UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO TECNOLÓGICO
CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE

MELHORAMENTO DO SINAL DE VOZ POR INIBIÇÃO LATERAL E MASCARAMENTO BINAURAL

EDIL JAMES DE JESUS NASCIMENTO

SÃO LUIS
2004

MELHORAMENTO DO SINAL DE VOZ POR INIBIÇÃO LATERAL E MASCARAMENTO BINAURAL

Por

EDIL JAMES DE JESUS NASCIMENTO

Orientador: Dr. Allan Kardec Dualibe Barros Filho

Dissertação de Mestrado submetida à Coordenação do curso de Pós-Graduação em Engenharia de Eletricidade da UFMA como parte dos requisitos para obtenção do título de mestre em Engenharia Elétrica.

Abril, 2004

MELHORAMENTO DO SINAL DE VOZ POR INIBIÇÃO LATERAL E MASCARAMENTO BINAURAL

MESTRADO

Área de Concentração: CIÊNCIAS DA COMPUTAÇÃO

EDIL JAMES DE JESUS NASCIMENTO

Orientador: Dr. Allan Kardec Dualibe Barros Filho

Curso de Pós-Graduação
Em Engenharia de Eletricidade da
Universidade Federal do Maranhão

MELHORAMENTO DO SINAL DE VOZ POR INIBIÇÃO LATERAL E MASCARAMENTO BINAURAL

EDIL JAMES DE JESUS NASCIMENTO

Dissertação aprovada em 02 de abril de 2004

Prof. Dr Allan Kardec Dualibe Barros Filho - UFMA/MA
(Orientador)

Prof. Dr Henio Henrique Aragão Rêgo - CEFET/MA
(Membro da Banca Examinadora)

Prof. Dr Zair Abdelouahab - UFMA/MA

DEDICATÓRIA

Dedico este trabalho às pessoas que me apoiaram e incentivaram:

- A minha mãe Maria Paula, que sempre me acompanhou com suas orações, palavras de ânimo e conselhos precisos.
- A minha querida esposa Ana Neri, cujo carinho e apoio foram à mola mestre para a conclusão deste trabalho.
- A meu irmão Edil Jarles, que mesmo estando longe não me deixou lutar sozinho nas dificuldades.
- A minha irmã Edilma (*in memorium*), por seu exemplo de vida.

AGRADECIMENTOS

- A Deus que é o autor e consumador de minha fé, sem o qual nada é possível.

- Ao professor Allan Kardec pela paciência, compreensão e apoio nos momentos mais difíceis.

- A minha mãe Maria Paula e meu irmão, Edil Jarles, pelas palavras encorajadoras.

- A minha esposa Ana Neri, pelo amor, dedicação, compreensão e pelas noites mal-dormidas me acompanhando durante a preparação deste trabalho.

- Aos Doutores e Amigos do INPE, em especial ao Dr. Eurico Rodrigues de Paula, ao Dr. Mangalathail Ali Abdu e ao Engenheiro Acácio Cunha Neto, pela compreensão e incentivo.

- Aos amigos do PIB, em especial a Fausto, Ricardo, Natália, Paulo Henrique, Maxwell e Márcio pelo companheirismo.

- Aos Amigos Geovane, Paulino, Ricardo Almeida, Valeska e Rejane por sua presença nos momentos mais importantes durante a concretização deste trabalho.

RESUMO

O sistema auditivo humano é capaz de realizar diferentes tarefas que seriam úteis em aplicações de engenharia. Uma delas é a habilidade de separar fontes sonoras, permitindo a um ouvinte “focar” uma única fonte sonora em um ambiente ruidoso. Grandes investimentos têm sido feitos no desenvolvimento de tecnologias aplicadas ao reconhecimento de voz, por meio de máquinas, em ambientes reais. Para isso, diferentes técnicas de processamento computacional têm sido propostas para a redução do ruído ambiente e melhoramento do sinal desejado em ambiente acústico complexo (cocktail party). Essas técnicas são motivadas pelo modelo do sistema auditivo humano em suas diferentes fases.

Neste trabalho, desenvolvemos um algoritmo para melhorar o processamento de um sinal de fala baseado no modelo auditivo binaural. Após receber os sinais misturados, por dois microfones, o algoritmo aumenta a inteligibilidade do sinal de maior energia de um dos receptores. Utilizando dois oradores e considerando que cada um está mais próximo de um dos receptores, fizemos uso dos conceitos de inibição lateral e mascaramento binaural, para recuperar o sinal de fala de maior energia de um dos receptores.

O algoritmo foi desenvolvido sob a plataforma matlab e comparado com um outro sem a utilização da inibição lateral na recuperação do sinal desejado. Os resultados, avaliados através do cálculo do erro relativo e da escala MOS, mostraram que a utilização da inibição lateral na recuperação do sinal, melhora o erro relativo entre o sinal desejado e o sinal recuperado e conseqüentemente a qualidade do sinal recuperado.

ABSTRACT

The human hearing system is capable to accomplish different tasks that would be useful in engineering applications. One of them is the ability to separate sound sources, allowing the listener to "focus" a single sound source in a noisy environment. Great investments have been made in the development of technologies applied to the voice recognition by machines in real environment. For that, different techniques of processing computational have been proposed, for reduction of the ambient noise and improvement of the signal desired in complex acoustic environment (cocktail party). The model of the human hearing system motivates those techniques in their different phases.

In this work, we developed an algorithm to improve the processing speech signal based on the binaural hearing model. After receiving the mixed signals, for two microphones, the algorithm increases the intelligibility of the signal of larger energy of one of the receivers. Using two speakers and considering that each one is closer of one of the microphones, we made use of the concepts of lateral inhibition and binaural masking, to recover the signal of speech of larger energy of one of the receivers.

The algorithm was developed in platform matlab and it was compared with another without use the lateral inhibition in the recovery of the desired signal. The results, appraised through the calculation of the relative error and of the scale MOS, showed that the use of the lateral inhibition in the recovery of the signal, improves the relative error between the desired signal and the recovered signal and consequently the quality of the recovered signal.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

SA – Sistema Auditivo

ICA - Independent Component Analysis

MATLAB – Matrix Laboratory

MOS - Mean Opinion Score

MB - Membrana Basilar

SAFIA - sound source Segregation based on estimating incident Angle of each Frequency component of Input signals Acquired by multiple microphones

FTD - Fourier Transform Discrete

ITD - Interaural Time Difference

IID - Interaural Intensity Difference

ASA - Auditory Scene Analysis

dB - Decibel

IL – Inibição Lateral

CAPÍTULO 1

INTRODUÇÃO

Os sons se originam do movimento ou vibração de um objeto. Esses movimentos causam variação na pressão do ar e dão origens as ondas sonoras. Essas ondas viajam através do ar e são caracterizadas como um estímulo transmitido ao ouvido.

Quando desejamos ouvir um som específico em um ambiente qualquer, o som que escutamos é formado pela mistura e reverberação do sinal desejado com outros sinais acústicos (interferências). Estes sinais neste caso são indesejáveis, principalmente quando diminuem a qualidade e a inteligibilidade do sinal que desejamos ouvir. O sistema auditivo (SA) trata essa questão de interferência, de forma automática, distinguindo cada som e permitindo-nos atentar para o sinal de maior interesse. Essa capacidade humana de atentar para um sinal, mesmo estando em ambiente acústico desfavorável, tem levado os pesquisadores ao estudo da psicoacústica[1].

Além do interesse puramente científico, de como o SA humano é capaz de detectar e separar fontes sonoras, diferentes aplicações têm se beneficiado com o estudo do SA, especialmente as telecomunicações, onde as atuais tecnologias de telefonia celular nos permitem acessar funcionalidades das operadoras por comando de voz. O problema é que o sinal transmitido através de um aparelho celular não consiste apenas do sinal de voz desejado, mas também contém ruídos e reverberações. Para a remoção dessas interferências diversos algoritmos têm sido desenvolvidos e o conhecimento do sistema binaural tem melhorado a qualidade desses algoritmos.

Durante as últimas décadas, a análise da cena auditiva tem sido estuda através de um ambiente denominado *cocktail party*: em uma sala existem muitas fontes sonoras misturadas e

[1] Estudo da percepção dos sons pelo ser humano, que visa buscar soluções para os problemas mais diversos relacionados à audição humana.

reverberadas: vozes, música, ruído de ar-condicionado, etc. A tarefa é separar uma ou mais fontes sonoras e melhorar sua inteligibilidade, tendo conhecimento apenas da mistura dos sinais.

Diversos autores têm proposto soluções para este problema. Alguns utilizam um único canal de entrada (sistema monaural), enquanto outros utilizam dois ou mais canais de entrada [1]. Algumas destas soluções envolvem o uso de características psicoacústicas do SA [2,3] e da harmonicidade da fala humana, através da extração da freqüência fundamental [4,5], por algoritmo tipo subtrativo [6], ou através da análise de componentes independentes (ICA)[3,7].

Neste trabalho, propomos um algoritmo simples que imita o SA na extração do sinal de maior energia de um dos receptores, a partir de uma mistura de dois sinais de fala. Para tanto, fizemos uso de um sistema binaural (dois canais de entrada), e de duas características psicoacústicas (inibição lateral e mascaramento binaural) do SA. Embora tenhamos utilizado somente misturas com dois sinais de fala, o algoritmo pode ser utilizado com misturas de vários sinais, uma vez que a inibição lateral enfatiza somente o sinal de maior energia em cada receptor. O algoritmo foi desenvolvido em plataforma MATLAB e os resultados experimentais foram avaliados de duas maneiras: através de medida objetiva - cálculo do erro relativo e através de medida subjetiva – onde dez ouvintes, utilizando a escala MOS (Mean Opinion Score), qualificaram o sinal.

Como estamos trabalhando com análise de cena auditiva não podemos deixar de mencionar o *cocktail party* que talvez seja o problema mais conhecido no estudo do sistema auditivo.

1.1 COCKTAIL PARTY

O *cocktail party* pode ser entendido como um ambiente que enfatiza o estudo da capacidade seletiva do ser humano em se fixar em um único sinal sonoro, mesmo estando em um local onde vários outros sinais estejam ocorrendo simultaneamente. Este estudo dá ênfase a um vasto e complexo quebra-cabeça, que tem sido montado com a intenção de explicar a interação entre o sinal sonoro, o sistema auditivo e o córtex cerebral.

Existem pelo menos dois problemas que envolvem o estudo do cocktail party.

- Primeiro, as misturas que chegam ao ouvido são versões reverberadas das fontes originais.

- Segundo, a resposta ao impulso de um ambiente muda de acordo com parâmetros tais como a disposição dos móveis, material que formam as paredes, temperatura, etc.

Não é conhecido como os sinais reverberados e as diferentes respostas ao impulso dos ambientes são combinados no cérebro, por isso diferentes pesquisas têm sido realizadas com o objetivo de apresentar uma solução satisfatória para o problema do cocktail party.

1.2 ESTADO DA ARTE

Quando olhamos para o sistema auditivo na sua funcionalidade, podemos identificá-lo como uma máquina de separar sons [1]. Esta máquina pode ser analisada sob duas perspectivas: monaural – utiliza somente um canal de entrada e binaural – utiliza dois canais de aquisição. Ao analisar essas perspectivas, os pesquisadores fazem uso da harmonicidade característica da fala humana, através da extração da freqüência fundamental por algoritmo tipo subtrativo [6] ou através da análise de componentes independentes [3]

Com o desejo de chegar aos mesmos resultados obtidos pelo aparelho auditivo (separação e identificação de uma fonte sonora), foram realizados estudos do funcionamento da cóclea que é responsável, em grande parte, pela nossa capacidade de identificar e separar sons.Os resultados destes estudos revelaram que essa capacidade está relacionada a nossa habilidade em identificar a freqüência fundamental de cada sinal. Fletcher [8], ao realizar seus experimentos, chegou à conclusão que o sistema auditivo periférico comportava-se como um banco de filtro passa-faixa com sobreposição de bandas, onde a membrana basilar (MB) fornece a base para este filtro. Cada parte da MB responde a uma faixa limitada de freqüência (filtro), onde o ponto de excitação máxima corresponde à freqüência central deste filtro.

Um método muito utilizado na análise monaural é o de algoritmo tipo subtrativo que se constitui um mecanismo tradicional para remoção de ruídos estacionários em sistemas de um único canal de entrada. Embora seja um mecanismo simples e computacionalmente eficiente, ele introduz no sinal melhorado um ruído residual. Para reduzir os efeitos desse ruído, vários trabalhos foram propostos. Alguns trabalhos incorporaram o modelo auditivo, explorando a característica do limiar de audibilidade [9,10]. Em seu trabalho, Virag [6] incorporou o modelo psicoacústico ao algoritmo de subtração espectral para reduzir o efeito do ruído residual. Ela realizou vários testes com diferentes tipos de ruídos e seus resultados diminuíram o ruído residual e mantiveram a distorção da fala em um nível aceitável.

No processamento binaural, os diferentes métodos são em geral bastante similares. Primeiramente eles fazem um processamento periférico para cada sinal de entrada (divisão em sub-bandas), para em seguida ocorrer à interação binaural entre os sinais dos dois receptores. O sistema de audição binaural facilita nossa capacidade de localizar, separar e identificar fontes sonoras devido às diferenças interaurais existentes entre os sons que chegam aos ouvidos.

Nakatani *et al.* [11] desenvolveram um sistema binaural de separação de feixes de fala, baseado na estrutura harmônica dos sinais. Primeiro eles extraíram os fragmentos harmônicos e seguida agruparam os fragmentos de acordo com a direção da fonte e a continuidade da freqüência fundamental. A direção foi obtida através do cálculo da diferença de fase (IPD) e da diferença de intensidade interaural (IID) entre as harmônicas correspondentes dos dois canais de entrada.

Em [3], Barros *e et al*, utilizaram Análise de Componentes Independentes, associado ao mascaramento auditivo e a filtros adaptativos para melhorar o sinal de fala embutido em uma ambiente ruidoso. Seu algoritmo baseou-se no conceito de harmonicidade da voz e a partir da determinação da freqüência fundamental (f_0) do sinal de entrada, eles usaram um filtro adaptativo centrado em f_0 e suas harmônicas. ICA foi usado para separar os sinais que formavam a mistura. Um módulo de decisão selecionou os componentes de maior energia do receptor. Em [7] Barros *e et al*, utilizaram o mesmo principio de [3] para obter sinais de um ambiente com reverberação. Neste artigo os autores realizaram dois tipos de simulações para

obter o sinal de maior energia dos canais de entrada. Uma foi a simulação computacional da mistura e a outra foi a aquisição de sinal em ambiente real. Os resultados foram avaliados através da escala MOS.

Aoki, *e et al* [2] desenvolveram um método binaural para separar um sinal de fala desejado de sons concorrentes, baseado na amplitude e fase do sinal de entrada. O método recebeu o nome de SAFIA (Sound Source Segregation Based on Estimating Incident Angle of each Frequency Component of Input Signals Acquired by Multiple Microphones). Nele os sinais recebidos por dois microfones foram transformados do domínio do tempo para o domínio da freqüência pela Transformada Discreta de Fourier (FDT). Para cada componente de freqüência foi calculada a diferença de amplitude e fase entre os receptores. Essa diferença foi utilizada para determinar qual o componente de freqüência vem da direção desejada e para reconstruir esses componentes como o sinal da fonte original. O resultados obtidos com o SAFIA mostraram que a melhor freqüência para diminuir a sobreposição dos sinais de entrada foi de 10 Hz.

Quando comparamos os dois métodos –monaural e binaural, verificamos que o monaural é restrito em aplicações de ambiente real e que o método binaural é mais promissor uma vez que disponibiliza mais informação sobre o sinal desejado e sobre o ruído, através das amostragens coletadas simultaneamente em diferentes pontos do ambiente.

1.3 DESCRIÇÃO DO PROBLEMA

Imagine que você está em uma sala onde duas pessoas estão falando simultaneamente. Você dispõe dois microfones colocados em locais diferentes da sala. Os microfones dão a você duas gravações de sinais no tempo, que denotamos por $x_1(t)$ e $x_2(t)$, sendo x_1, e x_2 as amplitudes e t um tempo determinado. Cada um dos sinais gravados é a somatória dos sinais de fala emitidos pelas duas pessoas, as quais denominamos de S_1 e S_2 Podemos expressar esta situação sob a forma linear:

$$x_1 = a_{11} s_1 + a_{12} s_2$$
$$x_2 = a_{21} s_1 + a_{22} s_1 \qquad (1)$$

onde a_{ij} são parâmetros que dependem do ambiente e da distância dos microfones aos oradores. A idéias é estimar S_1 e/ou S_2 usando somente os sinais gravados $x_1(t)$ e $x_2(t)$. A equação (1) pode ser expressa em forma de matriz:

$$X=A*S \qquad (2)$$

onde X corresponde as misturas, A representa a matriz de mistura, o $*$ denota a convolução dos sinais de entrada e S representa os sinais originais.

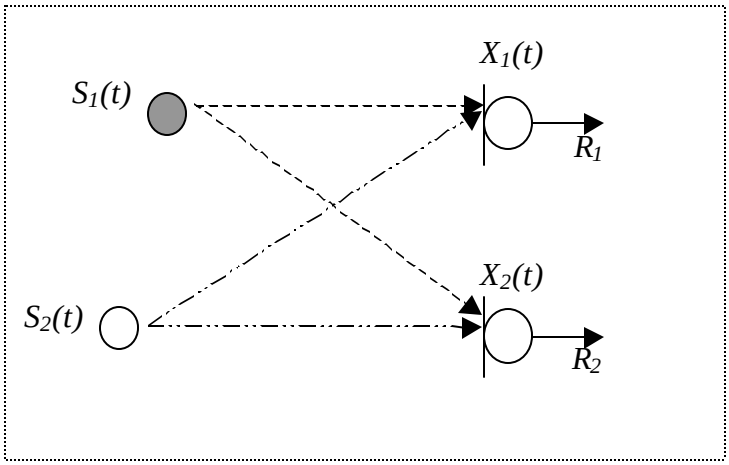


Figura 1.1 – Aquisição dos sinais pelos receptores

Observando a figura 1.1, podemos considerar que: o indivíduo S_1 está mais próximo do R_1 e o orador S_2 está mais próximo do R_2. Neste caso, o sinal adquirido por R_1 terá maior nível de energia da voz de S_1 do que de S_2, já com o R_2 ocorrerá o inverso. O objetivo é recuperar o sinal de maior nível de energia de um dos receptores. Nós consideramos que os sinais originais estão separados no domínio da freqüência, isto é, em uma certa banda de fr

freqüência do sinal misturado, existe somente uma fonte de sinal. A idéia é que se duas pessoas são diferentes, a freqüência fundamental de suas vozes também são.

1.4 ORGANIZAÇÃO DA TESE

Esta tese está dividida em cinco capítulos. No primeiro capítulo fizemos uma rápida abordagem do problema e apresentamos nosso objetivo: separação do sinal de maior energia de um receptor, a partir de uma mistura de sinais de fala. No capítulo dois, fazemos uma rápida abordagem sobre o sistema auditivo, uma vez que a solução proposta está intimamente ligada as características deste sistema. No terceiro capítulo descrevemos o algoritmo desenvolvido em suas quatro partes. No capítulo quatro apresentamos os resultados para as diversas simulações realizadas e no quinto capítulo fazemos a conclusão do trabalho desenvolvido.

CAPÍTULO 2

AUDIÇÃO HUMANA, MASCARAMENTO AUDITIVO, INIBIÇÃO LATERAL

2.1 SISTEMA AUDITIVO

A audição humana é formada por dois conjuntos idênticos e independentes (ouvidos), situados na lateral da cabeça, capazes de perceber variações de pressão do ar de 20Hz a 20KHz, sendo mais sensíveis às freqüências de 1kHz a 4kHz. A informação recebida por cada um dos ouvidos ajuda a separar e identificar uma fonte sonora através da diferença de tempo interaural (ITD) e da diferença de intensidade interaural (IID). A diferença binaural de tempo e de intensidade é de milésimos de segundos, mas o cérebro detecta essa diferença e descobre a direção do som. Os níveis perceptíveis de um som são medidos em decibéis, a tabela 2.1 mostra alguns níveis que variam de uma conversação normal até a sensação dolorosa causada por um sinal sonoro.

Tabela 2.1 – relação dB versus efeitos auditivos

DECIBÉIS	EMISSÕES E EFEITOS
50	Conversação normal
80	Ruído de tráfego intenso
100	Ruído de trens do metrô
120	Sensação de desconforto
140	Sensação dolorosa

A função do sistema auditivo é captar o estímulo externo (sinal acústico) e transmiti-lo ao cérebro, onde será feita a construção da imagem correspondente. O processo desde recebimento do sinal acústico até sua identificação no cérebro é feito através de uma seqüência de transformações de energia, iniciando pela sonora (som), passando pela mecânica

(vibração do tímpano e ossículos), hidráulica (movimento do líquido coclear) e finalizando com a energia elétrica dos impulsos nervosos que chegam ao cérebro.

2.2 ESTRUTURA DO SISTEMA AUDITIVO

A estrutura do ouvido, tanto pela sua complexidade, como pelo resultado operacional é um sistema perfeito, onde a energia acústica (ondas sonoras) é convertida em impulsos elétricos que são transmitidos, através do nervo auditivo, para o cérebro. O ouvido é dividido em três partes: externo, médio, e interno, como mostra a figura 2.1.

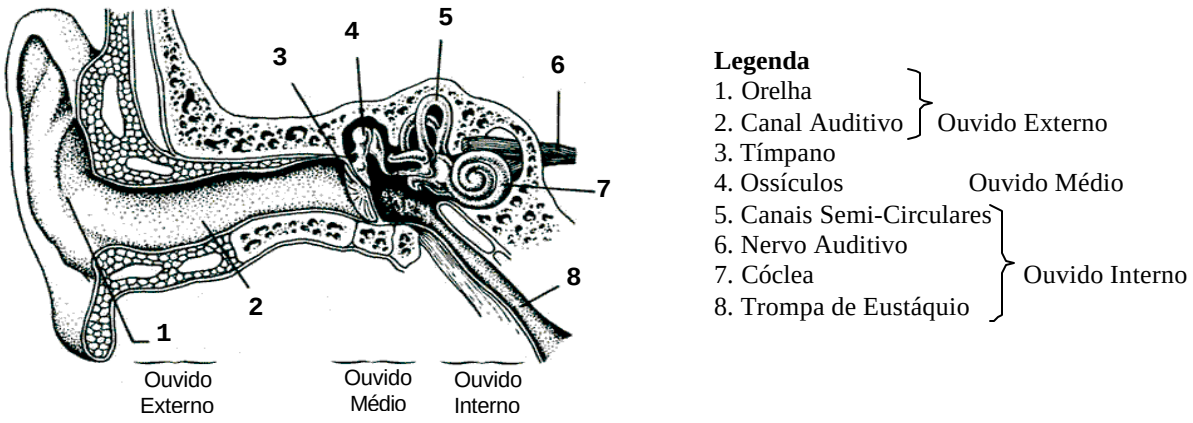


Figura 2.1 – Estrutura do ouvido

2.2.1 O OUVIDO EXTERNO

O ouvido externo é composto pela orelha (a parte visível do sistema auditivo), e pelo canal auditivo. A orelha captura os sons e ajuda na localização sonora, sendo mais sensível a sons que vêm da frente do ouvinte [1,8]. O canal auditivo é um tubo que tem uma de suas cavidades aberta para o meio externo e a outra fechada pela membrana timpânica, servindo como um guia do sinal até a o ouvido médio.

A função do ouvido externo é capturar o som ambiente e encaminhá-lo até o tímpano, uma membrana fina e elástica que faz a separação entre o ouvido externo e o ouvido médio, cujas vibrações, transformação da energia sonoras em mecânica, são equivalentes as freqüências do sinal recebido.

2.2.2 O OUVIDO MÉDIO

O ouvido médio é formado por um conjunto de três ossículos móveis, que transmitem a energia aplicada no tímpano, para a cóclea. Estes ossículos, denominados de martelo, bigorna e estribo, agem como um amplificador do sinal recebido transmitindo as vibrações do tímpano para a janela oval à entrada da cóclea. As funções do ouvido médio podem ser caracterizadas como:

1) transmitir os movimentos do tímpano - realizado através dos ossículos;

2) proteger o sistema auditivo dos efeitos danosos de sons muito altos – realizado pelo músculo do estribo, esta proteção aumenta a rigidez da cadeia ossicular, protegendo o ouvido de sons muito intensos.

2.2.3 O OUVIDO INTERNO

O ouvido interno, também chamado de labirinto, é formado por escavações no osso temporal, revestidas por membranas e preenchidas por líquido. Neste ouvido, encontra-se a parte mais importante do ouvido periférico (entre o pavilhão auditivo e o nervo auditivo), um órgão em forma de caracol chamado de cóclea, responsável em grande parte pela nossa capacidade de diferenciar e interpretar sons. A cóclea é um tubo dividido em três partes e enrolado em torno de uma área central, o modíolo. As três partes que a constituem denominam-se: escala vestibular, escala média e escala timpânica. A escala vestibular e a timpânica comunicam-se no ápice da cóclea por um orifício chamado de helicotrema e são preenchidas por um líquido denominado de perilinfa. As vibrações sonoras que atingem a janela oval, através do estribo, criam uma onda de propagação na perilinfa da escala

vestibular que, através do helicotrema, continua a se propagar na escala timpânica até ser dissipada na janela redonda. A escala média é separada da escala vestibular por uma membrana delgada chamada de *"Membrana de Reissner"* e da escala timpânica pela membrana basilar (MB) sobre a qual encontra-se o *"Órgão de Corti"*.

2.2.4 MEMBRANA BASILAR

A MB é uma fina membrana que transforma movimentos mecânicos em estímulos nervosos, os quais são transmitidos ao cérebro. A MB tem dimensões diferentes ao longo da cóclea, sendo mais estreita e rígida na base e tornando-se gradualmente mais larga e flexível. Devido a essas características, quando a MB esta exposta às ondas de propagação da perilinfa, ela vibra em resposta a excitação das diferentes freqüências, apresentando pontos de maior vibração, que correspondem a picos dentro de um espectro de freqüência. Altas freqüências produzem um deslocamento máximo na base da cóclea, enquanto baixas freqüências produzem um padrão de excitação que se desloca ao longo da membrana. A freqüência de um sinal que causa o máximo deslocamento em um determinado ponto da membrana é chamado de Freqüência Central (CF).

A vibração da MB estimula as células ciliares do Órgão de Corti que estão dispostas ao longo de toda a cóclea, sendo divididas em externas e internas. As células ciliares estão em contato com a membrana tectorial onde encontramos as terminações nervosas que irão transformar o movimento das células ciliares em impulsos elétricos. Estes impulsos são transmitidos, através dos nervos auditivos, para o córtex cerebral - que fará a interpretação dos sons.

Ao deslocar-se para cada componente de freqüência do som que chega ao ouvido, com espalhamentos em torno de um ponto central, a MB faz com que duas componentes próximas possam ou não ser ouvidas, é o que chamamos de mascaramento.

2.3 MASCARAMENTO

O mascaramento é provavelmente o fenômeno mais pesquisado na audição. Na prática é definido pela American Standards Association (ASA), como a quantidade (medida em dB's) ou o processo pelo qual o limiar de audibilidade de um som é elevado pela presença de outro som (ver em http://www.minidisc.org/MaskingPaper.html).

O mascaramento auditivo pode ser dividido em dois tipos: mascaramento em freqüência e mascaramento temporal. Mascaramento em freqüência é um fenômeno do domínio da freqüência que nos diz que quando dois sinais estão próximos em freqüência, o sinal de menor intensidade é mascarado pelo de maior intensidade, ou seja: dispõem-se de um sinal A com uma freqüência de 1000 Hz e um sinal B com freqüência de 1100 Hz e 18 dB abaixo do anterior, o sinal B não pode ser ouvido porque está próximo em freqüência de um som mais forte. Já o mascaramento temporal ocorre antes e depois de um som forte. Quando um som é mascarado depois de um som mais forte, ele é chamado pós-mascaramento, e se é mascarado antes o efeito é chamado pré-mascaramento. O pré-mascaramento ocorre por um curto momento (20 ms). O pós-mascaramento tem efeito de até 200 ms. O efeito do mascaramento temporal é atribuído à natureza ressonante da MB que não pode começar ou deixar de vibrar instantaneamente [12]. O fenômeno de mascaramento está associado ao limiar de audibilidade, que faz a relação entre o nível e a freqüência do sinal analisado.

2.3.1 LIMIAR DE AUDIBILIDADE

Os humanos podem ouvir uma faixa de freqüência que varia de 20 a 20KHz, mas isso não que dizer que podem ouvir todas as freqüências da mesma maneira. Baseado no nível do sinal e na freqüência, foi criada a curva de audibilidade. Este limiar de audibilidade não é o mesmo de pessoa para pessoa e, além disso, muda com a idade.

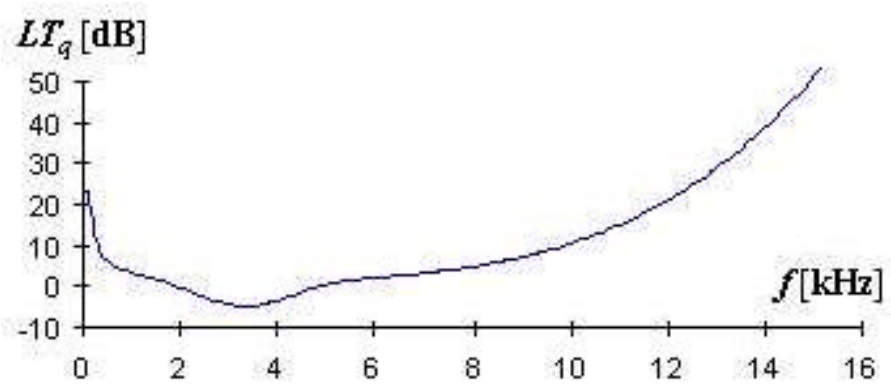


Figura 2.2 – Curva característica do Limiar de Audibilidade Humano

2.4 INIBIÇÃO LATERAL (IL)

Inibição Lateral é processo no qual uma estrutura neuronal salienta uma resposta forte em um conjunto de neurônios espacialmente organizados, inibindo os neurônios adjacentes. Sabe-se que uma rede deste tipo existe no sistema visual para detectar transições de intensidade luminosa numa imagem. Acredita-se que um esquema idêntico exista também no SA, onde a IL seria usada para melhorar a seletividade de freqüência, que é largamente determinada na cóclea, funcionando como um processo de "sintonia fina" na determinação da freqüência desejada. Este processo de sintonia fina reduz a quantidade de informação enviada para córtex cerebral, funcionando como um pré-processador que irá acentuar a diferença entre o sinal desejado e o ruído [13,14].

2.5 MASCARAMENTO BINAURAL

Mascaramento Binaural pode entendido como o mascaramento feito pelo cérebro, utilizando as informações dos dois ouvidos. O processamento e o mapeamento do som no cérebro não é conhecido mas, através de simulações realizadas, acredita-se que o mascaramento binaural ocorre quando a diferença de fase ou intensidade do sinal desejado não é igual à diferença de fase ou intensidade do ruído entre os ouvidos [8]. Tais diferenças ocorrem em situações reais onde quando a fonte do sinal desejado e a de ruído estão

localizadas em posições diferentes no espaço, como sugerido no *Cocktail Party.* A figura 2.3 ilustra algumas experiências realizadas com o mascaramento binaural. Para o caso (a), onde o mesmo sinal e ruído são aplicados em cada ouvido com a mesma fase e intensidade, o ouvinte tem dificuldade em identificar o sinal desejado. A mesma dificuldade ocorre quando os sinais são aplicados em um só ouvido (c). Verifica-se, no entanto, que somente invertendo a fase do sinal desejado, o ouvinte consegue identificá-lo (b). Da mesma forma, se o ruído for aplicado nos dois ouvidos e o sinal desejado em somente um, o ouvinte também consegue identificar o sinal desejado. Com isso observa-se que quando as relações entre os ouvidos (interaurais) são diferentes, um sinal é melhor bem identificado. Em nosso algoritmo este mascaramento foi trabalhado através da diferença de intensidade entre os sinais.

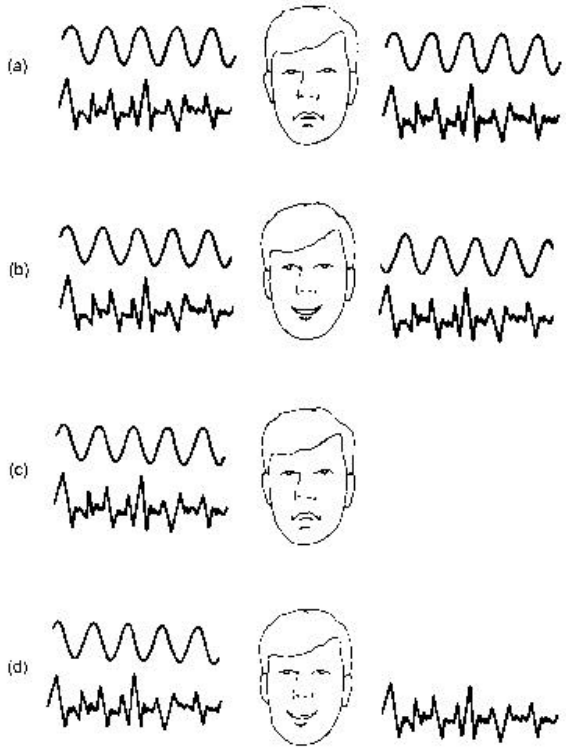


Figura 2.3 – Ilustração de duas situações em que ocorre mascaramento binaural. Em (a) e (c) o ouvinte tem dificuldade em identificar o sinal desejado, enquanto em (b) e (d) o ouvinte consegue identificar o sinal desejado, porque a relação interaural do sinal desejado e do ruído entre os ouvidos é diferente.

CAPÍTULO 3

MÉTODO

O algoritmo que desenvolvemos é bem simples e altamente motivado pela percepção do sistema auditivo. Nosso objetivo é melhorar o sinal de fala de maior energia de duas misturas convolutivas de duas fontes sonoras, adquiridas por dois microfones. Para atingir esse objetivo desenvolvemos, em ambiente matlab, um algoritmo que foi subdividido em quatro partes, como ilustrado na figura 3.1. Primeiramente criamos uma base de dados com frases ditas por dois homens e duas mulheres. Em seguida essas frases foram misturadas e convoluídas para simular uma aquisição real feita pela orelha externa. Como o SA humano é formado por dois sistemas independentes e a atividade do ouvido interno se assemelha a um banco de filtro passa-faixa, utilizamos um filtro passa-faixa para cada uma das misturas captadas pelos receptores, decompondo cada um dos sinais de entrada em várias bandas, usando [f_0, $2f_0$,..., nf_0] como freqüência central. Em seguida, utilizamos a característica psicoacústica de IL para acentuar a diferença entre os sinais de maior energia em cada uma das bandas. Os resultados obtidos com a IL foram utilizados como entrada do módulo de mascaramento binaural, o qual compara as bandas correspondentes, selecionando as bandas de maior energia, para em seguida sintetizá-las em novo vetor que conterá o sinal recuperado.

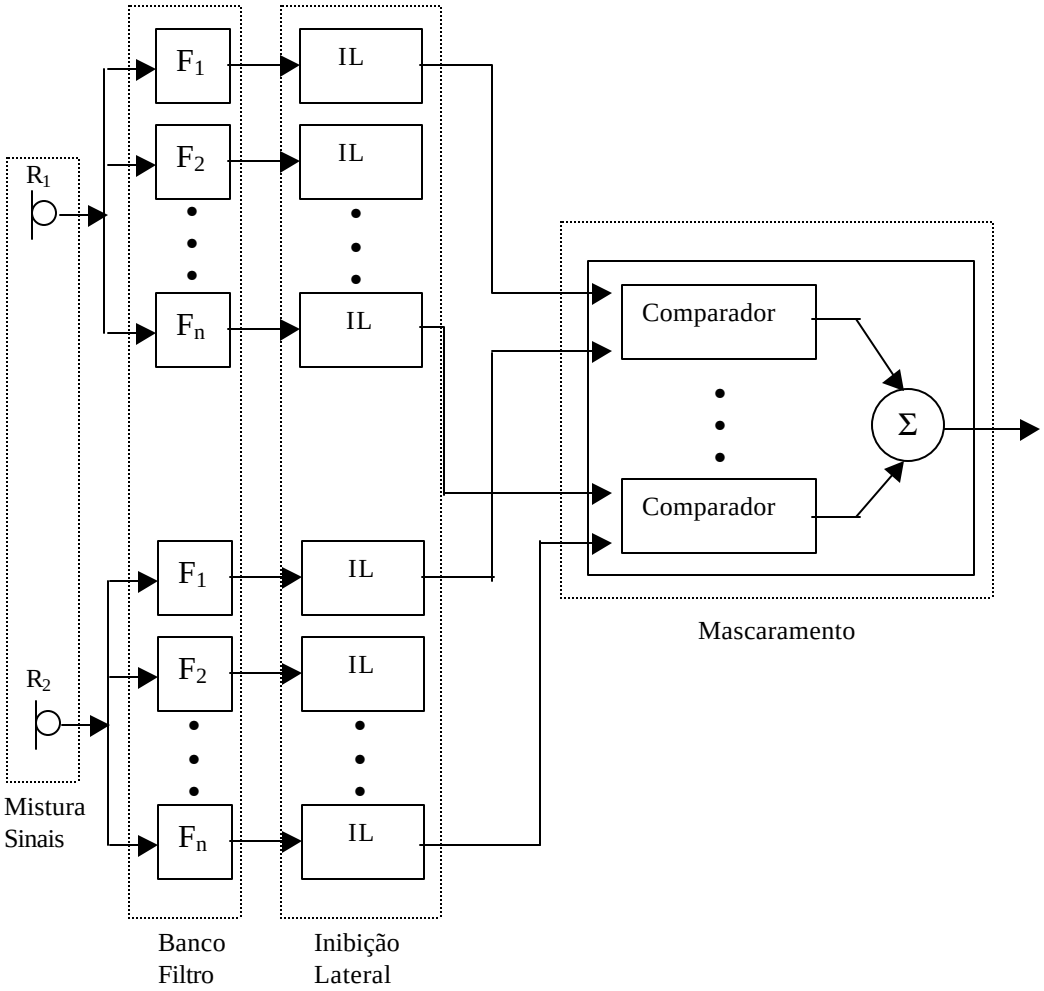


Figura 3.1 - Diagrama em bloco do algoritmo desenvolvido. Primeiro os sinais recebidos pelos microfones são processados por um banco de filtro passa-faixa. Em seguida a inibição lateral encontra os sinais de maior nível de energia em cada banda, para finalmente o módulo de mascaramento recuperar o sinal desejado.

3.1 Base de Dados e Mistura dos Sinais

A base de dados utilizada neste trabalho foi formada por textos em português gravados separadamente em ambiente fechado, com pouco ruído e obtidos a partir de quatro locutores, sendo dois homens e duas mulheres. Os sinais de áudio foram gravados com o programa *Nero wave editor*, utilizando um microfone de eletreto conectado a um computador Pentium III, processador de 1GHz. Os sinais foram gravados utilizando 16 bits por amostra, freqüência de amostragem de 8kHz e extensão tipo .wave. Quatro tipos de misturas foram realizadas com os

sinais da base de dados, para a combinação dos sinais homem x homem, homem x mulher e mulher x mulher.

Observando o desenho da figura 1.1, consideramos duas fontes de sinais no tempo t, onde os sinais originais $s = [s_1(t), s_2(t)]^T$, chegam em dois receptores $x(t) = [x_1(k), x_2(k)]^T$, cujos sinais correspondem a combinação reverberada das fontes originais.

$$x(t) = \int H(t)s(t+t)dt \qquad (3)$$

H modela a reverberação e a mistura dos sinais.

Para simular uma mistura instantânea, onde os sinais transmitidos chegam simultaneamente nos receptores, nós adicionamos os sinais originais através de uma matriz de mistura:

$$x(t) = As(t), \qquad (4)$$

A, corresponde a matriz de mistura , $s(t)$ corresponde ao sinal de entrada e $x(t)$ são os sinais dos receptores.

Em outra simulação realizada, procuramos verificar o desempenho do algoritmo em relação a sinais atrasados, desta forma, introduzimos dois tipos atrasos que chamamos de curto e longo. Para o atraso curto retardamos um sinal em relação ao outro de 5ms, ou seja, S_1 atinge R_2 5ms após S_2 e da mesma forma S_2 atinge R_1 5ms após S_1. A simulação do atraso longo, 30ms, foi semelhante ao atraso curto. No matlab estes atrasos foram conseguidos através da inserção de uma matriz de zeros, que para o atraso curto correspondeu a uma matriz de 40 pontos e para o atraso longo uma matriz de 240 pontos, uma vez que nossa freqüência de amostragem é de 8k Hz. Assim em cada receptor teremos:

$$x(t) = H(t)*s(t)\,, \qquad (5)$$

onde H modela o atraso entre os sinais e $*$ representa a convolução.

A última mistura realizada, teve como objetivo simular um ambiente real. Em um ambiente real acontece um fenômeno chamado de reverberação, que é formado pela sobreposição de ecos. Esses ecos acontecem porque o sinal transmitido é refletido pelas superfícies que compõem o ambiente. Dessa forma o sinal que chega em nossos ouvidos não é formado unicamente pelo som transmitido por uma fonte original, mas é composto pelas diferentes reflexões sofridas por este sinal, somado as reflexões de outras fontes existentes. Todas essas reflexões podem ser caracterizadas pela resposta do ambiente a um impulso. Assim, para simular um ambiente real, nós utilizamos um algoritmo que gera a resposta ao impulso de uma sala e convoluímos esta resposta com os sinais originais de cada orador.

$$x(t) = H(t)*s(t)\,, \qquad (6)$$

onde H , corresponde a resposta ao impulso da sala e o $*$ equivale a convolução.

3.2 Banco de Filtros

O sistema auditivo periférico comporta-se como um banco de filtro passa faixa , com sobreposição das bandas. A membrana basilar fornece a base para esse filtro, onde cada ponto desta membrana corresponde a uma freqüência central diferente [12].

Baseado nessa característica do ouvido periférico, dividimos os sinais dos receptores, conforme figura 3.2, usamos um banco de filtro passa-faixa $f(t)$ centrado na freqüência fundamental f_0 e suas harmônicas, de onde obtivemos valores intermediários $\varphi_{ik}(t)$ que são $x(t)$ filtrado entorno da freqüência de $k\,f_0(t),(k=1,2,\ldots N)$.

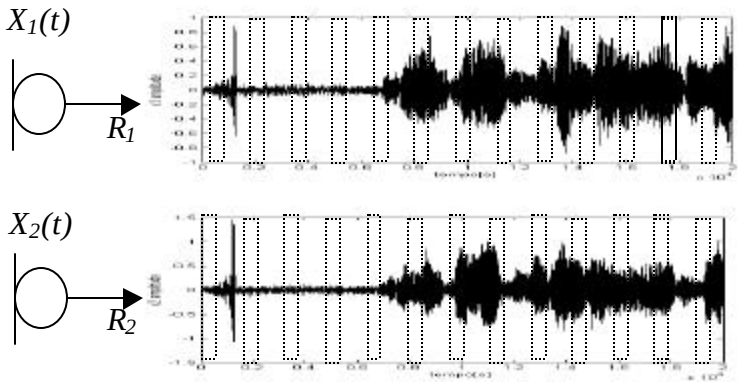


Figura 3.2 – Filtragem dos sinais misturados em n bandas.

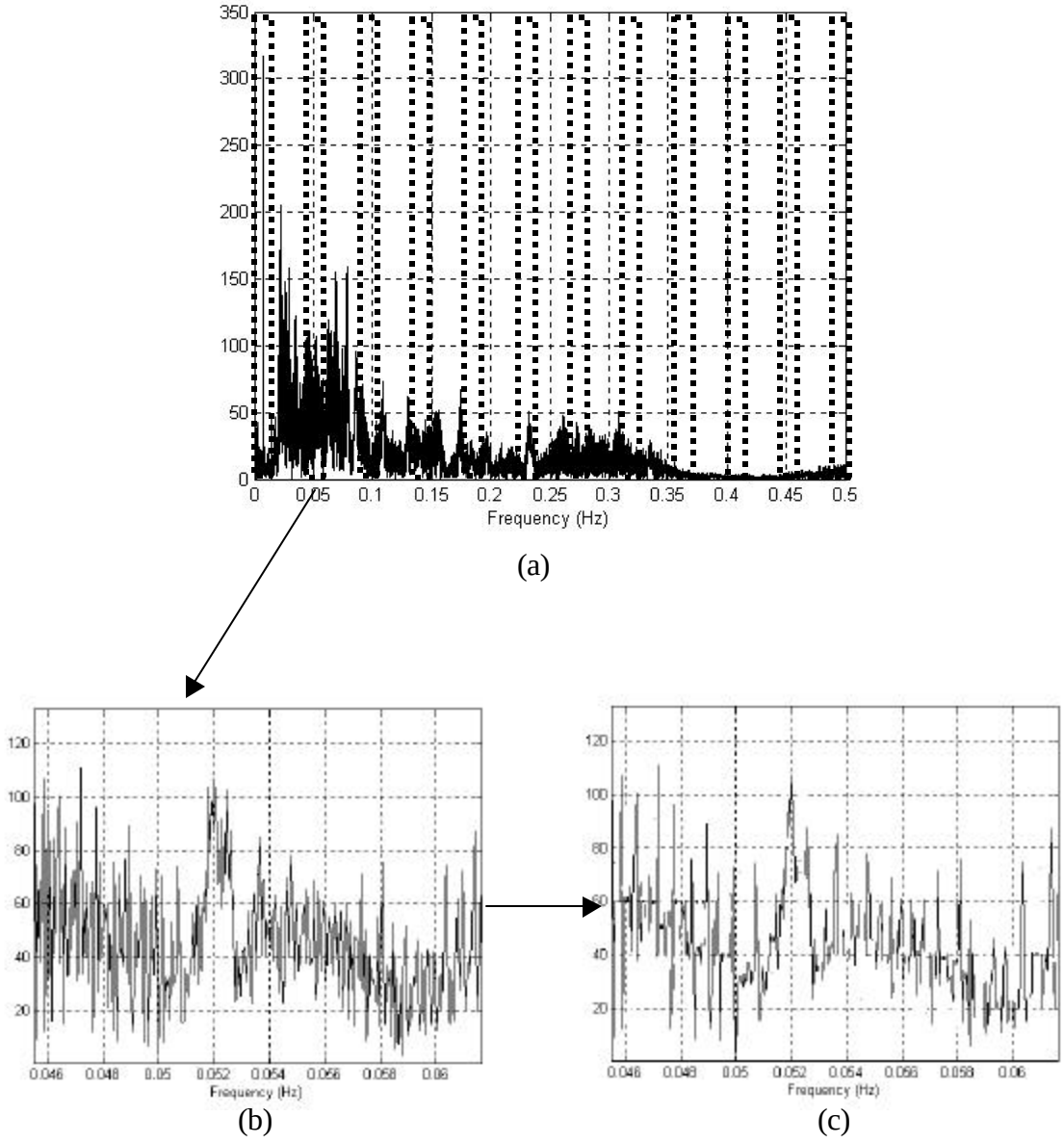
$$\varphi_{ik}(t) = \int f(t,k)x_i(t)dt \qquad (7)$$

onde N é o número de harmônicas de também de subbandas.

O filtro utilizado para dividir $x(t)$ em 72 bandas foi o *Butterworth* com freqüência f_0 de 25Hz.

3.3 Inibição Lateral

O módulo de IL foi usado para acentuar a diferença entre os sinais contidos na mesma banda. O algoritmo desenvolvido para este módulo faz uma varredura em cada banda de freqüência , através da função *argmax*, procurando em ordem decrescente, os sinais de maior nível de energia. Uma vez encontrado este sinal, seus vizinhos imediatos, da direita e esquerda, são zerados (inibidos). Após encontrar todos os sinais de maior energia, a banda é agora definida por $v(t)$. O objetivo da inibição lateral é minimizar a sobreposição de sinais. A figura 3.3 exemplifica o que foi dito.



(a)

(b) (c)

$$\delta_i(t) = \text{argmax}[\varphi_{ik}(t)] \qquad (8)$$

$$v(t) = [\delta_1(t), \ldots, \delta_n(t)] \qquad (9)$$

Figura 3.3 – Inibição lateral. (a) Sinal do receptor dividido em n bandas; (b) banda do sinal; (c) Sinal após utilizar a inibição lateral.

3.4 Mascaramento Binaural

O último módulo do sistema é responsável por comparar o nível de energia das bandas correspondentes de cada sinal de entrada. O módulo de mascaramento funciona como uma chave liga / desliga que irá armazenar em um vetor, a banda de maior energia correspondente ao receptor escolhido. Assim, se desejamos recuperar o sinal de maior energia de R_1, o algoritmo verifica quais as bandas correspondentes a este receptor apresentam maior nível de energia, comparado com os sinais do R_2 ,conforme expressão abaixo.

$$g_k(t) = \begin{cases} v_{1k}(t) & \text{se} \quad E(v_{1k}) > E(v_{2k}(t)) \\ 0 & \text{para os outros casos} \end{cases} \qquad (10)$$

onde $E(.)$ é um envelope estimador de energia, $v_{1k}(t)$ corresponde à banda selecionada, $g_k(t)$ é o vetor que armazena as amostras que formarão o sinal recuperado. Para finalizar, as bandas armazenadas em $g(t)$ são sintetizadas, formando o sinal recuperado.

$$y(t) = \sum_{k=1}^{M} g_k(t) \qquad (11)$$

CAPÍTULO 4

RESULTADOS E DISCUSSÃO

Neste capítulo apresentamos dois tipos de resultados: subjetivo e objetivo. Com estes resultados avaliamos o desempenho do algoritmo proposto, mostrado na figura 3.1, em relação a um algoritmo que não utiliza o módulo de inibição lateral, mostrado na figura 4.1. As simulações foram modeladas conforme a figura 1.1, sendo realizadas quatro tipos simulações (instantânea, com atraso curto, com atraso longo e com reverberação), utilizando a combinação de quatro sinais (homem x homem, homem x mulher e mulher x mulher). A tarefa foi melhorar o sinal de maior energia no receptor. O objetivo das simulações foi imitar a situação onde um orador está perto do ouvinte, mas há alguma interferência de fundo possivelmente causada por outros oradores. A figura 4.2 mostra um exemplo dos sinais de fala originais, da mistura entre eles e os resultados obtidos dos algoritmos com inibição lateral e sem inibição lateral, o anexo A contém amostras de todas as simulações realizadas.

Os resultados subjetivos foram obtidos através da utilização da escala MOS, tabela 4.1, para qualificar os sinal recuperado. Esta qualificação foi feita por dez ouvintes, que ouviram o sinal recuperado em média três vezes antes de conceitua-lo, o anexo B contém a tabelas com todos os valores atribuídos pelos ouvintes para as simulações. As tabelas de 4.2 a 4.9 mostram os valores obtidos para cada uma das misturas na comparação dos dois algoritmos: com e sem inibição lateral.

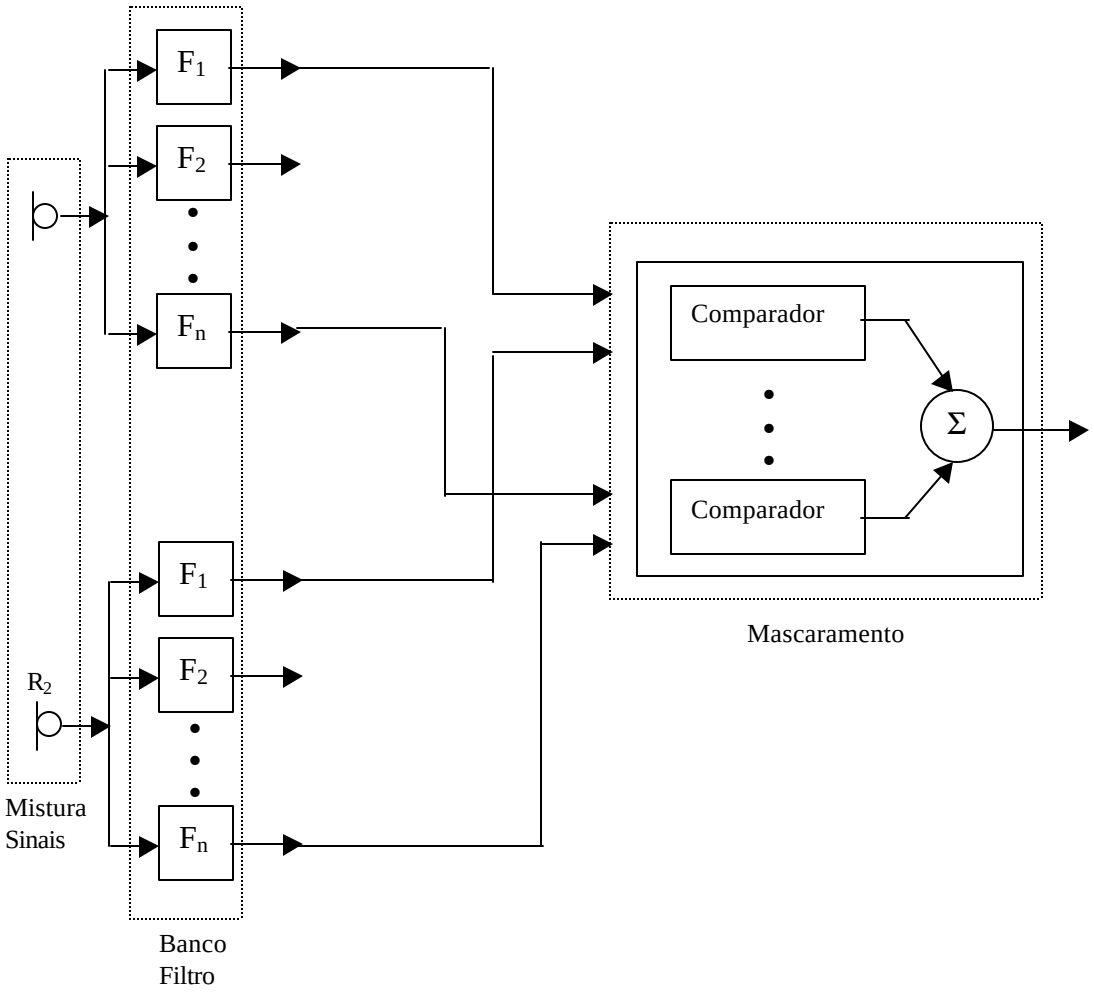
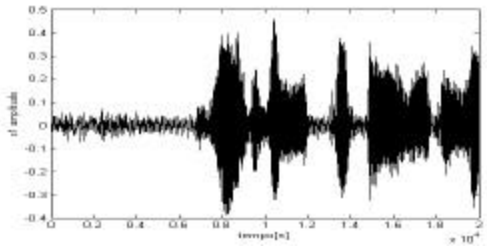


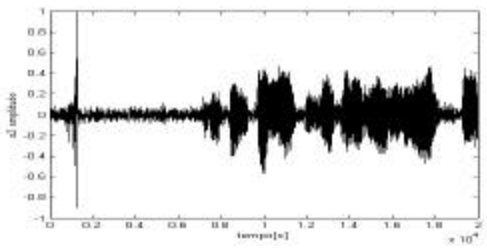
Figura 4.1 - Diagrama em bloco do algoritmo que recupera o sinal original sem o módulo de inibição lateral.

Tabela 4.1 – Referência da escala MOS

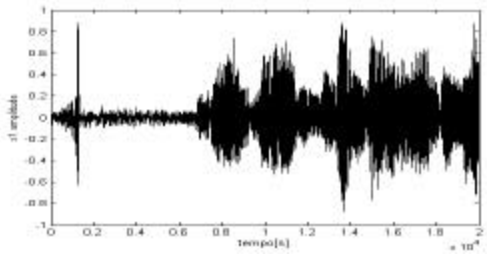
Qualidade do sinal	**Conceito**
Excelente	5
Boa	4
Razoável	3
Pobre	2
Ruim	1



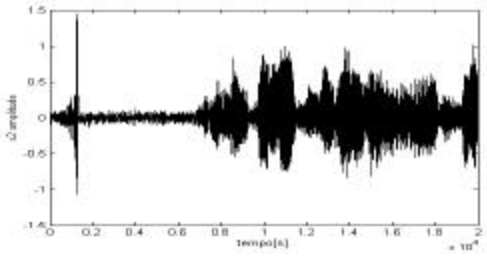
a) sinal de voz feminina1



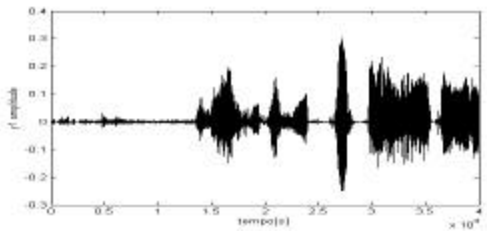
b) sinal de voz feminina2



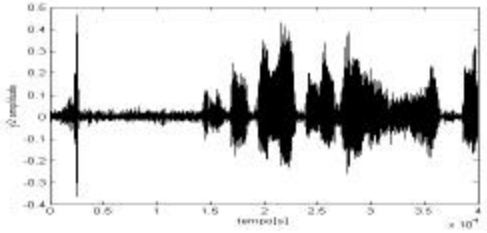
c) sinal de voz misturado



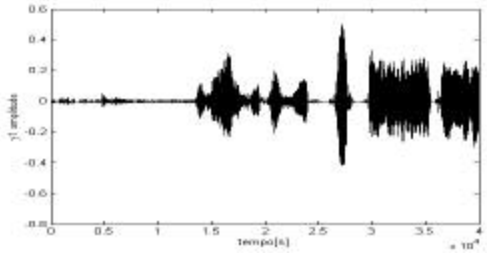
d) sinal de voz misturado



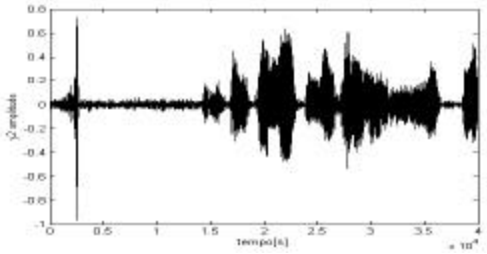
e) sinal de voz recuperado com IL



f) sinal de voz recuperado com IL



g) sinal de voz recuperado sem IL



h) sinal de voz recuperado sem IL

Figura 4.2 Amostragem dos sinais de fala originais, suas misturas e os sinais recuperados com e sem o módulo de inibição lateral, para os dois receptores.

Os resultados objetivos foram obtidos através do cálculo do erro relativo entre o sinal desejado e o sinal recuperado e entre o sinal desejado e a misturados sinais, conforme equação

$$\varepsilon = \sqrt{1/p \sum [(s_1 - s_2)/s_1]^2} \qquad (11)$$

onde p é o tamanho dos sinais, s_1 é o sinal de entrada e s_2 é o sinal recuperado.

Tabela 4.2 – Média dos Valores MOS para os Sinais Recuperados da Mistura Instantânea

	Homem x Homem		**Homem x Mulher**		**Mulher x Mulher**	
	H1	**H2**	**H1**	**M2**	**M1**	**M2**
MÉDIA MOS COM IL	4,2	4	4,4	4,1	3,9	3,8
MÉDIA MOS SEM IL	4	3,7	4,2	4	3,8	3,8

Tabela 4.3 – Tabela de Valores do Erro Relativo entre o Sinal Desejado e o Sinal Recuperado da Mistura Instantânea

	Homem x Homem		**Homem x Mulher**		**Mulher x Mulher**	
	H1	**H2**	**H1**	**M2**	**M1**	**M2**
Erro Sinal Recuperado com IL	0,1027	0,0936	0,1012	0,1090	0,0924	0,0908
Erro Sinal Recuperado sem IL	0,1817	0,1795	0,1674	0,2107	0,1803	0,1828
Erro Sinal Misturado	0,6516	0,5525	0,2898	1,2421	0,4569	0,7876

Tabela 4.4 – Média dos Valores MOS para os Sinais Recuperados da Mistura com Atraso Curto

	Homem x Homem		Homem x Mulher		Mulher x Mulher	
	H1	**H2**	**H1**	**M2**	**M1**	**M2**
MÉDIA MOS COM IL	3,7	3,6	3,8	3,6	3,6	3,5
MÉDIA MOS SEM IL	3,6	3,5	3,7	3,7	3,5	3,5

Tabela 4.5 – Tabela de Valores do Erro Relativo entre o Sinal Desejado e o Sinal Recuperado da Mistura com Atraso Curto

	Homem x Homem		Homem x Mulher		Mulher x Mulher	
	H1	**H2**	**H1**	**M2**	**M1**	**M2**
Erro Sinal Recuperado com IL	0,1195	0,1031	0,1126	0,1298	0,1054	0,1039
Erro Sinal Recuperado sem IL	0,2214	0,2014	0,2101	0,2596	0,2116	0,2125
Erro Sinal Misturado	0,6202	0,5525	0,2898	1,2416	0,4564	0,7875

Tabela 4.6 – Média dos valores MOS para os Sinais Recuperados da Mistura com Atraso Longo

	Homem x Homem		Homem x Mulher		Mulher x Mulher	
	H1	**H2**	**H1**	**M2**	**M1**	**M2**
MÉDIA MOS COM IL	3,6	3,4	3,5	3,3	3,2	3,1
MÉDIA MOS SEM IL	3,3	3,4	3,4	3,3	3,1	3,1

Tabela 4.7 – Tabela de Valores do Erro Relativo entre o Sinal Desejado e o Sinal Recuperado da Mistura com Atraso Longo

	Homem x Homem		Homem x Mulher		Mulher x Mulher	
	H1	**H2**	**H1**	**M2**	**M1**	**M2**
Erro Sinal Recuperado com IL	0,1454	0,1059	0,1161	0,1373	0,1219	0,1073
Erro Sinal Recuperado sem IL	0,2937	0,2136	0,2187	0,2753	0,2489	0,2235
Erro Sinal Misturado	0,6438	0,5525	0,2898	1,2416	0,4523	0,7875

Tabela 4.8 – Média dos valores MOS para os Sinais Recuperados da Mistura com Reverberação

	Homem x Homem		Homem x Mulher		Mulher x Mulher	
	H1	**H2**	**H**	**M**	**M1**	**M2**
MÉDIA MOS COM IL	3,6	3,4	3,6	3,5	3,5	3,4
MÉDIA MOS SEM IL	3,5	3,4	3,3	3,2	3,3	3,4

Tabela 4.9 – Tabela de Valores do Erro Relativo entre o Sinal Desejado e o Sinal Recuperado da Mistura com Reverberação

	Homem x Homem		Homem x Mulher		Mulher x Mulher	
	H1	**H2**	**H**	**M**	**M1**	**M2**
Erro Sinal Recuperado com IL	0,3001	0,2486	0,3145	0,2716	0,3417	0,3240
Erro Sinal Recuperado sem IL	0,8426	0,4334	0,6922	0,5763	0,6648	0,5162
Erro Sinal Misturado	1,3444	2,0160	1,0643	3,7257	1,5579	1,2134

4.3 DISCUSSÃO

A separação de sinais de um ambiente como o cocktail party, não é uma tarefa muito fácil, principalmente quando se deseja trabalhar em ambiente real. Isto ocorre devido a dificuldade em criar algoritmos que se adaptem a qualquer ambiente, uma vez que a reverberação e a resposta ao impulso, características de um ambiente real, variam de acordo com o ambiente. Observando esta dificuldade, nos espelhamos no sistema auditivo e utilizando as características psicoacústicas de inibição lateral e mascaramento binaural desenvolvemos um algoritmo que imita o SA na seleção do sinal de maior energia de um dos receptores.

Durante a execução deste trabalho várias simulações foram realizadas, cujos registros encontram-se nos apêndices A,B e C. Observando somente apêndice A, fica difícil perceber o melhoramento obtido com o algoritmo que utiliza inibição lateral em relação ao outro. Os resultados mostram que para as duas simulações (com e sem inibição lateral), os sinais recuperados são bastante semelhantes ao sinal original, ou seja, não é possível perceber quão mais próximo do sinal original conseguimos chegar utilizando a combinação: inibição lateral e mascaramento binaural. Esse melhoramento no sinal de saída pode ser percebido quando submetemos os resultados a percepção auditiva. O apêndice B e as tabelas 4.2, 4.4, 4.6 e 4.8 mostram os valores atribuídos por 10 ouvintes na avaliação dos resultados. Observando o gráfico do apêndice B, criado a partir da média dos valores dados pelos ouvintes, podemos perceber que o algoritmo que utilizou a inibição lateral como pré-processamnto, obteve melhor desempenho na recuperação do sinal original. Para validar ainda mais o algoritmo proposto, calculamos o erro relativo entre o sinal original e os sinais recuperados. Os valores obtidos nesse calculo, mostrados no apêndice C, confirmam que o resultado obtido com o algoritmo que utilizou a característica psicoacústica de inibição lateral possui menor erro relativo, isto é, está mais próximo do sinal desejado.

Quando verificamos o desempenho do algoritmo que utiliza somente o mascaramento binaural pudemos observar que as maiores perdas ocorrem quando o sinal original esta sujeito a atrasos e principalmente reverberações, o que também pode ser visto nos trabalhos [2], [3], [5], [10], [17], [30] e [32]. Pudemos observar também que quando o sinal está sujeito a reverberação, junto ao sinal recuperado percebe-se um som como de um grilo causado pelo ruído que não foi removido totalmente ou o suficiente para não ser percebido na saída. Esse

problema, também conhecido como *"ALIAS"*, ocorre devido a dificuldade em separar sinais próximos em freqüência. Para eliminar esse problema, introduzimos a inibição lateral em cada uma das bandas obtidas com filtro passa-faixa. Como a IL elimina os vizinhos mais próximos do sinal de maior energia, ocorre um distanciamento entre as freqüências e conseqüentemente diminui o efeito de *ALIAS*.

Além da inibição lateral, deve-se dar atenção especial aos filtros passa-faixa que provocará um som distorcido quanto maior for a diferença entre as freqüências centrais. Isso ocorre devido à falta de continuidade do sinal, ou seja, quanto menor o número de amostras pior será a qualidade do sinal recuperado. Em seu trabalho Aoki, *et al*. [2], conclui que trabalhando com freqüência central de 10Hz, consegue-se melhor resultado na recuperação do sinal. A dificuldade em usar freqüências de resolução muito baixas para a recuperação de sinais é que apesar de obtermos melhores resultados, perdemos no tempo de processamento computacional o que inviabiliza tais algoritmos para trabalharem em tempo real.

CAPÍTULO 5

CONCLUSÃO

A busca por sistemas de identificação mais perfeitos, que possam trabalhar em qualquer ambiente e em tempo real, tem direcionado os pesquisadores ao estudo do sistema auditivo, tendo como objetivo sua simulação computacional, uma vez que através deste sistema conseguimos identificar e localizar uma fonte sonora misturada a outras.

O objetivo deste trabalho foi melhorar o sinal de fala de maior energia de um dos receptores, através da redução do ruído. A redução de ruído em sinal de fala é uma linha de pesquisa bastante explorada que encontra em telecomunicações sua principal área de atuação, cujo objetivo é uma comunicação em tempo real com o mínimo de ruído possível. Já existem sistemas que realizam tais tarefas, como os utilizados pela telefonia móvel, onde é possível ao cliente escolher, através do comando de voz, algumas opções no menu de atendimento ao cliente. Mas este sistema, por exemplo, não funciona adequadamente quando existem várias fontes de sinais simultâneas, sendo necessário um melhor tratamento do sinal coletado para eliminar os sinais indesejáveis.

Nesta dissertação apresentamos um algoritmo que a partir da mistura de dois sinais de fala, melhora o sinal de maior energia de um dos receptores, através da utilização de duas características psicoacústicas: inibição lateral e mascaramento binaural. A inibição lateral foi usada para reduzir o ruído do sinal de saída, funcionando como um pré-processador do sinal que será analisado pelo sistema de decisão, que neste trabalho é o mascaramento binaural, .

Várias simulações foram realizadas utilizando mistura de vozes masculinas e femininas em quatro tipos de situações: mistura instantânea, mistura com atraso curto e com atraso longo e mistura com reverberação. Estas simulações foram realizadas em ambiente matlab e os resultados foram comparados com os resultados obtidos por um modelo convencional que utiliza somente o mascaramento binaural. Em todas as simulações realizadas verificamos que o algoritmo proposto obteve melhor desempenho tanto nos resultados subjetivos quanto nos objetivos, comprovando a eficiência do algoritmo proposto.

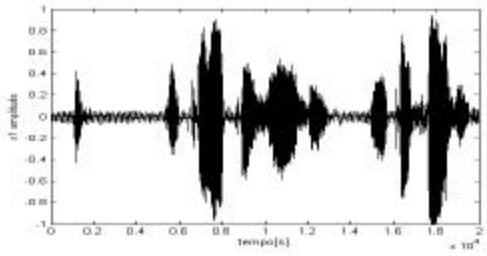
Por ser um algoritmo simples, de rápido processamento e por apresentar bons resultados, ele pode ser implementado, em trabalhos futuros, para operar em tempo real podendo ser acoplado como módulo de redução de ruído em sistemas de reconhecimento de fala on line e também em sistemas que integram visão e audição. Novos testes também podem ser feitos na busca do melhoramento deste algoritmo, como: utilizar um filtro adaptativo e aumentar o número de inibições realizadas em cada banda do filtro passa-faixa.

APÊNDICE A

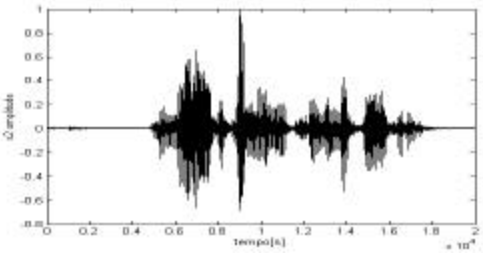
AMOSTRAS DOS SINAIS UTILIZADOS E RECUPERADOS

As figuras a seguir são amostras dos sinais utilizados e recuperados em cada uma das misturas realizadas.

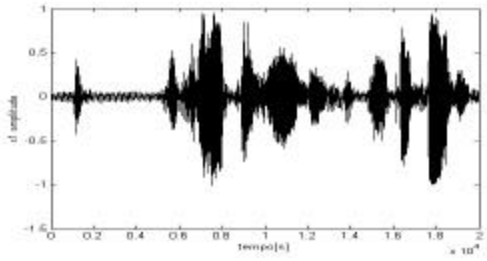
MISTURA INSTANTÂNEA: HOMEM X HOMEM



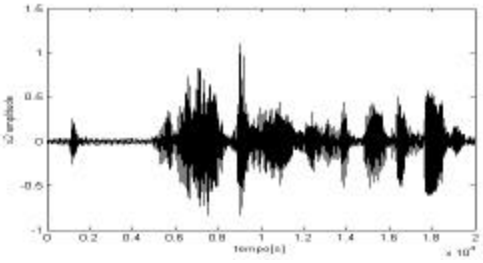
a) sinal de voz masculina1



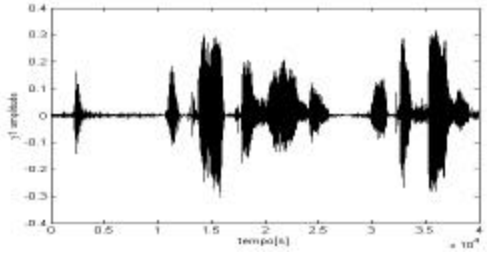
b) sinal de voz masculina2



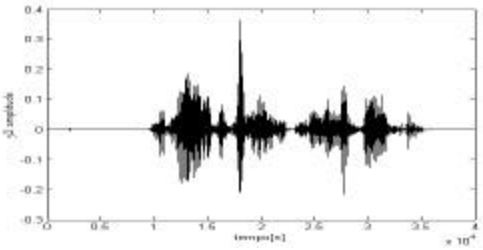
c) sinal de voz misturado



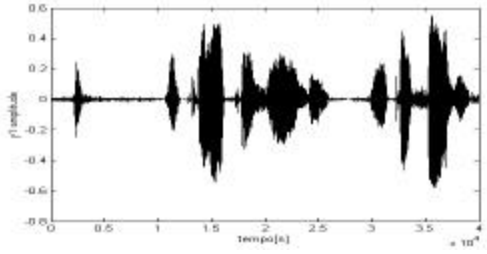
d) sinal de voz misturado



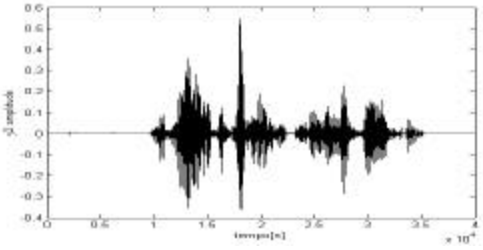
e) sinal de voz recuperado com IL



f) sinal de voz recuperado com IL



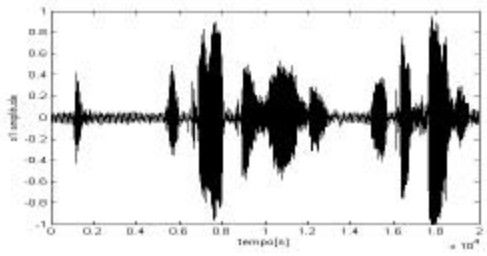
g) sinal de voz recuperado sem IL



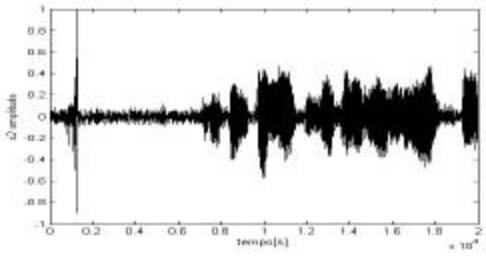
h) sinal de voz recuperado sem IL

Figura A.1 – Mistura dos Sinais Homem x Homem Utilizando Mistura Instantânea

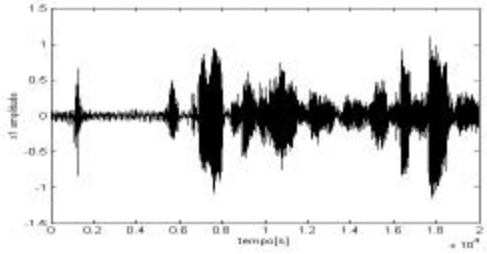
MISTURA INSTANTÂNEA: HOMEM X MULHER



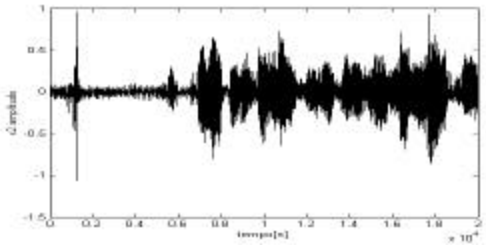
a) sinal de voz masculina



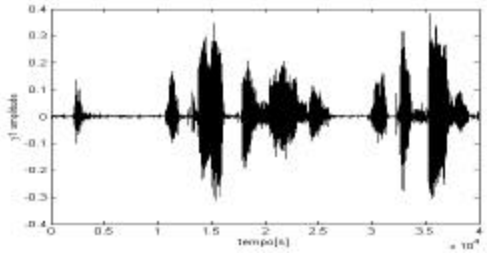
b) sinal de voz feminina



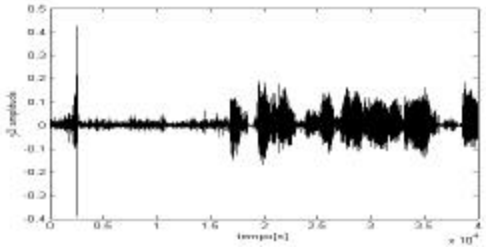
c) sinal de voz misturado



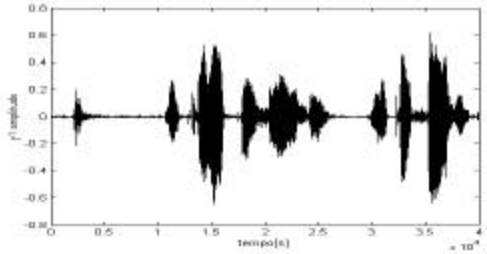
d) sinal de voz misturado



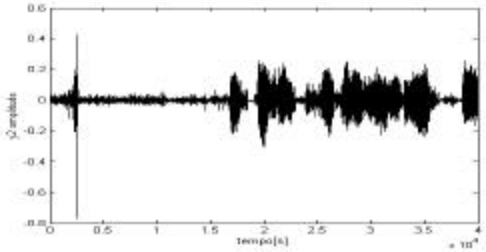
e) sinal de voz recuperado com IL



f) sinal de voz recuperado com IL



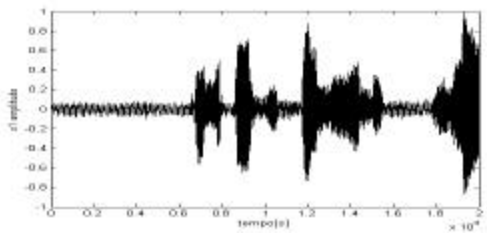
g) sinal de voz recuperado sem IL



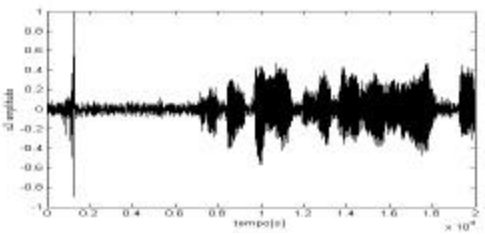
h) sinal de voz recuperado sem IL

Figura A 2 – Mistura dos Sinais Homem x Mulher Utilizando Mistura Instantânea.

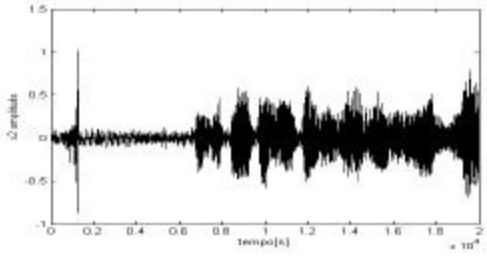
MISTURA INSTANTÂNEA: MULHER X MULHER



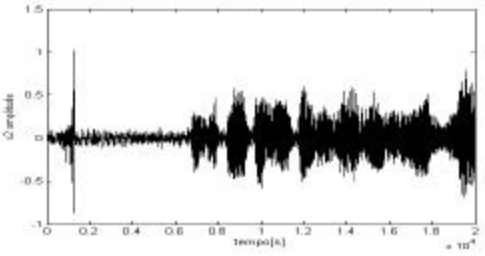
a) sinal de voz feminina1



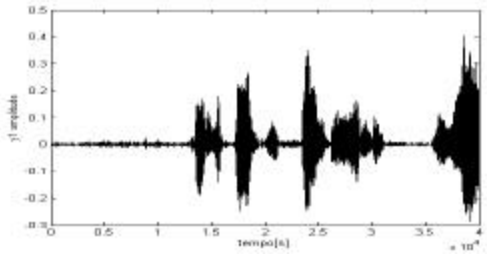
b) sinal de voz feminina2



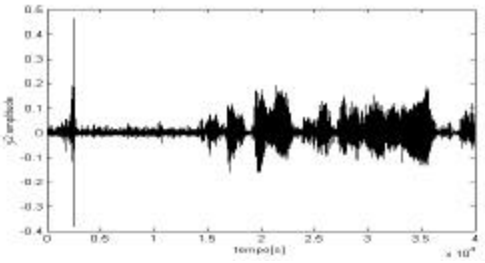
c) sinal de voz misturado



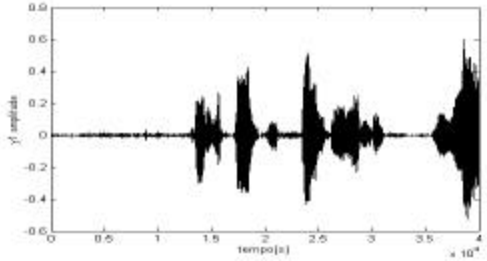
d) sinal de voz misturado



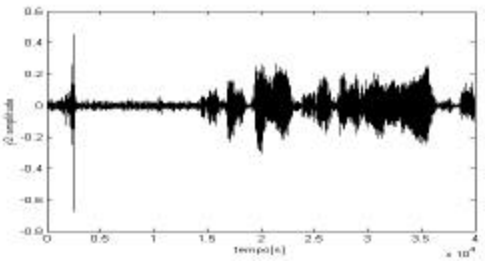
e) sinal de voz recuperado com IL



f) sinal de voz recuperado com IL



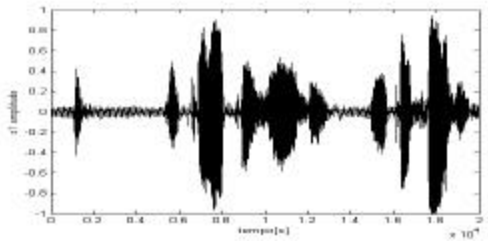
g) sinal de voz recuperado sem IL



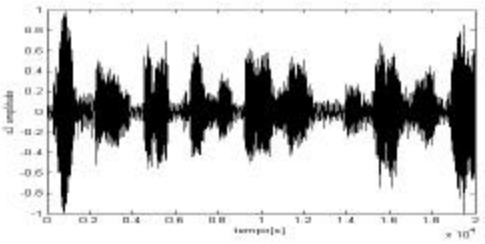
h) sinal de voz recuperado sem IL

Figura A 3 – Mistura dos Sinais Mulher x Mulher Utilizando Mistura Instantânea.

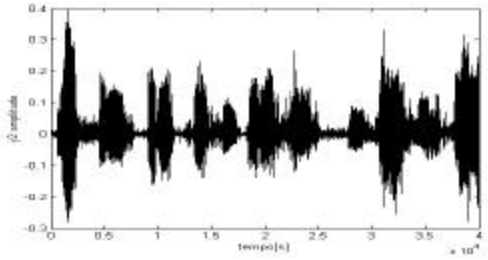
MISTURA COM ATRASO CURTO: HOMEM X HOMEM



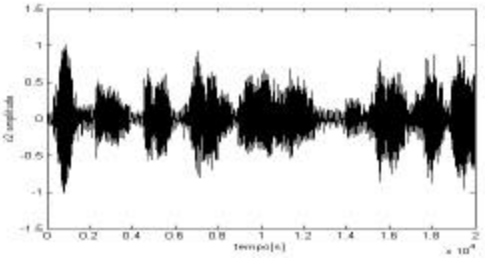
a) sinal de voz masculina1



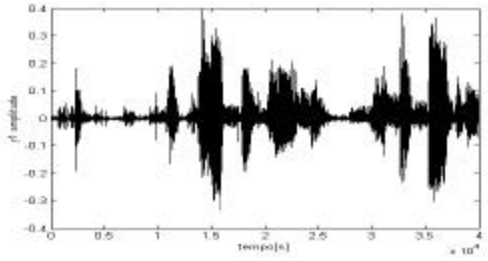
b) sinal de voz masculina2



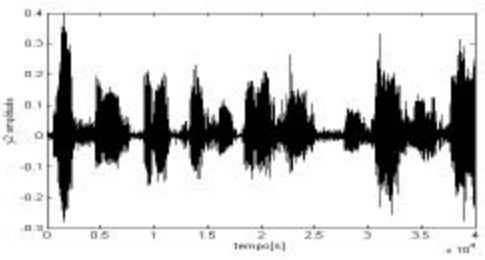
c) sinal de voz misturado



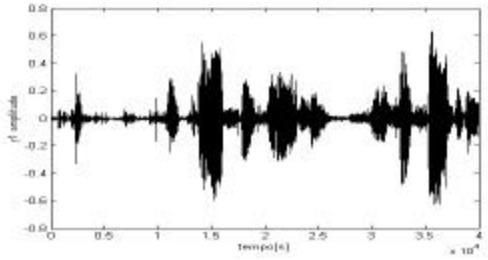
d) sinal de voz misturado



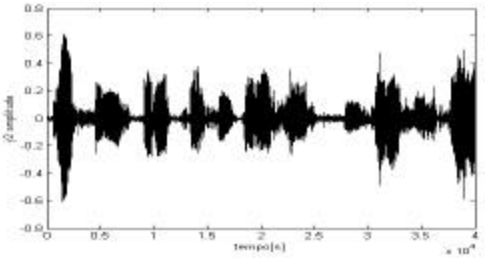
e) sinal de voz recuperado com IL



f) sinal de voz recuperado com IL



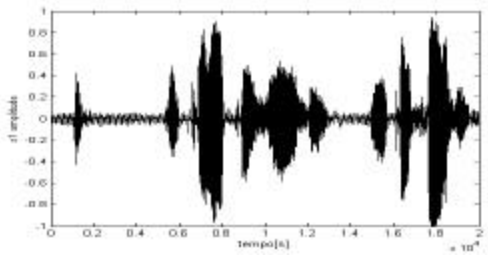
g) sinal de voz recuperado sem IL



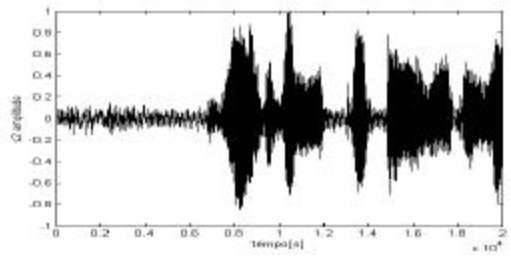
h) sinal de voz recuperado sem IL

Figura A 4 – Mistura dos Sinais Homem x Homem Utilizando Mistura com Atraso Curto.

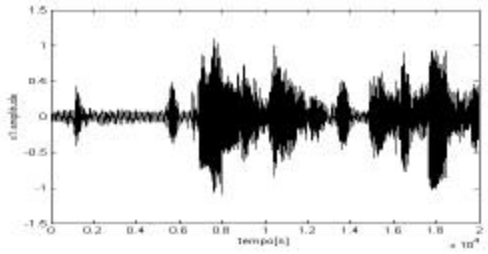
MISTURA COM ATRASO CURTO: HOMEM X MULHER



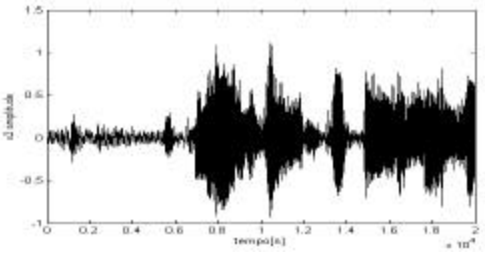
a) sinal de voz masculina



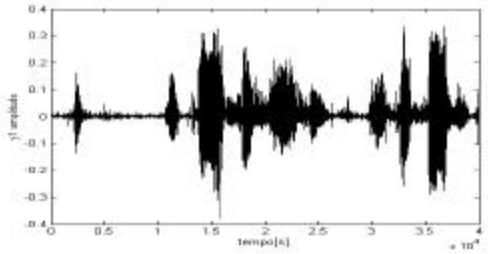
b) sinal de voz feminina



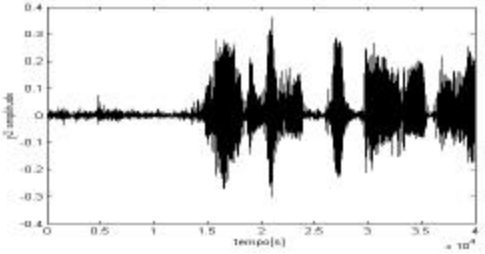
c) sinal de voz misturado



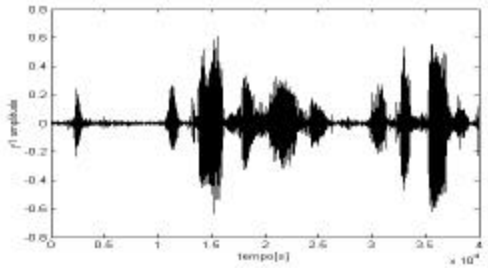
d) sinal de voz misturado



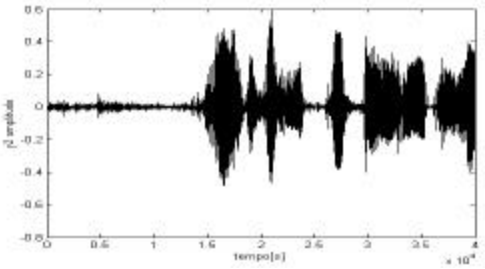
e) sinal de voz recuperado com IL



f) sinal de voz recuperado com IL



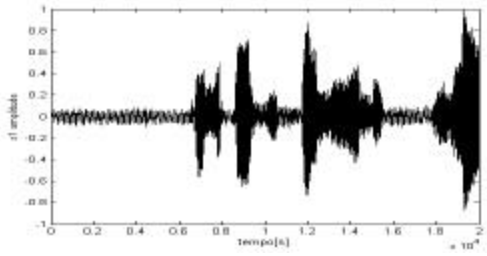
g) sinal de voz recuperado sem IL



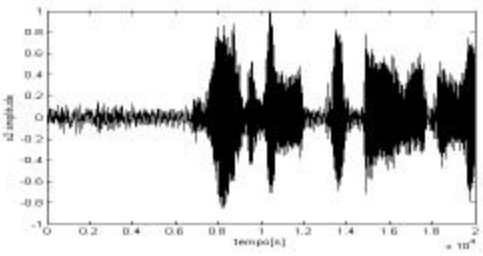
h) sinal de voz recuperado sem IL

Figura A 5 – Mistura dos Sinais Homem x Mulher Utilizando Mistura com Atraso Curto.

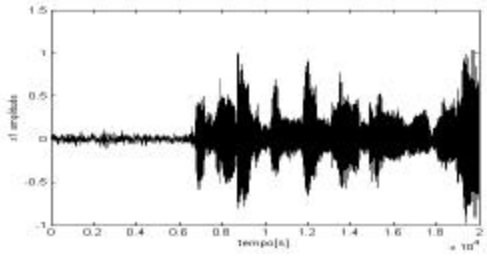
MISTURA COM ATRASO CURTO: MULHER x MULHER



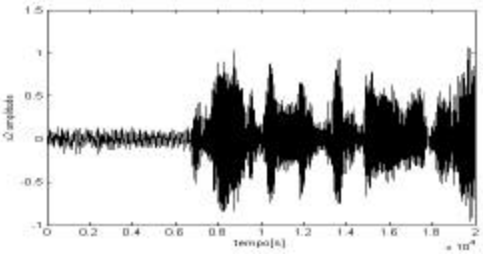
a) sinal de voz feminina1



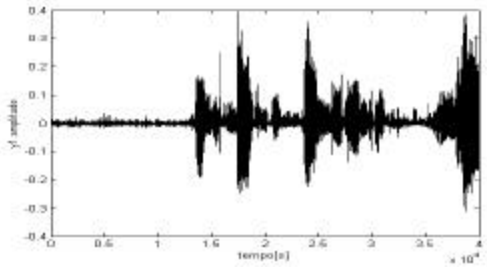
b) sinal de voz feminina2



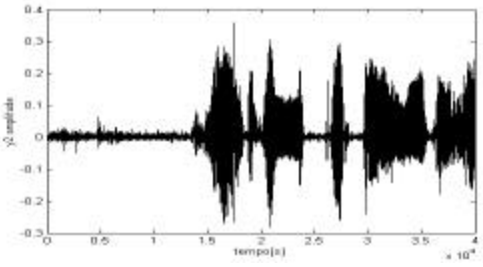
c) sinal de voz misturado



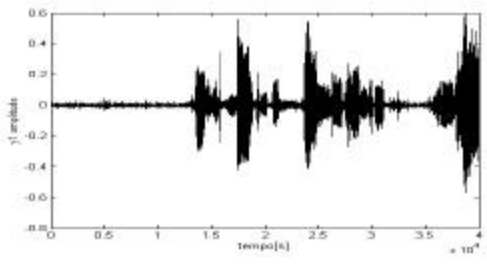
d) sinal de voz misturado



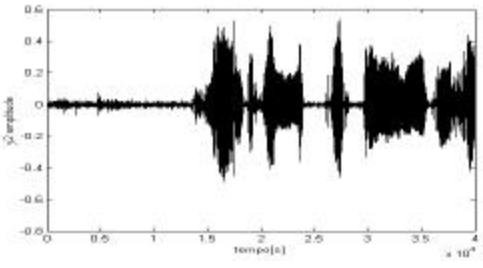
e) sinal de voz recuperado com IL



f) sinal de voz recuperado com IL



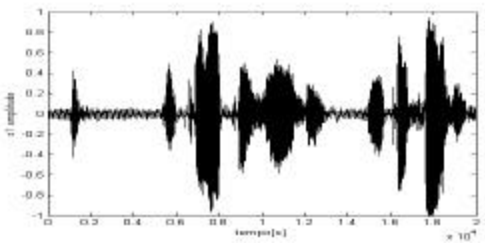
g) sinal de voz recuperado sem IL



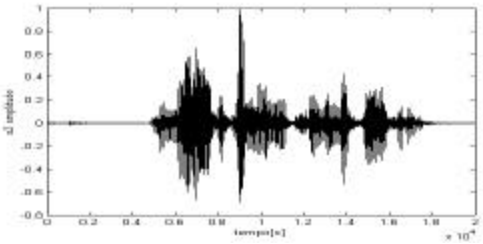
h) sinal de voz recuperado sem IL

Figura A 6 – Mistura dos Sinais Mulher x Mulher Utilizando Mistura com Atraso Curto.

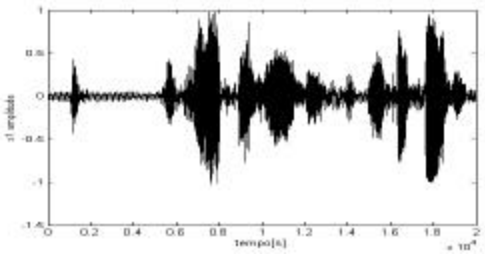
MISTURA COM ATRASO LONGO: HOMEM x HOMEM



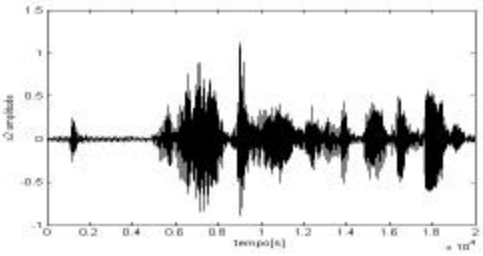
a) sinal de voz masculina1



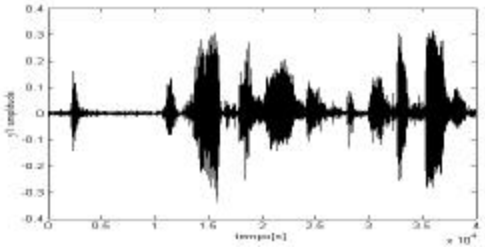
b) sinal de voz masculina2



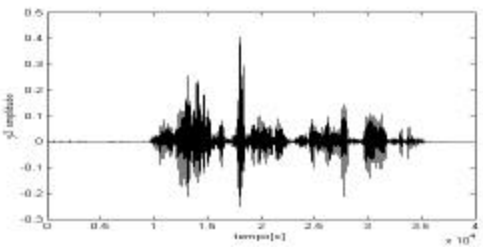
c) sinal de voz misturado



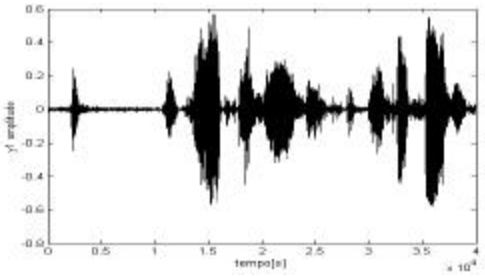
d) sinal de voz misturado



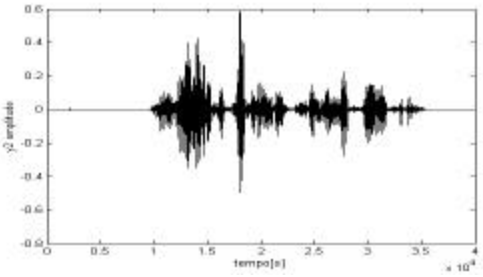
e) sinal de voz recuperado com IL



f) sinal de voz recuperado com IL



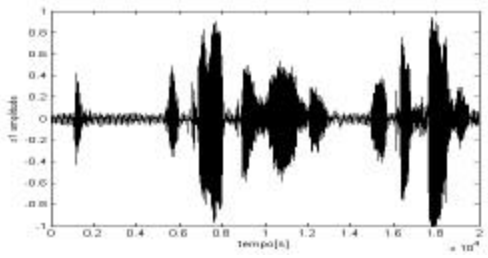
g) sinal de voz recuperado sem IL



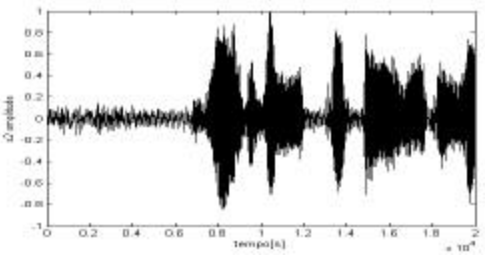
h) sinal de voz recuperado sem IL

Figura A 7 – Mistura dos Sinais Homem x Homem Utilizando Mistura com Atraso Longo.

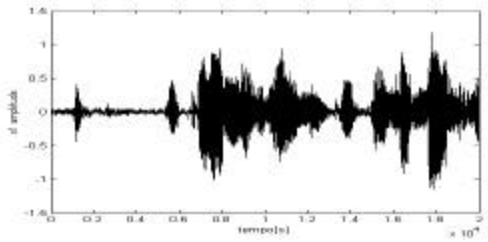
MISTURA COM ATRASO LONGO: HOMEM x MULHER



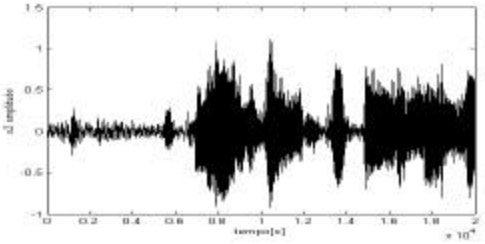
a) sinal de voz masculina



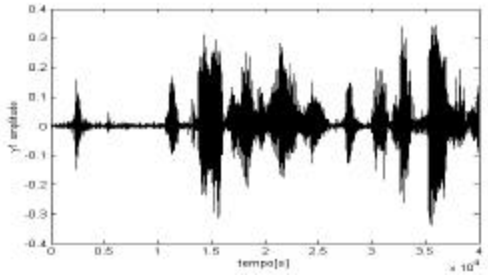
b) sinal de voz feminina



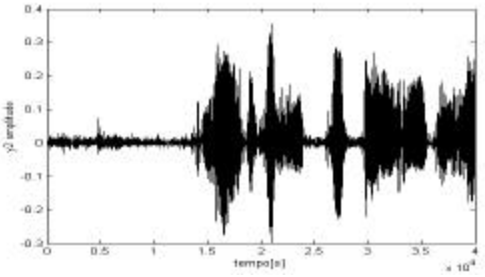
c) sinal de voz misturado



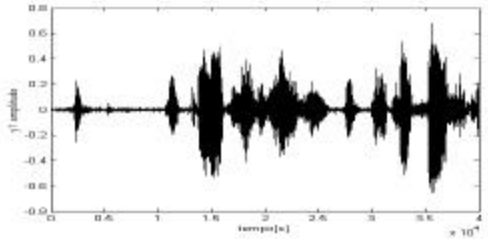
d) sinal de voz misturado



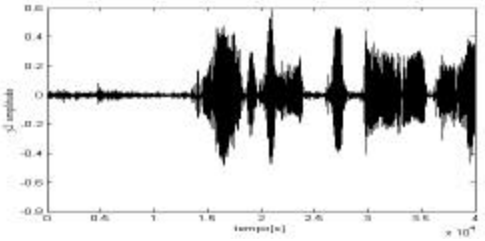
e) sinal de voz recuperado com IL



f) sinal de voz recuperado com IL



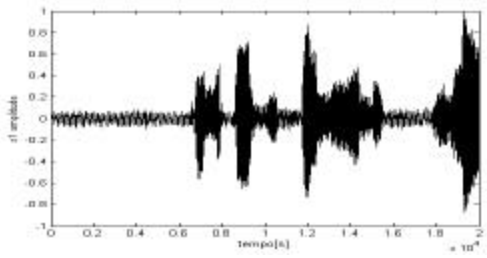
g) sinal de voz recuperado sem IL



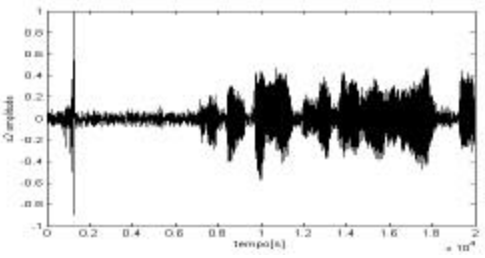
h) sinal de voz recuperado sem IL

Figura A 8 – Mistura dos Sinais Homem x Mulher Utilizando Mistura com Atraso Longo.

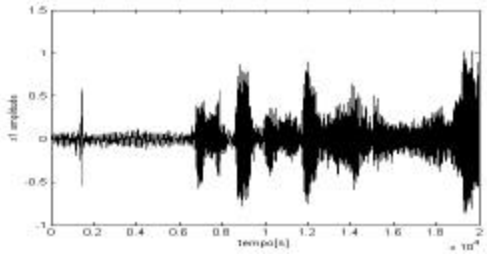
MISTURA COM ATRASO LONGO: MULHER x MULHER



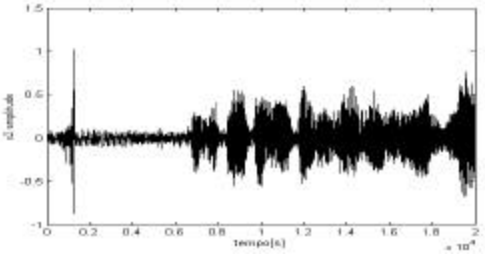
a) sinal de voz feminina1



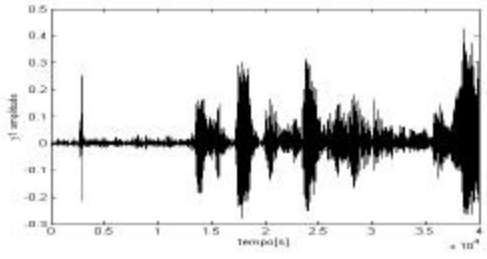
b) sinal de voz feminina2



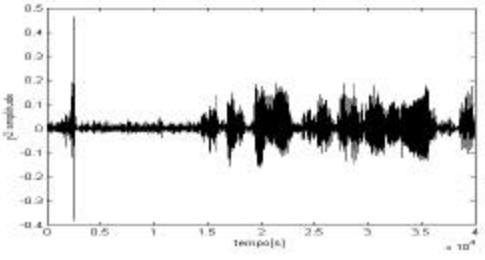
c) sinal de voz misturado



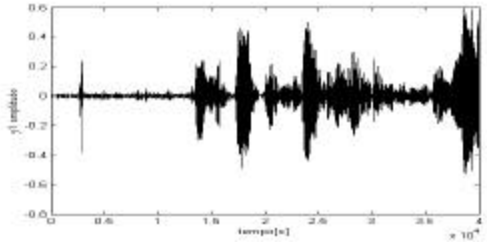
d) sinal de voz misturado



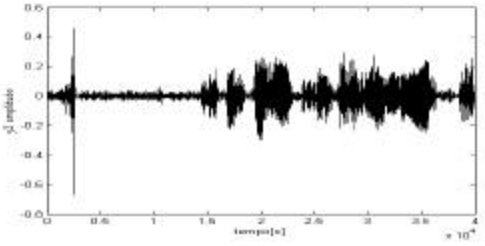
e) sinal de voz recuperado com IL



f) sinal de voz recuperado com IL



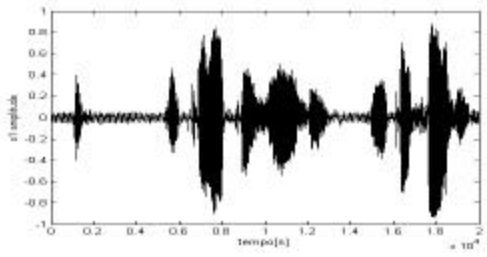
g) sinal de voz recuperado sem IL



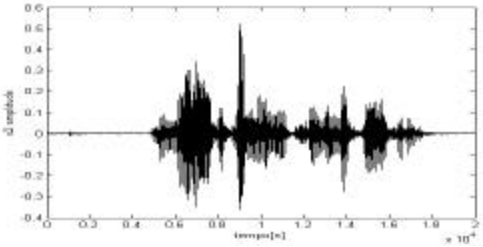
h) sinal de voz recuperado sem IL

Figura A 9 – Mistura dos Sinais Mulher x Mulher Utilizando Mistura com Atraso Longo.

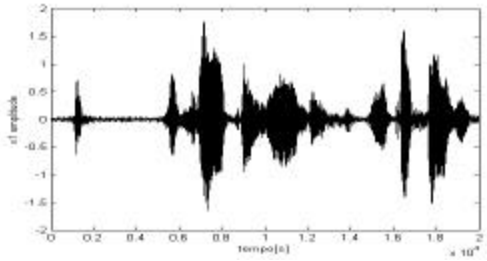
MISTURA COM REVERBERAÇÃO: HOMEM x HOMEM



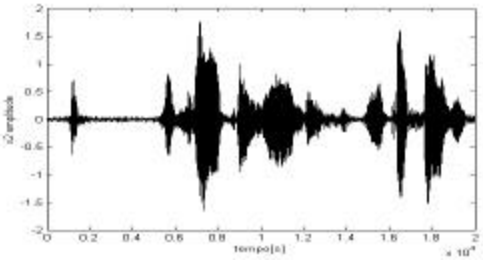
a) sinal de voz masculina1



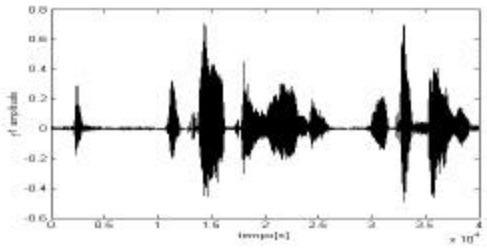
b) sinal de voz masculina2



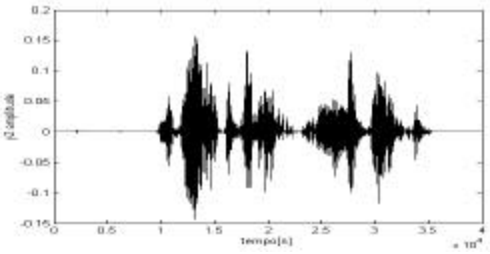
c) sinal de voz misturado



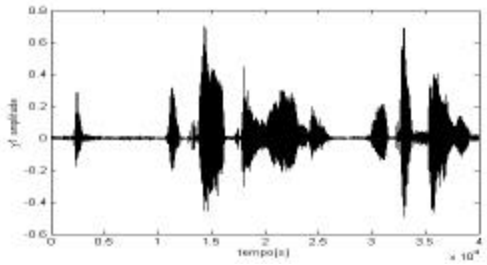
d) sinal de voz misturado



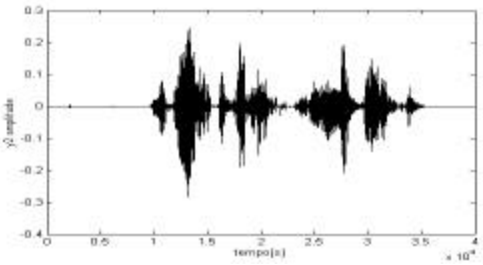
e) sinal de voz recuperado com IL



f) sinal de voz recuperado com IL



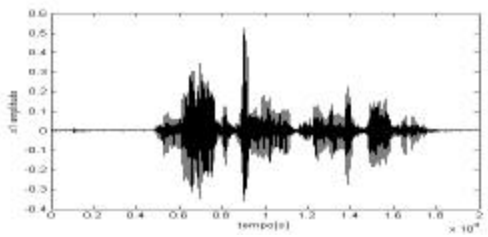
g) sinal de voz recuperado sem IL



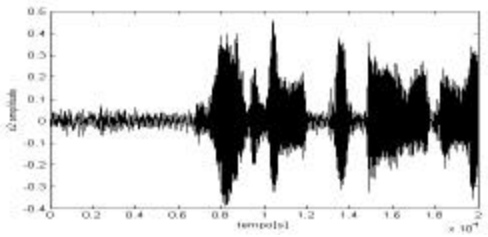
h) sinal de voz recuperado sem IL

Figura A 10 – Mistura dos Sinais Homem x Homem Utilizando Mistura com Reverberação.

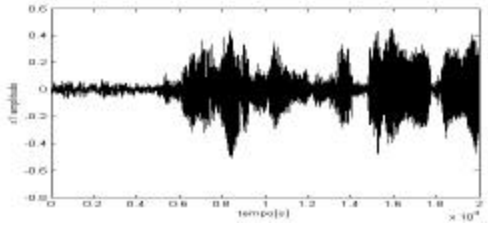
MISTURA COM REVERBERAÇÃO: HOMEM x MULHER



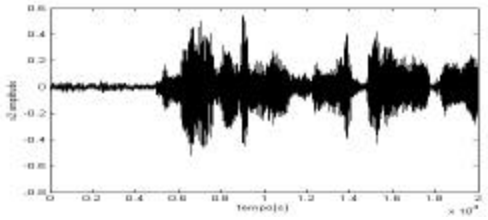
a) sinal de voz masculina



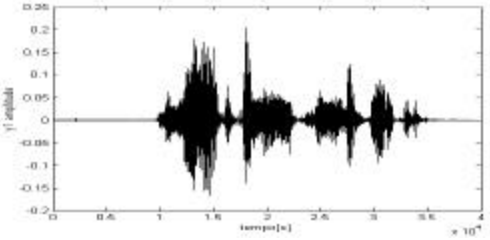
b) sinal de voz feminina



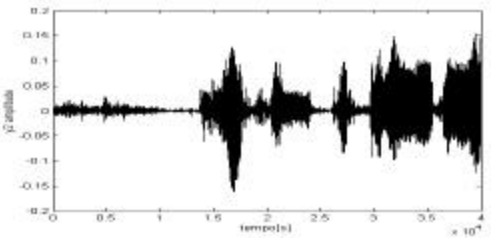
c) sinal de voz misturado



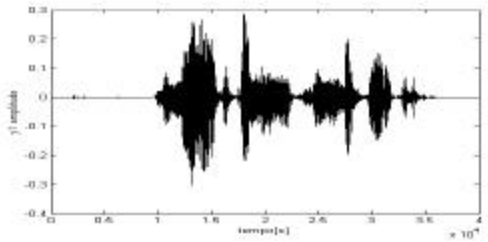
d) sinal de voz misturado



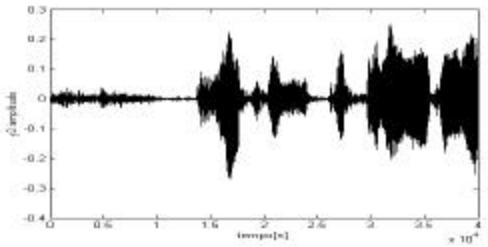
e) sinal de voz recuperado com IL



f) sinal de voz recuperado com IL



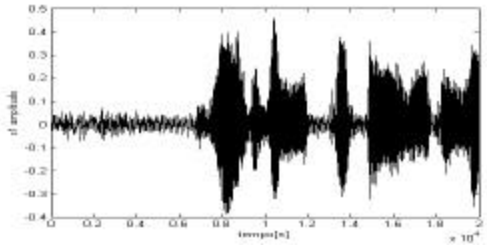
g) sinal de voz recuperado sem IL



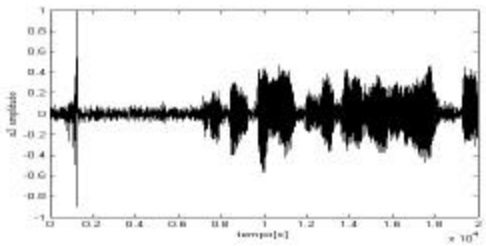
h) sinal de voz recuperado sem IL

Figura A 11 – Mistura dos Sinais Homem x Mulher Utilizando Mistura com Reverberação.

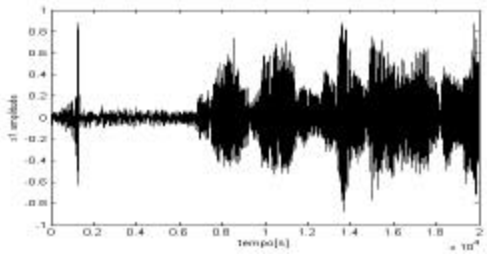
MISTURA COM REVERBERAÇÃO: MULHER x MULHER



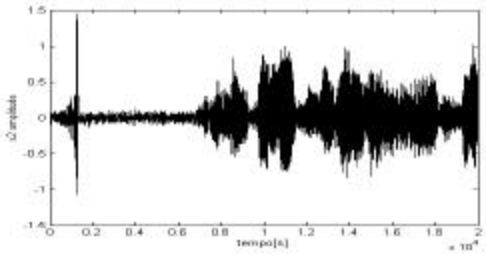
a) sinal de voz feminina1



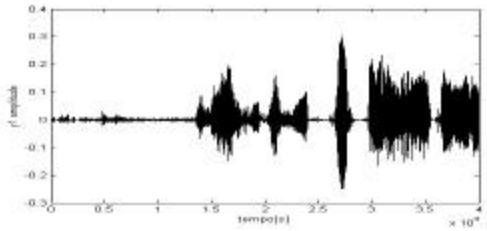
b) sinal de voz feminina2



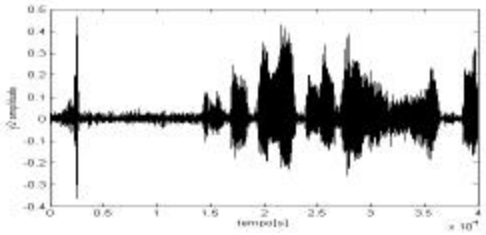
c) sinal de voz misturado



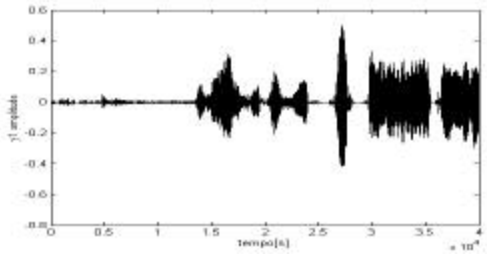
d) sinal de voz misturado



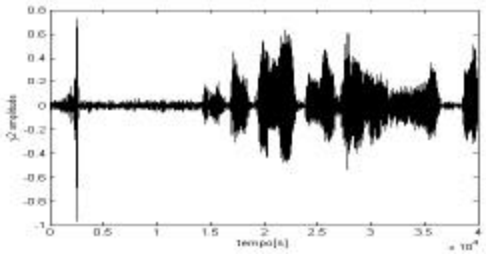
e) sinal de voz recuperado com IL



f) sinal de voz recuperado com IL



g) sinal de voz recuperado sem IL



h) sinal de voz recuperado sem IL

Figura A 12 – Mistura dos Sinais Mulher x Mulher Utilizando Mistura com Reverberação.

APÊNDICE B

VALORES MOS

A tabela a seguir contém a média e os valores da escala MOS, atribuídos por dez ouvintes para cada uma das simulações realizadas. Cada ouvinte escutou em média três vezes cada sinal recuperado, antes de conceitua-lo através da escala MOS. O gráfico mostra a relação entre os valores atribuídos pra cada simulação e entre os dois modelos: com e sem inibição lateral.

B1 - Valores MOS Atribuídos por 10 Ouvintes para o Sinal Recuperado das Mistura Instantânea e com Atraso Curto

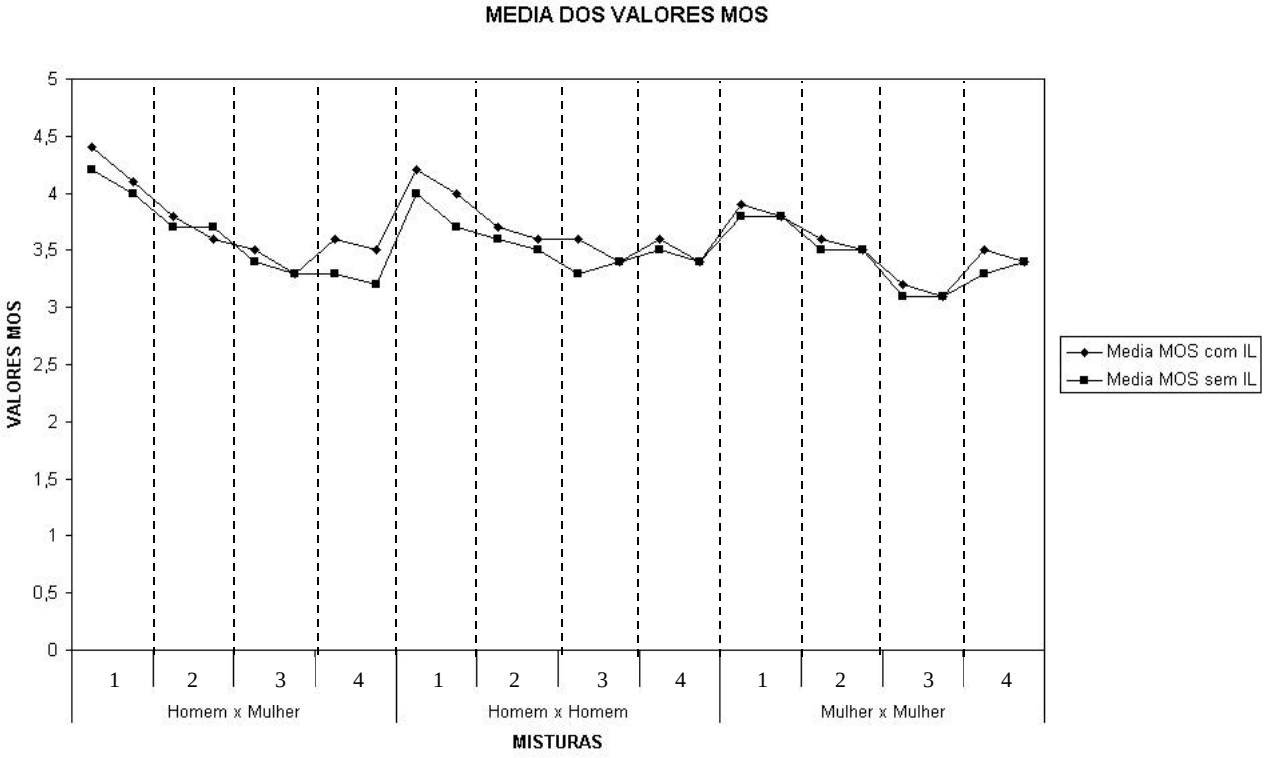
		HOMEM x MULHER				HOMEM x HOMEM				MULHER x MULHER			
	OUVINTES	Com inibição Lateral		Sem inibição lateral		Com inibição Lateral		Sem inibição lateral		Com inibição Lateral		Sem inibição lateral	
		s1	s2	s1	s2	s1	s2	s1	s2	s1	s2	s1	s2
MISTURA INSTANTÂNEA	1	5	4	4	4	4	4	4	4	4	4	4	4
	2	4	4	5	4	4	4	4	4	4	4	4	4
	3	4	5	4	4	5	4	4	4	4	3	4	4
	4	5	4	4	4	4	4	4	4	4	4	3	4
	5	4	4	4	4	4	5	4	3	4	4	4	4
	6	5	4	4	4	5	4	4	4	4	4	4	3
	7	4	4	5	4	4	4	4	4	4	4	4	4
	8	5	4	4	4	4	4	4	3	4	4	4	3
	9	4	4	4	4	4	4	4	3	4	4	4	4
	10	4	4	4	4	4	3	4	4	3	3	3	4
	Média MOS	4,4	4,1	4,2	4	4,2	4	4	3,7	3,9	3,8	3,8	3,8

		HOMEM x MULHER				HOMEM x HOMEM				MULHER x MULHER			
	OUVINTES	Com inibição Lateral		Sem inibição lateral		Com inibição Lateral		Sem inibição lateral		Com inibição Lateral		Sem inibição lateral	
		s1	s2	s1	s2	s1	s2	s1	s2	s1	s2	s1	s2
MISTURA COM ATRASO CURTO	1	3	3	4	4	4	4	3	3	3	3	3	3
	2	4	4	4	4	3	4	3	3	4	3	4	4
	3	4	3	4	3	4	3	4	4	4	3	4	3
	4	3	4	4	4	3	4	4	4	4	4	3	2
	5	4	4	3	4	3	4	3	4	3	4	4	4
	6	4	3	4	3	4	3	4	3	4	4	4	3
	7	4	4	4	4	4	3	4	4	3	3	3	4
	8	4	4	3	4	4	3	4	4	4	4	3	4
	9	4	3	4	3	4	4	3	3	4	3	4	4
	10	4	4	3	4	4	4	4	3	3	4	3	4
	Média MOS	3,8	3,6	3,7	3,7	3,7	3,6	3,6	3,5	3,6	3,5	3,5	3,5

B2 - Valores MOS Atribuídos por 10 Ouvintes para o Sinal Recuperado das Misturas com Atraso Longo e com Reverberação

		HOMEM x MULHER				HOMEM x HOMEM				MULHER x MULHER			
	OUVINTES	Com inibição Lateral		Sem inibição lateral		Com inibição Lateral		Sem inibição lateral		Com inibição Lateral		Sem inibição lateral	
		s1	s2	s1	s2	s1	s2	s1	s2	s1	s2	s1	s2
MISTURA COM ATRASO LONGO	1	4	3	4	3	4	3	4	4	2	2	3	3
	2	3	3	3	4	4	3	3	4	3	3	3	3
	3	4	3	4	3	4	4	3	3	3	3	3	3
	4	3	4	3	3	3	3	4	3	4	3	4	3
	5	4	3	4	4	4	3	3	3	3	3	3	3
	6	3	3	3	4	4	4	4	4	4	3	3	4
	7	3	4	3	3	4	3	3	3	4	4	3	3
	8	4	4	4	3	3	3	3	3	3	3	3	3
	9	3	3	3	3	3	4	3	4	3	3	3	3
	10	4	3	3	3	3	4	3	3	3	4	3	3
	Média MOS	3,5	3,3	3,4	3,3	3,6	3,4	3,3	3,4	3,2	3,1	3,1	3,1

		HOMEM x MULHER				HOMEM x HOMEM				MULHER x MULHER			
	OUVINTES	Com inibição Lateral		Sem inibição lateral		Com inibição Lateral		Sem inibição lateral		Com inibição Lateral		Sem inibição lateral	
		s1	s2	s1	s2	s1	s2	s1	s2	s1	s2	s1	s2
MISTURA COM REVERBERAÇÃO	1	4	4	3	3	4	4	3	3	3	3	3	4
	2	3	3	4	3	3	4	3	3	4	3	4	3
	3	4	3	3	3	5	4	3	3	3	3	3	3
	4	3	4	4	3	3	3	4	4	4	4	3	4
	5	4	3	3	4	4	3	4	3	3	3	4	4
	6	4	4	3	3	3	3	4	4	4	4	3	3
	7	3	4	4	4	4	4	3	4	3	4	4	3
	8	4	3	3	3	3	3	4	3	4	3	3	3
	9	3	3	3	3	4	3	3	4	4	4	3	4
	10	4	4	3	3	3	3	4	3	3	3	3	3
	Média MOS	3,6	3,5	3,3	3,2	3,6	3,4	3,5	3,4	3,5	3,4	3,3	3,4



B3 – Gráfico dos valores MOS

1 – Mistura Instantânea
2 – Mistura com Atraso Curto
3 – Mistura com Atraso Longo
4 – Mistura com Reverberação

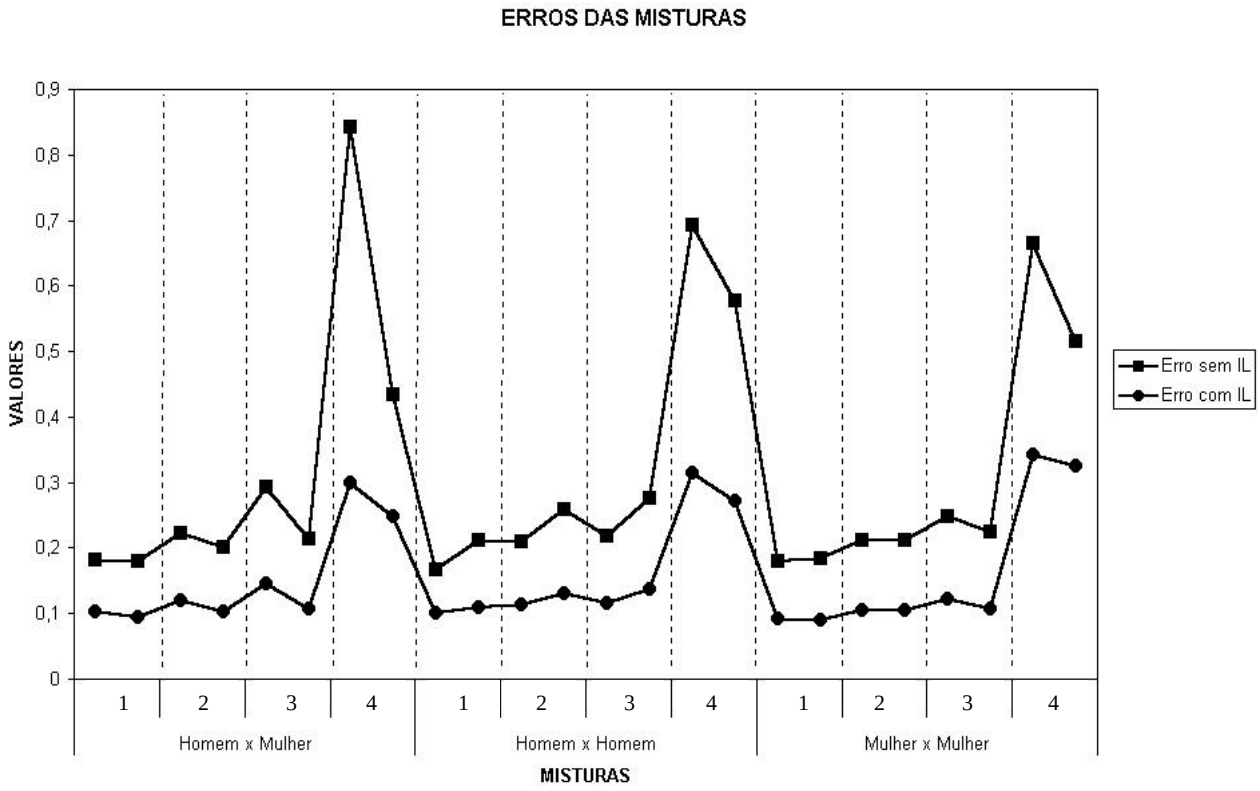
APÊNDICE C

ERROS DAS MISTURAS

Apresentamos os valores encontrados com o cálculo do erro relativo, entre o sinal desejado e o sinal recuperado para cada um dos tipos de misturas (instantânea, com atraso curto, com atraso longo e com reverberação), de duas formas: através da exposição em tabela e também em gráfico. Na exposição em tabela apresentamos também os valores do erro relativo entre o sinal desejado e a mistura.

C1 Valores do Erro Relativo das Misturas

Misturas		Homem x Homem		Homem x Mulher		Mulher x Mulher	
		H1	H2	H1	M2	M1	M2
Instantânea	Erro sinal Recuperado com IL	0,1027	0,0936	0,1012	0,1090	0,0924	0,0908
	Erro sinal Recuperado sem IL	0,1817	0,1795	0,1674	0,2107	0,1803	0,1828
	Erro sinal Misturado	0,6516	0,5525	0,2898	1,2421	0,4569	0,7876
Atraso Curto	Erro sinal Recuperado com IL	0,1195	0,1031	0,1126	0,1298	0,1054	0,1039
	Erro sinal Recuperado sem IL	0,2214	0,2014	0,2101	0,2596	0,2116	0,2125
	Erro sinal Misturado	0,6202	0,5525	0,2898	1,2416	0,4564	0,7875
Atraso Longo	Erro sinal Recuperado com IL	0,1454	0,1059	0,1161	0,1373	0,1219	0,1073
	Erro sinal Recuperado sem IL	0,2937	0,2136	0,2187	0,2753	0,2489	0,2235
	Erro sinal Misturado	0,6438	0,5525	0,2898	1,2416	0,4523	0,7875
Reverberada	Erro sinal Recuperado com IL	0,3001	0,2486	0,3145	0,2716	0,3417	0,324
	Erro sinal Recuperado sem IL	0,8426	0,4334	0,6922	0,5763	0,6648	0,5162
	Erro sinal Misturado	1,3444	2,0160	1,0643	3,7257	1,5579	1,2134



C2 Gráfico do Erro Relativo das Misturas

1 – Mistura Instantânea
2 – Mistura com Atraso Curto
3 – Mistura com Atraso Longo
4 – Mistura com Reverberação

Referências Bibliográficas

[1] Cheveigné, Alain *"The auditory system as a separation machine"* Proc. International Symposium on Hearing, in preparation 2000.

[2] Aoki Mariko, Okamoto M., Aoki S., Matsui H., Sakurai T. and Kaneda Y., *"Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones*" Acoust. Sci & Tech. 22, 2 (2001).

[3] Barros Allan K., Itakura F., Rytkowski T., Mansour Ali, Ohnishi Noboru *"Estimation of speech embedded in noisy environment using two microphones*". Proc. ICA'2000, Helsinki, Finland. v.1.pp 423-428.

[4] Lourens, T., Nakadai, K., Okuno, H. G. and Kitano,H. *"Selective attention by integration of vision and audition"* in Proc. of First IEEE-RAS International Conference on Humanoid Robot (Humanoid-2000).2000, IEEE/RSJ.

[5] Ottaviani L. and Rocchesso D. "*Separation of speech signal from complex auditory scenes*". Proceedings of the COST G-6 Conference on Digital Audio Effect (DAFX-01), Limerick, Ireland, December 6-8,2001.

[6] Virag, N. "Single *channel speech enhancement based on masking properties of the human auditory system"*. IEEE Tran. on Signal processing, Vol. 7, No.2, pp. 126-137, 1999.

[7] Barros Allan K., Itakura F., Rytkowski T., Mansour Ali, Ohnishi Noboru *"Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets*". IEEE Trans. on Neural Networks, Vol. 13, No. 4, pp 888-893, 2002.

[8] Moore Brian C.J. *"An introduction to the psychology of hearing*". Academic press 4th edition, 1997.

[9] Gustafsson S., Jax P., and Vary P., "*A novel psychoacustically motivated audio enhancement algorithm preserving background noise characteristcs"* In Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pages 397-400, 1998.

[10] Tsoukalas, D., Mourjopoulos J and Kokkinakis G. "*Speech enhancement based on audible noise suppresion"* IEEE Tran. Speech Audio Processing, 5:479, November 1997.

[11] Nakatani,T, Goto, M., Okono,H.G. "*Localization by harmonic structure and its application to harmonic sound stream segregation"*. Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceeding., 1996 IEEE International Conference on, Volume:2, 7-10 May 1996. Pages: 653 – 656 vol. 2.

[12] M. Christian Brown "*Audition*" Fundamental Neuroscience, chapter 27, 1999.

[13] Ifukube, Tohru and White, Robert L. *"A speech processor with lateral inhibition for an eight channel cochlea implant and its evaluation by subject testing"* Acoustics, Speech and Signal Processing, IEEE International Conference on ICASSP 86,, volume:11, Apr 1986

[14] Singh, L and Sridharan S. *"Speech enhancement using pre-processing"* TENCON '97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications'., Proceedings of IEEE , Volume: 2 , 2-4 Dec. 1997 Page(s): 755 -758 vol.2

[15] Antón, Emma Rodero *"El tono de la voz masculina y femenina en los informativos radiofónicos: un análisis comparativo"* Universidade Pontíficia de Salamanca. Congresso Internacional Mujeres, Hombres y Medios de Comunicación, Junta de Castilla y León, Valladolid, noviembre de 2001.

[16] Arons, B., *"A review of the cocktail party effect"*, Journal of the American Voice I/O Society. 1992.

[17] Berouti, M., Schwartz, R. and Makhou, J. *"Enhancement of speech corrupted by acoustic noise"* in Porc. IEEE ICASSP, Washington, DC, Apr.1979, pp. 208-211.

[18] Berthommier F. and Choi S. "*Evaluation of CASA and BSS models for sub band cocktail-party speech segregation*", Proceedings of ICSP' 01, Daejeon, Korea, 2001.

[19] Bodden M., "*Modeling human sound-source localization and cocktail-party-effect*" Acta Acoustical, vol1, pp. 43-55, 1993.

[20] Bodden, M. (1995) *"Binaural modeling and auditory scene analysis*. IEEE Signal Processing Society, 1995 Workshop on Appl. of Signal Processing to Audio and Acoustics, Mohonk Mountain House, New Paltz, NY

[21] Bregman, A., S. *"Auditory scene analysis"* Cambridge, MA: MIT press, 1990.

[22] Cunningham S., BG (2001). *"Localizing sound in rooms*" in Proceedings of the ACM SIGGRAPH and Euro graphics Campfire: Acoustic Rendering for Virtual Environments, Snowbird, Utah, 26-29 May 2001, 17-22.

[23] Ellis D. *"Speech recognition as a component in computational auditory scene analysis*", International Computer Science institute. Unpublished monograph. (4pp)

[24] Filho, Jozué Vieira *"Redução de ruído em sinais de voz usando critérios psicoacústicos"*, Universidade Estadual Paulista (DEE/FEI/UNESP). Simpósio Brasileiro de Telecomunicações, 2000, Gramado-RS.

[25] Fishbach, Alon *"Primary segmentation of auditory scenes"* Pattern Recognition, 1994. Vol. 3 - Conference C: Signal Processing, Proceedings of the 12th IAPR International Conference on , October 9-13, 1994 Page(s): 113 -117 vol.3

[26] Gold, B. and Morgan, N. *"Speech and audio signal processing"* John Wiely and Sons, 2000.

[27] Lyon R. F., "*A computational model of binaural localization and separation*" Proceedings of IEEE ICASSP, 1983

[28] Okuno Hiroshi G., Nakadai K., Lourens T. and Kitano H. "*Separating three simultaneous with two microphones by integration auditory and visual processing*". Proceedings of European Conference on Speech Processing (Euro speech 2001), to appear, Sep. 2001.

[29] Perdigão F., and Sá L., *'Modelo computacional da cóclea humana"*, ACÚSTICA'98 - Congresso Ibérico de Acústica, pp. 419-422, Lisbon, September, 1998

[30] Roman N., Wang D. L. and Brown G. J., "*Speech segregation based on sound localization*" *Proc. IJCNN*, pp. 2861-2866, 2001.

[31] Shields, P.W. and Campbell, D. R. *"Multi-Microphone noise cancellation for hearing aid performance"* Proc. ICASSP-97, IEEE Conference on Acoustics, Speech and Signal Processing, , pp 415-418

[32]Tsoukalas, D., Parakevas, M. and Mourjopoulos J. *"Speech enhancement using psychoacoustic criteria"* in Proc. IEEE ICASSP, Minneapolis, MN, Apr.1993, pp. 359-361.