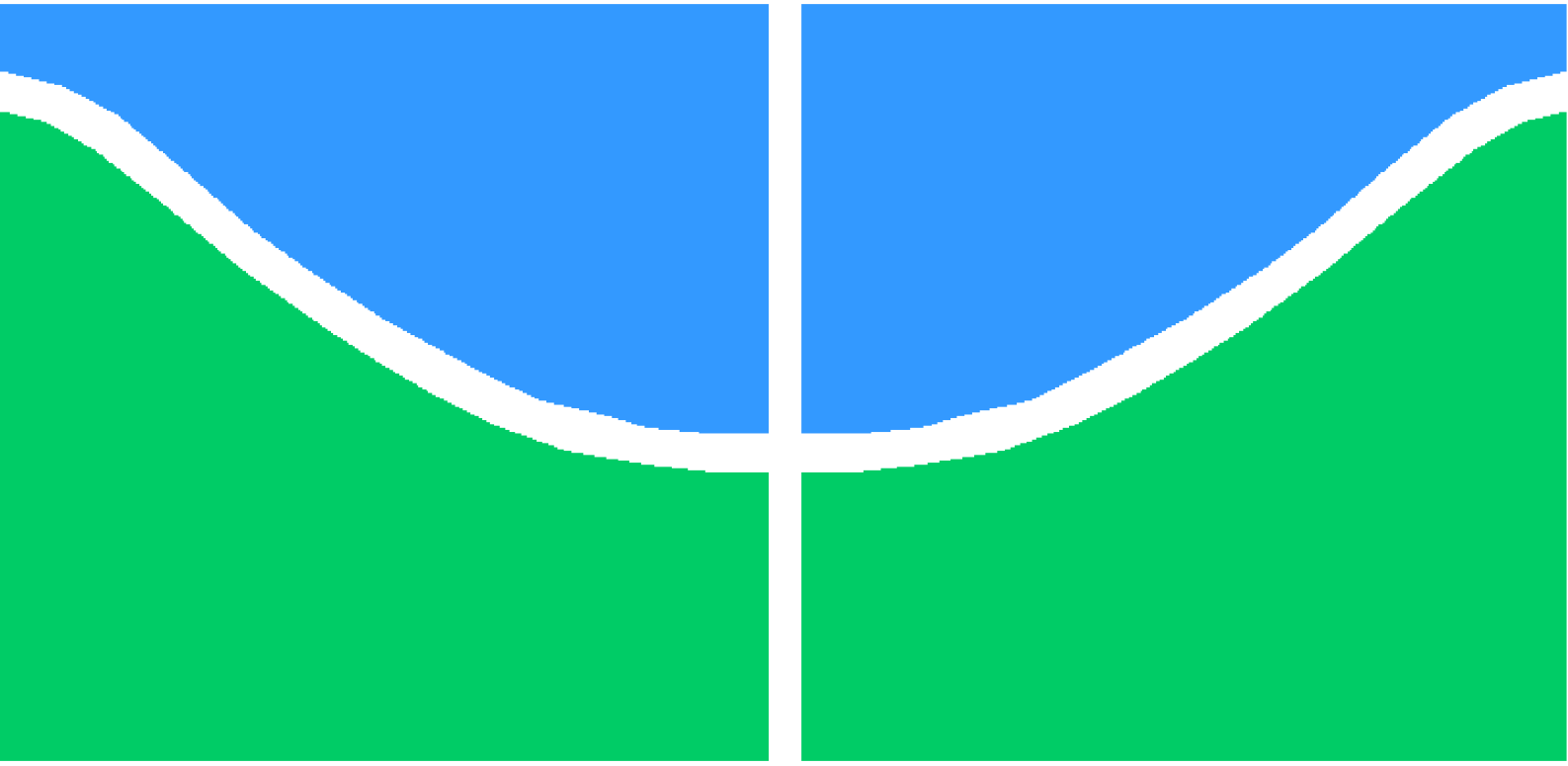


Universidade de Brasília - UnB

Faculdade UnB Gama - FGA

Engenharia Eletrônica

Sistema Computacional de Avaliação de Síntese Biaural para Sinais Sonoros de Fala com *Head Tracker*

Autor: Heitor Moraes Couto

Orientador: Dr. Marcelino Monteiro de Andrade

Brasília, DF

Dezembro de 2014

Heitor Moraes Couto

Sistema Computacional de Avaliação de Síntese Biaural para Sinais Sonoros de Fala com *Head Tracker*

Monografia submetida ao curso de graduação em (Engenharia Eletrônica) da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em (Engenharia Eletrônica).

Universidade de Brasília - UnB

Faculdade UnB Gama - FGA

Orientador: Dr. Marcelino Monteiro de Andrade

Coorientador: Dr. Márcio Henrique de Avelar Gomes

Brasília, DF

Dezembro de 2014

Heitor Moraes Couto

Sistema Computacional de Avaliação de Síntese Biaural para Sinais Sonoros de Fala com *Head Tracker*/ Heitor Moraes Couto. – Brasília, DF, Dezembro de 2014-

70 p. : il. (algumas color.) ; 30 cm.

Orientador: Dr. Marcelino Monteiro de Andrade

Trabalho de Conclusão de Curso – Universidade de Brasília - UnB
Faculdade UnB Gama - FGA , Dezembro de 2014.

1. Síntese Biaural. 2. HRTF. I. Dr. Marcelino Monteiro de Andrade. II. Universidade de Brasília. III. Faculdade UnB Gama. IV. Sistema Computacional de Avaliação de Síntese Biaural para Sinais Sonoros de Fala com *Head Tracker*

CDU 02:141:005.6

Heitor Moraes Couto

Sistema Computacional de Avaliação de Síntese Biaural para Sinais Sonoros de Fala com *Head Tracker*

Monografia submetida ao curso de graduação em (Engenharia Eletrônica) da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em (Engenharia Eletrônica).

Trabalho aprovado. Brasília, DF, Dezembro de 2014:

Dr. Marcelino Monteiro de Andrade
Orientador

Dr. Fernando William Cruz
Convidado 1

Dr. Henrique Gomes de Moura
Convidado 2

Brasília, DF
Dezembro de 2014

Agradecimentos

Agradeço aos professores Edson Júnior, Márcio Gomes e Marcelino Andrade pelas orientações no decorrer da minha graduação, que sem dúvida contribuíram para minha formação tanto profissional como pessoal.

Agradeço aos meus amigos e à minha família, sobretudo, aos meus pais, Mary Rose e José Rita, pelo apoio e motivação que me deram, desde a escolha da profissão Engenharia Eletrônica, passando pela gradução, até o presente momento.

Resumo

A utilização de áudio espacializado trás benefícios de inteligibilidade e identificação de falante em sistemas de telecomunicação. A auralização é o processo que espacializa áudio em 3D e é realizada através da síntese biaural, que consiste na convolução de uma fonte sonora com um par de HRTFs, gerando áudio biaural. O presente trabalho propõe o desenvolvimento de um sistema de avaliação de áudio biaural, sintetizado com fontes sonoras de fala e com bancos de HRTFs. Aplicações de síntese biaural foram implementadas, considerando-se apenas o plano horizontal. Um sistema de *head tracking* utilizando uma IMU também foi implementado. Foram desenvolvidas rotinas de testes de localização, para cenários com uma ou duas fontes sonoras, e com ou sem o uso de *head tracker*. Os testes de localização foram realizados com quatro sujeitos. Os resultados obtidos indicam que a síntese biaural confere aos sinais sonoros informações de diretividade suficientes para localização espacial no plano horizontal. Também indicam que a utilização do sistema de *head tracking* pode trazer benefícios para a localização de fontes sonoras em áudio biaural.

Palavras-chaves: Áudio Biaural. Auralização. HRTF. *Head Tracking*

Abstract

Spatialized audio brings benefits to intelligibility and to source localization in communication systems. Auralization is the process which spatializes audio in 3D and it is done by binaural synthesis, where a sound source is convolved with a pair of HRTFs to generate binaural audio. The present work proposes and develops a binaural audio evaluation system, where audio is synthesized with speech sound sources and with HRTF database. Binaural synthesis applications were implemented, considering only horizontal plane. A head tracking system using an IMU was also implemented. Localization tests routines for scenarios with one or two sound sources, and with our without use of head tracker, were developed. Tests were performed with four subjects. The results obtained indicate that binaural synthesis gives sufficient directional information to sound source localization in the horizontal plane. Results also indicate that use of a head tracking system can benefit binaural sound source localization.

Key-words: Binaural Audio. Auralization. HRTF. Head Tracking.

Lista de Figuras

Figura 1 – Fonógrafo de Thomas Edison 26
Figura 2 – Diferença entre audição binaural (a) e estereofonia (b) 27
Figura 3 – Padrão *surround* 5.1 27
Figura 4 – Manequim KEMAR 28
Figura 5 – Sistema de coordenadas relacionado às HRTFs 29
Figura 6 – Cone de confusão 30
Figura 7 – Exemplo de sistema de medição de HRTFs em câmara anecoica 31
Figura 8 – Exemplo de um par de HRIRs 32
Figura 9 – Par de HRTFs correspondente às HRIRs da Fig. (8) 32
Figura 10 – Princípio da auralização 33
Figura 11 – Auralização por meio da convolução 34
Figura 12 – Método de convolução *overlap-save*: segmentação da entrada e definição da saída 36
Figura 13 – Unidade inercial de medida *9DOF Razor Stick* da Sparkfun 37
Figura 14 – IMU baseada em três sensores. Adaptado de Ahmad et al. (2013) . . . 38
Figura 15 – Sistemas de coordenadas RPY 38
Figura 16 – Posições possíveis para posicionamento das fontes sonoras para teste de localização. 41
Figura 17 – Fluxograma do algoritmo de síntese biaural para uma fonte sonora . . 41
Figura 18 – Exemplo de cenário onde posição aparente da fonte sonora é alterada devido a movimentação de cabeça do ouvinte. Caixa verde indica posição da fonte sonora. Números em preto indicam posição real e números em azul indicam posição aparente. (a) Fonte sonora localizada na posição 3 (posição real); (b) Ouvinte com cabeça rotacionada 90° para a direita e fonte sonora posicionada na posição 1 (posição aparente) . . . 42
Figura 19 – Fluxograma do algoritmo de síntese biaural utilizando método de convolução por blocos *overlap-save* 43
Figura 20 – Unidade de aquisição e transmissão de sinais composta por arduino (caixa grande) e IMU (caixa pequena) 44
Figura 21 – *Head tracker*: fone de ouvido equipado com unidade inercial de medida 44
Figura 22 – Interface gráfica para realização dos testes de localização 47
Figura 23 – Botões de seleção de teste, botão de navegação e painel de informações da interface gráfica. (a) Teste 1 selecionado e botão de navegação no estado ativo “Avançar”. (b) Teste 2 selecionado e botão de navegação no estado ativo “Iniciar”. 48

Figura 24 – Etapas de navegação do Teste 1. (a) Seleção do Teste 1. (b) Descrição do treinamento do teste. (c) Indicação do fim do treinamento. (d) Descrição do procedimento do Teste 1. (e) Etapa de indicação da posição percebida por parte do ouvinte. (f) Indicação de conclusão do Teste 1 . 49
Figura 25 – Indicação visual da posição real da fonte sonora durante etapa de treinamento do Teste 1 50
Figura 26 – Procedimento para indicar posição percebida na interface. Depois de pressionar um botão, uma janela de confirmação é aberta 51
Figura 27 – Indicação da etapa de calibração do *head tracker* no Teste 3 51
Figura 28 – Indicação visual das posições reais das fontes sonoras durante etapa de treinamento do Teste 2 52
Figura 29 – Etapa do Teste 2 onde ouvinte indica posições percebidas das fontes sonoras. (a) Indicação da primeira fonte. (b) Indicação da segunda fonte. 53
Figura 30 – Indicação visual das posições reais das fontes sonoras durante etapa de treinamento do Teste 4. (a) Reprodução 1. (b) Reprodução 2. 54
Figura 31 – Histograma do erro de localização para o Teste 01 59
Figura 32 – Histograma do erro de localização para o Teste 03 60

Lista de Tabelas

Tabela 1 – Resultados obtidos com execução do Teste 01 para sujeito I 55
Tabela 2 – Resultados obtidos com execução do Teste 01 para sujeito II 56
Tabela 3 – Resultados obtidos com execução do Teste 01 para sujeito III 56
Tabela 4 – Resultados obtidos com execução do Teste 01 para sujeito IV 56
Tabela 5 – Resultados obtidos com execução do Teste 02 para sujeito I 56
Tabela 6 – Resultados obtidos com execução do Teste 02 para sujeito II 56
Tabela 7 – Resultados obtidos com execução do Teste 02 para sujeito III 57
Tabela 8 – Resultados obtidos com execução do Teste 02 para sujeito IV 57
Tabela 9 – Resultados obtidos com execução do Teste 03 para sujeito I 57
Tabela 10 – Resultados obtidos com execução do Teste 03 para sujeito II 57
Tabela 11 – Resultados obtidos com execução do Teste 03 para sujeito III 58
Tabela 12 – Resultados obtidos com execução do Teste 03 para sujeito IV 58
Tabela 13 – Resultados obtidos com execução do Teste 04 para sujeito I 58
Tabela 14 – Resultados obtidos com execução do Teste 04 para sujeito II 58
Tabela 15 – Resultados obtidos com execução do Teste 04 para sujeito III 58
Tabela 16 – Resultados obtidos com execução do Teste 04 para sujeito IV 59
Tabela 17 – Quantidade de acertos na percepção de localização por sujeito no Teste 02 60
Tabela 18 – Quantidade de acertos na percepção de localização por sujeito no Teste 04 60

Lista de abreviaturas e siglas

PC	*Personal Computer*
3D	Tridimensional
IP	*Internet Protocol*
VOIP	*Voice over IP*
HRTF	*Head Related Transfer Function*
HRIR	*Head Related Impulse Response*
IID	*interaural intensity difference*
ITD	*interaural time difference*
FIR	*Finite Impulse Response*
IMU	*Inertial Measurement Unit*
FFT	*Fast Fourier Transform*
DCM	*Direction Cosine Matrix*

Lista de símbolos

XVIII	Dezoito
XIX	Dezenove
XX	Vinte
θ	Ângulo azimutal
ϕ	Ângulo de elevação
p(t)	Saída da convolução de auralização
s(t)	Sinal de uma fonte sonora mono descrita no tempo.

Sumário

1 INTRODUÇÃO 21
1.1 Objetivos 23
1.2 Estrutura do trabalho 24

2 FUNDAMENTAÇÃO TEÓRICA 25
2.1 Som 25
2.2 Aspectos Históricos 25
2.3 Audição Binaural 29
2.4 HRTFs 30
2.5 Auralização 32
2.5.1 Síntese Biaural 33
2.5.2 Síntese Biaural em Tempo Real 34
2.5.2.1 Método de Convolução *overlap-save* 35
2.6 *Head Tracking* 36

3 MATERIAIS E MÉTODOS 39
3.1 Aparato Experimental 39
3.2 Metodologia 40
3.2.1 Desenvolvimento da Aplicação de Síntese Biaural 40
3.2.2 Desenvolvimento da Aplicação de *Head Tracking* 43
3.2.3 Desenvolvimento da Interface Gráfica 46
3.2.4 Procedimentos de Teste 47
3.2.4.1 Teste 1 e Teste 3 48
3.2.4.2 Teste 2 e Teste 4 52
3.3 Protocolo Experimental 53

4 RESULTADOS 55

5 DISCUSSÃO 63

6 CONCLUSÃO 65

Referências 67

1 Introdução

A qualidade de áudio desempenha papel importante nos sistemas de multimídia, como filmes, jogos e vídeo conferências. Uma qualidade de áudio ruim em um filme, por exemplo, pode gerar desconforto em quem assiste ao filme, não importando a qualidade do vídeo exibida.

Pesquisa realizada por Neuman, Crigler e Bove (1991) revela que um vídeo com maior qualidade de áudio é visto como sendo mais interessante e envolvente. Além disso, a melhora na qualidade de áudio provocou a sensação de que também houve uma melhora na qualidade gráfica do vídeo, mesmo esta última não tendo sido alterada.

Porém, mesmo com a importância destacada anteriormente, o desenvolvimento na área de áudio digital é mais novo e lento que o desenvolvimento na área de vídeos e imagens. Esse fato pode ser explicado pelo motivo de a visão ser considerada por muitos como o sentido de percepção mais relevante (BEGAULT, 2000). Um exemplo do avanço na área gráfica é a tecnologia de visualização em três dimensões, que até pouco tempo atrás estava presente apenas nos cinemas e hoje já se encontra em televisores e computadores comerciais (GOMES, 2012).

Com o avanço na área gráfica e também com o aumento de velocidade da internet, ferramentas de vídeo conferência para PC (*personal computer*) tornaram-se bastante populares. São utilizadas para atividades tais como reuniões de trabalho e encontros entre amigos e familiares (BALDIS, 2001).

Idealmente, um sistema de telecomunicação deve possibilitar aos seus usuários um meio de se comunicar que seja o mais natural possível, como se os usuários estivessem num mesmo ambiente (KANG; KIM, 1996). Entretanto, existem problemas relacionados ao áudio em teleconferências com mais de duas pessoas. Tais problemas podem comprometer a inteligibilidade das conversas entre os participantes. Dentre outros, destacam-se os seguintes problemas (YANKELOVICH et al., 2004):

- Alguns participantes não podem ser ouvidos;
- É difícil identificar quem está falando.

Kang e Kim (1996) mostram que aumento na qualidade de áudio diminui o esforço mental necessário para compreender o que se fala numa conferência. No mesmo sentido, Baldis (2001) mostra que a utilização de um sistema de áudio espacializado, ou seja, áudio bi (2D) ou tridimensional (3D), aumenta a inteligibilidade por parte de um participante, além de facilitar a identificação do falante.

Também buscando reduzir os problemas citados anteriormente, várias pesquisas buscam criar meios para conferência utilizando-se de áudio 3D, também chamado de áudio binaural. O trabalho de Kang e Kim (1996) propõe a criação de um sistema de teleconferência que utilize técnicas de auralização, buscando recriar virtualmente ambientes acústicos espaciais.

Rothbucher et al. (2011) também mostram o desenvolvimento de um sistema de teleconferência onde seus participantes são virtualmente posicionados em torno dos ouvintes utilizando técnicas de auralização. Este sistema ainda oferece suporte aos aparelhos convencionais de telefonia que são compatíveis com plataformas de voz sobre IP (VOIP). Desse modo, os participantes conectados à conferência por meio de software tem acesso ao áudio 3D e os que estão conectados por meio de um telefone comum recebem o áudio mono padrão do serviço.

Para reproduzir um ambiente sonoro 3D deve-se primeiramente entender como se ouve espacialmente. Tronco, ombros, cabeça e os ouvidos de um ouvinte interagem com uma onda sonora. Essa interação pode ser representada como um processo de filtragem linear do sinal sonoro, onde tal filtro é descrito pelas funções de transferência relacionadas à cabeça (HRTFs) (ROTHBUCHER et al., 2011), (KEYROUZ; DIEPOLD, 2007). Para cada posição no espaço 3D existe uma HRTF específica (CHANDA; PARK; KANG, 2006).

Desse modo, uma das técnicas para se sintetizar áudio binaural, conhecido como método biaural (FARIA, 2005), consiste em se filtrar um sinal sonoro com a HRTF correspondente a posição no espaço 3D em que se deseja virtualmente posicionar tal sinal (CHANDA; PARK; KANG, 2006). Percebe-se então que para sintetizar áudio 3D para várias posições, um banco de funções de transferência se faz necessário. Porém, o processo de obtenção dessas funções é demorado e necessita de equipamento especializado, fazendo com que apenas um arranjo de HRTFs para determinadas posições seja medido (KEYROUZ; DIEPOLD, 2006).

Outro aspecto que dificulta a utilização das HRTFs na síntese de áudio binaural é o fato de que cada pessoa possui uma característica física diferente, logo, as funções de transferência relacionadas à cabeça variam de pessoa para pessoa (CARTY; LAZZARINI, 2008). O trabalho realizado por Wenzel et al. (1993) revela que o uso de HRTFs não individualizadas na síntese de áudio 3D gera erros de percepção no que diz respeito à localização das fontes sonoras, principalmente quando se varia a localização de tais fontes no plano vertical. Resultado similar foi obtido por Hyder, Haun e Hoene (2010), que constatou ser mais fácil localizar a posição da fonte sonora quando esta está no plano horizontal, isto é, quando a fonte está na mesma altura que a cabeça do ouvinte.

Mesmo no plano horizontal ainda podem existir erros de percepção de localização, sendo mais frequente confundir se a fonte sonora está a frente ou se está atrás da cabeça. O uso de estímulos visuais pode ajudar a diminuir essa confusão, pois, desse modo, o

ouvinte consegue associar o som com a imagem. Por exemplo, se o ouvinte vê a imagem de um avião a sua frente, ele não irá perceber o som do avião como vindo de trás. Além dos estímulos visuais, um sistema que forneça *feedback* da posição da cabeça do ouvinte também ajuda na diminuição dos erros de percepção espacial. Esse *feedback* pode ser fornecido por um sistema de *head tracking* (FILIPANITS JR., 1994).

O papel do sistema de *head tracking* é monitorar a posição da cabeça do ouvinte para que, caso seja necessário, o sistema de síntese biaural atualize a posição relativa da fonte sonora. Numa situação em que uma fonte sonora está virtualmente posicionada a frente da cabeça do ouvinte, se ele vira a cabeça 90° para a direita, o sistema de síntese reposicionaria a fonte para a posição a esquerda do ouvinte.

Para o caso sistema de teleconferência com os participantes virtualmente posicionados no espaço acústico, o uso de um sistema de *head tracking* oferece a possibilidade de um participante se beneficiar do efeito *cocktail party*. Esse efeito corresponde a capacidade de um ouvinte focar atenção em um falante específico no meio de várias conversações simultâneas e sons do ambiente (ARONS, 1992).

Baseando-se no exposto anteriormente, propõe-se o desenvolvimento de um sistema para avaliar a qualidade de síntese de áudio biaural para fontes sonoras de fala com a utilização de diferentes bancos de HRTFs, por meio de testes subjetivos de localização. O sistema também avaliará a percepção de localização para um cenário com duas fontes sonoras de fala concorrentes, além de aferir se o uso de uma aplicação de *head tracking* traz benefícios para a inteligibilidade do áudio auralizado.

1.1 Objetivos

O objetivo deste trabalho é desenvolver um sistema de avaliação de qualidade e de inteligibilidade de áudios biaurais sintetizados com fontes sonoras de fala. Tal avaliação se dará pela verificação da percepção de localização das fontes sonoras em cenários com uma ou duas fontes, e com ou sem o uso de um sistema de *head tracking*. A auralzação se dará para posições apenas no plano horizontal, onde é mais fácil a localização. Para efeitos de simplificação, nenhum ambiente específico será considerado para o processo de auralização.

Os objetivos específicos do trabalho são:

- Desenvolver aplicação de síntese de áudio biaural para uma e para duas fontes sonoras;
- Desenvolver aplicação de *Head Tracking*;
- Desenvolver quatro rotinas de teste subjetivo de localização, sendo elas:

  - Teste de localização para o caso de áudio biaural sintetizado com uma fonte;
  - Teste de localização para o caso de áudio biaural sintetizado com duas fontes;
  - Teste de localização para o caso de uma fonte com sistema de *head tracking*;
  - Teste de localização para o caso de duas fontes com sistema de *head tracking*;

- Desenvolver uma interface gráfica para realização dos testes citados anteriormente.

1.2 Estrutura do trabalho

No capítulo 2 é feita uma abordagem teórica dos conceitos envolvidos no trabalho. Inicia-se com uma rápida descrição do que é o som. Depois se apresenta um breve histórico sobre áudio espacial na seção 2.2. Na seção 2.3 explica-se a audição biaural. Na seção 2.4 fala-se sobre as HRTFs e na seção 2.5 explica-se como se dá o processo de auralização, onde também se fala sobre síntese biaural em tempo real. Por fim, na seção 2.5 se comenta sobre sistema de *head tracking*.

O capítulo 3 se inicia com os materiais a serem utilizados no trabalho. Nas seções subsequentes é explicado o procedimento experimental realizado no trabalho, mostrando as etapas de desenvolvimento do sistema proposto e explicando como serão realizados os testes subjetivos de localização.

No capítulo 4 são mostrados os resultados obtidos com a realização dos testes propostos. No capítulo 5 é feita uma discussão dos resultados obtidos. Por fim, o capítulo 6 traz as conclusões obtidas com a realização do trabalho e as possibilidades de trabalhos futuros.

2 Fundamentação Teórica

2.1 Som

Som é um fenômeno ondulatório que resulta de variações da pressão, em torno da pressão atmosférica, no ar. Essas ondas sonoras se propagam longitudinalmente com velocidade de 344 m/s a 20 $°C$. Um processo que cause a propagação das ondas de pressão no ar é chamado de fonte sonora.

O número de oscilações por segundo do movimento vibratório do som define a frequência do mesmo, dada em *Hertz* (Hz), ou ciclos por segundo. Os seres humanos conseguem ouvir sons com frequência na faixa de 20 a 20 kHz.

Uma característica importante do som é a intensidade. A intensidade sonora define a quantidade de energia que uma onda sonora contém, o que se traduz em maior ou menor amplitude da onda (FERNANDES, 2005).

2.2 Aspectos Históricos

Os fundamentos teóricos relacionados à sistemas de áudio existem a muito tempo, porém, os conceitos necessários para criação física de tais sistemas foram estabelecidos somente no século XIX por nomes como Faraday, Henry, Ohm, Helmholtz e Lissajous.

Os trabalhos dos pesquisadores citados anteriormente levaram a invenção de um aparelho muito comum hoje em dia, o telefone. A invenção de Alexander Graham Bell em 1876 foi importante por estabelecer os princípios a respeito de transdutores de áudio, tanto para gravação como para reprodução, levando a evolução de microfones e auto-falantes (DAVIS, 2003).

Um ano mais tarde, o primeiro aparelho para gravar e reproduzir som foi inventado por Thomas Edison (BRUCK; GRUNDY; JOEL, 2013)(DAVIS, 2003). Mostrado na Fig. (1), o fonógrafo reproduzia som por apenas um alto-falante, sendo classificado então como um sistema de áudio monoaural, pois apresenta apenas uma fonte sonora (GOMES, 2012). Tal sistema oferece poucos elementos para percepção espacial (FARIA, 2005).

Na mesma época, mais especificamente no ano de 1881, Clement Ader conectou microfones espalhados pelo palco da Ópera de Paris à fones de ouvidos, um telefone em cada ouvido, espalhados por hotéis próximos à ópera. Desse modo, os ouvintes conseguiram perceber um efeito estéreo (BRUCK; GRUNDY; JOEL, 2013), ainda não propriamente descoberto, sendo o evento a primeira grande demonstração de áudio espacializado

Figura 1 – Fonógrafo de Thomas Edison

(DAVIS, 2003).

Esse efeito estéreo foi efetivamente explicado apenas em 1931 por Alan Blumlein, inventor Britânico que patenteou o estéreo (BRUCK; GRUNDY; JOEL, 2013)(DAVIS, 2003). Este sistema utiliza dois canais para reprodução de áudio. Com a utilização de tal sistema, pode-se criar uma "fonte sonora fantasma" entre os alto-falantes. Isso pode ser utilizado para ampliar a sensação de espacialização do som (FARIA, 2005).

Aqui vale explicar a diferença entre estereofonia e áudio binaural em relação à percepção espacial. A estereofonia se refere ao fato de os dois ouvidos detectarem um som similar vindo de direções diferentes, podendo conter certo atraso ou certa diferença de amplitude entre eles. Baseado na diferença de tempo que o sinal sonoro leva para chegar aos dois ouvidos, percebe-se a fonte sonora como estando numa direção (posição fantasma) no sentido do primeiro sinal a ser detectado. No caso do áudio binaural, um sinal sonoro gerado por uma fonte sonora chega aos ouvidos e dependendo de seu ângulo de incidência, haverá um atraso entre os dois ouvidos na detecção do sinal (RUMSEY; MCCORMICK, 2009). A Fig. (2) exemplifica a diferença explicada anteriormente.

Após sair de um canal para dois, os sistemas de áudio continuaram a evoluir, chegando aos sistemas multicanais, também conhecidos como sistemas de som envolvente (*surround*). O sistema *surround* expande o sistema monoaural e estéreo para duas ou três dimensões. Neste sistema, o ouvinte é envolvido pelo campo sonoro, o que permite criar ambientes de áudio mais realistas e complexos (FARIA, 2005).

Por criar um ambiente sonoro de maior imersão, o sistema *surround* passou a ser largamente utilizado em cinemas e *home theaters*, sendo padronizado pela norma ITU-R BS.775-1 (FARIA, 2005). A Fig. (3) mostra o padrão *surround* 5.1.

Até esse ponto, mostrou-se a evolução dos sistemas para espacialização de áudio, que criam ambientes sonoros envolventes e provocam uma sensação de espacialidade. No capítulo 1, também se falou sobre sistemas de auralização. Apesar de também ser

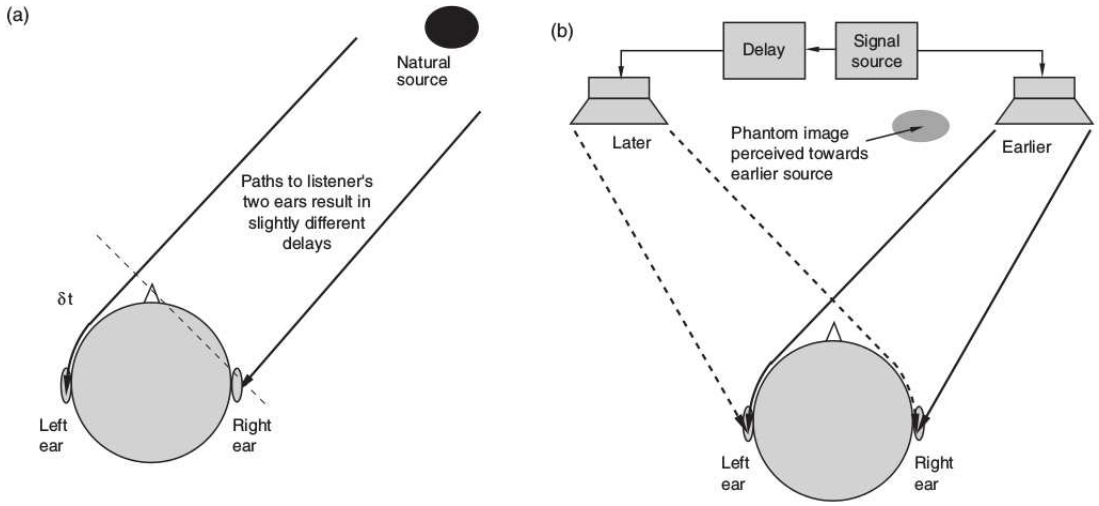


Figura 2 – Diferença entre audição binaural (a) e estereofonia (b) (RUMSEY; MCCORMICK, 2009)

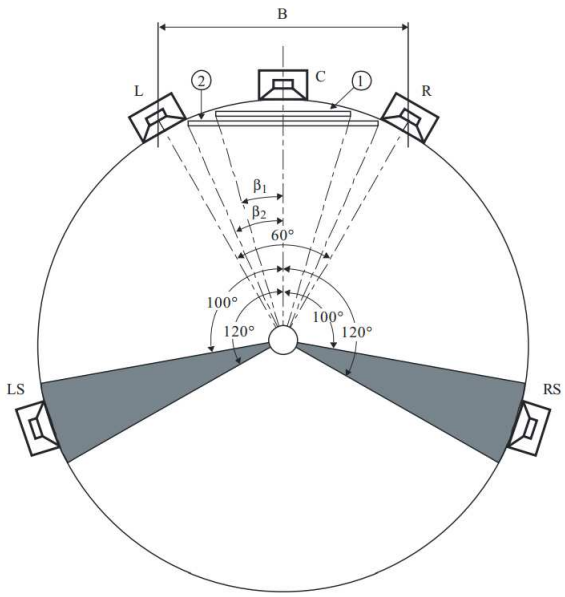


Figura 3 – Padrão *surround* 5.1 (ITU-R, 2012)

um sistema que trata de áudio 2D e 3D, os sistemas de auralização são utilizados para sintetizar ambientes sonoros mais fiéis à realidade, provendo percepção de diretividade e distância das fontes sonoras do ambiente simulado.

Os primeiros estudos que buscavam explicar o funcionamento da audição binaural são creditados a Wells e Venturi (final do século XVIII) (WADE; DEUTSCH, 2008). Segundo Paul (2009), no século XIX e no início do século XX, pesquisadores como Wheatstone, Steinhauser, Thompson e Lord Rayleigh também realizaram trabalhos na área, chegando a conclusão de que o fato de o ser humano possuir dois ouvidos, atuando como receptores de som, é determinante para localização e percepção de distância de fontes sonoras.

Baseando-se nesse conceito, cabeças artificiais com microfones nas posições dos ouvidos começaram a ser usadas para realizar gravações binaurais. Os primeiros estudos dessa área datam dos anos de 1930 (PAUL, 2009) e ao longo dos anos algumas cabeças artificiais se destacaram, como a *Neumann* (GENUIT; GIERLICH; BRAY, 1990) e a KEMAR (PAUL, 2009). A Fig. (4) mostra um manequim KEMAR.

Figura 4 – Manequim KEMAR (G.R.A.S. Sound & Vibration, 2006)

Um exemplo de aplicação de cabeças artificiais é o trabalho *Binaural Telephony*[1] (Telefonia Binaural), onde um manequim é posicionado numa mesa de reunião, simulando a presença da pessoa que participa da reunião por teleconferência. O objetivo é enviar à pessoa que o manequim representa o áudio que a forneça uma sensação de como se estivesse presente na sala da reunião.

Além de gravações binaurais, as cabeças artificiais também foram utilizadas para se medir HRTFs. O trabalho feito por Gardner e Martin (1994) apresenta medições de HRTFs realizadas num manequim KEMAR. Porém, como já dito na introdução deste trabalho, o uso de tais HRTFs na síntese biaural pode gerar erros de percepção de localização espacial. Por isso, a partir dos anos de 1990 novas estratégias para criação das cabeças artificiais passaram a ser utilizadas, seguindo uma tendência para individualização das HRTFs. O trabalho apresentado por Härmä et al. (2012) propõe uma personalização das funções de transferência por meio da seleção de uma HRTF com melhores resultados de localização.

[1] Trabalho realizado no instituto de pesquisa *Institute of Communication Systems and Data Processing* (IND), disponível em: <http://www.ind.rwth-aachen.de/en/research/speechaudio-communication/binaural-telephony/>

2.3 Audição Binaural

Os seres humanos conseguem localizar fontes sonoras no espaço devido a interação que uma onda sonora incidente de certa direção tem com tronco, ombros, cabeça e ouvidos. Essa interação se dá com a onda sendo difratada e refletida nessas partes do corpo humano. Logo, a onda sonora é distorcida linearmente, sendo essas distorções dependentes da direção de propagação da onda (VORLNDER, 2007).

Essa localização espacial de fontes sonoras leva em consideração três parâmetros (GOMES, 2012):

- Azimute: ângulo θ horizontal entre a fonte sonora e o centro da cabeça, medido no sentido horário;
- Elevação: ângulo ϕ vertical entre a fonte sonora e o centro da cabeça, onde valores positivos indicam posição acima do plano horizontal e valores negativos indicam posição abaixo do mesmo plano;
- Distância: distância entre a fonte sonora e a cabeça.

A Figura (5) mostra o sistema de coordenadas relacionado aos parâmetros citados acima. O ângulo azimutal θ varia de $0\,^\circ$ (a posição à frente da cabeça) até $360\,^\circ$, sendo $90\,^\circ$ a posição à direita da cabeça e $270\,^\circ$ a posição à esquerda da cabeça. E o ângulo de elevação ϕ varia de $-90\,^\circ$ (posição sob à cabeça) até $90\,^\circ$ (posição sobre a cabeça), onde $0\,^\circ$ corresponde ao plano horizontal mostrado na figura.

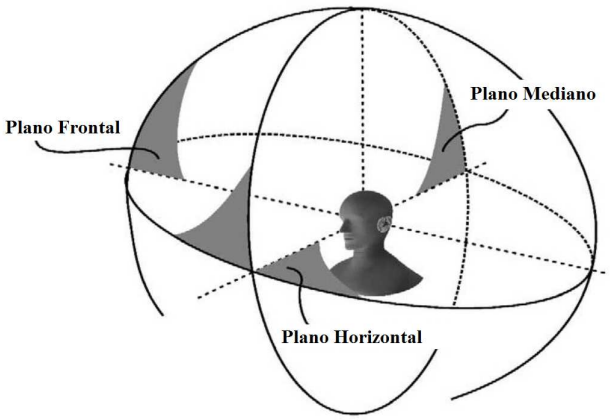


Figura 5 – Sistema de coordenadas relacionado às HRTFs (VORLNDER, 2007)

Como mencionado na seção anterior, o sinal de uma fonte sonora chega aos ouvidos com uma diferença temporal, pois a onda sonora pode levar mais tempo para atingir um ouvido do que o outro, dada a posição da fonte sonora. Também chega com uma diferença de amplitude, pois em altas frequências a cabeça pode atuar como uma barreira

para o som, gerando essa diferença de amplitude (RUMSEY, 2001). Tais diferenças são dependentes da posição da fonte sonora.

A diferença de tempo de chegada de uma onda sonora nos dois ouvidos é chamada de diferença de tempo interaural, do inglês *interaural time difference* (ITD). Já a diferença de nível de intensidade sonora, diferença de amplitude, nos ouvidos é chamada de *interaural intensity difference* (IID) (CHENG; WAKEFIELD, 2001).

Esses dois parâmetros, ITD e IID, conseguem fornecer informações de localização no plano horizontal. Porém, existe uma região na qual diferentes pontos no espaço podem gerar valores iguais de ITD e IID, o que geraria erros de percepção de localização. Essa região é chamada de cone de confusão. A Fig. (6) mostra um cone de confusão, onde a posição da fonte A pode ser confundida com a posição da fonte B e vice versa, o mesmo valendo para as posições C e D. Movimentos de cabeça podem minimizar o problema causado pelo cone de confusão.

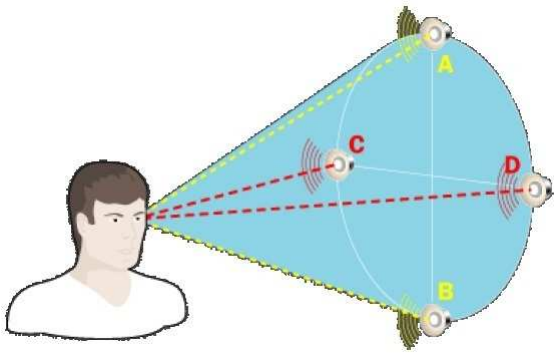


Figura 6 – Cone de confusão (WILSON, 2007)

Informações de localização a respeito da distância da fonte sonora à cabeça do ouvinte são fornecidas pelo nível de intensidade sonora, que é proporcional ao inverso do quadrado da distância, e também pela atenuação das componentes de alta frequência. E a elevação pode ser aferida a partir das alterações no espectro do sinal causadas pelas interações da onda sonora com o dorso do ouvinte (GOMES, 2012).

2.4 HRTFs

Ao incidir sobre um ouvinte, uma onda sonora proveniente de qualquer direção sofre efeitos de difração e reflexão provocados pela cabeça, ombros, tronco e ouvidos. Tais efeitos podem ser descritos por um filtro, chamado de função de transferência relacionada à cabeça, do inglês *head-related transfer function* (HRTF) (RUMSEY, 2001). Como as características antropométricas de cada indivíduo são diferentes, cada um e possui uma HRTF para cada posição de fonte sonora no espaço 3D (CHENG; WAKEFIELD, 2001). Cada HRTF na verdade é composta por um par de funções de transferência, uma para o ouvido esquerdo e outra para o ouvido direito.

Geralmente, HRTFs são medidas de pessoas ou manequins. O processo, usualmente realizado em câmaras anecoicas, consiste em colocar microfones na entrada do canal auditivo e medir as respostas ao impulso de estímulos sonoros reproduzidos por alto-falantes posicionados em torno do ouvinte (CHENG; WAKEFIELD, 2001). Os estímulos são gerados a partir de posições pré-definidas, gerando assim um banco de funções de transferência para as posições utilizadas no processo de medição. A Figura (7) exemplifica um sistema de medição de HRTFs em câmara anecoica. Como se pode ver na figura, os alto-falantes são posicionados em volta do ouvinte, o que garante a realização de medições para diferentes posições.

Figura 7 – Exemplo de sistema de medição de HRTFs em câmara anecoica (MASIERO, 2012)

Os trabalhos de Gardner e Martin (1994) e de Warusfel (2002) realizaram medições de HRTFs com o uso de manequins. Os bancos de funções de transferências obtidos são disponibilizados na versão temporal das HRTFs, as HRIRs (*head related impulse responses*), que podem ser utilizadas como um filtro FIR.

A Figura (8) mostra um par de HRIRs para uma fonte posicionada a 45°. Na figura, pode-se perceber os parâmetros ITD e IID. O som chega primeiro ao ouvido direito (HRIR em vermelho) e com menos intensidade (menor amplitude) ao ouvido esquerdo (HRIR em azul). Já a Figura (9) mostra o par de HRTFs correspondente para a mesma posição. Nesta última figura, também é possível notar a diferença de intensidade com que o som chega aos ouvidos esquerdo e direito.

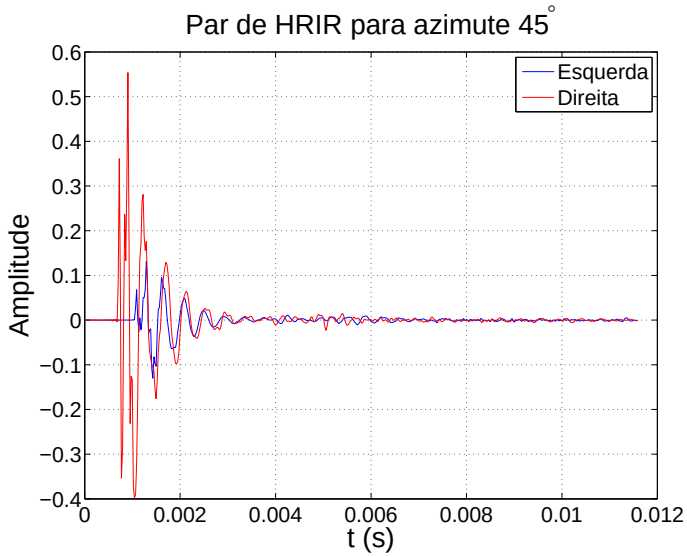


Figura 8 – Exemplo de um par de HRIRs

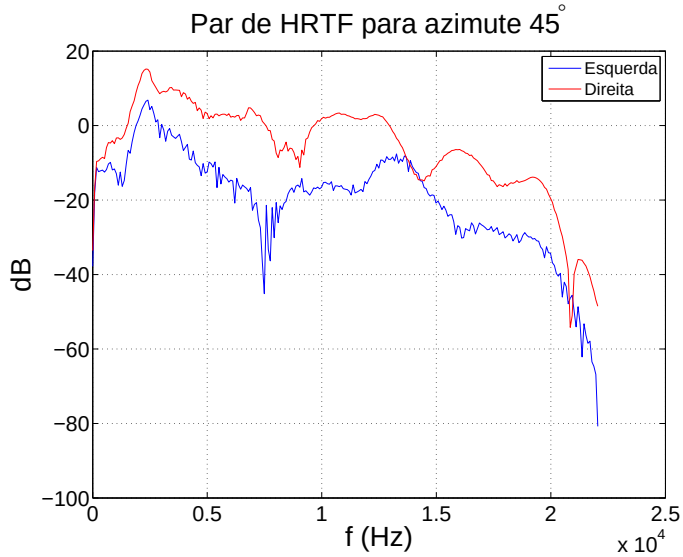


Figura 9 – Par de HRTFs correspondente às HRIRs da Fig. (8)

2.5 Auralização

Auralização é uma palavra utilizada com mesmo sentido que a palavra visualização, com a diferença de que a última se refere a visão e a primeira a audição (KLEINER; DALENBäCK; SVENSSON, 1991).

O processo da auralização consiste em criar arquivos de áudio a partir de dados numéricos, sejam eles simulados, medidos ou sintetizados (VORLNDER, 2007). A Figura (10) ilustra tal processo.

O processo se inicia com a descrição da fonte sonora. Um sinal sonoro é gravado ou criado, estando disponível em escala de amplitude, por exemplo. Então, tal sinal alimenta um caminho de transmissão, representado como uma função de transferência mensurada

ou simulada, que pode ser tratada como um filtro. O resultado dessa transmissão é um sinal perceptível e pronto para reprodução.

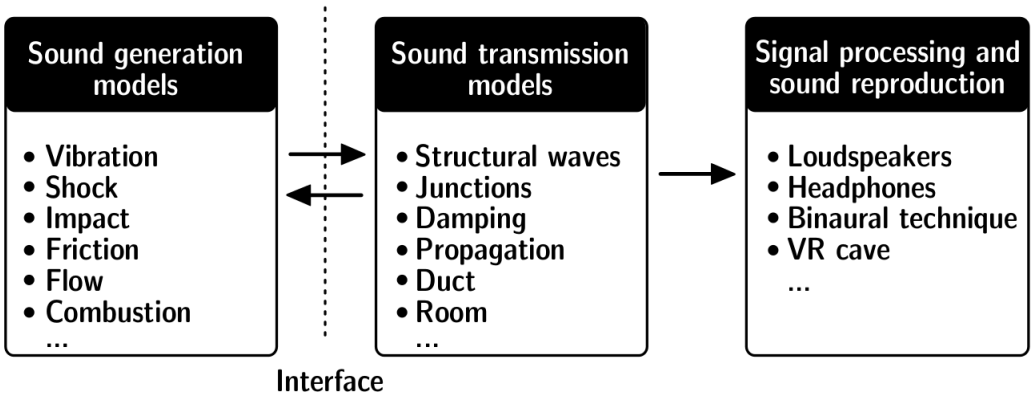


Figura 10 – Princípio da auralização (VORLNDER, 2007)

A auralização é dividida em duas categorias de síntese de áudio: biaural e multicanal. A síntese multicanal, que tem o objetivo de sintetizar áudio para matrizes de alto falantes e é frequentemente utilizada para simulação de ambientes acústicos para mais de uma pessoa (GOMES, 2012), não será abordada nesse trabalho.

2.5.1 Síntese Biaural

O objetivo da síntese biaural é pegar uma fonte sonora sem indicadores de diretividade e posicioná-la virtualmente no espaço 3D. Basicamente, consiste em se realizar a convolução do sinal de áudio com um par de HRTFs, uma para cada ouvido, e reproduzir a saída da convolução por meio de um fone de ouvido, como mostra a Fig. (11).

Como já citado anteriormente, as HRTFs descrevem o processo de filtragem que um sinal sonoro sofre ao incidir sobre o dorso de uma pessoa. Então, para sintetizar áudio biaural se filtra o sinal de uma fonte sonora por um filtro descrito por um par de HRTFs, processo que consiste na convolução mostrada na Fig. (11).

Além de se virtualizar a posição da fonte sonora, outro fator que se pode simular é o ambiente. Um ambiente também pode ser descrito por uma função de transferência. No caso do processo mostrado na Fig. (11), o ambiente simulado é o mesmo que o ambiente onde as HRTFs foram mensuradas, geralmente um ambiente anecoico. Caso se deseje virtualizar a posição de uma fonte sonora numa sala de reuniões, por exemplo, além de se convoluir o sinal da fonte sonora com um par de HRTFs, deve-se convoluir o sinal com a função de transferência que modele acusticamente a sala de reuniões.

A convolução pode ser realizada no domínio do tempo ou no domínio da frequência. No domínio do tempo é aplicada a convolução direta com a utilização de um filtro FIR. No caso da síntese biaural, o filtro FIR seria uma HRIR. Já no domínio da frequência, a convolução é realizada através da multiplicação do sinal sonoro, transformado para o

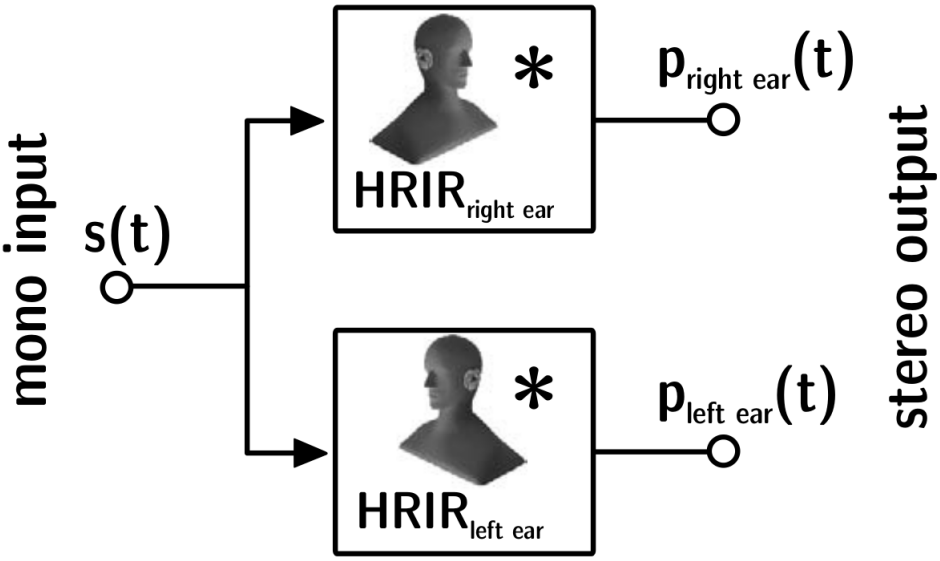


Figura 11 – Auralização por meio da convolução (VORLNDER, 2007)

domínio da frequência, pela a HRTF. O sinal sonoro é transformado para o domínio da frequência através da Transformada de Fourier. Computacionalmente, tal transformada é realizada pelo algoritmo da FFT, a Transformada Rápida de Fourier (BRIGHAM, 1974). Neste trabalho, utilizou-se a convolução no domínio do tempo.

Com um banco de HRTFs, como um dos citados anteriormente, é possível virtualizar a posição de uma fonte sonora mono nas direções que as funções foram mensuradas, realizando para isso a convolução da fonte sonora com um par de HRTFS. Essa operação é descrita pelas Eq. (2.1) e (2.2). E para sintetizar áudio binaural com diferentes fontes para diferentes posições, basta aplicar para cada fonte as Eq. (2.1) e (2.2) com as HRTFs respectivas a cada posição desejada e, então, mixar a saída da convolução de cada fonte em apenas uma saída estéreo.

$$p_{ouvidoesq}(t) = s(t) * HRTF_{ouvidoesq} \tag{2.1}$$

$$p_{ouvidodir}(t) = s(t) * HRTF_{ouvidodir} \tag{2.2}$$

2.5.2 Síntese Biaural em Tempo Real

O processo de síntese biaural mostrado na Fig. (11) e nas Eq. (2.1) e (2.2) simula o posicionamento de uma fonte sonora no espaço 3D estaticamente. Para simular tal posicionamento dinamicamente, ou seja, movimentar a fonte sonora ao redor do ouvinte, outras técnicas de convolução devem ser utilizadas. Uma dessas técnicas é o método *overlap-save*, um método de convolução por blocos.

Além disso, na síntese biaural em tempo real se considera o sinal de áudio contínuo no tempo. Desse modo, processar o sinal de uma vez não é possível, pois em tempo real não se pode esperar que todo o sinal de entrada seja amostrado e enviado para o canal de saída. Então, o sinal deve ser fragmentado em segmentos com duração de tempo iguais e processado bloco a bloco. Os resultados do processamento devem ser enviados sequencialmente à saída.

Processar o sinal bloco a bloco nos permite alterar o filtro para cada bloco. Assim é possível simular o posicionamento da fonte sonora dinamicamente ao redor do ouvinte. Para isso, é preciso utilizar o método de convolução em bloco citado anteriormente, o método *overlap-save*.

2.5.2.1 Método de Convolução *overlap-save*

O método de convolução por blocos *overlap-save* corresponde a realizar a convolução circular de tamanho L entre um bloco de sinal de tamanho L e uma resposta ao impulso de tamanho P, e identificar a saída correspondente à convolução linear. As saídas da convolução circular são então concatenadas para formar o sinal de saída. No caso da síntese binaural em tempo real, os segmentos de saída são sequencialmente enviados ao canal de saída.

A saída da convolução circular de um bloco de sinal $x_r[n]$, de tamanho L, com a resposta ao impulso $h[n]$, de tamanho P, onde $P < L$, tem seus primeiros $P-1$ elementos incorretos. O restante do resultado corresponde à saída da convolução linear para os dois sinais. Logo, um sinal $x[n]$ pode ser dividido em blocos de tamanho L, sendo que cada bloco sobrepõe o anterior em $P-1$ pontos. Essa sobreposição dá nome ao método, onde cada segmento de $x[n]$ consiste de $L-P+1$ novos pontos e $P-1$ pontos do segmento anterior. A Equação (2.3) mostra como se define cada segmento (OPPENHEIM; SCHAFFER; BUCK, 1998).

$$x_r[n] = x[n + r(L - P + 1) - P + 1], \qquad 0 \leq n \leq L - 1 \tag{2.3}$$

Seja $y_{rp}[n]$ a saída da convolução circular de cada bloco, define-se a saída final $y[n]$ da convolução por

$$y[n] = \sum_{r=0}^{\infty} y_r[n - r(L - P + 1) + P - 1], \tag{2.4}$$

com $y_r[n]$ definida por

$$y_r[n] = \begin{cases} y_{rp}[n], & P - 1 \leq n \leq L - 1 \\ 0, & \text{caso contrário} \end{cases} \tag{2.5}$$

A Figura (12) ilustra a segmentação do sinal de entrada definida pela Eq. (2.3). Também mostra como o sinal de saída $y[n]$ é definido a partir das saídas $y_{rp}[n]$.

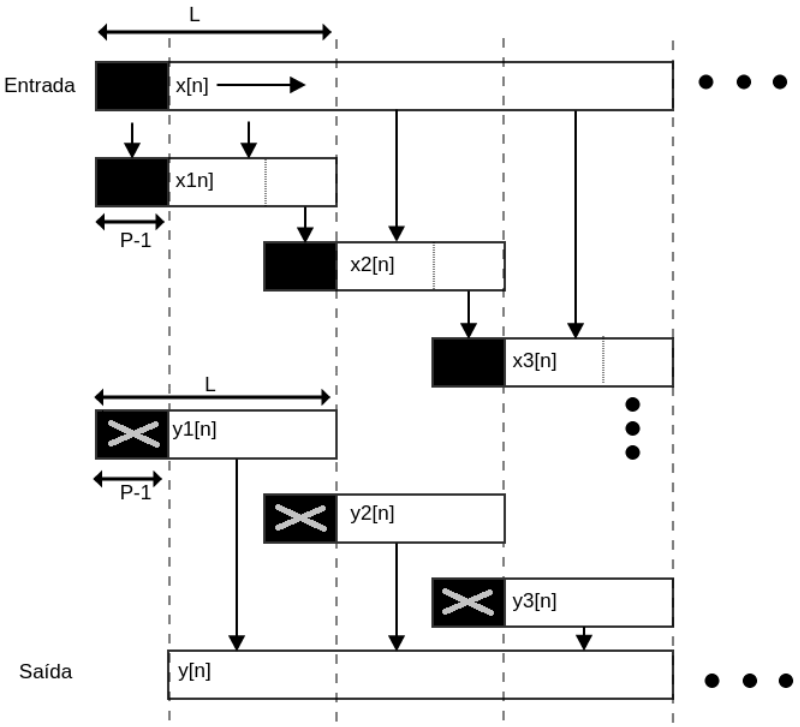


Figura 12 – Método de convolução *overlap-save*: segmentação da entrada e definição da saída

2.6 *Head Tracking*

O objetivo de um sistema de *head tracking* é monitorar a posição e os movimentos da cabeça de um sujeito. Aplicações de visão computacional, como reconhecimento facial e análise de expressão facial, utilizam sistemas de *head tracking* (CASCIA; SCLAROFF; ATHITSOS, 2000). Esses sistemas também são utilizados na área de realidade virtual (DEERING, 1992) e na área de jogos com detecção de movimentos (WANG et al., 2006).

Um sistema de *head tracking* pode ser desenvolvido por meio de uso de camêras e técnicas de processamento de imagens, onde o monitoramento é feito pelo processamento do vídeo que a câmera capta. Outra opção, a que será utilizada neste trabalho, é a utilização de unidades inerciais de medida, ou IMU, da sigla em inglês *inertial measurement unit*.

O uso de IMU para aplicações de *head tracking* se deve a evolução da tecnologia *Microelectromechanical systemas* (MEMS[2]) para construção de sensores inerciais. Essa

[2] MEMS são dispositivos compostos por partes móveis com tamanhos na faixa de μm a mm e que são produzidos por meio de processos de fotolitografia (PERLMUTTER; ROBIN, 2012)

evolução permitiu a diminuição do tamanho dos dispositivos inerciais, possibilitando a utilização de IMUs em diversas aplicações, como por exemplo celulares, brinquedos e armas (PERLMUTTER; ROBIN, 2012).

Segundo Ahmad et al. (2013), uma IMU é geralmente uitilizada para medir orientação, velocidade e força gravitacional. Existem dois tipos de IMU, as que possuem dois sensores e as que possuem três sensores. As que possuem dois sensores são compostas por acelerômetros, que medem aceleração inercial, e giroscópios, que medem velocidade angular. O terceiro sensor que compõe o segundo tipo de IMU é o magnetômetro, que mede a direção e magnitude de campos magnéticos, podendo funcionar como uma bússola.

A Figura (13) mostra uma IMU de três sensores com nove graus de liberdade, a *9DOF Razor Sticky* da Sparkfun[3], utilizada neste trabalho.

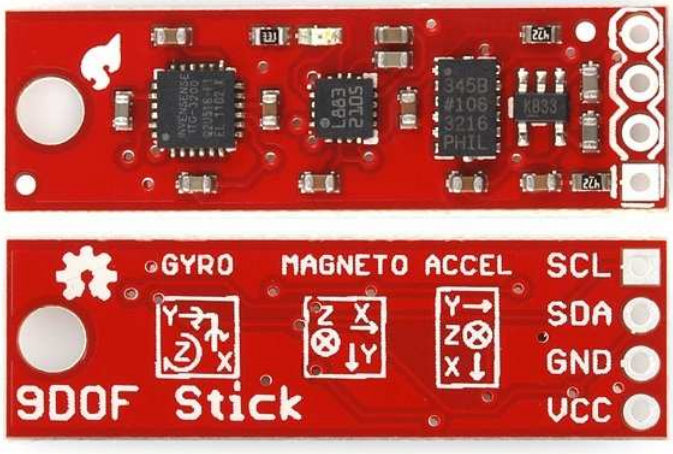


Figura 13 – Unidade inercial de medida *9DOF Razor Stick* da Sparkfun

A Figura (14) mostra o funcionamento simplificado de uma IMU de três sensores. Os três sensores geralmente possuem três graus de liberdade, definidos para os eixos x, y e z de coordenadas cartesianas. Isso totaliza nove graus de liberdade para a IMU. O bloco de fusão dos sensores mostrado na figura consiste na combinação dos dados dos sensores para obter medições mais precisas e corrigir erros de deriva do giroscópio (AHMAD et al., 2013).

[3] Disponível em: <https://www.sparkfun.com/products/10724>

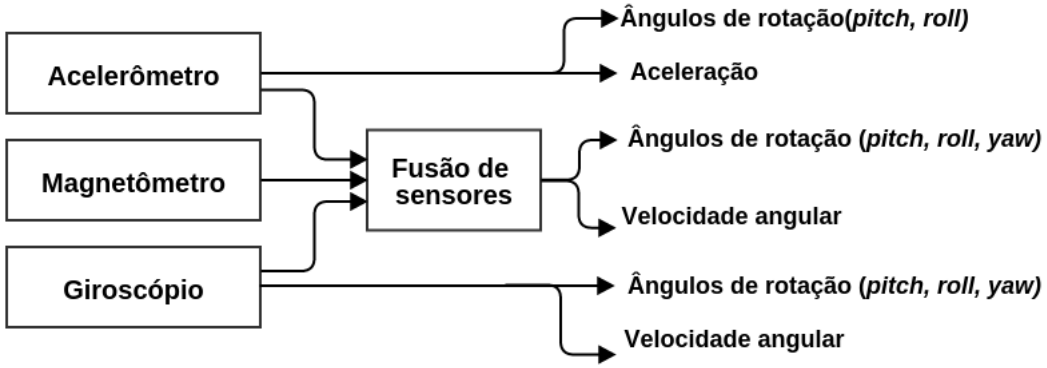


Figura 14 – IMU baseada em três sensores. Adaptado de Ahmad et al. (2013)

Como se pode ver na Fig. (14), a IMU mede tanto a aceleração e velocidade angular quanto os ângulos de rotação *yaw*, *roll* e *pitch*. Tais ângulos representam o sistema de coordenadas RPY(*Roll, Pitch, Yaw*) e descrevem a rotação no eixo z, no eixo x e no eixo y respectivamente (SANTANA, 2005). A Figura (15) ilustra o sistema de coordenadas RPY, considerando-se os movimentos da cabeça de uma pessoa.

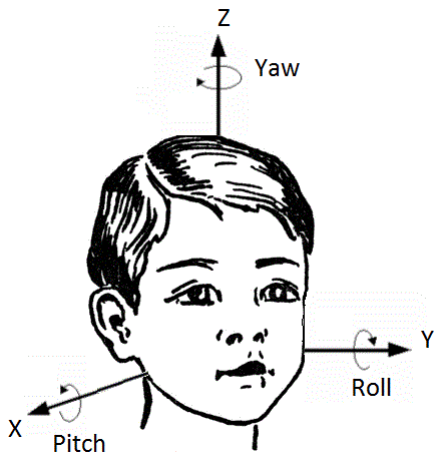


Figura 15 – Sistemas de coordenadas RPY

3 Materiais e Métodos

Este capítulo se inicia mostrando os materiais utilizados na realização do trabalho. A seguir são explicadas as etapas de desenvolvimento do sistema, que inclui a aplicação de síntese biaural, a aplicação de *head tracking*, a interface gráfica para realização dos testes propostos pelo trabalho.

3.1 Aparato Experimental

Para a realização do trabalho, os seguintes materiais foram utilizados:

- Matlab 7.14;
- Bancos de funções de transferência relacionadas à cabeça;
- Banco de fontes sonoras de fala;
- Unidade de medida inercial para aplicação de *head tracking*;
- Arduino Funduino Pro Mini também para aplicação de *head tracking*;
- Fone de ouvido.

Os bancos de HRTFs são necessários para a síntese biaural das fontes sonoras em diferentes posições. Diferentes bancos foram utilizados com o objetivo de possibilitar a seleção pelo ouvinte de diferentes HTRFs, buscando uma função que gere os melhores resultados. Os bancos selecionados foram os disponibilizados por Gardner e Martin (1994) e por Warusfel (2002). Tais bancos disponibilizam as funções de transferência no formato temporal, ou seja, HRIRs. Portanto, o processo de auralização será realizado por convolução direta do filtro FIR (HRIR) com o sinal de áudio de uma fonte sonora.

As fontes sonoras selecionadas para realização dos testes de localização são sinais de teste para sistemas telefônicos disponibilizadas pela International Telecomunication Union (ITU) (ITU-T, 1998). Esses sinais consistem em sinais de fala reais e são disponibilizados no formato *wave*. O formato *wave* é um formato de arquivo de áudio da Microsoft que contém o sinal de áudio amostrado e informações do sinal, tais como frequência de amostragem, número de bits por amostra e quantidade de canais (RUMSEY; MCCORMICK, 2009).

Tanto as funções de transferência relacionadas à cabeça como as fontes sonoras selecionadas possuem frequência de amostragem de 44,1 kHz, frequência amplamente

utilizada em sistemas de áudio (HAVELOCK; KUWANO; VORLAENDER, 2008). Com essa frequência de amostragem, considera-se para o processamento do áudio toda a faixa de frequências de som audível.

A síntese biaural foi desenvolvida, em MATLAB, para reprodução em fones de ouvido. O fone de ouvido foi equipado com uma unidade de medida inercial para o desenvolvimento do sistema de *head tracking*. E o arduino também faz parte da solução em *hardware* do sistema de monitoramento de posição da cabeça.

3.2 Metodologia

Nessa seção serão mostradas as etapas de desenvolvimento da aplicação de síntese biaural e do sistema de *head tracking*. Na sequência, será abordado o desenvolvimento da interface gráfica, onde serão mostrados os recursos que a mesma possui. E por fim, os testes de localização serão explicados, mostrando o funcionamento completo da interface.

3.2.1 Desenvolvimento da Aplicação de Síntese Biaural

Dois métodos de síntese biaural foram implementadas, um para ser utilizado nos testes sem o sistema de monitoramento da posição da cabeça e o outro método de síntese para ser utilizado em conjunto com o *head tracker*.

Para os testes sem o sistema de *head tracking*, a síntese biaural é realizada estaticamente, ou seja, as fontes sonoras são posicionadas em posições pré-definidas e não se movem em torno do ouvinte. Neste trabalho, são oito as posições possíveis para se localizar as fontes sonoras, todas no plano horizontal à cabeça do ouvinte. A Figura (16) ilustra as posições em torno da cabeça do ouvinte. As posições variam de 1 a 8 no sentido horário, ou de 0° a 315°, espaçadas de 45° entre si.

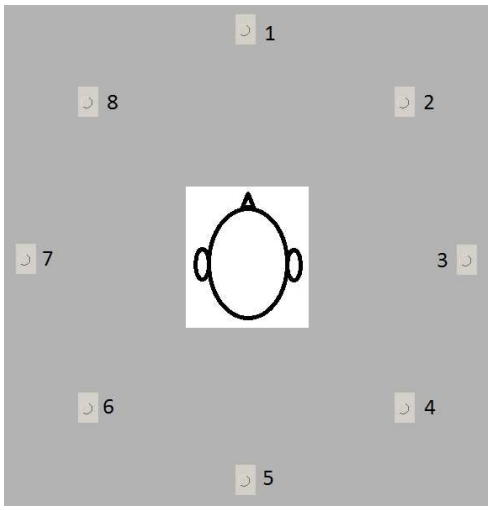


Figura 16 – Posições possíveis para posicionamento das fontes sonoras para teste de localização.

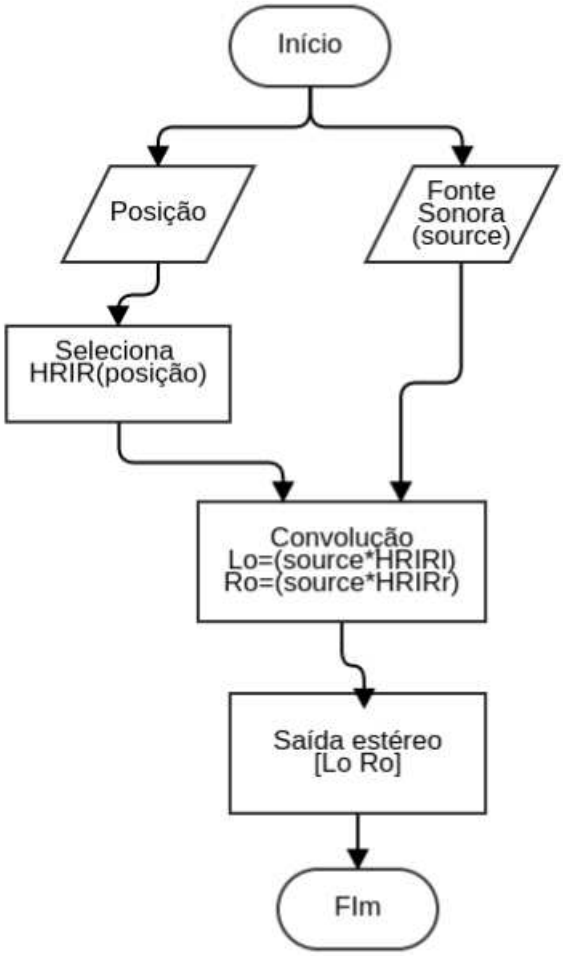
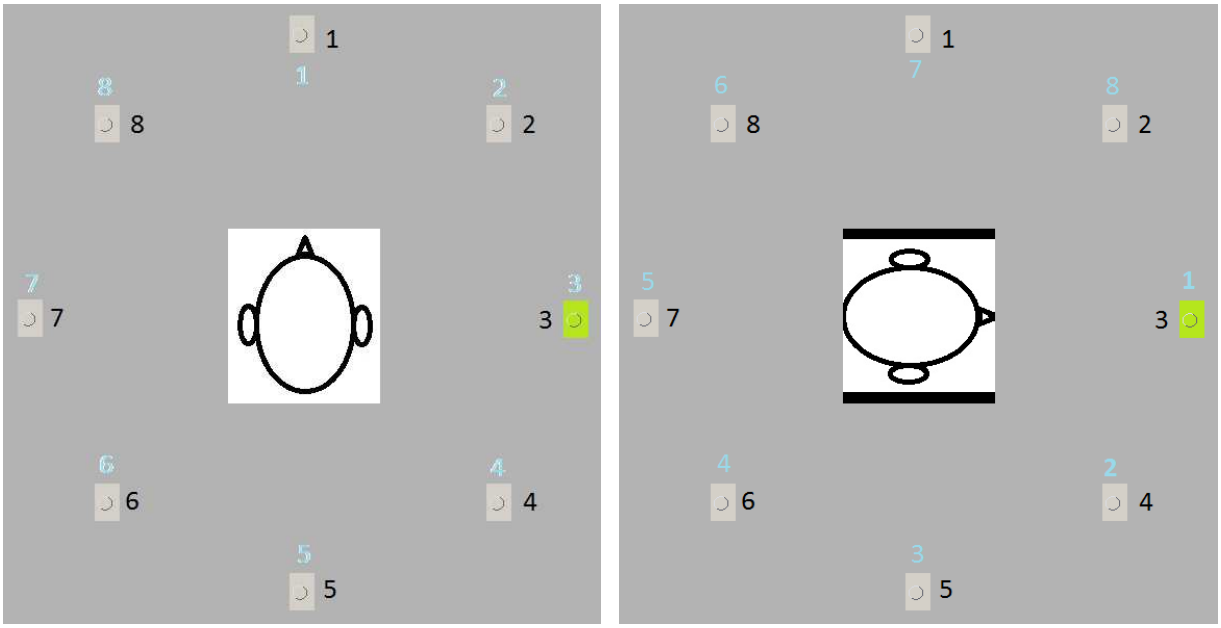


Figura 17 – Fluxograma do algoritmo de síntese biaural para uma fonte sonora

O segundo método de síntese biaural implementado utiliza o método *overlap-save* de convolução. Pretendeu-se com isso simular a reprodução em tempo real do áudio bi-

naural, por causa da utilização do sistema de *head tracking*. Nesse cenário, a fonte sonora é posicionada numa determinada posição em volta do ouvinte. Com a informação da posição da cabeça do ouvinte fornecida pela IMU, o sistema deve rearranjar o espaço sonoro quando o ouvinte mover a cabeça, isto é, a posição aparente da fonte sonora em relação ao ouvinte deve ser alterada dinamicamente.

A Figura (18) exemplifica essa situação. Na figura, os números em preto indicam a posição real da fonte e os números em azul indicam a posição aparente. A fonte sonora, indicada pela caixa verde na Fig. (18a), é posta na posição 3, considerada posição real da fonte. O ouvinte então percebe a fonte sonora a sua direita. Na Figura (18b), o ouvinte girou a cabeça 90° para a direita. Agora, o ouvinte deve perceber a fonte sonora a sua frente. Para isso, a aplicação de síntese biaural em tempo real deve atualizar a posição da fonte para a posição 1 (posição aparente).



(a) (b)

Figura 18 – Exemplo de cenário onde posição aparente da fonte sonora é alterada devido a movimentação de cabeça do ouvinte. Caixa verde indica posição da fonte sonora. Números em preto indicam posição real e números em azul indicam posição aparente. (a) Fonte sonora localizada na posição 3 (posição real); (b) Ouvinte com cabeça rotacionada 90° para a direita e fonte sonora posicionada na posição 1 (posição aparente)

O fluxograma do algoritmo implementado para a síntese biaural em tempo real utilizando o método de convolução *overlap-save* é mostrado na Fig. (19). O sinal de entrada é segmentado em n blocos. Ao menos o primeiro bloco é processado considerando-se a posição real onde a fonte sonora é posta. Cada bloco subsequente é processado de acordo com determinada posição aparente, que pode ser a real ou não, conforme a movimentação da cabeça do ouvinte. A convolução circular entre o n-ésimo bloco e sua respectiva posição

é então realizada. A saída (saída_n na figura) é definida como sendo os pontos da saída da convolução circular equivalentes à uma convolução linear. Na aplicação implementada, cada saída_n é enviada sequencialmente à placa de áudio para reprodução, simulando uma reprodução de áudio em tempo real.

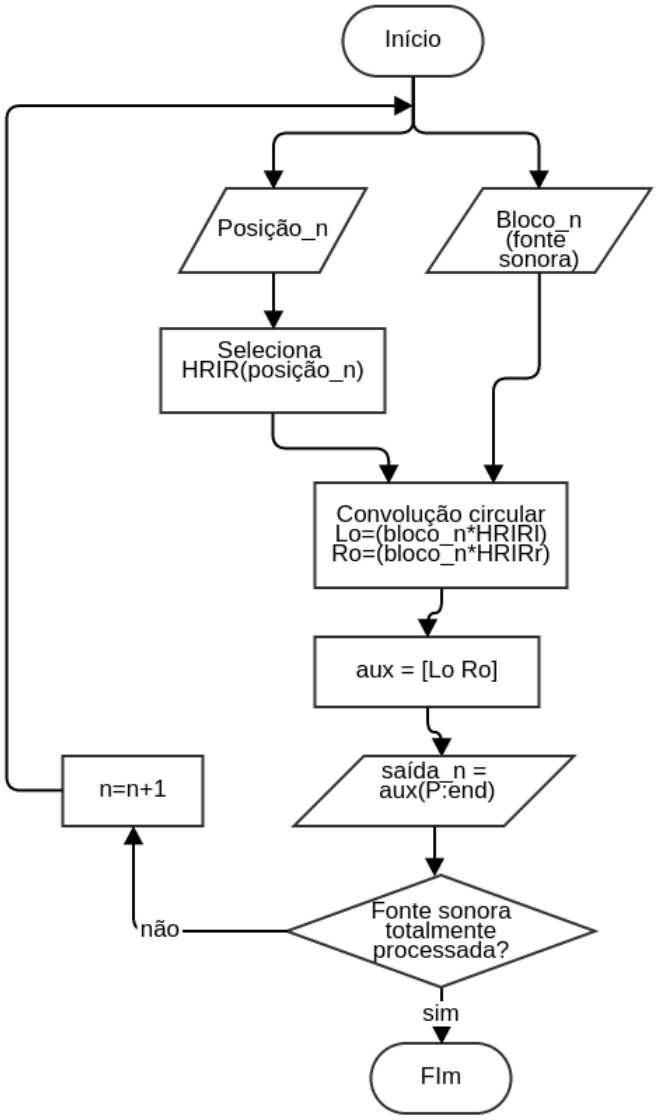


Figura 19 – Fluxograma do algoritmo de síntese biaural utilizando método de convolução por blocos *overlap-save*

3.2.2 Desenvolvimento da Aplicação de *Head Tracking*

O sistema de *head tracking* pode ser divido em duas partes, *hardware* e *software*. O *hardware* é composto por uma unidade de aquisição e transmissão de sinais e por um fone de ouvido. E o *software* é composto por um *firmware* no arduino e por uma aplicação em Matlab.

A unidade de aquisição e transmissão de sinais é composta por um arduino Funduino Pro Mini e uma unidade inercial de medida *9DOF Razor Stick* (Fig. (13)). O

equipamento é o mesmo utilizado por (LEITE et al., 2014) e é mostrada na Fig. (20). O arduino fica dentro da caixa maior, e a IMU *9DOF Razor Stick* fica dentro da caixa menor. A IMU foi anexada à haste de um fone de ouvido para monitorar a movimentação da cabeça de um ouvinte. O aparato montado é mostrado na Fig. (21).

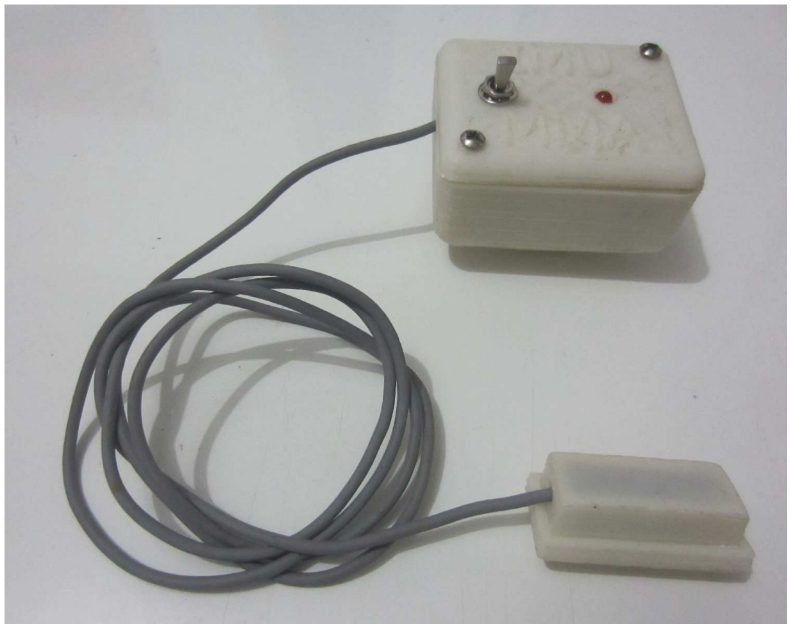


Figura 20 – Unidade de aquisição e transmissão de sinais composta por arduino (caixa grande) e IMU (caixa pequena) (LEITE et al., 2014)

Figura 21 – *Head tracker*: fone de ouvido equipado com unidade inercial de medida

O arduino se comunica com a placa de sensores por comunicação serial I^2C. Ele tem o papel de ler os dados dos sensores, processá-los e enviar para o computador os ângulos relativos ao movimento da cabeça do ouvinte. A comunicação do arduino com o computador também é feita via comunicação serial.

O *firmware* utilizado para o arduino é o *Razor AHRS v1.4.2*[1]. Este programa lê os sensores da IMU *9DOF Razor Stick*, realiza a fusão dos dados sensores (vide Fig. (14)) e envia via comunicação serial os ângulos de rotação *yaw*, *roll* e *pitch* da Fig. (15). Para realizar a fusão de sensores, o *firmware* utiliza um algoritmo DCM (*Direction Cosine Matrix*), baseado no trabalho realizado por Premerlani e Bizard (2009)[2].

Uma alteração no *firmware* foi realizada na função de envio dos ângulos para o computador. Ao invés de serem enviados os três ângulos citados anteriormente, somente o ângulo *yaw* é enviado. A alteração no envio de dados se deu porque, no presente trabalho, a virtualização das fontes sonoras se dará apenas no plano horizontal à cabeça do ouvinte. Isto quer dizer que apenas o deslocamento azimutal da cabeça é de interesse para o sistema, ou seja, o movimento de rotação no eixo z. Logo, apenas a informação do ângulo *yaw* é necessária para os testes com uso do sistema de *Head Tracking*.

O ângulo é enviado para o computador como uma *string* formatada: *"aSXXX.xx@"*. O caractere *a* é o indicador de início do valor enviado. O caractere @ é o indicador de término de envio do valor. O caractere S representa o sinal do valor, que pode ser negativo ou positivo. XXX representa a parte inteira do valor enviado e xx representa a parte fracionária. Os valores enviados pela IMU estão na seguinte faixa de valores: $+0.00$ a $+180.00$ e -0.00 a -180.00.

No Matlab, uma aplicação recebe a *string* formatada do arduino e retira o valor do ângulo da mensagem de acordo com a formatação explicada anteriormente. Inicialmente, uma calibração do *Head Tracker* é realizada, com o objetivo de identificar a posição da cabeça do ouvinte como a posição onde a face da pessoa aponte para a posição 1 da Fig. (16). Para atender a este objetivo, o ouvinte deve ficar com a cabeça parada no decorrer do processo de calibração.

A calibração consiste em coletar, por cinco segundos, valores do ângulo enviado pela unidade de aquisição e calcular a média desses valores coletados. Depois da etapa de calibração, os novos valores do ângulo de rotação do eixo z recebidos pela aplicação do computador são subtraídos do valor médio obtido durante a calibração. Além disso, caso o resultado da subtração seja negativo, é somado ao resultado o valor 360. Isso é feito para adequar os valores do ângulo de rotação à faixa de valores de 0 a 360 ° no sentido horário. Essa relação é definida por

$$\theta_H = \begin{cases} \theta_{yaw} - \theta_{Hm}, & (\theta_{yaw} - \theta_{Hm}) \geq 0 \\ \theta_{yaw} - \theta_{Hm} + 360, & (\theta_{yaw} - \theta_{Hm}) < 0 \end{cases} \tag{3.1}$$

onde θ_{yaw} corresponde ao ângulo de rotação do eixo z enviado para a apliação do Matlab

[1] Disponível em: <https://github.com/ptrbrtz/razor-9dof-ahrs>

[2] Sugere-se ao leitor leitura do artigo para entendimento da descrição do algoritmo DCM

pela unidade de aquisição e transmissão de sinais; θ_{Hm} corresponde ao valor médio obtido durante o processo de calibração e, θ_H corresponde ao ângulo de rotação no eixo z na faixa de 0 a 360 °. Como exemplo, na Fig. (18b) o valor de θ_H é 90 °, ou seja, o ouvinte virou a cabeça 90 ° para a direita.

Foi explicado na seção anterior que, na síntese biaural de tempo real, a posição aparente da fonte sonora deve mudar de acordo com a movimentação da cabeça do ouvinte. Considerando-se θ_H o ângulo da posição da cabeça e θ_S o ângulo da posição da fonte sonora, defini-se o ângulo θ_{ap} da posição aparente da fonte quando a cabeça é movimentada pela Eq. (3.2).

$$\theta_{ap} = \begin{cases} 360 - (\theta_H - \theta_S), & \theta_H > \theta_S \\ \theta_S - \theta_H, & \theta_H < \theta_S \\ \theta_H, & \theta_H = \theta_S \end{cases} \tag{3.2}$$

Com o novo ângulo θ_{ap}, a posição aparente da fonte sonora é alterada pela síntese biaural de tempo real, uma vez que o método de síntese utilizado nesse caso é o descrito pelo fluxograma da Fig. (19).

As possíveis posições aparentes não se resumem às oito possíveis posições reais de posicionamento das fontes sonoras da Fig. (16). Para dar a sensação de que a fonte se move continuamente em torno do ouvinte, é necessário que o sistema reconheça movimentos de cabeça que correspondam a ângulos menores que os 45 ° das oito posições mostrada na interface. O ideal seria que o menor ângulo possível fosse detectado. Isso não foi possível de se implementar, pois a resolução espacial dos bancos de HRTFs utilizados é de 5 °, isto é, tais bancos disponibilizam HRTFs para posições de 0 a 355 °, variando de 5 em 5 °. Portanto, o sistema reconhece movimentos correspondentes a 5 ° para reposicionar a fonte sonora.

3.2.3 Desenvolvimento da Interface Gráfica

O objetivo de se desenvolver uma interface gráfica foi criar um ambiente de testes que englobasse os testes de localização propostos neste trabalho (vide seção 3.2.4). A interface, também implementada em Matlab, é mostrada na Fig. (22).

Na interface, o usuário tem a opção de escolher entre quatro testes, selecionando para isso um dos quatro botões do tipo *check box* do canto superior esquerdo da interface. Abaixo dos botões dos testes, há o botão de navegação pelos testes. Na Figura (22), o botão está no estado desativado. Quando algum teste é selecionado, o botão passa para o estado ativado. Neste estado, a cor do botão é verde, podendo conter o texto “Iniciar” ou “Avançar”, dependendo do passo em que o usuário se encontra no teste. A Figura (23) mostra as duas configurações do botão de navegação no estado ativo.

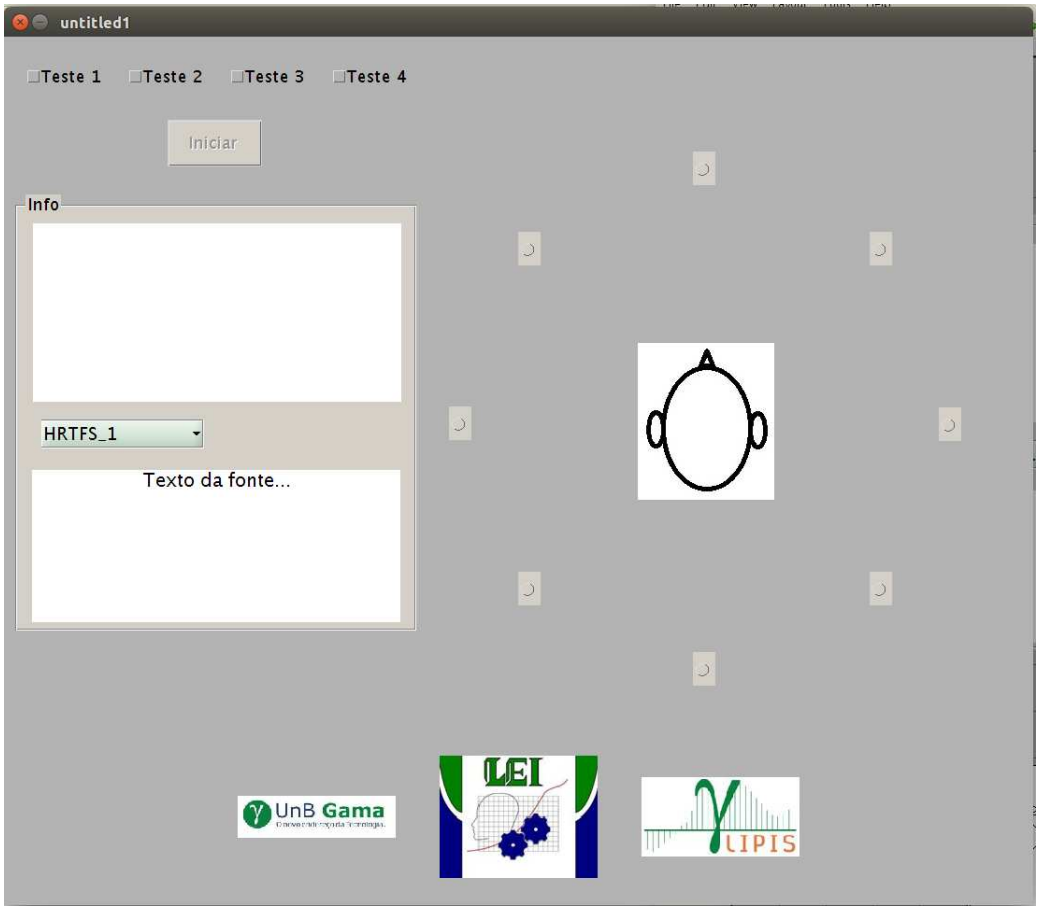


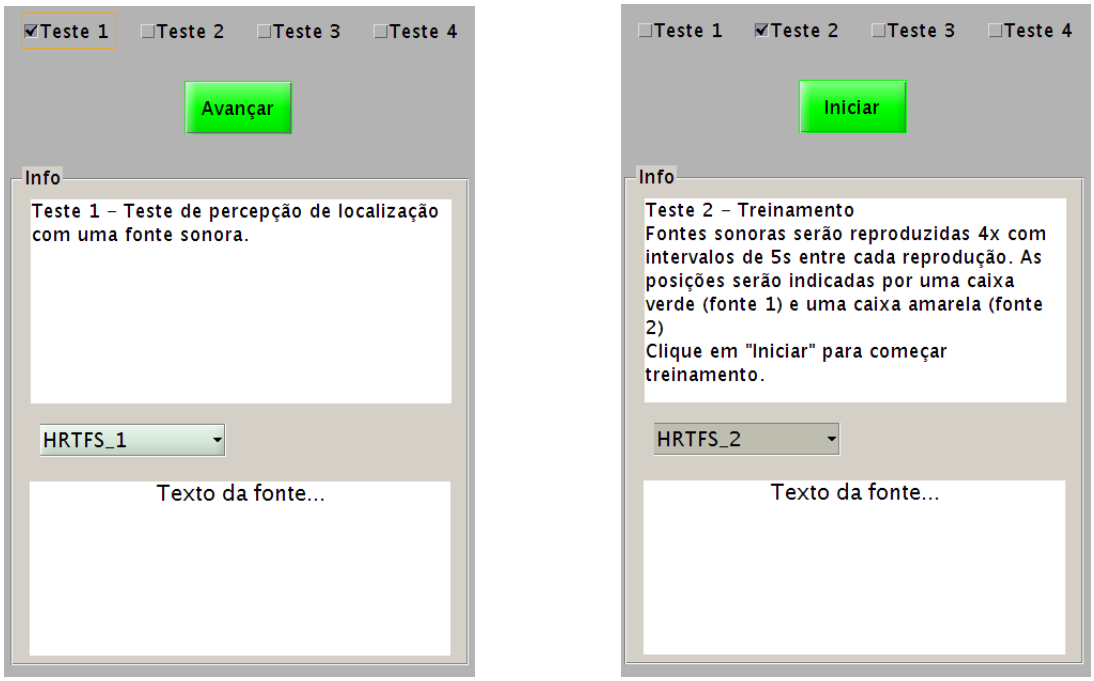
Figura 22 – Interface gráfica para realização dos testes de localização

Abaixo do botão de navegação se encontra um painel intitulado “Info”. Neste painel se encontram duas caixas de texto e um menu. Na primeira caixa de texto, informações a respeito do teste selecionado são mostradas ao usuário. Na Figura (23a), o texto informa do que se trata o teste 1. E na Figura (23b), o texto informa como será realizado o treinamento do teste 2. A segunda caixa de texto, como se pode ver na figura, tem o objetivo de mostrar o texto da fala da fonte sonora em execução. Esta funcionalidade não foi implementada no presente trabalho. O menu presente no painel de informações possibilita a seleção de diferentes bancos de HRTFs, como se pode ver nas Fig. (23a) e (23b).

No lado direito da interface da Fig. (22), são mostradas as posições possíveis (Fig. (16)) para posicionamento de fonte sonora em torno da cabeça de um ouvinte. Cada posição é corresponde a um botão do tipo *radio button*. Nesses botões o usuário pode indicar onde percebeu a localização de uma fonte sonora durante um teste.

3.2.4 Procedimentos de Teste

Quatro rotinas de testes de localização foram desenvolvidas. Elas podem ser divididas em dois grupos: rotinas de testes de localização em áudio binaural sem sistema de *head*



(a) (b)

Figura 23 – Botões de seleção de teste, botão de navegação e painel de informações da interface gráfica. (a) Teste 1 selecionado e botão de navegação no estado ativo "Avançar". (b) Teste 2 selecionado e botão de navegação no estado ativo "Iniciar".

tracking e rotinas de testes com sistema de *head tracking*.

As rotinas de testes sem o *head tracker* utilizam a aplicação de síntese biaural da Fig. (17). Na interface da Fig. (22), esses testes são o Teste 1 e o Teste 2. O Teste 1 é realizado com a auralização de apenas uma fonte sonora de fala. Já o Teste 2 é realizado com a auralização de duas fontes sonoras ao mesmo tempo.

Já as rotinas de teste com o sistema de *head tracking* utilizam a síntese biaural em tempo real da Fig. (19). Esses testes são o Teste 3 e o Teste 4 na interface, sendo o primeiro realizado com uma fonte sonora e o segundo realizado com duas fontes sonoras.

Para todos os testes, as posições reais possíveis para posicionamento das fontes sonoras são aquelas mostradas na Fig. (16). Antes de cada teste é realizada uma etapa de treinamento. Nesta etapa, a posição real da fonte sonora é mostrada visualmente na interface. Desse modo o ouvinte pode associar imagem e som, ficando mais fácil a distinção da posição da fonte sonora. Após a etapa de treinamento é realizado o teste de localização, onde a posição real da fonte sonora não é mostrada.

3.2.4.1 Teste 1 e Teste 3

O Teste 1 consiste num teste de localização com uma fonte sonora sem o sistema de *head tracking*. A Figura (24) mostra as etapas de navegação pelo teste que o usuário

faz durante a execução do teste.

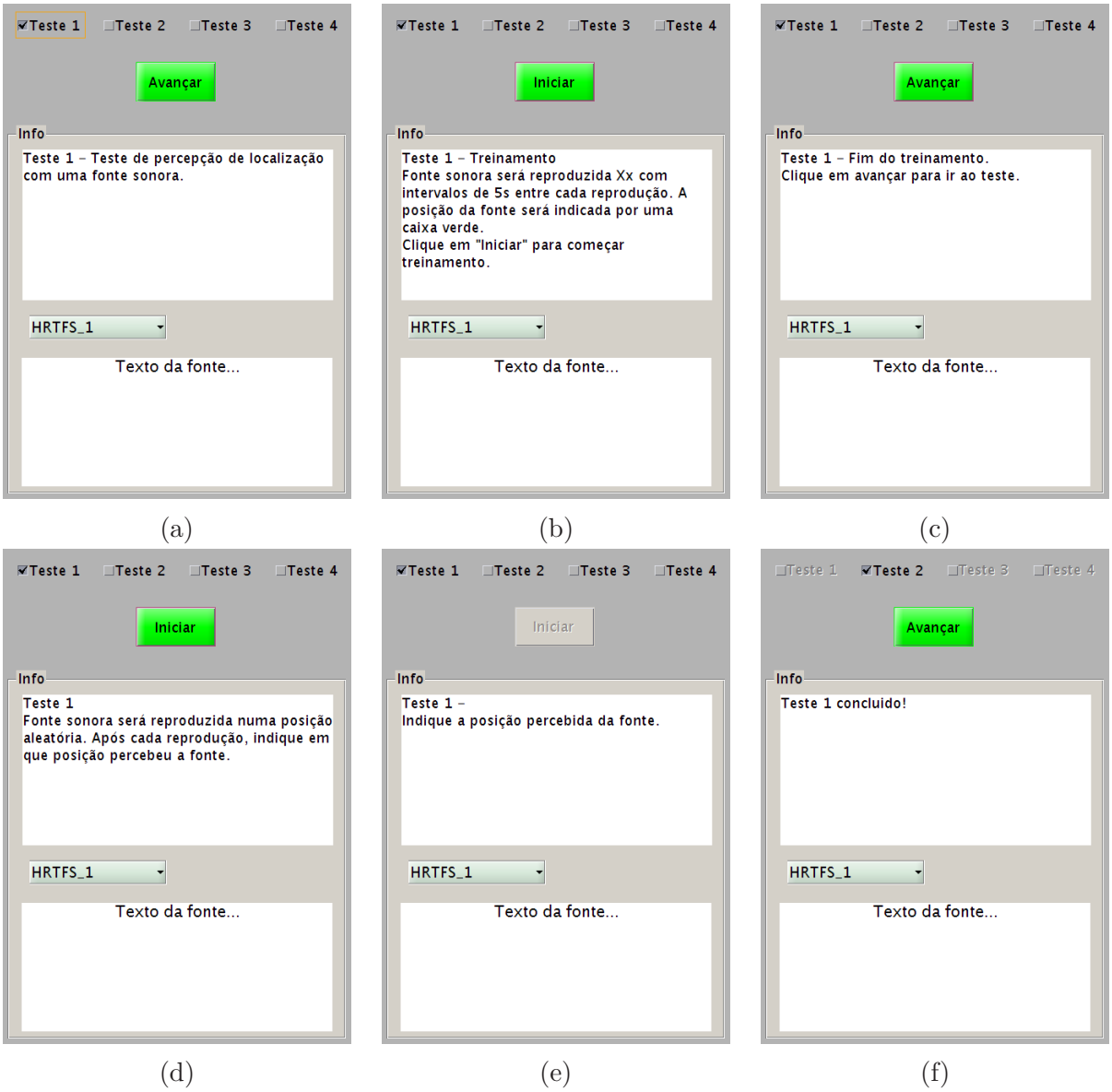


Figura 24 – Etapas de navegação do Teste 1. (a) Seleção do Teste 1. (b) Descrição do treinamento do teste. (c) Indicação do fim do treinamento. (d) Descrição do procedimento do Teste 1. (e) Etapa de indicação da posição percebida por parte do ouvinte. (f) Indicação de conclusão do Teste 1

Inicialmente, o ouvinte seleciona o Teste 1 (Fig. (24a). Então ele clica em “Avançar”. Feito isso, no painel de informação é descrito como será realizada a etapa de treinamento do teste (Fig. (24b)). Para iniciar o treinamento, o ouvinte deve apertar o botão “Iniciar”, mostrado na Fig. (24b).

No treinamento, uma fonte sonora é reproduzida quatro vezes em posições aleatórias. Durante cada reprodução, a posição em que a auralização se deu é mostrada por uma caixa verde, como se pode ver na Fig. (25). Quando o treinamento acaba, o ouvinte deve pressionar o botão “Avançar” (Fig. (24c) para que a interface mostre informações

sobre a execução do teste propriamente dito.

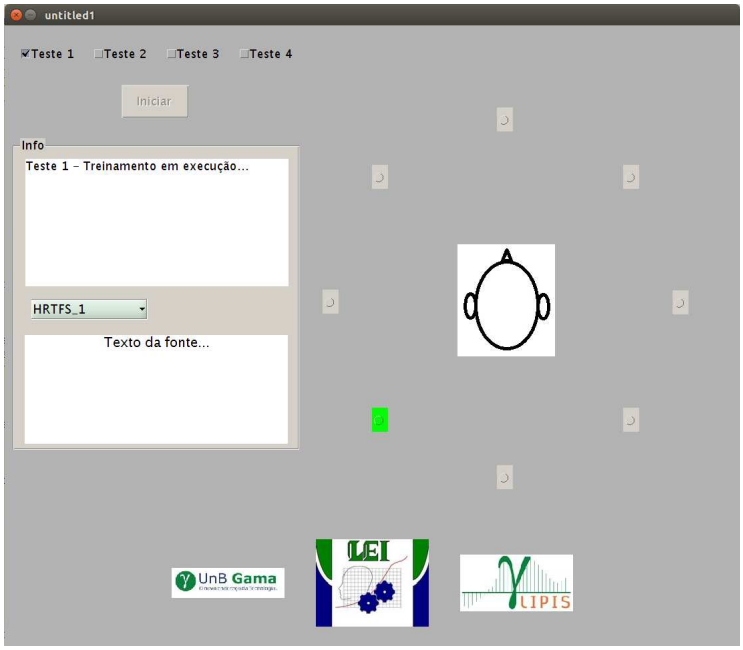


Figura 25 – Indicação visual da posição real da fonte sonora durante etapa de treinamento do Teste 1

Para começar o teste, o botão “Iniciar” deve ser pressionado (Fig. (24d). O teste consiste em se reproduzir, também quatro vezes e em posições aleatórias, uma fonte sonora de fala. Porém, diferentemente da etapa de treinamento, a posição real da fonte não será mostrada visualmente. Ao fim de cada reprodução, o ouvinte deve indicar em que posição percebeu a fonte sonora (Fig. (24e). Para isso, o ouvinte deve pressionar o botão que corresponda a posição de sua escolha. Ao pressionar o botão, uma janela de confirmação é aberta, como mostra a Fig. (26). Após o usuário confirmar a posição escolhida, caso a posição percebida coincida com a posição real da fonte na reprodução atual, a caixa da posição fica verde, indicando acerto do ouvinte. Caso contrário, a caixa fica vermelha, indicando erro.

O teste se encerra ao fim das quatro reproduções (Fig. (24f)). Então, o teste seguinte é automaticamente selecionado. Clicando no botão “Avançar” da Fig. (24f), o usuário inicia a realização do Teste 2.

Os procedimentos para realização do Teste 3 são similares aos procedimentos para realização do Teste 1. O Teste 3 contém uma etapa a mais que o Teste 1 antes do treinamento, a etapa de calibração do sistema de *head tracking*, mostrada na Fig. (27) e explicada na seção 3.2.2.

Como mencionado anteriormente, no Teste 3 o ouvinte tem a opção de movimentar a cabeça para tentar localizar mais facilmente a posição da fonte sonora (utilização do sistema de *head tracking*). Outra diferença entre o Teste 1 e o Teste 3 é que no último o

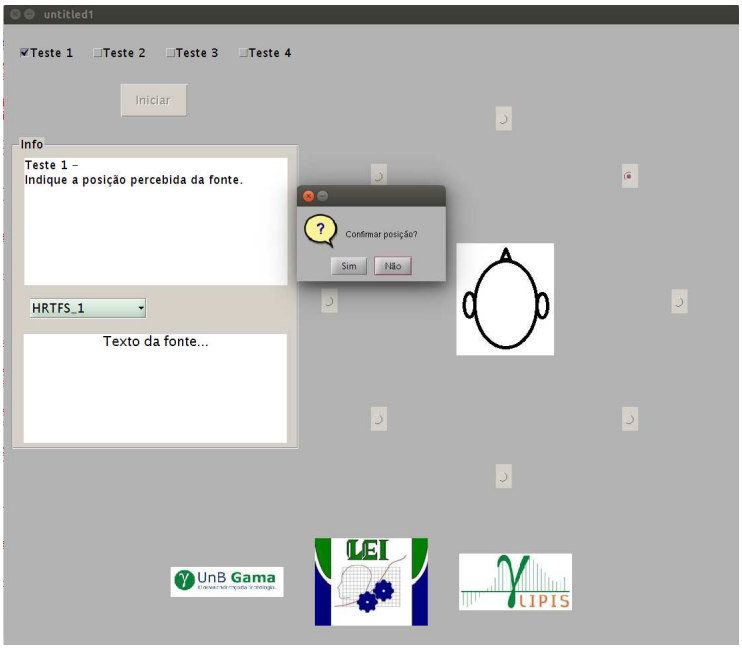


Figura 26 – Procedimento para indicar posição percebida na interface. Depois de pressionar um botão, uma janela de confirmação é aberta

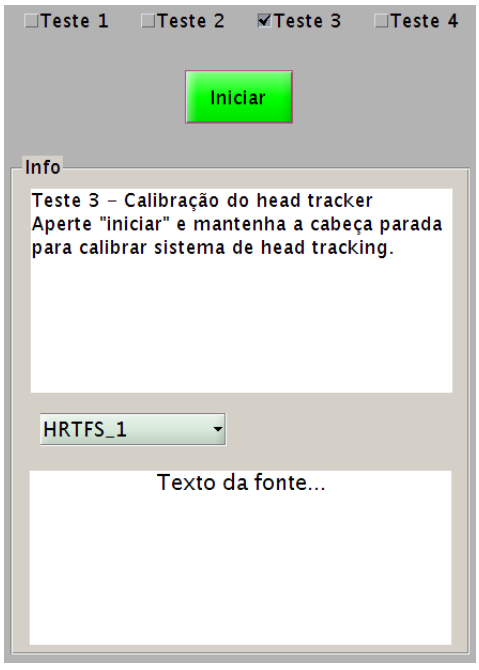


Figura 27 – Indicação da etapa de calibração do *head tracker* no Teste 3

tempo de reprodução de cada fonte sonora é maior. Isso é para propiciar tempo suficiente para o ouvinte movimentar a cabeça de um lado a outro a procura da fonte sonora.

Ao fim do Teste 3, o Teste 4 é automaticamente selecionado.

3.2.4.2 Teste 2 e Teste 4

O Teste 2 consiste em um teste de localização com duas fontes sonoras sem o sistema de *head tracking*. As etapas de navegação pelo teste são basicamente as mesmas que as mostradas na Fig. (24) para o Teste 1. O que muda são as informações mostradas no painel 'Info', que agora dizem respeito ao Teste 2.

Na etapa de treinamento do segundo teste, duas fontes sonoras são reproduzidas ao mesmo tempo quatro vezes. Em cada reprodução, a posição de cada fonte sonora é aleatoriamente escolhida. Uma fonte tem sua posição mostrada por uma caixa verde e a outra fonte, por uma caixa amarela. Isso é mostrado na Fig. (28).

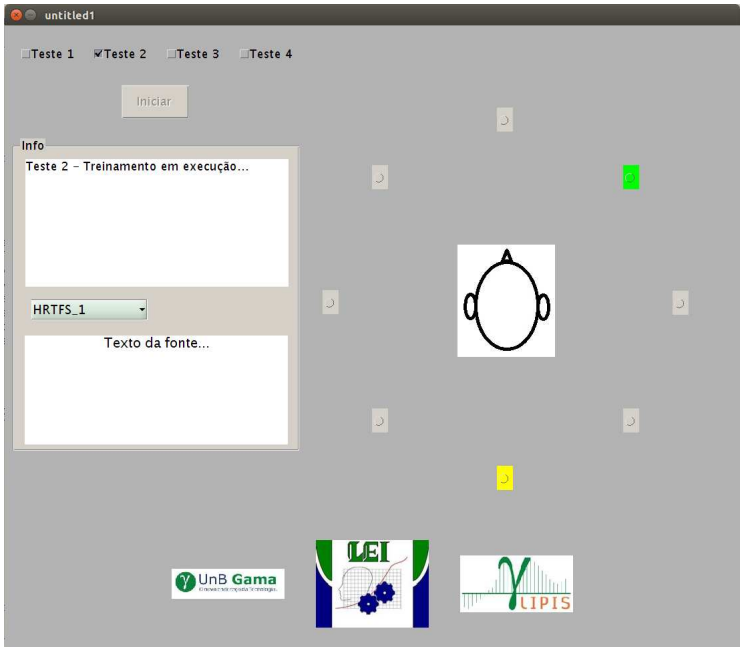


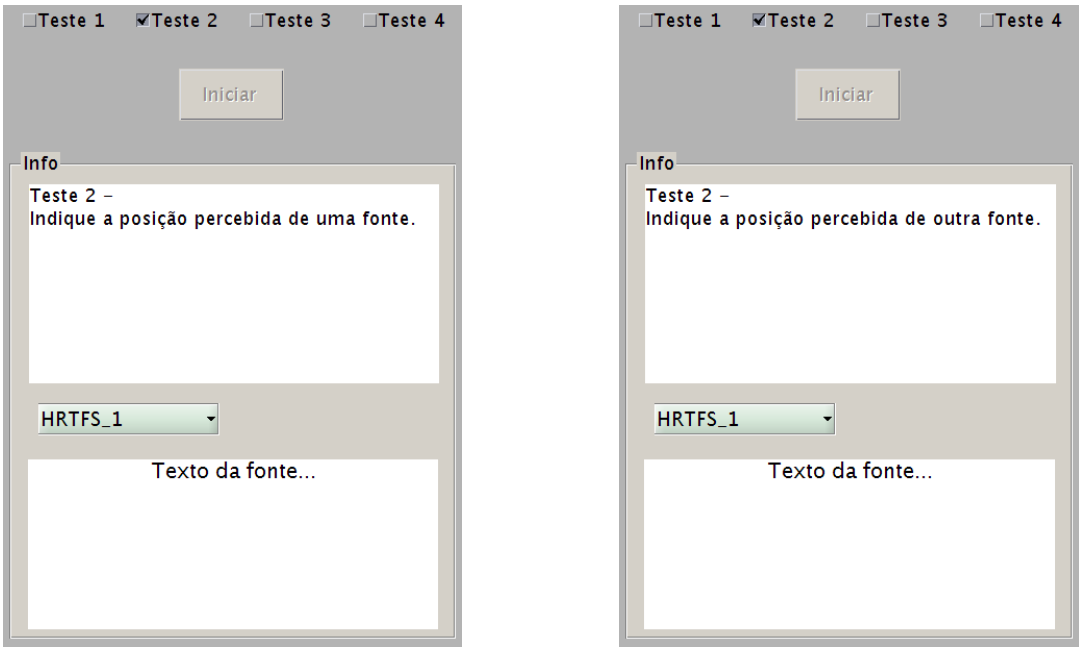
Figura 28 – Indicação visual das posições reais das fontes sonoras durante etapa de treinamento do Teste 2

A etapa após o treinamento é o teste em si. As fontes são reproduzidas quatro vezes em posições aleatórias, lembrando que na etapa de teste a posição não é visualmente indicada na interface. Após cada reprodução, o ouvinte é indagado a identificar a posição das duas fontes sonoras, uma de cada vez, como se pode ver na Fig. (29).

Ao fim das quatro reproduções, o Teste 2 se encerra. O Teste 3 é, então, automaticamente selecionado.

Em relação ao Teste 4, seus procedimentos são similares aos procedimentos para realização do Teste 2. Assim como no Teste 3, o Teste 4 contém uma etapa de calibração do *head tracker*. Além de o ouvinte ter a opção de movimentar a cabeça para tentar localizar mais facilmente a posição da fonte sonora com a utilização do sistema de *head tracking*, existe também uma diferença nas etapas dos treinamentos entre os testes 2 e 4.

No treinamento do Teste 4 são realizadas apenas duas reproduções. As posições



(a) (b)

Figura 29 – Etapa do Teste 2 onde ouvinte indica posições percebidas das fontes sonoras. (a) Indicação da primeira fonte. (b) Indicação da segunda fonte.

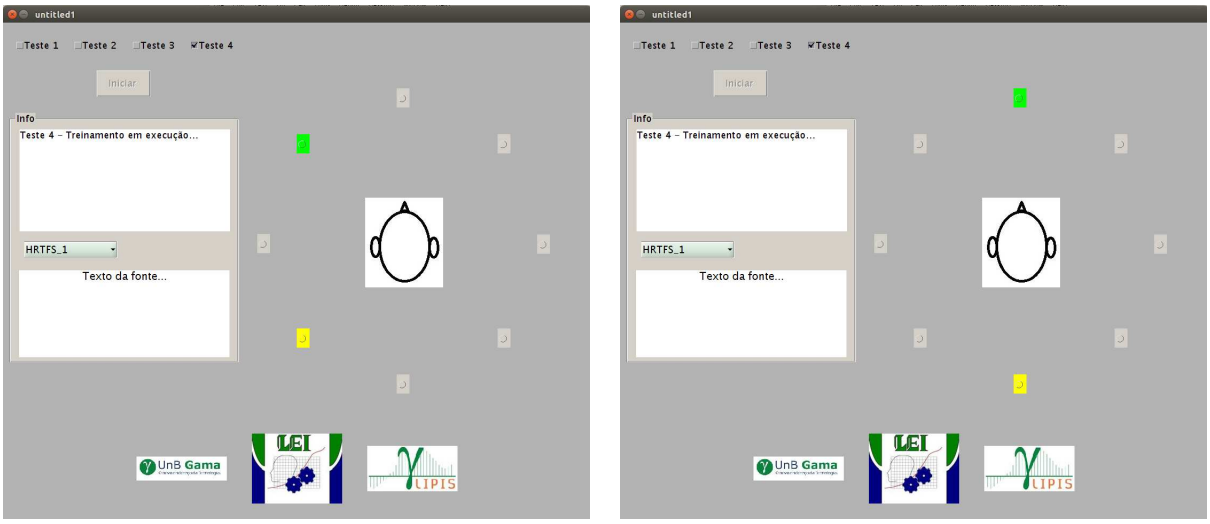
nesse caso não são aleatórias. Na primeira reprodução, as posições selecionadas são as mostradas na Fig. (30a). A Figura (30b) mostra as posições selecionadas para a segunda reprodução do treinamento. Essas posições foram selecionadas para mostrar ao ouvinte os benefícios de se utilizar o sistema de *head tracking*.

Como as posições das fontes na Fig. (30a) são equidistantes ao ouvido esquerdo do ouvinte, as duas estão na região do cone de confusão, explicado na seção 2.3. Por esse motivo, pode ser que o ouvinte perceba as duas fontes na mesma posição. Se o ouvinte girar a cabeça 90 ° para a esquerda ele perceberá a fonte da caixa verde à direita do ouvido direito e a fonte da caixa amarela à esquerda do ouvido esquerdo, distinguindo assim a posição de cada fonte. Situação parecida ocorre na Fig. (30b), onde o ouvinte pode não conseguir distinguir qual fonte está a sua frente e qual está atrás.

Os passos seguintes ao treinamento do Teste 4 são os mesmos realizados no Teste 2, lembrando é claro que o ouvinte pode movimentar a cabeça a procura da fonte sonora.

3.3 Protocolo Experimental

Antes da realização dos testes subjetivos de localização, primeiramente se verificou se as aplicações de síntese biaurais estavam realmente funcionando. Verificou-se também se o sistema de *head tracking* em conjunto com a síntese biaural em tempo real conseguia simular os efeitos da movimentação da cabeça sobre a percepção de localização de uma fonte sonora.



(a) (b)

Figura 30 – Indicação visual das posições reais das fontes sonoras durante etapa de treinamento do Teste 4. (a) Reprodução 1. (b) Reprodução 2.

Após verificar o correto funcionamento das aplicações de síntese biaural e do sistema de *head tracking*, e depois de finalizar o desenvolvimento da interface gráfica com as rotinas de teste, realizaram-se os testes subjetivos de localização com sujeitos.

Testou-se o sistema com quatro sujeitos. Cada sujeito realizou os quatro testes propostos, começando pelo Teste 1. Após a execução do Teste 1, foram anotadas as quatro posições reais em que o teste aleatoriamente posicionou uma fonte sonora. Também foram anotadas as posições onde o ouvinte percebeu a fonte sonora. O mesmo foi feito para o Teste 3. Para os testes 2 e 4, anotaram-se os pares de posições reais e os pares de posições percebidas.

O procedimento completo, isto é, a realização dos testes de 1 a 4, dura em média doze minutos.

4 Resultados

A verificação inicial das aplicações de auralização e do sistema de *head tracking* apresentou resultados que comprovaram o correto funcionamento dos sistemas desenvolvidos. Verificou-se que era possível perceber a virtualização da posição de uma fonte sonora com a utilização da síntese biaural da Fig. (17).

A síntese biaural em tempo real também foi verificada. Constatou-se que era possível movimentar uma fonte sonora em torno da cabeça do ouvinte de forma contínua e sem descontinuidades no som, dada a resolução espacial das HRTFs. A integração dessa síntese biaural com o sistema de *head tracking* também funcionou como esperado, isto é, permitia ao ouvinte movimentar a cabeça e perceber o reposicionamento dinâmico da fonte sonora em relação à posição da cabeça.

Após a verificação dos sistemas implementados e a partir da execução do protocolo de testes descrito no capítulo anterior, foram obtidas, para cada sujeito submetido ao teste, quatro tabelas referentes aos testes de 1 a 4. Cada tabela informa a posição de fonte sonora atribuída pela interface (posição real) durante a rotina de teste, e também a posição que o sujeito indicou como percebida. No caso das rotinas de testes para duas fontes sonoras, duas posições reais e duas percebidas são mostradas.

Neste ponto vale lembrar que são oito as posições reais possíveis para localização das fontes sonoras, sendo que essas posições vão de 1 (0 °, a frente da cabeça) a 8 (315 °) no sentido horário, variando 45 ° entre cada posição (vide Fig. (16)).

As Tabelas (1) a (4) mostram os resultados obtidos com a execução do Teste 01 para os quatro sujeitos. As Tabelas (5) a (8) mostram os resultados obtidos com a execução do Teste 02.

Passo do teste	**Posição real**	**Posição percebida**
Passo I	8	7
Passo II	5	2
Passo III	2	3
Passo IV	2	3

Tabela 1 – Resultados obtidos com execução do Teste 01 para sujeito I

Passo do teste	**Posição real**	**Posição percebida**
Passo I	3	7
Passo II	5	5
Passo III	4	4
Passo IV	1	5

Tabela 2 – Resultados obtidos com execução do Teste 01 para sujeito II

Passo do teste	**Posição real**	**Posição percebida**
Passo I	6	6
Passo II	4	4
Passo III	5	1
Passo IV	3	3

Tabela 3 – Resultados obtidos com execução do Teste 01 para sujeito III

Passo do teste	**Posição real**	**Posição percebida**
Passo I	7	7
Passo II	4	2
Passo III	3	3
Passo IV	3	3

Tabela 4 – Resultados obtidos com execução do Teste 01 para sujeito IV

Passo do teste	**Posições reais**	**Posições percebidas**
Passo I	3;5	3;6
Passo II	4;5	3;4
Passo III	7;5	7;8
Passo IV	5;8	7;8

Tabela 5 – Resultados obtidos com execução do Teste 02 para sujeito I

Passo do teste	**Posições reais**	**Posições percebidas**
Passo I	4;4	3;4
Passo II	3;8	7;3
Passo III	3;1	5;3
Passo IV	7;4	7;3

Tabela 6 – Resultados obtidos com execução do Teste 02 para sujeito II

Passo do teste	Posições reais	Posições percebidas
Passo I	8;7	7;6
Passo II	4;4	3;4
Passo III	4;3	3;4
Passo IV	5;5	1;5

Tabela 7 – Resultados obtidos com execução do Teste 02 para sujeito III

Passo do teste	Posições reais	Posições percebidas
Passo I	7;7	7;8
Passo II	2;3	3;4
Passo III	8;6	6;7
Passo IV	7;2	4;3

Tabela 8 – Resultados obtidos com execução do Teste 02 para sujeito IV

As Tabelas (9) a (12) mostram os resultados obtidos com a execução do Teste 03 para os quatro sujeitos. As Tabelas (13) a (16) mostram os resultados obtidos com a execução do Teste 04.

Passo do teste	Posição real	Posição percebida
Passo I	3	3
Passo II	7	7
Passo III	7	7
Passo IV		

Tabela 9 – Resultados obtidos com execução do Teste 03 para sujeito I

Passo do teste	Posição real	Posição percebida
Passo I	2	4
Passo II	4	5
Passo III	1	8
Passo IV	2	4

Tabela 10 – Resultados obtidos com execução do Teste 03 para sujeito II

Passo do teste	**Posição real**	**Posição percebida**
Passo I	7	6
Passo II	7	6
Passo III	6	6
Passo IV	4	4

Tabela 11 – Resultados obtidos com execução do Teste 03 para sujeito III

Passo do teste	**Posição real**	**Posição percebida**
Passo I	4	2
Passo II	6	7
Passo III	7	7
Passo IV	4	4

Tabela 12 – Resultados obtidos com execução do Teste 03 para sujeito IV

Passo do teste	**Posições reais**	**Posições percebidas**
Passo I	8;2	2;3
Passo II	5;4	2;3
Passo III	1;3	3;4
Passo IV	2;7	3;1

Tabela 13 – Resultados obtidos com execução do Teste 04 para sujeito I

Passo do teste	**Posições reais**	**Posições percebidas**
Passo I	8;1	7;3
Passo II	8;1	2;6
Passo III	7;3	7;3
Passo IV	8;2	7;3

Tabela 14 – Resultados obtidos com execução do Teste 04 para sujeito II

Passo do teste	**Posições reais**	**Posições percebidas**
Passo I	2;7	2;3
Passo II	4;2	2;5
Passo III	7;2	7;2
Passo IV	2;2	2;2

Tabela 15 – Resultados obtidos com execução do Teste 04 para sujeito III

Passo do teste	Posições reais	Posições percebidas
Passo I	2;3	2;4
Passo II	8;7	8;7
Passo III	3;1	4;5
Passo IV	2;4	1;2

Tabela 16 – Resultados obtidos com execução do Teste 04 para sujeito IV

Além das tabelas, para os testes 01 e 03, foram desenhados gráficos que mostram o histograma dos erros de localização considerando-se os quatro sujeitos. O erro é definido como sendo o módulo da subtração da posição real pela posição percebida em cada passo, e indica a distância, em posições, entre a posição real da fonte sonora e a posição percebida pelo sujeito. A Figura (31) mostra o gráfico para o Teste 01 e a Fig. (32) mostra o gráfico para o Teste 03.

Para os testes 02 e 04 não foram desenhados os histogramas de erro porque tais testes apresentam duas posições reais e duas posições percebidas de fonte sonora. E do modo como a interface está implementada, o sujeito não é indagado quanto a uma fonte específica. Apenas se pede que o sujeito indique uma posição onde ele percebeu alguma fonte. Desse modo, não se sabe qual a fonte relacionada a posição percebida pelo sujeito. Então, não tem sentido traçar o histograma de erro de distância entre as posições percebidas e as posições reais para os testes 02 e 04, pois nesse caso as distâncias poderiam ser entre duas fontes diferentes, o que invalidaria a análise.

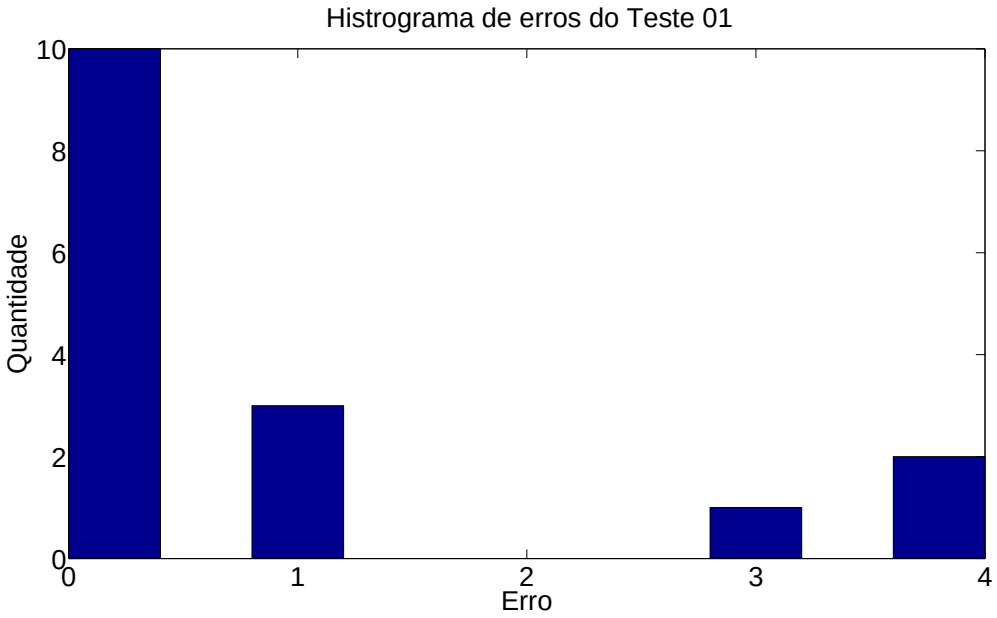


Figura 31 – Histograma do erro de localização para o Teste 01

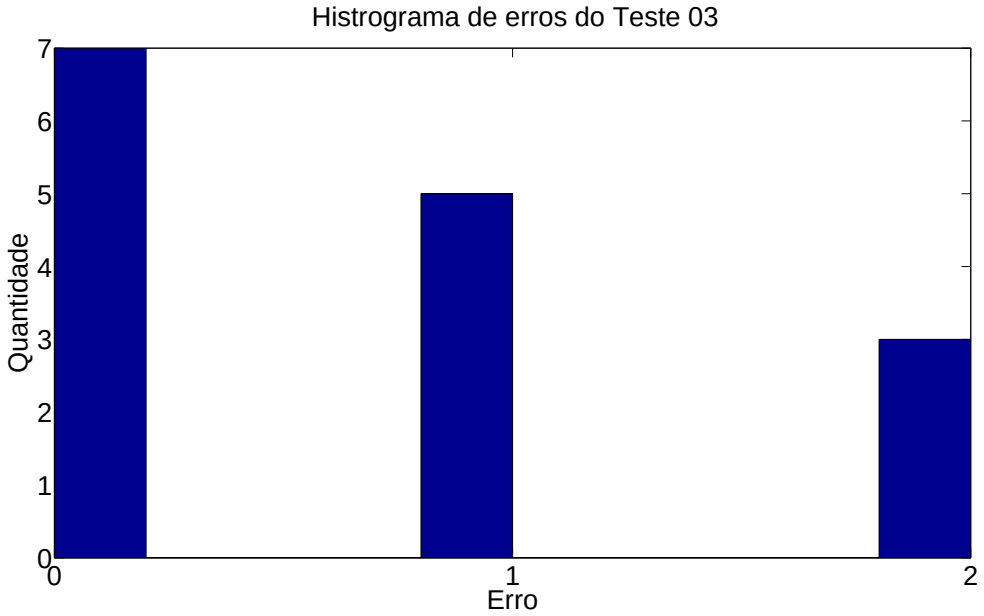


Figura 32 – Histograma do erro de localização para o Teste 03

Em relação aos resultados obtidos com a execução dos testes 02 e 04 (testes com duas fontes sonoras), pode-se tabelar a quantidade de acertos, isto é, a quantidade de posições percebidas iguais as posições reais. As Tabelas (17) e (18) mostram os acertos para os testes 02 e 04, respectivamente. Em cada tabela se mostra a quantidade de acertos de apenas uma das fontes num certo passo do teste, e a quantidade de vezes que o sujeito conseguiu identificar as duas fontes sonoras no teste.

Sujeito	**Acertos de apenas uma fonte**	**Acertos das duas fontes**
Sujeito I	3	1
Sujeito II	4	0
Sujeito III	3	1
Sujeito IV	3	0

Tabela 17 – Quantidade de acertos na percepção de localização por sujeito no Teste 02

Sujeito	**Acertos de apenas uma fonte**	**Acertos das duas fontes**
Sujeito I	2	0
Sujeito II	0	1
Sujeito III	1	3
Sujeito IV	2	1

Tabela 18 – Quantidade de acertos na percepção de localização por sujeito no Teste 04

Durante a execução dos testes, alguns sujeitos observaram que o sistema de *head tracking* apresentava uma pequena latência para rearranjar o espaço sonoro, o que significa que quando o sujeito movimentava a cabeça, o sistema não reposicionava instantaneamente a fonte sonora na posição aparente correta. Outra observação feita foi que o tempo de reprodução da fonte sonora com o *head tracker* era curto, não dando tempo para localizar corretamente a posição das fontes sonoras. Segundo os sujeitos, isso teve maior influência no teste com duas fontes. Uma fonte era identificada, mas a reprodução terminava antes de se conseguir localizar a segunda fonte.

5 Discussão

Analisando-se o histograma da Fig. (31), que compila os resultados obtidos com a execução do Teste 1, podemos notar que a quantidade de acertos (erro=0) na percepção de localização da fonte sonora foi de 62,5%, alta em comparação com os erros. Além disso, a figura também mostra que a maioria de erros é igual a um. Isto revela que o sujeito conseguiu identificar, ao menos, a região da posição real da fonte sonora.

Um dado interessante que a Fig. (31) mostra é que ocorreram dois erros com valor igual a 4. Observando as Tab. de (1) a (4), podemos ver que esses erros se originaram das posições 1 e 5 (0° e 180°), ou seja, os sujeitos não distinguiram corretamente se a fonte estava a frente ou atrás da cabeça. Esse tipo de erro, como dito na introdução deste trabalho, é comum quando a localização se dá apenas pelas informações provenientes da HRTF. Logo, apesar de o erro ser grande, é um erro que era esperado para os casos dessas posições.

Em relação ao Teste 03, segundo a Fig. (32), a quantidade de acertos diminuiu e a quantidade de erros iguais a dois aumentou em comparação ao Teste 1. Pode-se dizer que o erro igual a 2 é um erro grosseiro de localização, pois erro igual a dois corresponde a uma angulação de 90° entre a posição real da fonte sonora e a posição percebida. A quantidade de erros iguais a unidade também aumentou. Mesmo assim, pode-se notar que, com a utilização do sistema de *head tracking*, os erros iguais a 3 e a 4 não aconteceram. Isso indica que o *head tracker* trouxe, de alguma forma, benefícios para a localização da fonte sonora.

Comparando-se o desempenho dos sujeitos nos dois testes de localização para o caso de uma fonte sonora, a percepção de localização do sujeito II piorou consideravelmente do Teste 1 para o Teste 3. Por outro lado, o desempenho do sujeito I no teste melhorou de 0 para 100% de acertos. A maioria dos erros dos sujeitos III e IV no Teste 3 foram iguais a um. Nesse caso, o erro igual a unidade pode ser considerado pior que o mesmo erro no Teste 01, isso porque no Teste 03 o sujeito podia movimentar a cabeça para ajudar na localização da fonte sonora.

Analisando-se os dados relativos aos testes com duas fontes sonoras, a Tabela (17) mostra que o número de acertos das posições das duas fontes sonoras num passo do Teste 2 é baixo. Já o número de acertos de apenas uma fonte é considerável. Isso pode indicar que o sujeito entendeu melhor o que uma fonte dizia do que a outra, ou seja, o grau de inteligibilidade das duas fontes sonoras é baixo.

Já no Teste 4, com o uso do sistema de *head tracking*, o número de acertos das duas posições das fontes sonoras aumentou em relação ao Teste 2, o que também indica

benefícios advindos da utilização do sistema de monitoramento da posição da cabeça. Porém, a quantidade de acertos totais, considerando os acertos de apenas uma fonte e os acertos das duas fontes ao mesmo tempo, diminuiu. Portanto, pode-se dizer que a inteligibilidade, em relação às duas fontes, aumentou com a utilização do sistema de *head tracking*. Mas, no geral, não houve melhora na localização das fontes.

Nos testes 2 e 4, os sujeitos III e IV apresentaram melhora na percepção de localização de um teste para o outro. O que contrasta com o desempenho dos sujeitos I e II, que apresentaram maior dificuldade na localização das fontes com a utilização do sistema de *head tracking*. Isso pode indicar que tais sujeitos não se adaptaram bem à utilização do *head tracker* assim como o fizeram os sujeitos III e IV.

Ainda em relação aos testes com duas fontes sonoras, observando-se as Tab. de (5) a (8) e as Tab. de (13) a (16), vemos que, desconsiderando-se as posições acertadas pelos sujeitos, a maioria dos erros de localização são iguais a unidade. O que também significa que os ouvintes conseguiram, pelo menos, identificar a região da posição real da fonte sonora.

O problema de latência relatado pelos sujeitos que se submeteram aos testes se deve ao ambiente utilizado para se desenvolver o sistema de testes. Para o uso de uma aplicação de *head tracking*, era necessário que o processamento do áudio fosse em tempo real. No processo de desenvolvimento da aplicação de síntese biaural em tempo real para MATLAB, conseguiu-se processar o áudio com uma taxa de atualização suficientemente rápida para se ter um sinal de áudio de saída não distorcido.

Porém, ao integrar o sistema de *head tracking* à essa aplicação de síntese, a taxa de atualização se tornou um problema. O sistema não conseguia processar o áudio e atualizar o ângulo vindo da IMU com a mesma velocidade que processava o áudio quando o *head tracker* não era utilizado.

Para tentar contornar esse problema, aumentou-se o tamanho do bloco do sinal de entrada a ser processado. Dessa maneira haveria mais tempo para processar o áudio e atualizar o ângulo vindo da IMU. No entanto, o tamanho do bloco de sinal de áudio se tornou considerável. Desse modo, mesmo que o valor do ângulo seja atualizado rapidamente, para essa atualização chegar ao sinal de saída, o bloco anterior do sinal de áudio processado deve terminar de ser reproduzido. Devido ao tamanho dos blocos, o tempo para essa atualização ocorrer se torna apreciável. Por esse motivo, é percebida uma latência na movimentação aparente da fonte sonora quando o ouvinte movimenta a cabeça.

Outro fato que contribui para a ocorrência da latência descrita anteriormente, é o fato de o MATLAB ser um ambiente com várias camadas de abstração. Logo, não é possível garantir que uma aplicação do MATLAB seja executada em tempo real.

6 Conclusão

Com intuito de avaliar a qualidade de síntese biaural para fontes sonoras de fala com a utilização de bancos de funções de transferência relacionadas à cabeça, o presente trabalho buscou desenvolver um sistema para testes subjetivos de localização para uma ou duas fontes, com ou sem sistema de monitoramento da posição da cabeça.

Os resultados dos testes de localização para o caso de uma fonte sonora sem o uso de *head tracker* nos permitem concluir que a qualidade da síntese biaural utilizando um banco de HRTFs genérico é apreciável, uma vez que a quantidade de acertos nesses testes é alta. Além da quantidade de acertos, o fato dos ouvintes conseguirem identificar ao menos a região onde a fonte sonora estava posicionada também mostra a qualidade da síntese biaural. Com isso, pode-se dizer que a síntese biaural consegue fornecer informações de diretividade suficientes para uma boa localização espacial no plano horizontal (plano considerado neste trabalho).

Quanto ao teste com duas fontes sonoras, também sem o uso de *head tracker*, podemos concluir que há dificuldade na percepção de localização das duas fontes ao mesmo tempo. Porém, também se pode concluir que com a utilização do áudio binaural, o ouvinte consegue ao menos discernir uma das fontes sonoras de forma mais precisa, pois o número de acertos da localização de ao menos uma fonte sonora é alto.

No caso do teste com uma fonte sonora, o uso do *head tracker* propiciou diminuição dos maiores erros de localização em comparação ao teste que não utiliza o sistema. Porém, a quantidade de acertos diminuiu. Logo, apesar dos evidentes benefícios gerados por sua utilização, a eficiência do sistema no aumento da facilidade de localização das fontes sonoras não foi alta.

Em relação ao teste com duas fontes sonoras, o uso do sistema de *head tracking* nos leva a conclusões similares. Com a utilização do do sistema de monitoramento da cabeça, a localização das duas fontes ao mesmo tempo aumentou, mostrando os benefícios que a utilização do sistema trouxe. Porém, no geral, os acertos de localização das fontes sonoras diminuiu em comparação ao teste sem o uso de *head tracker*.

A eficácia do uso do sistema de *head tracking* não pode ser descartada por completo. Isso porque alguns fatores colaboraram para que o sistema não contribuísse como esperado nos testes. O fato de o sistema de testes ter sido implementado em Matlab prejudicou a execução em tempo real da síntese biaural com o sistema de *head tracking*. A latência observada na movimentação da posição aparente da fonte sonora em relação à movimentação da cabeça dos ouvintes é um elemento que contribuiu para a dificuldade de localização das fontes sonoras.

Além da latência observada, o tempo de reprodução dos áudios nos testes também pode ter influência no baixo rendimento do sistema de *head tracking*. Alguns sujeitos relataram que o tempo foi insuficiente para se localizar corretamente as fontes sonoras, sobretudo nos testes com duas fontes sonoras. Além disso, o treinamento antes dos testes também parece ter influenciado os resultados obtidos com a utilização do *head tracker*. De acordo com os resultados, alguns sujeitos apresentaram melhora com o uso do sistema. Porém, outros sujeitos apresentaram uma redução significativa na quantidade de acertos em comparação com os testes que não utilizam o sistema.

Com base nesses fatores, conclui-se que os resultados obtidos com a utilização do sistema de *head tracking* não permitem afirmar que tal sistema foi ineficiente para a localização das fontes sonoras. Isso porque seu funcionamento, no ambiente de desenvolvimento utilizado, não pode ser considerado como sendo pleno. Mesmo assim, os resultados mostram que a utilização de tal sistema pode trazer benefícios para a localização de fontes sonoras em áudio biaural.

A interface gráfica desenvolvida se mostra um bom ambiente para realização de testes de localização para áudio biaural com a opção de utilização de um sistema de *head tracking*. Considerando-se isso, sugere-se que a interface seja complementada para trabalhos futuros. Nos testes com duas fontes, ao invés de apenas pedir que o ouvinte indique as posições em que percebeu qualquer uma das fontes, é desejável que se pergunte ao ouvinte onde ele ouviu cada fonte sonora especificamente. Para isso, pode-se, por exemplo, mostrar o texto que é dito por cada fonte sonora ao perguntar em que posição tal fonte se encontra. Desse modo, a avaliação de inteligibilidade pode ser mais bem explorada, pois assim se pode avaliar o entendimento do ouvinte em relação às diferentes fontes sonoras. O treinamento e tempo de reprodução do áudio auralizado podem ser aumentados nos casos dos testes para uso do *head tracker*, visando melhorar a qualidade dos testes.

Também se sugere que o sistema de testes seja implementado num ambiente onde se tenha maior controle sobre o tempo de execução das tarefas da aplicação. Desse modo, o problema de latência pode ser extinto, o que permitiria uma melhor avaliação do uso de um sistema *head tracking* para localização de fontes sonoras em áudio biaural.

Referências

AHMAD, N. et al. Reviews on various inertial measurement unit (imu) sensor applications. *International Journal of Signal Processing Systems*, v. 1, n. 2, p. 256–262, dec 2013. Citado 3 vezes nas páginas 11, 37 e 38.

ARONS, B. A review of the cocktail party effect. *JOURNAL OF THE AMERICAN VOICE I/O SOCIETY*, v. 12, p. 35–50, 1992. Citado na página 23.

BALDIS, J. J. Effects of spatial audio on memory, comprehension, and preference during desktop conferences. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2001. (CHI '01), p. 166–173. Citado na página 21.

BEGAULT, D. R. *3-D Sound for Virtual Reality and Multimedia*. Moffett Field, California, USA: NASA, 2000. Citado na página 21.

BRIGHAM, E. O. *The Fast Fourier Transform*. 1st. ed. Englewood Cliffs, N. J.: Prentice-Hall, Incorporated, 1974. ISBN 0-13-307496-X. Citado na página 34.

BRUCK, J.; GRUNDY, A.; JOEL, I. *An Audio Timeline - A selection of significant events, inventions, products and their purveyors,from cylinder to DVD*. 2013. Disponível em: <http://www.aes.org/aeshc/docs/audio.history.timeline.html>. Citado 2 vezes nas páginas 25 e 26.

CARTY, B.; LAZZARINI, V. A rational hrtf interpolation approach for fast synthesis of moving sound. In: *Proc. 6th Linux Audio Conference*. [S.l.: s.n.], 2008. p. 28–35. Citado na página 22.

CASCIA, M. L.; SCLAROFF, S.; ATHITSOS, V. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, p. 322–336, 2000. Citado na página 36.

CHANDA, P.; PARK, S.; KANG, T.-I. A binaural synthesis with multiple sound sources based on spatial features of head-related transfer functions. In: *Neural Networks, 2006. IJCNN '06. International Joint Conference on*. [S.l.: s.n.], 2006. p. 1726–1730. Citado na página 22.

CHENG, C. I.; WAKEFIELD, G. H. Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. *J. Audio Eng. Soc*, v. 49, n. 4, p. 231–249, 2001. Disponível em: <http://www.aes.org/e-lib/browse.cfm?elib=10196>. Citado 2 vezes nas páginas 30 e 31.

DAVIS, M. F. History of spatial coding. *J. Audio Eng. Soc*, v. 51, n. 6, p. 554–569, 2003. Citado 2 vezes nas páginas 25 e 26.

DEERING, M. High resolution virtual reality. *SIGGRAPH Comput. Graph.*, ACM, New York, NY, USA, v. 26, n. 2, p. 195–202, jul. 1992. ISSN 0097-8930. Citado na página 36.

FARIA, R. R. A. *Auralização em ambientes audiovisuais imersivos*. Tese (Doutorado em Sistemas Eltrônicos) — Escola Politécnica, University of São Paulo, São Paulo, São Paulo, 2005. Citado 3 vezes nas páginas 22, 25 e 26.

FERNANDES, J. C. Acústica e ruídos. Apostila. 2005. Citado na página 25.

FILIPANITS JR., F. *Design and Implementation of an Auralization System with a Spectrum-Based Temporal Processing Optimization*. Dissertação (Mestrado) — University of Miami, may 1994. Citado na página 23.

GARDNER, B.; MARTIN, K. *hrtf measurements of an kemar dummy-head microphone*. [S.l.], 1994. Citado 3 vezes nas páginas 28, 31 e 39.

GENUIT, K.; GIERLICH, H. W.; BRAY, W. Development and use of binaural recording technique. In: *Audio Engineering Society Convention 89*. [S.l.: s.n.], 1990. Citado na página 28.

GOMES, D. A. R. *Criação e manipulação de áudio 3D em tempo real utilizando unidades de processamento gráfico (GPU)*. Dissertação (Mestrado em Informática) — Universidade de Brasília, Brasília, 2012. 184f. Citado 5 vezes nas páginas 21, 25, 29, 30 e 33.

G.R.A.S. Sound & Vibration. *KEMAR ® Manikin Type 45BA*. [S.l.], 2006. 8 p. Disponível em: <http://www.campbell-associates.co.uk/products/Gras/productdata/KEMAR-Manikin-Type-45BA.pdf>. Citado na página 28.

HAVELOCK, D.; KUWANO, S.; VORLAENDER, M. (Ed.). *Handbook of Signal Processing in Acoustic*. New York: Springer, 2008. Citado na página 40.

HYDER, M.; HAUN, M.; HOENE, C. Placing the participants of a spatial audio conference call. In: *Consumer Communications and Networking Conference (CCNC), 2010 7th IEEE*. [S.l.: s.n.], 2010. p. 1–7. Citado na página 22.

HäRMä, A. et al. Personalization of headphone spatialization based on the relative localization error in an auditory gaming interface. In: *Audio Engineering Society Convention 132*. [S.l.: s.n.], 2012. Citado na página 28.

ITU-R. *Multichannel stereophonic sound system with and without accompanying picture*. Geneva, 2012. 23 p. Citado na página 27.

ITU-T. *SERIES P: Telephone Transmission Quality, Telephone Installations, Local Line Networks. Artificial voices. Appendix I: Test signals*. [S.l.], 1998. 62 p. Citado na página 39.

KANG, S. H.; KIM, S. H. Realistic audio teleconferencing using binaural and auralization techniques. In: *ETRI Journal*. [S.l.: s.n.], 1996. vol. 18, n. 1, p. 41–51. Citado 2 vezes nas páginas 21 e 22.

KEYROUZ, F.; DIEPOLD, K. A rational hrtf interpolation approach for fast synthesis of moving sound. In: *Digital Signal Processing Workshop, 12th - Signal Processing Education Workshop, 4th*. [S.l.: s.n.], 2006. p. 222–226. Citado na página 22.

KEYROUZ, F.; DIEPOLD, K. Binaural source localization and spatial audio reproduction for telepresence applications. *Presence: Teleoper. Virtual Environ.*, MIT Press, Cambridge, MA, USA, v. 16, n. 5, p. 509–522, out. 2007. Citado na página 22.

KLEINER, M.; DALENBäCK, B.-I.; SVENSSON, P. Auralization-an overview. In: *Audio Engineering Society Convention 91*. [S.l.: s.n.], 1991. Citado na página 32.

LEITE, W. et al. Avaliação cinemática comparativa da marcha humana por meio de unidade inercial e sistema de video. In: *XXIV Congresso Brasileiro de Engenharia Biomédica – CBEB 2014*. [S.l.: s.n.], 2014. Citado na página 44.

MASIERO, B. S. *Individualized Binaural Technology: Measurement, Equalization and Perceptual Evaluation*. Tese (Doutorado) — Doctoral dissertation (German), Institute of Technical Acoustics, RWTH Aachen University, 2012. 177 pages. Citado na página 31.

NEUMAN, W. R.; CRIGLER, A. N.; BOVE, V. M. Television sound and viewer perceptions. 1991. Disponível em: <http://web.media.mit.edu/~vmb/papers/russ_sound.pdf>. Citado na página 21.

OPPENHEIM, A. V.; SCHAFFER, R. W.; BUCK, J. R. *Discrete-Time Signal Processing*. 2. ed. New Jersey, USA: Prentice Hall, 1998. Citado na página 35.

PAUL, S. Binaural recording technology: A historical review and possible future developments. *Acta Acustica united with Acustica*, v. 95, n. 5, p. 767–788, 2009. Citado 2 vezes nas páginas 27 e 28.

PERLMUTTER, M.; ROBIN, L. High-performance, low cost inertial mems: A market in motion! In: *Position Location and Navigation Symposium (PLANS), 2012 IEEE/ION*. [S.l.: s.n.], 2012. p. 225–229. Citado 2 vezes nas páginas 36 e 37.

PREMERLANI, W.; BIZARD, P. *Direction Cosine Matrix IMU: Theory*. 2009. Disponível em: <http://diydrones.com/profiles/blogs/dcm-imu-theory-first-draft>. Citado na página 45.

ROTHBUCHER, M. et al. Backwards compatible 3d audio conference server using hrtf synthesis and sip. In: *Signal-Image Technology and Internet-Based Systems (SITIS), 2011 Seventh International Conference on*. [S.l.: s.n.], 2011. p. 111–117. Citado na página 22.

RUMSEY, F. *Spatial Audio*. 1st. ed. [S.l.]: Focal Press, 2001. Citado na página 30.

RUMSEY, F.; MCCORMICK, T. *Sound and Recording*. 6. ed. Oxford, UK: Focal Press, 2009. Citado 3 vezes nas páginas 26, 27 e 39.

SANTANA, D. D. S. *Estimação De Trajetórias Terrestres Utilizando Unidade De Medição Inercial De Baixo Custo E Fusão Sensorial*. Dissertação (Mestrado) — Universidade de São Paulo, Nov 2005. Citado na página 38.

VORLNDER, M. *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2007. ISBN 3540488294, 9783540488293. Citado 4 vezes nas páginas 29, 32, 33 e 34.

WADE, N. J.; DEUTSCH, D. Binaural Hearing—Before and After the Stethophone. *Acoustics Today*, ASA, v. 4, n. 3, 2008. Citado na página 27.

WANG, S. et al. Face-tracking as an augmented input in video games: enhancing presence, role-playing and control. In: *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*. New York, NY, USA: ACM Press, 2006. p. 1097–1106. Citado na página 36.

WARUSFEL, O. *Listen HRTF Database*. 2002. Disponível em: <http://recherche.ircam.fr/equipes/salles/listen/index.html>. Citado 2 vezes nas páginas 31 e 39.

WENZEL, E. M. et al. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, v. 94, n. 1, p. 111–123, 1993. Citado na página 22.

WILSON, T. V. *How Virtual Surround Sound Works*. 2007. Disponível em: <http://electronics.howstuffworks.com/virtual-surround-sound.htm>. Citado na página 30.

YANKELOVICH, N. et al. Meeting central: Making distributed meetings more effective. In: *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*. New York, NY, USA: ACM, 2004. (CSCW '04), p. 419–428. Citado na página 21.