UNIVERSIDADE DE CAXIAS DO SUL
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

# SÍNTESE DE SOM TRIDIMENSIONAL

FERNANDA CARDOSO SETTI

CAXIAS DO SUL
2015

FERNANDA CARDOSO SETTI

# SÍNTESE DE SOM TRIDIMENSIONAL

Relatório apresentado na Disciplina de Trabalho de Conclusão de Curso I (AUT0248D) do curso de Engenharia de Controle e Automação na Universidade de Caxias do Sul.

Orientador Professor Guilherme Holsbach Costa, Eng. PhD.

CAXIAS DO SUL

2015

FERNANDA CARDOSO SETTI

# SÍNTESE DE SOM TRIDIMENSIONAL

Relatório apresentado na Disciplina de Trabalho de Conclusão de Curso I (AUT0248D) na Universidade de Caxias do Sul.

**Aprovado em 08/07/2015.**

**Banca Examinadora**

Professor Doutor Diorge Alex Báo Zambra
Universidade de Caxias do Sul – UCS

Professor Doutor Luciano Hennemann
Universidade de Caxias do Sul – UCS

## RESUMO

Este trabalho estuda fundamentos do áudio tridimensional do ponto de vista das Neurociências e do Processamento Digital de Sinais. Como motivação, busca-se um sistema que possa orientar geograficamente (em um plano horizontal) um usuário a partir de um estímulo de áudio aplicado por meio de fones de ouvido. Esse áudio é reproduzido de forma a ser percebido como sendo emitido de uma fonte sonora em uma posição espacial virtual. Para a síntese do som tridimensional, utiliza-se um conjunto de Funções de Transferência Relativas à Cabeça como um filtro para um sinal de áudio monoaural a ser processado. Como resultado preliminar, foi sintetizada uma fonte sonora com deslocamento circular em torno do usuário. O deslocamento dessa fonte ainda apresenta descontinuidades perceptíveis, mas avaliações preliminares indicam que a identificação da posição espacial da fonte sonora é satisfatória para a aplicação em questão. Por fim, uma proposta de trabalhos futuros é apresentada para o Trabalho de Conclusão de Curso II.



**Palavras-chave:** Áudio Binaural, Som tridimensional, Som 3D.

## ABSTRACT

This paper presents a study of tridimensional audio fundamentals from the point of view of Neurosciences and Digital Signals Processing. As motivation, it has been looked forward to develop a system to offer a geographical orientation (at the horizontal plane) to a user from an earphone audio stimulus. This audio is reproduced to be felt coming from a virtual special position of the sound source. For reproduction of tridimensional sound, a set of Head-Related Transfer Functions is used as a filter to the monaural audio to be processed. As preliminary results, it was built a circular movement sound source system around the user. The movement of this source still presents pretty noticeable discontinuities, but preliminary assessments indicate that the identification of the sound source is satisfactory for this application. Finally, a proposal for future works at the “Trabalho de Conclusão de Curso II” is presented.



**Keywords:** Binaural Audio, Tridimensional Sound, 3D Sound.

## LISTA DE ILUSTRAÇÕES

Figura 1 - Planos de referência .....13
Figura 2 - Ângulos de azimute e elevação .....14
Figura 3 – Sistema auditivo .....18
Figura 4 – Localização vertical do som .....20
Figura 5 – Retardo temporal interaural .....21
Figura 6 – Diferença de intensidade interaural .....22
Figura 7 - Estrutura de aquisição das HRTFs .....25
Figura 8 – Ilustração da ITD com ângulo de azimute igual a 5º (a) e 45º (b) .....28
Figura 9 - Ilustração das repostas ao impulso (a) e das respostas em frequência (b) das HRTFs das posições com ângulo de azimute 0º e elevação de 0º e 10º. .....29
Figura 10 - Método de interpolação bilinear .....30
Figura 11 – Esquema da filtragem para síntese de som tridimensional .....32
Figura 12 – Posições da fonte sonora no algoritmo implementado .....33
Figura 13 - Diagrama do algoritmo para síntese de som tridimensional .....34

## LISTA DE TABELAS

Tabela 1 – HRTFs medidas e passo em azimute para cada elevação……....25

## LISTA DE SIGLAS

| | |
|---|---|
| 3D | Tridimensional |
| HRIR | *Head-Related Impulse Response*<br>Função de Transferência Relativa à Cabeça |
| HRTF | *Head-Related Transfer Function*<br>Resposta ao Impulso Relativa à Cabeça |
| ITD | *Interaural Time Delay*<br>Diferença de Tempo Interaural |
| ILD | *Interaural Level Delay*<br>Diferença de Intensidade Interaural |

## SUMÁRIO

**1 INTRODUÇÃO ..... 10**
1.1 GERAÇÃO DE ESTÍMULOS ATRAVÉS DA AUDIÇÃO ..... 11
1.2 CONCEITOS E DEFINIÇÕES ..... 12
**1.2.1 Áudio Monoaural e Binaural ..... 12**
**1.2.2 Panoramização ..... 12**
**1.2.3 Sistema de Coordenadas Polares ..... 13**
1.3 OBJETIVOS DO TRABALHO ..... 14
**1.3.1 Objetivo Geral ..... 14**
**1.3.2 Objetivos Específicos ..... 15**
1.4 ORGANIZAÇÃO DO TRABALHO ..... 15

**2 NEUROCIÊNCIAS E O SOM ..... 17**
2.1 CONCEITOS REFERENTES AO SOM ..... 17
2.2 O SISTEMA AUDITIVO ..... 18
2.3 LOCALIZAÇÃO DO SOM ..... 19
**2.3.1 Localização no Plano Vertical ..... 19**
**2.3.2 Localização no Plano Horizontal ..... 20**

**3 ÁUDIO TRIDIMENSIONAL ..... 23**
3.1 FUNÇÕES DE TRANSFERÊNCIA RELATIVAS À CABEÇA ..... 24
**3.1.1 Conjunto de Gardner e Martin ..... 25**
**3.1.2 Características ..... 27**
**3.1.3 HRTFs para Síntese de Som 3D ..... 29**
**3.1.4 Avaliação de HRTFs ..... 31**
3.2 ALGORITMO PARA SÍNTESE DE SOM 3D ..... 31

**4 RESULTADOS PRELIMINARES E CONSIDERAÇÕES ..... 35**

**5 REFERÊNCIAS ..... 37**

# 1 INTRODUÇÃO

Enquanto alguns animais, como os pássaros, possuem um senso nato de direção, ou, como os morcegos, conseguem se localizar e capturar uma presa pela reflexão de ondas ultrassônicas emitidas por ele, o ser humano possui oito sentidos: audição, olfato, paladar, tato, visão, exterocepção, percepção cinestésica e equilibriocepção (BEAR *et al*, 2008). Intrigados com essa condição humana, um grupo de professores orientou pesquisas no Instituto de Ciências Cognitivas da Universidade de Osnabrück, na Alemanha, que resultaram em um projeto denominado feelSpace. Segundo publicação na internet, essas pesquisas buscam provar que o cérebro humano é flexível o suficiente para aprender novos sentidos e expandir suas formas de percepção do mundo (FEELSPACE, 2015). Geralmente os testes realizados nesse projeto envolvem dispositivos e sensores que agregam novas informações do ambiente, além daquelas que os oito sentidos humanos já são capazes de perceber. Nos primeiros experimentos, o projeto feelSpace focou na navegação, principalmente na orientação do corpo com relação ao campo magnético da Terra, criando um cinto com dispositivos vibratórios semelhantes aos encontrados em celulares. Uma bússola eletrônica identifica o Norte e os vibratórios são ativados indicando essa direção, de modo que a pessoa tenha informação constante da orientação espacial magnética em que se encontra.

Em uma matéria para o website Wired, Udo Wächter, um usuário que testou o dispositivo por seis semanas afirma que sua percepção de direção melhorou muito. Ele acredita ter adquirido a habilidade de criar mentalmente mapas da cidade, encontrar-se com maior facilidade e que passou a ter a sensação de que não se sentiria perdido mesmo em um lugar inteiramente novo (MIXED FEELINGS, 2015). A proposta do projeto feelSpace é justamente expor o ser humano a um estímulo externo até que ele se familiarize com as informações recebidas. O cinto criado foi uma das formas encontradas de capturar e transmitir esses estímulos ao ser humano, porém diversas outras maneiras podem ser exploradas para se obter o mesmo resultado.

## 1.1 GERAÇÃO DE ESTÍMULOS ATRAVÉS DA AUDIÇÃO

Nos seres humanos, a audição contribui para o senso de localização e equilíbrio e, embora alguns animais tenham esse sentido ainda mais apurado, é através da audição que pessoas com alguma deficiência visual conseguem significar e interagir com o ambiente quase que normalmente. O som também contribui para a maneira como o ser humano sente e percebe o mundo a sua volta, motivando emoções e sentimentos. Por isso, técnicas para manipulação do mesmo têm sido estudadas, de modo a gerar a sensação de realidade virtual da localização e movimento para, por exemplo, a indústria cinematográfica, simuladores e jogos eletrônicos.

O cérebro humano utiliza algumas indicações, como a comparação do sinal recebido por cada ouvido, para perceber e identificar os sons a sua volta, bem como sua localização no espaço. Para localizar a posição de uma fonte sonora, os sons são captados pelo sistema auditivo e levados ao cérebro, onde são decodificados. Além do formato e dimensões do conjunto formado pela cabeça, tronco e pina (ouvido externo) do indivíduo, que altera a maneira como os sons são recebidos, o cérebro também utiliza duas importantes indicações para localizar uma fonte sonora: a diferença de intensidade e a diferença de tempo – ou fase – dos sinais que chegam a cada ouvido.

Por meio de Funções de Transferência Relativas à Cabeça, comumente chamadas de HRTFs (*Head-Related Transfer Functions*), é possível filtrar um áudio de maneira a produzir outro que gera essa sensação de realidade virtual da localização da fonte sonora. As HRTFs são medidas com o auxílio de um manequim que simula como o conjunto cabeça-tronco-pina alteram o som e a filtragem através delas gera um estímulo binaural que pode ser transmitido através de fones de ouvido, como é o caso dos jogos eletrônicos, ou de caixas de som, como ocorre no cinema.

Percebe-se que, assim como no projeto feelSpace, é possível inserir estímulos externos através da audição que aprimorem a maneira como o ser humano percebe sua localização espacial. O algoritmo de um aplicativo para *smartphone* que filtra uma música ouvida com fones de ouvido, oferecendo ao usuário informações contínuas da posição em que se encontra, é uma maneira de gerar estímulos externos através da audição.

## 1.2 CONCEITOS E DEFINIÇÕES

Durante este trabalho, são utilizados alguns conceitos e termos que se encontram definidos nesta seção. Por ser um tema bastante específico, algumas convenções são adotadas para facilitar o entendimento. Essas convenções serão apresentadas a seguir, de forma que seja possível definir os objetivos do trabalho com o grau de formalismo adequado.

### 1.2.1 Áudio Monoaural e Binaural

**Áudio monoaural** é um som que apresenta apenas um canal e é reproduzido para um dos ouvidos. Entre os **áudios binaurais**, que apresentam dois canais, há os dióticos e os dicóticos, que significam um estímulo igual gerado para os dois ouvidos e estímulos diferentes em cada ouvido, respectivamente (GUNZI, 2008).

Um **áudio binaural** pode ser também chamado de som tridimensional ou som 3D. Um **sinal binaural** é qualquer estímulo diótico ou dicótico.

Neste trabalho, para facilitar o entendimento dos conceitos, utilizaremos o termo “áudio monoaural” para todos os sons que apresentam um canal gravado, inclusive quando esse mesmo canal é reproduzido para os dois ouvidos (áudios binaurais dióticos).

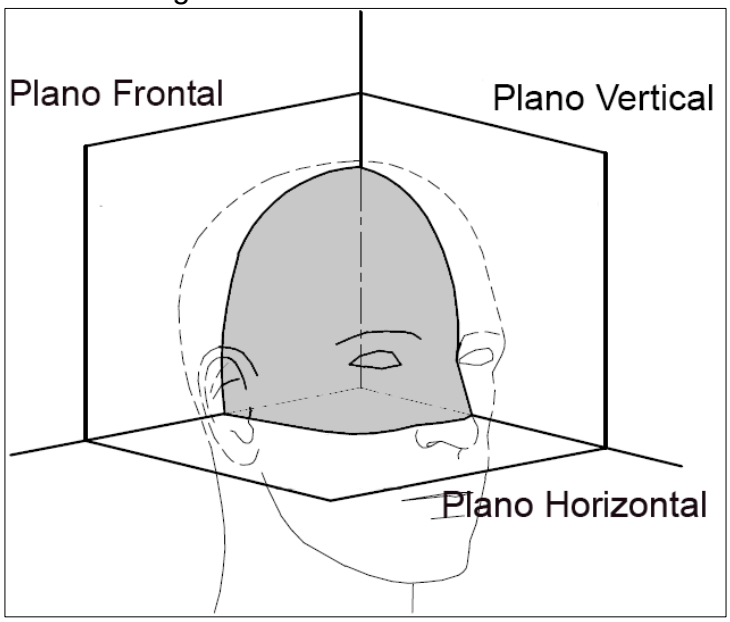
### 1.2.2 Panoramização

Panoramização, que pode ser classificada em pantofônica ou perifônica, é a definição de um método utilizado para transformar um áudio monoaural em binaural, ou seja, é o método de síntese do som 3D. A panoramização perifônica refere-se à técnica utilizada para gerar um áudio binaural com fontes sonoras posicionadas em uma esfera, enquanto a panoramização pantofônica refere-se à um método aplicado apenas no plano horizontal. A panoramização pantofônica foi o método utilizado durante este trabalho (GUNZI, 2008).

### 1.2.3 Sistema de Coordenadas Polares

Conforme a Figura 1, para este trabalho é adotado o sistema de coordenadas com os planos de referência: frontal, vertical e horizontal. O plano frontal corta a cabeça paralelamente à face passando pelo eixo imaginário que se dá pela reta que liga as duas orelhas. O plano vertical é o plano que corta ao meio a face, separando os dois lados simétricos da cabeça. O plano horizontal é o que corta horizontalmente a cabeça contendo o eixo imaginário que se dá pela reta que liga as duas orelhas.

Figura 1 - Planos de referência



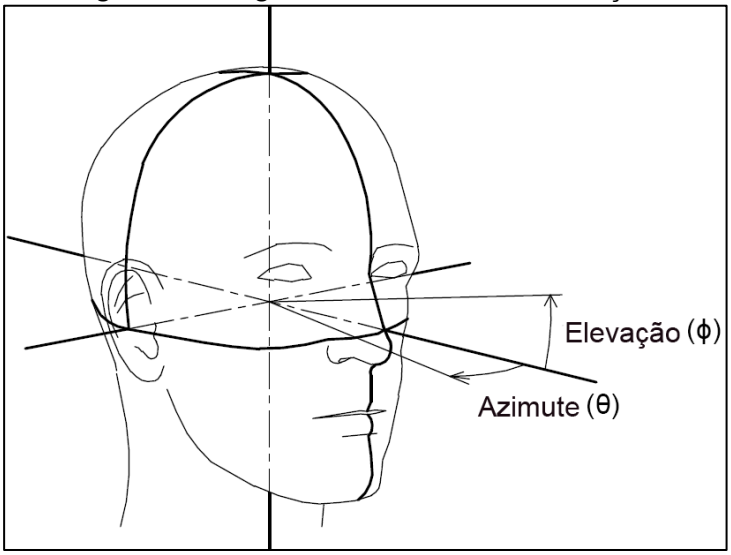

Fonte: Adaptado de Gunzi (2008)

Nesse sistema de coordenadas polar, qualquer ponto em uma esfera com centro na intersecção dos planos Horizontal, Frontal e Vertical e raio *r* pode ser descrito com a informação de dois ângulos: o ângulo de azimute e o ângulo de elevação. Os ângulos de azimute e elevação podem ser observados na Figura 2.

O **ângulo de azimute**, indicado nesse trabalho pela letra grega teta (θ), refere-se ao ângulo observado no plano horizontal. O ângulo de azimute é zero quando a fonte está exatamente em frente à cabeça, na direção do nariz da pessoa. O azimute é positivo no lado direito da cabeça e compreende a posição virtual da fonte sonora em 360º.

O **ângulo de elevação**, menos utilizado nesse trabalho, é indicado pela letra grega fi (ϕ). O ângulo de elevação é observado no plano vertical, é zero quando a fonte está na mesma altura da cabeça e positivo quando está acima do plano horizontal.

Figura 2 - Ângulos de azimute e elevação




Fonte: Adaptado de Gunzi (2008)

## 1.3 OBJETIVOS DO TRABALHO

Esta seção apresenta os principais objetivos desse trabalho, que pretende estudar um sistema de síntese de áudio tridimensional para fones de ouvido a partir de áudios monoaurais, pensando em utilizar esse sistema aplicado à geração de estímulos de localização geográfica do ouvinte no plano horizontal.

### 1.3.1 Objetivo Geral

Implementar um sistema de síntese de áudio binaural para fones de ouvido com o qual um usuário seja capaz de identificar corretamente a posição espacial da fonte de sinal sintetizada.

### 1.3.2 Objetivos Específicos

a) Entender os fundamentos do áudio 3D do ponto de vista das Neurociências e de Processamento de Sinais;
b) Implementar uma técnica de síntese de áudio 3D para fones de ouvido, em linguagem de alto nível; e
c) Avaliar o desempenho da técnica implementada quanto à probabilidade de identificação da posição da fonte pelo usuário.

## 1.4 ORGANIZAÇÃO DO TRABALHO

Este trabalho contribui com a concepção de uma forma de transmitir estímulos externos de direção através da síntese de áudio tridimensional a partir da avaliação do sistema auditivo. Essa técnica de filtragem para síntese de som binaural será construída para fones de ouvido, já que são acessórios muito utilizados por usuários de *smartphones*. O presente trabalho encontra-se organizado da seguinte forma:

Neste primeiro capítulo abordou-se:

- Justificativa e motivações;
- Conceitos e definições fundamentais utilizados durante o trabalho; e
- Objetivos e contribuições que o presente trabalho visa atingir.

No segundo capítulo “Neurociências e o Som” apresentam-se conceitos do ponto de vista das neurociências. São esses:

- Conceitos referentes ao som;
- Sistema auditivo e a percepção do som; e
- Indicações utilizadas pelo cérebro para identificar a localização de uma fonte sonora.

No terceiro capítulo “Áudio Tridimensional” apresenta-se:

- Funções de transferência relativas à cabeça;
- O conjunto de HRTFs de Gardner e Martin;
- Utilização das HRTFs para processamento de áudio tridimensional; e
- Algoritmo preliminar para síntese de áudio tridimensional.

Por fim, no Capítulo 4, “Conclusões e Resultados” são apresentados:

- Resultados obtidos com o algoritmo apresentado no Capítulo 3;

- Percepções referentes à implementação do algoritmo;
- Melhorias a serem desenvolvidas no algoritmo; e
- Trabalho e resultados que se espera construir ao longo do Trabalho de Conclusão de Curso II.

# 2 NEUROCIÊNCIAS E O SOM

Através da audição, muitas vezes é possível detectar a presença, localizar a origem e reconhecer um objeto, mesmo quando não se pode vê-lo. Dessa forma pessoas com deficiências visuais se comunicam com o mundo e, além das funções de comunicação, a audição diz respeito às percepções humanas, como para os músicos que fazem uso do som para transmitir sensações.

Neste capítulo, é apresentada a natureza do som, como ele é percebido pelo cérebro humano e como é possível fazer uso dos conceitos das neurociências para manipulação do som tridimensional.

## 2.1 CONCEITOS REFERENTES AO SOM

Segundo Bear *et al* (2008), basicamente tudo que pode mover moléculas do ar, tornando-as mais densas ou rarefeitas, produz algum som. Os ciclos de compressão e descompressão dessas moléculas são caracterizados por uma frequência – mensurada em hertz [Hz] –, que é definida pela fonte emissora deste som. O som é, então, uma onda que se propaga mecanicamente, tendo no ar uma velocidade em torno de 340 m/s.

Essa frequência de vibração distingue os diferentes sons que percebemos. O som de alta frequência, formado pelos tons mais agudos, é caracterizado por intervalos curtos de compressão e descompressão, enquanto o som de baixa frequência, formado por tons mais graves, possui um período maior nas ondas de pressão. O sistema auditivo humano possui a capacidade de detectar sons entre 20 Hz e 20.000 Hz, sendo essas frequências de sons graves e agudos, respectivamente (DO CARMO, 1999).

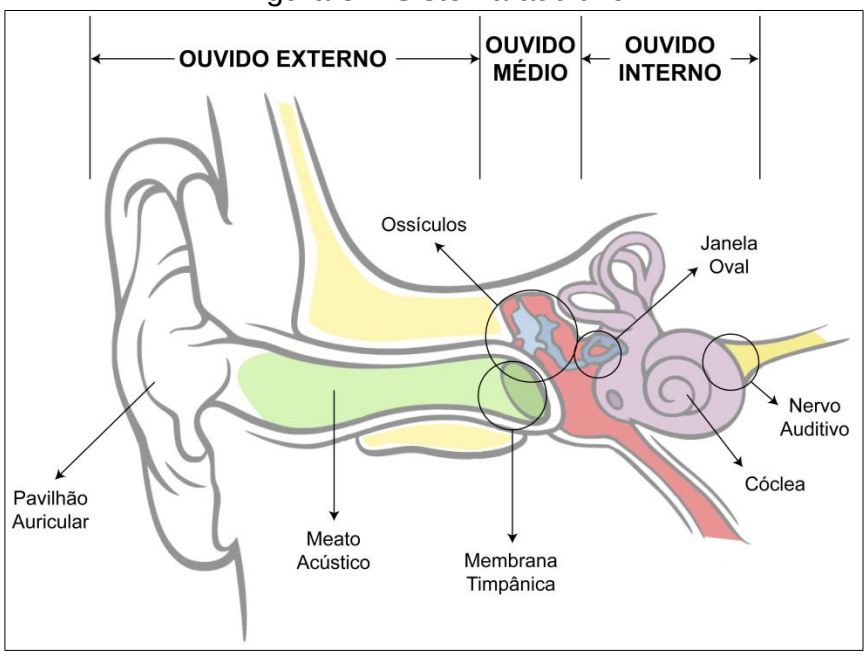
Todos os sons captados pelo sistema auditivo possuem uma combinação de frequências e intensidades que caracteriza aquele som e sua origem. Então, além da frequência, outra característica importante para o som é a intensidade, ou seja, o volume do ar deslocado nos ciclos de compressão e descompressão. A intensidade é a amplitude da onda de pressão do ar, o que determina o "volume" do som, sendo os sons de volume mais alto, os de maior intensidade. Bear *et al* (2008) destaca que o som mais intenso que uma pessoa ouve pode chegar a um trilhão de vezes a amplitude do som menos intenso possível de ser percebido. Como essa diferença é

muito grande, convenciona-se mensura-la na escala logarítmica de decibéis [dB], portanto o ser humano é capaz de ouvir sons entre 20 e 120 dB, sendo que próximo a 120 dB já se começa a sentir certo desconforto. Ainda, uma pessoa exposta continuamente a sons superiores a 85 dB pode ter perdas auditivas graduais e, quando exposta a sons superiores a 140 dB, sente dor e pode sofrer lesões, até mesmo irreversíveis, ao sistema auditivo (DO CARMO, 1999).

## 2.2 O SISTEMA AUDITIVO

O sistema auditivo é formado por diversos componentes que, apesar de estarem interligados, são comumente separados em três conjuntos com funções distintas: ouvido externo, ouvido médio e ouvido interno (BEAR *et al*, 2008). O ouvido externo é responsável por captar os sons e é formado pelo pavilhão e canal auditivo. O ouvido médio é formado pela membrana timpânica e ossículos, que transformam a vibração causada pelo som em movimentos mecânicos, de modo a ampliá-los para posterior decodificação. O ouvido interno é responsável por interpretar os movimentos mecânicos dos ossículos e enviá-los aos neurônios sensoriais, sendo formado principalmente pela janela oval e cóclea. Na Figura 3 é possível observar os três mecanismos e seus principais componentes.

Figura 3 – Sistema auditivo




Fonte: A autora (2015)

O som no cérebro é decodificado pelo nervo vestíbulo-coclear, na Figura 3 identificado como “Nervo Auditivo”, portanto alguns casos de surdez podem ser tratados utilizando mecanismos que agem diretamente sobre essa parte do sistema auditivo, de modo que o som não precise passar pelo ouvido médio. Cada neurônio tem uma frequência característica a qual ele é mais sensível e que o ser humano possui células com funções bastante específicas. Os neurônios binauriculares, por exemplo, recebem informações dos dois lados do tronco encefálico e, por isso, são importantes para a percepção de localização de fontes sonoras (BEAR *et al*, 2008).

Quanto à percepção do som, como mencionado, cada som percebido pelo cérebro contém intensidade, frequência e origem características. Segundo Bear *et al* (2008), a informação de intensidade se dá de acordo com o número de neurônios ativos e a frequência de disparo dos mesmos. Com relação às diferenças de frequência, o cérebro utiliza a tonotopia e a sincronia de fase para identificar a natureza do som. A percepção de origem do som será detalhada na seção seguinte, “Localização do Som”.

## 2.3 LOCALIZAÇÃO DO SOM

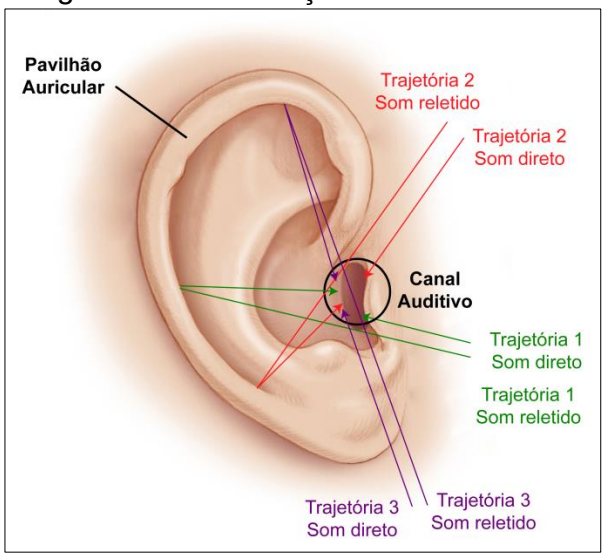
Identificar a origem de um determinado som é muito importante, por exemplo, quando uma pessoa precisa atravessar a rua ou quando alguém a chama. De acordo com Bear *et al* (2008), duas são as principais indicações que o cérebro utiliza para identificar a localização do som no plano horizontal: o retardo temporal interaural, comumente chamado de ITD (em inglês, *Interaural time difference*) e a diferença de intensidade interaural, chamada de ILD (em inglês, *Interaural level difference*). Esses dois processos caracterizam a Teoria Duplex de localização do som e ainda, segundo o autor, com apenas um dos ouvidos as pessoas são capazes de identificar a origem de um som no plano vertical, mas não no plano horizontal, o que sugere que seja feita uma comparação entre estímulos oriundos dos dois ouvidos quando se quer distinguir a origem de um som no plano horizontal.

### 2.3.1 Localização no Plano Vertical

Para localização do som no plano vertical, segundo Bear *et al* (2008), os neurônios fazem uso das deformações do som provindas das curvas sinuosas do

ouvido externo. Os retardos de tempo entre a via direta e a refletida se alteram de acordo com a posição da fonte sonora, conforme Figura 4.

Figura 4 – Localização vertical do som



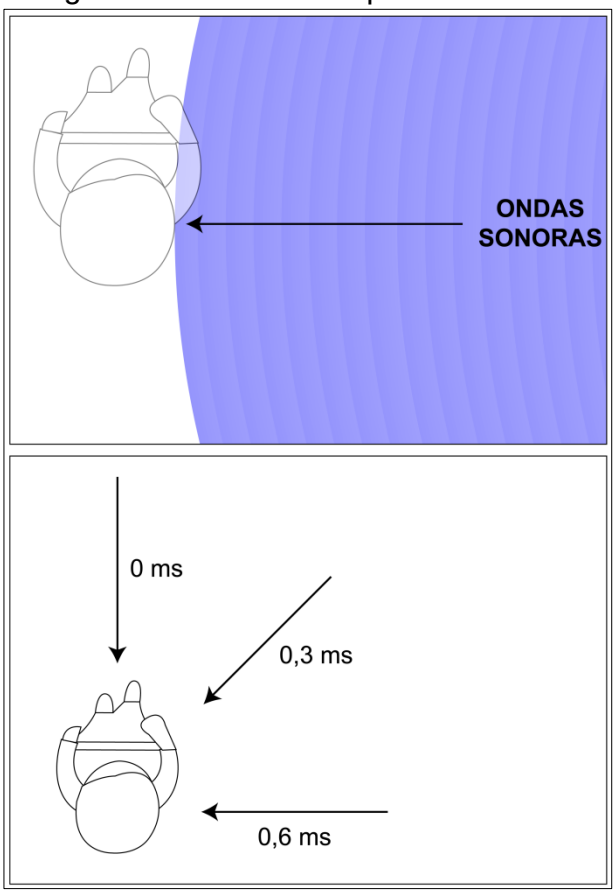

Fonte: A autora (2015)

### 2.3.2 Localização no Plano Horizontal

Os neurônios binauriculares, de acordo com Bear *et al* (2008), são responsáveis pela comparação e interpretação dos estímulos que alcançam os dois ouvidos e, por isso, pela identificação do som no plano horizontal. A localização do som no plano horizontal é dada pela Teoria Duplex de Localização do Som que consiste na comparação do instante de tempo (ITD) e da intensidade (ILD) do som que chega a cada orelha.

O retardo temporal interaural é uma importante dica para se localizar sons no plano horizontal. Ele acontece quando a fonte sonora está à esquerda ou à direita do ouvinte e o som chega a um ouvido com atraso em relação ao outro. Segundo Bear *et al* (2008) a distância média entre as orelhas é de 20 cm, o que geraria um atraso aproximado de 0,3 ms quando a origem do som está a +/-45º no plano horizontal e 0,6 ms quando a fonte está imediatamente a esquerda ou a direita do ouvinte considerando a posição 0º diretamente em frente ao indivíduo, de acordo com a Figura 5.

Figura 5 – Retardo temporal interaural




Fonte: A autora (2015)

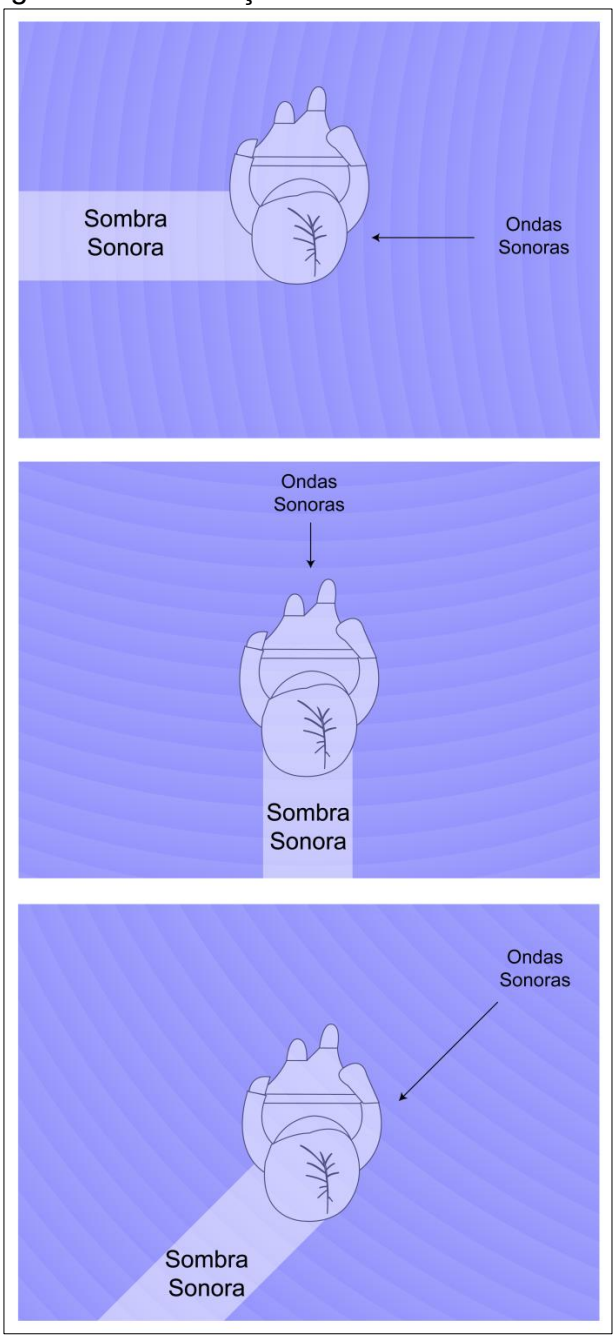
Quando o som ouvido é contínuo, ou seja, não é repentino, não é possível detectar o momento inicial quando as ondas chegaram aos ouvidos, portanto somente é possível comparar o atraso da fase de uma onda em relação à outra. Nesse caso, é possível utilizar a ITD para comprimentos de onda maiores do que 20 cm, que é a distância média entre as orelhas (BEAR *et al*, 2008).

Para sons de altas frequências, nesse caso considerados como sons de comprimentos de onda menores que a distância entre as orelhas, não é confiável utilizar a ITD. Portanto, para frequências superiores a aproximadamente 2.000 Hz, o cérebro utiliza a diferença de intensidade interaural.

A ILD consiste em um processo que utiliza uma relação direta entre a origem do som e a direção para onde a cabeça projeta o que Bear *et* al (2008) chama de sombra sonora para um dos ouvidos, conforme exemplos ilustrados na Figura 6. Devido a essa sombra sonora, em um dos ouvidos a intensidade do som que chega é menor que a que chega ao ouvido mais próximo da fonte. Essas sombras não

acontecem em sons de frequência mais baixa já que essas ondas sofrem difração e acabam contornando a cabeça.

Figura 6 – Diferença de intensidade interaural




Fonte: A autora (2015)

# 3 ÁUDIO TRIDIMENSIONAL

Normalmente os sons são monoaurais, ou seja, gravados com apenas um sinal que normalmente é transmitido por dois canais. Com o objetivo de se criar a ilusão de um ambiente cujos sons tenham diferentes origens, a síntese de um áudio 3D, ou áudio binaural, é realizada por meio da modificação de um sinal monoaural. Através das técnicas de geração de áudio binaural se busca manipular completamente a percepção espacial do ouvinte.

Diversos são os métodos utilizados na síntese do som 3D, cada um para uma determinada aplicação. Esses métodos podem ser divididos em dois grupos no que diz respeito à reprodução do som: reprodução em caixas acústicas e reprodução em fones de ouvido. Os métodos para geração em caixas acústicas ainda podem ser divididos nos que utilizam dois canais e nos que utilizam mais do que dois canais (FREELAND, 2001).

Com relação às técnicas utilizadas para síntese de som tridimensional, é preciso considerar as responsáveis pela localização espacial do som e as responsáveis pela sua ambientação. As responsáveis pela localização buscam, através da manipulação das indicações utilizadas pelo nosso cérebro, conforme visto na Seção 2.3, gerar sinais que simulem a origem de um determinado som. De acordo com Freeland (2001), as técnicas referentes à ambientação, no entanto, preocupam-se em simular os modificadores do som que são externos ao sistema cabeça-tronco-pina para cada posição da fonte sonora.

Pela vasta área de aplicação do assunto, se optou por focar na elaboração de um sistema que sintetize áudio tridimensional para reprodução em fones de ouvido. Também, por esse motivo, são tecidos poucos comentários referentes à reprodução ou sistemas para caixas acústicas, bem como não são apresenta-se soluções já utilizadas pelo mercado nessa área.

Para a síntese de som 3D a ser gerada por fones de ouvido é necessário prever as modificações sofridas pelo som como resultado das reflexões que ocorrem no tronco e na pina, bem como pela difração causada pela cabeça (FREELAND, 2001). As informações de como o conjunto cabeça-tronco-pina alteram o som podem ser representadas por funções de transferência que levam em consideração a posição da fonte sonora e cada um dos ouvidos. Essas funções de transferência

são conhecidas como Funções de Transferência Relativas à Cabeça (*Head-Related Transfer Functions*), referidas nesse trabalho pela sigla HRTF.

De acordo com Gunzi (2008), os modelos de HRTF podem ser individuais, ou seja, específicos para um indivíduo, ou não-individuais, que são medições realizadas com um manequim de dimensões médias. Ainda, podem ser obtidas por modelagem numérica — com simulações numéricas do indivíduo e na fonte — ou utilizando modelos físicos simplificados — buscando-se uma solução analítica de equações que preveem os fenômenos físicos envolvidos.

Neste capítulo serão explicadas as Funções de Transferência Relativas à Cabeça, bem como suas características e aspectos referentes ao conjunto aqui utilizado. Por fim, será apresentado o algoritmo utilizado para a geração de áudio binaural como uma prova de conceito a ser aprimorada durante o Trabalho de Conclusão de Curso II.

## 3.1 FUNÇÕES DE TRANSFERÊNCIA RELATIVAS À CABEÇA

Uma *Head-Related Transfer Function* (HRTF) é um conjunto de medições que prevê as modificações sofridas pelo som entre a fonte sonora e o conjunto cabeça-tronco-pina. HRTFs são necessárias para qualquer técnica de síntese de áudio tridimensional a ser reproduzido por fones de ouvido.

Apesar das HRTFs serem diferentes para cada pessoa, diversos modelos foram obtidos com a utilização de manequins que simulam um humano com dimensões médias. Gunzi (2008) realizou um estudo de análise e comparação entre três conjuntos de HRTF: conjunto adquirido por Gardner e Martin, conjunto do Projeto Listen e o conjunto da CIPIC. Gunzi (2008) concluiu na pesquisa que não há diferenças significativas entre os resultados obtidos para cada modelo.

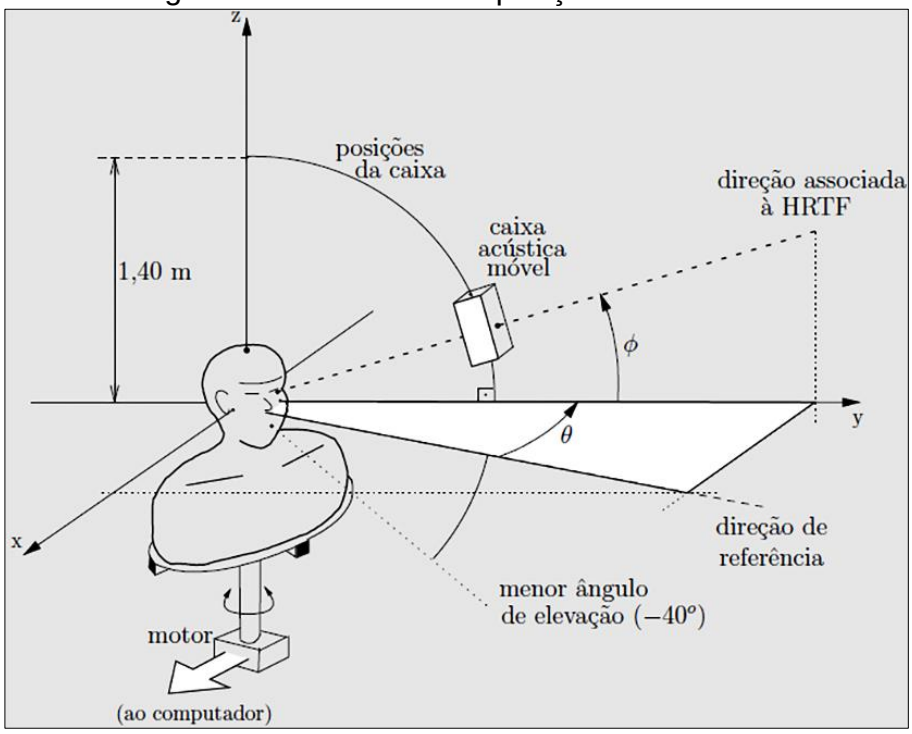
Para este trabalho, então, optou-se por utilizar o modelo obtido no MIT Media Lab e publicado por Gardner *et al* (1994) com o manequim KEMAR Dummy-Head Microphone, que foi disponibilizado para *download* em uma plataforma *online*. Nesta seção pretende-se explorar a maneira como esse conjunto foi adquirido e alguns aspectos relevantes do modelo para a síntese do som 3D.

### 3.1.1 Conjunto de Gardner e Martin

O conjunto de HRTFs utilizado nesse trabalho é o adquirido por Gardner *et al* (1994) e foi obtido em uma câmara anecoica utilizando uma cabeça artificial e um sistema de caixas acústicas. Com uma placa de aquisição de áudio acoplada a um computador adquiriu-se respostas ao impulso para determinadas posições de azimute e elevação da fonte sonora.

Quanto à configuração utilizada, o manequim é colocado em uma base giratória que garante o ângulo de azimute com alta precisão. A caixa acústica é fixada em um trilho que a mantém à distância fixa de 1,4 metros do manequim e varia sua posição garantindo um preciso ângulo de elevação entre -40º e 90º (FREELAND, 2001). Na Figura 7 é possível observar um esquema da estrutura utilizada para aquisição das HRTFs.

Figura 7 - Estrutura de aquisição das HRTFs




Fonte: Freeland (2001)

Com essa configuração, são feitas as medições das respostas ao impulso relacionadas às HRTFs. É gerada uma sequência de excitação chamada de ML (*Maximum Length*) e os sinais são simultaneamente gravados pelos microfones posicionados nos ouvidos do manequim, captando o som que provavelmente os

tímpanos captariam. Para cada posição obteve-se um número de 16383 amostras com frequência de amostragem igual a 44100 Hz que geravam as respostas impulsivas (HRIR – *Head-Related Impulse Response*) às respectivas HRTFs. Observou-se que grande parte das amostras poderia ser descartada, pois não agregava informação relevante ao modelo. As amostras iniciais que foram retiradas referiam-se ao atraso entre as operações de início da reprodução/início da gravação e as amostras finais, às reflexões causadas pela própria configuração. Sendo assim, têm-se, ao final, 512 amostras para cada uma das 710 posições medidas (GARDNER *et al*, 1994).

Com relação à quantidade de posições medidas, o ângulo de elevação foi determinado entre -40º a 90º com passo de 10º. Para cada ângulo de elevação estabeleceu-se um passo em azimute, conforme Tabela 1. Para 0º de elevação, ângulo utilizado para este trabalho, o passo medido foi de 5º.

Tabela 1 – HRTFs medidas e passo em azimute para cada elevação

| Elevação (ϕ) [º] | Número de Medidas | Passo do Azimute (θ) [º] |
|---|---|---|
| -40 | 56 | 6,43 |
| -30 | 60 | 6 |
| -20 | 72 | 5 |
| -10 | 72 | 5 |
| 0 | 72 | 5 |
| 10 | 72 | 5 |
| 20 | 72 | 5 |
| 30 | 60 | 6 |
| 40 | 56 | 6,43 |
| 50 | 45 | 8 |
| 60 | 36 | 10 |
| 70 | 24 | 15 |
| 80 | 12 | 30 |
| 90 | 1 | - |

Fonte: A autora (2015)

Quanto à cabeça artificial utilizada nas medições, o modelo é da fabricante KEMAR - *Knowles Electronics Mannequin for Acoustic Research*, DB-4004. Ainda, nesse manequim foram utilizados dois modelos de pina: para o ouvido esquerdo DB-061 e para o ouvido direito um modelo um pouco maior, DB-065, gerando dois modelos distintos. O conjunto do microfone e pré-amplificador é do modelo Etymotic ER-11 (GARDNER *et al*, 1994). Esse amplificador é conectado a uma placa Audiomedia com conversor A/D, controlada pelo computador.

É importante ressaltar que as medições foram realizadas com ângulo de azimute variando de 0 a 360º, pois os modelos de pina são diferentes. Caso não houvesse essa diferença, as medições poderiam ser realizadas em 180º de azimute e refletidas de maneira simétrica para os outros 180º, invertendo-se os canais. O banco de dados disponibilizado por Gardner *et al*, 1994 consiste, portanto, em dois conjuntos de medições distintos para dois tamanhos de pina.

Por fim, de acordo com Freeland (2001), uma série de fatores que poderiam influenciar a medição das HRTFs foi ponderada. Entre os fatores que poderiam prejudicar a fidelidade das medições estão as respostas da caixa acústica, dos amplificadores, da sala onde a estrutura foi instalada e do microfone. Os fones de ouvido utilizados pelo indivíduo também podem influenciar a precisão do conjunto de HRTFs, porém considerando-se o amplo universo de modelos e qualidade dos equipamentos, essa influência não foi prevista no conjunto.

Percebeu-se, então, que o único fator realmente crítico é a resposta da caixa acústica, por isso ela foi medida separadamente. Em um segundo banco de dados disponibilizado por Gardner *et al* (1994) essas HRIRs da caixa acústica foram subtraídas das HRIRs do modelo e percebeu-se que as a amplitude das HRIRs mostravam valores baixos em um tempo mais curto, por isso esse segundo banco de dados possui somente 128 amostras.

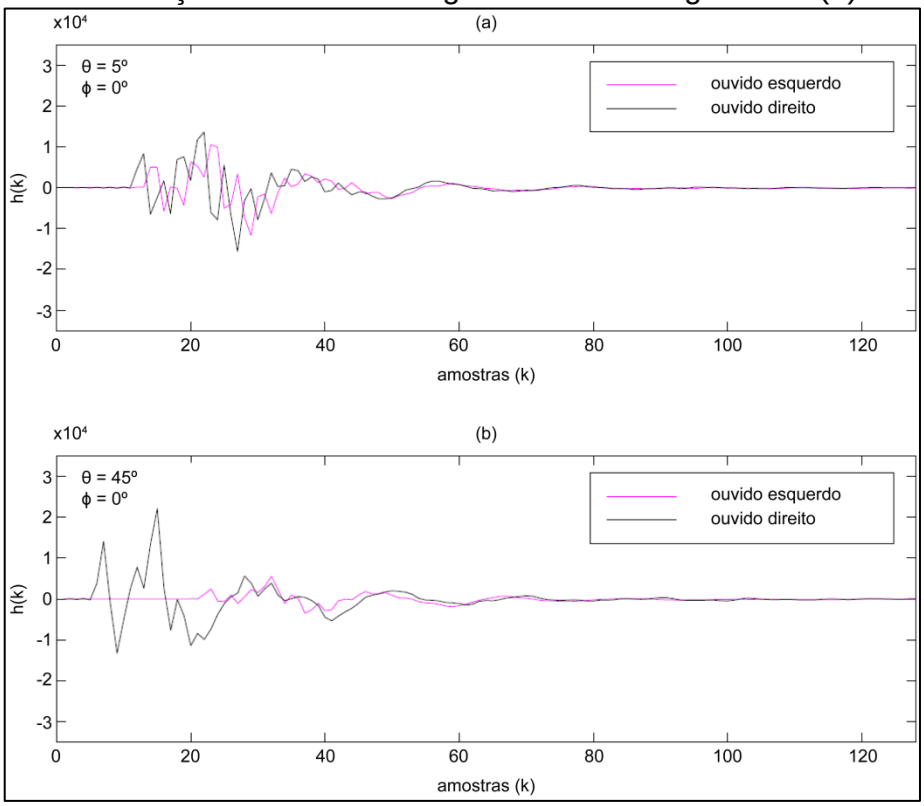
### 3.1.2 Características

Conforme visto na Seção 2.3, as duas principais indicações utilizadas pelo cérebro para identificar a origem de uma fonte sonora são a diferença de intensidade interaural (ILD) e a diferença de tempo interaural (ITD) do sinal recebido, além das reflexões no conjunto cabeça-tronco-pina para percepção do som no plano vertical. Embora as HRTFs variem bastante entre as pessoas devido à geometria do

conjunto cabeça-tronco-pina e ao modelo utilizado levar em conta as dimensões médias de uma pessoa, alguns padrões de percepção do som foram percebidos e registrados no estudo realizado por Freeland (2001).

A ITD é a diferença entre o tempo de chegada do som aos dois ouvidos e pode ser observada nas respostas ao impulso das HRTFs da Figura 8. Na Figura 8a é possível observar as HRIRs para o ângulo de azimute 5º e na Figura 8b, para 45º, ambas com elevação igual a 0º. Percebe-se que a diferença de tempo Interaural é muito maior entre os ouvidos no ângulo de 45º do que no outro caso.

Figura 8 – Ilustração da ITD com ângulo de azimute igual a 5º (a) e 45º (b)



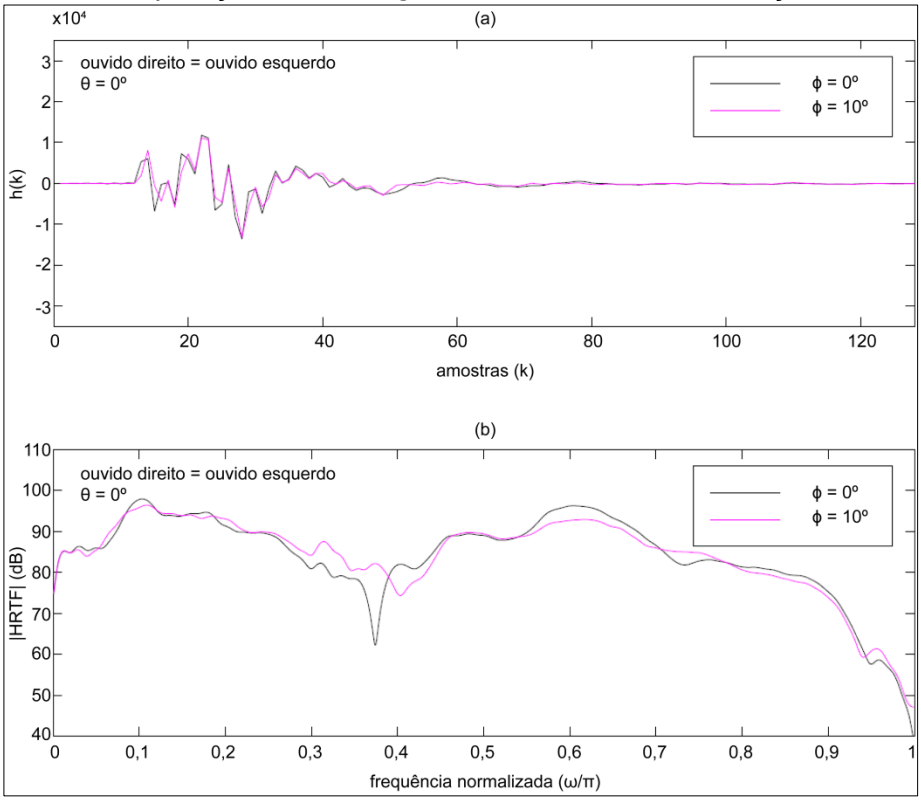

Fonte: A autora (2015)

Ainda na Figura 8 é possível observar o efeito da ILD, que indica a diferença da intensidade do som que chega a cada ouvido. Observa-se que quanto maior o ângulo de azimute, ou seja, quanto mais próxima a fonte sonora está de um dos lados da cabeça, maior é a “sombra sonora” (efeito abordado na Seção 2.3.2 e ilustrado na Figura 6) que a cabeça faz para um dos ouvidos.

Freeland (2001) comparou as respostas ao impulso também em casos com azimute fixado em 0º e variação do ângulo de elevação. Constatou, assim, que as

diferenças de tempo e intensidade são quase imperceptíveis nesses casos, quando a geometria do conjunto cabeça-tronco-pina e as reflexões ocasionadas por ele influenciam muito mais as HRIRs. Então, para localização da fonte sonora em variação do ângulo de elevação, utiliza-se o módulo da resposta em frequência. Na Figura 9a é possível observar o comportamento da ITD e ILD quando há variação do ângulo de elevação e na Figura 9b a resposta em frequência para esse caso, destacando a importância dos vales e picos do gráfico para identificar a posição da fonte sonora.

Figura 9 - Ilustração das repostas ao impulso (a) e das respostas em frequência (b) das HRTFs das posições com ângulo de azimute 0º e elevação de 0º e 10º.




Fonte: A autora (2015)

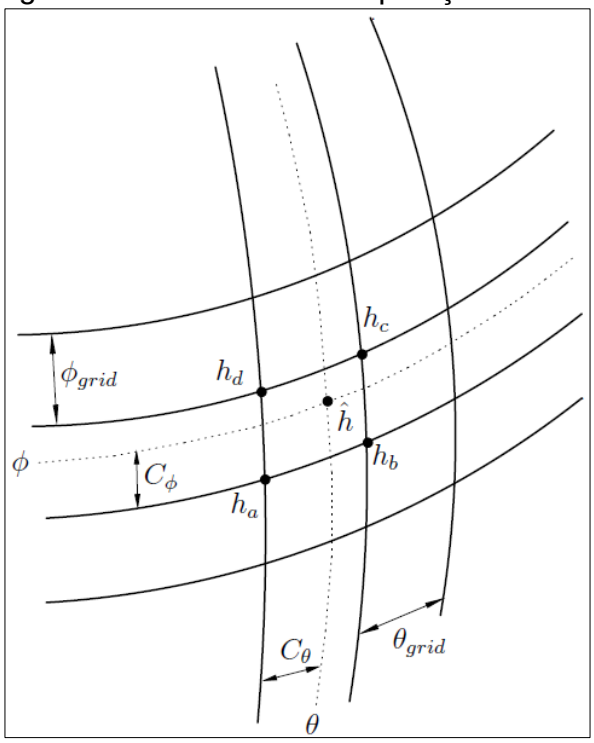
### 3.1.3 HRTFs para Síntese de Som 3D

Como mencionado anteriormente, as HRTFs são essenciais para a síntese do som tridimensional em fones de ouvido, pois suas respostas ao impulso são utilizadas como filtros que adicionam a informação de localização a um sinal qualquer. Cada ouvido (ou canal desse sinal), portanto, precisa de um filtro distinto

para que o cérebro compreenda as diferenças de intensidade e de tempo como aconteceria com qualquer som que a pessoa ouve naturalmente.

Para cada posição da fonte sonora obtêm-se duas HRIRs ou dois filtros distintos, uma para cada ouvido. Acaba não sendo possível gerar funções de transferência para todos os pontos no espaço, então quando há a necessidade de se gerar a sensação de movimento no som, pode ser necessário realizar uma interpolação utilizando os pontos medidos. Os métodos mais comuns de interpolação para panoramização perifônica são o método bilinear (ilustrado na Figura 10) e o método bilinear esférico, ambos com a possibilidade de implementação do domínio do tempo e no domínio da frequência (FREELAND, 2001).

Figura 10 - Método de interpolação bilinear




Fonte: Freeland (2001)

O método bilinear no domínio do tempo leva em consideração as quatro HRTFs medidas mais próximas do ponto onde se vai posicionar a fonte sonora virtual e é dada pela equação

$$\hat{h}(k) = (1 - c_\theta)(1 - c_\varphi\text{-})h_a(k) + c_\theta(1 - c_\varphi\text{-})h_b(k) + c_\theta c_\varphi h_c(k) + (1 - c_\theta)c_\varphi h_d(k), \quad (1)$$

onde $h_a(k), h_b(k), h_c(k)$ e $h_d(k)$ são os pares de HRTFs mais próximos da posição $\hat{h}(\theta, \varphi)$ e $c_\theta$ e $c_\varphi$ são os coeficientes da interpolação, dados por:

$$\begin{cases} c_\theta = \frac{C_\theta}{\theta_{grid}} \\ c_\varphi = \frac{C_\varphi}{\varphi_{grid}} \end{cases}, \tag{2}$$

sendo que $C_\theta$ e $C_\varphi$ indicam a posição em graus de $\hat{h}(k)$ e $\theta_{grid}$ e $\varphi_{grid}$ o passo dos quadrantes de azimute e elevação, conforme Figura 10 (FREELAND, 2001).

### 3.1.4 Avaliação de HRTFs

Em Freeland (2001) são tecidos comentários referentes à avaliação das HRTFs. Alguns métodos objetivos de avaliação, que envolvem a estimação de fatores como as diferenças de tempo e intensidade Interaural e a comparação com o modelo, já foram criados, no entanto geralmente são utilizados testes subjetivos para verificar a validade dos conjuntos.

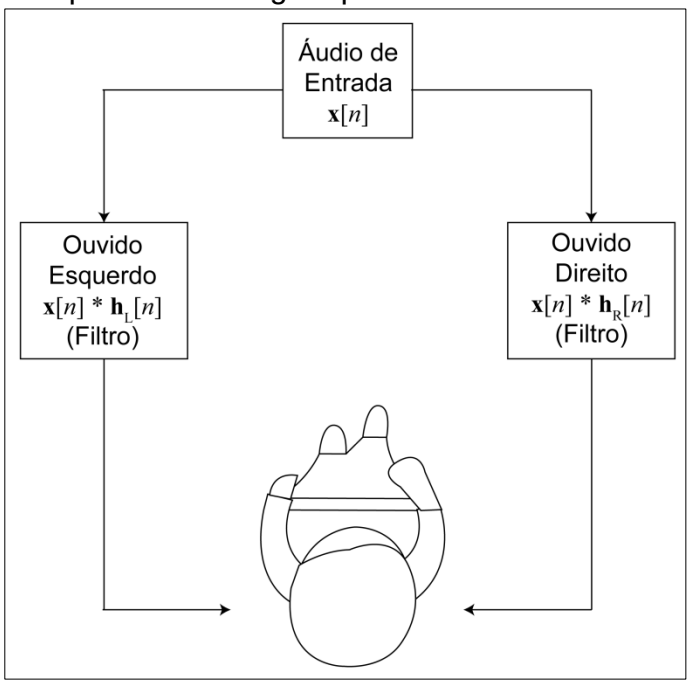
Os métodos subjetivos geralmente envolvem voluntários com fones de ouvido e venda nos olhos para eliminar quaisquer influencias visuais. Através dos fones de ouvido, geram-se então áudios-teste com a fonte sonora posicionada aleatoriamente e o voluntário é convidado a indicar qual é a direção da fonte sonora para cada áudio. A partir das respostas dos voluntários e de um método estatístico, avalia-se a validade do conjunto de HRTFs modelado (FREELAND, 2001).

## 3.2 ALGORITMO PARA SÍNTESE DE SOM 3D

Como prova dos conceitos apresentados até então no trabalho, elaborou-se um algoritmo para síntese de som tridimensional. Esse algoritmo consiste na leitura de um sinal de áudio, extração de um dos canais (quando esse sinal for estéreo) e filtragem desse canal para cada um dos ouvidos, gerando um áudio binaural que é gravado (visto que a implementação em questão não é em tempo-real) para ser reproduzido em fones de ouvido. O esquema do algoritmo de síntese do som é ilustrado na Figura 11, em que $\mathbf{h}_L[n]$ e $\mathbf{h}_R[n]$ são as respostas ao impulso referentes

às HRTFs dos ouvidos esquerdo e direito, respectivamente. O símbolo “*” corresponde à operação de convolução.

Figura 11 – Esquema da filtragem para síntese de som tridimensional

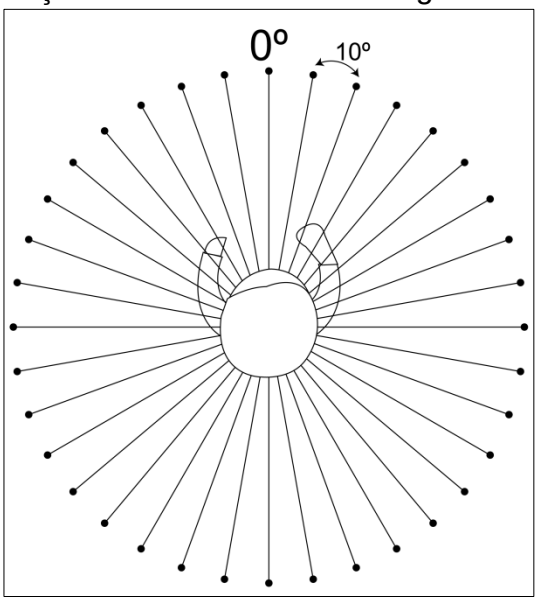



Fonte: A autora (2015)

Com relação ao áudio utilizado para geração do som tridimensional, optou-se por utilizar uma música qualquer que, quando lida pelo algoritmo, geralmente possui dois canais de informação. Somente um desses canais é selecionado para a filtragem, obtendo-se, portanto, um áudio monoaural que é duplicado para posterior processamento individual de cada ouvido. O sinal de áudio, então, é dividido em *frames* de 2 segundos.

Quanto às HRTFs utilizadas, o conjunto de dados utilizado é o completo, com 512 amostras. Definiu-se ângulo de elevação zero e variação do ângulo de azimute com passo de 10º, iniciando em 0º e realizando 360º, conforme Figura 12. Para cada posição da fonte sonora existem duas respostas impulsivas, uma para cada ouvido. Todas essas respostas são lidas e armazenadas no algoritmo. Ainda, vale ressaltar que se optou por utilizar as medições de resposta ao impulso dos dois modelos de pina, ou seja, as HRTFs foram testadas da maneira como foram medidas.

Figura 12 – Posições da fonte sonora no algoritmo implementado



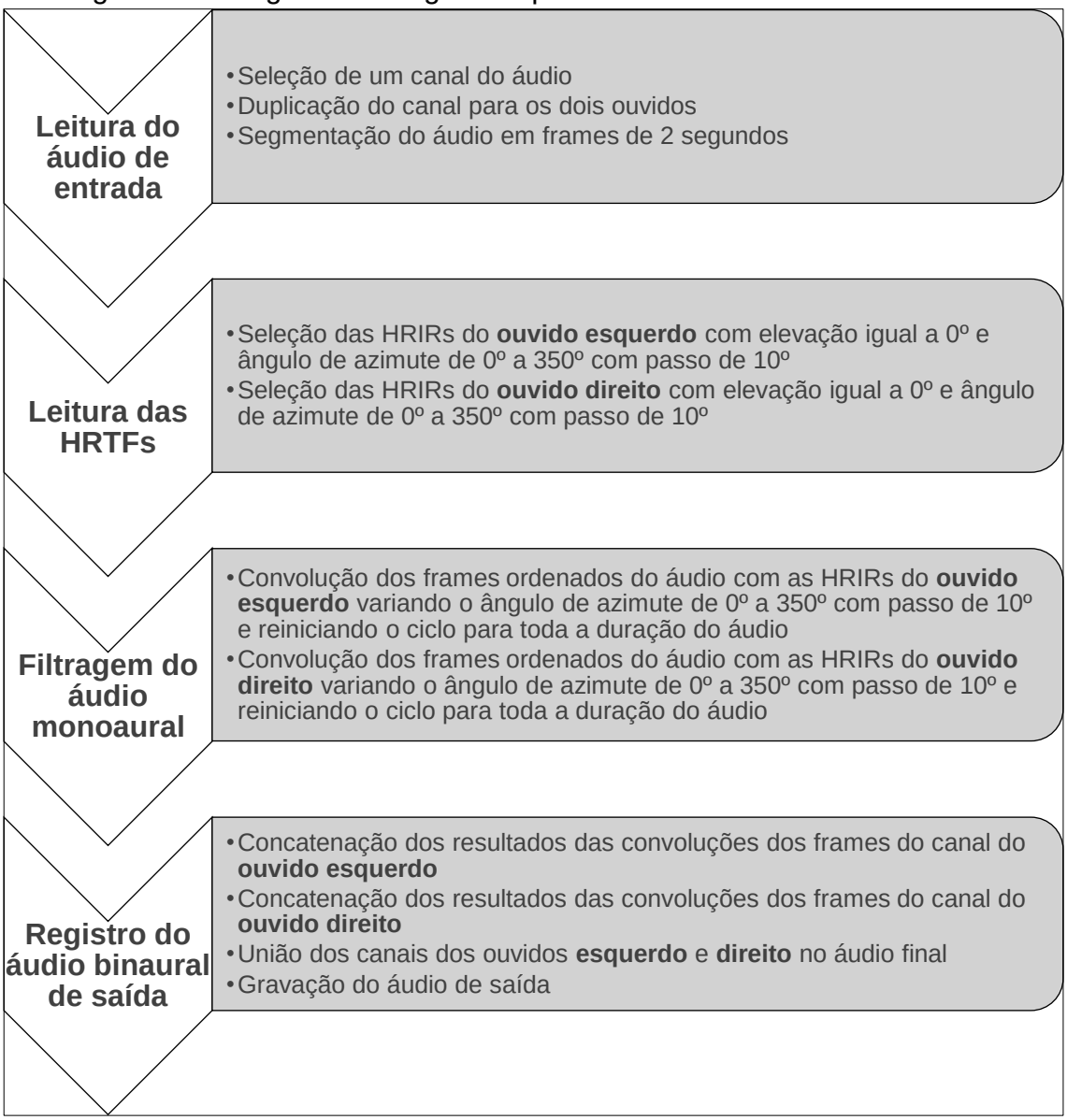

Fonte: A autora (2015)

O áudio a ser sintetizado e o conjunto de HRIRs são lidos e armazenados e então é realizada a filtragem para cada frame de dois segundos. A filtragem é obtida através da convolução no domínio do tempo dos frames ordenados do áudio de entrada com cada uma das respostas ao impulso das 35 posições de variação do ângulo de azimute da fonte sonora de cada canal (ouvido), reiniciando o ângulo de azimute em zero após 350º, durante toda a duração do áudio de entrada.

Por fim, o conjunto de convoluções forma um sinal filtrado com a informação de uma localização diferente da fonte sonora para cada dois segundos de áudio. Os resultados das convoluções dos frames com a HRIR da posição – ou seja, os frames de dois segundos já filtrados – são concatenados para formar novamente um único áudio. Como tínhamos filtros distintos para os dois ouvidos, os respectivos sinais são atribuídos para cada canal do áudio final gravado, gerando a sensação de movimento do som ao redor da cabeça do ouvinte.

Resumidamente, o algoritmo que sintetiza o som tridimensional está descrito no diagrama da Figura 13 e tem como passos-chave a leitura do áudio de entrada, leitura das HRTFs do modelo de Gardner *et al* (1994), filtragem do áudio monoaural e registro do áudio binaural de saída. Vale ressaltar que se optou por utilizar um diagrama com o objetivo de esclarecer a implementação do trabalho e facilitar o entendimento, já que um fluxograma do algoritmo não agregaria novas informações.

Figura 13 - Diagrama do algoritmo para síntese de som tridimensional




Fonte: A autora (2015)

# 4 RESULTADOS PRELIMINARES E CONSIDERAÇÕES

No decorrer deste trabalho, buscou-se entender fundamentos referentes a áudio tridimensional com abordagem na área das Neurociências e do Processamento de Sinais. Além disso, realizou-se a implementação de uma técnica de síntese de áudio 3D para fones de ouvido em linguagem de alto nível.

Do ponto de vista das Neurociências, estudou-se conceitos e características referentes ao som, ao sistema auditivo e à maneira como os seres humanos o percebem. Observou-se a importância da diferença interaural de tempo e de intensidade do som, além da geometria do conjunto cabeça-tronco-pina para que o ser humano identifique a localização de uma fonte sonora no espaço tridimensional.

Posteriormente, no que diz respeito ao Processamento de Sinais, realizou-se considerações referentes ao áudio binaural e também a aspectos relevantes da sua síntese e aplicação. Destacou-se que para a síntese de áudio binaural para reprodução em fones de ouvido é essencial utilizar-se um conjunto de funções de transferência que prevê as modificações sofridas pelo som desde a fonte sonora até a captação do ouvido humano. Essas funções de transferência são chamadas de HRTFs e servem como um filtro para o áudio monoaural ao qual se pretende aplicar a panoramização pantofônica.

No algoritmo construído como prova dos conceitos estudados, observou-se empiricamente que existe alguma dificuldade de distinguir determinadas posições, principalmente quando elas têm pequena diferença no ângulo de azimute. Isso se deve ao fato de que as HRTFs simulam um ser humano com dimensões médias, por isso algumas ações são previstas para melhorar a experiência do usuário que utiliza esse sistema. Uma dessas ações seria um breve treinamento prévio, quando é informada ao usuário a posição da fonte sonora para cada caso. Optou-se por implementar um algoritmo com passo definido na variação do azimute, porém para a sensação de movimento perfeito do som pode ser necessária a utilização de uma técnica de interpolação que simule os intervalos de ângulo para os quais não há uma HRIR medida. No algoritmo sintetizado podem ser percebidos cortes nítidos no áudio na transição entre os frames de dois segundos. Freeland (2001) relata cortes aparentemente similares e os atribuem à ausência de interpolação, entretanto, acredita-se que esses cortes podem ser eliminados de forma efetiva apenas com uma alteração temporal suave da resposta ao impulso dos filtros.

Com o algoritmo proposto é possível perceber aspectos sobre as técnicas de síntese de som tridimensional e entender a aplicação dos conceitos estudados. Para o Trabalho de Conclusão de Curso II propõe-se:

1. A formalização matemática do método proposto;
2. A implementação e avaliação do algoritmo utilizando o conjunto de HRIRs medido por Gardner *et al* (1994) que prevê as funções de transferência das caixas acústicas e que tem um número reduzido de amostras;
3. A escolha de uma das pinas para utilização em todo o algoritmo;
4. A formalização de uma solução para o problema da passagem abrupta entre os frames do áudio.

Conforme visto inicialmente, tem-se em vista o uso futuro desse algoritmo em um aplicativo para *smartphone* que forneça informação constante da posição do Norte geográfico, também limitado à simulação espacial de um determinado número de pontos localizados no plano horizontal à cabeça. Por isso, também no Trabalho de Conclusão de Curso II, é importante que seja formulado um método de avaliação do desempenho da técnica implementada quanto à probabilidade de o indivíduo ser capaz de identificar a posição da fonte sonora.

Por fim, esse trabalho visou contribuir ao trazer uma abordagem na área de Processamento de Sinais ainda não explorada na Universidade de Caxias do Sul, que é o processamento de sinais de áudio binaural. O estudo do processamento e síntese de áudio tridimensional é vasto no que diz respeito às formas de geração e reprodução do som, bem como na aplicação em aparelhos auditivos, por exemplo. Cada uma das diversas aplicações demanda conceitos distintos e específicos de neurociências, processamento de sinais e geometria analítica, sendo a última pouco vista no presente trabalho por ser ferramenta específica da síntese de som tridimensional quando a fonte de reprodução do som são caixas acústicas.

## 5 REFERÊNCIAS

BEAR, M. F.; CONNORS, B. W. PARADISO, Michael A. **Neurociências: desvendando o sistema nervoso**. 3. ed. Porto Alegre: Artmed, 2008.

DO CARMO, L. I. C. **Efeitos do ruído ambiental no organismo humano e suas manifestações auditivas**. 1999. Monografia de conclusão do Curso de Especialização em Audiologia Clínica. Centro de Especialização em Fonoaudiologia Clínica, Goiânia, 1999.

FREELAND, F. P. **Técnicas de Geração de Som Tridimensional**. 2001. Dissertação (Mestrado em Ciências em Engenharia Elétrica) – Universidade Federal do Rio de Janeiro, Programa de Pós-Graduação de Engenharia, Rio de Janeiro, 2001. il. color.

GARDNER, B.; MARTIN, K. **HRTF Mesurements of a KEMAR Dummy-Head Microphone**, Technical Report 280, MIT Media Lab., Cambridge, MA, 1994.

GUNZI, A. S. **Som Tridimensional: Métodos de Geração e Métodos de Reprodução**. 2008. Dissertação (Mestrado em Ciências em Engenharia Elétrica) – Universidade Federal do Rio de Janeiro, Programa de Pós-Graduação de Engenharia, Rio de Janeiro, 2008. il. color.

_____. **feelSpace**. Disponível em: <http://feelspace.cogsci.uni-osnabrueck.de/>. Acesso em 20 de abril 2015.

_____. **Mixed Feelings**. Disponível em: <http://archive.wired.com/wired/archive/15.04/esp.html>. Acesso em 20 de abril 2015.