV – Estimativa pela Máxima Verossimilhança

ETSI – *European Telecommunications Standards Institute*

F-E – *Front-End*

FFT – *Fast Fourier Transform*

FDP – Função Densidade de Probabilidade

GSM – Global System Mobile

G.712 – É um filtro com características definidas pela *International Telecommunication Union*

HTK – *Hidden Markov Model Toolkit*

HMM – *Hidden Markov Model*

ITU – *International Telecommunication Union*

IA – Inteligência Artificial

IHM – Interface Humano Máquina

Kist – *Korean Institute of Science and Technology*

LDC – *Linguistics Data Consortium*

LPC – *Linear Predictive Code*

Mail – Método de Ajuste Inicial Logístico

MFCC – *Mel Frequency Cepstral Coefficients*

Mirs – É um filtro com características definidas pela *International Telecommunication Union*

ML – *Maximum Likelihood*

MLE – *Maximum Likelihood Estimation*

MMQ – Método dos Mínimos Quadrados

MOS – *Mean Opinion Score*

MRL – Método da Regressão Linear

Pesq – *Perceptual Evaluation of Speech Quality*

RAF – Reconhecimento Automático de Fala

RAFR – Reconhecimento Automático de Fala Robusto

SNR – *Signal-Noise Rate*

SRAF – Sistema(s) de Reconhecimento Automático de Fala

SRAFR – Sistema(s) de Reconhecimento Automático de Fala Robusto(s)

STQ – *Speech Transmission and Quality, Aspects*

Susas – *Speech Under Simulated and Actual Stress*

SVM – *Support Vector Machine*

TA – Taxa de Acerto

TIDIGTS – É uma base de dados de dígitos conectados com amostragem de 20 kHz

TTS – *Text-to-Speech*

VAD – *Voice Activity Detection*

WI007 – *Front-End* do *Work Item* 007 do grupo Aurora

WI008 – *Front-End* do *Work Item* 008 do grupo Aurora

Lista de símbolos

3/4" – três quartos de polegada

A_n e K_n – parâmetros da solução numérica proposta

dB – decibel

$\in$ – Pertence

kHz – quilohertz

km/h – quilômetros por hora

Δ – precisão requerida

ε – tolerância desejada

tol – tolerância calculada

(x_n, y_n) – n-ésimo par ordenado

dsv – somatório dos desvios quadráticos

g – grama(s)

n – número inteiro positivo

N^* – Naturais (sem o zero)

y_i – i-ésimo ponto da curva que se deseja avaliar, para i = 1 até n

y_{iref} – i-ésimo ponto da curva de referência

m – metro

ms – milisegundo

mph – milhas por hora

vs. – versus

$\leq$ – Menor ou igual

$\geq$ – Maior ou igual

Σ – Somatório

] a, b [– intervalo aberto em a, b

Sumário

1. Introdução

Este capítulo introdutório está dividido em duas partes: a primeira oferece uma visão geral sobre o trabalho, apresentando o que foi produzido de forma resumida; a segunda parte faz uma abordagem geral para justificar a escolha do tema, apresentando os problemas do processo de comunicação humana, trazendo-os para o contexto do Reconhecimento Automático de Fala (RAF), incluindo questões que envolvem a área de Reconhecimento Automático de Fala Robusto (RAFR), foco central deste trabalho.

1.1 Visão do trabalho

Neste trabalho, com o objetivo de investigar sistemas de reconhecimento automático de fala robustos, duas metodologias de construção de base de dados foram produzidas.

A primeira metodologia tem como objetivo avaliar a adequação do material de treinamento utilizado na produção de bases de dados robustas, e determinar se as características de um determinado tipo de ruído estão suficientemente presentes na referida base de dados para o desempenho do sistema.

Essa metodologia foi produzida com base no projeto Aurora-1 [1], que é uma base de dados robusta construída por sinais de fala limpos e degradados por meio de oito tipos de ruídos a vários níveis.

A segunda metodologia foi elaborada para aprofundar as investigações sobre sistemas de reconhecimento automático de fala robusto. Neste sentido foi realizado um estudo sobre o efeito Lombard

(variabilidade da articulação do locutor a fim de se comunicar mais eficazmente quando na presença de ruído ambiental.) [2] e sugerida uma metodologia de construção de uma base de dados robusta que seja baseada neste efeito.

A metodologia de construção da base de dados robusta Lombard sugerida aqui, e que está apresentada no capítulo 3, foi aplicada e construída com características psico-acústicas, porém, devido ao elevado custo para a confecção de bases desse tipo, os resultados deste estudo ainda não foram finalizados. Desta forma, a parte referente ao treinamento e teste da base Lombard ficou para trabalhos futuros.

Além disso, pela facilidade de acesso a uma base de dados já concretizada cientificamente, este trabalho foi dirigido essencialmente para o estudo e avaliação sobre sistemas de reconhecimento automático de fala robustos baseados na degradação devido aos tipos e aos níveis de ruídos, como no projeto Aurora-1.

Os ruídos que foram usados para a construção da base de dados Aurora-1 são: trem suburbano (*subway*), multidão de pessoas (*babble*), carro (*car*), salão de exposição (*exhibition*), restaurante (*restaurant*), rua (*street*), aeroporto (*airport*) e estação de trem (*train-station*).

Este trabalho, no entanto, faz um estudo sobre quatro tipos de ruídos novos não presentes na base de dados Aurora-1, que são: corte de metal (*metal-cutting*), frente a um túnel de uma auto-estrada (*tunnel-front*), dentro do túnel (*tunnel-inside*) e multidão de crianças (*crowd-children*).

É necessário esclarecer que, para manter a compatibilidade com a base de dados Aurora-1, foram empregados aqui os termos na língua inglesa dentro dos parênteses.

Os níveis de relação sinal-ruído empregados para o treinamento foram: 5 dB, 10 dB, 15 dB e 20 dB; já os níveis de relação sinal-ruído empregados para o reconhecimento foram: -5 dB, 0 dB, 5 dB, 10 dB, 15 dB e 20 dB.

O treinamento foi constituído de 8.440 (oito mil quatrocentos e quarenta) sinais de fala (arquivos de sinais de fala de dígitos conectados), divididos em quatro conjuntos de sinais limpos mais quatro conjuntos de sinais degradados.

Para os testes foram tomados 4.004 (quatro mil e quatro) sinais de fala. A base Aurora-1 foi constituída de um conjunto de três testes denominados como: *testa*, *testb* e *testc*.

Desta forma, para fazer o estudo dos ruídos novos, apresenta-se uma metodologia de construção da base de dados que leva em consideração um novo teste chamado aqui de *testd*, produzido com os ruídos novos. Assim, além daqueles já presentes no projeto Aurora-1, foi produzido mais um teste.

O sistema de reconhecimento automático de fala empregado foi baseado na tecnologia DSR (*Distributed Speech Recognition*) e é composto dos *front-ends* WI007 [1] e WI008 [3]. A parte de reconhecimento (*back-end*) foi produzida com o HTK (*Hidden Markov Model Toolkit*) [4] que está incluído no pacote da base de dados do

projeto Aurora-1. O WI007, WI008 e demais tecnologias aplicadas neste trabalho estão resumidamente descritas no capítulo 2.

A configuração para o processamento dos sistemas (WI007 e WI008), foi a mesma definida no projeto Aurora-1: 16 estados por palavra, modelo *left-to-right* simples, e mistura de três Gaussianas por estado, com matriz diagonal covariância.

Assim sendo, aqui é apresentado o desenvolvimento de duas bases de dados robustas: a) uma baseada na degradação por ruídos; b) outra baseada na degradação pelo efeito Lombard. A metodologia de desenvolvimento dessas bases de dados e um procedimento de análise e avaliação de ruídos novos em uma base de dados robusta constituem as principais contribuições deste trabalho.

O procedimento de avaliação de ruídos novos em uma base de dados robusta foi criado a partir da análise da medida da qualidade da fala, com o uso da ferramenta Pesq (*Perceptual Evaluation Speech Quality*) [5], e da Taxa de Acerto (TA) dos sistemas de RAF empregados (WI007 e WI008), com uma função logística de três parâmetros proposta no trabalho de [6].

Este procedimento de análise e avaliação está descrito e aplicado no capítulo 5, quando se constata a necessidade de mudanças na base de dados robusta Aurora-1 para melhorar o desempenho dos sistemas em relação ao ruído *metal-cutting*, resultando em uma melhora significativa na taxa de acerto do sistema composto pelo WI007, com uma média de mais 3% de aumento na taxa de acerto para os níveis de 5 dB a 20 dB.

Em conjunto com o procedimento de análise e avaliação de ruídos novos na referida base de dados robusta, é apresentado o desenvolvimento matemático de um algoritmo que é a solução numérica para a função logística de três parâmetros, de difícil solução analítica, usada para modelar o comportamento dos sistemas WI007 e WI008 (curvas Pesq vs. TA) empregados aqui. A solução da função logística é feita a partir do intervalo definido para o parâmetro c, conforme Apêndice A.

Um método de ajuste inicial logístico (Mail) em conjunto com o Método dos Mínimos Quadrados (MMQ) [7][8] para ajustar as curvas Pesq vs. TA, para a avaliação do comportamento dos sistemas RAFR utilizados, também é uma das importantes contribuições deste trabalho.

1.2 Abordagem Geral

A maioria dos estudos nas décadas passadas foi elaborada sob bases de dados limpas e em ambientes silenciosos. De alguns anos para cá, no entanto, muitos trabalhos têm sido feitos com o objetivo de tornar a comunicação homem-máquina, por meio da fala, mais robusta.

A fala e a audição são alguns dos mecanismos mais importantes no processo de comunicação do ser humano. A boa comunicação é o resultado de muito treino desde o seu nascimento até que o homem atinja, por sua natural habilidade ou capacidade de aprendizado, o amadurecimento ou o aperfeiçoamento dos mecanismos que constituem o sistema de comunicação humana, ou seja, o trato vocal e a sua utilização com os devidos treinos executados pela sua rede neuronal - o cérebro.

Na fase de produção pode-se distinguir dois procedimentos: o primeiro é aquele em que o locutor transforma a informação que pretende transmitir em símbolos (fonemas) de uma estrutura linguística; o segundo procedimento consiste em transformar esses símbolos em unidades acústicas (sons). Os meios de transmissão levam as informações produzidas pelo trato vocal humano ao seu destino [9][10][11].

A Figura 1.1 esquematiza o processo de comunicação humana com as etapas de produção (transmissor), transmissão (canal) e recepção (receptor) dos sinais de fala.

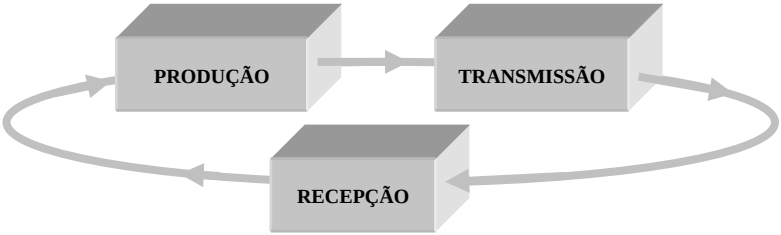


Figura 1.1: Um diagrama do processo de comunicação humana

Fora do contexto da comunicação natural, os sistemas automáticos de reconhecimento de fala tentam permitir que uma máquina, em uma simulação do ouvido, mediante a captura do som da fala por microfones, possa processar e interpretar, com boa precisão, palavras emitidas pelo trato vocal humano. [10][12]

Essas informações, contudo, são degradadas desde a produção até a recepção, devido aos próprios fatores que produzem a variabilidade acústica da fala [13][14], como o efeito Lombard, e devido ao meio ruidoso em que a informação é transmitida e percebida.

A variabilidade acústica intralocutor e entre locutores [15][16] e o ruído do ambiente são alguns dos maiores problemas que afetam a informação acústica produzida pelo trato vocal e, consequentemente, o reconhecimento da fala.

No caso das máquinas o problema é ainda mais complexo, pois não se trata apenas do reconhecimento do sinal acústico, há ainda as dimensões sintática, semântica e pragmática a considerar. Além disso, o sinal de fala pode ser afetado por ruídos de diversos tipos e níveis, provenientes do próprio meio em que a fala está sendo produzida, como, sons de outras falas, máquinas, ferramentas, automóveis, etc. [17]

É preciso ressaltar também que uma mesma fonte dificilmente produzirá ruídos idênticos em instantes de tempo diferentes. Esses ruídos podem variar com a mudança de ambiente, além do que a posição e a distância da fonte de ruído para o receptor também podem variar e afetar diferentemente um sinal de fala. [17][18]

De um modo geral esses ruídos são filtrados facilmente pelo ser humano. Na realidade, ouvidos normais têm uma espécie de banco de filtros que, com o auxílio da estrutura auditiva e do cérebro, além do processo de filtragem, conseguem fazer com que o sinal de interesse seja mais “enfatizado”, o que é fortemente auxiliado com a atenção dada pelo sistema audiovisual do ser humano ao áudio e fonte principal do som [10][19]. É fácil perceber, contudo, que, quando os sinais de ruído são de alto nível de energia em relação ao sinal de fala, torna-se difícil focar a atenção no sinal de interesse.

Outras características da comunicação humana também afetam a taxa de acerto na definição das palavras faladas. Por exemplo, em uma conversa entre duas pessoas há gestos, sorrisos, entre outros, que possuem um significado que não é percebido, senão pela visão ou tato.

Assim, uma máquina, para perceber uma comunicação com maior rigor, teria de contar com um modelo similar ao do ser humano para resolver esta questão, o que é um problema bastante complexo. Muitos pesquisadores procuram soluções baseadas em modelos estatísticos como os de Markov ou HMMs [20][21][22][23] e/ou na área da Inteligência Artificial (IA) [24], mais especificamente baseados em modelos multimodais [25].

Mesmo, porém, com a grande complexidade inerente ao processamento de fala, com algumas limitações os sistemas de reconhecimento automáticos de fala, hoje em dia, conseguem taxas de acertos elevadas, próximas a 100%, como em [12], [26], [27].

Para alcançar essas elevadas taxas de acerto, no entanto, esses sistemas precisam operar em ambientes silenciosos ou com uma SNR (*Signal-Noise Rate*) alta, do contrário, este desempenho cai acentuadamente, *dependendo do nível do ruído do meio em que a fala é produzida, transmitida e recebida.*

Devido a estes problemas, estudos sobre reconhecimento automático de fala robusto ou RAFR, como a caracterização de ruídos e a qualidade perceptual da fala, tornam-se importantes para diversas aplicações, por exemplo, em sistemas de *call center*, em aplicações para

dispositivos portáteis como celulares, em comandos para máquinas como robôs, etc. Ainda, porém, há muita distância entre a prática executada nos laboratórios de pesquisas em reconhecimento automático de fala para com a prática comercial de sistemas de RAFR. [28]

Assim, este trabalho foi desenvolvido sobre o tema Reconhecimento Automático de Fala, mais especificamente motivado na busca de contribuir com a área que envolve a construção de sistemas de RAFR.

É fundamental o desenvolvimento de pesquisas sobre a área de reconhecimento automático de fala robusto, por exemplo, [29] o qual estuda os efeitos do ruído de carro na comunicação móvel, e este trabalho, que apresenta um estudo sobre a caracterização de ruídos de diversos ambientes propondo uma metodologia de avaliação de ruídos novos na base de dados robusta.

A caracterização de ruídos neste trabalho revelou-se importante para a área de reconhecimento automático de fala robusto, pois a partir desses estudos foi avaliada a bases de dados para melhorar o desempenho dos sistemas em meio ruidoso, objetivo de boa parte dos trabalhos nos dias de hoje.

Neste sentido, em [6] é feito um trabalho de caracterização de ruídos, apresentando uma relação entre a qualidade percebida da fala, medida por meio da ferramenta Pesq [5], e a taxa de acerto de um sistema de reconhecimento de fala para dígitos conectados, com testes feitos sobre ruídos da base de dados Aurora-1. [1][30]

A relação entre a qualidade percebida dos sinais de fala e a taxa de acerto do sistema é importante, pois nos oferece a informação do quanto a inteligibilidade dos sinais de fala afeta a resposta do sistema. Dessa relação, pode-se verificar se as características que definem o comportamento de um ruído precisam estar presentes em uma base de dados robusta para treinamento para que o reconhecimento atinja uma melhor taxa de acerto. É neste objetivo que aqui é apresentada uma metodologia de avaliação de ruídos novos na base de dados robustas aproveitando o estudo feito em [6].

Desta forma, este trabalho apresenta uma extensão do estudo de [6], com quatro novos tipos de ruídos mencionados anteriormente, que não estão presentes na base Aurora-1 e que foram gravados nos seguintes ambientes (as fotos dos ambientes de gravação encontram-se no Apêndice B):

- No chão de uma fábrica enquanto cortava-se aço inox de espessura 3/4";
- Em frente a um túnel para passagem de veículos automotores a aproximadamente 10 m de distância da entrada do túnel;
- Dentro do túnel;
- Em um pátio de um colégio com uma multidão de crianças falando e gritando.

Neste sentido, foram realizados vários testes em diversos cenários para determinar:

- qual tipo de ruído teria maior influência na Taxa de Acerto (TA) dos sistemas;
- a relação Pesq vs. TA dos sinais de fala degradados;
- a relação SNR vs. Pesq;
- se algum ruído novo afeta de forma diferenciada a resposta de um sistema de RAF, com treinamento feito sob a base Aurora-1, e se suas características presentes na base de dados mudam essa diferenciação.

Concomitantemente, foi desenvolvida uma metodologia baseada em [6] e na construção da base Aurora-1 [1][30], buscando contribuir para o desempenho dos sistemas de reconhecimento automático de fala, na presença de ruído ambiental (RAFR), sob diversos tipos de ruídos e sob vários níveis de relação sinal-ruído.

Além da caracterização de novos ruídos em relação aos ruídos da Base Aurora-1 e a metodologia empregada para avaliação de ruídos novos nessas bases de dados, outras contribuições do trabalho são:

a) o modelo de avaliação de ruídos novos na base de dados robustas com base na resposta dos sistemas RAFR empregados;

b) a solução e a implementação numérica desenvolvida para o cálculo dos parâmetros da curva logística proposta em [6] igualmente usada nesta pesquisa para descrever o comportamento do sistema de reconhecimento automático de fala (SRAF) na relação Pesq vs. TA;

c) uma metodologia de ajuste das curvas Pesq vs. TA dos ruídos para um comparativo com os ruídos presentes na base de dados Aurora-1 ou outros tipos de ruídos.

Essas contribuições originais, obtidas com a produção deste trabalho, permitem a utilização de algumas técnicas elaboradas aqui em outras áreas, como a solução da função logística na estatística.

A maioria dos trabalhos da área de sistemas de reconhecimento automático de fala robusto visa hoje às aplicações da comunicação homem-máquina em ambientes reais. Desta forma, é importante ressaltar que, além de este estudo contribuir com aplicações de ambientes desse tipo, como DSR, as tecnologias que envolvem reconhecimento automático de fala robusto podem ser aproveitadas para muitas aplicações, pois não diferem das técnicas aplicadas em RAF, a não ser pelo local de processamento do *front-end* e/ou *back-end*.

Aplicações como, por exemplo, sistemas de comandos de mecanismos, envolvendo ou não a tecnologia distribuída, como controle de robôs, brinquedos, etc., podem se beneficiar de estudos como o desenvolvido aqui.

É importante ressaltar ainda que o estudo de ruídos nos quais se incluiu o ruído de corte de metal, teve motivação para o desenvolvimento de projetos de robôs que trabalham em chão de fábrica.

Os demais ruídos foram considerados para investigação de ruídos ambientes que ocorrem no dia a dia, porém não tão frequentes como os estudados na base Aurora-1. Apesar de esses ruídos não serem

tão frequentes, no entanto, existe a possibilidade da utilização de dispositivos que usam a tecnologia de RAF em ambientes dessa natureza.

Além disso, nada impede que um ambiente silencioso seja degradado pela ação de ruídos como o de corte de metal, entre outros. Por exemplo, caso algum serviço de obras da construção civil estiver sendo executado próximo ao local onde há um sistema de reconhecimento automático de fala, o desempenho do SRAF cairá sensivelmente, caso não tenha uma base de dados robusta bem equilibrada para resultar em um sistema *bem treinado.*

Hoje se constata a presença de ruídos de diversos tipos e níveis em quase todos os ambientes, em especial na vida urbana. Resulta daí a necessidade de uma investigação de ruídos mais completa do que a feita no projeto Aurora-1.

Assim, a tecnologia continua a amadurecer, e é evidente que muitas aplicações novas emergirão e se introduzirão na vida diária. O desafio de projetar uma máquina que se comunique através da fala em meio ruidoso similarmente a um ser humano é ainda uma das metas para o futuro. Tais realizações, hoje, são somente o primeiro passo para que seja possível uma comunicação fluente entre homem-máquina em "breve".

Este trabalho, desenvolvido a partir do estudo do estado da arte, está dividido em 6 capítulos.

No capítulo 1 foi apresentada uma introdução ao trabalho, no qual procurou-se expor uma visão geral da comunicação homem-

máquina e de sistemas de reconhecimento de fala robusto, apontando algumas contribuições.

O capítulo 2 descreve a fundamentação teórica, apresentando o estado da arte de sistemas de reconhecimento automáticos de fala robusto, fazendo um histórico de pesquisas sobre o tema e são mostradas algumas tecnologias e ferramentas, como DSR, HTK (*Hidden Markov Model Toolkit*) e o Pesq, que são utilizadas para a produção deste trabalho.

No capítulo 3 é apresentada a base de dados Aurora-1, a metodologia aplicada para elaborar um teste com quatro ruídos novos não presentes na referida base de dados e uma metodologia de construção de uma base de dados Lombard.

O capítulo 4 aborda os modelos para avaliação dos dados do trabalho, a função logística para descrever o comportamento das curvas Pesq vs. TA e o algoritmo numérico proposto para a solução dessa função. Também é apresentado o Método de Ajuste Inicial Logístico (Mail) para a estimação da função logística de três parâmetros comentada anteriormente.

No capítulo 5, são apresentados e analisados os resultados do teste de reconhecimento de fala com os ruídos novos em forma de tabelas e gráficos, sob os diversos cenários, com a aplicação do Mail para a estimação dos parâmetros da curva logística Pesq vs. TA dos sinais de fala degradados, fazendo inferências. Ainda no capítulo 5 são apresentados os resultados de um teste com o ruído *metal-cutting*

presente na base de dados, no qual obteve-se uma taxa média maior que 3% para o WI007, em relação à base de dados original (base Aurora-1).

As conclusões são apresentadas no capítulo 6, em que são feitas considerações a respeito do trabalho e sobre os resultados, salientando as contribuições e fazendo sugestões para trabalhos futuros.

Nos apêndices são exibidos alguns dados empregados no trabalho, tais como: Apêndice A, onde é apresentado o intervalo de localização para a determinação do parâmetro c; Apêndice B, em que constam as fotos dos ambientes de gravação dos ruídos novos para construção de mais um teste além dos que estão presentes na base de dados Aurora-1; Apêndice C, no qual são apresentadas as frases para a construção da base Lombard; Apêndice D, mostrando os resultados dos testes da base de dados Aurora-1; Apêndice E, que apresenta as médias das taxas de acerto de todos os testes; Apêndice F, no qual mostram-se algumas curvas de ajuste para uma melhor avaliação do comportamento dos sistemas, e no Apêndice G são apresentadas as curvas espectrais dos sinais de cada ruído depois de passarem pelo filtro G.712 (filtragem de telefonia digital).

2. Fundamentação teórica

Neste capítulo é apresentado resumidamente o estado da arte desde a década de 60 até os tempos atuais (ano de 2010). Além disso, são apresentadas as técnicas que foram estudadas como base para a produção deste trabalho.

2.1 Estado da arte

A partir dos estudos feitos para desenvolver esta pesquisa, observou-se que, após mais de quatro décadas de pesquisas, as tecnologias de reconhecimento automático de fala entraram finalmente no mercado, beneficiando os usuários em muitas maneiras.

Durante todo o curso do desenvolvimento de tais sistemas o conhecimento dos mecanismos de produção e percepção da fala foi empregado para estabelecer a fundamentação tecnológica para os reconhecedores da fala.

As técnicas que causaram maior avanço nesta área, entretanto, foram desenvolvidas nos anos 60 e 70 com a introdução da representação da fala baseada na análise LPC (*Linear Predictive Code*) [31][32], em métodos de análise *cepstral* [33][34][35], e principalmente nos anos 80, pela introdução de métodos estatísticos poderosos baseados nos modelos ocultos de Markov ou HMMs [20][36].

A Tabela 2.1, baseada no trabalho de [37], sumariza o estado da arte do desenvolvimento dos sistemas de reconhecimento automático de fala há mais de quatro décadas.

Tabela 2.1: Estado da arte dos SRAF desde a década de 60 [37]

1962 – 1967	**1967 – 1977**	**1977 - 1987**
. Vocabulário pequeno baseado na acústica fonética; . Palavras isoladas; . Banco de filtros; . Programação dinâmica.	. Vocabulário médio baseado em modelos; . Palavras isoladas; . Dígitos conectados; . Fala contínua; . Reconhecimento de padrão, LPC, e algoritmos de decodificação.	. Vocabulário grande baseado na estatística; . Palavras conectadas; . Fala contínua; . HMMs; . Modelos de linguagem estocásticos.
1987 - 1997	**1997 – 2007**	**2007 - 2010**
. Vocabulário grande baseado na estatística, sintaxe e semântica; . Fala contínua; . Compreensão da linguagem; . Rede finita de estados; . Aprendizagem estatística.	. Vocabulário muito grande baseado na estatística, sintaxe, semântica, multimodal e TTS; . Diálogo multimodal; . Aprendizagem de máquina, diálogo baseado na iniciativa/modo; . *Reconhecimento robusto;* . Aplicação em DSR; . Aplicação na robótica (humanoides);	. Vocabulário muito grande; . Diálogo multimodal; . Aprendizagem de máquina, diálogo baseado na iniciativa/modo; . Aplicação em DSR; . Aplicação na Robótica; . *Reconhecimento robusto baseado em téc. especializadas avançadas.**

**Estas técnicas especializadas envolvem reconstrução de características perdidas; processamento motivado por fisiologia mononaural e binaural, análise computacional da cena, reconhecimento de fala em alto nível de ruído, etc. [12][38][39]*

Analisando a Tabela 2.1 observa-se que, além da aplicação na tecnologia DSR, outra aplicação importante de reconhecimento automático de fala está presente na robótica, por exemplo, na construção do sentido da audição para os robôs humanoides. Hoje a Sony, Honda e a *Korean Institute of Science and Technology* (Kist), entre outras empresas, fabricam robôs humanoides que antes só eram vistos na ficção científica.

Um esquema dos “sentidos” do robô da Kist é apresentado na Figura 2.1 [40].

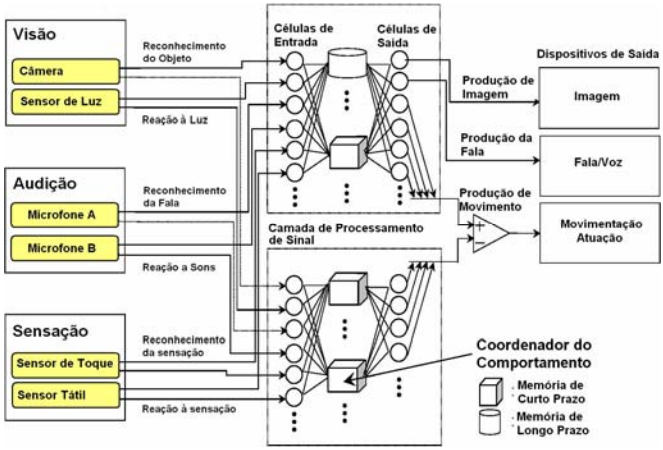


Figura 2.1: Esquema do robô humanoide da Kist [40]

Nesses robôs, a tecnologia da fala é empregada para que o mesmo possa reconhecer a fala, além de produzir a fala, o que dá ao robô o sentido da audição. Os “ouvidos” do robô são formados por dois microfones, um para o lado direito e outro para o lado esquerdo.

Hoje as pesquisas na área de reconhecimento de fala são feitas sobre diversas aplicações, desde brinquedos e serviços de autoatendimento por telefone até a construção de sentidos de robôs, como a audição.

Assim, pode-se afirmar que SRAF tem aplicações promissoras ao menos em cinco áreas: telecomunicações (por exemplo, em sistemas DSR), entretenimento (como em brinquedos), controle e automação (no controle de robôs, entre outras máquinas), saúde (no auxílio a

deficientes físicos e/ou visuais, entre outras aplicações), e na área de segurança (por exemplo, em reconhecimento de voz).

As aplicações de reconhecimento automático de fala, entretanto, ainda sofrem com a degradação dos sinais, ou seja, o ruído do meio em que a fala é produzida ainda é um problema para os SRAF chegarem a um nível mais confiável.

Neste sentido, é importante lembrar a relevância de trabalhos que foquem seus estudos na análise e avaliação das características de ruídos e sua influência em SRAF, como o *ruído de corte de metal***.** Este tipo de ruído, além de estar presente na área da construção civil, é frequente em ambientes de chão de fábrica, em que pode estar atuando um robô que opere com comandos de voz ou um escritório com um SRAF.

2.2 Reconhecimento automático de fala robusto

Os trabalhos que foram desenvolvidos nas décadas passadas sobre RAFR, bem como neste trabalho, usam técnicas que foram e ainda são aplicadas no reconhecimento automático de fala, como a FFT (*Fast Fourier Transform*) [41], amostragem dos sinais de fala, ML (*Maximum Likelihood*) [42], análise espectral [43][44], entre outras.

Estes trabalhos sobre reconhecimento automático de fala, contudo, na busca de melhorar a taxa de acerto de SRAF em ambientes ruidosos, vêm sendo aperfeiçoados pela capacidade atual de processamento das máquinas e/ou por novas técnicas que foram

desenvolvidas. Alguns desses trabalhos são apresentados a seguir, num sucinto histórico sobre RAFR.

2.2.1 Reconhecimento automático de fala nos anos 80

Como referido anteriormente, no início dos anos 80, apesar da ideia básica do modelo oculto de Markov (HMM) ser conhecida e compreendida, essa metodologia ainda não estava completamente difundida. A partir de meados dos anos 80 é que HMMs começaram a se transformar em um método popularmente empregado em RAF, devido a sua característica de modelagem temporal e da variabilidade acústica, além da melhora no desempenho computacional de máquinas e algoritmos.

Essa característica dos HMMs fizeram com que este modelo fosse usado de forma a melhorar sensivelmente as taxas de acertos de sistemas de reconhecimento automático de fala na época. Assim sendo, o HMM tornou-se o modelo preferido pela grande maioria dos pesquisadores. Por outro lado, além da variabilidade acústica e temporal dos sinais de fala, outros fatores causadores de problemas para a robustez de SRAF ainda estavam longe de serem resolvidos, mesmo com alguns trabalhos realizados sobre sinais de fala corrompidos por ruídos.

Por exemplo, em [45] foram empregados processos de filtragem para reduzir ruídos periódicos em sinais de fala. Em [46] foi apresentado o resultado de um estudo do efeito do ruído branco na comparação de dois vetores de coeficientes cepstrais, indicando o grupo que era mais robusto ao referido ruído. Em [47] é proposto um

algoritmo de redução de ruído, concluindo que tanto o nível de ruído ambiental como o estresse degrada e afeta os sinais de fala, diminuindo a taxa de reconhecimento de SRAF.

Desta forma, no final da década de 80 os HMMs deram a partida para o sucesso na questão da modelagem da variabilidade temporal e acústica dos sinais de fala. Sinais degradados pelo estresse e por ruídos ambientais, no entanto, entre outros fatores, ainda eram problemas que estavam começando a ser estudados.

2.2.2 Reconhecimento automático de fala nos anos 90

Na década de 90 muitos estudos foram realizados no intuito de fazer com que se aumentassem as taxas de acerto de sistemas de RAF.

Em [48] é apresentado um trabalho baseado na técnica de *word-spotting* (técnica empregada para detectar palavras-chave dentro de uma locução), para melhorar a taxa de reconhecimento em ambientes com uma relação sinal-ruído de até 5 dB.

Um novo método de processamento, baseado na audição binaural, com múltiplos microfones, foi proposto em [49].

No trabalho de [50] é proposto um algoritmo de compensação *cepstral* usando HMMs, no qual é assumido que a fala e o ruído são aditivos em energia e o ruído é estacionário.

Em [51] é feito um estudo sobre RAFR considerando a degradação da voz sobre várias topologias da tecnologia GSM (*Global System Mobile*).

Ao final da década de 90 um algoritmo com o emprego da técnica de detecção de atividade de voz (*Voice Activity Detection*, ou VAD) foi proposto em [52] para melhorar a robustez de RAF em ambientes ruidosos dentro de automóveis.

Na década de 90, portanto, além de consumar HMMs como excelentes modelos para as variabilidades acústicas e temporal da fala, outras técnicas foram usadas para melhorar a taxa de acerto de sistemas de RAF em meios ruidosos.

2.2.3 Reconhecimento automático de fala nos anos 2000

Nos anos 2000, apesar de contarmos com mais de uma década de utilização dos HMMs em vários trabalhos de estudo e pesquisas, até hoje nenhum paradigma novo mostrou-se forte o bastante para ser comparado aos resultados dos modelos de Markov. Pelo contrário, a maioria desses trabalhos foi e está sendo feito sob modelagem markoviana.

Apesar de [53] ter feito uso de SVM (*Support Vector Machine*) [54][55][56][57] para aplicar em reconhecimento automático de fala, este se limitou a utilizar SVM apenas para projetar modelos de mistura Gaussiana, para então fazer o processamento dos HMMs.

Trabalhos como o de [58], que apresenta um método para estimar dados de fala perdidos, e de [59], que propõe um modelo baseado na audição humana com banco de filtros, etc., utilizam HMMs como modelo para desenvolverem suas pesquisas.

Com a ideia de introduzir reconhecimento automático de fala em telefonia móvel, *Distributed Speech Recognition* (DSR), como em [1], [60], etc., os modelos usados para representação das palavras faladas também foram HMMs.

Assim, hoje em dia, além do uso continuado de HMMs, a área de reconhecimento automático de fala robusto está sendo desenvolvida com técnicas de redução e cancelamento de ruído [61], [62], assim como detecção de presença de sinal de fala [63], [64].

Além dessas técnicas, modelagens da audição periférica, como em [12], também estão sendo apresentadas para melhorar a performance de sistemas de RAF em ambientes ruidosos e para possibilitar cada vez mais que essa tecnologia seja empregada na vida real o mais próximo possível dos testes em laboratório.

Neste sentido, Richard M. Stern [28] (professor da *Robust Speech Recognition Group - Carnegie Mellon University*), argumenta que "*As speech recognition technology is transferred from the laboratory to the marketplace, robustness in recognition is becoming increasingly important…*", ou seja: enquanto a tecnologia do reconhecimento de fala é transferida do laboratório para o mercado, a robustez no reconhecimento está se tornando cada vez mais importante.

Desta forma, alguns outros trabalhos de pesquisa tem abordado tópicos importantes dentro do tema "Reconhecimento Automático de Fala Robusto", como em [39], [65], [66] e [67].

Em [39] é apresentado um estudo sobre recuperação de sinais de áudio levando em consideração os aspectos fisiológicos e perceptualmente motivados da audição humana.

Além dos trabalhos motivados pelos aspectos fisiológicos, algumas pesquisas focam seu interesse no campo da análise computacional da cena auditiva, ou *Computational Auditory Scene* (Casa) [65].

A análise computacional da cena auditiva é uma área interdisciplinar derivada do campo de análise da cena auditiva [65]. Este campo descreve e leva em conta a organização perceptual da audição no contexto dos sons simultâneos (misturas de sons), que busca simular o processo da separação de sinais feito pelo ouvido humano, como em [66] e [67].

Por outro lado, na busca de realçar a fala em meio ruidoso, trabalhos como [68] e [69], entre outros, utilizam-se de técnicas de redução de ruído e melhorar a qualidade da comunicação entre homem-máquina.

Nos trabalhos sobre reconhecimento automático de fala robusto, entretanto, envolvendo bases de dados robustas, como no Projeto Aurora-1 [1][30], não se observa uma metodologia de avaliação da base de dados robusta para a verificação e validação da mesma sob o ponto de vista de treinamento e reconhecimento para definir se a base de dados robusta é bem ou mal projetada.

No trabalho de [70], por exemplo, é proposta uma metodologia para projeto e aquisição de uma base de dados, porém, não

é levado em consideração o emprego de ruídos, apesar de alertarem para a necessidade de se observar o meio para o qual a aplicação de sistemas de RAF é voltada.

Desta forma, com lastro em estudos feitos do estado da arte, este trabalho, sob o tema "*Contribuições ao Reconhecimento Automático de Fala Robusto*", que apresenta uma metodologia de construção e avaliação para uma base de dados ruidosa, é uma contribuição importante aos demais trabalhos da área.

2.3 Técnicas de base para a produção do trabalho

Como uma das aplicações mais atuais e efetivas de reconhecimento de fala robusto, de alguns anos para hoje, residem sobre a comunicação móvel, são apresentados aqui, resumidamente, as técnicas envolvendo o presente trabalho e reconhecimento automático de fala robusto (RAFR) para DSR (*Distributed Speech Recognition*) [1][71].

Desta forma, neste capítulo é apresentado o projeto Aurora-1, de onde foi utilizada a base de dados para a produção deste trabalho; as respostas dos filtros G.712 [72][73] e MIRS [72]; o *Front-End* (F-E) versão 2.0 aqui chamado de WI007 [72], e o *Advanced Front-End* (A F-E) ou WI008 [3][74]; a tecnologia DSR [1][71]; a ferramenta Pesq (*Perceptual Evaluation Speech Quality*) [5] e o HTK (*Hidden Markov Model Toolkit*) [4].

2.3.1 O projeto Aurora-1

Com o objetivo de impulsionar a tecnologia DSR para reconhecimento automático de fala robusto, o projeto Aurora-1 (ETSI/ Aurora Project) com a *Ericsson Eurolab Germany* foi lançado em 25 de janeiro de 2000 (ETSI é a sigla para *European Telecommunications Standards Institute*).

Este projeto foi implementado a partir da conhecida base de dados TIDigits [75], adicionando ruídos ao conjunto de 12.444 (doze mil quatrocentos e quarenta e quatro) sinais de fala de dígitos conectados da referida base de dados. Como a base TIDigits foi gravada em uma taxa de amostragem de 20 kHz, a mesma foi convertida para 8 kHz, que foi a mesma adotada pelo projeto Aurora-1.

Segundo [76], o trabalho desenvolvido no projeto Aurora foi dividido em dois grupos de estudo e projeto:

- Advanced *Front-End* (AFE), grupo que é responsável pela definição dos *front-ends* e assuntos relacionados ao processamento de fala.
- Aplicações e protocolos (A&P), grupo que foi criado para considerar padrões de protocolos cliente-servidor para DSR.

O projeto Aurora-1 foi originalmente criado para estabelecer um padrão mundial para softwares de extração de características que formam o núcleo do *front-end* de um DSR.

O ETSI formalmente adotou essa atividade como *work item* 007 (WI007) e *work item* 008 (WI008) [77]:

- *ETSI DES/STQ WI007* (*DSR Front-End Feature Extraction Algorithm & Compression Algorithm*).
- *ETSI DES/STQ WI008* (*DSR Advanced Feature Extration Algorithm*).

Um número de um *work item*, é o número de um futuro padrão no programa de trabalho do ETSI, e antes desse número é colocado um número adicional do ETSI. A sigla *DES* no *work item ETSI DES/STQ* WI008 significa *Document ETSI Standard.* Pode haver um D (de *Document*) ou um *R* (de *Revision*) na numeração do *work item*, seguida das duas letras padrão (*ES*).

O STQ (*Speech Transmission and Quality Aspects*) é um comitê técnico do ETSI responsável pela qualidade de transmissão do discurso e multimeios [78].

2.3.2 A tecnologia DSR

A tecnologia de reconhecimento de fala distribuída, ou DSR, é um conceito que surgiu com a ideia de melhorar a robustez na comunicação móvel.

Na arquitetura de reconhecimento de fala distribuída, o front-end fica localizado nos terminais móveis que são conectados através da rede móvel para que um servidor efetue o reconhecimento.

Para permitir aplicações DSR no mercado, um padrão para o front-end foi necessário para assegurar a compatibilidade entre o terminal móvel e o reconhecedor remoto [72][79].

Em fevereiro de 2000 o grupo ETSI STQ- Aurora publicou o primeiro padrão DSR com um *front-end* baseado na extração de características *mel-cepstrais* [80].

Um diagrama de blocos da tecnologia DSR e seus componentes são sumarizados na Figura 2.2 [1].

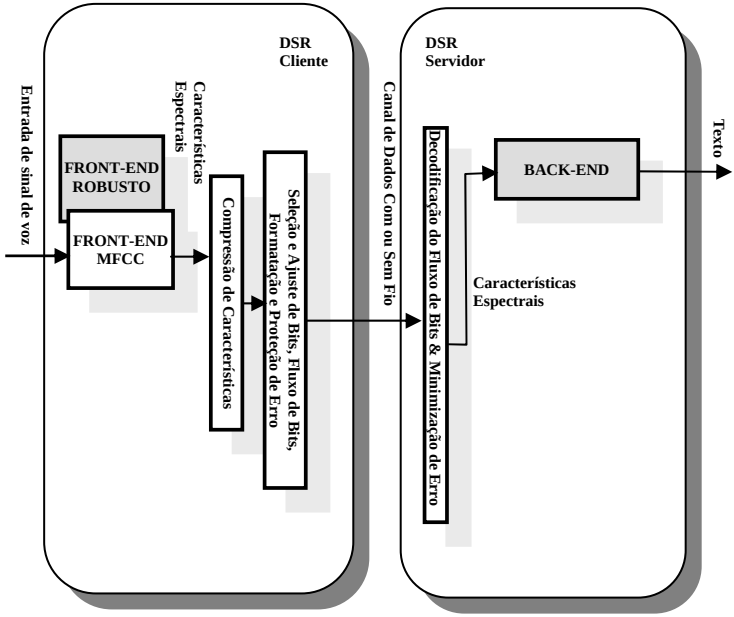


Figura 2.2: Um diagrama de blocos da tecnologia DSR [1]

Para otimizar os detalhes do algoritmo de extração de características, uma base de dados e uma estrutura experimental de referência foram estabelecidas. A base de dados para a tecnologia DSR

foi baseada na base original da TIDigits com filtragem controlada e a adição de ruído sobre o sinal de fala na faixa de 20 dB a -5 dB.

O esquema mostrado na Figura 2.3 indicando os blocos "*Front-End* MFCC" e "*Front-End* Robusto" exibe *front-ends* do tipo WI007 e WI008, respectivamente.

A extração de características baseada no *cepstrum* na escala *mel* [81] foi escolhida para o primeiro padrão em virtude de seu difundido uso em reconhecimento de fala. Na época, porém, foi reconhecido que um padrão ainda mais robusto para cenários ruidosos para extração de características deveria ser elaborado.

Desta forma, o grupo Aurora-1 trabalhou para a produção de um padrão avançado de um *front-end* que proporcionasse uma taxa de acerto superior, comparado ao padrão com a tecnologia *mel-cepstral*.

O bloco do *front-end* baseado nas características *mel-cepstrais* faz parte do padrão ETSI ES 201 108. O bloco do *front-end* robusto foi definido na época como um futuro padrão ETSI para DSR robusto. O bloco referente ao reconhecimento (*back-end*) é uma implementação proprietária [1].

No terminal móvel o sinal de fala é amostrado e parametrizado usando um algoritmo *mel-cepstral* para gerar 12 coeficientes *cepstrais* junto com o C_0 (coeficiente de ordem zero) e o parâmetro log da energia [1]. Esses parâmetros são então comprimidos [1][82] para obter uma taxa de dados mais baixa para a transmissão. Para ficar apropriado para redes sem fio, uma taxa de dados de 4800 b/s

foi escolhida. Então, os parâmetros comprimidos foram formatados em um *bit stream* definido para a transmissão. [1]

Assim, o *bit stream* é emitido em uma rede com ou sem fio, para um servidor remoto, onde os parâmetros recebidos com erros de transmissão são detectados e os parâmetros do *front-end* são descomprimidos para reconstituir as características *mel-cepstrais* do DSR. Em seguida esses dados são repassados para o reconhecimento, que fica na parte central do servidor. O reconhecedor (*back-end*) não faz parte do padrão.

2.3.2.1 Os front-ends WI007 e WI008

No conjunto de dados fornecidos com a base de dados Aurora-1, acompanham dois *front-ends*. Um deles é o *front-end* do conhecido HTK (*Hidden Markov Model Toolkit*). Outro *front-end* que acompanha a base de dados Aurora-1 é o F-E versão 2.0 ou WI007. O *front-end* WI008 pode ser obtido de [76] ou de [77].

Assim, como todos os experimentos do presente trabalho foram realizados com os *front-ends* WI007 e WI008, uma descrição geral deles também é encontrada nas Subseções 3.2.1.1 e 3.2.1.2, respectivamente.

2.3.2.1.1 O front-end WI007

O *front-end* WI007 foi desenvolvido pelo grupo STQ Aurora-1 no ano de 2000 a partir da necessidade de padronização das atividades do projeto DSR- Aurora-1. A intenção foi a de realizar reconhecimento

robusto em SRAF voltado para sistemas de comunicação móvel segundo o modelo DSR com um *front-end* mais avançado (WI008).

Como a padronização de um algoritmo mais avançado levaria um tempo mais longo, foi feita a padronização do WI007, que reunia somente uma análise *cepstral* dos sinais de fala, em que 13 (treze) coeficientes MFCCs (*Mel Frequency Cepstral Coefficients*), incluindo o coeficiente de ordem zero, foram determinados para um frame de fala de 25 ms de comprimento. Assim, cada vetor característico consistia de 14 componentes no total. A Figura 2.3 de [1] mostra um diagrama de blocos ilustrando o MFCC do WI007.

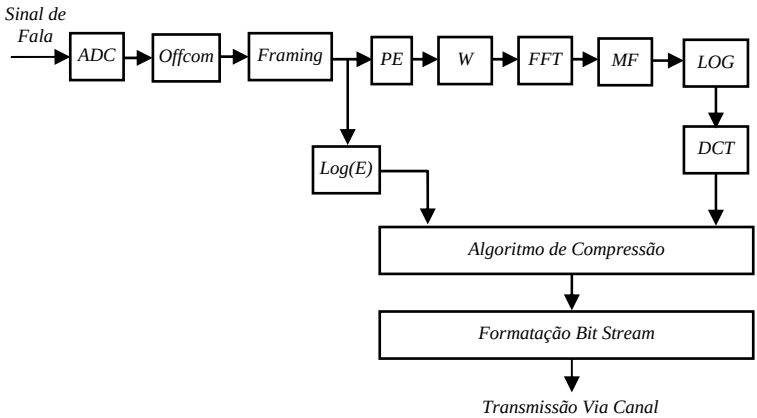


Figura 2.3: Front-End baseado no MFCC do WI007[1]

Onde,

ADC – Conversão Analógica para Digital;

Offcom – Compensação de Offset;

Framing – Seleção/enquadramento;

PE – Pré-ênfase;

Log(E) – Cálculo do log da energia;

W – Janelamento;

FFT – Transformada Rápida de Fourier;

MF – Filtro na Escala Mel;

LOG – Transformação não linear;

DCT – Transformada Discreta do Cosseno;

MFCC – *Mel Frequency Cepstral Coefficients*.

Alguns detalhes do esquema da análise *cepstral* do *front-end* WI007 estão sumarizados na Tabela 2.2 a seguir:

Tabela 2.2: Detalhes da análise *cepstral* do WI007 [1]

Característica	**Descrição**
Filtro: Pré-ênfase	$1\text{-}0.97z^{-1}$
Janelamento	Hamming
DFT	Algoritmo FFT baseado em filtro na escala *mel* com 23 faixas de frequência na escala de 64 Hz até a metade da frequência de amostragem.
OFFSET	Compensação com operação de filtragem.

Como se pôde observar até aqui, além da análise *cepstral*, um esquema de compressão faz parte do *front-end* WI007 para transferir parâmetros acústicos como uma *stream* de dados com uma taxa de 4.800 bits/s. [1][82]

Segundo [1] e [79], um esquema de quantização é usado para codificar os 14 coeficientes de cada frame com 44 bits. Ainda segundo [1] e [79], a quantização é baseada em um *codebook* no qual o conjunto de 14 componentes do vetor é separado em 7 subconjuntos com dois

coeficientes em cada um. Além disso, existem 7 *codebooks* para mapear cada subconjunto de componentes do vetor para uma entrada do *codebook* correspondente. Mais detalhes podem ser encontrados em [1].

2.3.2.1.2 O advanced front-end WI008

Na busca de padronizar um algoritmo robusto, o grupo STQ Aurora selecionou e padronizou um *front-end* avançado ou *Advanced Front-End* (AF-E WI008).

O WI008 combina o desempenho da tecnologia *mel-cepstral* com uma melhor resposta aos níveis de ruído de fundo (algoritmo de redução de ruído) e, portanto, com desempenho mais significativo para mais ambientes. [74]

O AF-E WI008 permite taxas de amostragem dos sinais de fala a 8 kHz, 11.025 kHz e 16 kHz, porém, neste trabalho, a taxa de amostragem empregada foi a de 8 kHz, pois, como já referido, o projeto Aurora-1 foi feito com esta taxa de amostragem.

O WI008 também conta com um outro algoritmo, de detecção da presença de voz, denominado VAD (*Voice Activity Detection*) ou (*Speech Activity Detection*). O VAD é uma técnica usada no processamento de fala que em aplicações de SRAF, DSR, etc., podem auxiliar na habilitação de comandos durante o processo.

Desta forma, o WI008 é apropriado para ambientes que são típicos do uso de telefones móveis, com seus respectivos ruídos de fundo, como os apresentados na base de dados Aurora-1 [1].

Na Figura 2.4, baseada em [3], é apresentado um esquema de extração de características dos sinais de fala com o uso das técnicas do WI008. Como é possível observar, o WI008 também usa os parâmetros *mel-cepstrais* (MFCC – *Mel Frequency Cepstral Coefficients*).

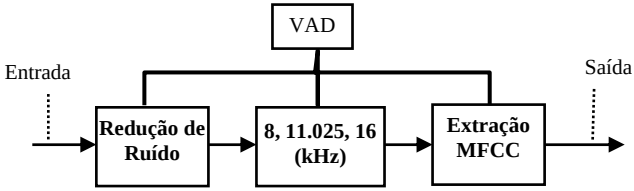


Figura 2.4: Front-End baseado no WI008 [74]

Para finalizar, segundo [3], o melhor desempenho do WI008 em comparação com o WI007, portanto, deve-se principalmente às aplicações dos algoritmos de redução de ruído. O VAD pode auxiliar na minimização dos custos de máquina durante o processamento.

2.3.3 HTK – *Hidden Markov Model Toolkit*

Nesta seção vamos abordar brevemente *Hidden Markov Model Toolkit* (HTK), que foi usado como *back-end* no projeto Aurora-1, e, consequentemente, foi usado neste trabalho.

No HTK, assim como na maioria dos sistemas de reconhecimento automático de fala, geralmente assume-se que o sinal de fala é uma realização de alguma mensagem codificada como uma sequência de alguns símbolos.

A Figura 2.5 de [4] ilustra o processo de codificação e decodificação de um sinal de fala no HTK.

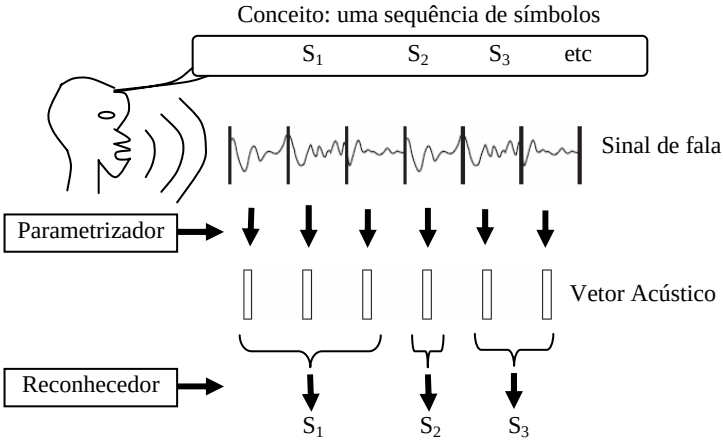


Figura 2.5: Codificação e decodificação dos sinais de fala no HTK [4]

Para efetuar a operação reversa de reconhecer a sequência de símbolos dada uma locução falada, a forma de onda contínua do discurso é convertida primeiramente para uma sequência de parâmetros de vetores de fala [4].

A tarefa do reconhecedor é efetuar um mapeamento entre sequências de vetores de fala e as sequências de base de símbolos procuradas.

No HTK é suposto que uma sequência de vetores de fala observados, que correspondem a cada símbolo, é gerada por um HMM (*Hidden Markov Model*) [83][84].

A Figura 2.6 de [4] mostra o projeto dos modelos HMMs no HTK.

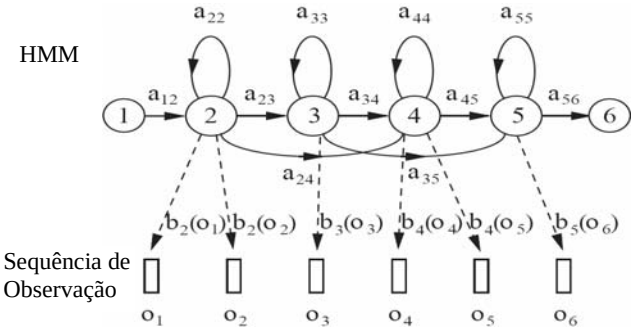


Figura 2.6: Um esquema HMM típico no HTK [4]

O HTK reúne um conjunto de ferramentas para a construção e manipulação de modelos ocultos de Markov ou HMMs.

Primeiramente, as ferramentas de treinamento do HTK são usadas para estimar os parâmetros de um conjunto de HMMs empregando as locuções de treinamento e suas transcrições fonéticas associadas. Em segundo lugar, as locuções a serem reconhecidas são transcritas para texto por meio das ferramentas de reconhecimento do HTK.

Segundo [4], para um melhor entendimento do HTK, é importante ter a compreensão de alguns princípios básicos sobre HMMs. No trabalho de Rabiner [20] é encontrado um bom tutorial sobre *Hidden Markov Models*.

Desta forma, pode-se conseguir uma visão geral do conjunto de ferramentas do HTK e ter uma noção mais aprofundada de como o treinamento e o reconhecimento no HTK funcionam e são organizados.

A Figura 2.7 de [4], mostra o esquema de treinamento e reconhecimento do HTK, com os dois estágios de processamento envolvidos.

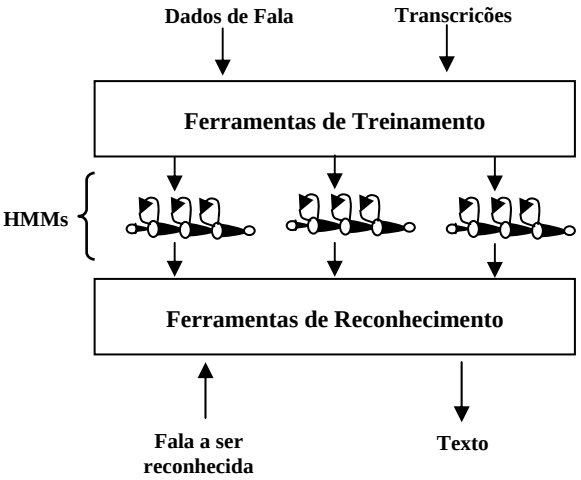


Figura 2.7: Processo de treinamento e reconhecimento no HTK [4]

Assim, contanto que um número suficiente de exemplos representativos de cada fala possa ser coletado, um HMM pode ser construído de tal forma que modele “todas” as fontes de variabilidade inerentes do discurso real. [4]

Para alcançar um melhor desempenho, todavia, procura-se contar com um grande número de modelos acústicos. Cabe ressaltar que os fatores como o ruído, sotaques, fala espontânea, etc., ainda hoje são

problemas para sistemas automáticos de reconhecimento de fala, e o HTK não é imune a esses fatores.

Para o desenvolvimento deste trabalho foi utilizado o sistema proveniente da base de dados Aurora-1, que tem a seguinte configuração: 16 estados por palavra, modelo *left-to-right* simples, e mistura de três Gaussianas por estado, com matriz diagonal covariância para reconhecimento de palavras isoladas. Esta configuração, para possibilitar comparativos, não foi alterada.

2.3.4 PESQ – *Perceptual Evaluation of Speech Quality*

Além do HTK, outra ferramenta empregada neste trabalho foi o Pesq (*Perceptual Evaluation of Speech Quality*) [5]. O Pesq é um algoritmo de avaliação que fornece estimativas de pontuação para a qualidade percebida da fala. Por exemplo, em uma rede telefônica, pode-se comparar a entrada do sinal de áudio (referência) com o sinal correspondente degradado na saída.

O Pesq tem como princípio tomar o sinal degradado e o sinal de referência individualmente, filtrando e promovendo um nivelamento com as características de transferência de um dispositivo de recepção.

Basicamente o Pesq pode usar dois filtros de funções diferentes: um filtro de banda estreita e um filtro de banda larga [5]. No Pesq os sinais são alinhados no tempo para compensar os deslocamentos pequenos que podem ocorrer devido a diferentes algoritmos de codificação.

Um diagrama de blocos do Pesq, com uma pequena adequação, é mostrado na Figura 2.8 e esquematiza esse processo. O diagrama original pode ser visto em [86].

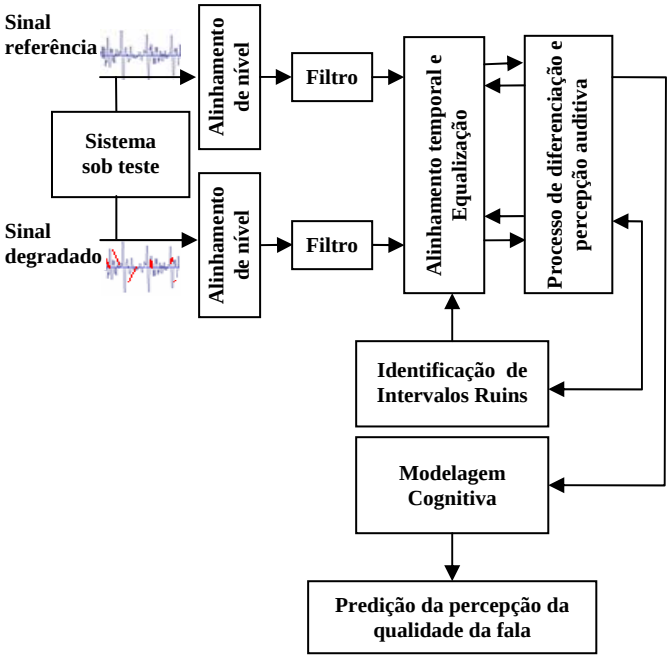


Figura 2.8: Diagrama de blocos do Pesq [86]

Para descrever as distorções que são percebidas por um ouvinte humano o modelo transforma os dois sinais alinhados e filtrados para o domínio da frequência. Subtraindo as duas representações do sinal, uma estimativa das diferenças audíveis é derivada.

As diferenças audíveis são acumuladas dependendo se forem diferentes por uma distorção adicionada ao sinal ou se após a transmissão é detectado falta de parte do sinal (modelo cognitivo).

Na recomendação P.862 [5] observa-se que os valores Pesq se situam em uma faixa de pontuação que vai de -0.5 a 4.5, além de ser indicado que o Pesq seja adotado para avaliação da qualidade da fala para uma banda de frequência de 3.1 kHz (banda de telefonia: 300 a 3400 Hz).

No projeto Aurora-1 e, portanto, neste trabalho, é usada uma banda de frequência que vai de 300-3400 Hz, ou seja, exatamente a banda de frequência requerida por [5].

A contagem Pesq corresponde à contagem MOS (*Mean Opinion Score*) para uma faixa de valores entre 1.0 e 4.5, que é a escala normal de valores do MOS encontrados em um teste ACR (*Absolute Category Rating*) de qualidade [5], quando as condições de teste devem ser apresentadas em ordem aleatória e somente uma vez.

A recomendação P.10/G.100 de julho de 2006 [87] define ACR precisamente como um método de teste em que algumas pessoas são inquiridas a expressar seu julgamento usando uma escala absoluta de qualidade (excelente, bom, e assim sucessivamente).

Esta escala MOS (1.0 a 4.5) pode surpreender, porém na prática observa-se que os valores Pesq médios de vários sinais de fala de um mesmo experimento (contaminados com o mesmo tipo de ruído e mesmo SNR e seguindo a recomendação ITU P.862), não ultrapassam as fronteiras da faixa de pontuação MOS. E, portanto, não atinge a

pontuação ACR excelente, no caso de considerar este valor qualitativo como uma pontuação MOS sem degradação em relação ao sinal de referência. [87]

O aspecto colocado anteriormente é ratificado por [5], informando que a contagem final do Pesq é uma combinação linear do valor médio do sinal degradado e do valor assimétrico médio desse sinal.

Embora a escala de contagem Pesq seja de -0.5 a 4.5, para a maioria dos casos a escala de saída é uma qualidade de escuta MOS como uma contagem entre 1.0 e 4.5, sendo a escala normal de valores MOS encontrados em uma experiência ACR. [5]

Ainda segundo [5], é possível usar Pesq para avaliar a qualidade de sistemas que processam sinais de fala na presença de ruído ambiental ou de fundo (por exemplo, carro, rua, restaurantes, escolas, etc.). As gravações do ruído devem ser passadas através de um filtro apropriado para o teste.

Além disso, para que o Pesq tome em consideração a degradação subjetiva em um contexto de teste ACR, devido ao ruído e a todas as distorções da codificação, o sinal original (referência) usado com Pesq deve estar limpo, mas o ruído, obviamente, deve ser adicionado antes que os sinais sejam passados pelo sistema sob teste. [5]

A Tabela 2.3 mostra a escala de opinião da medida Pesq-MOS usada em [5] e que também é adotada aqui.

Tabela 2.3: Escala de opinião Pesq-MOS [5]

Qualidade do Sinal de Fala	Pontuação
Excelente	5
Bom	4
Satisfatório	3
Pobre	2
Ruim	1

Desta forma, neste trabalho, os valores Pesq médios obtidos de todos os sinais de fala limpos (sinais referência) em relação aos sinais ruidosos (sinais degradados) sob cada tipo de ruído são os valores adotados para a contagem Pesq-MOS, e, em decorrência, para todo processo de análise.

Para finalizar deve ficar claro que, na pontuação Pesq-MOS empregada para medir a qualidade da fala aqui, o procedimento foi utilizado para os sinais não processados, ou seja, antes de serem submetidos pelos *Fronts-Ends*: WI007 e WI008.

3. Base de dados

Neste capítulo é descrita a base de dados utilizada neste trabalho (base Aurora-1) e também é apresentada uma metodologia para a produção de uma base de dados Lombard.

A base para os testes feitos é formada pela base de dados do projeto Aurora-1, com locuções que contemplam vários tipos de ruídos. Nela foi acrescentado um conjunto de testes com locuções degradadas por quatro novos tipos de ruídos. A produção dessas locuções, bem como a sua análise, constitui uma das contribuições deste trabalho.

A metodologia de construção de uma base de dados para o estudo do efeito Lombard é descrita, e, embora este estudo ainda não tenha sido finalizado, a base de dados foi concluída, e constitui uma importante contribuição para as pesquisas na área de reconhecimento robusto de fala.

Um bom desempenho para SRAFR pode ser conseguido por uma extração apropriada de características dos sinais ruidosos no *front-end* e/ou pela adaptação dos sinais de referência à situação de degradação do sinal. [88][89]

Dessa forma, uma base de dados ruidosa com um bom projeto de treinamento e os conjuntos de testes podem ser tomados para determinar o desempenho do sistema de reconhecimento de fala, *e esta é a principal motivação para este trabalho.*

3.1 A base de dados Aurora-1

A base de dados do projeto Aurora-1 foi implementada a partir da base de dados TIDigits, adicionando ruídos ao conjunto de 12.444 (doze mil, quatrocentos e quarenta e quatro) sinais de fala de dígitos conectados da referida base de dados e convertida da taxa de amostragem de 20 kHz para uma taxa de 8 kHz (*downsampled*).

Além da base de dados para treinamento, o projeto Aurora-1 é constituído de um conjunto de três testes para reconhecimento, denominados *testa*, *testb* e *testc*.

Ambas as partes, de treino e de testes, são compostas por locuções produzidas sob duas condições: limpas (dados *clean*) e múltiplas (dados degradados por ruídos).

Para a formação da base de dados robusta (base Aurora-1), a degradação dos sinais de fala foi feita artificialmente e acusticamente com a adição de 8 (oito) tipos de ruídos, a seis níveis de SNR (-5 dB, 0 dB, 5 dB, 10 dB, 15 dB e 20 dB).

Depois do processo de *downsampling* e mixagem dos ruídos, um filtro G.712 [73] foi aplicado para que a frequência do sinal degradado ficasse na banda de 300-3400 Hz, que é a banda de frequência característica de telefonia digital.

Dois padrões de frequência foram usados, as quais têm suas definições dadas pela Recomendação ITU G.712 de 1996, correspondendo, portanto, a dois filtros empregados no projeto da base de dados Aurora-1: o filtro G.712 e o filtro Mirs. As abreviações G.712 e Mirs foram introduzidas como referências para esses filtros.

Na Figura 3.1 pode-se observar a resposta em frequência do filtro G.712.

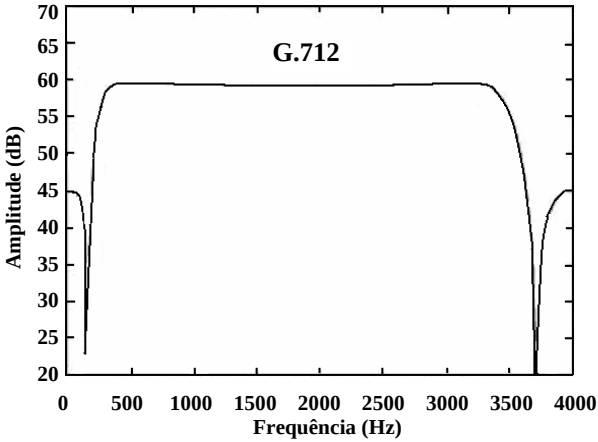


Figura 3.1: Resposta em frequência do filtro G.712 [43]

Na Figura 3.2 pode-se observar a resposta em frequência do filtro Mirs.

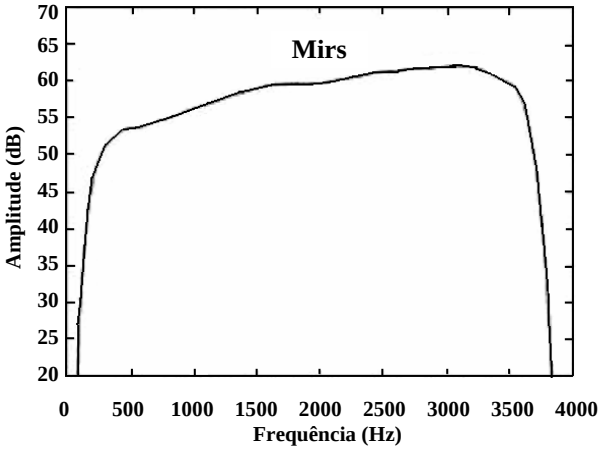


Figura 3.2: Resposta em frequência do filtro Mirs [1]

O filtro Mirs pode ser visto como tendo uma frequência característica que simula o comportamento de um terminal de telecomunicação, no qual se encontram as exigências para a resposta em frequência de entrada como especificado, por exemplo, para a tecnologia GSM (*Global System Mobile*).

A diferença principal é a característica de uma curva plana do G.712 na escala entre 300 e 3400 hertz em que o Mirs mostra uma característica de inclinação crescente com uma atenuação em baixas frequências. Ambos os tipos de filtros foram produzidos com base nos módulos do software [90].

Para a base de dados Aurora-1, os sinais de ruído foram selecionados para representar vários cenários, para as mais prováveis aplicações para terminais de telecomunicações.

Desta forma, os ruídos foram gravados em diferentes lugares [1], conforme pode ser observado a seguir:

- Suburban train (*Subway*);
- Crowd of people (*Babble*);
- Car;
- Exhibition hall (*Exhibition*);
- Restaurant;
- Street;
- Airport;
- Train-station (*Train*).

Para a confecção do teste com os ruídos novos foram usados quatro cenários diferentes para as gravações dos mesmos, como descrito na sequência.

- Corte de Metal – Aço Inox ¾" (*Metal-Cutting*);
- Em frente a um túnel de uma autoestrada movimentada (*Tunnel-Front*);
- Dentro do Túnel da referida autoestrada (*Tunnel-Inside*);
- Multidão de Crianças (*Crowd-Children*).

É preciso ressaltar que tanto para os ruídos da base Aurora-1 quanto para os ruídos novos apresentados anteriormente, os termos especificados entre parênteses são os que foram utilizados neste trabalho. A ausência de parênteses significa que o próprio nome dado ao ruído foi utilizado.

O projeto para produção dos dados de treinamento e para os dados de teste, incluindo o *testd*, com os ruídos novos mostrados anteriormente, é apresentado nas Seções 3.1.1 e 3.1.2, respectivamente, a seguir.

3.1.1 Produção dos dados de treinamento

O conjunto de dados obtidos da TIDigits para formar a base de dados Aurora-1, como abordado anteriormente, foi separado em um subconjunto de dados para treinamento e um subconjunto de dados para teste.

O treinamento foi constituído de 8.440 (oito mil, quatrocentos e quarenta) sinais de fala. Para os testes foram tomados 4.004 (quatro mil e quatro) sinais. Desta forma, a proporção de dados para treinamento e testes para a totalidade do conjunto de sinais de fala foi de, aproximadamente, 68% e 32%, respectivamente.

Os dados para treinamento foram considerados em dois modos:

- treinamento sob condições limpas (*training under clean data*);
- treinamento sob condições múltiplas (*training under multi-condition data*).

O treinamento sob condições limpas corresponde aos 8.440 (oito mil quatrocentos e quarenta) sinais de fala limpos referidos anteriormente, amostrados a 8 kHz e filtrados com as características do filtro G.712.

O treinamento sob condições múltiplas foi produzido com quatro conjuntos de sinais de fala limpos (*clean1*, *clean2*, *clean3* e *clean4*), mais quatro conjuntos de sinais de fala degradados pela adição artificial de ruídos a quatro níveis de SNR: 20 dB, 15 dB, 10 dB e 5 dB, totalizando os 8.440 (oito mil quatrocentos e quarenta) arquivos de fala já mencionados.

Os ruídos foram usados para definir conjuntos denominados de N1, N2, N3 e N4 (Noise1, Noise2, Noise3 e Noise4, respectivamente), e cada um foi constituído dos quatro níveis de SNR já mencionados.

Cada conjunto limpo (*clean1*, *clean2*, *clean3* e *clean4*), foi constituído de 422 (quatrocentos e vinte e dois) arquivos de sinais de fala de dígitos conectados.

Cada nível de SNR dentro dos conjuntos N1, N2, N3 e N4 também foi formado por 422 (quatrocentos e vinte e dois) arquivos de sinais de fala, porém degradados por um tipo de ruído no seu respectivo SNR.

No projeto da base de dados Aurora-1, conforme especificações observadas de [1], [3] e [74], foram usados quatro tipos de ruídos para o treinamento, a saber:

- *subway* (N1);
- *babble* (N2);
- *car* (N3);
- *exhibition* (N4).

Desta forma, 20 (vinte) diferentes condições foram tomadas como entrada para o treinamento em condições múltiplas [1]. Neste trabalho esses dados também foram usados para treino.

3.1.2 Produção dos dados de reconhecimento

O conjunto de dados da TIDigits tomado para testes no projeto Aurora-1 foi separado em três outros conjuntos. Estes conjuntos, conforme já mencionado, foram formados por três testes denominados *testa*, *testb* e *testc*.

Os ruídos dentro de cada teste, da mesma forma que no treino, também foram separados em quatro subconjuntos, denominados de N1, N2, N3 e N4 (Noise1, Noise2, Noise3 e Noise4, respectivamente), porém com 6 (seis) níveis de SNR para cada conjunto (tipo) de ruído.

O *testc*, no entanto,foi constituído apenas com os conjuntos N1 e N2. Assim como no treino, os dados para teste também foram considerados em dois modos:

- teste sob condições limpas (*test under clean data*).
- teste sob condições múltiplas (*test under multi-condition data*).

Os testes sob condições limpas correspondem a sinais de fala limpos, amostrados a 8 kHz filtrados com as características do filtro G.712, separados em conjuntos denominados *clean1*, *clean2*, *clean3* e *clean4*, obviamente com conteúdos diferentes dos conjuntos com o mesmo nome do treinamento.

Os testes sob multicondições foram produzidos com os quatro conjuntos de sinais de fala limpo *clean1*, *clean2*, *clean3* e *clean4* descritos anteriormente, degradados pela adição artificial de ruídos a 6 (seis) níveis de SNR: 20dB, 15dB, 10dB, 5dB, 0dB e -5dB.

Os níveis de relação sinal-ruído foram aplicados para definir seis subconjuntos dentro de N1, N2, N3 e N4, denominados de SNR_XX (onde XX é o nível numérico do ruído).

Cada conjunto limpo *clean1*, *clean2*, *clean3* e *clean4*, e cada conjunto degradado por ruídos a seis níveis de SNR, foram constituídos de 1.001 (mil e um) arquivos de sinais de fala de dígitos conectados.

Desta forma, com os três testes do projeto da base de dados Aurora-1, mais o novo teste, neste trabalho foram usados quatro conjuntos de dados de teste para o reconhecimento, os quais são apresentados nas subseções que seguem.

3.1.2.1 Produção do testa

Para o *testa*, os 4 (quatro) conjuntos N1, N2, N3 e N4, foram produzidos com a adição artificial de ruídos nos respectivos conjuntos de sinais limpos (*clean data*):

- *subway* (N1);
- *babble* (N2);
- *car* (N3);
- *exhibition* (N4).

Cada um dos conjuntos N1, N2, N3 e N4 foi constituído de sinais de fala com seis níveis de SNR: -5dB, 0db, 5dB, 10dB, 15dB e 20dB.

Os sinais de fala limpos usados para a formação de N1, N2, N3 e N4 foram *clean1*, *clean2*, *clean3* e *clean4*, respectivamente, e foram constituídos de 1.001 (mil e um) sinais distintos entre si.

Além dos conjuntos degradados (N1 a N4), os conjuntos de fala limpos (*clean1* a *clean4*) também fazem parte do teste.

Desta forma, cada conjunto dos ruídos (N1, N2, N3 e N4) também foi constituído do mesmo número de sinais de fala.

Então, pode-se afirmar que o *testa* foi constituído de 4 conjuntos de sinais de fala limpos (*clean1*, *clean2*, *clean3* e *clean4*), mais 4 conjuntos de sinais de fala degradados por ruídos (N1, N2, N3 e N4).

A combinação desses conjuntos degradados a seis níveis de SNR, conforme mencionado anteriormente, determinam o processamento ("custo de máquina") de 24.024 (vinte e quatro mil e vinte e quatro) sinais de fala degradados mais 4.004 (quatro mil e quatro) processos referentes aos sinais de fala limpos, num total de 28.028 (vinte e oito mil e vinte e oito) processos para o *testa*.

Os demais testes (*testb*, *testc* e *testd*) seguiram as mesmas características do *testa*, portanto a produção deles é apresentada resumidamente a seguir.

3.1.2.2 Produção do testb

O *testb* foi constituído com as mesmas características do *testa*, porém com diferentes tipos de ruído. Os ruídos usados para o *testb* foram:

- *restaurant* (N1);
- *street* (N2);
- *airport* (N3);
- *train-station* (N4).

3.1.2.3 Produção do testc

O *testc* foi feito com o conjunto N1 (ruído *subway*) do *testa* e o conjunto N2 (ruído *street*) do *testb*.

Outra diferença entre o *testc* para os demais, é que ele foi constituído aplicando-se um filtro de característica Mirs [1].

A intenção em fazer o *testc* foi considerar outras frequências, aplicando Mirs ao invés do filtro G.712, para simular a influência de terminais móveis de telefonia celular com essas frequências diferentes. Desta forma foi construída a base de dados Aurora-1, e a Figura 3.3 sumariza como ela foi produzida.

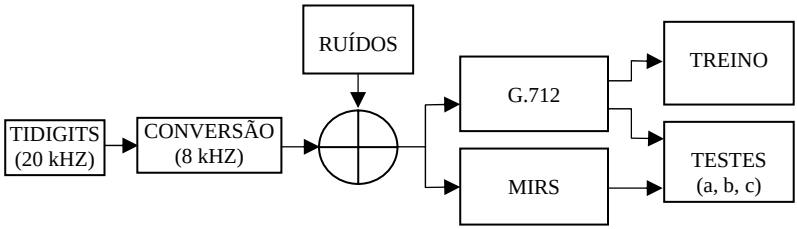


Figura 3.3: Um diagrama de blocos do projeto Aurora-1

3.1.2.4 Produção do testd

Os dados do *testd* produzidos para este trabalho foram elaborados com a mesma metodologia para a produção do *testa* e *testb*, porém os ruídos empregados para a degradação dos sinais de fala foram diferentes dos usados nesses testes.

Os seguintes tipos de ruídos foram usados para a produção do *testd*:

- *Metal-Cutting*;
- *Tunnel-Front*;
- *Tunnel-Inside*;
- *Crowd-Children*.

Em automação industrial o ruído *Metal-Cutting* é de especial interesse em aplicações que fazem uso de SRAF, uma vez que este é um ruído comum no chão de fábrica. Isto pode ser pouco relevante em outras áreas, mas pensando em sistemas de RAF para IHM, a inclusão deste e de outros ruídos característicos deste ambiente podem se mostrar importantes.

O ruído *Metal-Cutting* foi gravado dentro de um galpão enquanto uma serra poli-corte efetuava o corte de barras de aço inox de ¾ de polegada.

O microfone foi colocado a uma distância aproximada de 3 m da máquina, no interior de um galpão com cerca de 10 m de largura, 30 m de comprimento e 6 m de altura.

O ruído *Tunnel-Front* foi obtido em uma autoestrada movimentada, a cerca de 15 m da entrada do túnel e 2 m do início de sua lateral.

O ruído *Tunnel-Inside* foi obtido aproximadamente na metade do seu comprimento.

O ruído *Crowd-Children* foi gravado em um colégio para alunos com idade escolar entre 5 e 12 anos de idade, pouco antes da entrada em sala de aula.

Todos os ruídos foram gravados com um computador *laptop*, usando software de gravação e edição de áudio profissional, com taxa de amostragem de 16 kHz com uma quantização de 16 bits, mono.

Depois da gravação dos ruídos, a taxa de amostragem foi convertida para 8 kHz, conforme feito para a base de dados Aurora-1.

Em todos os ambientes de gravação – galpão, em frente ao túnel, dentro do túnel e no colégio – o microfone foi colocado a aproximadamente 2 m de altura.

No Apêndice B são apresentadas fotos dos ambientes de gravação dos quatro ruídos novos.

Um dos problemas enfrentados durante as gravações dos ruídos foi a saturação, resolvido checando o dial de volume do gravador.

O microfone utilizado foi do tipo omni-direcional flexível com fone de ouvido (*headset* ajustável) desenvolvido especialmente para aplicações multimídia, computadores interativos, reconhecimento de voz, estações de escuta e laboratórios de linguagem. As características do microfone são apresentadas na Tabela 3.1.

Tabela 3.1
Especificações do microfone utilizado para a produção do testd

Especificações
Impedância: 32 ohms
Sensibilidade: 96 dB +/- 4dB
Frequência: 20 Hz a 20 kHz
Padrão polar: omni-direcional

A produção dos dados para a realização do *testd*, ou seja, com os novos ruídos, foi feita a partir da mixagem dos mesmos aos áudios limpos (*clean1*, *clean2*, *clean3* e *clean4*) da base de dados Aurora-1.

As mixagens para treino e teste foram feitas com scripts executados no Matlab e verificados com auxílio de ferramentas de edição de áudio profissionais. Ambos os testes mostraram que a mixagem foi realizada com sucesso com todos os ruídos e em todos os níveis de SNR. Por outro lado, diferentemente do procedimento de mixagem acústica feita no Projeto Aurora-1, aqui este processo foi realizado computacionalmente.

Desta forma, a base de dados para o *testd* foi produzida sob 4 (quatro) tipos de ruídos a 6 (seis) níveis cada: -5 dB, 0 dB, 5 dB, 10 dB, 15 dB e 20 dB.

Antes da mixagem os ruídos foram passados por um filtro com as características do G.712, com banda de frequência da telefonia digital (300-3.400 Hz). Desta forma os dados foram mantidos com as mesmas características dos demais testes do projeto Aurora-1.

A partir dos dados de teste prontos e utilizando a base de dados Aurora-1, o presente trabalho foi desenvolvido usando o *front-end* WI007 [1] [74] e o *Advanced Front-End* WI008 [3].

3.2 Base de dados lombard

Como abordado no início deste capítulo, foi projetada e construída uma base de dados para estudo do efeito Lombard. Nesta seção são mostrados os detalhes deste trabalho.

3.2.1 Motivação

Segundo [2], com o avanço da tecnologia, a comunicação com dispositivos móveis ocorre em vários ambientes, como salão de exposições, carros, etc., porém o uso dessa tecnologia em vários cenários implica uma maior dificuldade de comunicação devido aos tipos e níveis de ruídos presentes nos diferentes ambientes de comunicação.

Para tentar compensar de algum modo esta dificuldade, a forma de fala é mudada. Esta mudança é conhecida como *Efeito Lombard (EL)*, que pode ser definido como a variabilidade da articulação do locutor a fim de se comunicar mais eficazmente quando

na presença de ruído ambiental. Este é um efeito psicológico do ruído no locutor. [2]

No tratamento do ruído ambiental esquemas de redução de ruído e realce da fala empregam várias técnicas para melhorar a taxa de acerto de sistemas.

Até agora, todavia, segundo [2], na comunidade de processamento de fala, o efeito Lombard foi suposto ser uniforme para todos os tipos e níveis de ruído. Desta forma, este aspecto é uma das motivações para apresentar uma metodologia de construção de uma base de dados para análise da fala Lombard em SRAF.

3.2.2 Estudos sobre a fala Lombard

O efeito Lombard é um tópico de pesquisa emergente na comunidade de fala e conta com diversos estudos preliminares na literatura [2].

Segundo [2], muitas das pesquisas recentes sobre o efeito Lombard estão descritas em [91] e [92].

Ainda segundo [2], entretanto, análises sobre as características acústicas e fonéticas do discurso Lombard têm sido menos frequentes. Uma análise acústica e fonética detalhada do discurso sob diferentes tipos de estresse, incluindo o EL, estresse físico, de carga de trabalho e emoção, foi realizada por Hansen em [93].

Porém, a construção de uma base de dados robusta que leve em consideração o EL não é fácil de ser realizada.

Neste sentido, a base de dados usada em [2] faz parte do "*Speech Under Simulated and Actual Stress*" (SUSAS), detalhes dos quais podem ser encontrados em [94], conforme indicado em [2], e disponível no consórcio de dados linguístico, ou "*Linguistics Data Consortium*" (LDC) [95].

Um estudo similar ao de [2] também foi desenvolvido por [96], no qual o discurso produzido em voz alta e Lombard em ambiente simulado tipo cabine de piloto foi usado para análise. Conforme consta em [2], as análises acústicas e perceptuais também foram feitas por Summers et al. [97].

Os estudos citados mostraram que sob o EL a duração das vogais aumenta e que a inclinação espectral diminui, implicando um aumento das componentes de alta frequência. Um aumento no pitch e posição do primeiro formante também ocorrem. [2]

Também há migração de energia de baixa e de alta frequência para uma faixa média para vogais, e movimentos da baixa para as altas bandas. Além disso, diferenças entre os locutores homens e mulheres foram observadas em [98].

Ainda na percepção de [2], outro aspecto de interesse no discurso Lombard é a inteligibilidade, explicitando que um estudo do discurso para este aspecto sob o EL foi feito por [99], [100], e [101].

Estes estudos revelam que quando a voz é apresentada em uma situação constante de estresse, a inteligibilidade do discurso Lombard aumenta até um determinado nível de ruído, e diminui abruptamente quando a fala em voz alta se torna gritada.

As pesquisas referidas também mostram que o retorno auditivo do discurso é necessário para manter a inteligibilidade da fala Lombard que, segundo [2], é vital, porque a finalidade primária do EL é aumentar a inteligibilidade de uma comunicação com outros locutores, em meio ruidoso.

Em [2] foi realizado um trabalho de análise do discurso Lombard sob três tipos diferentes de ruído: ruído em um carro que viaja a 65 mph (aproximadamente 104 km/h) em uma estrada, com a metade das janelas abertas; ruído cor-de-rosa; e ruído de uma grande multidão de pessoas.

Conforme comentado anteriormente, [2] utilizou para seus experimentos uma base de dados provenientes do consórcio de dados linguístico, ou LDC, porém uma base de dados no português do Brasil e estudos do EL para a pronúncia em língua não nativa parecem ainda não estarem disponíveis.

Além disso, a construção de uma base de dados Lombard é muito mais difícil de ser realizada do que uma base em ambiente silencioso, ou base robusta a partir de tipos e níveis de ruídos. Para a construção de uma base Lombard, além de ter de conseguir locutores para fazer as gravações, ainda é necessário convencê-los a escutar os tipos e níveis de ruídos a que são submetidos, tornando essa tarefa mais difícil.

Desta forma, neste trabalho é proposta uma metodologia para construção de uma base Lombard, e, empregando os quatro tipos de

ruídos novos propostos aqui, esta metodologia foi aplicada com a colaboração de 44 locutores: 22 homens e 22 mulheres.

3.2.3 Uma metodologia para a construção de uma base Lombard

Nas pesquisas de [2], [102] e [103], entre outros, é observado que a taxa de reconhecimento de SRAF não é afetada somente pelo ruído, mas também pelo EL. Isolar a influência deste efeito, contudo, não é uma tarefa trivial, pois o locutor deve falar em um ambiente ruidoso, mas a gravação deve conter apenas o áudio da voz, sem o ruído.

Em outras palavras, o interesse da medida Lombard deve focar as diferenças de pronúncia do locutor [2][96] devido à presença de ruído ambiental. Nesta seção será descrita a metodologia utilizada para a construção de uma base de dados Lombard para o português do Brasil e para uma base de dados Lombard na língua inglesa pronunciada por locutores não nativos.

a) Metodologia

Inicialmente pensou-se em fazer com que os locutores ouvissem os ruídos através de fones de ouvido. Desta forma seria fácil capturar apenas a voz, uma vez que o ruído não seria captado pelo microfone.

Esta ideia, entretanto, foi descartada pela dificuldade de medir a potência do ruído na saída dos fones de ouvido. Outro fator que determinou o abandono desta ideia foi a percepção de que ouvir ruído através de fones de ouvido pode ser diferente de ouvi-lo em ambientes reais.

Uma segunda ideia foi a de utilizar microfones de garganta, que são bastante imunes ao ruído ambiente. Como alguns testes preliminares mostraram bons resultados, esta opção foi adotada no presente trabalho. Além disso, outros cuidados foram tomados em relação ao ambiente de gravação, e ao processo de geração dos ruídos.

b) Ambiente de gravação

As gravações foram realizadas em uma sala semianecoica, com revestimento de espuma de alta densidade em todas as faces da sala, à exceção do piso. A Figura 3.4 mostra uma foto desse ambiente.

Figura 3.4: Sala semianecoica usada para produzir a base Lombard

Apesar de a Figura 3.4 mostrar somente as partes não reverberantes, o piso não é revestido com espumas, é de concreto, ou seja, uma sala semianecoica.

Objetivando dar uma característica o mais real possível aos sons emitidos para a escuta dos ruídos pelos locutores, foi produzido um aparato com três caixas acústicas, posicionadas com base na tecnologia

surround [102] para garantir que a pressão sonora (em dB) chegasse aos níveis desejados.

Desta forma, duas caixas foram posicionadas na altura dos ouvidos, sobre um pedestal móvel, a menos de meio metro da parte de trás da cabeça de cada locutor. Uma terceira caixa acústica foi posicionada em linha às demais caixas, na parte central. Assim, foi simulado o som dos ruídos provenientes de variadas direções, simultaneamente, dentro do ambiente de gravação.

A Figura 3.5 exibe o esquema montado para realizar os procedimentos de emissão de ruído para produção do EL.

Figura 3.5: Montagem do esquema acústico de emissão de ruído para o efeito Lombard

Conforme se pode perceber na Figura 3.5, o esquema montado possibilitou o nivelamento da altura das caixas acústicas, tanto em termos de altura como a distância de onde se posicionava o locutor.

c) Microfone

Como referido anteriormente, as gravações foram realizadas com um microfone de garganta. O modelo utilizado foi um dispositivo para comunicação tátil da Motorola, tipo *Talkabout*, modelo T5025.

O microfone de garganta para o *Talkabout* Motorola, modelo T5025, têm a característica de captar a voz do locutor em meio ruidoso e ao mesmo tempo impedir que o ruído degrade totalmente a fala.

Na experiência que foi feita neste trabalho, empregando este microfone, chegou-se a obter gravações com redução de ruído de aproximadamente 90%.

O microfone de garganta para o *Talkabout* Motorola, modelo T5025 é apresentado na Figura 3.6.

Figura 3.6: Microfone de garganta utilizado para produção da base Lombard

Entretanto, há modelos mais sofisticados, como o modelo de [103] mostrado na Figura 3.7, que segundo [103] pode apresentar melhor desempenho.

Figura 3.7: Microfone de garganta da Sanwa Technology [103]

No modelo de microfone de garganta para o *Talkabout* Motorola há a necessidade do sinal ser ativado pelo canal do radiocomunicador, o que provoca ruídos não desejados nos sinais de fala. Além disso, caso não opere com ativação de sinal, o microfone captura a voz com uma sensibilidade menor, e para vozes muito baixas ele acaba distorcendo o sinal.

O microfone de [103], mostrado na Figura 3.7, promete um desempenho superior a todos os microfones táteis atuais, mas ao menos na época em que foram produzidos os experimentos não era fácil de ser adquirido.

Ademais, os resultados obtidos com o microfone para o *Talkabout* Motorola foram em geral satisfatórios para o objetivo de

deixar como contribuição uma metodologia de construção de uma base de dados Lombard.

d) Locutores e material gravado

Para a construção da base Lombard foram gravadas falas de 22 locutores do sexo masculino e 22 do sexo feminino, ambos nativos da língua brasileira. Cada locutor leu 13 vezes 20 frases foneticamente balanceadas de [104] para o português falado no Brasil.

Além das frases em português os locutores também leram, 13 vezes, 20 frases em inglês, formadas por dígitos conectados da Base Aurora-1. Essas gravações em outra língua possibilitam estudos acerca do EL quando comparado à fala na língua nativa e não nativa.

Tanto as frases no português brasileiro quanto as sequências de dígitos conectados da base de dados Aurora-1 foram selecionadas pelo seu conteúdo fonético e pelo comprimento. No Apêndice C são apresentadas as 20 frases para o português e as 20 frases para o inglês.

e) Ruídos e procedimentos de gravação

Para constituir um ambiente ruidoso nos experimentos de gravação e obter o efeito Lombard nos locutores foram utilizados quatro tipos de ruídos:

- *metal-cutting*;
- *tunnel-front*;
- *tunnel-inside*;
- *crowd-children*.

Esses ruídos foram emitidos a três níveis: baixo (correspondendo a 70 dB), médio (correspondendo a 80 dB) e alto (correspondendo a 90 dB).

Para medir os níveis de SNR foi utilizado um aparelho de medição digital de pressão sonora (“decibelímetro” digital).

A medida da pressão sonora foi feita para cada nível e tipo de ruído, próximo ao ouvido de cada locutor, regulando a altura do pedestal mostrado anteriormente na Figura 3.5.

Na sequência, após colocar o microfone de garganta no locutor regulou-se a altura do pedestal e efetuou-se a medida da pressão sonora emitida pelo ruído.

A Figura 3.8 mostra o medidor de pressão sonora ao lado do microfone de garganta utilizado.

Figura 3.8: “Decibelímetro” e microfone de garganta utilizado nas gravações

Para calibrar o nível de pressão sonora, o dial de volume do equipamento utilizado (*laptop computer*) para emitir os ruídos era modificado. Assim que o nível de pressão sonora desejado era estabelecido, as gravações começavam *sem interrupção*, até completar as 20 frases em português e as 20 frases em inglês.

A interrupção não era realizada porque exigiria um tempo maior de permanência dos locutores no local de gravação. Desta forma, foi decidido efetuar as gravações até o final de todas as frases, para depois editar os sinais de fala, recortando cada trecho do sinal de interesse e nomeando-o.

Apesar do custo em termos de trabalho de edição de áudio, isso facilitou que os locutores mostrassem mais ânimo para aceitar realizar as gravações, pois o fator tempo, que foi em média de 50 minutos, sempre foi questionado no momento de aceitar ou não ser voluntário, mesmo que para isso eles recebessem uma "deliciosa barra de chocolate de 170 g", como foi o caso.

Isso posto, depois de todas as gravações realizadas, os sinais de fala foram editados, resultando em 44 locutores vezes 40 frases vezes 13 repetições, num total de 22.880 locuções.

As partes referentes ao treinamento e ao teste da fala Lombard foram divididas na proporção de 70% e 30%, respectivamente. Esta proporção se aproxima das quantidades do projeto feito para a base de dados Aurora-1, que foi de cerca de 68% para treino e 32% para teste.

Na realidade, no Brasil existe uma lacuna muito grande nas experiências feitas para estudos de sistemas de reconhecimento de fala,

sejam eles realizados sob base de dados limpa, robusta com degradação de ruídos ou com degradação devido ao efeito Lombard. Essa lacuna é provocada pela falta de uma base de dados única, que seja pública, e que os pesquisadores da área no Brasil possam usar para comparar os seus resultados.

É devido a este detalhe que esta metodologia de construção de uma base de dados Lombard é deixada para ser utilizada caso algum estudo sobre esta área seja feito. Assim, ao menos pode-se ter pesquisas feitas com a mesma metodologia na construção da base de dados.

Desta forma, optou-se por trabalhar com a base de dados Aurora-1, que é uma base de dados robusta baseada na degradação de ruídos e que é disponibilizada ao público, e da qual é possível tirar conclusões mais seguras.

A aplicação da metodologia para construção de uma base Lombard ficará neste trabalho como uma sugestão para trabalhos futuros, que poderão vir da própria base de dados aqui construída.

4. Modelos para avaliação dos dados

Conforme exposto anteriormente, este trabalho visa a contribuir com a área de reconhecimento automático de fala robusto, fazendo um estudo do comportamento de quatro tipos de ruídos não presentes na base de dados Aurora-1.

Promover um estudo e análise do comportamento de ruídos sem ter algum objetivo que o justifique – não faria sentido. Neste trabalho, entretanto, esse estudo é conduzido de forma a permitir uma análise e avaliação de uma base de dados robusta, concluindo se o sistema está bem ou mal treinado para um determinado tipo de ruído.

No projeto de construção da base de dados Aurora-1, assim como em muitos trabalhos na área de RAFR, como [12][13][89], etc., não informam quais os procedimentos adotados para a *avaliação das referidas bases de dados*. Esses trabalhos definem se a base está bem ou mal treinada pelo simples desempenho do sistema, o que leva a acreditar que uma metodologia organizada e baseada em critérios matemáticos e/ou estatísticos ou não existe ou não está disponível publicamente.

Desta forma, neste trabalho apresenta-se uma metodologia que adota um modelo matemático/computacional baseado no ajuste de curvas pelo método dos mínimos quadrados, que é mais rigoroso e preciso do que simplesmente avaliar os resultados do sistema para cada tipo de ruído ou conjunto deles.

Assim, estudos dessa natureza são muito importantes para a área de RAFR, e neste capítulo são apresentadas as ferramentas

matemáticas que são empregadas para a proposta de metodologia de avaliação de bases de dados ruidosas, como a base Aurora-1.

Para fazer análise dos dados de saída dos sistemas usados neste trabalho (WI007 e WI008), três modelos de curva foram adotados:

a) um modelo mostra o comportamento mediante a qualidade dos sinais de fala e taxa de acerto, ou seja: curvas Pesq vs. TA. Este modelo é considerado aqui o mais importante devido à relação da qualidade dos sinais de fala no comportamento dessas curvas ser usado para a caracterização dos ruídos novos e, consequentemente, a metodologia de avaliação de bases robustas, que é um dos objetivos mais importantes desta Pesquisa.

b) o segundo modelo procura esboçar o comportamento dos sistemas diante do nível de degradação e da qualidade dos sinais de fala: curvas SNR vs. Pesq.

c) o terceiro modelo apresenta o comportamento dos sistemas diante do nível de degradação dos sinais de fala e da Taxa de Acerto: curvas SNR vs. TA. Este modelo e as curvas SNR vs. Pesq serão usados para comparação com as curvas dos ruídos da base Aurora-1 para ajudar a validar a metodologia empregada na produção do teste com os ruídos novos.

Além disso, uma função logística de três parâmetros proposta no trabalho de [6] é usada para modelar as curvas Pesq vs. TA do comportamento dos sistemas para todos os ruídos.

Neste capítulo, na seção 4.1, além de apresentar e estudar essa função logística, na subseção 4.1.2 é proposta uma solução numérica para determinação de seus parâmetros. Ainda, na subseção 4.1.3, é apresentado um Método de Ajuste Inicial Logístico (Mail) que, mediante a solução da referida função logística e com o uso dos mínimos quadrados [7][8], define as curvas de ajuste da relação Pesq vs. TA.

As curvas de ajustes Pesq vs. TA dos sistemas obtidas pelo Mail são empregadas mais adiante (no capítulo 5) para analisar o comportamento dos ruídos e concluir se a base de dados está ou não bem treinada para os ruídos novos, considerando uma precisão desejada.

4.1 Modelo logístico para a curva Pesq vs. TA

Para traçar as curvas Pesq vs. TA, assim como as demais curvas, basta usar os pontos experimentais obtidos na saída do reconhecedor, após cada teste, para cada *front-end* usado.

Caso, no entanto, seja necessário estimar algum valor fora da faixa dos pontos experimentais, é preciso fazer uso de algum método de ajuste que possibilite calcular esses valores.

Ao processar os dados experimentais constatou-se, como em [6], que a tendência desses dados pode ser aproximada por uma curva logística. Assim, seria possível definir dois tipos de comportamento:

1. para valores Pesq altos, uma variação no valor desta pontuação não muda de forma significativa o valor da taxa de acertos;

2. para valores baixos da pontuação Pesq, uma pequena alteração provoca uma grande diferença no desempenho do sistema.

A região de interesse nesta análise seria o "cotovelo" da curva, em que se nota uma mudança de um comportamento para o outro.

Isto seria útil, por exemplo, como mais uma alternativa em sistemas de *call center*, que poderiam identificar automaticamente o desempenho esperado do sistema de reconhecimento e tomar a decisão de solicitar ao usuário que ligue novamente ou transferir para um atendente humano, se o sinal estiver muito ruim.

Em [6] foi sugerida uma função logística para relacionar os valores Pesq com a taxa de acerto de SRAF. No trabalho de [6] foi comentado que a referida função logística é mais adequada ao comportamento Pesq vs. TA do que uma função polinomial de 4ª ordem do trabalho de Sun e seus colegas, apresentado em [105].

Na presente pesquisa, é aproveitado o modelo matemático dessa função logística proposta em [6] para fazer o estudo da caracterização dos ruídos novos em relação aos ruídos da base de dados Aurora-1 e, consequentemente, analisar a base de dados ruidosa para quatro novos ruídos.

Sabe-se que as curvas logísticas mais comumente empregadas para representar a resposta de experimentos são compostas de funções a dois parâmetros, entretanto, [6] fez uso de uma função logística com o acréscimo de um terceiro parâmetro que possibilitasse um ajuste fino no deslocamento vertical da curva logística para sistemas de RAF.

A função logística proveniente de [6], e usada aqui, é apresentada na equação (4.1).

$$f(x) = \left(\frac{1}{1 + e^{b - a.x}} - c \right).100 \qquad (4.1)$$

Os valores Pesq correspondem aos valores de x, assim como os valores TA correspondem aos valores de $f(x)$.

4.1.1 Característica dos parâmetros da função logística

Como se pode notar da equação (4.1), $f(x)$ é constituída de três parâmetros. O parâmetro ***a*** dá a inclinação da curva, o parâmetro ***b*** controla o deslocamento horizontal, enquanto o parâmetro ***c*** influi num leve deslocamento vertical.

As curvas exibidas na Figura 4.1, construídas com o parâmetro ***a*** variável e os parâmetros ***b*** = 10.5 e ***c*** = $2x10^{-5}$ constantes, mostram que o parâmetro ***a*** define a característica de inclinação da curva logística da equação (4.1).

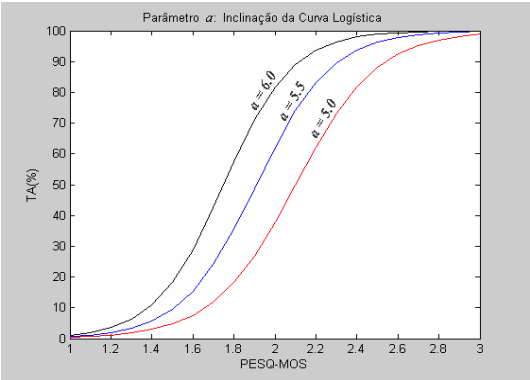


Figura 4.1: Efeito do parâmetro *a* na característica da curva logística

Olhando a Figura 4.1 da esquerda para a direita, pode-se observar que a primeira curva, para $\boldsymbol{a} = 6$, a segunda curva com $\boldsymbol{a} = 5.5$ e a terceira curva para $\boldsymbol{a} = 5.0$, resultam em inclinações diferentes.

Assim sendo, se considerar as curvas da Figura 4.1 como as respostas de três SRAF distintos, pode-se afirmar que no sistema com $\boldsymbol{a} = 6.0$, para um mesmo valor Pesq, se obtem um melhor desempenho. Desta forma, fisicamente o valor do parâmetro $\boldsymbol{a}$ tem um importante significado. Na realidade, é desejado que o cotovelo superior tenha a queda abrupta nos menores valores Pesq, ou seja, para valores de $\boldsymbol{a}$ maiores quanto possível.

O significado físico do parâmetro $\boldsymbol{b}$, é igualmente importante, pois como este parâmetro desloca a curva horizontalmente, influi diretamente na relação Pesq vs. TA, e, consequentemente, seu valor também afeta o desempenho do sistema.

A Figura 4.2 mostra o comportamento de três curvas com o parâmetro b variável.

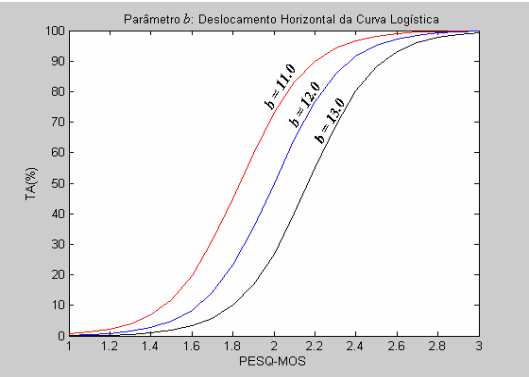


Figura 4.2: Efeito do parâmetro *b* na característica da curva logística

A primeira curva, da Figura 4.2, da esquerda para a direita foi obtida com o parâmetro b = 11, a segunda com o parâmetro b = 12 e a terceira com o parâmetro b = 13. Os demais parâmetros (a = 6.0 e c = $2x10^{-5}$) foram mantidos constantes.

Assim, analogamente à análise feita para o parâmetro anterior, se forem consideradas as curvas da Figura 4.2 como as respostas de três SRAF distintos, pode-se concluir que no sistema com menor valor escalar (***b*** = 11) se obtém um melhor desempenho.

A Figura 4.3 mostra o comportamento de três curvas logísticas com a variação do parâmetro ***c,*** mantendo ***a*** = 5.0 e ***b*** = 7.0 constantes.

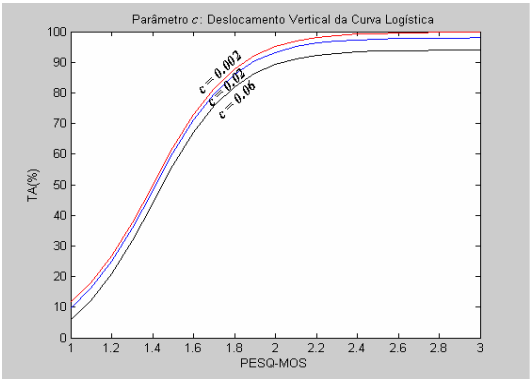


Figura 4.3: Efeito do parâmetro *c* na característica da curva logística

Analisando a Figura 4.3 para o parâmetro ***c***, da mesma forma que na análise dos demais parâmetros, pode-se notar que a primeira curva, construída com ***c*** = 0.002, a segunda curva com ***c*** dez vezes maior, e a terceira curva com c = 0.06, diferenciam-se por um deslocamento vertical fino.

Esse deslocamento vertical fino corresponde fisicamente à taxa de acerto do sistema de RAF, e é aproveitado para se obter uma melhor aproximação dos pontos experimentais e fazer inferências. Observando, porém, as curvas da Figura 4.3 como a resposta de sistemas distintos, pode-se afirmar que quanto menor for o valor de ***c*** maior é a taxa de reconhecimento para um mesmo valor Pesq.

Consequentemente há uma relação importante entre os três parâmetros da equação (4.1), em que a mudança de valor de um deles afetará os demais e, portanto, há um valor simultâneo para cada parâmetro que representa melhor o comportamento de um SRAF.

Assim, a equação (4.1) é empregada neste trabalho para traçar as curvas Pesq vs. TA de todos os testes e pode-se predizer valores dos sistemas usados (WI007 e WI008).

Como abordado no início desse capítulo, é necessário identificar quais os valores dos três parâmetros que, simultaneamente, proporcionam a melhor curva de ajuste para cada sistema empregado. E, para obter essa curva de ajuste, na subseção 4.1.2, é proposto um algoritmo para a solução numérica da equação (4.1), e na sequência, na subseção 4.1.3, é apresentado um modelo matemático/computacional para definir as curvas de ajuste.

4.1.2 Algoritmo para solução numérica da função logística

Conforme mencionado anteriormente, a função logística que é usada neste trabalho é sugerida em [6], porém os autores não apresentam o método aplicado para obter os valores dos três parâmetros da função.

Diante disso, aqui é apresentada uma contribuição ao trabalho de [6], propondo um algoritmo para o cálculo numérico desses parâmetros. O algoritmo para a solução numérica da curva logística da equação (4.1) é simples, levando em consideração que o parâmetro ***c*** de (4.1), observada a taxa de reconhecimento para vários valores experimentais, é próximo ao valor zero para sinais muito degradados (Pesq próximo de 1.0).

Como a tendência dos valores experimentais da taxa de reconhecimento varia na escala de 0% a 100%, e como o parâmetro ***c*** dá a característica do deslocamento vertical da curva logística, este parâmetro corresponde diretamente a valores da escala da taxa de reconhecimento. Isto pode ser observado na Figura 4.3 apresentada na subseção 4.1.1.

Além desse raciocínio, no entanto, torna-se necessário procurar um modelo matemático que, a partir da equação (4.1), possibilite calcular todos os parâmetros. Como ***c*** é um parâmetro que pode ser considerado o ponto de partida para computar os demais, pode-se descrever o desenvolvimento matemático mostrado a seguir.

Reescrevendo a equação (4.1),

$$f(x) = \left(\frac{1}{1 + e^{b - a.x}} - c \right).100 \tag{4.1}$$

De (4.1) pode-se obter (4.2),

$$b - ax = \ln\left(\frac{100}{f(x) + 100\,c} - 1 \right) \tag{4.2}$$

A equação (4.2) ainda precisa de uma solução para os parâmetros ***a***, ***b*** e ***c***. Assim, para solucionar (4.2), são usados três pares ordenados de um experimento qualquer de um SRAF, ou seja: (x_1, y_1), (x_2, y_2) e (x_3, y_3), onde $y_n = f(x_n)$.

Desta forma, aplicando-se a devida correspondência dos pontos de y_n em (4.2), se obtém a equação (4.3):

$$b - ax_n = \ln\left(\frac{100}{y_n + 100\ c} - 1\right) \tag{4.3}$$

Na sequência, para os pares (x_1, y_1) , (x_2, y_2) e (x_3, y_3), o lado direito de (4.3) é renomeado de forma a se obter termos reduzidos para o desenvolvimento da solução de (4.1), conforme descrito nas equações (4.4a), (4.4b), (4.4c) e (4.5a), (4.5b), (4.5c), a seguir.

$$A_1 = \ln\left(\frac{100}{y_1 + 100.c} - 1\right) \tag{4.4a}$$

$$A_2 = \ln\left(\frac{100}{y_2 + 100.c} - 1\right) \tag{4.4b}$$

$$A_3 = \ln\left(\frac{100}{y_3 + 100.c} - 1\right) \tag{4.4c}$$

As equações (4.5a), (4.5b) e (4.5c) vêm da substituição das equações (4.4) na equação (4.3), como segue:

$$b - ax_1 = A_1 \tag{4.5a}$$

$$b - ax_2 = A_2 \tag{4.5b}$$

$$b - ax_3 = A_3 \tag{4.5c}$$

Logo, fazendo operações no sistema acima, pode-se descrever os parâmetros ***a e b*** como nas equações (4.6) e (4.7):

$$a = \frac{A_2 - A_1}{x_1 - x_2} \tag{4.6}$$

$$b = \frac{A_1 - \frac{x_1}{x_2}.A_2}{1 - \frac{x_1}{x_2}} \tag{4.7}$$

Agora é preciso definir uma equação para o parâmetro ***c***. Desta forma, aplicando (4.6) e (4.7) na equação (4.5c), se obtém a equação (4.8):

$$\frac{A_1 - \frac{x_1}{x_2}A_2}{1 - \frac{x_1}{x_2}} - \frac{A_2 - A_1}{x_1 - x_2}x_3 = A_3 \tag{4.8}$$

Considerando as equações 4.9a, 4.9b e 4.9c e aplicando-as em (4.8), é obtida a equação (4.10):

$$K_1 = \frac{x_1}{x_2} \tag{4.9a}$$

$$K_2 = 1 - \frac{x_1}{x_2} \tag{4.9b}$$

$$K_3 = \frac{x_3}{x_1 - x_2} \tag{4.9c}$$

$$\frac{A_1 - A_2K_1}{K_2} + (A_1 - A_2)K_3 - A_3 = 0 \tag{4.10}$$

É preciso ressaltar que as substituições feitas pelas equações (4.4) e (4.9), apesar de não serem necessárias, foram uma opção para se chegar a uma equação mais compacta, mostrada na equação (4.10). Além disso, observe-se que a equação (4.10) não apresenta o parâmetro ***c***. É preciso lembrar, porém, que ***c*** está presente nas equações que definem A_1, A_2 e A_3, portanto ***c*** está implicitamente presente nos termos do lado esquerdo de (4.10), e os parâmetros ***a*** e ***b*** não estão presentes.

Desta forma, a equação (4.10) pode parecer o ponto final da solução de (4.1), mas é apenas uma equação geral para resolver a equação (4.1) proposta em [6], bastando antes substituir os valores de A_n e K_n ($n \in$ N*, $1 \leq n \leq 3$).

Resolver analiticamente (4.10), entretanto, não é trivial. Ao fazer as substituições A_n e K_n em (4.10), é observado uma equação não linear, de difícil solução, conforme mostrado no Apêndice A. No Apêndice A, entretanto, é descrito o raciocínio matemático que define os limites (intervalo dos reais) para a solução do parâmetro ***c***, onde $-1 < \boldsymbol{c} < 1$.

Por outro lado, como comentado anteriormente, com as substituições feitas nos termos A_n e K_n, o lado esquerdo da equação (4.10) ficará somente em função do parâmetro ***c*** da equação (4.1).

Logo, pode-se aplicar um algoritmo, por exemplo - Newton-Raphson (NR) -, usando (4.10) para calcular um valor de ***c*** que corresponda a uma tolerância (ε) desejada, buscando a identidade do lado direito de (4.10). O valor inicial de ε deve estar dentro do intervalo aberto i_1, i_2 =]-1, 1[. A Figura 4.4 mostra um fluxograma da solução

numérica proposta aqui para a equação (4.1), usando (4.10), considerando o algoritmo de NR.

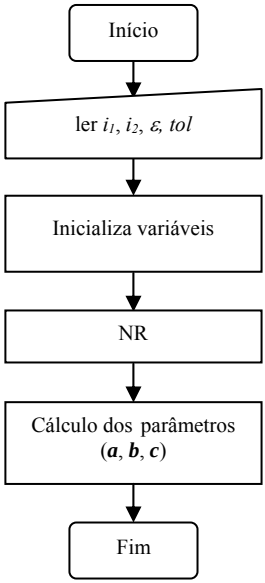


Figura 4.4: Fluxograma da solução numérica proposta para a função logística

Cabe ressaltar que outras soluções podem existir, assim como outras curvas logísticas podem ser usadas para representar o comportamento Pesq vs. TA, porém a ideia da solução numérica de (4.1) apresentada aqui é fácil de ser compreendida e aplicada.

Além disso, o raciocínio do método numérico proposto para solucionar (4.1) pode ser aplicado em outros casos similares, e em outras funções em outras áreas do conhecimento, bastando fazer alguns ajustes, caso necessário.

Por exemplo, em [106] é empregada uma função logística similar para controle dos movimentos de um robô.

O grau de dificuldade para a solução da função logística de [106] pode ser minimizado com o auxílio do algoritmo proposto neste trabalho, no entanto [106] usa uma função logística com dois parâmetros, e, desta forma, a solução usando a *Estimativa pela Máxima Verossimilhança* (EMV) [107][108] pode ser aplicada com mais facilidade do que no caso de uma função logística de três parâmetros.

O método numérico proposto aqui é mais simples do que a aplicação do EMV, pois não usa derivadas parciais. Além disso, o EMV também usa método numérico para estimar valores dos parâmetros, aplicando o algoritmo de Newton Raphson [109][110] que, por sua vez, também depende do cálculo de sucessivas derivadas até a precisão desejada ou chegar no número de iterações permitidas. Entretanto, como a equação (4.10) é uma função monótona, a solução por NR é uma ótima opção.

Ainda, o algoritmo numérico desenvolvido nesta pesquisa, além de simples, converge rapidamente, dependendo apenas da escolha de um valor inicial para o parâmetro $\boldsymbol{c}$, dentro do intervalo definido para os reais, que pode ser facilmente definido conforme Apêndice A.

Desta forma, como os parâmetros da equação (4.1) podem ser obtidos pelo algoritmo proposto na subseção 4.1.2, torna-se necessário utilizar um método para definir as curvas de ajuste para analisar o comportamento dos sistemas WI007 e WI008.

Assim, na subseção 4.1.3 é apresentado um modelo denominado Método de Ajuste Inicial Logístico (Mail) baseado na solução de (4.1) em conjunto com o método dos mínimos quadrados que

permite definir a referida curva de ajuste com uma determinada precisão.

4.1.3 Método de ajuste inicial logístico (Mail)

Muito embora a função mostrada na equação (4.1) possa representar os valores Pesq vs. TA, esta função é apenas uma aproximação para os resultados práticos de sistemas de RAF. Além disso, dependendo dos valores dos parâmetros ***a***, ***b*** e ***c*** da equação (4.1) a função pode não se ajustar muito bem aos valores práticos obtidos.

Assim, para traçar uma curva de ajuste, além da necessidade de se obter os parâmetros de (4.1) precisa-se, mediante algum método de ajuste, definir a curva que mais se aproxima das respostas práticas das saídas do SRAF.

A partir dos pontos experimentais o método mais indicado para a definição dos parâmetros da curva de (4.1), é o método da *Estimação pela Máxima Verossimilhança* (EMV) [107][108], já que o *Método da Regressão Linear* (MRL) [111], que usa o *Método dos Mínimos Quadrados* (MMQ) [6][7][112], não é indicado.

Por outro lado, o método da Estimação pela Máxima Verossimilhança ou *Maximum Likelihood Estimation* (MLE) [113], para poucos pontos, conforme orientado nas especificações do Minitab [114], que é um software estatístico feito por especialistas da área, não é muito preciso.

Além disso, vale registrar que as funções logísticas que esses métodos costumam processar têm apenas dois parâmetros e somente um

termo, ou seja, não existe um segundo termo na função logística, como o parâmetro ***c*** de (4.1).

Este aspecto dessas funções logísticas de dois parâmetros resulta em uma facilidade matemática enorme em relação à equação (4.1), pois ficam fáceis de serem produzidas as Funções Densidade de Probabilidade (FDP) [115][116][117] e as derivadas parciais para a EMV.

Por outro lado, o parâmetro ***c*** permite um ajuste fino na relação Pesq vs. TA que é fundamental para o ajuste dessa função aos pontos experimentais. Conforme visto na subseção 4.1.2, porém, ao tentar solucionar a equação (4.1) constata-se que se trata de um problema não linear de difícil solução.

Além disso, o número de pontos que se tem em cada teste é pequeno para a aplicação do EMV (apenas 6 pontos). E, se usarmos o MRL para obter a solução de (4.1) com o MMQ, não é obtido o valor do parâmetro ***c***.

Em [118] é sugerido um modelo logístico de três parâmetros denominado ML3, que calcula a probabilidade de um indivíduo j com habilidade μ_j responder corretamente o item i, e é chamado de função resposta do item.

Os três parâmetros dessa função são: a) dificuldade, b) discriminação e c) probabilidade de resposta correta por indivíduos de baixa habilidade. Como, no entanto, os parâmetros do ML3 são conhecidos, torna-se fácil determinar o comportamento da curva logística.

Desta forma, como não foi encontrado na literatura um método que solucione a equação (4.1) satisfatoriamente para obter a melhor curva de ajuste para a avaliação do comportamento dos ruídos, foi adotado o Método de Ajuste Inicial Logístico (Mail) proposto neste trabalho, que usa a solução numérica para a equação (4.1) proposta na subseção 4.1.2.

O nome Mail deriva da necessidade do uso de uma curva de ajuste inicial escolhida de forma subjetiva.

Neste viés, é possível usar a solução numérica de (4.1), proposta neste trabalho, para determinar os três parâmetros, ***a***, ***b*** e ***c***, e, através do MMQ definir a melhor curva para uma precisão requerida, definida aqui como Δ.

Assim, no método proposto (Mail) calcula-se os parâmetros da função logística de (4.1) sem a necessidade de fazer cálculos de derivadas ou transformação linear como no MRL e no EMV, usando somente três pontos e as equações apresentadas na seção 4.2.

É preciso ressaltar que, para o *Método de Ajuste Inicial Logístico* (Mail) funcionar corretamente, como abordado anteriormente, há a necessidade de se fazer um ajuste inicial logístico, que é subjetivo.

Isto significa que é necessário fornecer os três primeiros pontos definidos subjetivamente como sendo os melhores para que a curva de ajuste logística tenha a menor soma dos desvios quadráticos, ou o menor valor da aplicação do MMQ.

A soma dos desvios quadráticos, ou MMQ, utilizada neste estudo é calculada conforme a equação (4.11), e deve ser menor ou igual ao desvio máximo Δ.

$$dsv = \sum_{i=1}^{n} (y_i - y_{ref})^2 \qquad (4.11)$$

Onde:

dsv = somatório dos desvios quadráticos;

n = número de pontos;

y_i = pontos da curva que se deseja avaliar;

y_{ref} = pontos da curva de referência.

Desta forma, quando $dsv \leq \Delta$, está definida a melhor curva logística para os pontos experimentais em questão.

Obviamente que, para obter os três pontos para o funcionamento inicial do Mail, não é necessário que a curva passe exatamente nos pontos experimentais.

Se, caso a melhor curva logística intuitivamente observada tenha de passar entre pontos considerados próximos, um bom ponto de partida é calcular a média aritmética entre esses pontos.

Em testes comparativos entre o Mail, MRL e EMV, usando o Minitab [114], devido ao fato de os valores da soma dos desvios quadráticos das curvas do Mail serem menores, foi observado que ocorreu uma melhor aproximação da curva de ajuste obtida pela aplicação do Mail do que pela aplicação do MRL e EMV.

A Figura 4.5 apresenta um algoritmo, na forma de fluxograma, para aplicação do Mail para N pontos.

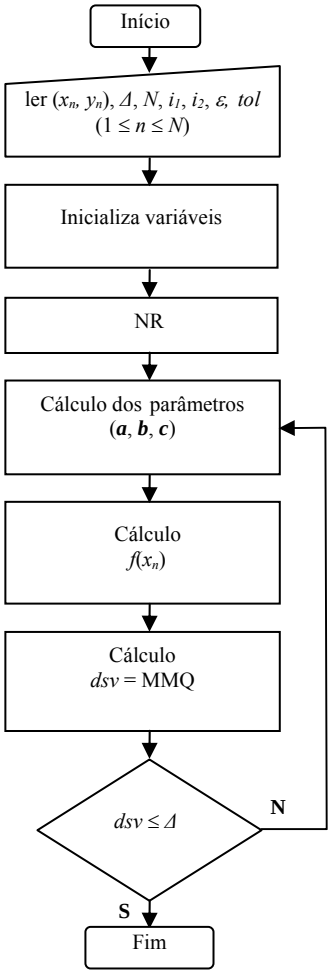


Figura 4.5: Fluxograma do Mail

É preciso deixar claro que para garantir que o laço não seja repetido indefinidamente um procedimento, como limite de iterações, deve ser definido.

Nos testes comparativos entre o Mail, MRL e EMV expressados no parágrafo anterior, deve-se considerar, também, que na Estimação pela Máxima Verossimilhança, além de se ter poucos pontos, ao usar a equação (4.1) há o problema da complexidade matemática que envolve a solução e diferenciação parcial provenientes dessa equação. E, devido a esta complexidade, também pelo EMV torna-se necessário solucionar a equação (4.1) por método numérico.

Caso, porém, ***c*** seja inicialmente desprezado por ser um valor pequeno na solução de (4.1), os parâmetros ***a*** e ***b*** podem ser facilmente obtidos pelo MRL, porém ***c*** ainda fica indeterminado, a não ser que ***a*** e ***b*** sejam aplicados em (4.1) para a obtenção do parâmetro ***c***, o que não é tão preciso.

4.2 Modelo da curva SNR vs. Pesq

Como mencionado anteriormente, um dos modelos para auxiliar na análise do comportamento dos ruídos usados neste trabalho, procura esboçar a característica dos sistemas diante da qualidade dos sinais de fala limpos (Pesq) para com os mesmos sinais, porém degradados pelos níveis de ruído empregados para os testes. Desta forma, a relação SNR vs. Pesq é obtida a partir do sinal de fala limpo (sinal de referência) versus o sinal degradado, em todos os quatro testes experimentais realizados.

A Figura 4.6 mostra a característica típica do comportamento da relação sinal-ruído para a qualidade dos sinais de fala: curva SNR vs. Pesq usando os ruídos do *testa* e do *testd*.

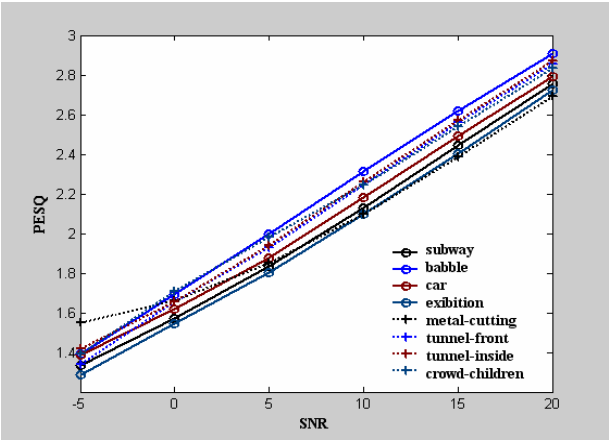


Figura 4.6: Característica das curvas SNR vs. Pesq

Pode-se perceber das curvas da Figura 4.6, que existe uma tendência linear no comportamento da relação SNR vs. Pesq. Entretanto, a análise do comportamento dessas curvas é feita com maior rigor no capítulo 5.

4.3 Modelo da curva SNR vs. TA

Outro modelo para auxiliar na análise do comportamento dos ruídos novos, procura esboçar a característica dos sistemas ante o nível de degradação dos sinais de fala para com a taxa de acerto do sistema.

Assim sendo, a relação SNR vs. TA é obtida a partir do sinal de fala degradado versus a taxa de acerto, também em todos os quatro testes experimentais.

É preciso deixar claro que os modelos empregados neste estudo já vêm sendo utilizados em outras pesquisas, tal como em [6],

[74], entre outros. Assim, as contribuições feitas aqui não são especificamente os modelos apresentados para a análise e avaliação dos resultados, mas sim a metodologia de avaliação envolvendo esses modelos com um maior rigor matemático e computacional.

A metodologia empregada aqui, portanto, é uma das contribuições deste trabalho. Outras contribuições são: o algoritmo para a solução numérica da equação (4.1), que foi apresentado anteriormente na subseção 4.1.2; o modelo Mail apresentado na seção 4.1.3 para ajuste da curva logística.

Para finalizar, todos os modelos apresentados aqui são aplicados no capítulo 5, em que os resultados também são mostrados. A metodologia de avaliação foi aplicada aos ruídos novos e resultou na mudança da base de dados na qual foi incluído o ruído *metal-cutting*, melhorando a taxa de acerto para as locuções de teste, no sistema WI007, em mais de 3%.

5. Resultados

Neste capítulo primeiramente são apresentados os resultados da taxa de acerto do *testd* (devido aos ruídos novos) na forma de tabelas, tanto para o WI007 quanto para o WI008, sob os diversos cenários. É preciso ressaltar que o modelo adotado para o presente trabalho é independente de locutor.

Os resultados de cada teste da base Aurora-1 (*testa*, *testb* e *testc*), na forma de tabelas, são apresentados no Apêndice D. Esses resultados foram obtidos a partir do processamento dos dados neste trabalho e são iguais aos apresentados pela equipe do projeto da base de dados Aurora, indicando que os procedimentos adotados aqui foram corretos.

Depois das tabelas com os resultados dos sistemas empregando os ruídos do *testd*, são apresentados os gráficos da relação SNR vs. TA, Pesq vs. TA e SNR vs. Pesq. Alguns comentários são feitos paralelamente a cada resultado apresentado.

A partir das curvas Pesq vs. TA é realizada a caracterização dos ruídos em grupamento de famílias. Por meio da caracterização dos ruídos em famílias pode-se visualizar e analisar melhor como é a resposta dos SRAF para os diferentes tipos e níveis de ruídos e se fazer inferências sobre a base de dados robusta.

Assim, conforme mencionado anteriormente, analisando o comportamento de grupamentos de ruídos, pode-se definir testes de calibração e/ou mudanças na base de dados robusta para se obter melhores taxas de acerto.

É preciso lembrar que, conforme explicitado no capítulo 3, o treinamento e o reconhecimento foram feitos com os sinais de fala limpos (condições limpas) e com a adição de ruídos da base Aurora-1 (condição múltipla), para todos os testes. O *testd*, como já referido, foi produzido com os ruídos novos.

5.1 Taxas de acerto sob diversos cenários

Nesta seção os resultados das taxas de acerto na forma de tabelas são apresentados, sob algumas características nem sempre encontradas em conjunto nos trabalhos da área de processamento de fala.

Observando os trabalhos de [1][74], entre outros, percebe-se que os resultados tabelados são dispostos de modo a analisar apenas os tipos e níveis de ruídos, sem a preocupação de dispor dados por faixa de degradação e as médias dessas faixas.

Além disso, aqui é proposta uma tabela que separa em campos distintos com cores diferenciando os dados, sejam esses dados provenientes dos testes, dos níveis de ruídos ou das faixas de degradação dos sinais de fala.

Aqui, portanto, é dada uma pequena contribuição para melhorar um pouco a apresentação dos dados no formato tabelado das taxas de acerto diante dos níveis e faixas de SNR, para que se tenha uma melhor clareza ao observar e analisar essa forma de apresentar os dados.

O entendimento das tabelas de taxas de acerto mostradas nesta seção e no Apêndice D é simples, destacando-se que elas estão divididas em três cores para realçar a representação dos dados:

- Em branco, no topo da primeira e última coluna, é informado o tipo de sistema (*front-end* e *back-end*, respectivamente). O cenário, juntamente com os tipos de teste, são especificados na cor cinza no campo entre essas colunas.
- Na cor cinza também estão os dados referentes às taxas de acerto, por tipo de ruído, e por teste/ruídos.
- A degradação, por nível e por faixa, está na coluna do lado esquerdo, abaixo da especificação do tipo de *front-end*.
- Na cor azul-claro estão as médias das taxas de reconhecimento por nível e faixa de SNR (última coluna da direita). As médias por tipo de ruído estão nas quatro colunas alinhadas aos tipos (de ruídos) especificados. Já as médias por faixa de SNR situam-se na última coluna inferior direita.

Desta forma, seguindo o modelo proposto na seção 5.1, na subseção 5.1.1 as Tabelas 5.1a e 5.1b apresentam os resultados das taxas de acerto para o *testd* sob condições limpas.

5.1.1 Testd sob condições limpas usando WI007 e WI008

a) Usando WI007

Tabela 5.1a
Taxa de acerto para o *testd* usando F-E WI007 e treinamento sob condições limpas

WI007 FRONT-END		**Testd – Treinamento sob Condições Limpas Taxa de Acerto (%) Por Tipo de Ruído**				**HTK BACK-END**	
Degradação		***Metal-Cutting***	***Tunnel-Front***	***Tunnel-Inside***	***Crowd-Children***	**Média (%)**	
Nível de SNR (dB)	**Clean**	98,93	99,00	98,96	99,20	**99,02**	**Por Nível**
	20	90,29	94,86	95,90	94,13	**93,80**	
	15	85,29	86,21	83,61	88,72	**85,96**	
	10	75,39	65,15	53,93	70,26	**66,18**	
	5	57,68	34,04	23,74	42,88	**39,59**	
	0	34,62	10,76	9,22	20,54	**18,79**	
	-5	13,01	6,54	8,04	10,94	**9,63**	
		Média (%)					
Faixa de SNR (dB)	**15 a 20**	87,79	90,54	89,76	91,43	**89,88**	**Por Faixa**
	10 a 20	83,66	82,07	77,81	84,37	**81,98**	
	5 a 20	77,16	70,07	64,30	74,00	**71,38**	
	0 a 20	68,65	58,20	53,28	63,31	**60,86**	
	-5 a 20	59,38	49,59	45,74	54,58	**52,32**	

b) Usando WI008

Tabela 5.1b
Taxa de acerto para o *testd* usando F-E WI008 e treinamento sob condições limpas

WI008 FRONT-END		**Testd – Treinamento sob Condições Limpas Taxa de Acerto (%) Por Tipo de Ruído**				**HTK BACK-END**	
Degradação		***Metal-Cutting***	***Tunnel-Front***	***Tunnel-Inside***	***Crowd-Children***	**Média (%)**	
Nível de SNR (dB)	**Clean**	99,08	99,00	99,05	99,23	**99,09**	**Por Nível**
	20	92,02	96,48	97,95	95,31	**95,44**	
	15	89,09	96,82	97,38	95,12	**94,60**	
	10	85,50	95,26	96,60	93,07	**92,61**	
	5	79,23	90,31	92,92	85,05	**86,88**	
	0	67,79	74,04	75,14	68,09	**71,27**	
	-5	43,68	44,02	33,26	43,60	**41,14**	
		Média (%)					
Faixa de SNR (dB)	**15 a 20**	90,56	96,65	97,67	95,22	**95,02**	**Por Faixa**
	10 a 20	88,87	96,19	97,31	94,50	**94,22**	
	5 a 20	86,46	94,72	96,21	92,14	**92,38**	
	0 a 20	82,73	90,58	92,00	87,33	**88,16**	
	-5 a 20	76,22	82,82	82,21	80,04	**80,32**	

Na Tabela 5.1a, para treinamento em condições limpas e o teste feito com os ruídos novos usando o WI007, pode-se perceber que há uma maior queda na taxa de acerto comparando com a Tabela 5.1b, que usa o WI008. Este aspecto já era esperado, devido às técnicas de redução de ruído empregadas pelo WI008. É preciso lembrar que o WI007 não utiliza essas técnicas, com a extração de características dos sinais de fala sendo feita pelo MFCC.

Pode-se perceber das Tabelas 5.1a e 5.1b que, sem a presença de ruído ambiental (sem degradação ou Clean) e sob treinamento em condições limpas, todos os ruídos novos apresentam resultados similares àqueles dos ruídos dos testes da base Aurora-1, apresentados no Apêndice D, onde a média da taxa de acerto é superior a 99%. Conforme esperado, quando a degradação dos sinais de fala do *testd* é elevada do nível de 20 dB até o nível de -5 dB com os ruídos novos, observa-se uma queda na taxa de acerto que se acentua com o aumento da degradação dos sinais de fala do ***testd***. Este aspecto também é verificado para cada faixa SNR de degradação desses sinais de fala.

É preciso observar que os dados "*Clean*" apresentam diferentes respostas, devido ao fato que as locuções são diferentes para cada tipo de ruído.

5.1.2 Testd sob condições múltiplas usando WI007 e WI008

a) Usando WI007

Tabela 5.2a
Taxa de acerto para o *testd* usando F-E WI007 e treinamento sob condições múltiplas

WI007 FRONT-END		**Testd – Treinamento sob Condições Múltiplas Taxa de Acerto (%) Por Tipo de Ruído**				**HTK BACK-END**	
Degradação		***Metal-Cutting***	***Tunnel-Front***	***Tunnel-Inside***	***Crowd-Children***	**Média (%)**	
Nível de SNR (dB)	**Clean**	98,68	98,52	98,39	98,49	**98,52**	**Por Nível**
	20	93,69	96,70	97,62	95,34	**95,84**	
	15	92,18	97,19	97,38	95,93	**95,67**	
	10	90,54	94,93	96,63	94,44	**94,14**	
	5	85,90	85,93	92,05	88,35	**88,06**	
	0	77,40	61,57	68,24	75,61	**70,71**	
	-5	55,39	25,77	25,40	53,29	**39,96**	
		Média (%)					
Faixa de SNR (dB)	**15 a 20**	92,94	96,95	97,50	95,64	**95,75**	**Por Faixa**
	10 a 20	92,14	96,27	97,21	95,24	**95,21**	
	5 a 20	90,58	93,69	95,92	93,52	**93,43**	
	0 a 20	87,94	87,26	90,38	89,93	**88,88**	
	-5 a 20	82,52	77,02	79,55	83,83	**80,73**	

b) Usando WI008

Tabela 5.2b
Taxa de acerto para o *testd* usando F-E WI008 e treinamento sob condições múltiplas

WI008 FRONT-END		Testd – Treinamento sob Condições Múltiplas Taxa de Acerto (%) Por Tipo de Ruído				HTK BACK-END	
Degradação		*Metal-Cutting*	*Tunnel-Front*	*Tunnel-Inside*	*Crowd-Children*	Média (%)	
Nível de SNR (dB)	Clean	99,02	98,79	98,93	99,14	**98,97**	Por Nível
	20	95,02	96,79	98,10	95,80	**96,43**	
	15	93,69	97,40	97,74	96,33	**96,29**	
	10	92,67	96,76	97,38	95,12	**95,48**	
	5	88,44	94,68	95,57	89,78	**92,12**	
	0	80,62	84,13	87,13	78,87	**82,69**	
	-5	58,79	59,55	51,25	57,89	**56,87**	
		Média (%)					
Faixa de SNR (dB)	15 a 20	94,36	97,10	97,92	96,07	**96,36**	Por Faixa
	10 a 20	93,79	96,98	97,74	95,75	**96,07**	
	5 a 20	92,46	96,41	97,20	94,26	**95,08**	
	0 a 20	90,09	93,95	95,18	91,18	**92,60**	
	-5 a 20	84,87	88,22	87,86	85,63	**86,65**	

5.1.3 Curvas SNR vs. TA para o testd

As curvas SNR vs. TA informam o desempenho do sistema em relação ao nível de degradação dos sinais de fala e para cada tipo de ruído.

Assim sendo, são apresentadas em apenas um gráfico as curvas dos quatro ruídos novos para poder fazer um comparativo do desempenho do sistema em termos de taxa de acerto, por nível de SNR e por tipo de ruído, levando em consideração cada cenário.

As Figuras 5.1a e 5.1b mostram as curvas SNR vs. TA dos quatro ruídos novos, sob condições limpas, usando o WI007 e o WI008, respectivamente.

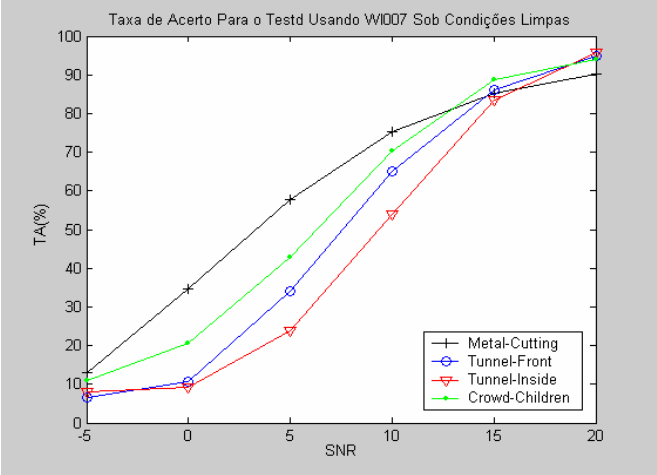


Figura 5.1a: SNR vs. TA usando WI007 sob condições limpas

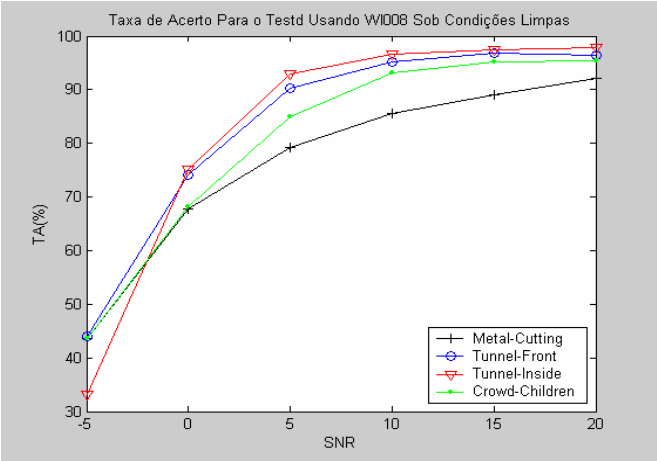


Figura 5.1b: SNR vs. TA usando WI008 sob condições limpas

Pode-se observar na Figura 5.1a que o ruído *tunnel-inside*, para SNR < 15 dB, é o que mais afeta a taxa de acerto, mas a partir da faixa de valores SNR > 15 dB, o ruído *metal-cutting* é o que mais afeta a taxa de acerto do sistema.

Na Figura 5.1b, a partir de valores para SNR > 0 dB, observa-se que o ruído *metal-cutting* é o que mais afeta a taxa de acerto do SRAF para toda a faixa Pesq.

Esta característica do ruído *metal-cutting* também é verificada nas Figuras 5.2a e 5.2b, quando o sistema processa os sinais de fala sob condições múltiplas.

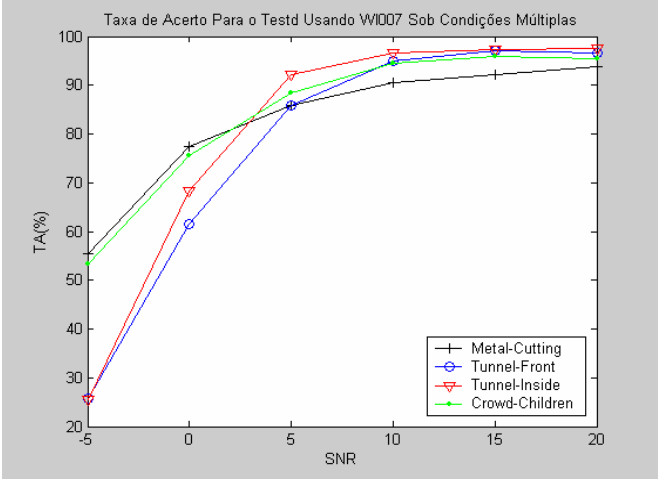


Figura 5.2a: SNR vs. TA usando WI007 sob condições múltiplas

Na Figura 5.2b, a partir de valores para SNR > 0 dB, pode-se observar que o ruído *metal-cutting* em média também é o que mais afeta a taxa de acerto do SRAF.

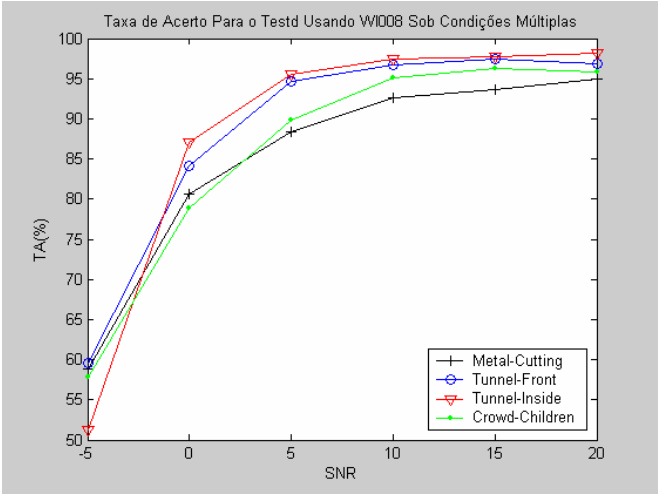


Figura 5.2b: SNR vs. TA usando WI008 sob condições múltiplas

Desta forma, observa-se das Figuras 5.1 e 5.2, que em média, o ruído *metal-cutting*, *para uma base de dados robusta*, é o que mais degrada os sinais de fala, *independentemente do front-end usado.*

Este aspecto é muito importante, pois os sistemas de reconhecimento automáticos de fala robustos, em especial para o ruído *metal-cutting*, dependendo da precisão requerida, têm que melhorar o projeto da base de dados ruidosa.

5.2 Caracterização dos ruídos novos pela relação Pesq vs. TA

Nesta seção, são usados os resultados na forma gráfica da relação Pesq vs. TA para fazer uma investigação comparativa dos testes da base Aurora-1 (*testa* e *testb*) com o novo teste proposto, para

caracterizar os quatro ruídos novos que compõem o *testd* em relação à base de dados robusta.

Desta forma, aqui é sugerida a referida metodologia de avaliação da base de dados ruidosa, a partir da caracterização dos ruídos novos, baseada na referida relação Pesq vs. TA.

Para que a base de dados robusta possa ser avaliada em relação aos ruídos novos, além de usar um comparativo das curvas Pesq vs. TA de cada ruído novo com os ruídos da base Aurora-1, é feita uma avaliação qualitativa subjetiva que busca extrair do referido comparativo os grupamentos de ruídos que tenham comportamento similar, da mesma forma que em [6], porém dentro de um critério matemático/computacional. Este critério está definido na subseção 5.2.1. Os ruídos que têm comportamento similar são considerados pertencentes a uma mesma família ou classe de ruídos como em [6].

A partir da definição dos grupos, famílias ou classe de ruídos com comportamento similar, as curvas Pesq vs. TA dessas famílias também são produzidas e apresentadas na forma gráfica. E, na sequência dessa apresentação gráfica, uma tabela com os pontos experimentais dessas curvas é elaborada.

Observando a curva Pesq vs. TA de cada grupo de famílias, são identificadas as curvas dos ruídos da base Aurora-1 que, por sua vez, são tomadas como referência. Todas as curvas Pesq vs. TA dos ruídos são produzidas com os pontos experimentais.

Desta forma, com base nesses pontos (experimentais) e com o Método de Ajuste Inicial Logístico proposto na subseção 5.1.3, são

produzidas as curvas de ajuste para cada ruído. A partir da curva de ajuste dos ruídos, é calculado o MMQ de cada curva em relação à(s) curva(s) de referência. Este cálculo é mostrado em um gráfico no qual todas as curvas são apresentadas (curvas dos ruídos novos em relação à curva referência).

Baseada no cálculo do referido MMQ de cada curva dos ruídos novos para a curva do ruído referência, uma análise conclusiva sobre a definição qualitativa da família de ruídos é empreendida. Assim, somente serão considerados pertencentes à família de ruídos aqueles que o MMQ ficar dentro de um determinado valor desejado (critério definido como um desvio máximo de 5% ou $dsv \leq 0{,}05$).

Para finalizar a metodologia de análise do comportamento dos ruídos novos em relação aos ruídos da base Aurora-1, um quadro com os valores provenientes da aplicação do Método dos Mínimos Quadrados (MMQ) para todos os ruídos da família definida qualitativamente é apresentado.

Isto posto, baseado no grupamento dos ruídos novos que formam família com os ruídos da base de dados Aurora-1 é que se decidiu pela necessidade ou não de mudanças na base de dados ruidosa.

Assim, esta metodologia foi aplicada para os ruídos do *testd* e, como pode-se notar na subseção 5.2.3, o ruído *metal-cutting*, que não formou família com o critério de $dsv \leq 5\%$, foi adicionado à base de dados, melhorando a performance do sistema WI007. Os demais ruídos do *testd*, como formaram família com os ruídos presentes na base de dados Aurora-1, não foram inseridos na base de dados.

Um diagrama de blocos da metodologia de avaliação dos ruídos novos na base de dados robusta é apresentado na Figura 5.3, a seguir.

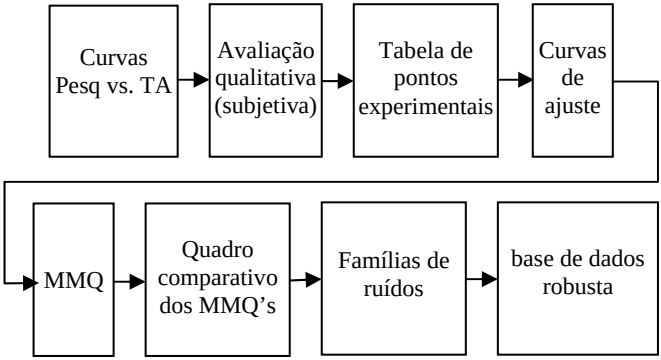


Figura 5.3: Diagrama de blocos da metodologia de avaliação dos ruídos novos na base de dados robusta

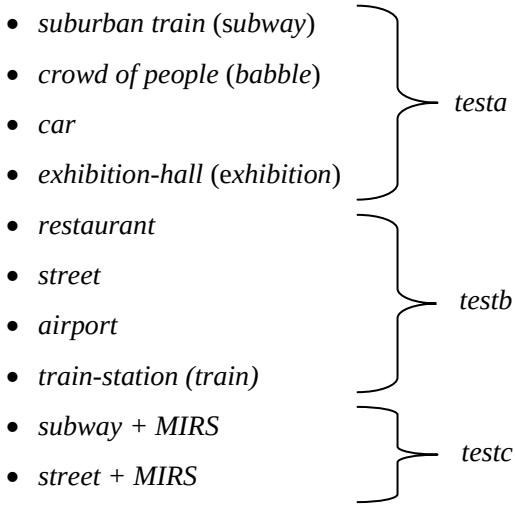
As curvas Pesq vs. TA, conforme supracitado, são produzidas com os dados experimentais, considerando três testes (*testa, testb* e *testd*), e para os cenários:

- *treinamento em condições limpas;*
- *treinamento em condições múltiplas.*

A avaliação qualitativa é feita sobre as curvas Pesq vs. TA, e, a partir dessa avaliação e da tabela de pontos experimentais são obtidas as curvas de ajuste pelo Mail. Na seqüência é produzido os valores de MMQ entre as curvas de ajuste dos ruídos do *testd* e as curvas dos ruídos do "*testa*" e "*testb*" da base de dados Aurora-1.

Um quadro comparativo dos MMQ's de cada curva auxilia na comparação dos valores para a determinação dos ruídos que formam família, e assim fazer inferência sobre a base de dados robusta.

É preciso lembrar que a base de dados Aurora-1 foi produzida com os testes: *testa, testb* e *testc*. O esquema a seguir mostra uma visão geral dos ruídos da base Aurora-1:



Conforme apresentado anteriormente, o *testc* foi produzido com o ruído *subway* e *street*, com ambos sendo submetidos ao filtro Mirs.

O *testd* foi produzido com os ruídos novos:

- *metal-cutting*
- *tunnel-front*
- *tunnel-inside*
- *crowd-children*

A nomenclatura na língua inglesa foi adotada aqui devido à maior facilidade de trabalhar com termos mais curtos. Além disso, mostra-se mais adequada aos termos usados para os ruídos da base de dados Aurora-1.

Em relação aos resultados tabelados, as curvas Pesq vs. TA (e SNR vs. Pesq apresentadas na subseção 5.3) fornecem mais informações sobre o comportamento dos ruídos de cada teste perante a qualidade dos sinais de fala.

Todos os resultados apresentados estão baseados nos dados de saída dos SRAF utilizados: com o uso do *front-end* WI007 e do *front-end* WI008.

Em cada figura nos comparativos gráficos são apresentadas legendas das curvas dos ruídos dos testes da base Aurora-1 e do *testd* para facilitar a visualização, informando o tipo de traçado de cada curva.

Além disso, importa lembrar que os valores Pesq, ferramenta apresentada no capítulo 2, são produzidos com a referência (sinal limpo) em relação ao sinal degradado, antes de passarem pelos *front-ends* mencionados.

Assim, nas seções que seguem, as curvas de cada teste (*testa* e *testb*) da base Aurora-1 com o *testd* são mostradas para facilitar o comparativo do comportamento dos ruídos novos em relação aos demais ruídos e a caracterização em grupos e/ou famílias.

Conforme mencionado anteriormente, a caracterização em grupos e/ou famílias de ruídos pode, a princípio, não ser considerada importante para melhorar a taxa de reconhecimento de sistemas de RAFR. Estudos desse tipo, no entanto, podem auxiliar na avaliação e construção da base de dados robusta, e, desta forma, indiretamente contribuir para uma melhor performance do sistema, e isto é mostrado neste trabalho.

Por exemplo, quando as curvas Pesq vs. TA do comportamento dos ruídos forem diferentes em relação a uma determinada precisão requerida, pode-se misturar à base de dados o ruído que mais afeta o sistema, como o ruído *metal-cutting* comentado antes, e fazer a devida calibração da base de dados para que a taxa de acerto melhore. Sabe-se, porém, que isso não garante o sucesso na melhoria da resposta do sistema, pois nem sempre basta apenas acrescentar os ruídos à base de dados, embora esta seja a principal característica na construção de SRAF robustos.

A base de dados Aurora-1, foi construída com essa característica de construção de SRAF robustos, ou seja, com a mistura de ruídos aos sinais de fala que constituem a base de dados, contudo, no projeto da referida base de dados (Aurora-1) não apresentam um procedimento metodológico de avaliação e construção (da base de dados) fundamentado em modelos matemáticos, como proposto aqui.

5.2.1 Usando o testd e treinamento sob condições limpas

O objetivo neste tópico é definir quais dos ruídos novos do *testd* apresentam comportamento similar aos da base de dados Aurora-1, identificando se pertencem ao mesmo grupo, classe ou família de ruídos. Além disso, pode-se determinar, em relação aos valores Pesq, qual desses ruídos afeta mais a resposta dos sistemas de RAF usados.

Para caracterizar os ruídos em famílias ou classes de ruídos com comportamento similar, são considerados dois modelos de avaliação: 1. modelo de avaliação qualitativo; 2. modelo de avaliação quantitativo.

1. É definido aqui como modelo de avaliação qualitativo, que é subjetivo, o modelo formado pela visualização gráfica das curvas de ajuste dos ruídos. Assim, caso haja uma semelhança no comportamento conclui-se que os ruídos podem formar uma família ou um grupo com características similares.

2. O modelo de avaliação quantitativo é aplicado sobre a resposta da avaliação qualitativa. Dessa forma, pode-se definir com melhor precisão as famílias baseado na avaliação do MMQ.

Explicando: no modelo quantitativo é considerada a proximidade das curvas de ajuste de cada ruído novo (ruídos do *testd*) com a(s) curva(s) de ajuste do(s) ruído(s) de referência (ruídos pertencentes à base de dados Aurora-1) em uma margem de erro de 5% obtida da aplicação do Método dos Mínimos Quadrados (MMQ). Então, por definição, é considerada aqui uma curva **A** com comportamento familiar à curva referência **B** se, e somente se, os pontos da curva de ajuste de **A** apresentarem o valor da soma dos desvios quadrático menor ou igual a 5% ($dsv \leq 0{,}05$) em relação aos pontos da curva referência **B**.

O valor $dsv \leq 0{,}05$ é apenas uma idéia da precisão requerida para confirmar ou não a hipótese de uma curva **A** ter comportamento similar a uma curva **B**. Usando o bom senso, porém, nada impede que esse valor seja modificado para auxiliar na definição da construção de bases de dados robustas.

Para determinar a curva de ajuste é aplicado o método proposto na subseção 5.1.3, Método de Ajuste Inicial Logístico (Mail**)**. Assim, os parâmetros das curvas de ajuste de cada ruído serão obtidos e torna-se possível fazer inferências. Neste sentido, a partir dos pontos experimentais das curvas dos ruídos, é obtida a "melhor" curva de ajuste (respeitada a precisão definida), e, consequentemente, calculados os parâmetros ***a***, ***b*** e ***c*** correspondentes, para medir com mais precisão, com o MMQ, o quanto essas curvas se comportam de modo similar.

Nas sequência são apresentadas as curvas Pesq vs. TA da resposta dos sistemas para cada *front-end* (WI007 e WI008), sob os diversos cenários, e são feitas inferências sobre o comportamento dos ruídos novos empregando o Mail.

5.2.1.1 Usando o testa - testd e WI007

Neste item são apresentadas as curvas dos ruídos do *testa* e do *testd* com o uso do WI007 e condições limpas. Primeiramente é analisado o comportamento das curvas dos ruídos de forma qualitativa; na sequência, a partir desta análise, são identificadas em um único gráfico as curvas que têm comportamento similar para todos os valores Pesq. A partir do gráfico com as curvas de comportamento similar, verificadas qualitativamente, são definidas pelo Mail as curvas de ajuste para cada ruído.

Desta forma, a partir das curvas de ajuste, é calculado o MMQ de cada uma dessas curvas para cada ruído novo em relação àqueles da base de dados Aurora-1, com o critério $dsv \leq 0{,}05$. Assim é decidido se a avaliação qualitativa é aceitável.

A Figura 5.4a mostra o comportamento das curvas Pesq vs. TA dos ruídos do *testa* da base Aurora-1 com os ruídos do *testd*, depois do processo de reconhecimento sob treinamento em condições limpas, usando o WI007.

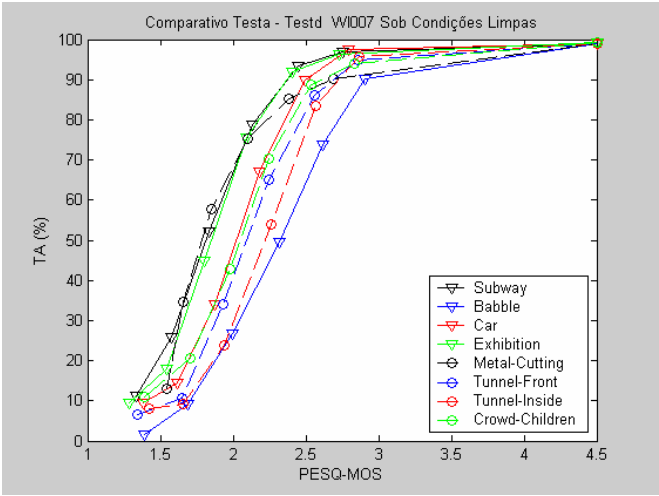


Figura 5.4a: Comparativo testa-testd usando WI007 sob condições limpas

Tomando como base a avaliação qualitativa comentada anteriormente, pode-se dizer que, com a atuação do WI007 sob condições limpas, os ruídos *car*, *tunnel-front* e *crowd-children* podem formar uma família de ruídos.

A Figura 5.4b, apresenta as curvas dos ruídos *car*, *tunnel-front* e *crowd-children* que se avalia, qualitativamente, terem comportamento similar para formar uma família.

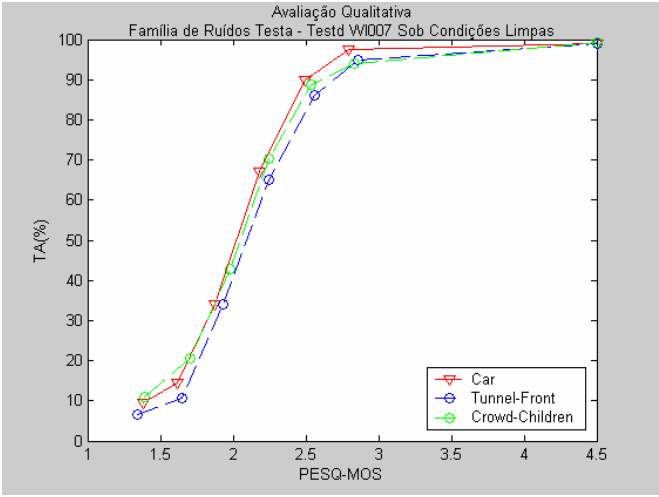


Figura 5.4b: Família de ruídos da avaliação qualitativa do testa-testd usando WI007 sob condições limpas

Seguindo esta mesma forma de avaliação qualitativa, pode-se afirmar que os ruídos *subway* e *exihbition* também formam uma família de ruídos com comportamento parecido. Para não sair do objetivo proposto neste trabalho, entretanto, o foco do estudo é dado apenas sobre aos ruídos novos do *testd*, tomando como referência os ruídos da base Aurora-1 do *testa* e *testb*.

Consequentemente, são definidos os ruídos do *testd* que formam uma família com os ruídos da referida base de dados. Nada impede, contudo, que a metodologia aplicada aqui seja aproveitada para fazer novos comparativos e definir novas famílias de ruídos.

Desta forma, utilizando o Mail é possível, a partir dos pontos experimentais, se obter os valores aproximados dos parâmetros ***a***, ***b*** e ***c***, que definem a curva de ajuste usando a equação (5.1).

Assim, a partir dos pontos experimentais das curvas dos ruídos é obtida a melhor curva de ajuste, e, consequentemente, é possível calcular os parâmetros ***a***, ***b*** e ***c*** correspondentes, para medir com mais precisão o quanto essas curvas se comportam de modo similar.

Na Tabela 5.3a são apresentados os valores de seis pontos experimentais das curvas Pesq vs. TA dos ruídos *car*, *tunnel-front* e *crowd-children*, usando o WI007 sob condições limpas, e, a partir desses pontos, é definida qual a melhor curva de ajuste para cada um desses ruídos. Na referida tabela e nas demais, o ponto Pesq = 4.5 foi excluído pelo fato de não pertencer ao conjunto de pontos resultantes dos sinais de fala degradados, portanto é um valor obtido para o sinal limpo e permanecendo constante para todos os testes feitos.

Tabela 5.3a
Pontos experimentais das curvas Pesq vs. TA dos ruídos car, tunnel-front e crowd-children usando o WI007 sob condições limpas

Car		**Tunnel-Front**		**Crowd-Children**	
TA(%)	**Pesq**	**TA(%)**	**Pesq**	**TA(%)**	**Pesq**
9.39	1.386	6.54	1.340	10.94	1.392
14.46	1.618	10.76	1.653	20.54	1.708
34.09	1.878	34.04	1.931	42.88	1.983
67.01	2.181	65.15	2.246	70.26	2.244
90.04	2.493	86.21	2.559	88.72	2.537
97.41	2.793	94.86	2.856	94.13	2.834

Assim, como explicitado anteriormente, a partir dos dados da Tabela 5.4a, aplicando Mail, que usa o MMQ para medir os desvios quadráticos das distâncias entre os pontos, obtém-se a melhor curva para o ruído *car*, conforme mostra a Figura 5.5a.

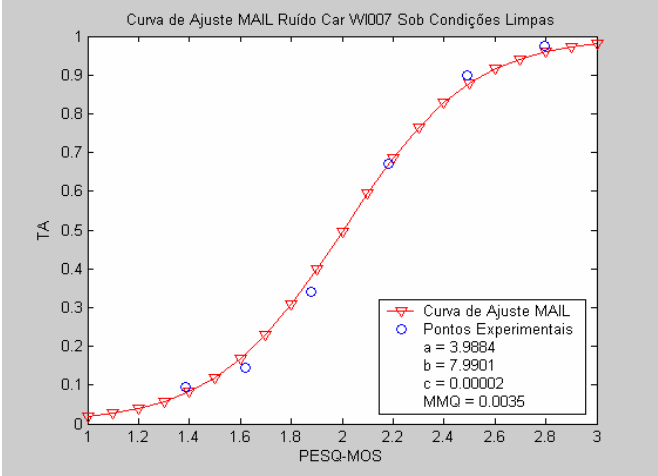


Figura 5.5a: Curva de ajuste pelo Mail para o ruído car usando WI007 sob condições limpas

Por curiosidade, é mostrado na Figura 5.5b a mesma curva de ajuste, porém produzida através da aplicação do *Método da Regressão Linear* (MRL).

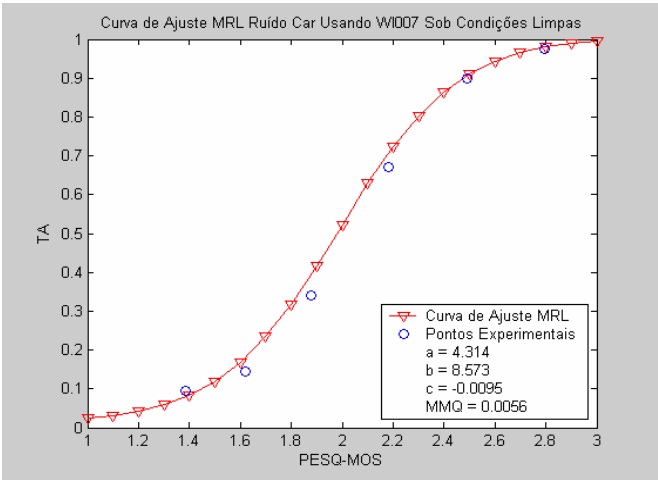


Figura 5.5b: Curva de ajuste pela regressão linear para o ruído car usando WI007 sob condições limpas

Os parâmetros ***a*** e ***b*** da Figura 5.5b foram obtidos diretamente do MRL. Já o ***c*** foi obtido aplicando-se ***a*** e ***b*** na equação (5.1). No ajuste pelo MRL, o valor do cálculo do MMQ também é apresentado.

Observando os valores da soma dos desvios quadráticos das curvas das Figuras 5.5a e 5.5b, pode-se notar que há uma melhor aproximação da curva de ajuste obtida pela aplicação do Mail do que pela aplicação do MRL. Esta diferença entre os métodos se dá porque na regressão linear não foi considerado o valor do parâmetro ***c***, além de outros aspectos comentados na subseção 5.1.3.

Assim, como temos definida pelo Mail a melhor curva de ajuste do ruído *car*, pode-se determinar com melhor precisão se este, com os demais ruídos (*tunnel-front* e *crowd-children*), formam uma família.

Desta forma, como já referido, a partir dos dados da Tabela 5.4a, aplicando Mail, conforme feito com o ruído *car*, pode-se obter uma curva de ajuste para o ruído *tunnel-front* e *crowd-children*.

As Figuras 5.5c e 5.5d mostram as curvas de ajuste do ruído *tunnel-front* e *crowd-children*, respectivamente.

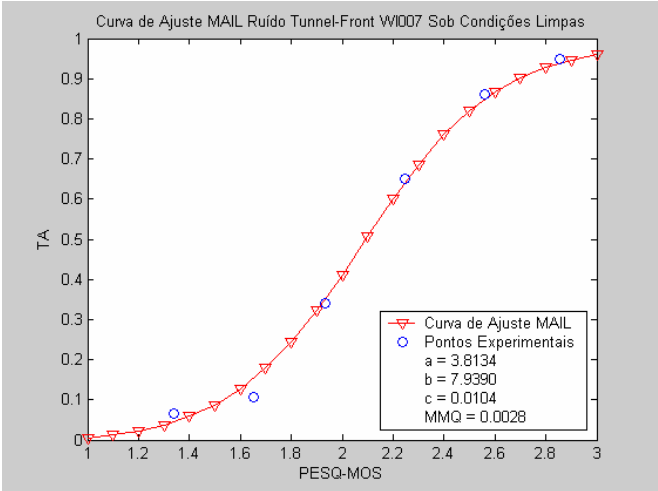


Figura 5.5c: Curva de ajuste pelo Mail para o ruído tunnel-front usando WI007 sob condições limpas

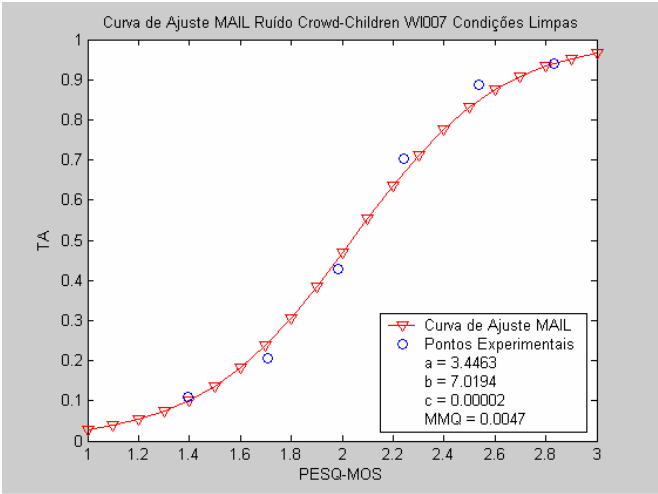


Figura 5.5d: Curva de ajuste pelo Mail para o ruído crowd-children usando WI007 sob condições limpas

A partir das curvas de ajuste dos ruídos *tunnel-front* e *crowd-children* pode-se calcular, pelo MMQ, a semelhança que cada uma dessas curvas tem com a curva do ruído *car*.

A Figura 5.5e mostra o traçado das curvas de ajuste dos ruídos *car*, *tunnel-front* e *crowd-children* em um só gráfico, onde, MMQ-TF e MMQ-CC são os resultados dos modelos dos desvios quadráticos dos ruídos *tunnel-front* e *crowd-children*, respectivamente. Cada curva do ruído de referência é distinguida pelo símbolo triangular.

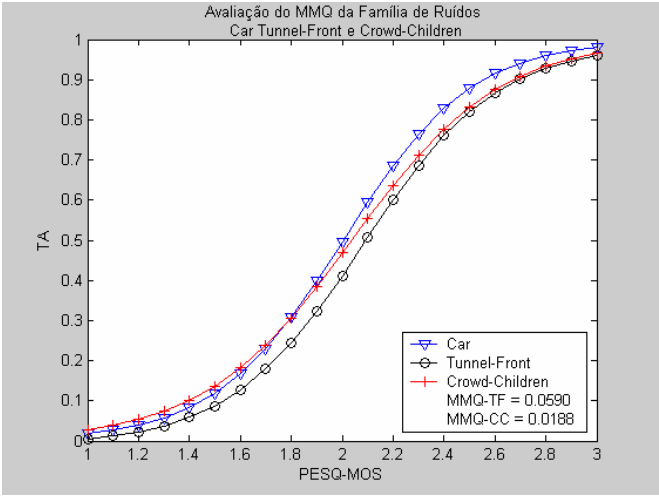


Figura 5.5e: Avaliação do MMQ da família de ruídos car, tunnel-front e crowd-children usando WI007 sob condições limpas

Analisando a Figura 5.5e, e considerando a definição de família para o valor $dsv \leq 0{,}05$, conclui-se que somente o ruído *crowd-children* e o ruído *car* formam uma família de ruídos com comportamento similar.

Obviamente, para um valor menos restrito dos desvios quadráticos pode-se incluir o ruído *tunnel-front*. Por outro lado, se são considerados os valores Pesq > 2, os valores dos desvios quadráticos passam a ser: MMQ-TF = 0.0425 e MMQ-CC = 0.0170.

Estes resultados indicam que a formação de uma família de ruídos depende da faixa de valores Pesq que se pretende analisar. É evidente que esses valores Pesq dependem do nível de degradação (SNR) de cada sinal.

Finalizando a análise sobre o comportamento dos ruídos novos da Figura 5.5e, com o uso do WI007 sob condições limpas, são apresentados na Tabela 5.3b os valores dos desvios quadráticos entre as curvas de ajuste dos ruídos *tunnel-front e crowd-children* para o ruído *car*.

Tabela 5.3b
Desvios quadráticos das curvas de ajuste dos ruídos tunnel-front e crowd-children para o ruído car usando WI007 sob condições limpas

Ruídos Novos	**Tunnel-Front**	**Crowd-Children**
Ruídos Base Aurora-1	**MMQ**	**MMQ**
Car	0.0590	0.0188

Desta forma, observando a Tabela 5.3b e seguindo o critério $dsv \leq 0{,}05$, pode-se confirmar que para o WI007 sob condições limpas tem-se a definição de apenas uma família: *car* e *crowd-children*.

Faz-se necessário lembrar que, neste trabalho, foram analisadas as curvas para os 6 (seis) valores Pesq experimentais e para 21 (vinte e um) pontos dos valores das curvas de ajuste no intervalo $1{,}0 \leq Pesq \leq 3{,}0$.

5.2.1.2 Usando o testa - testd e WI008

A Figura 5.6a mostra o comparativo das curvas Pesq vs. TA, dos ruídos do *testa* da base Aurora-1 com os ruídos do *testd*, depois do processo de reconhecimento sob treinamento em condições limpas, usando o WI008.

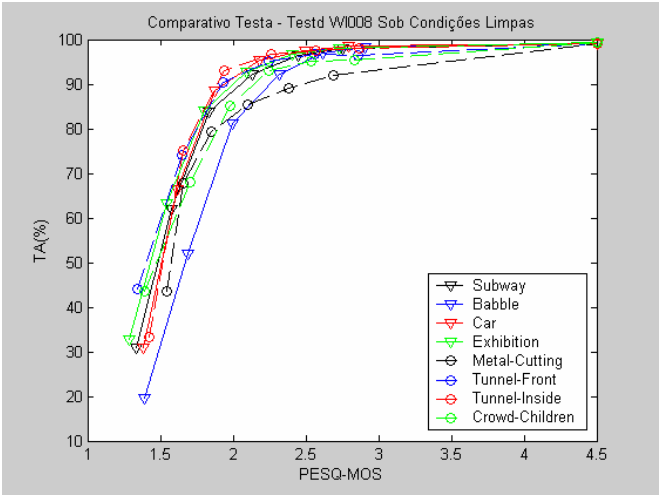


Figura 5.6a: Comparativo testa-testd usando WI008 sob condições limpas

Tomando como base a avaliação qualitativa feita para a família de ruídos definida anteriormente, pode-se concluir que, com a atuação do WI008, os ruídos *subway*, *car*, *exhibition*, *tunnel-front e tunnel-inside* podem formar uma família de ruídos.

A Figura 5.6b mostra as curvas experimentais dos ruídos *subway*, *car*, *exhibition*, *tunnel-front e tunnel-inside*, usando o *front-end* WI008 sob condições limpas e que, segundo o critério qualitativo adotado aqui, formam uma família.

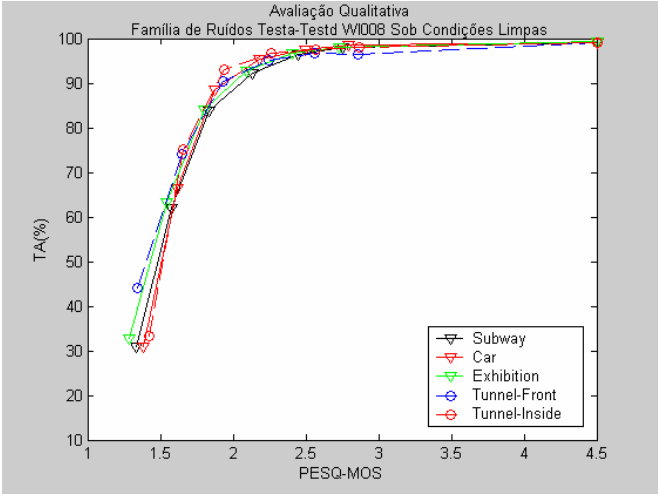


Figura 5.6b: Família de ruídos da avaliação qualitativa do testa-testd usando WI008 sob condições limpas

A partir dos pontos experimentais das curvas dos ruídos da Figura 5.6b (com uso do WI008) foi aplicado a mesma metodologia usada anteriormente para definir se os ruídos *tunnel-front* e *tunnel-inside* formam uma família com os ruídos da base Aurora-1, ou seja, com os ruídos *subway*, *car* e *exhibition*.

Assim, na Tabela 5.4a são apresentados os valores dos seis pontos experimentais (com degradação) das curvas Pesq vs. TA dos ruídos *subway*, *car*, *exhibition*, *tunnel-front e tunnel-inside*, usando o WI008 sob condições limpas, e, a partir desses pontos, é definido qual a melhor curva de ajuste para cada um desses ruídos.

Tabela 5.4a
Pontos experimentais das curvas Pesq vs. TA dos ruídos subway, car, exhibition, tunnel-front e tunnel-inside usando o WI008 sob condições limpas

Tunnel-Front		Tunnel-Inside	
TA(%)	**Pesq**	**TA(%)**	**Pesq**
96.48	1.340	97.95	1.421
96.82	1.653	97.38	1.655
95.26	1.931	96.60	1.942
90.31	2.246	92.92	2.260
74.04	2.559	75.14	2.572
44.02	2.856	33.26	2.870

Subway		Car		Exhibition	
TA(%)	**Pesq**	**TA(%)**	**Pesq**	**TA(%)**	**Pesq**
97.91	1.333	98.48	1.386	97.90	1.289
96.41	1.573	97.58	1.618	96.82	1.544
92.23	1.833	95.29	1.878	92.78	1.804
83.82	2.131	88.49	2.181	84.05	2.095
61.93	2.444	66.42	2.493	63.28	2.404
30.86	2.756	30.84	2.793	32.86	2.726

A seguir, nas Figuras 5.7a a 5.7e são apresentadas as curvas de ajuste para a hipótese da formação de família de ruídos *subway*, *car*, *exhibition*, *tunnel-front e tunnel-inside*, *respectivamente*, em relação aos pontos experimentais, exibindo também os parâmetros e o resultado da aplicação do MMQ.

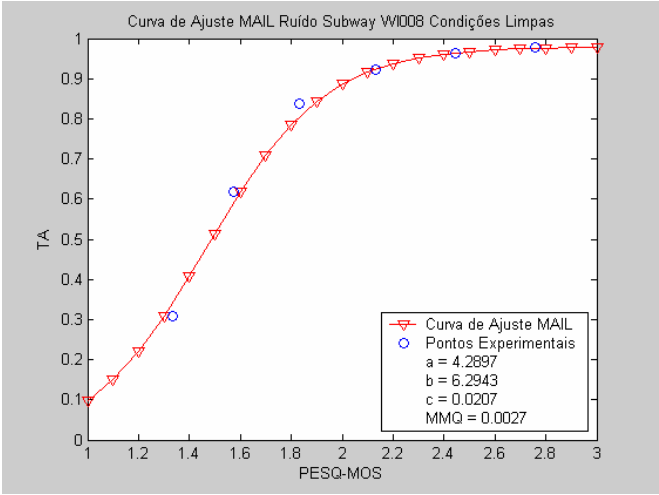


Figura 5.7a: Curva de ajuste pelo Mail para o ruído subway usando WI008 sob condições limpas

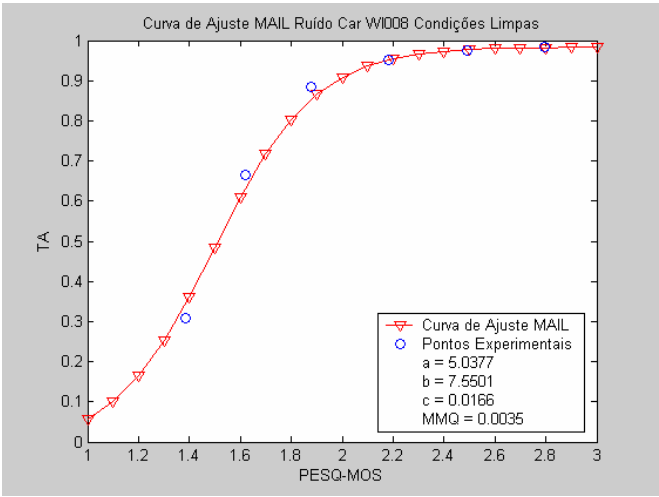


Figura 5.7b: Curva de ajuste pelo Mail para o ruído car usando WI008 sob condições limpas

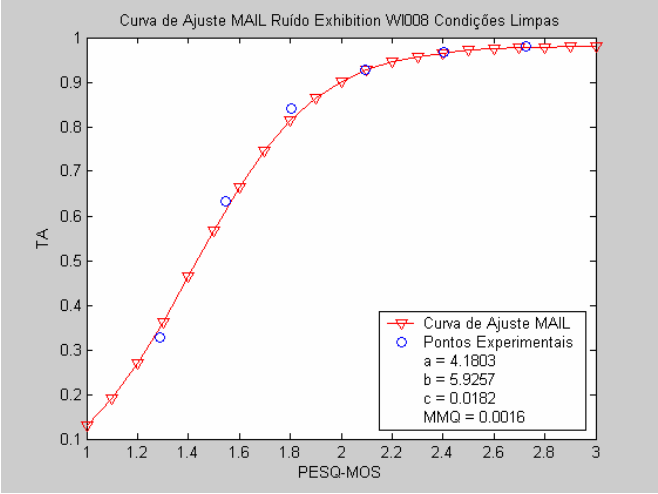


Figura 5.7c: Curva de ajuste pelo Mail para o ruído exhibition usando WI008 sob condições limpas

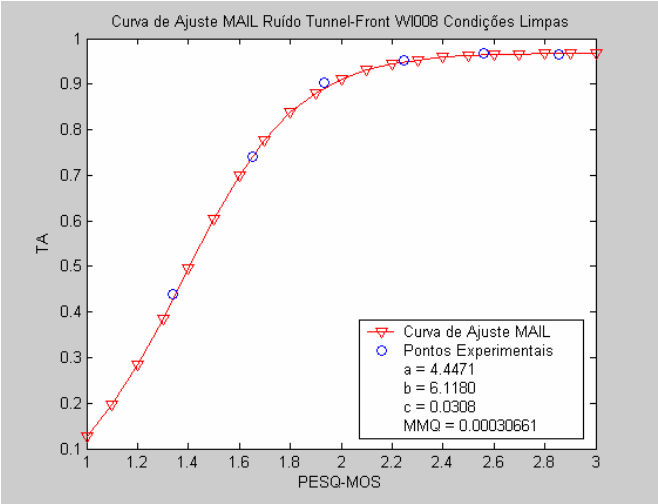


Figura 5.7d: Curva de ajuste pelo Mail para o ruído tunnel-front usando WI008 sob condições limpas

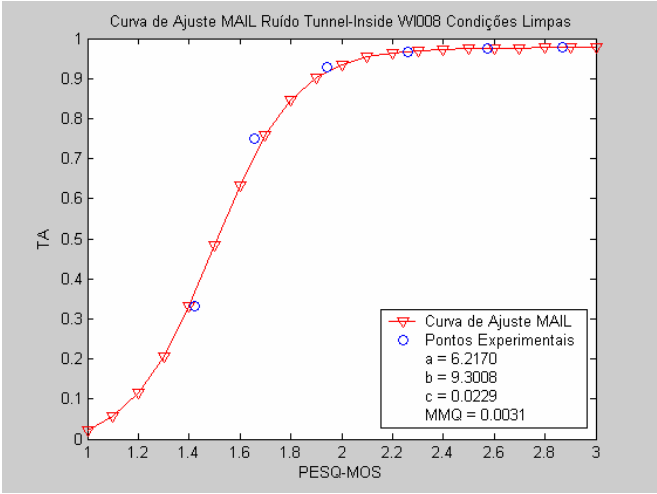


Figura 5.7e: Curva de ajuste pelo Mail para o ruído tunnel-inside usando WI008 sob condições limpas

Conforme modelo adotado anteriormente, a partir das curvas de ajuste dos ruídos pode-se calcular, pelo MMQ, a semelhança que cada uma das curvas dos ruídos *tunnel-front* e *tunnel-inside* tem com as curvas dos ruídos *subway*, *car* e *exhibition.*

Desta forma, como há três curvas de referência, são apresentados três gráficos para verificar a possibilidade de os ruídos novos serem familiares ao comportamento de algum dos três ruídos da base Aurora-1 supramencionados.

Assim, a Figura 5.7f, mostra o traçado das curvas de ajuste da família de ruídos *subway*, *tunnel-front* e *tunnel-inside*. Onde, MMQ-TF e MMQ-TI são os resultados do modelo dos desvios quadráticos dos ruídos *tunnel-front* e *tunnel-inside*, respectivamente.

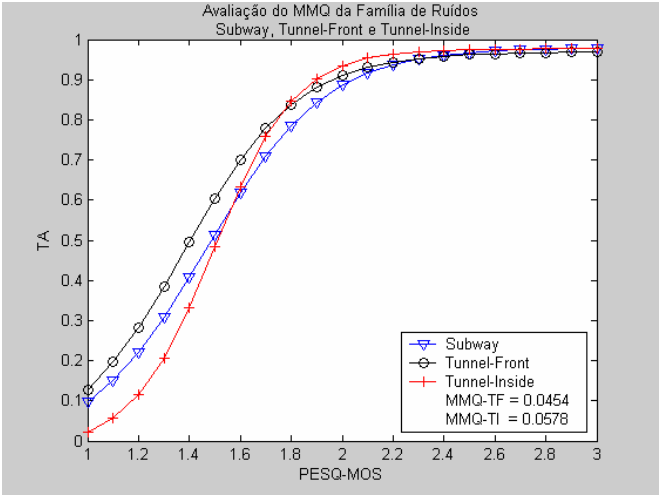


Figura 5.7f: Avaliação do MMQ da família de ruídos subway, tunnel-front e tunnel-inside usando WI008 sob condições limpas

Analisando a Figura 5.7f, e considerando a definição de família para o valor $dsv \leq 0{,}05$, conclui-se que somente o ruído *tunnel-front* e o ruído *subway* formam uma família de ruídos com comportamento similar. Porém, a mesma observação feita anteriormente sobre o rigor do valor $dsv \leq 0{,}05$, vale aqui.

A Figura 5.7g mostra o traçado das curvas de ajuste dos ruídos *car*, *tunnel-front* e *tunnel-inside*.

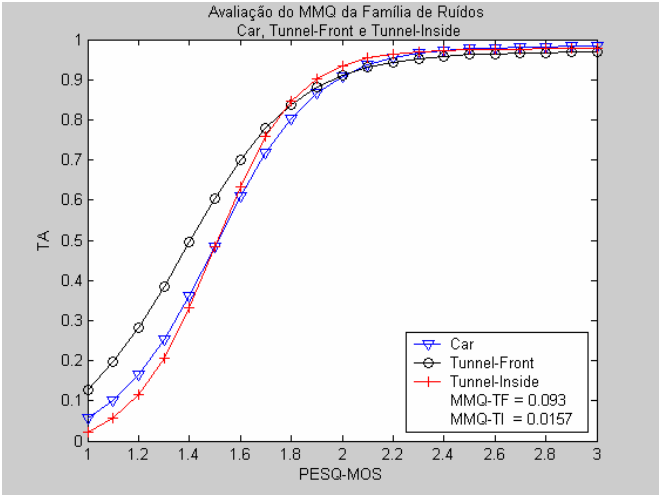


Figura 5.7g: Avaliação do MMQ da família de ruídos car, tunnel-front e tunnel-inside usando WI008 sob condições limpas

Na Figura 5.7g observa-se que o ruído *tunnel-inside* é o único que satisfaz o critério *dsv* ≤ 0,05, assim, podendo-se concluir que o comportamento dos ruídos *car* e *tunnel-inside* são familiares.

Por outro lado, o ruído *exhibition*, segundo o critério adotado, tem comportamento similar ao ruído *tunnel-front*. Isto pode ser constatado na Figura 5.7h, que mostra um MMQ-TF = 0.0072 e um MMQ-TI = 0.1091 em relação à curva do ruído *exhibition*.

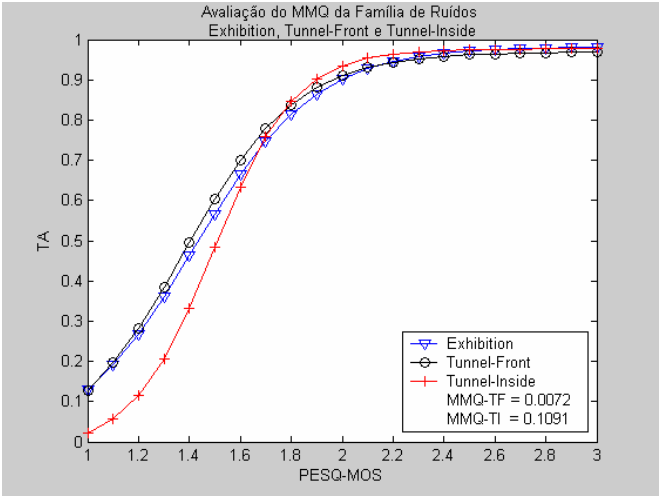


Figura 5.7h: Avaliação do MMQ da família de ruídos exhibition, tunnel-front e tunnel-inside usando WI008 sob condições limpas

Este aspecto, aliado à formação da família *subway* e *tunnel-front* definida anteriormente leva a acreditar que esses dois ruídos (*exhibition* e *tunnel-front)* formam uma família com o ruído *subway*. Isto pode ser provado aplicando-se o MMQ para a curva de ajuste dos ruídos *subway*, *exhibition* (curva referência) e *tunnel-front*, conforme mostra a Figura 5.7i.

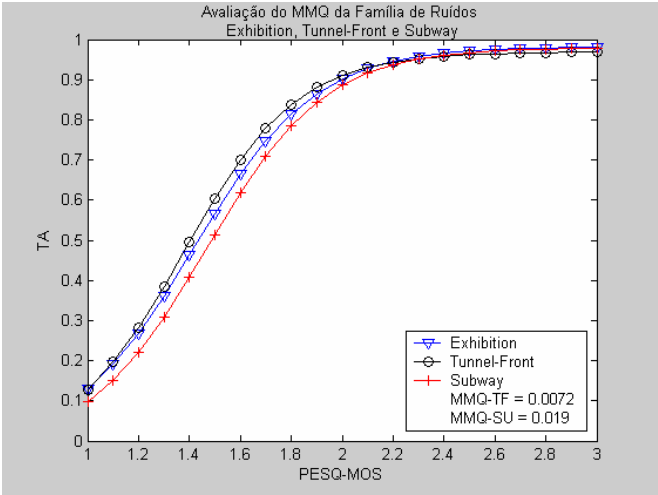


Figura 5.7i: Avaliação do MMQ da família de ruídos exhibition, tunnel-front e subway usando WI008 sob condições limpas

Finalizando a análise sobre o comportamento dos ruídos com o uso do WI008 sob condições limpas, na Tabela 5.4b são apresentados os valores dos desvios quadráticos dos ruídos *tunnel-front e tunnel-inside* para os ruídos *subway*, *car* e *exhibition*.

Tabela 5.4b
Desvios quadráticos *das curvas de ajuste dos ruídos tunnel-front e tunnel-inside* para os ruídos *subway*, *car* e *exhibition* usando WI008 sob condições limpas

Ruídos Novos	**Tunnel-Front**	**Tunnel-Inside**
Ruídos Base Aurora-1	**MMQ**	**MMQ**
Subway	0.0454	0.0578
Car	0.0930	0.0157
Exhibition	0.0072	0.1091

Desta forma, numa perspectiva geral, observando a Tabela 5.4b e seguindo o critério *dsv* ≤ 0,05, pode-se confirmar que para o WI008 sob condições limpas temos a definição de duas famílias: 1) *subway*, *exhibition* e *tunnel-front*; 2) *car* e *tunnel-inside*.

5.2.1.3 Usando o testb - testd e WI007

Com base em análise anterior, os ruídos do *testb* pertencentes à base de dados Aurora-1 são: *restaurant*, *street*, *airport* e *train-station* (*train*).

Assim, neste tópico serão repetidos os procedimentos do tópico anterior, para que se possa identificar se, entre os ruídos novos (*testd*) comparados aos ruídos do *testb*, há formação de famílias de ruídos com comportamento similar. Para tanto, é seguido o mesmo critério, ou seja, *dsv* ≤ 0,05, e, portanto, as curvas dos ruídos do *testb* e do *testd* usando WI007 sob condições limpas também são apresentadas.

Neste contexto, é analisado o comportamento das curvas dos ruídos do *testb* e *testd* de forma qualitativa e, a partir desta análise, identificado em um único gráfico as curvas que têm comportamento similar para todos os valores Pesq. A Figura 5.8a apresenta o comportamento comparativo das curvas Pesq vs. TA, dos ruídos do *testb* da base Aurora-1 com os ruídos do *testd*, depois do processamento sob treinamento em condições limpas, usando o WI007.

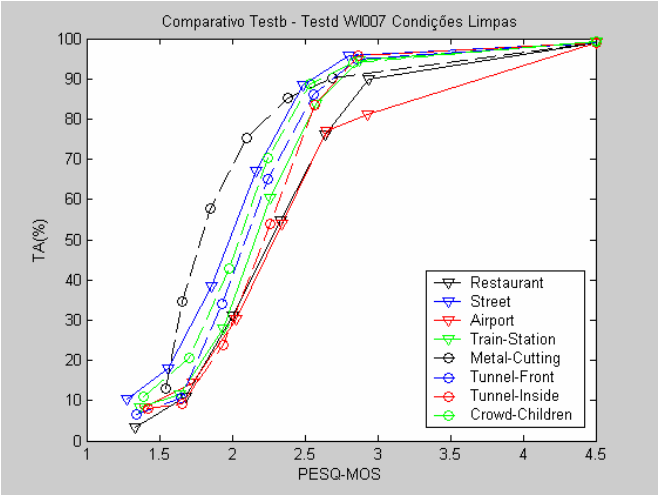


Figura 5.8a: Comparativo testb-testd usando WI007 sob condições limpas

Na Figura 5.8a, do comparativo entre o *testd* e o *testb*, também pode-se observar alguns grupos de ruídos com comportamento similar.

É importante lembrar que a escolha inicial do grupo de ruídos com comportamento similar é subjetiva. Assim, para toda a faixa Pesq da Figura 5.8a de forma qualitativa, pode-se afirmar que o ruído *tunnel-front* praticamente tangencia a forma dos ruídos *train-station*, *street* e *crowd-children*, podendo todos eles formarem um grupamento familiar de ruídos.

A curva do ruído *tunnel-inside*, entretanto, parece estar próxima o bastante para ser incluída nos testes e identificar se pode fazer parte desse mesmo grupo. Esta decisão de incluir o ruído *tunnel-inside*

neste momento será avaliada mais adiante. A Figura 5.8b mostra as curvas dos ruídos *train-station*, *street*, *tunnel-front*, *tunnel-inside* e *crowd-children* que aqui foram avaliadas qualitativamente como terem possibilidade de formarem uma família.

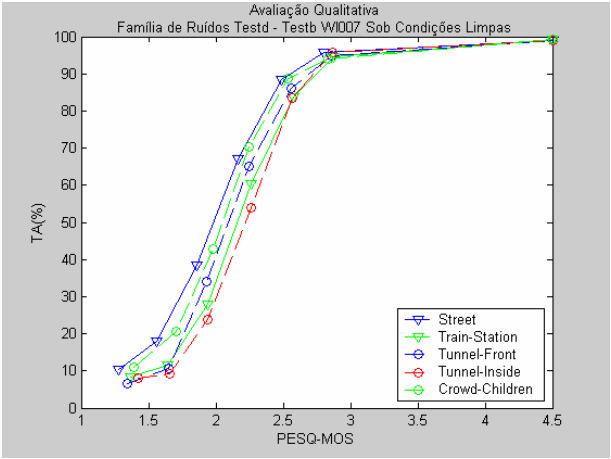


Figura 5.8b: Família de ruídos da avaliação qualitativa do testd-testb usando WI007

Da mesma forma que antes, utilizando Mail e o MMQ pode-se, a partir dos pontos experimentais, obter os valores aproximados dos parâmetros ***a***, ***b*** e ***c***, que definem a curva de ajuste de cada um dos ruídos da Figura 5.8b.

Na Tabela 5.5a são apresentados os valores dos seis pontos experimentais das curvas Pesq vs. TA dos ruídos *train-station*, *street*, *tunnel-front*, *tunnel-inside* e *crowd-children*, usando o WI007 sob condições limpas, e, a partir desses pontos, define-se qual a melhor curva de ajuste para cada um desses ruídos.

Tabela 5.5a
Pontos experimentais das curvas Pesq vs. TA dos ruídos train-station, street, tunnel-front, tunnel-inside e crowd-children usando WI007 sob condições limpas

Street		**Train-Station**	
TA(%)	**Pesq**	**TA(%)**	**Pesq**
10.46	1.281	8.45	1.368
17.84	1.561	11.57	1.638
38.45	1.859	27.92	1.937
67.11	2.169	60.29	2.263
88.45	2.488	83.65	2.577
95.74	2.800	94.72	2.873

Tunnel-Front		**Tunnel-Inside**		**Crowd-Children**	
TA(%)	**Pesq**	**TA(%)**	**Pesq**	**TA(%)**	**Pesq**
6.54	1.340	8.04	1.421	10.94	1.392
10.76	1.653	9.22	1.655	20.54	1.708
34.04	1.931	23.74	1.942	42.88	1.983
65.15	2.246	53.93	2.260	70.26	2.244
86.21	2.559	83.61	2.572	88.72	2.537
94.86	2.856	95.90	2.870	94.13	2.834

Assim, como já abordado, a partir dos dados da Tabela 5.5a, aplicando o Mail, e usando o MMQ para medir os desvios quadráticos das distâncias entre os pontos, obtém-se a melhor curva para o ruído *street*, conforme mostra a Figura 5.8c.

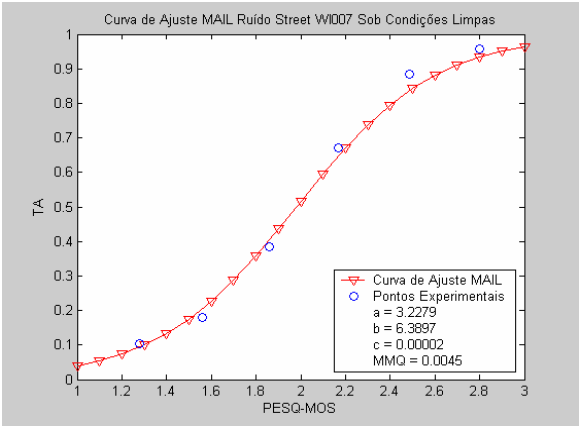


Figura 5.8c: Curva de ajuste pelo Mail para o ruído street usando WI007 sob condições limpas

Desta forma, como se tem definida pelo Mail a curva de ajuste do ruído *street*, pode-se determinar com melhor precisão se este, com os demais ruídos (*tunnel-front*, *tunnel-inside* e *crowd-children*), formam uma família.

Antes, porém, é preciso definir a curva de ajuste do ruído *train-station*. A Figura 5.8d, a seguir, mostra a curva de ajuste para este ruído, com os respectivos valores dos parâmetros e do MMQ calculado em relação aos pontos experimentais.

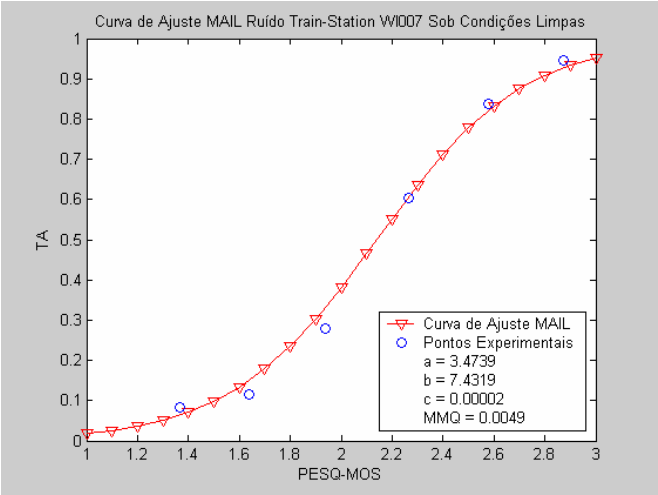


Figura 5.8d: Curva de ajuste pelo Mail para o ruído train-station usando WI007 sob condições limpas

As curvas de ajuste para os ruídos *tunnel-front*, *tunnel-inside* e *crowd-children*, para o WI007 sob condições limpas já foram definidas antes, à exceção do ruído *tunnel-inside*, que tem a curva de ajuste apresentada na Figura 5.8e, a seguir.

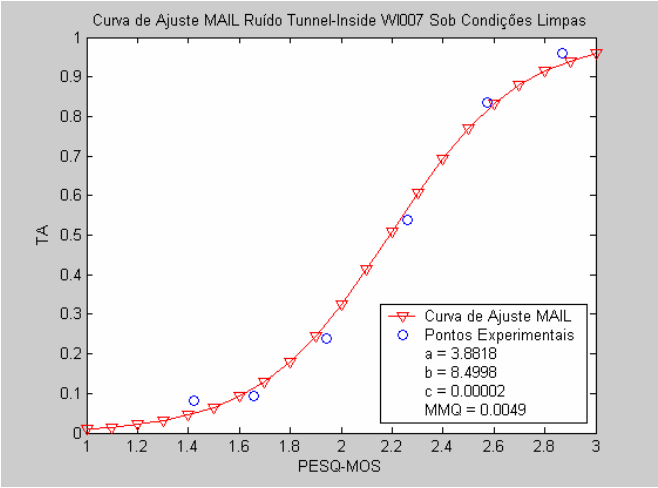


Figura 5.8e: Curva de ajuste pelo Mail para o ruído tunnel-inside usando WI007 sob condições limpas

Conforme procedimento anterior, pode-se calcular, pelo MMQ, a semelhança que cada uma dessas curvas tem com a curva dos ruídos de referência: *street* e *train-station*.

A Figura 5.8f mostra o traçado das curvas de ajuste dos ruídos *street*, *tunnel-front*, *tunnel-inside* e *crowd-children* em um só gráfico, onde MMQ-TF, MMQ-TI e MMQ-CC são os resultados do modelo dos desvios quadráticos dos ruídos *tunnel-front*, *tunnel-inside* e *crowd-children*, respectivamente.

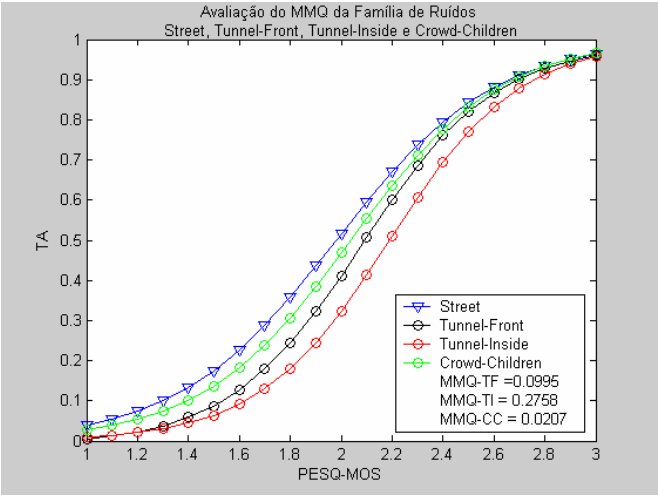


Figura 5.8f: Avaliação do MMQ da família de ruídos street, tunnel-front, tunnel-inside e crowd-children usando WI007 sob condições limpas

Analisando a Figura 5.8f, e considerando a definição de família para o valor $dsv \leq 0{,}05$, conclui-se que somente o ruído *crowd-children* e o ruído *street* formam uma família com comportamento similar.

Igualmente, observando os resultados dos valores do MMQ mostrados na Figura 5.8f, pode-se perceber que a curva do ruído *tunnel-inside* é a que mais se distancia do comportamento da curva do ruído *street*.

Para o ruído *tunnel-front*, constata-se uma maior aproximação com o ruído *street* do que com o ruído *tunnel-inside*, porém não se enquadra dentro do critério adotado ($dsv \leq 0{,}05$).

Por outro lado, com o ruído *train-station*, a característica da família definida anteriormente muda sensivelmente.

Conforme pode-se detectar na Figura 5.8g, *crowd-children* deixa de fazer parte da família constituída por *train-station*, dando lugar aos ruídos *tunnel-front* e *tunnel-inside*, que não formaram família com o ruído *street*.

Pelo critério adotado fica fácil perceber que os ruídos *street* e *train-station* não têm a curva de ajuste próxima uma da outra, assim, não formam uma família entre sí.

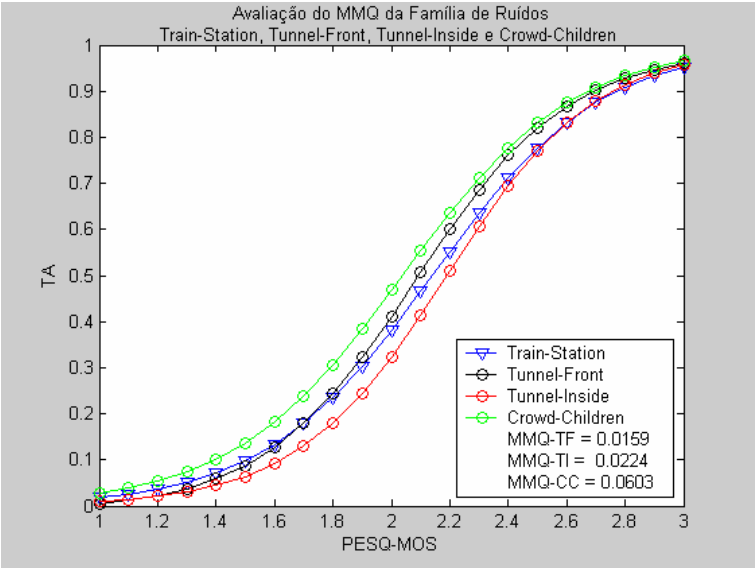


Figura 5.8g: Avaliação do MMQ da família de ruídos train-station, tunnel-front, tunnel-inside e crowd-children usando WI007 sob condições limpas

Finalizando a análise do comportamento dos ruídos do *testb* com o *testd* com uso do WI007 sob condições limpas, são apresentados

na Tabela 5.5b os valores dos desvios quadráticos dos ruídos *tunnel-front, tunnel-inside e crowd-children* para o ruído *street e train-Station.*

Tabela 5.5b
Desvios quadráticos das curvas de ajuste dos ruídos *tunnel-front, tunnel-inside* e *crowd-children* para os ruídos *street e train-station* usando WI007 sob condições limpas

Ruídos Novos	**Tunnel-Front**	**Tunnel-Inside**	**Crowd-Children**
Ruídos Base Aurora-1	**MMQ**	**MMQ**	**MMQ**
Street	0.0995	0.2758	0.0207
Train-Station	0.0159	0.0224	0.0603

Desta forma, observando a Tabela 5.5b e seguindo o critério $dsv \leq 0{,}05$, pode-se confirmar que para o WI007 sob condições limpas, tendo em conta os ruídos do *testb* e *testd*, se tem a definição de mais duas famílias de ruídos: 1) *street* e *crowd-children*; 2) *train-station, tunnel-front* e *tunnel-inside*.

5.2.1.4 Usando testb - testd e WI008

A Figura 5.9a mostra o comportamento das curvas Pesq vs. TA, dos ruídos do *testb* da base Aurora-1 com os ruídos do *testd*, depois do processo de reconhecimento, usando o WI008 sob treinamento em condições limpas.

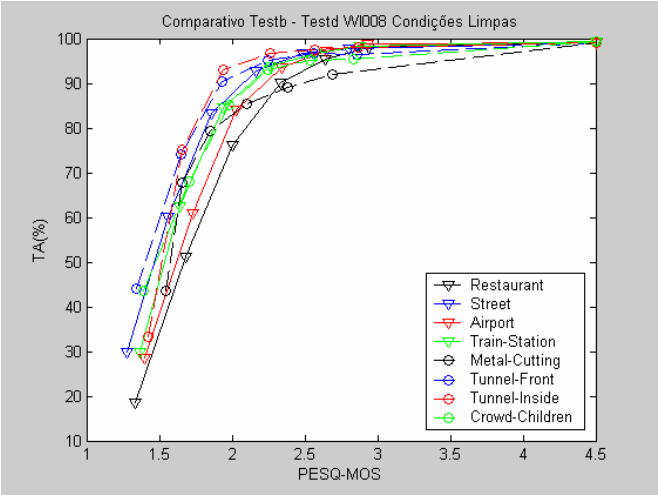


Figura 5.9a: Comparativo testb-testd usando WI008 sob condições limpas

Tomando como base a avaliação qualitativa feita para as famílias de ruídos anteriormente definidas, pode-se argumentar que, com a atuação do WI008, os ruídos *street*, *train-station*, *tunnel-front*, *tunnel-inside* e *crowd-children* podem formar uma família de ruídos.

A Figura 5.9b, apresenta as curvas experimentais dos ruídos *street*, *train-station*, *tunnel-front*, *tunnel-inside* e *crowd-children*, usando o *front-end* WI008 sob condições limpas e que, segundo os critérios qualitativos adotados aqui, formam uma família.

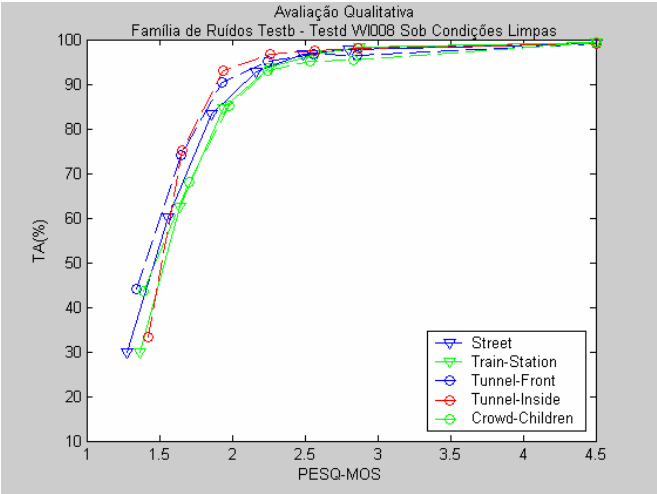


Figura 5.9b: Famílias de ruídos do comparativo testb – testd usando WI008 sob condições limpas

Na Tabela 5.6a são apresentados os valores de seis pontos experimentais das curvas Pesq vs. TA dos ruídos *street*, *train-station*, *tunnel-front*, *tunnel-inside* e *crowd-children*, usando o WI008 sob condições limpas, e, a partir desses pontos, se define qual a melhor curva de ajuste para cada um desses ruídos.

Tabela 5.6a
Pontos experimentais das curvas Pesq vs. TA dos ruídos *street*, *train-station*, *tunnel-front*, *tunnel-inside* e *crowd-children* usando WI008 sob condições limpas

Street		**Train-Station**	
TA(%)	**Pesq**	**TA(%)**	**Pesq**
29.87	1.281	29.96	1.368
60.07	1.561	62.57	1.638
83.28	1.859	84.57	1.937
92.78	2.169	93.77	2.263
96.74	2.488	96.73	2.577
97.64	2.800	98.36	2.873

Tunnel-Front		**Tunnel-Inside**		**Crowd-Children**	
TA(%)	**Pesq**	**TA(%)**	**Pesq**	**TA(%)**	**Pesq**
96.48	1.340	97.95	1.421	43.60	1.392
96.82	1.653	97.38	1.655	68.09	1.708
95.26	1.931	96.60	1.942	85.05	1.983
90.31	2.246	92.92	2.260	93.07	2.244
74.04	2.559	75.14	2.572	95.12	2.537
44.02	2.856	33.26	2.870	95.31	2.834

Uma vez que anteriormente, na subseção 5.2.1.2, foram definidas as curvas de ajuste dos ruídos *tunnel-front* e *tunnel-inside* para o WI008 sob condições limpas, torna-se desnecessário repeti-las.

Em prosseguimento, nas Figuras 5.10a, 5.10b e 5.10c, as curvas de ajuste dos ruídos *street*, *train-station*, e *crowd-children* são apresentadas sob o mesmo modelo anterior, ou seja, apresentando o valor dos parâmetros e o resultado da aplicação da soma dos desvios quadráticos ou MMQ em relação aos pontos experimentais.

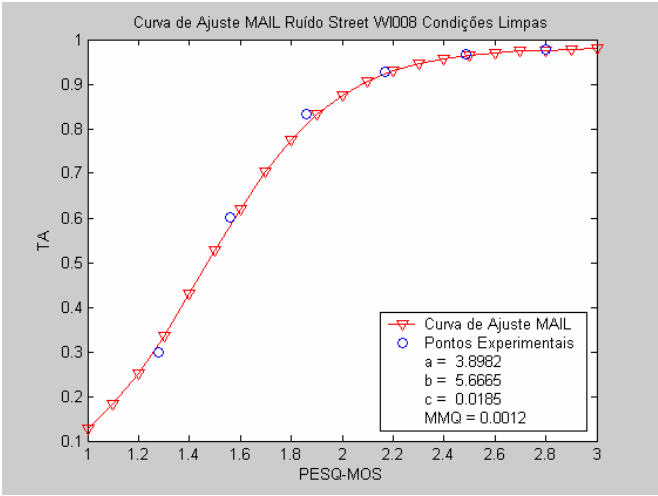


Figura 5.10a: Curva de ajuste pelo Mail para o ruído street usando WI008 sob condições limpas

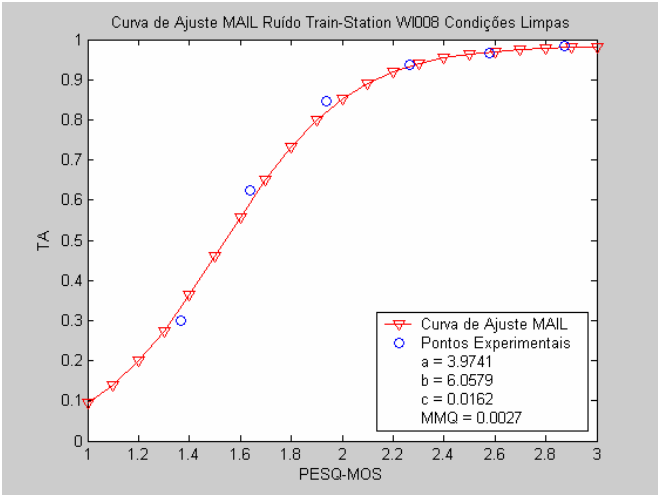


Figura 5.10b: Curva de ajuste pelo Mail para o ruído train-station usando WI008 sob condições limpas

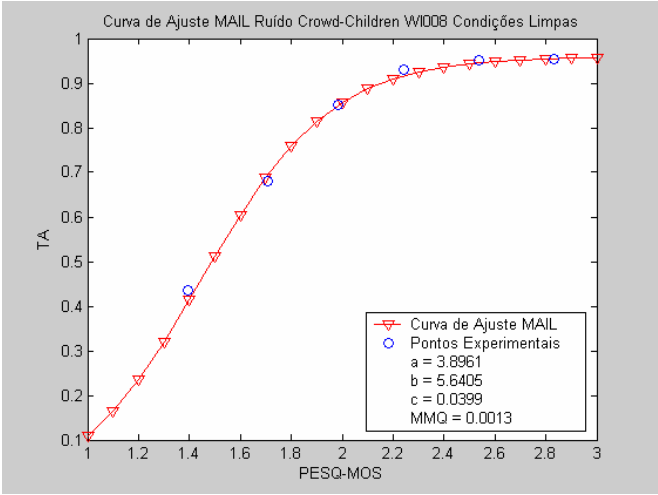


Figura 5.10c: Curva de Juste pelo Mail para o ruído crowd-children usando WI008 sob condições limpas

A partir das curvas de ajuste dos ruídos, pode-se calcular, pelo MMQ, a distância que cada uma das curvas dos ruídos *tunnel-front*, *tunnel-inside* e *crowd-children* tem das curvas dos ruídos referência: *street* e *train-station.*

Desta forma, como existem duas curvas de referência, são apresentados dois gráficos para verificar a hipótese de os ruídos novos serem familiares ao comportamento de algum dos dois ruídos da base Aurora-1, ou seja: *street* e/ou *train-station*.

A Figura 5.10d mostra o traçado da curva de ajuste do ruído *street* com os demais ruídos (*tunnel-front*, *tunnel-inside* e *crowd-children*).

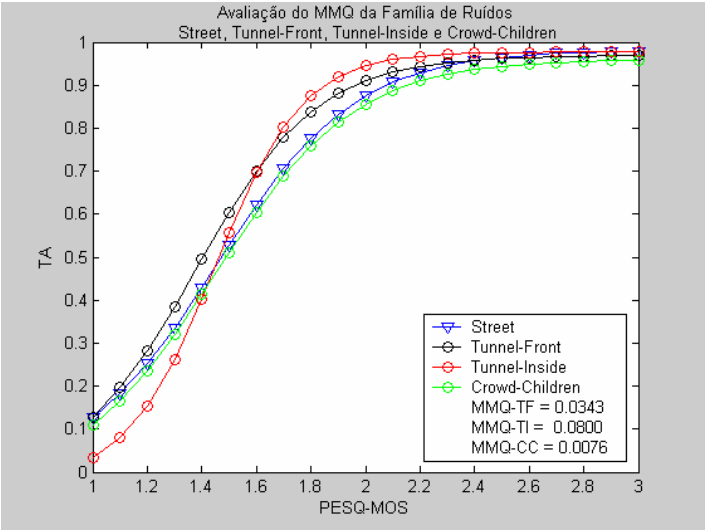


Figura 5.10d: Avaliação do MMQ da família de ruídos street, tunnel-front, tunnel-inside e crowd-children usando WI008 sob condições limpas

Analisando a Figura 5.10d, e considerando o sistema WI008 sob condições limpas, com o critério $dsv \leq 0{,}05$, conclui-se que os ruídos *tunnel-front* e *crowd-children* formam uma família de ruídos com comportamento similar ao ruído *street*.

Conforme pode ser observado na Figura 5.10e, contudo, para o ruído *train-station*, ainda com o uso do WI008 sob condições limpas, e adotanto o mesmo critério, conclui-se que não há a formação de família de ruídos. A Figura 5.10e apresenta o traçado da curva de ajuste do ruído *train-station* com os mesmos ruídos (*tunnel-front*, *tunnel-inside* e *crowd-children*).

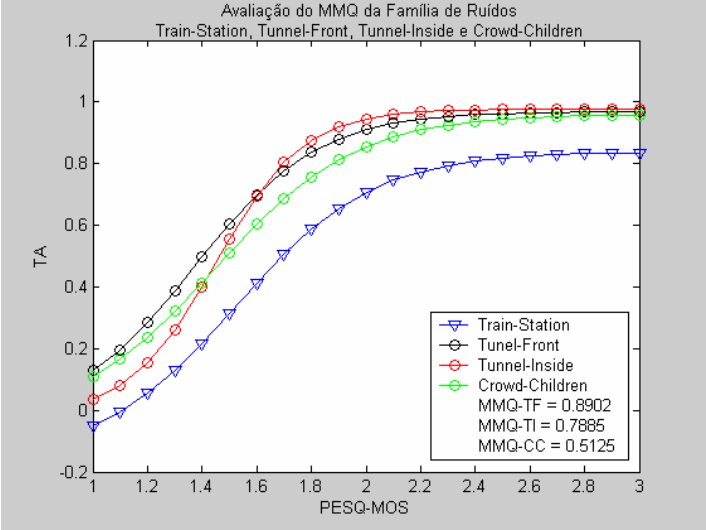


Figura 5.10e: Avaliação do MMQ da família de ruídos train-station, tunnel-front, tunnel-inside e crowd-children usando WI008 sob condições limpas

Finalizando a análise sobre o comportamento dos ruídos com o uso do WI008 sob condições limpas para o comparativo dos ruídos do *testb* e *testd*, são apresentados na Tabela 5.6b os valores dos desvios quadráticos dos ruídos *tunnel-front*, *tunnel-inside* e *crowd-children* para os ruídos *street* e *train-station*.

Tabela 5.6b
Desvios quadráticos das curvas de ajuste dos ruídos tunnel-front, tunnel-inside e crowd-children para os ruídos Street e Train-Station usando WI008 sob condições limpas

Ruídos Novos	**Tunnel-Front**	**Tunnel-Inside**	**Crowd-Children**
Ruídos Base Aurora-1	**MMQ**	**MMQ**	**MMQ**
Street	0.0343	0.0800	0.0076
Train-Station	0.8902	0.7885	0.5125

Assim, observando a Tabela 5.6b e seguindo o critério $dsv \leq 0{,}05$, pode-se confirmar que para o WI008 sob condições limpas para o comparativo entre os ruídos do *testb* e *testd*, tem-se a definição de mais uma família: *street*, *tunnel-front* e *crowd-children*.

5.2.2 Usando o testd e treinamento sob condições múltiplas

O objetivo desta subseção é a análise do comportamento dos ruídos novos sob teste em múltiplas condições para verificar, com o mesmo procedimento feito para condições limpas, se há formação de grupamentos de ruídos com características similares. Assim sendo, são apresentadas as curvas Pesq vs. TA dos ruídos novos em conjunto com os testes (*testa* e *testb*) da base Aurora-1.

Assim, a partir dessas curvas e das definições de família de ruídos feitas em tópicos anteriores, os ruídos novos são avaliados quanto à base de dados robusta empregada e aos resultados obtidos.

5.2.2.1 Usando o testa – testd e WI007

A Figura 5.11a mostra o comparativo das curvas Pesq vs. TA, dos ruídos do *testa* da base Aurora-1 com os ruídos do *testd*, após o processo de reconhecimento sob treinamento em condições múltiplas, usando o WI007.

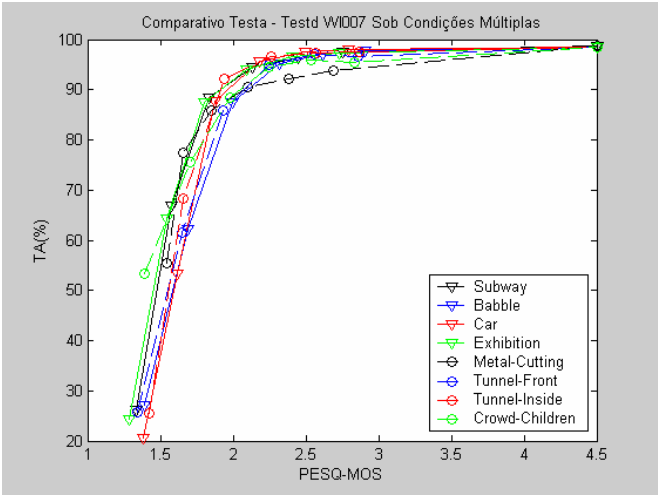


Figura 5.11a: Comparativo testa-testd usando WI007 sob condições múltiplas

Pode-se concluir pela observação da Figura 5.11a que, quando o teste é feito sob condições múltiplas, há uma maior aproximação das curvas dos ruídos, e, desta forma, quase todos formam um grupamento com comportamento similar. Na avaliação qualitativa feita aqui o ruído *metal-cutting*, contudo, ainda não pertence a nenhum grupamento, pois para a faixa $1,5 < Pesq < 3,0$ mostra valores de taxa de acerto distantes em relação aos demais ruídos.

Para avaliar a hipótese lançada de que o ruído *metal-cutting* não faz parte de nenhum grupo de ruídos, a partir desse ponto este aspecto é avaliado. A Figura 5.11b apresenta as curvas Pesq vs. TA dos ruídos do *testa* da base Aurora-1 com a curva do ruído *metal-cutting*.

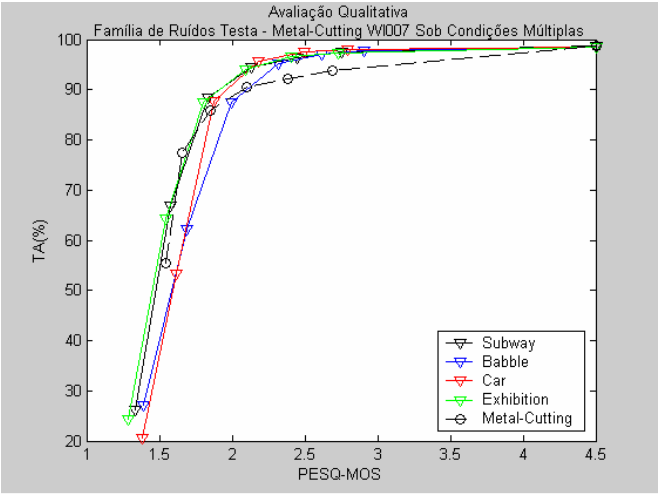


Figura 5.11b: Família de ruídos da avaliação qualitativa do testa-metal-cutting usando WI007 sob condições múltiplas

Conforme pode-se observar na Figura 5.11b, o ruído *metal-cutting* apresenta um comportamento similar às curvas dos ruídos *subway* e *exhibition* somente para uma faixa Pesq-MOS muito pequena: 1,5 < Pesq < 2,0.

É oportuno lembrar que o ruído *metal-cutting* foi o único que não apareceu nas famílias de ruídos definidas anteriormente. Para comprovação, entretanto, de que o ruído *metal-cutting*, considerando os critérios adotados, precisa ter suas características presentes na base de dados robusta, todos os passos da metodologia serão aplicados.

Nesse viés, a metodologia de avaliação da base de dados ruidosa proposta neste trabalho será aplicada para calcular o valor dos desvios quadráticos somente para o ruído *metal-cutting* em relação aos demais ruídos da base de dados Aurora-1.

Na Tabela 5.7a são apresentados os valores dos pontos experimentais das curvas Pesq vs. TA dos ruídos do *testa*: *subway*, *babble*, *car*, *exhibition* e *mettal-cutting*.

Tabela 5.7a
Pontos experimentais das curvas Pesq vs. TA dos ruídos subway, babble, car, exhibition e mettal-cutting usando WI007 sob condições múltiplas

Subway		**Babble**	
TA(%)	**Pesq**	**TA(%)**	**Pesq**
26.13	1.333	27.18	1.391
66.9	1.573	62.15	1.691
88.36	1.833	87.55	1.997
94.44	2.131	95.28	2.315
96.47	2.444	97.04	2.617
97.61	2.756	97.73	2.910

Car		**Exhibition**		**Metal-Cutting**	
TA(%)	**Pesq**	**TA(%)**	**Pesq**	**TA(%)**	**Pesq**
20.58	1.386	24.34	1.289	93.69	1.548
53.44	1.618	64.36	1.544	92.18	1.660
87.8	1.878	87.60	1.804	90.54	1.849
95.74	2.181	94.11	2.095	85.90	2.099
97.61	2.493	96.67	2.404	77.40	2.385
98.03	2.793	97.41	2.726	55.39	2.690

A partir dos pontos experimentais da Tabela 5.7a são produzidas as curvas de ajuste pelo Mail. Desta forma, nas Figuras 5.12a, 5.12b, 5.12c, 5.12d e 5.12e são apresentadas as curvas de ajuste dos ruídos *metal-cutting*, *subway*, *babble*, *car* e *exhibition*, para o WI007 sob condições múltiplas, sob o mesmo modelo anterior, ou seja, apresentando o valor dos parâmetros e o resultado da aplicação da soma dos desvios quadráticos ou MMQ em relação aos pontos experimentais.

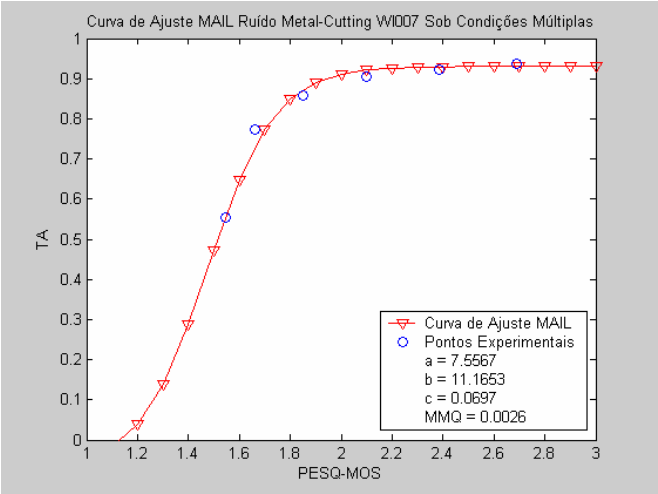


Figura 5.12a: Curva de ajuste pelo Mail para o ruído metal-cutting usando WI007 sob condições múltiplas

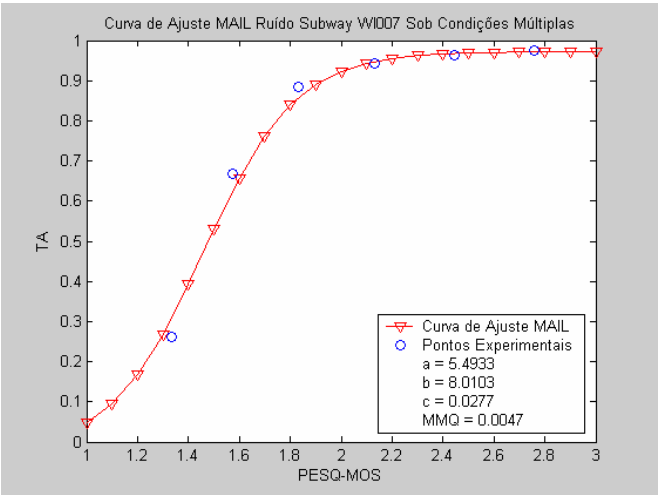


Figura 5.12b: Curva de ajuste pelo Mail para o ruído subway usando WI007 sob condições múltiplas

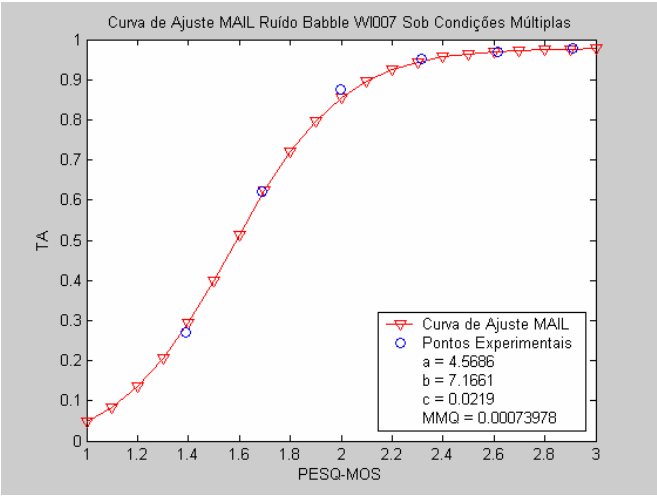


Figura 5.12c: Curva de ajuste pelo Mail para o ruído babble usando WI007 sob condições múltiplas

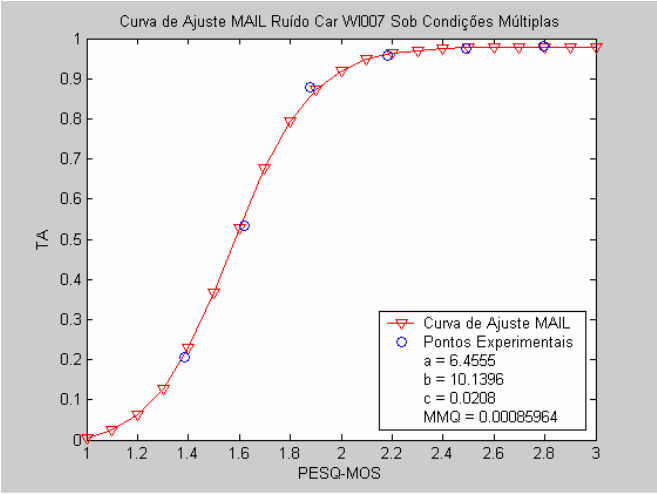


Figura 5.12d: Curva de ajuste pelo Mail para o ruído car usando WI007 sob condições múltiplas

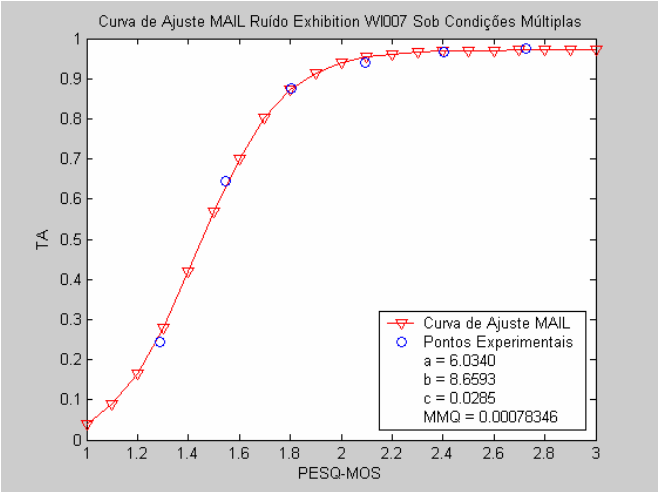


Figura 5.12e: Curva de ajuste pelo Mail para o ruído exhibition usando WI007 sob condições múltiplas

A partir das curvas de ajuste dos ruídos pode-se calcular, pelo MMQ, a distância da curva do ruído *metal-cutting* para cada uma das curvas dos ruídos referência: *subway*, *babble*, *car* e *exhibition*. Assim, como existem três curvas de referência, são apresentadas as três curvas em figuras distintas para confirmar a hipótese do ruído *metal-cutting* não formar família com outro tipo de ruído, mesmo sob condições múltiplas, ou seja, aqui está sendo aplicando o método para negar a formação de grupos de ruídos com o ruído *metal-cutting*.

A Figura 5.12f mostra o traçado da curva de ajuste do ruído *subway* com o ruído *metal-cutting*.

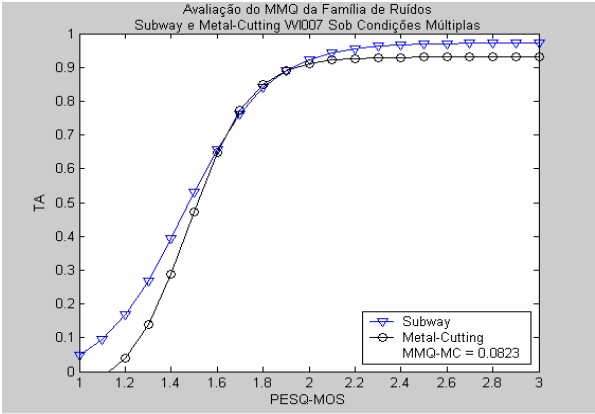


Figura 5.12f: Avaliação do MMQ da família de ruídos subway e metal-cutting usando WI007 sob condições múltiplas

Como pode-se perceber da Figura 5.12f, a curva do ruído *metal-cutting* com a do ruído *subway* são similares apenas para uma faixa Pesq pequena, no cotovelo superior e o DESVIO > 5%.

A Figura 5.12g apresenta o traçado da curva de ajuste do ruído *babble* com o ruído *metal-cutting*.

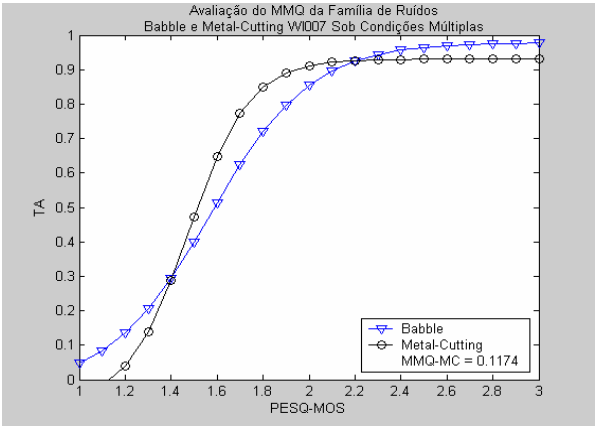


Figura 5.12g: Avaliação do MMQ da família de ruídos babble e metal-cutting usando WI007 sob condições múltiplas

Como também pode-se constatar da Figura 5.12g, a curva do ruído *metal-cutting* com a do ruído *babble* não são similares e o MMQ = 0.1175, ou seja, dsv > 5%.

As Figuras 5.12h e 5.12i mostram o traçado da curva de ajuste do ruído *car* e do ruído *exhibition* com o ruído *metal-cutting*, respectivamente.

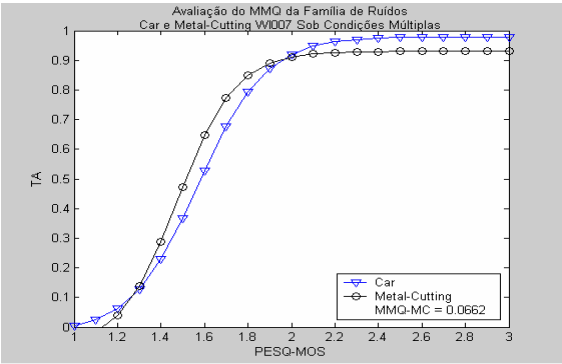


Figura 5.12h: Avaliação do MMQ da família de ruídos car e metal-cutting usando WI007 sob condições múltiplas

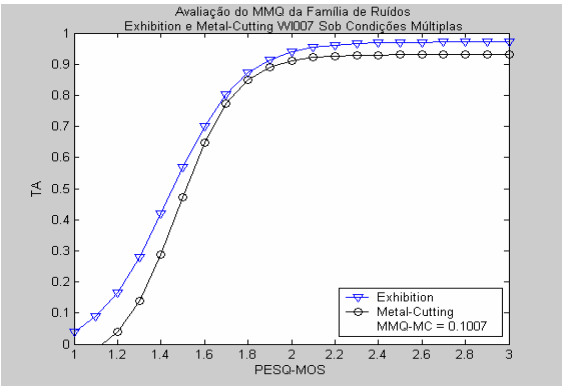


Figura 5.12i: Avaliação do MMQ da família de ruídos exhibition e metal-Cutting usando WI007 sob condições múltiplas

Além das figuras anteriores (Figuras 5.12f e Figura 5.12g), pode-se perceber analisando as Figura 5.12h e 5.12i, que a curva do ruído *metal-cutting* com a do ruído *car* e com o ruído *exhibition*, conforme cálculo do MMQ, também não são similares a ponto de formar uma família de ruídos com comportamento similar.

Isso posto, ao calcular o MMQ da curva do ruído *metal-cutting* para as curvas dos ruídos *subway*, *babble*, *car* e *exhibition*, observa-se um maior valor dos desvios quadráticos, acima dos 5% (*dsv* > 0,05) definido anteriormente. Isto confirma a hipótese da avaliação qualitativa, de que o ruído *metal-cutting*, para o sistema WI007, não tem comportamento similar a nenhum outro ruído da base de dados Aurora-1, mesmo com o *testd* sendo executado sob treinamento em condições múltiplas.

A Tabela 5.7b mostra os valores do MMQ do ruído *metal-cutting* em relação aos ruídos do *testa* com o uso do WI007 sob condições múltiplas.

Tabela 5.7b
Desvios quadráticos das curvas de ajuste do ruído metal-cutting para os ruídos subway, babble, car e exhibition usando WI007 sob condições múltiplas

Ruídos Novos	**Metal-Cutting**
Ruídos Base Aurora-1	**MMQ**
Subway	0.0823
Babble	0.1174
Car	0.0662
Exhibition	0.1007

Esse comportamento do ruído *metal-cutting* pode ser observado em todos os resultados, inclusive os tabelados, o que também leva a crer que esse tipo de ruído carece de mais estudos dentro do

contexto da base de dados ruidosa utilizada. É necessário verificar, entretanto, qual é o comportamento desse ruído com o uso do WI008 sob as mesmas condições, o que é feito no item a seguir.

5.2.2.2 Usando o testa – testd e WI008

A Figura 5.13a mostra o comportamento das curvas Pesq vs. TA, dos ruídos do *testa* da base Aurora-1 com os ruídos do *testd*, após o processo de reconhecimento sob treinamento em condições múltiplas usando o WI008.

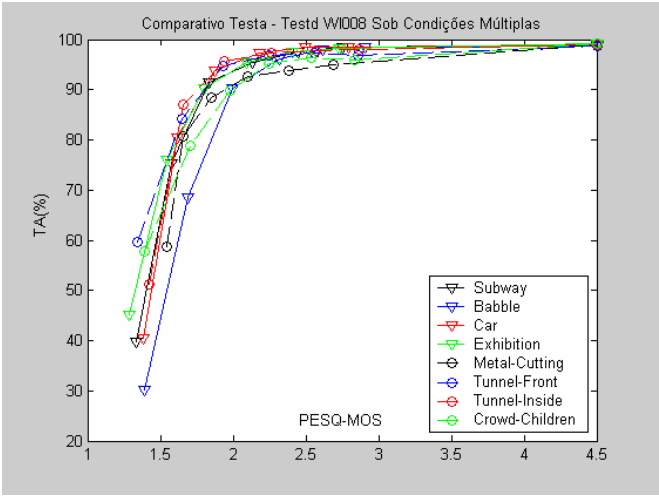


Figura 5.13a: Comparativo testa-testd usando WI008 sob condições múltiplas

Assim, para melhorar os estudos sobre o ruído *metal-cutting* e descobrir se este ruído se comporta dentro do critério adotado, pode-se avaliar, ainda, o seu comportamento para um sistema com o uso do WI008 sob condições múltiplas. Desta forma, após submeter o sinal de

ruído às características do WI008, que tem filtros de redução de ruído, é avaliado o comportamento sob a base de dados ruidosa.

Na Figura 5.13a observa-se que o comportamento dos ruídos adotando o WI008 sob condições múltiplas é semelhante quando é usado o WI007 nas mesmas condições, porém no WI008 o traçado das curvas em geral, especialmente para a faixa mais degradada (Pesq < 2), muda sensivelmente, com a elevação da taxa de acerto. Este fato é explicado pela presença no WI008 dos filtros de redução de ruído.

Para o ruído *metal-cutting* a ação do WI008 também aumenta a taxa de acerto para os níveis de degradação maiores dos sinais de fala, porém no geral o comportamento desse ruído é quase o mesmo ao do uso do WI007 sob as mesmas condições.

Para facilitar a análise e se verificar com mais facilidade as explicações anteriores, na Figura 5.13b é apresentada uma comparação entre as curvas de ajuste do ruído *metal-cutting* do WI007 com a do WI008, ambos sob condições múltiplas.

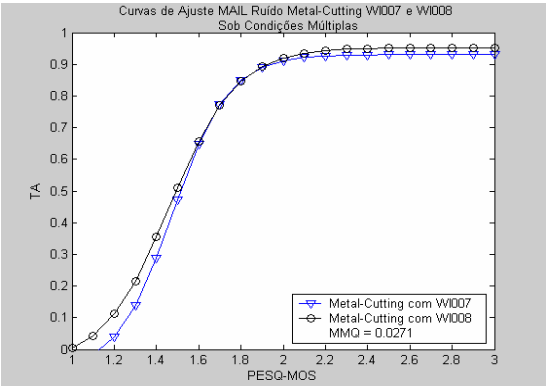


Figura 5.13b: Comparativo entre as curvas do ruído metal-cutting usando WI007 e WI008 sob condições múltiplas

Como é possível observar da Figura 5.13b, a curva do ruído *metal-cutting* praticamente não mudou o seu comportamento do sistema que usa o WI007 para o sistema que usa o WI008. Aplicando o MMQ, pode-se notar que a soma dos desvios quadráticos é igual a 2,71%, abaixo dos 5% do critério estabelecido para verificar a semelhança no comportamento das curvas, conforme a metodologia proposta.

Assim pode-se concluir que, da Figura 5.13a, quase todos os ruídos novos formam um grupamento com comportamento similar, à exceção do ruído *metal-cutting* na faixa 1,5 < Pesq < 3,0, onde os valores de taxa de acerto são sensivelmente diferentes em relação aos demais ruídos.

Considerando, então, a análise qualitativa e o reconhecimento sob múltiplas condições usando o WI008, para toda a faixa Pesq da Figura 5.13a, conclui-se que os ruídos *subway*, *car*, *exhibition*, *tunnel-front*, *tunnel-inside* e *crowd-children* formam um grupo de ruídos com o comportamento muito parecido para as curvas Pesq vs. TA.

Por outro lado, com base no critério adotado ($dsv \leq 0{,}05$), o ruído *metal-cutting* não forma família com nenhum dos demais ruídos.

Desta forma, como o comportamento do ruído *metal-cutting* e dos demais ruídos não mudou a ponto de ser feita uma avaliação qualitativa diferente da anterior, não será aplicada a metodologia de avaliação da base ruidosa para o sistema com o WI008 sob condições múltiplas. Ademais, acredita-se que todos os procedimentos feitos até aqui explicam suficientemente como é realizada a aplicação da metodologia proposta.

No Apêndice F encontram-se, as curvas comparativas dos ruídos do *testa* e do ruído *metal-cutting* para o WI007 e para o WI008 sob condições múltiplas em uma mesma figura para facilitar comparações.

Além disso, ainda no apêndice F pode-se analisar as curvas de ajuste dos ruídos do *testa* e do ruído *metal-cutting* para o WI008 sob condições múltiplas, com os respectivos resultados dos cálculos do MMQ e constatar que não é necessária a aplicação da metodologia neste caso.

Mesmo com o teste sendo feito sob condições múltiplas, portanto, o ruído *metal-cutting* mostra que suas características na base de dados robusta, considerando o critério de $dsv \leq 0{,}05$, podem não estar presentes o suficiente para melhorar a taxa de acerto em relação aos demais ruídos.

Desta forma, observando todos os resultados encontrados até aqui, confirma-se que o fato de se usar uma base de dados robusta bem projetada leva à obtenção de melhores resultados nas taxas de acerto e em um desempenho mais robusto do SRAF quando o reconhecimento de fala se faz necessário em meio ruidoso. A base de dados, no entanto, precisa estar bem projetada. Assim, conforme já referido, na subseção 5.2.3 é apresentado um estudo do ruído *metal-cutting* presente na base de dados Aurora-1.

Antes, porém, nos itens 5.2.2.3 e 5.2.2.4 apresenta-se resumidamente a avaliação para os ruídos novos em relação aos ruídos

do *testb*, tanto para o WI007 como para o WI008, em condições múltiplas.

5.2.2.3 Usando o testb – testd e WI007

Em relação ao comparativo com os ruídos do *testb* e *testd*, são apresentadas algumas considerações somente com base no critério qualitativo. Cabe lembrar aqui que, para a avaliação dos ruídos novos, em relação à base de dados ruidosa (base Aurora-1), o mais importante é a comparação com os ruídos que constituem esta base de dados, que foi realizado na subseção anterior.

Na Figura 5.14, portanto, mostra-se o comparativo das curvas Pesq vs. TA, dos ruídos do *testb* da base Aurora-1 com os ruídos do *testd*, depois do processo de reconhecimento sob treinamento em condições múltiplas, usando o WI007.

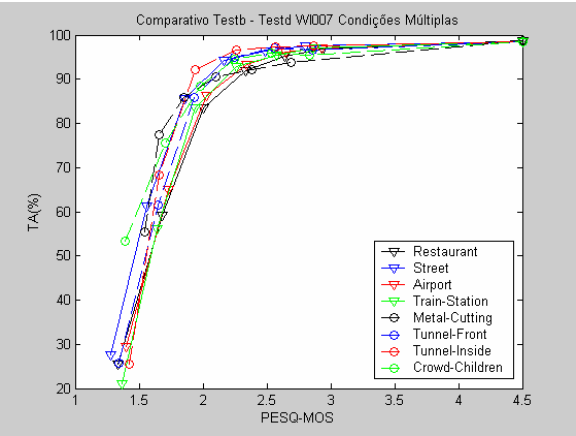


Figura 5.14: Comparativo testb-testd usando WI007 sob condições múltiplas

Pode-se desprender da Figura 5.14 que, igualmente, quase todos os ruídos formam um grupamento com um comportamento similar, à exceção do ruído *metal-cutting*, que para a faixa 2,0 < Pesq < 3,5, mostra valores de taxa de acerto sensivelmente abaixo dos valores dos demais ruídos.

Assim, para toda a faixa Pesq-MOS da Figura 5.14, observa-se que os ruídos *restaurant*, *street*, *airport*, *train-station*, *tunnel-front*, *tunnel-inside* e *crowd-children* formam um grupo de ruídos com o comportamento parecido para as curvas Pesq vs. TA.

Desta forma, também para os ruídos do *testb* e uso do WI007 sob condições múltiplas, revela-se que o fato de se usar uma base de dados robusta equilibrada implica a obtenção de melhores resultados nas taxas de acerto e um desempenho mais robusto do SRAF quando o reconhecimento de fala se faz necessário em meio ruidoso.

E, da mesma maneira que no comparativo dos ruídos do *testa* sob as mesmas condições de processamento, os resultados anteriores também mostram um comportamento diferente para o ruído *metal-cutting*. Como mencionado antes, este aspecto vem sendo observado em todos os resultados.

5.2.2.4 Usando o testb – testd e WI008

A Figura 5.15 mostra o comparativo das curvas Pesq vs. TA, dos ruídos do *testb* da base Aurora-1 com os ruídos do *testd*, após o processo de reconhecimento sob treinamento em condições múltiplas, usando o WI008.

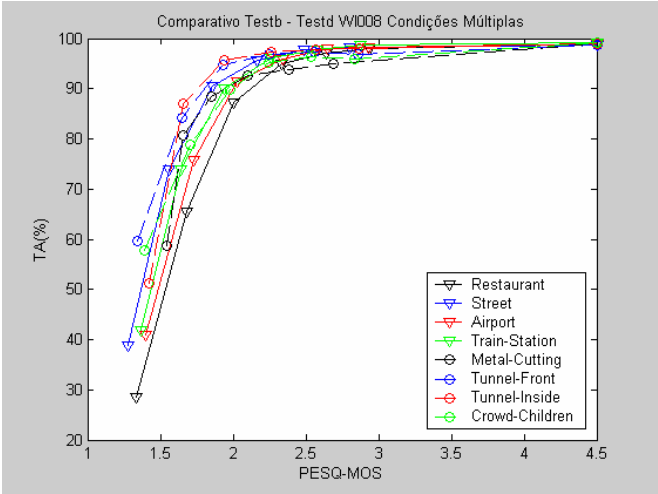


Figura 5.15: Comparativo testb-testd usando WI008 sob condições múltiplas

Observa-se na Figura 5.15 que o comportamento dos ruídos empregando o WI008 sob condições múltiplas também é praticamente o mesmo de quando é usado o WI007 nas mesmas condições de processamento, porém os valores da taxa de acerto do ruído *metal-cutting* para a faixa 2,0 < Pesq < 3,5 ficam um pouco abaixo das demais curvas.

Assim, considerando os valores das taxas de acerto para a faixa Pesq > 2,0, pode-se concluir, da Figura 5.15, que quase todos os ruídos formam um grupamento com comportamento similar, à exceção do ruído *metal-cutting*.

Como pode-se perceber, até aqui foram estudados os ruídos novos estabelecendo um comparativo com os ruídos de todos os testes, porém, não foi realizada a comparação com os ruídos do *testc* da base de dados Aurora-1.

Os ruídos do *testc*, porém, como descrito anteriormente, são os mesmos encontrados no *testa* (*subway*) e no *testb* (*street*) da referida base, passados pelo filtro Mirs. Desta forma, como um comparativo dos ruídos novos com os ruídos do *testc* não acrescentará algo novo, o comparativo com o *testc* não foi realizado.

Na realidade, para verificar se os ruídos novos têm suas características representadas pela base de dados robusta (neste caso a base Aurora-1), o comparativo com os ruídos que constituem o *testb* e o *testc* não é importante, os ruídos que formam a base de dados estes sim são importantes e, neste caso, são os mesmos ruídos do *testa*, ou seja, *subway*, *babble*, *car* e *exhibition*.

5.2.3 Análise do ruído metal-cutting na base de dados

Nesta subseção a metodologia de avaliação de base de dados robusta finalmente é aplicada em sua parte conclusiva: a modificação da base de dados robusta inserindo as características de um determinado tipo de ruído que não satisfaça os critérios definidos, no caso, a soma dos desvios quadráticos ser inferior a 5%.

Uma série de comparativos foi feita anteriormente para chegar à conclusão que o ruído *metal-cutting* carecia de maiores estudos. Como foi possível perceber, este tipo de ruído não teve comportamento similar dentro do critério estabelecido com nenhum outro ruído presente na base de dados Aurora-1.

Assim, como *metal-cutting* foi o único ruído que não formou grupamento de família com comportamento similar dentro do critério

adotado, especialmente com os ruídos da base de dados, neste item são apresentados os resultados do comportamento dos sistemas (WI007 e WI008) usando as características deste ruído na base de dados Aurora-1.

Os procedimentos adotados para a mixagem do referido ruído à base de dados Aurora-1 foram os mesmos anteriormente explicitados.

Por facilidade, o procedimento de mixagem foi feito substituindo o ruído *babble*, da base de dados, pelo ruído *metal-cutting*, que, com as mesmas locuções de *babble*, teve suas características presentes no treinamento.

Os resultados para o ruído *metal-cutting* inserido no material de treinamento, devido à aplicação do sistema WI007 sob condições múltiplas, encontra-se na Tabela 5.8a.

Tabela 5.8a
Taxa de acerto para o F-E WI007 e treinamento sob condições múltiplas usando o ruído *metal-cutting* na base de dados

WI007 FRONT-END		**Taxa de Acerto (%) Usando Ruído Metal Cutting**		**HTK BACK-END**
Degradação		**Base Aurora Original (BAO)**	**Base Aurora com Metal-Cutting (BMC)**	**BMC - BAO**
Nível de SNR (dB)	**Clean**	98,68	98,68	-
	20	93,69	95,21	1,52
	15	92,18	94,34	2,16
	10	90,54	94,13	3,59
	5	85,90	91,16	5,26
	0	77,40	83,86	6,46
	-5	55,39	58,55	3,16
Média (20 a -5dB)		**82,51**	**85,24**	**3,69**

Depois do processamento, todos os resultados foram iguais aos testes anteriores, à exceção do ruído *metal-cutting*, ou seja, a modificação na base Aurora-1 alterou os resultados de taxa de acerto devido ao grupo de locuções modificadas na base de dados, mostrando que o procedimento foi bem-sucedido.

Analisando a Tabela 5.8a, é possível constatar que com o uso do F-E WI007 e treinamento em condições múltiplas, a taxa de acerto, em média, para o ruído *metal-cutting*, teve um aumento considerável, depois de incluir as características desse ruído na base de dados.

Conforme pode-se observar da Tabela 5.8a, para uma SNR = 5 dB, foi alcançada uma melhoria na taxa de acerto de 5,26%; para uma SNR = 10 dB a taxa de acerto melhorou 3,59%; para SNR = 15 dB houve uma melhora de 2,16%; e para SNR = 20 dB a taxa de acerto melhorou 1,52%. Também, pode-se notar que a taxa de acerto melhorou em 3,16% e 6,46% para uma SNR de -5 dB e 0 dB, respectivamente.

Para condições nas quais não há degradação nos sinais de teste (*Clean*), o resultado foi o mesmo, evidentemente devido ao uso do mesmo sinal de fala comparado à base de dados sem variação devido à ruídos.

Esse resultado permite concluir que a metodologia de avaliação da base de dados robusta, proposta neste trabalho, pode ser aplicada para verificação do balanceamento da mesma em termos dos ruídos que a constituem. Esta é uma ferramenta muito importante para Sistemas de Reconhecimento de Fala Robustos.

Por outro lado, analisando a Tabela 5.8b, observa-se que com o uso do F-E WI008 e treinamento em condições múltiplas, para o ruído *metal-cutting*, houve uma queda na taxa de acerto do sistema em todos os níveis de degradação.

Tabela 5.8b
Taxa de acerto para o F-E WI008 e treinamento sob condições múltiplas usando o ruído *metal-cutting* na base de dados

WI008 FRONT-END		**Taxa de Acerto (%) Usando Ruído Metal Cutting**		**HTK BACK-END**
Degradação		**Base Aurora Original (BAO)**	**Base Aurora com Metal-Cutting (BMC)**	**BMC - BAO**
Nível de SNR (dB)	**Clean**	**99,02**	**99,02**	-
	20	**95,02**	**95,02**	-3,37
	15	**93,69**	**93,69**	-4,57
	10	**92,67**	**86,49**	-6,18
	5	**88,44**	**80,22**	-8,22
	0	**80,62**	**70,11**	-10,51
	-5	**58,79**	**50,79**	-8,00
Média (20 a -5dB)		**84,87**	**78,06**	**-6,81**

Desta forma, a base de dados "dopada" com o ruído *metal-cutting*, usando o WI008 merece mais investigações. Esta investigação pode ser feita a partir da passagem do material de treinamento pelo mesmo tipo de filtro usado no WI008, entre outros estudos sobre a atuação dos filtros de redução de ruído especificamente sobre o de corte de metal.

Utilizando o WI007 (Tabela 5.8a), observa-se que a taxa de acerto é superior ao sistema que usa a base de dados original (Aurora-1) com o W008 (Tabela 5.8b) em todos os níveis de degradação, com uma

média de taxa de acerto igual a 86,2% do WI007 contra 84,87% do WI008. Este aspecto reforça a tese de que, com as características do ruído *metal-cutting* inseridas na base de dados, houve uma sensível melhora nas taxas de acerto do WI007, inclusive em relação ao uso do WI008 sem a presença dessas características no material de treino.

Para finalizar, levando-se em consideração o uso do WI007 na faixa de SNR de 5 a 20 dB, pode-se inferir da Tabela 5.8a que a taxa de acerto, para o ruído *metal-cutting* presente na base de dados, aumentou em média 3,69%.

5.3 Relação SNR vs. Pesq

Nesta seção são apresentados em gráficos os resultados da relação SNR vs. Pesq. Esses gráficos são produzidos com os dados experimentais, para os três testes (*testa*, *testb* e *testd*), fazendo um comparativo entre os testes da base Aurora-1 (*testa* e *testb*) e o teste produzido com os ruídos novos (*testd*).

5.3.1 Comparativo SNR vs. Pesq entre o testa e o testd

Ao levantar as curvas SNR vs. Pesq pode-se observar o comportamento dos ruídos e, devido às características apresentadas anteriormente, em especial do ruído *metal-cutting*, fazer inferências.

Desta forma, a análise do comportamento desses ruídos poderá ser feita com maior rigor, a respeito da influência do nível e ruídos empregados para degradar os sinais de fala. A Figura 5.16 mostra o comparativo da relação SNR vs. Pesq, do *testa* com o *testd*.

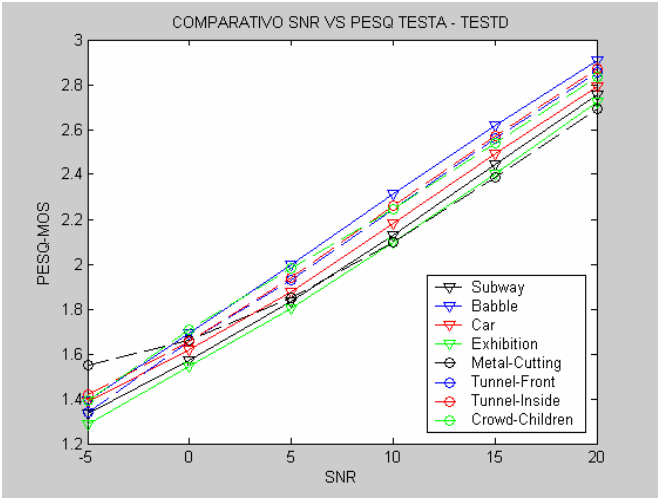


Figura 5.16: Comparativo SNR vs. Pesq testa-testd

No comparativo SNR vs. Pesq apresentado na Figura 5.16, observa-se que há uma tendência linear para o comportamento de todos os ruídos, ou seja, à medida que a SNR diminui, a qualidade do sinal de fala aumenta de forma aproximadamente linear.

Além disso, pode-se constatar que o ruído *metal-cutting* é o que mais degradou os sinais de fala para um valor SNR > 5 dB.

5.3.2 Comparativo SNR vs. Pesq entre o testb e o testd

No comparativo SNR vs. Pesq do *testb* com o *testd*, também observa-se que há uma tendência linear para o comportamento de todos os ruídos, e que à proporção que a SNR diminui, obviamente a qualidade do sinal de fala também diminui.

Além disso, pode-se observar que o ruído *metal-cutting* continua sendo o que mais degradou os sinais de fala para um valor SNR > 5dB.

A Figura 5.17 mostra o comparativo da relação SNR vs. Pesq, dos ruídos do *testb* com os ruídos do *testd*.

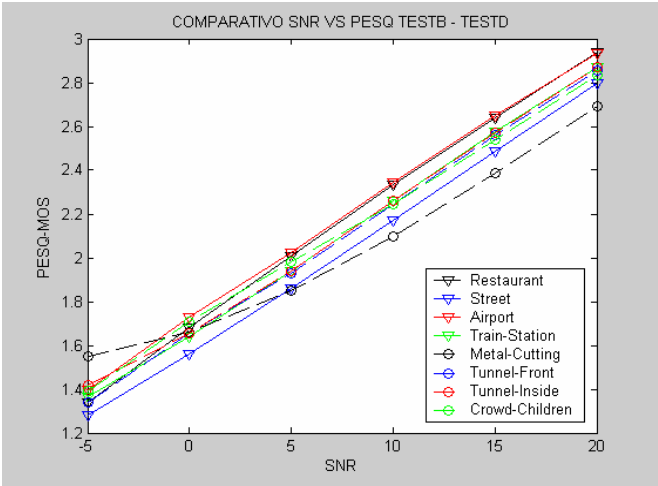


Figura 5.17: Comparativo SNR vs. Pesq testb-testd

Para finalizar a análise sobre os resultados mostrados aqui, chama-se a atenção para o fato de que diferentes tipos de ruídos produzem diferentes taxas de acerto em SRAF. Com isto pode-se concluir que, além dos níveis de relação sinal-ruído (SNR), os sistemas de reconhecimento de fala também são afetados devido ao tipo, como é o caso do ruído *metal-cutting*.

6. Conclusões

Neste trabalho foram estudados sistemas de reconhecimento automático de fala robustos, baseados na degradação por tipos e níveis de ruídos, assim como pelo efeito psicológico (efeito Lombard).

Desta forma, sugeriu-se uma metodologia de construção de uma base de dados Lombard, e foi mostrado como é possível elaborá-la. Ainda, depois do desenvolvimento da referida metodologia, foi construída uma base de dados Lombard com 22 locutores homens e 22 mulheres para futuros estudos.

Neste estudo, porém, os esforços foram focados no objetivo de construir uma metodologia de estudo e avaliação sobre sistemas de reconhecimento de fala robusto baseado na degradação dos sinais de fala por tipos e níveis de ruídos.

Assim, com o objetivo de investigar a construção de sistemas dessa natureza, degradados por ruídos, foram reunidos em um teste chamado *testd*, quatro ruídos novos que não estão presentes na base de dados Aurora-1.

Conforme pode-se observar no presente trabalho, foram feitas exaustivas comparações dos ruídos novos com os ruídos da base de dados Aurora-1, e a consequente caracterização dos mesmos em famílias para a avaliação da referida base de dados.

A influência dos sistemas com redução de ruído e com os testes sendo feitos sob condições múltiplas podem mudar sensivelmente a resposta de sistemas de RAF robustos.

A metodologia de avaliação da base de dados robusta proposta não se restringe apenas às bases produzidas por degradação de ruídos, podendo ser aplicada também as provenientes da degradação Lombard.

Observando os resultados, de uma maneira geral, pode-se afirmar que os ruídos novos apresentaram comportamento similar aos demais ruídos presentes na base Aurora-1, tanto para o WI007 quanto para o WI008, não descaracterizando os dados do novo teste para nenhum dos quatro ruídos novos propostos.

Este aspecto é positivo, pois fornece uma informação que permite concluir que o teste com os ruídos novos foi bem-sucedido, diante dos valores observados nas tabelas com as taxas de acerto apresentadas no capítulo 5, assim como no comportamento dos sistemas nas curvas SNR vs. TA , Pesq vs. TA e SNR vs. Pesq.

Analisando os resultados e os tipos de ruídos de cada teste, pode-se perceber que, em condições totalmente não degradadas (teste e treino limpos), para o WI007, todos os testes apresentaram taxa de acerto acima de 99%. Este resultado, apesar de não contribuir de forma significativa para o trabalho tratado aqui, fornece a informação de que o *testd* foi feito de forma coerente com os demais, no reconhecimento de fala, sem a presença de ruído ambiental.

De cada nível de SNR determinando a degradação dos sinais de fala no teste com os ruídos novos, pode-se perceber que os ruídos novos se comportam de modo similar aos ruídos presentes na base Aurora-1 em termos de afetar as taxas de acerto dos sistemas de RAF

empregados. Este aspecto pode ser constatado nas curvas Pesq vs. TA, entre outras apresentadas neste estudo.

A característica citada no parágrafo anterior também é observada quando a degradação dos sinais de fala de cada teste é reduzida em 5 dB. Neste sentido pode-se verificar que o teste com os ruídos novos também apresentam características similares às dos outros ruídos para ambas as faixas de degradação dos sinais de fala.

Nos comparativos SNR vs. Pesq foi observado que há uma tendência linear para o comportamento de todos os ruídos. Além disso, pôde-se constatar que para todos os testes, o ruído *metal-cutting* é o que mais degradou os sinais de fala, especialmente nas faixas de SNR acima de 15 dB.

Para todos os testes, a partir do valor SNR = 5 dB, foi possível perceber uma melhora acentuada na taxa de reconhecimento dos sistemas empregados.

Paralelamente, observou-se que diferentes tipos de ruídos produzem diferentes taxas de acerto em SRAF. Isso significa que, além dos níveis de relação sinal-ruído, os tipos também afetam sistemas de reconhecimento de fala. Este aspecto ficou comprovado especialmente com o ruído *metal-cutting*.

Analisando as curvas da relação Pesq vs. TA apresentadas no capítulo 5, independentemente do tipo de ruído, fica evidente que a taxa de reconhecimento cai abruptamente quando o Pesq é menor que 2, isto é, a SNR é menor que 5 dB.

Isso posto, torna-se necessário ressaltar que este trabalho apresentou algumas contribuições originais importantes para SRAF robustos, tais como:

- *testd* com ruídos novos para a base Aurora-1;
- uma metodologia de estudo e avaliação de ruídos para uma base de dados robusta;
- uma metodologia para construção de uma base de dados Lombard;
- um algoritmo para a solução da função logística, equação (5.1), de três parâmetros proposta em [6];
- um método de ajuste logístico (Mail) das curvas Pesq vs. TA, para estudo do comportamento de ruídos na degradação de sinais de fala e a influência nas taxas de acerto, com vantagens sobre o emprego do MRL e EMV;
- Tabelas de apresentação de dados.

A caracterização em grupos e/ou famílias de ruídos pode auxiliar na construção de bases de dados robustas, e, desta forma, indiretamente contribuir para um melhor desempenho dos sistemas.

Por exemplo, com o uso da metodologia proposta aqui, verificou-se que o ruído *metal-cutting* necessitava de estudos mais aprofundados. E, para um nível de precisão dos resultados, a base de dados robusta foi modificada para conter as características do comportamento da degradação dos sinais de fala.

Neste sentido, nos testes feitos, conforme apresentado na subseção 6.2.3, com a presença dos sinais degradados pelo ruído *metal-cutting* na base de dados robusta, obteve-se uma melhora média na taxa de acerto do sistema WI007 em mais de 3% para os níveis de degradação de 5 dB a 20 dB.

Ao se concluir este trabalho deixa-se como contribuição principal a metodologia empregada para estudo e avaliação de bases de dados em SRAF robustos, juntamente com as técnicas desenvolvidas aqui.

Ademais, a solução da equação (5.1), para a função logística proposta em [6], e o método de ajuste Mail, são contribuições que vão além do foco deste trabalho, pois o modelo matemático de solução proposto, além de simples, pode ser aplicado, com os devidos ajustes, em problemas de outras áreas de conhecimento, como na Estatística [118].

6.1 Perspectivas para trabalhos futuros

Ao realizar este trabalho sobre SRAF robustos, outros estudos foram iniciados para uma melhor investigação no objetivo de produzir alguma contribuição nova, por menor que fosse.

Um desses trabalhos foi um estudo sobre reconhecimento de palavras-chave (*word spotting*) e que pode ser empregado para melhorar a taxa de acerto em sistemas DSR e/ou de RAF.

Desta forma, um dos trabalhos que pode-se sugerir na área de RAF robustos para o futuro é o estudo de *word spotting* para obter melhores taxas de acerto nesses tipos de sistemas.

De forma complementar, durante o desenvolvimento deste trabalho foi produzido um sistema de reconhecimento de fala, para poucas palavras, na linguagem C/C++ e também no HTK.

O sistema desenvolvido na linguagem C/C++ focou o reconhecimento de comandos para robôs e o dicionário foi constituído de apenas três palavras (*siga*, *pare* e *volte*). O sistema com o uso do HTK foi construído com um dicionário de 20 palavras.

Os testes feitos com um robô do laboratório do s2i/DAS/UFSC resultaram no controle de partida, parada e retorno, e funcionou corretamente, com taxa de acerto de 100%. Ressalte-se, porém, que além de o dicionário conter poucas palavras, os comandos foram dados via arquivos de voz, ou seja, sem a presença de ruído ambiental.

O sistema produzido com o HTK foi projetado e executado com o uso de monofones e trifones, para o pequeno vocabulário de 20 palavras, entre elas 10 dígitos conectados que podiam ser expressos em sequência.

Os SRAF desenvolvidos no HTK para trifones (fones dependentes de contexto) e para monofones foram testados e obteve-se uma taxa de acerto superior a 95%. No caso dos testes com trifones, contudo, os resultados foram superiores aos dos monofones em aproximadamente 10%.

O sistema para trifones também foi testado em tempo real e com a presença de ruído ambiental. Esse teste, obviamente apresentou uma queda na taxa de acerto.

Neste sentido, sugere-se a produção de um SRAF com base em monofones e trifones para proferir comandos para que um robô execute tarefas cotidianas, como pegar um objeto e deixá-lo em um ambiente-alvo. Desta forma, partindo de estudos sobre SRAF robusto como o desenvolvido neste trabalho, pode-se estudar a tecnologia de fala e explorar os efeitos dos ruídos dos meios em que o robô transitar.

Como prosseguimento do presente trabalho, propõe-se um estudo sobre os ruídos novos e sua influência na fala Lombard. Sabe-se que ruídos não são os únicos fatores que afetam a qualidade da fala. Segundo [2], em certos casos as falas Lombard são as que mais afetam o desempenho em sistemas de RAF.

Como a variabilidade acústica [14] é um fator de erro predominante na comunicação homem-máquina, e como estudos sobre este aspecto estão sendo intensamente desenvolvidos, outra sugestão para trabalhos futuros é a realização de uma investigação do efeito do ruído para falantes não nativos, tanto para o efeito Lombard quanto para os efeitos dos níveis (SNR) e tipos de ruídos. Trabalhos deste tipo de pesquisa poderiam ser aplicados, principalmente, em sistemas de telecomunicações internacionais.

Além disso, um estudo para definir se é viável determinar uma família de ruídos apenas comparando os valores paramétricos ***a***, ***b***, e/ou ***c*** merece uma investigação.

Um estudo que também pode-se sugerir, no intuito de obter uma melhor taxa de acerto dos sistemas de RAF usados na base de dados Aurora-1, é sobre o desempenho nos modelos de treinamento.

Para o presente estudo, por exemplo, foi usado o sistema bem treinado do projeto Aurora-1, porém outra pesquisa pode ser feita com o objetivo de testar mudanças na parte de treino e propor algum outro modelo/projeto para a base de dados ruidosa do projeto Aurora, como por exemplo, a presença na base ruidosa de características do ruído *metal-cutting*.

Da mesma forma, pode-se produzir uma base de dados com os ruídos da base Aurora-1 mais o ruído *metal-cutting*, conforme indicado na subseção 6.3.2, porém após a mixagem dos ruídos, realizar uma filtragem desses sinais com os mesmos tipos de filtros de redução de ruídos usados no WI008.

Assim, com este procedimento torna-se possível verificar, tanto para o ruído *metal-cutting* como para os ruídos do projeto Aurora-1, entre outros, no caso particular do sistema com o uso do WI008, se há um melhor desempenho. Aplicando-se, porém, a metodologia de avaliação de bases de dados robustas proposta, pode-se ganhar tempo e reduzir esforços.

Para finalizar, é preciso ressaltar que todas as contribuições e sugestões apresentadas neste trabalho têm o objetivo de auxiliar na construção e avaliação, especialmente sobre o material de treinamento, de sistemas de reconhecimento de fala robustos. Nada impede, entretanto, que, com as devidas adaptações, as ideias expostas aqui

sejam aproveitadas para fazer estudos sobre base de dados limpas, isto é, em sistemas de reconhecimento de fala sem a presença de ruído ambiental, como a metodologia de avaliação da base de dados proposta e aplicada no capítulo 5.

Referências

[1] ETSI DSR Applications and Protocols Working Group AU/335/01, "New Aurora Activity for Standardization of a Front-End Extension for Tonal Language Recognition and Speech Reconstruction", June 2001.

[2] Hansen, J. H. L.; Varadarajan, V. "Analysis and Compensation of Lombard Speech Across Noise Type and Levels With Application to In-Set/Out-of-Set Speaker Recognition," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 17, No. 2, February 2009.

[3] AU/225/00. "Baseline Results for subset of SpeechDat-Car Finnish Database for ETSI STQ WI008 Advanced Front-end Evaluation", Nokia, Jan 2000.

[4] Young, Steve, et al. "Hidden Markov Model Toolkit - HTK Book", Version 3.4, Cambridge University Engineering Department. December 2006.

[5] International Telecommunication Union. "ITU-T Recommendation P.862," ITU-T Recommendations, Series P: Telephone Transmission Quality, Telephone Installations, Local Line Networks, Printed in Switzerland, Geneva, 2001.

[6] F. J. Fraga; C. A. Ynoguti; A. G. Chiovato. "Further Investigations on the Relationship between Objective Measures of Speech Quality and Speech Recognition Rates in Noisy Environments." *Proceedings of the ICSLP'2006*, Pittsburgh, Pennsylvania, p. 185-188, 2006.

[7] Å. Björck. “Numerical Methods for Least Squares Problems”, *SIAM*, 1996.

[8] J. Wolberg. “Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments”, *Springer*, 2005.

[9] Rabiner, L. R; Juang, B. H. “Fundamentals of Speech Recognition”, *Prentice-Hall International, Inc*, 1993.

[10] Huang, X. et al. “Spoken Language Processing: A Guide to Theory, Algorithm, and System Development”, *Prentice Hall*, 2001.

[11] Davis, S. B. “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-28, No. 4, p. 357-366, August 1980.

[12] Y. Shao, Z. Jin; D. Wang; S. Srinivasan. “An Auditory-Based Feature for Robust Speech Recognition”, *Proc. ICASSP-2009*, pp. 4.625-4.628.

[13] Tsuge, S. et al. “Study of Intra-Speaker’s Speech Variability Over Long and Short Time Periods for Speech Recognition”, *ICASSP 2006*, p. 397-400.

[14] Tsuge, S. et al. “Study of Relationships between Intra-speaker's Speech Variability and Speech Recognition Performance”, *IEEE/ISPACS2006*, Japan, 2006, p. 41-44.

[15] Barrault, L. et al. “Characterizing Feature Variability in Automatic Speech Recognition Systems”, *ICASSP 2006*, pp. V1029-V1032.

[16] Liang, M.-S.; Lyu, R.-Y.; Chiang, Y.-C. “Phonetic Transcription Using Speech Recognition Technique Considering Variations in Pronunciation”, *ICASSP 2007*, p. 109-112.

[17] White, R. G. et al. “Noise and Vibration”. *Ellis Horwood Limited*, 1982.

[18] Walshaw, A. C. “Mechanical Vibrations With Applications”. *Ellis Horwood Limited*, 1984.

[19] Rabiner, L. R. et al. “Digital Processing of Speech Signals”, *Prentice Hall*, 1978.

[20] Rabiner, L. R. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, *Proceedings of the IEEE*, Vol. 77, No. 2, p. 257-286, February 1989.

[21] Brémaud, P. “Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues”, *Springer*, 1999.

[22] Chou, W.; Juang, B. H. “Pattern Recognition in Speech and Language Processing”, *CRC Press*, USA, 2003.

[23] Picone, J. “Fundamentals of Speech Recognition: a Short Course”, *Institute for Signal and Information Processing (ISIP)*. Department of Electrical and Computer Engineering, Mississippi State University and Texas Instruments. May, 1996.

[24] Haykin, S. “Redes Neurais, Princípios e Prática”, P. Alegre, Bookman, 2ª Edição, 2004.

[25] Oviatt, S. “User-Centered Modeling and Evaluation of Multimodal Interfaces”, *Proc. IEEE*, Vol. 91, No 9, p. 1457-1468, September 2003.

[26] Borgström, B. J.; Alwan, A. “Utilizing Compressibility in Reconstructing Spectrographic Data, With Applications to Noise Robust ASR”, *IEEE Signal Processing Letters*, Vol. 16, No. 5, p. 398-401, May 2009.

[27] Windmann, S.; Haeb-Umbach, R. “Parameter Estimation of a State-Space Model of Noise for Robust Speech Recognition”, *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 17, No. 8, p. 1.577-1.590, November, 2009.

[28] http://www.cs.cmu.edu/~robust (último acesso em: 14 jan. 2010).

[29] Lockwood, P.; Boudy, J.; Blanchet, M. “Non-Linear Spectral Subtraction (NSS) and Hidden Markov Models for Robust Speech Recognition in Car Noise Environments”, *IEEE Transc*., Matra Communication, France, 1992.

[30] http://portal.etsi.org/stq/kta/DSR/dsr.asp (último acesso em: 15 out. 2009).

[31] Hai, J.; Joo, E. M. “Improved Linear Predictive Coding Method for Speech Recognition”, *IEEE/ICICS-PCM*, p. 1.614-1.618, 15-18 December, 2003.

[32] Tang, Y. Y.; Li, T.; Suenl, C. Y. “VLSI Arrays for Speech Processing with Linear Predictive Coding”, *IEEE Transc*., 1994.

[33] Bogert, B. P.; Healy, M. J. R.; Tukey, J. W. “The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking Time Series Analysis”, *M. Rosenblatt Ed*., 1963, p. 209–243.

[34] Oppenheim, A. V.; Schafer, R. W. “From Frequency to Quefrency: A History of the Cepstrum”, *IEEE Signal Processing Magazine*, p. 95-106, September 2004.

[35] Oppenheim, A. V.; Schafer, R. W. “Homomorphic Analysis of Speech”, *IEEE Transactions on Audio and Electroacoustics* Vol. Au-16, No. 2 p. 221-226, June 1968.

[36] Rabiner, L. R.; Juang, B. H. “An Introduction to Hidden Markov Models”, *IEEE ASSP Magazine*, p. 4-16, January 1986.

[37] Juang, B. H.; Rabiner, L. R. “Automatic Speech Recognition – A Brief History of the Technology Development. Georgia Institute of Technology”, Atlanta Rutgers University and the University of California, Santa Bárbara, August, 10, 2004.

[38] Gajic´, B.; Paliwal, K. K. “Robust Speech Recognition in Noisy Environments Based on Subband Spectral Centroid Histograms”, *IEEE Transac. on Audio, Speech, and Language Processing*, Vol. 14, No. 2, p. 600-608, March 2006.

[39] Wolfe, P. J.; Godsill, S. J. “Perceptually Motivated Approaches to Music Restoration”, University of Cambridge, Signal Processing Group, Department of Engineering, Journal of New Music Research 2001, Vol. 30, No. 1, p. 83–92.

[40] www.waseda.jp/top/index-e.html (último acesso em: julho/2007).

[41] Duhamel, P.; Vetterli, M. “Fast Fourier Transform: A Tutorial Rewiew and a State of the Art”, *IEEE Signal Processing*, p. 259-299, 1990

[42] Juang, B. H.; Levinson, S. E.; Sondhi, M. M. “Maximum Likelihood Estimation for Multivariate Mixture Observations of Markov Chains”, *IEEE Trans. Information Theory*, Vol. It-32, No. 2, p. 307-309, March 1986.

[43] Kim, W.; Hansen, J. H. L. “Time–Frequency Correlation-Based Missing-Feature Reconstruction for Robust Speech Recognition in Band-Restricted Conditions”. IEEE Trans. on Audio, Speech, and Language Processing, Vol. 17, Nº. 7, Sep. 2009, p. 1.292–1.304.

[44] Park, K. Y.; Kim, H. S. “Narrowband to wideband conversion of speech using GMM-based transformation”. Proc. ICASSP’00, Jun. 2000, p. 1.847–1.850.

[45] Cox, R.; V.; Malah, D. “A Technique for Perceptually Reducing Periodically Structured Noise in Speech,” *IEEE,* Bell Laboratories Murray Hill, New Jersey, 1981.

[46] Mansour D.; Hang, B. H. “A Family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition”, *IEEE*, AT&T Bell Labs, 1988.

[47] Hansen, J. H. L.; Clements, M. A. “Stress Compensation and Noise Reduction Algorithms for Robust Speech Recognition”, *IEEE Trans*c., 1989.

[48] Takebayashi, Y.; Tsuboi, H.; Kanazawa, H. “A Robust Speech Recognition System Using Word-Spotting With Noise Immunity Learning”, *IEEE Transac*. Toshiba Corporation, Research & Development Center, Japan, 1991.

[49] Sullivan, T.; M.; Stern, R., M. “Multi-Microphone Correlation-Based Processing for Robust Speech Recognition”, *IEEE Transc*., Department of Electrical and Computer Engineering School of Computer Science Carnegie Mellon University, Pittsburgh, 1993.

[50] Yang, R.; Haavisto, P. “An Improved Noise Compensation Algorithm for Speech Recognition in Noise”, *IEEE Transc*., p. 49-52, 1996.

[51] Salonidis, T., Digalakis, V., “Robust Speech Recognition for Multiple Topological Scenarios of the GSM Mobile Phone System”, *IEEE Transc*., Technical University GREECE, 1998.

[52] Shozakai, M.; Nakamura, S.; Shikano, K.; “Robust Speech Recognition in Car Environments”, *IEEE – Intenational Conf. on Acoustic Speech, and Signal Processing*, Vol. 1, p. 269-272, 1998.

[53] Ganapathiraju A. “Support Vector Machines for Speech Recognition”, Dissertation for Degree of Doctor of Philosophy. Department of Electrical and Computer Engineering, Mississippi State University, May 2002.

[54] Cristianini, N.; Shawe-Taylor, J. “An Introduction to Support Vector Machines”. Cambridge University Press, 2005.

[55] V. Vapnik. “The Nature of Statistical Learning Theory”, New York, *Springer-Verlag*, 1995.

[56] Salomon, J. “Support Vector Machines for Phoneme Classification”. Master of Science School of Artificial Intelligence Division of Informatics University of Edinburgh, 2001.

[57] Abe, S. “Support Vector Machines for Pattern Classification”. Kobe University, *Springer*, 2005.

[58] Kim, L. Y.; Cho, H.Y.; Oh, Y.H. “Missing Data Techniques Using Voicing Probability for Robust Automatic Speech Recognition”, *IEEE - Electronics Letters*, Vol. 37 No. 11, May 2001.

[59] Jing, Z.; Johnson, M. H. “Auditory-Modeling Inspired Methods of Feature Extraction for Robust Automatic Speech Recognition”, *IEEE-Transc.*, 2002.

[60] Hacioglu, K.; Pellon, B. “A Distributed Architecture For Robust Automatic Speech Recognition”, *ICASSP/2003*.

[61] Chen, J.; Benesty, J.; Huang, Y.; Doclo, S. “New insights into the noise reduction Wiener filter”. IEEE Trans. Audio, Speech, and Language Process., vol.14, nº 4, p.1.218–1.234, July 2006.

[62] Reuven, G.; Gannot, S.; Cohen, I. “Multichannel Acoustic Echo Cancellation and Noise Reduction in Reverberant Environments Using The Transfer-Function GSC”, *ICASSP/2007*, Vol I, p. 81-84.

[63] Kim, J.; Lee, H.; Ryu, W.; et al. “Improved Noise Reduction with Packet Loss Recovery Based on Post-Filtering over IP Networks”. IEICE Trans. Commun., Vol.E91–B, Nº.3, March 2008, p. 975-979.

[64] Doclo, S.; Moonen, M.; Clippel; E. D. “Combined acoustic echo and noise reduction using GSVD-based optimal filtering”. In IEEE Int. Conf. Acoust. Speech and Sig. Proc. (ICASSP’00), June 2000, p. 1.061-1.064.

[65] Wang, D.; Brown, G. J. “Computational Auditory Scene Analysis: Principles, Algorithms, and Applications”, *Wiley/IEEE Press*, New York, 2006 (Book Review).

[66] Li, P.; Guan, Y.; Xu, B.; Liu, W. “Monaural Speech Separation Based on Computational Auditory Scene Analysis and Objective Quality Assessment of Speech”, *IEEE Computer Society*, Proc. of the First International Conference on Innovative Computing, Information and Control (ICICIC'06).

[67] Okuno, H. G.; Nakadai, K. “Computational Auditory Scene Analysis and Its Application to Robot Audition”, *IEEE Speech Communication and Microphone Arrays*. HSCMA 2008, p. 124 – 127.

[68] Souden, M.; Benesty, J.; Affes, S. “Microphone Arrays For Noise Reduction With Low Signal Distortion in Room Acoustics”, *ICASSP 2008*, p. 77 – 80.

[69] Akhtar, M. T.; Mitsuhashi, W. “Robust Adaptive Algorithm For Active Noise Control of Impulsive Noise Acoustics”, *ICASSP 2009*, p. 261 – 264.

[70] Morais, E.; Viera, J. M.; Arantes, P. et al. “Metodologias para Projeto e Aquisição de uma Base de Dados Lingüísticos Visando ao Treinamento e à Avaliação de Sistemas de Reconhecimento de Fala”, XXV Congresso da Sociedade Brasileira de Computação, 22 a 29 de julho de 2005, São Leopoldo, RS.

[71] AU/301/01 v3. “Advanced DSR Front-end: Definition of required performance characteristics”. STQ Aurora DSR Working Group, October 2001.

[72] Hirsch, H.G.; Pearce, D. “The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions”, ISCA ITRW ASR2000, Paris, France, September 2000.

[73] International Telecommunication Union. “ITU-T Recommendation G.712,” ITU-T Recommendations, Series G: Transmission Systems and Media, Digital Systems and Networks, Printed in Switzerland, Geneva, 2001.

[74] Pearce, D. “Enabling New Speech Driven Services for Mobile Devices: An overview of the ETSI Standards Activities for Distributed Speech Recognition Front-Ends,” AVIOS 2000: The Speech Applications Conference, May 22-24, 2000. San Jose, CA, USA.

[75] Leonard, R. G. “A database for speaker-independent digit recognition”, *Proc. IEEE International Conference on Acustics, Speech, and Signal Processing*, 1984, vol. 3, p. 4.211-4.214.

[76] http://www.etsi.org/aurora/ (último acesso em: 29 out. 2009).

[77] http://www.elda.org/ (último acesso em: 03 mar. 2009).

[78] http://portal.etsi.org/stq/Summary.asp (último acesso em: 04 mar. 2009).

[79] AU/384/02. “DSR Front End LVCSR Evaluation”. Aurora Working Group, France, December 06, 2002.

[80] ETSI ES 201 108 v1.1.2. “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms”, April 2000.

[81] Becchetti, C. B. et al. “Speech Recognition – Theory and C++ Implementation”, *John Wiley & Sons Ltda*, New York, 1999.

[82] AU/372/00 v7. “Overview of Evaluation Criteria for Advanced Distributed Speech”. STQ Aurora DSR Working Group, October 2001.

[83] Rabiner, L. R.; Juang, B. H. “Fundamentals of Speech Recognition”, *Prentice-Hall International*, *Inc*, USA, 1993.

[84] Ynoguti, C. A. “Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov”. Tese de Doutorado, Decom/FEEC/Unicamp, maio de 1999.

[85] http://htk.eng.cam.ac.uk/ (último acesso em: 14 jan. 2010).

[86] http://www.opticom.de/technology/pesq.html (último acesso em: 14 jan. 2010).

[87] International Telecommunication Union. “ITU-T Recommendation P.10/G.100,” ITU-T Recommendations, Series P: Telephone Transmission Quality, Telephone Installations, Local Line Networks, Printed in Switzerland, Geneva, 2006.

[88] Zhu, D.; Huo, Q. “Irrelevant variability normalization based HMM Training Using MAP Estimation of Feature Transforms for Robust Speech Recognition”, *ICASSP-2008*, p. 4.717-4.720.

[89] Huo, Q.; Zhu, D. “Robust Speech Recognition Based on Structured Modeling, Irrelevant Variability Normalization and Unsupervised Online Adaptation”, *ICASSP/2009*, p. 4.637 – 4.640, 2009.

[90] ITU-T Software Tool Library, Group on Software Tools, Geneva, August 2005.

[91] Lane, H. L.; Tranel, B.; Sisson, C. “Regulation of voice communication by sensory dynamics”, *J. Acoust. Soc. Amer*., vol. 32, p. 451–454, 1970.

[92] Lane, H. L.; Tranel, B. “The Lombard sign and the role of hearing in speech”, *J. Speech Hear*. Res., vol. 14, pp. 677–709, 1971.

[93] Hansen, J. H. L. “Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition”, Ph.D. dissertation, School of Elect. Eng., Georgia Inst. of Technol., Atlanta, 1988.

[94] Hansen, J. H. L.; Bou-Ghazale, S. “Getting started with Susas: A speech under simulated and actual stress database”, *Proc. Eurospeech’ 97*, Rhodes, Greece, Sep. 1997, vol. 4, p. 1.743–1.746.

[95] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S78 (Último acesso em: 16 fev. 2010).

[96] Stanton, B. J.; Jamieson, L. H.; Allen, G. D. “Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions”, *ICASSP-1988*, p. 331–334.

[97] Summers, W.V. et al. “Effects of noise on speech production : Acoustical and perceptual analyses,” *J. Acous. Soc. Amer.*, pp. 917–928, Sep. 1988.

[98] Junqua, J. C. “The Lombard reflex and its role on human listeners and automatic speech recognizers,” *J. Acoust. Soc. Amer.*, vol. 93, p. 510–524, Jan. 1993.

[99] Pickett, J. M. “Effects of vocal force on the intelligibilty of speech sounds”, *J. Acous. Soc. Amer*., pp. 902–905, Sep. 1956.

[100] Dreher, J.; Neil, J. “Effects of ambient noise on speaker intelligibility for words and phrases”, J. Acous. Soc. Amer., p. 1.320–1.323, Dec. 1957.

[101] Ladgefoged, P. “Three Areas of Experimental Phonetics”. London, U.K.: Oxford Univ. Press.

[102] Li, Z.; Duraiswami, R.; Davis, L. S. “Recording and Reproducing High Order Surround Auditory Scenes for Mixed and Augmented Reality”, Computer Society, Proceedings of the Third IEEE and ACM International Symposium on Mixed and Augmented Reality (Ismar 2004).

[103] http://www.coolest-gadgets.com/20090323/sanwa-throat-microphone/ (último acesso em: 30 out. 2009).

[104] 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. P. 23-41. Dezembro de 1992.

[105] Sun, H.; Shue, L.; Chen, J. “Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech”, Icassp’2004. Montreal, Canada, p. 865-868.

[106] Munasinghe, S. R.; Nakamura, M. “Hyperbolic Tangent Function Based Force-Position Compliant Controller for Robotic Devices”. International Conference on Control, Automation and Systems, Oct. 17-20, 2007 in Coex, Seoul, Korea.

[107] Agresti, A. “Categorical Data Analysis”, JW&S, INC, Secound Edition, 2002.

[108] Cramer, J. S. “Econometric Applications of Maximum Likelihood Methods”, Cambridge University Press, 1986.

[109] Campos, F. F. “Algoritmos Numéricos”. Rio de Janeiro, *Livros Técnicos e Científicos Editora S.A*, 2a Ed., 2007.

[110] Barroso, L. C. et al. “Cálculo Numérico – Aspectos Teóricos e Computacionais”. São Paulo, *Editora Makron Books*, 2ª ed., 1996.

[111] Chatterjee, S.; Hadi, A. S.; Price, B. “Regression Analysis by Example”, *JW&S Publication*, Third Edition, 2000.

[112] Souza, M. J. F. “Ajuste de Curvas Pelo Método dos Mínimos Quadrados”. Departamento de Computação, Universidade Federal de Ouro Preto, notas de aula, Métodos Numéricos, 2???.

[113] Aldrich, J. "Maximum Likelihood", *Statistical Science*, 1997.

[114] http://www.minitabbrasil.com.br (último acesso em: 14 dez. 2009).

[115] Soares, J. F. et. al. “Introdução à Estatística”, Belo Horizonte, *Livros Técnicos e Científicos Editora*, 1991.

[116] Barbetta, P. A. “Estatística Aplicada às Ciências Sociais”, Florianópolis, *Editora da UFSC*, 5ª edição, 2002.

[117] Montgomery, D. C.; Runger, G. C. “Estatística Aplicada e Probabilidade Para Engenheiros”, *LTC*, 2ª edição, 2003.

[118] Andrade, D. F.; Tavares, H. R.; Valle, R. C. “Teoria da Resposta ao Item: Conceitos e Aplicações”, Sinape, 2000.

Apêndice A

Intervalo para a solução do parâmetro *c*

Neste apêndice é definido o intervalo para a solução da equação (4.10), intervalo para a determinação do parâmetro ***c***, considerando o campo dos reais e o intervalo teórico de variação mínimo e máximo da taxa de acerto (TA) ou (y_n) dos sistemas de reconhecimento automático de fala, isto é: $0 \leq y_n \leq 100$. Também é apresentada a derivada de (4.10) utilizada no algoritmo de Newton-Raphson.

Da equação (4.10) pode-se obter a equação (A1):

$$\frac{1}{K_2}\ln\left(\frac{100}{y_1+100C}-1\right)-\frac{K_1}{K_2}\ln\left(\frac{100}{y_2+100C}-1\right)+K_3\left(\ln\left(\frac{100}{y_1+100C}-1\right)-\ln\left(\frac{100}{y_2+100C}-1\right)\right)-\ln\left(\frac{100}{y_3+100C}-1\right)=0 \quad \text{(A1)}$$

Analisando a equação (A1) para os reais, temos:

$$y_n+100.C>0 \quad \text{logo:} \quad C>-\frac{y_n}{100} \quad \text{(A2)}$$

$$\frac{100}{y_n+100.C}>1 \quad \text{logo:} \quad C<1-\frac{y_n}{100} \quad \text{(A3)}$$

De (A2) e (A3) vem (A4):

$$-\frac{y_n}{100}<C<1-\frac{y_n}{100} \quad \text{(A4)}$$

Assim, para os extremos de y_n, $\boldsymbol{c}_{y=0} =]0, 1[$ e $\boldsymbol{c}_{y=100} =]-1, 0[$, pode-se dizer que para qualquer valor y_n de (4.10) ou (A1) obtemos a solução no intervalo $\boldsymbol{c} \in]-1, 1[$ ou $-1 < \boldsymbol{c} < 1$.

Dado que $y_3 > y_2 > y_1$, para os sistemas de reconhecimento automático de fala temos o intervalo:

$$-\frac{y_1}{100}<C<1-\frac{y_3}{100} \quad \text{(A5)}$$

Logo, para determinar um valor inicial e aplicar um método numérico para a solução de (4.10), pode-se atribuir um valor de ***c*** dentro do intervalo de (A5). No caso de uma função monótona, o Método de Newton-Raphison (MNR) é um método eficaz e mais indicado.

Para a aplicação do MNR, por exemplo, no Matlab, é necessário derivar a equação (4.10) em relação ao parâmetro ***c***. A equação (4.10) é apresentada com modificações no formato em (A6), e a respectiva derivada é apresentada na equação (A7).

$$f(c)=((1/k_2)+k_3).(\log((100/(y_1+100c))-1))+ ((-k_1/k_2)-k_3).(\log((100/(y_2+100c))-1))-\log((100/(y_3+100c))-1) \quad (A6)$$

$$df(c)=((1/k_2)+k_3).(-10000/((y_1+100c).(100-y_1-100c)))+ ((-k_1/k_2)-k_3).(-10000/((y_2+100c).(100-y_2-100c)))- (-10000/((y_3+100c)(100-y_3-100c))) \quad (A7)$$

Para escrever a derivada de f(c) em termos mais compactos/distintos, denominando o primeiro termo de (A6) como T_1, o segundo como T_2 e o terceiro como T_3, temos:

$$T_1 = ((1/k_2)+k_3).(\log((100/(y_1+100c))-1)) \quad (A8a)$$

$$T_2 = ((-k_1/k_2)-k_3).(\log((100/(y_2+100c))-1)) \quad (A8b)$$

$$T_3 = -\log((100/(y_3+100c))-1) \quad (A8c)$$

Definindo a derivada de cada termo das equações (A8), em relação ao parâmetro ***c***, como dT_1, dT_2, dT_3, temos as equações (A9):

$$dT_1 = ((1/k_2)+k_3).(-10000/((y_1+100.c).(100-y_1-100.c))) \quad (A9a)$$

$$dT_2 = ((-k_1/k_2)-k_3).(-10000/((y_2+100.c).(100-y_2-100.c))) \quad (A9b)$$

$$dT_3 = -\log((100/(y_3+100.c))-1) \quad (A9c)$$

Desta forma, para simplificar, a derivada de (A6) pode ser escrita como a equação (A10):

$$df(c) = dT_1 + dT_2 + dT_3 \quad \text{(A10)}$$

Apêndice B

Ambientes de gravação dos ruídos do testd

Figura B1a: Galpão onde foi gravado o ruído *metal-cutting*

Figura B1b: Local onde foi gravado o ruído *tunnel-front*

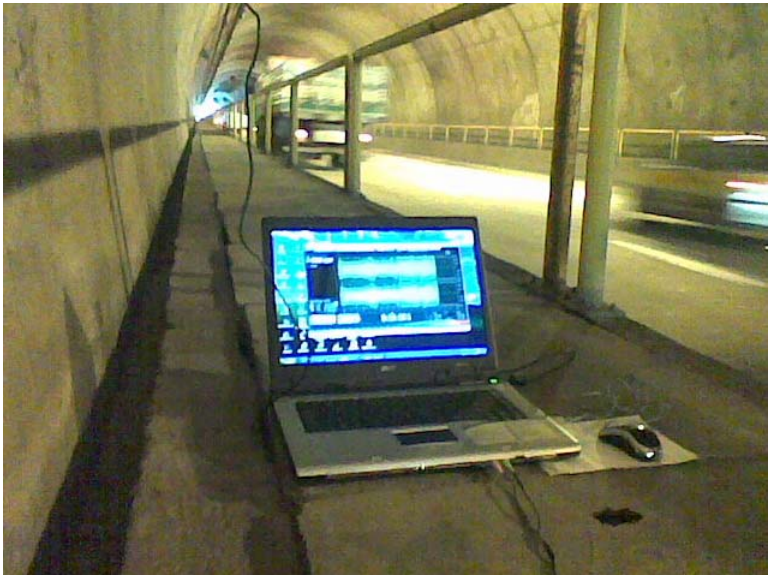


Figura B1c: Local onde foi gravado o ruído *tunnel-inside*

Figura B1d: Local onde foi gravado o ruído *crowd-children*

Apêndice C

Frases para produção da base de dados Lombard

Tabela C1: Frases do Português Brasileiro [104]

Número da Locução	Locução
1	Os maiores picos da Terra ficam debaixo d'água.
2	A inauguração da vila é quarta-feira.
3	Só vota quem tiver o título de eleitor.
4	É fundamental buscar a razão da existência.
5	A temperatura só é boa mais cedo.
6	Em muitas regiões a população está diminuindo.
7	Nunca se pode ficar em cima do muro.
8	Pra quem vê de fora o panorama é desolador.
9	É bom te ver colhendo flores.
10	Eu me banho no lago ao amanhecer.
11	As crianças conheceram o filhote de ema.
12	A bolsa de valores ficou em baixa.
13	O Congresso volta atrás em sua palavra.
14	A médica receitou que eles mudassem de clima.
15	Não é permitido fumar no interior do ônibus.
16	A apresentação foi cancelada por causa do som.
17	Uma garota foi presa ontem à noite.
18	O prato do dia é couve com atum.
19	Eu viajarei ao Canadá amanhã.
20	A balsa é o meio de transporte daqui.

Tabela C2: Frases do Inglês obtidas da base de dados Aurora-1 [1]

Número da Locução	**Locução**
1	Oh
2	Z
3	1
4	2
5	3
6	4
7	5
8	6
9	7
10	8
11	9
12	8 6 Z 1 1 6 2
13	O O 2 1 6 4 1
14	4 3 O 6 5 7 1
15	9 8 Z 7 4 3 7
16	7 Z 4 6 9 8 9
17	Z 4 7 4 5 6 3
18	9 9 8 9 2 7 6
19	2 4 7 3 4 8 Z
20	5 Z 5 5 3 9 Z

Apêndice D

Resultados em tabelas dos testes da base Aurora-1

1 Resultados do testa usando o WI007

Tabela D1a
Taxa de acerto para o testa usando F-E WI007 e treinamento sob condições limpas

WI007 FRONT-END		Testa – Condições Limpas Taxa de Acerto (%) Por Tipo de Ruído				HTK BACK-END	
Degradação		Subway	Babble	Car	Exhibition	Média (%)	
Nível de SNR (dB)	Clean	98,93	99,00	98,96	99,20	99,02	Por Nível
	20	97,05	90,15	97,41	96,39	95,25	
	15	93,49	73,76	90,04	92,04	87,33	
	10	78,72	49,43	67,01	75,66	67,71	
	5	52,20	26,81	34,09	44,83	39,48	
	0	26,01	9,28	14,46	18,05	16,95	
	-5	11,18	1,57	9,39	9,60	7,94	
		Média (%)					
Faixa de SNR (dB)	15 a 20	95,27	81,96	93,73	94,22	91,29	Por Faixa
	10 a 20	89,75	71,11	84,82	88,03	83,43	
	5 a 20	80,37	60,04	72,14	77,23	72,44	
	0 a 20	69,49	49,89	60,60	65,39	61,34	
	-5 a 20	59,78	41,83	52,07	56,10	52,44	

Tabela D1b
Taxa de acerto para o testa usando F-E WI007 e treinamento sob condições múltiplas

WI007 FRONT-END		Testa – Condições Múltiplas Taxa de Acerto (%) Por Tipo de Ruído				HTK BACK-END	
Degradação		Subway	Babble	Car	Exhibition	Média (%)	
Nível de SNR (dB)	Clean	98,68	98,52	98,39	98,49	98,52	Por Nível
	20	97,61	97,73	98,03	97,41	97,70	
	15	96,47	97,04	97,61	96,67	96,95	
	10	94,44	95,28	95,74	94,11	94,89	
	5	88,36	87,55	87,8	87,6	87,83	
	0	66,9	62,15	53,44	64,36	61,71	
	-5	26,13	27,18	20,58	24,34	24,56	
		Média (%)					
Faixa de SNR (dB)	15 a 20	97,04	97,39	97,82	97,04	97,32	Por Faixa
	10 a 20	96,17	96,68	97,13	96,06	96,51	
	5 a 20	94,22	94,40	94,80	93,95	94,34	
	0 a 20	88,76	87,95	86,52	88,03	87,82	
	-5 a 20	78,32	77,82	75,53	77,42	77,27	

2 Resultados do testb usando o WI007

Tabela D2a
Taxa de acerto para o testb usando F-E WI007 e treinamento sob condições limpas

WI007 FRONT-END		**Testb – Condições Limpas Taxa de Acerto (%) Por Tipo de Ruído**				**HTK BACK-END**	
Degradação		**Restaurant**	**Street**	**Airport**	**Train-station**	**Média (%)**	
Nível de SNR (dB)	**Clean**	98,93	99,00	98,96	99,20	99,02	**Por Nível**
	20	89,99	95,74	90,64	94,72	92,77	
	15	76,24	88,45	77,01	83,65	81,34	
	10	54,77	67,11	53,86	60,29	59,01	
	5	31,01	38,45	30,33	27,92	31,93	
	0	10,96	17,84	14,41	11,57	13,70	
	-5	3,47	10,46	8,23	8,45	7,65	
		Média (%)					
Faixa de SNR (dB)	**15 a 20**	83,12	92,10	83,83	89,19	87,06	**Por Faixa**
	10 a 20	73,67	83,77	73,84	79,55	77,71	
	5 a 20	63,00	72,44	62,96	66,65	66,26	
	0 a 20	52,59	61,52	53,25	55,63	55,75	
	-5 a 20	44,41	53,01	45,75	47,77	47,73	

Tabela D2b
Taxa de acerto para o testb usando F-E WI007 e treinamento sob condições múltiplas

WI007 FRONT-END		Testb – Condições Limpas Taxa de Acerto (%) Por Tipo de Ruído				HTK BACK-END	
Degradação		Restaurant	Street	Airport	Train-station	Média (%)	
Nível de SNR (dB)	Clean	98,68	98,52	98,39	98,49	98,52	Por Nível
	20	96,87	97,58	97,44	97,01	97,23	
	15	95,3	96,31	96,12	95,53	95,82	
	10	91,96	94,35	93,29	92,87	93,12	
	5	83,54	85,61	86,25	83,52	84,73	
	0	59,29	61,34	65,11	56,12	60,47	
	-5	25,51	27,6	29,41	21,07	25,90	
		Média (%)					
Faixa de SNR (dB)	15 a 20	96,09	96,95	96,78	96,27	96,52	Por Faixa
	10 a 20	94,71	96,08	95,62	95,14	95,39	
	5 a 20	91,92	93,46	93,28	92,23	92,72	
	0 a 20	85,39	87,04	87,64	85,01	86,27	
	-5 a 20	75,41	77,13	77,94	74,35	76,21	

3 Resultados do testc usando o WI007

Tabela D3a
Taxa de acerto para o testc usando F-E WI007 e treinamento sob condições limpas

WI007 FRONT-END		**Testc – Condições Limpas Taxa de Acerto (%) Por Tipo de Ruído**		**HTK BACK-END**	
Degradação		**Subway (MIRS)**	**Street (MIRS)**	**Média (%)**	
Nível de SNR (dB)	**Clean**	99,14	98,97	99,06	**Por Nível**
	20	93,49	95,13	94,31	
	15	86,77	88,91	87,84	
	10	73,90	74,40	74,15	
	5	51,27	49,21	50,24	
	0	25,42	22,91	24,17	
	-5	11,82	11,15	11,49	
		Média (%)			
Faixa de SNR (dB)	**15 a 20**	90,13	92,02	91,08	**Por Faixa**
	10 a 20	84,72	86,15	85,43	
	5 a 20	76,36	76,91	76,64	
	0 a 20	66,17	66,11	66,14	
	-5 a 20	57,11	56,95	57,03	

Tabela D3b
Taxa de acerto para o testc usando F-E WI007 e treinamento sob condições múltiplas

WI007 FRONT-END		**Testc – Condições Limpas Taxa de Acerto (%) Por Tipo de Ruído**		**HTK BACK-END**	
Degradação		**Subway (MIRS)**	**Street (MIRS)**	**Média (%)**	
Nível de SNR (dB)	**Clean**	98,50	98,58	98,54	**Por Nível**
	20	97,30	96,55	96,93	
	15	96,35	95,53	95,94	
	10	93,34	92,50	92,92	
	5	82,41	82,53	82,47	
	0	46,82	54,44	50,63	
	-5	18,91	24,24	21,58	
		Média (%)			
Faixa de SNR (dB)	**15 a 20**	96,83	96,04	96,43	**Por Faixa**
	10 a 20	95,66	94,86	95,26	
	5 a 20	92,35	91,78	92,06	
	0 a 20	83,24	84,31	83,78	
	-5 a 20	72,52	74,30	73,41	

4 Resultados do testa usando o WI008

Tabela D4a
Taxa de acerto para o testa usando F-E WI008 e treinamento sob condições limpas

WI008 FRONT-END		**Testa – Condições Limpas Taxa de Acerto (%) Por Tipo de Ruído**				**HTK BACK-END**	
Degradação		**Subway**	**Babble**	**Car**	**Exhibition**	**Média (%)**	
Nível de SNR (dB)	**Clean**	99,08	99,00	99,05	99,23	99,09	**Por Nível**
	20	97,91	98,31	98,48	97,90	98,15	
	15	96,41	96,89	97,58	96,82	96,93	
	10	92,23	92,35	95,29	92,78	93,16	
	5	83,82	81,08	88,49	84,05	84,36	
	0	61,93	51,90	66,42	63,28	60,88	
	-5	30,86	19,71	30,84	32,86	28,57	
		Média (%)					
Faixa de SNR (dB)	**15 a 20**	97,16	97,60	98,03	97,36	97,54	**Por Faixa**
	10 a 20	95,52	95,85	97,12	95,83	96,08	
	5 a 20	92,59	92,16	94,96	92,89	93,15	
	0 a 20	86,46	84,11	89,25	86,97	86,70	
	-5 a 20	77,19	73,37	79,52	77,95	77,01	

Tabela D4b
Taxa de acerto para o testa usando F-E WI008 e treinamento sob condições múltiplas

WI008 FRONT-END		**Testa – Condições Múltiplas Taxa de Acerto (%) Por Tipo de Ruído**				**HTK BACK-END**	
Degradação		**Subway**	**Babble**	**Car**	**Exhibition**	**Média (%)**	
Nível de SNR (dB)	**Clean**	99,02	98,79	98,93	99,14	98,97	**Por Nível**
	20	98,59	98,58	98,57	98,27	98,50	
	15	97,51	97,97	98,39	97,56	97,86	
	10	95,43	96,04	97,35	95,40	96,06	
	5	91,43	90,21	93,83	90,25	91,43	
	0	75,41	68,50	80,52	76,09	75,13	
	-5	39,73	30,23	40,47	45,26	38,92	
		Média (%)					
Faixa de SNR (dB)	**15 a 20**	98,05	98,28	98,48	97,92	98,18	**Por Faixa**
	10 a 20	97,18	97,53	98,10	97,08	97,47	
	5 a 20	95,74	95,70	97,04	95,37	95,96	
	0 a 20	91,67	90,26	93,73	91,51	91,79	
	-5 a 20	83,02	80,25	84,85	83,80	82,98	

5 Resultados do testb usando o WI008

Tabela D5a
Taxa de acerto para o testb usando F-E WI008 e treinamento sob condições limpas

WI008 FRONT-END		**Testb – Condições Limpas Taxa de Acerto (%) Por Tipo de Ruído**				**HTK BACK-END**	
Degradação		**Restaurant**	**Street**	**Airport**	**Train-station**	**Média (%)**	
Nível de SNR (dB)	**Clean**	99,08	99,00	99,05	99,23	99,09	**Por Nível**
	20	97,97	97,64	98,91	98,36	98,22	
	15	95,33	96,74	97,11	96,73	96,48	
	10	90,08	92,78	93,47	93,77	92,53	
	5	76,27	83,28	84,07	84,57	82,05	
	0	51,09	60,07	60,99	62,57	58,68	
	-5	18,67	29,87	28,54	29,96	26,76	
		Média (%)					
Faixa de SNR (dB)	**15 a 20**	96,65	97,19	98,01	97,55	97,35	**Por Faixa**
	10 a 20	94,46	95,72	96,50	96,29	95,74	
	5 a 20	89,91	92,61	93,39	93,36	92,32	
	0 a 20	82,15	86,10	86,91	87,20	85,59	
	-5 a 20	71,57	76,73	77,18	77,66	75,79	

Tabela D5b
Taxa de acerto para o testb usando F-E WI008 e treinamento sob condições múltiplas

WI008 FRONT-END		Testb – Condições Limpas Taxa de Acerto (%) Por Tipo de Ruído				HTK BACK-END	
Degradação		Restaurant	Street	Airport	Train-station	Média (%)	
Nível de SNR (dB)	Clean	99,02	98,79	98,93	99,14	98,97	Por Nível
	20	98,10	98,13	89,14	98,83	96,05	
	15	96,99	97,82	98,06	97,78	97,66	
	10	94,87	95,56	95,94	96,11	95,62	
	5	87,17	90,42	91,44	90,16	89,80	
	0	65,52	73,97	75,75	73,99	72,31	
	-5	28,58	38,88	40,89	41,90	37,56	
		Média (%)					
Faixa de SNR (dB)	15 a 20	97,55	97,98	93,60	98,31	96,86	Por Faixa
	10 a 20	96,65	97,17	94,38	97,57	96,44	
	5 a 20	94,28	95,48	93,65	95,72	94,78	
	0 a 20	88,53	91,18	90,07	91,37	90,29	
	-5 a 20	78,54	82,46	81,87	83,13	81,50	

6 Resultados do testc usando o WI008

Tabela D6a
Taxa de acerto para o testc usando F-E WI008 e treinamento sob condições limpas

WI008 FRONT-END		**Testc – Condições Limpas Taxa de Acerto (%) Por Tipo de Ruído**		**HTK BACK-END**	
Degradação		**Subway (MIRS)**	**Street (MIRS)**	**Média (%)**	
Nível de SNR (dB)	**Clean**	99,08	99,03	99,06	**Por Nível**
	20	97,36	97,70	97,53	
	15	95,33	95,77	95,55	
	10	90,24	90,69	90,47	
	5	79,03	78,17	78,60	
	0	51,73	52,09	51,91	
	-5	24,62	25,57	25,10	
		Média (%)			
Faixa de SNR (dB)	**15 a 20**	96,35	96,74	96,54	**Por Faixa**
	10 a 20	94,31	94,72	94,52	
	5 a 20	90,49	90,58	90,54	
	0 a 20	82,74	82,88	82,81	
	-5 a 20	73,05	73,33	73,19	

Tabela D6b
Taxa de acerto para o testc usando F-E WI008 e treinamento sob condições múltiplas

WI008 FRONT-END		**Testc – Condições Limpas Taxa de Acerto (%) Por Tipo de Ruído**		**HTK BACK-END**	
Degradação		**Subway (MIRS)**	**Street (MIRS)**	**Média (%)**	
Nível de SNR (dB)	**Clean**	98,99	98,82	98,91	**Por Nível**
	20	98,10	97,97	98,04	
	15	97,54	97,70	97,62	
	10	95,58	95,22	95,40	
	5	88,85	87,45	88,15	
	0	67,06	65,66	66,36	
	-5	30,49	30,83	30,66	
		Média (%)			
Faixa de SNR (dB)	**15 a 20**	97,82	97,84	97,83	**Por Faixa**
	10 a 20	97,07	96,96	97,02	
	5 a 20	95,02	94,59	94,80	
	0 a 20	89,43	88,80	89,11	
	-5 a 20	79,60	79,14	79,37	

Apêndice E

Médias das taxas de acerto de todos os testes

1 Médias das taxas de acerto por teste

Com base nos dados das tabelas de todos os testes pode-se calcular, em média, o quanto a taxa de acerto é afetada por cada tipo de teste, nos níveis e faixas de SNR.

1.1 Médias por teste para treinamento em condições limpas

As Tabelas E1a e E1b mostram, respectivamente, a média da taxa de acerto de todos os testes com o uso do WI007 e WI008, sob treinamento em condições limpas.

Tabela E1a
Média da taxa de acerto (%) de todos os testes usando F-E WI007 e treinamento em condições limpas

	WI007	**Testa**	**Testb**	**Testc**	**Testd**
		Média % Por Nível da Degradação			
Nível de SNR (dB)	**clean**	99,02	99,02	99,06	99,02
	20	95,25	92,77	94,31	93,80
	15	87,33	81,34	87,84	85,96
	10	67,71	59,01	74,15	66,18
	5	39,48	31,93	50,24	39,59
	0	16,95	13,70	24,17	18,79
	-5	7,93	7,65	11,49	9,63
		Média % Por Redução da Degradação			
Faixa de SNR (dB)	**15 a 20**	91,29	87,06	91,08	89,88
	10 a 20	83,43	77,71	85,44	81,98
	5 a 20	72,44	66,26	76,64	71,38
	0 a 20	61,34	55,75	66,14	60,86
	-5 a 20	52,44	47,74	57,03	52,32

Tabela E1b
Média da taxa de acerto (%) de todos os testes usando F-E WI008
e treinamento em condições limpas

	WI008	Testa	Testb	Testc	Testd
		Média % Por Nível da Degradação			
Nível de SNR (dB)	clean	99,09	99,09	99,06	99,09
	20	98,15	98,22	97,53	95,44
	15	96,93	96,48	95,55	94,60
	10	93,16	92,53	90,47	92,61
	5	84,36	82,05	78,60	86,88
	0	60,88	58,68	51,91	71,27
	-5	28,57	26,76	25,10	41,14
		Média % Por Redução da Degradação			
Faixa de SNR (dB)	15 a 20	97,54	97,35	96,55	95,03
	10 a 20	96,08	95,74	94,52	94,22
	5 a 20	93,15	92,32	90,54	92,38
	0 a 20	86,70	85,59	82,81	88,16
	-5 a 20	77,01	75,79	73,19	80,32

1.2 Médias por teste para treinamento em condições múltiplas

As Tabelas E1c e E1d apresentam, respectivamente, a média da taxa de acerto de todos os testes com o uso do WI007 e WI008, sob treinamento em condições múltiplas.

Tabela E1c
Média da taxa de acerto (%) de todos os testes usando F-E WI007 e treinamento em condições múltiplas

	WI007	Testa	Testb	Testc	Testd
		Média % Por Nível da Degradação			
Nível de SNR (dB)	clean	98,52	98,52	98,54	98,52
	20	97,70	97,22	96,93	95,84
	15	96,95	95,82	95,94	95,67
	10	94,89	93,12	92,92	94,14
	5	87,83	84,73	82,47	88,06
	0	61,71	60,47	50,63	70,71
	-5	24,56	25,90	21,58	39,96
		Média % Por Redução da Degradação			
Faixa de SNR (dB)	15 a 20	97,32	96,52	96,44	95,76
	10 a 20	96,51	95,39	95,26	95,22
	5 a 20	94,34	92,72	92,07	93,43
	0 a 20	87,82	86,27	83,78	88,88
	-5 a 20	77,27	76,21	73,41	80,73

Tabela E1d
Média da taxa de acerto (%) de todos os testes usando F-E WI008 e treinamento em condições múltiplas

	WI008	Testa	Testb	Testc	Testd
		Média % Por Nível da Degradação			
Nível de SNR (dB)	clean	98,97	98,97	98,91	98,97
	20	98,50	96,05	98,04	96,43
	15	97,86	97,66	97,62	96,29
	10	96,06	95,62	95,40	95,48
	5	91,43	89,80	88,15	92,12
	0	75,13	72,31	66,36	82,69
	-5	38,92	37,56	30,66	56,87
		Média % Por Redução da Degradação			
Faixa de SNR (dB)	15 a 20	98,18	96,86	97,83	96,36
	10 a 20	97,47	96,44	97,02	96,07
	5 a 20	95,96	94,78	94,81	95,08
	0 a 20	91,79	90,29	89,12	92,60
	-5 a 20	82,99	81,50	79,37	86,65

Apêndice F

Curvas do testa e ruído metal-cutting em condições múltiplas

Na Figura F1 encontra-se as curvas comparativas dos ruídos do *testa* e do ruído *metal-cutting* para o WI007 e para o WI008 sob condições múltiplas**.** A resposta do WI007 está identificada com a letra "a)", enquanto a resposta do sistema que usa o WI008 está identificada com a letra "b)".

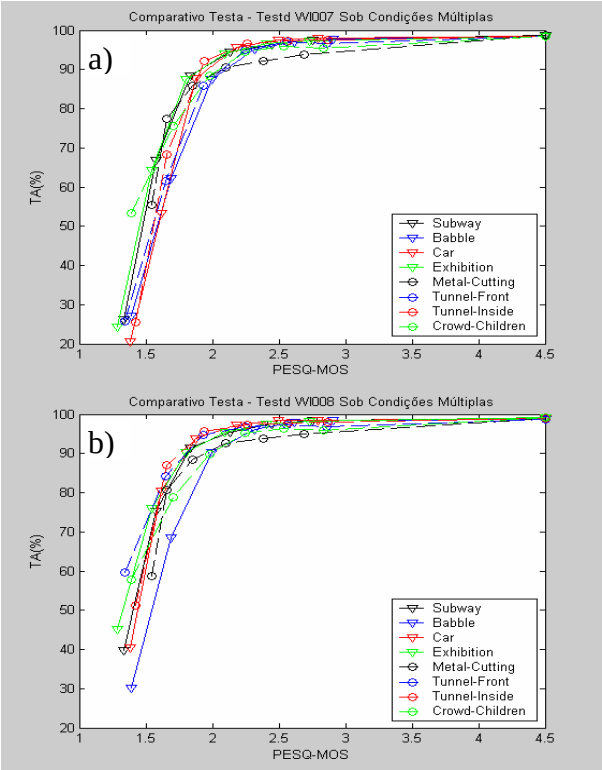


Figura F1: Comparativo testa-testd usando WI007 (a) e WI008 (b) sob condições múltiplas

Nas Figuras F2 a F6, são apresentadas, respectivamente, as curvas de ajuste do ruído *metal-cutting*, *subway*, *babble*, *car* e *exhibition* para o WI008 sob condições múltiplas.

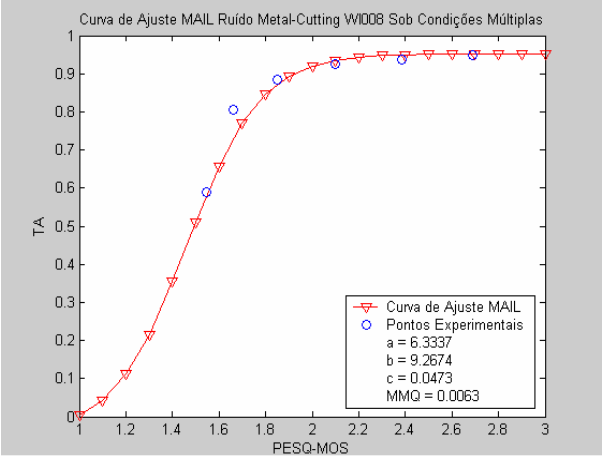


Figura F2: Curva de ajuste do ruído *metal-cutting* usando WI008 sob condições múltiplas

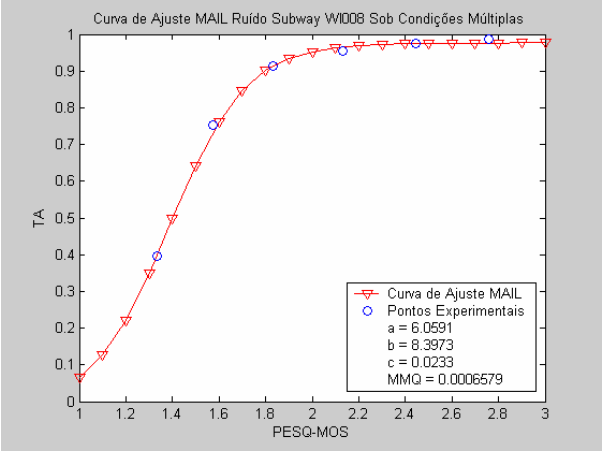


Figura F3: Curva de ajuste do ruído *subway* usando WI008 sob condições múltiplas

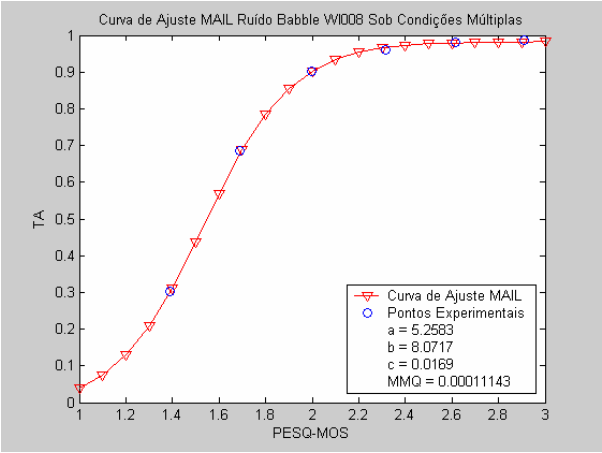


Figura F4: Curva de ajuste do ruído *babble* usando WI008 sob condições múltiplas

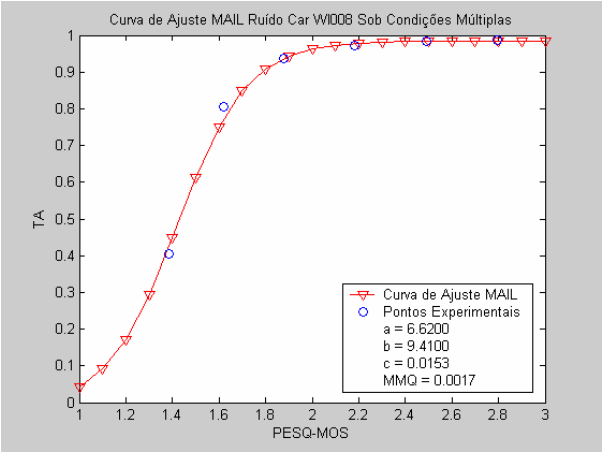


Figura F5: Curva de ajuste do ruído *car* usando WI008 sob condições múltiplas

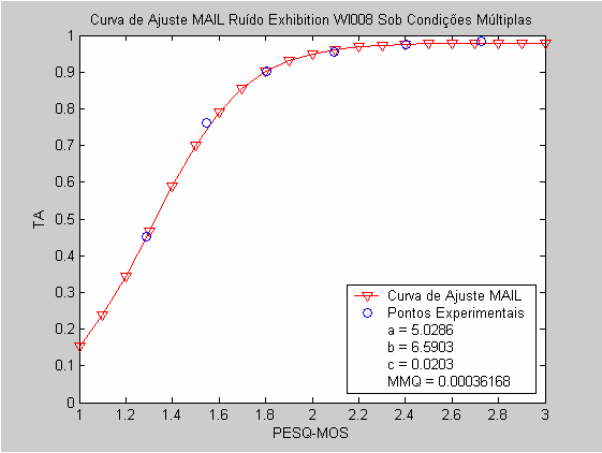


Figura F6: Curva de ajuste do ruído *exhibition* usando WI008 sob condições múltiplas

Apêndice G

Análise espectral dos ruídos

Neste apêndice é apresentada uma breve discussão sobre a análise espectral dos sinais dos ruídos da base Aurora-1 e dos ruídos novos.

Observando as curvas espectrais de cada ruído pode-se perceber que alguns apresentam comportamentos semelhantes antes mesmo de sofrerem qualquer tratamento.

Este aspecto induz a fazer um comparativo das características espectrais de cada ruído dos testes da base Aurora-1 com os ruídos novos.

As Figuras G1 a G8 apresentam o espectro de frequência dos ruídos da base Aurora *subway*, *babble*, *car*, *exhibition*, *airport*, *street*, *restaurant* e *train-station*, respectivamente.

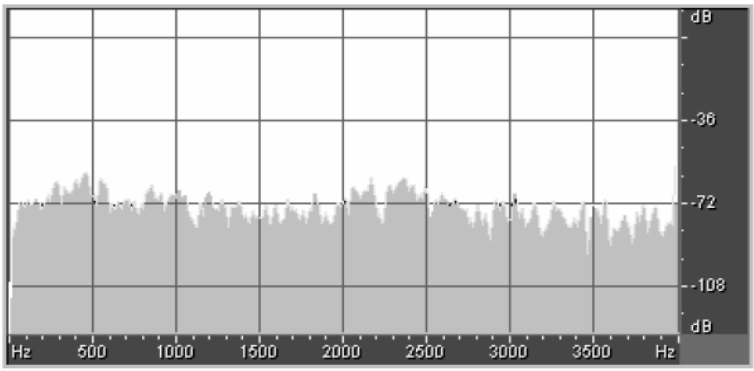


Figura G1: Espectro do ruído subway

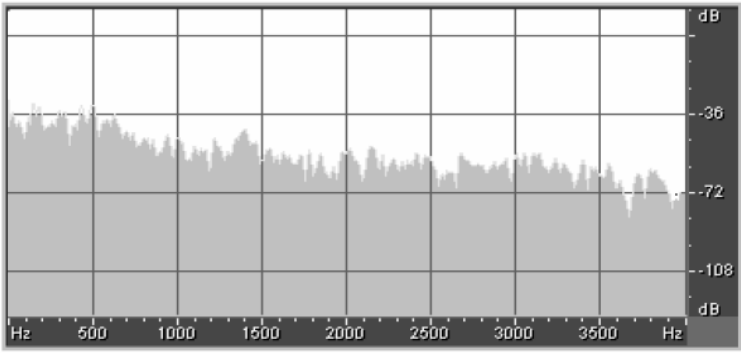


Figura G2: Espectro do ruído babble

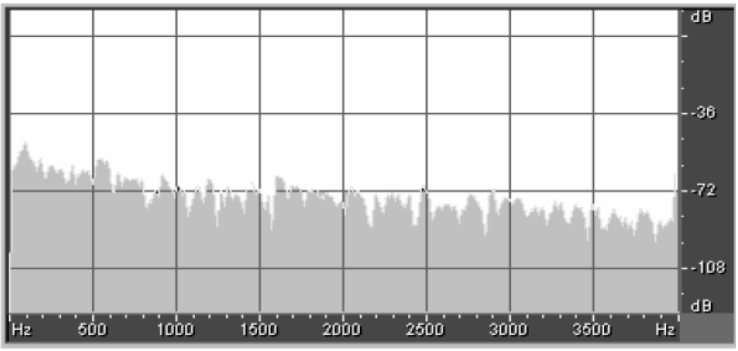


Figura G3: Espectro do ruído car

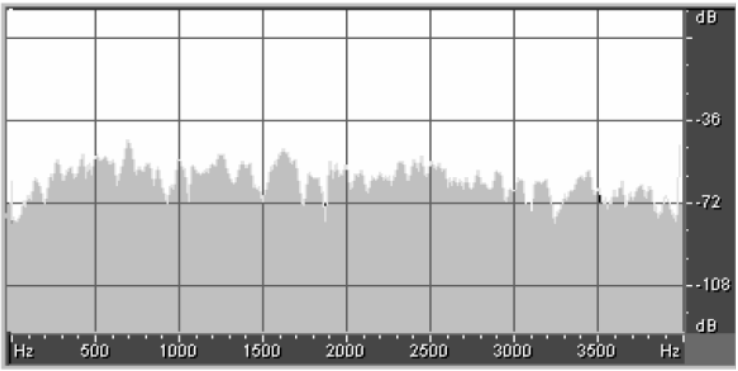


Figura G4: Espectro do ruído exhibition

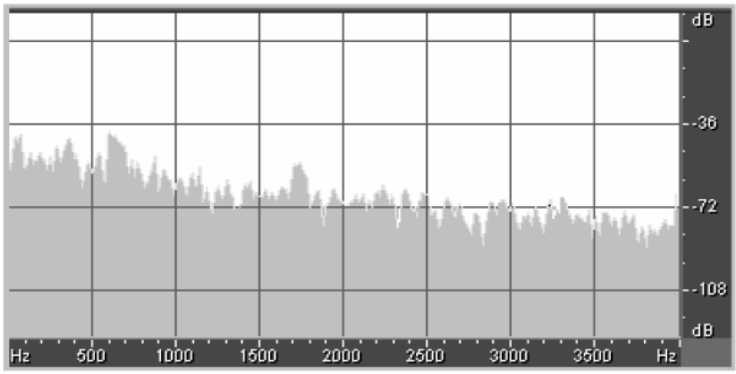


Figura G5: Espectro do ruído airport

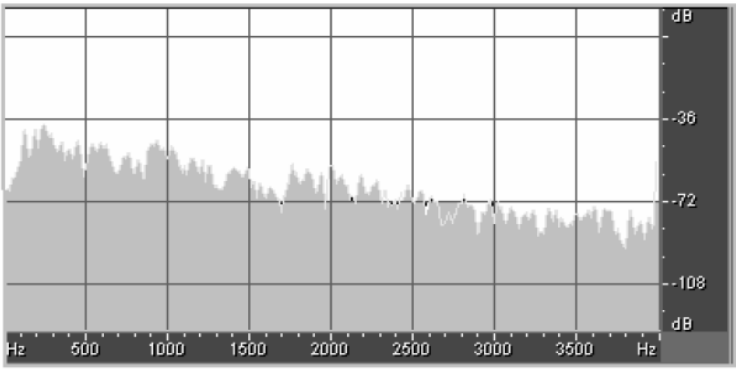


Figura G6: Espectro do ruído street

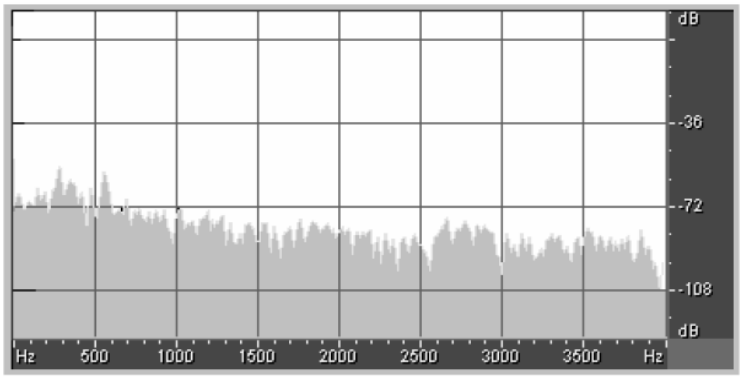


Figura G7: Espectro do ruído restaurant

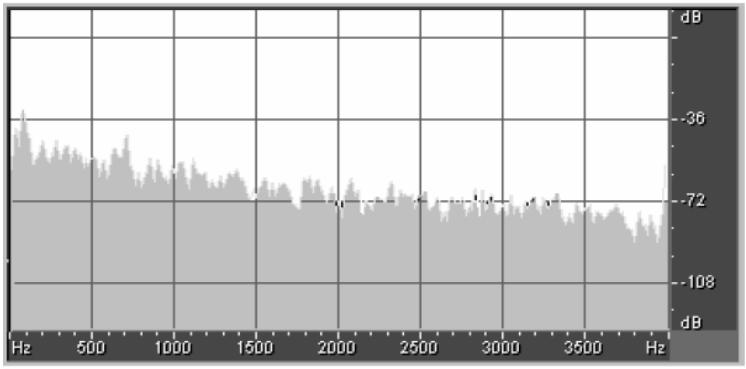


Figura G8: Espectro do ruído train-station

As Figuras G9 a G12 mostram o espectro de frequência dos ruídos novos: *metal-cutting*, *tunnel-front*, *tunnel-inside* e *crowd-children*, respectivamente.

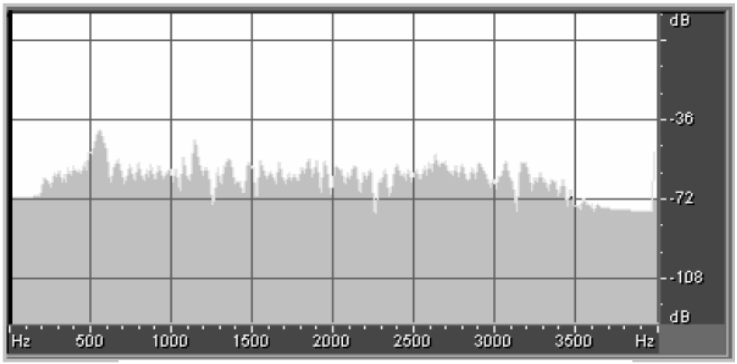


Figura G9: Espectro do ruído metal-cutting

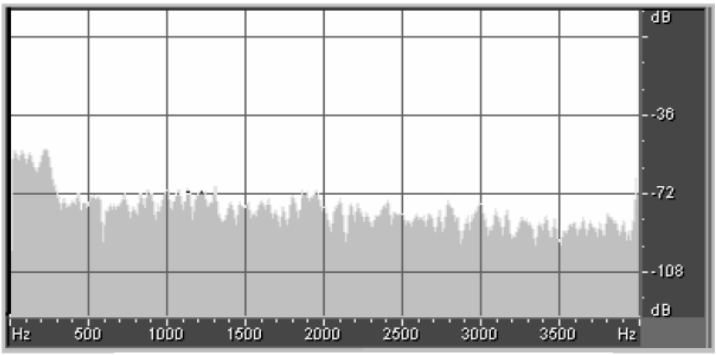


Figura G10: Espectro do ruído tunnel-front

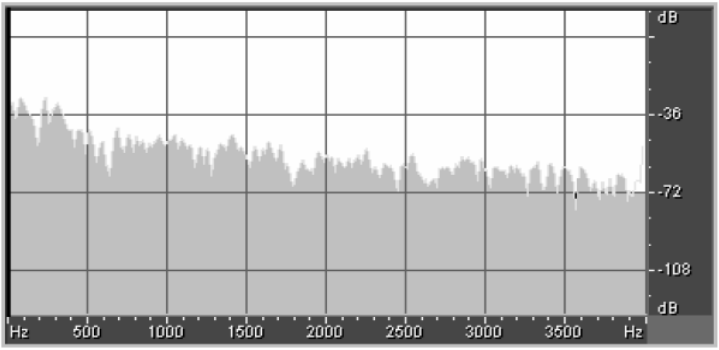


Figura G11: Espectro do ruído tunnel-inside

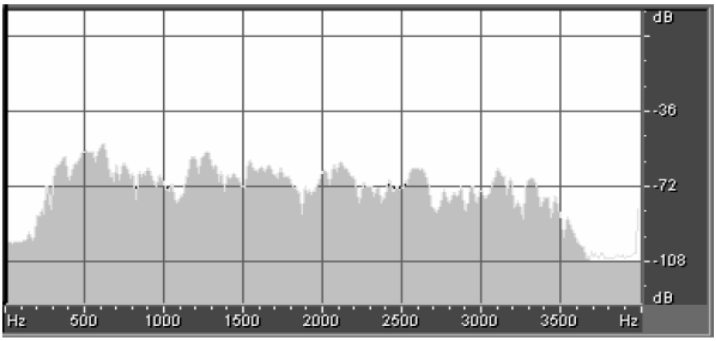


Figura G12: Espectro do ruído crowd-children

Observando os espectros dos ruídos conclui-se que a maior parte da energia dos sinais concentra-se na região de baixa frequência. Além disso, do ponto de vista espectral, antes de qualquer tratamento, alguns sinais de ruído aparentam ser bem similares mesmo que sejam gravados em ambientes totalmente diferentes. Sabe-se, entretanto, que a partir da passagem desses sinais de ruídos pelos filtros, o espectro de frequência muda sensivelmente, e quando são passados pela banda de frequência de telefonia digital, todos os sinais se concentram nessa banda.