Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 14º Congresso de Engenharia de Áudio
20ª Convenção Nacional da AES Brasil
17 a 19 de Maio de 2016, São Paulo, SP

Qual o futuro do MP3? Áudio espacial e codificação orientada a objetos

Bruno Masiero[1]

[1] Universidade Estadual de Campinas (Unicamp),
Faculdade de Engenharia Elétrica e Computação, Departamento de Comunicações
Campinas, SP, 13083-852, Brasil

masiero@unicamp.br

RESUMO

Os avanços tecnológicos das últimas décadas resultaram em uma mudança significativa na nossa relação com a mídia. A popularização da Internet permitiu que uma enorme quantidade de documentos de áudio estejam disponíveis a qualquer hora e de praticamente qualquer lugar, uma condição que só foi atingida graças a técnicas de codificação que permitem a compactação de arquivos de áudio, sendo a codificação MP3 a porta estandarte deste processo. Por outro lado, a crescente popularização de sistemas surround para uso doméstico e do uso de fones-de-ouvido associados a telefones portáteis com grande capacidade de processamento permite a integração da informação espacial como um elemento integrante do processo e do produto artístico. Isto significa que se faz necessário técnicas de codificação pensadas para arquivos de áudio espacial. Neste documento iremos apresentar as principais técnicas de captura e de reprodução de áudio espacial, que são essenciais para se entender os requisitos de um formato de codificação de áudio espacial e ao fim iremos discutir uma proposta para atender a estes requisitos, a codificação baseada em objetos.

0 INTRODUÇÃO

A música está intimamente ligada com a evolução humana, sendo, em muitas culturas, um elemento fundamental em cerimônias, festividades e rituais religiosos. Apesar de cada cultura ter sua própria interpretação do que seria música, um fato comum ao evento musical é que a apreciação da música era um evento *ao vivo*, ou seja, para apreciar a música o ouvinte precisava necessariamente estar presente no local e instante que a música esta sendo executada.

Mas isto começou a mudar no fim do século XIX com novas invenções como o *Théâtrophone*, que permitia que performances de óperas ou concertos orquestrais pudessem ser escutados pelo telefone, dispositivo que

havia sido patenteado em 1876 por Alexander Graham Bell. Isto era realizado colocando-se vários telefones em uma sala ou teatro que eram conectados as telefones (operados por moedinhas) geralmente disponíveis em hotéis e bares. Percebe-se que a ideia de vender música a distância já nascia aí, apesar de que ainda limitada a execuções ao vivo.

Mas tudo isto mudou com o *Fonógrafo*, inventado em 1877 por Thomas Edison. Esta invenção deu início a uma verdadeira revolução na forma como a música poderia ser "consumida", já que o fonógrafo permitia que a música pudesse ser registrada (gravada) para posterior reprodução, possivelmente em outro local. Ou seja, esta invenção decretou o fim da exclusividade da música ao vivo e permitiu a criação de uma nova indústria, a indústria fonográfica [1].

As primeiras gravações apresentavam uma qualidade sonora muito baixa, tanto que, após sua invenção, Thomas Edison abandou o projeto do fonógrafo por dez até que uma nova tecnologia foi desenvolvida que permitisse uma gravação de melhor qualidade [1]. Ainda assim, nas primeiras décadas da indústria fonográfica a qualidade da reprodução era tão baixa e, por outro lado, a possibilidade de escutar música em casa sem a presença de músicos tão incrível que neste momento não importava o fato de o sistema ter um único alto-falante e, portanto, não reproduzir as características espaciais do som que poderiam ser apreciadas durante uma performance ao vivo.

Conforme as técnicas de gravação e transmissão[1] foram avançando e o custo dos transdutores caiu, a evolução lógica foi incluir informação espacial ao conteúdo musical. Note que o Théâtrophone já permitia o uso de duas linhas telefônicas (ao dobro do preço), para serem escutadas uma com cada ouvido, dando assim uma certa sensação de espacialidade.

Em de 1931 Alan Blumlein patenteou um sistema que é considerado a gênese do que hoje conhecemos como estéreo [2]. Neste documento Blumlein sugere, entre muitas outras coisas, a gravação com dois microfones e a posterior reprodução com dois alto-falantes. Muitos experimentos e tentativas foram feitas por cientistas, engenheiros e audiófilos para desenvolver um sistema de reprodução de alta-fidelidade com dois canais. Mas foi só por volta da década de 1960 que sistemas estereofônicos[2] se tornaram economicamente viáveis e que a gravação em estéreo se tornou o padrão *de facto* da indústria fonográfica, o que permanece sendo até os dias de hoje.

Após a introdução do cinema falado, a indústria cinematográfica sempre se manteve na vanguarda da reprodução sonora em alta-fidelidade e logo assimilou o uso da estereofonia. Inicialmente, as gravações eram feitas em dois ou mais canais para facilitar a associação da imagem com o som, mas logo cinemas passaram a usar a reprodução multi-canal, muitas vezes colocando o diálogo em um canal ligado a um alto-falante no centro da tela e a trilha sonora ligada a dois alto-falantes nos cantos da tela. Com o passar do tempo foram adicionados mais canais para incrementar a sensação de espacialidade até a indústria cinematográfica chegar ao seu padrão *de facto*, o som *surround* ou 5.1, que depois ganhou as casas com o advento dos *home theatres*.

Mas a busca pela sensação de imersão no espaço sonoro, ou seja, pelo realismo, não parou por aí. Novos sistemas de reprodução sonora continuaram a ser desenvolvidos, como por exemplo o Ambisonics [3] e o *wave field synthesis* (WFS) [4]. Estes sistemas de reprodução espacial requerem um número elevado de canais de transmissão, sendo, portanto, sistemas cada vez mais complexos e que requerem técnicas de gravação também cada vez mais complexas.

A indústria fonográfica sofreu uma nova guinada com a era digital e com o advento do compartilhamento de arquivos de áudio pela internet, que resultou em um novo modelo de consumo de música, que não é mais comprada em uma mídia física como o vinil ou o CD, mas sim é baixada diretamente da rede de computadores. Como esta nova forma de consumo possui uma restrição intrínseca na taxa de transmissão de dados da rede, se tornou necessário a redução do tamanho dos arquivos de áudio. Por esta razão técnicas de compressão de áudio com perdas, como a famosa MP3, se popularizaram para permitir o envio de arquivos de áudio por canais com taxas de transmissão reduzidas.

Pelo que foi apresentado, nota-se um claro desencontro entre técnicas de reprodução espacial mais modernas e os novos hábitos de consumo de conteúdo musical, já que para transmitir o elevado número de canais de áudio requeridos pelas novas técnicas se faz necessária uma taxa de transmissão muito alta, o que ainda continua sendo uma situação restritiva, ainda mais com o aumento do uso da internet por redes de telefonia celular.

Ou seja, se faz necessário uma forma de compactação eficiente dos arquivos de áudio espacial. Por outro lado, como acima mencionado, o uso cada vez mais frequente de telefones celulares para acesso à internet também resulta num incremento no número de usuários que escutam música por fones de ouvido.

Um contraponto ao aumento desenfreado no número de canais é o sistema biauricular [5], que assume que são necessário apenas dois canais de gravação e reprodução (cada canal reproduzindo diretamente em um ouvido através do uso de fones-de-ouvido), uma vez que nós possuímos apenas dois ouvidos. Este sistema já era estudo, por exemplo, por Blumlein na década de 1930 e continuou como objeto de curiosidade de muitos estudioso, tendo tido algumas tentativas frustradas de

[1] Estações comerciais de rádio começam a transmitir conteúdo musical e de notícias no início da década de 1920, após o fim da Primeira Guerra Mundial.

[2] Do grego, *stereo* significa "sólido". Ou seja, na concepção do nome, o sistema estéreo não é limitado a dois canais, mas se tratava, sim, de um sistema de reprodução sonora espacial. Ao menos esta era a impressão quando saímos da monofonia!

comercialização. Atualmente, contudo, o sistema biauricular vive um período de renascimento.

Com tudo que foi exposto, percebe-se que há uma nova demanda para a codificação de áudio que deve conseguir codificar gravações com conteúdo espacial de forma a ser transmitida em canais com baixas taxas de transmissão e cujo mesmo arquivo possa ser reproduzido através de diferentes técnicas de reprodução. Tendo em vista estes requisitos o *Moving Picture Expert Group* (MPEG) propôs um novo paradigma de codificação, uma codificação baseada em objetos sonoros ao invés de canais de áudio com o intuito de atender à demanda de reduzir a taxa de informação sendo transmitida e ao mesmo tempo permitir que o arquivo de áudio possa ser reproduzido da melhor maneira possível em sistemas de reprodução de diferentes complexidades, ou seja, que o sistema de áudio sejam escaláveis.

Este artigo pretende dar uma visão geral sobre as principais técnicas de gravação e reprodução sonora para nortear uma discussão sobre os requisitos da codificação de áudio espacial orientada a objetos, ou, em inglês, *Spatial Audio Object Coding* (SAOC).

1 CAPTAÇÃO ESPACIAL DO SOM

A captação e o armazenamento de um evento sonoro só se faz possível com o uso de um transdutor que converta a onda sonora em uma onda em algum outro meio. O primeiro dispositivo a realizar a transdução de uma onda sonora em uma onda elétrica foi desenvolvido na década de 1870 por David Hughes [6] e proporcionou uma melhoria tão considerável na transdução do som que só a partir de sua invenção que a telefonia, a radiodifusão e a indústria fonográfica puderam ganhar corpo.

Durante a primeira metade do século XX as gravações eram feitas usando-se, predominantemente, apenas um microfone. Realizar a gravação com um único microfone é equivalente a dizer que a onda sonora de interesse foi amostrada em um único ponto. Se quisermos melhorar a informação espacial uma maneira seria amostrar o espaço em mais pontos, ou seja, usar mais microfones [7]. Por exemplo, em sua patente de 1931, Blumlein já previa o uso de dois microfones para a gravação estereofônica [2].

Com a popularização do estéreo e posterior desenvolvimento de outras técnicas de reprodução espacial (ver seção 2) diversos outros tipos de arranjos[3] de microfones foram desenvolvidos para atender a diferentes propósitos. Ainda assim, podemos classificar estes arranjos de microfones para captação espacial entre dois paradigmas principais: arranjos compactos ou arranjos distribuídos. Dentro do primeiro grupo destacamos ainda um tipo de arranjo especial para captação biauricular, que usa um manequim (com cabeça e torço) provido de dois microfones onde seriam as entradas do canais auditivos.

[3]Um arranjo de microfones é um agrupamento ordenado de dois ou mais microfones.

1.1 Captação biauricular

O nosso sistema auditivo se faz valer do fato de que possuímos duas orelhas (posicionadas em lados opostos da cabeça) para determinar a direção de chegada de uma frente de onda sonora, o que é conhecido por audição biauricular [8]. Uma frente de onda ao atingir nossa cabeça irá gerar um evento sonoro diferente em cada ouvido, dependente do ângulo de chegada desta onda.

Blauert define a tecnologia biauricular da seguinte maneira [8]: "A tecnologia biauricular são métodos que envolvem o sinal acústico captados pelos dois ouvidos de um ouvinte para algum objetivo prático, como, por exemplo, gravação, análise, síntese, processamento, avaliação e reprodução destes sinais."

A ideia principal é que todos os eventos auditivos percebidas por nós são extraídos dos sinais pelos nossos dois ouvidos. Se formos capazes de gravar os sinais que adentram nosso canal auditivo, os sinais biauriculares, e posteriormente reproduzir exatamente os mesmos sinas na mesma posição (com o uso, por exemplo, de fones-de-ouvido), o evento auditivo gerado por esta reprodução deve ser o mesmo que aquele causado durante a gravação.

É importante ressaltar que a percepção espacial está intimamente ligada às características anatômicas do ouvinte, representadas pela função de transferência anatômica (*head-related transfer function*, HRTF) [9, 10]. Portanto, a gravação biauricular deveria, idealmente, ser feita com a própria cabeça do ouvinte, o que, venhamos e convenhamos, não é sempre praticável.

Gravações biauriculares acabam sendo, em geral, feitas com uma cabeça artificial, ou manequim [11]. No entanto, apesar de algumas experimentações, como por exemplo o álbum *Street Hassle* de Lou Reed, lançado em 1978 como o primeiro disco produzido comercialmente com tecnologia biauricular, esta tecnologia não logrou sucesso comercial e ficou relegada a cientistas e entusiastas. Apenas recentemente, com a popularização do uso de fones-de-ouvido por causa da miniaturização dos sistemas de áudio portátil é que a técnica biauricular tem vivido uma renascença.

1.2 Arranjos compactos de dois canais

A audição biauricular usa dois parâmetros para estimar a direção de chegada da onda: a diferença de amplitude entre os ouvidos (*interaural level difference*, ILD) ou a diferença do tempo tempo de chegada[4] da onda aos ouvidos (*interaural time difference*, ITD). Como veremos na seção 2.3.1, a reprodução estereofônica busca reproduzir estes parâmetros de forma a posicionar fontes virtuais entre os alto-falantes.

Os arranjos para gravação estereofônicas buscam justamente capturar uma diferença de fase ou amplitude entre os dois canais de gravação. Note que este tipo de arranjo apresenta os microfones agrupados de forma a

[4]Uma diferença no tempo de chegada pode ser interpretado como uma diferença na fase do espectro deste sinal.

captar o som em uma pequena região do espaço, por isso o nome *compacto*.

Listamos abaixo os três tipos de arranjos para gravação estereofônica mais usuais.

Diferença de intensidade Também conhecido como arranjo X-Y ou par de Blumlein, este arranjo usa dois microfones direcionais (com diretividade do tipo cardioide ou figura-de-oito) posicionados perpendicularmente entre si e com suas membranas justapostas. Esta configuração dá ganhos diferentes para o sinal de acordo com sua direção de chegada por causa de sua diretividade não ser omnidirecional. Isto significa que os dois canais resultantes irão apresentar diferença de amplitude, mas não de fase (por estarem justapostos) [12].

Diferença de tempo de chegada Também conhecido como arranjo A-B, este arranjo usa dois microfones em paralelo e distantes (20 a 50 cm) um do outro. Esta configuração irá gerar dois sinais com diferença de nível sonoro e principalmente de tempo de chega. Ou seja, os dois canais resultantes irão apresentar diferença de fase em seu espectro [12].

Diferença diretividade Também conhecido como arranjo M-S, de *mid/side*, este arranjo usa um microfone omnidirecional e outro com diretividade figura-de-oito justapostos. Diferentemente das configuração anteriores, neste caso os canais de reprodução não são os mesmos da gravação, mas sim a combinação linear (some e subtração) dos dois sinais gravados. Este tipo de gravação permite uma maior flexibilidade para o pós-processamento do sinal, além de apresentar compatibilidade direta com a reprodução monofônica [12].

1.3 Arranjos compactos multicanais

A técnica de reprodução *surround*, ou 5.1, se popularizou nas últimas décadas, se tornando o padrão *de facto* da indústria cinematográfica. Em geral as faixa de áudio e efeitos sonoros de um filme são mixadas em pós-produção, onde as fontes sonoras são distribuídas no espaço usando a técnica de panorama (seção 2.3.1). Mas para situações onde se deseja gravar uma cena sonora para reprodução em 5.1 sem edição ou pós-processamento usa-se um arranjo com cinco microfones tipo cardioides, cada um apontando para cada uma das posições padronizadas para os alto-falantes no arranjo 5.1 (seção 2.3.1) [13]. Note que se utilizarmos apenas os sinais dos microfones apontando para -45° e 45° temos uma configuração X-Y e, portanto, compatibilidade com a reprodução em estéreo.

Analisando os arranjos compactos de dois canais da seção anterior verifica-se que os arranjos X-Y e A-B buscam reproduzir características de como o sistema auditivo humano extrai informações espaciais (ITD e ILD) enquanto o arranjo M-S busca descrever a variação local do campo acústico, uma vez que a diretividade figura-de-oito pode ser interpretada como a primeira derivada de um microfone omnidirecional em uma dada direção.

Podemos imaginar que queremos refinar a descrição da variação espacial do campo sonoro. Para isto, seria necessário posicionar três microfones figura-de-oito ortogonalmente entre si (frente-trás, direita-esquerda, cima-baixo), além do microfone omnidirecional. Mas esta é uma construção difícil de ser feita.

Michael Gerzon, um matemático que trabalhava no campo da física quântica, propôs uma maneira de solucionar este problema usando uma ferramenta matemática do seu campo de trabalho, a decomposição em harmônicas esféricas (DHE). Esta ferramenta permitia transformar uma gravação feita com um arranjo de microfones em forma de um tetraedro em uma gravação que teria sido feita com um microfone do tipo M-S expandido discutido acima.

O primeiro arranjo tetraédrico de microfones, batizado de *Soundfield microphone* começou a ser comercializado em 1978. Os sinais gerados pelo arranjo são conhecidos por formato-A enquanto que os sinais obtidos após a transformação pela DHE são conhecidos por formato-B. O formato-B é então pós-processado e eventualmente distribuído para alto-falantes. Gerzon também propôs uma maneira de realizar este pós-processamento baseada na DHE, que deu o nome de Ambisonics (seção 2.3.1).

Apesar de o soundfield ser construído com apenas quatro microfones, a teoria do DHE pode ser aplicada para arranjos com um maior número de microfones, permitindo extrair-se um maior número de harmônicas e com isso realizar uma melhor descrição da variação espacial do campo acústico. Esta técnica de gravação vem ganhando cada vez mais prestígio e novos microfones vêm sendo desenvolvidos para este fim. Um desdes arranjos compactos está disponível comercialmente, o *Eigenmike*, com 32 microfones distribuídos em uma esfera.

1.4 Arranjos distribuídos

Vimos que arranjos compactos são projetados para representar localmente algum parâmetro do campo sonoro. Já arranjos distribuídos, como o próprio nome diz, estão distribuídos pelo espaço e, portanto, permitem amostrar uma maior região do campo sonoro. Este tipo de arranjo é bastante usado para imageamento acústico [14] e imageamento sísmico [15], mas praticamente não é usado pela indústria fonográfica.

Para gravações de grandes grupos tocando ao vivo, como orquestras sinfônica, técnicos de som geralmente lançam mão de uma espécie de arranjo distribuído, posicionando diversos microfones espalhados pelo orquestra. Já quando a gravação é feita em estúdio, é comum o instrumentos serem gravados individualmente ou em pequenos grupos (close miking). Neste segundo caso as trilhas gravadas são mixadas posteriormente (pós-produção) e a maneira como são mixadas levam em conta o tipo de sistema que deverá ser usado para reprodução.

Interessante é notar que esta técnica de captação individual é a que mais se aproxima do conceito de objetos sonoros explorado pela codificação SAOC, já que cada instrumento ou naipe é considerado um objeto sonoro.

2 REPRODUÇÃO ESPACIAL DO SOM

Passamos agora da captação para o outro extremo da cadeia fonográfica, a reprodução, que por muitas décadas era feita usando-se uma única fonte acústica. Inicialmente estas fontes eram feitas usando-se uma corneta metálica ligada a uma membrana que era excitada por um objeto pontiagudo. Alguns fonógrafos também usavam um sistema de ar comprimido para amplificar o som produzido. Mas foi só após o desenvolvimento dos alto-falantes de bobina móvel que se tornou possível a reprodução sonora de alta fidelidade.

O próximo passo na busca da melhoria destes sistemas foi introduzir *espacialidade* ao som reproduzido. Diversos tentativas foram feitas para desenvolver um sistema de reprodução de alta-fidelidade com dois (ou mais) canais. Mas foi só por volta da década de 1960 que sistemas estereofônicos (com dois canais) se tornaram economicamente viáveis.

Invariavelmente, as técnicas de reprodução continuaram sua evolução através do aumento do número de canais, o que faz necessário o uso de arranjos com múltiplos alto-falantes. Assim como fizemos com os arranjos de microfones, podemos classificar os arranjos de alto-falantes em arranjos compactos ou distribuídos, sendo a segunda opção a mais comumente utilizado.

Os sinais reproduzidos pelos arranjos distribuídos podem ser classificados sob três paradigmas distintos: biauricular, panorama ou síntese de campo.

Como já apresentado na seção 1.1, a reprodução biauricular apresenta sons diferentes para cada ouvido do ouvinte, garantindo assim a reprodução dos parâmetros interaurais. Este tipo de reprodução é geralmente feito através de fones-de-ouvido, que podem ser interpretados como um arranjo compacto individual (de dois alto-falantes). Também é possível a reprodução de sinais biauriculares através de dois ou mais alto-falantes, mas para este fim é necessário o uso de um banco de filtros para cancelamento de diafonia (*crosstalk cancellation*, CTC) [16].

Os sistemas tipo panorama buscam recriar em uma área do espaço (o *sweet spot*) diferenças interaurais (ILD e ITD) de forma que o ouvinte ali presente tenha uma impressão de espacialidade. Por outro lado, sistemas tipo síntese de campo buscam efetivamente recriar dentro do ambiente de reprodução o campo sonoro existente no local de gravação.

2.1 Reprodução individual

A reprodução de gravações biauriculares pode ser feita de duas maneiras: com fones-de-ouvido ou com alto-falantes.

Fones-de-ouvido A reprodução através de fones de ouvido é, em primeira instância, mais simples, uma vez que a separação de canal obtida desta maneira é quase ideal. No entanto, cada fone-de-ouvido possui uma resposta em frequência distinta, que, ainda por cima, varia de acordo com a geometria do ouvido externo do ouvinte e também com o posicionamento do dispositivo [17]. Desta maneira, é desejável que a equalização do fone-de-ouvido seja individualizada [18].

Um aspecto crítico para a qualidade da reprodução biauricular é que estamos, mesmo que inconscientemente, constantemente nos movimentando. No entanto, a disposição da cena sonora apresentada continua fixa em relação à posição do fone-de-ouvido. Isto significa que ao nos movimentarmos as fontes sonoras irão se mover conosco ao invés de ficarem fixas no espaço, podendo resultar no desaparecimento da ilusão acústica criada pela reprodução biauricular.

Para solucionar este problema se faz necessário o uso de um sistema dinâmico, que rastreie a posição da cabeça do ouvinte e compense seus movimentos. Isto é praticável com sinais biauricular que tenham sido auralizados [19], de forma a permitir que a posição das fontes seja atualizada [20]. Mais uma vez motivando um sistema de armazenamento de áudio orientado a objeto que realize a auralização da cena sonora no fim da cadeia de reprodução.

CTC Se os sinais biauriculares fossem reproduzidos através de dois alto-falantes teríamos o efeito de diafonia, ou seja, o sinal esquerdo, que deveria ser ouvido apenas pelo ouvido esquerdo, ao ser reproduzido por um alto-falante à esquerda do ouvinte atingirá primeiro o ouvido esquerdo, mas depois também o direito e vice-versa para o sinal do ouvido direito. Esta mistura dos sinais resulta na distorção dos parâmetros interaurais contidos no sinal e no consequente desaparecimento da sensação de espacialidade contida no sinal biauricular.

Para anular este efeito de diafonia são necessários filtros CTC. Estes filtros misturam os dois sinais biauriculares gerando o que chamamos de sinais transaurais. Os sinais transaurais são então reproduzidos por alto-falantes e ao chegar aos ouvidos do espectador eles interagem de forma que o sinal resultante é o próprio sinal biauricular [21, 22, 23].

Estes filtros, no entanto, dependem da posição dos alto-falantes em relação ao ouvinte e da própria HRTF (que é individual). Se filtros não individualizados forem usados a qualidade da localização obtida com o sistema fica comprometida [24, 25].

Como mencionado na seção anterior estamos constantemente nos movendo. Isto implica que a direção dos alto-falantes em relação à cabeça do ouvinte também está em constante variação e por isso os filtros CTC precisam ser constantemente atualizados levando em conta a posição do ouvinte em relação ao arranjo de alto-falantes [20, 26].

2.2 Arranjos compactos

Chamamos de arranjos compactos de alto-falantes dispositivos com um grande número de transdutores posicionados muito próximos entre si. Geralmente estes transdutores são fixados sobre uma esfera rígida [27, 28] ou sobre sólidos platônicos (como o dodecaedro ou o icosaedro) [29] e permitem direcionar o feixe sonoro, criando fontes móveis e super-diretivas.

Estes dispositivos são muito usados para caracterização acústica de salas, como o cálculo do tempo de reverberação ou o coeficiente de transmissão de paredes [30]. No entanto, eles não costumam ser usados para reprodução musical e por esta razão não serão aqui discutidos em detalhes.

2.3 Arranjos distribuídos

Os sistemas de reprodução espacial mais frequentemente utilizados são arranjos distribuídos de dois ou mais alto-falantes. O termo "distribuídos" é usado porque nesta configuração os alto-falantes estão distribuídos ao redor dos ouvinte. Este tipo de sistema geralmente possui um *sweet spot*, um ponto (geralmente no centro do arranjo) que garante a melhor qualidade de reprodução espacial.

Os sistemas de reprodução com arranjos distribuídos se baseiam em dois paradigmas: panorama e síntese de campo, que serão discutidos a seguir.

2.3.1 Panorama

Como já discutido na seção 1.2, nós localizamos sons no plano transversal através dos parâmetros interaurais ILD e ITD. A dependência destes dois parâmetros foi denominada de "teoria duplex" por Lord Rayleigh, primeiro cientista a descrever este efeito. Posteriormente determinou-se que a ITD é utilizada para localização de baixas frequências (até ≈1500 Hz) e a ILD para a localização de altas frequências (a partir de ≈700 Hz).

Algumas das técnicas de captação discutidas na seção 1.2 buscam reproduzir esta diferença de amplitude e fase entre os dois canais gravados, que serão posteriormente reproduzidos por dois alto-falantes. Mas quando é usada a técnica de *close miking*, como é possível "posicionar" esta fonte sonora entre os alto-falantes?

A técnica mais utilizada para este fim é o panorama, que envia o mesmo sinal com ganhos distintos para cada falante. Quando são usados dois alto-falantes esta técnica ganha o nome de panorama estéreo, ou simplesmente estéreo.

Estéreo Vamos assumir que tenhamos dois alto-falantes posicionados em duas pontas de um triângulo equilátero e o ouvinte se encontra na outra ponta. A técnica de panorama consiste em enviar o mesmo sinal com ganhos diferentes para cada alto-falante, de forma a posicionar o som entre aos falantes.

Mas por que percebemos um evento sonoro entre os falantes e não dois eventos sonoros distintos?

Por causa da geometria da cabeça do ouvinte, o som gerado pela fonte à esquerda do ouvinte irá chegar primeiro ao seu ouvido esquerdo e depois ao seu ouvido direito. A cabeça do ouvinte também irá cria uma sombra acústica tal que o som chegará ao ouvido esquerdo com maior intensidade que no ouvido direito. O inverso é válido para a fonte à direita do ouvinte.

Suponhamos que o sinal enviado para à fonte da esquerda tem um ganho de 3 dB em relação à fonte da direita. Para altas frequências a sombra gerada pela cabeça no ouvido oposto (maior do que 10 dB para frequências acima de 2 kHz) faz com que o som direto predomine. Neste caso, o som no ouvido esquerdo será 3 dB maior que o do ouvido direito, gerando portanto uma ILD que será interpretada como um som vindo de uma direção mais à esquerda.

Mas e para as baixas frequências onda a sombra acústica não é tão pronunciada? Neste caso o sinal gerado por cada fonte irá se combinar em cada um dos ouvidos de forma vetorial, resultando em um sinal de mesma amplitude mas com diferença de fase [31]. Ou seja, a diferença de amplitude resulta, para baixas frequências, em uma ITD, que é justamente o parâmetro relevante para o nosso sistema auditivo nesta faixa de frequência.

Mas como definir qual a relação de ganho necessária para posicionar a fonte em uma dada direção? Testes subjetivos [32] mostraram que a curva que melhor estima a direção ϕ percebida para uma dada combinação de ganhos g_R e g_L é

$$\frac{\tan\phi}{\tan\phi_0} = \frac{g_R - g_L}{g_R + g_L}. \tag{1}$$

Aqui assume-se que o ouvinte esteja centrado em relação aos falantes e que estes estão posicionados em um ângulo de ϕ_0 em relação à direção de visada do ouvinte.

Esta técnica possui a vantagem de ser de muito fácil implementação, no entanto, as fontes virtuais geradas por panorama são percebidas como largas e apresentam bastante coloração espectral [33]. Felizmente, esta coloração só é percebida em ambientes anecoicos. Quando ouvido dentro de uma sala a reverberação acaba mascarando este efeito, resultando em um sistema simples e com boa localização, provavelmente a razão de seu grande sucesso.

Surround Este é o termo utilizado para as expansões do panorama de amplitude para um formato envolvente e comercialmente viável, mais frequentemente utilizado pela industria cinematográfica. Como não houve uma padronização neste sentido, muitos sistemas diferentes foram desenvolvidos e comercializados, diferindo principalmente em como a faixa sonora era codificada e gravada/sincronizada nos rolos de filme. No que tange o arranjo, ouve uma convergência pelo formato conhecido por 5.1, com uma fonte no centro (0°), duas fontes em ±45° e duas fontes em ±110°.

Esta geometria do arranjo acabou sendo escolhida porque pares de falantes na lateral do ouvinte (por exem-

plo em 45° e 135°) não permitem realizar a tradução de ILD em ITD, ou seja, não é possível posicionar fontes laterais. O surround compensa este fato colocando os falantes posteriores mais abertos, ou seja, mais próximos ao eixo dos ouvidos, gerando assim uma fonte estável nas laterais. Além disto, um canal central é usado para garantir que diálogos seja ouvidos sempre vindo da tela, mesmo que o ouvinte não esteja mais no sweet spot.

VBAP Pulkki estendeu o conceito de panorama para arranjos tridimensionais. Ele sugeriu que o trio de alto-falantes mais próximos da direção da fonte virtual deveriam ser selecionados, deixando o restante dos falantes do arranjo inativos. O ganho destas três fontes era então calculado como os ganhos que permitiam descrever a direção da fonte virtual como combinação linear da direção das três fontes ativas, daí o nome *vector-base amplitude panning* (VBAP).

Note que a argumentação de que pares de fontes próximos do eixo dos ouvidos não permitem gerar fontes virtuais entre eles também é válida para pares na vertical. Pulkki já havia notado isto, mas ele argumenta que, apesar de a percepção de elevação variar de ouvinte para ouvinte, ela ainda assim fica sempre limitada dentro da área delimitada pelas três fontes ativas [34].

Ambisonics Enquanto VBAP propõem usar três alto-falantes por direção, a técnica Ambisonics propõem o uso de todos os alto-falantes de um arranjo distribuído para sintetizar uma fonte virtual.

Conforme discutido na seção 1.3, microfones compactos permitem descrever o campo sonoro ao seu redor em função de harmônicas esféricas, que são funções próprias (base ortonormal) da onda acústica quando representada em coordenadas esféricas [35].

Para determinar o peso de cada alto-falante do arranjo Gerzon assumiu que estes se comportavam como uma fonte de onda plana, aproveitando o fato de que a onda plana possui uma representação analítica em harmônicas esféricas, dependente apenas da posição do falante no arranjo. Bastava então ponderar os alto-falante tal que, para cada harmônica, a soma das influências de todos alto-falantes fosse igual ao formato-B, ou seja, ao valor da harmônica esférica gravada com o arranjo compacto de microfones [36].

Como as harmônicas esféricas são um conjunto de bases ortogonais, eu posso truncar meu resultado na ordem necessária, definida pelo menor número de microfones ou alto-falantes [37, 38]. Gerzon sugeriu que quando truncado em primeira ordem[5], esta técnica funcionaria baseada em princípios psicoacústicos, ou seja, seria capaz de gerar parâmetros interaurais de forma similar à técnica de panorama [3].

[5] A ordem 0 possui uma harmônica enquanto a ordem 1 possui três harmônicas distintas, resultando em um total de quatro harmônicas para um sistema de ordem 1.

No entanto, conforme se aumenta a quantidade de harmônicas usadas para descrever o campo sonoro, este sistema promove uma transição suave do paradigma de panorama para o paradigma de síntese de campo sonoro [38].

2.3.2 Síntese de campo

Enquanto técnicas baseadas em panorama tentar recriar parâmetros interaurais, as técnicas de síntese de campo buscam justamente recriar um campo sonoro com a maior fidelidade possível. Note que até agora todas estas técnicas são capazes apenas de variar a direção de chegada percebida para a fonte virtual, mas não a sua distância do ouvinte, já que todas elas assumem que as fontes se encontram em campo distante e atuam como geradores de ondas planas.

Já as técnicas de síntese de campo, como são capazes de controlar o campo sonoro e consequentemente a curvatura da frente de onda, permitem sintetizar (em parte do interior do arranjo) fontes pontuais a distâncias menores que a dos próprios alto-falantes [39].

No entanto, estas técnicas se baseiam na suposição de haver um infinito número de fontes no arranjo. No momento que voltamos à realidade e construímos tais sistemas com um número limitado de alto-falantes, problemas mundanos começam a aparecer, como uma área limitada onde o campo sonoro é fiel ao planejado (sweet spot) e o surgimento de artefatos na reprodução em altas frequências (dependente da distância entre as fontes) [40].

Higher Order Ambisonics Com já foi dito, Ambisonics permite uma cadeia completa de gravação, transmissão e reprodução *escalável* de som espacial, tal que conforme aumentamos a ordem máxima das harmônicas esféricas utilizadas realizamos uma transição entre panorama e síntese.

Isto significa dizer que o Ambisonics de alta ordem (*Higher Order Ambisonics*, HOA) nada mais é do que a técnica Ambisonics descrita na seção anterior, mas truncada em uma ordem de harmônicos esféricos mais elevada. E assim como sua versão de baixa ordem ela permite tanto a mixagem de sons (geração de fontes virtuais) quanto a gravação de uma performance com um arranjo concêntrico de microfones.

O sweet spot relacionado a esta técnica se torna maior conforme aumentamos a ordem máxima do sistema. É interessante notar que o sweet spot de sistemas Ambisonics se localizam sempre no centro do arranjo [41].

Wave Field Synthesis Vimos que o HOA é baseado em sinais captados com arranjos concêntricos, que captam como o som chega a um dado ponto e cuja reprodução busca fornecer sinais para os alto-falantes que, quando combinados, reproduzam no centro do arranjo o mesmo campo sonoro que foi captado pelo arranjo de microfones.

Já a técnica de síntese de campo sonoro (*wave field syntesis*, WFS) é baseada no princípio de Huygens, que postula que uma frente de onda pode ser decomposta na soma de infinitas fontes pontuais [4]. Ou seja, se eu conheço o campo sonoro que incide sobre o lado externo de um volume, eu posso reproduzir este mesmo campo na parte interna deste volume distribuindo infinitas fontes pontuais pela superfície limítrofe deste volume e fazendo que cada uma delas emita um sinal equivalente ao sinal gerado pelo campo sonoro externo naquele ponto infinitesimal.

Ou seja, diferentemente do HOA, que amostra o campo sonoro em uma região compacta do espaço, o WFS assume que o campo sonoro será amostrado em todo contorno do meu espaço de reprodução. Esta suposição não é viável na praticamente, pois precisaríamos de um arranjo distribuído de grandes proporções. Por esta razão, o WFS é usado somente com sinais auralizados, ou seja, para campos sonoros virtuais. Isto o torna um sistema praticamente "nativo" para reprodução de arquivos com compactação orientada a objetos.

Infelizmente, apesar de teoricamente o WFS permitir uma reprodução muito fiel do campo acústico simulado, isto só acontece de fato se o arranjo utilizado for suficientemente denso, o que o torna este sistema caro e ainda não muito viável para uso cotidiano.

3 CODIFICAÇÃO ORIENTADA A OBJETOS

Como vimos nas seções anteriores, existe uma vasta gama de sistemas de gravação e reprodução de áudio espacial e a maioria destes sistemas só são compatíveis com seus sistemas duais, ou seja, uma gravação com um arranjo compacto do tipo X-Y deve ser reproduzida com um arranjo distribuído tipo estéreo. Para poder ser reproduzido com outro tipo de sistema, como VBAP ou Ambisonics, seria necessário um processamento conhecido como "up-mixing", que busca extrair as informações espaciais do arquivo original e sintetizar um novo arquivo compatível com um outro sistema de reprodução.

Como a tendência da reprodução de áudio doméstica é a oferta de soluções proprietárias baseada em um ou vários dos paradigmas discutidos neste artigo, é importante criar um tipo de arquivo que possa ser reproduzido em qualquer tipo de sistema. Isto significa que os arquivos devem fornecer todas as informações necessárias para que a cena acústica desejada possa ser sintetizada pelo próprio cliente, que deve fornecer informação sobre as características de seu sistema de reprodução.

E é justamente nesta direção que caminha o mais novo formato de compressão de áudio proposto pela MPEG, o formato SAOC, que propõe uma codificação orientada a objetos sonoros.

Este novo paradigma de codificação é completamente diferente de todos outros formatos de codificação já propostos pela MPEG, como o *MPEG Surround*, que realiza um "down-mix" de um sinal 5.1 para um sinal estereofônico, que é então comprimido para transmissão, além de gerar uma sequência de metadados espaciais com baixa taxa de informação (que é posteriormente usada para tentar reconstituir o arquivo multi-canal). O fato é que o MPEG Surround foi desenvolvido sob o paradigma de múltiplos canais, e não de objetos sonoros.

ISO/IEC 23003-2:2010 Esta norma padroniza as interfaces do SAOC, ditando os rumos que o desenvolvimento deve seguir e garantindo compatibilidade entre diferentes soluções.

O SAOC utiliza o conceito de *objetos sonoros*, isto é, sinais monofônicos não-correlacionados, que são então compactados conjunta ou separadamente. Além disto, uma sequência de metadados espaciais com baixa taxa de informação também é adicionada ao fluxo de dados. Estas informações são então utilizadas pelo receptor para gerar os sinais para cada canal de saída de acordo com o sistema de reprodução disponível informado pelo usuário.

Mas e se não tenho as gravações individuais de cada objeto da cena? É aqui que entram novos paradigmas de processamento de sinais, que não foram padronizados e ainda são objetos de pesquisa. Estes novos algoritmos serão baseados em modelos da audição espacial que possuirão como entrada gravações feitas com arranjos de microfones, sejam eles densos ou esparsos. Um exemplo nesta direção é o algoritmo DirAc, que extrai estes parâmetros de gravações feitas com microfones do tipo soundfield ou eigenmic [42, 43].

No lado de entrada do codec será necessário desenvolver algoritmos que permitam a separação cega de fontes e extração de informações espaciais da cena sonora. Do lado de saída serão necessário desenvolver sintetizadores robustos capazes de fornecer uma reprodução de qualidade para qualquer tipo de geometria de fontes.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Randall Stross, "The Incredible Talking Machine," *TIME*, 2010.

[2] Alan Dower Blumlein, "Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems," 1931.

[3] Michael Gerzon, "Surround-sound psychoacoustics: Criteria for the design of matrix and discrete surround-sound systems," *Wireless World*, vol. 80, no. December, pp. 483–486, 1974.

[4] A. J. Berkhout, "A Holographic Approach to Acoustic Control," *Journal of the Audio Engineering Society*, vol. 36, no. 12, pp. 977–995, 1988.

[5] Henrik Møller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 3-4, pp. 171–218, 1992.

[6] Anton Huurdeman, *The Worldwide History of Telecommunications*, John Wiley & Sons, 2003.

[7] Philip A. Nelson and Stephen J Elliott, *Active Control of Sound*, Academic Press, San Diego, CA, 3rd edition, 1992.

[8] Jens Blauert, *Spatial hearing: the psychophysics of human sound localization*, MIT Press, 1997.

[9] Henrik Møller, Michael Friis Sørensen, Dorte Hammershøi, and Clemen Boje Jensen, "Head-Related Transfer Functions of Human Subjects," *J. Audio Eng. Soc.*, vol. 43, no. 5, pp. 300–310, 1995.

[10] William M. Hartmann, "How We Localize Sound," *Physics Today*, vol. 52, no. 11, pp. 24, nov 1999.

[11] Stephan Paul, "Binaural Recording Technology: A Historical Review and Possible Future Developments," *Acta Acustica united with Acustica*, vol. 95, no. 5, pp. 767–788, sep 2009.

[12] Ville Pulkki and Matti Karjalainen, *Communication Acoustics*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.

[13] Bill Whiston, "Microphone systems used for Surround Sound pickup - And their use at Wimbledon tennis and the Proms," *EBU Technical Review*, , no. Q1, pp. 1–9, 2008.

[14] Vítor Heloiz Nascimento, Bruno Sanches Masiero, and Flávio P. Ribeiro, "Acoustic Imaging Using the Kronecker Array Transform," in *Signals and Images: Advances and Results in Speech, Estimation, Compression, Recognition, Filtering, and Processing*, Rosangela Fernandes Coelho, Vitor Heloiz Nascimento, Ricardo Lopes de Queiroz, Joao Marcos Travassos Romano, and Charles Casimiro Cavalcante, Eds., pp. 153–178. CRC Press, 2015.

[15] Lawrence C. Wood and Sven Treitel, "Seismic Signal Processing," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 649–661, 1975.

[16] Jerald L Bauck and Duane H Cooper, "Generalized Transaural Stereo," in *93rd AES Convention*, San Francisco, USA, 1992.

[17] Bruno Sanches Masiero and Janina Fels, "Perceptually Robust Headphone Equalization for Binaural Reproduction," in *130th AES Convention*, London, England, 2011, pp. 1–7.

[18] Josefa Oberem, Bruno Sanches Masiero, and Janina Fels, "Authenticity and naturalness of binaural reproduction via headphones regarding different equalization methods," *Proceedings of AIA-DAGA 2013 Conference on Acoustics*, 2013.

[19] Michael Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality (RWTHedition)*, Springer, 2007.

[20] Tobias Lentz, "Dynamic crosstalk cancellation for binaural synthesis in virtual reality environments," *J. Audio Eng. Soc.*, vol. 54, no. 4, pp. 283–294, 2006.

[21] B. B. Bauer, "Stereophonic Earphones and Binaural Loudspeakers," *J. Audio Eng. Soc.*, vol. 9, no. 2, pp. 148–151, 1961.

[22] Jerald L Bauck and Duane H Cooper, "On Transaural Stereo for Auralization," in *95th AES Convention*, New York, 1993, vol. 3728.

[23] Henrik Møller, "Reproduction of artificial-head recordings through loudspeakers," *J. Audio Eng. Soc.*, vol. 37, no. 1, pp. 2–5, 1989.

[24] Michael A Akeroyd, John Chambers, David Bullock, Alan R Palmer, A Quentin Summerfield, Philip A. Nelson, and Stuart Gatehouse, "The binaural performance of a cross-talk cancellation system with matched or mismatched setup and playback acoustics," *J. Acoust. Soc. Am.*, vol. 121, no. 2, pp. 1056—-1069, 2007.

[25] Piotr Majdak, Bruno Sanches Masiero, and Janina Fels, "Sound localization in individualized and non-individualized crosstalk cancellation systems.," *Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2055–2068, apr 2013.

[26] Bruno Sanches Masiero and Michael Vorländer, "A Framework for the Calculation of Dynamic Crosstalk Cancellation Filters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1345–1354, sep 2014.

[27] Martin Pollow, Gottfried K Behler, and Bruno Sanches Masiero, "Measuring directivities of natural sound sources with a spherical microphone array," in *Proceedings of Ambisonics Symposium*, Graz, AT, jun 2009, IEM Graz.

[28] Alexander Mattioli Pasqual, *Sound Directivity Control in a 3-D Space by a Compact Spherical Loudspeaker Array*, Phd, Universidade Estadual de Campinas, 2010.

[29] Franz Zotter, *Analysis and Synthesis of Sound-Radiation with Spherical Arrays*, Ph.d., University of Music and Performing Arts, Graz, Austria, jan 2009.

[30] Internacional Standard, "ISO 3382:1997 - Acoustics - Measurement of the reverberation time of rooms with reference to other acoustical parameters," 1997.

[31] Klaus Wendt, *Das Richtungshören bei der Überlagerung zweier Schallfelder bei Intensitäts- und Laufzeitstereophonie*, Phd, RWTH Aachen, 1963.

[32] JC Bennett, K Barker, and FO Edeko, "A new approach to the assessment of stereophonic sound system performance," *J. Audio Eng. Soc*, vol. 33, no. 5, pp. 314–321, 1985.

[33] Ville Pulkki, *SPATIAL SOUND GENERATION AND PERCEPTION BY AMPLITUDE PANNING TECHNIQUES*, Ph.D. thesis, Helsinki University of Technology, 2001.

[34] Ville Pulkki, "Localization of amplitude-panned virtual sources II: Two-and three-dimensional panning," *J. Audio Eng. Soc.*, vol. 49, no. 9, pp. 753–767, 2001.

[35] Angelo Farina, Ralph Glasgal, Enrico Armelloni, and Anders Torger, "Ambiophonic Principles for the Recording and Reproduction of Surround Sound for Music," in *19th AES Conference on Surround Sound, Techniques, Technology and Perception*, Schloss Elmau, Germany, 2001.

[36] Bruno Sanches Masiero and Michael Vorländer, "SPATIAL AUDIO REPRODUCTION METHODS FOR VIRTUAL REALITY," in *Anais del 42o Congreso Español de Acústica*, Cáceres, Spain, 2011.

[37] Michael A. Gerzon, "Ambisonics in multichannel broadcasting and video," *J. Audio Eng. Soc.*, vol. 33, no. 11, pp. 859–871, 1985.

[38] Jérôme Daniel, Rozenn Nicol, and Sébastien Moreau, "Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging," in *114th AES Conference*, Amsterdam, Netherlands, 2003, pp. 2764–2778.

[39] Filippo Maria Fazi, *Sound Field Reproduction*, Ph.D. thesis, Uuiversity of Southampton, 2010.

[40] M Kolundzija, C Faller, and Martin Vetterli, "Sound Field Reconstruction: An Improved Approach For Wave Field Synthesis," in *Preprint 126th Conv. Aud. Eng. Soc*, Munich, Germany, 2009.

[41] Sébastien Moreau, Jérôme Daniel, and Stéphanie Bertet, "3D sound field recording with higher order ambisonics-objective measurements and validation of spherical microphone," in *120th AES Convention*, 2006, vol. 5, pp. 1–24.

[42] Juha Vilkamo, Tapio Lokki, and Ville Pulkki, "Directional audio coding: Virtual microphone-based synthesis and subjective evaluation," *AES: Journal of the Audio Engineering Society*, vol. 57, no. 9, pp. 709–724, 2009.

[43] Jürgen Herre, Cornelia Falch, Dirk Mahane, Giovanni Del Galdo, Markus Kallinger, and Oliver Thiergart, "Interactive teleconferencing combining spatial audio object coding and DirAC technology," *Journal of the Audio Engineering Society*, vol. 59, no. 12, pp. 924–935, 2012.