UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

JOEL AUGUSTO LUFT

SEPARAÇÃO DE SINAIS DE ÁUDIO ATRAVÉS DA DENSIDADE DE POTÊNCIA ESPECTRAL DIRECIONAL

Porto Alegre
2019

JOEL AUGUSTO LUFT

SEPARAÇÃO DE SINAIS DE ÁUDIO ATRAVÉS DA DENSIDADE DE POTÊNCIA ESPECTRAL DIRECIONAL

Tese de doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Rio Grande do Sul como parte dos requisitos para a obtenção do título de Doutor em Engenharia Elétrica.

Área de concentração: Engenharia de Computação

ORIENTADOR: Prof. Dr. Altamiro Amadeu Susin

Porto Alegre
2019

JOEL AUGUSTO LUFT

SEPARAÇÃO DE SINAIS DE ÁUDIO ATRAVÉS DA DENSIDADE DE POTÊNCIA ESPECTRAL DIRECIONAL

Esta tese foi julgada adequada para a obtenção do título de Doutor em Engenharia Elétrica e aprovada em sua forma final pelo Orientador e pela Banca Examinadora.

Orientador: _____
Prof. Dr. Altamiro Amadeu Susin, UFRGS
Doutor pela Polytechnique National Institute of Grenoble – Grenoble, France

Banca Examinadora:

Prof. Dr. Claudio Rosito Jung, PPGC-UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul – Porto Alegre, Brasil

Prof. Dr. Márcio Holsbach Costa, UFRGS
Doutor pela Universidade Federal de Santa Catarina – Florianópolis, Brasil

Prof. Dr. Gilson Inácio Wirth, UFRGS
Doutor pela University of Dortmund – Dortmund, Germany

Prof. Dr. Valner João Brusamarello, UFRGS
Doutor pela Universidade Federal de Santa Catarina – Florianópolis, Brasil

Prof. Dr. Edison Pignaton de Freitas, UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul – Porto Alegre, Brasil.

Coordenador do PPGEE: _____
Prof. Dr. João Manoel Gomes da Silva Jr.

Porto Alegre, agosto de 2019.

AGRADECIMENTOS

Quero agradecer a Deus, por ter abençoado todos os dias da minha vida e me dar forças para seguir sempre em frente.

Gostaria de agradecer a Eliane, minha esposa, pelo permanente apoio e paciência que sempre demonstrou no decorrer do trabalho. Agradeço ainda o amor e compreensão demostrados diante dos imensos sacrifícios necessários durante esse período.

À minha filha pelos momentos de alegria proporcionados.

A meus pais e todos da minha família sempre presentes e compreensivos.

Agradeço ao meu orientador Dr. Altamiro Amadeu Susin pela oportunidade, confiança depositada e motivação.

Ao Programa de Pós-Graduação em Engenharia Elétrica, PPGEE, pela oportunidade de realização de trabalhos em minha área de pesquisa.

Ao IFRS pelo apoio para realização deste doutorado.

Aos colegas do IFRS se dispuseram a assumir meus encargos durante o período que me afastei para realização do trabalho. Em especial ao professor Claudio Enrique Fernandéz Rodríguez que prontamente colaborou nos momentos finais de defesa da tese.

A todos os colegas do LAPSI , em especial ao Dr. Marcelo Negreiros e Dr. Fábio Irigon Pereira pelo apoio durante todo o trabalho.

RESUMO

A separação das fontes sonoras é uma das principais preocupações em muitas aplicações, como sistemas de comunicação, aparelhos auditivos, reconhecimento de fala, etc. Frequentemente, o número de fontes a serem separadas excede o número de microfones, tornando importante lidar com os chamados casos subdeterminados. Este trabalho propõe novos métodos para separar sinais de áudio com base na estimativa da Densidade Espectral de Potência (PSD), usando a diretividade de *beamformers* para estimar a PSD de cada fonte sonora. O primeiro método proposto usa a combinação de restrição não negativa com solução de mínimos quadrados para obter a densidade de potência espectral na direção de interesse. O segundo método tolera a correlação entre as fontes, diferentemente das abordagens anteriores na literatura, que tratam apenas da separação de sinais de fontes não correlacionadas. Além disso, é proposta uma terceira abordagem em que o número de fontes excede o número de microfones. Nos métodos propostos presume-se que as funções de transferência acústica (ATFs) entre fontes de som e microfones sejam conhecidas. Como as ATFs geralmente não estão disponíveis e são difíceis de obter em casos reais, as Funções de Transferência Relativa (RTFs), que podem ser obtidas diretamente dos sinais dos microfones, são usadas obtendo resultados semelhantes. Neste trabalho, também são propostos métodos para estimar as RTFs quando existem várias fontes. Um método utiliza detecção de picos em histogramas suavizados e ponderados pelas estimas de PSDs. Outro utiliza o algoritmo Fuzzy C-Means (wFCM) para enfatizar pistas confiáveis no processo de agrupamento, empregando pesos baseados na distribuição de probabilidade da fala, a qual é bem descrita pela Distribuição Laplaciana (LD). Os resultados da simulação mostram que os métodos propostos superam outras abordagens e também suportam correlação dos sinais podendo lidar com configurações subdeterminadas.

Palavras-chave: Separação de sinais, Arranjo de microfones, Densidade de Potência Espectral, Função de Transferência Relativa.

ABSTRACT

Sound sources separation is a main concern for many applications such as communication systems, hearing aids, speech recognition, etc. Frequently, the number of sources to be separated exceeds the number of microphones, and it is important to deal with the so-called underdetermined cases. This work proposes new methods for separating audio signals based on the Power Spectral Density (PSD) estimation using the directivity of the beamformers to estimate the PSD of each sound source. The first proposed method uses the combination of non-negative constraint with least squares solution to obtain the spectral power density in the direction of interest. The second method tolerates correlation between the sources, differently from previous approaches in the literature that address only the separation of signals from uncorrelated sources. Additionally, a third approach where the number of sources exceeds the number of microphones is proposed. The Acoustic Transfer Functions (ATFs) between sound sources and microphones are assumed to be known in the proposed methods. Since ATFs are often unavailable and are hard to obtain in real cases, the Relative Transfer Functions (RTFs), which can be obtained directly from the microphone signals, are used with similar results. In this work, we also propose methods to estimate the RTFs when there are several sources. One uses peak detection in smoothed histograms weighted by the PSD estimates. Another uses a weighted Fuzzy C-Means (wFCM) algorithm to emphasize reliable clues in the clustering process employing weights based on the speech probability distribution, which is well described by Laplacian Distribution (LD). The simulation results show that the proposed methods outperform other approaches and, also, support signal correlation and can handle underdetermined configurations.

Keywords: Source separation, beamforming, Power Spectral Density, Relative Transfer Function.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ATF	Acoustical Transfer Function
DOA	Direction of Arrival
DS	Delay-and-Sum
DUET	Degenerate Un-mixing Estimation Technique
DPC	Directional PSDs estimation with Correlation
DPCM	Directional PSDs estimation with Correlation restricted to M sources
DPCMb	Directional PSDs estimation with Correlation restricted to M source with beamformer filtering
DPNN	Directional PSDs estimation with Non-Negative restrictions
ESS	Exponential Sine Sweep
FCM	Fuzzy c-Means
FDUET	Frequency dependente DUET
GSC	Generalized Sidelobe Canceller
HRTF	Head-Related Transfer Function
HRIR	Head-Related Impulse Response
ICA	Independent Component Analysis
LS	Least Square
MMSE	Minimum Mean Square Erros
MSL	Maximum Length Sequence
MWF	Multichannel Wiener Filter
MVDR	Minimum Variance Distortionless Response
NMF	Non-negative Matrix Factorization
NNLS	Nonnegative Least Squares
PESQ	Perceptual Evaluation of Speech Quality
PSD	Power Spectral Density
RLFSI	Robust Least-Squares Frequency-Invariant

RTF Relative Transfer Function

SDR Signal-to-Distortion Ratio

SDW-MWF Speech-Dstortion-Weighted Multichannel Wiener Filter

SIR Signal-to-Interference Ratio

STOI Short-Time Objective Intelligibility

STFT Short-Time Fourier Transform

W-DO W-Disjoint Ortogonality

wFCM Weighted Fuzzy c-Means

wFCM-L Weighted Fuzzy c-Means with Laplacian Distribution

wFDUET Weighted Frequency dependente DUET

SUMÁRIO

1 INTRODUÇÃO

O processamento de áudio faz referência a qualquer processamento envolvendo algum tipo de som ou simplesmente um sinal dentro da faixa de frequência audível pelo ser humano. Neste contexto são inúmeras as possibilidades de processamento e análise envolvendo aquisição, processamento, armazenamento, transmissão e reprodução dos sons. Com relação ao processamento espacial faz-se referência, além do som, à localização no espaço onde o som é gerado ou reproduzido. São muitas as aplicações envolvendo o processamento espacial, por exemplo: sistemas de áudio e videoconferência, reconhecimento de voz, gravações de áudio de alta qualidade, sistemas de realidade virtual e realidade aumentada, jogos, reprodução de ambientes virtuais em filmes e shows, reprodução de sons em equipamentos portáteis, separação de sinais em próteses auditivas reconhecidamente incômodas por amplificar os sinais de todo ambiente com perda de diretividade, etc.

A detecção do som envolve o processamento de sons captados por microfones onde o conteúdo do sinal de áudio de interesse pode estar imerso em outros sons que corrompem ou mascaram o som de interesse. Para isso o processamento envolve separação dos sons emitidos e uma das possibilidades é incorporar o conhecimento espacial, ou seja, utilizar a característica de que cada elemento do som a ser processado tem origem em diferentes posições no espaço. A audição humana trata de maneira bastante eficiente essas condições e uma das razões é a habilidade de identificar o local de origem do som possibilitando o entendimento e separação de diferentes sons emitidos simultaneamente. Na reprodução do som, o interesse principal é modificar a percepção da localização espacial onde se encontra a fonte sonora. Isso se deve a uma característica importante do sistema auditivo humano que é a capacidade de monitorar eventos em todas as direções pela identificação da posição da origem da fonte sonora. A geração de ambientes virtuais, onde se procura reproduzir no espaço a localização da fonte sonora, é amplamente utilizada como na reprodução de música em dois canais (som estéreo) e de filmes com múltiplos canais (Dolby 5.1 e 7.1 canais). Estes sistemas utilizam fontes de som também distribuídas espacialmente pelo ambiente para sintetizar sons no espaço. Este trabalho apresenta aspectos da separação espacial de sinais sonoros visando sua utilização em sistemas onde o número de fontes sonoras pode ser maior que o de microfones.

A separação de sinais de áudio é um problema complexo e não se tem soluções definitivas para isso. Basicamente dois aspectos caracterizam o problema: cego/não cego, que se caracteriza pelo conhecimento ou não do processo de mistura dos sinais ou determinado/subdeterminado que está relacionado ao número de fontes de som e de microfones disponíveis (VINCENT *et al.*, 2014). O objetivo principal da separação de sinais de áudio é estimar as fontes sonoras a partir do som captado em microfones. Este trabalho propõe métodos para serem utilizados no processo de separação de sinais de áudio utilizando arranjos de microfones baseados no conhecimento da localização espacial da fonte sonora. Algumas aplicações podem restringir o número de microfones, o que pode ser uma limitação e reforça a importância no caso subdeterminado, onde o número de fontes a serem separadas excede o número de microfones. O trabalho se concentra em soluções com reduzido número de microfones visando também possibilitar implementações binaurais. Arranjos de microfones são utilizados para este fim, pois a combinação dos sinais obtidos de um arranjo de microfones resulta em uma resposta que depende da posição espacial da origem do som. Esta combinação que depende da quantidade e da disposição dos microfones é conhecida como formador de feixe ou *beamformer*. Esses tipos de arranjos imitam com alguma semelhança as propriedades do sistema auditivo humano que também se baseia nas diferenças de propagação do som entre os ouvidos (BREEBART; FALLER, 2007).

Os métodos apresentados com o uso de *beamformers* pressupõem o conhecimento da localização da fonte sonora ou da direção de chegada (DOA - *Direction of Arrival*) representada pela Função de Transferência Acústica (ATF - *Acoustic Transfer Function*) e assumem que os sinais envolvidos: sinal desejado, sinais interferentes e ruídos são mutuamente não correlacionados.

Neste trabalho, são propostos métodos para separação de sinais utilizando o mesmo princípio proposto por Hioka (HIOKA *et al.*, 2013), no qual a estimativa das PSDs das fontes sonoras é feita baseada nos ganhos de diretividade dos *beamformers*, ou seja, usa-se a propriedade dos *beamformers* de apresentar respostas que dependem da posição espacial das fontes. Os métodos propostos superam as limitações apresentadas pelo método de Hioka e apresentam melhor desempenho na separação dos sinais.

Devido às características não estacionárias e da esparsidade dos sinais de áudio, a análise e o processamento dos sinais são realizados do domínio da Transformada de Fourier de Curta Duração (STFT - *Short Time Fourier Transform*), o que implica no janelamento do sinal com efeitos já conhecidos no sinal transformado (OPPENHEIM; SCHAFER, 2009). A condição em que a representação em tempo-frequência de duas fontes de sinal não se sobrepõe, chamada W-DO (W-*Disjoint Ortogonality*), é uma aproximação comprovada para o caso de sinais de voz e música no domínio da STFT (RICKARD; YILMAZ, 2002; YILMAZ; RICKARD, 2004), porém, não é possível assumir a perfeita não correlação entre os sinais neste caso. Assumindo que os sinais podem estar correlacionados, a

solução em (HIOKA *et al.*, 2013) não está correta porque negligencia a correlação cruzada entre os sinais.

A primeira contribuição deste trabalho é a apresentação de um método que aborda o problema aplicando restrições às soluções. Uma abordagem de minimização de mínimos quadrados com restrição Não Negativa é usada para resolver a formulação de ganho de Diretividade das PSD, esta abordagem permite que a limitação do número de fontes seja ultrapassada sem a instabilidade numérica até então existente, este método foi denominado DPNN (Directional PSDs estimation with Non-Negative restrictions). Esta abordagem encontra uma solução que minimiza o erro quadrático e ao mesmo tempo restringe a solução a valores não negativos, que está de acordo com o sentido físico das grandezas estimadas. Esta solução também pressupõe a não correlação entre os sinais a serem separados na sua formulação. Também diferentemente do proposto por Hioka (HIOKA *et al.*, 2013) que utiliza o filtro de Wiener, calculado com as estimativas de PSD, na saída dos *beamformers* esta nova proposta aplica o filtro aos sinais cujas magnitudes são obtidas pela raiz quadrada da estimativa das PSDs em conjunto com fase da saída dos *beamformers*.

Também é proposto um método para separar sinais assumindo explicitamente a existência de correlação entre fontes, ao contrário do pressuposto de não correlação entre os sinas assumido nos demais métodos já citados. O método proposto é baseado nas PSDs das saídas dos *beamformers* e nas PSDs cruzadas (Cross-PSDs) das saídas dos *beamformers* e é chamado DPC (Directional PSDs estimation with Correlation). Além de assumir a correlação, no método proposto o método também permite estimar a amplitude e a fase dos sinais antes da filtragem para rejeição das fontes interferentes.

Com aumento do número de fontes de sinal a correlação tende a aumentar mas a aproximação W-DO não deixa de ser válida devido à esparsidade dos sinais, pode-se então aplicar o método DPC apenas sobre as fontes dominantes e obter a separação no caso subdeterminado. Este método considera que o número de fontes dominantes não excede o numero de Microfones, o método é denominado DPCM (Directional PSDs estimation with Correlation restricted to M sources).

Os métodos propostos utilizam as Funções de Transferência Acústica (ATFs) entre fontes e microfones. As estimativas das ATFs em situações práticas nem sempre é possível e são difíceis de obter. Muitos métodos de separação utilizam as Funções de Transferência Relativa (RTFs), isto é, a relação de funções de transferência acústica entre dois sensores, em vez de ATFs, porque podem ser obtidas a partir dos microfones (GANNOT *et al.*, 2017). Por esse motivo, a aplicabilidade das RTFs nos métodos propostos é demonstrada. A utilização das RTFs nos métodos apresenta resultados equivalentes às ATFs.

Duas propostas para estimativa das RTFs para várias fontes também foram desenvolvidas neste trabalho. Um dos métodos proposto é baseado em histogramas ponderados

suavizados de estimativa da RTF instantânea, similar à abordagem DUET (YILMAZ; RICKARD, 2004), mas sem o uso de atraso fixo para todas as frequências, e com utilização de pesos determinados por um novo processo iterativo com realimentação das estimativas das PSDs das fontes; as PSDs são determinadas pelos métodos de separação propostos. O outro método proposto é baseado no agrupamento ponderado *Fuzzy C-Means* (wFCM). Neste método, enfatizam-se características significativas aplicando a distribuição Laplaciana para determinação de pesos, este método foi chamado wFCM-L. Os métodos propostos são avaliados em termos de erros de estimativa das RTFs e separação de fontes.

Os métodos propostos para separação de sinais de áudio assumindo correlação entre as fontes, DPC e DPCM, o uso das RTFs aplicada aos métodos propostos e o método de estimativa das RTFs por aglomeração, wFCM-L, compõem um artigo [1] já submetido para publicação em periódico.

[1] LUFT, J. A.; PEREIRA, F. I.; SUSIN, A. **Directional sound source separation with relative transfer function estimation**. Submetido à publicação no Journal of Audio Engineering Society. 2019

2 TRABALHOS RELACIONADOS

Existem várias abordagens para o problema da separação de sinais de áudio. A utilização de arranjos de microfones é uma boa alternativa para separar o som quando as características espaciais estão envolvidas, pois eles têm respostas diferentes de acordo com a direção ou posição da localização da fonte de som (BENESTY; CHEN; HUANG, 2008). A separação também pode ser estimada sem conhecimento prévio das fontes e de suas posições como na análise de componentes independentes (ICA – *Independent Component Analysis*), porém, o número de fontes está restrito ao número de microfones (BENESTY; CHEN; HUANG, 2008). Em outra abordagem, a separação pode ser feita por técnicas de fatoração de matrizes não negativas (NMF – *Non-negative Matrix Factorization*) que não possui limitação do número de fontes, mas requer treinamento ou processamento *offline* de todo o sinal (YOSHII *et al.*, 2013; JODER *et al.*, 2012; FENG; KOWALSKI, 2017). As técnicas de mascaramento binário, mais utilizado do caso de separação cega, tratam o problema subdeterminado e assumem que apenas uma fonte está ativa durante um determinado tempo e em determinada frequência, utilizando essa premissa para extrair a fonte predominante (YILMAZ; RICKARD, 2004).

O termo *beamformer* se refere a um filtro espacial usado na saída de um arranjo de microfones que é dependente da frequência e da direção dos sinais e podem ser classificados como independentes dos dados ou dependentes dos dados (HIDRI; MEDDEB; AMIRI, 2012). Os *beamformers* independentes dos dados, também chamados de *beamformers* fixos ou *beamformers* determinísticos não dependem dos sinais captados nos microfones e são projetados em função da resposta desejada para cada direção. Os *beamformers* dependentes dos dados utilizam propriedades estatísticas dos sinais para otimizar os filtros de acordo com algum critério.

Nos *beamformers* fixos a posição das fontes dos sinais determina os parâmetros dos filtros, esta solução é determinística, mas tem desempenho de separação limitado, o exemplo principal é o *Delay-and-Sum Beamforming* (DS) (BENESTY; CHEN; HUANG, 2008; VAN TREES, 2002; HAYKIN; LIU, 2009; BRANDSTEIN; WARD, 2001) onde os sinais dos microfones são atrasados de modo que sinais originados de determinada direção sejam somados em fase. Outras abordagens utilizam algoritmos de minimização para obter

os coeficientes que aproximam a resposta desejada como nos *beamformers* superdiretivos (DOCLO; MOONEN, 2007) e nos *beamformers* invariantes a frequência (RLFSI -Robust Least-Squares Frequency-Invariant)(MABANDE; SCHAD; KELLERMANN, 2009; BARFUSS *et al.*, 2015).

Outras técnicas usam comportamento estatístico dos sinais captados como correlação ou densidade de potência espectral e utilizam essas medidas para melhorar os sinais desejados enquanto rejeitam os sinais indesejados. Nesta categoria está o *beamformer* de Resposta sem Distorção de Variância Mínima (MVDR - *Minimum Variance Distortionless Response*) (GANNOT; COHEN, 2008) ou filtragem espacial adaptativa como no *Generalized Sidelobe Canceller* (GSC) (GRIFFITHS; JIM, 1982). Estas abordagens podem ser estendidas com versões para solução binaural (HADAD; GANNOT; DOCLO, 2012; HADAD; DOCLO; GANNOT, 2016; ZOHOURIAN; MARTIN, 2016). Outra abordagem é a utilização do Filtro de Wiener Multicanal (MWF - *Multichannel Wiener Filter*) também baseada nas características estatísticas de segunda ordem dos sinais observados e é utilizada para filtrar os ruídos e sinais interferentes (GANNOT; COHEN, 2008). O MWH é a melhor aproximação no sentido do Mínimo Erro Quadrático Médio (MMSE - Minimum Mean Square Error) e pode se visto e decomposto em um *beamformer* MVDR seguido de um filtro Wiener (BRANDSTEIN; WARD, 2001). O MWF pode ser generalizado considerando o compromisso entre redução de ruído e distorção da voz pelo SDW-MWF (Speech Distortion Weighted MWF) (DOCLO *et al.*, 2007; HAYKIN; LIU, 2009) e é também aplicado à separação binaural (KUKLASINSKI; JENSEN, 2017). Essas técnicas requerem informações espaciais das fontes de som e detectores de fala (VAD - *Voice Activity Detectors*) para identificar quando os sinais desejados estão ativos.

A recente introdução de *deep learning* à separação de fala supervisionada acelerou o progresso e aumentando o desempenho da separação, com grandes avanços em separar a fala de outros sinais diferentes ou separação dependente do locutor. Entretanto a situação independente do locutor com múltiplos locutores ainda é um problema que requer avanços, (WANG; CHEN, 2018)

O desempenho dos algoritmos de extração de sinal nas soluções dependentes dos dados depende criticamente das estimativas das estatísticas dos componentes de sinal desejados e indesejados. A implementação desses algoritmos exige a estimativa constante dos componentes desejados ou não desejados, no entanto, todos esses métodos ainda sofrem com o fato de que as estimativas de interferência e ruído não podem ser atualizadas enquanto a fonte alvo está ativa, de modo que elas são propensas a falhas com ruído não estacionário e interferência como é o caso da fala humana (DOCLO *et al.*, 2015).

O uso de *beamformers* para separação de múltiplas fontes para o caso subdeterminado, onde o número de fontes separáveis pode ser maior que o número de microfones, foi abordado por Hioka em (HIOKA *et al.*, 2013) onde propõe um método para estimar as densidades de potência espectral (PSDs - *Power Spectral Densities*) das fontes para

aplicação de pós-filtragem com filtro de Wiener. O método proposto por Hioka permite a separação de no máximo $M(M-1)+1$ fontes para M microfones e esse limite diminui dependendo do posicionamento dos microfones e das fontes sonoras. Também apresenta problemas de condicionamento que resulta em erros de estimativa e obtenção de valores de PSDs negativas, o que está em desacordo com a interpretação física da grandeza. Este método foi estendido em (HIOKA; NIWA, 2014) para estimativa da PSD de ruídos incoerentes, ou seja, onde uma distribuição isotrópica da energia sonora pode ser assumida. Também é apresentada proposta para o projeto dos *beamformers* associados ao método do Hioka para reduzir o problema de condicionamento (NIWA; HIOKA; KOBAYASHI, 2019).

A Tabela 1 apresenta resultados de separação de fontes de de métodos relacionados. A comparação é apresentada com resultados de separação em termos de melhoria de SIR e SDR, descritos na seção 3.6, onde valores maiores indicam melhor separação. Nesta tabela, Luft,2019 apresenta resultados dos métodos propostos neste trabalho.

Tabela 1 – Comparação dos métodos de separação

	Método	**N/M**	**Anecoico SIR/SDR (dB)**	**Reverberante SIR/SDR,(dB)**
Winter,2004	Hierarchical cluster	3/2	– / –	9,9 / 6,5,@130ms
Araki, 2007	Kmeans	4/3	15,5 / 9	12 / 7,5,@128ms
Jafari, 2013	Kmeans	4/3	17 / –	16 / –,@128ms
Hioka, 2013	PSD direcional	3/3	17 / 15	10 / 12,@300ms
Atcheson,2014	sFCM	3/2	17 / –	14 / –,@150ms
Feng, 2017	NMF	3/2	-	4 / 2,5,@$\sim$130ms 5 / 2,8,@ 250ms
Luft, 2019	DPC/DPCM	4/3	24.6 / 14	15,2 / 6,9@100ms 9,7 / 2,6,@200ms

Apesar das RTFs poderem ser obtidas diretamente dos microfones, a metodologia para obter as RTFs para múltiplas fontes ainda não está bem estabelecida. As técnicas atuais de estimativa de RTF têm limitações: como a necessidade de detectores eficientes de fala (VAD - *Voice Activity Detector*) para determinar quando as fontes de interesse estão ativas (MARKOVICH; GANNOT; COHEN, 2009; COHEN, 2004); restrição ao caso determinado, onde o número de fontes menor que o número de microfones, e sem ruído (DELEFORGE; GANNOT; KELLERMANN, 2015); não resultam em RTF individual por fonte (TASESKA; HABETS, 2015); ou precisam de treinamento (BAO *et al.*, 2013). Nos métodos para o caso subdeterminado, os autores não apresentam explicitamente os erros de estimação. Eles apresentam a estimativa de RTF como um estágio na separação dos sinais e geralmente apresentam apenas os resultados finais da separação de fontes (YILMAZ; RICKARD, 2004; WINTER *et al.*, 2004; SAWADA; ARAKI; MAKINO, 2011; ITO; ARAKI; NAKATANI, 2015).

3 REVISÃO TEÓRICA

Neste capítulo é formalizado o problema, como também apresentado as variáveis e condições envolvidas na separação das fontes de sinal. Também são mostrados *beamformers* utilizados na separação de sinais, baseadas em arranjos de microfones. Também são apresentadas a principais métricas utilizadas para a avaliação de sistemas de separação de sinais de áudio.

3.1 Definição do Problema

Nesta sessão o problema de separação da fonte sonora multicanal é formulado como na Figura 1. Supondo N fontes de som diferentes, cujos sinais são capturados por um arranjo de M microfones, o sinal do microfone m no domínio do tempo pode ser formulado com o seguinte modelo de mistura:

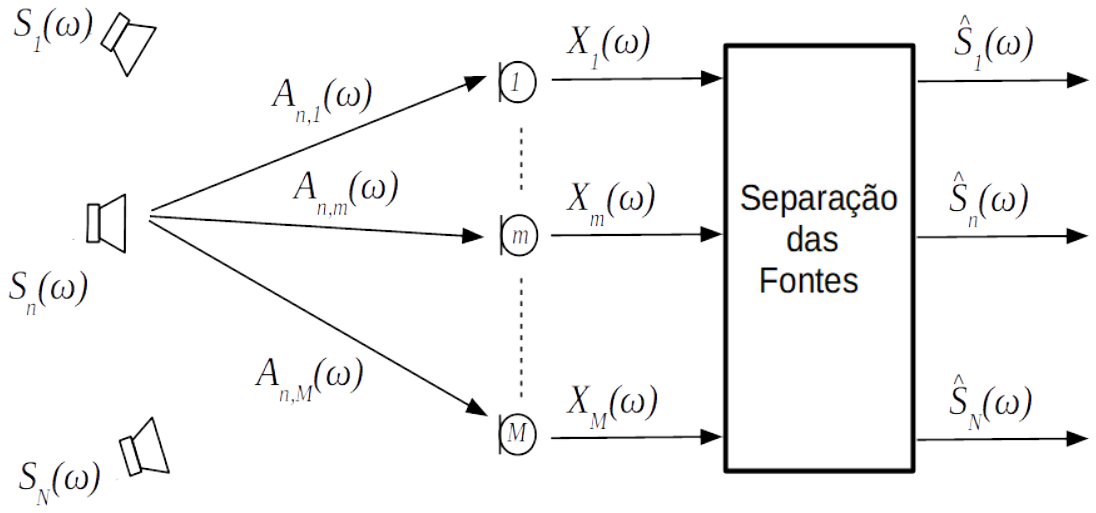
$$x_m(t) = \sum_{n=1}^{N} a_{n,m}(t) * s_n(t) + v_m(t) \tag{1}$$

onde $s_n(t)$ é a fonte de sinal n, $v_m(t)$ é a contribuição de todas as fontes de ruído e $a_{n,m}(t)$ corresponde à resposta ao impulso entre a fonte n e o microfone m e $*$ significa convolução. Esta resposta ao impulso inclui as características do ambiente e a resposta do microfone e sua representação no domínio frequência pela transformada de Fourier é denominada Função de transferência Acústica (ATF - *Acoustic Transfer Function*).

Assume-se um cenário estático onde os sinais recebidos pelos microfones podem ser aproximados no domínio STFT (*Short-Time Fourier Transform*) como uma multiplicação de um sinal anecoico com uma ATF invariante no tempo e que a aplicação de uma transformada de Fourier de curta duração em (1) pode ser aproximada no domínio tempo-frequência pela chamada aproximação de banda estreita (KELLERMANN; BUCHNER, 2003; KOWALSKI; VINCENT; GRIBONVAL, 2010) como

$$X_m(k,\omega) \approx \sum_{n=1}^{N} A_{n,m}(\omega) S_n(k,\omega) + V_m(k,\omega) \tag{2}$$

Figura 1 – Formulação do problema.



Fonte: do autor

onde ω é a frequência, k é o índice da janela de amostragem e $X_m(k,\omega)$, $A_{n,m}(\omega)$, $S_n(k,\omega)$ and $V_m(k,\omega)$ são as representações no domínio frequência de $x_m(t)$, $a_{n,m}(t)$, $s_n(t)$ e $v_m(t)$, respectivamente. Sempre que aplicável, os sinais envolvidos neste trabalho são tratados no domínio da frequência, o índice da janela de amostragem k será omitido e o termo de ruído é desconsiderado. Desta forma os sinais das fontes sonoras ficam

$$\mathbf{s}(\omega) = [S_1(\omega), S_2(\omega), ..., S_M(\omega)]^T \tag{3}$$

onde ^{T} é a transposta, as funções de transferência acústica

$$\mathbf{A}_n(\omega) = [A_{n,1}(\omega), A_{n,2}(\omega), ..., A_{n,M}(\omega)]^T \tag{4}$$

e os sinais na saída dos microfones

$$\mathbf{x}(\omega) = [X_1(\omega), X_2(\omega), ..., X_M(\omega)]^T \tag{5}$$

.

O problema abordado neste trabalho é estimar cada fonte $S_n(\omega)$ com base nos sinais observados $X_m(\omega)$ enquanto assumem-se alguns pré-requisitos para permitir a aplicabilidade dos métodos. As fontes de som estão em diferentes posições espaciais localizadas ao redor dos microfones e chegam aos microfones de diferentes direções. A separação será feita com base na suposição de que a posição espacial de fontes e microfones ou as ATFs, e a geometria do conjunto de microfones (localização relativa dos microfones entre si) são conhecidos. É importante notar que a ATF, que inclui as características do ambiente e a resposta do microfone, geralmente, não é conhecida em situações práticas, pois requer o conhecimento de todo o ambiente ou um procedimento de medição para

obter essa função. Outra consideração, se não mencionada de outra forma, é que as fontes sonoras são mutuamente não correlacionadas.

$$E[S_n(\omega)S^*_{n'}(\omega)] = 0 \quad \text{if } n \neq n' \tag{6}$$

onde $E[.]$ é o operador de esperança.

3.2 Características dos sinais

O primeiro ponto a considerar está relacionado às características das fontes de sinal. O principal foco deste trabalho é o tratamento de sinais de áudio, principalmente sinais de voz. Estes sinais têm algumas propriedades que devem ser consideradas para o problema em questão. Os sinais se caracterizam por ser um sinal de banda larga abrangendo o espectro de frequência que varia aproximadamente de 20Hz a 20kHz. Algumas propriedades relevantes a serem consideradas para separação de sinais são (MAKINO; SAWADA; LEE, 2007):

- Sinais de fala originados de diferentes locutores e originados de diferentes localizações em um ambiente acústico podem ser considerados estatisticamente independentes.
- Os sinais de voz são quase estacionários para curtos períodos de tempo (10ms) mas são não estacionários para longos períodos.
- Cada fala apresenta estrutura temporal única (diferentes funções de autocorrelação) para períodos menores que 1 segundo.

Essas características remetem ao tratamento e análise dos sinais de áudio em uma janela pequena de tempo em que as propriedades temporais de energia e de correlação podem ser consideradas fixas para intervalos de tempo da ordem de 10ms a 30ms (RABINER; SCHAFER, 1978).

Outra consideração importante é definição de esparsidade onde quanto mais amostras nulas contiver um sinal, mais esparsas as amostras serão, o que significa que os sinais se sobrepõem com pouca frequência (BLIN; ARAKI; MAKINO, 2003). A suposição de esparsidade significa que apenas algumas amostras são significativamente diferentes de zero. Se as fontes não forem esparsas em seu domínio original (por exemplo, o domínio de tempo para sinais de áudio), elas podem ser esparsas em um domínio transformado (por exemplo, o domínio de Fourier, transformação wavelet) (TAN, 2005). Sinais de áudio como voz e música são esparsos no domínio de Fourier e essa esparsidade é descrita pelo conceito da W-*Disjoint Orthogonality* (W-DO)(RICKARD; YILMAZ, 2002). Duas funções s_1 e s_2 são chamadas W-DO se, para uma determinada função janela W, as STFTs de s_1 e s_2 são disjuntas.

A STFT de s_j é definida como

$$S_j(k,\omega) = \int_{-\infty}^{\infty} W(t-k)s_j(t)e^{-\omega t}dt \tag{7}$$

onde a função janela $W(t-k)$ determina o intervalo de tempo que o sinal é analisado, com k definindo o índice de tempo em que a janela de análise é aplicada.

A W-DO pode ser colocada de forma concisa como

$$S_1(k,\omega)S_2(k,\omega) = 0, \forall k,\omega \tag{8}$$

A representação no domínio da STFT quando a voz está ativa raramente é zero em desacordo com (8), porém, devido à esparsidade da fala ocorre que, em um determinado intervalo de tempo, a contribuição de um locutor é geralmente bem maior que a do outro. Em (YILMAZ; RICKARD, 2004) é demonstrado que se podem considerar os sinais da fala como aproximadamente W-DO. O tamanho da janela usado na STFT influencia no quanto os sinais são considerados W-DO e foi constatado em avaliações experimentais que janelas entre 32 ms e 128 ms com janela de Hanning apresentaram os melhores resultados (BLIN; ARAKI; MAKINO, 2003; YILMAZ; RICKARD, 2004).

3.3 Estimação da PSD

Para um processo estacionário a PSD descreve o conteúdo espectral de um sinal. A densidade de potência espectral (PSD - Power Spectral Density) de $x(t)$ é a Transformada de Fourier de sua autocorrelação. Sendo a função de autocorrelação de um sinal $x(t)$ definida como (PAPOULIS; PILLAI, 2002)

$$R(\tau) = E[x(t-\tau)x^*(t)] \tag{9}$$

a densidade de potência espectral (PSD - Power Spectral Density) é

$$\phi_{xx}(\omega) = \int_{-\infty}^{-\infty} R(\tau)e^{-j\omega\tau}d\tau \tag{10}$$

Como já mencionado, os sinais de áudio apresentam características não estacionárias e a PSD não deve ser calculada para intervalos grandes de tempo. A forma mais simples de estimar a PSD é simplesmente utilizar a STFT, mas fica muito sujeita à interferência de ruídos. Alguns métodos utilizam a média de janelas consecutivas ou sobrepostas (BARTLETT, 1948; WELCH, 1967). Outra alternativa, que é a forma empregada neste trabalho, é a suavização recursiva (MARTIN, 2001; HIOKA *et al.*, 2013; LI *et al.*, 2016)

calculada de acordo com

$$\Phi_{XX}(k,\omega) = \alpha\Phi_{XX}(k-1,\omega) + (1-\alpha)|X(k,\omega)|^2 \tag{11}$$

onde ω é a frequência, k é o índice da janela de amostragem, $X(k,\omega)$ é a STFT e α determina o grau de suavização.

3.4 Função de transferência acústica (ATF)

As Funções de Transferência Acústica descrevem no domínio das frequências as respostas ao impulso entre a fonte de sinal e o sinal amostrado, incluindo as respostas do ambiente e dos microfones. As técnicas de medida mais utilizadas em áudio são baseadas em varredura em frequência ou utilizam geradores de ruído branco pseudo-aleatório (MSL - *Maximum Length Sequence*) (HOLTERS; CORBACH; ZÖLZER, 2009). Para medição das HRIRs duas abordagens são utilizadas sendo que a técnica de varredura exponencial em frequência (ESS – *Exponential Sine Sweep*) apresenta vantagens em relação à MLS principalmente relacionada à melhor relação sinal ruído por ser menos sensível a distorções do sistema e variações temporais (FARINA, 2000, 2007).

A dificuldade de obtenção dessas funções em situações práticas leva a utilização de aproximações em função das posições relativas entre as fontes sonoras e os microfones. Um modelo comumente usado para a ATF é o vetor de direção (GANNOT *et al.*, 2017), que assume que os sons chegam aos microfones como ondas planas vindas da direção da fonte e todas com a mesma atenuação. Neste caso, o modelo assumindo um arranjo de microfones omnidirecionais é

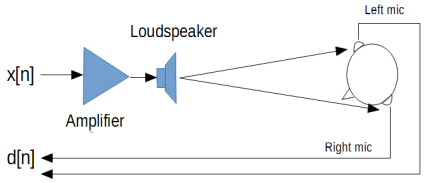
$$A_{n,m}(\omega) = e^{-j.\omega.d_{n,m}/c} \tag{12}$$

onde $d_{n,m}$ é a distância da fonte n para o microfone m e c é a velocidade do som. Esta solução pressupõe o conhecimento das direções das fontes sonoras que podem ser obtidas por técnicas de direção de chegada (DOA) (BRANDSTEIN; WARD, 2001).

Quando se tratam de aplicações binaurais, em que a cabeça de um indivíduo está no caminho acústico com os microfones geralmente colocados próximos aos ouvidos, a função de transferência relacionada à cabeça (HRTF- *Head Related Transfer Function*) representa a resposta ao impulso proveniente de algum local específico no espaço para cada orelha (CHENG; WAKEFIELD, 1999). Estas medidas são geralmente realizadas em câmaras anecoicas e existem bases de dados disponíveis com dezenas de modelos de cabeça diferentes (ALGAZI *et al.*, 2001). As funções de transferência da cabeça são obtidas para centenas de posições ao redor da cabeça. Devido às características de uma pessoa, como o tamanho e a forma da cabeça, da orelha externa e do tronco, a HRTF varia para cada pessoa, direção e distância da fonte de som. A HRTF é a Transformada

de Fourier da resposta ao impulso relacionada à cabeça (HRIR - *Head-Related Impulse Response*). Esta função de transferência é obtida experimentalmente gerando sinais sonoros um som numa posição espacial específica em relação a uma cabeça e adquirindo os sons que chega aos microfones localizados nas orelhas. Pode ser feito diretamente com microfones colocados nos ouvidos de uma pessoa ou em um manequim com cabeça e torso utilizando as técnicas citadas para medição das ATFs.

Figura 2 – Sistema para medição da HRIR/HRTF



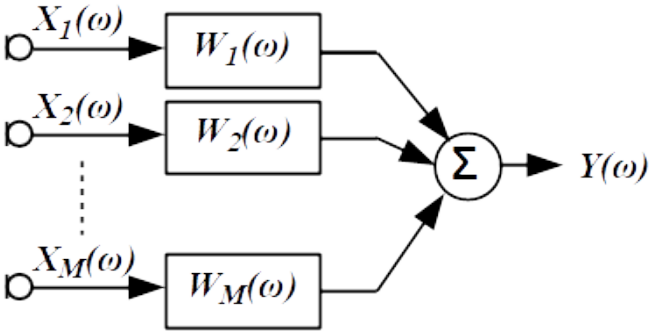
Fonte: do autor

Com relação à metodologia para determinação da HRTF ainda não existe um padrão de procedimento para medição dessas funções que são complexas e de difícil validação. Em (ANDREOPOULOU; BEGAULT; KATZ, 2015) foram comparados medições feitas para a mesma cabeça (manequim) em 10 laboratórios diferentes e, apesar de serem feitos com a mesma cabeça, os resultados revelaram variações de magnitude espectral até 12,5 dB para bandas de frequência abaixo de 6 kHz e até 23 dB para frequências altas, assim como grandes assimetrias espectrais entre o lado esquerdo e direito para o conteúdo de alta frequência e variações de tempo interaural (ITD - *Interaural Time Diference*) de 235 μs, excedendo valores percebidos pelo homem. Também sugerem que estudos subjetivos ainda devem ser feitos para avaliar a relevâncias dessas diferenças e também estudos são necessários para determinar a origem dessas variações. O sistema básico empregado para fazer medições de resposta de impulso relacionadas à cabeça é apresentado na Figura 2.

3.5 Arranjos de microfones para separação de sinais

Nesta seção é apresentado o uso de *beamformers* para filtragem espacial, onde a filtragem esta relacionada com a posição espacial de origem dos sinais envolvidos. Um arranjo de microfones consiste de um conjunto de microfones posicionado de forma a extrair parâmetros ou algum sinal explorando suas características e informações espaciais. A maioria dos problemas envolvendo arranjos de microfones utiliza a estrutura da Figura 3 onde os sinais obtidos dos microfones são filtrados e somados. Estes filtros são otimizados por algum critério e a saída desse arranjo é denominada *beamformer* ou formador de feixe. Geralmente, deseja-se filtrar espacialmente de modo que um sinal de uma determinada direção é reforçado por uma combinação construtiva e o ruído de outros ângulos é rejeitado por interferência destrutiva.

Figura 3 – Estrutura do *beamformer*.



Fonte: do autor

Pode-se definir a saída do *beamformer* como

$$Y_(\omega) = \sum_{m=1}^{M} W_m(\omega) X_m(\omega) \tag{13}$$

onde W_m são os coeficientes dos filtros ou, em notação matricial,

$$\mathbf{y}(\omega) = [Y_1(\omega), Y_2(\omega), ..., Y_M(\omega)]^T \tag{14}$$

onde ^{T} é a transposta.

$$\mathbf{w}(\omega) = [W_1(\omega), W_2(\omega), ..., W_M(\omega)]^T \tag{15}$$

onde ^{T} é a transposta e a saída do filtro fica

$$\mathbf{y}(\omega) = \mathbf{w}^T(\omega)\mathbf{x}(\omega). \tag{16}$$

Nas subseções seguintes são apresentados alguns tipos de *beamformers* que variam basicamente na maneira que os coeficientes $\mathbf{w}$ são definidos. A utilização dos *beamformers* neste trabalho está relacionada às diferenças de ganho que eles apresentam em função da posição espacial do sinal filtrado. Este trabalho não está baseado especificamente no desempenho dos *beamformers* como ferramenta de separação de sinais; apesar dos *beamformers* serem utilizados para rejeitar fontes interferentes, a separação dos sinais neste trabalho não é feita pelos *beamformers*, eles apenas serão utilizados para modificar de forma conhecida o processo de mixagem das fontes sonoras que serão separadas. A princípio qualquer *beamformer* apresentado a seguir pode ser usado nos métodos propostos.

3.5.1 *Beamformer* fixo

O objetivo de um *beamformer* fixo ou *beamformer* independente dos sinais é obter foco na fonte de som baseado na sua posição espacial e reduzir os sons não provenientes da mesma direção que a fonte de interesse. Diferentes tipos de *beamformers* fixos são possíveis tais como *Delay-and-Sum beamformers*, *Superdirective beamformers* e *frequency invariant beamformers*.

3.5.1.1 Delay-and-Sum beamformer (DS)

Os sinais nos microfones de um arranjo chegam a cada um deles com uma pequena diferença de tempo que é função da geometria do arranjo e da direção de chegada do som. A técnica DS consiste em duas etapas: primeiro atrasa o sinal em cada microfone de modo que todos os sinais fiquem em fase e em seguida são somados obtendo o máximo de ganho na direção especificada.

Considerando um caso simples com microfones omnidirecionais recebendo um som direto de uma distância d_{nm} o problema pode ser simplificado e a função de transferência passa a ser um atraso puro

$$A_{n,m}(\omega) = e^{-j.\omega.d_{n,m}/c} \tag{17}$$

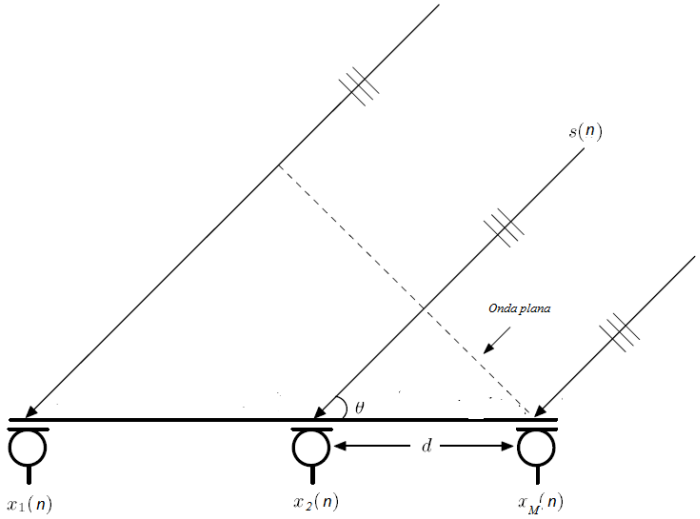
onde c é a velocidade do som.

Este modelo está exemplificado de forma pela Figura 4 e os coeficientes do filtro são:

$$W_{m,\theta}(\omega) = e^{jw(m-1)d\ cos(\theta)/c} \tag{18}$$

onde c é a velocidade de propagação do som e d é a distância entre os microfones.

Figura 4 – Arranjo de microfones linear com onda plana incidente.



Fonte: do autor

3.5.1.2 Beamformer fixo superdiretivo

A resposta do formador de feixes pode ser otimizada considerando certa resposta desejada, ou seja, minimizar o erro entre a saída real do *beamformer* e a saída desejada. A resposta do *beamformer* é:

$$D(\theta, \omega) = W(\theta, \omega)^H A(\theta, \omega) \tag{19}$$

onde $A(\theta, \omega)$ é a ATF.

O erro entre a resposta do *beamformer* $D(\theta, \omega)$ e uma resposta desejada $D_d(\theta, \omega)$ é

$$\xi(\theta, \omega) = D(\theta, \omega) - D_d(\theta, \omega) \tag{20}$$

A função a ser minimizada por mínimos quadrados (LS - *Least Square*) fica

$$J_{LS}(\omega) = \int_\theta V(\theta, \omega) |\xi(\theta, \omega)|^2 d\theta \tag{21}$$

$$\min_W J_{LS}(\omega) \tag{22}$$

onde θ é a região espacial de interesse, $V(\theta, \omega)$ é uma função de ponderação para enfatizar ou reduzir a importância de certas direções. Os coeficientes do filtro ficam (BENESTY; CHEN; HUANG, 2008; LAI; NORDHOLM; LEUNG, 2017)

$$Q(\omega) = \int_\theta V(\theta, \omega) A(\theta, \omega) A^H(\theta, \omega) d\theta \tag{23}$$

$$q(\omega) = \int_\theta V(\theta, \omega) D_d(\theta, \omega) A(\theta, \omega) d\theta \tag{24}$$

$$W(\omega) = Q^{-1}(\omega) q(\omega) \tag{25}$$

Considerando L direções de interesse, os coeficientes do filtro para a direção l ficam (MAAZAOUI; GRENIER; ABED-MERAIM, 2012)

$$W_l(\omega) = conj\left(\frac{(R_{AA})^{-1} A_l(\omega)}{A_l(\omega)^H (R_{AA})^{-1} A_l(\omega)}\right) \tag{26}$$

onde $A_l(\omega)$ é a ATF na direção l e $R_{AA} = \frac{1}{L} \sum_1^L A_l(\omega) A_l^H(\omega)$.

Os *beamformers* superdiretivos são sensível ao ruído e aos desvios das características assumidas do microfone (ganho, fase e posição), especialmente em baixas frequências, mas podem ser minimizadas com restrições e avaliando o compromisso entre diretividade e robustez (BRANDSTEIN; WARD, 2001; DOCLO *et al.*, 2007)

$$W_l(\omega) = conj(\frac{(R_{AA} + \mu I)^{-1} A_l(\omega)}{A_l(\omega)^H (R_{AA} + \mu I)^{-1} A_l(\omega)}) \tag{27}$$

onde o parâmetro μ controla a robustez, quanto maior μ mais robusto e menos diretivo.

3.5.2 *Beamformer* de Mínima Variância sem Distorção

O *beamformer* de mínima variância sem distorção (MVDR - *Minimum Variance Distortionless Response*) (BENESTY; CHEN; HUANG, 2008; HABETS *et al.*, 2010), também conhecido por *beamformer* superdiretivo, objetiva minimizar a potência da saída sujeita a uma restrição linear na direção desejada. A ideia é determinar o conjunto de coeficientes W que minimiza a potêencia de saída, $E[Y^2(\omega)]$, com a restrição de que o sinal na direção desejado não é afetado.

$$E[|Y(\omega)|^2] = E[W(\omega)^H X(\omega) X(\omega)^H W(\omega)] = W(\omega)^H \Phi_{XX} W(\omega), \tag{28}$$

onde $\Phi_{XX} = E[X(\omega)X(\omega)^H]$ é a matriz de correlação dos sinais recebidos nos microfones.

$$\min_{W(\omega)} W(\omega)^H \Phi_{XX} W \quad \text{sujeito a} \quad W(\omega)^H A = 1 \tag{29}$$

onde $W(\omega)$ é um conjunto de M vetores de coeficientes que multiplicam os sinais de entrada para obter a saída Y e A é o vetor de direção com a função de transferência para cada microfone na direção desejada. Este problema de otimização tem a seguinte solução (BENESTY; CHEN; HUANG, 2008):

$$W = \frac{\Phi_{XX}^{-1} A}{A^H \Phi_{XX}^{-1} A} \tag{30}$$

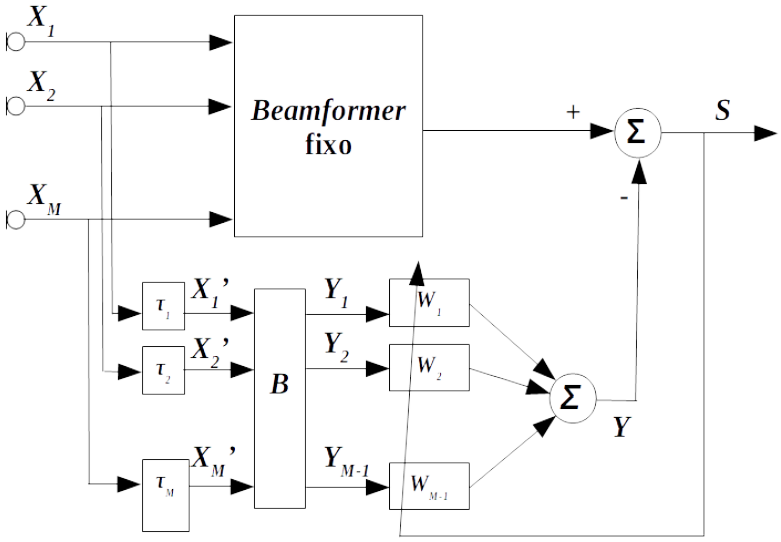
Na prática a matriz de correlação é obtida utilizando apenas os sinais interferentes ou ruído, ou seja, a matriz é calculada utilizando amostras do sinal de entrada quando o sinal de interessa não está ativo. Numa situação ideal a solução é a mesma utilizando somente o ruído ou o sinal completo porém se torna muito sensível a erros no vetor de direção e na determinação da matriz, que é estimada com uma amostragem finita dos sinais de entrada (EHRENBERG *et al.*, 2010).

3.5.3 Generalized Side-lobe Canceller

O *beamformer* conhecido por *Generalized Side-lobe Canceller* (GSC) (GRIFFITHS; JIM, 1982) é um *beamformer* adaptativo que usa uma versão do sinal interferente para o cancelamento. Este algoritmo é composto de 3 partes: um *beamformer* fixo, um modulo de bloqueio e um filtro adaptativo. O diagrama esquemático é apresentado na Figura 5

O *beamformer* fixo pode ser feito com algum dos métodos já apresentados, por exemplo, *Delay-and-Sum*. O módulo de matriz de bloqueio gera um sinal de referência que

Figura 5 – Diagrama esquemático do GSC



Fonte: do autor

não inclui o sinal na direção de interesse. Pode ser implementado utilizando o princípio do *beamformer delay-and-sum* aplicando os atrasos nos sinais de entrada e com os sinais atrasados utilizar a matriz de bloqueio

$$B = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \tag{31}$$

$$Y(k) = BX'(k) \tag{32}$$

onde $X'(k)$ são os sinais dos microfones devidamente atrasados por $\tau_1, \tau_2, ...\tau_M$ que são os atrasos relativos do sinal de interesse e os microfones.

O sinal $Y(k)$ passa pelo filtro adaptativo de modo a cancelar a interferência da saída do *beamformer*.

$$Y'(k) = W(k)Y(k) \tag{33}$$

O ajuste dos coeficientes $W(k)$ é baseado no algoritmo de mínimo erro quadrático médio (*Least Mean Square* – LMS), que recalcula os coeficientes do filtro de forma a minimizar o sinal de erro.

$$W(k+1) = W(k) + 2\mu Y'(k)S(k) \tag{34}$$

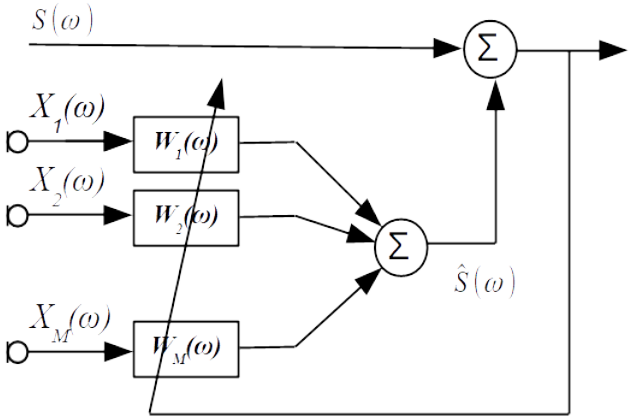
onde μ é coeficiente de convergência e $S(k)$ é o sinal de saída do *beamformer* GSC. O coeficiente μ geralmente é normalizado pela potência do sinal. O algoritmo LMS pode ser aplicado de maneira eficiente e estável processando os sinais no domínio frequência

(SHYNK, 1992).

3.5.4 *Multichanel* Wiener *filter*

A abordagem do filtro de Wiener multicanal (GANNOT; COHEN, 2008; BRANDSTEIN; WARD, 2001) baseia-se em determinar o vetor de coeficientes ótimo $W_m(\omega)$ no sentido do Mínimo Erro Quadrático Médio (MMSE - Minimum Mean Square Erros) entre o sinal desejado $S(\omega)$ e a saída do filtro $\hat{S}(\omega)$, como mostrado na Figura 6

Figura 6 – *Multichannel* Wiener *Filter*.



Fonte: do autor

$$W = \underset{W}{\operatorname{argmin}}\, E[|S_m - Y|^2] = \underset{W}{\operatorname{argmin}}\, E[|S_m - W^H X|^2] \tag{35}$$

A solução é (HAYKIN; LIU, 2009)

$$W = (\phi_{SS} A A^H + \Phi_{VV})^{-1} \phi_{SS} A \tag{36}$$

onde $\phi_{ss} = E[|S|^2]$ é a PSD do sinal desejado e $\Phi_{vv} = E[VV^H]$ é a matriz de correlação do ruído.

O MWF pode ser implementado como um *beamformer* MVDR seguido de um filtro de Wiener ficando:

$$W = \left(\frac{\Phi_{XX}^{-1} A}{A^H \Phi_{XX}^{-1} A}\right)\left(\frac{\phi_{SS}}{\phi_{SS} + \phi_{VV}}\right) \tag{37}$$

O MWF pode ser generalizado levando em consideração o compromisso entre redução de ruído e distorção do sinal desejado resultando no *Speech Distortion Weighted MWF* (DOCLO *et al.*, 2007)

$$W = \left(\frac{\Phi_{XX}^{-1} A}{A^H \Phi_{XX}^{-1} A}\right)\left(\frac{\phi_{SS}}{\phi_{SS} + \mu\phi_{VV}}\right) \tag{38}$$

Quanto menor for o fator μ escolhido, menor será a distorção da fala resultante. E se

$\mu = 1$ o MWF (37) é obtido. Se $\mu > 1$, o nível de ruído residual será reduzido com o aumento da distorção da fala.

3.6 Métricas de avaliação

No campo de separação de sinais um aspecto muito importante é a avaliação do resultado da separação e da qualidade dos sinais de áudio ou voz. A qualidade da separação pode ser medida objetivamente comparando o sinal separado com o sinal de referência ou de forma subjetiva ouvindo o sinal obtido. Entre várias métricas objetivas amplamente usadas para avaliação do desempenho (HU; LOIZOU, 2008),(GANNOT *et al.*, 2017), neste trabalho foram utilizadas as seguintes métricas: *Signal-to-Interference Ratio* (SIR) e *Signal-to-Distortion Ratio* (SDR) (VINCENT; GRIBONVAL; FEVOTTE, 2006), que medem a redução dos sinais indesejados e o sinal de interesse, PESQ (RIX *et al.*, 2001), que mede a percepção da qualidade geral da fala em banda larga e STOI (TAAL *et al.*, 2011), que estima a inteligibilidade do sinal de fala.

As medidas *Signal-to-Interference Ratio* (SIR) e *Signal-to-Distortion Ratio* (SDR) são definidas a seguir

$$SDR = 10log_{10}\frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2} \tag{39}$$

$$SIR = 10log_{10}\frac{\|s_{target}\|^2}{\|e_{interf}\|^2} \tag{40}$$

onde s_{target} é uma versão do sinal de origem modificada, e e_{interf}, e_{noise} e e_{artif} são, respectivamente, interferências, ruído e artefatos. Elas foram calculadas utilizando a ferramenta para MATLAB chamada BSS EVAL que está disponível *online* (FÉVOTTE; GRIBONVAL; VINCENT, 2007)

A medida PESQ (Perceptual Evaluation of Speech Quality): segue a recomendação da International Telecommunication Union (ITU) (ITU-T P.862.2) e estima a avaliação subjetiva no contexto da fala em banda larga, ou seja, sinais com uma largura de banda de áudio que se estende de 50 a 7000 Hz. A pontuação da PESQ é mapeada para uma escala com valor único no intervalo de –0,5 a 4,5.

A STOI é uma medida de inteligibilidade objetiva que tem uma forte relação monotônica com os escores de inteligibilidade de vários testes de escuta . Os valores de STOI variam de 0 a 1. A métrica STOI também demonstrou alta correlação com o rWERR (Relative Word Error Rate Reduction) e, portanto, pode ser útil na avaliação do desempenho de novos algoritmos de um sistema reconhecimento automático da fala (MOORE; PARADA; NAYLOR, 2017).

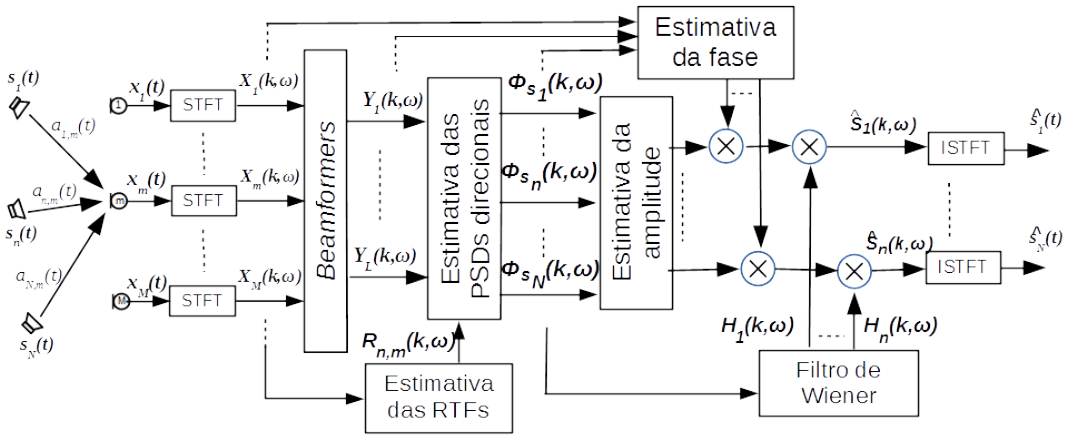
4 SEPARAÇÃO DE SINAIS PELA DENSIDADE DE POTÊNCIA ESPECTRAL DA SAÍDA DE *BEAMFORMERS*

Os métodos desenvolvidos neste trabalho para separação de fontes sonoras se baseia em estimar a densidade de potência espectral (PSD) nas direções das fontes sonoras e, a partir dessa estimativa, obter o sinal de interesse. A determinação das PSDs é feita utilizando a propriedade dos *beamformers* apresentarem ganhos diferentes para diferentes direções no espaço. Esta solução pressupõe o conhecimento das funções de transferência acústica (ATFs) entre as fontes sonoras e os microfones ou, para uma solução aproximada, conhecer as direções das fontes sonoras, as quais podem ser estimadas por técnicas de estimativa da direção de chegada existentes (BRANDSTEIN; WARD, 2001; DIBIASE, 2000). Considerando conhecidas as direções das fontes sonoras e o arranjo de microfones, os métodos propostos podem separar fontes de som e com número de fontes superior ao número de microfones (caso subdeterminado). O uso da diretividade dos *beamformers* foi inicialmente proposto por Hioka (HIOKA *et al.*, 2013) resultando na separação de um número de fontes inferior a $M(M-1)+1$ sendo M o número de microfones. A solução utiliza as PSDs obtidas para determinar filtros que aplicados aos sinais captados rejeitam as fontes interferentes.

Como já mencionado, as soluções atuais estimando as PSD a partir da direcionalidade dos *beamformers* apresentam problemas como soluções incoerentes com a interpretação física (PSDs Negativas) e limitações no número de fontes separáveis. Este trabalho aborda o problema visando superar essas limitações com a proposta de três novos métodos para separação se fontes baseadas nas estimativas das PSDs de *beamformers*. O diagrama em bloco do processo de separação onde os métodos propostos estão inseridos é apresentado na Figura 7. A primeira contribuição propõe obter as PSDs por técnicas de minimização de mínimos quadrados com restrição não negativa (NNLS - *Nonnegative Least Squares*). O segundo método proposto relaxa a suposição de que as fontes sonoras são não correlacionadas; a estimativa das PSDs direcionais é formulada considerando explicitamente a correlação de fontes através das PSDs cruzadas dos *beamformers*. A terceira proposta apresentada é baseada nas esparsidade dos sinais sonoros e assume que o número de fontes dominantes simultâneas não excede o número de microfones e formula a estimativa

de PSDs apenas para as fontes dominantes, tratando um problema subdeterminado como um problema determinado.

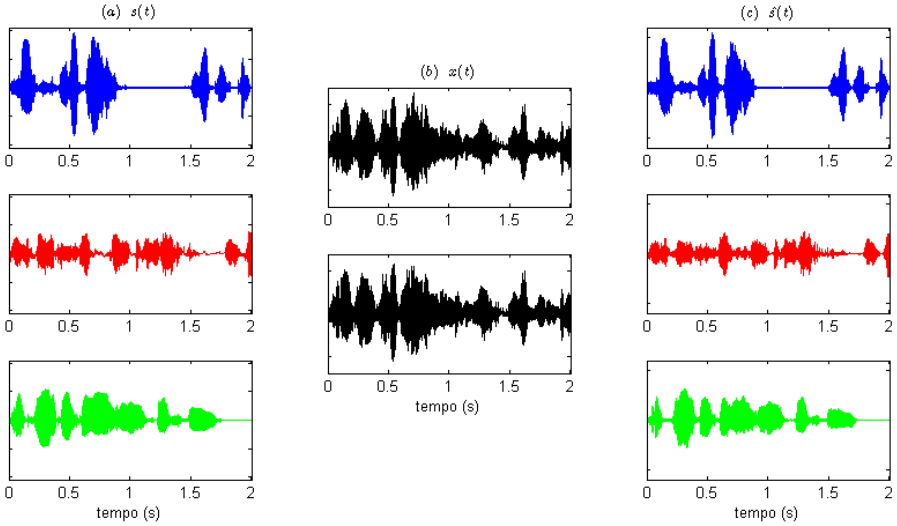
Figura 7 – Diagrama em blocos dos métodos de separação propostos.



Fonte: do autor

Para visualizar as diversas etapas do processo de separação a Figura 8 apresenta um exemplo de separação envolvendo três fontes e dois microfones simulando um ambiente anecoico. Na coluna (a) são mostrados os sinais de origem que serão misturados pelas ATFs, gerando os sinais nos microfones mostrados na coluna (b). Após processados pelos métodos de separação os sinais estimados na saída do processo são apresentados na coluna (c). Este exemplo foi gerado utilizando o método DPNN apresentado na seção 4.2.

Figura 8 – Sinais no processo de separação. (a) Sinais das fontes sonoras, (b) sinais nos microfones e (c) sinais das fontes sonoras estimados

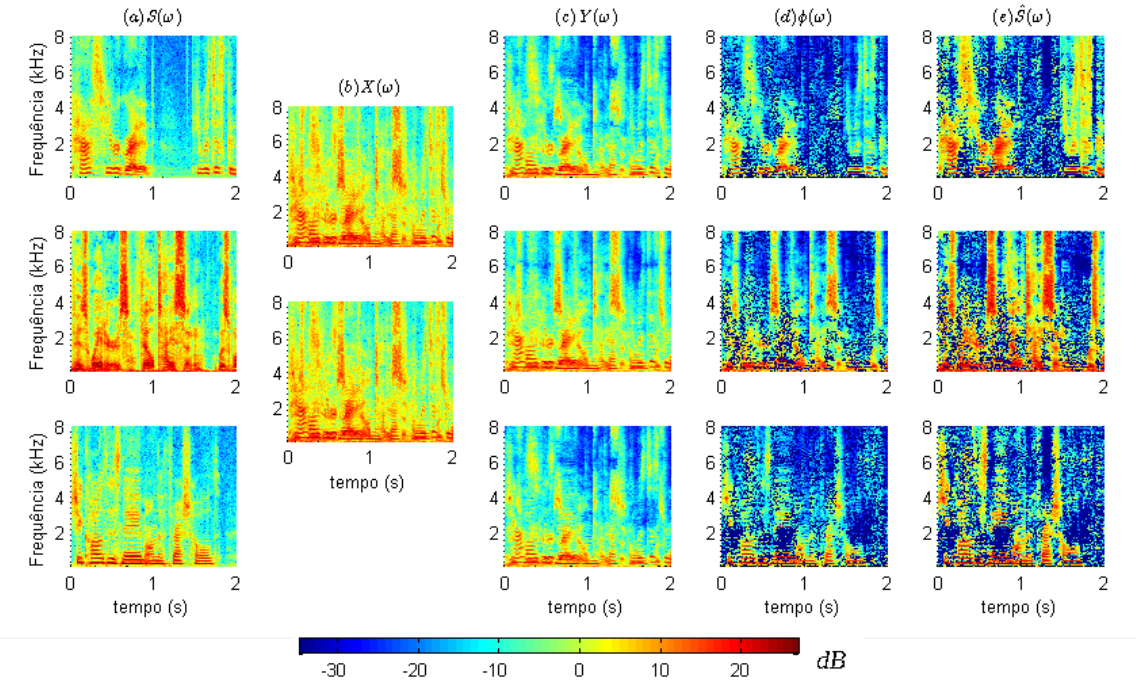


Fonte: do autor

Na Figura 9 são apresentados os sinais na forma de espectrograma onde os sinais estão no domínio da STFT. Nesta figura alem dos sinais apresentados na Figura 8: (a) sinal das

fontes, (b) sinal nos microfones e (e) sinal estimado, também são apresentados em (c) os sinais de saída dos *beamformers* e em (d) as PSDs estimadas para cada fonte.

Figura 9 – sinais observados no processo de separação.(a) Sinais das fontes sonoras, (b) sinais nos microfones, (c) sinais nas saídas dos *beamformers*, (d) PSDs estimadas e (e) sinais das fontes sonoras estimados



Fonte: do autor

4.1 Estimativa da potência das fontes sonoras baseada na diretividade dos *beamformers*

A densidade de potência espectral de uma fonte sonora pode ser determinada com a resolução de um sistema formado pela combinação de ganhos de diversos *beamformers* focados nas direções das fontes de som (HIOKA *et al.*, 2013).

A partir de um arranjo de M microfones podem-se definir diversos *beamformers* com diferentes direcionalidades. Supondo N o número de fontes de sinal determinam-se L *beamformers* ($L \geq N$) direcionados em L diferentes direções. Desta forma a saída de cada *beamformer* l é dada por

$$Y_l(\omega) = \sum_{m=1}^{M} W_{l,m}(\omega) X_m(\omega) \tag{41}$$

onde $W_{l,m}(\omega)$ são os coeficientes dos filtros de cada microfone m para o *beamformer* l e $X_m(\omega)$ é o sinal captado no microfone m. Considerando a função de transferência

$A_{n,m}(\omega)$ entre cada fonte sonora $S_n(\omega)$ e os microfones, substitui-se (2) em (41)

$$Y_l(\omega) = \sum_{m=1}^{M}\sum_{n=1}^{N} W_{l,m}(\omega)A_{n,m}(\omega)S_n(\omega) + \sum_{m=1}^{M} W_{l,m}(\omega)V_m(\omega) \tag{42}$$

$$Y_l(\omega) = \sum_{n=1}^{N} D_{l,n}(\omega)S_n(\omega) + \sum_{m=1}^{M} W_{l,m}(\omega)V_m(\omega) \tag{43}$$

onde $D_{l,n}(\omega)$ é a função de direção do *beamformer* l para a direção da fonte n e representa o ganho entre cada fonte n na saída do *beamformers* l.

$$D_{l,n}(\omega) = \sum_{m=1}^{M} W_{l,m}(\omega)A_{n,m}(\omega) \tag{44}$$

A PSD da saída do *beamformer* na direção l é dada por

$$\phi_{Yl}(\omega) = E[Y_l(\omega)Y_l(\omega)^*] \tag{45}$$

Assumindo que as fontes e o ruído são mutuamente não correlacionadas, desconsiderando o termo de ruído que não está associado a uma direção específica, a PSD da saída de cada *beamformer* é dada por (HIOKA *et al.*, 2013):

$$\phi_{Y_l}(\omega) = \sum_{n=1}^{N} |D_{l,n}(\omega)|^2 \phi_{S_n}(\omega) \tag{46}$$

onde $\phi_{S_n}(\omega)$ é a PSD da fonte n.

Definindo pelo menos L *beamformers* diferentes é possível com a Equação (46) estimar a PSD de cada fonte $\phi_{S_n}(\omega)$ resolvendo o sistema de L equações abaixo

$$\begin{bmatrix} \phi_{Y_1} \\ \phi_{Y_2} \\ \vdots \\ \phi_{Y_L} \end{bmatrix} = \begin{bmatrix} |D_{1,1}|^2 & |D_{1,2}|^2 & \cdots & |D_{1,N}|^2 \\ |D_{2,1}|^2 & |D_{2,2}|^2 & \cdots & |D_{2,N}|^2 \\ \vdots & \vdots & \ddots & \vdots \\ |D_{L,1}|^2 & |D_{L,2}|^2 & \cdots & |D_{L,N}|^2 \end{bmatrix} \begin{bmatrix} \phi_{S_1} \\ \phi_{S_2} \\ \vdots \\ \phi_{S_N} \end{bmatrix} \tag{47}$$

A variável ω foi omitida mas o sistema acima deve ser resolvido para cada frequência. Na forma matricial a Equação (47) fica

$$\Phi_Y(\omega) = D(\omega)\Phi_S(\omega) \tag{48}$$

Hioka propõe a solução do sistema (48) por

$$\Phi_S(\omega) = D^{-1}(\omega)\Phi_Y(\omega) \tag{49}$$

sendo que $D^{-1}(\omega)$ é a inversa da matriz quadrada $D(\omega)$ ou, no caso de $L > N$, é a matriz pseudo-inversa de Moore-Penrose. Esta solução pode resultar em valores negativos de PSD, neste caso, os valores são substituídos pelos seus absolutos (HIOKA *et al.*, 2013).

Este método apresenta alguns problemas relacionados com o condicionamento da matriz $D(\omega)$. Hioka demonstrou que se $N > M(M-1)+1$ o método falha devido ao mau condicionamento da matriz tornando o sistema indefinido o que resulta em estimativas instáveis. Este limite piora com a escolha inadequada da geometria do arranjo de microfones devido a simetrias na distribuição dos microfones. Também existem problemas em baixas frequências, devido à resposta nestas frequências ser muito plana e em altas frequências com alias espacial com ganhos iguais em diferentes frequências.

Nas próximas seções serão apresentados os novos métodos propostos para superar essas limitações.

4.2 Estimativa não negativa por mínimos quadrados das PSDs

Nesta seção é apresentado um novo método, chamado DPNN, para a separação da fonte sonora, que se baseia na técnica de mínimos quadrados com restrição não negativa aplicada à estimativa de PSD com diretividade dos *beamformers*. Nesta proposta a solução do problema continua sendo formulado considerando a não correlação entre os sinais, porém, como esta é uma aproximação, busca-se uma solução, com o menor erro no sentido de mínimos quadrados, mas restringindo a solução a valores que tenham sentido numa implementação real. Este método foi apresentado em (LUFT; PEREIRA; SUSIN, 2017) (Apêndice A) aplicado à separação de sons binaurais. O novo método, baseado na técnica de mínimos quadrados não negativos, aplicada à estimativa de PSD supera os problemas apresentados na proposta de Hioka (HIOKA *et al.*, 2013). Nesta abordagem, o problema formulado em (48) é resolvido minimizando com o critério de mínimos quadrados e com restrição de solução não negativa.

$$\min_{\Phi_S} \|D\Phi_S - \Phi_Y\|^2 \quad \text{subject to } \Phi_S \geq 0 \tag{50}$$

onde $\|.\|$ é a norma L_2. Este problema pode ser resolvido usando o método proposto em (LAWSON; HANSON, 1974). Alternativas para este método podem ser encontradas em (CHEN; PLEMMONS, 2010).

Como o problema pode ser subdeterminado ($N > M$) ou mal condicionado podendo-se obter diferentes soluções ou causar instabilidades numéricas. Aplicando restrição não negativa e resolvendo por minimização de mínimos quadrados evitam-se soluções não realistas (PSD negativa) o que não é contemplado pelo proposto por Hioka, que às vezes resultam em estimativas negativas de PSD.

4.3 Estimativa das PSDs direcionais com fontes correlacionadas

Nesta seção, são apresentados novos método para estimar a PSD de fontes que consideram a possibilidade de uma correlação entre as fontes denominados DPC e DPCM. As STFTs de dois sinais da fala são ditas aproximadamente W-DO (W-disjoint orthogonals) (RICKARD; YILMAZ, 2002) (YILMAZ; RICKARD, 2004). Isso significa que a sua representação de tempo e frequência se sobrepõe e a condição de perfeita não correlação entre os sinais não pode ser assumida, diferente do proposto por Hioka (HIOKA *et al.*, 2013). Os sinais de áudio são, em grande parte, sinais não estacionários, por esta razão as PSDs observadas são calculadas durante um período de tempo que deve ser curto o suficiente para considerar a fala como um sinal estacionário. Apesar das características esparsas dos sinais de voz, é plausível que em curtos períodos de tempo, com ocorrências simultâneas, possa haver certa correlação entre os sinais, ou seja, aproximadamente W-DO. Além de assumir a correlação, no método proposto pode-se estimar a amplitude e a fase dos sinais antes da filtragem no final do processo.

Assumindo essa hipótese, a solução em (46) não está correta, pois negligencia as PSDs cruzadas entre as fontes. Pode-se generalizar a solução usando as PSDs cruzadas das saídas dos *beamformers*

$$\phi_{Y_{l_1,l_2}}(\omega) = E[Y_{l_1}(\omega)Y^*_{l_2}(\omega)] \tag{51}$$

onde $Y_{l_1}(\omega)$ e $Y_{l_2}(\omega)$ são as saídas de dois *beamformers*. Expandindo (51) com a Equação (43) e desprezando o termo de ruído que não esta relacionado a nenhuma direção específica

$$\phi_{Y_{l_1 l_2}}(\omega) = E\left[\sum_{n=1}^{N} D_{l_1,n}(\omega)S_n(\omega)\sum_{n'=1}^{N} D^*_{l_2,n'}(\omega)S^*_{n'}(\omega)\right] \tag{52}$$

então

$$\phi_{Y_{l_1 l_2}}(\omega) = \sum_{n=1}^{N}\sum_{n'=1}^{N} D_{l_1,n}(\omega)D^*_{l_2,n'}(\omega)E\left[S_n(\omega)S^*_{n'}(\omega)\right]. \tag{53}$$

Utilizando $L \geq N$ *beamformers* diferentes, onde N é o número de fontes, tem-se um conjunto de L^2 equações que devem ser resolvidas para cada frequência ω. Apresentando a formulação em forma de vetores e matrizes

$$\mathbf{s}(\omega) = [S_1(\omega), S_2(\omega), ..., S_N(\omega)]^T \tag{54}$$

e

$$\mathbf{y}(\omega) = [Y_1(\omega), Y_2(\omega), ..., Y_L(\omega)]^T \tag{55}$$

onde ^{T} é a transposta, $S_n(\omega)$ é a fonte de sinal n e $Y_l(\omega)$ é a saída do *beamformer* l.

$$\Phi_S(\omega) = E[\mathbf{s}(\omega) \otimes \mathbf{s}^*(\omega)] \tag{56}$$

e

$$\Phi_Y(\omega) = E[\mathbf{y}(\omega) \otimes \mathbf{y}^*(\omega)] \tag{57}$$

onde * é o conjugado e $\otimes$ é o produto de Kronecker e $E[]$ é o valor esperado, que, devido a não estacionariedade dos sinas de áudio pode ser calculado por (11), $\Phi_S(\omega)$ é uma matriz N^2 x 1 com as PSD e as PSDs cruzadas das fontes $S_n(\omega)$ e $\Phi_Y(\omega)$ é uma matriz L^2 x 1 com as PSDs e as PSDs cruzadas das saídas $Y_l(\omega)$ dos *beamformers*,

$$\mathbf{d}_l(\omega) = [D_{l,1}(\omega), D_{l,2}(\omega), ..., D_{l,N}(\omega)] \tag{58}$$

onde $D_{l,n}(\omega)$ é o ganho entre a fonte n e a saída do *beamformer* l. Com esses vetores obtêm-se

$$\mathbf{D}(\omega) = [\mathbf{d}_1(\omega), \mathbf{d}_2(\omega), ..., \mathbf{d}_L(\omega)]^T \tag{59}$$

A matriz L^2xN^2 envolvendo os ganhos cruzados dos *beamformers* é

$$\mathbf{G}(\omega) = \mathbf{D}(\omega) \otimes \mathbf{D}^*(\omega) \tag{60}$$

onde * é o conjugado, $\otimes$ é o produto de Kronecker. Finalmente, para estimar a PSD de cada fonte, precisa-se resolver

$$\Phi_Y(\omega) = \mathbf{G}(\omega)\Phi_S(\omega) \tag{61}$$

Omitindo ω e exemplificando para o caso de $L = 2$ and $N = 2$, a Equação (61) fica

$$\begin{bmatrix} \phi_{Y_1} \\ \phi_{Y_{1,2}} \\ \phi_{Y_{2,1}} \\ \phi_{Y_2} \end{bmatrix} = \begin{bmatrix} |D_{1,1}|^2 & D_{1,1}D_{1,2}^* & D_{2,1}D_{1,1}^* & |D_{1,2}|^2 \\ D_{1,1}D_{2,1}^* & D_{1,1}D_{2,2}^* & D_{1,2}D_{2,1}^* & D_{1,2}D_{2,2}^* \\ D_{2,1}D_{1,1}^* & D_{2,1}D_{1,2}^* & D_{2,2}D_{1,1}^* & D_{2,2}D_{1,2}^* \\ |D_{2,1}|^2 & D_{2,1}D_{2,2}^* & D_{2,1}D_{2,1}^* & |D_{2,2}|^2 \end{bmatrix} \begin{bmatrix} \phi_{S_1} \\ \phi_{S_{1,2}} \\ \phi_{S_{2,1}} \\ \phi_{S_2} \end{bmatrix} \tag{62}$$

.

Esta solução permite estimar a amplitude das fontes sonoras, a partir das PSDS de casa sinal, e também a fase relativa das fontes sonoras, através das correlações cruzadas entre cada par de fontes $\phi_{S_{n,n'}}(\omega)$ cujos argumentos são as diferenças de fase entre cada par.

O número de microfones não entra explicitamente no método DPC, e pode-se aplicar a formulação do DPC a qualquer número de fontes, mas a separação quando o número de fontes é maior que o número de microfones, caso subdeterminado, depende de vários fatores (HIOKA *et al.*, 2013). As situações em que elas restringem o número de fontes são a geometria da matriz de microfones, a posição das fontes e a escolha apropriada dos *beamformers* cujas escolhas dependem da posição das fontes. A escolha adequada dessas configurações continua sendo um problema em aberto.

Tratar o problema subdeterminado explorando a esparsidade das fontes de som é um procedimento padrão em várias abordagens de separação (YILMAZ; RICKARD, 2004; WINTER *et al.*, 2004; ARAKI; SAWADA; MAKINO, 2007; JAFARI *et al.*, 2013). A suposição de esparsidade significa que apenas alguns coeficientes das fontes são significativamente diferentes de zero. A partir dessa suposição, a proposta consiste em aplicar o DPC às M fontes dominantes, onde M é o número de microfones, chamando este método de DPCM. Primeiro, aplica-se o DPC ao número total de fontes N e selecionam-se as fontes M com maior PSD. Supondo que apenas as fontes M estejam ativas, reaplica-se o DPC para essas fontes e forçam-se as outras como zero.

4.4 Estimativa das PSDs direcionais com a Função de Transferência Relativa

Esta seção apresenta a estimativa de PSD direcional usando as Funções de Transferência Relativa. As RTFs são versões normalizadas das ATFs, ou seja, a ATF para um microfone de referência dividida por uma combinação linear das ATFs com outros microfones (DELEFORGE; GANNOT; KELLERMANN, 2015). Diferentemente das ATFs, cujas medições se baseiam em sinais transmitidos e recebidos sincronizados, que muitas vezes não estão disponíveis em situações práticas, as RTFs podem ser estimadas diretamente dos sinais observados nos microfones sem o conhecimento prévio da localização das fontes ou das características do ambiente.

As técnicas descritas nas seções 4.1 a 4.3 usam as ATFs entre as fontes e os microfones ou uma aproximação, como em (12), associada à posição espacial das fontes. A aplicabilidade das RTFs, substituindo as ATFs, será demostrada para esses métodos e como isso afeta o resultado final. A definição da RTF, assumindo microfone 1 como referência, é definida como(GANNOT; BURSHTEIN; WEINSTEIN, 2001)

$$R_{n,m}(\omega) = \frac{A_{n,m}(\omega)}{A_{n,1}(\omega)} \tag{63}$$

onde $A_{n,m}$ é a Função de Transferência Acústica entre a fonte n e o microfone m, ou em notação vetorial

$$\mathbf{r}_n(\omega) = \left\{1, \frac{A_{n,2}(\omega)}{A_{n,1}(\omega)}, ..., \frac{A_{n,M}(\omega)}{A_{n,1}(\omega)}\right\} \tag{64}$$

Se, em (45), as ATFs são substituídas pelas RTFs, então o ganho direcional relativo do *beamformer* l na direção da fonte n fica

$$D'_{l,n}(\omega) = \sum_{m=1}^{M} W_{l,m}(\omega) R_{n,m}(\omega) \tag{65}$$

onde $W_{l,m}(\omega)$ é o coeficiente do *beamformer* l para o microfone m. De (63) e (65) obtém-

se

$$D'_{l,n}(\omega) = \sum_{m=1}^{M} W_{l,m}(\omega) \frac{A_{n,m}(\omega)}{A_{n,1}(\omega)} \tag{66}$$

então

$$D_{l,n}(\omega) = A_{n,1}(\omega) D'_{l,n}(\omega). \tag{67}$$

Substituindo (67) em (46) resulta em

$$\phi_{Y_l}(\omega) = \sum_{n=1}^{N} |D'_{l,n}(\omega)|^2 |A_{n,1}(\omega)|^2 \phi_{S_n}(\omega). \tag{68}$$

Definindo

$$\phi_{S'_l}(\omega) = |A_{n,1}(\omega)|^2 \phi_{S_n}(\omega) \tag{69}$$

a solução do problema apresentado em (48) pode ser realizada resolvendo

$$\Phi_Y(\omega) = \mathbf{D}'(\omega) \Phi_{S'}(\omega). \tag{70}$$

Como resultado, os valores estimados das PSDs $\Phi_{S'}(\omega)$ das fontes são relativas ao microfone de referência ($m = 1$). As fontes estimadas $s'_n(\omega)$ são versões filtradas das fontes reais $s_n(\omega)$ ou, em outras palavras, elas são as fontes reais filtradas pelas ATFs referentes ao microfone de referência ou, ainda, são as fontes amostradas pelo microfone de referência.

Aplicando a mesma demonstração ao método com correlação das fontes apresentada na seção 4.3 o resultado é

$$\Phi_Y(\omega) = \mathbf{G}'(\omega) \Phi_{S'}(\omega) \tag{71}$$

onde $\mathbf{G}'(\omega)$ é construído usando $D'_{l,n}(\omega)$ obtido em (65). As fontes estimadas $s'_n(\omega)$ também serão relativas ao microfone de referência ($m = 1$).

4.5 Estimativa dos sinais a partir das PSDs

Nesta seção é colocado o procedimento para recuperação dos sinais a partir das PSDs obtidas pelos métodos apresentados. Com as PSDs dos sinais é calculado um pós-filtro de Wiener para cada sinal de interesse. Aplicado o filtro à saída de um *beamformer* obtemos o sinal estimado (HIOKA *et al.*, 2013). O pós-filtro de Wiener mascara aos sinais interferentes presentes no sinal de saída do *beamformer*. O pós-filtro de Wiener é calculado conforme:

$$H_n(\omega) = \frac{\phi_{S_n}(\omega)}{\sum_{n=1}^{N} \phi_{S_n}(\omega)} \tag{72}$$

A estimativa do sinal de saída é então obtido por

$$\hat{S}_n(\omega) = H_n(\omega)Y_n(\omega) \tag{73}$$

onde $Y_n(\omega)$ é a saída do *beamformer* focado na direção da fonte n.

Como alternativa para melhorar a estimativa do sinal é proposto aplicar o filtro de Wiener em uma versão aproximada do sinal, que, utilizando o método DPNN, é obtida usando a raiz quadrada das PSDs, calculadas em (50), como magnitude do sinal em conjunto com a fase do sinal na saída do *beamformer* (LUFT; PEREIRA; SUSIN, 2017).

$$Ph_n(\omega) = \arg[Y_n(\omega)] \tag{74}$$

em que Ph_n replica a fase do *beamformer* $Y_n(\omega)$ (13).

No caso das estimativa pelo métodos DPC ou DPCM, a magnitude é a raiz quadrada das PSDs, calculadas em (61), e a fase Ph_n é obtida a partir das fases das PSDs cruzadas entre cada par de sinais $\arg[\phi_{S_{n,n'}}(\omega)]$, também obtidas em (61). Com as informações de amplitudes e fases relativas dos sinais, uma vez que os valores combinados nos microfones são conhecidos, pode-se estimar os valores individuais de cada sinal.

O valor aproximado do sinal combina a raiz quadrada da PSD estimada com a fase calculada.

$$\hat{S}'_n(\omega) = \sqrt{\phi_{S_n}(\omega)}e^{-iPh_n(\omega)} \tag{75}$$

onde $\phi_{S_n}(\omega)$ é a estimativa da PSD do sinal de interesse. Este sinal $\hat{S}'_n(\omega)$ é filtrado pelo pós-filtro de Wiener (72) obtendo a estimativa final.

$$\hat{S}_n(\omega) = H_n(\omega)\hat{S}'_n(\omega) \tag{76}$$

A Transformada inversa de Fourier é então aplicada a $\hat{S}_n(\omega)$.

5 ESTIMATIVA DA FUNÇÃO DE TRANSFERÊNCIA RELATIVA PARA MÚLTIPLAS FONTES

Neste capítulo são propostas metodologias para estimar as RTFs de múltiplas fontes com solução também aplicável na condição subdeterminada avaliando-as em termos de erros de estimativa e separação de fontes. Conforme mostrado na seção anterior, os métodos de separação propostos neste trabalho podem usar as RTFs para separar as fontes sonoras.

A Função de Transferência Relativa (RTF) apresenta a relação entre as funções de transferência acústica de dois sensores para uma determinada fonte de sinal. As RTFs são definidas como uma versão normalizada das ATFs, ou seja, a ATF relacionada a um determinado microfone é dividida por uma combinação linear das ATFs de outros microfones (DELEFORGE; GANNOT; KELLERMANN, 2015). A RTF descreve o acoplamento entre os microfones como uma resposta a uma determinada fonte. Diferentemente das ATFs, cujas medições se baseiam em sinais transmitidos e recebidos sincronizados, que muitas vezes não estão disponíveis em situações práticas, as RTFs podem ser estimadas diretamente dos sinais observados nos microfones sem conhecimento prévio da localização das fontes ou das características do ambiente.

As metodologias de estimativa apresentadas assumem que em determinada combinação de tempo e frequência uma fonte tem uma potência considerável e a contribuição de todas as outras fontes nesse ponto de tempo-frequência é muito menor. Essa observação é a base para os algoritmos de estimação e a propriedade esparsa dos sinais sonoros torna essa abordagem atraente (ARAKI; SAWADA; MAKINO, 2007). A suposição de esparsidade implica que apenas algumas características dos sinais das fontes são significativamente diferentes de zero. Os sinais das fontes podem não ser esparsos em seu domínio original, como é o caso dos sinais de áudio no domínio do tempo, porém, eles podem ser esparsos em um domínio transformado, como ocorre com os sinais de áudio no domínio de Fourier (YILMAZ; RICKARD, 2004).

Existem várias técnicas para encontrar a fonte dominante com base no maior valor de ocorrência: histogramas (YILMAZ; RICKARD, 2004), *k-means clustering* (ARAKI *et al.*, 2007), *Singular Value Decomposition* (BAO *et al.*, 2013), *Hierarchical Clustering*

(WINTER *et al.*, 2004), *Fuzzy c-Means* (JAFARI *et al.*, 2011, 2013; ATCHESON *et al.*, 2014). Eles geralmente são mais focados em descobrir as fontes dominantes do que explicitamente em obter as funções de transferência do processo de mixagem.

A relação entre dois microfones utilizando os sinais em (2) e negligenciando o ruído fica

$$\frac{X_m(k,\omega)}{X_1(k,\omega)} \approx \frac{\sum_{n=1}^{N} A_{n,m}(\omega) S_n(k,\omega)}{\sum_{n'=1}^{N} A_{n',1}(\omega) S_{n'}(k,\omega)} \tag{77}$$

A partir da Equação acima considerando a dominância de uma das fontes pode-se observar que

$$\frac{X_m(k,\omega)}{X_1(k,\omega)} \approx \frac{A_{n,m}(\omega)}{A_{n,1}(\omega)} \quad \text{se} \quad S_{n'} = 0 \quad \forall \quad n' \neq n \tag{78}$$

Ou seja, de acordo com a Equação (63) sempre que uma única fonte n estiver ativa ou dominante sobre as outras

$$R_{n,m}(k,\omega) \approx \frac{X_m(k,\omega)}{X_1(k,\omega)} \tag{79}$$

onde $R_{n,m}(k,\omega)$ é o valor instantâneo da RTF para a fonte dominante n obtida no *frame* k para a frequência ω.

Nas próximas seções são propostos dois métodos para estimativa das RTFs. O primeiro é baseada em histogramas ponderados e utiliza como pesos as PSDs estimadas das fontes. O segundo utiliza a técnica de aglomeração (*clustering*) baseada em *weighted fuzzy c-means* e assume uma distribuição laplaciana dos sinais.

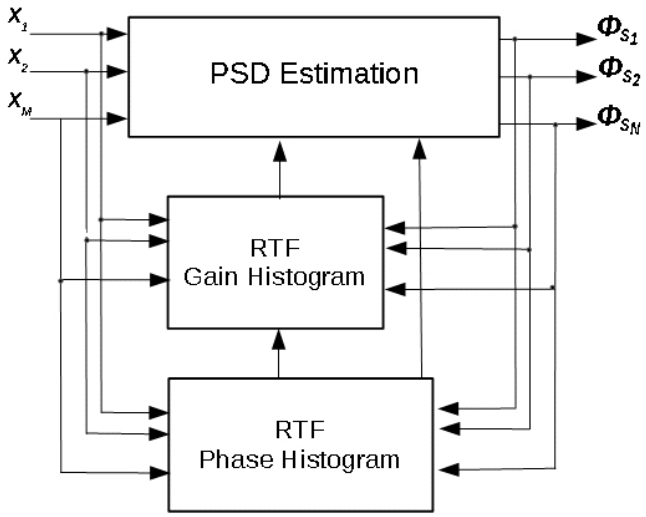
5.1 Estimativa das RTFs com histogramas

A nova metodologia é apresentada na Figura 10 e consiste em usar as estimativas das PSDs das fontes de sinal para determinar pesos em histogramas de estimativa instantânea das RTFs. Encontrando os valores de maior ocorrência na Equação (79) durante certo período de tempo, encontra-se os valores das RTFs das fontes. Este princípio é usado em outros métodos de estimação (YILMAZ; RICKARD, 2004; BAO *et al.*, 2013), que diferem basicamente na maneira de determinar os valores de maior ocorrência, e são mais focados para descobrir a fonte dominante do que explicitamente obter a função de transferência do processo de mistura. O método de estimação apresentado neste trabalho, chamado wFDUET, é baseado em DUET (YILMAZ; RICKARD, 2004) detectando picos em histogramas suavizados. Diferentemente de DUET, constroem-se histogramas suavizados do ganho e da fase para cada frequência a partir dos valores obtidos por (79). Definindo os valores instantâneos de ganho e fase:

$$\tilde{\mu}_{n,m}(k,\omega) := |R_{n,m}(k,\omega)| \tag{80}$$

$$\tilde{\rho}_{n,m}(k,\omega) := arg(R_{n,m}(k,\omega)). \tag{81}$$

Figura 10 – Diagrama em blocos do método wFDUET de estimativa das RTFs.



Fonte: do autor

Diversos fatores como ruídos, número de fontes, localização, intensidade e esparsidade das fontes afetam e podem dificultar essas estimativas causando erros nos valores medidos e na permutação os canais. A fim de minimizar esses erros e melhorar a robustez, é proposto o uso de estimativas PSD das fontes para realimentar o processo. Primeiro estima-se a fase das RTFs construindo um histograma ponderado para cada frequência ω da seguinte forma

$$H_{n,m}(\rho,\omega) = \sum_{k=1}^{K} W_{\rho,n,m}(k,\omega)|X_1(k,\omega)X_m(k,\omega)| \tag{82}$$

onde,

$$W_{\rho,n,m}(k,\omega) = \begin{cases} w_n(k,\omega), & \text{se} \quad |\tilde{\rho}_{n,m}(k,\omega) - \rho| < p \\ 0, & \text{caso contrário,} \end{cases} \tag{83}$$

$w_n(k,\omega)$ é o parâmetro de peso, inicialmente definido como 1 e posteriormente modificado de acordo com as PSDs das fontes, ρ é a fase apresentada no histograma e p é um parâmetro de resolução do histograma. Além do parâmetro de peso proposto neste trabalho o histograma também é ponderado por um termo associado à amplitude dos sinais dos microfones, sugerido em (YILMAZ; RICKARD, 2004). Neste ponto, tem-se um conjunto de histogramas, um para cada frequência, que devem ser suavizado, o que pode ser feito por um filtro de média móvel ou semelhante, tendo o cuidado de compensar o atraso do filtro. Cada pico corresponde a uma fonte e a localização do pico corresponde à fase associada das RTFs. Este procedimento tem um problema de permutação associado, isto é, não é possível saber com qual fonte cada fase está associada. Neste trabalho, é usado um método que utiliza a direção de chegada (DOA), ou seja, o ângulo de direção de cada

fonte de sinal em relação ao arranjo de microfones (SAWADA *et al.*, 2003).

A direção de chegada é calculada considerando a configuração apresentada na Figura 11 e pode ser calculada por (GANNOT; COHEN, 2008)

$$\theta = asin\left(\frac{c\tilde{\rho}_{n,m}(k,\omega)}{4\pi\omega d}\right) \tag{84}$$

onde c é a velocidade do som, d é a distância entre os microfones e o centro do arranjo de microfones. Para frequência abaixo do limite de alias espacial, esta equação é válida, já acima, sinais que chegam de mais de uma direção e têm o mesmo valor de fase. Neste trabalho o valor DOA de baixa frequência é extrapolado para todas as frequências. A permutação é feita de acordo com essa direção, ou seja, cada fase estimada é associada a fonte como valor de θ mais próximo, resultando em uma fase estimada para cada fonte $\hat{\rho}_{n,m}(\omega)$.

O segundo passo é a determinação dos ganhos das RTFs, o que é feito de maneira similar à estimativa de fase usando histogramas suavizados. A estimativa dos ganhos apresenta o mesmo problema de permutação mas, geralmente, não contém informações de direção, já que os microfones estão próximos e as diferenças de ganho devido à atenuação da propagação são muito pequenas. Para isso, um histograma ponderado de ganhos é construído para cada fonte usando as informações de direção já determinadas. Para cada fonte, é construído um histograma

$$H_{n,m}(\mu,\omega) = \sum_{k=1}^{K} W_{\mu,n,m}(k,\omega)|X_1(k,\omega)X_m(k,\omega)| \tag{85}$$

onde,

$$W_{\mu,n,m}(k,\omega) = \begin{cases} \tilde{w}_{n,m}(k,\omega), & \text{se} \quad |\hat{\mu}_{n,m}(k,\omega) - \mu| < q \\ 0, & \text{caso contrário} \end{cases} \tag{86}$$

onde μ é o valor de ganho representado no histograma e q é um parâmetro da resolução do histograma e $w_{n,M}(k,\omega)$ é um peso determinado por

$$\tilde{w}_{n,m}(k,\omega) = \frac{\sum_{n=1}^{N} \frac{1}{|\rho_{n,m}(k,\omega) - \hat{\rho}_{n,m}(\omega)| + \epsilon}}{\frac{1}{\rho_{n,m}(k,\omega) - \hat{\rho}_{n,m}(\omega)| + \epsilon}} \tag{87}$$

onde $\hat{\rho}_{n,m}(\omega)$ é a fase estimada e ϵ é um pequeno valor de regularização para evitar divisão zero. Esta função foi proposta de modo a colocar peso 1 quando a função tem fase igual à fonte de interesse e zero se estiver associada às outras fontes. Após suavizar os histogramas, as posições dos picos correspondem aos valores estimados de ganho $\hat{\mu}_{n,m}(\omega)$. As Funções de Transferência Relativa estimadas para cada fonte são

$$\hat{R}_{n,m}(\omega) = \hat{\mu}_{n,m}(\omega)e^{-i\hat{\rho}_{n,m}(\omega)}. \tag{88}$$

A RTF estimada é usada para calcular os valores em (65) e determinar as PSDs de cada fonte ϕ_{S_n}. Com os valores de PSD das fontes, o processo de estimativa das RTFs é repetido modificando os pesos em (83) que foram inicialmente definidos como 1 usando os valores obtidos em (72)

$$w_n(k,\omega) = H_n(k,\omega) \tag{89}$$

Em vez de construir apenas um histograma para determinar a posição dos N maiores picos, um histograma é construído para cada fonte usando $w_n(k,\omega)$ e são determinados os picos em cada um deles.

5.2 Estimativa das RTFs com wFCM

Nesta seção é proposto um método, chamado wFCM-L, para estimar a RTF de múltiplas fontes com base no agrupamento *weighted Fuzzy C-Means* (wFCM), nesta proposta os pesos são caracterizados pela Distribuição Laplaciana dos sinais de voz. Os algoritmos *fuzzy c-mean* provaram ser eficazes na separação (JAFARI *et al.*, 2013; ATCHESON *et al.*, 2014) e na localização das fontes (KÜHNE; TOGNERI; NORDHOLM, 2009). No método proposto, $R_k(\omega)$ representa a Função de Transferência Relativa instantânea entre dois microfones, o índice de microfone m é omitido para simplificar. Os $R_k(\omega)$ são agrupados com o algoritmo wFCM para encontrar elementos $\mu_{kn}(\omega)$ que variam entre 0 e 1 e especificam o grau de associação $R_k(\omega)$ ao cluster n. No wFCM, o agrupamento é alcançado minimizando a função de custo (MIYAMOTO; INOKUCHI; KURODA, 2006)

$$J_{wFCM}(\omega) = \sum_{k=1}^{K}\sum_{n=1}^{N} w_{kn}(\omega)\mu_{kn}(\omega)^q \Delta_{kn}(\omega) \tag{90}$$

onde $\Delta_{kn}(\omega) = ||R_k(\omega) - C_n(\omega)||^2$ é a distância Euclidiana entre R_k e o centro do *cluster* C_n, N é o número de *clusters*, K é o número de *frames*, $w_{kn}(\omega)$ é o peso associado a R_k ou à medida de distância $\Delta_{kn}(\omega)$, q é um parâmetro de fuzzificação positivo para controlar a sobreposição de clusters. Partindo de um particionamento aleatório para $\mu_{kn}(\omega)$ e ganhos unitários $w_{kn}(\omega) = 1$, esta função pode ser resolvida iterativamente até convergir.

$$C_n(\omega) = \frac{\sum_{k=1}^{K} w_{kn}(\omega)\mu_{kn}^q(\omega) * R_k(\omega)}{\sum_{k=1}^{K} w_{kn}(\omega)\mu_{kn}^q(\omega)} \tag{91}$$

e

$$\mu_{n,k}(\omega) = \frac{1}{\sum_{j=1}^{N}\left(\frac{\Delta_{kn}(\omega)}{\Delta_{kj}(\omega)}\right)^{\frac{2}{q-1}}} \tag{92}$$

os centroides finais correspondem às estimativas do vetor de características da RTF. O critério de término é atendido quando a variação da função de custo $J_{wFCM}(\omega)$ entre iterações atingir um valor de tolerância ϵ.

5.2.1 Determinação do vetor de caracteríscas

Pesquisas anteriores já identificaram que a relação entre as amplitudes e as diferenças de fase das observações são características apropriadas para agrupamento na separação de fontes (ARAKI *et al.*, 2007). Uma vez que as fontes estão localizadas em diferentes localizações espaciais e que as relações de amplitude e as diferenças de fase fornecem informações geométricas em função da localização relativa entre as fontes e os sensores, essas características se adéquam para uma separação efetiva. Definindo os valores instantâneos de ganho e fase

$$R^{\rho}_{n,m}(k,\omega) := |R_{n,m}(k,\omega)| \tag{93}$$

$$R^{\theta}_{n,m}(k,\omega) := arg(R_{n,m}(k,\omega)) \tag{94}$$

O vetor de características usado no algoritmo wFCM é definido como

$$R_k(\omega) = [R^{\rho}_{n,m}(k,\omega), R^{\theta}_{n,m}(k,\omega)] \tag{95}$$

5.2.2 Determinação dos pesos

No FCM convencional, todos os objetos têm uma distribuição uniforme, diferentemente do wFCM, onde os pesos têm a função de aumentar a importância de amostras mais representativas para formar os *clusters*. Sabe-se que os sinais de fala não são estacionários e têm uma distribuição estatística mais complexa. Vários fatores, como ruídos, número de fontes, localização, intensidade, reverberação e dispersão das fontes podem afetar os valores medidos. Estudos anteriores propuseram a ponderação de dados em favor de pontos confiáveis no cálculo dos centroides dos *clusters* (KÜHNE; TOGNERI; NORDHOLM, 2009, 2010; ATCHESON *et al.*, 2014; HOLLICK *et al.*, 2014).

Como já mostrado, a RTF é obtida encontrando o centro do *cluster* onde os valores calculados por (78) tendem a se concentrar em torno dele. O valor instantâneo da RTF obtido em (79) não depende das características do sinal da fonte que é anulado na razão entre os sinais dos microfones, porém, sinais interferentes de outras fontes caso ocorram simultaneamente podem interferir no valor medido. Considerando que os sinais interferentes que provocam a dispersão dessas medidas são essencialmente sinais de fala, é proposto ponderar o processo de agrupamento baseado na distribuição de probabilidade da fala. De acordo com (GAZOR; ZHANG, 2003) o sinal de fala ativo pode ser caracterizado pela Distribuição Laplaciana (LD)

$$f(x) = \frac{1}{2a} e^{-\frac{|x-c|}{a}} \tag{96}$$

onde $|.|$ é o valor absoluto e a é o desvio médio absoluto do valor médio c. Durante o processo de aglomeração os pesos $w_{n,k}(\omega)$ devem ser calculados em cada iteração de acordo com as partições $\mu_{n,k}(\omega)$ dessa forma o parâmetro a é calculado como desvio da

distância média ponderada

$$a_{n,k}(\omega) = \frac{\sum_{k=1}^{K} \mu_{n,k}(\omega)|\Delta_{kn}(\omega)|}{\sum_{k=1}^{K} \mu_{n,k}(\omega)} \tag{97}$$

e os pesos usados na próxima iteração são

$$w_{n,k}(\omega) = e^{-\frac{|\Delta_{kn}(\omega)|}{a_{n,k}(\omega)}} \tag{98}$$

O Algoritmo 1 abaixo resume as etapas para estimativa das RTFs.

Algoritmo 1: Estimando as RTFs com wFCM

Entrada: X, q, ϵ
Saída: C,μ
início
  Inicializar $\mu_{n,k}(\omega)$ com valores aleatórios (0 a 1) $(\sum_n \mu_{n,k}(\omega)) = 1$);
  Inicializar $w_{n,k}(\omega)$ com 1 ;
  Determinar o vetor $R_k(\omega)$ com 95;
  repita
    Calcular centroides $C_n(\omega)$ (91);
    Calcular partições $\mu_{n,k}(\omega)$ (92);
    Calcular pesos $w_{n,k}(\omega)$ (98);
    Calcular funçao de custo $J_{wFCM}(\omega)$ (90);
  até *variação de* $J_{wFCM}(\omega)$ <ϵ;
fim

O algoritmo deve ser repetido para todas as frequências ω. Neste ponto do processo, obtêm-se os N centroides para cada frequência

$$C_n(\omega) = [C_n^{\rho}(\omega), C_n^{\theta}(\omega)] \tag{99}$$

e a RTF estimada para cada *bin* de frequência é

$$\hat{R}_n(\omega) = C_n^{\rho}(\omega) e^{i * C_n^{\theta}(\omega)} \tag{100}$$

5.2.3 Permutação

O procedimento apresentado na seção anterior tem uma permutação associada ao longo das frequências, ou seja, o *cluster* n em alguma frequência não é necessariamente da mesma fonte que o *cluster* n em outra frequência. Neste trabalho é utilizada a estimativa de direção de chegada (DOA), ou seja, a direção de cada fonte em relação ao arranjo

de microfones, para corrigir a permutação (SAWADA *et al.*, 2003). As direções de chegada são calculadas considerando a configuração do arranjo de microfones apresentada na Figura 11 e é calculada por

$$\theta_n(\omega) = asin\left(\frac{C_n^{\theta}(\omega)c}{4\pi\omega d}\right) \tag{101}$$

onde c é a velocidade do som, d é a distância entre os microfones e o centro do arranjo de microfones. Esta equação é válida para a frequência abaixo do limite do *alias* espacial, acima do qual o sinal que chega de mais de uma direção tem o mesmo valor de fase e essas direções devem ser avaliadas. Neste trabalho o valor DOA de baixas frequências é extrapolado para todas as frequências e permutado de acordo com essa direção.

Para determinar a direção das N fontes, os valores de θ_n são agrupados usando o mesmo procedimento descrito na seção 5 com os pesos $w_{n,k}(\omega)$ iguais a um. Os vetores de características são então permutados para a direção mais próxima obtendo as funções de transferência relativa estimadas para cada fonte

$$\hat{R}_n(\omega) = \hat{C}_n^{\rho}(\omega)e^{i*\hat{C}_n^{\theta}(\omega)} \tag{102}$$

onde $\hat{C}_n(\omega)$ são as características permutadas.

6 IMPLEMENTAÇÃO E RESULTADOS

Nesta seção, será demonstrada através de simulações a efetividade dos métodos propostos. Os métodos DPNN, DPC e DPCM são avaliados e comparados com outras soluções através de métricas objetivas de separação, inteligibilidade e qualidade. Também é apresentado o uso de RTFs demonstrando sua aplicabilidade em substituição das ATFs nos métodos propostos. Por fim os métodos propostos para estimativa das RTFs para múltiplas fontes são avaliados e testados na separação de sinas.

6.1 Configurações das implementações

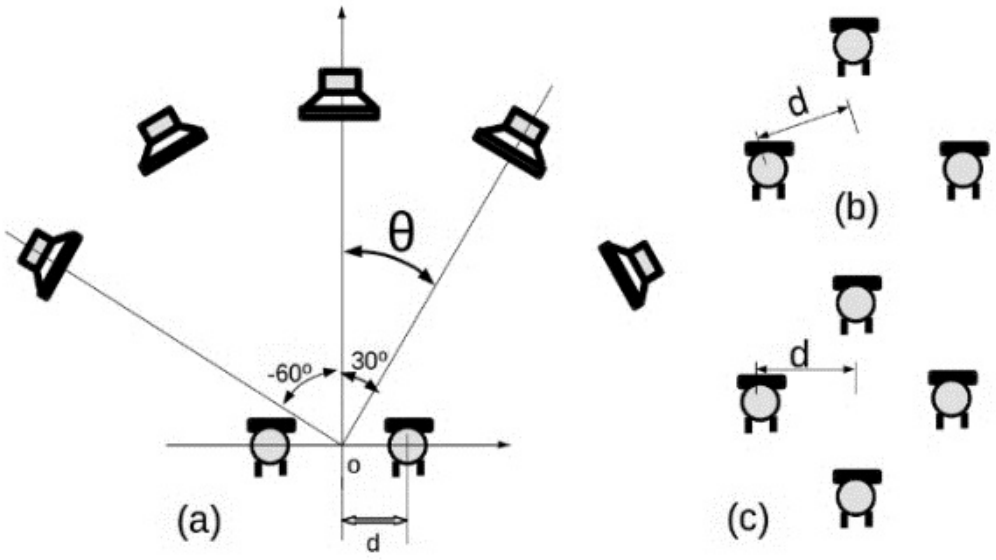
Os métodos propostos foram simulados em MATLAB nas condições apresentadas nesta seção. Todas as simulações foram realizadas considerando microfones omnidirecionais. Foram usados arranjos de microfones com $M = 2$, $M = 3$ e $M = 4$, onde M é o número de microfones, arranjados conforme a Figura 11, com os microfones dispostos de forma circular, com a distância de cada microfone para o centro do arranjo de $d = 2$cm. Os arranjos de microfones foram escolhido com geometria e dimensões semelhantes a outros trabalhos relacionados (WINTER *et al.*, 2004; ATCHESON *et al.*, 2014; HIOKA *et al.*, 2013; FENG; KOWALSKI, 2017) permitindo uma melhor comparação. Os microfones e as fontes sonoras estão todos colocados em um mesmo plano e a direção das fontes especificadas por um ângulo θ conforme a Figura 11. Para cada arranjo de M microfones, o número de fontes de sinal foi variado de 2 a 6. Estas fontes foram dispostas em frente aos microfones formando um semicírculo de raio 1.2 m em torno do centro do arranjo de microfones. A Tabela 2 mostra os ângulos das fontes sonoras para cada quantidade de fontes usadas na simulação. As fontes de som foram selecionadas da base de dados *LibriSpeech ASR corpus* (PANAYOTOV *et al.*, 2015).

Nas simulações os *beamformers* foram projetados usando a técnica *Delay-and-Sum*. Os coeficientes foram calculados de modo que os *beamformers* tivessem ganho máximo na direção das fontes.

$$W_{m,\theta}(\omega) = e^{j\omega d_{m,\theta}/c} \tag{103}$$

onde $d_{m,\theta}$ é a distância do microfone m até a até a reta perpendicular a a direção θ que

Figura 11 – Configuração das fontes sonoras e arranjos de microfones (a)N=5 and M=2, (b)M=3 and (c)M=4.



Fonte: do autor

Tabela 2 – Ângulo das fontes sonoras

Número de fontes	Ângulos θ (graus)
2	30,-30
3	45,0,-45
4	54,18,-18,54
5	60,30,0,-30,-60
6	64.3,36.6,12.8-12.8,-36.6,-64.3

passa pelo microfone mais distante da fonte.

Todas as simulações foram feitas para um período de 10 s de áudio. Os sinais de áudio foram amostrados a uma taxa de 16 kHz. Os métodos foram aplicados para janelas de análise de 512 amostras, o que equivale a 32ms de áudio, com deslocamento da janela de 256 amostras. Foi utilizada a janela de Hanning como janela de análise. O ambiente é uma sala simulada com dimensões de 3,5m x 4,2m x 2,6m e o centro do arranjo de microfones na posição [1,65m, 1,5m, 1,1m]. As simulações são realizadas em três cenários diferentes: um ambiente anecoico ideal, reverberante $T_{60} = 100$ ms e reverberante $T_{60} = 200$ ms. As respostas ao impulso de sala sintética (RIR - *Room Impulse Response*) foram obtidas aplicando a técnica ISM (Image Source Model) (LEHMANN; JOHANSSON, 2008) usando uma implementação online para MATLAB disponível em (LEHMANN, 2012). Os sinais em simulação são gerados usando a resposta completa do impulso da sala, mas no processo de separação, é usada uma versão truncada (512 pontos).

6.2 Métricas de avaliação

As métricas utilizadas para avaliar o desempenho foram as métricas subjetivas descritas na seção 3.6, sendo apresentadas as diferenças entre os valores de entrada e de saída do processo de separação. São apresentadas a melhoria da Relação Sinal-Interferência (SIR) e a melhoria da Relação Sinal-Distorção (SDR) (VINCENT; GRIBONVAL; FEVOTTE, 2006). Os valores foram obtidos utilizando a ferramenta MATLAB denominada BSS EVAL, distribuída online (FÉVOTTE; GRIBONVAL; VINCENT, 2007). Avaliação de inteligibilidade foi obtida pela STOI (TAAL *et al.*, 2011) e calculada pelo código MATLAB disponível *online* em (TAAL, 2011). Para avaliação da qualidade foi utilizada PESQ (RIX *et al.*, 2001) com os códigos disponíveis em (ITU, 2006).

A melhoria SIR, SDR, STOI e PESQ para a enésima fonte é definida como

$$\Delta SIR_n = SIR_{o,n} - SIR_{i,n} \tag{104}$$

$$\Delta SDR_n = SDR_{o,n} - SDR_{i,n} \tag{105}$$

$$\Delta STOI_n = STOI_{o,n} - STOI_{i,n} \tag{106}$$

$$\Delta PESQ_n = PESQ_{o,n} - PESQ_{i,n} \tag{107}$$

onde $SIR_{i,n}$ $SDR_{i,n}$, $STOI_{i,n}$ e $PESQ_{i,n}$ são os valores obtidas com os sinais dos microfones e $SIR_{o,n}$, $SDR_{o,n}$, $STOI_{o,n}$ e $PESQ_{o,n}$ não obtidos na saída do sistema. As avaliações utilizaram a banda de frequência na faixa de 200 Hz a 7000 Hz e cada valor apresentado na nos gráficos representa o valor médio de ΔSIR, ΔSDR $\Delta STOI$ e $\Delta PESQ$ das N fontes sonoras usadas nas simulações.

Para avaliar a estimativa de RTF, usou-se o ângulo Hermitiano entre um vetor RTF verdadeiro $r_n(\omega)$ e o vetor RTF estimado $\hat{r}_n(\omega)$, computado como

$$\Theta_n(\omega) = arccos \frac{|r_n^H(\omega)\hat{r}_n(\omega)|}{||r_n(\omega)||^2||\hat{r}_n(\omega)||^2} \tag{108}$$

onde $(.)^H$ é o transposta do conjugado, $|.|$ é o valor absoluto e $\|.\|^2$ é a norma Euclidiana. Esta medida de erro foi adotada recentemente para avaliar o desvio da estimativa da RTF (TASESKA; HABETS, 2015; VARZANDEH; TASESKA; HABETS, 2017; GÖßLING; DOCLO, 2018).

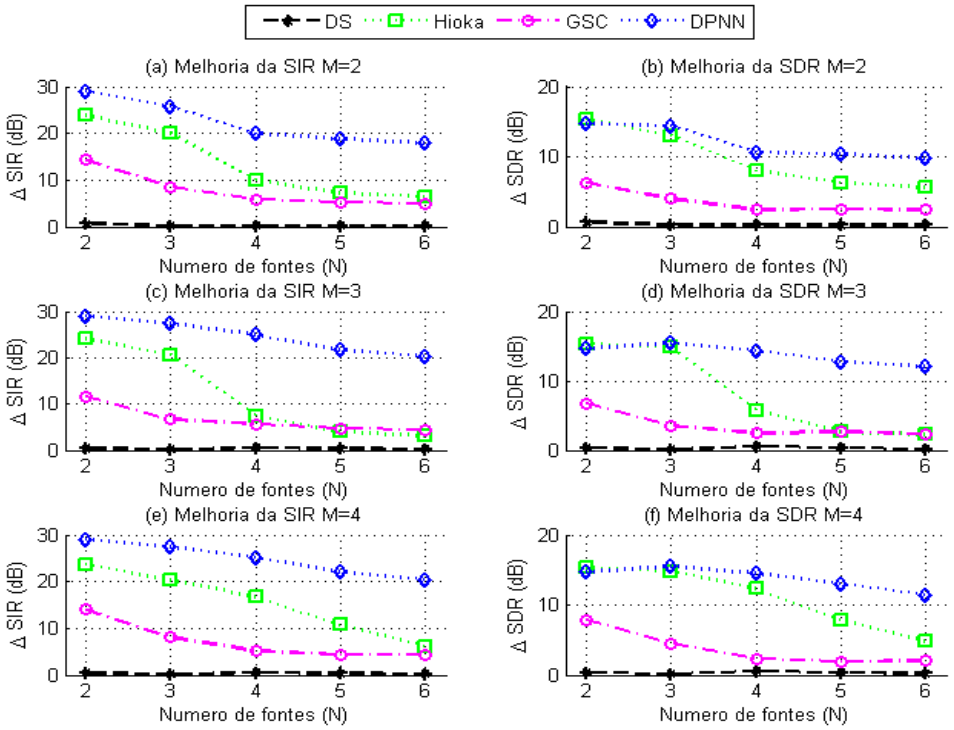
6.3 Separação das fontes

Nesta seção os métodos propostos para separação de fontes, DPNN, DPC e DPCM, são avaliados e comparados com outros métodos: Delay and sum , GSC e método proposto por Hioka. As avaliações são feitas através de simulações em MATLAB utilizando

segmentos de áudio devidamente mixados. O método DPCM utiliza as M fontes dominantes detectadas pelo método DPC.

O primeiro método avaliado é o DPNN, as figuras 12, 13 e 14 apresentam os valores das métricas objetivas de avaliação da separação. O desempenho do método proposto é apresentado comparando com outras abordagens. Como referência básica é colocado o *Delay-and-Sum*, indicado nos gráficos como DS que, neste caso, apresenta a saída dos *beamformers* quando direcionados para a fonte de interesse, os coeficientes DS são calculados conforme (103). O segundo método utilizado na comparação é o GSC, conforme apresentado na seção 3.5.3. O método utiliza filtros adaptativos baseados no algoritmo LMS (SHYNK, 1992). Para o processo adaptativo é importante que os coeficientes não sejam adaptados quando a fonte de interesse está ativa. Como já mencionado, existe a necessidade de eficientes detectores de voz (VADs) para identificar quando a fonte está ativa. Foi utilizado nas simulações um VAD ideal, ou seja, foi utilizado o sinal original (não disponível por ser um ambiente real) para bloquear o ajuste dos filtros quando o sinal está ativo. O terceiro método comparado é o proposto por Hioka (HIOKA *et al.*, 2013) que é o estado da arte utilizando estimativa de PSD pela direcionalidade de *beamformers*.

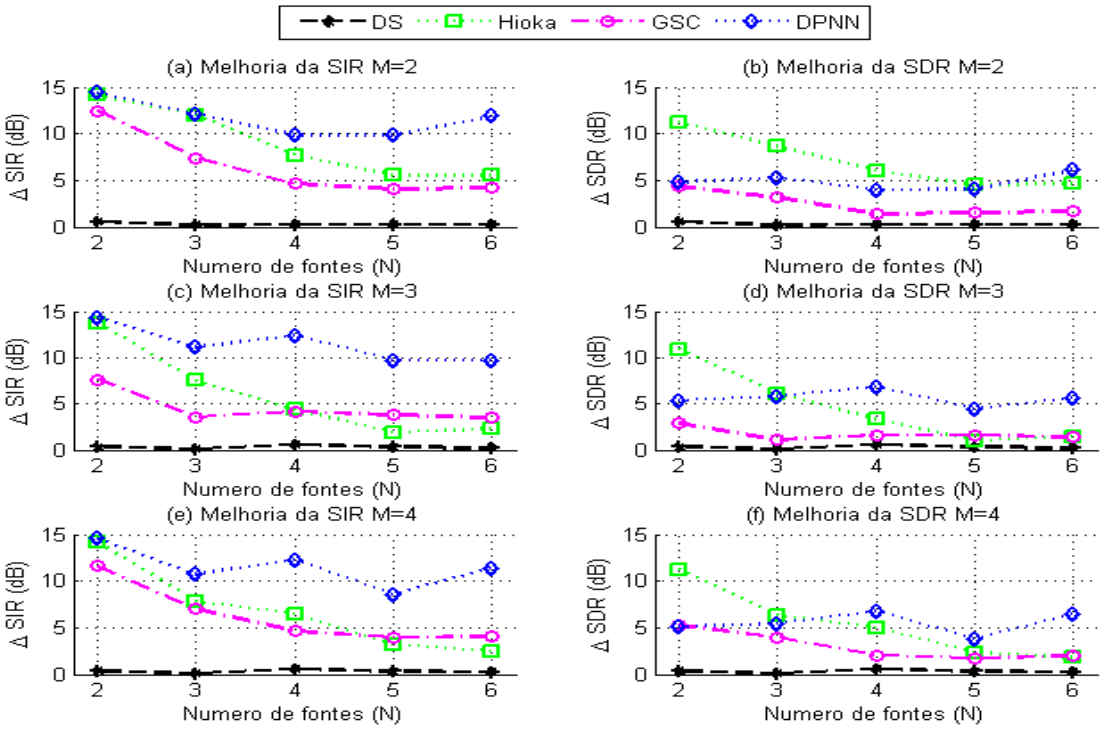
Figura 12 – Melhoria da SIR e SDR simulado em ambiente anecoico variando o número de fontes de sinal N de 2 a 6 e o número de microfones M de 2 a 4.



Fonte: do autor

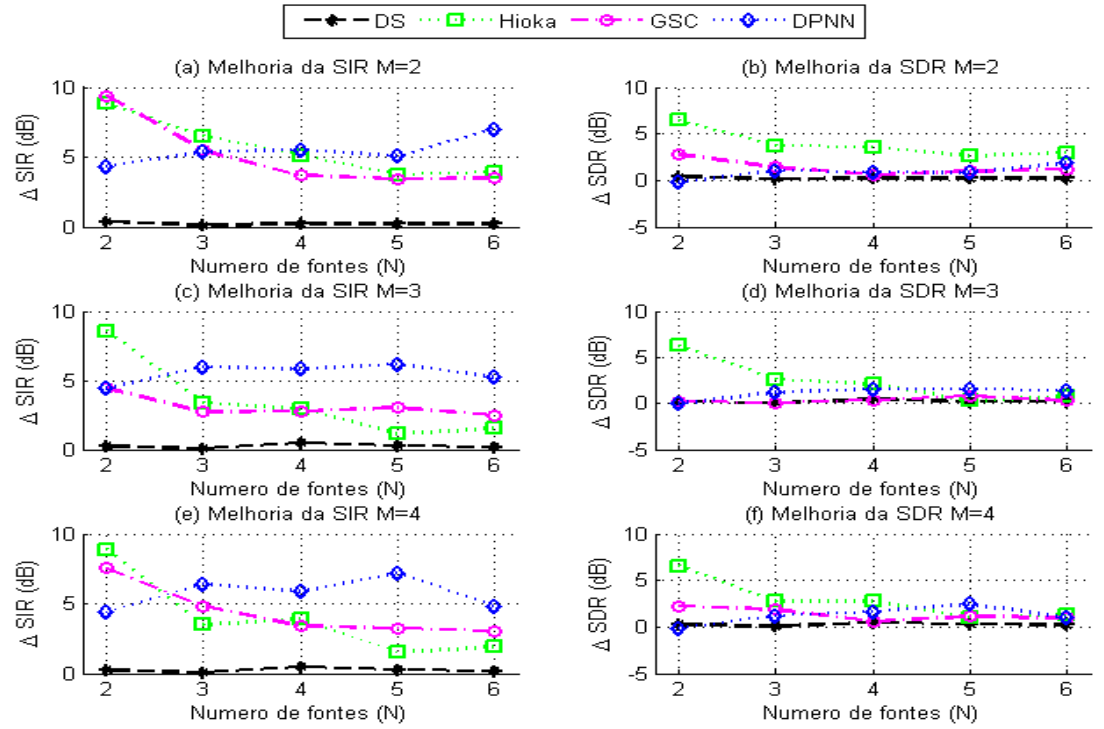
Os resultados apresentados na Figura 12 mostram que o método proposto supera os outros métodos para todas as combinações de microfones, fontes e ambientes simulados.

Figura 13 – Melhoria da SIR e SDR simulado em ambiente ecoico $RT_{60} = 100$ms, variando o número de fontes de sinal N de 2 a 6 e o número de microfones M de 2 a 4.



Fonte: do autor

Figura 14 – ΔSIR e ΔSDR simulados em ambiente ecoico $RT_{60} = 200$ms, variando o número de fontes de sinal N de 2 a 6 e o número de microfones M de 2 a 4.

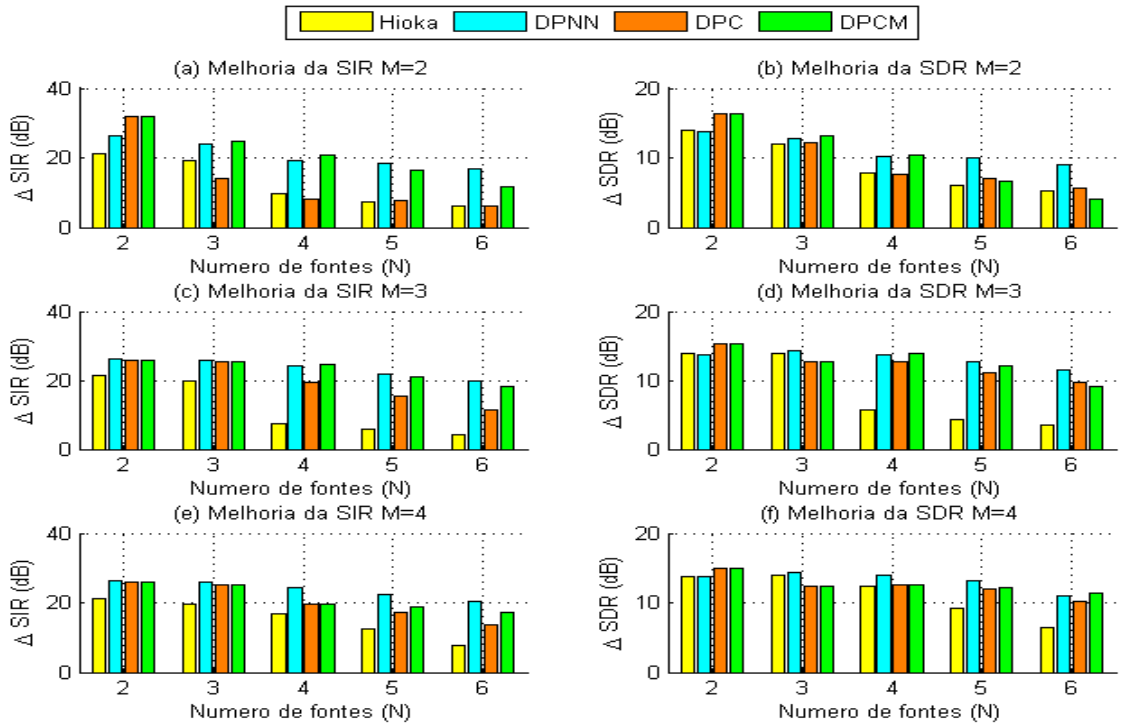


Fonte: do autor

As simulações utilizadas apresentam algumas características que afetam o condicionamento da matriz de diretividade D: sinais de banda larga, presença de baixas frequências, simetrias na geometria do arranjo do microfone e posição da fonte. O desempenho permanece superior e estável mesmo quando o método Hioka atinge o limite superior e falha. Já em ambiente com reverberação, figuras 13 e 14, os níveis de SIR continuam superiores aos outros métodos mas com uma diminuição na relação sinal-distorção, ficando abaixo do Hioka em algumas situações, principalmente nos casos onde o número de microfones é menor que o de fontes.

Nas próximas simulações, os três métodos propostos, DPNN, DPC e DPCM serão comparados entre si e com o método proposto por Hioka. As simulações no caso anecoico aparecem na Figura 15. De acordo com esta Figura a separação do método proposto DPNN e DPCM, em termos de ΔSIR e ΔSDR foi maior que no método proposto por Hioka em todas as simulações, já o DPC foi superior na maioria das simulações e muito semelhante nas outras. O DPNN apresentou melhor separação que o DPC, mas o DPCM apresenta desempenho superior ao DPNN para o número de fontes menor que o de microfones e desempenho similar com o número de fontes superior ao dos microfones (caso subdeterminado).

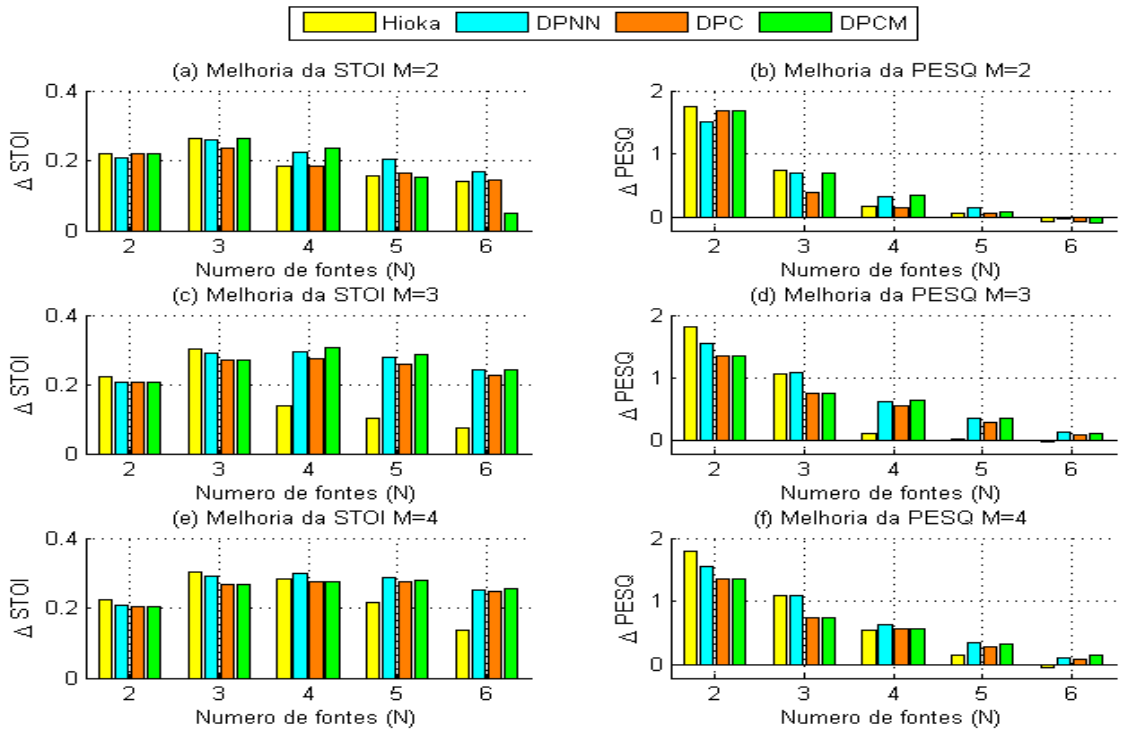
Figura 15 – ΔSIR e ΔSDR simulados em ambiente anecoico variando o número de fontes de sinal de 2 a 6 e o número de microfones de 2 a 4.



Fonte: do autor

As mesmas simulações também foram avaliadas em termos das métricas STOI e PESQ para avaliar as variações na inteligibilidade e na qualidade dos sinais obtidos, os resultados são apresentados na Figura 16.

Figura 16 – Variação da STOI e PESQ simulados em ambiente anecoico variando o número de fontes de sinal de 2 a 6 e o número de microfones de 2 a 4.



Fonte: do autor

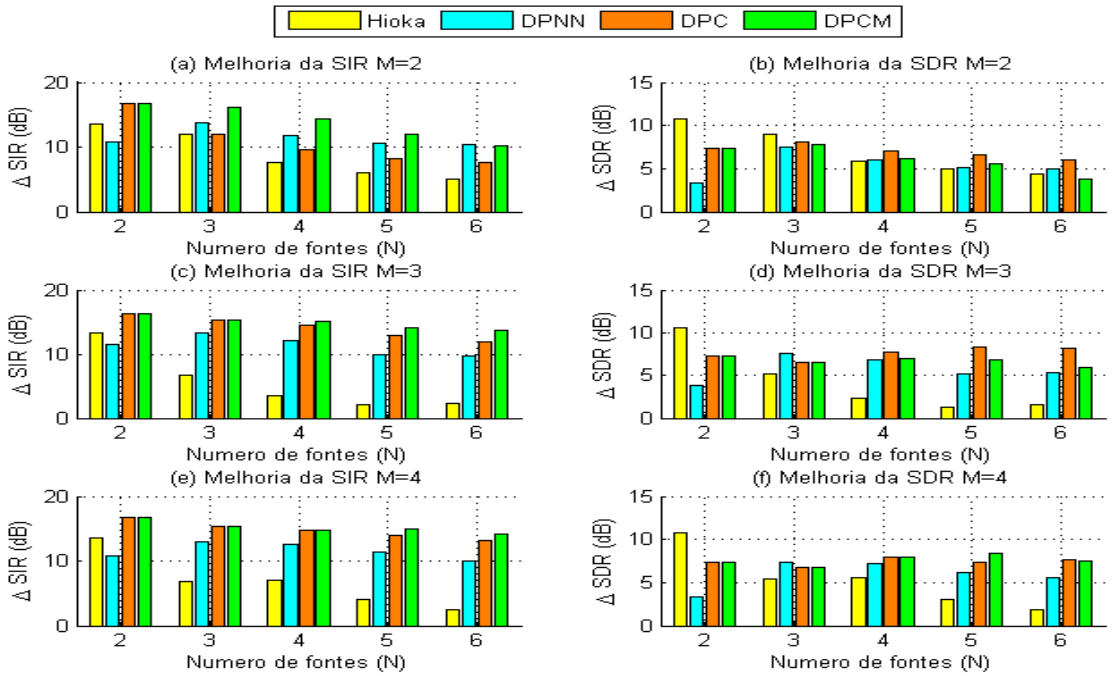
De acordo com os valores de STOI obtidos todos os métodos apresentam melhoria na inteligibilidade e os três métodos propostos DPNN, DPC e DPCM têm desempenho superior ao método do Hioka em quase todas as simulações. Para os casos em que o número de fontes é menor ou igual ao número de microfones o desempenho do Hioka é levemente superior aos outros métodos. Os métodos propostos foram bastante semelhantes entre si, superando 0.2 e chegando a .3, ou seja, uma melhoria de 20% a 30%. No caso subdeterminado todos superaram o método Hioka, a exceção da simulação para 2 microfones e 6 fontes no qual o DPCM teve desempenho inferior.

As avaliações relacionadas à percepção da qualidade dos sinais usando PESQ indicam melhorias em todos os métodos mas percebe-se claramente o decréscimo no valor com o aumento do número de fontes. Importante observar que os valores são sempre positivos indicando melhora na percepção de qualidade, diferenças maiores que 0.2 PESQ são claramente percebidas por ouvintes (SERVETTI; MARTIN, 2005).

As figuras 17 e 18 apresentam os resultados das simulações em ambientes ecoicos (reverberantes) ($RT_{60} = 100$ms e $RT_{60} = 200$ms). Nestas condições os métodos DPC e DPCM apresentaram níveis de separação maiores que DPN e Hioka. Isto pode ser explicado porque a suposição de que as fontes são W-DO é violada em ambiente ecoico (YILMAZ; RICKARD, 2004), indicando que o aumento da correlação piora o desempenho dos métodos Hioka e DPNN que assumem essa premissa em sua formulação. Também não foi percebida diferença significante entre os métodos DPC e DPCM nessas condições,

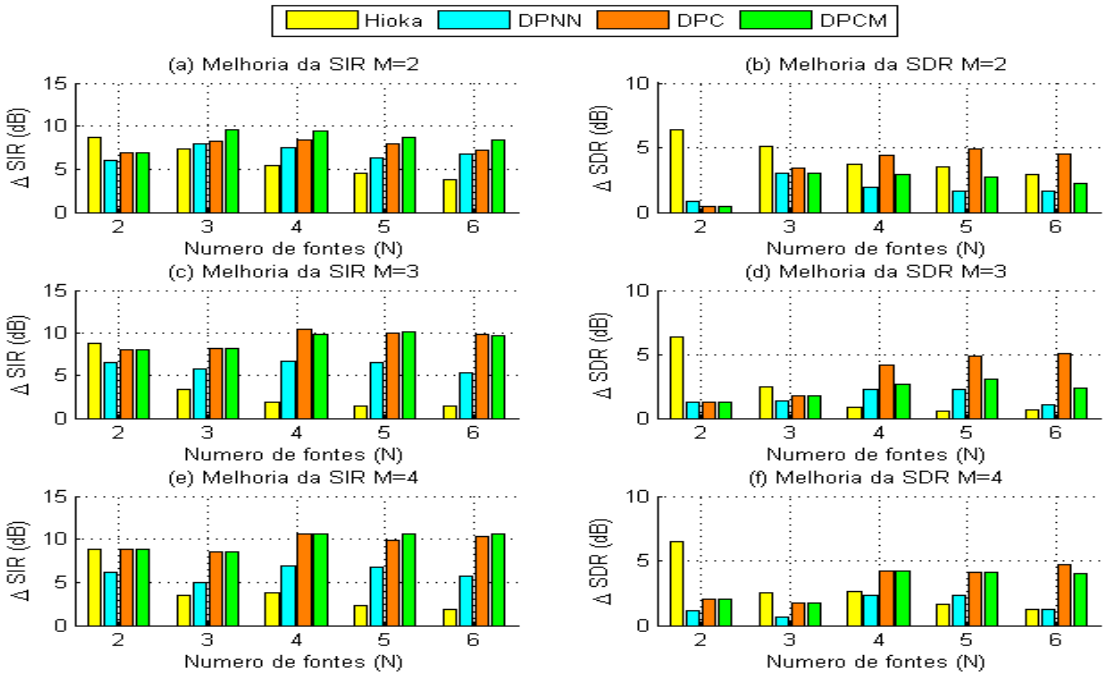
o que se explica pela diminuição da esparsidade devido à reverberação com o aumento do número de fontes presentes simultaneamente.

Figura 17 – Variação da SIR e da SDR simulados em ambiente ecoico com $RT_{60} = 100$ms variando o número de fontes de sinal de 2 a 6 e o número de microfones de 2 a 4.



Fonte: do autor

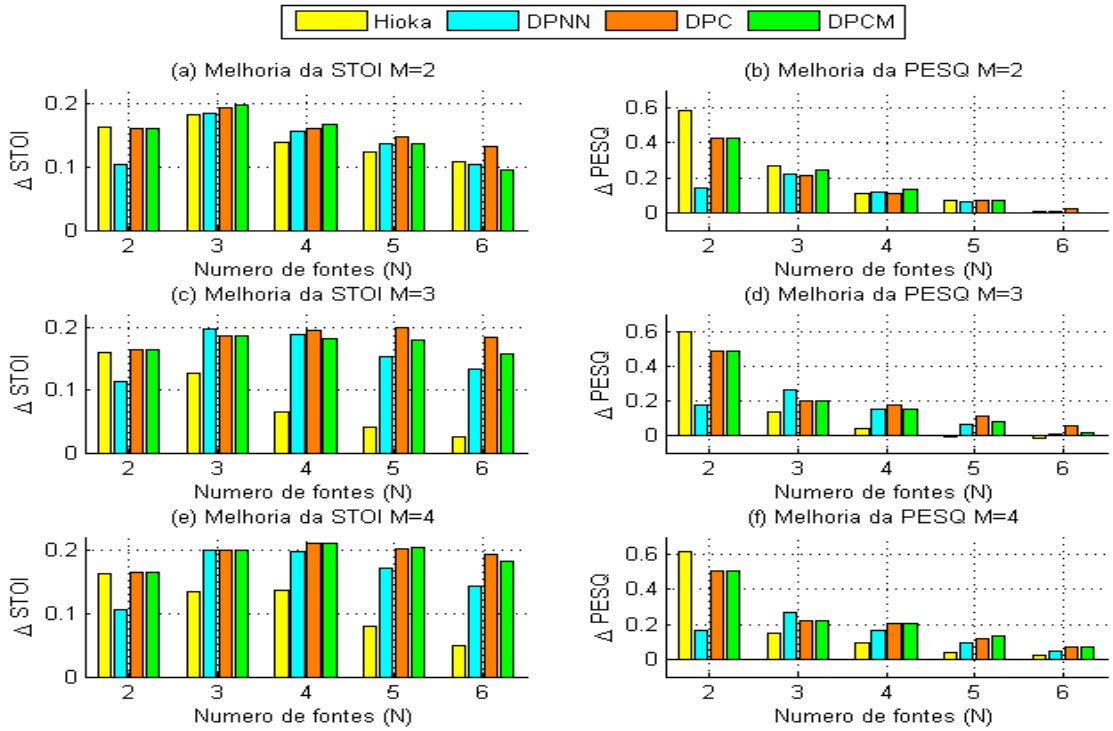
Figura 18 – Melhoria da SIR e SDR simulados em ambiente ecoico com $RT_{60} = 200$ms variando o número de fontes de sinal de 2 a 6 e o número de microfones de 2 a 4.



Fonte: do autor

As simulações em ambientes reverberantes também foram avaliadas com STOI e PESQ para avaliar as variações na inteligibilidade e na qualidade dos sinais obtidos e os resultados são apresentados nas figuras 19 e 20. Houve um ganho menor de STOI em todos os métodos com o aumento da reverberação. DPC e DPCM tiveram desempenho superior em relação ao DPNN, diferentemente do caso anecoico onde os valores de STOI nesses métodos eram equivalentes. Os métodos propostos também superam Hioka principalmente com o aumento do número de fontes.

Figura 19 – Variações de STOI e PESQ simulados em ambiente ecoico com $RT_{60} = 100$ms variando o número de fontes de sinal de 2 a 6 e o número de microfones de 2 a 4.

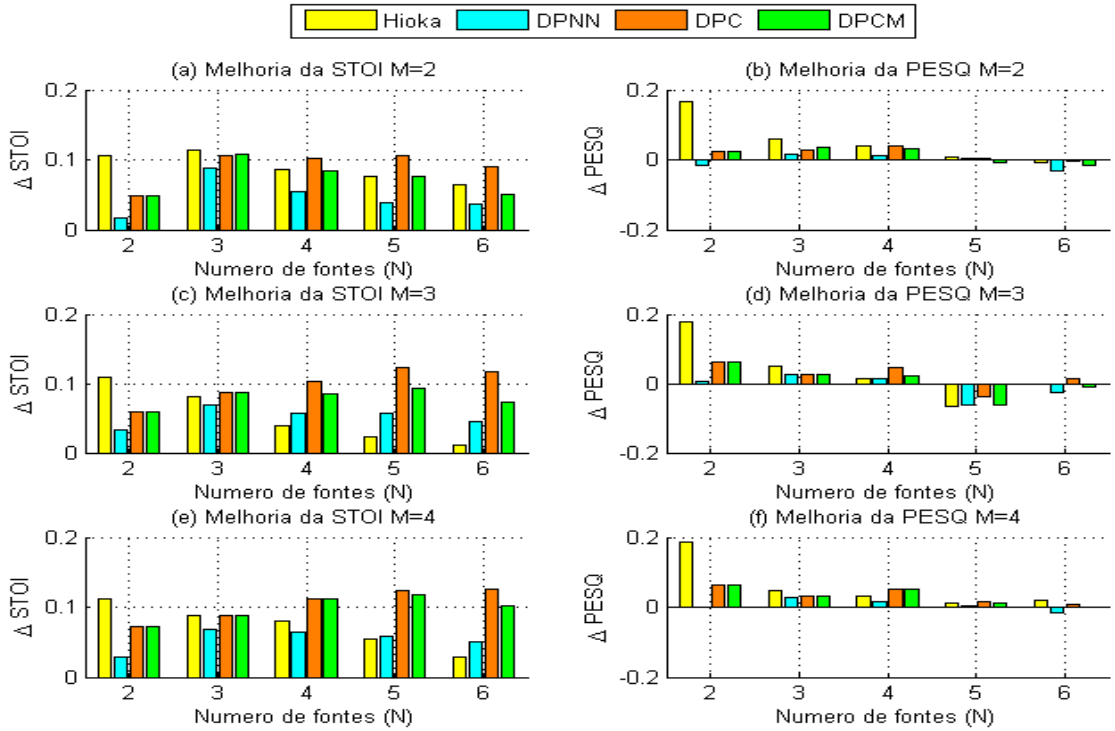


Fonte: do autor

Com relação ao PESQ nos casos reverberantes a tendência de queda com o aumento do número de fontes se manteve. A melhoria de PESQ foi pouco significativa em todos os métodos simulados na situação ecoica, com poucas situações superando a variação perceptível de 0.2 PESQ. Apesar do baixo desempenho em melhoria de PESQ em nenhum caso simulado ocorreu piora perceptível na avaliação de qualidade havendo uma diminuição máxima menor que 0.1 PESQ, só verificada com o número máximo simulado de 6 fontes.

As figuras 21 e 22 apresentam os mesmos valores de SIR e SDR já mostrados para o caso de 3 microfones porem é acrescida a informação do desvio padrão das medidas. Da mesma forma a informação de desvio padrão é adicionada ao caso de 3 microfones nas condições ecoicas com os resultados apresentados nas figuras 23, 24, 25 e 26. A vaiação em termos do desvio padrão se mantém constante independente do número de fontes sonoras. O método DPCM apresentou uma variação de desvio um pouco superior

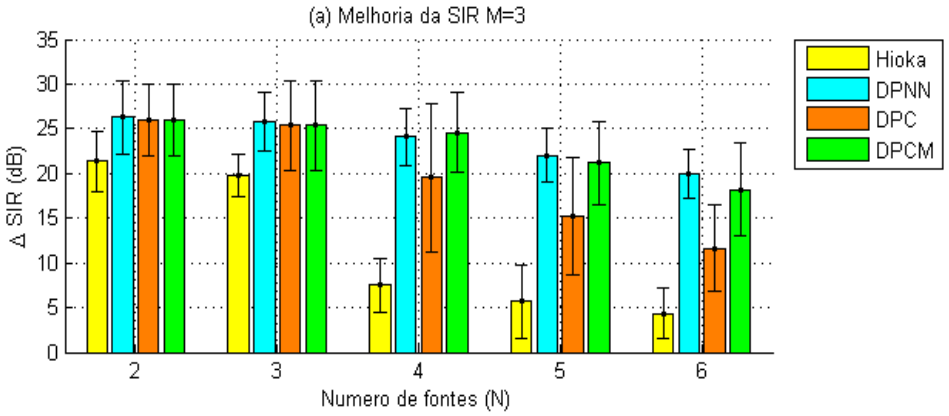
Figura 20 – Melhoria da STOI e PESQ simulados em ambiente ecoico com $RT_{60} = 200$ms variando o número de fontes de sinal de 2 a 6 e o número de microfones de 2 a 4.



Fonte: do autor

aos demais métodos no cano anecoico. Os métodos propostos apresentaram desvio maior que o método Hioka mas devido a melhoria apresentada os métodos propostos apresentam vantagem no desempenho principalmente número de fontes maior que o de microfones e em casos ecoicos.

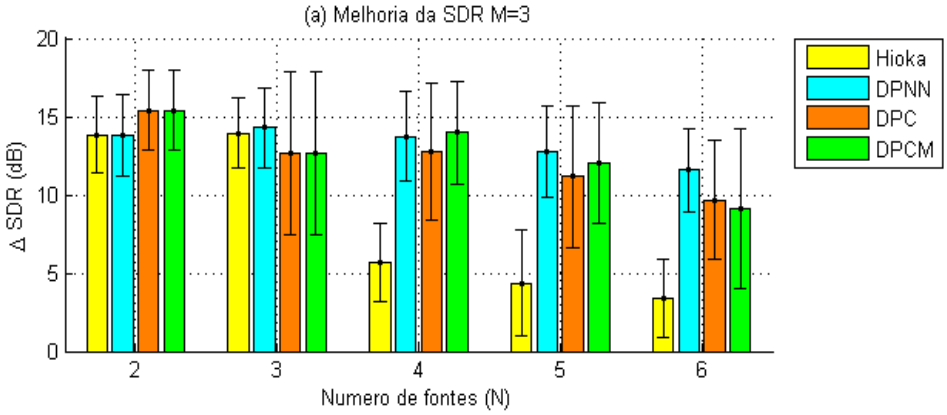
Figura 21 – ΔSIR com desvio padrão simulados em ambiente anecoico variando o número de fontes de sinal de 2 a 6 e o número de microfones igual a 3.



Fonte: do autor

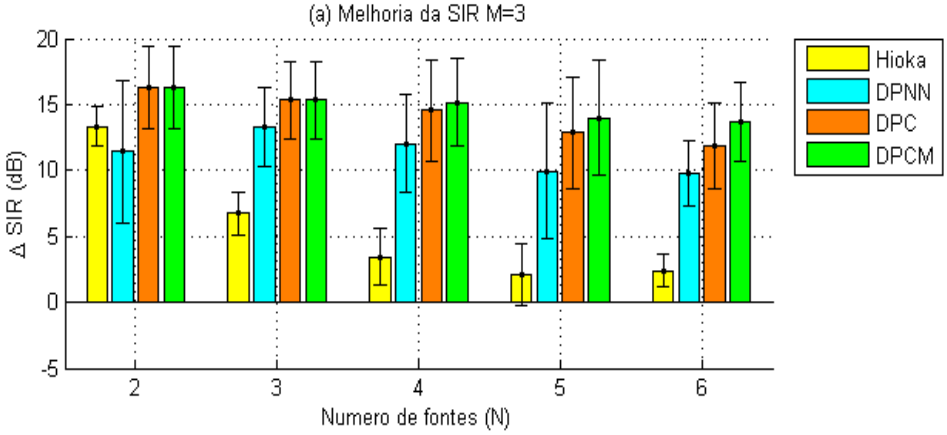
Outra avaliação realizada foi no procedimento de recuperação dos sinais a partir das PSDs estimadas. No método DPCM O sinal filtrado pelo pós-filtro de Wiener é uma estimativa a partir da amplitude e fase das PSDs estimadas, diferente do método em [(HIOKA

Figura 22 – ΔSDR com desvio padrão simulados em ambiente anecoico variando o número de fontes de sinal de 2 a 6 e o número de microfones igual a 3.



Fonte: do autor

Figura 23 – Variação da SIR com desvio padrão simulados em ambiente ecoico com $RT_{60} = 100$ms variando o número de fontes de sinal de 2 a 6 e o número de microfones igual a 3.

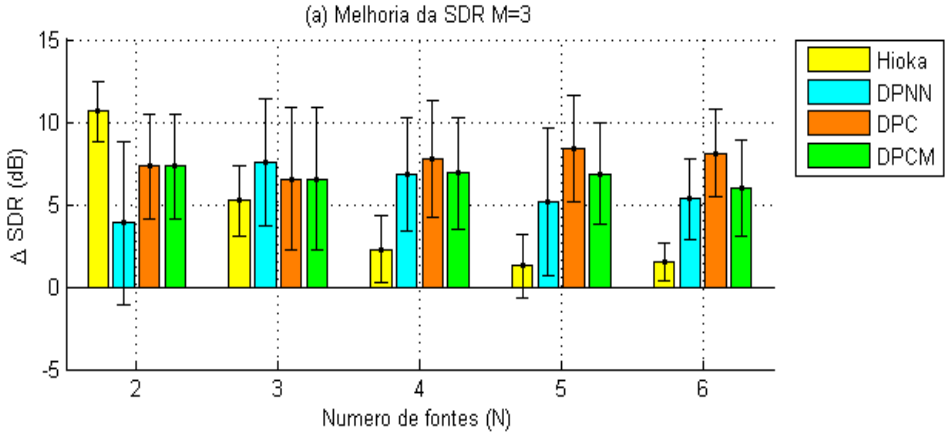


Fonte: do autor

et al., 2013)] que filtra diretamente a saída dos *beamformers*. O método DPCM comparado com uma variação do DPCM em que o pós-filtro é aplicado diretamente na saída dos *beamformers*. Esta comparação é apresentada nas figuras 27, 28 e 29 e é mostrada com o nome de DPCMb (Directional PSDs estimation with Correlation restricted to M source with beamformer filtering). Esta variação apresenta pior desempenho que o DPCM nos casos anecoicos mas ainda com desempenho superior ao proposto por Hioka. Esta variação se mostrou eficaz na melhoria da SDR principalmente em ambiente ecoico tendo também um menor desvio padrão.

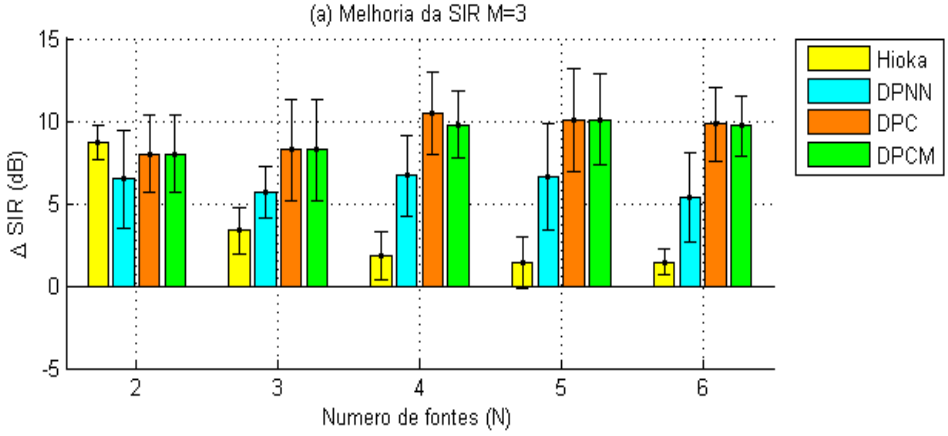
Na condição anecoica os métodos DPNN e DPCM apresentaram resultados semelhantes e superiores aos demais métodos comparados. Neste caso o método DPNN apresenta vantagens pela sua simplicidade e menor carga computacional. Nas simulações com com reverberação o DPNN teve desempenho inferior aos métodos DPC e DPCM. Nas condi-

Figura 24 – Variação da SDR com desvio padrão simulados em ambiente ecoico com $RT_{60} = 100$ms variando o número de fontes de sinal de 2 a 6 e o número de microfones igual a 3.



Fonte: do autor

Figura 25 – Variação da SIR com desvio padrão simulados em ambiente ecoico com $RT_{60} = 200$ms variando o número de fontes de sinal de 2 a 6 e o número de microfones igual a 3.

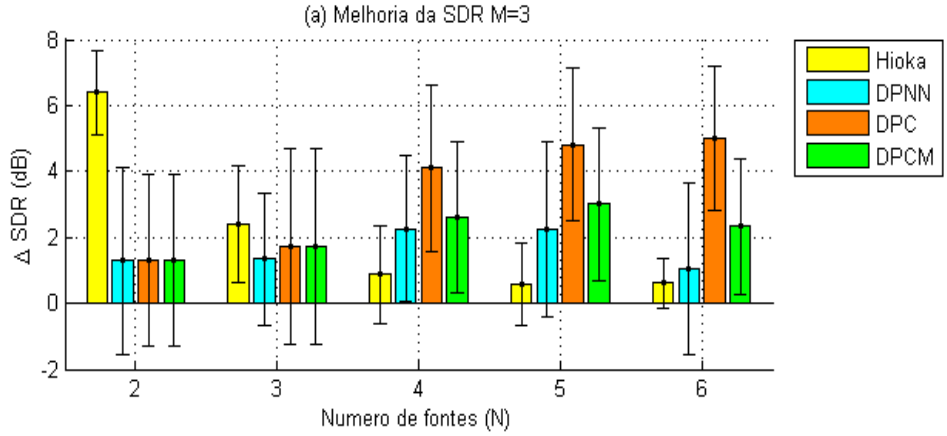


Fonte: do autor

ções ecoicas os métodos DPC e DPCM tem desempenho semelhante, sendo o DPC um pouco superior e com custo computacional menor. Em todos as situações a SDR pode ser melhorada, com o compromisso de uma diminuição na SIR se utilizar a pós-filtragem diretamente na saída dos *beamformer*. Se for considerada a necessidade de operar em todas condições, com e sem reverberação, a melhor escolha é o DPCM que obteve boa separação em todas as condições.

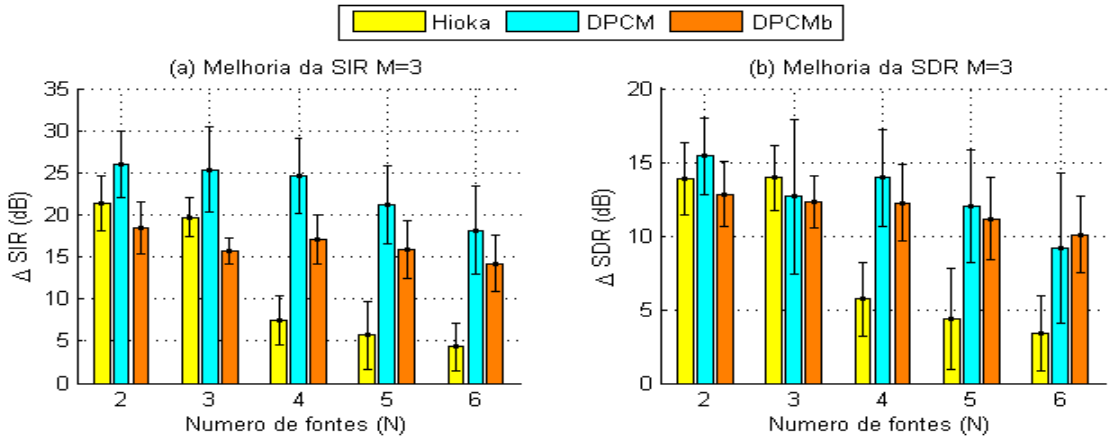
Para avaliar a aplicabilidade das Funções de Transferência Relativa os métodos propostos foram comparados usando as ATFs e as RTFs e os resultados sem reverberação são apresentados na Figura 30. Os valores indicados como DPNN, DPC e DPCM usam funções de transferência acústica como descrito na seção 4.3. As simulações indicadas como DPNNr, DPCr e DPCMr usam funções de transferência relativas de acordo com a

Figura 26 – Variação da SDR com desvio padrão simulados em ambiente ecoico com $RT_{60} = 200$ms variando o número de fontes de sinal de 2 a 6 e o número de microfones igual a 3.



Fonte: do autor

Figura 27 – Comparação do método DPCM ΔSIR e ΔSDR com a saída obtida pela filtragem da saída dos *beamformers*. Simulação em ambiente anecoico variando o número de fontes de sinal de 2 a 6 e o número de microfones de 2 a 4.



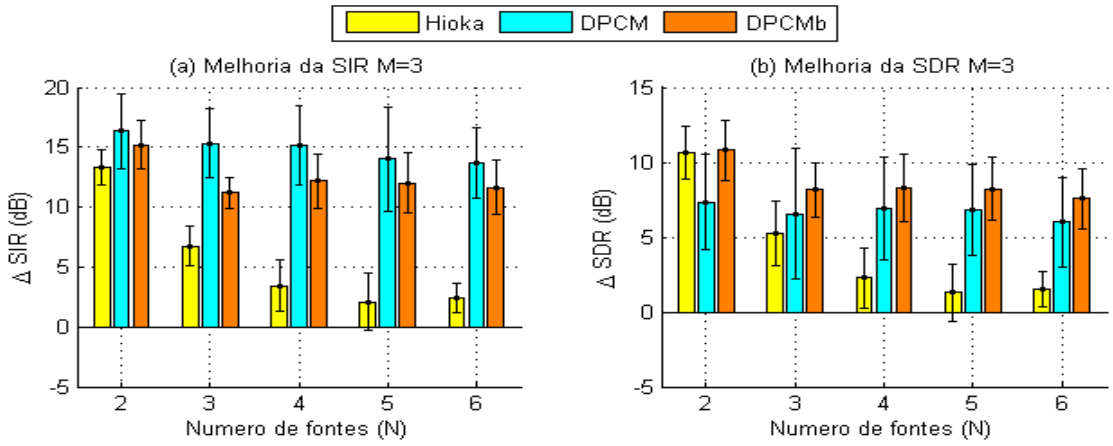
Fonte: do autor

seção 4.4.

Na simulação em ambiente anecoico o uso das ATFs e das RTFs apresentaram resultados de separação em termos da ΔSIR e da ΔSDR equivalentes. Comparando DPNN com DPNNr verificas-se os mesmos níveis de separação e da mesma forma comparando DPC com DPCr e DPCM com DPCMr também tiveram a mesma performance.

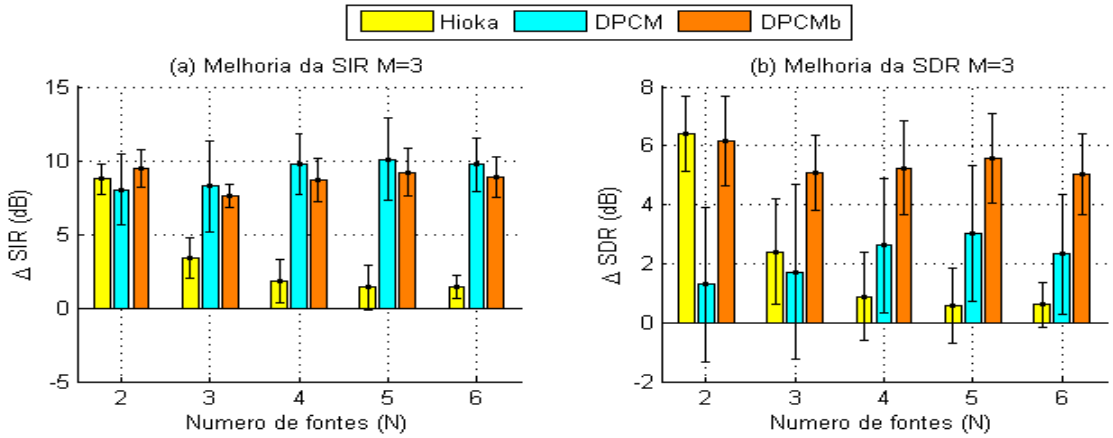
Nos ambiente ecoicos para $RT_{60} = 100$ms e $RT_{60} = 200$ms, apresentado nas Figuras 31 e 32, obteve-se algumas diferenças que se explicam pelo uso de funções de transferência truncadas; as janelas de análise são bem mais curtas que as respostas ao impulso do ambiente. Os sinais foram gerados utilizando as funções completas e nas separações foram utilizadas as funções truncadas. Na maioria das simulações com reverberação a separação utilizando as RTFs apresentou melhor desempenho que a separação utilizando

Figura 28 – Comparação do método DPCM ΔSIR e ΔSDR com a saída obtida pela filtragem da saída dos *beamformers*. Simulação em ambiente ecoico com $RT_{60} = 100$ms variando o número de fontes de sinal de 2 a 6 e o número de microfones de 2 a 4.



Fonte: do autor

Figura 29 – Comparação do método DPCM ΔSIR e ΔSDR com a saída obtida pela filtragem da saída dos *beamformers*. Simulação em ambiente ecoico com $RT_{60} = 200$ms variando o número de fontes de sinal de 2 a 6 e o número de microfones de 2 a 4.



Fonte: do autor

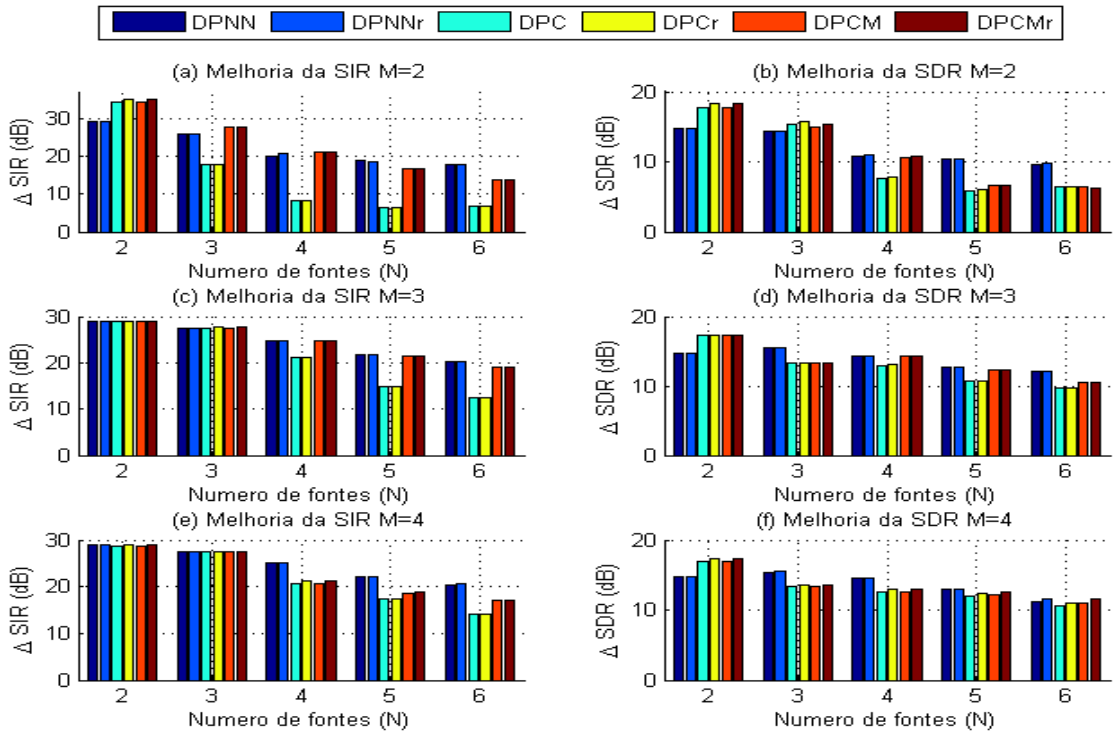
as ATFs.

6.4 Estimativa das RTFs

Diferentemente das ATFs, as RTFs podem ser obtidas diretamente dos microfones sem um procedimento de medição específico. Como demonstrado na seção anterior, o uso das RTFs apresenta desempenho semelhante ao uso das ATFs para a separação de fontes. Nesta seção, serão avaliados os métodos propostos para estimar as RTF e o desempenho dos valores estimados na separação de origem.

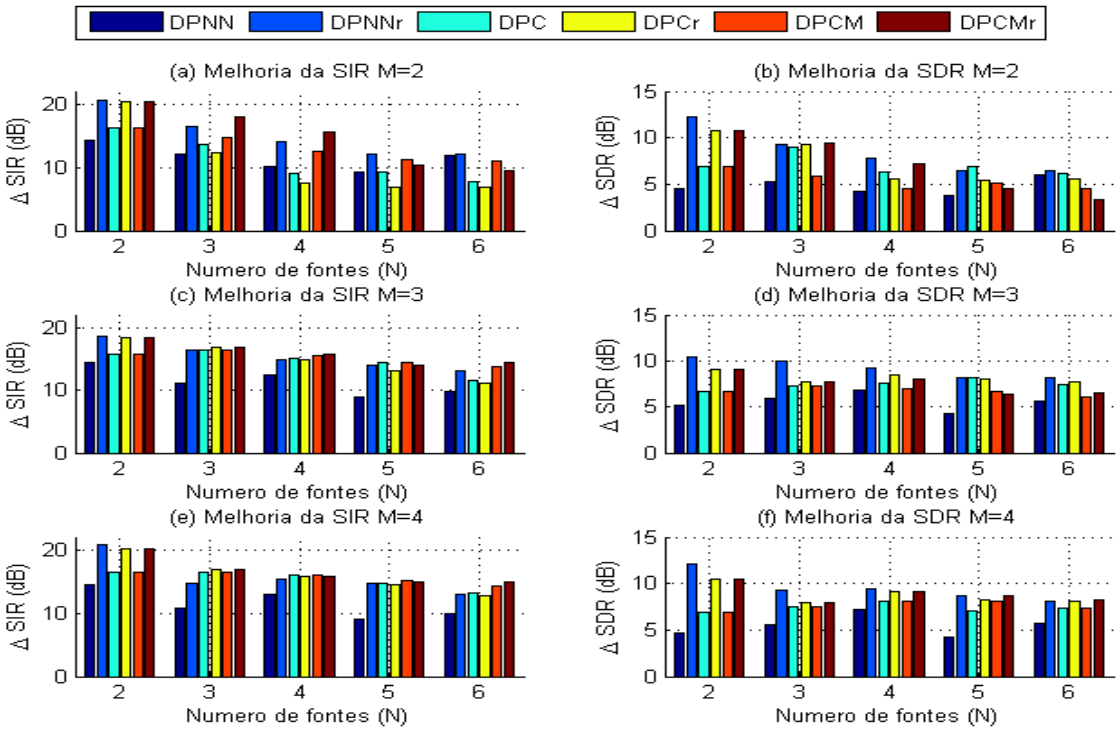
As próximas simulações avaliam as estimativas das RTFs e o desempenho dos valores estimados na separação das fontes. Primeiro, é apresentado como exemplo a estimativa

Figura 30 – Comparação dos valores de ΔSIR e ΔSDR utilizando ATFs e RTFs nos métodos propostos em ambiente anecoico



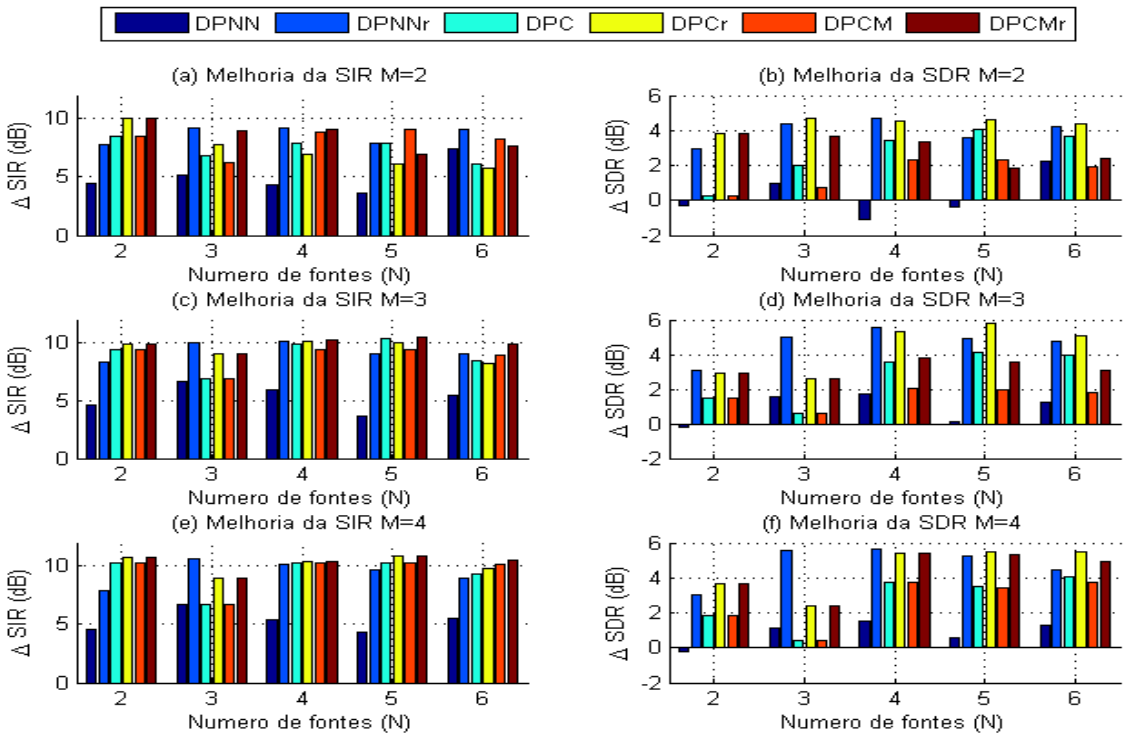
Fonte: do autor

Figura 31 – Separação usando ATFs e RTFs para ambiente ecoico $RT_{60} = 100$ms



Fonte: do autor

Figura 32 – Separação usando ATFs e RTFs para ambiente ecoico $RT_{60} = 200$ms

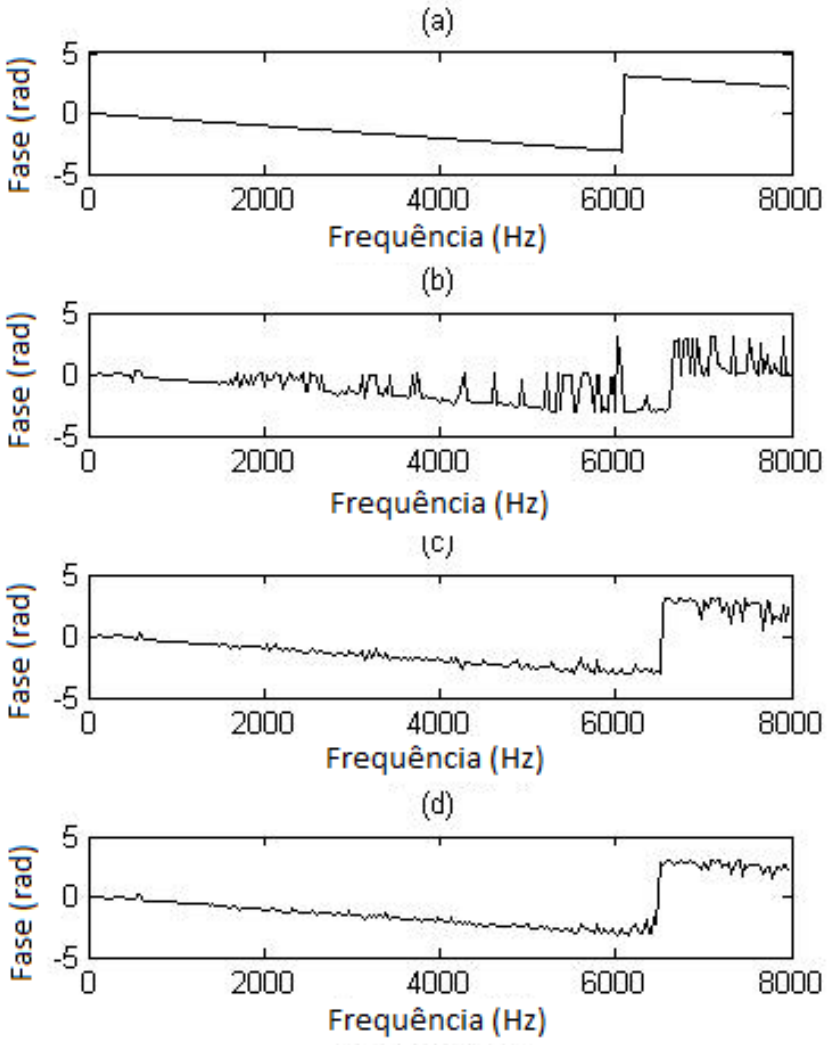


Fonte: do autor

para um cenário de 3 fontes. As resoluções dos histogramas são de 2000 pontos e a função de suavização usa 40 pontos de comprimento. O objetivo é demonstrar que o uso das PSDs pode melhorar a estimativa das RTFs. A Figura 33 apresenta a fase da RTF estimada antes e depois de re-estimar as RTFs ponderados pelos PSDs. É possível observar de forma subjetiva, apenas para observar as modificações na estimativa, que, após a aplicação dos pesos obtidos das PSDs, a RTF apresenta menor diferença em relação à RTF real. Em várias simulações, pode-se avaliar que repetir o procedimento sucessivamente não tem uma melhora aparente na estimativa.

Esse comportamento pode estar relacionado ao fato de que o procedimento de ponderação com os pesos não afeta a posição dos picos no histograma, apenas facilita a localização, servindo para torná-los mais visíveis, uma vez que a influência de fontes detectadas de outra direção é atenuada. A Figura 34 é um exemplo de como os pesos das PSDs afetam a estimativa das RTFs. Neste exemplo, são apresentados os histogramas de fase para a frequência de 1125 Hz usando 2 microfones e 3 fontes de acordo com a Figura 11. O primeiro histograma (Figura 34a) usa pesos com valor 1, nos outros histogramas é recalculado com peso em (83), um para cada fonte. Neste exemplo o pico perto de 1 rad na Figura 34a não está correto, a posição correta é a posição detectada na Figura 34b. A posição dos picos para cada fonte é muito mais evidente do que no primeiro histograma. Nos testes subsequentes serão realizadas medidas objetivas dos erros de estimação das RTFs.

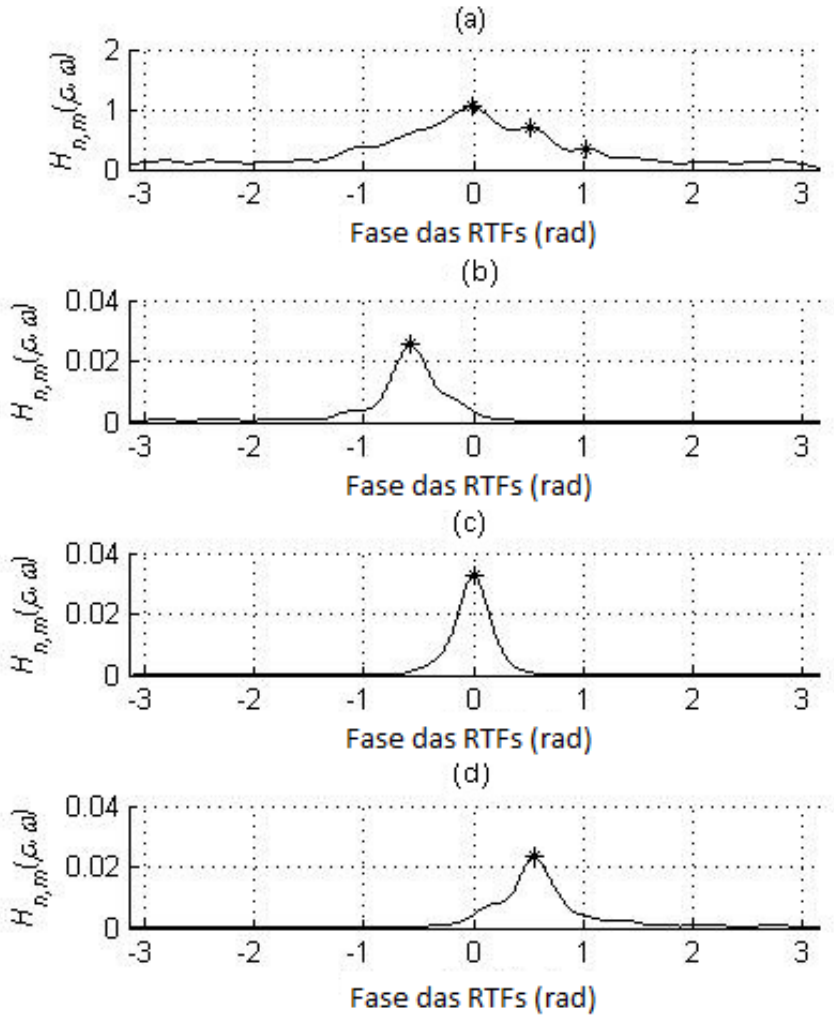
Figura 33 – Exemplo da estimativa de fase da RTF para uma fonte. (a) fase correta, (b) com pesos iguais a 1, (c) pesos calculados com as PSDs estimadas com as RTF obtida em (b) e (d) pesos calculados com as PSDs estimadas em (c)



Fonte: do autor

As simulações para estimação das RTFs foram realizadas em um ambiente anecoico utilizando o cenário de dois microfones. A simulação utiliza quadros de 512 amostras e deslocamento de 128 quadros usando uma janela de análise de Hanning. Os métodos propostos, wFDUET e wFCM-L, foram comparados simulando a separação com o método DPCM. A comparação foi feita com um método baseado em DUET (YILMAZ; RICKARD, 2004) detectando picos em histogramas suavizados de $R_k(\omega)$, diferentemente de DUET, um histograma é construído para cada frequência, como descrito na seção 5 porém com os pesos com valor 1, referido como FDUET. Também foi comparado com o método de aglomeração (*clustering*) baseado no FCM (JAFARI *et al.*, 2013), semelhante

Figura 34 – Exemplo da estimativa da fase das RTFs. (a) picos detectados com pesos iguais a 1, (b), (c) e (d) são os histogramas para cada uma das fonte, obtidos usando as PSDs e RTFs estimadas em (a), a posição do pico de cada histograma representa a fase da RTF associada a cada fonte



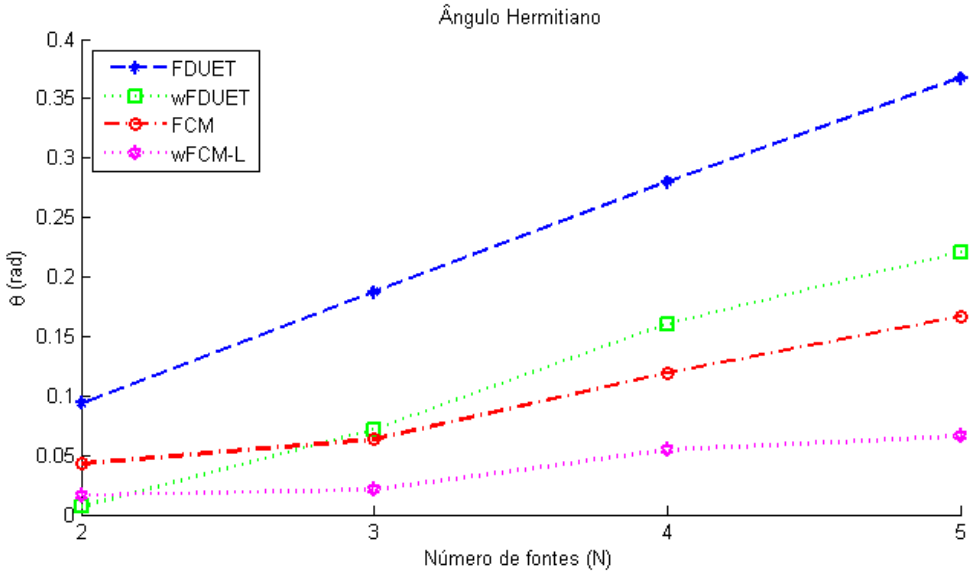
Fonte: do autor

ao wFCM apresentado na seção 5 porém onde os pesos $w_{kn}(\omega)$ possuem valor 1. O parâmetro q do FCM e do wFCM foi definido como 2. Todos os métodos usam o mesmo procedimento de permutação descrito na seção 5.2.3.

As figuras 35, 35 e 35 apresentam os erros na estimativa das RTFs para 2 microfones com o número de fontes variando de $N = 2$ a $N = 5$, elas apresentam respectivamente simulações para o caso anecoico, $RT_{60} = 100$ ms e $RT_{60} = 200$ ms. Os erros são estimados pelo no ângulo hermitiano calculado por (108), quanto menor o ângulo menor o erro. Os erros apresentados são a média sobre todas as frequências e fontes para os diferentes métodos. De acordo com a medição de erro pelo ângulo Hermitiano, o método

wFDUET diminui o erro calculado em relação ao método FDUET, mas não alcança o desempenho dos métodos baseados em aglomeração baseados no FCM. O método proposto, wFCM-L, baseado na distribuição laplaciana, apresenta o menor erro entre todos os métodos comparados.

Figura 35 – Ângulo Hermitiniano entre o vetor das RTFs e o vetor das RTFs estimadas (valor médio para todas as frequências e as fontes), simulação em ambiente anecoico

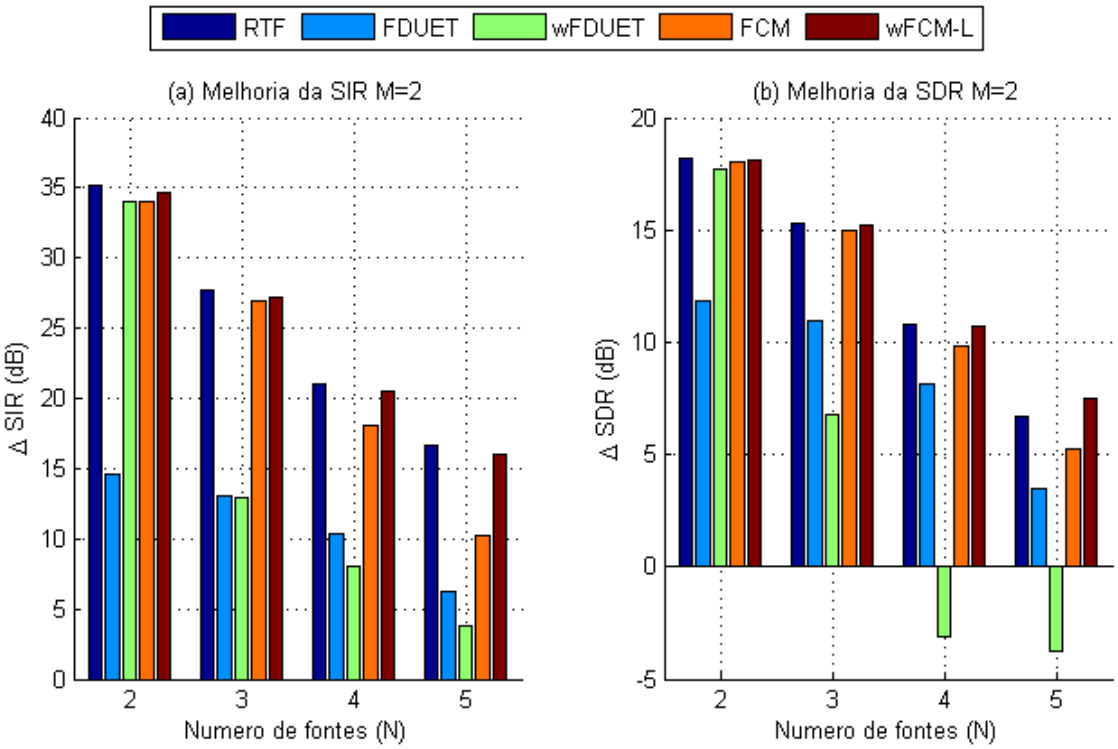


Fonte: do autor

Foram realizadas simulações envolvendo a separação das fontes utilizando as RTFs estimadas com o método de separação proposto DPCM. A Figura 36 apresenta a separação em termos de ΔSIR e ΔSDR na condição anecoica. Nas figuras 38 e 40 são apresentados os resultados da separação com RTFs estimados em simulações com reverberação $RT_{60} = 100$ ms e $RT_{60} = 200$ ms.

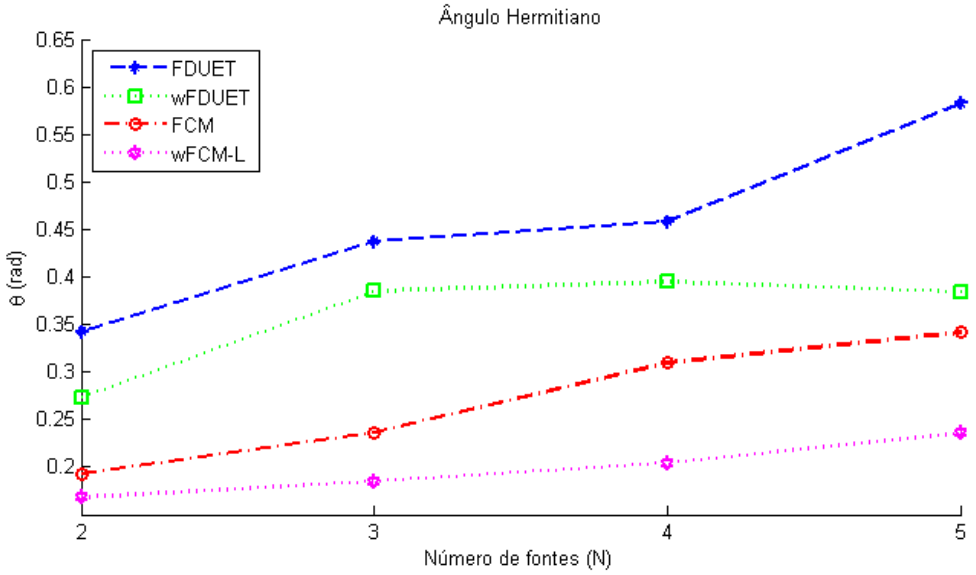
As sinais separados com as RTFs estimadas foram comparados com os sinais separados usando as RTFs verdadeiras. É importante considerar que as RTFs verdadeiras utilizada nos casos com reverberação são versões truncadas das RTFs completas, os sinais são gerados com as funções de transferência completas, mas a separação é feita considerando apenas parcela da resposta ao impulso dentro da janela de amostragem. Comparando o erro nas estimativas apresentadas nas figuras 35, 37 e 39 verifica-se que a diminuição no erro da estimativa reflete no aumento da separação apresentado nas figuras 36, 38 e 40, com exceção do método wFDUET que apresentou discrepâncias neste critério. No caso do wFDUET, foi verificado, analisando as RFTs estimadas,que ocorriam erros elevados nos ganhos $\hat{\mu}_{n,m}(\omega)$ das RTF estimadas. Já o método wFCM-L apresentou o melhor desempenho em todas as simulações com separação similar a obtida com a RTF verdadeira.

Figura 36 – Variação de SIR e SDR usando as RTFS estimadas em ambiente anecoico



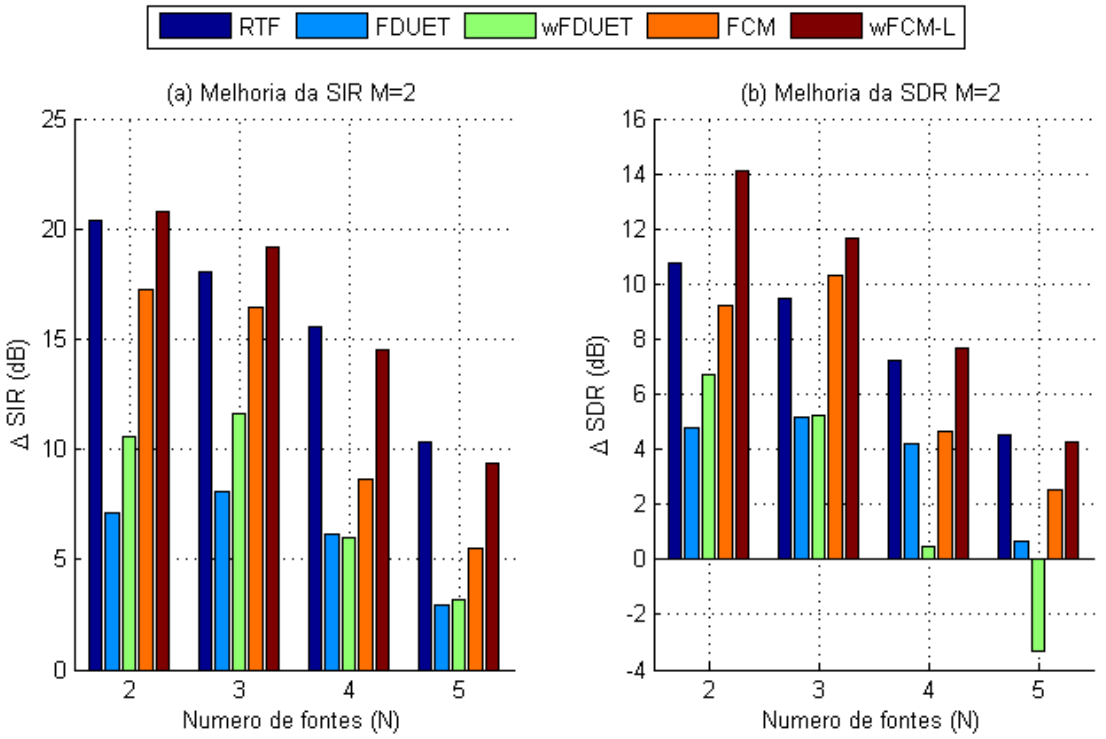
Fonte: do autor

Figura 37 – Ângulo Hermitiniano entre vetors das RTFs e o vetor das RTFs estimadas (valor médio para todas as frequencias e as fontes), simulção em ambiente ecoico $RT_{60} =$ 100ms



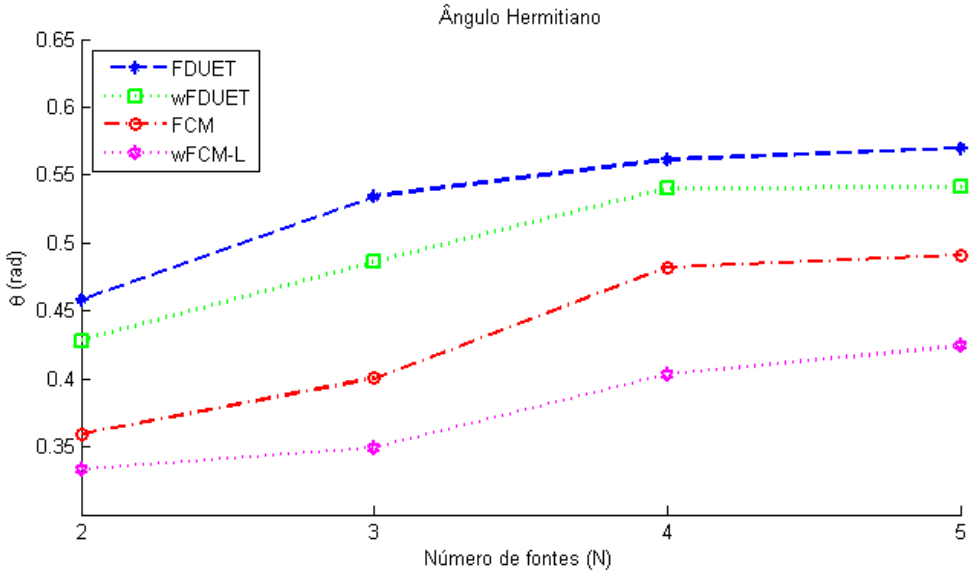
Fonte: do autor

Figura 38 – Variação de SIR e SDR usando as RTFS estimadas em ambiente ecoico $RT_{60} = 100$ms



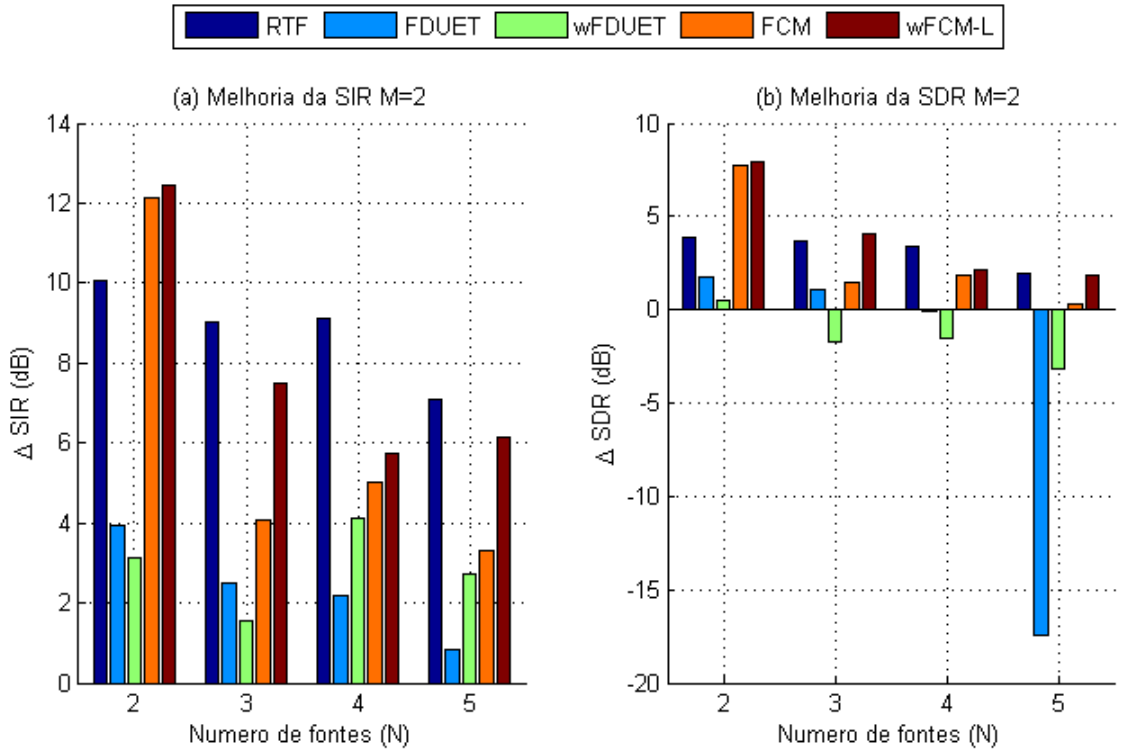
Fonte: do autor

Figura 39 – Ângulo Hermitiniano entre o vetor das RTFs e o vetor das RTFs estimadas (valor médio para todas as frequências e as fontes), simulação em ambiente ecoico $RT_{60} = 200$ms



Fonte: do autor

Figura 40 – Variação de SIR e SDR usando as RTFS estimadas em ambiente ecoico $RT_{60} = 200$ms



Fonte: do autor

7 CONCLUSÃO

Este trabalho propõe métodos baseados na diretividade de *beamformers* para serem utilizados no processo de separação de sinais de áudio utilizando arranjos de microfones e o conhecimento da localização espacial da fonte sonora. O objetivo principal da separação de sinais de áudio é estimar as fontes sonoras a partir do som captado em microfones. Algumas soluções restringem o número de microfones, o que pode ser uma limitação e reforça a importância no caso subdeterminado, onde o número de fontes a serem separadas excede o número de microfones. O trabalho se concentra em soluções aplicáveis ao caso subdeterminado e também supõe que as fontes são localizadas em direções diferentes no espaço considerando número reduzido de microfones visando possibilitar implementações binaurais. Arranjos de microfones são utilizados para este fim, pois a combinação dos sinais obtidos de um arranjo de microfones resulta em uma resposta que depende da posição espacial da origem do som. Esta combinação que depende da quantidade e da disposição dos microfones é conhecida como formador de feixe ou *beamformer*. Neste trabalho, são propostos métodos para separação de sinais utilizando o mesmo princípio proposto por Hioka (HIOKA *et al.*, 2013), no qual a estimativa das PSDs das fontes sonoras é feita baseada nos ganhos de diretividade dos *beamformers*, ou seja, usa-se a propriedade dos *beamformers* de apresentar respostas que dependem da posição espacial das fontes.

A primeira contribuição deste trabalho utiliza uma abordagem de minimização de mínimos quadrados com restrição Não Negativa é usada para resolver a formulação de ganho de Diretividade das PSD, esta abordagem permite que a limitação do número de fontes seja superada sem a instabilidade numérica até então existente, como no proposto por Hioka; este método foi denominado DPNN (Directional PSDs estimation with Non-Negative restrictions). Esta abordagem encontra uma solução que minimiza o erro quadrático e ao mesmo tempo restringe a solução a valores não negativos, que está de acordo com o sentido físico das grandezas estimadas. O desempenho permanece superior e com estabilidade numérica mesmo quando o método Hioka atinge o limite superior e falha. Já em ambiente com reverberação os níveis de relação sinal-interferência (SIR) continuam superiores aos outros métodos mas com uma diminuição na relação sinal-distorção (SDR)

ficando abaixo do Hioka neste quesito em algumas situações, principalmente nos casos onde o número de microfones é menor que o de fontes.

Outro método proposto, chamado DPC (Directional PSDs estimation with Correlation), foi desenvolvido assumindo a hipótese da existência de correlação entre as fontes sonoras. Com janelas de análise de curta duração, devido a não estacionariedade do sinal da fala, o resultado das STFTs são sinais ditos aproximadamente W-DO (W-disjoint orthogonals) (RICKARD; YILMAZ, 2002) (YILMAZ; RICKARD, 2004). Isso significa que a sua representação de tempo e frequência se sobrepõe e a condição de perfeita não correlação entre os sinais não pode ser assumida, Com uso das PSDs direcionais e considerando explicitamente a correlação dos sinais através das PSDs cruzadas dos sinais de saída dos *beamformers*, permite-se relevar a restrição de não correlação entre os sinais. O método proposto é utiliza as PSDs das saídas dos *beamformers* e as PSDs cruzadas das saídas dos *beamformers* para estimar as PSDs de cada fonte sonora. As simulações demonstram seu maior desempenho em comparação com os métodos que negligenciam a correlação mesmo nos casos subdeterminados e, principalmente, com a presença de reverberação.

Como os sinais são aproximadamente W-DO devido à esparsidade dos sinais, podemos assumir que temos poucas fontes dominantes, com essa premissa, foi proposto o método DPCM (Directional PSDs estimation with Correlation restricted to M sources) que aborda a solução de caso subdeterminado, assumindo ativas apenas as M fontes dominantes (M igual ao número de microfones). As simulações mostram também que o método apresenta melhores resultados do que considerando todas as fontes ativas em ambientes anecoicos. Em condição reverberante, DPC e DPCM apresentam desempenho similar.

Os métodos de separação propostos neste trabalho usam as ATFs entre as fontes e os microfones ou uma aproximação associada à posição espacial das fontes. As estimativas das ATFs em situações práticas nem sempre é possível e são difíceis de obter. Como alternativa para essa limitação, muitos métodos de separação utilizam as Funções de Transferência Relativa (RTFs), isto é, a relação de funções de transferência acústica entre os sensores, em vez das ATFs, porque podem ser obtidas a partir dos microfones. Foi demonstrado que os métodos propostos quando empregados utilizando as RTFs resultam na separação as fontes relativas a um microfone de referência, ou seja, são estimativas das fontes da forma como são amostradas pelo microfone de referência. As simulações demostraram que o uso das RTFs apresentou resultados equivalentes às ATFs.

Apesar de as RTFs poderem ser estimadas diretamente dos sinais observados nos microfones, sem conhecimento prévio da localização das fontes ou das características do ambiente, e serem uma alternativa viável para uso na separação de sinais com os métodos propostos, não se encontram trabalhos na literatura específicos para obtenção de RTFs para o caso de múltiplas fontes associadas ao caso subdeterminado. As abordagens exis-

tentes são mais focadas em descobrir as fontes dominantes e não explicitamente em obter as funções de transferência do processo de mixagem.

Duas propostas para estimativa das RTFs associadas a múltiplas fontes no caso subdeterminado foram desenvolvidas neste trabalho. Um dos métodos propostos é baseado em histogramas ponderados suavizados de estimativa da RTF instantânea e utiliza pesos determinados pelas estimativas das PSDs das fontes sendo denominado wFDUET (Weighted Frequency dependente DUET) O outro método proposto é baseado no agrupamento ponderado *Fuzzy C-Means* (wFCM). Neste método, enfatizam-se características significativas aplicando a distribuição Laplaciana, característica dos sinais de fala, para determinação de pesos, este método foi chamado wFCM-L (Weighted Fuzzy c-Means with Laplacian Distribution). Apesar da utilização das estimativas das PSDs como pesos resultarem em diminuição do erro de estimativa das RTFs. O método wFDUET não apresentou resultados satisfatórios quando utilizados para separação dos sinais. O método wFCM-L apresentou os menores erros nas estimativas de RTFs e com resultados, em termos de separação, semelhantes ao uso das RTFs verdadeiras.

Os trabalhos futuros relacionados à separação das fontes sonoras incluem estender os métodos para tratar o ruído difuso, avaliação do desempenho computacional, emprego de diferentes arranjos de microfones e diferentes *beamformers*. As estimativas das RTFs também apresentam diversas possibilidades de estudos como: a melhoria nos procedimentos de permutação, a avaliação de outros tipos de distribuição para determinação dos pesos e o emprego de diferentes métricas de distância no processo de aglomeração.

REFERÊNCIAS

ALGAZI, V. *et al.* The CIPIC HRTF database. *In*: IEEE WORKSHOP ON THE APPLICATIONS OF SIGNAL PROCESSING TO AUDIO AND ACOUSTICS, 2001, New Platz. **Proceedings [...]** Piscataway: IEEE, 2001. p.99–102.

ANDREOPOULOU, A.; BEGAULT, D. R.; KATZ, B. F. G. Inter-laboratory round robin HRTF measurement comparison. **IEEE Journal of Selected Topics in Signal Processing**, New York, v.9, n.5, p.895–906, Aug. 2015.

ARAKI, S.; SAWADA, H.; MAKINO, S. K-means based underdetermined blind speech separation. *In*: MAKINO, S.; SAWADA, H.; LEE, T.-W. (ed.). **Blind speech separation**. Dordrecht: Springer Netherlands, 2007. p.243–270.

ARAKI, S. *et al.* Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. **Signal Processing**, [S.l.], v.87, n.8, p.1833 – 1847, 2007.

ATCHESON, M. *et al.* On the use of contextual time-frequency information for full-band clustering-based convolutive blind source separation. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2014, Florence. **Proceedings [...]** New York: IEEE, 2014. p.2114–2118.

BAO, G. *et al.* A compressed sensing approach to blind separation of speech mixture based on a two-layer sparsity model. **IEEE Transactions on Audio, Speech, and Language Processing**, Piscataway, v.21, n.5, p.899–906, May 2013.

BARFUSS, H. *et al.* HRTF-based robust least-squares frequency-invariant beamforming. *In*: IEEE WORKSHOP ON APPLICATIONS OF SIGNAL PROCESSING TO AUDIO AND ACOUSTICS (WASPAA), 2015, New Paltz. **Proceedings [...]** Piscataway: IEEE, 2015. p.1–5.

BARTLETT, M. S. Smoothing periodograms from time-series with continuous spectra. **Nature**, London, v.161, n.4096, p.686–687, May 1948.

BENESTY, J.; CHEN, J.; HUANG, Y. **Microphone array aignal processing**. Berlin: Springer, 2008. (Springer Topics in Signal Processing, v.1).

BLIN, A.; ARAKI, S.; MAKINO, S. Blind source separation when speech signals outnumber sensors using a sparseness-mixing matrix estimation (SMME). *In*: INTERNATIONAL WORKSHOP ON ACOUSTIC ECHO AND NOISE CONTROL, 2003, Kyoto. **Proceedings [...]** [S.l.: s.n.], 2003. p.211–214.

BRANDSTEIN, M.; WARD, D. **Microphone arrays signal processing techniques and applications**. 1.ed. New York: Springer-Verlag, 2001.

BREEBART, J.; FALLER, C. **Spatial audio processing MPEG surround and other applications**. England: John Wiley & Sons Ltd, 2007.

CHEN, D.; PLEMMONS, R. J. Nonnegativity constraints in numerical analysis. *In*: BULTHEEL, A.; COOLS, R. (ed.). **The birth of numerical analysis**. Singapore: WORLD SCIENTIFIC, 2010. p.109–139.

CHENG, C. I.; WAKEFIELD, G. H. Introduction to head-related transfer functions (HRTFs): representations of hrtfs in time, frequency, and space. *In*: AUDIO ENGINEERING SOCIETY CONVENTION, 107., 1999, New York. **Proceedings [...]** New York: AES, 1999.

COHEN, I. Relative transfer function identification using speech signals. **IEEE Transactions on Speech and Audio Processing**, New York, v.12, n.5, p.451–459, Sept. 2004.

DELEFORGE, A.; GANNOT, S.; KELLERMANN, W. Towards a generalization of relative transfer functions to more than one source. *In*: EUROPEAN SIGNAL PROCESSING CONFERENCE (EUSIPCO), 23., 2015, Nice. **Proceedings [...]** New York: IEEE, 2015. p.419–423.

DIBIASE, J. H. **A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays**. 2000. Tese (Doutorado em Engenharia). Division of Engineering at Brown University, Providence, May. 2000.

DOCLO, S.; MOONEN, M. Superdirective beamforming robust against microphone mismatch. **IEEE Transactions on Audio, Speech, and Language Processing**, Piscataway, v.15, n.2, p.617–631, Feb. 2007.

DOCLO, S. *et al.* Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction. **Speech Communication**, [S.l.], v.49, n.7-8, p.636–656, 2007.

DOCLO, S. *et al.* Multichannel signal enhancement algorithms for assisted listening devices- Exploiting spatial diversity using multiple microphones. **IEEE Signal Processing Magazine**, New York, v.32, n.2, p.18–30, 2015.

EHRENBERG, L. *et al.* Sensitivity analysis of MVDR and MPDR beamformers. *In*: IEEE CONVENTION OF ELECTRICAL AND ELECTRONICS ENGINEERS IN ISRAEL, 26., 2010, Eliat. **Proceedings [...]** New York: IEEE, 2010. p.416–420.

FARINA, A. Simultaneous measurement of impulse response and distortion with a swept-sine technique. *In*: AUDIO ENGINEERING SOCIETY CONVENTION, 108., 2000, Paris. **Proceedings [...]** New York: AES, 2000.

FARINA, A. Advancements in impulse response measurements by sine sweeps. *In*: AUDIO ENGINEERING SOCIETY CONVENTION, 122., 2007, Vienna. **Proceedings [...]** New York: AES, 2007.

FENG, F.; KOWALSKI, M. Sparsity and low-rank amplitude based blind source separation. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2017, New Orleans. **Proceedings [...]** New York: IEEE, 2017. p.571–575.

FÉVOTTE, C.; GRIBONVAL, R.; VINCENT, E. **BSS_EVAL Toolbox for performance measurement in (blind) source separation**. 2007. 19 p. Disponível em: <http://www.irisa.fr/metiss/bss_eval>. Acesso em: 30 Jun. 2019.

GANNOT, S.; BURSHTEIN, D.; WEINSTEIN, E. Signal enhancement using beamforming and nonstationarity with applications to speech. **IEEE Transactions on Signal Processing**, New York, v.49, n.8, p.1614–1626, Aug. 2001.

GANNOT, S.; COHEN, I. Adaptive beamforming and postfiltering. *In*: BENESTY, J.; SONDHI, M. M.; HUANG, Y. A. (ed.). **Springer handbook of speech processing**. Heidelberg: Springer Berlin Heidelberg, 2008. p.945–978.

GANNOT, S. *et al.* A consolidated perspective on multimicrophone speech enhancement and source separation. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Piscataway, v.25, n.4, p.692–730, Apr. 2017.

GAZOR, S.; ZHANG, W. Speech probability distribution. **IEEE Signal Processing Letters**, New York, v.10, n.7, p.204–207, Jul. 2003.

GRIFFITHS, L.; JIM, C. An alternative approach to linearly constrained adaptive beamforming. **IEEE Transactions on Antennas and Propagation**, New York, v.30, n.1, p.27–34, Jan. 1982.

GÖßLING, N.; DOCLO, S. Relative transfer function estimation exploiting spatially separated microphones in a diffuse noise field. *In*: INTERNATIONAL WORKSHOP ON ACOUSTIC SIGNAL ENHANCEMENT (IWAENC), 16., 2018, Tokyo. **Proceedings [...]** New York: IEEE, 2018. p.146–150.

HABETS, E. *et al.* New insights into the MVDR beamformer in room acoustics. **IEEE Transactions on Audio, Speech, and Language Processing**, Piscataway, v.18, n.1, p.158–170, Jan. 2010.

HADAD, E.; DOCLO, S.; GANNOT, S. A generalized binaural MVDR beamformer with interferer relative transfer function preservation. *In*: EUROPEAN SIGNAL PROCESSING CONFERENCE (EUSIPCO), 24., 2016, Budapest. **Proceedings [...]** New York: IEEE, 2016. p.1643–1647.

HADAD, E.; GANNOT, S.; DOCLO, S. Binaural linearly constrained minimum variance beamformer for hearing aid applications. *In*: INTERNATIONAL WORKSHOP ON ACOUSTIC SIGNAL ENHANCEMENT (IWAENC), 2012, Aachen. **Proceedings [...]** Offenbach: VDE Verlag, 2012. p.1–4.

HAYKIN, S.; LIU, K. J. R. **Handbook on array processing and sensor networks**. New Jersey: John Wiley & Sons, Inc., 2009.

HIDRI, A.; MEDDEB, S.; AMIRI, H. About multichannel speech signal extraction and separation techniques. **Journal of Signal and Information Processing**, [S.l.], v.03, n.02, p.238–247, 2012.

HIOKA, Y.; NIWA, K. PSD estimation in beamspace for source separation in a diffuse noise field. *In*: INTERNATIONAL WORKSHOP ON ACOUSTIC SIGNAL ENHANCEMENT (IWAENC), 14., 2014, Juan-les-Pins. **Proceedings [...]** New York: IEEE, 2014. p.85–88.

HIOKA, Y. *et al.* Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain. **IEEE Transactions on Audio, Speech, and Language Processing**, Piscataway, v.21, n.6, p.1240–1250, 2013.

HOLLICK, J. *et al.* Source number estimation in reverberant conditions via full-band weighted, adaptive fuzzy c-means clustering. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2014, Florence. **Proceedings [...]** New York: IEEE, 2014. p.7450–7454.

HOLTERS, M.; CORBACH, T.; ZÖLZER, U. Impulse response measurement techniques and their applicability in the real world. *In*: INTERNATIONAL CONFERENCE ON DIGITAL AUDIO EFFECTS DAFX-09, 12., 2009, Como. **Proceedings [...]** [S.l.: s.n.], 2009. p.108–112.

HU, Y.; LOIZOU, P. C. Evaluation of objective quality measures for speech enhancement. **IEEE Transactions on Audio, Speech, and Language Processing**, Piscataway, v.16, n.1, p.229–238, Jan. 2008.

ITO, N.; ARAKI, S.; NAKATANI, T. Permutation-free clustering of relative transfer function features for blind source separation. *In*: EUROPEAN SIGNAL PROCESSING CONFERENCE (EUSIPCO), 23., 2015, Nice. **Proceedings [...]** New York: IEEE, 2015. p.409–413.

ITU. **P.862 - Revised Annex A - Reference implementations and conformance testing for ITU-T Recs P.862, P.862.1 and P.862.2**. 2006. Disponível em: <https://www.itu.int/rec/T-REC-P.862-200511-I!Amd2/en>. Acesso em: 10 May 2019.

JAFARI, I. *et al.* Underdetermined blind source separation with fuzzy clustering for arbitrarily arranged sensors. *In*: ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION, INTERSPEECH, 2011, Florence. **Proceedings [...]** [S.l.: s.n.], 2011. p.1753–1756.

JAFARI, I. *et al.* Evaluations on underdetermined blind source separation in adverse environments using time-frequency masking. **EURASIP Journal on Advances in Signal Processing**, Heidelberg, v.2013, n.1, p.162, Oct. 2013.

JODER, C. *et al.* Real-time speech separation by semi-supervised nonnegative matrix factorization. *In*: INTERNATIONAL CONFERENCE ON LATENT VARIABLE ANALYSIS AND SIGNAL SEPARATION, LVA/ICA, 10., 2012, Berlin. **Proceedings [...]** Springer Berlin Heidelberg, 2012. p.322–329.

KELLERMANN, W.; BUCHNER, H. Wideband algorithms versus narrowband algorithms for adaptive filtering in the DFT domain. *In*: ASILOMAR CONFERENCE ON SIGNALS, SYSTEMS COMPUTERS, 37., 2003, Pacific Grove. **Proceedings [...]** New York: IEEE, 2003. v.2, p.1278–128.

KOWALSKI, M.; VINCENT, E.; GRIBONVAL, R. Beyond the narrowband approximation: wideband convex methods for under-determined reverberant audio source separation. **IEEE Transactions on Audio, Speech, and Language Processing**, Piscataway, v.18, n.7, p.1818–1829, Sept. 2010.

KUKLASINSKI, A.; JENSEN, J. Multichannel wiener filters in binaural and bilateral hearing aids — speech intelligibility improvement and robustness to DOA errors. **Journal of the Audio Engineering Society**, New York, v.65, n.1/2, p.8–16, 2017.

KÜHNE, M.; TOGNERI, R.; NORDHOLM, S. Robust source localization in reverberant environments based on weighted fuzzy clustering. **IEEE Signal Processing Letters**, New York, v.16, n.2, p.85–85, Feb. 2009.

KÜHNE, M.; TOGNERI, R.; NORDHOLM, S. A novel fuzzy clustering algorithm using observation weighting and context information for reverberant blind speech separation. **Signal Processing**, [S.l.], v.90, n.2, p.653 – 669, 2010.

LAI, C. C.; NORDHOLM, S. E.; LEUNG, Y. H. **A study into the design of steerable microphone arrays**. Singapore: Springer Singapore, 2017. (SpringerBriefs in Electrical and Computer Engineering).

LAWSON, C. L.; HANSON, R. J. **Solving least-squares problems**. Upper Saddle River: Prentice Hall, 1974.

LEHMANN, E. A. **Image-source model MATLAB code implementation**. 2012. Disponível em: <http://www.eric-lehmann.com>. Acesso em: 16 May 2019.

LEHMANN, E. A.; JOHANSSON, A. M. Prediction of energy decay in room impulse responses simulated with an image-source model. **The Journal of the Acoustical Society of America**, Woodbury, v.124, n.1, p.269–277, 2008.

LI, X. *et al.* Non-stationary noise power spectral density estimation based on regional statistics. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2016, Shanghai. **Proceedings [...]** New York: IEEE, 2016. p.181–185.

LUFT, J. A.; PEREIRA, F. I.; SUSIN, A. Binaural sound source separation based on directional power spectral densities. *In*: AUDIO ENGINEERING SOCIETY CONVENTION, 143., 2017, New York. **Proceedings [...]** New York: AES, 2017.

MAAZAOUI, M.; GRENIER, Y.; ABED-MERAIM, K. From binaural to multimicrophone blind source separation using fixed beamforming with HRTFs. *In*: INTERNATIONAL CONFERENCE ON SYSTEMS, SIGNALS AND IMAGE PROCESSING (IWSSIP), 19., 2012, Vienna. **Proceedings [...]** New York: IEEE, 2012. p.480–483.

MABANDE, E.; SCHAD, A.; KELLERMANN, W. Design of robust superdirective beamformers as a convex optimization problem. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 2009, Taipei. **Proceedings [...]** New York: IEEE, 2009. p.77–80.

MAKINO, S.; SAWADA, H.; LEE, T.-W. (ed.). **Blind speech separation**. Dordrecht: Springer, 2007.

MARKOVICH, S.; GANNOT, S.; COHEN, I. Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. **IEEE**

Transactions on Audio, Speech, and Language Processing, Piscataway, v.17, n.6, p.1071–1086, Aug. 2009.

MARTIN, R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. **IEEE Transactions on Speech and Audio Processing**, New York, v.9, n.5, p.504–512, Jul. 2001.

MIYAMOTO, S.; INOKUCHI, R.; KURODA, Y. Possibilistic and fuzzy c-means clustering with weighted objects. *In*: IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS, 2006, Vancouver. **Proceedings [...]** Piscataway: IEEE, 2006. p.869–874.

MOORE, A.; PARADA, P. P.; NAYLOR, P. "Speech enhancement for robust automatic speech recognition: evaluation using a baseline system and instrumental measures. **Computer Speech & Language**, [S.l.], v.46, p.574 – 584, 2017.

NIWA, K.; HIOKA, Y.; KOBAYASHI, K. Microphone array source enhancement using subtractive PSD estimation model. **Applied Acoustics**, [S.l.], v.143, p.239 – 249, 2019.

OPPENHEIM, A. V.; SCHAFER, R. W. **Discrete-time signal processing**. 3rd.ed. Upper Saddle River: Prentice Hall Press, 2009.

PANAYOTOV, V. *et al.* Librispeech: an asr corpus based on public domain audio books. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2015, Brisbane. **Proceedings [...]** New York: IEEE, 2015. p.5206–5210.

PAPOULIS, A.; PILLAI, S. U. **Probability, random variables, and stochastic processes**. 4.ed. New York: McGrawHill, 2002.

RABINER, L. R.; SCHAFER, R. W. **Digital processing of speech signals**. New Jersey: Prentice-Hall, 1978. (Prentice-Hall signal processing series).

RICKARD, S.; YILMAZ, O. On the approximate W-disjoint orthogonality of speech. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 2002, Orlando. **Proceedings [...]** New York: IEEE, 2002. v.1, p.I–529–I–532.

RIX, A. W. *et al.* Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 2001, Salt Lake City. **Proceedings [...]** New York: IEEE, 2001. v.2, p.749–752.

SAWADA, H.; ARAKI, S.; MAKINO, S. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. **IEEE Transactions on Audio, Speech, and Language Processing**, Piscataway, v.19, n.3, p.516–527, Mar. 2011.

SAWADA, H. *et al.* A robust approach to the permutation problem of frequency-domain blind source separation. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. (ICASSP), 2003, Hong Kong. **Proceedings [...]** New York: IEEE, 2003. v.5, p.V–381.

SERVETTI, A.; MARTIN, J. C. D. Error tolerant MAC extension for speech communications over 802.11 WLANs. *In*: IEEE VEHICULAR TECHNOLOGY CONFERENCE, 61., 2005, Stockholm. **Proceedings [...]** New York: IEEE, 2005. v.4, p.2330–2334.

SHYNK, J. Frequency-domain and multirate adaptive filtering. **IEEE Signal Processing Magazine**, New York, v.9, n.1, p.14–37, Jan. 1992.

TAAL, C. H. **STOI – short-time objective intelligibility MATLAB code implementation**. 2011. Disponível em: <http://www.ceestaal.nl/code/>. Acesso em: 16 May 2019.

TAAL, C. H. *et al.* An algorithm for intelligibility prediction of time–frequency weighted noisy speech. **IEEE Transactions on Audio, Speech, and Language Processing**, Piscataway, v.19, n.7, p.2125–2136, Sept. 2011.

TAN, V. Y. F. A study of the effect of source sparsity for various transforms on blind audio source separation performance. *In*: WORKSHOP ON SIGNAL PROCESSING WITH ADAPTIVE SPARSE STRUCTURED REPRESENTATIONS (SPARS'05), 2005, Rennes, France. **Proceedings [...]** [S.l.: s.n.], 2005.

TASESKA, M.; HABETS, E. A. P. Relative transfer function estimation exploiting instantaneous signals and the signal subspace. *In*: EUROPEAN SIGNAL PROCESSING CONFERENCE (EUSIPCO), 23., 2015, Nice. **Proceedings [...]** New York: IEEE, 2015. p.404–408.

VAN TREES, H. L. **Optimum array processing part IV of detection, estimation, and modulation theory**. New York: Jonh Wiley & Sons, Inc, 2002.

VARZANDEH, R.; TASESKA, M.; HABETS, E. A. P. An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation. *In*: HANDS-FREE SPEECH COMMUNICATIONS AND MICROPHONE ARRAYS (HSCMA), 2017, San Francisco. **Proceedings [...]** New York: IEEE, 2017. p.11–15.

VINCENT, E.; GRIBONVAL, R.; FEVOTTE, C. Performance measurement in blind audio source separation. **IEEE Transactions on Audio, Speech and Language Processing**, Piscataway, v.14, n.4, p.1462–1469, Jul. 2006.

VINCENT, E. *et al.* From blind to guided audio source separation: how models and side information can improve the separation of sound. **IEEE Signal Processing Magazine**, New York, v.31, n.3, p.107–115, May 2014.

WANG, D.; CHEN, J. Supervised Speech Separation Based on Deep Learning: an overview. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Piscataway, v.26, n.10, p.1702–1726, Oct. 2018.

WELCH, P. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. **IEEE Transactions on Audio and Electroacoustics**, New York, v.15, n.2, p.70–73, Jun. 1967.

WINTER, S. *et al.* Overcomplete bss for convolutive mixtures based on hierarchical clustering. *In*: PUNTONET, C. G.; PRIETO, A. (ed.). **Independent component analysis and blind signal separation**. Heidelberg: Springer Berlin Heidelberg, 2004. p.652–660.

YILMAZ, O.; RICKARD, S. Blind separation of speech mixtures via time-frequency masking. **IEEE Transactions on Signal Processing**, New York, v.52, n.7, p.1830–1847, Jul. 2004.

YOSHII, K. *et al.* Beyond nmf: time-domain audio source separation without phase reconstruction. *In*: CONFERENCE OF THE INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL (ISMIR), 14., 2013, Curitiba. **Proceedings [...]** [S.l.: s.n.], 2013. n.14, p.369–374.

ZOHOURIAN, M.; MARTIN, R. Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2016, Shanghai. **Proceedings [...]** New York: IEEE, 2016. v.2016-May, n.1, p.430–434.

APÊNDICE A: AES 2017 *CONVENTION PAPER*

Audio Engineering Society

Convention Paper

Presented at the 143rd Convention
2017 October 18–21, New York, NY, USA

Binaural sound source separation based on directional power spectral densities

Joel Augusto Luft[1,2], Fabio Irigon Pereira[2], and Altamiro Amadeu Susin[2]

[1] *Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul, Canoas, RS, Brazil*
[2] *Universidade Federal do Rio Grande do Sul, DELET, PPGEE, Porto Alegre, RS, Brazil*

Correspondence should be addressed to Joel Augusto Luft (joel.luft@ufrgs.br)

ABSTRACT

Microphone arrays are a common choice to be used in spatial sound source separation. In this paper, a new method for binaural source separation is presented. The separation is performed using the spatial position of sound source, the Head-Related Transfer Function and the Power Spectral Density of fixed beamformers. A non-negative constrained least-squares minimization approach is used to solve the Head-Related Transfer Function based directivity gain formulation and the Power Spectral Density is used as a magnitude estimation of the sound sources. Simulation examples are presented to demonstrate the performance of the proposed algorithm.

1 Introduction

The separation of audio signals is a complex problem and there are no definitive solutions to this. Basically two aspects characterize the problem: blind / non-blind, characterized by knowledge of the process of signal mixing or determinate / indeterminate that is related to the number of sound sources and microphones available [1]. The separation or enhancement of sound sources from one or more sound streams remains an important signal processing problem, involved in several different applications, such as: audio and video communication systems, voice recognition, audio recording, assistance to the visually impaired and separation of signals in hearing aids.

The main goal of source separation is to estimate the sound sources from the sound arriving at the microphones. There are several approaches to the problem of separation or enhancement of audio signals. The conventional beamformers that use fixed spatial filtering to obtain signals use signals position as a filter parameter, this solution is deterministic but have limited separation performance, the main example is Delay-and-Sum Beamforming (DS) [2].

Another class of conventional techniques known as Linear Constraint Minimum Variance (LCMV) beamformers uses statistic behavior of noise/interference to enhance the desired signals while reject the interfering signals as in Minimum Variance Distortionless Response (MVDR) [3] or adaptive spatial filtering methods such as Generalized Sidelobe Canceller (GSC) [4], these techniques require good spatial information of sound sources or voice activity detectors to identify when the desired signals are active.

The separation can also be estimated without previous knowledge of the sources or their positions as in Independent Component Analysis (ICA) where the number of sound sources is restricted to the number of microphones used [2]. In another approach the separation

is done by Non-negative Matrix Factorization (NMF) techniques with no limitation in the number of sources but requires training or off-line processing, where the whole signal is known [5][6].

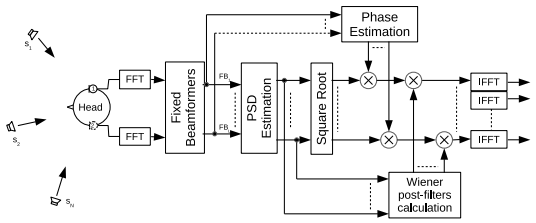


Fig. 1: Flow diagram of proposed method.

This work treats the problem of separation when the spatial position of sound source is involved. The transfer functions from sound sources to the microphones are assumed to be known. The sound sources are assumed to be positioned around the microphones. In the case of binaural source separation, the transfer function from source to microphone is represented by the Head-Related Transfer Function (HRTF). The HRTF is defined as a free-field transfer function from a point in a space to a point in the listener's outer ear, normally the entrance to the ear canal [7]. Due to the features of a person like size and shape of the head, the outer ear and the torso the HRTF varies for each ear, person, direction and distance of the source of sound. The HRTF is the Fourier Transform of the Head-Related Impulse Response (HRIR) which can be obtained experimentally generating a sound in a specific spatial position relative to a head, acquiring the sound that arrives in microphones located at the ears and calculating the transfer function. It can be done directly to a person or to a mannequin.

The microphone arrays continue to be a good alternative to separate sound when the spatial features are involved since they have different responses according to direction or position of sound source location. Hioka [8] proposed a method to separate sound sources based on Power Spectral Density (PSD) estimation using the directivity of beamformers to estimate the PSD of each source and using a wiener post-filter to obtain the signal source.

This paper presents a new method, as illustrated in Fig. 1. It starts with the approach of Hioka [8] and applies a combination of non-negative least-squares constraints to obtain PSD estimations and use the estimated PSD to reconstruct the signal instead of filtering the mixed signal. The proposed method is able to achieve a higher separation and exceeds the maximum limit of separable sources presented by Hioka.

The paper is structured as follows. Section 2 presents the current approaches to solve the problem of Sound Source Separation. The method proposed by Hioka [8] is described in more details and presents a new approach based only on nonnegative values of the PSD. Simulation results are presented in section 3 and the paper finishes with conclusions and further work in section 4.

2 Methods

2.1 Problem formulation

Assuming N different sound sources whose signals are captured by M microphones. Each source n is located at any direction and at a certain distance $d_{n,m}$ from each microphone m. The signal of the m-th microphone in frequency domain can be expressed as

$$X_m(\omega) = \sum_{n=1}^{N} A_{n,m}(\omega) S_n(\omega) \tag{1}$$

where w is the frequency, S_n is the n-th sound source and $A_{n,m}(\omega)$ is the transfer function from source n to microphone m. The transfer function includes the environment characteristics and microphone response.

The problem consists in estimating the PSDs of each source S_n based on measured PSD of observed signals $X_m(\omega)$. This will be made based on the assumption that the sources directions are known and that the sound sources are mutually uncorrelated resulting

$$E[S_n(\omega) S_{n'}(\omega)] = 0 \quad \text{if } n \neq n' \tag{2}$$

where $E[.]$ is the expectation operator.

2.2 Signal Estimation Using a Combination of Beamformers Gains

In this section the method for sound source separation proposed by Hioka [8] is presented in more detail. The proposed solution for separation of sound sources is based on estimating the PSDs of the sources in the directions of interest and calculates a filter to obtain the selected signal. The number of possible sources

to be separated by this method is $M(M-1)+1$ where M is the number of microphones. Having defined the number of sources and their directions it is possible to estimate the powers in each direction. The directions of sound sources can be obtained by techniques of estimation of the direction of arrival [9]. Assuming N the number of signal sources we should determine L fixed beamformers ($L \geq N$) targeted in L different directions. In this way, the output of each beamformer l is given by

$$Y_l(\omega) = \sum_{m=1}^{M} W_{l,m}(\omega) X_m(\omega) \tag{3}$$

where $W_{l,m}(\omega)$ is the filter coefficient of microphone m for beamformer l. The beamformer output can be rewrite substituting (1) in (3) as

$$Y_l(\omega) = \sum_{m=1}^{M} \sum_{n=1}^{N} W_{l,m}(\omega) A_{n,m}(\omega) S_n(\omega) \tag{4}$$

$$Y_l(\omega) = \sum_{n=1}^{N} D_{l,n}(\omega) S_n(\omega) \tag{5}$$

where $D_{l,n}(\omega)$ is the direction function of beamformer l to the direction of source n. The PSD of beamformer output target in the direction l is given by

$$\phi_{Yl}(\omega) = E[Y_l(\omega) Y_l(\omega)^*] \tag{6}$$

Assuming that the sources are mutually uncorrelated the PSD of each beamformer output can be approximated [8] by

$$\phi_{Y_l}(\omega) = \sum_{n=1}^{N} |D_{l,n}(\omega)|^2 \phi_{S_n}(\omega) \tag{7}$$

Choosing at least N different beamformers it is possible from (7) to estimate PSD of each source $\phi_{S_n}(\omega)$ solving a set of simultaneous L equations

$$\begin{bmatrix} \phi_{Y_1} \\ \phi_{Y_2} \\ \vdots \\ \phi_{Y_L} \end{bmatrix} = \begin{bmatrix} |D_{1,1}|^2 & |D_{1,2}|^2 & \cdots & |D_{1,N}|^2 \\ |D_{2,1}|^2 & |D_{2,2}|^2 & \cdots & |D_{2,N}|^2 \\ \vdots & \vdots & \ddots & \vdots \\ |D_{L,1}|^2 & |D_{L,2}|^2 & \cdots & |D_{L,N}|^2 \end{bmatrix} \begin{bmatrix} \phi_{S_1} \\ \phi_{S_2} \\ \vdots \\ \phi_{S_N} \end{bmatrix} \tag{8}$$

The variable w was omitted but the equation above (7) must be solved for each frequency bin. The equation (8) in matrix form is

$$\Phi_Y(\omega) = D(\omega) \Phi_S(\omega) \tag{9}$$

Current solution proposed by Hioka solves the equation (9) by

$$\Phi_S(\omega) = D^{-1}(\omega) \Phi_Y(\omega) \tag{10}$$

being that $D^{-1}(\omega)$ is the inverse of the square matrix $D(\omega)$ if $L = N$ or is the Moore-Penrose pseudo-inverse matrix if $L > N$. This method may result in unrealistic negative values to source PSD. Considering that PSD may not be negative they are substituted by their absolute values. After estimating the PSD for each angle the output signal to the desired direction n can be obtained by applying the Wiener post-filter to the output of the beamformer.

$$H_n(\omega) = \frac{\phi_{S_n}(\omega)}{\sum_{n=1}^{N} \phi_{S_n}(\omega)} \tag{11}$$

$$\hat{S}_n(\omega) = H_n(\omega) Y(\omega) \tag{12}$$

Then, the sound source signal is obtained by applying the inverse Fourier Transform to the filtered source.

This method presents some problems related to conditioning of matrix $D(\omega)$. Hioka demonstrated that if $N > M(M-1)+1$ $D(\omega)$ is rank deficient and in this case the method fails because it is ill-conditioned, the simultaneous equations becomes indefinite, which may give an unstable estimate. This problem also appears in cases of: inappropriate microphone array geometry due to certain symmetries in the microphone distribution, low frequencies where the shape of directivity gain is too flat and spatial alias where the gain is equal at different frequencies. To overcome the limitations a new method to solve the problem is presented in section 2.4.

2.3 Fixed beamforming with HRTFs

In the case of binaural signal separation the microphones are fixed in a head usually located in-the-ear or behind-the-ear. The Head Related Transfer Functions (HRTFs) or the transfer function from sound source at l-th direction to the M microphones is defined as

$$\mathbf{a_l}(\omega) = [a_{l,1}(\omega), ..., a_{l,M}(\omega)]^T \tag{13}$$

To apply the separation method is necessary at least N different beamformers to the directions of the sources. Directive beamformer can be obtained by least square techniques [2] and the coefficients $W_{l,m}(\omega)$ of equation (3) must be determined to obtain the desired response. In this case we used a conventional super directive beamformer solution[10] [11] where a unit gain in the direction of the *l-th* source is desired and the gain to other directions is minimized. The filter coefficients are obtained by

$$W_l(\omega) = conj(\frac{(R_{\mathbf{aa}} + \mu I)^{-1}\mathbf{a}_l(\omega)}{\mathbf{a}_l(\omega)^H(R_{\mathbf{aa}} + \mu I)^{-1}\mathbf{a}_l(\omega)}) \quad (14)$$

where, $R_{\mathbf{aa}} = \frac{1}{L}\sum_1^L \mathbf{a}_l(\omega)\mathbf{a}_l^H(\omega)$ and the weighting parameter μ is to control the robustness against noise and location errors. The larger μ, the larger the robustness of the beamformer, but the smaller its directivity.

2.4 Nonnegative Least Square Directional Estimation of PSD

In this section a new method for sound source separation is presented, which is based on Nonnegative Least Square technique applied to PSD estimation proposed by Hioka [8]. The new method overcomes the problems presented in Hioka proposal. We propose to use a method to solve the system restricting the solution to be only nonnegative values. In this approach the formulated problem in (9) is solved minimizing with criterion of least squares and with restriction of nonnegative solution.

$$\min_{\Phi_S} \|D\Phi_S - \Phi_Y\|^2 \quad \text{subject to } \Phi_S \geq 0 \quad (15)$$

where $\|.\|$ denotes the L_2 norm operator. In this paper the least square problem is solved using the method proposed in [12]. Some alternative algorithms can be found in [13]. Since we can have an under-determined problem ($N > M$) or ill-conditioned situations we may have different solutions causing numerical instabilities. Applying the nonnegative constraint and least square minimization avoid unrealistic negative solutions what is not contemplated by Hioka.

The second proposal is to obtain the source signal form the source PSD estimation instead of using the filtered fixed beamformer output. We use the square root of power density as the source magnitude and the phase of fixed beamformer output as the source phase.

$$Ph_n(\omega) = \arg[Y_n(\omega)] \quad (16)$$

where Ph_n is the phase of the fixed beamformer (3) target to the direction of source n. The signal source is obtained multiplying this phase to the square root of PSD

$$\hat{S}'_n(\omega) = Ph_n(\omega) \times \sqrt{\phi_{S_n}(\omega)} \quad (17)$$

where $\phi_{S_n}(\omega)$ is the PSD source estimation from the new method. This new signal $\hat{S}'_n(\omega)$ is then filtered by the Wiener post-filter (11) calculated with the nonnegative estimated PSD.

$$\hat{S}_n(\omega) = HN_n(\omega)\hat{S}'_n(\omega) \quad (18)$$

Finally, the inverse Fourier Transform is applied to $\hat{S}_n(\omega)$.

3 Results and discussion

The proposal was evaluated by computer simulation. The HRTF are sampled at CIPIC database [14] and Acoustics Research Institute (ARI) Behind-The-Ear HRTF database [15]. The direction of arrival of the target signal was assumed to be known for the experiments, i.e., no localization algorithm was applied. Simulation was made considering an ideal anechoic environment. In all simulations the microphones and sound sources were placed at the same plane. The directions of sound sources are presented as an angle of incidence θ. The simulation was evaluated for a period of 3 seconds and the audio signals are sampled at 16kHz. The methods were applied in frames of 512 samples and frame shift of 256 samples. Signals S4 and S5 are band limited to 4kHz.

The metrics used to evaluate the performance was Signal-to-Interference Ratio (SIR) improvement and Signal-to-Distortion Ratio (SDR) improvement [16] measured within the MATLAB toolbox named BSS EVAL distributed online under the GNU Public License [17]. The improvement in SIR and SDR for *n-th* source are defined as

$$\Delta SIR_n = SIR_{o,n} - SIR_{i,n} \quad (19)$$

$$\Delta SDR_n = SDR_{o,n} - SDR_{i,n} \quad (20)$$

where $SIR_{i,n}$ and $SDR_{i,n}$ are the measurements at the microphones and $SIR_{o,n}$ and $SDR_{o,n}$ are at the output of the system. The measurement used signals with wide range band from 200 Hz to 7000 Hz.

In the simulations the number of microphones was $M = 2$ and the separation methods were simulated by varying the number of sources N from 2 to 8. For each simulation the sources were distributed uniformly in front of the array of microphones forming a semicircle of radius 1.2 m. The angle $\theta = 0$ Degrees represent the position in front of the head. Table 1 shows the angles of the sound sources for each number of sources used and the sound sources used are described in table 2. They are used in the order presented on table, for N sources, S1 to SN are used. The RMS value of all sources are the same.

Table 1: Angles of sources

N	Angles of sources θ (degrees)
2	30,-30
3	45,0,-45
4	54,18,-18,54
5	60,30,0,-30,-60
6	64.3,36.6,12.8-12.8,-36.6,-64.3
7	67.5,45,22.5,0,-22.5,-45,-67.5
8	70,50,30,10,-10,-30,-50,-70

Table 2: Sources description

S_n	Description	S_n	Description
S1	male speech	S5	female speech(4k)
S2	female speech	S6	music (male vocal)
S3	music (rock)	S7	music (guitar)
S4	male speech(4k)	S8	street sound

In the simulations a white Gaussian isotropic noise was added as background noise to each microphone signal. The proposed method was tested in different background noise levels. The figure 2 shows the average values of ΔSDR e ΔSIR for $N = 3$ using ARI-BTE HRTF database. This test shows that a SNR greater then 20dB does not significantly degrades the performance.

The proposed method is compared to the fixed beamformer, GSC and to the method proposed by Hioka [8]. The fixed beamformer is determined from equation (14) and using μ set to $min(diag(R_{aa}))$. In the simulations

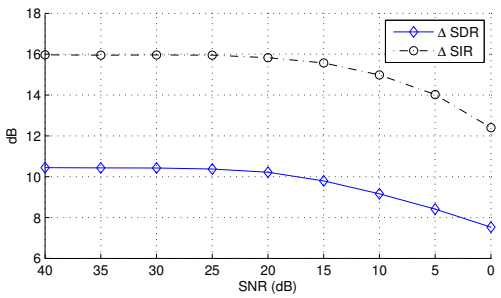


Fig. 2: Average SIR and SDR improvements varying the background noise for $N = 3$ (ARI-BTE HRTF database)

the GSC was implemented in the frequency domain and processed with frame size of 512 samples. The simulations were evaluated with a background noise equivalent to a 20dB SNR.

All signal sources were separated and the values of ΔSIR and ΔSDR were measured according to (19)(20). The simulation results are presented in Figures 3, 4, 5 and 6. Each point represents the average value of ΔSIR and ΔSDR of the N sound sources used in the simulation, as described by equations (21) and (22), where ΔSIR_l and ΔSDR_l are the values of the separated signal source l.

$$\Delta SIR(N) = \frac{1}{N}\sum_{l=1}^{N} \Delta SIR_l \tag{21}$$

$$\Delta SDR(N) = \frac{1}{N}\sum_{l=1}^{N} \Delta SDR_l \tag{22}$$

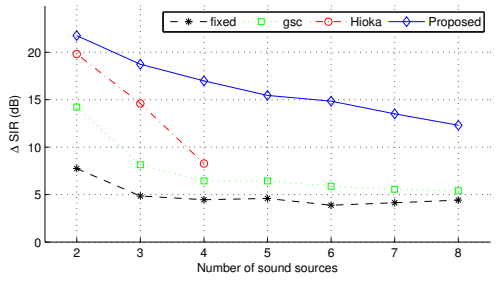


Fig. 3: Average SIR improvement (CIPIC HRTF database)

The results presented in previous section show that the proposed method outperforms the other methods

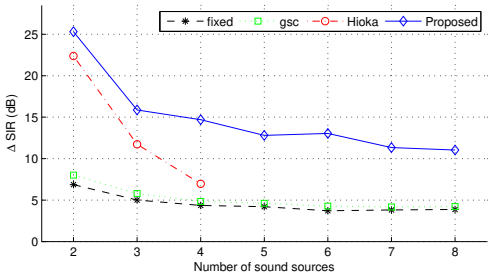


Fig. 4: Average SIR improvement (ARI-BTE HRTF database)

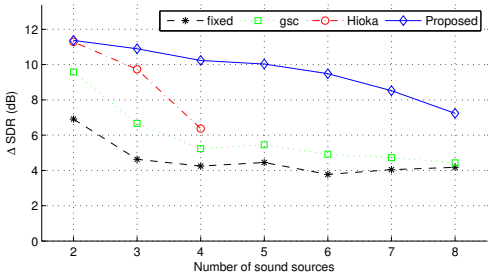


Fig. 5: Average SDR improvement (CIPIC HRTF database)

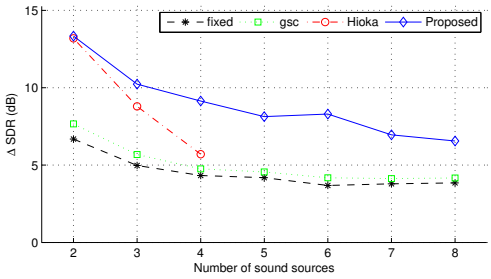


Fig. 6: Average SDR improvement (ARI-BTE HRTF database)

in terms of SIR improvement for all combination of sound source simulated. The proposed solution present superior and stable separation performance compares to the others methods even when the Hioka solution fails. To the case of two microphones Hioka demonstrate that your method is capable to separate only three sources. The proposed method still work for all quantity of sound sources tested.

4 Summary

We present a method to separate sound sources in a binaural condition using the directionality presented by the HRTFs. A Nonnegative Least Square Directional Estimation of PSD method was presented and compared to state-of-the-art approaches for under-determined sound source separation. The proposed method has shown better separation results than current approaches in all simulations. It still works even when the number of sources is bigger than the maximum theoretical value presented by Hioka [8]. Future work includes the evaluation of alternative minimization methods, strategies to improve phase recovery, analysis of influence of variation in HRTFs, variations in directions of arrival and evaluation in reverberant environments.

References

[1] Vincent, E., Bertin, N., Gribonval, R., and Bimbot, F., "From Blind to Guided Audio Source Separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, 31(3), pp. 107–115, 2014, ISSN 1053-5888, doi:10.1109/MSP.2013.2297440.

[2] Benesty, J., Chen, J., and Huang, Y., *Microphone array Signal Processing*, Springer, Berlin, Germany, 2008.

[3] Habets, E. a. P., Benesty, J., Cohen, I., Gannot, S., and Dmochowski, J., "New Insights Into the MVDR Beamformer in Room Acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1), pp. 158–170, 2010, ISSN 1558-7916, doi:10.1109/TASL.2009.2024731.

[4] Griffiths, L. and Jim, C., "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, 30(1), pp. 27–34, 1982, ISSN 0096-1973, doi:10.1109/TAP.1982.1142739.

[5] Yoshii, K., Tomioka, R., Mochihashi, D., and Goto, M., "Beyond NMF: Time-Domain Audio Source Separation without Phase Reconstruction," in *ISMIR*, 2013.

[6] Joder, C., Weninger, F., Eyben, F., Virette, D., and Schuller, B., *Real-Time Speech Separation*

by Semi-supervised Nonnegative Matrix Factorization, pp. 322–329, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, ISBN 978-3-642-28551-6, doi:10.1007/978-3-642-28551-6_40.

[7] Cheng, C. I. and Wakefield, G. H., "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space," in *Audio Engineering Society Convention 107*, 1999.

[8] Hioka, Y., Furuya, K., Kobayashi, K., Niwa, K., and Haneda, Y., "Underdetermined Sound Source Separation Using Power Spectrum Density Estimated by Combination of Directivity Gain," *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6), pp. 1240–1250, 2013, ISSN 1558-7916, doi:10.1109/TASL.2013.2248715.

[9] Brandstein, M. and Ward, D., *Microphone Arrays Signal Processing Techniques and Applications, 1st ed.*, Springer, New York, 2001.

[10] Doclo, S. and Moonen, M., "Superdirective Beamforming Robust Against Microphone Mismatch," *IEEE Transactions on Audio, Speech and Language Processing*, 15(2), pp. 617–631, 2007, ISSN 1558-7916, doi:10.1109/TASL.2006.881676.

[11] Maazaoui, M., Grenier, Y., and Abed-Meraim, K., "From binaural to multimicrophone blind source separation using fixed beamforming with HRTFs," in *2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 480–483, 2012, ISSN 2157-8672.

[12] Lawson, C. L. and Hanson, R. J., *Solving Least-Squares Problems*, Prentice Hall, Upper Saddle River, NJ, 1974.

[13] Chen, D. and Plemmons, R. J., "Nonnegativity constraints in numerical analysis," in *The Birth of Numerical Analysis*, pp. 109–139, WORLD SCIENTIFIC, 2009, ISBN 9789812836250, doi:10.1142/9789812836267_0008.

[14] Algazi, V., Duda, R., Thompson, D., and Avendano, C., "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, October, pp. 99–102, IEEE, 2001, ISBN 0-7803-7126-7, ISSN 10164383, doi:10.1109/ASPAA.2001.969552.

[15] "Acoustics Research Institute, Vienna, Austria," 2017, available in: https://www.sofaconventions.org.

[16] Vincent, E., Gribonval, R., and Fevotte, C., "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, 14(4), pp. 1462–1469, 2006, ISSN 1558-7916, doi:10.1109/TSA.2005.858005.

[17] Févotte, C., Gribonval, R., and Vincent, E., "BSS_EVAL Toolbox User Guide – Revision 2.0," Technical report, IRISA, 2005.