

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Escola de Engenharia
Programa de Pós-Graduação em Engenharia Elétrica

Guilherme Márcio de Melo Campos Fonte Boa

**UM ESTUDO SOBRE A SEPARAÇÃO CEGA DA FALA EM
SALAS REVERBERANTES**

Belo Horizonte
2025

Guilherme Márcio de Melo Campos Fonte Bôa

**UM ESTUDO SOBRE A SEPARAÇÃO CEGA DA FALA EM
SALAS REVERBERANTES**

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Engenharia Elétrica.

Linha de pesquisa: Inteligência Computacional.

Orientador: Prof. Dr. Adriano Vilela Barbosa

Coorientador: Prof. Dr. Hani Camille Yehia

Belo Horizonte
2025

F682e

Fonte Bôa, Guilherme Márcio de Melo Campos.

Um estudo sobre a separação cega da fala em salas reverberantes
[recurso eletrônico] / Guilherme Márcio de Melo Campos Fonte Bôa. -
2025.

1 recurso online (128 f. : il., color.) : pdf.

Orientador: Adriano Vilela Barbosa.

Coorientador: Hani Camille Yehia.

Dissertação (mestrado) - Universidade Federal de Minas Gerais,
Escola de Engenharia.

Inclui bibliografia.

1. Engenharia elétrica - Teses. 2. Inteligência computacional - Teses.
3. Fala - Teses. 4. Microfone - Teses. I. Barbosa, Adriano Vilela. II. Yehia,
Hani Camille. III. Universidade Federal de Minas Gerais. Escola de
Engenharia. IV. Título.

CDU: 621.3(043)



UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENGENHARIA
COLEGIADO DO CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

"Um Estudo sobre a Separação Cega da Fala em Salas Reverberantes"

Guilherme Márcio de Melo Campos Fonte Bôa

Dissertação de Mestrado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do grau de Mestre em Engenharia Elétrica.

Aprovada em 31 de outubro de 2025.

Por:

Prof. Dr. Adriano Vilela Barbosa
DELT (UFMG) - Orientador

Prof. Dr. Hani Camille Yehia
DELT (UFMG)

Prof. Ph.D. Eduardo Mazoni Andrade Marçal Mendes
DELT (UFMG)

Prof. Dr. Leandro Freitas de Abreu
DELT (UFMG)

Prof. Dr. Maurilio Nunes Vieira
DELT (aposentado) (UFMG)



Documento assinado eletronicamente por **Adriano Vilela Barbosa, Professor do Magistério Superior**, em 31/10/2025, às 17:24, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).

Documento assinado eletronicamente por **Leandro Freitas de Abreu, Professor do**



Magistério Superior, em 31/10/2025, às 17:25, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Eduardo Mazoni Andrade Marcal Mendes, Professor do Magistério Superior**, em 31/10/2025, às 17:25, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Hani Camille Yehia, Professor do Magistério Superior**, em 31/10/2025, às 17:27, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Maurilio Nunes Vieira, Professor do Magistério Superior**, em 31/10/2025, às 17:30, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site

[https://sei.ufmg.br/sei/controlador_externo.php?](https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0)

[acao=documento_conferir&id_orgao_acesso_externo=0](https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0), informando o código verificador **4682342** e o código CRC **E211DCBC**.

Dedico este trabalho à minha mãe, ao meu irmão George, e à Érica, por todo carinho e apoio incondicional.

AGRADECIMENTOS

Gratidão talvez seja um sentimento difícil de se manter durante uma fase ruim. Contudo, a resiliência em permanecer firme em meus propósitos, com a maturidade de saber que “tudo isso passa, meu filho” e com o suporte das boas pessoas, fez-me emergir em meio à penumbra da tristeza e ingratidão. Assim, antes de tudo, agradeço Àquele em quem tenho fé, por ter colocado em minha vida as pessoas que mencionarei aqui e que, à sua maneira, influenciaram este trabalho.

Primeiro, agradeço à minha mãe por todo o amor, cuidado e apoio; pelas lições de vida, por me fazer uma pessoa mais positiva, pelos inúmeros sacrifícios para sempre garantir uma boa educação e pelo exemplo de pessoa. Agradeço ao George, que, além de ser meu irmão, é também meu melhor amigo, e que sempre me ajuda com suas palavras e sua sagacidade, além de ter auxiliado financeiramente parte dos meus estudos. Agradeço à Érica pelo carinho, companheirismo e empatia, por me lembrar diariamente de como é bom estar ao seu lado, além de ter me ajudado ativamente nos experimentos e por ter composto os trechos de fala feminina usados neste trabalho.

Agradeço ao meu orientador, professor Adriano, pelos ensinamentos, pelo esforço em me ajudar a melhorar, pelo exemplo de profissional, por confiar em mim para integrar uma banca de Trabalho de Conclusão de Curso e pelas boas conversas sobre os mais diversos assuntos. Agradeço também ao meu coorientador, professor Hani, pela sabedoria, pelo auxílio durante a realização deste trabalho e pelas conversas, ensinamentos e caronas. A ambos, muito obrigado por demonstrarem tamanha humanidade e paciência nos momentos mais delicados.

Agradeço a cada um dos professores do Programa de Pós-Graduação em Engenharia Elétrica que lecionaram para mim e sempre se mostraram solícitos: professor Hani, professor Adriano, professor Eduardo Mazoni, professor Braga, professor Frederico Coelho, professor Frederico Guimarães, professor Michel Bessani e professor Aguirre. As disciplinas ministradas contribuíram significativamente, ainda que de forma indireta, para a realização deste estudo.

Agradeço às pessoas que trabalham na secretaria do PPGEe pela proatividade em solucionar os mais diversos problemas burocráticos.

Agradeço também aos meus colegas do CEFALA (Centro de Estudos da Fala, Acústica, Linguagem e Música): ao Rafael, pelas conversas e pela ajuda com o Mac Mini;

ao Kaio, pelo auxílio nos experimentos preliminares; e ao Felipe, ao Aldo, à Michelle e à dona Vera, pelas conversas.

Agradeço aos meus amigos de longa data, Victor, Marcos e Ítalo. Mesmo à distância, as conversas e risadas me ajudaram a superar os problemas e influenciaram positivamente a realização deste trabalho.

Quero ainda agradecer às várias pessoas que me ajudaram, de certo modo, nessa fase da minha vida: ao Eduardo da escola KIND; ao meu sobrinho João Pedro; aos meus cunhados Denival, Elisângela, Vanessa e Thainara; e à minha amiga Aline.

Agradeço à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) e ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pelo apoio financeiro ao longo do mestrado.

Por fim, meus agradecimentos aos meus sogros, com quem, infelizmente, pude conviver por poucos anos. Ao senhor José Luiz, pelo carinho, respeito e pelas duradouras conversas. À dona Irene, pelo carinho, pela confiança e por me mostrar que, mesmo diante das mais variadas adversidades impostas pela vida, nunca devemos deixar de aproveitar nosso bem maior: a vida.

*“E o futuro é uma astronave
Que tentamos pilotar
Não tem tempo, nem piedade
Nem tem hora de chegar
Sem pedir licença
Muda nossa vida
E depois convida
A rir ou chorar.”*

Trecho da música Aquarela, Toquinho.

RESUMO

A separação de fontes é uma etapa fundamental para o processamento eficaz de sinais de fala em diversas aplicações, como: assistentes virtuais, sistemas de videoconferência e dispositivos de assistência auditiva. Nesses contextos, os sinais captados são frequentemente compostos por misturas de múltiplas fontes de fala e ruído ambiente, o que torna a recuperação dos sinais originais um desafio. Métodos de separação de sinais são usados para filtrar sinais de fala indesejados ou recuperá-los a partir de misturas. As tarefas de separação tornam-se desafiadoras no caso de sinais que foram misturados em salas reverberantes, devido à natureza convolutiva das misturas resultantes. As técnicas de separação comumente usadas nesses cenários incluem a Análise em Componentes Independentes no Domínio da Frequência (FDICA) e a Análise de Vetores Independentes (IVA), que utilizam diferentes métodos de otimização, como gradiente natural, iteração de ponto fixo (*fast*) e funções auxiliares. Fatores como o tipo de algoritmo e o posicionamento dos microfones podem afetar o desempenho da separação para diferentes cenários acústicos. O primeiro devido ao modelo e otimização adotados, e, o segundo, devido à diversidade da informação espacial. Este estudo visou investigar e comparar esses fatores para condições experimentais específicas. Para isso, seis métodos de separação cega de fala para misturas convolutivas (três FDICA e três IVA) foram aplicados às misturas de duas fontes, gravadas em quatro salas com diferentes tempos de reverberação. O desempenho dos métodos foi então comparado utilizando-se as seguintes métricas: Razão Sinal-Interferência (SIR), Razão Sinal-Distorção Invariante à Escala (SI-SDR) e Avaliação Perceptiva da Qualidade da Fala (PESQ). Nesse caso, foi realizada a análise com bloqueio de todos os fatores, exceto do fator método. A SI-SDR e a PESQ, por terem resultados com menor variabilidade, também foram usadas para avaliar e comparar, em cada sala, como a distância entre os microfones (0 cm, 4 cm e 57 cm) e a distância média entre fontes e microfones (75 cm, 140 cm, 210 cm) afetaram o desempenho de cada um dos três métodos IVAs, escolhidos a partir da primeira análise. Para a comparação dos algoritmos, a SIR apresentou resultados inconclusivos. A comparação da SI-SDR e PESQ mostrou que o fastFDICA foi estatisticamente inferior aos outros métodos. Além disso, o IVAng superou as técnicas FDICA e não apresentou diferença estatística em relação a outros métodos IVA em um nível de significância de 5%. Para a comparação do posicionamento, os resultados mostraram que o desempenho dependeu mais das informações espaciais do que dos algoritmos. Para cada sala, aumentar a distância entre os microfones melhorou o desempenho, embora esse efeito tenha diminuído com tempos de reverberação mais altos. Variações nas distâncias entre microfones e fontes não produziram efeitos estatisticamente significativos em nenhum tempo de reverberação específico.

Palavras-chave: separação cega da fala; reverberação; análise em componentes independentes; análise em vetores independentes; arranjo de microfones.

ABSTRACT

Source separation is a fundamental step for the effective processing of speech signals in various applications, such as virtual assistants, videoconferencing systems, and hearing-assistance devices. In these contexts, the captured signals are often composed of mixtures of multiple speech sources and ambient noise, which makes recovering the original signals challenging. Signal separation methods are used either to filter out undesired speech signals or to recover them from mixtures. Separation tasks become particularly difficult when signals have been mixed in reverberant rooms due to the convolutive nature of the resulting mixtures. Techniques commonly used in such scenarios include Frequency-Domain Independent Component Analysis (FDICA) and Independent Vector Analysis (IVA), which rely on different optimization approaches such as natural gradient, fixed-point iteration (fast), and auxiliary functions. Factors such as the type of algorithm and the placement of microphones can affect separation performance in different acoustic scenarios—the former due to the adopted model and optimization strategy, and the latter due to the diversity of spatial information. This study aimed to investigate and compare these factors under specific experimental conditions. To this end, six blind speech separation methods for convolutive mixtures (three FDICA and three IVA) were applied to two-source mixtures recorded in four rooms with different reverberation times. The performance of the methods was then compared using the following metrics: Signal-to-Interference Ratio (SIR), Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), and Perceptual Evaluation of Speech Quality (PESQ). In this stage, all factors except the method factor were blocked. Because SI-SDR and PESQ exhibited lower variability, they were also used to evaluate and compare, within each room, how microphone spacing (0 cm, 4 cm, and 57 cm) and the average distance between sources and microphones (75 cm, 140 cm, 210 cm) affected the performance of each of the three IVA methods selected from the initial analysis. Regarding algorithm comparison, SIR results were inconclusive. The SI-SDR and PESQ comparisons showed that fastFDICA was statistically inferior to the other methods. Moreover, IVAng outperformed the FDICA techniques and showed no statistically significant difference from the other IVA methods at a 5% significance level. As for the evaluation of microphone placement, the results indicated that performance depended more on spatial information than on the specific algorithms. For each room, increasing the distance between microphones improved performance, although this effect diminished at higher reverberation times. Variations in the distances between microphones and sources did not produce statistically significant effects at any specific reverberation time.

Keywords: blind speech separation; reverberation; independent component analysis; independent vector analysis; microphone array.

LISTA DE ILUSTRAÇÕES

Figura 1 – Sinal de fala em diferentes representações.	30
Figura 2 – Processo de formação das misturas.	31
Figura 3 – Ilustração do processo de captação de uma fonte de fala em um ambiente real.	32
Figura 4 – Processo de separação das misturas.	36
Figura 5 – Histogramas das misturas comparados às distribuições gaussianas ajus- tadas.	40
Figura 6 – Dispersões de duas variáveis aleatórias gaussianas e laplacianas.	41
Figura 7 – Dispersão das misturas com o vetor de separação, e o histograma relativo comparado à distribuição geradora.	42
Figura 8 – Modelo gerativo da FDICA.	46
Figura 9 – Distribuição laplaciana bivariada esfericamente simétrica.	47
Figura 10 – Histograma da parte real e imaginária.	47
Figura 11 – Modelo gerativo da IVA.	49
Figura 12 – Diagrama da separação.	52
Figura 13 – Espectro de frequência do sinal de fala para avaliação da perda de informação causada pela compressão em “mp3”.	66
Figura 14 – Faixa de áudio estéreo das 16 combinações.	67
Figura 15 – Trecho inicial da faixa de áudio do sinal TSP.	68
Figura 16 – Foto do estúdio e sua planta com as dimensões.	69
Figura 17 – Foto do laboratório CEFALA e sua planta com as dimensões.	69
Figura 18 – Foto da sala 2419 e sua planta com as dimensões.	69
Figura 19 – Foto da sala de seminários e sua planta com as dimensões.	70
Figura 20 – Monitores de áudio usados nos experimentos.	70
Figura 21 – Cabo P2 estéreo 2xP10 mono.	71
Figura 22 – Microfone usado nos experimentos.	71
Figura 23 – Diagrama polar e resposta em frequência do microfone.	71
Figura 24 – Interface de áudio usada.	72
Figura 25 – Cabo <i>FireWire</i> utilizado na comunicação entre interface e computador.	72
Figura 26 – Disposição geral entre os equipamentos durante o experimento.	72
Figura 27 – Foto de cada ensaio realizado no estúdio.	74
Figura 28 – Foto de cada ensaio realizado no CEFALA.	75

Figura 29 – Foto de cada ensaio realizado na sala 2419.	76
Figura 30 – Foto de cada ensaio realizado na sala de seminários.	77
Figura 31 – Procedimento adotado para estimar o tempo de reverberação a partir de cada AIR.	79
Figura 32 – AIRs normalizadas dos ensaios 6, 15, 24 e 33.	79
Figura 33 – Curvas de energia instantânea.	81
Figura 34 – Curva e análise de energia instantânea da sala de seminários.	82
Figura 35 – Box-plots das estimativas da reverberação de cada ambiente.	83
Figura 36 – Espectrogramas do sinal original “aM” e do sinal original “iF”.	91
Figura 37 – Espectrogramas das misturas do sinal “aM” com o sinal “iF” para o ensaio “P4” em cada ambiente.	92
Figura 38 – Espectrogramas do sinal “aM” extraído da mistura no estúdio.	93
Figura 39 – Espectrogramas do sinal “iF” extraído da mistura no CEFALA.	94
Figura 40 – Espectrograma do sinal “aM” extraído da mistura na sala 2419.	94
Figura 41 – Espectrogramas do sinal “iF” extraído da mistura na sala de seminários.	95
Figura 42 – Gráfico de barras para cada método e ambiente, avaliados sob a SIR.	96
Figura 43 – Interação entre método e ensaio, avaliados sob a SIR.	99
Figura 44 – Gráfico quantil-quantil dos resíduos do modelo multifatorial de segunda ordem para a SIR.	101
Figura 45 – Gráfico de barras entre método e ambiente, avaliados sob a SDR.	102
Figura 46 – Interação entre método e ensaio, avaliados sob a SDR.	103
Figura 47 – Gráfico quantil-quantil dos resíduos do modelo ANOVA para o fator “método”, avaliado sob a SDR.	105
Figura 48 – Ganho médio de SDR por canal para variações na distância entre microfones e fontes em cada sala.	109
Figura 49 – Ganho médio de PESQ por canal para variações na distância entre microfones e fontes em cada sala.	109
Figura 50 – Ganho médio de SDR por canal para variações na distância entre microfones em cada sala.	110
Figura 51 – Ganho médio de PESQ por canal para variações na distância entre microfones em cada sala.	111
Figura 52 – Ganho médio de SI-SDR para variações de distância entre fonte e microfone em cada sala. Algoritmo: IVAng.	124
Figura 53 – Ganho médio de SI-SDR para variações de distância entre microfones em cada sala. Algoritmo: IVAng.	125
Figura 54 – Ganho médio de PESQ para variações de distância entre fonte e micro- fone em cada sala. Algoritmo: IVAng.	125
Figura 55 – Ganho médio de PESQ para variações de distância entre microfones em cada sala. Algoritmo: IVAng.	126

Figura 56 – Ganho médio de SI-SDR para variações de distância entre fonte e microfone em cada sala. Algoritmo: fastIVA.	126
Figura 57 – Ganho médio de SI-SDR para variações de distância entre microfones em cada sala. Algoritmo: fastIVA.	127
Figura 58 – Ganho médio de PESQ para variações de distância entre fonte e microfone em cada sala. Algoritmo: fastIVA.	127
Figura 59 – Ganho médio de PESQ para variações de distância entre microfones em cada sala. Algoritmo: fastIVA.	128

LISTA DE TABELAS

Tabela 1 – Nome de cada elocução (sinais das fontes).	65
Tabela 2 – Nome de cada combinação de elocuições.	67
Tabela 3 – Parâmetros físicos dos ensaios realizados no estúdio.	74
Tabela 4 – Parâmetros físicos dos ensaios realizados no CEFALA.	75
Tabela 5 – Parâmetros físicos dos ensaios realizados na sala 2419.	76
Tabela 6 – Parâmetros físicos dos ensaios realizados na sala de seminários 1012.	77
Tabela 7 – Nome de cada um dos 9 ensaios de acordo com os valores de L_2 e L_3	78
Tabela 8 – Valores médios e desvios padrão das estimativas do tempo de reverberação.	84
Tabela 9 – Todas configurações de parâmetros dos métodos.	87
Tabela 10 – Melhores configurações de parâmetros para cada método pelo critério de seleção.	90
Tabela 11 – Resultados para o ensaio “P4” em cada sala para os métodos instantâneos.	90
Tabela 12 – ANOVA do modelo multifatorial de terceira ordem para a SIR. A sigla GL representa “Graus de liberdade”.	100
Tabela 13 – ANOVA do modelo multifatorial de segunda ordem para a SIR. A sigla SQ se refere à soma dos quadrados e a MQ, média dos quadrados.	101
Tabela 14 – ANOVA do modelo multifatorial de terceira ordem para a SDR. A sigla GL representa “Graus de liberdade”.	104
Tabela 15 – ANOVA com blocagem de fatores, contendo apenas “método” como fator para a SDR. As sigla GL e SQ representam, respectivamente “Graus de liberdade” e “Soma dos quadrados”.	104
Tabela 16 – Teste de Tukey para o fator “método”, avaliados pela SDR.	106

LISTA DE ABREVIATURAS E SIGLAS

AIR	Acoustic Impulse Response
ANOVA	Analysis of Variance
ASR	Automatic Speech Recognition
BSS	Blind Source Separation
CEFALA	Centro de Estudos da Fala, Acústica, Linguagem e Música
DSP	Digital Signal Processing
FDICA	Frequency-Domain Independent Component Analysis
FIR	Finite Impulse Response
HEAD	Hybrid Exact Approximative Diagonalization
ICA	Independent Component Analysis
IID	Independente Identicamente Distribuído
IIR	Infinite Impulse Response
ILRMA	Independent Low-Rank Matrix Analysis
INCA	Independent Component Analysis
INFOMAX	Information Maximization
IVA	Independent Vector Analysis
MA	Moving Average
MDP	Minimal Distortion Principle
MLE	Maximum Likelihood Estimation
MNMF	Multichannel Nonnegative Matrix Factorization
MOS	Mean Opinion Score

NMF	Nonnegative Matrix Factorization
PCA	Principal Component Analysis
PESQ	Perceptual Evaluation of Speech Quality
PSQM	Perceptual Speech Quality Measurement
RAM	Random Access Memory
RIR	Room Impulse Response
RT ₆₀	Reverberation Time
SAR	Signal-to-Artifacts Ratio
SDR	Signal-to-Distortion Ratio
SIR	Signal-to-Interference Ratio
SI-SAR	Scale-Invariant Signal-to-Artifacts Ratio
SI-SDR	Scale-Invariant Signal-to-Distortion Ratio
SI-SIR	Scale-Invariant Signal-to-Interference Ratio
SPL	Sound Pressure Level
STFT	Short-Time Fourier Transform
SVD	Singular Value Decomposition
CLT	Central Limit Theorem
TF	Time-Frequency
TSP	Time-Stretched Pulse
UFMG	Universidade Federal de Minas Gerais
UIT-T	International Telecommunication Union – Telecommunication Standardization Sector

LISTA DE SÍMBOLOS

$\mathbf{A}$	Matriz de mistura
$\mathbf{A}_f$	Matriz de mistura dos componentes de frequência f
$\mathbf{A}_h$	Tensor de mistura convolutiva
$\mathbf{B}$	Matriz de separação ideal
$\mathbb{C}$	Conjunto dos números complexos
C_d	Termo constante da distribuição laplaciana multivariável
$\mathbf{C}_{i,n}$	Matriz de covariância ponderada
$\mathbf{e}_n$	Linha n da matriz identidade
$\mathbb{E}$	Operador de valor esperado
e_{alvo}	Sinal original
e_{artf}	Resíduo de artefatos
e_{interf}	Resíduo de interferências
e_{res}	Resíduo de distorções
f	Frequência
H	Entropia
$h(k)$	Resposta ao impulso acústico
$h_{mn}(k)$	Resposta ao impulso acústico entre o microfone m e fonte n
$\mathbf{I}$	Matriz identidade
I	Número total de componentes de frequência
$\mathcal{J}$	Função de custo
J	Número total de quadros de análise

K	Número de amostras de um sinal
k	Índice da amostra
L	Comprimento do filtro FIR
L_1	Distância entre alto-falantes
L_2	Distância entre microfones
L_3	Distância entre fonte e microfone
M	Número de microfones
N	Número de fontes
$\mathbf{P}$	Matriz de permutação
Q	Número de amostras do quadro (hiperparâmetro)
Q_{func}	Função de janelamento
Q_{hop}	Porcentagem de sobreposição de quadros
$\mathbb{R}$	Conjunto dos números reais
$\mathbf{s}$	Vetor de sinais das fontes no tempo
$\mathbf{s}_n$	Sinal de fonte emitido pela fonte n
$\mathbf{s}_f$	Vetor de sinais de fontes na frequência
$\hat{\mathbf{s}}$	Estimação do vetor de sinais das fontes
$\hat{\mathbf{s}}_i$	Vetor estimado de sinais das fontes do i -ésimo componente de frequência
$\mathcal{U}$	Distribuição uniforme
$\mathbf{U}$	Matriz de separação no espaço descorrelacionado
$\mathbf{U}_i$	Matriz de separação no espaço descorrelacionado do i -ésimo componente de frequência
$\mathbf{V}$	Matriz de descorrelação
$\mathbf{V}_i$	Matriz de descorrelação do i -ésimo componente de frequência
$\mathcal{W}$	Tensor de todas matrizes de separação
$\mathbf{W}_{amb}$	Estimação da matriz de separação

$\mathbf{W}$	Matriz de separação
$\mathbf{w}_n$	Linha n da matriz de separação
$\mathbf{W}_i$	Matriz de separação do i -ésimo componente de frequência
$\mathbf{x}$	Vetor de sinais de mistura no tempo
$\mathbf{x}_m$	Sinal de mistura obtida no microfone m
$\mathbf{x}_f$	Vetor de sinais de mistura dos componentes de frequência f
$\mathbf{X}$	Tensor de todas as misturas na frequência
$\mathbf{y}$	Vetor latente de sinais das fontes no tempo
$\mathbf{y}_n$	Sinal separado referente à fonte n
$\mathbf{y}_i$	Vetor latente de sinais das fontes do i -ésimo componente de frequência
$\mathbf{y}_{j,n}$	Vetor tempo-frequência multicanal
$\mathbf{Y}$	Tensor de todos sinais latentes das fontes na frequência
$\mathbf{z}$	Variável descorrelacionada
α	Atenuação devido à propagação da onda acústica
γ	Variável escalar de adaptação da otimização por iteração de ponto fixo
η	Taxa de aprendizagem
Λ	Matriz diagonal de autovalores
$\boldsymbol{\mu}$	Vetor de médias
σ	Desvio padrão
τ	Atraso de propagação
ϕ	Valores possíveis de $\mathbf{y}_n$
$*$	Operação de convolução discreta

SUMÁRIO

1	INTRODUÇÃO	22
1.1	Contextualização	22
1.2	Separação Cega de Fala - História	24
1.3	Justificativa	26
1.4	Objetivos Geral e Específicos	26
1.5	Organização da Dissertação	27
2	FONTES, MISTURAS E SEPARAÇÃO DE SINAIS	28
2.1	Sinais das Fontes	28
2.1.1	Sinal de fala	29
2.2	Sistema de Mistura	31
2.2.1	Misturas Instantâneas	31
2.2.2	Misturas Convolutivas	32
2.2.3	Modelos e Parâmetros Acústicos	34
2.3	Separação de Sinais	36
2.4	Conclusões do capítulo	38
3	ALGORITMOS DE SEPARAÇÃO	39
3.1	Análise em Componentes Independentes	39
3.1.1	ICA por Curtose	41
3.1.2	ICA por Neguentropia	43
3.1.3	ICA por Máxima Verossimilhança	44
3.2	ICA no Domínio da Frequência	45
3.2.1	Representação Tempo-Frequência	45
3.2.2	Ambiguidade de Permutação	47
3.2.3	Ambiguidade de Escala	48
3.3	Análise em Vetores Independentes	49
3.4	Otimização	50
3.4.1	Gradiente Natural	50
3.4.2	Iteração de Ponto Fixo	51
3.4.3	Funções Auxiliares	51
3.5	Algoritmos de Separação Cega	52
3.5.1	ICA Curtose	52
3.5.2	ICA Neguentropia	52
3.5.3	ICA ML	54
3.5.4	FDICAng	54
3.5.5	fastFDICA	54
3.5.6	FDICAaux	57

3.5.7	IVAng	58
3.5.8	fastIVA	59
3.5.9	IVAaux	60
3.6	Medidas de Desempenho	61
3.6.1	Medidas Objetivas	61
3.6.2	Medidas Subjetivas	62
3.7	Conclusão do capítulo	63
4	METODOLOGIA	64
4.1	Condições Experimentais	64
4.2	Etapas pré experimento	64
4.2.1	Geração de Frases	64
4.2.2	Edição	66
4.2.3	Modelagem Acústica	68
4.3	Experimentos	68
4.3.1	Ambientes	68
4.3.2	Equipamentos	70
4.3.3	Ensaio	72
4.3.4	Estimação da Reverberação	78
4.4	Implementações	84
4.4.1	Algoritmos de Separação	84
4.4.2	Métricas de Avaliação	85
4.4.3	Execução dos Algoritmos	85
4.5	Conclusões do capítulo	88
5	RESULTADOS E DISCUSSÃO	89
5.1	Considerações Iniciais	89
5.2	Avaliações Gerais	91
5.3	Estudo Comparativo dos Métodos	96
5.3.1	Análises da SIR	96
5.3.2	Análises da SDR	102
5.3.3	PESQ8, SDR e SIR	106
5.3.4	Análises da SIR para fator único	107
5.4	Efeitos do Posicionamento dos Microfones	108
5.4.1	Distância entre microfones e fontes	108
5.4.2	Distância entre microfones	110
5.5	Conclusões do capítulo	112
6	CONCLUSÃO	113
6.1	Trabalhos Futuros	114
6.2	Publicações	115

REFERÊNCIAS	116
Apêndice A Implementações dos métodos auxiliares	122
Apêndice B Resultados do Posicionamento	124

Capítulo 1

Introdução

1.1 Contextualização

A separação de sinais misturados é uma necessidade recorrente em diferentes áreas. Em muitas situações práticas, várias fontes de informação são captadas simultaneamente por sensores, resultando em sinais sobrepostos, cuja interpretação direta se torna inviável. Assim, a tarefa de decompor esses sinais em suas componentes originais é essencial para aplicações que vão desde a astronomia até a medicina [Cardoso et al., 2002, Barros and Cichocki, 2001].

A relevância do problema de separação pode ser observada em diferentes áreas: na análise de sinais fisiológicos, em que atividades cerebrais distintas precisam ser isoladas a partir de registros de eletroencefalogramas [Makeig et al., 1996]; em sistemas de comunicação sem fio, onde canais sobrepostos ou interferências devem ser identificados e mitigados [van der Veen and Paulraj, 1996]; e no processamento de imagens hiperespectrais, em que padrões espectrais específicos de diferentes materiais são separados para uma classificação precisa [Bioucas-Dias et al., 2012]. Esses exemplos evidenciam que a separação de fontes não é apenas um problema teórico, mas uma ferramenta crucial para transformar sinais complexos em informações úteis.

Entre as áreas em que a separação de sinais se mostra relevante, a fala representa uma aplicação singular. Diferentemente de registros biomédicos ou espectros de imagens, os sinais de fala estão quase sempre inseridos em cenários cotidianos, repletos de interferências sonoras e sujeitos a diversos fenômenos acústicos. Nesses ambientes, múltiplas fontes acústicas se sobrepõem, o que compromete a inteligibilidade e dificulta a extração de informações. Esse desafio é classicamente descrito pelo chamado “Problema do Coquetel” (em inglês, *Cocktail Party Problem*) [Cherry, 1953], que ilustra a habilidade humana de concentrar-se em apenas um locutor em meio a várias vozes e ruídos concorrentes. Enquanto o sistema auditivo humano consegue realizar essa tarefa de forma simples, a reprodução computacional desse mecanismo ainda é uma tarefa em aberto devido à natureza dinâmica dos cenários acústicos e à variabilidade dos sinais de fala. Por essa

razão, a separação de sinais de fala é um eixo fundamental de pesquisa em Processamento de Sinais e Inteligência Artificial, devido ao impacto direto em aplicações modernas, como assistentes de voz, sistemas de videoconferência, dispositivos de acessibilidade auditiva e tecnologias de comunicação homem-máquina.

Os desafios da separação da fala não decorrem apenas da presença de múltiplas fontes, mas também das propriedades acústicas do ambiente. Fenômenos como a reverberação, causada por múltiplas reflexões sonoras, assim como a presença de ruídos de fundo não estacionários, comprometem a clareza do sinal, reduzindo sua inteligibilidade [Huang et al., 2001, Kuttruff, 2009]. Além disso, a sobreposição de diferentes locutores apresenta uma dificuldade adicional: ao contrário de outras modalidades de sinal, os sinais de fala compartilham características espectrais semelhantes, o que dificulta a distinção entre eles. Essas condições tornam a tarefa de separar e reconhecer a fala particularmente desafiadora, especialmente em sistemas artificiais que não dispõem dos mesmos mecanismos cognitivos do sistema auditivo humano.

A relevância prática da separação de fala manifesta-se em uma ampla gama de aplicações que vão muito além do reconhecimento automático de fala. Sistemas de comunicação em tempo real, como videoconferências, dependem da recuperação clara da voz para garantir a inteligibilidade e a qualidade da interação. Da mesma forma, dispositivos de assistência auditiva e tecnologias de acessibilidade exigem algoritmos capazes de destacar o locutor desejado em ambientes ruidosos, oferecendo uma melhor experiência aos usuários com perda auditiva. Ainda, assistentes virtuais e interfaces homem-máquina necessitam de sinais de fala limpos para interpretar comandos com precisão, enquanto aplicações em segurança, monitoramento e saúde demandam a separação eficaz de vozes sobrepostas para a extração de informações críticas. Assim, a tarefa de separar sinais de fala não é apenas um desafio científico, mas também um requisito central para a evolução de diversas tecnologias modernas [Wang and Chen, 2018].

O primeiro esforço sistemático para lidar com a fala em ambientes ruidosos concentrou-se no chamado Aprimoramento de Fala (*Speech Enhancement*). Nessa abordagem, o objetivo é realçar a voz de interesse em meio ao ruído de fundo, melhorando sua qualidade perceptiva e inteligibilidade por meio da modelagem e atenuação do ruído interferente que corrompe o sinal de fala. Embora eficazes em condições relativamente simples, esses métodos apresentam limitações em cenários mais complexos, especialmente quando o ruído é não estacionário ou quando múltiplos locutores falam simultaneamente, situações em que a simples melhoria de sinal já não é suficiente.

Essas limitações motivaram o avanço de técnicas de separação de sinais de fala que buscam identificar e recuperar as diferentes fontes presentes em um ambiente acústico. Inicialmente, tais métodos dependiam de informações adicionais, como o número de fontes ou estatísticas do ruído, para orientar o processo de decomposição. No entanto, em cenários práticos, essas informações raramente estão disponíveis, o que levou ao

desenvolvimento da chamada Separação Cega de Fala. Nesse caso, assume-se que pouco ou nenhum conhecimento prévio é disponível sobre as fontes ou sobre o meio de propagação, cabendo ao algoritmo explorar propriedades intrínsecas dos sinais de fala e do ambiente para realizar a separação. Assim, tais métodos são capazes de extrair os sinais de fala a partir apenas dos sinais misturados, embora ainda dependam de certas premissas para que a separação seja viável.

Ao longo da evolução da Separação de Sinais de Fala, o problema passou a integrar, além dos métodos estatísticos clássicos, as abordagens modernas baseadas em aprendizado de máquina, consolidando-se como uma área importante, com aplicações relevantes. Apesar disso, as abordagens baseadas em aprendizado de máquina não caracterizam um cenário de separação cega, que é o foco deste estudo.

1.2 Separação Cega de Fala - História

A Separação Cega de Fala pertence à área da Separação Cega de Fontes (BSS - *Blind Source Separation*), a qual generaliza o processo de separação para diferentes tipos de sinais. Ambas as áreas se desenvolveram ao longo dos anos conjuntamente, permitindo que técnicas aplicadas a um tipo de sinal pudessem ser utilizadas em outros tipos de sinais, com as modificações dependendo de cada problema. Com isso, além de sinais de fala, os métodos de separação podem ser aplicados aos sinais de música, de eletroencefalografia, de vibração, entre outros. As técnicas podem ser categorizadas em três grupos diferentes, que relacionam o número de fontes (N) com o número de sensores (M). As técnicas determinadas são aquelas aplicadas aos problemas em que o número de fontes e sensores é igual ($M = N$). Os casos subdeterminados e sobredeterminados ocorrem quando $M < N$ e $M > N$, respectivamente. Por conveniência, apenas os dois primeiros casos são comentados nesta seção, com maior ênfase nos casos determinados.

Os primeiros trabalhos em BSS iniciaram-se na metade da década de 1980 e no início da década de 1990, como o trabalho de Jutten e Herault, em que, por meio da inspiração biológica, propuseram um algoritmo neuromimético para distinguir fontes, além de mencionar a separação de sinais com componentes independentes [Jutten and Herault, 1988]. No estudo de Cardoso, automatrizes contendo informações sobre cumulantes de alta ordem são usadas para separar sinais não gaussianos [Cardoso, 1990]. Um outro trabalho precursor importante foi o de Jutten e Herault, no qual se aplicou um algoritmo adaptativo neural para separação cega de fontes por Análise em Componentes Independentes (ICA - *Independent Component Analysis*), antes chamada de INCA [Jutten and Herault, 1991]. Esse estudo, apesar de introduzir o conceito básico do ICA, não o demonstrou com o formalismo matemático necessário. Tais provas foram tratadas no estudo de Comon, que é provavelmente um dos trabalhos mais importantes em ICA [Comon, 1994]. Nesse estudo, foi introduzido o conceito de ICA matematicamente, demonstrado que a não gaussianidade

promove independência, apresentado que ICA é uma extensão da Análise em Componentes Principais (PCA - *Principal Component Analysis*), e foi proposto um algoritmo relacionando não gaussianidade e rotação no espaço geométrico formado pelos dados por meio da Decomposição em Valores Singulares (SVD - *Singular Value Decomposition*).

A partir do estudo de Comon, vários trabalhos renomados foram publicados. Em 1995, no estudo de Bell e Sejnowski, foi proposto um algoritmo para separação de fontes por meio da maximização da informação (Infomax) [Bell and Sejnowski, 1995]. Ainda no mesmo ano, Amari et al. propuseram um algoritmo de separação cega de sinais sem necessidade de calcular inversão matricial a cada iteração, chamado de método do gradiente natural [Amari et al., 1995]. Em 1996, Cardoso e Souloumiac mostraram que a diagonalização simultânea de automatrizes resultava em separação em componentes independentes [Cardoso and Souloumiac, 1996]. O algoritmo da fastICA (ICA rápida) foi desenvolvido em 1997 por Hyvärinen e Oja e leva esse nome devido à sua velocidade de convergência, mantendo a estabilidade durante o processo de otimização [Hyvärinen and Oja, 1997].

O primeiro algoritmo de separação de sinais misturados pelo processo convolutivo foi chamado de ICA no domínio da frequência (FDICA - *Frequency-Domain Independent Component Analysis*) e foi apresentado no estudo de Smaragdis [Smaragdis, 1998]. Desde então, até o ano de 2006, muitos estudos focaram em aplicar novas abordagens na FDICA, seja para a separação de fontes [Bingham and Hyvärinen, 2000, Buchner et al., 2004] ou para resolver os problemas decorrentes da FDICA, como a ambiguidade de permutação [Anemüller and Kollmeier, 2000, Sawada et al., 2004] e a ambiguidade de escala [Matsuoka and Nakashima, 2001, Murata et al., 2001].

Em 2006, dois artigos publicados independentemente trataram de modelos multivariados para as fontes de sinais [Hiroe, 2006, Kim et al., 2006]. Um desses estudos foi responsável pelo conceito da Análise em Vetores Independentes (IVA - *Independent Vector Analysis*). Em 2007, Lee et al. apresentaram a fastIVA [Lee et al., 2007]. O desenvolvimento da FDICA e da IVA otimizados por funções auxiliares foram apresentados em diferentes estudos [Ono and Miyabe, 2010, Ono, 2011]. Em 2012, Ono apresentou uma nova formulação da função de custo do FDICA por meio da modelagem espectral local gaussiana com variância variante no tempo [Ono, 2012].

Paralelamente, a área das abordagens subdeterminadas de separação de sinais também se desenvolveu. Virtanen aplicou aos espectrogramas de misturas a Fatoração em Matrizes Não Negativas (NMF - *Nonnegative Matrix Factorization*), conceito desenvolvido no trabalho de Lee e Seung [Virtanen, 2004, Lee and Seung, 1999]. Os trabalhos decorrentes trataram de problemas subdeterminados [Virtanen, 2007, Duong et al., 2010], mas também de modo generalizado pelo NMF multicanal (MNMF - *Multichannel Non-negative Matrix Factorization*) [Ozerov and Févotte, 2010, Sawada et al., 2013].

A união entre o IVA e o MNMF foi proposta no estudo de Kitamura et al., de modo que o MNMF aumentasse a flexibilidade harmônica do IVA, sendo esse método chamado

de Análise em Matrizes Independentes de Baixo Posto (ILRMA - *Independent Low-Rank Matrix Analysis*) [Kitamura et al., 2016]. Tal abordagem permitiu separações de sinais de naturezas diferentes, como sinais musicais e fala, em casos determinados.

Atualmente, muitos estudos são desenvolvidos para separação de sinais com uso de Aprendizagem Profunda (aprendizagem supervisionada) [Chetupali and Habets, 2022] e [Ansari et al., 2023] sendo, nesses casos, separações não cegas. Entretanto, ainda são realizados estudos baseados em processamento estatístico de sinais, como o desenvolvido por Kellermann, em que foi proposto um algoritmo para o IVA mais rápido que o de iteração por ponto fixo [Brandel and Kellermann, 2021], e o estudo de Yatabe e Kitamura, no qual foi proposto um algoritmo para análise de vetores harmônicos [Yatabe and Kitamura, 2021].

1.3 Justificativa

Um dos desafios na pesquisa em Separação Cega de Fala está na obtenção das misturas. Apesar dos avanços nas metodologias de separação, grande parte dos estudos ainda se baseia em misturas sintéticas ou simuladas. Além disso, alguns trabalhos utilizam apenas uma amostragem dos sinais-estímulo para formar misturas em ambientes reais, considerando uma única configuração espacial de fontes e microfones por sala. Essas abordagens limitam a generalização dos resultados para cenários variados, onde múltiplas fontes de fala, diferentes posicionamentos e efeitos de reverberação coexistem, afetando significativamente o desempenho dos algoritmos. Portanto, torna-se essencial investigar como esses fatores influenciam a separação cega de fala em condições mais diversificadas, fornecendo diretrizes para o desenvolvimento de sistemas robustos e eficazes em aplicações práticas.

1.4 Objetivos Geral e Específicos

Este trabalho teve como objetivo principal a investigação de como o desempenho da separação cega da fala é afetado pelo tipo de algoritmo, pela distância entre fontes e microfones e pela distância entre os microfones, com uso de gravações feitas em salas reais com diferentes tempos de reverberação.

Os objetivos específicos foram:

- Gravar e editar as elocuições a serem utilizadas como sinais das fontes
- Implementar os algoritmos de separação cega da fala a serem comparados
- Selecionar dentre várias configurações de parâmetros, o melhor algoritmo de cada tipo para ir à etapa comparativa

- Estimar o tempo de reverberação de cada ambiente envolvido na experimentação, por meio da estimação das respostas ao impulso acústico
- Comparar o desempenho dos algoritmos de separação cega da fala para cada ambiente
- Comparar o desempenho dos algoritmos de separação cega da fala para diferentes distâncias entre os microfones e as fontes
- Comparar o desempenho dos algoritmos de separação cega da fala para diferentes distâncias entre os microfones.

1.5 Organização da Dissertação

A dissertação está estruturada em seis capítulos. Neste capítulo, foram apresentadas a contextualização e a história da separação cega de sinais de fala, além da justificativa e objetivos. O Capítulo 2 apresenta os fundamentos relacionados a fontes, misturas e separação de sinais, fornecendo a base conceitual necessária para compreender os desafios e abordagens da área. O Capítulo 3 discute os principais algoritmos de separação, bem como as métricas utilizadas para avaliar o desempenho desses métodos. No Capítulo 4, são detalhadas a metodologia adotada, os experimentos realizados e os procedimentos para coleta e processamento dos sinais, além da estimação do tempo de reverberação. O Capítulo 5 apresenta a análise e discussão dos resultados obtidos para diferentes fatores para diferentes métricas. Finalmente, o Capítulo 6 reúne as conclusões gerais e sugestões para pesquisas futuras.

Capítulo 2

Fontes, Misturas e Separação de Sinais

A compreensão do processo de separação de sinais exige, primeiramente, o entendimento de como as misturas são formadas. Esse conhecimento permite a adoção de premissas adequadas para a separação. Sendo assim, neste capítulo são apresentados os conceitos teóricos fundamentais que envolvem a separação de sinais de fala. Primeiro, são descritas as características dos sinais das fontes e suas representações. Em seguida, são discutidos os modelos de mistura e os parâmetros acústicos que influenciam sua formação. Por fim, é introduzido o modelo matemático da separação, que servirá de base para os métodos abordados nos próximos capítulos.

2.1 Sinais das Fontes

Os sinais das fontes, também chamados de sinais originais, são aqueles produzidos individualmente por cada emissor e livres de interferências de outras fontes. Em um processo de mistura, esses sinais constituem as entradas do sistema. Uma forma de compreensão é pela composição musical em estúdio: cada instrumento é gravado separadamente e, posteriormente, as gravações são combinadas para formar a música. Nesse caso, cada gravação isolada corresponde a um sinal de fonte.

O problema de separação para o caso da música seria sua decomposição, recuperando assim as gravações isoladas de cada instrumento. Na separação cega de sinais, esses sinais das fontes são desconhecidos e, além disso, não se sabe como eles foram combinados.

A descrição apresentada até aqui é geral e pode ser aplicada a diferentes tipos de sinais em diversas áreas. Nos próximos parágrafos, será dado destaque ao caso particular da fala, que apresenta propriedades específicas, sendo estas fundamentais para a modelagem do problema de separação.

2.1.1 Sinal de fala

O sinal de fala é um som natural que carrega muita informação. Cada fonema do idioma possui energia elevada em regiões diferentes no espectro de frequências. Sendo assim, a sequência de diversos fonemas de uma palavra ou frase pronunciada faz com que a fala seja um sinal altamente não estacionário. Entretanto, para análises de 20 ms, o trato vocal exibe variações articulatórias lentas e, assim, pode-se considerar que o sinal de fala é estacionário ou *quasi*-estacionário [Rabiner and Schafer, 1978]. As figuras 1a e 1b mostram as representações desse sinal no domínio do tempo e no domínio tempo-frequência.

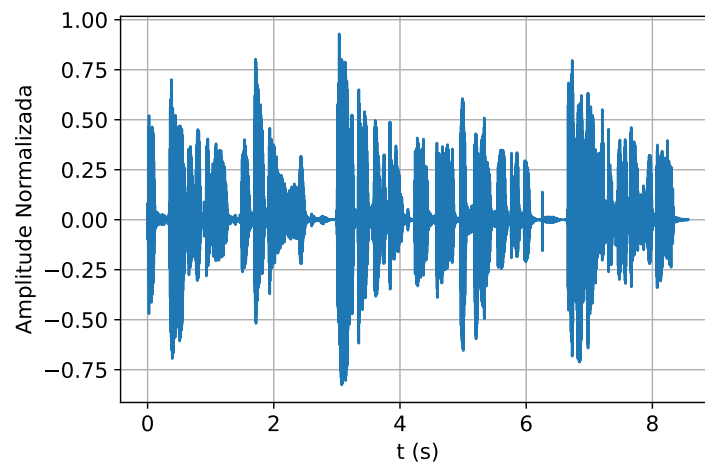
No caso da separação de fala, cada fonte a ser separada é atribuída a um falante ou ao alto-falante que reproduz uma gravação de fala. Uma premissa simplificadora comumente usada é supor que a fonte é pontual, permitindo, assim, o uso de simplificações para o modelo gerativo das misturas, o qual representa como as misturas são geradas pelos sinais das fontes [Vincent et al., 2018, Duong et al., 2010].

Na separação cega da fala, é necessário realizar algumas suposições sobre os sinais de entrada, i.e., os sinais das fontes. Algumas técnicas a serem apresentadas no Capítulo 3 necessitam que o sinal de entrada seja modelado por uma função de densidade de probabilidade. Especificamente para o sinal de fala, distribuições supergaussianas são consideradas *a priori*, dependendo da representação usada [Makino et al., 2007]. As distribuições supergaussianas, também chamadas de leptocúrticas, são caracterizadas por caudas pesadas e pelo afinamento do pico da distribuição. No domínio do tempo, usualmente o sinal de fala é modelado por uma distribuição laplaciana univariada, definida por

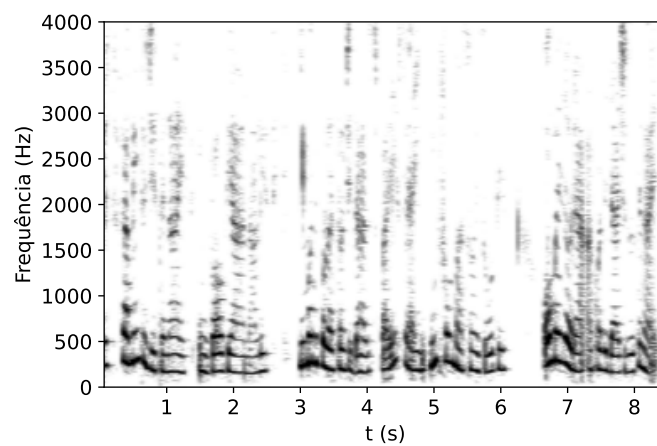
$$p_X(x) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right), \quad (2.1)$$

em que μ é a média, e b é o parâmetro de controle da dispersão, o qual é definido em termos do desvio padrão σ como $b = \sigma/\sqrt{2}$. A Figura 1c mostra o histograma de um sinal de fala real, e a distribuição laplaciana parametrizada pelas estatísticas obtidas desse mesmo sinal.

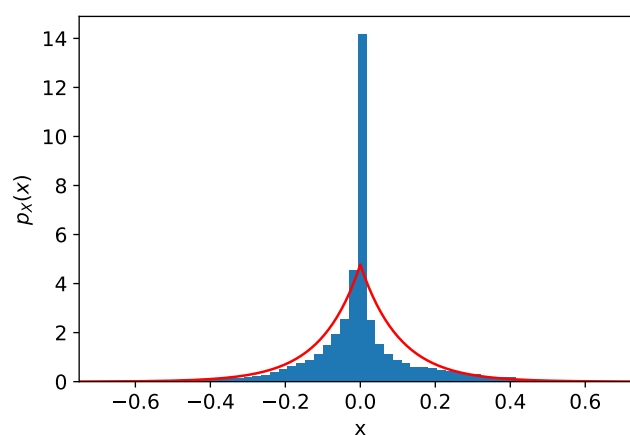
Por conveniência, os detalhes da modelagem dos sinais de fala no domínio tempo-frequência serão apresentados no Capítulo 3.



(a) Sinal de fala no tempo.



(b) Representação tempo-frequência do sinal de fala.



(c) Em azul, o histograma do sinal de fala, e de vermelho, a função densidade de probabilidade relativa.

Figura 1 – Sinal de fala: no domínio do tempo, no domínio tempo-frequência, histograma e distribuição laplaciana relativa. Elocução: “*Processamento de sinais envolve a análise e manipulação de dados para extrair informações úteis, ou melhorar a qualidade dos sinais.*”. Fonte: Autoria própria.

2.2 Sistema de Mistura

O sistema de mistura é responsável por combinar os sinais das fontes (entradas), resultando em misturas (saídas). Isso pode ser visto pela perspectiva de Sinais e Sistemas, como mostra a Figura 2. Nesse caso, considera-se que o processo é de múltiplas entradas e múltiplas saídas, as quais são geradas pelo sistema.



Figura 2 – Processo de formação das misturas. O sistema efetua operações nos N sinais de entrada (fontes) resultando em M sinais de saída (misturas). Fonte: Autoria própria.

Na separação cega, tanto os sinais das fontes quanto os parâmetros do sistema de mistura são desconhecidos. Para viabilizar o processo, são feitas suposições que tornem o problema tratável. Entre as mais usuais, destacam-se a linearidade do sistema de mistura e a suposição de que seus parâmetros permanecem constantes ao longo do tempo, ou seja, que o sistema seja invariante no tempo. Tais premissas permitem simplificar de forma significativa a modelagem.

Um aspecto relevante a ser considerado é se o sistema de mistura realiza operações nos sinais que envolvem atrasos. Em outras palavras, deve-se verificar se os sinais de mistura (saídas) resultam de combinações instantâneas ou convolutivas dos sinais de entrada. Isso será discutido a seguir com mais detalhes.

2.2.1 Misturas Instantâneas

Nas misturas instantâneas, o sistema de mistura não aplica atrasos aos sinais, de modo que cada saída é simplesmente uma combinação linear das fontes. Nesse caso, o modelo gerativo dos dados (ou das misturas) é descrito por

$$\mathbf{x} = \mathbf{A}\mathbf{s} , \quad (2.2)$$

em que $\mathbf{x}$ é a matriz de M sinais de mistura $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M]^T$ de tamanho $M \times K$, $\mathbf{s}$ é a matriz dos N sinais das fontes $[\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_N]^T$ de tamanho $N \times K$, $\mathbf{A}$ é o sistema de mistura ou matriz de mistura, com tamanho $M \times N$, e K é o número de amostras dos sinais.

O sistema de mistura instantânea no tempo é caracterizado pela matriz $\mathbf{A}$, composta por coeficientes escalares e reais, ou seja, $a_{mn} \in \mathbb{R}$. Nesse caso, cada sinal observado é a soma de ponderações das fontes. Para ilustrar o modelo de misturas instantâneas, considere o caso com duas fontes ($N = 2$), duas misturas ($M = 2$) e k o índice da amostra. Seja $\mathbf{s}(k)$ o vetor das amostras dos sinais das fontes no instante k , e $\mathbf{A}$ a matriz de mistura, dados, respectivamente, por

$$\mathbf{s}(k) = \begin{bmatrix} s_1(k) \\ s_2(k) \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

De acordo com a equação 2.2, o vetor de misturas observadas em cada instante k é dado por

$$\mathbf{A}\mathbf{s}(k) = \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} = \begin{bmatrix} a_{11}s_1(k) + a_{12}s_2(k) \\ a_{21}s_1(k) + a_{22}s_2(k) \end{bmatrix}.$$

2.2.2 Misturas Convolutivas

Os sinais de fala, em situações reais, interagem acusticamente com o ambiente. As misturas, nesse caso, são os sinais captados por cada microfone, contendo não apenas as falas originais, mas também as modificações impostas pelo meio acústico. Os efeitos mais relevantes incluem a perda de energia durante a propagação do sinal pelo meio e a presença de réplicas do sinal original, atrasadas e atenuadas, resultantes das reflexões e da absorção da energia da onda em um meio de propagação heterogêneo, como mostra a Figura 3.

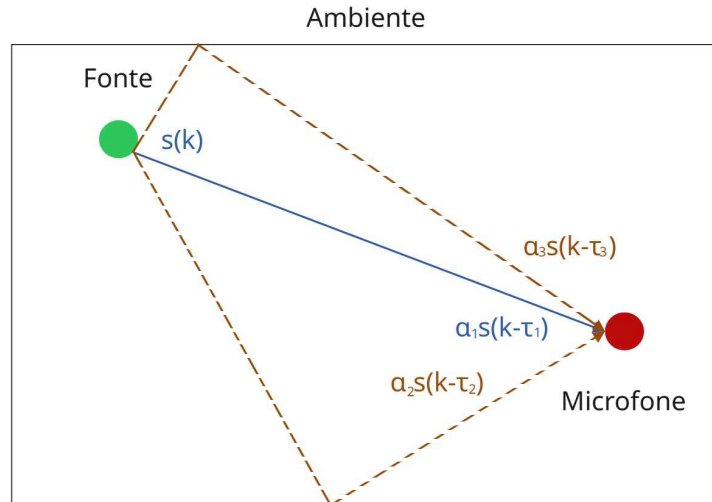


Figura 3 – Ilustração do processo de captação de uma fonte de fala em um ambiente real. A linha contínua azul representa o caminho direto de propagação, e as linhas tracejadas marrons representam as réplicas atrasadas geradas pelas reflexões.

Fonte: Autoria própria.

A linha contínua azul na Figura 3 representa a propagação do caminho direto, sem reflexões, sujeita à atenuação e ao atraso da propagação, indicados por α_1 e τ_1 , respectivamente. As linhas tracejadas em marrom correspondem aos sinais refletidos, também chamados de sinais de múltiplos caminhos. Nesses casos, α_2 e α_3 representam não apenas a atenuação pela propagação, mas também a absorção de energia pelas superfícies refletoras. Além disso, pode-se considerar que o atraso do caminho direto é menor do que o atraso dos caminhos refletidos, ou seja, $\tau_1 < \{\tau_2, \tau_3\}$.

Apesar do processo ser de Resposta ao Impulso Infinita (IIR - *Infinite Impulse Response*), para um melhor tratamento matemático, os efeitos apresentados podem ser modelados como um processo de média móvel (MA - *Moving Average*) [Vincent et al., 2018, Kuttruff, 2009], o qual é um filtro de Resposta ao Impulso Finita (FIR - *Finite Impulse Response*) [Aguirre, 2015].

O processo apresentado na Figura 3 é, na prática, contínuo no tempo e no espaço. Apesar disso, algumas suposições simplificadoras são adotadas para que se obtenha um modelo dinâmico simples relacionado à fonte e ao microfone. Essas suposições são: a fonte é pontual no espaço e o tempo e o espaço são discretizados. Portanto, o que antes era um processo convolutivo e contínuo torna-se convolutivo e discreto. Assim, após as simplificações, observa-se que, para o sinal da fonte indicado por $s(k)$, o microfone amostra as versões atrasadas e atenuadas desse sinal, ou seja,

$$x(k) = \sum_l \alpha_l s(k - \tau_l) , \quad (2.3)$$

em que $x(k)$ é o sinal amostrado pelo microfone. Essa relação pode ser representada como uma convolução discreta entre o sinal da fonte e o filtro FIR $h(k)$, que modela a resposta ao impulso entre a fonte e o microfone. Assim, considerando que o filtro possui L coeficientes, o sinal amostrado pelo microfone é dado por

$$x(k) = \sum_{l=0}^{L-1} h(l)s(k-l) = h(k) * s(k) , \quad (2.4)$$

em que o símbolo $*$ representa a operação de convolução discreta.

Para o caso com um microfone e N fontes espalhadas no ambiente, existem N filtros FIR que modelam a resposta ao impulso entre o microfone e cada fonte. Portanto, deve-se somar a contribuição de cada fonte, e isso é calculado por

$$x(k) = \sum_{n=1}^N s_n(k) * h_n(k) = \sum_{n=1}^N \sum_{l=0}^{L-1} h_n(l) s_n(k-l) , \quad (2.5)$$

em que $s_n(k)$ e $h_n(k)$ representam o sinal da fonte n e a resposta ao impulso entre o microfone e a fonte n , respectivamente.

Por fim, apresenta-se o caso geral de múltiplas entradas e múltiplas saídas, com M microfones e N fontes. Cada par fonte-microfone possui uma resposta ao impulso diferente. Então, para o microfone m , o sinal amostrado é dado por

$$x_m(k) = \sum_{n=1}^N s_n(k) * h_{mn}(k) = \sum_{n=1}^N \sum_{l=0}^{L-1} h_{mn}(l) s_n(k-l), \quad m = 1, 2, \dots, M, \quad (2.6)$$

em que $h_{mn}(k)$ é o filtro FIR que modela a resposta ao impulso entre o microfone m e a fonte n .

A notação da Equação 2.2 pode ser modificada para este caso, ao supor que o elemento a_{mn} da matriz de mistura $\mathbf{A}_h$ corresponde à resposta ao impulso $h_{mn}(k)$; assim, o modelo linear gerativo das misturas convolutivas é definido por

$$\mathbf{x} = \mathbf{A}_h * \mathbf{s}, \quad (2.7)$$

em que $\mathbf{A}_h$ é um tensor $M \times N \times L$, sendo L o número de parâmetros do filtro FIR. Assim, para o caso com duas fontes e dois microfones, a mistura obtida pelo microfone 1 é dada por

$$x_1(k) = h_{11}(k) * s_1(k) + h_{12}(k) * s_2(k). \quad (2.8)$$

É importante destacar que esse modelo é válido sob a premissa de que fontes e microfones são pontuais no espaço. Nessa configuração, a matriz de mistura também pode ser referida como modelo espacial.

Como será abordado no Capítulo 3, os métodos de separação de misturas convolutivas operam no domínio tempo-frequência, e assim, as operações de convolução no tempo tornam-se multiplicações na frequência. O modelo gerativo que define a operação de mistura instantânea na frequência é dado por

$$\mathbf{x}_f = \mathbf{A}_f \mathbf{s}_f, \quad (2.9)$$

em que f é o componente de frequência dos sinais de fonte e mistura, e a matriz $\mathbf{A}_f$ representa o sistema de mistura para o componente f , sendo que seus valores são complexos, ou seja, $a_{fmn} \in \mathbb{C}$.

2.2.3 Modelos e Parâmetros Acústicos

Como explicado anteriormente, a matriz de mistura é composta por filtros responsáveis pelo processo de mistura convolutiva entre os sinais das fontes. Cada elemento

está associado a um par fonte–microfone que define uma função de transferência, a qual caracteriza a forma como os sinais acústicos se propagam de uma fonte específica até um microfone específico [Makino et al., 2007]. Esses filtros são chamados de Respostas ao Impulso da Sala (RIR – *Room Impulse Response*) e, em um sentido mais geral, para qualquer tipo de ambiente em que o sinal acústico se propaga, podem ser denominados Respostas ao Impulso Acústico (AIR – *Acoustic Impulse Response*).

As AIRs são altamente sensíveis à geometria da sala e ao posicionamento espacial das fontes e microfones no ambiente [Wang et al., 2021]. Pequenas mudanças na posição das fontes e microfones podem modificar essas respostas de maneiras sutis ou significativas, dependendo das propriedades acústicas do ambiente, como a presença de superfícies refletoras, o formato da sala e até mesmo a presença de objetos. Como resultado, essas variações impactam diretamente o modelo de mistura espacial, alterando as características dos sinais de fala observados.

Algoritmos de separação cega da fala separam os sinais por meio da estimação de matrizes de separação. Portanto, esses métodos dependem da informação espacial entre as fontes e os microfones. Melhorias no desempenho do algoritmo têm sido observadas quanto mais próximo o arranjo dos microfones estiver do ideal, embora encontrar esse posicionamento não seja uma tarefa trivial [Wang et al., 2021]. Em certas configurações, o sistema pode não capturar informação espacial suficiente para uma separação eficiente, tornando difícil recuperar os sinais individuais.

Uma abordagem utilizada para modelar as AIRs envolve o uso de um sinal chamado Pulso Esticado no Tempo (TSP – *Time-Stretched Pulse*) [Aoshima, 1981, Suzuki et al., 1995]. O TSP é um sinal de excitação, projetado no domínio da frequência, que se estende no tempo de forma controlada e pode ser considerado como uma varredura contínua de frequência. Ao excitar o sistema com tal sinal e observar a resposta no microfone, a AIR pode ser calculada por meio da convolução do sinal recebido no microfone com o sinal TSP reverso no domínio do tempo [Suzuki et al., 1995]. Consequentemente, ao calcular a AIR, pode-se caracterizar o ambiente de forma precisa. Assim, sendo uma técnica eficaz e simples, ela é amplamente utilizada para a caracterização acústica de salas e ambientes [Kim et al., 1994].

De forma análoga às respostas ao impulso na análise de sistemas dinâmicos, uma AIR que decai rapidamente caracteriza um processo com dinâmica rápida, no qual a energia se dissipa em pouco tempo. Como a AIR pode ser representada por um filtro FIR, dado um comprimento L , processos mais lentos apresentam um número maior de coeficientes relevantes (não desprezíveis). Assim, o comprimento do filtro está diretamente relacionado a um parâmetro fundamental do ambiente: o tempo de reverberação. Esse parâmetro, também conhecido como RT_{60} , corresponde ao intervalo de tempo em que a AIR sofre uma atenuação de 60 dB em relação ao seu valor máximo (pico da resposta) [Vincent et al., 2018, Kuttruff, 2009]. Portanto, a partir da AIR estimada, é possível calcular também o

tempo de reverberação [Schroeder, 1965].

2.3 Separação de Sinais

As seções anteriores apresentaram os fundamentos sobre os sinais das fontes e o sistema de mistura, descrevendo como os sinais são combinados ao se propagarem em um ambiente. A etapa de separação está diretamente relacionada a esses conceitos, pois consiste no processo inverso: dado o conjunto de sinais misturados captados pelos microfones, busca-se recuperar os sinais originais das fontes, como mostra a Figura 4. Assim, o objetivo não é estimar diretamente os parâmetros do sistema de mistura, mas sim os do sistema de separação, cuja função é reverter, de forma exata ou aproximada, os efeitos da mistura. Para tanto, é fundamental compreender a estrutura da mistura e as propriedades dos sinais das fontes, de modo a formular e aplicar os métodos de separação.

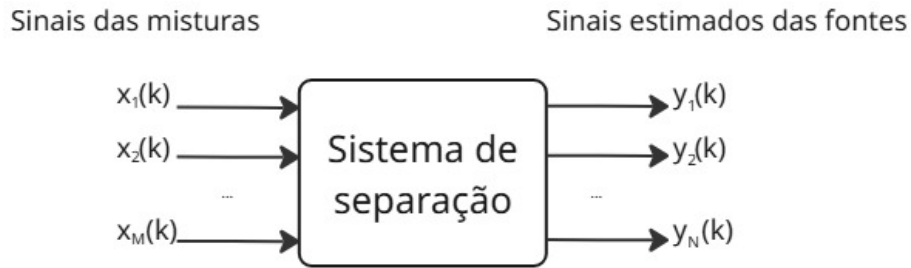


Figura 4 – Processo de separação das misturas. O sistema efetua operações nos M sinais de entrada (misturas) resultando em N sinais de saída (sinais estimados das fontes). Fonte: Autoria própria.

Do ponto de vista de aprendizado supervisionado, a separação poderia ser formulada como o treinamento de um modelo (como uma rede neural) que, a partir de sinais misturados, aprende a produzir diretamente estimativas das fontes originais. Para isso, seriam necessários pares de dados compostos por misturas e suas respectivas fontes, o que permitiria ao modelo capturar mapeamentos complexos sem depender explicitamente da estrutura do sistema de mistura. No entanto, a indisponibilidade dos sinais das fontes e do modelo do sistema de mistura em muitos cenários práticos motiva o estudo de métodos cegos, nos quais apenas as misturas são conhecidas, e o processo de separação precisa ser inferido sem acesso direto às fontes.

Na separação cega da fala, como mencionado anteriormente, a relação entre o número de fontes (N) e o número de microfones (M) define o tipo de problema de separação e, consequentemente, o algoritmo aplicável. Há três casos possíveis: determinado ($M = N$), subdeterminado ($M < N$) e sobredeterminado ($M > N$). Neste trabalho, considera-se apenas o caso determinado.

A partir do modelo gerativo da Equação 2.2 ($\mathbf{x} = \mathbf{A}\mathbf{s}$), a separação consiste em estimar os elementos da matriz de mistura inversa, denominada matriz de separação.

Alguns métodos também estimam o modelo espectral dos sinais das fontes [Kitamura et al., 2016, Ozerov and Févotte, 2010], mas esses não são abordados neste trabalho. Assim, ao adotar o modelo de mistura da Equação 2.2, a forma vetorial da separação de fontes é expressa por

$$\mathbf{y} = \mathbf{\Lambda} \mathbf{P} \mathbf{W} \mathbf{x} , \quad (2.10)$$

em que $\mathbf{W}$ é uma aproximação de $\mathbf{B} = \mathbf{A}^{-1}$, considerada a matriz de separação ideal. A matriz $\mathbf{W}$ possui em cada uma de suas linhas, um vetor de separação, ou seja

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \vdots \\ \mathbf{w}_N^T \end{bmatrix} , \quad (2.11)$$

em que cada $\mathbf{w}_n^T$ extrai das misturas um sinal-fonte.

Após a estimação de $\mathbf{B}$, os sinais das fontes estimados $\mathbf{y}$ são recuperados a partir das misturas $\mathbf{x}$.

O processo de inicialização da matriz de separação pode resultar em sinais separados com permutação de índices; por exemplo, para duas fontes ($N = 2$), pode ocorrer que a estimativa da fonte $\mathbf{s}_1$ seja dada por $\mathbf{y}_2$ e a estimativa de $\mathbf{s}_2$ seja dada por $\mathbf{y}_1$. Para corrigir essa ambiguidade, utiliza-se a matriz $\mathbf{P}$, uma matriz de posto completo com elementos 0 ou 1, chamada matriz de permutação. Além disso, a estimação pode alterar a escala dos sinais separados em relação aos sinais originais. Essa é a chamada ambiguidade de escala, cuja correção é realizada pela matriz diagonal $\mathbf{\Lambda}$.

No Capítulo 3, o modelo de separação convolutiva e as ambiguidades associadas serão apresentados em maiores detalhes.

2.4 Conclusões do capítulo

Nesse capítulo, foram apresentados os conceitos fundamentais para a compreensão do problema de separação cega de sinais de fala. Inicialmente, discutiram-se as propriedades dos sinais das fontes, com destaque para o caso da fala, que apresenta características estatísticas específicas relevantes para a modelagem. Em seguida, foram introduzidos os modelos de mistura, tanto instantâneos quanto convolutivos, ressaltando como a reverberação influencia diretamente na formação das misturas observadas. Por fim, foi abordado o modelo matemático da separação, no qual a tarefa consiste em estimar os parâmetros da matriz de separação a partir das misturas, levando em conta as ambiguidades de permutação e de escala. Este capítulo é a base para os algoritmos de separação cega da fala, apresentados no Capítulo 3.

Capítulo 3

Algoritmos de Separação

Neste capítulo, são apresentadas as abordagens de separação para misturas instantâneas e convolutivas, assim como as derivações matemáticas. Os métodos de otimização são explicados brevemente, para que seja mostrado como o tipo de otimização e a abordagem resultam em diferentes algoritmos de separação cega da fala. Por fim, são apresentadas as medidas de desempenho dos algoritmos de separação.

3.1 Análise em Componentes Independentes

Na separação de sinais, a Análise em Componentes Independentes (ICA - *Independent Component Analysis*) tem sido a abordagem mais utilizada ao longo dos anos. Rigorosamente, ela é um modelo estatístico de variáveis latentes, ou seja, os sinais das fontes são ocultos. Assim, por meio da suposição de um modelo gerativo adequado (Equação 2.2) e da adoção de premissas, pode-se realizar a separação de sinais ou extração de atributos [Hyvärinen et al., 2001].

A aplicação da ICA tem como finalidade recuperar sinais independentes a partir de misturas observadas. Para isso, é fundamental definir as características estatísticas que distinguem os sinais de saída (misturas) das fontes originais.

Supõe-se, para simplificação, que todos os sinais de entrada compartilham a mesma distribuição de probabilidade não gaussiana, são estatisticamente independentes e que o processo de mistura é instantâneo. Nesse contexto, cada sinal de entrada pode ser modelado como um processo estocástico, cujas amostras IID (independentes e identicamente distribuídas) são extraídas a partir da distribuição adotada. Supondo que existam N fontes e M microfones, tal que $N = M$, e considerando o modelo gerativo da Equação 2.2, as linhas da matriz quadrada de mistura $\mathbf{A}$ realizam combinações lineares dos sinais das fontes $\mathbf{s}(k)$. Sendo assim, desde que $\mathbf{A}$ não seja uma matriz identidade escalada e/ou permutada, e as fontes não sejam gaussianas, a mistura resultante terá uma distribuição diferente daquela adotada para os sinais das fontes [Hyvärinen et al., 2001, Kay, 2006]. Em consequência disso, pelo Teorema do Limite Central (CLT – Central Limit Theorem), cada

mistura $\mathbf{x}_m$, sendo uma soma ponderada de muitas variáveis independentes com variância finita, tenderá assintoticamente a uma distribuição aproximadamente gaussiana quando $N \rightarrow \infty$. A Figura 5 mostra os histogramas para misturas de sinais com distribuição laplaciana e como o aumento do número de fontes N modifica a distribuição da mistura, dado que os pesos dos dados são semelhantes.

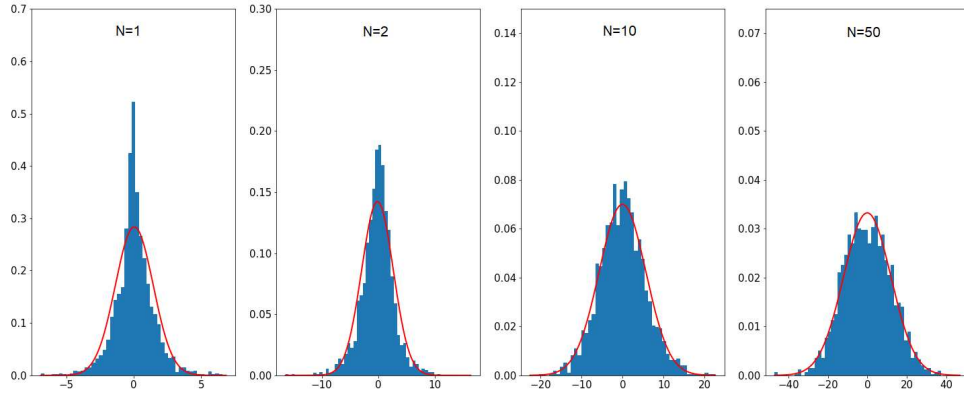


Figura 5 – Histogramas das misturas (azul) comparados às distribuições gaussianas ajustadas (vermelho). Foram consideradas $N = 1, 2, 10, 50$ fontes com distribuição laplaciana, e os vetores de mistura $\mathbf{a}_n$ foram amostrados de $\mathcal{U}(-2, 2)$. Fonte: Autoria própria.

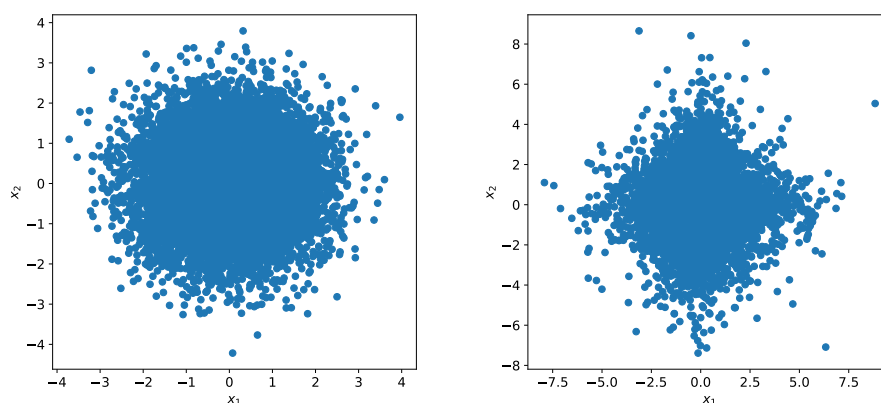
Ao comparar os histogramas para $N = 1$ e $N = 2$, nota-se que o primeiro apresenta maior afastamento de uma distribuição gaussiana em relação ao segundo. Para $N = 10$ e $N = 50$, os histogramas mostram uma aproximação mais evidente e são melhor ajustados por uma distribuição gaussiana. Esse efeito indica que, quanto menos gaussiano for um sinal separado, melhor será a estimativa da fonte independente correspondente. Dessa forma, observa-se uma relação direta entre não gaussianidade e independência estatística [Comon, 1994, Hyvärinen et al., 2001, Sawada et al., 2018].

A matriz de separação $\mathbf{W}$ (Equação 2.10) transforma linearmente as misturas, de modo que geometricamente, as projeções das misturas sob os vetores formados pelas linhas de $\mathbf{W}$ resultam em sinais com distribuições não gaussianas. A Figura 7 mostra esse processo para duas misturas de duas variáveis aleatórias laplacianas apresentadas nas figuras 7a, 7c e 7e. Em cada caso, o vetor de separação $\mathbf{w}_1$ (em preto) realiza a combinação linear das misturas $\mathbf{x}$, resultando na estimativa da fonte $\mathbf{y}_1$. As figuras 7b, 7d e 7f mostram os histogramas da estimativa da fonte $\mathbf{y}_1$ para cada caso. Além disso, comparam-se tais histogramas com a distribuição laplaciana geradora do sinal da fonte $\mathbf{s}_1$. Portanto, pode-se afirmar que a ICA estima os sinais das fontes, fazendo com que a transformação linear das misturas resulte em sinais cada vez menos gaussianos.

Embora a ICA não consiga separar fontes gaussianas, ainda é possível realizar a separação se apenas uma fonte seguir essa distribuição. O caso extremo ocorre quando dois ou mais sinais das fontes são gaussianos. Nesse caso a ICA falha, pois qualquer vetor de projeção produz uma distribuição gaussiana. Isso é ilustrado geometricamente

na Figura 6, que mostra duas misturas, uma com duas variáveis aleatórias gaussianas (Figura 6a) e outra com duas variáveis aleatórias laplacianas (Figura 6b). Ao comparar as dispersões, observa-se que, na Figura 6b, pode-se encontrar vetores cujos dados projetados resultam em distribuições não gaussianas, enquanto na Figura 6a, qualquer direção do vetor de separação resultará em um sinal com distribuição gaussiana. Portanto, exceto no caso gaussiano, a decorrelação entre os sinais de fonte não garante independência, tornando a ICA uma generalização da Análise em Componentes Principais (PCA - *Principal Component Analysis*) [Hyvärinen et al., 2001]. Portanto, a premissa central da ICA é que os sinais das fontes sejam independentes e não gaussianos, de modo que a busca por não gaussianidade nos dados conduza à independência.

Na ICA, as funções contraste são responsáveis por quantificar a não gaussianidade das estimativas realizadas pela matriz de separação [Comon, 1994]. Existem várias funções desse tipo, como: estatísticas de alta ordem, entropia, divergentes e informação mútua [Hyvärinen et al., 2001]. Neste trabalho foram utilizadas a curtose, a neguentropia e a estimação por máxima verossimilhança, as quais são descritas a seguir.



(a) Dispersão de duas variáveis aleatórias gaussianas. (b) Dispersão de duas variáveis aleatórias laplacianas.

Figura 6 – Dispersões de duas variáveis aleatórias gaussianas (esquerda) e laplacianas (direita). Fonte: Autoria própria.

3.1.1 ICA por Curtose

O uso de estatísticas de alta ordem como medida de não gaussianidade parte do fato de que a distribuição gaussiana é completamente caracterizada apenas pela média e pela variância (primeiro e segundo momentos centrais). Nesse caso, todos os momentos centrais ímpares são nulos, e os momentos centrais pares podem ser expressos unicamente em função da variância, de modo que os momentos de ordem superior não trazem informação adicional sobre a distribuição.

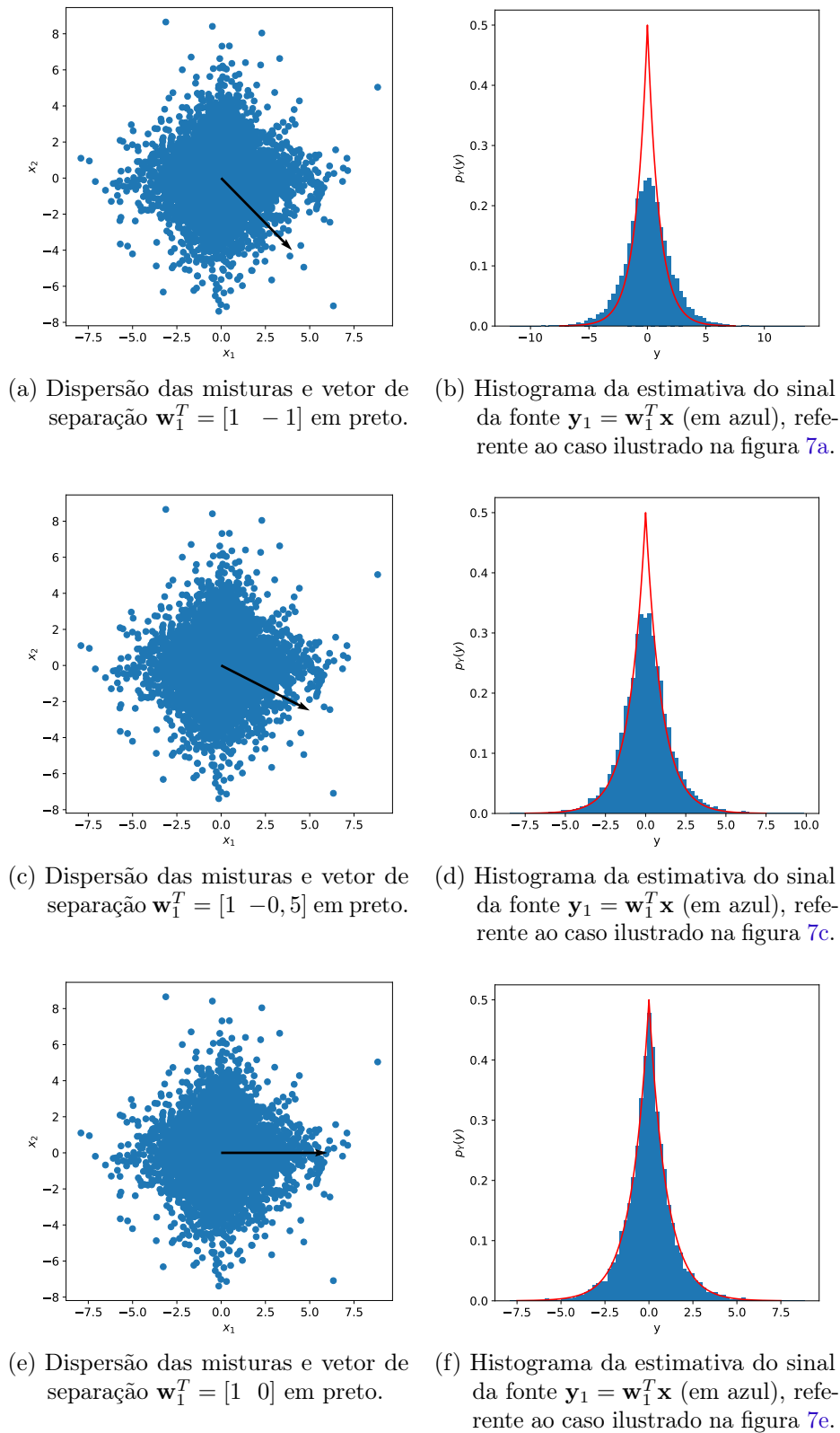


Figura 7 – À esquerda, a dispersão das misturas (círculos em azul) juntamente com o vetor de separação (seta em preto e reescalada para melhor visualização). À direita, o histograma obtido a partir do produto interno entre o vetor de separação e as misturas (barras em azul), comparado à distribuição laplaciana correspondente a cada fonte (linha contínua em vermelho). Fonte: Autoria própria.

A curtose é uma estatística que caracteriza o achatamento e a cauda da distribuição, e seu valor é calculado, para uma variável aleatória $\mathbf{y}_n$, por

$$\text{kurt}(\mathbf{y}_n) = \frac{\mathbb{E}\{(\mathbf{y}_n - \mu)^4\}}{\mathbb{E}\{(\mathbf{y}_n - \mu)^2\}^2}, \quad (3.1)$$

em que μ é o primeiro momento central da variável $\mathbf{y}_n$. As distribuições podem ser categorizadas pelo valor da curtose, sendo as subgaussianas para curtoses negativas, supergaussianas para curtoses positivas. A distribuição gaussiana possui curtose nula. Assim, ao propor uma função que é o valor absoluto da curtose ($f(\mathbf{y}_n) = |\text{kurt}(\mathbf{y}_n)|$), ou função quadrática ($f(\mathbf{y}_n) = \text{kurt}(\mathbf{y}_n)^2$), o ICA se torna um problema de otimização, pois a maximização da função $f(\mathbf{y}_n)$ faz com que os sinais estimados $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}$ sejam mutuamente independentes.

Os parâmetros a serem otimizados são as linhas da matriz de separação $\mathbf{W}$, em que cada uma extrai um componente independente. O ICA para o valor absoluto da curtose é, portanto, o seguinte problema de otimização:

$$\max_{\mathbf{w}_n} |\text{kurt}(\mathbf{w}_n^T \mathbf{x})|, \quad (3.2)$$

em que $\mathbf{w}_n^T$ é a linha n da matriz de separação, $\mathbf{x}$ é o vetor de misturas, e $\mathbf{y}_n = \mathbf{w}_n^T \mathbf{x}$.

Na prática, a curtose possui a desvantagem de ser muito sensível aos valores anômalos, fazendo com que o vetor representado pela linha da matriz de separação aponte mais para a direção do *outlier* do que para a melhor direção dos dados. Consequentemente, a estimação não é robusta e, portanto, são necessárias outras abordagens para calcular as componentes independentes [Comon, 1994, Hyvärinen et al., 2001].

3.1.2 ICA por Neguentropia

A entropia é uma medida de aleatoriedade, de modo que, quanto maior seu valor, menos previsível é a variável. Ela é definida por

$$H(\mathbf{y}_n) = - \int p_{\mathbf{y}_n}(\phi) \log p_{\mathbf{y}_n}(\phi) d\phi, \quad (3.3)$$

em que $p_{\mathbf{y}_n}$ é a função densidade de probabilidade de $\mathbf{y}_n$. Um resultado fundamental da Teoria da Informação afirma que, dentre todas as distribuições de uma variável aleatória com variâncias iguais, a variável gaussiana é a que possui maior entropia. Essas afirmações são úteis para a ICA, pois com elas é possível obter uma forma alternativa à curtose para obter sinais não gaussianos [Hyvärinen et al., 2001].

A neguentropia é uma medida de não gaussianidade derivada da entropia (Equação 3.3), a qual relaciona as entropias de uma variável aleatória gaussiana com uma variável aleatória de distribuição arbitrária, sendo definida por

$$J(\mathbf{y}_n) = H(\mathbf{y}_{n_{gauss}}) - H(\mathbf{y}_n) , \quad (3.4)$$

em que $H(\mathbf{y}_{n_{gauss}})$ é a entropia da distribuição gaussiana definida pela média e variância das observações de $\mathbf{y}_n$. Com isso, a neguentropia $J(\mathbf{y}_n)$ será sempre não negativa, e será nula se, e somente se, $\mathbf{y}_n$ tiver distribuição gaussiana.

A desvantagem de se utilizar as equações acima na ICA é que nem sempre $\mathbf{y}_n$ tem uma distribuição paramétrica, dificultando o cálculo de $H(\mathbf{y}_n)$ com exatidão. A aproximação da neguentropia por momentos polinomiais sofre da alta sensibilidade em relação aos valores anômalos, pois utilizam cumulantes de alta ordem, como ocorre na curtose. Assim, a neguentropia é aproximada por momentos não polinomiais, com a finalidade de atenuar o efeito dos *outliers*. Uma aproximação da neguentropia pelo momento não polinomial é definida por

$$J(\mathbf{y}_n) = [\mathbb{E}\{G(\mathbf{y}_n)\} - \mathbb{E}\{G(v)\}]^2 , \quad (3.5)$$

em que $G(\cdot)$ é uma transformação da variável aleatória, e é dada por $G(\mathbf{y}_n) = \frac{1}{a_1} \log[\cosh(a_1 \mathbf{y}_n)]$, com a_1 sendo hiperparâmetro constante e positivo, v uma variável aleatória normal, e $\mathbb{E}$ o operador de valor esperado [Hyvärinen et al., 2001].

Os componentes independentes são obtidos ao maximizar a neguentropia, sendo o problema de otimização definido por

$$\max_{\mathbf{w}_n} [\mathbb{E}\{G(\mathbf{w}_n^T \mathbf{x})\} - \mathbb{E}\{G(v)\}]^2 . \quad (3.6)$$

3.1.3 ICA por Máxima Verossimilhança

No problema de separação cega da fala, o sinal de fala independente pode ser modelado por uma função densidade de probabilidade supergaussiana. A premissa de atribuir uma distribuição específica *a priori* para os sinais, permite aplicar o conceito de maximizar a probabilidade de se observar um conjunto de amostras (variáveis latentes), dado o conjunto de parâmetros (matriz de separação). Essa é a Estimção por Máxima Verossimilhança (MLE - *Maximum Likelihood Estimation*) e, na separação de sinais de fala, é a abordagem mais utilizada para obter a função de custo, devido à sua flexibilidade para agregar distribuições e conceitos diferentes, que estão fora do escopo deste trabalho. Para maiores detalhes, ver [Sawada et al., 2018]. A ICA por MLE possui algumas características importantes:

- A MLE comprovadamente minimiza a informação mútua entre as variáveis de saída, a qual é a principal medida teórica de independência de variáveis.
- A função de custo da ICA por MLE pode também ser encontrada com aplicação do divergente de Kullback-Leibler.

- O efeito produzido pela MLE é o mesmo da Infomax em redes neurais, em que a entropia das variáveis da saída é maximizada, aumentando o fluxo de informação e, conseqüentemente, a informação mútua entre a saída e a entrada.

O desenvolvimento da função de custo da ICA por MLE é abrangente e foi abordado em alguns estudos [Bell and Sejnowski, 1995, Sawada et al., 2018]. Aqui será apresentada diretamente a função de custo, definida por

$$\mathcal{J}(\mathbf{W}) = J \left[\sum_{n=1}^N \mathbb{E}\{-\log[p_n(\mathbf{w}_n^T \mathbf{x})]\} - 2 \log |\det \mathbf{W}| \right], \quad (3.7)$$

em que $\mathbf{W}$ é a matriz de separação, $\mathbf{w}_n^T$ é a linha da matriz que calcula o n -ésimo componente independente, J é o número de amostras, e $p_n(\mathbf{w}_n^T \mathbf{x})$ é a distribuição assumida para o componente independente n . Nesse caso, os valores ótimos de $\mathbf{W}$ são calculados pela minimização do negativo da log-verossimilhança.

3.2 ICA no Domínio da Frequência

3.2.1 Representação Tempo-Frequência

Os sinais captados por microfones em um ambiente reverberante são misturas convolutivas, e as abordagens anteriores não são capazes de separar os sinais, pois os métodos são aplicados ao caso de misturas instantâneas. Métodos de separação no domínio do tempo de misturas convolutivas existem. No entanto, esses métodos são pouco explorados por dois motivos: a elevada complexidade computacional envolvida e o fato de que, em situações reais, o processo de branqueamento das saídas frequentemente atua como um mínimo local, impedindo a convergência adequada do algoritmo [Smaragdis, 1998]. Dessa forma, torna-se necessária a busca por outras representações que permitam a separação de sinais.

Ao representar o processo de mistura no domínio tempo-frequência (TF - *time-frequency*) fazendo-se uso da Transformada de Fourier de Tempo Curto (STFT - *Short-Time Fourier Transform*), faz com que as misturas convolutivas no tempo sejam representadas como misturas instantâneas na frequência. Com isso, o modelo de misturas na frequência, ilustrado na Figura 8, pode ser representado pela Equação 2.9 [Makino et al., 2007]. A ICA para esse caso é chamada de ICA no Domínio da Frequência (FDICA - *Frequency-Domain Independent Component Analysis*) [Smaragdis, 1998]. A diferença é que, nesse caso, resolve-se não apenas um problema de separação instantânea, mas vários, sendo que o tamanho do quadro de análise da STFT é o que define a quantidade de problemas de separação pela ICA. Desse modo, assume-se que cada quadro seja uma realização da variável aleatória em uma dada frequência. Portanto, resolve-se a ICA para cada componente de frequência,

resultando em uma matriz de separação $\mathbf{W}_i$ associada àquela componente de frequência. A separação para a i -ésima componente é definida de forma similar à Equação 2.10, ou seja,

$$\mathbf{y}_i = \mathbf{\Lambda}_i \mathbf{P}_i \mathbf{W}_i \mathbf{x}_i, \quad (3.8)$$

em que $\mathbf{y}_i$ é a estimativa dos sinais das fontes na i -ésima componente de frequência; $\mathbf{\Lambda}_i$ e $\mathbf{P}_i$ são as matrizes de solução das ambiguidades de escala e permutação, respectivamente; $\mathbf{W}_i$ é a matriz de separação; e $\mathbf{x}_i$ é composta pelas misturas na i -ésima componente de frequência.

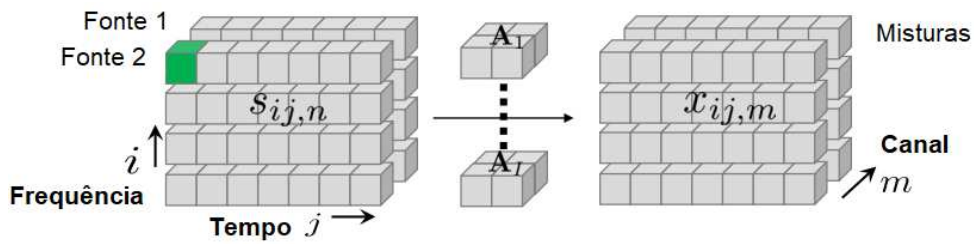


Figura 8 – Modelo gerativo da FDICA. Fonte: Modificado de [Kitamura et al., 2016].

Na FDICA, os sinais de misturas, os sinais das fontes e os elementos da matriz de separação são números complexos. Logo, o modelo estatístico para o sinal de fala representado no domínio da frequência deve ser uma distribuição bivariada, pois ela modela a parte real e imaginária (ambos números reais). Em geral, é utilizada a distribuição laplaciana bivariada esfericamente simétrica com função densidade de probabilidade definida por

$$p(\mathbf{x}) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}\|}{\sigma}\right), \quad (3.9)$$

em que $\boldsymbol{\mu}$ é o vetor de médias, σ é o desvio padrão, que neste caso é um valor escalar devido à simetria da distribuição e $\|\cdot\|$ representa a norma L_2 . O gráfico dessa função é apresentado na Figura 9, com $\sigma = 1$ e $\boldsymbol{\mu} = \mathbf{0}$.

A função de custo da FDICA pode ser derivada por MLE, como foi realizado na seção anterior, e é apresentada pela mesma Equação 3.7, com a diferença de que a função de custo é definida para cada componente de frequência, e a distribuição é supergaussiana bivariada. A Figura 10 mostra o histograma das partes real e imaginária da representação tempo-frequência de um sinal de fala real. Como pode ser observado, os histogramas têm características de distribuições supergaussianas.

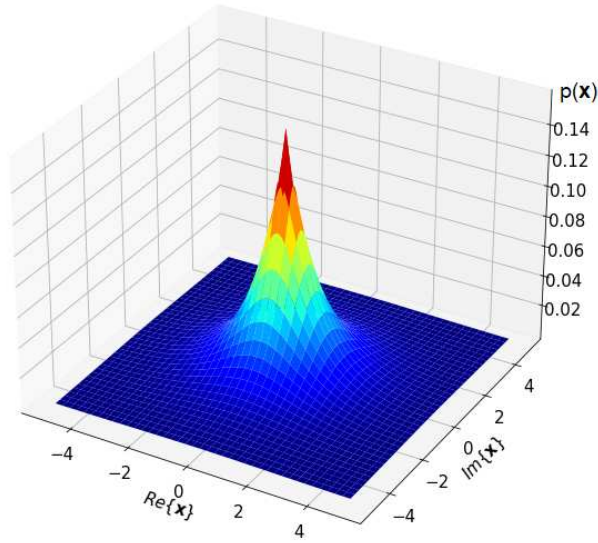


Figura 9 – Distribuição laplaciana bivariada esfericamente simétrica com $\sigma = 1$ e $\mu = 0$.
Fonte: Autoria própria.

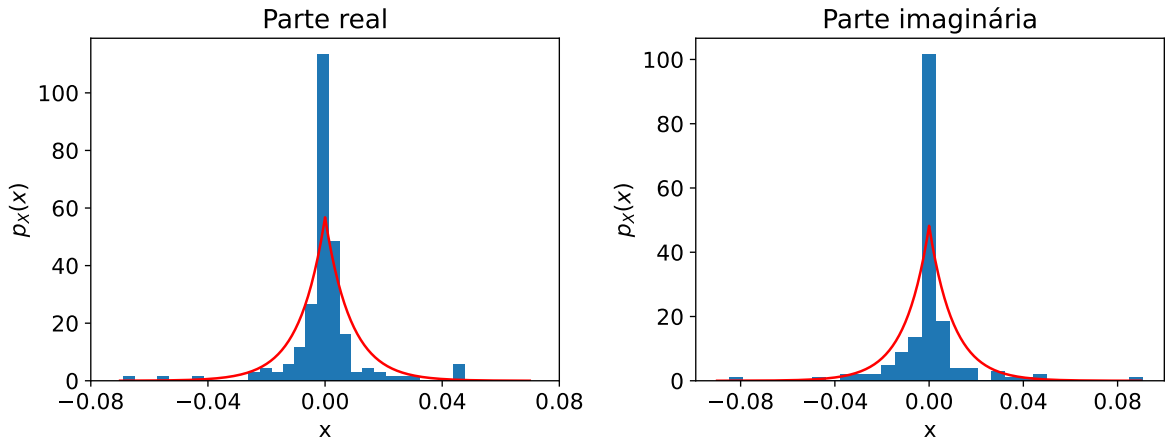


Figura 10 – Histograma da parte real e imaginária de um sinal de fala real (Figura 1a) representado no domínio TF na componente de 397 Hz. Fonte: Autoria própria.

3.2.2 Ambiguidade de Permutação

Como já foi explicado, a inicialização da matriz de separação pode ocasionar a permutação dos componentes independentes, como abordado no Capítulo 2. Para misturas instantâneas no tempo, essa ambiguidade não resulta em problemas, mas o mesmo não pode ser afirmado para o caso do FDICA, onde cada componente de frequência define um problema de separação diferente. Apesar de serem problemas diferentes, cada componente independente $\mathbf{y}_{in}$ é uma estimativa do sinal da fonte $\mathbf{s}_{in}$ para o componente de frequência i , em que $i = \{1, 2, \dots, I\}$. Assim, são I problemas de ICA em que a fonte $\mathbf{s}_{in}$ pode ser estimada de N formas, ou seja, de $\mathbf{s}_{i1}$ até $\mathbf{s}_{iN}$. Em consequência disso, ao reconstruir o sinal $\mathbf{y}_n$ contendo I componentes de frequência pela inversa da STFT, o sinal resultante

da síntese poderá conter em cada frequência, componentes referentes às outras fontes, gerando, assim, um sinal distorcido.

Ao longo dos anos, várias abordagens foram propostas para resolver o problema de permutação no FDICA. Todas elas partem do fato de que existem atributos semelhantes para os componentes independentes da mesma fonte, ao longo dos componentes de frequência. A correlação entre frequências (*interfrequency correlation*) [Anemüller and Kollmeier, 2000] é uma abordagem fundamental para eliminar (ou pelo menos reduzir) a distorção no sinal reconstruído, ocasionada pela ambiguidade de permutação. Nesse método, afirma-se que o envelope do componente de frequência i da fonte n possui maior correlação com envelopes de outros componentes de frequências $\{1, \dots, i - 1, i + 1, \dots, I\}$ referentes à mesma fonte n do que com as outras fontes. Outras abordagens estão fora do escopo deste trabalho, mas podem ser encontradas outro estudo [Sawada et al., 2004].

3.2.3 Ambiguidade de Escala

Durante o processo de otimização, os sinais de mistura são projetados no vetor formado pelas linhas da matriz de separação de modo a obter uma projeção que resulte na maior verossimilhança ou não gaussianidade. Assim, apenas a direção do vetor é importante na extração do componente independente. Os valores arbitrários de módulo e sentido do vetor, apesar de não serem fundamentais na separação, alteram a escala dos sinais separados. Essa ambiguidade de escala é resolvida pela matriz diagonal $\mathbf{\Lambda}$ na Equação 2.10. Nas misturas instantâneas no tempo, os sinais separados estarão apenas com um fator de escala de diferença em relação aos sinais das fontes originais, sendo necessários N escalares reais para a correção. No domínio da frequência, a ambiguidade de escala ocasiona modificação da magnitude e fase ao longo de todo o espectro de frequências, resultando em sinais reconstruídos com componentes de frequências atenuados ou amplificados, ou seja, completamente distorcidos. Nesse caso, são necessários $N \times I$ valores complexos para correção.

A correção da ambiguidade de escala foi proposta por duas abordagens diferentes: Princípio da Distorção Mínima (MDP - *Minimal Distortion Principle*) [Matsuoka and Nakashima, 2001] e Retroprojeção (*projection back*) [Murata et al., 2001]. Os detalhes de ambos os desenvolvimentos não são abordados neste trabalho e a explicação será feita apenas para o MDP, visto que o segundo método não foi aplicado neste estudo.

O principal resultado do estudo que desenvolveu o MDP foi que, dada uma matriz de separação $\mathbf{W}_{amb}$ válida (otimizada) e com correção de permutação, o problema de escala é corrigido por

$$\mathbf{W} = \text{diag}(\mathbf{W}_{amb}^{-1})\mathbf{W}_{amb} , \quad (3.10)$$

em que $\mathbf{W}_{amb}$ é a matriz de separação sem correção da ambiguidade de escala.

3.3 Análise em Vetores Independentes

Na FDICA, a etapa de pós-processamento é necessária para solucionar as ambiguidades de permutação e escala, provocando elevação do custo computacional. Ao supor que as fontes geram variáveis estocásticas escalares, como mostrado na Figura 8, os sinais complexos separados da mesma fonte e em cada frequência não compartilham informação durante a etapa de otimização, em consequência da representação por bases ortogonais da Transformada de Fourier. Uma suposição importante na área de separação de sinais é a de que a fonte é modelada não por uma distribuição bivariada em cada componente de frequência, mas por uma única distribuição multivariada com $2I$ dimensões (parte real e imaginária) para cada fonte. De acordo com esse modelo, as fontes geram variáveis estocásticas vetoriais, como mostra a Figura 11. Assim, a decomposição neste caso não é em componentes independentes, mas em vetores independentes, abordagem à qual é nomeada como Análise em Vetores Independentes (IVA - *Independent Vector Analysis*). A IVA possui três características que a difere da FDICA:

- Além de ser aplicada no caso determinado de separação cega, aplica-se diretamente no caso sobredeterminado ($N > M$).
- A IVA mantém as correlações das componentes de frequência da mesma fonte e separa aqueles componentes de fontes diferentes.
- No pós-processamento, a IVA necessita apenas da solução da ambiguidade de escala, sendo livre da ambiguidade de permutação.

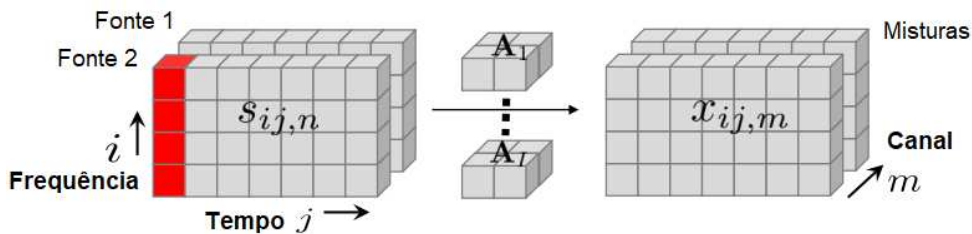


Figura 11 – Modelo gerativo da IVA. Fonte: Modificado de [Kitamura et al., 2016].

Apesar de haver uma única distribuição multivariada para cada fonte, as matrizes de separação na IVA ainda são definidas para cada componente de frequência, mas apenas uma única função de custo é otimizada para todo o problema de separação. O desenvolvimento da função a ser otimizada é apresentado em [Lee et al., 2007, Guo et al., 2023], e por conveniência, será apresentada apenas a função de custo obtida por meio da MLE, dada por

$$\mathcal{J}(\mathcal{W}) = \sum_{j=1}^J \sum_{n=1}^N -\log p(\mathbf{y}_{j,n}) - 2J \sum_{i=1}^I \log |\det \mathbf{W}_i|, \quad (3.11)$$

em que $\mathcal{W} = \{\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_I\}$ é o conjunto das matrizes de separação otimizadas conjuntamente, e $\mathbf{y}_{j,n}$ é a realização j do vetor aleatório referente à fonte n .

O termo $p(\mathbf{y}_{j,n})$ é uma distribuição multivariada, a qual deve ser supergaussiana para a separação cega da fala. Em geral, assume-se que seja uma distribuição laplaciana multivariada esfericamente simétrica, definida por

$$p(\mathbf{y}_{j,n}) = C_d \exp(-\|\mathbf{y}_{j,n}\|_2) , \quad (3.12)$$

em que $\|\cdot\|_2$ representa a norma L_2 do vetor, e C_d é uma constante dependente do número de dimensões $d = 2I$ da distribuição que, em geral, pode ser considerada unitária durante a otimização [Makino et al., 2007].

A separação pelo IVA ocorre quando a função de custo da Equação 3.11 é minimizada (minimização do negativo da log-verossimilhança).

3.4 Otimização

No problema de separação cega da fala, o tipo de método de otimização das funções de custo apresentadas resulta em diferentes algoritmos de separação. Aqui serão abordados de forma breve três diferentes métodos de otimização. Os detalhes de cada método podem ser vistos nos trabalhos referenciados.

3.4.1 Gradiente Natural

Em otimização, o método do gradiente descendente é uma abordagem simples para minimizar a função objetivo. Nele, a cada atualização da variável de decisão (como a matriz de separação), utiliza-se a direção oposta ao gradiente da função-custo, que corresponde ao sentido de maior declínio local no espaço cartesiano. A magnitude do gradiente atua como fator proporcional no reajuste da variável, e o processo é interrompido quando se alcança um número máximo de iterações ou algum critério de convergência.

Apesar do gradiente descendente ser um método simples, em separação cega de sinais, a atualização envolve cálculo de matrizes inversas. Mesmo que a matriz de separação numa dada iteração não seja singular, calcular várias matrizes inversas a cada recursão pode ser inviável. Entretanto, um estudo propôs uma transformação do gradiente na perspectiva da Geometria da Informação, sob a suposição de que a matriz de mistura $\mathbf{A}$ seja invertível. A transformação é a multiplicação do termo de atualização por $\mathbf{W}^T \mathbf{W}$, fazendo com que o gradiente desça na direção mais íngreme no espaço riemanniano [Amari et al., 1995]. Além disso, o procedimento tem a propriedade de equivariância, em que a transformação dos dados é igual à transformação do parâmetro [Cardoso and Laheld, 1996]. A transformação proposta elimina a necessidade de realizar a inversão matricial, acelerando

a convergência do algoritmo. No entanto, essa modificação não produz soluções melhores do que a abordagem baseada em inversão matricial. Por se tratar de uma otimização baseada em gradiente, o método ainda depende da escolha adequada da taxa de aprendizado e permanece sujeito a possíveis instabilidades.

3.4.2 Iteração de Ponto Fixo

Os métodos de separação cega de sinais, nos quais as funções objetivo são otimizadas por iteração de ponto fixo (ou Newton-Raphson), costumam receber o prefixo *fast* em seus nomes [Hyvärinen and Oja, 2000]. Isso parte do fato de que essa otimização leva poucas iterações para convergir, além de ser estável e livre de hiperparâmetros como a taxa de aprendizagem [Hyvärinen and Oja, 1997].

Apesar das qualidades mencionadas, a abordagem assume que os dados estejam descorrelacionados e que as linhas da matriz de separação sejam de norma unitária ($\|\mathbf{w}\| = 1$). A descorrelação dos sinais demanda a etapa de pré-processamento com aplicação da PCA e armazenamento das matrizes responsáveis pela descorrelação dos dados. Além disso, as linhas da matriz de separação serem de norma unitária restringe o espaço de busca a uma hipersfera de raio unitário. Isso acelera a velocidade de convergência, ao custo da qualidade das soluções.

Após a otimização, a etapa de pós-processamento deve primeiro aplicar a transformação inversa da PCA para obter as matrizes de separação adequadas ao problema real, e depois, solucionar as ambiguidades de permutação e/ou de escala.

3.4.3 Funções Auxiliares

Na separação cega de sinais, a otimização por funções auxiliares promove um remapeamento da variável de decisão envolvida na função contraste, isto é, a função que avalia a não gaussianidade, de modo que a nova função, chamada de auxiliar, tenha propriedades matemáticas adequadas para otimização. Uma função auxiliar amplamente usada é a quadrática, pois ela garante que o mínimo global da função auxiliar seja encontrado a cada iteração. Assim, ao minimizar a variável auxiliar, garante-se que a variável de decisão original também seja minimizada [Ono, 2011, 2012].

A solução pela função auxiliar quadrática envolve resolver o problema de Diagonalização de Aproximação Exata Híbrida (HEAD), também chamada de Projeção Iterativa [Yeredor, 2009]. Esse método garante diagonalizar a matriz de covariância ponderada (função auxiliar). Os algoritmos apresentados na Seção 3.5 abordarão esse problema para cada função de custo.

3.5 Algoritmos de Separação Cega

As etapas de cada algoritmo de separação cega da fala, para misturas instantâneas no tempo e misturas instantâneas na frequência, são apresentadas a seguir. A Figura 12 apresenta os passos básicos para a separação.

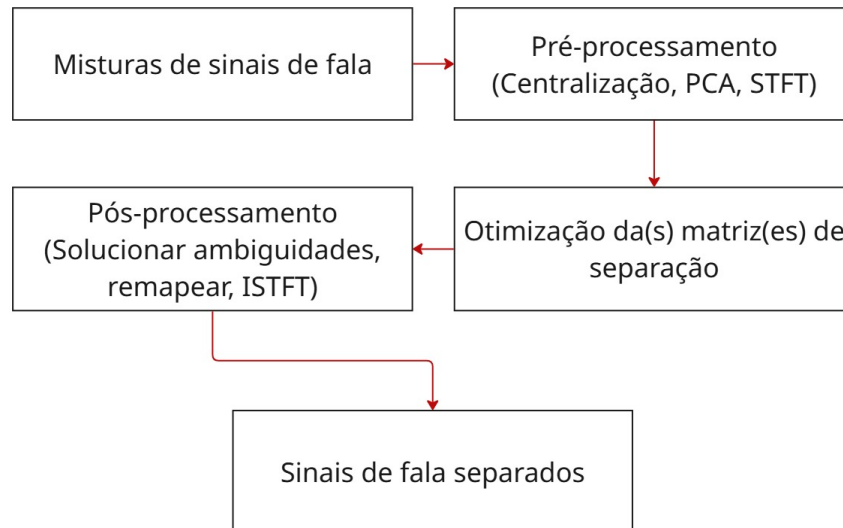


Figura 12 – Diagrama com os passos básicos para a separação de sinais de fala. Fonte: Autoria própria.

3.5.1 ICA Curtose

A ICA por curtose e otimizada pelo gradiente descendente, com pré-processamento, é apresentada pelo Algoritmo 1 [Comon, 1994, Hyvärinen et al., 2001]. Cada vetor de separação é estimado separadamente.

3.5.2 ICA Neguentropia

A ICA por neguentropia otimizada pelo gradiente descendente com pré-processamento, é apresentada pelo Algoritmo 2 [Comon, 1994, Hyvärinen et al., 2001]. Como anteriormente, cada vetor de separação é estimado a cada instante.

Algoritmo 1: ICA por Curtose

Entrada: Misturas $\mathbf{x}(t)$ **Pré processamento**

1. Centralização das misturas pela média.
2. Descorrelação das misturas pelo PCA: $\mathbf{z} = \mathbf{V}\mathbf{x}$

OtimizaçãoInicialização: $p \leftarrow 1$; $\mathbf{W} = \mathbf{I}$; η ;

3. Selecionar a linha p da matriz de separação.
4. $\mathbf{w}_p \leftarrow \mathbf{w}_p + \eta \text{sign}(\text{kurt}(\mathbf{w}_p^T \mathbf{z}))[\mathbb{E}\{\mathbf{z}(\mathbf{w}_p^T \mathbf{z})\} - 3\mathbf{w}_p]$
5. Se $p \geq 2$, ortogonalizar pelo método deflacionário (Gram-Schmidt):

$$\mathbf{w}_p \leftarrow \mathbf{w}_p - \sum_{j=1}^{p-1} (\mathbf{w}_p^T \mathbf{w}_j) \mathbf{w}_j$$

6. $\mathbf{w}_p \leftarrow \mathbf{w}_p / \|\mathbf{w}_p\|$
 7. Se $\mathbf{w}_p$ não houver convergido, voltar no passo 4.
 8. $p \leftarrow p + 1$ e voltar no passo 3.
 9. Calcular separação: $\mathbf{y}(t) \leftarrow \mathbf{W}\mathbf{z}$
-

Algoritmo 2: ICA por Neguentropia

Entrada: Misturas $\mathbf{x}(t)$ **Pré processamento**

1. Centralização das misturas pela média.
2. Descorrelação das misturas pelo PCA: $\mathbf{z} = \mathbf{V}\mathbf{x}$

OtimizaçãoInicialização: $p \leftarrow 1$; $\mathbf{W} = \mathbf{I}$; $\gamma \leftarrow 1$;

3. Selecionar a linha p da matriz de separação.
4. $\gamma \leftarrow \mathbb{E}\{\ln(\cosh(\mathbf{w}_p^T \mathbf{z}))\} - 0,162$
5. $\mathbf{w}_p \leftarrow \mathbf{w}_p + \gamma \mathbb{E}\{\mathbf{z} \tanh(\mathbf{w}_p^T \mathbf{z})\}$
6. Se $p \geq 2$, ortogonalizar pelo método deflacionário (Gram-Schmidt):

$$\mathbf{w}_p \leftarrow \mathbf{w}_p - \sum_{j=1}^{p-1} (\mathbf{w}_p^T \mathbf{w}_j) \mathbf{w}_j$$

7. $\mathbf{w}_p \leftarrow \mathbf{w}_p / \|\mathbf{w}_p\|$
 8. Se $\mathbf{w}_p$ não houver convergido, voltar no passo 4.
 9. $p \leftarrow p + 1$ e voltar no passo 3.
 10. Calcular separação: $\mathbf{y}(t) \leftarrow \mathbf{W}\mathbf{z}$
-

3.5.3 ICA ML

A ICA por máxima verossimilhança otimizada pelo gradiente descendente com pré-processamento, é apresentada pelo Algoritmo 3. Esse algoritmo também é chamado de algoritmo de Bell-Sejnowski [Bell and Sejnowski, 1995].

3.5.4 FDICAng

A FDICA por máxima verossimilhança otimizada pelo gradiente natural é apresentada pelo Algoritmo 4 [Smaragdis, 1998, Amari et al., 1995, Sawada et al., 2018].

3.5.5 fastFDICA

A FDICA por máxima verossimilhança otimizada pela iteração por ponto fixo é apresentada pelo Algoritmo 5 [Hyvärinen and Oja, 2000, Smaragdis, 1998, Sawada et al., 2018].

Algoritmo 3: ICA por Máxima Verossimilhança

Entrada: Misturas $\mathbf{x}(t)$ **Pré processamento**

1. Centralização das misturas pela média.
2. Descorrelação das misturas pelo PCA: $\mathbf{z} = \mathbf{V}\mathbf{x}$

OtimizaçãoInicialização: $\mathbf{W} = \mathbf{I}$; η ;

3. $\mathbf{W} \leftarrow \mathbf{W} + \eta([\mathbf{W}^T]^{-1} + \mathbb{E}\{-2 \tanh(\mathbf{W}\mathbf{z})\mathbf{z}^T\})$
 4. Para cada linha p de $\mathbf{W}$:

$$\mathbf{w}_p \leftarrow \frac{\mathbf{w}_p}{\|\mathbf{w}_p\|}$$
 5. Se $\mathbf{W}$ não houver convergido ou limite de iterações não alcançado, voltar no passo 3.
 6. Calcular separação: $\mathbf{y}(t) \leftarrow \mathbf{W}\mathbf{z}$
-

Algoritmo 4: FDICA por Máxima Verossimilhança - Gradiente Natural

Entrada: Misturas $\mathbf{x}(t)$ Hiperparâmetros: η ; Q ; Q_{hop} ; Q_{func} ;**Pré processamento**

1. $\mathbf{X} \leftarrow \text{STFT}(\mathbf{x}(t), Q, Q_{hop}, Q_{func})$ em que $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_I\}$.

OtimizaçãoInicialização: $\mathcal{W} = \{\mathbf{W}_1 \leftarrow \mathbf{I}, \dots, \mathbf{W}_I \leftarrow \mathbf{I}\}$;Para cada componente de frequência i :

2. Para cada fonte n calcular: $\mathbf{y}_{i,n} = \mathbf{w}_{i,n}^H \mathbf{x}_i$ e $g(\mathbf{y}_{i,n}) = \mathbf{y}_{i,n} / (2|\mathbf{y}_{i,n}|)$;
3. Atualizar: $\mathbf{W}_i \leftarrow \mathbf{W}_i - \eta[\mathbb{E}\{\mathbf{g}(\mathbf{y}_i)\mathbf{y}_i^H\} - \mathbf{I}]\cdot\mathbf{W}_i$
4. Se $\mathbf{W}_i$ não houver convergido ou limite de iterações não alcançado, voltar no passo 2.

Pós processamento

5. Calcular separação: $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_I | \mathcal{W}, \mathbf{X}\}$.
 6. Corrigir permutação: $\mathcal{W} \leftarrow \text{perm}(\mathcal{W}, \mathbf{Y})$.
 7. Corrigir escala: $\mathcal{W} \leftarrow \text{escal}(\mathcal{W})$.
 8. Calcular separação sem ambiguidade: $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_I | \mathcal{W}, \mathbf{X}\}$.
 9. Calcular sinais separados no tempo: $\mathbf{y}(t) \leftarrow \text{ISTFT}(\mathbf{Y}, Q, Q_{hop}, Q_{func})$
-

Algoritmo 5: FDICA por Máxima Verossimilhança - Iteração por Ponto Fixo

Entrada: Misturas $\mathbf{x}(t)$ Hiperparâmetros: $Q; Q_{hop}; Q_{func};$ **Pré processamento**

1. $\mathbf{X} \leftarrow \text{STFT}(\mathbf{x}(t), Q, Q_{hop}, Q_{func})$ em que $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_I\}$.
2. Para cada componente i de frequência:
Centralizar os dados $\mathbf{x}_i$.
Descorrelacionar os dados pelo PCA: $\mathbf{z}_i \leftarrow \mathbf{V}_i \mathbf{x}_i$.

OtimizaçãoInicialização: $\mathcal{U} = \{\mathbf{U}_1 \leftarrow \mathbf{I}, \dots, \mathbf{U}_I \leftarrow \mathbf{I}\};$ Para cada componente de frequência i e fonte n :

3. Calcular separação: $\mathbf{y}_{i,n}(j) \leftarrow \mathbf{u}_{i,n}^H \mathbf{z}_i(j)$
4. Calcular derivada do contraste: $g(\mathbf{y}_{i,n}) \leftarrow \mathbf{y}_{i,n} / (2|\mathbf{y}_{i,n}|)$
5. Calcular derivada segunda do contraste: $g'(\mathbf{y}_{i,n}) \leftarrow 1 / (4|\mathbf{y}_{i,n}|)$
6. Atualização: $\mathbf{u}_{i,n} \leftarrow \mathbb{E}\{g(\mathbf{y}_{i,n})\mathbf{z}\} - \mathbb{E}\{g'(\mathbf{y}_{i,n})\}\mathbf{u}_{i,n}$
7. Se $n \geq 2$, ortogonalizar pelo método deflacionário (Gram-Schmidt):

$$\mathbf{u}_{i,n} \leftarrow \mathbf{u}_{i,n} - \sum_{k=1}^{n-1} (\mathbf{u}_{i,k}^H \mathbf{u}_{i,n}) \mathbf{u}_{i,k}$$

8. Normalizar: $\mathbf{u}_{i,n} \leftarrow \mathbf{u}_{i,n} / \|\mathbf{u}_{i,n}\|$
9. Se $\mathbf{U}_i$ não houver convergido ou limite de iterações não alcançado, voltar no passo 3.

Pós processamento

10. Remapear as matrizes de separação: $\mathcal{W} \leftarrow \{\mathbf{W}_1 \leftarrow \mathbf{U}_1 \mathbf{V}_1, \dots, \mathbf{W}_I \leftarrow \mathbf{U}_I \mathbf{V}_I\}$
 11. Calcular separação das misturas originais: $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_I | \mathcal{W}, \mathbf{X}\}$.
 12. Corrigir permutação: $\mathcal{W} \leftarrow \text{perm}(\mathcal{W}, \mathbf{Y})$.
 13. Corrigir escala: $\mathcal{W} \leftarrow \text{escal}(\mathcal{W})$.
 14. Calcular separação sem ambiguidades: $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_I | \mathcal{W}, \mathbf{X}\}$.
 15. Calcular sinais temporais separados: $\mathbf{y}(t) \leftarrow \text{ISTFT}(\mathbf{Y}, Q, Q_{hop}, Q_{func})$
-

3.5.6 FDICAaux

A FDICA por máxima verossimilhança otimizada por função auxiliar é apresentada pelo Algoritmo 6 [Ono and Miyabe, 2010, Sawada et al., 2018].

Algoritmo 6: FDICA por Máxima Verossimilhança - Função Auxiliar

Entrada: Misturas $\mathbf{x}(t)$

Hiperparâmetros: Q ; Q_{hop} ; Q_{func} ;

Pré processamento

1. $\mathbf{X} \leftarrow \text{STFT}(\mathbf{x}(t), Q, Q_{hop}, Q_{func})$ em que $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_I\}$.

Otimização

Inicialização: $\mathcal{W} = \{\mathbf{W}_1 \leftarrow \mathbf{I}, \dots, \mathbf{W}_I \leftarrow \mathbf{I}\}$;

Para cada componente de frequência i e fonte n :

2. Calcular separação: $\mathbf{y}_{i,n}(j) \leftarrow \mathbf{w}_{i,n}^H \mathbf{x}_i(j)$
3. Calcular matriz de covariância ponderada de cada fonte:

$$\mathbf{C}_{i,n} \leftarrow \mathbb{E} \left[\frac{1}{2|\mathbf{y}_{i,n}|} \mathbf{x}_i \mathbf{x}_i^H \right]$$

4. Resolver o problema HEAD (diagonalização conjunta) para $\mathbf{C}_{i,1}, \dots, \mathbf{C}_{i,N}$, com $\mathbf{e}_n$ sendo o vetor unitário com o n -ésimo elemento igual a 1.

$$\mathbf{w}_{i,n} \leftarrow (\mathbf{W}_i \mathbf{C}_{i,n})^{-1} \mathbf{e}_n$$

5. Atualização:

$$\mathbf{w}_{i,n} \leftarrow \mathbf{w}_{i,n} / \sqrt{\mathbf{w}_{i,n}^H \mathbf{C}_{i,n} \mathbf{w}_{i,n}}$$

6. Se $\mathbf{W}_i$ não houver convergido ou limite de iterações não alcançado, voltar no passo 2.

Pós processamento

7. Calcular separação das misturas originais: $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_I | \mathcal{W}, \mathbf{X}\}$.
 8. Corrigir permutação: $\mathcal{W} \leftarrow \text{perm}(\mathcal{W}, \mathbf{Y})$.
 9. Corrigir escala: $\mathcal{W} \leftarrow \text{escal}(\mathcal{W})$.
 10. Calcular separação sem ambiguidades: $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_I | \mathcal{W}, \mathbf{X}\}$.
 11. Calcular sinais temporais separados: $\mathbf{y}(t) \leftarrow \text{ISTFT}(\mathbf{Y}, Q, Q_{hop}, Q_{func})$
-

3.5.7 IVAng

O algoritmo da IVA por máxima verossimilhança otimizada pelo gradiente natural é apresentada pelo Algoritmo 7 [Kim et al., 2006, Sawada et al., 2018].

Algoritmo 7: IVA por Máxima Verossimilhança - Gradiente Natural

Entrada: Misturas $\mathbf{x}(t)$

Hiperparâmetros: η ; Q ; Q_{hop} ; Q_{func} ;

Pré processamento

1. $\mathbf{X} \leftarrow \text{STFT}(\mathbf{x}(t), Q, Q_{hop}, Q_{func})$ em que $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_I\}$.

Otimização

Inicialização: $\mathcal{W} = \{\mathbf{W}_1 \leftarrow \mathbf{I}, \dots, \mathbf{W}_I \leftarrow \mathbf{I}\}$;

Para cada componente i de frequência e fonte n :

2. Calcular: $\mathbf{y}_{i,n} = \mathbf{w}_{i,n}^H \mathbf{x}_i$ e $g(\mathbf{y}_n) \leftarrow \mathbf{y}_n / (2\sqrt{\sum_{i=1}^I |\mathbf{y}_{i,n}|^2})$;
3. Atualizar: $\mathbf{W}_i \leftarrow \mathbf{W}_i - \eta[\mathbb{E}\{g(\mathbf{y}_n)\mathbf{y}_n^H\} - \mathbf{I}]\mathbf{W}_i$
4. Se $\mathbf{W}_i$ não houver convergido ou limite de iterações não alcançado, voltar no passo 2.

Pós processamento

5. Calcular separação: $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_I | \mathcal{W}, \mathbf{X}\}$.
 6. Corrigir escala: $\mathcal{W} \leftarrow \text{escal}(\mathcal{W})$.
 7. Calcular separação sem ambiguidade: $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_I | \mathcal{W}, \mathbf{X}\}$.
 8. Calcular sinais separados no tempo: $\mathbf{y}(t) \leftarrow \text{ISTFT}(\mathbf{Y}, Q, Q_{hop}, Q_{func})$
-

3.5.8 fastIVA

A IVA por máxima verossimilhança otimizada por iteração de ponto fixo é apresentada pelo Algoritmo 8 [Lee et al., 2007, Sawada et al., 2018].

Algoritmo 8: IVA por Máxima Verossimilhança - Iteração por Ponto Fixo

Entrada: Misturas $\mathbf{x}(t)$

Hiperparâmetros: Q ; Q_{hop} ; Q_{func} ;

Pré processamento

1. $\mathbf{X} \leftarrow \text{STFT}(\mathbf{x}(t), Q, Q_{hop}, Q_{func})$ em que $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_I\}$.
2. Para cada componente i de frequência:
Centralizar os dados $\mathbf{x}_i$.
Descorrelacionar os dados pelo PCA: $\mathbf{z}_i \leftarrow \mathbf{V}_i \mathbf{x}_i$.

Otimização

Inicialização: $\mathcal{U} = \{\mathbf{U}_1 \leftarrow \mathbf{I}, \dots, \mathbf{U}_I \leftarrow \mathbf{I}\}$;

Para cada componente de frequência i e fonte n :

3. Calcular separação: $\mathbf{y}_{i,n} \leftarrow \mathbf{u}_{i,n}^H \mathbf{z}_i$
4. Calcular valor da derivada do contraste: $g(\mathbf{y}_n) \leftarrow 1/(2\sqrt{\sum_{i=1}^I |\mathbf{y}_{i,n}|^2})$
5. Calcular valor da derivada segunda do contraste:
 $g'(\mathbf{y}_n) \leftarrow -1/\left(4\left(\sqrt{\sum_{i=1}^I |\mathbf{y}_{i,n}|^2}\right)^3\right)$
6. Atualização: $\mathbf{u}_{i,n} \leftarrow \mathbb{E}\{g(\mathbf{y}_n) + \|\mathbf{y}_n\|^2 g'(\mathbf{y}_n)\} \mathbf{u}_{i,n} - \mathbb{E}\{\mathbf{y}_{i,n}^* g'(\mathbf{y}_n) \mathbf{z}_i\}$
7. Ortogonalizar:
$$\mathbf{U}_i \leftarrow (\mathbf{U}_i \mathbf{U}_i^H)^{-1/2} \mathbf{U}_i$$
8. Normalizar: $\mathbf{u}_{i,n} \leftarrow \mathbf{u}_{i,n}/\|\mathbf{u}_{i,n}\|$
9. Se $\mathbf{U}_i$ não houver convergido ou limite de iterações não alcançado, voltar no passo 3.

Pós processamento

9. Remapear as matrizes de separação: $\mathcal{W} \leftarrow \{\mathbf{W}_1 \leftarrow \mathbf{U}_1 \mathbf{V}_1, \dots, \mathbf{W}_I \leftarrow \mathbf{U}_I \mathbf{V}_I\}$
 10. Calcular separação das misturas originais: $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_I | \mathcal{W}, \mathbf{X}\}$.
 11. Corrigir escala: $\mathcal{W} \leftarrow \text{escal}(\mathcal{W})$.
 12. Calcular separação sem ambiguidades: $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_I | \mathcal{W}, \mathbf{X}\}$.
 13. Calcular sinais temporais separados: $\mathbf{y}(t) \leftarrow \text{ISTFT}(\mathbf{Y}, Q, Q_{hop}, Q_{func})$
-

3.5.9 IVAaux

A IVA por máxima verossimilhança otimizada pelo gradiente natural é apresentada pelo Algoritmo 9 Ono [2011], Sawada et al. [2018].

Algoritmo 9: IVA por Máxima Verossimilhança - Função Auxiliar

Entrada: Misturas $\mathbf{x}(t)$

Hiperparâmetros: Q ; Q_{hop} ; Q_{func} ;

Pré processamento

1. $\mathbf{X} \leftarrow \text{STFT}(\mathbf{x}(t), Q, Q_{hop}, Q_{func})$ em que $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_I\}$.

Otimização

Inicialização: $\mathcal{W} = \{\mathbf{W}_1 \leftarrow \mathbf{I}, \dots, \mathbf{W}_I \leftarrow \mathbf{I}\}$;

Para cada componente de frequência i e fonte n :

2. Calcular separação: $\mathbf{y}_{i,n}(j) \leftarrow \mathbf{w}_{i,n}^H \mathbf{x}_i(j)$
3. Calcular atividade da fonte (variável auxiliar):

$$r_{j,n} \leftarrow \sqrt{\sum_{i=1}^I |\mathbf{w}_{i,n}^H \mathbf{x}_{i,j}|^2}$$

4. Calcular matriz de covariância ponderada de cada fonte:

$$\mathbf{C}_{i,n} \leftarrow \mathbb{E} \left[\frac{1}{2r_{j,n}} \mathbf{x}_i \mathbf{x}_i^H \right]$$

5. Resolver o problema HEAD (diagonalização conjunta) para $\mathbf{C}_{i,1}, \dots, \mathbf{C}_{i,N}$, com $\mathbf{e}_n$ sendo o vetor unitário com o n -ésimo elemento igual a 1.

$$\mathbf{w}_{i,n} \leftarrow (\mathbf{W}_i \mathbf{C}_{i,n})^{-1} \mathbf{e}_n$$

6. Atualização: $\mathbf{w}_{i,n} \leftarrow \mathbf{w}_{i,n} / \sqrt{\mathbf{w}_{i,n}^H \mathbf{C}_{i,n} \mathbf{w}_{i,n}}$

7. Se $\mathbf{W}_i$ não houver convergido ou limite de iterações não alcançado, voltar no passo 2.

Pós processamento

7. Calcular separação das misturas originais: $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_I | \mathcal{W}, \mathbf{X}\}$.
 8. Corrigir escala: $\mathcal{W} \leftarrow \text{escal}(\mathcal{W})$.
 9. Calcular separação sem ambiguidades: $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_I | \mathcal{W}, \mathbf{X}\}$.
 10. Calcular sinais temporais separados: $\mathbf{y}(t) \leftarrow \text{ISTFT}(\mathbf{Y}, Q, Q_{hop}, Q_{func})$
-

3.6 Medidas de Desempenho

A avaliação dos algoritmos de separação pode ser realizada de várias formas. Uma abordagem simples consiste em comparar auditivamente os sinais separados com a mistura gravada. Embora essa estratégia não invalide o resultado, ela é subjetiva, pois depende da percepção de cada avaliador. Como consequência, apresenta baixa reprodutibilidade e reduz o rigor da análise.

As medidas de desempenho têm como objetivo quantificar o resultado de um método de separação. As métricas são específicas para dois cenários diferentes, os quais se distinguem de acordo com os sinais envolvidos no cálculo. No primeiro caso, os sinais disponíveis para comparação são os sinais das fontes (originais) e os sinais separados. No segundo caso, a comparação é realizada entre as próprias misturas e os sinais separados. Em geral, a última abordagem é utilizada para os casos em que os sinais das fontes são desconhecidos. Entretanto, o primeiro caso foi adotado neste trabalho por permitir um estudo mais controlado e detalhado das propriedades do método, permitindo uma avaliação mais precisa de seu desempenho. Em alguns trabalhos, são apresentadas métricas usadas nos casos em que não se tem acesso aos sinais originais [Makino et al., 2007].

3.6.1 Medidas Objetivas

As medidas de desempenho objetivas são aquelas calculadas sem parâmetros subjetivos envolvidos, de modo que são medidas exatas e completamente baseadas nos sinais. Na área de separação de sinais de fala, o sinal recuperado pode apresentar diferentes tipos de degradação, como ruído, interferências de outras fontes de fala e artefatos introduzidos por etapas do processamento, por exemplo, o janelamento, processamento em banda estreita, convolução circular devido à truncagem da resposta ao impulso ou erros na estimação da matriz de separação. De forma geral, todas essas degradações são agrupadas sob o termo *distorção*.

Ao longo dos anos, três métricas principais foram empregadas para avaliar o desempenho de algoritmos de separação: a Razão Sinal-Interferência (SIR - *Signal-to-Interference Ratio*), a Razão Sinal-Artefato (SAR - *Signal-to-Artifact Ratio*) e a Razão Sinal-Distorção (SDR - *Signal-to-Distortion Ratio*), sendo esta última uma combinação das duas primeiras [Vincent et al., 2006].

Posteriormente, essas métricas foram reformuladas em versões invariantes à escala (SI - *Scale Invariant*), resultando nas medidas SI-SIR, SI-SAR e SI-SDR [LeRoux et al., 2019].

A SDR quantifica a razão entre o sinal original e todas as fontes de distorções possíveis, dentre essas: ruído, artefatos e interferências. Dado que o sinal original é s , o sinal separado é $\hat{s}$ e $e_{alvo} = \alpha s$, afirma-se que s e $\hat{s}$ se diferem por um termo residual de interferência, ruído e artefato, ou seja, $e_{res} = e_{interf} + e_{artf}$ (ruído incluso em artefatos)

[LeRoux et al., 2019]. Sendo assim, seu cálculo é definido por

$$SDR = 10 \log_{10} \left\{ \frac{\|e_{alvo}\|^2}{\|e_{res}\|^2} \right\}. \quad (3.13)$$

Similarmente, a SIR e a SAR são as razões que relacionam a potência do sinal original com a potência da interferência (e_{interf}) e dos artefatos (e_{artf}), respectivamente. Na separação de fontes, a SIR pode ser interpretada como a quantificação do grau de interferência que uma fonte tem sobre a estimativa de outra fonte, após o processo de separação. Por outro lado, a SDR além de quantificar essa interferência, ainda incorpora a medição do grau de corrompimento do sinal alvo pelos artefatos e ruído. Em geral, não se conhece a informação do ruído, o qual é atribuído aos artefatos. Sendo assim, artefatos é tudo o que corrompe o sinal, com exceção das interferências. Portanto, ao calcular a SDR no problema de separação, calcula-se a qualidade do sinal separado.

Neste trabalho, utilizou-se a SIR por ser uma métrica que avalia a separação dos sinais e a SI-SDR, pois essa quantifica a qualidade dos sinais separados [Vincent et al., 2006]. Ao longo deste trabalho, o acrônimo SDR refere-se ao SI-SDR, enquanto a SIR refere-se à versão variante à escala.

3.6.2 Medidas Subjetivas

No contexto de separação de sinais de fala, o sinal separado pode passar por uma avaliação da inteligibilidade. Assim, um ouvinte pode atribuir uma nota de 0 a 5, em que 5 e 4 é uma fala de alta qualidade; 3 a 4 é suficiente para a comunicação; 1,5 a 3 é pobre em inteligibilidade, e abaixo disso a fala é ruim e incompreensível [Huang et al., 2001]. Ao realizar esse teste com 20 a 60 sujeitos não treinados, obtém-se a média dos valores, o que é chamado de Pontuação Média de Opinião (MOS - *Mean Opinion Score*).

A realização de avaliações da maneira proposta é inviável, pois, para um único trecho de fala, são necessárias a opinião de vários ouvintes. Devido a esse impedimento, em 1994, o Setor de Padronização de Telecomunicação da União Internacional de Telecomunicações (ITU-T - *International Telecommunication Union - Telecommunication Standardization Sector*) criou a Medida Perceptiva da Qualidade da Fala (PSQM - *Perceptual Speech Quality Measure*) para avaliação da inteligibilidade da fala pela escala Bark [Zwicker, 1961]. Essa escala psicoacústica modela a percepção auditiva humana, de modo que a faixa audível (20 Hz a 20 kHz) é dividida em 24 bandas de frequências críticas, que representam a distinção sonora realizada pelo ouvido interno humano. Em 2001, a UIT-T substituiu o PSQM pela Avaliação Perceptiva da Qualidade da Fala (PESQ - *Perceptual Evaluation of Speech Quality*) [Rix et al., 2001]. Os detalhes da implementação e funcionamento da PESQ está fora do escopo deste trabalho. No entanto, é importante mencionar que pela escala Bark, foi possível a criação de dois métodos PESQ: o PESQ8 e o PESQ16. O primeiro

avalia a inteligibilidade da fala para sinais amostrados até 8 kHz e o segundo, até 16 kHz.

Apesar desta seção apresentar metodologias subjetivas para avaliação dos métodos de separação cega da fala, apenas o MOS é realmente subjetivo. O PESQ é uma medida objetiva, com parâmetros que aproximam à audição humana (escala Bark), fazendo com que seu resultado seja próximo do MOS para um grupo real de avaliadores. Portanto, os valores resultantes da PESQ são os mesmos da MOS, como foi apresentado anteriormente. Assim, ao ser aplicada a um determinado sinal de fala, a PESQ assume valores mais elevados quando o sinal apresenta alta inteligibilidade e valores mais baixos quando a inteligibilidade é reduzida.

3.7 Conclusão do capítulo

Neste capítulo foram abordadas as principais técnicas de separação de sinais em misturas instantâneas e convolutivas, incluindo suas bases matemáticas. Foram discutidos os métodos de otimização empregados e como diferentes escolhas de abordagem e otimização conduzem a distintos algoritmos de separação cega da fala. Por fim, foram apresentadas as medidas de desempenho utilizadas para avaliar os algoritmos, permitindo assim as análises experimentais apresentadas no Capítulo 5.

Capítulo 4

Metodologia

Neste capítulo, são descritas inicialmente todas as etapas para obtenção dos sinais de fala, bem como a gravação e edição dos áudios para os experimentos. Logo após, aborda-se a seleção das salas, a gravação das misturas e os equipamentos utilizados nos ensaios. O processo de estimação do tempo de reverberação de cada sala é apresentado em seguida. As aplicações dos algoritmos de separação da fala com os parâmetros específicos e as métricas de avaliação são apresentadas ao final.

4.1 Condições Experimentais

Previamente a qualquer etapa prática deste trabalho, optou-se por fixar alguns parâmetros experimentais. Por questões de simplicidade e capacidade de generalização, o número de fontes (N) é igual ao de microfones (M), sendo assim $N = M = 2$.

4.2 Etapas pré experimento

4.2.1 Geração de Frases

A pronúncia de uma determinada frase, ainda que dita pelo mesmo locutor, pode sofrer variações de velocidade, volume e entonação [Ladefoged, 2001]. Para evitar essa variabilidade, gravou-se as frases e as reproduziu nos alto-falantes durante os experimentos. Esse procedimento é importante na etapa experimental, porque a variabilidade da fala pode alterar o desempenho dos métodos e, com isso, comprometer a análise dos resultados. As elocuições gravadas foram frases obtidas por meio do ChatGPT3, uma ferramenta de inteligência artificial criada pela OpenIA [OpenAI, 2023]. A mensagem de *prompt* solicitou ao agente gerar frases curtas explicando o que era um determinado assunto relacionado ao trabalho, os quais foram: áudio, Processamento Digital de Sinais (DSP - *Digital Signal Processing*), ICA e Fatoração de Matrizes Não-negativas (NMF - *Nonnegative Matrix Factorization*). Assim, as frases resultantes foram as seguintes:

- 1 - “Áudio refere-se aos sinais sonoros capturados, armazenados, transmitidos ou reproduzidos eletronicamente, usados para comunicação, entretenimento ou outras finalidades.”
- 2 - “Processamento de sinais envolve a análise e manipulação de dados para extrair informações úteis, ou melhorar a qualidade dos sinais.”
- 3 - “A análise de componentes independentes é uma técnica que separa um conjunto de sinais misturados, em componentes que são estatisticamente independentes.”
- 4 - “NMF é uma técnica de decomposição matricial, a qual representa dados como a combinação de matrizes não negativas, útil em aplicações como processamento de áudio e imagem.”

As frases acima foram gravadas por dois locutores: um masculino e outro feminino. A gravação foi realizada por um aplicativo do *smartphone* modelo A22 da Samsung, no qual foi possível configurar a frequência de amostragem para 44100 Hz. A Tabela 1 apresenta o nome de cada elocução. Além disso, antes de realizar a gravação, foi aferido o nível de ruído do ambiente usando tanto um medidor de nível de pressão sonora (SPL - *Sound Pressure Level*) da Minipa [Minipa], quanto um aplicativo da Splend Apps. A gravação ocorreu durante a madrugada em uma sala fechada por cortinas e microfone próximo da boca, de modo que o ruído ambiente e a reverberação afetassem minimamente a gravação. O nível de ruído medido pelo medidor foi de 29 dB SPL, e no aplicativo, 31 dB SPL em média. As medições de nível do ruído ambiente apresentadas no restante do capítulo foram obtidas somente pelo aplicativo.

Tabela 1 – Nome de cada elocução (sinais das fontes).

Nome	Elocução (Sinais das fontes)
aM	Elocução sobre “áudio” do falante masculino
aF	Elocução sobre “áudio” do falante feminino
dM	Elocução sobre “DSP” do falante masculino
dF	Elocução sobre “DSP” do falante feminino
iM	Elocução sobre “ICA” do falante masculino
iF	Elocução sobre “ICA” do falante feminino
nM	Elocução sobre “NMF” do falante masculino
nF	Elocução sobre “NMF” do falante feminino

4.2.2 Edição

As oito elocuições obtidas isoladamente foram utilizadas para construir os estímulos usados no experimento de gravação das misturas, descrito nas próximas subseções. Como este trabalho trata do caso determinado 2×2 (i.e., duas fontes e dois microfones), foi necessário definir pares de elocuições para serem usadas como as duas fontes de cada mistura. Para evitar problemas de sincronismo durante a reprodução das elocuições de cada par, o método adotado foi o de construir sinais de áudio estéreo onde as elocuições individuais de cada par são usadas como os canais esquerdo e direito do sinal estéreo.

Essas edições foram realizadas no *software* Audacity. Inicialmente, a tarefa foi de conversão dos arquivos da extensão “mp3” para “wav”, e depois selecionou-se o tempo de início e término de cada sentença. Apesar de os sinais originais terem sido gravados pelo aplicativo e posteriormente salvos em formato “mp3”, a Figura 13 mostra que essa compressão não afetou significativamente as componentes do sinal de fala, sendo perceptível apenas a atenuação das frequências acima de 16 kHz. Esse teste foi realizado gravando-se um sinal de fala em formato “wav”, com taxa de amostragem de 44,1 kHz, e posteriormente convertendo-o para o formato comprimido “mp3” em 64 kbps.

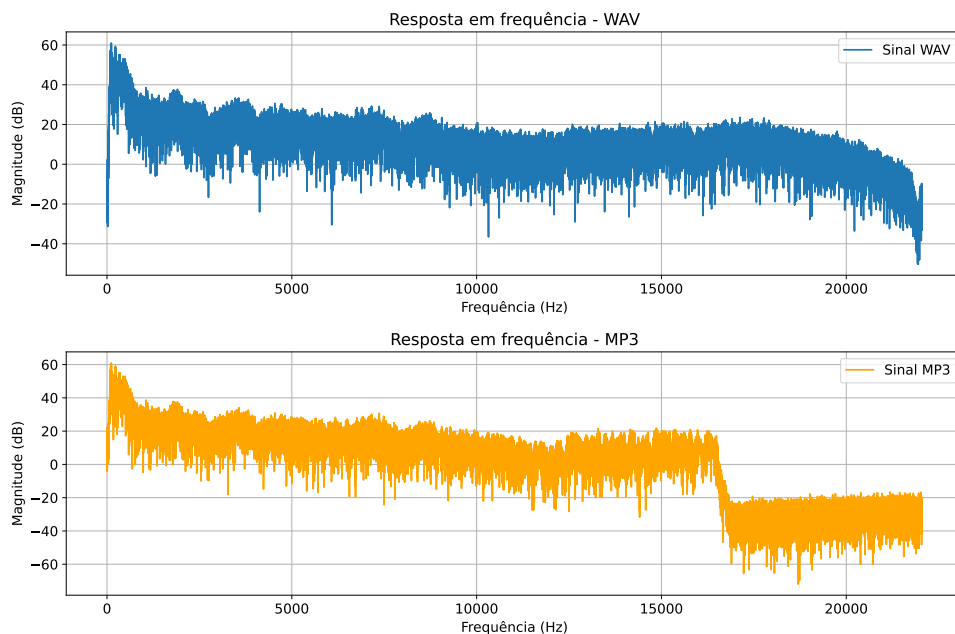


Figura 13 – Espectro de frequência do sinal de fala para avaliação da perda de informação causada pela compressão em “mp3”. Fonte: Autoria própria.

Como há 8 elocuições isoladas disponíveis, é possível definir $8 \times 8 = 64$ pares de elocuições. No entanto, apenas pares de elocuições provenientes de locutores diferentes foram definidos, uma vez que misturas de elocuições de um mesmo locutor não foram avaliadas neste trabalho. Isso reduz o número de pares possíveis para $4 \times 4 = 16$. Cada uma destas combinações foi então usada para construir um único sinal de áudio estéreo a ser usado no experimento de gravação das misturas, sendo o sujeito masculino sempre

reproduzido no canal esquerdo e o feminino no canal direito. Entre cada combinação, foi inserido um trecho de silêncio de 1 segundo, visando não apenas facilitar a edição após os experimentos, mas também anular o efeito de reverberação entre falas consecutivas. Essa etapa permitiu que os ensaios fossem contínuos, além de alcançar sincronismo fixo do instante inicial de cada combinação. O tempo total da faixa foi de 3:09 minutos, e ela é apresentada na Figura 14. Cada nome das combinações é apresentada na Tabela 2.

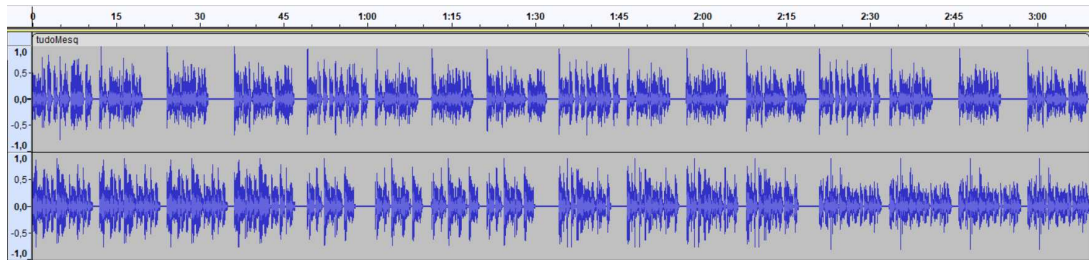


Figura 14 – Faixa de áudio estéreo das 16 combinações. A faixa do canal esquerdo (superior) é referente à fala masculina, e a do direito (inferior), feminina.
Fonte: Autoria própria.

Tabela 2 – Nome de cada combinação de elocuições.

Combinação	Nome	Falante masculino	Falante feminino
1	aMaF	Elocução “áudio”	Elocução “áudio”
2	dMaF	Elocução “DSP”	Elocução “áudio”
3	iMaF	Elocução “ICA”	Elocução “áudio”
4	nMaF	Elocução “NMF”	Elocução “áudio”
5	aMdF	Elocução “áudio”	Elocução “DSP”
6	dMdF	Elocução “DSP”	Elocução “DSP”
7	iMdF	Elocução “ICA”	Elocução “DSP”
8	nMdF	Elocução “NMF”	Elocução “DSP”
9	aMiF	Elocução “áudio”	Elocução “ICA”
10	dMiF	Elocução “DSP”	Elocução “ICA”
11	iMiF	Elocução “ICA”	Elocução “ICA”
12	nMiF	Elocução “NMF”	Elocução “ICA”
13	aMnF	Elocução “áudio”	Elocução “NMF”
14	dMnF	Elocução “DSP”	Elocução “NMF”
15	iMnF	Elocução “ICA”	Elocução “NMF”
16	nMnF	Elocução “NMF”	Elocução “NMF”

4.2.3 Modelagem Acústica

A modelagem da resposta ao impulso acústico foi realizada por meio da emissão do sinal TSP [Aoshima, 1981, Suzuki et al., 1995]. A sintetização do sinal foi feita no *software* MATLAB, por meio de um *script* [Yehia, 1998]. O sinal TSP obtido foi de 1,2 segundos com largura de banda de 20 Hz a 20000 Hz, ou seja, toda a faixa audível. Como anteriormente, o sinal TSP foi editado no Audacity, de forma que a faixa completa contivesse o pulso em ambos os canais estéreo, com a diferença de tempo de aproximadamente 1,2 segundos entre um pulso e outro. Isso permitiu que cada alto-falante reproduzisse o sinal separadamente, e todas as AIRs pudessem ser obtidas. O trecho inicial do sinal TSP é apresentado na Figura 15.

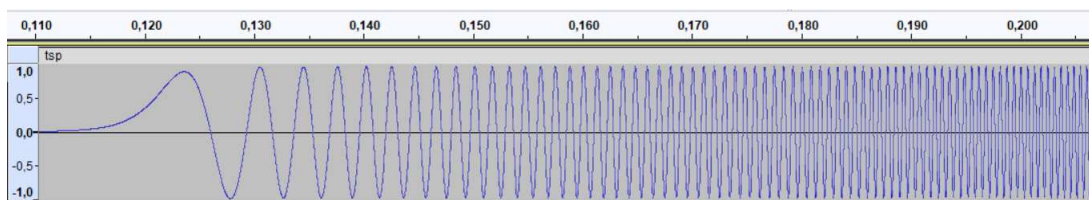


Figura 15 – Trecho inicial da faixa de áudio do sinal TSP. Fonte: Autoria própria.

4.3 Experimentos

4.3.1 Ambientes

Os experimentos foram conduzidos em quatro ambientes diferentes na Escola de Engenharia da UFMG. O primeiro foi o estúdio do CEFALA mostrado na Figura 16, uma sala com isolamento acústica, acarpetado nas laterais e no chão, além de espuma acústica no teto. O segundo local foi no próprio laboratório CEFALA (Figura 17), uma sala em formato de L sem tratamento acústico. O terceiro ambiente foi a sala 2419, apresentada na Figura 18, uma sala de aula com poucos elementos absorvedores. Por último, a sala de seminários 1012 (Figura 19). Nela há um pouco de tratamento acústico, sendo chão com carpete, teto com espuma absorvedora e cortinas. Todos esses ambientes possuem um determinado tempo de reverberação, o qual foi estimado e é apresentado em uma seção adiante. Como cada ambiente possui valor diferente nesse parâmetro acústico, os desempenhos dos métodos de separação puderam ser avaliados para cada sala.

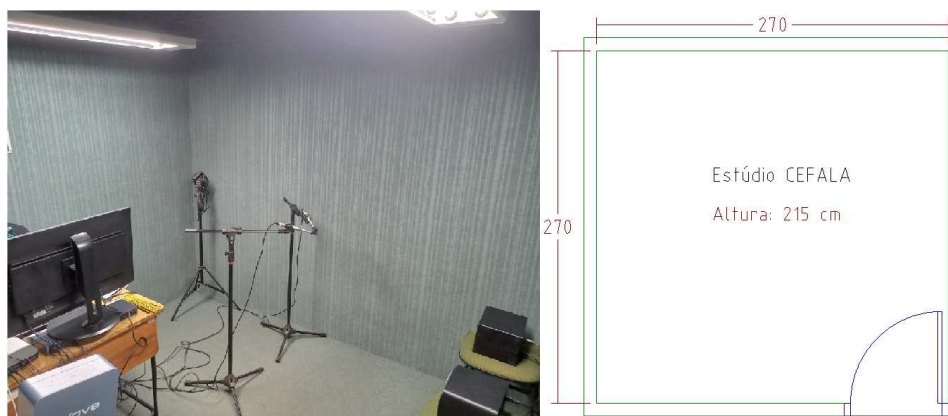


Figura 16 – Foto do estúdio (esquerda) e sua planta com as dimensões em centímetros (direita). Fonte: Autoria própria.

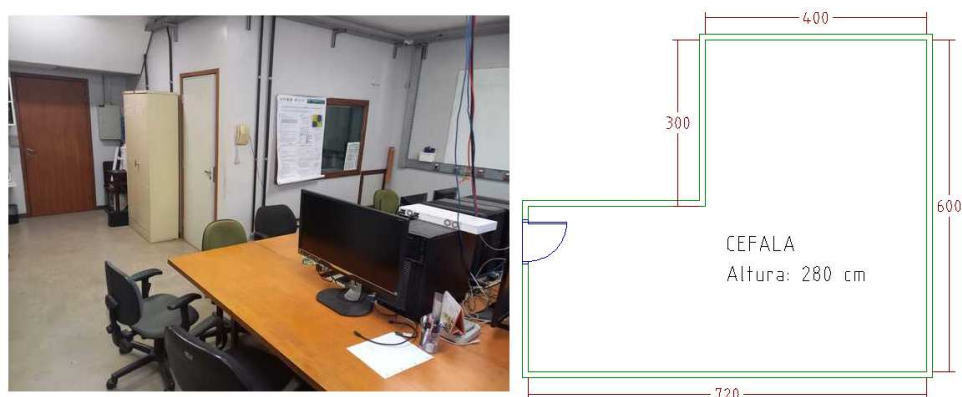


Figura 17 – Foto do laboratório CEFALA (esquerda) e sua planta com as dimensões em centímetros (direita). Fonte: Autoria própria.

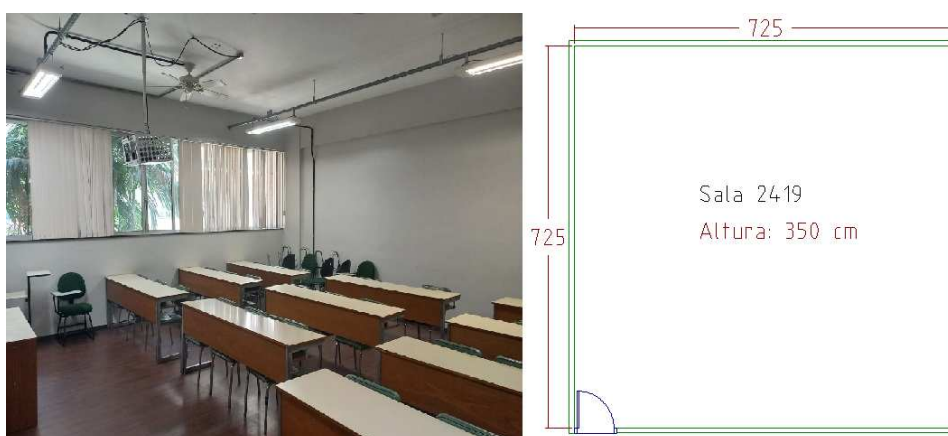


Figura 18 – Foto da sala 2419 (esquerda) e sua planta com as dimensões em centímetros (direita). Fonte: Autoria própria.

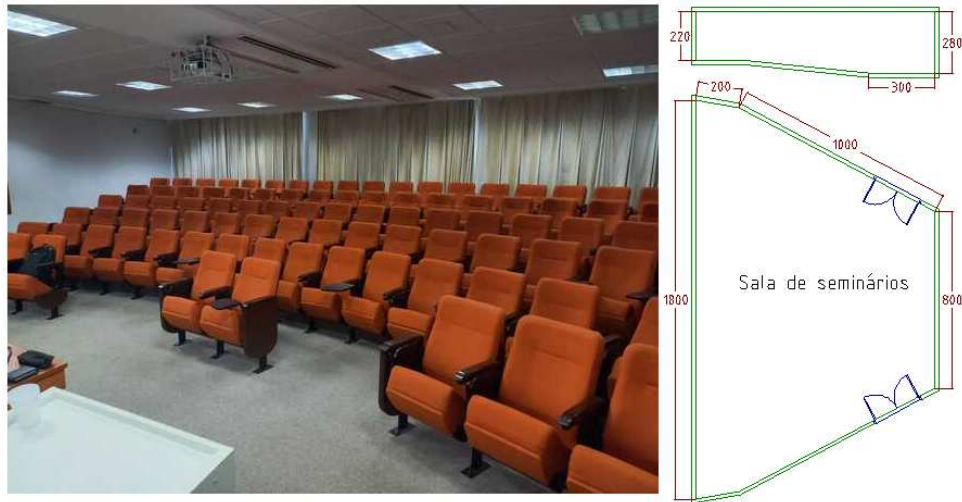


Figura 19 – Foto da sala de seminários (esquerda) e sua vista lateral (direita superior) e planta (direita inferior) com as dimensões em centímetros. Fonte: Autoria própria.

4.3.2 Equipamentos

A realização dos experimentos demandou equipamentos do laboratório CEFALA, e de uso pessoal. Como a quantidade de fontes deste trabalho é igual a 2, foi utilizada a mesma quantidade de monitores de áudio, os quais possuem a resposta em frequência mais plana do que as caixas de som simples. As falas masculinas foram reproduzidas arbitrariamente pelo monitor da M-Audio modelo Studiophile BX5 (Figura 20-esquerda), com resposta em frequência de 56 Hz a 20 kHz segundo o manual. As falas femininas foram emitidas pelo monitor da M-Audio modelo BX5a (Figura 20-direita), com resposta em frequência de 56 Hz a 22 kHz.



Figura 20 – Monitores de áudio usados nos experimentos. À esquerda o monitor usado para falas masculinas, à direita o que foi usado para falas femininas. Fonte: Autoria própria.

Ambos os monitores foram conectados ao *laptop* pessoal por um cabo com conexões P2 estéreo e dois P10 mono Figura 21, de modo que a primeira extremidade fosse conectada no computador, e as outras duas nos monitores. Dessa forma foi possível reproduzir as

faixas mencionadas nas seções anteriores (fala e TSP), em que no canal esquerdo foi a fala masculina e no direito, a fala feminina.



Figura 21 – Cabo P2 estéreo 2xP10 mono. Fonte:[Cabos, 2024].

O sistema de captação consistiu em dois microfones omnidirecionais Behringer ECM8000 Figura 22, com diagrama polar e resposta em frequência apresentados na Figura 23. Ambos os microfones foram conectados via cabos P10 mono a uma interface de áudio da Presonus modelo Firestudio Project Figura 24, disponível no laboratório.



Figura 22 – Microfone usado nos experimentos. Fonte: Autoria própria.

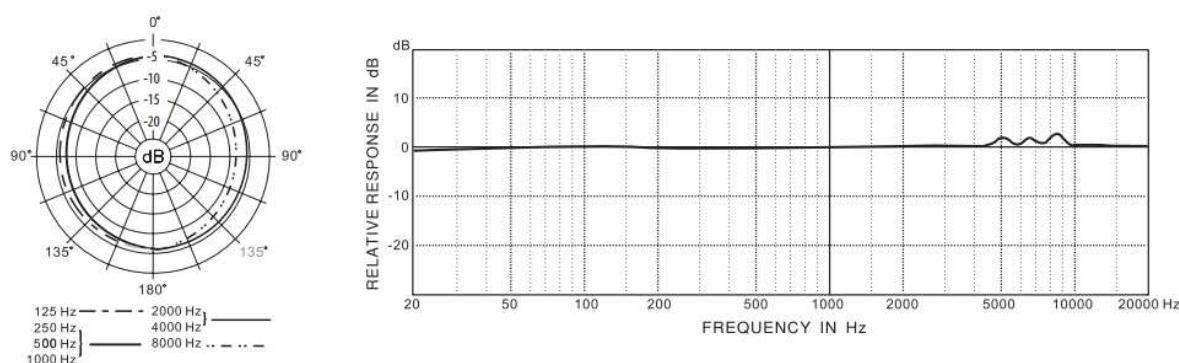


Figura 23 – Diagrama polar e resposta em frequência do microfone da Figura 22. Fonte: [Behringer, 2024].

Por fim, a interface de áudio foi conectada ao Mini MAC do CEFALA pelo cabo *FireWire* (Figura 25), responsável por transmitir os sinais ao Audacity, *software* o qual foi usado para gravação.



Figura 24 – Interface de áudio usada. Fonte: [Presonus, 2024].



Figura 25 – Cabo *FireWire* utilizado na comunicação entre interface e computador. Fonte: [Shopee, 2024].

4.3.3 Ensaios

A Figura 26 apresenta a disposição geral dos equipamentos. Durante a execução de cada experimento, os alto-falantes (fontes) ficaram a uma distância fixa de $L_1 = 57$ cm, e ambos a 55 cm do chão. Entretanto, para avaliar o desempenho dos métodos de separação e como o posicionamento dos microfones interfere no desempenho, a distância entre os microfones e a distância destes em relação às fontes foram alteradas.

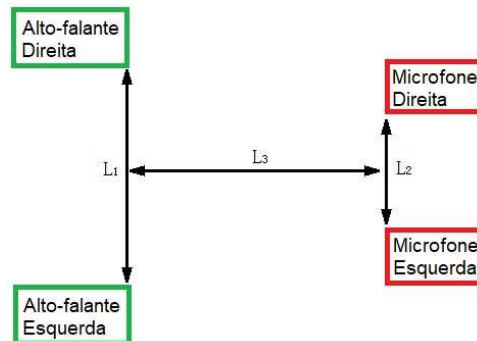


Figura 26 – Disposição geral entre os equipamentos durante o experimento. Fonte: Autoria própria.

A distância entre os microfones correspondeu a três valores possíveis de L_2 : 0 cm (orientação 90°), 4 cm e 57 cm. O primeiro valor refere-se às misturas altamente correlacionadas (aproximadamente o mesmo ponto espacial), com microfones orientados ortogonalmente, o que resulta em diferentes atenuações no sinal captado. O segundo caso é abordado em vários estudos [Sawada et al., 2018], e permite incorporar métodos de separação em sistemas embarcados. O último caso foi simplesmente a situação de microfones alinhados com cada monitor de áudio.

Embora os microfones sejam omnidirecionais e, portanto, pouco sensíveis à orientação, o arranjo em 90° foi adotado para verificar se, mesmo estando aproximadamente no mesmo ponto espacial, poderiam ocorrer diferenças sutis nas misturas captadas, capazes de influenciar os métodos de separação cega da fala.

A distância dos microfones em relação às fontes foi a medida direta entre o microfone e o monitor de áudio do mesmo lado (esquerdo ou direito). Os valores da distância L_3 puderam ser separados em três categorias diferentes: longa distância ($L_3 > 2$ m); média distância ($1 \text{ m} \leq L_3 \leq 2 \text{ m}$); e curta distância ($L_3 < 1 \text{ m}$). Ao calcular a média dos valores de L_3 em todos os ensaios e salas, obtêm-se valores médios aproximados de 210 cm para a longa distância, 140 cm para a média distância e 75 cm para a curta distância. Assim, a avaliação de desempenho dos métodos de separação para a proximidade dos microfones em relação às fontes pôde ser realizada. Ao longo deste trabalho, cada categoria foi referida pelo valor médio correspondente.

Alguns pontos importantes devem ser destacados. Primeiro, a distância específica de 57 cm entre os monitores de áudio foi atribuída devido ao primeiro experimento realizado no estúdio do CEFALA. Por ser um ambiente pequeno, houve restrição na distância entre as caixas de som, e assim, manteve-se esse valor. Isso leva ao segundo ponto: esse foi o único ambiente em que houve alteração do valor da categoria de longa distância, em razão das restrições físicas da sala. Por fim, ressalta-se que a distância dos microfones para o chão foi de aproximadamente 1 m para todos os experimentos.

A combinação de cada valor dos parâmetros espaciais levou à execução de nove ensaios diferentes. Assim, para cada mudança, seja de distância entre microfones e/ou de distância entre microfones e monitores de áudio, foi reproduzido o sinal estéreo com as falas e o sinal TSP. O primeiro teve como objetivo obter as duas misturas, e o segundo, permitiu obter as quatro AIRs entre cada emissor e receptor. As tabelas 3, 4, 5, e 6 resumem todos os ensaios executados em cada ambiente, e a Tabela 7 apresenta o nome de cada ensaio.

Tabela 3 – Parâmetros físicos dos ensaios realizados no estúdio.

Estúdio	Parâmetros Experimentais								
Ensaio	1	2	3	4	5	6	7	8	9
L_3 : Mic.-Fonte (cm)	159	153	148	102	100	96	65	62	52
L_2 : Mic.-Mic. (cm)	0 (90°)	4	57	0 (90°)	4	57	0 (90°)	4	57
Mic.-Chão (cm)	104	103	103	103	102	101	103	102	103
Fonte-Chão (cm)	55								
L_1 : Fonte-Fonte (cm)	57								
Ruído Ambiente	30 dB SPL								

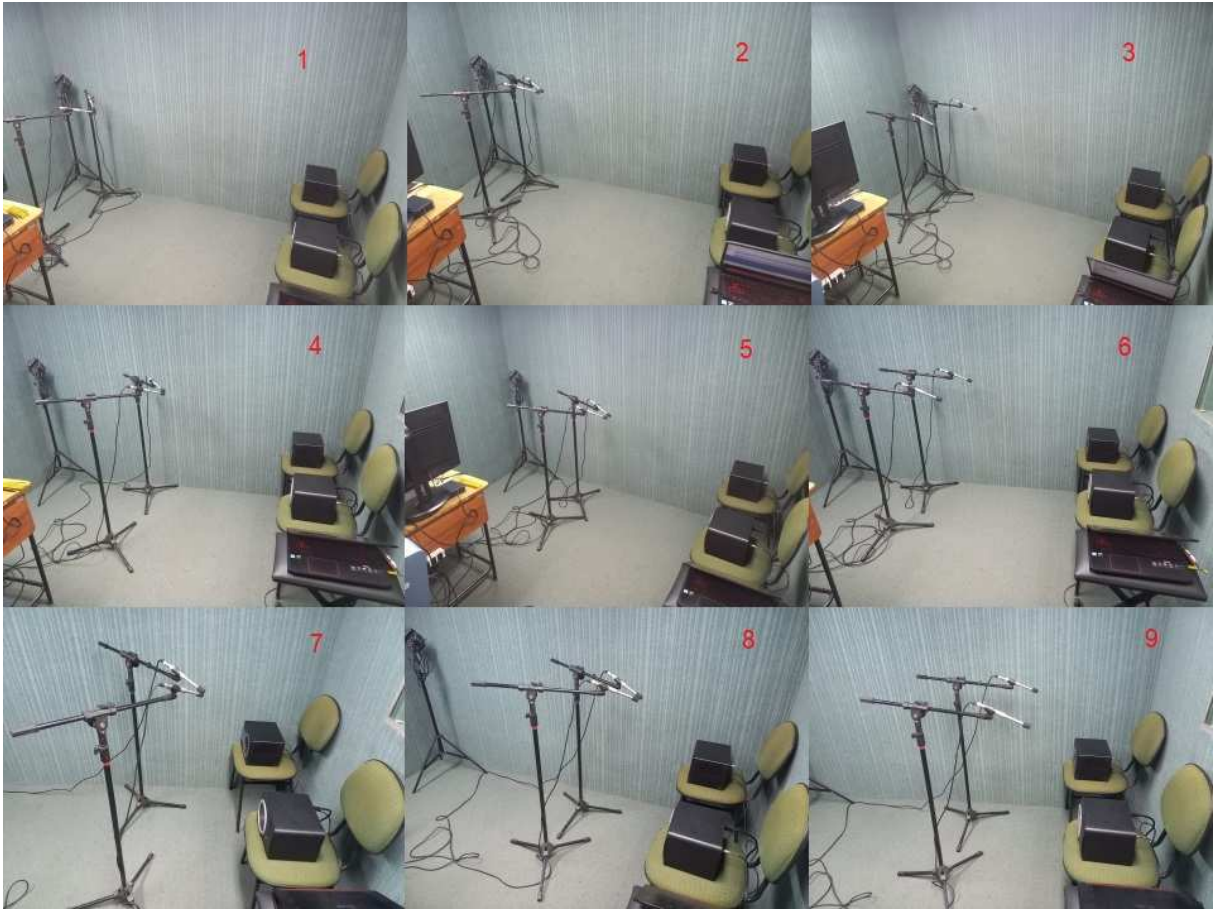


Figura 27 – Foto de cada ensaio realizado no estúdio com suas respectivas numerações, em vermelho. Fonte: Autoria própria.

Tabela 4 – Parâmetros físicos dos ensaios realizados no CEFALA.

CEFALA	Parâmetros Experimentais								
Ensaio	10	11	12	13	14	15	16	17	18
L_3 : Mic.-Fonte (cm)	225	215	205	160	160	145	80	75	75
L_2 : Mic.-Mic. (cm)	0 (90°)	4	57	0 (90°)	4	57	0 (90°)	4	57
Mic.-Chão (cm)	107								
Fonte-Chão (cm)	55								
L_1 : Fonte-Fonte (cm)	57								
Ruído Ambiente	41 dB SPL								



Figura 28 – Foto de cada ensaio realizado no CEFALA com suas respectivas numerações, em vermelho. Fonte: Autoria própria.

Tabela 5 – Parâmetros físicos dos ensaios realizados na sala 2419.

Sala 2419	Parâmetros Experimentais								
Ensaio	19	20	21	22	23	24	25	26	27
L_3 : Mic.-Fonte (cm)	235	230	225	150	155	155	86	87	78
L_2 : Mic.-Mic. (cm)	0 (90°)	4	57	0 (90°)	4	57	0 (90°)	4	57
Mic.-Chão (cm)	107								
Fonte-Chão (cm)	55								
L_1 : Fonte-Fonte (cm)	57								
Ruído Ambiente	35 dB SPL								



Figura 29 – Foto de cada ensaio realizado na sala 2419 com suas respectivas numerações, em vermelho. Fonte: Autoria própria.

Tabela 6 – Parâmetros físicos dos ensaios realizados na sala de seminários 1012.

Sala de Seminários	Parâmetros Experimentais								
Ensaio	28	29	30	31	32	33	34	35	36
L_3 : Mic.-Fonte (cm)	245	243	240	143	143	136	80	75	72
L_2 : Mic.-Mic. (cm)	0 (90°)	4	57	0 (90°)	4	57	0 (90°)	4	57
Mic.-Chão (cm)	106								
Fonte-Chão (cm)	55								
L_1 : Fonte-Fonte (cm)	57								
Ruído Ambiente	31 dB SPL								

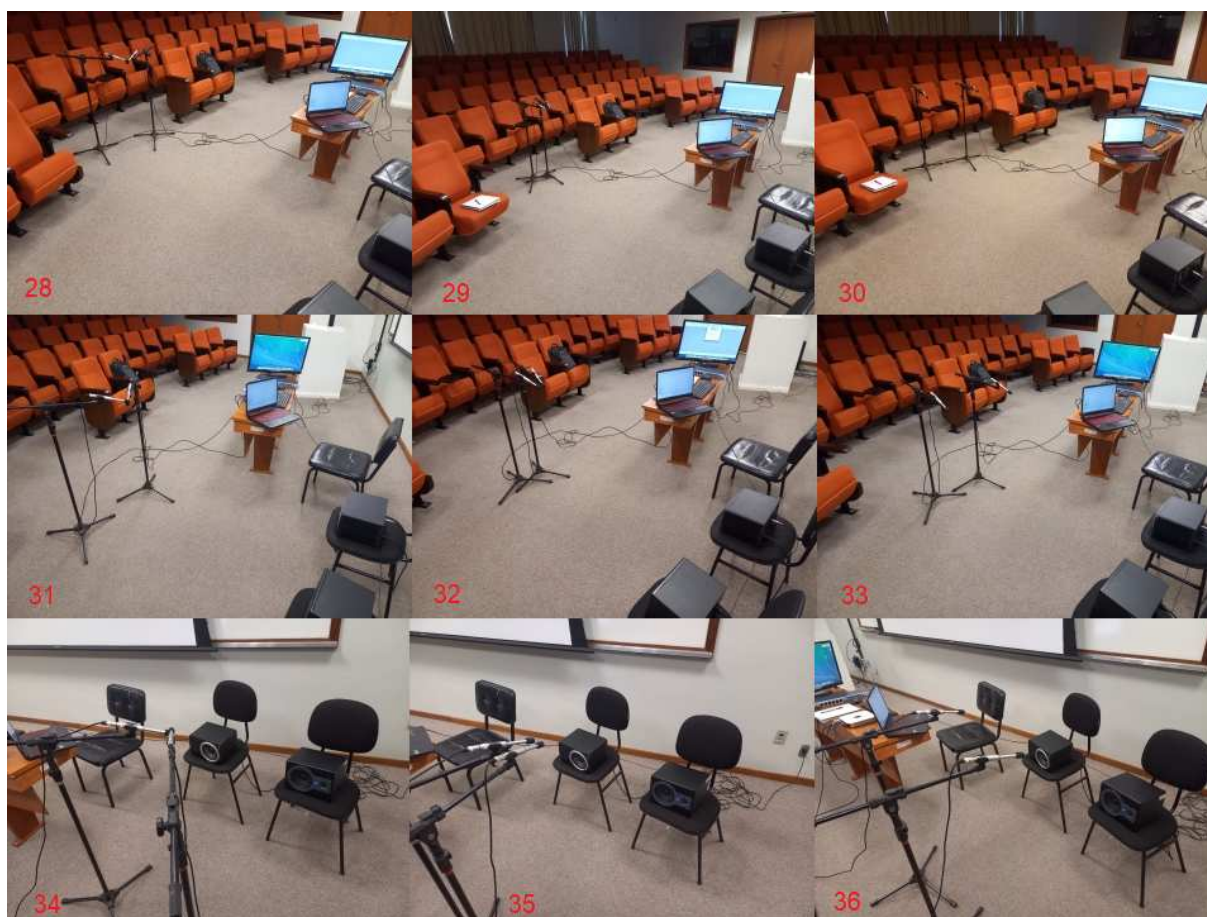


Figura 30 – Foto de cada ensaio realizado na sala de seminários com suas respectivas numerações, em vermelho. Fonte: Autoria própria.

Tabela 7 – Nome de cada um dos 9 ensaios de acordo com os valores de L_2 e L_3 .

Nome	L_2	L_3
P90	0 cm (90°)	$L_3 < 1$ m
M90	0 cm (90°)	$1 \text{ m} < L_3 < 2 \text{ m}$
L90	0 cm (90°)	$L_3 > 2$ m
P4	4 cm	$L_3 < 1$ m
M4	4 cm	$1 \text{ m} < L_3 < 2 \text{ m}$
L4	4 cm	$L_3 > 2$ m
P57	57 cm	$L_3 < 1$ m
M57	57 cm	$1 \text{ m} < L_3 < 2 \text{ m}$
L57	57 cm	$L_3 > 2$ m

Após a execução dos experimentos, os sinais misturados foram editados no Audacity. Para cada ensaio, as 16 misturas estéreo resultantes foram armazenadas para utilização na etapa de separação.

4.3.4 Estimação da Reverberação

Os sinais TSP obtidos de cada ensaio, permitiram estimar as AIRs, as quais podem ser utilizadas tanto para simular e validar o sistema de mistura, quanto para obter parâmetros acústicos do ambiente. Todas as etapas a seguir foram feitas em Python 3.9, no ambiente Jupyter Notebook [Project Jupyter, 2015].

Neste trabalho recorreu-se às AIRs calculadas pelo sinal TSP para estimar o tempo de reverberação do local de experimentação. Para cada ensaio realizado, 4 sinais TSP foram captados, totalizando 36 sinais TSP em cada sala. Cada um desses sinais foi convolvido com a versão temporalmente reversa do TSP original, obtendo-se assim a respectiva AIR [Suzuki et al., 1995].

Como cada AIR foi obtida em diferentes pontos do espaço, elas diferem entre si, mas compartilham características acústicas semelhantes por terem sido obtidas na mesma sala. Dessa forma, foi possível estimar o tempo de reverberação do ambiente. A Figura 32 apresenta quatro AIRs obtidas de ambientes diferentes.

A estimação do tempo de reverberação foi realizada com base no método proposto por Schröder [Schroeder, 1965]. Os procedimentos a seguir foram aplicados a todas as AIRs e são resumidos na Figura 31.

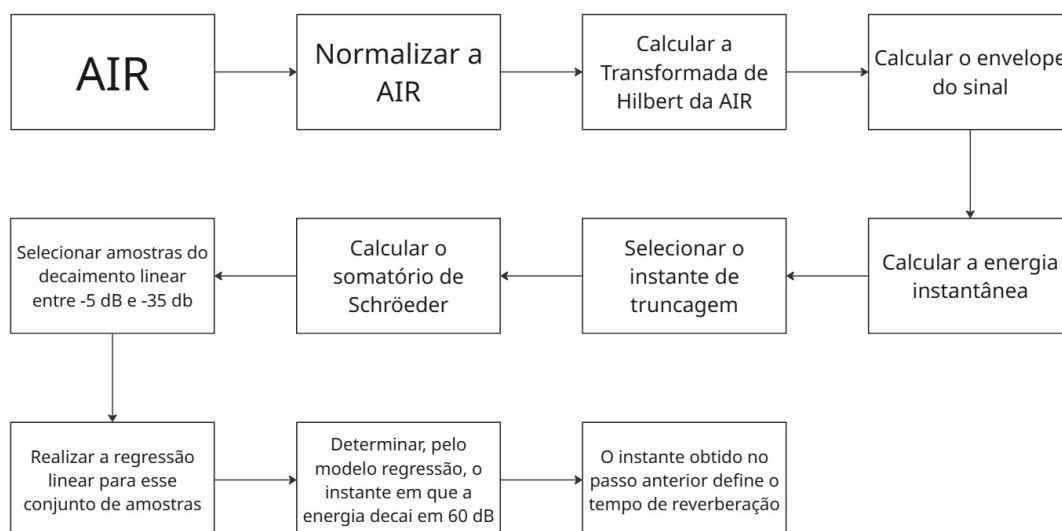


Figura 31 – Procedimento adotado para estimar o tempo de reverberação a partir de cada AIR. Fonte: Autoria própria.

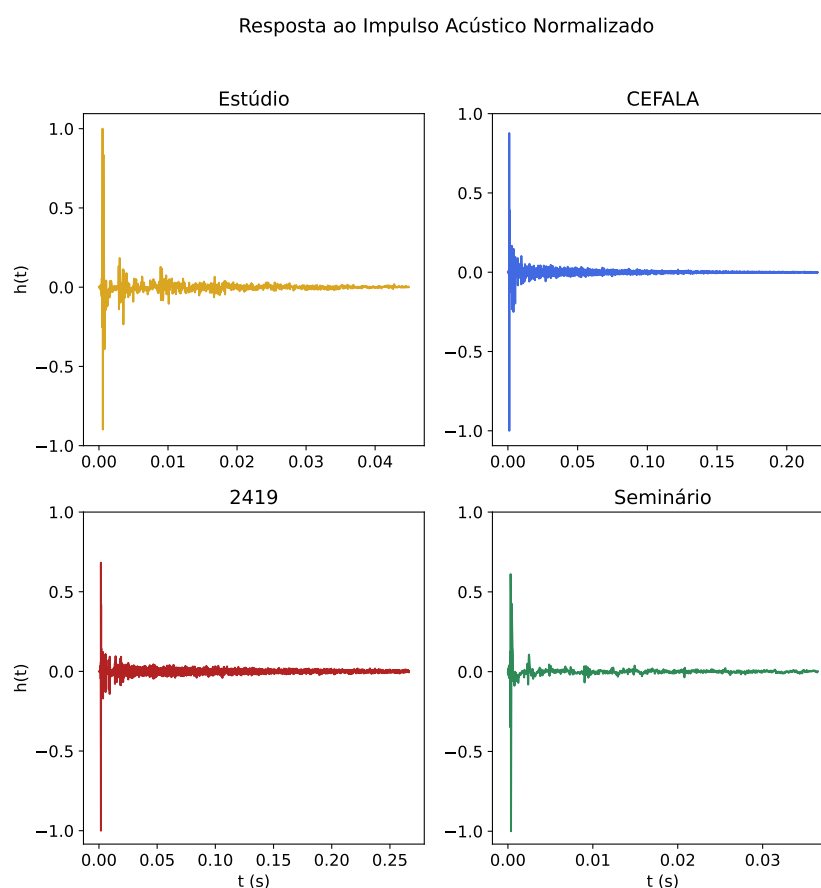


Figura 32 – AIRs normalizadas dos ensaios 6, 15, 24 e 33, as quais representam a resposta ao impulso entre monitor de áudio da direita e microfone da esquerda. Fonte: Autoria própria.

Inicialmente, normalizou-se a AIR $h(k)$ pelo seu maior valor absoluto dado por

$$h_{norm}(k) = \frac{h(k)}{\max(|h(k)|)} . \quad (4.1)$$

Depois, calculou-se a Transformada de Hilbert de $h_{norm}(k)$, resultando no sinal analítico $h_a(k)$ [Oppenheim et al., 1999]. Em seguida, obteve-se o envelope do sinal analítico ao calcular seu valor absoluto

$$h_{env}(k) = |h_a(k)| . \quad (4.2)$$

A partir do envelope, foi possível calcular a função da energia acústica instantânea [Kuttruff, 2009, International Organization for Standardization, 2008], definida por

$$E_h(k) = 10 \log \left(\frac{h_{env}^2(k)}{\max(h_{env}^2(k))} \right) , \quad (4.3)$$

a qual apresenta a energia da AIR para cada amostra. A Figura 33 apresenta essa função para cada uma das AIRs da Figura 32. Esses perfis de decaimento evidenciam que os ambientes apresentam propriedades acústicas distintas. Na sala 2419 e no CEFALA, observa-se um decaimento de energia mais lento, associado a tempos de reverberação elevados e à presença de filtros FIR mais extensos. Por outro lado, o estúdio e a sala de seminários exibem um decaimento mais rápido, coerente com tempos de reverberação reduzidos e filtros FIR com menor número de coeficientes. Dessa forma, em tarefas de separação cega de sinais, ambientes com altos tempos de reverberação tendem a dificultar a obtenção de uma boa separação [Duong et al., 2010].

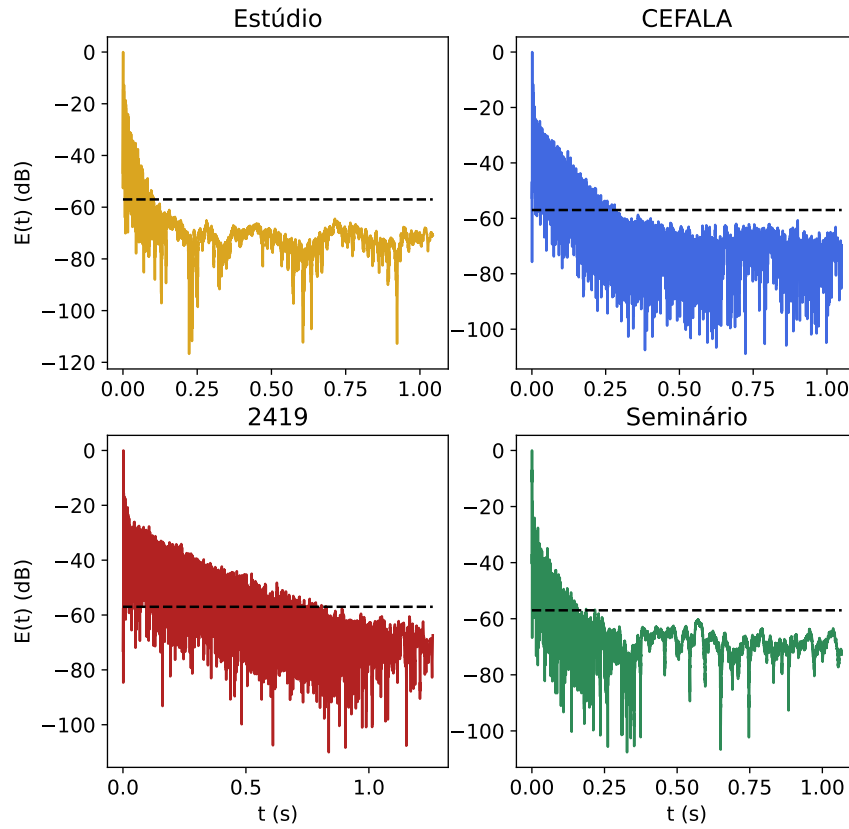


Figura 33 – Curvas de energia instantânea relativas às AIRs da Figura 32 com o nível de ruído tracejado. Fonte: Autoria própria.

Em um dado instante de tempo, a energia da AIR é igual ou menor do que a energia do ruído de medição, o qual foi aproximadamente igual a -55 dB. Portanto, deve-se analisar o instante em que $E_h(k)$ deve ser truncado. Apesar de existirem alguns métodos que realizam essa tarefa de forma automática para *softwares* de medição acústica em tempo real, cada uma das funções $E_h(k)$ de cada AIR foi analisada, e a seleção do instante de truncagem foi feita manualmente. A Figura 34 mostra diferentes instantes de truncagem da curva de energia instantânea, representados pelas linhas verticais pontilhadas de vermelho, magenta e azul.

O primeiro instante de truncagem (em vermelho) é inadequado, pois descarta grande parte de $E_h(k)$. O segundo instante (em magenta) também é inadequado, já que inclui tanto $E_h(k)$ quanto o ruído. Por outro lado, o terceiro instante (em azul) abrange aproximadamente toda a parte relevante de $E_h(k)$, excluindo do cálculo futuro o trecho corrompido pela energia do ruído de medição.

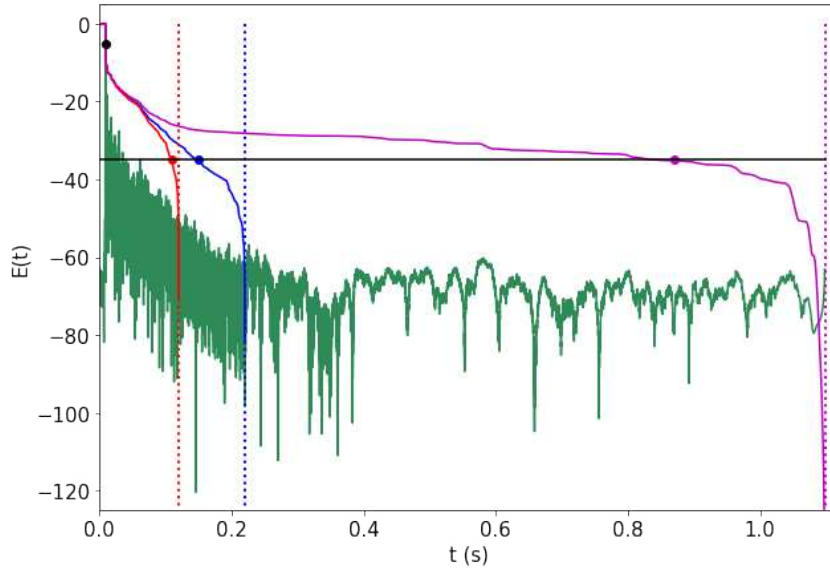


Figura 34 – Curva de energia instantânea da sala de seminários (em verde). São indicados os instantes de truncagem (linhas verticais pontilhadas) e os respectivos somatórios de Schröder, representados pelas curvas contínuas em vermelho, azul e magenta. O círculo preto indica o instante em que a energia acumulada atinge -5 dB, enquanto os demais círculos correspondem aos instantes em que cada curva atinge -35 dB, nível indicado pela linha horizontal preta. Fonte: Autoria própria.

Após a truncagem, calculou-se a integral (somatório) de Schröder [Schroeder, 1965, Kuttruff, 2009, International Organization for Standardization, 2008] definida na forma discreta por

$$E_{Sch}(k) = 10 \log_{10} \left(\frac{\sum_{i=k}^{T-1} h_{env}^2(i)}{\sum_{i=0}^{T-1} h_{env}^2(i)} \right), \quad (4.4)$$

que apresenta como a energia acumulada decai ao longo do tempo. Na Figura 34, esses decaimentos referentes a cada instante de truncagem são representados pelas linhas contínuas vermelha, azul e magenta.

De acordo com a norma ISO 3382 [International Organization for Standardization, 2008], a estimativa do tempo de reverberação requer uma faixa dinâmica mínima de 75 dB, considerando o intervalo de decaimento entre -5 dB e -65 dB, de modo a evitar as flutuações iniciais. Além disso, o nível de ruído de fundo deve ser, no máximo, -75 dB, garantindo uma diferença mínima de 10 dB em relação à faixa analisada. No entanto, a maioria dos instrumentos de medição não apresenta sensibilidade e precisão suficientes para atender a esses requisitos, o que dificulta a estimativa direta do RT_{60} .

Nos experimentos realizados, o ruído de medição foi de aproximadamente -55 dB, impossibilitando o cálculo direto do RT_{60} . Assim, empregou-se uma estimativa indireta baseada em intervalos menores de decaimento como 10 dB, 20 dB e 30 dB, os quais

correspondem a RT_{10} , RT_{20} e RT_{30} , respectivamente. Neste trabalho, optou-se pelo RT_{30} , cuja faixa dinâmica necessária é de 45 dB. Para essa estimativa, foram selecionadas as amostras de nível correspondentes ao intervalo em que a curva de energia acumulada decai de -5 dB a -35 dB (círculos preto e azul na Figura 34), e ajustou-se um modelo de regressão linear sob a premissa de que o decaimento é aproximadamente linear no logaritmo da energia acumulada, ao longo do tempo.

O valor de RT_{60} é determinado pelo instante em que o modelo de regressão linear obtido na etapa anterior atinge um nível igual ou inferior a -60 dB. Nesse ponto observa-se a importância do processo de truncagem. Se a truncagem for realizada no final da curva de energia instantânea, parte da curva conterá somente energia do ruído, ocasionando a superestimação do tempo de reverberação devido à soma da energia do ruído na Equação 4.4. Se a truncagem ocorrer antes da energia da resposta ao impulso alcançar o nível de ruído, a estimativa do tempo de reverberação é subestimada, uma vez que parte da energia acústica não é considerada. As curvas de decaimento da energia que resultam em superestimação e subestimação do tempo de reverberação, são mostradas na Figura 34 em linhas contínuas magenta e vermelha, respectivamente. A curva de decaimento de energia acumulada em azul é adequada para uma boa estimativa do tempo de reverberação.

Todos esses procedimentos foram efetuados para todas as AIRs obtidas de cada local de experimentação, totalizando 144 estimativas, sendo 36 para cada uma das 4 salas. A Figura 35 mostra os gráficos *boxplot* relativos à estimativa do RT_{60} de cada ambiente, e a Tabela 8 mostra os valores médios e desvios padrão dessas estimativas.

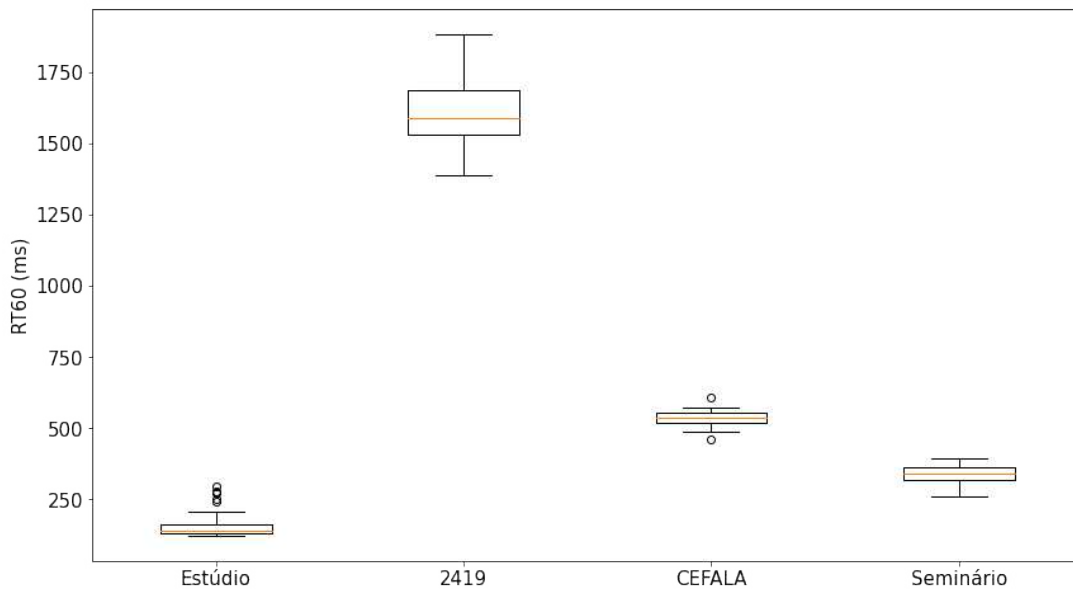


Figura 35 – *Box-plots* das estimativas da reverberação de cada ambiente. Fonte: Autoria própria.

Tabela 8 – Valores médios e desvios padrão das estimativas do tempo de reverberação.

Sala	Tempo de reverberação estimado
Estúdio	$162,33 \pm 50,86$ ms
2419	$1605,88 \pm 129,48$ ms
CEFALA	$533,86 \pm 28,03$ ms
Seminário	$337,71 \pm 35,04$ ms

4.4 Implementações

Todas as implementações feitas neste trabalho foram realizadas em Python 3.9, pelo Jupyter Notebook. Os arquivos com as implementações podem ser acessados pelo [link](#). No Apêndice A é apresentado o que são, e como foram implementados os *scripts* auxiliares.

4.4.1 Algoritmos de Separação

Ao todo, foram implementados nove algoritmos diferentes de separação cega da fala, sendo três para casos instantâneos e seis para casos convolutivos. Desses últimos, três foram FDICAs e três foram IVAs, com diferenças nas metodologias de otimização. Os parágrafos a seguir explicam como cada um dos métodos foi implementado. Todos os parâmetros dos métodos podem ser vistos na Tabela 9.

Os três primeiros algoritmos implementados foram a ICA por curtose (ICA_kurt), a ICA por negentropia (ICA_negen) e a ICA por MLE (ICA_ML). Esses métodos trabalharam com os dados no domínio do tempo para misturas instantâneas. Seguiu-se os passos apresentados por [Hyvärinen et al., 2001] para a implementação de cada um deles. Na ICA por curtose, fez-se uso da função de contraste da curtose (Equação 3.1) para encontrar o argumento que maximize a Equação 3.2, a qual foi otimizada via gradiente descendente. Similarmente, na ICA por negentropia, o problema de otimização da Equação 3.6 foi obtido pela função contraste da negentropia e otimizada por gradiente descendente. Na ICA por máxima verossimilhança, utilizou-se o gradiente descendente para otimizar a função de custo na Equação 3.7, a qual foi obtida pela abordagem da Estimação de Máxima Verossimilhança. Além disso, ressalta-se que enquanto nesse último algoritmo os dados utilizados no processo de otimização foram as misturas originais, nos dois primeiros os dados foram decorrelacionados por PCA, como pode ser observado em [Hyvärinen et al., 2001].

O primeiro grupo de algoritmos de separação de misturas convolutivas implementados foram as FDICAs, sendo elas FDICA por gradiente natural (FDICAnat), a FDICA por iteração de ponto fixo (fastFDICA) e a FDICA por função auxiliar (FDICAaux). As funções de custo foram calculadas da mesma maneira para os três métodos, e neste caso, elas foram obtidas pelo princípio da Infomax (máxima verossimilhança). Os três algoritmos

se diferenciam por sua forma de otimizar, fazendo com que cada um dos métodos possua parâmetros e/ou procedimentos específicos, tornando-os diferentes.

O último grupo de métodos correspondeu às IVAs, ou seja, IVA por gradiente natural (IVAng), IVA por iteração de ponto fixo (fastIVA) e IVA por função auxiliar (IVAaux). A função de custo dos três métodos foi calculada por MLE. A única diferença em parâmetros em relação ao grupo anterior é que não há necessidade de solucionar a ambiguidade da permutação.

4.4.2 Métricas de Avaliação

Neste trabalho, foram implementadas duas métricas objetivas de avaliação de desempenho dos métodos de separação: SI-SDR e SIR [LeRoux et al., 2019, Vincent et al., 2006]. Nesse caso, foi feito uso dos sinais originais, e assim, devido ao processo de edição das misturas, foi necessário ajustar a fase entre o sinal original e o sinal a ser avaliado. Esse problema foi solucionado via correlação cruzada, adicionando zeros no início do sinal de referência, de modo que o pico de correlação fosse para o atraso de 0 amostras.

Outra medida de desempenho utilizada nesse trabalho foi a PESQ, apesar de não ter sido implementada. O código-fonte foi obtido pelo site da ITU-T [International Telecommunication Union, 2001]. Em todos os resultados obtidos, os valores representam o ganho médio. Assim, primeiro foram obtidos os valores das métricas para as misturas, e depois, para os sinais separados. Com esses valores, aplicou-se a subtração para avaliar o quanto de ganho se teve após a separação. Como cada fonte exibiu um valor, foi aplicada a média aritmética, obtendo assim a métrica final. Um exemplo é apresentado a seguir para o caso do SDR.

$$\begin{aligned} SDR_{y_1} - SDR_{x_1} &= SDR_1 , \\ SDR_{y_2} - SDR_{x_2} &= SDR_2 , \\ \frac{1}{2}(SDR_1 + SDR_2) &= SDR_{final} . \end{aligned}$$

4.4.3 Execução dos Algoritmos

A configuração dos parâmetros dos métodos de separação foi uma importante etapa deste trabalho, pois permitiu avaliar qual é o desempenho deles diante de tais modificações e compará-los. A Tabela 9 mostra como cada método foi parametrizado, totalizando um valor de 36 testes dentre todos os algoritmos. Essas mesmas variações foram aplicadas a todos os ambientes onde foram realizados os experimentos. Os nomes das colunas são: “Configuração” o qual é referente ao número da especificação adotada, “Método” é o método de separação aplicado, “nIter” é o número de iterações, “Fsn” é a nova frequência de amostragem em hertz, sendo que todos os sinais originais foram amostrados em 44100 Hz,

“ η ” é a taxa de aprendizagem, Q é a quantidade de amostras usadas na análise para cada quadro, Q_{hop} é a porcentagem de sobreposição do próximo quadro sobre o quadro de análise anterior, “Perm.” é o valor booleano referente à solução da ambiguidade de permutação e “Escala” é o valor booleano referente à solução da ambiguidade de escala. O parâmetro de controle da função densidade de probabilidade “ α ” foi escolhido para todos os testes como sendo 0, ou seja, a suposição de que os sinais da fonte possuem distribuição laplaciana. Os valores especificados foram provenientes de simulações executadas anteriormente, até observar, principalmente, a convergência.

A duração total da execução de todas as configurações apresentadas e avaliadas para cada ambiente, foi em média aproximadamente 24 h, necessitando assim de 4 a 5 dias para a execução completa no computador com processador Intel Core i5 9300H 2,40 GHz, com 24 GB de memória RAM, no sistema operacional Windows 10, com a conexão Wi-Fi desconectada.

Tabela 9 – Todas configurações de parâmetros dos métodos.

Configuração	Método	nIter	F_s (Hz)	η	Q	Q_{hop} (%)	Perm.	Escala
1	ICA_kurt	100	44100	0,5	-	-	-	-
2	ICA_negen	100	44100	0,5	-	-	-	-
3	ICA_ML	100	44100	0,5	-	-	-	-
4	FDICAnat	150	16000	0,2	2048	75	True	True
5	FDICAnat	100	16000	0,1	2048	75	True	True
6	FDICAnat	100	16000	0,2	4096	75	True	True
7	FDICAnat	100	8000	0,2	2048	75	True	True
8	FDICAnat	100	16000	0,2	2048	50	True	True
9	FDICAnat	100	16000	0,2	2048	75	False	False
10	FDICAnat	100	16000	0,2	2048	75	False	True
11	fastFDICA	200	16000	-	2048	75	True	True
12	fastFDICA	100	16000	-	2048	75	True	True
13	fastFDICA	200	16000	-	4096	75	True	True
14	fastFDICA	200	8000	-	2048	75	True	True
15	fastFDICA	200	16000	-	2048	50	True	True
16	fastFDICA	200	16000	-	2048	75	False	False
17	fastFDICA	200	16000	-	2048	75	False	True
18	FDICAaux	100	16000	-	2048	75	True	True
19	FDICAaux	100	16000	-	4096	75	True	True
20	FDICAaux	100	8000	-	2048	75	True	True
21	FDICAaux	100	16000	-	2048	50	True	True
22	FDICAaux	100	16000	-	2048	75	False	False
23	FDICAaux	100	16000	-	2048	75	False	True
24	IVAng	500	16000	0,2	2048	75	-	True
25	IVAng	100	16000	0,1	2048	75	-	True
26	IVAng	500	16000	0,2	4096	75	-	True
27	IVAng	500	8000	0,2	2048	75	-	True
28	IVAng	500	16000	0,2	2048	50	-	True
29	fastIVA	100	16000	-	2048	75	-	True
30	fastIVA	100	16000	-	4096	75	-	True
31	fastIVA	100	8000	-	2048	75	-	True
32	fastIVA	100	16000	-	2048	50	-	True
33	IVAaux	100	16000	-	2048	75	-	True
34	IVAaux	100	16000	-	4096	75	-	True
35	IVAaux	100	8000	-	2048	75	-	True
36	IVAaux	100	16000	-	2048	50	-	True

4.5 Conclusões do capítulo

Neste capítulo, foram apresentadas as etapas metodológicas essenciais para a realização dos experimentos de separação de sinais de fala. Para esse fim, foram apresentados os procedimentos de pré-experimento, incluindo a geração e edição das frases e a preparação do sinal TSP, que garantem a padronização e o sincronismo dos estímulos utilizados. Em seguida, foram detalhados os experimentos, abordando os diferentes ambientes, os ensaios realizados, os equipamentos utilizados e a estimação da reverberação, os quais são importantes aspectos que podem interferir no desempenho da separação. Por fim, foram descritas as implementações dos algoritmos de separação e as métricas de avaliação, responsáveis por quantificar de maneira objetiva o desempenho dos métodos propostos, permitindo assim a interpretação dos resultados apresentados no Capítulo 5.

Capítulo 5

Resultados e Discussão

Neste capítulo, são apresentados e discutidos os resultados obtidos nos experimentos. Inicialmente, descreve-se o processo de seleção dos melhores parâmetros dos algoritmos, com base em seu desempenho. Em seguida, realiza-se uma análise geral dos sinais por meio de espectrogramas. Posteriormente, são apresentadas as análises fatoriais de cada métrica. Por fim, discutem-se os resultados relacionados ao posicionamento dos microfones.

5.1 Considerações Iniciais

A comparação de todos os algoritmos de separação, incluindo suas alterações nos parâmetros, é inviável devido à quantidade de variações de um mesmo método, como mostra a Tabela 9, o que dificultaria o comparativo entre cada algoritmo de separação para cada sala reverberante analisada. Tendo em vista esse problema, foi aplicado um procedimento de seleção da melhor combinação de parâmetros de cada algoritmo. Esse critério foi baseado na média da métrica de desempenho de cada configuração de parâmetro (36), para cada: métrica de desempenho (4), ensaio (9) e ambiente (4). O cálculo da média foi baseado nas 16 misturas. A configuração que obtivesse a melhor média de uma dada métrica, em um ensaio e ambiente específicos, receberia a nota máxima de 35, e assim por diante até a configuração de pior média, com nota 0. Ao final, todas as notas de cada configuração foram somadas para escolher quais delas obtiveram os melhores resultados globais. A Tabela 10 mostra as melhores variações de parâmetros de cada método.

Inicialmente, analisariam-se os 9 métodos de separação implementados, 3 de misturas instantâneas e 6 de misturas convolutivas. Entretanto, após a seleção dos melhores parâmetros, optou-se por não realizar o estudo comparativo com os métodos aplicados para misturas instantâneas, devido ao baixo desempenho em relação aos métodos de separação para misturas convolutivas. Apesar disso, a Tabela 11 apresenta os resultados desses métodos para o ensaio “P4” em todos os ambientes, para três métricas distintas.

Tabela 10 – Melhores configurações de parâmetros para cada método pelo critério de seleção.

Método	nIter	F_s	η	Tam.Quad.	Sobrep.Quad.	Perm.	Escala
ICA_kurt	100	44100	0,5	-	-	-	-
ICA_negen	100	44100	0,5	-	-	-	-
ICA_ML	100	44100	0,5	-	-	-	-
FDICAnat	150	16000	0,2	2048	75%	True	True
fastFDICA	200	16000	-	2048	75%	True	True
FDICAaux	100	16000	-	2048	75%	True	True
IVAng	500	8000	0,2	2048	75%	-	True
fastIVA	100	8000	-	2048	75%	-	True
IVAaux	100	8000	-	2048	75%	-	True

Tabela 11 – Resultados para o ensaio “P4” em cada sala para os métodos instantâneos.

Sala	ICA_kurt		
	SDR (dB)	SIR (dB)	PESQ8 (dB)
Estúdio	-1,14 ± 0,73	3,25 ± 2,32	-0,15 ± 0,02
CEFALA	-1,65 ± 0,58	3,75 ± 1,37	-0,16 ± 0,02
Seminário	-1,47 ± 0,92	6,80 ± 2,55	-0,13 ± 0,02
2419	-1,12 ± 1,03	3,2 ± 1,98	-0,23 ± 0,03
Sala	ICA_negen		
	SDR (dB)	SIR (dB)	PESQ8 (dB)
Estúdio	-1,17 ± 0,47	2,84 ± 2,64	-0,15 ± 0,02
CEFALA	-1,66 ± 0,59	3,70 ± 1,92	-0,16 ± 0,03
Seminário	-1,47 ± 0,64	6,84 ± 1,82	-0,13 ± 0,02
2419	-1,13 ± 1,12	2,55 ± 1,72	-0,23 ± 0,02
Sala	ICA_ML		
	SDR (dB)	SIR (dB)	PESQ8 (dB)
Estúdio	-1,17 ± 0,96	3,65 ± 2,57	-0,15 ± 0,02
CEFALA	-1,66 ± 0,75	3,95 ± 2,36	-0,16 ± 0,02
Seminário	-1,48 ± 0,85	6,35 ± 2,41	-0,13 ± 0,02
2419	-1,14 ± 1,25	2,35 ± 2,12	-0,24 ± 0,04

Para analisar os métodos selecionados, é necessário escolher as métricas às quais a comparação ocorreu. Dentre as quatro métricas calculadas, serão abordadas três (SI-SDR, SIR e PESQ8), uma vez que os valores da PESQ16 foram muito semelhantes aos da PESQ8.

5.2 Avaliações Gerais

Nesta seção, as comparações dos algoritmos foram realizadas por meio de simples inspeções dos espectrogramas dos sinais originais, das misturas e dos sinais separados, além de uma comparação de forma audível. Como a avaliação de todos os casos seria impraticável, optou-se por fixar o ensaio P4 (Tabela 7) e a mistura “aMiF” (Tabela 2), ou seja, fala masculina sobre áudio (“aM”) e feminina sobre ICA (“iF”). Os arquivos de áudio podem ser acessados por este [link](#).

A comparação gráfica permitiu avaliar como os sinais originais foram transformados em misturas pelo ambiente e como os métodos de separação extraem tais sinais. A Figura 36 apresenta os espectrogramas dos sinais originais. Alguns exemplos de misturas obtidas para cada ambiente podem ser analisados na Figura 37. Ao inspecionar as figuras, observa-se o efeito gerado pela mistura no ambiente, e como os formantes da fala masculina e feminina interagem, tanto no tempo quanto na frequência. Além disso, o espectrograma da mistura na sala 2419 exhibe o efeito de *blur* (em português, borrão) ao longo do eixo de tempo, que é consequência do elevado tempo de reverberação nesse ambiente. Isso fica mais fraco nos outros ambientes, os quais possuem RT_{60} muito inferior à sala 2419.

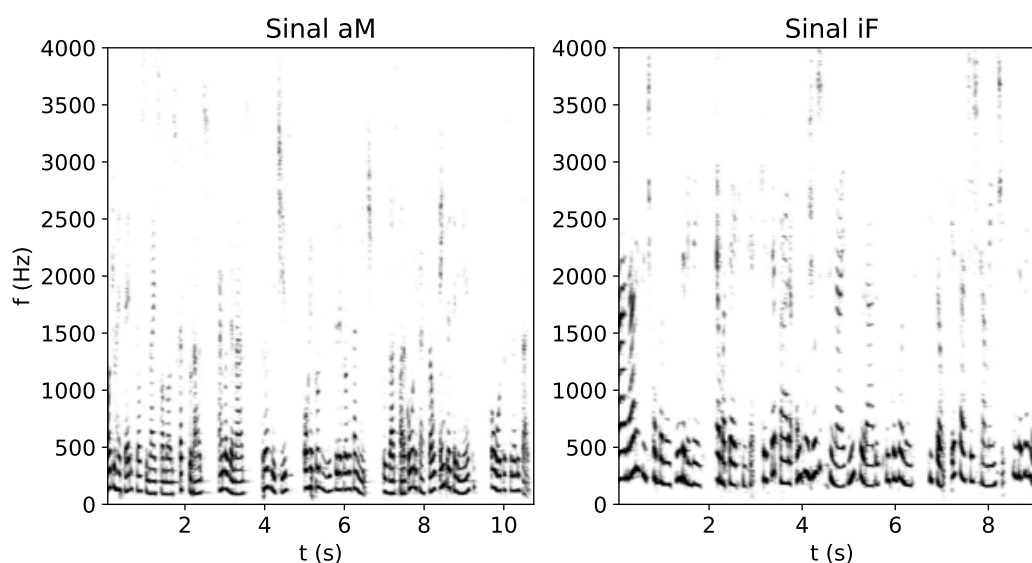


Figura 36 – Espectrogramas do sinal original “aM” (esquerda) e do sinal original “iF” (direita). Fonte: Autoria própria.

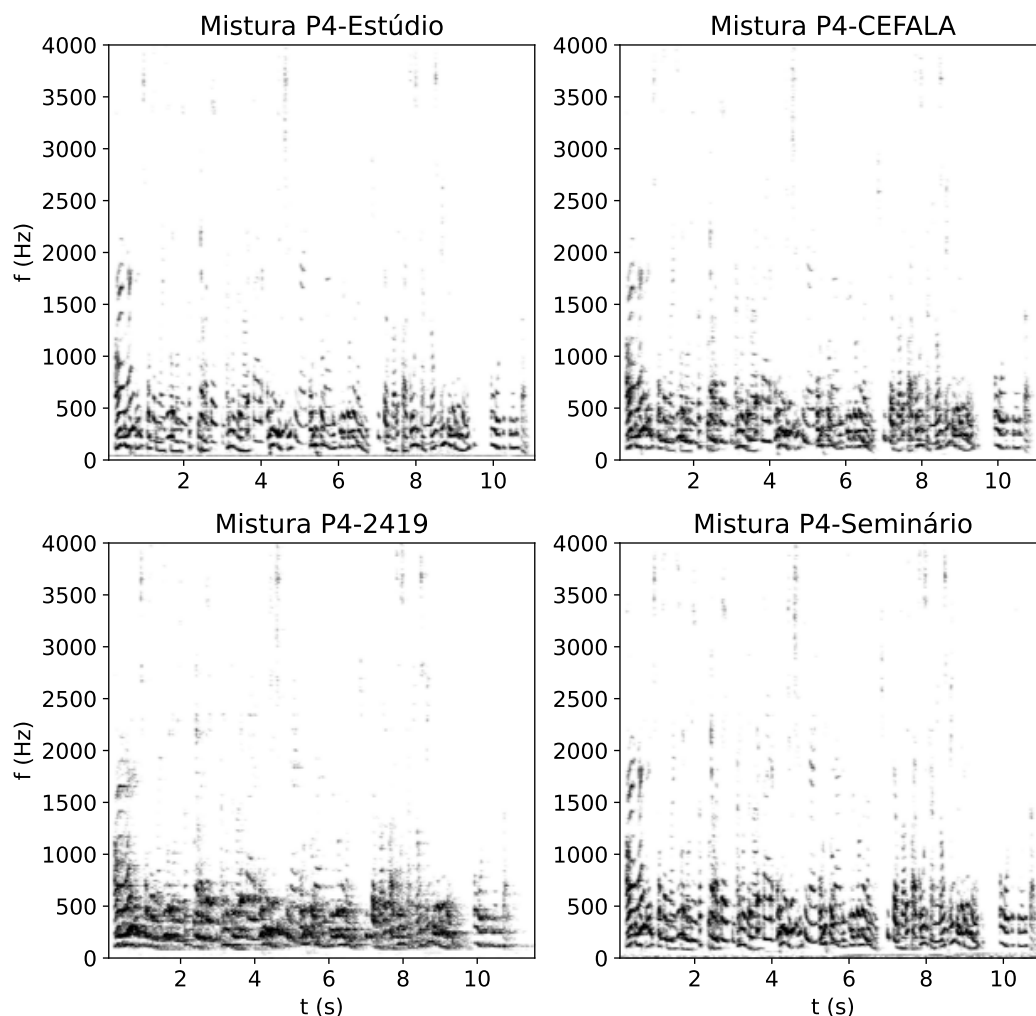


Figura 37 – Espectrogramas das misturas do sinal “aM” com o sinal “iF” para o ensaio “P4” em cada ambiente. Fonte: Autoria própria.

Após a aplicação dos métodos de separação, realizou-se uma avaliação inicial subjetiva por meio da escuta de cada sinal separado. Além disso, os espectrogramas dos sinais foram analisados e apresentados apenas em alguns casos de interesse.

Primeiramente, para o estúdio, o método fastFDICA exibiu muitos artefatos sonoros similares à microfonia, além de distorcer o timbre de voz. Os outros métodos tiveram bom desempenho em separar e manter o timbre de cada falante, em especial os métodos fastIVA, FDICAaux e IVAaux, em que não se escutou a fala do canal interferente, diferente dos métodos IVAng e FDICAnat, os quais separaram os sinais, mas ainda com a fala interferente perceptível. A Figura 38 mostra dois espectrogramas dos sinais aM (reamostrados a 8 kHz) separados pela FDICAaux e fastFDICA. Apesar de ainda haver interferência da fala feminina em ambos os casos, pode ser visto que, no primeiro caso, isso é menos

acentuado do que no segundo, em que há mais interferência do outro canal, além da ativação constante de um componente de frequência próxima a 3500 Hz, mostrando que a estimação do modelo espacial foi ruim para esse componente.

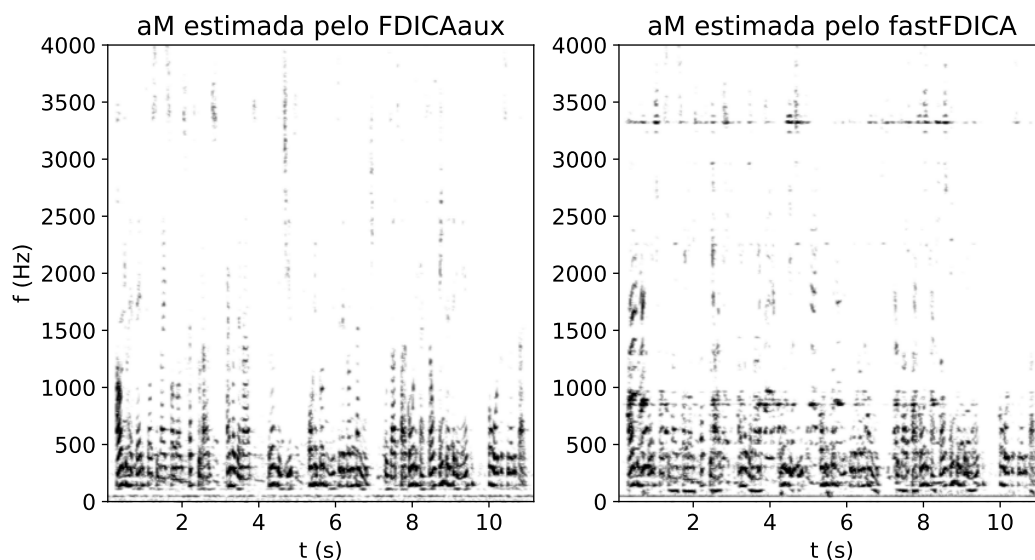


Figura 38 – Espectrogramas do sinal “aM” extraído da mistura no estúdio. À esquerda, separação pela FDICAaux, e à direita, pela fastFDICA. Fonte: Autoria própria.

Para o experimento no CEFALA, após escutar os sinais separados, observou-se que, em geral, a separação foi pior do que para a separação de misturas do estúdio. Primeiramente, a fastFDICA, a FDICAaux e a FDICAnat não separaram os sinais adequadamente, sendo que o primeiro ainda distorceu os timbres dos locutores. Além disso, os métodos IVA tiveram bom desempenho de separação comparados aos demais, sendo a fastIVA o melhor dentre eles. A IVAng apesar de não ter tido desempenho igual à fastIVA, foi o que teve menos artefatos sonoros, assim como a IVAaux. A Figura 39 mostra as faixas “iF” separadas pelos métodos FDICAnat e fastIVA. Além disso, como as falas não tiveram a mesma duração, a partir dos 10 segundos há apenas fala masculina. A partir deste instante, observa-se que a fastIVA atenua fortemente a energia dos componentes ao longo de todo o espectro, principalmente de 1000 Hz a 3000 Hz. Assim, ao comparar com o espectrograma do sinal separado pela FDICAnat, conclui-se que a fastIVA foi consideravelmente superior.

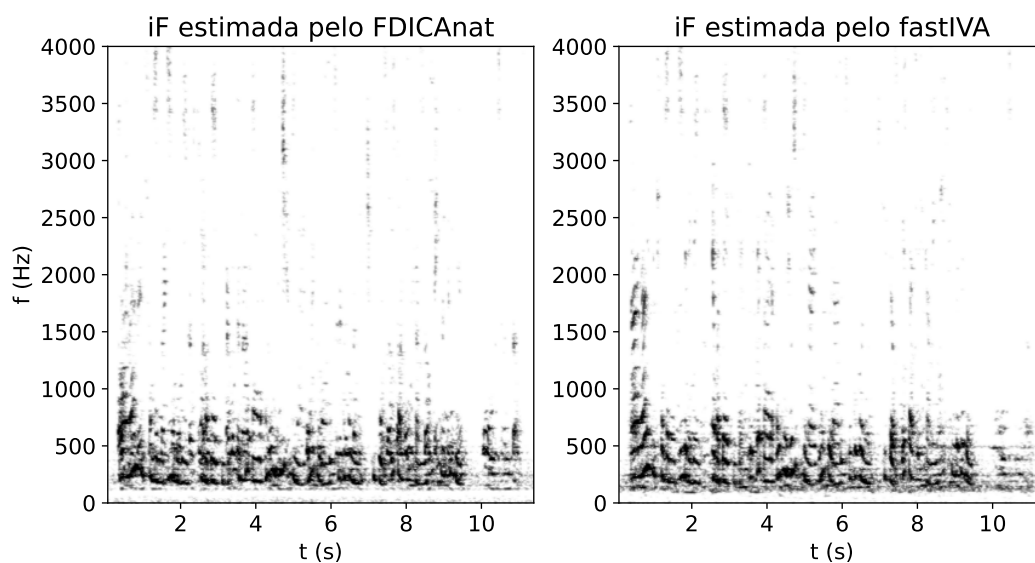


Figura 39 – Espectrogramas do sinal “iF” extraído da mistura no CEFALA. À esquerda, separação pela FDICAnat, e à direita, pela fastIVA. Fonte: Autoria própria.

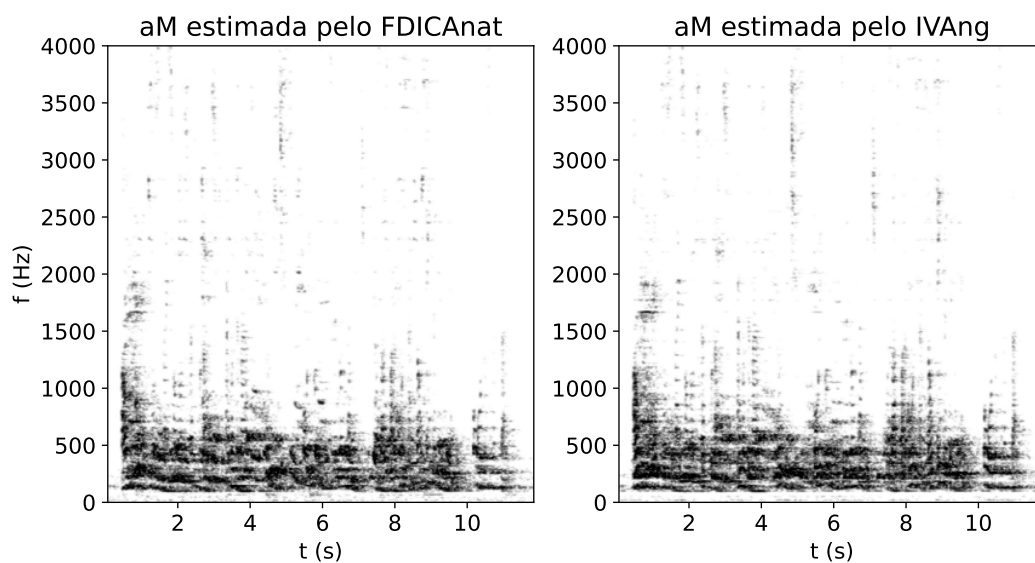


Figura 40 – Espectrograma do sinal “aM” extraído da mistura na sala 2419. À esquerda, separação pela FDICAnat, e à direita, pela IVAng. Fonte: Autoria própria.

Após a escuta das faixas separadas das misturas da sala 2419, observou-se que os métodos apresentaram um desempenho consideravelmente inferior em comparação com os obtidos no estúdio e CEFALA. Assim, foi constatado que os métodos fastIVA, fastFDICA, FDICAnat, FDICAaux não separaram de forma adequada os sinais. Os métodos IVAng e IVAaux foram os melhores para esse ambiente, sendo que o primeiro distorceu menos os sinais do que o segundo. A Figura 40 mostra os espectrogramas dos sinais “iF”, separados pelos métodos FDICAnat e IVAng. Como pode ser observado, o IVAng recupera melhor a

energia do espectro original do que o FDICAnat, em que os componentes ainda exibem muita sobreposição.

Na sala de seminários, ao ouvir os sinais separados, constatou-se que os métodos tiveram desempenho de separação melhor do que para os ambientes CEFALA e 2419. Com exceção da fastFDICA, todos os métodos realizaram boa separação. Dentre esses, os melhores foram IVAaux e FDICAaux, que resultaram em faixas separadas bem compreensíveis, com pouca interferência e distorções. A Figura 41 mostra os espectrogramas para ambos os métodos. Como pode-se observar, apesar de ambos os métodos terem sido eficazes, o IVAaux teve melhor desempenho, pois recuperou melhor os componentes da fala feminina das misturas. Além disso, a atenuação dos componentes da fala masculina entre os métodos fica clara a partir dos 10 s (apenas fala masculina), em que o IVAaux atenua quase por completo a energia dos componentes.

Em diversos espectrogramas dos sinais separados apresentados ao longo desta seção, observa-se a presença de linhas horizontais que não aparecem nos espectrogramas dos sinais originais, como mostrados na Figura 36. Uma hipótese para esse efeito baseia-se no fato de que os métodos de separação de misturas convolutivas empregados neste estudo são de banda estreita, ou seja, aplicados individualmente a cada componente de frequência. Dessa forma, uma estimativa imprecisa da matriz de separação pode resultar em uma separação inadequada em determinadas componentes de frequência, tendo assim, sobreposições das fontes naquela componente.

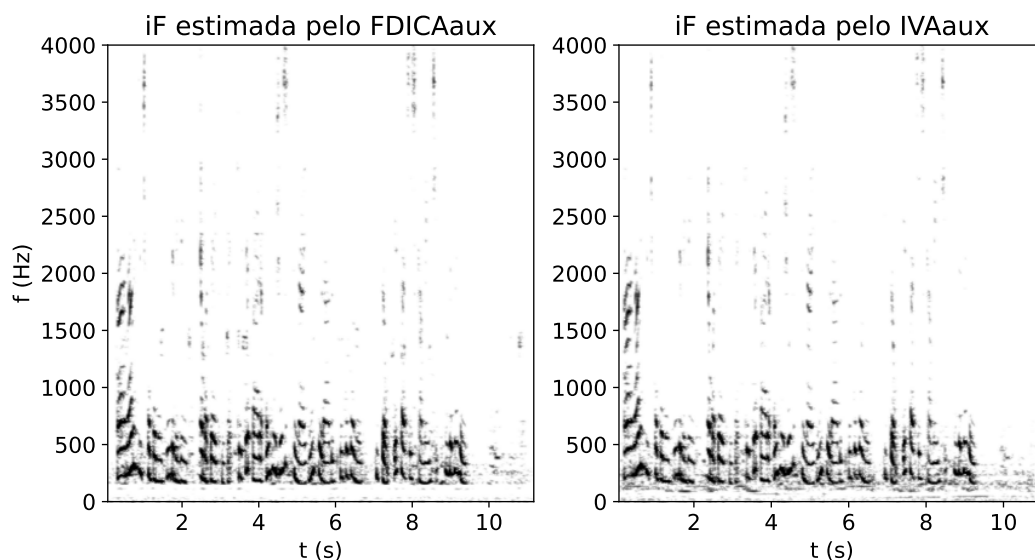


Figura 41 – Espectrogramas do sinal “iF” extraído da mistura na sala de seminários. À esquerda, separação pela FDICAaux, e à direita, pela IVAaux. Fonte: Autoria própria.

5.3 Estudo Comparativo dos Métodos

Nesta etapa do estudo, buscou-se investigar a influência de diferentes condições experimentais sobre o desempenho dos métodos de separação cega da fala. O objetivo dessa análise é identificar quais fatores exercem maior impacto sobre as métricas de desempenho de forma quantitativa.

Para realizar essa investigação, aplicou-se uma Análise de Variância (ANOVA) com múltiplos fatores, de forma a avaliar os efeitos principais e as possíveis interações entre os fatores. Essa abordagem estatística possibilita determinar, de forma objetiva, a relevância de cada fator e suas combinações na variação observada das métricas de desempenho. Os fatores considerados no experimento foram o método, o ensaio e o ambiente, enquanto as variáveis resposta analisadas foram SIR, SDR e PESQ8. As subseções seguintes apresentam as análises dos efeitos de cada fator sobre as métricas mencionadas.

5.3.1 Análises da SIR

A avaliação da variável resposta SIR, envolveu a etapa de agrupamento pela média de cada ensaio realizado. Assim, analisaram-se 216 linhas, cada uma representando a média de um método (6), ensaio (9) e ambiente (4) específicos. Primeiramente, foi obtido o gráfico de barras, de modo que foi possível avaliar preliminarmente e visualmente o comportamento dos fatores, para depois realizar os testes estatísticos.

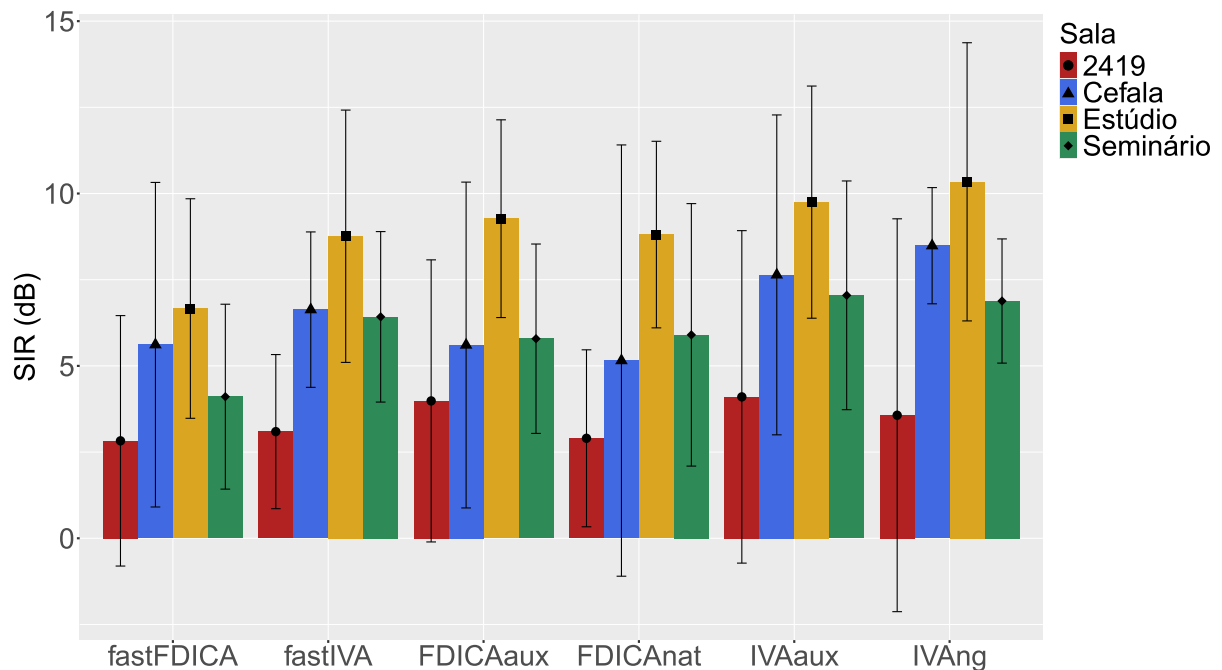


Figura 42 – Gráfico de barras para cada método e ambiente, avaliados sob a SIR. Fonte: Autoria própria.

A Figura 42 mostra o comportamento da SIR em função do método de separação para cada ambiente. Em média, independente da posição dos microfones, os algoritmos possuíram melhor desempenho no estúdio, o qual é o ambiente com menor tempo de reverberação (Tabela 8). Observa-se que, para a sala de seminários (barras verdes), a SIR foi muito próxima em média dos resultados apresentados no CEFALA, mesmo sendo um ambiente com tratamento acústico e tempo de reverberação menor, tendo uma diferença de aproximadamente 0,5 dB.

As barras em vermelho mostram que todos os métodos tiveram uma queda de desempenho de separação para a sala de maior tempo de reverberação (sala 2419). Essa redução geral de desempenho pode ser devido a causas específicas, tais como sobreposição de componentes tempo-frequência e artefatos decorrentes da convolução circular, consequência da truncagem do filtro IIR pelo FIR em salas reverberantes [Makino et al., 2007]. A sobreposição é uma consequência natural da reverberação: quanto maior o RT_{60} , mais os reflexos atenuados e atrasados da mesma fonte permanecem com energia significativa. Em uma representação tempo-frequência, quadros contendo múltiplos reflexos de uma mesma fonte podem apresentar sobreposição de componentes de frequência. Isso pode reduzir a não gaussianidade suposta para a fonte, levando à redução do desempenho da separação cega.

Como discutido no Capítulo 2, a reverberação é um processo de múltiplos caminhos e pode ser modelada por um filtro FIR de tamanho L . Assim, o quadro de análise necessita que o número de amostras seja de tamanho igual ou maior a L , para que não ocorra o efeito de convolução circular. Apesar do aumento do número de amostras do quadro de análise melhorar a resolução na frequência, e consequentemente melhorar o modelo do filtro FIR pelo aumento do número de parâmetros, isso reduz a resolução no tempo, problema esse que pode ser amenizado com a redução do número de amostras de deslizamento do quadro de análise (*frame shift*). Contudo, essa solução não resolve outro problema associado ao uso de quadros de análise muito longos, que está relacionado à não estacionariedade do sinal de fala em intervalos de tempo superiores a 20 ms. Tal problema pode ser melhor explicado com o seguinte exemplo: o quadro de análise da STFT para misturas da sala 2419, de modo a mitigar o efeito da convolução circular, deve ser de no mínimo 12800 amostras para frequência de amostragem de 8 kHz. Porém, tal quadro não deveria ultrapassar 160 amostras, pois fora dessa faixa, o sinal já é considerado não estacionário. Assim, corrigir a convolução circular com janelas longas em casos de ambientes reverberantes significa abarcar em um mesmo quadro, diversos fonemas com energia em diferentes regiões do espectro, como um trecho de fala contendo vogais (formantes de baixa e média frequência), consoantes fricativas (alta frequência) e plosivas (aproximadamente todo espectro). Assim, o desempenho da separação seria reduzido, dado que os algoritmos de separação no domínio da frequência necessitam tanto da informação de magnitude quanto de fase das misturas. A seleção do valor subótimo de amostras do quadro de análise é um problema de otimização

e de decisão multicritério, pois leva a objetivos conflitantes, o que está fora do escopo deste trabalho.

A partir da Figura 42, observa-se que o método fastFDICA possui em média desempenho reduzido, comparado aos outros métodos, assim como dentre as IVAs, a fastIVA é a que possui o pior desempenho. Esses métodos otimizam os valores do sistema de mistura no espaço descorrelacionado das amostras obtido pela PCA, sob restrição das linhas da matriz de mistura serem formadas por vetores normalizados. Sendo assim, limitar o espaço de busca para uma hipersfera de raio unitário acelera a velocidade de convergência do algoritmo (chamados de algoritmos *fast*) e promove maior estabilidade (inexistência do hiperparâmetro da taxa de aprendizagem), em troca da limitação de desempenho.

Os métodos IVAaux e IVAng tiveram desempenhos similares e melhores do que suas versões univariadas (FDICAaux e FDICAnat). Provavelmente, esta melhoria está relacionada com a modelagem do vetor de frequências de forma multivariada, o que mantém as correlações entre as componentes de sinais provenientes da mesma fonte, e melhora a separação entre componentes de fontes independentes. Assim, diferentemente da versão univariada, não há necessidade do pós-processamento para solução da ambiguidade de permutação.

Outra análise envolve o desempenho dos métodos para os diferentes ensaios realizados, representado pela Figura 43. Verificou-se que cada método teve desempenhos dependentes do tipo de ensaio, como por exemplo, a FDICAaux foi melhor para o “P4” (curva amarela) e pior para o “M90” (curva verde escuro). Novamente, em média, os IVAs foram melhores do que os FDICAs, o que mostra que as metodologias possuem diferenças que serão avaliadas.

Uma outra constatação importante é que os métodos possuem SIR baixos em alguns casos específicos. As menores SIR dos métodos, em média, foram para a “L90”, “M90” e “P90”, e as maiores para os microfones a 4 cm de distância. Para o caso dos microfones com distância de 57 cm, apenas no ensaio “P57” os métodos tiveram alta SIR comparado aos casos “M57” e “L57”. Esses resultados indicam que, embora a distância entre o microfone e a fonte possa impactar o desempenho dos métodos, sua influência é inferior à da distância entre os microfones.

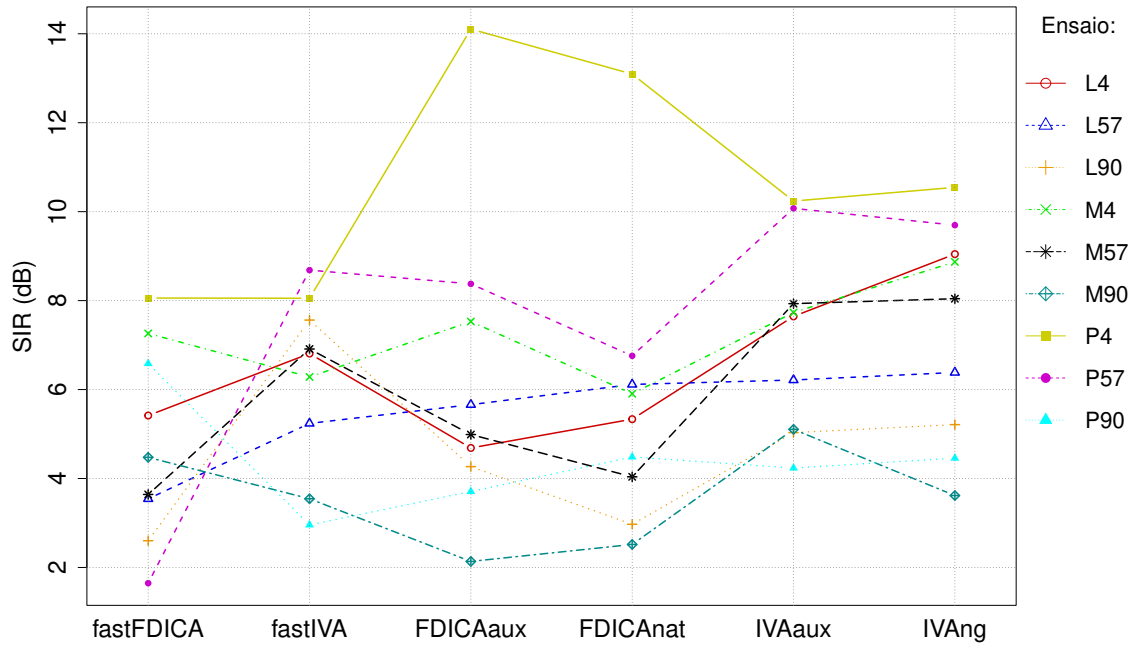


Figura 43 – Interação entre método e ensaio, avaliados sob a SIR. Fonte: Autoria própria.

A diferença de desempenho nos ensaios não é isoladamente devido ao comportamento dos métodos, mas pode também ser pelo formato do arranjo de microfones. O pior desempenho foi para casos em que os microfones estavam na distância de 0 cm e em orientação de 90°. Nesse caso, os sinais estão sendo amostrados em pontos espaciais muito próximos, resultando em correlações elevadas para cada componente de frequência, ou seja, tanto a magnitude quanto a fase são similares. As magnitudes são altamente correlacionadas não apenas devido aos microfones estarem próximos, mas pela característica de omnidirecionalidade do equipamento. A alta correlação entre as fases é explicada pela falta de diferença angular azimutal em relação às fontes emissoras, o que é reduzido com o aumento da distância entre os microfones. Consequentemente, seria natural pensar que quanto maior a distância entre os microfones, menos correlacionadas estariam as fases das misturas, porém, com apenas uma exceção, constatou-se que os ensaios em que os microfones estavam a 57 cm tiveram desempenho pior do que aqueles a 4 cm. Uma possível razão para esse comportamento é o *aliasing* espacial. Na representação da mistura no domínio tempo-frequência, cada componente corresponde a uma frequência específica do espectro, com comprimento de onda λ fixo. Ao amostrar um sinal senoidal de comprimento de onda λ em pontos diferentes no espaço, de modo que os microfones estejam a uma distância igual ou maior do que $\lambda/2$, resulta em ambiguidade de fase, o chamado *aliasing* espacial [Trees, 2002]. Esse efeito pode afetar a estimação do sistema de separação, o que é aprendido pelos algoritmos de separação, prejudicando assim o desempenho. Supondo que a velocidade do som seja igual a 340 m/s, para a distância entre os microfones de 4 cm e 57 cm, o *aliasing* espacial ocorre para frequências acima de 4,25 kHz e 298 Hz, respectivamente. Para a distância de 0 cm, em teoria, não ocorre esse efeito.

A fim de investigar estatisticamente a influência de cada fator, o modelo inicial do ANOVA considerou a interação dos fatores até em terceira ordem para avaliar a média dos quadrados, o qual atribuiu-se o nome de “modelo completo”. A Tabela 12 mostra os valores da soma dos quadrados e da média quadrática de cada fator principal e suas interações.

Tabela 12 – ANOVA do modelo multifatorial de terceira ordem para a SIR. A sigla GL representa “Graus de liberdade”.

Fatores	GL	Soma dos quadrados	Média dos quadrados
método	5	156,0	31,21
ensaio	8	871,8	108,98
ambiente	3	829,9	276,62
método:ensaio	40	459,6	11,49
método:ambiente	15	58,4	3,89
ensaio:ambiente	24	788,0	32,83
mét:ens:amb	120	493,1	4,11

A média dos quadrados mostra que a SIR é mais influenciada pelo ambiente e ensaio do que pelo método de separação usado, o que era esperado pelas análises anteriores. Além disso, a interação “ensaio:ambiente” possui influência similar ao fator “método” na variável resposta. Todos os outros fatores que interagem com o fator método de 2^a ou 3^a ordem resultaram em baixa média quadrática. Isso levou à realização de uma outra ANOVA, sendo o novo modelo composto pelas interações de até segunda ordem, o qual foi usado para avaliar estatisticamente os fatores envolvidos.

Os resultados da ANOVA para o novo modelo podem ser vistos na Tabela 13. Pelo valor obtido do coeficiente de determinação R^2 , esse modelo foi capaz de explicar 86,51% da variância da SIR. Além disso, com exceção do fator de interação “método:ambiente”, todos os outros resultaram em uma estatística F elevada. Assim, pode-se afirmar que dentre esses fatores (principais e de interação) existem níveis em que a diferença na média global do modelo é diferente de zero. Assim, pelo teste de hipóteses, rejeita-se as hipóteses nulas de que os fatores (com exceção de “método:ambiente”) não afetam a SIR, ao nível de significância de 5%. Por fim, o uso do teste de Shapiro-Wilk [Shapiro and Wilk, 1965] permitiu verificar que os resíduos são considerados normais (p-valor = 0.278), o que pode ser observado pelos pontos dentro dos limiares da normalidade no gráfico quantil-quantil na Figura 44.

Tabela 13 – ANOVA do modelo multifatorial de segunda ordem para a SIR. A sigla SQ se refere à soma dos quadrados e a MQ, média dos quadrados.

Fatores	GL	SQ	MQ	Valor F	P(>F)
método	5	156,0	31,21	7,595	3,14e-06
ensaio	8	871,8	108,98	26,52	< 2e-16
ambiente	3	829,9	276,62	67,318	< 2e-16
método:ensaio	40	459,6	11,49	2,796	8,72e-06
método:ambiente	15	58,4	3,89	0,947	0,52
ensaio:ambiente	24	788,0	32,83	7,990	1,82e-15
Resíduos	120	493,1	4,11		

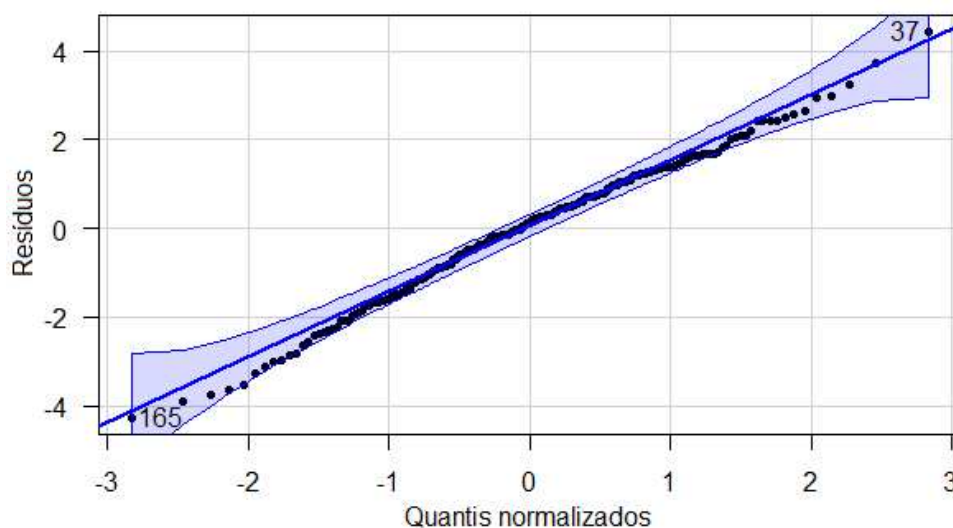


Figura 44 – Gráfico quantil-quantil dos resíduos do modelo multifatorial de segunda ordem para a SIR. Fonte: Autoria própria.

Um dos objetivos deste trabalho é a comparação dos métodos, entretanto, a presença de fatores interagindo prejudica a comparação. Embora o fator principal “método” tenha apresentado, no Teste de Fligner-Killeen para homocedasticidade, aceitação da hipótese nula ($p = 0,9625$), indicando variâncias homogêneas, o Teste de Tukey [Tukey, 1949] (todos contra todos) também resultou na aceitação da hipótese nula. Isso significa que não houve diferenças estatisticamente significativas entre os efeitos dos níveis do fator “método”, considerando um nível de significância de 5%. Portanto, conclui-se que para a medida de desempenho SIR, deve-se analisar os métodos para cada um dos casos, fixando um nível para o “ambiente” e “ensaio”, pois, além dos efeitos principais, ainda há os efeitos de interação, tornando a análise comparativa complexa.

5.3.2 Análises da SDR

Analogamente à SIR, a SDR também foi analisada seguindo os mesmos procedimentos. Assim, foi obtido o gráfico de barras entre método e ambiente, apresentado na Figura 45. Diferente da métrica anterior, a SDR teve um padrão de comportamento semelhante para cada método (com exceção da fastFDICA), o que pode ser verificado pelas barras no gráfico. Outro ponto importante é que observa-se novamente a relação entre redução de desempenho, com o aumento do RT_{60} , que, apesar de ter sido abordado anteriormente, para a SDR é mais evidente.

Os métodos que obtiveram o melhor desempenho, em média, foram as IVAs, seguido pela FDICAaux e FDICAnat, que tiveram medidas de desempenho próximas. A fastFDICA foi consideravelmente inferior a todos os outros cinco métodos, e em todos os casos, distorceu muito os sinais separados durante o processo de separação, ao ponto de que as SDR foram negativas em todos os casos.

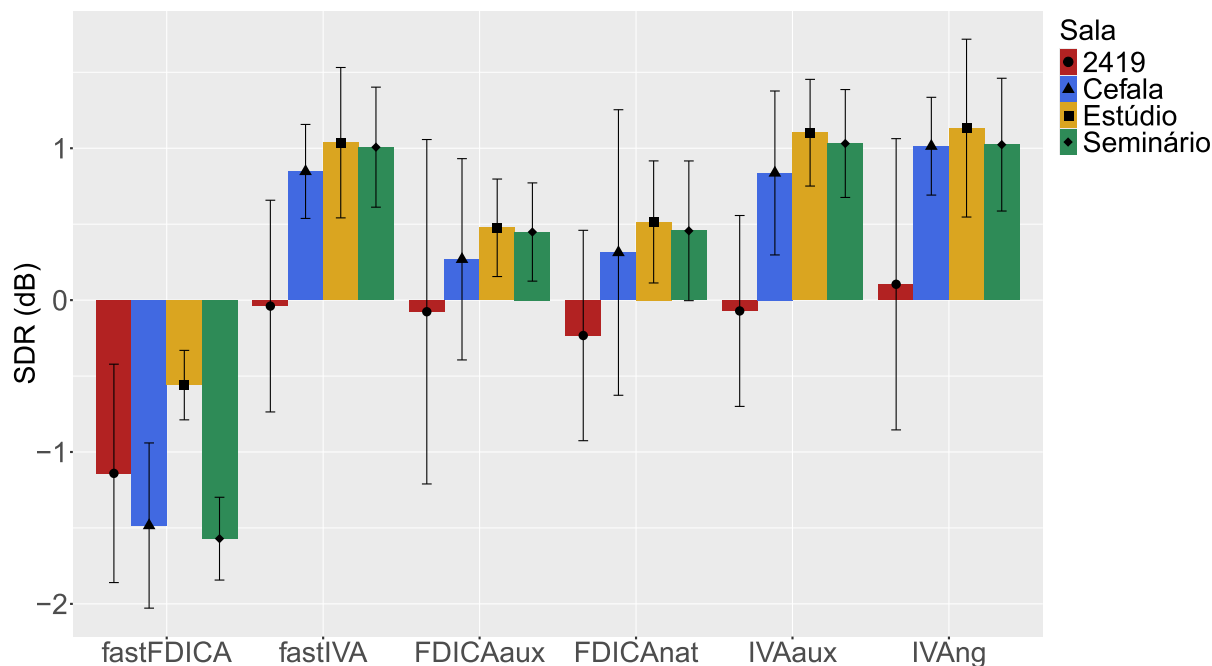


Figura 45 – Gráfico de barras entre método e ambiente, avaliados sob a SDR. Fonte: Autoria própria.

Outro importante gráfico de análise foi o de interação entre os métodos e os ensaios, mostrado na Figura 46. O mesmo comportamento observado nos gráficos anteriores foi verificado novamente. Com exceção do comportamento anômalo para os métodos FDICAaux e FDICAnat no caso “P4”, os quais alcançaram desempenhos competitivos com os IVAs, para todos os outros casos, a comparação preliminar dos métodos realizada para o gráfico de barras entre método e ambiente também é válida.

Ao comparar as Figuras 43 e 46, conclui-se que a estrutura do ensaio interfere na variável resposta. Nos ensaios em que os microfones estiveram quase no mesmo ponto espacial e com orientação de 90°, os métodos tiveram piores desempenhos. Por outro

lado, constatou-se que microfones a 4 cm e a 57 cm tiveram desempenhos próximos. Por essas duas constatações, afirma-se que, provavelmente, a elevada correlação entre as componentes de frequência das misturas seja mais efetiva na redução da SDR, do que o *aliasing* espacial, o qual afeta a estimação do modelo espacial. Em seções futuras deste capítulo, é comparado estatisticamente como a distância entre os microfones e a distância entre as fontes e microfones influenciam no desempenho da separação.

Apesar das diferenças da SDR observadas para os métodos, é importante comentar sobre os valores dessa métrica. Como a SDR é a relação entre a potência do sinal original com a potência do termo residual contendo interferências, artefatos e ruído, seu valor será sempre o “pior” caso (menor valor). Nesse trabalho, não foram utilizadas misturas geradas artificialmente por AIRs modeladas em ambientes reais, mas sim misturas captadas diretamente no ambiente real, com presença de ruído. Esse ruído de medição do microfone, apesar de estar presente nas misturas, não foi retirado, e os métodos separaram os sinais sem filtragem. Isso, possivelmente, pode ter influenciado na estimação das matrizes de separação e, conseqüentemente, no baixo valor da SDR. Um estudo futuro apresentado no Capítulo 6 desse trabalho menciona a possibilidade de utilizar, por exemplo, filtragem ou suavização ótima para melhorar a qualidade das misturas.

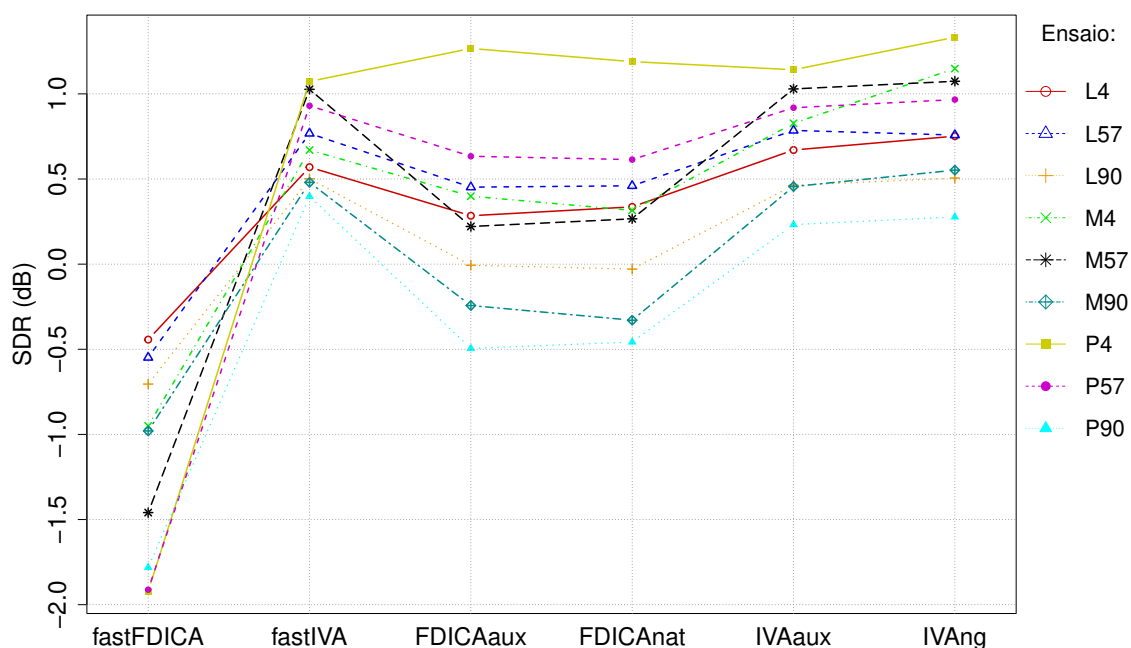


Figura 46 – Interação entre método e ensaio, avaliados sob a SDR. Fonte: Autoria própria.

O fato de o desempenho dos métodos terem apresentado comportamentos similares, independente do ambiente e ensaio, mostra que a interação entre esses fatores é fraca, e que possivelmente os fatores principais têm maiores efeitos significativos na média global. Para verificar isto, o modelo completo da ANOVA contou primeiramente com termos de primeira à terceira ordem.

A Tabela 14 mostra a soma dos quadrados e a média dos quadrados do modelo completo da ANOVA. Como era esperado pelas análises preliminares, os fatores de interação resultaram em uma baixa média dos quadrados em relação aos fatores principais, concluindo-se que tais efeitos podem ser tratados como resíduos do modelo. Além disso, diferente da SIR, o fator “método” teve consideravelmente o maior efeito na média global do modelo e, portanto, optou-se por criar um modelo com apenas este fator para analisar o desempenho dos métodos.

Tabela 14 – ANOVA do modelo multifatorial de terceira ordem para a SDR. A sigla GL representa “Graus de liberdade”.

Fatores	GL	Soma dos quadrados	Média dos quadrados
método	5	101,96	20,391
ensaio	8	16,12	2,015
ambiente	3	21,65	7,218
método:ensaio	40	19,91	0,498
método:ambiente	15	10,04	0,669
ensaio:ambiente	24	13,17	0,549
mét:ens:amb	120	15,47	0,129

A Tabela 15 mostra os valores da soma dos quadrados e média dos quadrados. A um nível de significância de 5%, rejeitou-se a hipótese nula de que o desempenho dos métodos não possui efeito estatisticamente significativo na média global do modelo. Assim, para alguns métodos, o efeito é significativo e deve ser analisado.

Antes de prosseguir para a análise comparativa, aplicou-se o Teste de Shapiro-Wilk para avaliar a normalidade dos resíduos, e o Teste de Fligner-Killeen para testar a homogeneidade de variâncias para o fator “método”. O primeiro resultou em p-valor elevado, e igual a 0,5689, aceitando a hipótese nula de que os resíduos são normais ao nível de significância de 5%. A Figura 47 apresenta o gráfico quantil-quantil, em que pode ser verificado que os resíduos são normais. No teste de homoscedasticidade, o p-valor foi de 0,1567 e, portanto, ao nível de significância de 5%, aceita-se que entre os grupos o fator “método” possua variância uniforme.

Tabela 15 – ANOVA com blocagem de fatores, contendo apenas “método” como fator para a SDR. As sigla GL e SQ representam, respectivamente “Graus de liberdade” e “Soma dos quadrados”.

Fatores	GL	SQ	Média dos quadrados	Valor F	P(>F)
método	5	101,96	20.391	44.44	< 2e-16
Resíduos	210	96.36	0.459		

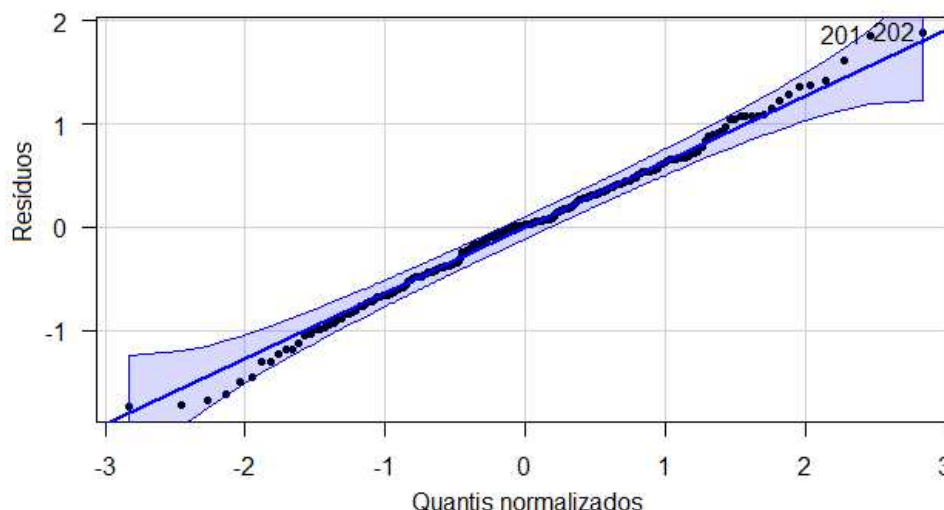


Figura 47 – Gráfico quantil-quantil dos resíduos do modelo ANOVA para o fator “método”, avaliado sob a SDR. Fonte: Autoria própria.

Como não há um método de separação de referência, fez-se uso do Teste de Tukey e todos os métodos foram comparados entre eles. A Tabela 16 mostra os resultados para a diferença de efeito de cada método. A última coluna mostra o valor da probabilidade de aceitar a hipótese nula, de que não há diferença entre os métodos. Os valores em **negrito** mostram as probabilidades que foram menores do que o nível de significância estatística adotado (5%), levando à aceitação da hipótese alternativa.

A partir da Tabela 16, pode-se fazer algumas constatações. Ao observar os testes 11, 12, 13, 14 e 15, que avaliaram as hipóteses da *fastFDICA* em comparação com os demais métodos, observa-se que este apresentou desempenho inferior aos outros. Isso mostra que o processo combinado de descorrelacionar as misturas, restringir o espaço de busca durante a otimização, e ainda pós-processar os sinais separados para atenuar o efeito da ambiguidade de permutação, gera muita distorção, prejudicando a SDR.

Apesar das *FDICAaux* e *FDICAnat* terem desempenhos semelhantes (teste 4), é importante lembrar que otimizar uma função auxiliar quadrática tem maior vantagem do que selecionar a taxa de aprendizagem ideal. Enquanto o primeiro garante estabilidade e convergência, o segundo, além de não garantir estabilidade, pode ter velocidade de convergência reduzida, o que é desfavorável para o caso de processamento em tempo real.

Diferente do caso da *fastFDICA*, a *fastIVA* foi muito superior comparado à metodologia de componentes univariada. Apesar de ainda utilizar a PCA para descorrelacionar as misturas e restringir o espaço de busca, apenas por manter a correlação entre componentes de frequência da mesma fonte por meio dos vetores independentes, resulta em acréscimo de SDR. Nos testes 5 e 7 não rejeitaram as hipóteses nulas ao nível de significância de 5%, apresentando p-valores até 3% acima desse limite em comparação aos outros métodos *FDICA*.

Entre as IVAs, todos tiveram desempenhos semelhantes (teste 1, 2, e 3). Contudo,

o método IVAng apresentou ainda uma ligeira superioridade em termos de SDR. Os métodos possuem procedimentos diferentes, e.g., pré-processamento, escolha da taxa de aprendizagem, escolha da função auxiliar. Isso faz com que a escolha do algoritmo seja realizada tendo como objetivo manter a estabilidade e convergência durante a otimização, além de reduzir a complexidade dos algoritmos.

Tabela 16 – Teste de Tukey para o fator “método”, avaliados pela SDR.

Teste	Hipótese nula	Valor t	P(> t)
1	IVAaux - fastIVA = 0	0,074	1,00
2	fastIVA - IVAng = 0	-0,661	0,99
3	IVAaux - IVAng = 0	-0,587	0,99
4	FDICAaux - FDICAnat = 0	0,102	1,00
5	fastIVA - FDICAaux = 0	2,716	0,08
6	IVAaux - FDICAaux = 0	2,791	0,06
7	fastIVA - FDICAnat = 0	2,818	0,06
8	IVAaux - FDICAnat = 0	2,893	0,04
9	FDICAnat - IVAng = 0	-3,479	0,01
10	FDICAaux - IVAng = 0	-3,378	0,01
11	fastFDICA - FDICAnat = 0	-9,092	< 0,001
12	fastFDICA - IVAng = 0	-12,572	< 0,001
13	FDICAaux - fastFDICA = 0	9,194	< 0,001
14	fastIVA - fastFDICA = 0	11,911	< 0,001
15	IVAaux - fastFDICA = 0	11,985	< 0,001

5.3.3 PESQ8, SDR e SIR

Os resultados da PESQ8 foram similares aos da SDR. Ao recorrer à definição da PESQ8, pode-se concluir que a inteligibilidade é afetada tanto pela interferência quanto pelos artefatos e ruído. Apesar disso, a PESQ8 se relaciona diretamente com a métrica que é a combinação desses efeitos prejudiciais aos sinais processados. Essa relação se mostra tão evidente que, assim como a SDR, a PESQ8 não exibiu valores elevados, de modo que o primeiro e terceiro quartil para todas as misturas envolvidas na análise foram -0,02 e 0,26, respectivamente. Essa baixa variabilidade em relação à escala MOS, a qual varia de 0 a 5, mostra que o ganho médio de pontuação de inteligibilidade por canal (fonte) foi baixo, o que não está relacionado somente com a capacidade de separação dos métodos, mas também com o quanto de artefatos (e distorções) o processo induz ao sinal separado. Assim, mesmo que a PESQ8 seja baixa, é possível que haja boa separação.

A relação existente entre ambas as medidas de desempenho leva à conclusão de que a qualidade do sinal separado (medida pela SDR) impacta diretamente na inteligibilidade do sinal de fala (medida pela PESQ8). Isso não pode ser confundido com uma separação ruim (medida pela SIR), em que cada método busca realizar, podendo ou não deteriorar a qualidade do sinal separado, impactando na inteligibilidade. Assim, usar a PESQ8 e a SDR, mesmo que sejam métricas com objetivos diferentes, nesse caso pode gerar resultados redundantes. Portanto, a escolha de qual medida usar será dependente do problema.

5.3.4 Análises da SIR para fator único

Os resultados da SIR foram fortemente influenciados pelo ambiente e estrutura dos ensaios, além de suas interações de segunda ordem, inviabilizando a blocagem de fatores, como feito para a SDR e PESQ8. A abordagem que se segue foi realizada para cada caso específico de ambiente e ensaio, em que se compara os métodos de separação para a SIR, e, diferente da análise anterior, sem agrupamento dos resultados pela média (16 valores por método). Ao todo foram 36 casos específicos de comparação dos métodos, e em cada caso, os níveis do fator “método” foram comparados pelo Teste de Tukey. Os testes de hipóteses tiveram intervalo de confiança de 95%. Devido à grande quantidade de comparações e repetições de resultados, serão apenas comentados os resultados obtidos.

A avaliação dos métodos nos ambientes “CEFALA” e “2419” mostrou que, para todos os ensaios, não há diferença estatística das abordagens utilizadas, ou seja, os níveis do fator “método” não exibem influência na variável de saída SIR. Em todos os casos, observou-se baixo valor do R^2 , além de esporádicas violações da normalidade dos resíduos e da homocedasticidade.

A comparação dos métodos para o ambiente “Estúdio” apresentou alguns resultados diferentes dos anteriores. Para o ensaio “L57”, o teste de hipóteses apresentou diferenças estatisticamente significativas entre a FDICAnat e fastFDICA, sendo o primeiro exibindo valores superiores ao segundo. No ensaio “L90”, a FDICAnat foi estatisticamente inferior à fastIVA, e no “P57” observou-se inferioridade estatística da fastFDICA em relação à IVAaux e IVAng. No restante dos ensaios, os métodos não apresentaram diferenças estatisticamente significativas.

No ambiente “Seminários”, o único ensaio em que houve diferenças significativas foi no “M57”, em que as IVAs foram estatisticamente superiores à FDICAnat. Além disso, a FDICAaux foi inferior à IVAng e IVAaux.

As análises desta seção mostraram que o ganho médio de SIR por sinal separado pelos métodos é estatisticamente semelhante. Contudo, observou-se que, quando há diferenças, os métodos IVAs têm superioridade, pois não foram estatisticamente inferiores a nenhuma FDICA. O comportamento dessa métrica mostra que, apesar de haver similaridades nos resultados, cada método separa de forma diferente, ao otimizar o melhor vetor de projeção dos dados para cada fonte. Porém, apesar de ocorrer a minimização da função de custo em

todos os casos (SIR aumenta), as distorções impostas pelas abordagens exibem diferenças significativas, como mostrado na seção anterior.

5.4 Efeitos do Posicionamento dos Microfones

Além das comparações dos algoritmos, este estudo teve o objetivo de analisar o impacto do posicionamento dos microfones no desempenho da separação de fontes, sem a intenção de determinar a configuração ideal. Consequentemente, ao avaliar apenas o efeito da distância do arranjo em relação às fontes, as distâncias entre os microfones foram agrupadas. De forma similar, ao examinar o efeito do espaçamento entre microfones, as distâncias até as fontes também foram agrupadas. Essa investigação foi realizada apenas para os algoritmos IVA e somente nas métricas SDR e PESQ8. Isso porque as IVAs obtiveram melhores resultados, e as métricas SDR e PESQ8 terem sido mais importantes no estudo comparativo, como foi abordado anteriormente.

Como todos os resultados de cada algoritmo foram semelhantes, apenas aqueles referentes ao IVAaux são apresentados e discutidos. Os resultados para o IVAng e o fastIVA são mostrados no Apêndice B. As comparações estatísticas que se seguem avaliaram o desempenho do IVAaux para diferentes posições dos microfones em salas com diferentes valores de RT_{60} , com nível de significância estatística de 5%, utilizando o Teste de Tukey [Tukey, 1949]. Assim, foi possível comparar os efeitos de cada nível dos fatores isoladamente (para uma sala específica), bem como a comparação entre os seus níveis para diferentes tempos de reverberação.

Nesta seção, é apresentada primeiro a comparação dos efeitos dos níveis do primeiro fator, ou seja, as distâncias do arranjo em relação às fontes. Em seguida, são analisados os efeitos dos níveis do segundo fator (distâncias entre microfones).

5.4.1 Distância entre microfones e fontes

As Figuras 48 e 49 mostram como as medidas de desempenho SDR e PESQ8, respectivamente, se comportaram diante das variações no posicionamento entre fontes e microfones.

Primeiramente, para a SDR nas salas “Estúdio” e “Seminário”, alterar a distância entre os microfones e as fontes não resultou em mudanças significativas no desempenho dentro de cada sala separadamente. O mesmo foi observado na sala “2419”, onde a variação dessas distâncias não resultou em diferenças de desempenho. Um comportamento diferente foi observado no ambiente “CEFALA”, onde o posicionamento mais distante ($L_3 = 210$ cm) foi estatisticamente inferior aos outros dois.

Na Figura 49, observou-se o mesmo comportamento para a métrica PESQ8. No caso da sala “CEFALA”, esse comportamento foi ainda mais evidente devido à menor variabilidade em comparação com a primeira métrica.

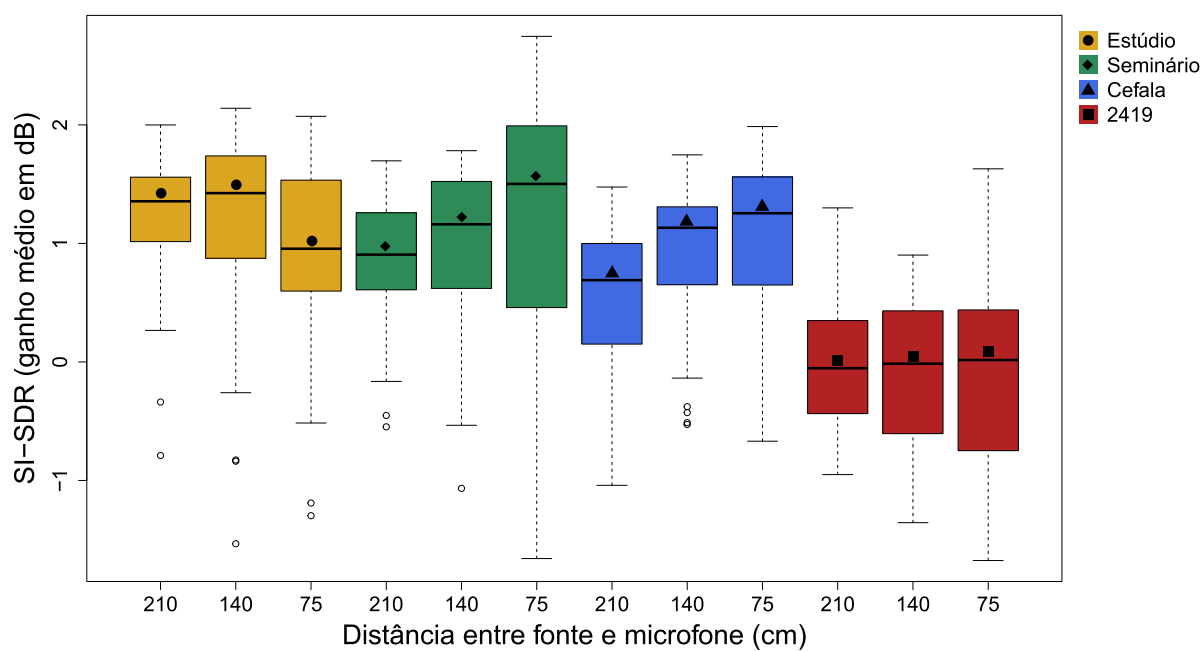


Figura 48 – Ganho médio de SDR por canal para variações na distância entre microfones e fontes em cada sala. Fonte: Autoria própria.

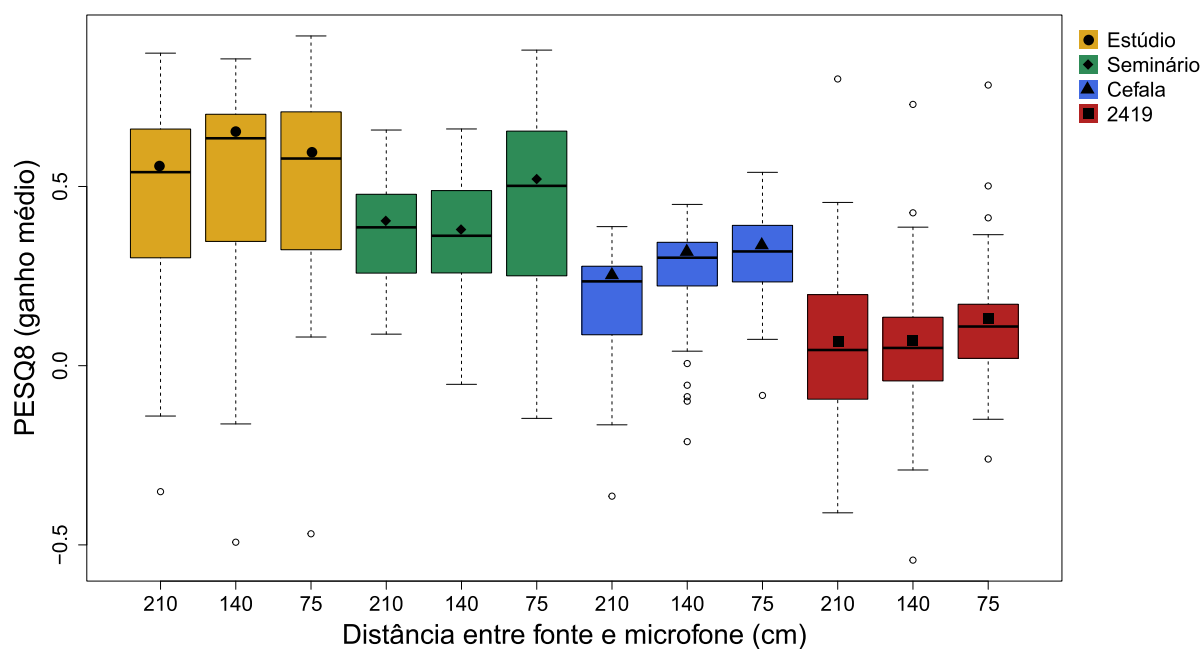


Figura 49 – Ganho médio de PESQ por canal para variações na distância entre microfones e fontes em cada sala. Fonte: Autoria própria.

Ambas as medidas de desempenho mostraram redução em seus valores com o aumento do tempo de reverberação, independentemente da distância entre o arranjo de microfones e as fontes.

Nas salas analisadas neste trabalho, o tempo de reverberação teve um impacto maior no desempenho dos algoritmos de separação do que a distância.

5.4.2 Distância entre microfones

As Figuras 50 e 51 mostram o comportamento das métricas de desempenho, SDR e PESQ8, em relação às variações no posicionamento entre os microfones. Pode-se observar que a alteração na posição relativa entre os microfones influenciou diretamente o desempenho do algoritmo de separação. De modo geral, para as métricas SDR e PESQ8, observa-se que o desempenho da separação é reduzido nos casos em que os microfones estão posicionados muito próximos entre si, especialmente quando essa distância é de 0 cm.

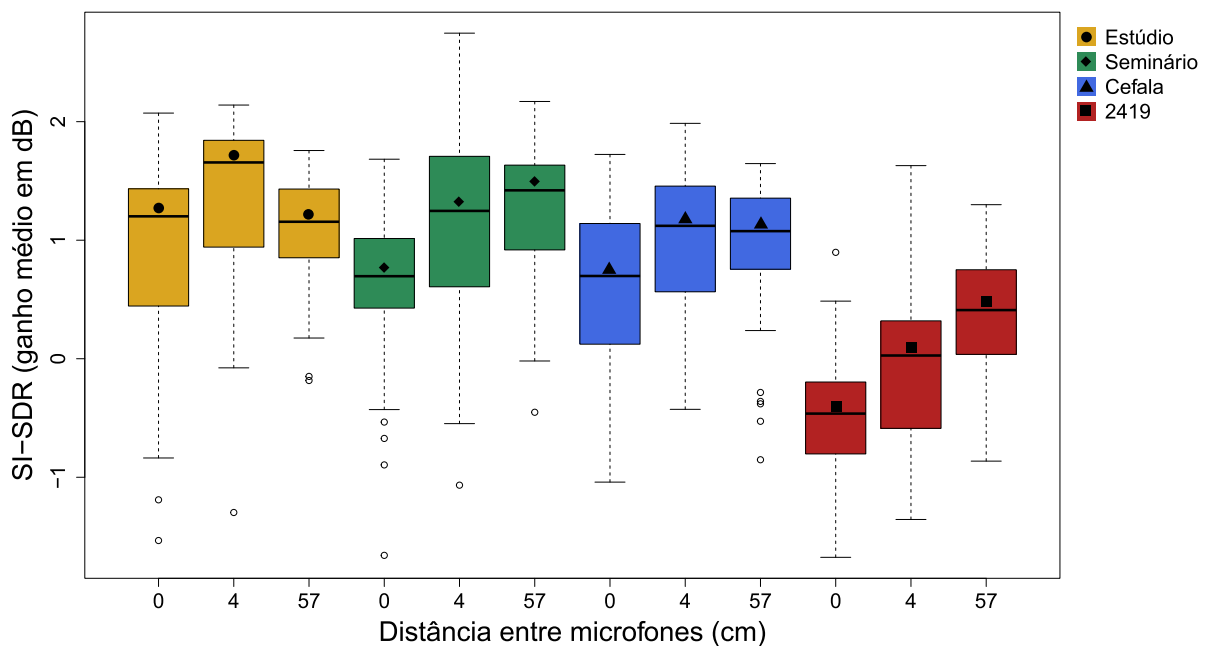


Figura 50 – Ganho médio de SDR por canal para variações na distância entre microfones em cada sala. Fonte: Autoria própria.

Essa queda de desempenho pode ser explicada por dois aspectos fundamentais que influenciam os algoritmos de separação e que estão relacionados entre si: o modelo espacial das misturas e a representação tempo-frequência dos sinais.

Quanto ao primeiro aspecto, sabe-se que o sistema de mistura é determinado pelas AIRs da sala. Quando os microfones estão localizados no mesmo ponto espacial ou muito

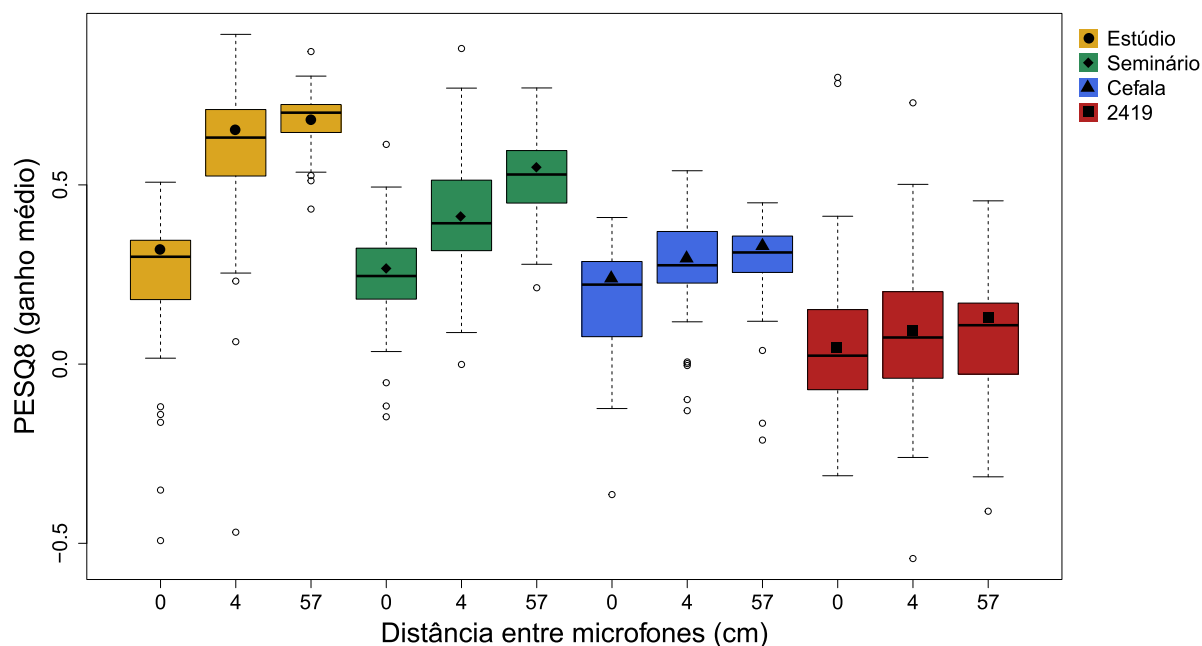


Figura 51 – Ganho médio de PESQ por canal para variações na distância entre microfones em cada sala. Fonte: Autoria própria.

próximos entre si, suas AIRs tendem a ser bastante semelhantes, o que pode resultar em matrizes de mistura com características próximas da singularidade no domínio da frequência. Esse fenômeno dificulta a estimativa do sistema de separação, comprometendo a eficácia do algoritmo.

O segundo aspecto está relacionado a como a separação é realizada no domínio da frequência, o que depende tanto da magnitude quanto da fase das misturas captadas. Como os microfones estão praticamente no mesmo ponto espacial, a amostragem espacial dos sinais é muito semelhante para ambos, resultando em misturas altamente correlacionadas tanto em magnitude quanto em fase. Essa forte correlação corrobora a baixa diversidade espacial das informações disponíveis para o algoritmo e, consequentemente, limita sua capacidade de identificar e separar as fontes de forma eficiente.

Nas três primeiras salas analisadas, a métrica SDR apresentou pequenas diferenças no desempenho da separação, dependendo da distância entre os microfones, como observado na sala “CEFALA”, onde a variação entre as distâncias de 0 cm e 4 cm não apresentou diferenças significativas (embora entre 0 cm e 57 cm tenha sido estatisticamente diferente). No entanto, esse efeito foi mais evidente na sala “2419”, onde as diferenças de desempenho entre diferentes distâncias se tornaram mais significativas, sugerindo uma maior sensibilidade do ambiente às variações no arranjo dos sensores.

Por outro lado, para a métrica PESQ8, o comportamento observado foi mais claro e oposto, mostrando que as métricas podem apresentar comportamentos distintos. Para as

três primeiras salas, observou-se que maiores espaçamentos entre os microfones resultaram em melhor desempenho para a IVAaux. No entanto, esse efeito mostrou-se dependente do tempo de reverberação RT_{60} ; à medida que ele aumentou, a influência da distância entre os microfones foi reduzida. Na sala “2419”, em particular, independentemente da configuração espacial dos microfones, não foram observadas diferenças estatisticamente significativas para os fatores analisados, sugerindo que a consequência do alto tempo de reverberação é a neutralização da influência do espaçamento entre os microfones.

De forma geral, os resultados corroboram os achados da subseção anterior, indicando que a reverberação tem um impacto significativo e generalizado na redução do desempenho da separação, tanto para as métricas SDR quanto PESQ8.

5.5 Conclusões do capítulo

Neste capítulo, foram apresentados os resultados da separação cega de sinais de fala. De modo geral, foi mostrado que o desempenho dos métodos de separação cega de fontes depende fortemente das condições acústicas e do posicionamento dos microfones, sendo a reverberação o principal fator de degradação. Ambientes com menor RT_{60} , como o estúdio, apresentaram os melhores resultados em todas as métricas analisadas (SIR, SDR e PESQ8). Entre os métodos avaliados, as IVAs (IVAaux, fastIVA e IVAng) destacaram-se, enquanto a fastFDICA teve o pior desempenho. Quanto ao arranjo dos microfones, observou-se que distâncias muito pequenas entre eles resultaram em forte correlação entre as misturas e pior separação.

Capítulo 6

Conclusão

O presente estudo investigou como diversos fatores influenciam o desempenho da separação cega da fala em salas reverberantes, uma importante etapa de processamento de sinais presente em diversos dispositivos para várias finalidades. Os fatores avaliados nesse estudo foram: algoritmos de separação, distância entre fontes e microfones, e distância entre microfones, todos avaliados em diferentes salas, com tempos de reverberação diferentes.

A comparação dos algoritmos foi realizada para investigar se algum método poderia ser superior em dados cenários acústicos. Assim, observou-se que as três abordagens para misturas instantâneas nomeadas como ICA_kurt, ICA_negen e ICA_ML, não obtiveram bons resultados devido à natureza convolutiva das misturas. Nas abordagens no domínio da frequência, a fastFDICA teve valores reduzidos comparados às FDICAs e às IVAs para as métricas SDR e PESQ8, ao nível de significância estatística de 5%. Entretanto, as FDICAs tiveram desempenho abaixo comparado às IVAs, mostrando que o modelo de fontes multivariável (vetor aleatório) é superior à suposição do modelo de fontes univariada (componente aleatório). Entre às IVAs, não foram observadas diferenças estatísticas entre os algoritmos fastIVA, IVAng e IVAaux. No geral, todos os algoritmos tiveram deterioração no desempenho com o aumento do tempo de reverberação.

A justificativa para comparação do posicionamento dos microfones foi devido à dependência dos algoritmos pela informação espacial, o que pode influenciar no desempenho da separação. Os resultados mostram que, em geral, o desempenho da separação cega de fala foi reduzido com o aumento do tempo de reverberação do ambiente. Com exceção da sala “Estúdio”, mudanças na distância entre microfones e fontes não apresentaram diferenças estatisticamente significativas para ambas as métricas avaliadas. A mudança na distância entre os microfones mostrou que, para a SDR, microfones espacialmente próximos resultam em misturas altamente correlacionadas, o que impede uma separação adequada. Isso se tornou mais evidente para a métrica de desempenho PESQ8, onde essa diferença diminuiu com o aumento do RT_{60} .

Dessa forma, as evidências levantadas neste estudo demonstram que a Separação Cega da Fala em ambientes reverberantes é um processo complexo, e depende da escolha

do algoritmo, das condições acústicas e da distribuição espacial das fontes e microfones. As abordagens baseadas em modelos convolutivos, particularmente as IVAs, tiveram uma superioridade perante às FDICAs e aos métodos de separação para misturas instantâneas, reforçando a validade do modelo de fontes multivariável. A dependência mais crítica observada foi em relação ao tempo de reverberação, em que o aumento deste parâmetro levou a uma deterioração de desempenho generalizada, independentemente do algoritmo ou posicionamento. Contudo, demonstrou-se que a otimização do arranjo de microfones é crucial, pois distâncias muito curtas introduzem alta correlação nas misturas, prejudicando a separação, enquanto um posicionamento adequado pode atenuar a degradação em salas específicas.

Dessa forma, os resultados obtidos neste trabalho permitem concluir que o desempenho da separação cega de fala em salas reverberantes depende do tipo de algoritmo, do tempo de reverberação, além da posição dos microfones. Os métodos baseados em modelos convolutivos, especialmente as IVAs, mostraram melhor desempenho, enquanto as FDICAs tiveram resultados inferiores. Também foi verificado que o aumento do tempo de reverberação reduz significativamente a qualidade da separação, independentemente do algoritmo e do posicionamento. Assim, conclui-se que a melhoria da separação cega da fala em aplicações reais, reside na escolha do algoritmo e no posicionamento adequado dos microfones, o que pode melhorar o desempenho em determinadas salas, dependendo do tempo de reverberação.

6.1 Trabalhos Futuros

O problema de separação pode ser visto como um problema de representação de sinais (modelo espectral) e escolha de estrutura (sistema de misturas). Assim, uma possibilidade de estudo futuro seria a criação de uma abordagem que pudesse lidar com esses problemas de forma automática, para fontes em movimento.

Outro estudo seria a aplicação de filtragem ótima para redução do ruído nas misturas, de modo a investigar se tal pré processamento pode interferir no desempenho dos métodos.

Uma outra investigação seria adicionar outros métodos de separação que utilizam, por exemplo, o NMF para representar os sinais, como no ILRMA e MNMF.

Como os modelos espaciais foram obtidos pela emissão do TSP, um estudo possível é a utilização dessas AIRs para treinar redes profundas (aprendizagem supervisionada) para separar sinais de misturas obtidas nas salas dos experimentos.

6.2 Publicações

Bôa, G. M. M. C. F.; Yehia, H. C.; Barbosa, A. V. Separação Cega da Fala: Estudo Comparativo de Algoritmos em Salas Reverberantes. Simpósio Brasileiro de Automação Inteligente, São João del-Rei, 2025.

Bôa, G. M. M. C. F.; Yehia, H. C.; Barbosa, A. V. Blind speech separation in reverberant rooms: effects of microphone positioning. 54th International Congress & Exposition on Noise Control Engineering (Inter-Noise), São Paulo, 2025.

REFERÊNCIAS

- L. A. Aguirre. *Introdução à Identificação de Sistemas. Técnicas Lineares e não Lineares Aplicadas a Sistemas. Teoria e Aplicação*. Editora UFMG, 4^a edição edition, 2015.
- S. Amari, A. Cichocki, and H. Yang. A new learning algorithm for blind source separation. *Proceedings of the 8th International Conference on Neural Information Processing Systems*, 1995.
- J. Anemüller and B. Kollmeier. Amplitude modulation decorrelation for convolutive blind source separation. *Proceedings of the 2nd International Workshop on Independent Component Analysis and Blind Signal Separation*, 2000.
- S. Ansari, K. Alnajjar, T. Khater, S. Mahmoud, and A. Hussain. A robust hybrid neural network architecture for blind source separation of speech signals exploiting deep learning. *IEEE Access*, VOL. 11, 2023.
- N. Aoshima. Computer-generated pulse signal applied for sound measurement. *Journal of the Acoustical Society of America*, 69, 1484–1488, 1981.
- A. K. Barros and A. Cichocki. Extraction of specific signals with temporal structure. *Neural Computation*, 13(9):1995–2003, 2001.
- Behringer. Behringer ecm8000. https://www.manualpdf.com.br/behringer/ecm8000/manual#google_vignette, 2024. Acesso em: 12 de dezembro 2024.
- A. Bell and T. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 1995.
- E. Bingham and A. Hyvärinen. A fast fixed-point algorithm for independent component analysis of complex valued signals. *International Journal of Neural Systems*, 2000.
- J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):354–379, 2012. doi: 10.1109/JSTARS.2012.2194696.

- A. Brandel and W. Kellermann. Fasteriva: update rules for independent vector analysis based on negentropy and the majorize-minimize principle. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2021.
- H. Buchner, R. Aichner, and W. Kellermann. Trinicon: a versatile framework for multi-channel blind signal processing. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.
- C. Cabos. P2 estéreo para 2xp10 mono. <https://www.cirilocabos.com.br/cabo-p2-estereo-para-2-p10-mono-por-metro/p>, 2024. Acesso em: 12 de dezembro 2024.
- J.-F. Cardoso. Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1990.
- J.-F. Cardoso and B. H. Laheld. Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, 44(12):3017–3030, 1996.
- J.-F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 1996.
- J.-F. Cardoso, H. Snoussi, J. Delabrouille, and G. Patanchon. Blind separation of noisy gaussian stationary sources. application to cosmic microwave background imaging. In *Proceedings of the 11th European Signal Processing Conference (EUSIPCO 2002)*, 2002.
- E. Cherry. Some experiments on the recognition of speech, with one and with two ear. *Journal of the Acoustical Society of America*, 25(5), 975-979, 1953.
- S. Chetupali and E. Habets. Speech separation for an unknown number of speakers using transformers with encoder-decoder attractors. *International Speech Communication Association Annual Conference*, 2022.
- P. Comon. Independent component analysis, a new concept? *Signal Processing, Elsevier*, 1994.
- N. Duong, E. Vincent, and R. Grimbonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, VOL. 18, NO. 7, 2010.
- R. Guo, Z. Luo, and M. Li. A survey of optimization methods for independent vector analysis in audio source separation. *MDPI Sensors*, 2023.
- A. Hiroe. Solution of permutation problem in frequency domain ica, using multivariate probability density functions. *Independent Component Analysis and Blind Signal Separation – ICA 2006, Springer*, 2006.

- X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall, 2001.
- A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 1997.
- A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks* 13, page 411–430, 2000.
- A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.
- International Organization for Standardization. ISO 3382-2:2008 Acoustics — Measurement of room acoustic parameters — Part 2: Reverberation time in ordinary rooms. <https://www.iso.org/standard/36201.html>, 2008. Geneva: ISO.
- International Telecommunication Union. PESQ Script Files. <https://www.itu.int/rec/T-REC-P.862>, 2001. Acesso em: 05 de julho 2024.
- C. Jutten and J. Herault. A neuromimetic solution for the problem of sources discrimination. *Machines Numériques*, 1988.
- C. Jutten and J. Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing, Elsevier*, 1991.
- S. M. Kay. *Intuitive Probability and Random Processes using MATLAB*. Springer, 1st edition, 2006.
- H.-Y. Kim, F. Asano, Y. Suzuki, and T. Sone. Measurement of long impulse response by using a time-stretched pulse. *Research Institute of Electrical Communication, Tohoku University*, 1994.
- T. Kim, T. Eltoft, and T.-W. Lee. Independent vector analysis: An extension of ica to multivariate components. *Independent Component Analysis and Blind Signal Separation – ICA 2006, Springer*, 2006.
- D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE Transactions on Audio, Speech, and Language Processing*, 2016.
- H. Kuttruff. *Room Acoustics*. CRC Press, London, 5th edition, 2009.
- P. Ladefoged. *Vowels and Consonants: An Introduction to the Sounds of Languages*. Blackwell, 2001.
- D. Lee and H. Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 1999.

- I. Lee, T. Kim, and T.-W. Lee. Fast fixed-point independent vector analysis algorithms for convolutive blind source separation. *Signal Processing 87 Elsevier*, page 1859–1871, 2007.
- J. LeRoux, S. Wisdom, H. Erdogan, and J. Hershey. Sdr–half-baked or well done? *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 226–230, 2019.
- S. Makeig, A. J. Bell, T.-P. Jung, and T. J. Sejnowski. Independent component analysis of electroencephalographic data. In *Advances in Neural Information Processing Systems (NIPS 8)*, pages 145–151. MIT Press, 1996.
- S. Makino, T.-W. Lee, and H. Sawada. *Blind Speech Separation*. Springer, Kyoto, 2007.
- K. Matsuoka and S. Nakashima. Minimal distortion principle for blind source separation. *Proceedings of the 3rd International Workshop on Independent Component Analysis and Blind Signal Separation*, 2001.
- Minipa. DecibellÍmetro digital sonômetro digital digital sound level meter msl-1355b.
- N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing 41, Elsevier*, 2001.
- N. Ono. Stable and fast update rules for independent vector analysis based on auxiliary function technique. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011.
- N. Ono. Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions. *Proceedings of The Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA/ASC)*, 2012.
- N. Ono and S. Miyabe. Auxiliary-function-based independent component analysis for super-gaussian sources. *Latent Variable Analysis and Signal Separation (LVA/ICA)*, Springer, 2010.
- OpenAI. Chatgpt (versão gpt-3). <https://chat.openai.com>, 2023. Acesso em: 29 maio 2024.
- A. V. Oppenheim, R. W. Schaffer, and J. R. Buck. *Discrete-Time Signal Processing*. Prentice Hall, 2nd edition, 1999.
- A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing, VOL. 18, NO. 3*, 2010.

- Presonus. Presonus firestudio project. <https://www.sonology.shop/product/presonus-firestudio-project.html>, 2024. Acesso em: 12 de dezembro 2024.
- Project Jupyter. Jupyter notebook, 2015. URL <https://jupyter.org>.
- L. Rabiner and R. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.
- A. Rix, J. Beerends, M. Hollier, and A. Hekstra. Perceptual evaluation of speech quality (pesq) – a new method for speech quality assessment of telephone networks and codecs. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.
- H. Sawada, R. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Transactions on Speech and Audio Processing*, VOL. 12, NO. 5, 2004.
- H. Sawada, H. Kameoka, S. Araki, and N. Ueda. Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on Audio, Speech, and Language Processing*, VOL. 21, NO. 5, 2013.
- H. Sawada, N. Ono, H. Kameoka, and D. Kitamura. Blind audio source separation on tensor representation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.
- M. Schroeder. New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(3), pages 409–412, 1965.
- S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965. doi: 10.1093/biomet/52.3-4.591. URL <https://doi.org/10.1093/biomet/52.3-4.591>.
- Shopee. Cabo firewire. <https://shopee.com.br/Cabo-Firewire-6-Pin-macho-X-6-Pin-macho-%284%29-i.848885601.18942573050>, 2024. Acesso em: 12 de dezembro 2024.
- P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 1998.
- Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone. An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses. *J. Acoust. Soc. Am.*, vol. 97, no. 2, pages 1119–1123, 1995.
- H. L. V. Trees. *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. Wiley-Interscience, 2002.
- J. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, 5(2), pages 99–114, 1949.

- A.-J. van der Veen and A. Paulraj. An analytical constant modulus algorithm. *IEEE Transactions on Signal Processing*, 44(5):1136–1155, 1996. doi: 10.1109/78.502327.
- E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Speech and Audio Processing*, 14:1462–1469, 2006.
- E. Vincent, T. Virtanen, and S. Gannot. *Audio Source Separation and Speech Enhancement*. Wiley, 2018.
- T. Virtanen. Separation of sound sources by convolutive sparse coding. *ITRW on Statistical and Perceptual Audio Processing*, 2004.
- T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, VOL. 15, NO. 3, 2007.
- D. Wang and J. Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018. doi: 10.1109/TASLP.2018.2842159.
- Q. Wang, S. Low, Z. Li, and K. Yiu. Sensor placement optimization of blind source separation in a wireless acoustic sensor network via hybrid descent methods. *Applied Acoustics 188 (2022)*, 2021.
- K. Yatabe and D. Kitamura. Determined bss based on time-frequency masking and its application to harmonic vector analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, VOL. 29, 2021.
- H. Yehia. Gerador de tsp. Script do MATLAB, 1998.
- A. Yeredor. On hybrid exact-approximate joint diagonalization. *3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, 2009.
- E. Zwicker. Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America*, 33(2):248–249, 1961.

Apêndice A

Implementações dos métodos auxiliares

Os métodos auxiliares foram um conjunto de funções criadas para reduzir e facilitar as implementações dos métodos de separação, os quais dependem de uma série de procedimentos. Todas as funções são descritas a seguir, e seus respectivos códigos armazenados localmente no arquivo “util.py”.

A função “length_match” recebe dois sinais de tamanhos diferentes e retorna ambos sinais com tamanhos iguais, tarefa a qual é realizada usando a adição de zeros (*zero padding*) ao fim do sinal de menor comprimento.

A “norm_sig” é uma função que recebe um sinal e o retorna normalizado pelo maior valor absoluto, de modo que o restrinja ao intervalo de -1 a +1.

A função “changeFs” recebe como parâmetros um sinal, sua atual frequência de amostragem e a nova frequência de amostragem, e retorna o sinal reamostrado. Isso foi feito utilizando a filtragem polifásica, devido à sua baixa complexidade e sua redução do consumo de memória em relação aos processos diretos de interpolação e decimação [Oppenheim et al., 1999]. Utilizou-se a função “resample_poly”, do módulo “scipy.signal”.

A “corr” é uma função que recebe duas sequências de variáveis aleatórias, e retorna o coeficiente de correlação de Pearson entre elas.

A função “perm_solver” é responsável por resolver o problema de permutação. Os parâmetros são o vetor de matrizes de separação de cada componente de frequência e as estimativas dos sinais das fontes (sinais separados) na representação tempo-frequência. Apesar de haver várias formas de resolver esse problema, escolheu-se a abordagem de correlação dos sinais para frequências adjacentes [Sawada et al., 2004]. Assim, quando ocorre permutação, as linhas da matriz de separação devem ser permutadas, resultando em uma representação tempo-frequência com menor influência da outra fonte e, conseqüentemente, menos distorção e melhor separação.

A resolução do problema de escala foi realizada pela função “scale_solver”. Ela recebe como parâmetro as matrizes de separação em cada frequência e, por meio do princípio da mínima distorção, retorna as matrizes de separação com escalas ajustadas.

A função “pca” foi implementada para realizar o PCA. Apesar desse método ser

muito utilizado para reduzir a dimensionalidade dos dados, aqui ele foi empregado para descorrelacionar os dados, o que é essencial em alguns métodos de separação. Dado isso, a função possui como parâmetro de entrada os sinais das misturas $\mathbf{X}$, sejam temporais ($\mathbf{X} \in \mathbb{R}$) ou na representação tempo-frequência ($\mathbf{X} \in \mathbb{C}$), e retorna os sinais descorrelacionados.

Por fim, as funções “stft” e “istft” se referem à transformação dos sinais multivariados no domínio do tempo para suas representações tempo-frequência pela Transformada de Fourier de Tempo Curto (STFT, do inglês, *Short Time Fourier Transform*), e as transformadas inversas, respectivamente. Fez-se uso das funções de mesmo nome do módulo “scipy.signal”, que realizam a tarefa apenas para um sinal. Dentre vários parâmetros das funções implementadas, apenas quatro foram essenciais neste trabalho: sinal temporal multivariado ou STFT multivariada (no caso da “istft”), a frequência de amostragem, o número de amostras para um quadro de análise e o número de amostras sobrepostas entre um quadro e outro. Além disso, ambas as funções utilizaram janelas de Hanning para síntese e análise, as quais não foram alteradas ao longo da execução do trabalho.

Apêndice B

Resultados do Posicionamento

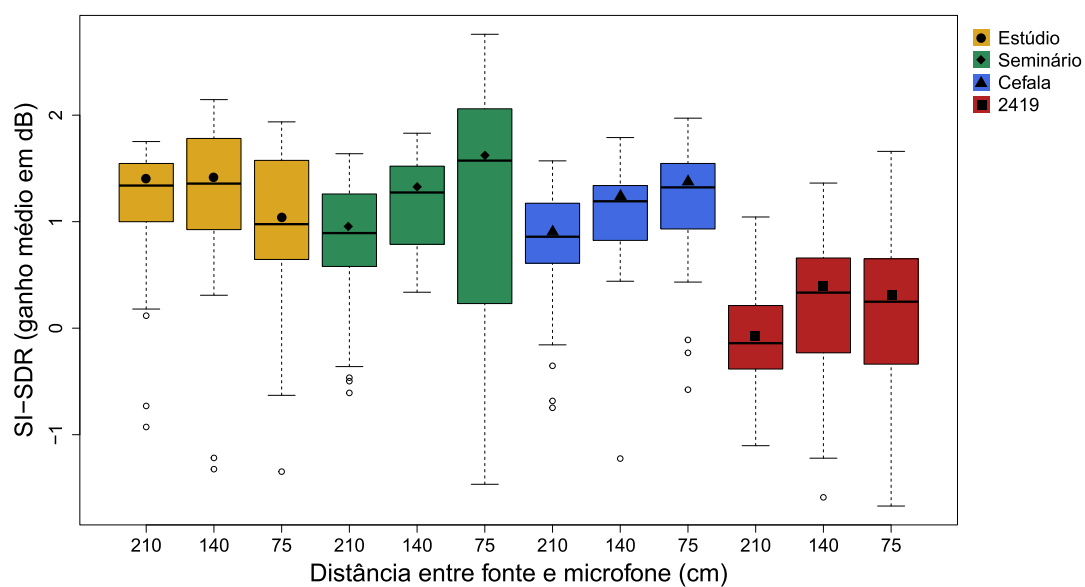


Figura 52 – Ganho médio de SI-SDR para variações de distância entre fonte e microfone em cada sala. Algoritmo: IVAng. Fonte: Autoria própria.

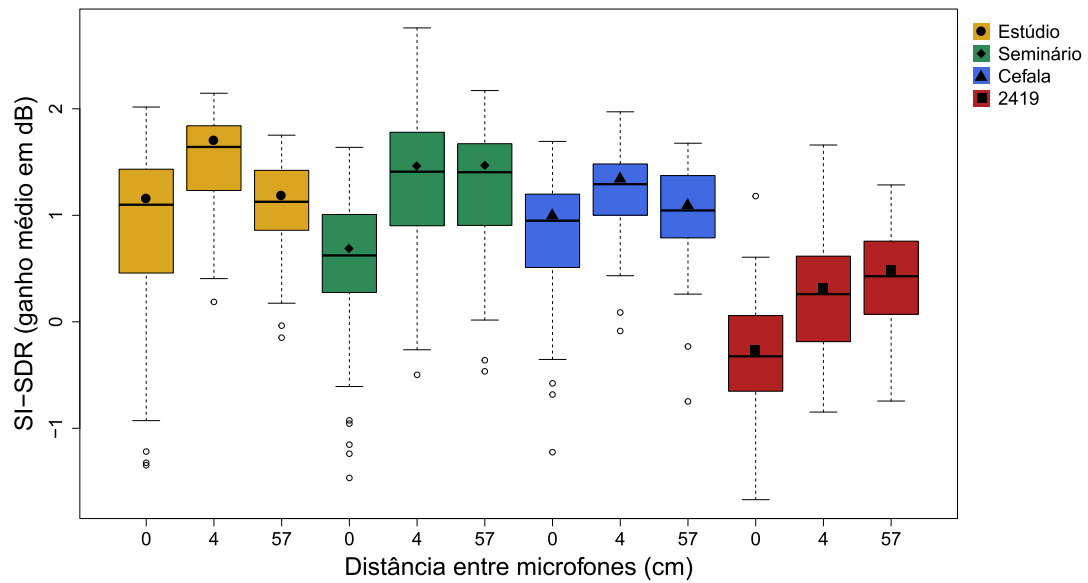


Figura 53 – Ganho médio de SI-SDR para variações de distância entre microfones em cada sala. Algoritmo: IVAng. Fonte: Autoria própria.

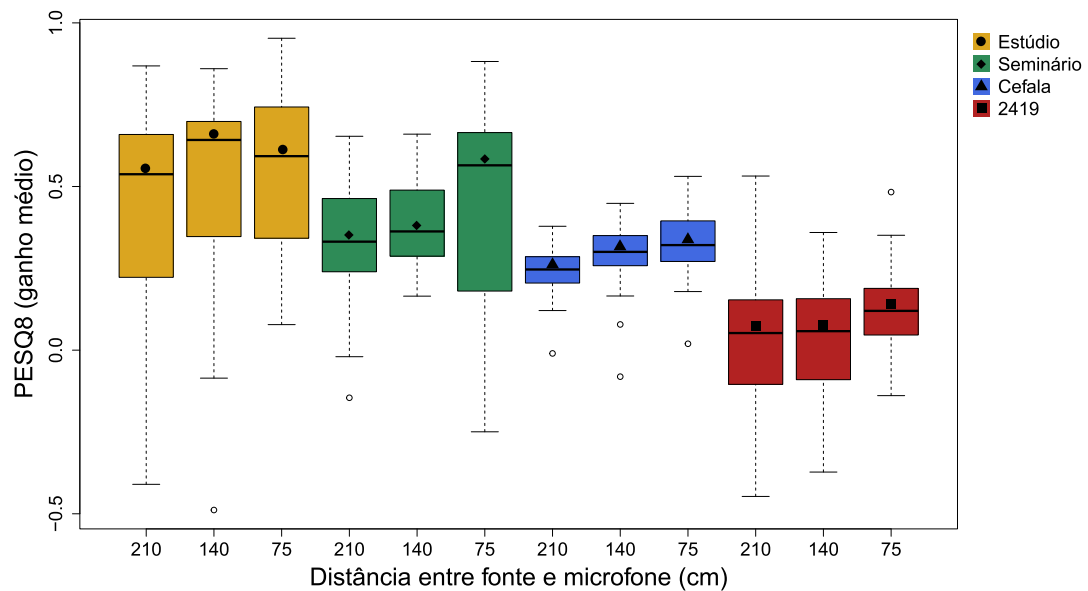


Figura 54 – Ganho médio de PESQ para variações de distância entre fonte e microfone em cada sala. Algoritmo: IVAng. Fonte: Autoria própria.

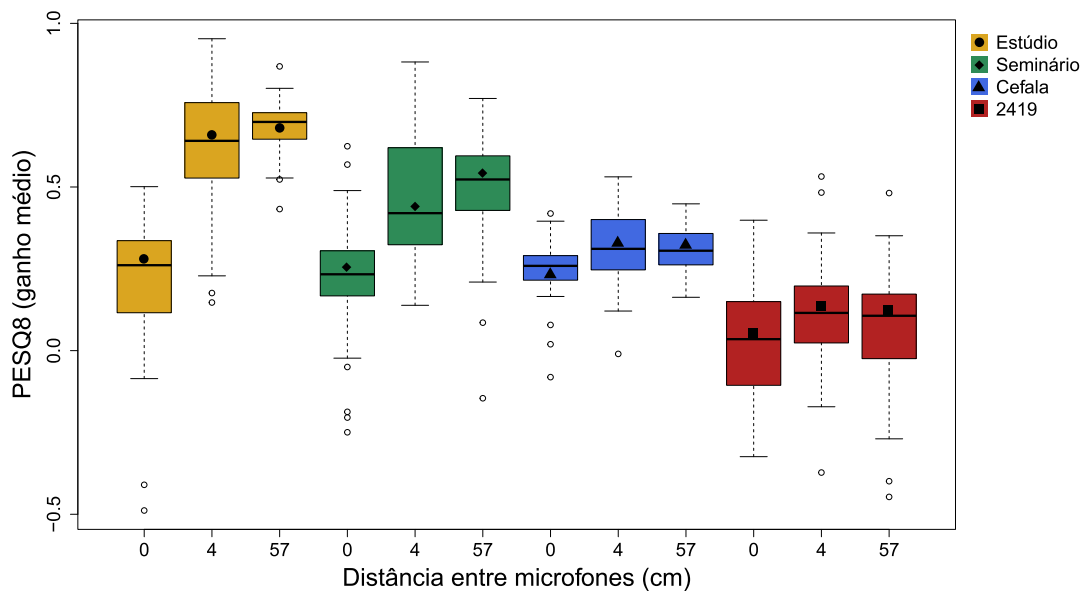


Figura 55 – Ganho médio de PESQ para variações de distância entre microfones em cada sala. Algoritmo: IVAng. Fonte: Autoria própria.

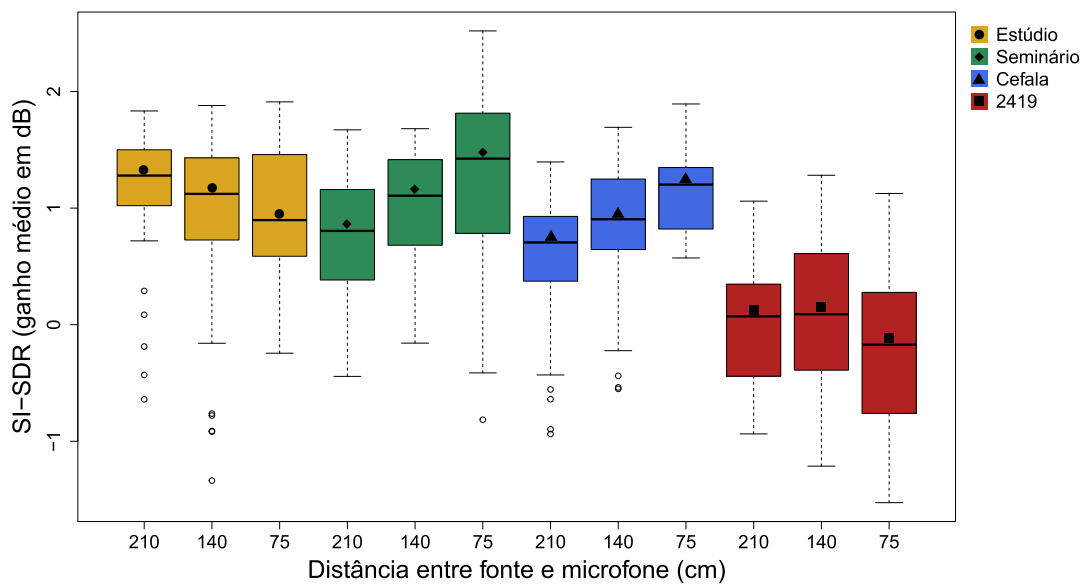


Figura 56 – Ganho médio de SI-SDR para variações de distância entre fonte e microfone em cada sala. Algoritmo: fastIVA. Fonte: Autoria própria.

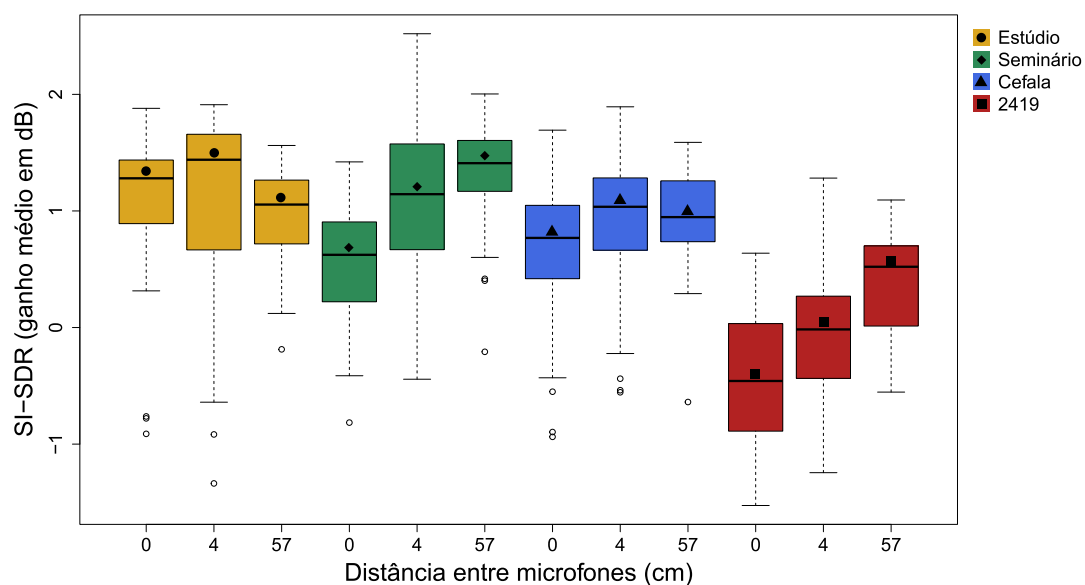


Figura 57 – Ganho médio de SI-SDR para variações de distância entre microfones em cada sala. Algoritmo: fastIVA. Fonte: Autoria própria.

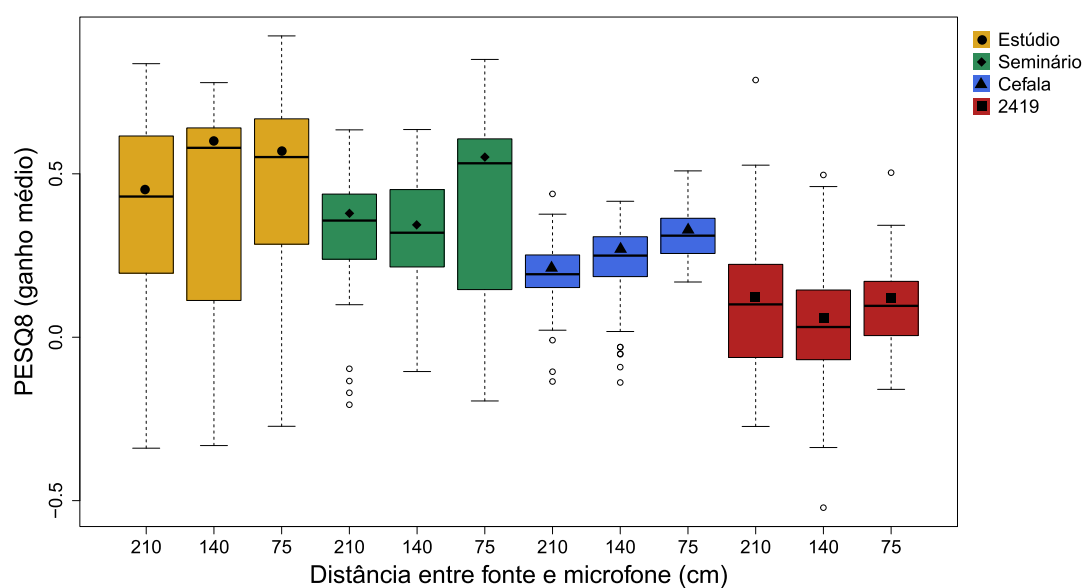


Figura 58 – Ganho médio de PESQ para variações de distância entre fonte e microfone em cada sala. Algoritmo: fastIVA.

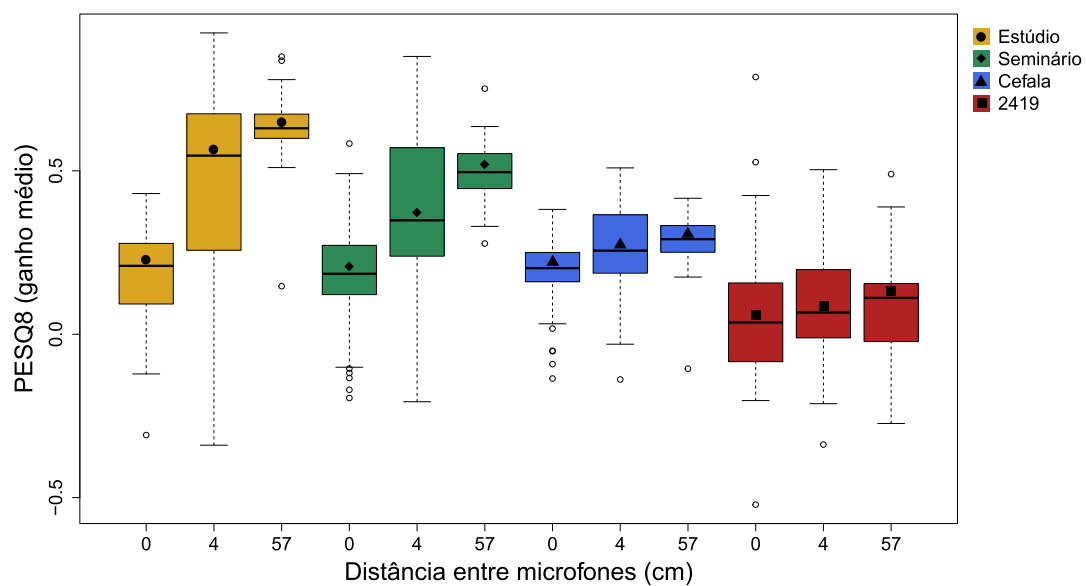


Figura 59 – Ganho médio de PESQ para variações de distância entre microfones em cada sala. Algoritmo: fastIVA. Fonte: Autoria própria.