

Universidade Federal de Minas Gerais
Programa de Pós-Graduação em Engenharia Elétrica

Um Estudo sobre Limitações de Técnicas de Mascaramento Espectral na Separação Cega de Sinais de Voz Reverberados

Gustavo Fernandes Rodrigues

Tese de doutorado submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do Título de Doutor em Engenharia Elétrica.

Orientador: Hani Camille Yehia

Belo Horizonte
Maio/2008

A Deus,
aos meus pais,
a Aletea,
e a Sofia.

Resumo

O objetivo deste trabalho é analisar e verificar as limitações de técnicas de mascaramento binário no domínio tempo-frequência para o problema de separação cega de fontes (*Blind Source Separation*) a partir de misturas convoluídas. Tais técnicas baseiam-se na esparsidade dos sinais de voz e consideram que, em uma mistura de fontes independentes, é improvável encontrar sinais ativos em um mesmo instante de tempo e frequência. Neste trabalho, inicialmente, é verificado o desempenho do algoritmo *DUET* para misturas convoluídas. A utilização da informação de fase para a separação de fontes é avaliada através da análise de desempenho de um algoritmo proposto que considera apenas a informação de amplitude relativa para a estimação das máscaras espectrais. Observa-se que, no caso específico da separação de dois sinais de voz, a utilização da informação de fase não influi no desempenho do algoritmo. No próximo passo, analisa-se as limitações da técnica de mascaramento espectral em função do tempo de reverberação dos sinais que formam a mistura a ser separada. Quando se conhece as máscaras ideais, obtém-se uma separação da ordem de 9 *dB*, para o caso de um ambiente com tempo de reverberação inferior a 300 *ms*. A partir daí, à medida que o tempo de reverberação aumenta, os sinais que formam as misturas analisadas espalham-se sobre o plano tempo-frequência, reduzindo progressivamente o desempenho do processo de separação. A seguir, dado que encontrar máscaras ideais com base em apenas uma mistura é ainda um problema em aberto, mede-se a queda de desempenho do processo de separação em função da distância entre a máscara ideal e a máscara efetivamente usada. Os resultados encontrados indicam uma queda de desempenho de 3 *dB* quando aproximadamente 10% dos bits da máscara ideal estão invertidos. Finalmente, são realizadas análises preliminares para encontrar a máscara ideal com base na negentropia, na curtose e na energia dos sinais separados.

Abstract

The objective of this study is to analyze the limitations of techniques for blind source separation (*BSS*) of convolved mixtures based on time-frequency domain binary masking. These techniques are based on the sparsity of speech signals and assume that, in a mixture of independent speech signals, for each time and frequency, it is unlikely to find more than one active source. In this study, initially, the performance of the *DUET* algorithm is analyzed for convolved mixtures. The use of phase information for source separation is evaluated through the analysis of performance of an algorithm proposed which uses only relative amplitude information for the estimation of spectral masks. It was verified that, for the specific case of the separation of two speech signals, the use of phase information does not affect the algorithm performance. In the next step, the limitations of the spectrum masking technique is analyzed as a function of the reverberation time of the signals that compose the mixture to be separated. When the ideal masks are known, a separation of about 9 *dB* is obtained for the case of an environment with reverberation time less than 300 *ms*. From this point on, as the reverberation time increases, the signals that compose the mixture spread over the time-frequency plane, progressively reducing the separation process performance. Next, given that finding ideal masks based on a single mixture is still an open problem, the performance loss of the separation process is measured as a function of the distance between the ideal mask and the mask effectively used. The results found show a performance loss of 3 *dB* when approximately 10% of the bits of the ideal mask are inverted. Finally, preliminary analyses are carried out to find the ideal mask based on negentropy, kurtosis and on the energy of the separated signals.

Sumário

Resumo	i
Abstract	ii
Lista de Abreviaturas e Siglas	xviii
Lista de Símbolos	xx
1 Introdução	1
1.1 Objetivo Gerais	1
1.2 Formulação do problema de <i>ICA</i>	2
1.3 Objetivos Específicos	3
1.4 Aplicações	4
1.5 Organização do trabalho	5
2 Algoritmos de separação cega de sinais	7
2.1 Mistura instantânea ou convoluída	7
2.1.1 Misturas sobre-determinadas, sub-determinadas ou com a mesma quantidade de fontes e sensores	8
2.2 Modelos de separação	8
2.2.1 Sistemas de fase não-mínima	9
2.3 Solução no domínio do tempo ou da frequência	10
2.4 Breve histórico dos algoritmos <i>BSS/ICA</i> para misturas instantâneas . . .	11
2.5 História dos algoritmos <i>BSS/ICA</i> para misturas convoluídas	13
2.5.1 Métodos baseados em estatística de ordem superior (<i>HOS</i>)	14
2.5.2 Métodos baseados em estatística de segunda ordem (<i>SOS</i>)	16
2.5.3 Análise de cena auditiva	18
2.5.4 Fontes esparsas no domínio do tempo-frequência	19
2.6 Algoritmos propostos para separação de fontes usando apenas uma mistura	20
2.6.1 Método de aprendizagem da máscara ideal utilizando <i>HMM</i>	21

2.6.2	Método de obtenção da máscara ideal utilizando uma abordagem de aprendizagem espectral	23
2.6.3	Abordagem probabilística para a separação de fontes utilizando uma única mistura	24
2.6.4	Estimação de uma máscara não-binária para a separação de fontes utilizando uma única mistura	27
2.7	Algoritmos de separação cega de sinais fonte	30
2.8	Considerações finais	32
3	Mascaramento Espectral de Sinais	33
3.1	Mascaramento ideal no domínio tempo-frequência	34
3.2	Fontes esparsas e o Algoritmo <i>DUET</i>	35
3.3	Estimação da amplitude e atraso relativos	37
3.4	Definição do histograma e separação das fontes	38
3.5	Medida de desempenho	39
3.6	Análise de desempenho do algoritmo <i>DUET</i>	40
3.7	Eliminação da informação de fase (atraso relativo)	42
3.8	Análise do tamanho dos quadros utilizados para separação dos sinais via mascaramento espectral	43
3.9	Análise das máscaras ideais	44
3.10	Separação de misturas instantâneas considerando apenas amplitudes diferentes entre as fontes e sensores	47
3.11	Mistura dos sinais fonte considerando amplitudes (ganhos) e atrasos diferentes entre as fontes e sensores	49
3.12	Misturas dos sinais fonte através da simulação de respostas ao impulso do ambiente	50
3.13	Gravações em campo aberto (ambiente com pouca reverberação)	52
3.14	Gravações em ambiente com reverberação	53
3.15	Considerações finais	54
4	Limitações do mascaramento espectral	55
4.1	Análise da máscara binária ideal para separação de fontes no domínio tempo-frequência	55
4.2	Limitações de métodos de mascaramento espectral para sinais reverberados	57
4.3	Gravações em ambientes reais com reverberação	63
4.4	Influência do tempo de reverberação no mascaramento espectral	74
4.5	Testes de percepção	74

4.6	Testes subjetivos para avaliação de desempenho do algoritmo <i>DUET</i> e da versão simplificada proposta	78
4.7	Testes subjetivos para avaliação da inteligibilidade de sinais de voz recuperados através da aplicação da máscara ideal sobre a mistura de dois sinais de voz considerando diversos tempos de reverberação	81
4.8	Algoritmos propostos para separação de fontes usando apenas uma mistura	82
4.9	Medidas de não-gaussianidade	83
4.9.1	Curtose	83
4.9.2	Negentropia	84
4.10	Análise das medidas de curtose e negentropia para quadros de sinais de voz separados e misturados.	86
4.11	Considerações finais	91
5	Conclusão	93
	Bibliografia	96

Lista de Figuras

2.1	Rede <i>feedforward</i> de um sistema 2×2 de misturas convoluídas.	9
2.2	Rede <i>feedback</i> de um sistema 2×2 de misturas convoluídas.	9
2.3	Técnica de mascaramento aplicada à separação de uma fonte utilizando apenas uma mistura (Roweis, 2000).	21
2.4	Exemplos de sugestões clássicas utilizadas para o agrupamento de frequências de uma única fonte baseada na distribuição da energia. A figura à esquerda mostra a existência de uma estrutura com múltiplos harmônicos. A figura do centro exemplifica casos onde as frequências de uma determinada região aumentam ou diminuem sua energia de forma simultânea e abrupta. A figura à direita exemplifica regiões onde a energia de determinadas frequências apresenta uma determinada tendência em relação ao aumento ou diminuição das frequências ao longo do tempo (Roweis, 2000).	22
2.5	O espectrograma à esquerda mostra uma segmentação ótima (obtida considerando que os sinais são conhecidos) de dois sinais de voz, onde os dois locutores estão representados pelas cores cinza e preto. À direita pode-se observar o espectrograma estimado para os dois locutores utilizando o algoritmo proposto por Bach e Jordan (2004).	24
2.6	Funções base obtidas através da análise em componentes independentes (<i>ICA</i>) para locutores masculinos (a) e femininos (b). Estas funções foram obtidas através do algoritmo de <i>ICA</i> proposto por Lee, Jung, Lee, e Lee (2000) para segmentos de voz contendo 64 amostras e através da seleção de 32 funções base a partir de um conjunto de 64 (Lee et al., 2000).	26

2.7	Modelos gerados a partir da mistura de sinais fonte. (a) Uma mistura é gerada a partir de dois sinais de voz com características diferentes. (b) O sinal de uma das fontes é gerado a partir da soma ponderada dos coeficientes (c_{ik}^t) das funções base ($\mathbf{b}_{ik}$). (c) Histograma dos coeficientes obtidos. Geralmente apresentam uma distribuição super-gaussiana. As distribuições são modeladas utilizando uma função de densidade gaussiana da seguinte forma: $p(c_{ik}^t) \propto \exp(- c_{ik}^t ^q)$. Pode-se observar da esquerda para a direita que a distribuição se torna mais super-gaussiana (Jang e Lee, 2002).	27
2.8	Espectros de potência de uma mistura de dois sinais de voz e dos dois sinais que formam a mistura.	29
3.1	Espectrograma de dois sinais de voz.	35
3.2	As três figuras acima correspondem aos dois espectrogramas dos sinais de voz s_1 e s_2 normalizados com energia unitária e ao produto dos mesmos. As figuras de cima para baixo correspondem, respectivamente, aos espectrogramas dos sinais de voz: $ s_1(w, t) $, $ s_2(w, t) $ e $ s_1(w, t)s_2(w, t) $. Pode-se observar na figura de baixo ($ s_1(w, t)s_2(w, t) $) que os valores encontrados satisfazem aproximadamente a medida de WDO .	36
3.3	Diagrama do algoritmo <i>DUET</i> para o caso de misturas com três fontes.	38
3.4	Exemplos de histogramas obtidos utilizando o <i>DUET</i> . Cada pico no histograma corresponde a uma fonte de acordo com os seus respectivos valores de amplitude e atraso relativos. A figura (a) representa uma fonte com valores de $(\alpha_j, \delta_j) = (0, 2; 0, 5)$. A figura (b) representa os picos obtidos de acordo com os parâmetros de amplitude e atraso relativos para uma mistura de seis fontes. A figura (c) representa os picos obtidos de acordo com os parâmetros de amplitude e atraso relativos para uma mistura de dez fontes (Yilmaz e Rickard, 2004a).	39

3.5	Comparação de histogramas para misturas anecóica e com reverberação. Na coluna da esquerda, de cima para baixo, temos histogramas individuais e em conjunto para uma mistura gravada em ambiente anecóico de três fontes colocadas a 0°, 90° e 180° graus de posição em relação aos microfones. O histograma em conjunto contendo todas as fontes da mistura corresponde basicamente a soma dos histogramas individuais. Na coluna da direita, de cima para baixo, temos também histogramas individuais e em conjunto para uma mistura gravada em ambiente com reverberação de três fontes colocadas a 0°, 90° e 180° graus de posição em relação aos microfones. No histograma da mistura anecóica pode-se observar que os picos estão bem definidos (esquerda, centro e direita). Já na mistura com reverberação não é possível identificar os picos devido à sobreposição dos mesmos (Yilmaz e Rickard, 2004a).	41
3.6	Histograma obtido para uma gravação de dois locutores em um ambiente real com reverberação que pode ser encontrada no banco de dados disponibilizado por Te-Won Lee (Yilmaz e Rickard, 2004a; Lee, 1998).	42
3.7	Relação sinal-distorção média ($\overline{SDR}$) obtida após a aplicação de máscaras ideais, para diferentes tamanhos de quadros (com sobreposição de quadros equivalente a 50% do tamanho dos mesmos), sobre dois sinais de voz separados.	43
3.8	Relação sinal-distorção média ($\overline{SDR}$) obtida após a aplicação de máscaras ideais, para diferentes tamanhos de quadros (com sobreposição de quadros equivalente a 50% do tamanho dos mesmos), sobre uma mistura (soma de dois sinais de voz).	44
3.9	Relação sinal-distorção média ($\overline{SDR}$) obtida após a recuperação dos sinais fonte utilizando apenas a informação de amplitude relativa, ao invés de utilizar as informações de amplitude e atraso relativos para a formação de clusters, como acontece no algoritmo <i>DUET</i> (Yilmaz e Rickard, 2004a), para diferentes tamanhos de quadros sobre uma mistura instantânea de dois sinais de voz.	45
3.10	Imagens contendo as máscaras ideais binárias para as fontes s_1 e s_2 para os 100 primeiros quadros dos sinais de voz analisados.	45
3.11	Histograma dos valores de atenuação relativa dos sinais fonte para misturas instantâneas considerando apenas valores diferentes de atenuações entre as fontes e sensores.	46

3.12	Medidas de distorção (<i>SDR</i>) do sinal recuperado para as fontes s_1 e s_2 considerando vários intervalos a partir do ponto crítico (-0,46) entre os dois picos utilizando passos de 0,01 entre os intervalos. (a) Medidas de distorção obtidas para a fonte s_1 a partir da redução dos intervalos considerados. (b) Medidas de distorção obtidas para a fonte s_2 a partir da redução dos intervalos considerados.	47
3.13	Histograma dos valores de atenuação relativa dos sinais fonte para misturas instantâneas considerando apenas valores diferentes de atenuações entre as fontes e sensores.	48
3.14	Histograma dos valores de atenuação relativa dos sinais fonte para misturas considerando atenuações e atrasos diferentes entre as fontes e sensores. . .	50
3.15	Respostas ao impulso simuladas através do método de imagens da fonte. A figura h_{11} representa a resposta ao impulso da fonte 1 em relação à posição do microfone 1. A figura h_{12} representa a resposta ao impulso da fonte 2 em relação à posição do microfone 1. A figura h_{21} representa a resposta ao impulso da fonte 1 em relação à posição do microfone 2. A figura h_{22} representa a resposta ao impulso da fonte 2 em relação à posição do microfone 2.	51
3.16	Histograma dos valores de atenuação dos sinais fonte para misturas obtidas através da simulação de respostas ao impulso do ambiente.	51
3.17	Gravações realizadas em campo aberto com pouca reverberação.	52
3.18	Histograma dos valores de atenuação dos sinais fonte para misturas gravadas em campo aberto (sem reverberação).	52
3.19	Gravações realizadas em sala com reverberação.	53
3.20	Histograma dos valores de atenuação dos sinais fonte para misturas gravadas em um ambiente com reverberação.	53
4.1	Histograma referente a quantidade de valores iguais a um para cada quadro pertencente a máscara ideal de dois sinais de voz (obtida para dois sinais de voz, utilizando um quadro com 512 amostras e frequência de amostragem igual a 10000 Hz).	56
4.2	Valores médios das relações sinal-distorção do sinal fonte obtidos após introdução gradativa e aleatória de erros na máscara ideal aplicada sobre uma mistura de dois sinais de voz (somados artificialmente). Para obtenção das máscaras ideais utilizou-se quadros com 512 amostras, desta forma a quantidade de erro máxima corresponde a 256. Pode-se notar que à medida que aumenta-se a quantidade de erros a relação sinal-distorção diminui.	56

4.3	Respostas ao impulso simuladas de ambiente com volumes diferentes, porém com mesmo tempo de reverberação ($T_{60} = 600 \text{ ms}$). (a) Resposta ao impulso simulada de um ambiente com as seguintes dimensões $12,5 \text{ m} \times 7,5 \text{ m} \times 5,0 \text{ m}$) e com alta absorção das paredes (coeficiente de absorção igual a 0,4). (b) Resposta ao impulso simulada de um ambiente com as seguintes dimensões $3,0 \text{ m} \times 2,0 \text{ m} \times 2,0 \text{ m}$) e com baixa absorção das paredes (coeficiente de absorção igual a 0,1).	58
4.4	Ambiente considerado para obtenção das respostas ao impulso simuladas (Dimensões: $6,5 \text{ m} \times 3,25 \text{ m} \times 2,5 \text{ m}$).	58
4.5	Respostas ao impulso simuladas considerando o ambiente anecóico. (a) Respostas ao impulso simuladas. (b) Espectros das respostas ao impulso simuladas.	59
4.6	Respostas ao impulso simuladas considerando reflexão em apenas uma parede do ambiente com coeficiente de absorção igual a 0.01. (a) Respostas ao impulso simuladas. (b) Espectros das respostas ao impulso simuladas.	59
4.7	Respostas ao impulso simuladas considerando reflexão em apenas duas paredes do ambiente (Ax1 e Ax2) com coeficiente de absorção igual a 0.01. (a) Respostas ao impulso simuladas. (b) Espectros das respostas ao impulso simuladas.	60
4.8	Respostas ao impulso simuladas considerando reflexão em apenas três paredes do ambiente (Ax1, Ax2 e Ay1) com coeficiente de absorção igual a 0.01. (a) Respostas ao impulso simuladas. (b) Espectros das respostas ao impulso simuladas.	60
4.9	Respostas ao impulso simuladas considerando reflexão em apenas quatro paredes do ambiente (Ax1, Ax2, Ay1 e Ay2) com coeficiente de absorção igual a 0.01. (a) Respostas ao impulso simuladas. (b) Espectros das respostas ao impulso simuladas.	61
4.10	Respostas ao impulso simuladas considerando reflexão em cinco paredes do ambiente (Ax1, Ax2, Ay1, Ay2 e Az1) com coeficiente de absorção igual a 0.01. (a) Respostas ao impulso simuladas. (b) Espectros das respostas ao impulso simuladas.	61
4.11	Respostas ao impulso obtidas considerando reflexão em todas as paredes, com baixo coeficiente de absorção das paredes (cerca de 0.08). (a) Respostas ao impulso simuladas. (b) Espectros das respostas ao impulso simuladas.	62
4.12	Relação sinal-distorção dos sinais recuperados ao aplicarmos uma máscara binária para tempos de reverberação variados. A qualidade dos sinais recuperados foi avaliada para as seguintes situações: i) aplicação da máscara ideal direto sobre os sinais fonte; ii) aplicação da máscara ideal direto sobre os sinais misturados; iii) aplicação da máscara obtida para o caso do ambiente anecóico direto sobre os sinais fonte e iv) aplicação da máscara obtida para o caso do ambiente anecóico direto sobre os sinais misturados.	63

4.13	Pulso estendido no tempo (<i>TSP</i>) utilizado para as medições das respostas ao impulso dos ambientes.	64
4.14	Ambientes onde foram realizadas as gravações dos sinais fonte. (a) Ambiente 1: estúdio vazio com pouca reverberação (as paredes são revestidas de carpete). (b) Ambiente 2: sala vazia com reverberação (paredes revestidas de cimento e tinta).	65
4.15	Figura comparativa contendo as duas respostas ao impulso obtidas para os ambientes (mencionados na Figura 4.14) com tempos de reverberação iguais a 100 <i>ms</i> (Ambiente 1) e 1100 <i>ms</i> (Ambiente 2).	66
4.16	Gravações de dois sinais de voz <i>s1</i> e <i>s2</i> (uma voz masculina e uma voz feminina) realizadas no ambiente 1 com tempo de reverberação igual a 100 <i>ms</i> . De cima para baixo as figuras correspondem, respectivamente: i) mistura dos sinais de voz <i>s1</i> e <i>s2</i> ; ii) sinal <i>s1</i> recuperado ao aplicarmos a máscara ideal sobre a mistura; iii) relação sinal-distorção do sinal <i>s1</i> recuperado em relação ao sinal <i>s1</i> gravado separadamente ($\overline{SDR} = 9$ <i>dB</i>); iv) sinal <i>s2</i> recuperado ao aplicarmos a máscara ideal sobre a mistura; v) relação sinal-distorção do sinal <i>s2</i> recuperado em relação ao sinal <i>s2</i> gravado separadamente ($\overline{SDR} = 8$ <i>dB</i>).	67
4.17	Gravações de dois sinais <i>s1</i> (voz masculina) e <i>s3</i> (música) realizadas no ambiente 1 com tempo de reverberação igual a 100 <i>ms</i> . De cima para baixo as figuras correspondem, respectivamente: i) mistura dos sinais de voz <i>s1</i> e música <i>s3</i> ; ii) sinal <i>s1</i> recuperado ao aplicarmos a máscara ideal sobre a mistura; iii) relação sinal-distorção do sinal <i>s1</i> recuperado em relação ao sinal <i>s1</i> gravado separadamente ($\overline{SDR} = 8$ <i>dB</i>); iv) sinal <i>s3</i> recuperado ao aplicarmos a máscara ideal sobre a mistura ; iv) relação sinal-distorção do sinal <i>s3</i> recuperado em relação ao sinal <i>s3</i> gravado separadamente ($\overline{SDR} = 21$ <i>dB</i>).	68
4.18	Gravações de dois sinais <i>s1</i> (voz masculina) e <i>s4</i> (ruído branco) realizadas no ambiente 1 com tempo de reverberação igual a 100 <i>ms</i> . De cima para baixo as figuras correspondem, respectivamente: i) mistura dos sinais de voz <i>s1</i> e o ruído branco <i>s4</i> ; ii) sinal <i>s1</i> recuperado ao aplicarmos a máscara ideal sobre a mistura; iii) relação sinal-distorção do sinal <i>s1</i> recuperado em relação ao sinal <i>s1</i> gravado separadamente ($\overline{SDR} = 4$ <i>dB</i>); iv) sinal <i>s4</i> recuperado ao aplicarmos a máscara ideal sobre a mistura ; iv) relação sinal-distorção do sinal <i>s3</i> recuperado em relação ao sinal <i>s4</i> gravado separadamente ($\overline{SDR} = 34$ <i>dB</i>).	69

- 4.19 Gravações de dois sinais de voz $s1$ e $s2$ (uma voz masculina e uma voz feminina) realizadas no ambiente 2 com tempo de reverberação igual a 1100 ms . De cima para baixo as figuras correspondem, respectivamente: i) mistura dos sinais de voz $s1$ e $s2$; ii) sinal $s1$ recuperado ao aplicarmos a máscara ideal sobre a mistura; iii) relação sinal-distorção do sinal $s1$ recuperado em relação ao sinal $s1$ gravado separadamente ($\overline{SDR} = 8\text{ dB}$); iv) sinal $s2$ recuperado ao aplicarmos a máscara ideal sobre a mistura; iv) relação sinal-distorção do sinal $s2$ recuperado em relação ao sinal $s2$ gravado separadamente ($\overline{SDR} = 4\text{ dB}$). 70
- 4.20 Gravações de dois sinais $s1$ (voz masculina) e $s3$ (música) realizadas no ambiente 2 com tempo de reverberação igual a 1100 ms . De cima para baixo as figuras correspondem, respectivamente: i) mistura dos sinais de voz $s1$ e música $s3$; ii) sinal $s1$ recuperado ao aplicarmos a máscara ideal sobre a mistura; iii) relação sinal-distorção do sinal $s1$ recuperado em relação ao sinal $s1$ gravado separadamente ($\overline{SDR} = 8\text{ dB}$); iv) sinal $s3$ recuperado ao aplicarmos a máscara ideal sobre a mistura ; iv) relação sinal-distorção do sinal $s3$ recuperado em relação ao sinal $s3$ gravado separadamente ($\overline{SDR} = 11\text{ dB}$). 71
- 4.21 Gravações de dois sinais $s1$ (voz masculina) e $s4$ (ruído branco) realizadas no ambiente 2 com tempo de reverberação igual a 1100 ms . De cima para baixo as figuras correspondem, respectivamente: i) mistura dos sinais de voz $s1$ e o ruído branco $s4$; ii) sinal $s1$ recuperado ao aplicarmos a máscara ideal sobre a mistura; iii) relação sinal-distorção do sinal $s1$ recuperado em relação ao sinal $s1$ gravado separadamente ($\overline{SDR} = 4\text{ dB}$); iv) sinal $s4$ recuperado ao aplicarmos a máscara ideal sobre a mistura ; iv) relação sinal-distorção do sinal $s3$ recuperado em relação ao sinal $s4$ gravado separadamente ($\overline{SDR} = 18\text{ dB}$). 72

4.22	Espectrogramas e curvas de decaimento da energia obtidas para um sinal <i>TSP</i> . O tempo de reverberação (T_{60}) foi especificado de acordo com o tempo de decaimento do nível do sinal de -5 dB até -25 dB . (a) Espectrograma obtido para o caso de simulação de um ambiente anecóico. (b) Curva de decaimento de energia obtido para o caso de simulação de um ambiente anecóico. (c) Espectrograma obtido considerando a simulação de um ambiente com tempo de reverberação igual a 33 ms . (d) Curva de decaimento de energia considerando a simulação de um ambiente com tempo de reverberação igual a 33 ms . (e) Espectrograma obtido considerando a simulação de um ambiente com tempo de reverberação igual a 40 ms . (f) Curva de decaimento de energia considerando a simulação de um ambiente com tempo de reverberação igual a 40 ms . (g) Espectrograma obtido considerando a simulação de um ambiente com tempo de reverberação igual a 182 ms . (h) Curva de decaimento de energia considerando a simulação de um ambiente com tempo de reverberação igual a 182 ms	75
4.23	Espectrogramas e curvas de decaimento da energia obtidas para um sinal <i>TSP</i> . (a) Espectrograma obtido considerando a simulação de um ambiente com tempo de reverberação igual a 290 ms . (b) Curva de decaimento de energia considerando a simulação de um ambiente com tempo de reverberação igual a 290 ms . (c) Espectrograma obtido considerando a simulação de um ambiente com tempo de reverberação igual a 980 ms . (d) Curva de decaimento de energia considerando a simulação de um ambiente com tempo de reverberação igual a 980 ms . (e) Espectrograma obtido considerando a simulação de um ambiente com tempo de reverberação igual a 1750 ms . (f) Curva de decaimento de energia considerando a simulação de um ambiente com tempo de reverberação igual a 1750 ms	76
4.24	Espectrogramas de um sinal <i>TSP</i> obtido através de gravações em ambientes reais. (a) Espectrograma obtido para um ambiente com tempo de reverberação igual a 100 ms . (b) Espectrograma obtido para um ambiente com tempo de reverberação igual a 1100 ms	77
4.25	Resultado do teste subjetivo individual de avaliação de inteligibilidade de sinais de voz separados, a partir de vários tipos de misturas (instantâneas, com atrasos, convoluídas artificialmente, gravações reais em ambiente aberto e gravações em ambiente com reverberação), utilizando o algoritmo <i>DUET</i> e uma versão simplificada proposta baseada somente na informação de amplitude relativa. Nos testes foram utilizados dez avaliadores. A inteligibilidade das frases foi expressa pelo percentual de acertos definido pela razão entre a quantidade de palavras corretamente percebidas e a quantidade de palavras presentes nas frases, conforme mostrado na Tabela 4.3. .	79

4.26	Resultado do teste subjetivo individual de avaliação de inteligibilidade de sinais de voz separados através de uma máscara binária espectral ideal considerando diversos tempos de reverberação. Nos testes foram utilizados dez avaliadores. A inteligibilidade das frases foi expressa pelo percentual de acertos definido pela razão entre a quantidade de palavras corretamente percebidas e a quantidade de palavras presentes nas frases, conforme mostrado na Tabela 4.4. A curva pontilhada apresenta as percentagens médias de acertos dos avaliadores referentes as sentenças reproduzidas de acordo com o tempo de reverberação. As retas cruzando os pontos apresentam o desvio padrão dos resultados obtidos em relação a cada tempo de reverberação analisado.	82
4.27	Função de densidade de probabilidade gaussiana, super-gaussiana e sub-gaussiana. Pode-se verificar que a distribuição gaussiana apresenta valor de curtose igual a zero, a distribuição super-gaussiana apresenta valor igual a 11,28 e a distribuição sub-gaussiana valor igual a -1,25.	84
4.28	Valores de curtose, negentropia e energia obtidos para os quadros de dois sinais de voz separados e para a mistura (soma) dos mesmos. (a) Valores de Curtose. Para a análise dos valores obtidos considerou-se o valor da curtose ao quadrado, afim de evitar valores negativos. Pode-se verificar neste caso, que 45% dos quadros analisados da soma dos sinais apresentam valores de curtose ao quadrado maiores que os valores máximos obtidos entre os quadros dos sinais separados. E observamos também que apenas 10,4% dos quadros analisados da soma dos sinais apresentam valores de curtose ao quadrado menores que os valores de mínimo encontrados entre os quadros dos sinais separados. (b) Valores de Negentropia. Ao verificar os resultados percebe-se que cerca de 85% dos quadros analisados da soma dos sinais apresentam valores de negentropia menores que os valores de mínimo encontrados entre os quadros dos sinais separados. Observa-se também que aproximadamente 3,8% dos quadros analisados da soma dos sinais apresentam valores de negentropia maiores que os valores máximos obtidos entre os quadros dos sinais separados. (c) Valores de Energia. Como era de se esperar 100% dos quadros analisados da soma dos sinais apresentam valores de energia maiores que os valores de máximo encontrados entre os quadros dos sinais separados.	87

4.29	As curvas tracejadas correspondem aos menores valores do quadrado da curtose e da negentropia, e aos maiores valores de energia entre os quadros dos dois sinais de voz separados. As curvas contínuas correspondem aos valores do quadrado da curtose, negentropia e energia para a mistura (soma) dos dois sinais. (a) Valores relativos ao quadrado da curtose. O coeficiente de correlação linear entre a curva contendo os valores mínimos de curtose ao quadrado relativos aos quadros dos sinais separados e a curva contendo o valores de curtose ao quadrado dos quadros da mistura dos sinais é de 0,22. (b) Valores de negentropia. O coeficiente de correlação linear entre a curva contendo os valores mínimos de negentropia relativos aos quadros dos sinais separados e a curva contendo os valores de negentropia dos quadros da mistura dos sinais é de 0,91. (c) Valores de energia. O coeficiente de correlação linear entre a curva contendo o valores máximos de energia relativos aos quadros dos sinais separados e a curva contendo os valores de energia dos quadros da mistura dos sinais é de 0,95.	88
4.30	Diferença entre o valor mínimo da negentropia de cada quadro entre os sinais separados e o valor da negentropia de cada quadro da mistura (soma) dos dois sinais.	90

Lista de Tabelas

2.1	Algoritmos de separação cega de sinais.	31
3.1	Medidas de distorção dos sinais recuperados para misturas instantâneas obtidas através da matriz mostrada na Equação 3.13.	47
3.2	Medidas de distorção dos sinais recuperados para misturas instantâneas obtidas através da matriz mostrada na Equação 3.15.	48
3.3	Medidas de distorção obtidas para misturas com ganhos e atrasos diferentes entre as fontes e sensores.	50
3.4	Medidas de distorção obtidas para misturas dos sinais fonte através da simulação de respostas ao impulso do ambiente.	51
3.5	Medidas de distorção obtidas em um ambiente real com pouca reverberação.	52
3.6	Medidas de distorção obtidas em um ambiente com reverberação.	54
4.1	Variação do tempo de reverberação a medida que se aumenta a quantidade de paredes com reflexão no ambiente simulado.	62
4.2	Aplicação da máscara ideal sobre uma mistura de sinais gravados em ambientes com tempos de reverberação iguais a 100 e 1100 <i>ms</i>	73
4.3	Sequência de frases reproduzidas durante o primeiro experimento para avaliação de inteligibilidade de sinais de voz recuperados utilizando técnicas de mascaramento espectral.	80
4.4	Sequência de frases reproduzidas durante o segundo experimento para avaliação de inteligibilidade de sinais de voz recuperados utilizando uma máscara binária ideal a partir de misturas convoluídas com diferentes tempos de reverberação.	83
4.5	Percentual da relação sinal-distorção referente aos quadros dos sinais de voz estimados em comparação com cada quadro do sinal original.	90

Lista de Abreviaturas e Siglas

ASA	Análise de cena auditiva (Auditory Scene Analysis)
BBS	Separação cega de fontes (Blind Source Separation)
CASA	Análise de cena auditiva computacional (Computational Auditory Scene Analysis)
CDMA	Acesso múltiplo por divisão de código (Code Division Multiple Access)
DESPRIT	DUET-ESPRIT
DFT	Transformada discreta de Fourier (Discrete Fourier Transform)
DOA	Medida de direção do sinal (Direction of Arrival)
DUET	Degenerate Unmixing Estimation Technique
EASI	Equivariant Adaptive Separation via Independence
EEG	Eletroencefalograma (Electroencephalogram)
ECG	Eletrocardiograma (Electrocardiogram)
FastICA	Algoritmo rápido de análise em componentes independentes
FFT	Transformada rápida de Fourier (Fast Fourier Transform)
FIR	Resposta ao impulso finita (Finite Impulse Response)
fMRI	Ressonância magnética funcional (Functional Magnetic Resonance Imaging)
FOBI	Identificação cega de quarta ordem (Fourth-Order Blind Identification)
HMM	Modelos Ocultos de Markov (Hidden Markov Models)
HOS	Estatística de ordem superior (High Order Statistics)
ICA	Análise em componentes independentes (Independent Component Analysis)
IID	Diferença de intensidade interaural (Interaural Intensity Difference)
IIR	Resposta ao impulso infinita (Infinite Impulse Response)
ITD	Diferença de tempo interaural (Interaural Time Difference)
INFOMAX	Algoritmo de separação cega de fontes baseado na maximização da informação mútua entre sinais de entrada e de saída
JADE	Joint approximate diagonalisation of eigenmatrices
MAP	Máxima probabilidade a posteriori (Maximum a Posteriori Probability)
MEG	Magneto-eletroencefalograma (Magneto-encephalogram)

MFCC	Coeficientes mel-cepstrais (Mel-Frequency Cepstral Coefficients)
MIMO	Sistema de múltiplas entradas e múltiplas saídas (Multiple Input Multiple Output System)
PDF	Função de densidade de probabilidade (Probability Density Function)
SDR	Relação sinal-distorção (Source to Distortion Ratio)
SOBI	Identificação cega de segunda ordem (Second Order Blind Identification)
SOS	Estatística de segunda ordem (Second Order Statistics)
TIMIT	Banco de dados de voz (TIMIT Speech Database)
WDO	Medida de esparsidade dos sinais fonte (W-disjoint Orthogonality)

Lista de Símbolos

s	Vetor contendo sinal fonte
x	Vetor contendo mistura de fontes
y	Vetor contendo sinal separado
f(.)	Função de mistura
g(.)	Função de separação
A	Matriz de mistura
W	Matriz de separação
h	Resposta ao impulso do ambiente em relação a uma fonte e sensor
n	Vetor contendo sinal de ruído
λ_p	Corresponde ao ganho da fonte p em relação ao microfone
B	Matriz de funções base
O	Matriz inversa da matriz de funções base ($O = B^{-1}$)
c	Vetor contendo os coeficientes das funções base
μ	Média
σ	Desvio Padrão
θ	Conjunto de parâmetros estimados para uma distribuição não-gaussiana simétrica
M	Matriz contendo máscaras ideais binárias
α_j	Parâmetro correspondente à amplitude relativa da fonte j em relação à fonte considerada como referência
δ_j	Parâmetro correspondente ao atraso relativo da fonte j em relação à fonte considerada como referência
A_x	Corresponde à distância entre as paredes laterais do ambiente simulado referentes ao comprimento do mesmo
A_y	Corresponde à distância entre as paredes frontais do ambiente simulado referentes à largura do mesmo
A_z	Corresponde à distância entre as paredes do ambiente simulado referentes ao teto e piso do mesmo

$Curtose(y)$	Medida de curtose
$Entropia(y)$	Medida de entropia
$J(y)$	Medida de negentropia
G_i	Função não-quadrática

Capítulo 1

Introdução

1.1 Objetivo Gerais

O objetivo deste trabalho é analisar e verificar as limitações da técnica de mascaramento no domínio do tempo e frequência para o problema de separação cega de fontes (*Blind Source Separation*) utilizando misturas convoluídas.

A motivação está relacionada ao fato de existirem muitas situações reais onde fontes independentes estão misturadas e deseja-se separar ou estimar estes sinais sem conhecimento das características do canal de transmissão.

Algumas técnicas de separação cega de sinais baseadas no mascaramento espectral se baseiam na esparsidade dos sinais fonte. Os sinais de voz podem ser considerados esparsos, ou seja, em um determinado instante de tempo e frequência somente uma das fontes estará ativa. Aproveitando-se desta consideração pode-se separar os sinais fonte através da aplicação de uma máscara ideal binária no domínio do tempo e frequência. O algoritmo *DUET* (*Degenerate Unmixing Estimation Technique*) visa a estimação de máscaras ideais binárias para a separação cega dos sinais fonte. Em um primeiro momento, neste trabalho, foi verificado o desempenho do algoritmo *DUET* para os seguintes tipos de misturas: instantâneas, com atrasos e convoluídas. Na etapa seguinte foi avaliado as limitações da técnica de mascaramento espectral em função do aumento do tempo de reverberação. Finalmente, dado que o problema de encontrar máscaras ideais com base em apenas uma mistura ainda é um problema em aberto, avaliou-se o desempenho do processo de separação em função da distância entre a máscara ideal e a máscara efetivamente usada. Análises preliminares para encontrar a máscara ideal com base em medidas de negentropia, curtose e energia também foram avaliadas.

1.2 Formulação do problema de *ICA*

Separação cega de sinais (*BSS*) é uma área de estudo onde as fontes $s_i(k)$ devem ser recuperadas a partir dos sinais observados $x_i(k)$. Uma consideração comum para resolver este problema é considerar que os sinais são inerentemente independentes. Neste caso, as soluções encontradas pertencem à área de análise em componentes independentes (*ICA*).

Uma importante aplicação da separação cega de sinais é a separação de sinais de voz. As primeiras demonstrações de separação cega de fontes consideravam as misturas como sendo instantâneas, não considerando a presença de atrasos e reverberação. Um exemplo clássico é o problema denominado *cocktail party*, assim cunhado por Cherry e Collin (1953). Neste problema, várias pessoas podem estar conversando em um determinado ambiente e, ao captarmos o sinal de áudio em pontos específicos, obtemos sinais constituídos por misturas de diversos sinais de voz. O objetivo da separação cega de fontes (*BSS*) é separar os sinais independentes originais (sinais de voz) a partir das misturas observadas.

Os problemas de separação cega de sinais e análise em componentes independentes podem ser definidos seguindo uma notação matemática. Vamos considerar a medição de n sinais $x_1(k), \dots, x_n(k)$ que correspondem a misturas de n fontes independentes $s_1(k), \dots, s_n(k)$. O objetivo consiste em obter n sinais de saída $y_1(k), \dots, y_n(k)$ a partir dos sinais observados, de maneira que as fontes sejam recuperadas. Considere $\mathbf{x}(k) = [x_1(k), x_2(k), \dots, x_n(k)]$ e assumamos similarmente o mesmo para $\mathbf{s}(k)$ e $\mathbf{y}(k)$. Assumamos também que a função de mistura seja dada por $\mathbf{f}(\cdot)$ e a função de separação seja dada por $\mathbf{g}(\cdot)$. Desta forma, podemos escrever:

$$\mathbf{x}(k) = \mathbf{f}(\mathbf{s}(k)), \quad (1.1)$$

$$\mathbf{y}(k) = \mathbf{g}(\mathbf{x}(k)). \quad (1.2)$$

A resolução do problema consiste em estimarmos uma função $g(\cdot)$ a partir de $\mathbf{x}(k)$, sem que $f(\cdot)$ e $\mathbf{s}(k)$ sejam conhecidos. Uma solução seria considerar $g(\cdot) = f^{-1}(\cdot)$. Para o caso da matriz de mistura ser linear, invariante no tempo e instantânea, $\mathbf{x}(k) = \mathbf{A}\mathbf{s}(k)$, onde $\mathbf{A}$ é uma matriz constante $n \times n$. Se a matriz $\mathbf{A}$ é inversível temos que $\mathbf{W} = \mathbf{A}^{-1}$, onde $\mathbf{W}$ é a matriz de separação. Os sinais de saída são obtidos através de $\mathbf{y}(k) = \mathbf{W}\mathbf{x}(k)$. Entretanto, a escala dos sinais e a ordem em que os sinais recuperados aparecem não é importante, uma vez que não há referências para comparação. Com isto, $\mathbf{W}$ não precisa ser igual a $\mathbf{A}^{-1}$, mas deve ser equivalente sob o ponto de vista da separação de informações (Carvalho, 2000).

Para misturas instantâneas, sem a presença de ruído, existe um sistema inversível capaz de recuperar as fontes de forma quase perfeita, a menos de uma permutação da ordem

e de um fator de escala (Vincent, Gribonval, e F'evotte, 2006). Mas, no caso de misturas convoluídas, o problema se torna bem mais difícil. Para obter a separação cega de misturas convoluídas, vários métodos foram propostos por Bell e Sejnowski (1995); Araki, Makino, Sawada, e Mukai (2004); Torkkola (1996), entre outros. Os algoritmos para os problemas de separação de fontes convoluídas são encontrados no domínio do tempo ou da frequência. Algumas abordagens no domínio do tempo consideram os sistemas de mistura como sendo filtros *FIR* (*finite impulse response*) e o algoritmo de separação cega tem como objetivo a estimação de filtros inversos capazes de separar os sinais fonte. No domínio da frequência, os problemas envolvendo misturas convoluídas tornam-se mais simples permitindo a aplicação de algoritmos utilizados em *ICA* para misturas instantâneas. Embora existam sistemas de separação cega de sinais de voz gravados em ambientes acústicos reais, pode-se verificar que o desempenho destes algoritmos ainda tem muito para ser melhorado (Araki et al., 2004).

1.3 Objetivos Específicos

Imagine o efeito do som produzido por vários pianos tocados simultaneamente. Se cada pianista estiver tocando de forma aleatória é muito difícil identificar qual nota vem de cada piano. Mas, caso eles estejam tocando uma melodia, o processo de separação torna-se mais fácil devido à estrutura harmônica da música. Qualquer sistema de audição biológico ou artificial deve ser capaz de extrair objetos acústicos individuais com o objetivo de tornar possível a localização, remoção do ruído e reconhecimento do som. Bregman (1990) chamou este processo de análise de cena auditiva (*ASA*). Por sua vez, o processo de separação de fontes de forma computacional através da análise de sinais gravados por microfones é chamado de análise de cena auditiva computacional (*CASA*). Sistemas baseados na análise de cena auditiva computacional exploram o processo de mascaramento que ocorre na audição humana, onde um som de intensidade menor é mascarado por um outro de intensidade maior dentro de uma mesma banda crítica. Uma classe de algoritmos para a separação de sinais de voz utiliza uma máscara no domínio do tempo-frequência. Esta técnica se baseia na esparsidade dos sinais de voz e considera que, em uma mistura de diferentes fontes, é improvável encontrar sinais ativos em um mesmo instante de tempo e frequência. Desta forma, uma fonte sonora específica pode ser recuperada aplicando uma máscara binária sobre uma mistura acústica a cada instante de tempo e frequência. As regiões onde uma determinada fonte está ativa, ou seja, apresenta energia maior que as demais fontes, recebem peso igual a 1 e as demais regiões recebem peso igual a zero. Os sinais separados por estas máscaras binárias apresentam boa qualidade sonora e inteligibilidade.

Yilmaz e Rickard (2004a) propuseram um sistema de estimação cega de máscaras no domínio do tempo-frequência para a separação de sinais de voz denominado *DUET*, o qual é similar em alguns aspectos ao sistema proposto por Roman, Wang, e Brown (2003). O *DUET* foi proposto considerando um ambiente anecóico, contendo somente atrasos. No caso de ambientes com reverberação, o desempenho do algoritmo é bem inferior ao obtido para misturas contendo somente atrasos. O *DUET* foi desenvolvido para trabalhar em casos sub-determinados, onde existam mais fontes do que sensores. As máscaras binárias são estimadas a partir de duas misturas e através de dois parâmetros denominados amplitude relativa e atraso relativo.

Neste trabalho é verificado o desempenho do algoritmo *DUET* para misturas convoluídas, assim como suas limitações. A utilização da informação de fase (atraso relativo) para a separação de fontes é avaliada através da análise de desempenho de uma variação do *DUET* que considera apenas a informação de amplitude relativa para a estimação das máscaras ideais.

Uma das vantagens dos métodos de separação de sinais fonte via mascaramento espectral está relacionada com a capacidade destes algoritmos atuarem em casos sub-determinados, ou seja, situações onde há mais fontes que sensores. Uma situação ainda mais extrema ocorre quando há apenas uma mistura e duas ou mais fontes. Esta situação é um problema de difícil solução e ainda em aberto para o caso de misturas envolvendo sinais de voz, principalmente para o caso de misturas convoluídas. As limitações das técnicas de mascaramento espectral, a influência do tempo de reverberação na separação de sinais fonte e as dificuldades encontradas para a separação de fontes utilizando apenas uma mistura são analisadas e discutidas.

1.4 Aplicações

A técnica de separação cega de sinais encontra aplicações em diversas áreas, tais como processamento de sinais de voz, deficiência auditiva, processamento de sinais médicos, comunicações digitais, sensoriamento remoto, e análise de séries temporais. A seguir são descritas algumas destas áreas e aplicações:

- Processamento de sinais de voz: o desempenho de um sistema de reconhecimento de fala pode ser fortemente afetado pela presença de ruídos e interferência de outros sinais de voz. Métodos de separação cega de sinais são aplicados a sistemas de reconhecimento de fala para melhorar a relação sinal-ruído e eliminar a interferência de outros sinais de voz (Sanchez, 2002).
- Deficiência auditiva: pessoas com alguns tipos de deficiência auditiva têm dificuldade

de concentrar-se em apenas uma voz em um ambiente sob o efeito do problema denominado *cocktail party*. A separação cega de fontes pode ser aplicada para separar o sinal de voz de uma determinada pessoa. Os aparelhos de audição convencionais são constituídos por um amplificador e um filtro auditivo que não são muito eficazes em um ambiente ruidoso, pois o dispositivo amplifica tanto o sinal desejado quanto o ruído de forma indiscriminada. Técnicas de separação cega também são aplicadas para eliminação do ruído (Chazan, Medan, e Shvadron, 1988).

- Processamento de sinais médicos: a separação cega de fontes possui grande aplicação na área médica, principalmente no tratamento de sinais provenientes de eletroencefalograma (*EEG*) (Makeig, Bell, Jung, e Sejnowski, 1998), eletrocardiograma (*ECG*) (Barros, Mansour, e Ohnishi, 1998), magneto-eletroencefalograma (*MEG*) (Vigario, Sarela, Jousmaki, e Oja, 1998) e processamento de imagens de ressonância magnética funcional (*fMRI*) (McKeown, Makeig, Brown, Jung, Kindermann, Bell, Iragui, e Sejnowski, 1998). Como exemplo de aplicação podemos mencionar o uso de algoritmos *BSS/ICA* em sinais de eletroencefalograma (*EEG*). Os sinais *EEG* são corrompidos por artefatos produzidos pelos eletrodos, músculos ou até mesmo pela respiração. A utilização de simples filtros não consegue eliminar estes artefatos, porém algoritmos de *BSS/ICA* são capazes de eliminar a interferência.
- Comunicações digitais: as técnicas de *BSS/ICA* são aplicadas, por exemplo, na separação de sinais de rádio de canais devanecentes (Torkkola, 1998) e na separação de misturas convoluídas por sistemas *CDMA* (*Code Division Multiple Access*) (Cristescu, Ristaniemi, Joutsensalo, e Karhunen, 2000).
- Sensoriamento Remoto: o sensoriamento remoto é a ciência que visa adquirir informações a respeito de materiais, áreas ou fenômenos, sem contato físico. No sensoriamento remoto as medições dão origem a imagens digitais para o propósito de interpretação. Os métodos de *BSS/ICA* são utilizados para o tratamento e extração de informações destas imagens digitais (Fiori, 2002).
- Análise de séries temporais: a aplicação de *BSS/ICA* em séries temporais inclui a análise de séries financeiras (Kiviluoto e Oja, 1998) e de séries temporais geofísicas (Aires, Chedin, e Nadal, 2000; Sanchez, 2002).

1.5 Organização do trabalho

O conteúdo dos capítulos seguintes está distribuído da seguinte forma:

- **Capítulo 2:** são apresentadas diferentes teorias e soluções para a separação cega de sinais. Uma revisão cronológica dos diferentes algoritmos de separação cega de fontes existentes na literatura é apresentada.
- **Capítulo 3:** é apresentado um estudo sobre a utilização do mascaramento no domínio tempo-frequência para a separação de sinais de voz. O desempenho do algoritmo *DUET* é avaliado para o caso de misturas de sinais de voz reverberados. É proposto um método para separação de sinais baseado no algoritmo *DUET* e na utilização apenas da informação de amplitude relativa para estimação das máscaras ideais via mascaramento no domínio do tempo e frequência.
- **Capítulo 4:** partindo do pressuposto que as máscaras ideais sejam conhecidas, uma série de análises foram realizadas para verificar as limitações das técnicas de mascaramento espectral, a influência do tempo de reverberação na separação de sinais fonte via mascaramento espectral e as dificuldades encontradas para a separação de fontes utilizando apenas uma mistura são analisadas.
- **Capítulo 5:** neste capítulo são apresentadas as conclusões do estudo realizado neste trabalho.

Capítulo 2

Algoritmos de separação cega de sinais

Os modelos utilizados para representar misturas de sinais fazem uso de aproximações e simplificações de acordo com a natureza e propriedades do processo através do qual as misturas são produzidas. Este capítulo tem como objetivo apresentar diversas abordagens e algoritmos existentes na literatura referentes à área de separação cega de fontes, sugere-se ao leitor que estiver familiarizado com o assunto, ir direto para a leitura do capítulo seguinte.

Existem várias abordagens para o problema de separação cega de sinais de acordo com os seguintes tópicos:

- modelos de misturas instantâneas ou convoluídas;
- métodos de separação;
- solução no domínio do tempo ou da frequência.

2.1 Mistura instantânea ou convoluída

Um modelo muito comum adotado por vários pesquisadores para representar o fenômeno físico de mistura de fontes consiste em considerar o processo de mistura como sendo linear, invariante no tempo e instantâneo. O ruído é considerado aditivo e independente em cada sensor. Os sinais observados são considerados como sendo uma soma escalar das fontes somadas ao ruído captado pelos sensores:

$$x_i(k) = \sum_{j=1}^m h_{ij}s_j(k) + n_i(k), \quad (2.1)$$

onde os termos h_{ij} representam as respostas ao impulso do ambiente em relação a cada fonte e microfone, as seqüências s_j representam as fontes de sinais amostrados e as seqüências n_i representam as fontes de ruído amostradas.

O processo de mistura instantâneo não corresponde a muitas situações reais onde pode-se verificar a presença de atrasos e reflexões dos sinais fonte na mistura. Um processo de misturas convoluídas pode ser obtido através da convolução da resposta ao impulso do ambiente em relação a cada uma das fontes e sensores:

$$x_i(k) = \sum_{j=1}^m \sum_{l=0}^p h_{ij}(l) s_j(k-l) + n_i(k). \quad (2.2)$$

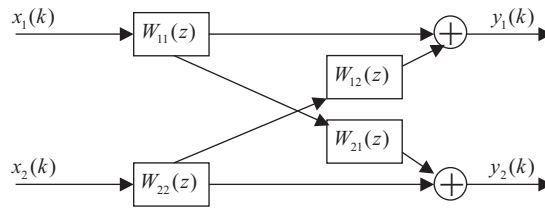
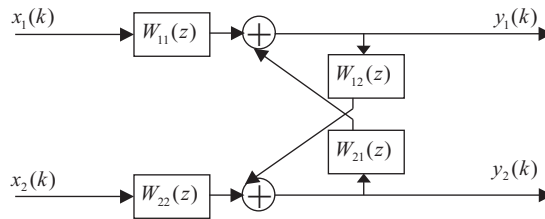
2.1.1 Misturas sobre-determinadas, sub-determinadas ou com a mesma quantidade de fontes e sensores

Uma questão importante nos sistemas de separação refere-se à relação entre a quantidade de fontes e de sensores existentes. Normalmente considera-se o número de sensores igual ou maior que o número de fontes. Quando o número de fontes é igual ao número de sensores, a mistura é denominada quadrada. Os casos onde o número de fontes é menor ou maior que o número de sensores corresponde, respectivamente, aos casos em que o sistema é sub ou sobre-determinado. No caso de sistemas sub-determinados, métodos lineares podem não ser capazes de recuperar as fontes.

2.2 Modelos de separação

Os parâmetros utilizados para a separação de fontes são estimados com base nos modelos utilizados para a determinação do sistema inverso. A maioria dos algoritmos de separação cega de sinais considera o modelo de mistura como tendo uma resposta ao impulso finita (*FIR*). Os sistemas de separação mais comuns incluem estruturas de redes do tipo *feedforward* ou *feedback*. Várias funções de contraste (como função de correlação e funções não-lineares) são utilizadas para satisfazer um critério de separação. A partir disto, os sinais são reconstruídos (Hyvarinen e Oja, 2000). O principal objetivo consiste na separação dos sinais e não em uma perfeita recuperação dos mesmos. Embora a reconstrução perfeita do sinal seja desejável, ela não é o foco principal. Por isto, em alguns casos pode não ser necessário identificar o sistema inverso que consiga recuperar perfeitamente as fontes. Ao invés disto, é suficiente estimar-se filtros de separação W que removam a interferência entre as fontes introduzida pelo processo de mistura. Estes filtros de separação podem ter uma arquitetura do tipo *feedback*, onde normalmente utilizam-se filtros *IIR* (reposta ao impulso infinita) ou *feedforward* onde geralmente utilizam filtros *FIR* (reposta ao impulso finita). Estas arquiteturas estão representadas nas figuras 2.1 e 2.2 para um sistema 2×2 (duas fontes e dois sensores).

As redes *feedforward* e *feedback* também podem ser combinadas para formar uma rede

Figura 2.1: Rede *feedforward* de um sistema 2×2 de misturas convoluídas.Figura 2.2: Rede *feedback* de um sistema 2×2 de misturas convoluídas.

híbrida, onde uma topologia *feedforward* é seguida por uma arquitetura do tipo *feedback* (Choi e Cichocki, 1997, 1999).

As redes do tipo *feedforward* apresentam como vantagem o fato de permitirem a criação de um sistema inverso mais generalizado para a separação dos sinais de tal forma a obter uma solução de *ICA* mais aproximada para um sistema de fase não-mínima (Choi, Cichocki, Park, e Lee, 2005). Um filtro de fase não-mínima pode ser expresso como produto de um filtro de fase mínima com um filtro passa-tudo (*all-pass*). A inversa de um filtro de fase não-mínima é então dada pelo produto da inversa do filtro de fase mínima, que é estável e causal, pela inversa do filtro passa-tudo que é avançada no tempo.

2.2.1 Sistemas de fase não-mínima

A estabilidade e causalidade de um sistema inverso depende da localização de seus pólos e zeros. Para um sistema de fase mínima todos os seus zeros e pólos estão dentro do círculo unitário. Um sistema de fase mínima apresenta um sistema inverso causal e estável. Um sistema de fase máxima possui todos os seus zeros fora do círculo unitário, sendo que um sistema de fase não-mínima apresenta somente alguns zeros fora do círculo (Hashemi e Hammond, 1996). Para um sistema de fase não-mínima existem dois tipos de sistemas inversos: i) estável e não-causal e ii) causal e instável (Lee, Goudeseune, e Hasegawa-Johnson, 2003).

O ideal seria obtermos um sistema inverso que fosse causal e estável. Sistemas que são tanto estáveis quanto causais devem possuir todos os seus pólos dentro do círculo unitário

(i.e. fase mínima). Um pólo que está dentro do círculo unitário no plano z contribui com um termo exponencialmente decrescente lateral direito para a resposta ao impulso. Não podemos ter um pólo fora do círculo unitário, uma vez que a transformada inversa de um pólo localizado fora do círculo contribuiria com um termo exponencial crescente lateral direito, o qual não é estável ou com um termo exponencial decrescente lateral esquerdo, o qual não é causal (Haykin e Veen, 1999).

De acordo com as medições realizadas por Neely e Allen (1979), pode-se constatar que a resposta ao impulso de um ambiente é geralmente de fase não-mínima. A resposta ao impulso para um ambiente pequeno (com as dimensões aproximadamente iguais a $4,00 \times 3,35 \times 2,10$ metros) constitui um sistema de fase mínima somente se os coeficientes de reflexão assumem valores abaixo de aproximadamente 0,37 (sendo que coeficientes típicos de reflexão nas paredes de um ambiente estão entre 0,7 e 0,9). Variações nas posições da fonte e microfone podem provocar pequenas alterações (para mais ou menos) destes coeficientes. Neely e Allen (1979) observaram que, para um ambiente com tamanho específico e posições fixas de microfones e fontes, existe um limiar para os coeficientes de reflexão. Quando os coeficientes de reflexão (das seis paredes) estão abaixo deste limiar, a resposta ao impulso apresenta um comportamento de um sistema de fase mínima. Mas, se os coeficientes estão acima deste limiar, o comportamento da resposta ao impulso corresponde à de um sistema de fase não-mínima. Este limiar de reflexão não é constante para um determinado ambiente, sofrendo uma diminuição nos casos onde (Neely e Allen, 1979): (i) as paredes mais próximas estão equidistantes da fonte e microfone; (ii) aumenta-se a distância entre a fonte e microfone. No domínio do tempo, por inspeção visual da resposta ao impulso do ambiente, pode se determinar se ela é de fase mínima. Normalmente para um sistema de fase mínima a primeira amostra (depois do atraso inicial) é muito maior que as demais amostras e a resposta deve diminuir rapidamente (Neely e Allen, 1979).

Quando um sinal de voz é filtrado pela resposta ao impulso de um sistema de fase mínima e em seguida por sua inversa, podemos verificar que o sinal obtido é idêntico ao original. Entretanto, quando um sinal de voz é filtrado pela resposta ao impulso de um sistema de fase não-mínima e filtrado novamente por um sistema inverso de fase mínima (que não corresponde exatamente ao sistema inverso), o sinal obtido apresentará distorções em relação ao sinal original (Neely e Allen, 1979).

2.3 Solução no domínio do tempo ou da frequência

A solução para o problema de separação pode ser desenvolvida no domínio do tempo ou da frequência. No domínio do tempo, os coeficientes da matriz de mistura representam filtros que serão convoluídos com as fontes (Sanchez, 2002). O processo de misturas

convoluídas pode ser simplificado transformando as misturas para o domínio da frequência. A convolução linear no domínio do tempo pode ser escrita no domínio da frequência como uma multiplicação:

$$X(w) = H(w)S(w) + N(w). \quad (2.3)$$

Para cada componente de frequência, $w = 2\pi f$, $H(w)$ é uma matriz complexa de ordem $m \times n$, $X(w)$ e $N(w)$ são vetores complexos de ordem $n \times 1$, e similarmente $S(w)$ é um vetor complexo de ordem $n \times 1$. A transformação para o domínio da frequência é obtida utilizando a transformada discreta de Fourier (*DFT*). Utilizando a transformada rápida de Fourier (*FFT*) a convolução pode ser implementada de forma eficiente no domínio da frequência.

Uma das vantagens da separação cega de sinais no domínio da frequência para o caso de misturas convoluídas é o fato do problema poder ser reduzido a uma mistura instantânea para cada componente de frequência. Embora este fato simplifique a tarefa de separação de misturas convoluídas, deve-se destacar também algumas desvantagens. Os sinais no domínio da frequência obtidos a partir da *DFT* apresentam valores complexos. Nem todos os algoritmos de separação instantânea são desenvolvidos para trabalhar com sinais contendo valores complexos (Pedersen, Larsen, Kjems, e Parra, 2007). No domínio da frequência pode-se não ter dados suficientes disponíveis para avaliar-se a independência estatística. Para alguns algoritmos é necessário que o tamanho do quadro utilizado na obtenção da transformada discreta de Fourier seja muito maior que o comprimento da resposta ao impulso do ambiente (Parra e Spence, 2000). Ao utilizar quadros deste tamanho obtém-se poucas amostras por frequência interferindo nas considerações sobre a independência das fontes (Araki, Mukai, Makino, Nishikawa, e Saruwatari, 2003b). Outro problema que ocorre quando trabalhamos no domínio da frequência é o da permutação e escala. Como os problemas de misturas convoluídas são tratados para cada frequência como sendo um problema separado, os sinais fonte de cada frequência podem ser estimados de forma arbitrária, tanto em termos de permutação das componentes de frequência quanto de escala.

2.4 Breve histórico dos algoritmos *BSS/ICA* para misturas instantâneas

Não consegue-se definir, com precisão, o início exato das pesquisas envolvendo separação cega de fontes (*BSS*) e análise em componentes independentes (*ICA*) (Douglas, 2003; Fiori, 2002). Ambos os problemas possuem fortes conexões com problemas relacionados à área

de estatística, cujas origens podem ser rastreadas a partir do trabalho desenvolvido por Thurstone (1931). A idéia de separar variáveis aleatórias misturadas linearmente é uma área muito antiga na estatística. As medidas de não-gaussianidade propostas em trabalhos de estatística são muito similares às medidas de curtose e negentropia utilizadas em *ICA*. Darmois (1953) na década de 50 e Rao (1966) na década de 60 publicaram trabalhos sobre caracterização de variáveis aleatórias em misturas lineares (Fiori, 2002). Apesar de nenhum algoritmo ser proposto nestes artigos, pode-se verificar fatos interessantes que foram abordados posteriormente pela teoria de *ICA* como, por exemplo, o fato dos resultados mostrarem que, para um sistema ser completamente identificável, deve existir no máximo uma fonte gaussiana.

Os algoritmos de deconvolução cega de sinais começaram a surgir por volta dos anos 80 com aplicações nas áreas de geofísica e comunicações digitais (Sato, 1975; Godfrey, 1978; Wiggins, 1978; Benveniste, Goursat, e Ruget, 1980). Um dos primeiros trabalhos a ser publicado foi desenvolvido por Benveniste et al. (1980). Embora este trabalho estivesse relacionado com deconvolução cega de fontes, vale ressaltar que o método utilizado para a estimação das equações é aproximadamente o mesmo utilizado nas técnicas de *ICA* com mistura instantânea.

Na comunidade de processamento de sinais, Héroult e Jutten (1986) foram os pioneiros a abordar problemas referentes à análise de componentes independentes e propuseram uma rede neural adaptativa para este fim (Jutten e Héroult, 1991). Esta abordagem foi posteriormente aperfeiçoada por Cichocki e Unbehauen (1996). Dinç e Bar-Ness (1992) foram os dois primeiros a estabelecer procedimentos simples para a extração de informação útil de um sinal a partir de misturas lineares desta informação, sem a necessidade de treinamento de dados. Cardoso (1989) desenvolveu um algoritmo denominado *FOBI* (*Fourth-Order Blind Identification*) que utiliza momentos de quarta ordem para realizar a separação de fontes. Comon (1994) mostrou que a informação mútua era a medida mais natural de independência e, que, para maximizar a medida de não-gaussianidade, bastava minimizar a medida de informação mútua entre os sinais de saída da rede. A formalização e desenvolvimento da teoria básica referente à análise em componentes independentes foi proposta por Comon (1994).

Cardoso e Laheld (1996) desenvolveu um algoritmo de separação baseado em uma expressão recursiva para obtenção da matriz de separação. O algoritmo recebeu o nome de *EASI* (*Equivariant Adaptive Separation via Independence*).

Na publicação de Bell e Sejnowski (1995) é sugerido um método para maximização da informação mútua entre as entradas e as saídas da rede neural proposta. Este algoritmo foi denominado *Infomax*, sendo um dos algoritmos mais difundidos devido à sua simplicidade e velocidade de convergência (Knezevic, 2004). Uma importante melhoria na eficiência do

algoritmo *Infomax* é proposta por Amari, Cichocki, e Yang (1996) que utilizam uma regra de aprendizagem obtida através do método do gradiente natural descendente, sob o ponto de vista da geometria da informação desenvolvida por Amari (1985). O método proposto por Bell e Sejnowski (1995) permitiu o desenvolvimento de uma série de algoritmos em *ICA*. Uma das vantagens do algoritmo é a possibilidade de separar um número grande de fontes. Bell e Sejnowski (1995) em seu trabalho chegaram a separar com sucesso mais de dez fontes (Smaragdis, 1997; Hyvarinen e Oja, 2000).

A idéia de não-gaussianidade das fontes foi explorada por Hyvarinen e Oja (1997) para o desenvolvimento do algoritmo denominado *FastICA*.

Em 1995, Puntonet, Prieto, Jutten, Rodriguez-Alvarez, e Ortega (1995) propuseram um método geométrico para *ICA* baseado nas propriedades dos espaços vetoriais das fontes e misturas. O método apresenta como principal vantagem a simplicidade de implementação.

Belouchrani, Meraim, Cardoso, e Moulines (1997) apresentam um algoritmo chamado de *SOBI* (*Second Order Blind Identification*) baseado em estatística de segunda ordem. O método apresenta como principal vantagem uma grande imunidade à presença de ruído (Choi et al., 2005).

Delfosse e Loubaton (1995) propõem um método denominado "*deflation*" baseado em estatística de ordem superior. O algoritmo proposto extrai as fontes componente por componente. O método foi posteriormente aperfeiçoado por Thawonmas, Cichocki, e Amari (1998) que utilizaram redes neurais.

A utilização de redes neurais para a obtenção da matriz de separação e das fontes foi proposta inicialmente por Héroult e Jutten (1986). Posteriormente, uma série de trabalhos foi desenvolvida usando a teoria de redes neurais (Yang e Amari, 1997; Girolami e Fyfe, 1997; Karhunen, Hyvarinen, Vigario, Hurri, e Oja, 1997; Hyvarinen e Oja, 1998; Fiori e Celledoni, 2004).

Os primeiros algoritmos de *BSS* e *ICA* propostos somente apresentavam soluções para os casos onde o número de fontes era igual ao de sensores e para casos sobre-determinados. A primeira solução para o caso sub-determinado foi apresentada por Belouchrani e Cardoso (1994), sendo proposto um algoritmo baseado em uma abordagem probabilística denominado *MAP* (*maximum a posteriori probability*).

2.5 História dos algoritmos *BSS/ICA* para misturas convoluídas

Os métodos de *BSS* e *ICA* utilizados em misturas instantâneas não funcionam quando aplicados a misturas contendo atrasos e reflexões das fontes. Para separação de misturas

convoluídas, novas técnicas devem ser aplicadas. A primeira motivação para separação de misturas convoluídas foi aplicada ao problema de separação de fontes de voz (*Cocktail Party*). A maioria dos algoritmos para os problemas de separação de fontes convoluídas pode ser encontrada no domínio do tempo ou da frequência, embora existam algoritmos recentes que utilizam um banco de filtros de sub-bandas (Choi et al., 2005).

Os algoritmos de separação cega de fontes são baseados em diferentes considerações sobre as fontes e o processo de mistura. Em geral, assume-se que as fontes são independentes ou, pelo menos, decorrelacionadas. O critério de separação pode ser definido em métodos baseados em estatística de ordem superior (*HOS* - *Higher Order Statistics*) e em métodos baseados em estatística de segunda ordem (*SOS* - *Second Order Statistics*). Uma série de algoritmos requer restrições mais fortes a respeito das características estatísticas das fontes. Como exemplo, alguns algoritmos requerem que as fontes sejam esparsas no domínio tempo-frequência, o que significa que os espectros das fontes raramente se sobrepõem (Yilmaz e Rickard, 2004a; Araki, Makino, Aichner, Nishikawa, e Saruwatari, 2003a; Blin, Araki, e Makino, 2005). De maneira similar, outros algoritmos de separação de misturas acústicas exploram a capacidade do sistema auditivo de identificar as relações entre objetos sonoros e a percepção de grupos ou sucessões de objetos sonoros com propriedades características. Este processo de percepção é chamado de análise de cena auditiva (*ASA* - *Auditory Scene Analysis*) (Bregman, 1990; Hu e Wang, 2007; Brown e Wang, 2005; Hu e Wang, 2001).

2.5.1 Métodos baseados em estatística de ordem superior (*HOS*)

A separação de fontes baseadas em estatística de ordem superior assume que as fontes sejam independentes. Muitos algoritmos são baseados na minimização da dependência de segunda e quarta ordem entre os sinais fonte. A separação de fontes baseadas em cumulantes tem sido proposta por vários autores. A separação de misturas convoluídas por meio de cumulantes de quarta ordem pode ser encontrada em (Comon e Rota, 2003; Capdevielle, Serviere, e Lacoume, 1995; Icart e Gautier, 1996; Hornillo-Mellado, Puntonet, Martin-Clemente, e Rodríguez-Álvarez, 2004). Um algoritmo baseado em cumulante e aplicado no domínio da frequência é o *JADE* (*Joint Approximate Diagonalisation of Eigenmatrices*) (Cardoso e Souloumiac, 1993). Algoritmos para a separação de misturas convoluídas baseados em curtose são mostrados em (Prasad, Saruwatari, e Shikano, 2003; Thomas, Deville, e Hosseini, 2006).

Alguns algoritmos utilizam indiretamente estatística de ordem superior para a separação de misturas convoluídas usando funções não-lineares para o cancelamento de momentos cruzados (*cross-moments*) (Mansour, Barros, e Ohnishi, 2000). O primeiro algoritmo para a separação de fontes convoluídas utilizando esta abordagem foi proposto por Jut-

ten, Thi, Dijkstra, Vittoz, e Caelen (1991). O algoritmo desenvolvido era aplicado para o caso de dois sensores e duas fontes, com a resposta ao impulso do ambiente considerada como sendo *FIR*. Back e Tsoi (1994) aplicaram o algoritmo proposto por Jutten e Héroult (1991) no domínio da frequência. Posteriormente o mesmo algoritmo foi aperfeiçoado por Charkani e Deville (1997). No caso destes algoritmos, o processo de mistura é considerado como sendo de fase mínima e utiliza-se uma rede com arquitetura do tipo *feedback*. Na separação de dois sinais de voz com dois microfones, o modelo proposto por Jutten et al. (1991) falha se os dois locutores estão localizados do mesmo lado, pois os filtros de separação não podem ser obtidos sem o atraso de uma das fontes e os filtros *FIR* também devem ser causais (Pedersen et al., 2007).

Métodos de separação cega de fontes baseados em teoria da informação propõem a separação de sinais através da maximização da entropia dos sinais de saída da rede. Uma das primeiras soluções foi desenvolvida por Bell e Sejnowski (1995), sendo proposta uma arquitetura de rede do tipo *feedforward* e utilização de filtros *FIR* (*Infomax*). O algoritmo *Infomax* melhorou em termos de eficiência utilizando o chamado gradiente natural ao invés de utilizar o gradiente absoluto (Amari et al., 1996; Amari, Cichocki, e Yang, 1997). *Infomax* é o método paramétrico mais comum e foi posteriormente modificado por Torkkola (1996) que empregou uma rede similar com ambas as arquiteturas do tipo *feedforward* e *feedback* para a separação de misturas convoluídas.

Um método mais eficiente em termos computacionais para a estimação de sistemas inversos consiste em representar os sinais no domínio da frequência. Smaragdis (1997) percebeu que, trabalhando no domínio da frequência, os problemas envolvendo misturas convoluídas tornam-se mais simples permitindo a aplicação de algoritmos utilizados em *ICA* para misturas instantâneas. Em Smaragdis (1997) é apresentado um algoritmo que utiliza o *Infomax*. As regras de aprendizagem são reformuladas para os dados no domínio da frequência e usadas para a aprendizagem de um processo de separação para cada componente de frequência. Um dos maiores problemas desta abordagem está relacionada à permutação e escala dos sinais recuperados que podem resultar numa estimação incorreta das fontes. Apesar de um grande número de heurísticas terem sido apresentadas (Anemuller e Kollmeier, 2003) este problema ainda está em aberto (O’Grady, Pearlmutter, e Rickard, 2005b).

Outros algoritmos utilizam uma abordagem baseada na medida de máxima verossimilhança (*Maximum Likelihood - ML*). Parra, Spence, e Vries (1998) apresentam um algoritmo que combina *ICA* com técnicas de modelagem de sinais baseadas em medida de máxima verossimilhança para misturas convoluídas. A abordagem probabilística denominada *MAP* (*maximum a posteriori probability*) é semelhante à abordagem baseada na medida de máxima verossimilhança. Nos métodos denominados de *MAP*, informações

a priori sobre os parâmetros do modelo são levadas em consideração. Algoritmos que utilizam esta abordagem na separação de misturas convoluídas podem ser encontrados em (Lambert, 1996; Lambert e Nikias, 1995; Olsson e Hansen, 2004a,b; Winter, Sawada, Araki, e Makino, 2004; Winter, Kellermann, Sawada, e Makino, 2007). Lambert (1995) propôs uma álgebra envolvendo uma matriz *FIR* que pode ser usada para problemas com misturas convoluídas. A idéia do algoritmo consiste em estender a álgebra envolvendo matrizes de escalares para matrizes de filtros (no domínio do tempo) ou polinômios (no domínio da frequência) (Lambert, 1996; O'Grady et al., 2005b). Westner e Bove (1999) sugerem um método de separação de misturas convoluídas para o caso sobre-determinado (mais sensores que fontes) baseado no algoritmo proposto por Lambert (1996). A técnica apresentada envolve a utilização de respostas ao impulso de um ambiente no qual as observações são produzidas, a aplicação de filtros *FIR* no processo de mistura e as respostas ao impulso inversas. A motivação para usar mais sensores que fontes está relacionada com a qualidade da separação: resultados mostram que a relação sinal-ruído das fontes separadas melhora com o aumento do número de sensores. Esta abordagem (onde as respostas ao impulso são medidas) diverge dos algoritmos de separação cega de fontes; entretanto, indicam a quantidade de informação necessária para a estimação de fontes convoluídas.

2.5.2 Métodos baseados em estatística de segunda ordem (SOS)

Em algumas situações a separação pode ser obtida utilizando estatística de segunda ordem e assumindo que as fontes não sejam correlacionadas. Estes métodos assumem considerações como a não-estacionariedade das fontes ou que os sistemas sejam de fase mínima (Lindgren e Broman, 1998). Os sistemas baseados em estatística de segunda ordem apresentam como vantagens: o fato de serem menos sensíveis ao ruído, necessitarem de menor quantidade de dados para separação dos sinais fonte, menor complexidade dos algoritmos e melhor eficiência computacional (Schobben e Sommen, 1999). Gerven e Compennolle (1995) mostraram que dois sinais fonte podem ser separados através da decorrelação, se o sistema de mistura é de fase mínima. Considera-se que o processo de mistura possa ser simulado através da aplicação de filtros *FIR*. Estes filtros *FIR* devem ser causais e possuir inversas estáveis. Estas condições não são satisfeitas se o processo de mistura for de fase não-mínima.

Em 1992, Feder, Weinstein, e Oppenheim (1992) desenvolveram um algoritmo iterativo no domínio do tempo para um sistema 2×2 de misturas convoluídas baseado na decorrelação dos sinais de saída. O algoritmo proposto apresentava um alto custo computacional. Em 1996, Lindgren e Veen (1996) propuseram um modelo similar ao de Feder et al. (1992), porém mais eficiente, baseado somente em estatística de segunda ordem.

Estas técnicas propostas são requeridas para a determinação de um processo de mistura inverso e têm sido aplicadas a vários métodos de separação de fontes convoluídas. Estes algoritmos tinham como limitação o fato de funcionarem apenas para sistemas 2×2 . Chan (1997) propôs um algoritmo baseado em decorrelação do sinal de saída e em estatística de segunda ordem para a separação de sistemas mais complexos do tipo $n \times n$.

O fato de muitos sinais serem não-estacionários pode ser utilizado na separação cega de fontes. Sinais de voz podem ser considerados não-estacionários para intervalos de tempo superiores a 10 *ms* (Murata, Ikeda, e Ziehe, 2001). A variância temporal relacionada à não-estacionariedade das fontes fornece informação adicional para a separação. Em contrapartida, a mudança de localização das fontes geralmente complica a separação dos sinais devido à variação no tempo do canal de mistura (Pedersen et al., 2007). A partir destas abordagens, muitos pesquisadores propuseram novos algoritmos como podemos verificar em Murata et al. (2001); Ikeda e Murata (1999). Murata et al. (2001) desenvolveram um algoritmo no domínio da frequência baseado no algoritmo *SOBI* utilizado para misturas instantâneas. Para resolver o problema da permutação é proposto um método baseado em uma estrutura temporal dos sinais que explora a não-estacionariedade do sinal de voz. As componentes separadas para cada frequência são combinadas com componentes de outras frequências que possuem um envelope similar durante um período de tempo especificado. Para obtermos a separação de fontes utilizando estatística de segunda ordem é necessário minimizarmos as correlações cruzadas. Estatística de segunda ordem no domínio da frequência é explorada utilizando o espectro de potência cruzada, através da minimização da potência cruzada. Um algoritmo eficiente no domínio da frequência foi proposto por Parra e Spence (2000). Wang, Chambers, e Sanei (2003a) propõem um critério de separação ligeiramente diferente e que proporciona uma convergência mais rápida do que o algoritmo proposto por Parra e Spence (2000).

Yin e Sommen (1999) apresentam um algoritmo baseado na consideração de não-estacionariedade das fontes e na simplificação do modelo de mistura. O modelo proposto assume que as respostas ao impulso do ambiente entre uma fonte e dois microfones muito próximos (localizados a distâncias menores que 10 *cm*) são iguais. O modelo considera apenas o sinal percorrido pelo caminho direto, os sinais reverberados são considerados ruídos (Pedersen et al., 2007; Yin e Sommen, 2000). Olsson e Hansen (2004a) mostraram que, com o conhecimento *a priori* sobre sinais de voz e a quasi-periodicidade dos sinais, a separação pode ser obtida a partir de um modelo de espaço de estados (*state-space*) não-estacionário e linear. Buchner, Aichner, e Kellermann (2003) propõem um algoritmo para a separação de fontes baseado em três tipos de abordagens: na não-brancura (*non-whiteness*), na não-estacionariedade e na não-gaussianidade das fontes. Os métodos que consideram as duas primeiras abordagens normalmente utilizam estatística de

segunda ordem e os métodos que são baseados na não-gaussianidade utilizam usualmente estatística de ordem superior. O algoritmo proposto combina os três tipos de abordagem usando métodos de estatística de segunda ordem e de ordem superior. Aplicações desta abordagem em separação de misturas convoluídas podem ser encontradas em Aichner, Buchner, Yan, e Kellermann (2004); Buchner, Aichner, e Kellermann (2004).

Um sinal pode ser chamado de ciclo-estacionário quando a função de distribuição cumulativa é invariante em relação a mudanças no tempo para algum período T . Muitos sinais utilizados em meios de comunicação apresentam a propriedade de ciclo-estacionariedade (Pedersen et al., 2007). Os sinais de voz podem, em algumas situações, ser considerados aproximadamente ciclo-estacionários (Sharnsunder e Giannakis, 1997). Wang, Jafari, Sanei, e Chambers (2003b) utilizam o critério de ciclo-estacionariedade para melhorar a performance de separação das fontes. Bradaric, Petropulu, e Diamantaras (2003) propõem um algoritmo de separação de fontes $n \times n$ com múltiplas entradas e múltiplas saídas (*MIMO*) para o caso especial dos sinais de entrada serem ciclo-estacionários. A ciclo-estacionariedade é utilizada para resolver o problema da permutação de frequências.

Os sinais de voz, assim como muitos sinais naturais, são temporalmente correlacionados. Este tipo de informação pode ser utilizado para a separação de fontes. A correlação temporal dos sinais pode ser utilizada para reduzir o problema de misturas convoluídas para um problema de misturas instantâneas (Mansour, Jutten, e Loubaton, 1996; Gorokhov e Loubaton, 1997). Mansour et al. (1996) mostram que, usando apenas estatística de segunda ordem, mas mais sensores que fontes, a mistura convoluída pode ser tratada como uma mistura instantânea.

2.5.3 Análise de cena auditiva

A área de estudo denominada de análise de cena auditiva (*Auditory Scene Analysis - ASA*) foi proposta por Bregman (1990) e, através do estudo do sistema auditivo humano, tem como objetivo a segregação de fontes sonoras. Esta técnica motivou o surgimento de métodos computacionais denominados de análise de cena auditiva computacional (*Computational Auditory Scene Analysis - CASA*) (Brown e Wang, 2005; Hu e Wang, 2007; Nix, Kleinschmidt, e Hohmann, 2003; Rutkowski, Cichocki, e Barros, 2001). Estes métodos, que exploram o fenômeno de mascaramento auditivo e a percepção do sistema auditivo do sinal dominante (com maior energia), motivaram o surgimento de técnicas que utilizam o mascaramento no domínio tempo-frequência para a separação de fontes sonoras (Weintraub, 1984; Hu e Wang, 2006, 2001). Com base no mascaramento auditivo humano, Roman et al. (2003) propõem uma abordagem de aprendizagem supervisionada para a segregação de sinais de voz, na qual o sinal de interesse é separado dos outros sinais fonte utilizando informações sobre a localização espacial: diferença de tempo interaural

(*ITD*) e diferença de intensidade interaural (*IID*). Para uma determinada configuração espacial podemos observar agrupamentos em torno dos valores estimados de *ITD* e *IID*. Roman et al. (2003) propõem um sistema de classificação de padrões para a estimação de máscaras espectrais visando a separação de sinais fonte (Roman e Wang, 2004). As técnicas de mascaramento no domínio do tempo-freqüência podem ser aperfeiçoadas pela utilização de filtros baseados no mascaramento perceptivo e limiares da audição humana (Pedersen et al., 2007; Guddeti e Guddeti, 2005).

Alguns algoritmos utilizam a freqüência fundamental do sinal de fala (*pitch*) para a separação de sinais de voz (Barros, Rutkowski, Mansour, Itakura, e Ohnishi, 2002; Tordini e Piazza, 2002). Estes métodos permitem a separação de sinais de voz nos casos onde os sinais apresentam diferentes valores de freqüência fundamental. A extração de múltiplas freqüências fundamentais em uma mistura convoluída não é um problema trivial. Tordini e Piazza (2002) propõem um algoritmo para a separação de misturas convoluídas baseado na extração da freqüência fundamental (*pitch*) e na teoria Bayesiana utilizando informação *a priori*. Uma outra abordagem de separação de fontes utilizando informação *a priori*, extração do *pitch* e de coeficientes mel-cepstrais (*MFCC*) é proposta por Gandhi e Hasegawa-Johnson (2004). O algoritmo foi aplicado em uma mistura de dois sinais de voz (uma voz feminina e outra masculina). Um *codebook* com coeficientes mel-cepstrais, obtidos a partir dos sinais originais, é utilizado na separação das fontes. Furukawa, Hioka, Ema, e Hamada (2003) propõem a utilização do *pitch* para especificar o número de sinais fonte presentes na mistura, como uma etapa de pré-processamento na separação dos sinais no domínio da freqüência (Pedersen et al., 2007). Barros et al. (2002) propõem um método para separação de sinais com mais fontes que sensores. O algoritmo proposto combina *ICA* com técnicas de análise de cena auditiva computacional, como extração de *pitch* e utilização de filtros auditivos. Bancos de filtros são utilizados para modelar o funcionamento da cóclea (Barros et al., 2002; Rutkowski et al., 2001).

2.5.4 Fontes esparsas no domínio do tempo-freqüência

Em muitos casos, o número de fontes é maior que o número de sensores (i.e. microfones disponíveis). Diversos métodos de separação cega de sinais são limitados pelo número de sensores disponíveis. Uma técnica que permite separar um número de fontes maior que o número de sensores se baseia na esparsidade dos sinais (O’Grady, Pearlmutter, e Rickard, 2005a). Assumindo que os sinais de voz são esparsos, ou seja, não se sobrepõem no domínio tempo-freqüência, podemos separá-los utilizando técnicas de mascaramento. Um dos primeiros algoritmos práticos, baseado na esparsidade dos sinais para separação de fontes em ambientes sem reverberação, mas contendo atrasos dos sinais, foi o *DUET* (*Degenerate Unmixing Estimation Technique*) inicialmente proposto por Jourjine, Rickard, e Yilmaz

(2000) e aperfeiçoado por Yilmaz e Rickard (2004a).

Uma máscara pode ser aplicada no domínio tempo-frequência para atenuar a energia do sinal de interferência e priorizar a energia do sinal de interesse. Em muitos casos a separação é obtida utilizando uma máscara binária (Rickard, Melia, e Fearon, 2005; Rickard e Yilmaz, 2002; Yilmaz e Rickard, 2004b). Entretanto, estes algoritmos foram desenvolvidos considerando um ambiente anecóico, contendo somente atrasos. No caso de ambientes com reverberação, o desempenho dos algoritmos é bem inferior ao obtido para misturas contendo somente atrasos (Jourjine et al., 2000). Para misturas convoluídas, Araki et al. (2003a) propõem um método de separação de fontes baseado em *ICA* e em mascaramento no domínio tempo-frequência. Normalmente os métodos que utilizam mascaramento no domínio tempo-frequência não usam mais que duas misturas para obter as máscaras. As máscaras binárias são aplicadas em apenas uma mistura (gravação). Rickard et al. (2005) apresentam um método de mascaramento denominado *DESPRIT* que utiliza informação de mais de duas misturas para a separação dos sinais fonte. Entretanto, os sinais recuperados após a aplicação das máscaras binárias apresentam artefatos (ruído musical). Para reduzir o efeito causado por estes artefatos são propostos algoritmos que utilizam máscaras não binárias (Raki, Makino, Sawada, e Mukai, 2005).

2.6 Algoritmos propostos para separação de fontes usando apenas uma mistura

No caso da separação com base em um número de misturas menor que o número de fontes, considerações mais específicas devem ser assumidas para a separação das fontes. No caso de problemas envolvendo pelo menos dois sensores (duas misturas) progressos têm sido alcançados utilizando considerações a respeito da esparsidade das fontes através do mascaramento no domínio tempo-frequência (Jourjine et al., 2000). Porém, para o caso de sinais gravados em ambientes reais, ou seja, misturas contendo reverberação, o desempenho dos algoritmos de separação cega de sinais ainda tem muito para ser melhorado. No caso mais extremo onde temos apenas uma mistura e duas ou mais fontes, temos um problema de difícil solução e ainda em aberto para o caso de sinais complexos como são os sinais de voz. Neste caso, simples considerações estatísticas relacionadas aos sinais fonte não são suficientes para a separação.

Vários algoritmos têm sido propostos na literatura para a separação de uma ou mais fontes utilizando apenas uma mistura (Bach e Jordan, 2004; Weiss e Ellis, 2006; Gandhi e Hasegawa-Johnson, 2004; Reddy e Raj, 2004a,b; Shashanka, Raj, e Smaragdis, 2007; Roweis, 2000). A maioria destas abordagens para a separação dos sinais fonte visa a estimação de máscaras ideais no domínio tempo-frequência. Estas máscaras conseguem

construir com alta qualidade uma variedade de sinais e têm sido amplamente aplicadas na etapa de pré-processamento de sistemas de reconhecimento de voz. A introdução de erros causados pela estimação das máscaras ideais gera, entretanto, uma degradação gradual no desempenho de sistemas de reconhecimento de voz (Roman et al., 2003).

2.6.1 Método de aprendizagem da máscara ideal utilizando *HMM*

Sistemas baseados na análise de cena auditiva computacional utilizam substancial conhecimento do sistema auditivo humano e características psicoacústicas. Porém, em situações acústicas reais é difícil para os sistemas incorporar uma grande quantidade de conhecimento estatístico sobre o problema em questão. Por outro lado, pesquisadores que trabalham com análise de componentes independentes e separação cega de sinais têm desenvolvido técnicas que trabalham com várias misturas considerando um ambiente de mistura simplificado e poderosas técnicas estatísticas de separação cega. Estes algoritmos (inclusive aqueles desenvolvidos para atuar no caso de mais fontes que sensores) não podem separar fontes no caso de existir apenas uma mistura (Roweis, 2000).

Roweis (2000) propõe uma técnica chamada de refiltragem capaz de recuperar sinais fonte através da aplicação de uma máscara em uma única mistura e discute a aplicação de algoritmos estatísticos no processo de aprendizagem da máscara ideal. O algoritmo considera que a energia do sinal em diferentes instantes de tempo e frequência apresenta informações independentes a respeito dos sinais fonte. Isto sugere a análise do espectrograma do sinal que mostra a energia em função do tempo e da frequência. A idéia básica de refiltragem consiste em recuperar os sinais fonte através da aplicação de uma máscara sobre as componentes de frequência do sinal misturado. Esta técnica representa a base de muitos sistemas de separação que utilizam abordagem do tipo *CASA*, conforme mostra a Figura 2.3. Diferentes fontes podem ser separadas através da aplicação de diferentes máscaras.

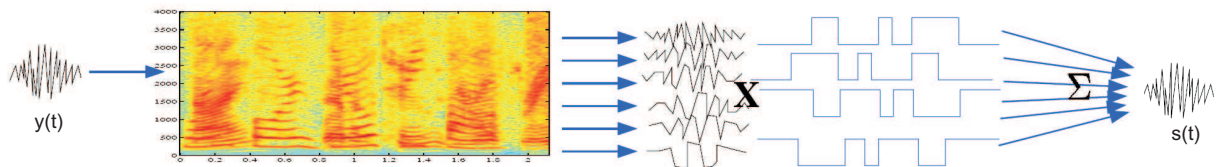


Figura 2.3: Técnica de mascaramento aplicada à separação de uma fonte utilizando apenas uma mistura (Roweis, 2000).

A separação dos sinais fonte utilizando apenas um microfone é possível se as máscaras utilizadas são bem escolhidas. Desta forma, o problema consiste em determinar computa-

cionalmente as máscaras ideais a partir de uma única mistura. O objetivo consiste em especificar quais são as regiões do espectrograma que pertencem ao mesmo objeto auditivo. Os sinais de voz apresentam um conjunto de regularidades em relação à forma como a energia é distribuída ao longo do domínio tempo-frequência. Sugestões para o agrupamento de frequências baseadas nestas regularidades vêm sendo estudadas há muitos anos e utilizadas em muitos sistemas do tipo *CASA* (Hu e Wang, 2006, 2007). Pode-se observar através da análise do espectrograma várias sugestões para o agrupamento de determinadas regiões, tais como: i) existência de uma estrutura com múltiplos harmônicos; ii) casos onde as frequências de uma determinada região aumentam ou diminuem sua energia de forma simultânea e iii) regiões onde a energia de determinadas frequências movem-se em um determinado sentido simultaneamente. A Figura 2.4 mostra exemplos destas sugestões bastante utilizadas para o agrupamento de frequências de uma única fonte baseado na distribuição da energia.

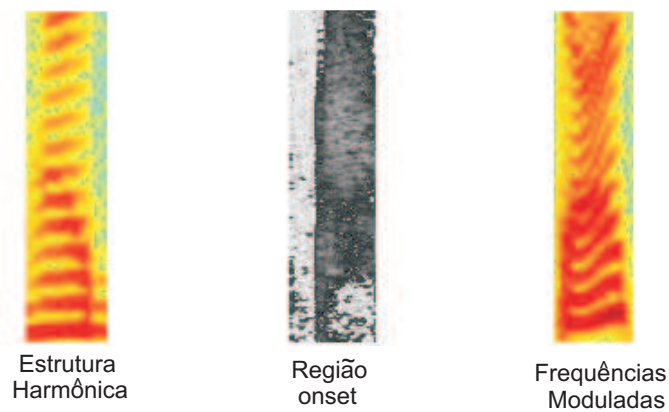


Figura 2.4: Exemplos de sugestões clássicas utilizadas para o agrupamento de frequências de uma única fonte baseada na distribuição da energia. A figura à esquerda mostra a existência de uma estrutura com múltiplos harmônicos. A figura do centro exemplifica casos onde as frequências de uma determinada região aumentam ou diminuem sua energia de forma simultânea e abrupta. A figura à direita exemplifica regiões onde a energia de determinadas frequências apresenta uma determinada tendência em relação ao aumento ou diminuição das frequências ao longo do tempo (Roweis, 2000).

Existem vários sistemas de reconhecimento de padrões que podem ser aplicados para aproveitar a informação fornecida a partir do espectrograma do sinal. Estes métodos podem ser classificados como supervisionados ou não supervisionados. Os métodos não-supervisionados utilizam modelos de aprendizagem de fontes isoladas e tentam explicar os dados referentes à mistura como sendo causados pela interação dos modelos das fontes individuais. Já os métodos supervisionados, os quais determinam modelos de agrupamento em uma mistura acústica, necessitam de dados pré-identificados da mistura de entrada para o processo de aprendizagem.

Roweis (2000) propõe a utilização de um algoritmo de aprendizagem juntamente com

a abordagem de mascaramento do sinal. O algoritmo proposto utiliza um modelo *HMM* (*Hidden Markov Model*) dependente de locutor cujos resultados indicam que a aprendizagem automática no problema de separação de fontes utilizando um microfone pode ser possível. O sistema proposto não necessita também da informação do número de fontes presentes na mistura. Cada máscara proposta pelo sistema tem um determinado valor de probabilidade, grupos contendo altos valores de probabilidades podem ser considerados como fontes. O sistema retorna uma coleção de possíveis máscaras e suas respectivas probabilidades e com base nisto, decide o número de fontes presentes na cena auditiva.

2.6.2 Método de obtenção da máscara ideal utilizando uma abordagem de aprendizagem espectral

A abordagem proposta por Bach e Jordan (2004) explora a esparsidade das fontes, resolvendo o problema de separação de fontes através da segmentação no domínio tempo-frequência. É proposto um algoritmo de aprendizagem que otimiza um critério de segmentação baseado em características espectrais dos sinais. Os sinais para o treinamento do algoritmo mencionado são obtidos através da mistura de sinais de voz normalizados e separados.

As características espectrais utilizadas são sugestões clássicas propostas por estudos na área de percepção acústica. Cada uma destas sugestões está associada com um tempo específico de duração, considerando que os dados foram amostrados a uma taxa de 5500 *Hz*: pequeno (menor que cinco quadros), médio (entre 10 a 20 quadros) e grande (relacionadas a todos os quadros). Nenhuma destas características usadas isoladamente é capaz de separar os sinais fonte, daí a necessidade de combiná-las (Bach e Jordan, 2004).

Considerando sugestões para sinais não-harmônicos, duas características são exploradas: continuidade e fatos comuns. A característica de continuidade considera que dois pontos no domínio tempo-frequência pertencem ao mesmo segmento se eles assumem valores próximos tanto no tempo quanto na frequência. A outra característica relacionada ao comportamento comum do espectro mostra que elementos que possuem a mesma variação no tempo provavelmente pertencem à mesma fonte. Estas características normalmente estão relacionadas ao início (*onset*) ou término (*offset*) de um determinado agrupamento de frequências ao se analisar uma sub-banda do espectrograma.

Ao analisar-se as estruturas harmônicas, principalmente para o caso de sinais de voz, muitas informações podem ser obtidas através da estimação da frequência fundamental. De acordo com as características espectrais apresentadas o sistema proposto constrói matrizes de afinidades (similaridades) que são utilizadas para definir o segmento espectral. Em particular esta abordagem define matrizes de afinidade parametrizadas e utiliza um algoritmo de aprendizagem para ajustar os parâmetros.

Bach e Jordan (2004) realizaram experimentos utilizando dados de quatro diferentes locutores contendo sinais de voz com duração de três segundos. Um total de 28 parâmetros foram estimados pelo algoritmo proposto por Bach e Jordan (2004). Para os testes foram utilizadas misturas com cinco locutores diferentes dos que participaram do treinamento. A Figura 2.5 mostra um exemplo de segmentação obtida para dois locutores do conjunto de testes. A imagem à esquerda mostra a segmentação ideal obtida a partir do conhecimento dos dois sinais de voz separados e a imagem à direita revela a segmentação obtida pelo algoritmo. Apesar de algumas componentes do locutor (cujo espectrograma está representado pela cor preta) estarem ausentes, o desempenho do algoritmo permite obter sinais audíveis e com qualidade razoável.

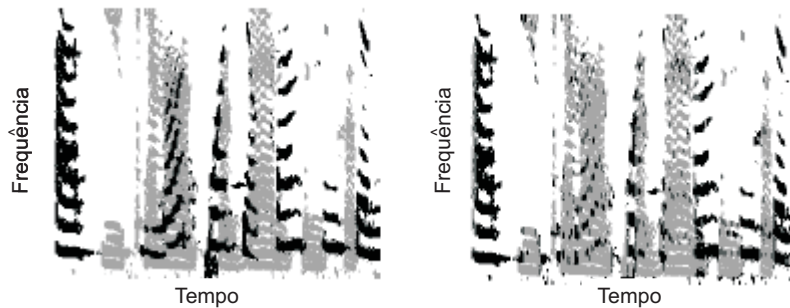


Figura 2.5: O espectrograma à esquerda mostra uma segmentação ótima (obtida considerando que os sinais são conhecidos) de dois sinais de voz, onde os dois locutores estão representados pelas cores cinza e preto. À direita pode-se observar o espectrograma estimado para os dois locutores utilizando o algoritmo proposto por Bach e Jordan (2004).

É importante salientar que o método apresentado não requer conhecimento *a priori* dos locutores, porém requer que os dois locutores tenham valores de frequência fundamental bem distintos. Outra desvantagem do algoritmo apresentado é a alta complexidade computacional. O código fonte foi desenvolvido em Matlab e em linguagem C, possuindo um tempo de processamento de aproximadamente trinta minutos para separar quatro segundos de fala utilizando um processador com 1.8 GHz e 1 GB de RAM (Bach e Jordan, 2004).

2.6.3 Abordagem probabilística para a separação de fontes utilizando uma única mistura

A idéia principal desta abordagem consiste em explorar a estrutura temporal das fontes através da aprendizagem de um conjunto de funções base no domínio do tempo capaz de representar as fontes de uma maneira eficiente. Um algoritmo de aprendizagem baseado

na medida de máxima probabilidade dada uma única mistura e um conjunto de funções base foi proposto por Jang e Lee (2002). Para cada ponto no tempo são inferidos os sinais fonte e seus fatores de contribuição. Esta inferência é possível devido ao conhecimento *a priori* das funções e de seus respectivos coeficientes de ponderação (Jang e Lee, 2002).

Para formular o problema assume-se que o sinal observado x^t corresponde a uma soma de P sinais fonte independentes:

$$x^t = \lambda_1 s_1^t + \lambda_2 s_2^t + \dots + \lambda_P s_P^t, \quad (2.4)$$

onde s_i^t corresponde ao t -enésimo valor da amostra da i -enésima fonte e λ_p corresponde ao ganho de cada fonte que é fixo no tempo. O objetivo consiste em recuperar todas as amostras s_i^t , dado as amostras de uma única mistura x^t .

Como já colocado, análise em Componente Independentes tem se mostrado uma técnica eficiente na codificação de padrões, incluindo imagens e sinais de voz (Lee e Jang, 2001; Lee et al., 2000). *ICA* considera que o sinal de voz possa ser decomposto em funções base com seus respectivos coeficientes de ponderação. As funções base podem ser adaptadas pela regra de aprendizagem padrão de *ICA*. Esta técnica foi empregada por Lee et al. (2000) para a aprendizagem de funções base de um sinal de voz. As funções base e os coeficientes aprendidos pela *ICA* constituem uma representação eficiente de uma fonte sonora no domínio do tempo.

O método proposto por Jang e Lee (2002) utiliza as funções base estimadas através da *ICA* para a separação das fontes a partir de uma única mistura. Estas funções são estimadas de tal forma que os coeficientes sejam independentes e a informação redundante minimizada. As funções são obtidas *a priori* a partir de um conjunto de dados de treinamento e estes dados são utilizados para separar as fontes sonoras dos dados de teste. Lee et al. (2000) utilizaram um método de análise em componentes independentes para estimar as funções base e os coeficientes das fontes para locutores masculinos e femininos, respectivamente. As características aprendidas mostraram diferenças significativas no domínio da frequência e do tempo. Os resultados obtidos sugerem que o sinal feminino apresenta um espalhamento muito maior no domínio do tempo, conforme mostra a Figura 2.6.

Os sinais fonte originais são considerados como uma superposição linear e ponderada de funções base, sendo descritos da seguinte forma:

$$s_i^t = \sum_{k=1}^M \mathbf{b}_{ik} c_{ik}^t = \mathbf{B}_i \mathbf{c}_i^t, \quad (2.5)$$

onde M corresponde ao número de funções base, $\mathbf{b}_{ik}$ é a k -enésima função base da i -enésima fonte e c_{ik}^t corresponde aos respectivos coeficientes de ponderação.

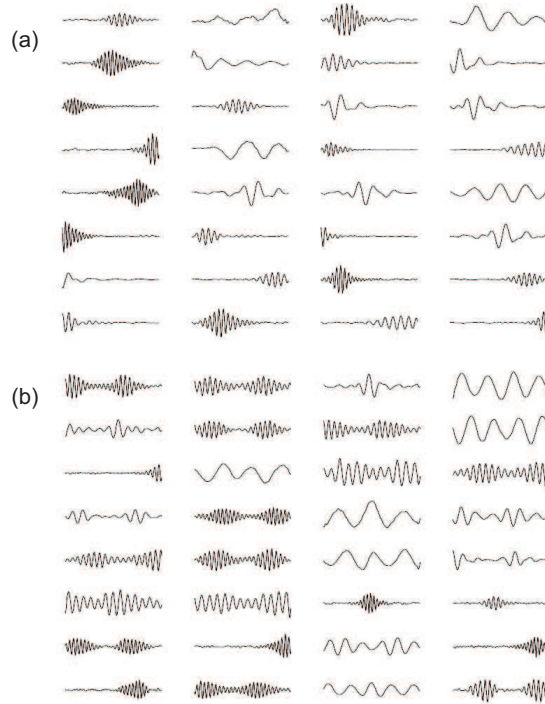


Figura 2.6: Funções base obtidas através da análise em componentes independentes (*ICA*) para locutores masculinos (a) e femininos (b). Estas funções foram obtidas através do algoritmo de *ICA* proposto por Lee et al. (2000) para segmentos de voz contendo 64 amostras e através da seleção de 32 funções base a partir de um conjunto de 64 (Lee et al., 2000).

A inversa da matriz de funções base, $O_i = B_i^{-1}$ referente aos filtros *ICA* transforma os segmentos analisados nos vetores dos coeficientes: $\mathbf{c}_i^t = O_i \mathbf{s}_i^t$. O algoritmo *ICA* irá buscar uma transformação linear que proporcione componentes tão independentes quanto possível, maximizando o produto das densidades marginais das componentes transformadas a partir de um conjunto dado de treinamento:

$$\mathbf{O}_i^* = \arg \max_{\mathbf{O}_i} \prod_t Pr(\mathbf{s}_i^t | \mathbf{O}_i) = \arg \max_{\mathbf{O}_i} \prod_t \prod_k Pr(c_{ik}^t). \quad (2.6)$$

O histograma dos coeficientes dos dados reais revela que a distribuição apresenta um pico afiado e uma longa cauda conforme verificado na Figura 2.7(c). O algoritmo propõe a utilização de uma distribuição gaussiana generalizada que permita estimar com eficiência uma distribuição não-gaussiana simétrica através do ajuste do expoente q do conjunto de parâmetros θ da seguinte forma:

$$p(c|\theta) \propto \exp \left[- \left| \frac{c - \mu}{\sigma} \right|^q \right], \quad \theta = \{\mu, \sigma, q\}, \quad (2.7)$$

onde $\mu = E[c]$, e σ corresponde ao desvio padrão de c . Através do algoritmo de *ICA*, as funções e seus individuais parâmetros θ_{ik} são obtidos a partir de dados de treinamento

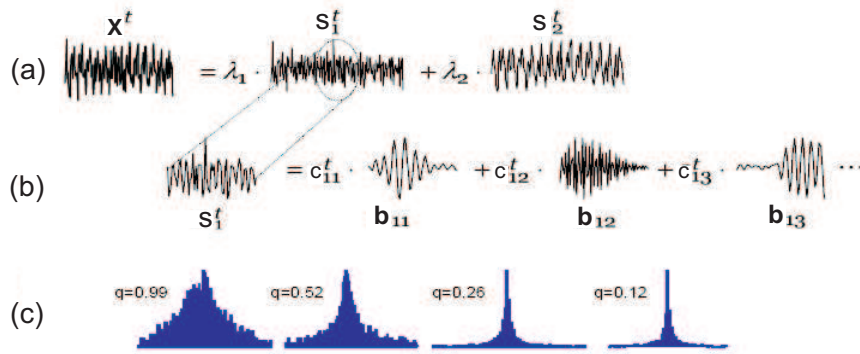


Figura 2.7: Modelos gerados a partir da mistura de sinais fonte. (a) Uma mistura é gerada a partir de dois sinais de voz com características diferentes. (b) O sinal de uma das fontes é gerado a partir da soma ponderada dos coeficientes (c_{ik}^t) das funções base (b_{ik}). (c) Histograma dos coeficientes obtidos. Geralmente apresentam uma distribuição super-gaussiana. As distribuições são modeladas utilizando uma função de densidade gaussiana da seguinte forma: $p(c_{ik}^t) \propto \exp(-|c_{ik}^t|^q)$. Pode-se observar da esquerda para a direita que a distribuição se torna mais super-gaussiana (Jang e Lee, 2002).

que são utilizados como informação *a priori* pelo algoritmo de separação das fontes.

O algoritmo recupera os sinais fonte, iterativamente através do método do gradiente ascendente, maximizando a probabilidade dos sinais separados, calculada através do uso das funções base e das funções de densidade de probabilidade dos coeficientes obtidos. Experimentos mostram ser possível a separação de dois sinais fonte diferentes para misturas de sinais de áudio (*rock* e *jazz*) e dois sinais de voz (sendo uma voz masculina e outra feminina) (Jang e Lee, 2002).

2.6.4 Estimação de uma máscara não-binária para a separação de fontes utilizando uma única mistura

Conforme descrito até agora, a maioria dos métodos apresentados para separação de fontes utilizando apenas uma mistura no domínio tempo-freqüência utiliza máscaras binárias (*hard masks*). Reddy e Raj (2004b) propõem um algoritmo para separação de fontes utilizando uma máscara não-binária. O sinal de voz reconstruído pode ser estimado a partir do espectro de potência do locutor de interesse. Os resultados experimentais encontrados são melhores que os obtidos quando se utiliza a máscara binária. A seguir segue o detalhamento do algoritmo proposto por Reddy e Raj (2004b).

Considere que $s_1(t)$ e $s_2(t)$ como sendo sinais de voz gerados por dois locutores falando simultaneamente em um mesmo microfone. O sinal misturado ($x(t)$) captado pelo microfone corresponde à soma dos dois sinais:

$$x(t) = s_1(t) + s_2(t). \quad (2.8)$$

O espectro de potência do sinal $s_1(t)$ pode ser representado por $S_1(w)$:

$$S_1(w) = |\mathfrak{F}(s_1(t))|^2, \quad (2.9)$$

onde $\mathfrak{F}$ representa a transformada de Fourier. Similarmente $S_2(w)$ e $X(w)$ correspondem aos espectros de potência de $s_2(t)$ e $x(t)$, respectivamente. Considerando que $s_1(t)$ e $s_2(t)$ são decorrelacionados, temos que:

$$X(w) = S_1(w) + S_2(w). \quad (2.10)$$

A relação mostrada na Equação 2.10 é válida estritamente para termos longos e não é garantida para medições do espectro de potência a partir da análise de quadros de comprimento finito.

Considerando que $x(w)$, $s_1(w)$ e $s_2(w)$ representem o logaritmo de $X(w)$, $S_1(w)$ e $S_2(w)$, respectivamente, a partir da Equação 2.10 temos:

$$x(w) = \log(e^{s_1(w)} + e^{s_2(w)}), \quad (2.11)$$

que pode ser escrita da seguinte forma:

$$x(w) = \max(s_1(w), s_2(w)) + \log(1 + e^{(\min(s_1(w), s_2(w)) - \max(s_1(w), s_2(w)))}). \quad (2.12)$$

A potência espectral para qualquer frequência de um sinal de voz misturado é geralmente dominada por um locutor, o que nos permite dizer que:

$$x(w) \approx \max(s_1(w), s_2(w)). \quad (2.13)$$

O tamanho do quadro utilizado deve ter uma boa relação de compromisso entre as equações 2.10 e 2.13. Reddy e Raj (2004b) propõem a utilização de quadros com tamanho igual a 25 *ms*. A Figura 2.8 mostra o logaritmo do espectro de potência de um segmento com 25 *ms* de um sinal misturado com dois locutores, bem como o logaritmo do espectro de potência dos sinais originais. Pode-se observar que, em geral, o valor do logaritmo do espectro de potência do sinal misturado é muito próximo do maior valor de potência espectral entre os sinais separados, apesar de não ser sempre exatamente o mesmo valor.

Reddy e Raj (2004b) modelaram a distribuição do logaritmo do espectro de potência para qualquer locutor através de uma mistura de densidade gaussiana.

Considerando que s_1 e s_2 representem o logaritmo do espectro de potência para os locutores 1 e 2, respectivamente, a distribuição de s_1 para o locutor 1 pode ser representada por:

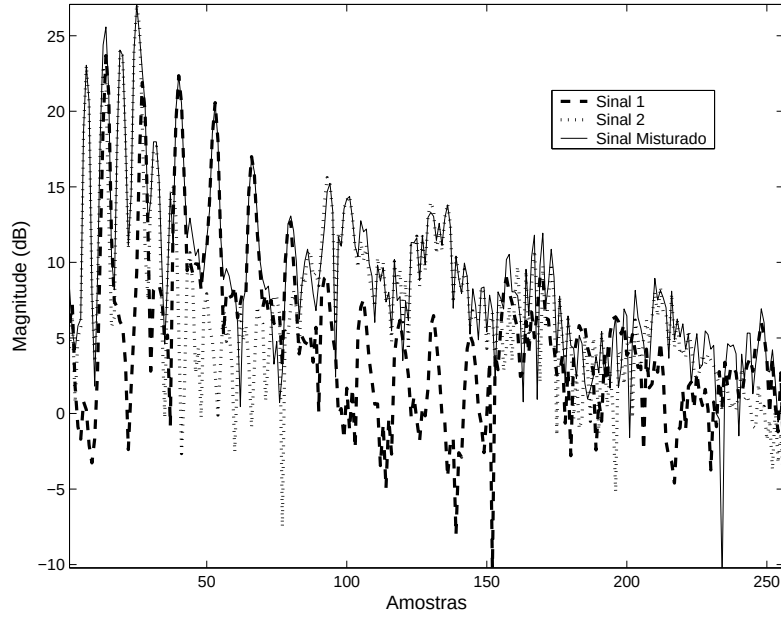


Figura 2.8: Espectros de potência de uma mistura de dois sinais de voz e dos dois sinais que formam a mistura.

$$P(s_1) = \sum_{k_{s_1}=1}^{k_{s_1}} P_{s_1}(K_{s_1}) \prod_{d=1}^D \eta(s_{1d}, \mu_{k_{s_1},d}^{s_1}, v_{k_{s_1},d}^{s_1}), \quad (2.14)$$

onde, K_{s_1} corresponde ao número de gaussianas utilizada na mistura, $P_{s_1}(K_{s_1})$ representa a probabilidade *a priori* da k_{s_1} -enésima gaussianas, D representa a dimensão do vetor de potência espectral (s_1), s_{1d} representa a d -enésima dimensão de s_1 . A média e variância da k_{s_1} -enésima gaussianas da mistura correspondem respectivamente a $\mu_{k_{s_1},d}^{s_1}$ e $v_{k_{s_1},d}^{s_1}$. A função de densidade gaussianas com média $\mu_{k_{s_1},d}^{s_1}$ e variância $v_{k_{s_1},d}^{s_1}$ em s_{1d} é dada por $\eta(s_{1d}, \mu_{k_{s_1},d}^{s_1}, v_{k_{s_1},d}^{s_1})$. A função de distribuição de s_2 ($P(s_2)$) pode ser explicada de forma análoga. Os parâmetros de $P(s_1)$ e $P(s_2)$ são ajustados de acordo com um banco de dados de sinais de voz dos dois locutores.

Para a separação dos sinais fonte é considerado o valor obtido de $P(s_{1d} = x_d|x)$ como sendo os valores de uma máscara não-binária (*soft mask*) que identifica a contribuição do locutor 1 em relação ao espectro do sinal misturado x . Considerando que m_{s_1} corresponda à máscara não-binária para o locutor 1, temos que a máscara para o locutor 2 é definida por: $1 - m_{s_1}$. O detalhamento do algoritmo pode ser encontrado em (Reddy e Raj, 2004b).

Na etapa de treinamento das misturas de distribuições gaussianas foi utilizada uma hora de gravações para dois locutores, sendo um masculino e outro feminino. As misturas de densidades gaussianas com 256 gaussianas foram estimadas através do logaritmo do espectro de cada locutor. Além dos dados de treinamento, dez minutos de gravação para

cada locutor foram utilizados para testes. Os resultados obtidos pelo algoritmo proposto por Reddy e Raj (2004b) apresentaram uma melhora em torno de 3 a 5 dB quando comparados a outros métodos que utilizam máscaras binárias (Reddy e Raj, 2004b).

2.7 Algoritmos de separação cega de sinais fonte

Uma série de algoritmos de separação cega de sinais fonte pode ser encontrada na literatura conforme descrito neste capítulo. Porém, muitos algoritmos consideram o processo de mistura como sendo instantânea. Misturas instantâneas são produzidas artificialmente pelo computador e não correspondem a processos reais, onde pode-se verificar a presença de atrasos e reverberação. Vários algoritmos foram desenvolvidos para separar sinais fonte a partir de misturas convoluídas, mas o desempenho destes algoritmos ainda têm muito para ser melhorado. No caso de separação de sinais de voz reverberados algoritmos que utilizam o mascaramento espectral no domínio do tempo e frequência para a separação dos sinais são propostos, conforme mostrado na Tabela 2.1. Estes algoritmo tem apresentado um desempenho superior aos demais para o caso específico de separação cega de sinais de voz reverberados. A Tabela 2.1 apresenta um resumo dos principais algoritmos de separação cega de sinais fonte mencionados neste capítulo.

Algoritmos	Tipos de Misturas	Características
<i>Infomax</i> (Bell e Sejnowski, 1995).	Instantâneas	Algoritmo de separação cega de fontes baseado na maximização da informação mútua entre sinais de entrada e de saída.
<i>FOBI</i> - Identificação cega de quarta ordem (Cardoso, 1989).	Instantâneas	Algoritmo utiliza momentos de quarta ordem para realizar a separação de fontes.
<i>EASI</i> - <i>Equivariant Adaptive Separation via Independence</i> (Cardoso e Laheld, 1996).	Instantâneas	Algoritmo de separação baseado em uma expressão recursiva para obtenção da matriz de separação.
<i>FastICA</i> - Algoritmo rápido de análise em componentes independentes (Hyvarinen e Oja, 1997).	Instantâneas	O algoritmo explora medidas de não-gaussianidade para a separação das fontes.
Puntonet et al. (1995) propuseram um método geométrico.	Instantâneas	Método geométrico para <i>ICA</i> baseado nas propriedades dos espaços vetoriais das fontes e misturas.
<i>SOBI</i> - Identificação cega de segunda ordem (Belouchrani et al., 1997).	Instantâneas	Algoritmo baseado em estatística de segunda ordem.
<i>Deflation</i> (Delfosse e Loubaton, 1995).	Instantâneas	Algoritmo baseado em estatística de ordem superior. O método proposto extrai as fontes componente por componente.
<i>MAP</i> - Máxima probabilidade a posteriori (Belouchrani e Cardoso, 1994).	Instantâneas	Algoritmo baseado em uma abordagem probabilística e uma das primeiras soluções propostas para o caso sub-determinado.
<i>JADE</i> - <i>Joint approximate diagonalisation of eigenmatrices</i> (Cardoso e Souloumiac, 1993).	Convoluídas	Um algoritmo baseado em cumulante e aplicado no domínio da frequência.
<i>Infomax</i> foi modificado por Torkkola (1996).	Convoluídas	O algoritmo modificado utiliza uma rede similar com ambas as arquiteturas do tipo <i>feedforward</i> e <i>feedback</i> para a separação de misturas convoluídas.
Smaragdis (1997) propôs um algoritmo que utiliza o <i>Infomax</i> .	Convoluídas	As regras de aprendizagem são reformuladas para os dados no domínio da frequência.
Parra et al. (1998) propõem algoritmo baseado na medida de máxima verossimilhança.	Convoluídas	O algoritmo combina <i>ICA</i> com técnicas de modelagem de sinais baseadas em medida de máxima verossimilhança para misturas convoluídas.
Murata et al. (2001) desenvolveram um algoritmo no domínio da frequência baseado no algoritmo <i>SOBI</i> .	Convoluídas	O algoritmo para resolver o problema da permutação propõe um método baseado em uma estrutura temporal dos sinais que explora a não-estacionariedade do sinal de voz.
Parra e Spence (2000) propõem um algoritmo baseado em estatística de segunda ordem no domínio da frequência.	Convoluídas	O algoritmo utiliza o espectro de potência cruzada para a separação das fontes através da minimização da potência cruzada.
Yin e Sommen (1999) apresentam um algoritmo baseado na consideração de não-estacionariedade das fontes e na simplificação do modelo de mistura (Cardoso e Souloumiac, 1993).	Convoluídas	O modelo proposto assume que as respostas ao impulso do ambiente entre uma fonte e dois microfones muito próximos são iguais.
Buchner et al. (2003) propõem um algoritmo baseado na não-brancura, na não-estacionariedade e na não-gaussianidade das fontes.	Convoluídas	O algoritmo proposto combina os três tipos de abordagem usando métodos de estatística de segunda ordem e de ordem superior.
Bradaric et al. (2003) propõem um algoritmo de separação de fontes $n \times n$ com múltiplas entradas e múltiplas saídas (<i>MIMO</i>).	Convoluídas	O algoritmo utiliza a ciclo-estacionariedade dos sinais de entrada para resolver o problema da permutação de frequências.
Roman et al. (2003) propõem um sistema de classificação de padrões para a estimação de máscaras espectrais visando a separação de sinais fonte.	Convoluídas	O algoritmo apresenta uma abordagem de aprendizagem supervisionada para a segregação de sinais de voz utilizando informações sobre a localização espacial (<i>ITD</i>) e diferença de intensidade interaural (<i>IID</i>).
Tordini e Piazza (2002) propõem um algoritmo baseado na frequência fundamental.	Convoluídas	O algoritmo baseia-se na extração da frequência fundamental (<i>pitch</i>) e na teoria Bayesiana utilizando informação <i>a priori</i> .
Gandhi e Hasegawa-Johnson (2004) propõem a extração do <i>pitch</i> e de coeficientes mel-cepstrais (<i>MFCC</i>) para a separação de sinais.	Convoluídas	O algoritmo foi aplicado em uma mistura de dois sinais de voz (uma voz feminina e outra masculina). Um <i>codebook</i> com coeficientes mel-cepstrais, obtidos a partir dos sinais originais, é utilizado na separação das fontes.
Barros et al. (2002) propõem um método para a separação de sinais com mais fontes que sensores.	Convoluídas	O algoritmo proposto combina <i>ICA</i> com técnicas de análise de cena auditiva computacional, como extração de <i>pitch</i> e utilização de filtros auditivos.
<i>DUET</i> - <i>Degenerate Unmixing Estimation Technique</i> (Jourjine et al., 2000).	Convoluídas	O algoritmo assume que os sinais de voz são esparsos, ou seja, não se sobrepõem no domínio tempo-frequência, podemos separá-los utilizando técnicas de mascaramento.
<i>DESPRIT</i> - <i>DUET-ESPIRIT</i> Rickard et al. (2005).	Convoluídas	O algoritmo propõem um método de mascaramento que utiliza informação de mais de duas misturas para a separação dos sinais fonte.
Roweis (2000) propõe uma técnica chamada de refiltragem capaz de recuperar sinais fonte através da aplicação de uma máscara usando apenas uma única mistura.	Convoluídas	O algoritmo propõem a aplicação de métodos estatísticos (<i>HMM</i>) no processo de aprendizagem da máscara ideal. O algoritmo considera que a energia do sinal em diferentes instantes de tempo e frequência apresenta informações independentes a respeito dos sinais fonte.
Bach e Jordan (2004) explora a esparsidade das fontes, resolvendo o problema de separação de fontes, através da aplicação de uma máscara usando apenas uma única mistura.	Convoluídas	É proposto um algoritmo de aprendizagem que otimiza um critério de segmentação baseado em características espectrais dos sinais.
Jang e Lee (2002) propõem um algoritmo de aprendizagem baseado na medida de máxima probabilidade dada uma única mistura e um conjunto de funções base.	Convoluídas	Para cada ponto no tempo são inferidos os sinais fonte e seus fatores de contribuição. Esta inferência é possível devido ao conhecimento <i>a priori</i> das funções base e de seus respectivos coeficientes de ponderação.
Reddy e Raj (2004b) propõem um algoritmo para separação de fontes utilizando uma máscara não-binária dada uma única mistura.	Convoluídas	O sinal de voz reconstruído pode ser estimado a partir do espectro de potência do locutor de interesse. Os resultados experimentais encontrados são melhores que os obtidos quando se utiliza a máscara binária.

Tabela 2.1: Algoritmos de separação cega de sinais.

2.8 Considerações finais

Este capítulo teve como objetivo apresentar diversas abordagens e algoritmos existentes na literatura referentes à área de separação cega de fontes. Algumas diferenças entre os algoritmos de separação de fontes instantâneas e convoluídas foram discutidas. No caso de misturas instantâneas pode-se perceber a existência de várias técnicas de *BSS/ICA*. Os resultados obtidos pelos algoritmos de misturas instantâneas apresentam uma separação de fontes quase perfeita. Entretanto em situações reais, envolvendo misturas convoluídas, não existem muitos algoritmos em comparação com a quantidade de métodos propostos para misturas instantâneas. O desempenho dos métodos de separação cega de fontes convoluídas existentes ainda é baixo e possui muitas limitações. Um caso mais difícil e ainda em aberto é a separação cega de fontes utilizando apenas uma mistura. Algoritmos de separação de sinais de voz utilizando apenas uma mistura são apresentados neste capítulo.

Capítulo 3

Mascaramento Espectral de Sinais

O ser humano possui a capacidade de identificar e perceber uma fonte acústica específica junto a uma mistura contendo diversas fontes, como ocorre no problema conhecido como *cocktail party*. O sistema auditivo possui um mecanismo capaz de identificar as relações entre objetos sonoros e a percepção de grupos ou sucessões de objetos sonoros com propriedades características. Este processo de percepção, como já mencionado anteriormente, é chamado de análise de cena auditiva (*ASA*). A análise de cena auditiva inspirou a criação de muitos sistemas computacionais utilizados para a separação de sons, originando um novo campo de pesquisa chamado de análise de cena auditiva computacional (*CASA*) (Brown e Wang, 2005; Hu e Wang, 2006).

Em situações naturais, raramente encontramos fontes sonoras ativas num mesmo instante de tempo e frequência. Um ponto importante em muitos sistemas (*CASA*) é o mascaramento no domínio tempo-frequência. Uma fonte sonora específica pode ser recuperada aplicando uma máscara com pesos sobre uma mistura acústica a cada instante de tempo e frequência. As regiões dominadas pela fonte em questão recebem pesos maiores que as regiões onde predominam as outras fontes sonoras da mistura analisada. As máscaras podem ser binárias ou podem assumir valores reais. O uso da máscara binária é motivada pelo processo de mascaramento que ocorre na audição humana, onde um som de intensidade menor é mascarado por um outro de intensidade maior dentro de uma mesma banda crítica. Com o propósito de separação de sinais de voz, Hu e Wang (2001) propuseram a utilização de uma máscara binária ideal. Dado um sinal de voz $s(t, f)$ e um ruído $n(t, f)$, onde t e f representam o instante de tempo e frequência, respectivamente, o mascaramento binário ideal $m(t, f)$ pode ser obtido através da seguinte expressão:

$$m(t, f) = \begin{cases} 1 & \text{se } s(t, f) > n(t, f), \\ 0 & \text{caso contrário.} \end{cases} \quad (3.1)$$

Uma abordagem similar foi adotada por Jourjine et al. (2000), que observaram uma

tendência, apresentada por diferentes sinais de voz, a serem ortogonais numa representação no domínio tempo-freqüência com alta resolução, sendo possível a separação dos sinais através de uma máscara binária. Um grande número de trabalhos tem demonstrado que o sinal de fala reconstruído por uma máscara ideal binária é intelegível quando extraído de uma mistura de dois ou mais locutores (Brown e Wang, 2005; Roman et al., 2003; Roman e Wang, 2004; Roweis, 2000). O mascaramento ideal também tem sido utilizado para melhorar de forma substancial a inteligibilidade de um sinal de voz corrompido por ruído (Roman et al., 2003).

3.1 Mascaramento ideal no domínio tempo-freqüência

O mascaramento ideal tem como objetivo determinar as regiões no domínio tempo-freqüência cuja energia do sinal de interesse seja maior que o sinal de interferência. A máscara ideal binária, considerada neste trabalho, é uma matriz binária que assume valor igual 1 quando a energia do sinal de interesse é maior que a energia do sinal de interferência para uma freqüência específica em um determinado instante de tempo, assumindo valor igual a zero, caso contrário. Ao aplicarmos uma máscara ideal a uma mistura instantânea de dois sinais de voz, observamos que o sinal obtido é perfeitamente audível, com boa qualidade e sem nenhuma interferência do outro sinal. Um dos objetivos deste trabalho é avaliar a separação obtida de um sinal de voz quando aplicamos a máscara binária ideal a uma mistura de sinais considerando a presença de atrasos e reverberação.

Para a determinação da máscara ideal os sinais fonte (s_1 e s_2) são transformados para o domínio da freqüência, obtendo seus respectivos espectrogramas:

$$\begin{aligned} s_1(t) &\rightarrow S_1(w, t), \\ s_2(t) &\rightarrow S_2(w, t), \end{aligned} \tag{3.2}$$

onde w representa a freqüência angular e t representa o instante de tempo do quadro de voz analisado. O mascaramento binário pode ser determinado através da comparação da magnitude dos dois espectrogramas, mostrados na Figura 3.1.

As máscaras ideais binárias (M_1 e M_2) foram obtidas da seguinte forma:

$$\begin{aligned} M_1 &= 1, \quad \text{para } |S_1(w, t)| > |S_2(w, t)|, \\ M_2 &= 1, \quad \text{para } |S_2(w, t)| > |S_1(w, t)|, \end{aligned} \tag{3.3}$$

os demais valores das máscaras correspondem a zero. As máscaras obtidas são aplicadas diretamente sobre as misturas dos sinais de voz no domínio tempo-freqüência.

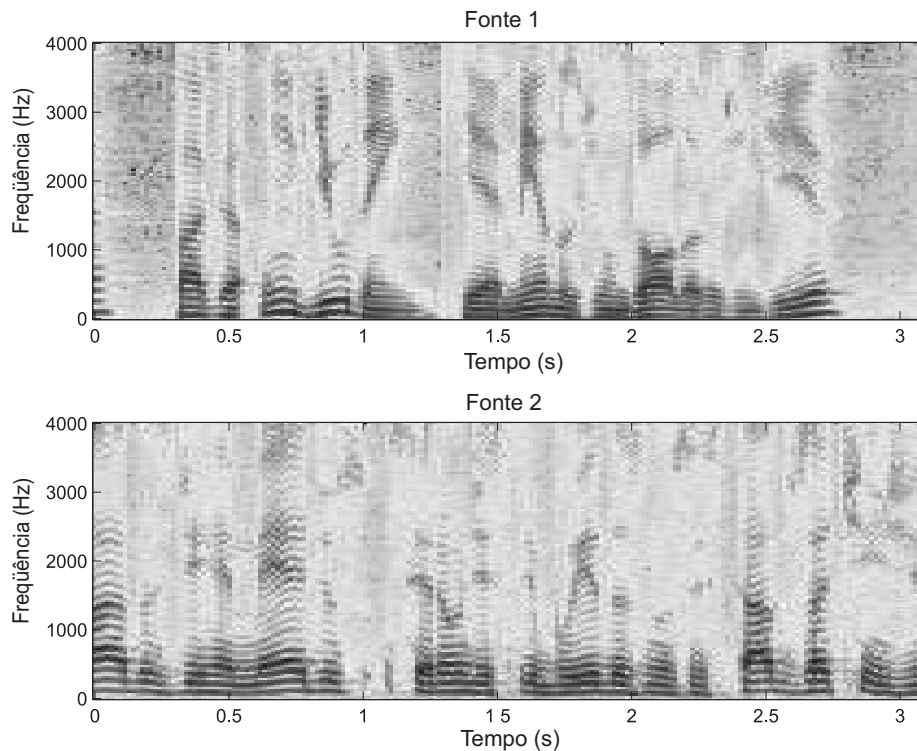


Figura 3.1: Espectrograma de dois sinais de voz.

3.2 Fontes esparsas e o Algoritmo *DUET*

Recentemente, vários métodos baseados no mascaramento no domínio tempo-frequência para a separação de sinais fonte foram propostos (Yilmaz e Rickard, 2004b; Araki et al., 2003a). A separação é baseada na não sobreposição do espectro das fontes no domínio tempo-frequência. Se os sinais fonte são suficientemente esparsos, ou seja, a maioria das amostras do sinal são quase zero, podemos considerar em uma mistura que no máximo uma componente de frequência estará ativa para cada instante de tempo e frequência. Um dos primeiros algoritmos práticos baseado nesta consideração, proposto para a separação de fontes em ambientes sem reverberação, foi o *DUET* (*Degenerate Unmixing Estimation Technique*) (Jourjine et al., 2000; Yilmaz e Rickard, 2004b). Este método propõe a extração de sinais fonte a partir da aplicação de uma máscara binária no domínio tempo-frequência. O método proposto é baseado na medida chamada de "*W-disjoint orthogonality (WDO)*" (Yilmaz e Rickard, 2004b). A medida *WDO* baseia-se na consideração de que o sinal de voz é esparsos tanto no domínio tempo como da frequência e que apenas uma fonte sonora está ativa em um determinado instante de tempo e frequência. Pode-se verificar que esta propriedade é aproximadamente satisfeita para o caso onde os sinais fonte são sinais de voz (Blin et al., 2005). Conhecendo qual fonte está ativa para cada componente de frequência num determinado instante de tempo podemos separar as

componentes de cada fonte correspondente utilizando uma máscara binária. Investigações a respeito da propriedade de esparsidade dos sinais de voz são detalhadas em Yilmaz e Rickard (2004b), considerando gravações em uma câmara anecóica e em Blin et al. (2005) para o caso de gravações realizadas em um ambiente com tempos de reverberação variando entre 130 *ms* e 200 *ms*. Considerando uma mistura de dois sinais de voz, s_1 e s_2 , a medida *WDO* assume que o produto destes sinais no domínio do tempo e frequência seja igual a zero:

$$\hat{s}_1(w, t)\hat{s}_2(w, t) = 0, \quad \forall \quad w, t. \quad (3.4)$$

A Equação 3.4 é apenas aproximadamente satisfeita para o caso de sinais simultâneos de voz porque a representação dos sinais de voz ativos no domínio do tempo-frequência não é nula. Entretanto o sinal de voz é esparso e uma pequena porcentagem dos coeficientes dos sinais no domínio do tempo representam a maior parte da energia. Ou seja, para diferentes sinais de voz podemos dizer que é improvável encontrar sinais de voz diferentes e ativos em um mesmo instante de tempo e frequência, conforme pode-se verificar na Figura 3.2.

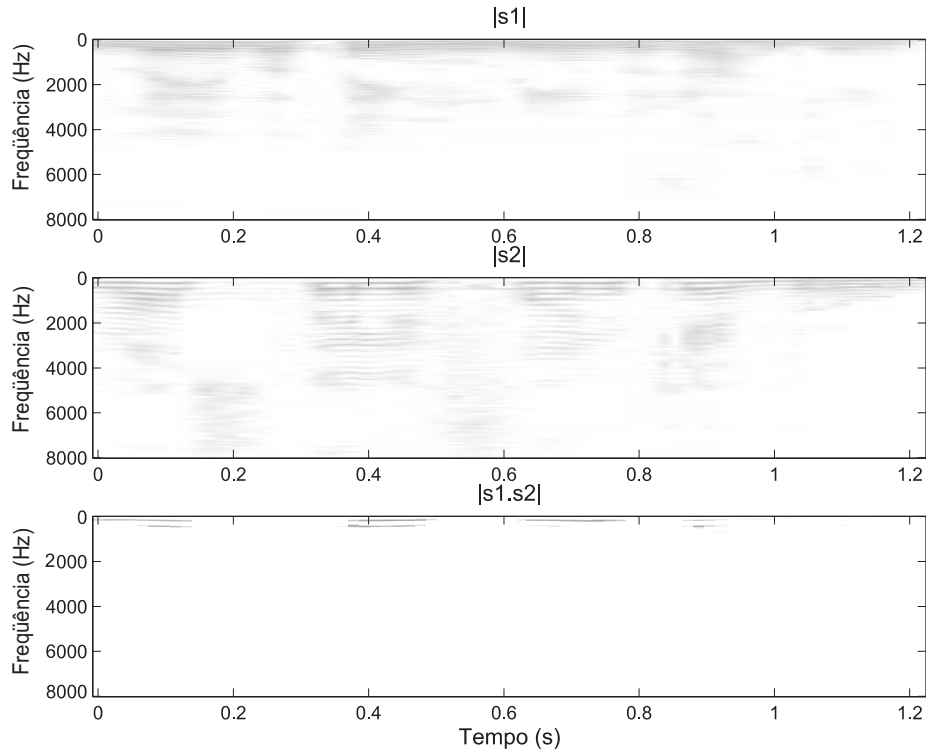


Figura 3.2: As três figuras acima correspondem aos dois espectrogramas dos sinais de voz s_1 e s_2 normalizados com energia unitária e ao produto dos mesmos. As figuras de cima para baixo correspondem, respectivamente, aos espectrogramas dos sinais de voz: $|s_1(w, t)|$, $|s_2(w, t)|$ e $|s_1(w, t)s_2(w, t)|$. Pode-se observar na figura de baixo ($|s_1(w, t)s_2(w, t)|$) que os valores encontrados satisfazem aproximadamente a medida de *WDO*.

Em sistemas do tipo "CASA" a técnica de mascaramento no domínio tempo-freqüência tem sido aplicada há anos (Bregman, 1990; Brown e Wang, 2005). Uma vantagem de utilizarmos uma máscara binária é que somente uma decisão binária deve ser feita. Esta decisão é baseada em técnicas de agrupamento ou na medida de direção do sinal (*direction of arrival* - *DOA*). No caso do algoritmo *DUET*, proposto por Yilmaz e Rickard (2004a), dois parâmetros, chamados de atenuação e tempo de atraso relativos, são utilizados para a formação de *clusters* e obtenção das máscaras binárias. Entretanto a utilização de máscaras binárias provoca uma descontinuidade do sinal devido ao preenchimento de zeros ao extrairmos o sinal fonte gerando uma distorção do sinal recuperado e também a presença de ruído (*loud musical noise*) (Araki et al., 2003a). Outra desvantagem do algoritmo *DUET* é o fato de que, quando aplicado sobre misturas convoluídas, o desempenho deste método cai significativamente. Devido à reverberação, o parâmetro de atraso relativo obtido através da fase do sinal sofre várias alterações não permitindo a identificação de agrupamentos (*clusters*) e prejudicando a separação dos sinais.

3.3 Estimação da amplitude e atraso relativos

Considere uma medição de um par de sensores sem a presença de reverberação. A primeira mistura ($x_1(t)$), considerada como sendo a referência, tem os valores de atenuação e atraso nulos. As duas misturas podem ser expressas da seguinte forma (Yilmaz e Rickard, 2004b):

$$x_1(t) = \sum_{j=1}^N s_j(t), \quad (3.5)$$

$$x_2(t) = \sum_{j=1}^N \alpha_j s_j(t - \delta_j), \quad (3.6)$$

onde $s_j(t)$, $j = 1, \dots, N$, correspondem às N fontes, δ_j é o atraso relativo entre o segundo sensor e a fonte j , e α_j é o fator de atenuação relativo aos percursos entre as fontes e o segundo sensor, conforme ilustrado na Figura 3.3.

No domínio tempo-freqüência tem-se então:

$$\begin{bmatrix} \hat{x}_1(w, t) \\ \hat{x}_2(w, t) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ \alpha_1 e^{-i w \delta_1} & \dots & \alpha_N e^{-i w \delta_N} \end{bmatrix} \begin{bmatrix} \hat{s}_1(w, t) \\ \vdots \\ \hat{s}_N(w, t) \end{bmatrix}. \quad (3.7)$$

Considerando que as fontes são esparsas e que apenas uma das N fontes está ativa (assume valor diferente de zero) para uma dada freqüência (w) e instante de tempo (t), a Equação 3.7 pode ser simplificada:

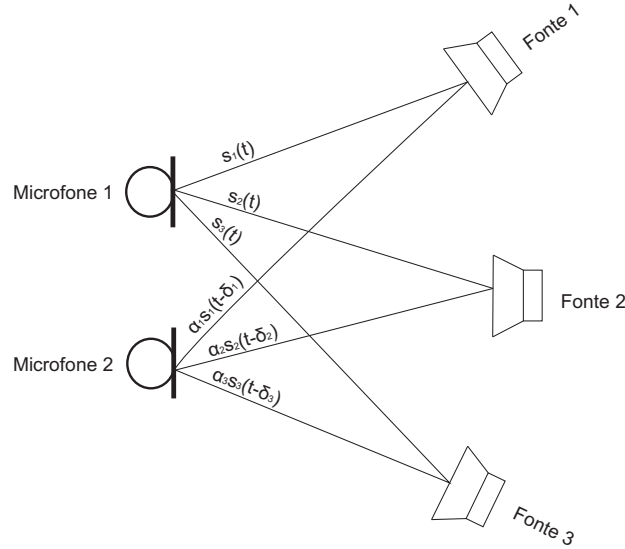


Figura 3.3: Diagrama do algoritmo *DUET* para o caso de misturas com três fontes.

$$\begin{bmatrix} \hat{x}_1(w, t) \\ \hat{x}_2(w, t) \end{bmatrix} = \begin{bmatrix} 1 \\ \alpha_j e^{-iw\delta_j} \end{bmatrix} \hat{s}_j(w, t), \quad (3.8)$$

para um dado valor j . A amplitude e o atraso relativo associados a uma determinada frequência e instante de tempo podem ser expressos por:

$$(\alpha_{w,t}, \delta_{w,t}) = \left(\left| \frac{\hat{x}_2(w, t)}{\hat{x}_1(w, t)} \right|, \frac{1}{w} \angle \frac{\hat{x}_1(w, t)}{\hat{x}_2(w, t)} \right), \quad (3.9)$$

onde $\angle ae^{i\phi} = \phi$, $-\pi \leq \phi \leq \pi$.

3.4 Definição do histograma e separação das fontes

No processo de separação das fontes é identificada cada componente de frequência (w) em um determinado instante de tempo (t), através dos valores obtidos de amplitude relativa ($\alpha_{w,t}$) e atraso relativo ($\delta_{w,t}$). No algoritmo *DUET* é utilizado como parâmetro de amplitude relativa: $\alpha_{w,t}^* = (\alpha_{w,t}) - (1/\alpha_{w,t})$, com o objetivo de tornar os valores presentes no eixo relativo à amplitude simétricos. É construído então um histograma bidimensional cujos picos correspondem aos valores de amplitude e atraso relativos de cada uma das fontes. Através do agrupamento dos pontos no domínio tempo-frequência pertencentes às mesmas fontes é possível a construção de máscaras binárias utilizadas para a separação dos sinais.

O histograma possui N picos, os quais correspondem à quantidade de fontes presentes na mistura, centrados em $(\alpha_j \text{ e } \delta_j)$, para $j = 1, 2, \dots, N$. A altura dos picos nos histogra-

mas é proporcional à a energia das componentes de frequência das fontes correspondentes presentes no sinal misturado. O histograma é obtido de acordo com a energia das componentes de frequência das fontes em função dos valores de amplitude relativa e atraso relativo, conforme a expressão abaixo:

$$h(\alpha, \delta) = \sum_{w,t} |\hat{x}_1(w, t) \hat{x}_2(w, t)|; \quad (3.10)$$

onde o termo h representa o histograma ponderado de energia em relação aos parâmetros de amplitude e atraso relativos referentes a dois sinais fonte.

Para a separação das fontes são criadas máscaras no domínio tempo-frequência correspondentes a cada pico do histograma: a máscara de uma fonte específica receberá valores iguais a um, nos instantes de tempo e frequência pertencentes ao pico da respectiva fonte. As máscaras binárias obtidas são então aplicadas a uma única mistura para produzir a representação no domínio do tempo-frequência dos sinais fonte originais. Na Figura 3.4 pode-se observar os histogramas obtidos para o caso de uma fonte apenas e misturas com seis e dez fontes.

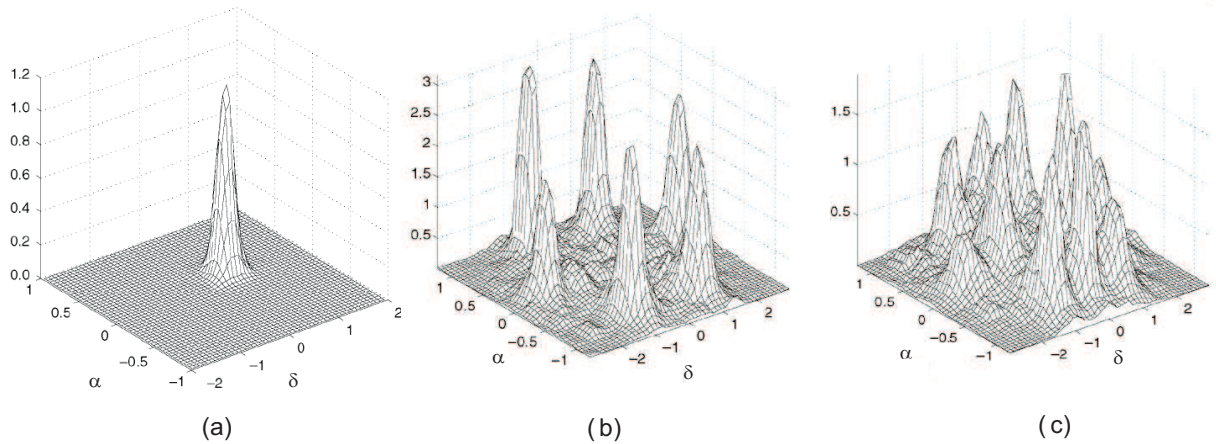


Figura 3.4: Exemplos de histogramas obtidos utilizando o *DUET*. Cada pico no histograma corresponde a uma fonte de acordo com os seus respectivos valores de amplitude e atraso relativos. A figura (a) representa uma fonte com valores de $(\alpha_j, \delta_j) = (0, 2; 0, 5)$. A figura (b) representa os picos obtidos de acordo com os parâmetros de amplitude e atraso relativos para uma mistura de seis fontes. A figura (c) representa os picos obtidos de acordo com os parâmetros de amplitude e atraso relativos para uma mistura de dez fontes (Yilmaz e Rickard, 2004a).

3.5 Medida de desempenho

O desempenho dos algoritmos de separação de fontes propostos na literatura apresenta limitações. Estas limitações estão relacionadas às variações de ganho e distorção provocadas pelo algoritmo no sinal estimado. Em determinadas situações, pode-se permitir um

grau maior ou menor de distorção no sinal recuperado. Em aplicações envolvendo sinais de voz, por exemplo, distorções causadas no sinal durante o processo de recuperação das fontes podem ser permitidas, desde que o sinal de voz continue inteligível.

Para análise de desempenho da recuperação dos sinais fonte será utilizada a medida de relação sinal-distorção (SDR) dada por:

$$SDR = 10 \log_{10} \left(\frac{\sum_k s_i^2(k)}{\sum_k (s_i(k) - \hat{s}_i(k))^2} \right), \quad (3.11)$$

onde s_i corresponde ao sinal original da fonte i e $\hat{s}_i$ representa o sinal estimado da fonte i . Quanto maior for o valor da relação sinal-distorção (SDR) melhor é a separação obtida. Se o sinal obtido for idêntico ao original então a medida de SDR assume valor igual a ∞ .

A distorção aceitável é um parâmetro de projeto e depende da aplicação. Conforme será verificado nos testes realizados, pode-se constatar que, ao aplicar uma máscara ideal a uma mistura instantânea obtém-se uma $\overline{SDR}$ média de 12 dB . O sinal apresenta distorção perceptível quando comparado ao original, porém apresenta boa qualidade sonora em termos de inteligibilidade.

3.6 Análise de desempenho do algoritmo *DUET*

Para a análise de desempenho do algoritmo *DUET*, Yilmaz e Rickard (2004a) propõem uma série de experimentos utilizando misturas sintéticas e gravações realizadas em ambientes anecóicos e com reverberação.

No caso de misturas sintéticas, Yilmaz e Rickard (2004a) realizaram testes utilizando uma mistura de seis fontes (foi utilizado o banco de dados *TIMIT* para escolha das fontes) usando os seguintes valores de atenuação e atraso relativos (1;-2), (3/2;-2), (3/2;1), (1;2), (2/3;1) e (2/3;-1). O histograma obtido com a mistura das seis fontes é mostrado na Figura 3.4(b). Os sinais separados apresentaram uma relação sinal-distorção média de 6 dB . Os sinais fonte recuperados apresentaram distorções, mas em termos de inteligibilidade possuíam boa qualidade sonora. Para verificar o limite do algoritmo *DUET*, Yilmaz e Rickard (2004a) realizaram também testes com uma mistura de dez fontes. Os sinais obtidos apresentaram uma relação sinal-distorção média de apenas 2 dB . Neste caso, os sinais recuperados apresentaram muitas distorções comprometendo a inteligibilidade dos mesmos (Yilmaz e Rickard, 2004a).

Yilmaz e Rickard (2004a) avaliaram ainda o desempenho do algoritmo *DUET* para o caso de um ambiente anecóico, utilizando uma mistura de dois sinais de voz gravados com taxa de amostragem de 16000 Hz . Para as gravações foram utilizados microfones omni-direcionais (espaçados de uma distância de 1,75 cm) e os sinais de voz (fontes) foram colocados em várias posições ao redor dos microfones sobre um semi-círculo com

1,5 *m* de raio. Foram realizados vários testes de acordo com as posições das fontes. Os dois sinais recuperados apresentaram uma relação sinal-distorção (*SDR*) média em torno de 15 *dB*. O pior resultado (*SDR* = 12 *dB*) aconteceu quando as fontes foram colocadas nas posições de 0° e 30° graus em relação aos microfones. A razão para esta diminuição do desempenho deve-se ao fato dos picos dos histogramas começarem a se sobrepor dificultando a separação dos sinais fonte.

Ao mudarmos do ambiente anecóico para um ambiente com reverberação, pode-se observar que o desempenho do algoritmo cai bastante. Isto se deve ao fato do modelo ser baseado em uma mistura anecóica. Um ambiente reverberante faz com que, do ponto de vista do microfone, a mesma fonte venha de diversas direções e com diversos atrasos relativos. Yilmaz e Rickard (2004a) realizaram gravações em um escritório com tempo de reverberação de 500 *ms*. A mistura obtida apresentava três fontes colocadas a 0°, 90° e 180° graus de posição em relação aos microfones. Embora os sinais separados apresentassem distorção e baixa inteligibilidade, pode-se dizer que alguma separação ainda é obtida de acordo com Yilmaz e Rickard (2004a). A Figura 3.5 compara os histogramas individuais e em conjunto para as gravações realizadas em ambiente anecóico e com reverberação contendo três fontes em três diferentes posições. É possível observar que, na gravação realizada em ambiente anecóico, pode-se identificar os agrupamentos de fontes (*clusters*) em regiões distintas. Já no caso da mistura em ambiente com reverberação, os agrupamentos das fontes (*clusters*) estão espalhados e se sobrepõem uns aos outros.

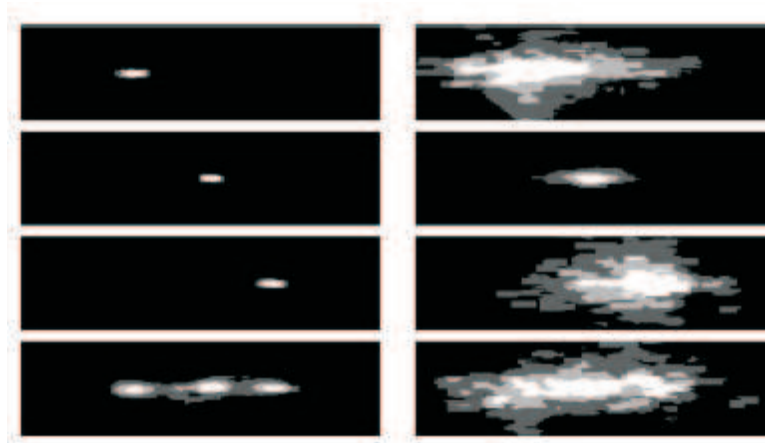


Figura 3.5: Comparação de histogramas para misturas anecóica e com reverberação. Na coluna da esquerda, de cima para baixo, temos histogramas individuais e em conjunto para uma mistura gravada em ambiente anecóico de três fontes colocadas a 0°, 90° e 180° graus de posição em relação aos microfones. O histograma em conjunto contendo todas as fontes da mistura corresponde basicamente a soma dos histogramas individuais. Na coluna da direita, de cima para baixo, temos também histogramas individuais e em conjunto para uma mistura gravada em ambiente com reverberação de três fontes colocadas a 0°, 90° e 180° graus de posição em relação aos microfones. No histograma da mistura anecóica pode-se observar que os picos estão bem definidos (esquerda, centro e direita). Já na mistura com reverberação não é possível identificar os picos devido à sobreposição dos mesmos (Yilmaz e Rickard, 2004a).

A Figura 3.6 mostra o histograma obtido para uma gravação de dois locutores em um ambiente real com reverberação, a qual pode ser encontrada no banco de dados disponibilizado por Te-Won Lee (Lee, 1998). As gravações foram realizadas em um escritório, com as seguintes dimensões: $5,5\text{ m} \times 8,0\text{ m} \times 2,5\text{ m}$. Os microfones foram colocados a uma distância de 120 cm dos locutores e os sinais foram gravados com uma taxa de amostragem de 16000 Hz . Pode-se observar a existência de vários picos, demonstrando que métodos baseados somente em amplitude e atrasos relativos não são eficientes na separação de fontes convoluídas.

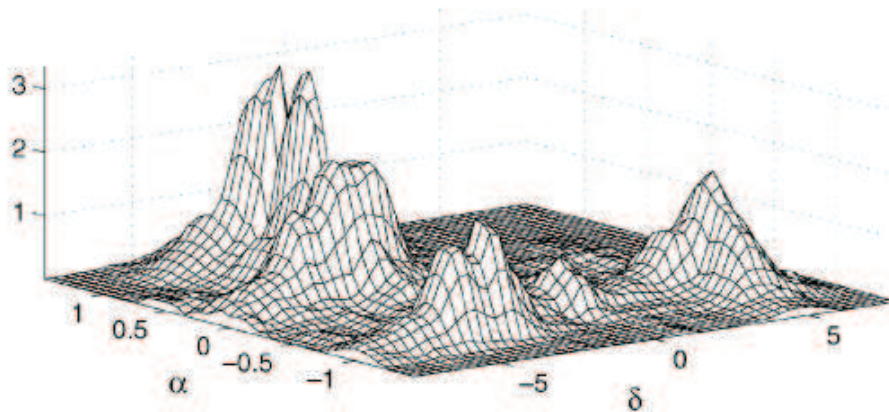


Figura 3.6: Histograma obtido para uma gravação de dois locutores em um ambiente real com reverberação que pode ser encontrada no banco de dados disponibilizado por Te-Won Lee (Yilmaz e Rickard, 2004a; Lee, 1998).

3.7 Eliminação da informação de fase (atraso relativo)

Uma inspeção dos histogramas da Figura 3.5 mostra que a informação de amplitude relativa é a que permite uma separação efetiva das fontes. Para verificar a influência da informação de fase (atraso relativo) na separação de sinais de voz, considerando misturas instantâneas, com atrasos e com reverberação, é proposto um algoritmo que não leva em consideração a informação de fase para a estimação das máscaras ideais binárias.

O algoritmo proposto utiliza apenas a informação de amplitude relativa ($\alpha_{w,t}$) para a definição do histograma, descartando a informação referente ao parâmetro denominado de atraso relativo ($\delta_{w,t}$). Neste caso o histograma é obtido de acordo com a energia das componentes de frequência das fontes em função apenas dos valores de amplitude relativa, conforme a expressão abaixo:

$$h(\alpha) = \sum_{w,t} |\hat{x}_1(w,t)\hat{x}_2(w,t)|; \quad (3.12)$$

onde o termo h representa o histograma ponderado de energia em relação ao parâmetro de amplitude relativa referente a dois sinais fonte.

Os testes realizados na Seção 3.9 mostram que a utilização da informação de fase para a separação das fontes (considerando uma mistura com duas fontes) não implica em diferença dos resultados obtidos. Em determinadas situações a utilização da informação de fase pode até proporcionar um desempenho ligeiramente inferior.

3.8 Análise do tamanho dos quadros utilizados para separação dos sinais via mascaramento espectral

De forma a determinar os limites de desempenho da separação de fontes baseadas em máscaras espectrais, o primeiro passo é verificar a relação do tamanho do quadro utilizado para a separação de sinais via mascaramento no domínio do tempo-freqüência e a relação sinal-distorção (SDR). Os sinais de voz utilizados para os testes realizados nesta seção possuem taxa de amostragem igual a 10000 Hz . Para a obtenção do sinal no domínio tempo-freqüência, os sinais foram divididos em quadros com diferentes tamanhos: 128 , 256, ... e 16384 amostras.

Através da comparação dos espectros de dois sinais de voz no domínio tempo-freqüência, foram obtidas máscaras ideais utilizando diferentes tamanhos de quadros. Foi analisada a distorção causada pelas máscaras ideais quando as aplicamos diretamente sobre os sinais fonte. As medidas de distorção médias referentes aos sinais recuperados utilizando estas máscaras são mostradas na Figura 3.7. Pode-se verificar que a relação sinal-distorção máxima é obtida quando utilizamos quadros de 1024 amostras e possui SDR média de 12 dB .

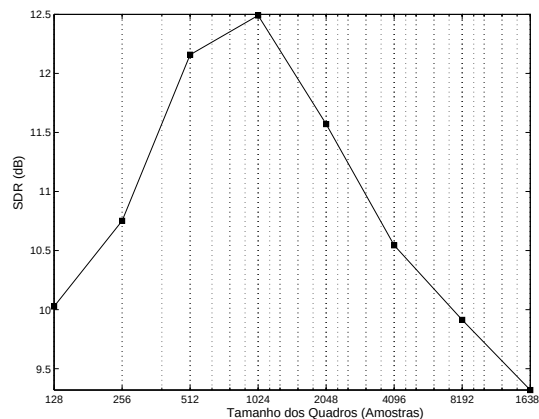


Figura 3.7: Relação sinal-distorção média ($\overline{SDR}$) obtida após a aplicação de máscaras ideais, para diferentes tamanhos de quadros (com sobreposição de quadros equivalente a 50% do tamanho dos mesmos), sobre dois sinais de voz separados.

O desempenho das máscaras ideais também foi avaliado para o caso de uma mistura contendo a soma dos dois sinais de voz, sendo aplicadas sobre esta mistura máscaras binárias ideais utilizando diferentes tamanhos de quadros. Na Figura 3.8 pode-se verificar que os sinais foram separados e, conforme esperado, apresentaram um desempenho inferior quando comparado com a aplicação das máscaras ideais diretamente sobre os sinais fonte originais. Porém, pode-se verificar que, novamente a relação sinal-distorção máxima é obtida quando utilizamos quadros de 1024 amostras e possui SDR média de 9 dB .

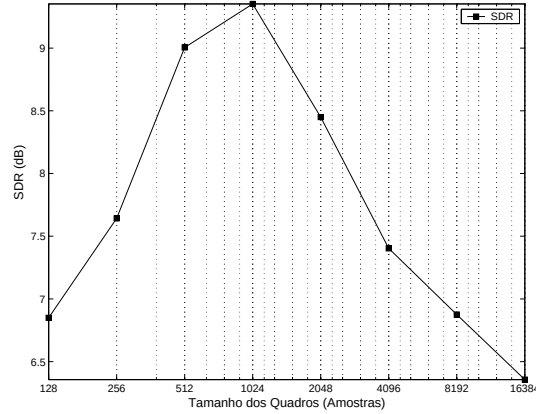


Figura 3.8: Relação sinal-distorção média ($\overline{SDR}$) obtida após a aplicação de máscaras ideais, para diferentes tamanhos de quadros (com sobreposição de quadros equivalente a 50% do tamanho dos mesmos), sobre uma mistura (soma de dois sinais de voz).

Para avaliação do tamanho dos quadros foi utilizada a seguinte matriz de mistura instantânea:

$$A = \begin{bmatrix} 0,9 & 0,8 \\ 0,6 & 0,7 \end{bmatrix}. \quad (3.13)$$

A separação da mistura instantânea foi realizada através do mascaramento no domínio tempo-freqüência, sendo utilizada apenas a informação de amplitude relativa para a estimação das máscaras ideais, ao invés de utilizar as informações de amplitude e fase relativas para a formação de clusters, como acontece no algoritmo *DUET* (Yilmaz e Rickard, 2004a). Na Figura 3.9 verifica-se que a relação sinal-distorção apresenta valor máximo de SDR média em torno de 8,7 dB quando utilizamos também um quadro com 1024 amostras.

3.9 Análise das máscaras ideais

O objetivo desta seção é analisar a separação dos sinais fonte utilizando uma máscara ideal e a separação obtida utilizando apenas informação da amplitude relativa do sinal,

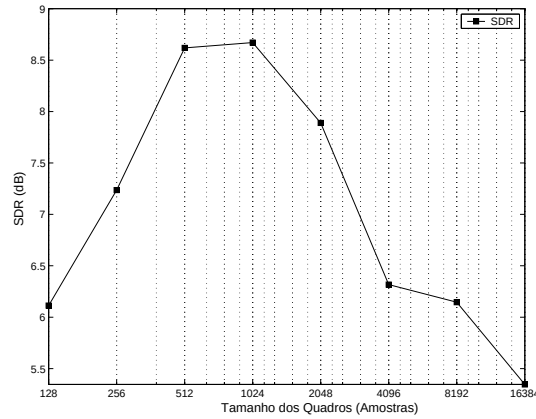


Figura 3.9: Relação sinal-distorção média ($\overline{SDR}$) obtida após a recuperação dos sinais fonte utilizando apenas a informação de amplitude relativa, ao invés de utilizar as informações de amplitude e atraso relativos para a formação de clusters, como acontece no algoritmo *DUET* (Yilmaz e Rickard, 2004a), para diferentes tamanhos de quadros sobre uma mistura instantânea de dois sinais de voz.

não considerando a informação de fase (atraso relativo). A matriz de mistura utilizada é novamente a mostrada na Equação 3.13.

A Figura 3.10 mostra um exemplo de máscaras ideais binárias referentes a dois sinais de voz. Para a obtenção das máscaras, os sinais de voz foram divididos em quadros de tamanho igual a 1024 amostras com sobreposição de 512 amostras. Para determinar a transformada de Fourier foram utilizados 1024 pontos.

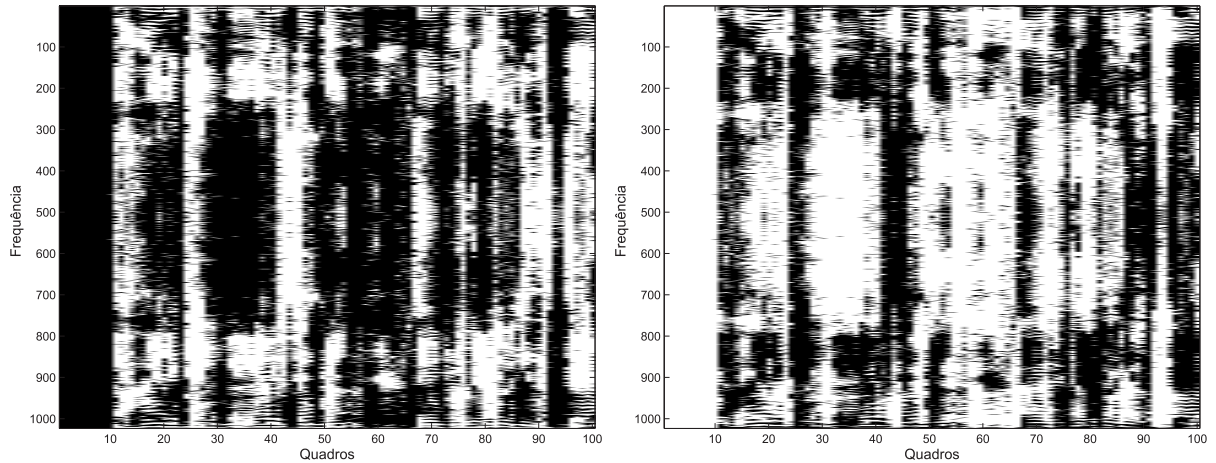


Figura 3.10: Imagens contendo as máscaras ideais binárias para as fontes s_1 e s_2 para os 100 primeiros quadros dos sinais de voz analisados.

Na Figura 3.11 observa-se a formação de dois clusters, sendo que os valores de atenuação relativa das fontes s_1 e s_2 estão próximos aos valores correspondentes aos picos $-0,83$ e $-0,28$. Para a obtenção das máscaras, primeiramente determina-se os valores dos picos e, posteriormente, foi especificado o ponto crítico $(-0,46)$ entre os dois picos para a separação dos sinais de acordo com a variância de cada um dos picos mencionados:

$$Pc = \mu_1(\sigma_2^2/(\sigma_1^2 + \sigma_2^2)) + \mu_2(\sigma_1^2/(\sigma_1^2 + \sigma_2^2)); \quad (3.14)$$

onde Pc representa o ponto crítico, μ_i representa o valor correspondente a média referente ao pico i e σ_i^2 representa a variância referente ao pico i .

A partir deste ponto foi realizada uma varredura e obtida a relação sinal-distorção (SDR) para cada um dos intervalos analisados. O primeiro intervalo analisado considera que uma das fontes está ativa quando os valores de amplitude são maiores que $-0,46$ e que a outra fonte está ativa quando os valores de amplitude são menores que $-0,46$. A partir deste ponto, os intervalos foram sendo reduzidos (com passo de $0,01$) até atingir o limite dos valores de máximo e mínimo de amplitude relativa. De acordo com a Figura 3.12 pode-se verificar que a recuperação dos sinais apresentou bom desempenho quando utilizamos o ponto crítico para limite de separação, ou seja, ao considerar que uma das fontes está ativa para valores de amplitude menores que $-0,46$ e que a outra fonte está ativa quando os valores de amplitude são maiores que $-0,46$.

Através da Tabela 3.1 observa-se um desempenho inferior do método proposto ($\overline{SDR} = 11 \text{ dB}$) ao recuperar as fontes s_1 e s_2 em comparação com o desempenho quando aplicamos as máscaras ideais ($\overline{SDR} = 12 \text{ dB}$). Porém, os sinais recuperados pelo método proposto apresentam boa qualidade e inteligibilidade. Vale lembrar que não é possível aplicar a máscara ideal quando não se conhece os sinais separados. O método proposto (utilizando apenas a amplitude relativa) atua sobre os sinais misturados e apresenta um desempenho inferior de aproximadamente 1 dB em relação ao obtido com a máscara ideal.

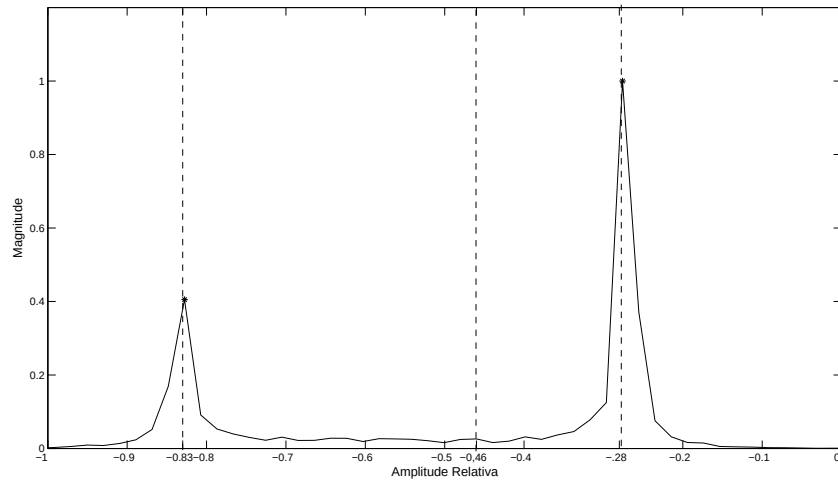
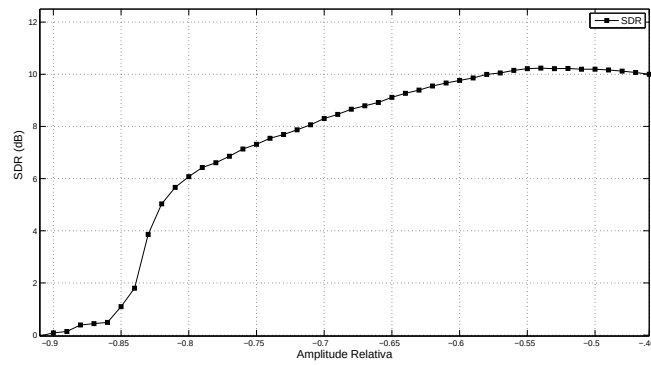
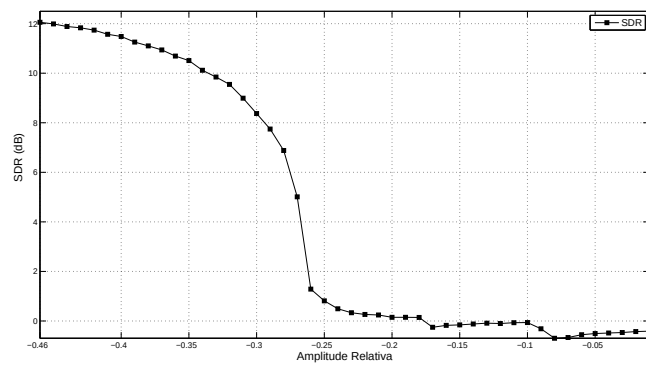


Figura 3.11: Histograma dos valores de atenuação relativa dos sinais fonte para misturas instantâneas considerando apenas valores diferentes de atenuações entre as fontes e sensores.



(a)



(b)

Figura 3.12: Medidas de distorção (SDR) do sinal recuperado para as fontes s_1 e s_2 considerando vários intervalos a partir do ponto crítico (-0,46) entre os dois picos utilizando passos de 0,01 entre os intervalos. (a) Medidas de distorção obtidas para a fonte s_1 a partir da redução dos intervalos considerados. (b) Medidas de distorção obtidas para a fonte s_2 a partir da redução dos intervalos considerados.

Tabela 3.1: Medidas de distorção dos sinais recuperados para misturas instantâneas obtidas através da matriz mostrada na Equação 3.13.

<i>Método</i>		$\hat{s}_1$	$\hat{s}_2$	Média
Máscara Ideal	SDR (dB)	12	12	12
Amplitude Relativa	SDR (dB)	12	10	11

3.10 Separação de misturas instantâneas considerando apenas amplitudes diferentes entre as fontes e sensores

O próximo passo é analisar a separação de fontes de uma mistura instantânea de dois sinais de voz utilizando diferentes algoritmos de *BSS/ICA*. Para cada um dos dois sinais

utilizados foi obtida uma máscara binária ideal. O objetivo é analisar a separação de sinais de voz utilizando o algoritmo *DUET* e o método proposto que utiliza apenas informação da amplitude relativa do sinal com técnicas clássicas de *ICA*: *Infomax*, *Fastica*, *Jade* e um algoritmo de *ICA* no domínio da frequência proposto por Ikeda e Murata (1999) (utilizando uma janela Hanning com 1024 amostras e sobreposição de 512 amostras). A Tabela 3.2 mostra os valores de *SDR* obtidos para a mistura em questão.

A matriz de mistura utilizada foi:

$$A = \begin{bmatrix} 0,9 & 0,8 \\ 0,7 & 0,9 \end{bmatrix}. \quad (3.15)$$

Tabela 3.2: Medidas de distorção dos sinais recuperados para misturas instantâneas obtidas através da matriz mostrada na Equação 3.15.

<i>Método</i>		$\hat{s}_1$	$\hat{s}_2$	Média
Máscara Ideal	<i>SDR (dB)</i>	13	15	14
Amplitude Relativa	<i>SDR (dB)</i>	11	11	11
<i>DUET</i>	<i>SDR (dB)</i>	11	11	11
<i>ICA</i> (Domínio da Frequência)	<i>SDR (dB)</i>	2	6	4
Infomax	<i>SDR (dB)</i>	45	32	39
<i>FastICA</i>	<i>SDR (dB)</i>	45	46	46
<i>Jade</i>	<i>SDR (dB)</i>	48	47	48

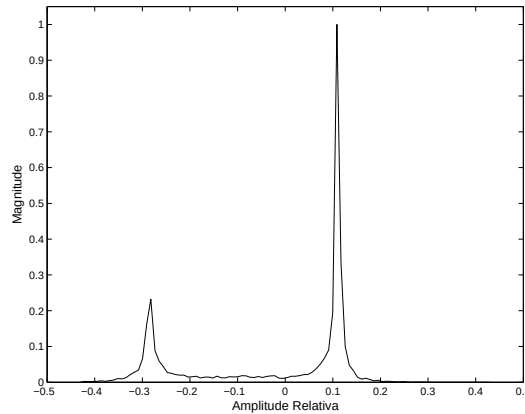


Figura 3.13: Histograma dos valores de atenuação relativa dos sinais fonte para misturas instantâneas considerando apenas valores diferentes de atenuações entre as fontes e sensores.

Para o caso de misturas instantâneas, de acordo com a Tabela 3.2, pode-se observar que os algoritmos *Infomax*, *FastICA* e *Jade* apresentam resultados muito melhores de separação (acima de 39 *dB*). Outro ponto importante a ser analisado é o fato dos sinais

recuperados através de *ICA* não apresentam quase nenhum artefato ou distorção. Ao utilizar os algoritmos baseados no domínio tempo-frequência nota-se um aumento significativo de distorções nos sinais separados, embora o sinal de voz continue intelegível e com boa qualidade sonora. Pode-se verificar na Figura 3.13 que os valores de atenuação relativa da fonte s_1 estão próximos ao valor $-0,28$ enquanto que os valores da fonte s_2 estão próximos ao valor $0,11$. Na obtenção da máscara para a recuperação dos sinais, considera-se que s_1 está ativa quando os valores de amplitude são menores que o ponto crítico entre os dois picos $0,09$ e s_2 está ativo quando os valores de amplitude são maiores que $0,09$. A separação dos sinais através da amplitude relativa, apresenta um desempenho médio igual ao obtido pelo algoritmo *DUET* (i.e. separação através da amplitude e do atraso relativos) em termos da relação sinal-distorção ($\overline{SDR} = 11 \text{ dB}$). Assim, a informação referente ao atraso relativo não influencia na busca pela máscara ideal. Para fins de referência, ao aplicar a máscara ideal obtém-se uma relação sinal-distorção média de 14 dB . Os algoritmos *Infomax*, *FastICA* e *Jade* foram desenvolvidos para lidarem com misturas instantâneas. No caso de misturas convoluídas o desempenho destes algoritmos reduz significativamente. Como o objetivo deste trabalho é verificar as limitações de técnicas de mascaramento espectral para o caso de misturas convoluídas, nas próximas seções será avaliado o desempenho do *DUET* e do método simplificado proposto denominado amplitude relativa.

3.11 Mistura dos sinais fonte considerando amplitudes (ganhos) e atrasos diferentes entre as fontes e sensores

Seguindo adiante na análise, além de valores de atenuação diferentes, foram inseridos atrasos nas misturas dos sinais fonte.

$$X_1 = 0,9s_1(t) + 0,5s_1(t-11) + 0,2s_1(t-21) + 0,8s_2(t) + 0,4s_2(t-11) + 0,3s_2(t-21), \quad (3.16)$$

$$X_2 = 0,7s_1(t) + 0,4s_1(t-11) + 0,3s_1(t-21) + 0,9s_2(t) + 0,4s_2(t-11) + 0,3s_2(t-21). \quad (3.17)$$

Na Figura 3.14 verifica-se ainda a presença de clusters em torno dos valores de amplitude relativa referentes às fontes s_1 e s_2 . O método proposto (amplitude relativa) apresenta um desempenho ligeiramente melhor que o algoritmo *DUET*, com uma relação

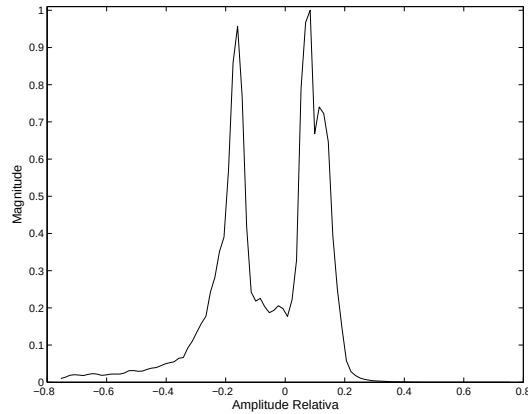


Figura 3.14: Histograma dos valores de atenuação relativa dos sinais fonte para misturas considerando atenuações e atrasos diferentes entre as fontes e sensores.

Tabela 3.3: Medidas de distorção obtidas para misturas com ganhos e atrasos diferentes entre as fontes e sensores.

<i>Método</i>		$\hat{s}_1$	$\hat{s}_2$	Média
Amplitude Relativa	<i>SDR (dB)</i>	4	6	5
<i>DUET</i>	<i>SDR (dB)</i>	2	3	3

signal-distorção média em torno de 5 *dB*, contra 3 *dB* atingidos pelo *DUET*, conforme mostrado na Tabela 3.3.

3.12 Misturas dos sinais fonte através da simulação de respostas ao impulso do ambiente

Uma vez analisados sinais com poucos atrasos, o passo seguinte é analisar a separação dos sinais fonte para ambientes reverberantes. A reverberação e as características de absorção do sinal sonoro pelo ambiente presentes nos sinais gravados pelos microfones podem ser simuladas através da convolução da resposta ao impulso finita (*FIR*) do modelo usado para o ambiente (em relação a cada fonte e sensor) e os sinais originais das fontes. A modelagem das respostas ao impulso do ambiente foi obtida através do método baseado nas imagens das fontes (Rindel, 2000; Savioja, 2000). As respostas ao impulso do ambiente simuladas em relação a cada microfone e fonte são mostradas na Figura 3.15.

Nos testes realizados, nenhum dos dois algoritmos (amplitude relativa e *DUET*) conseguiu uma boa separação dos sinais fonte, atingindo relações sinal-distorção da ordem de $SDR = -2$ *dB*, conforme mostrado na Tabela 3.4. Na Figura 3.16 pode-se perceber que a separação dos picos não é nítida. Assim, o uso da informação de amplitude relativa,

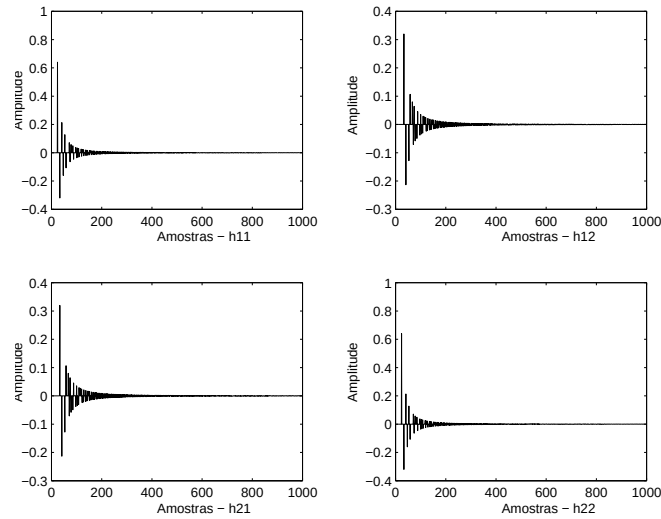


Figura 3.15: Respostas ao impulso simuladas através do método de imagens da fonte. A figura h_{11} representa a resposta ao impulso da fonte 1 em relação à posição do microfone 1. A figura h_{12} representa a resposta ao impulso da fonte 2 em relação à posição do microfone 1. A figura h_{21} representa a resposta ao impulso da fonte 1 em relação à posição do microfone 2. A figura h_{22} representa a resposta ao impulso da fonte 2 em relação à posição do microfone 2.

Tabela 3.4: Medidas de distorção obtidas para misturas dos sinais fonte através da simulação de respostas ao impulso do ambiente.

<i>Método</i>		$\hat{s}_1$	$\hat{s}_2$	Média
Amplitude Relativa	<i>SDR (dB)</i>	-2	0	-1
<i>DUET</i>	<i>SDR (dB)</i>	-2	-1	-2

somada ou não à informação de atraso relativo, não permite a separação de sinais em ambientes reverberantes, mesmo para tempos de reverberação curtos, da ordem de dezenas de milisegundos.

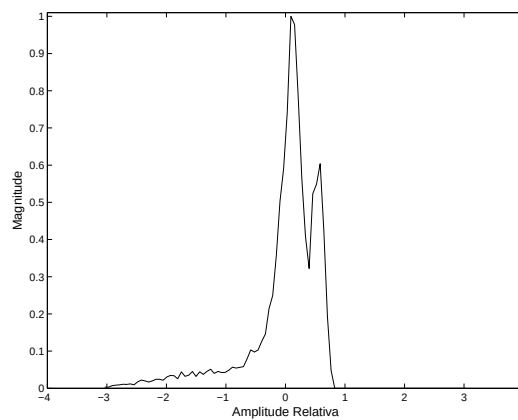


Figura 3.16: Histograma dos valores de atenuação dos sinais fonte para misturas obtidas através da simulação de respostas ao impulso do ambiente.

3.13 Gravações em campo aberto (ambiente com pouca reverberação)

Após a análise através de simulações, foram realizados testes utilizando gravações reais. As gravações foram obtidas em campo aberto ¹, com baixo nível de ruído e com mínima reverberação, conforme ilustrado na Figura 3.17.

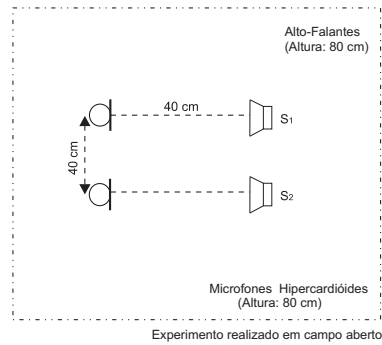


Figura 3.17: Gravações realizadas em campo aberto com pouca reverberação.

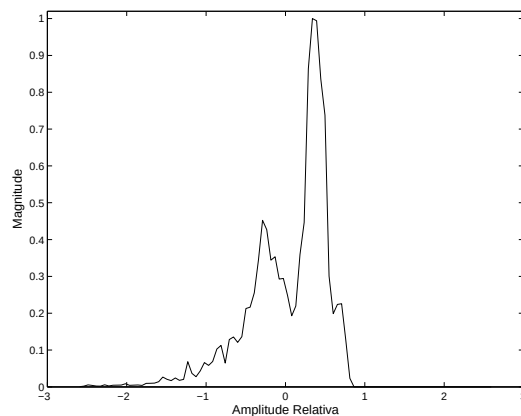


Figura 3.18: Histograma dos valores de atenuação dos sinais fonte para misturas gravadas em campo aberto (sem reverberação).

Tabela 3.5: Medidas de distorção obtidas em um ambiente real com pouca reverberação.

<i>Método</i>		$\hat{s}_1$	$\hat{s}_2$	Média
Amplitude Relativa	<i>SDR (dB)</i>	-4	-1	-3
<i>DUET</i>	<i>SDR (dB)</i>	-4	-2	-3

¹Gravações realizadas no centro de um campo de futebol gramado, em um domingo, à 6 horas da manhã.

Pode-se perceber na Figura 3.18 a formação de clusters referentes às fontes s_1 e s_2 . Porém, a Tabela 3.5 mostra que não é possível separar as fontes através do uso da amplitude relativa, somado ou não a informações de atraso relativo (*DUET*).

3.14 Gravações em ambiente com reverberação

Finalmente, os sinais foram gravados em um ambiente real com reverberação. As posições das fontes e sensores, bem como as dimensões do ambiente estão descritas na Figura 3.19.

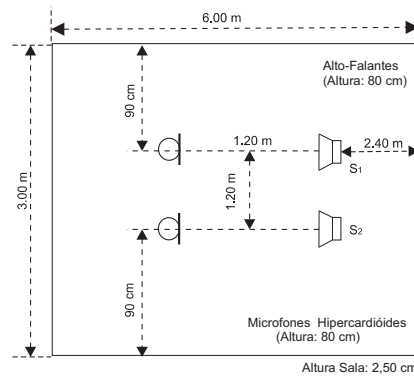


Figura 3.19: Gravações realizadas em sala com reverberação.

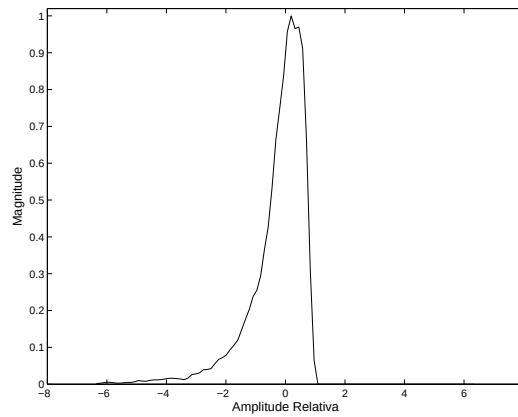


Figura 3.20: Histograma dos valores de atenuação dos sinais fonte para misturas gravadas em um ambiente com reverberação.

Na presença de reverberação não é possível nem mesmo definir claramente os picos relativos às atenuações das fontes s_1 e s_2 , conforme mostrado na Figura 3.20. Assim, no caso de sinais reverberados também não é possível a separação dos sinais com base na amplitude relativa, somada ou não a informações de atraso relativo.

Tabela 3.6: Medidas de distorção obtidas em um ambiente com reverberação.

<i>Método</i>		$\hat{s}_1$	$\hat{s}_2$	Média
Amplitude Relativa	<i>SDR (dB)</i>	-5	-2	-4
<i>DUET</i>	<i>SDR (dB)</i>	-5	-2	-4

3.15 Considerações finais

Neste capítulo é feita uma abordagem sobre o mascaramento espectral de sinais de voz no domínio do tempo-frequência. Inicialmente, os tamanhos dos quadros utilizados para a separação de fontes via mascaramento espectral são analisados. A seguir apresenta-se o algoritmo *DUET*, o qual é um método de estimação cega de máscaras binárias, a partir de duas misturas de sinais de voz, para a separação de fontes. O método foi desenvolvido para atuar em casos sub-determinados e em misturas contendo apenas atrasos, ou seja, sem reverberação. O algoritmo é capaz de separar misturas anecóicas, considerando apenas atrasos presentes na mistura, mas não funciona no caso de misturas reverberadas. Para a estimação das máscaras ideais binárias o *DUET* utiliza dois parâmetros chamados de amplitude relativa e atraso relativo. Um método baseado apenas na informação de amplitude relativa, descartando a informação referente ao parâmetro denominado de atraso relativo, foi implementado com desempenho equivalente ao *DUET*. Ao final, conclui-se que métodos baseados na informação de amplitude relativa, com ou sem adição de informação de atraso relativo, não são capazes de separar misturas realizadas em ambiente reverberante. O objetivo do próximo capítulo é determinar até que ponto é possível separar tais misturas quando se conhece a máscara ideal.

Capítulo 4

Limitações do mascaramento espectral

Quando o problema de separação de fontes é sobre-determinado (existem mais misturas que fontes) considerações genéricas tais como independência das fontes podem ser utilizadas para a tarefa de separação. Entretanto, muitos casos são sub-determinados, ou seja, existem mais fontes que misturas. Neste caso, considerações mais específicas devem ser assumidas para a separação das fontes. No caso de problemas envolvendo pelo menos dois sensores (duas misturas) progressos têm sido alcançados utilizando considerações a respeito da esparsidade das fontes através do mascaramento no domínio tempo e frequência (Bach e Jordan, 2004; Jourjine et al., 2000). Porém, considerando o caso de misturas convoluídas, o desempenho destes algoritmos ainda tem muito para ser melhorado. Visando verificar a viabilidade e as limitações dos métodos baseados em mascaramento espectral para a separação cega de fontes convoluídas, uma série de análises é discutida neste capítulo.

4.1 Análise da máscara binária ideal para separação de fontes no domínio tempo-frequência

Ao analisar a máscara ideal (obtida para dois sinais de voz, utilizando um quadro com 512 amostras e frequência de amostragem igual a 10000 Hz), pode-se observar através da distribuição dos valores iguais a um, obtidos para cada quadro das máscaras binárias ideais dos sinais de voz, que não existe uma quantidade específica ou predominante de "uns ou zeros", conforme mostra a Figura 4.1.

Para verificar a performance da máscara ideal na separação de dois sinais fonte (somados artificialmente) foram realizados vários testes a partir da introdução de erros na máscara ideal. Os erros foram introduzidos de forma aleatória, sendo aumentados gradativamente ("bit a bit"). Antes de cada acréscimo de erro, foram geradas 1000 máscaras

diferentes. No gráfico da Figura 4.2, pode-se verificar que à medida que se afasta da máscara ideal (ou à medida que se aumenta a quantidade de erros) a relação sinal-distorção diminui. Ao medir a queda de desempenho do processo de separação em função da distância entre a máscara ideal e a máscara efetivamente usada, observa-se uma queda de desempenho de 3 *dB* quando aproximadamente 10% dos bits da máscara ideal estão invertidos.

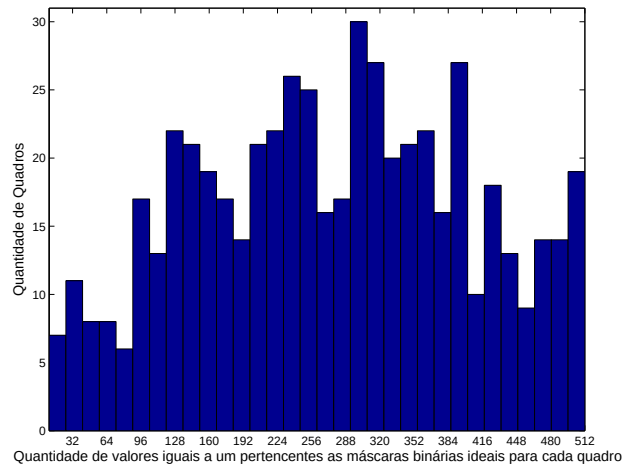


Figura 4.1: Histograma referente a quantidade de valores iguais a um para cada quadro pertencente a máscara ideal de dois sinais de voz (obtida para dois sinais de voz, utilizando um quadro com 512 amostras e frequência de amostragem igual a 10000 *Hz*).

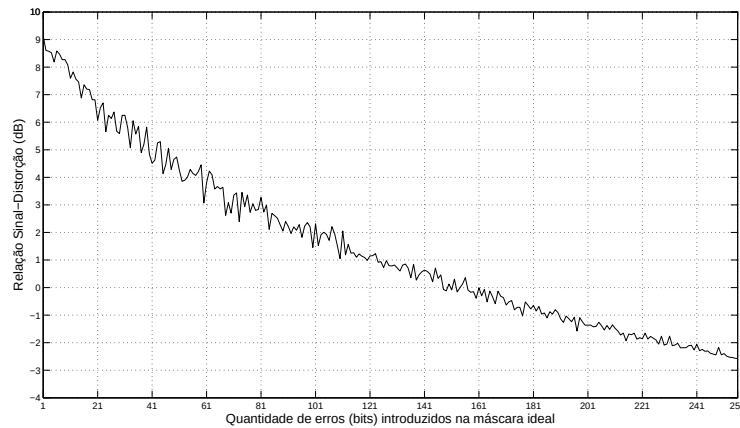


Figura 4.2: Valores médios das relações sinal-distorção do sinal fonte obtidos após introdução gradativa e aleatória de erros na máscara ideal aplicada sobre uma mistura de dois sinais de voz (somados artificialmente). Para obtenção das máscaras ideais utilizou-se quadros com 512 amostras, desta forma a quantidade de erro máxima corresponde a 256. Pode-se notar que à medida que aumenta-se a quantidade de erros a relação sinal-distorção diminui.

4.2 Limitações de métodos de mascaramento espectral para sinais reverberados

Para o estudo das implicações causadas pelo aumento do tempo de reverberação no processo de separação de fontes via mascaramento no domínio tempo-frequência foram simuladas e obtidas diversas respostas ao impulso do ambiente com vários tempos de reverberação. A mistura dos sinais foi obtida através da seguinte expressão:

$$x(t) = s_1(t) * h_{11}(t) + s_2(t) * h_{12}(t); \quad (4.1)$$

onde o símbolo "*" representa a operação de convolução, $s_1(t)$ e $s_2(t)$ representam os sinais fonte e $h_{11}(t)$ e $h_{12}(t)$ representam, respectivamente, a resposta ao impulso simulada para a fonte (s_1) em relação ao microfone e para a fonte (s_2) em relação ao mesmo microfone.

Wallace Clement Sabine, por volta de 1900, introduziu pela primeira vez a definição do tempo de duração da reverberação (T_{60}), tendo desenvolvido a Equação 4.2, onde o tempo de reverberação é calculado em função do volume do ambiente, da superfície da área das paredes e dos respectivos coeficientes de absorção das paredes:

$$RT_{60} = 0.161 \left(\frac{V}{A} \right), \quad (4.2)$$

onde V corresponde ao volume do ambiente (m^3) e A representa uma janela de absorção imaginária (m^2) que é calculada através da soma das áreas das paredes multiplicadas pelos respectivos coeficientes de absorção. O tempo de reverberação pode ser definido como sendo o tempo necessário para o nível de pressão do som diminuir 60 *dB* devido a atenuação provocada pela absorção da energia sonora em cada reflexão ocorrida nas paredes do ambiente. Desta forma, pode-se obter tempos de reverberação diferentes dependendo das dimensões do ambiente e dos coeficientes de absorção.

O tempo de reverberação é dado em função do volume do ambiente e da absorção das paredes. Espectrogramas obtidos em ambientes diferentes e com mesmo tempo de reverberação podem ser significativamente diferentes. Por exemplo, pode-se observar que em um ambiente de grande volume e elevada absorção nas paredes pode ter seu tempo de reverberação similar a um ambiente de pequeno volume e baixa absorção nas paredes. Porém, no segundo caso, o espectrograma tende a ser mais "denso". A Figura 4.3 mostra as respostas ao impulso simuladas obtidas para o caso de um ambiente com grande volume e elevada absorção das paredes e para o caso de um ambiente com pequeno volume e baixa absorção das paredes. Entretanto, análises preliminares feitas para os casos acima não apresentaram variação de desempenho de separação através das máscaras ideais.

A Figura 4.4 mostra as dimensões do ambiente considerado para obtenção das respos-

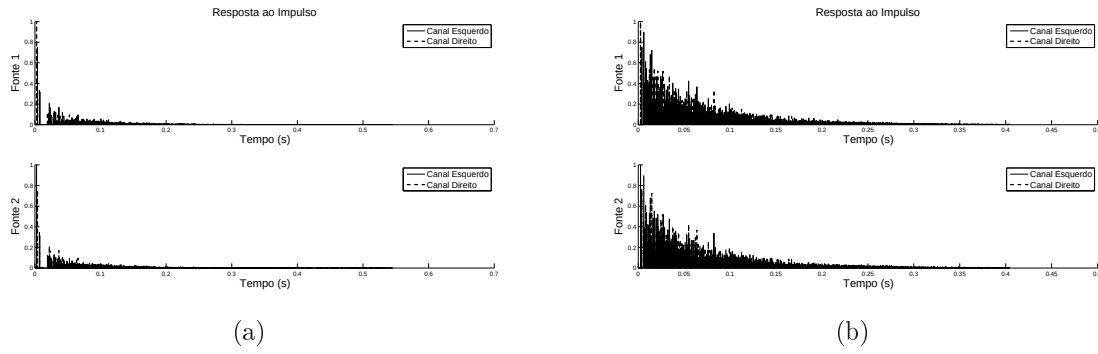


Figura 4.3: Respostas ao impulso simuladas de ambiente com volumes diferentes, porém com mesmo tempo de reverberação ($T_{60} = 600 \text{ ms}$). (a) Resposta ao impulso simulada de um ambiente com as seguintes dimensões $12,5 \text{ m} \times 7,5 \text{ m} \times 5,0 \text{ m}$ e com alta absorção das paredes (coeficiente de absorção igual a 0,4). (b) Resposta ao impulso simulada de um ambiente com as seguintes dimensões $3,0 \text{ m} \times 2,0 \text{ m} \times 2,0 \text{ m}$ e com baixa absorção das paredes (coeficiente de absorção igual a 0,1).

tas ao impulso simuladas, bem como as posições das fontes e sensores. Foi considerado um ambiente com as seguintes dimensões: $6,5 \text{ m}$ de comprimento, $3,25 \text{ m}$ de largura e $2,5 \text{ m}$ de altura. Para obtenção das respostas ao impulso simuladas foi utilizado o software desenvolvido em MatLab por Campbell, Palomäki, e Brown (2007).

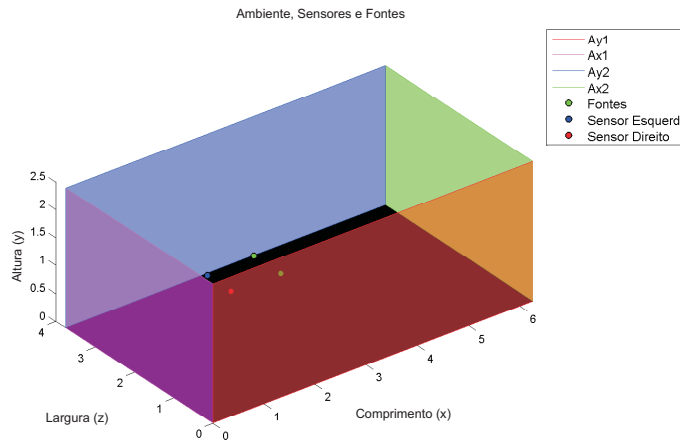


Figura 4.4: Ambiente considerado para obtenção das respostas ao impulso simuladas (Dimensões: $6,5 \text{ m} \times 3,25 \text{ m} \times 2,5 \text{ m}$).

De acordo com as dimensões do ambiente e posições das fontes e sensores mostradas na Figura 4.4, foram simuladas as seguintes respostas ao impulso e obtidos os respectivos espectros de frequência: resposta ao impulso obtida considerando o ambiente anecóico, reflexão em apenas uma parede, reflexão em duas paredes, reflexão em três paredes, reflexão em quatro paredes, reflexão em cinco paredes, reflexão em todas as paredes,

conforme mostrado nas figuras 4.5, 4.6, 4.7, 4.8, 4.9, 4.10 e 4.11 e reflexão em um ambiente com o dobro das dimensões.

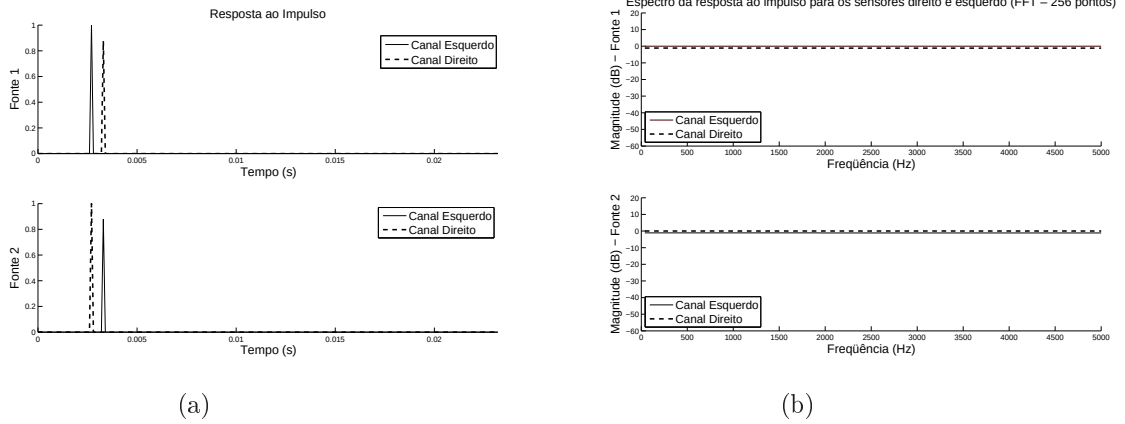


Figura 4.5: Respostas ao impulso simuladas considerando o ambiente anecóico. (a) Respostas ao impulso simuladas. (b) Espectros das respostas ao impulso simuladas.

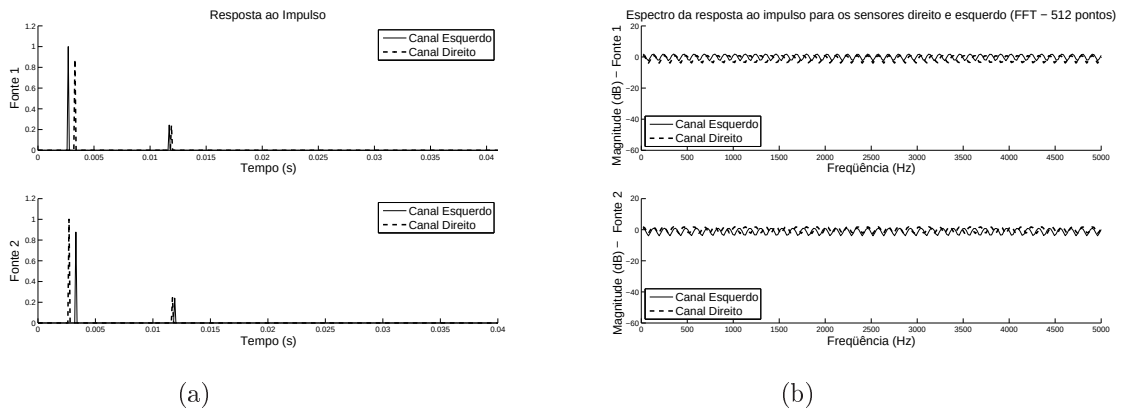


Figura 4.6: Respostas ao impulso simuladas considerando reflexão em apenas uma parede do ambiente com coeficiente de absorção igual a 0.01. (a) Respostas ao impulso simuladas. (b) Espectros das respostas ao impulso simuladas.

A relação sinal-distorção dos sinais recuperados ao aplicarmos uma máscara binária foi avaliada considerando vários tempos de reverberação, conforme mostra a Figura 4.12. Primeiramente considerou-se o ambiente anecóico e posteriormente o tempo de reverberação foi sendo aumentado (à medida que aumentava-se a quantidade de paredes com reflexão no ambiente simulado). O tempo de reverberação para cada um dos casos considerados é mostrado na Tabela 4.1.

A qualidade dos sinais recuperados foi avaliada para as seguintes situações: i) aplicação da máscara ideal direto sobre os sinais fonte; ii) aplicação da máscara ideal direto sobre os sinais misturados; iii) aplicação da máscara obtida para o caso do ambiente anecóico

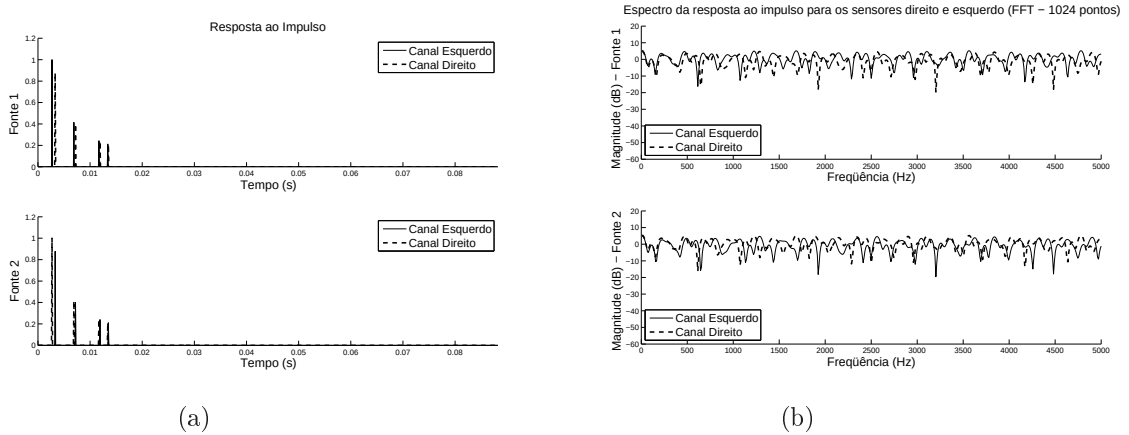


Figura 4.7: Respostas ao impulso simuladas considerando reflexão em apenas duas paredes do ambiente (Ax1 e Ax2) com coeficiente de absorção igual a 0.01. (a) Respostas ao impulso simuladas. (b) Espectros das respostas ao impulso simuladas.

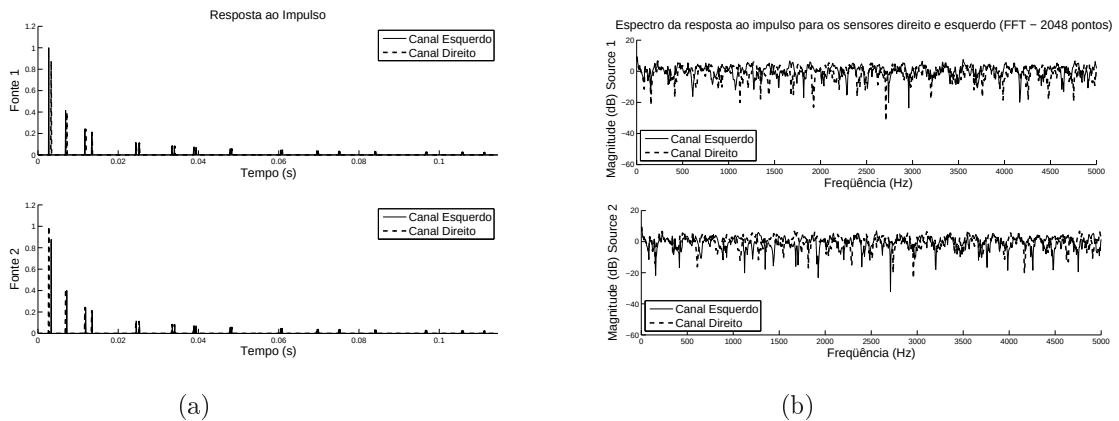


Figura 4.8: Respostas ao impulso simuladas considerando reflexão em apenas três paredes do ambiente (Ax1, Ax2 e Ay1) com coeficiente de absorção igual a 0.01. (a) Respostas ao impulso simuladas. (b) Espectros das respostas ao impulso simuladas.

direto sobre os sinais fonte e iv) aplicação da máscara obtida para o caso do ambiente anecóico direto sobre os sinais misturados.

Na primeira situação foi aplicada a máscara ideal diretamente sobre cada sinal fonte convoluído com a resposta ao impulso do ambiente em relação à posição da fonte e sensor. Pode-se verificar pela Figura 4.12 que ao obter-se a máscara para o caso do ambiente anecóico, a relação sinal-distorção média obtida foi de 12 dB, apresentando o sinal boa inteligibilidade. Aumentando o tempo de reverberação até 300 ms pode-se verificar que a relação sinal-distorção diminui muito pouco (1 dB). Ao analisar a qualidade do sinal recuperado para tempos de reverberação longo (3,9 s) pode-se concluir que, mesmo ocorrendo uma redução da relação sinal-distorção para 7 dB, o sinal recuperado ainda continua com boa inteligibilidade.

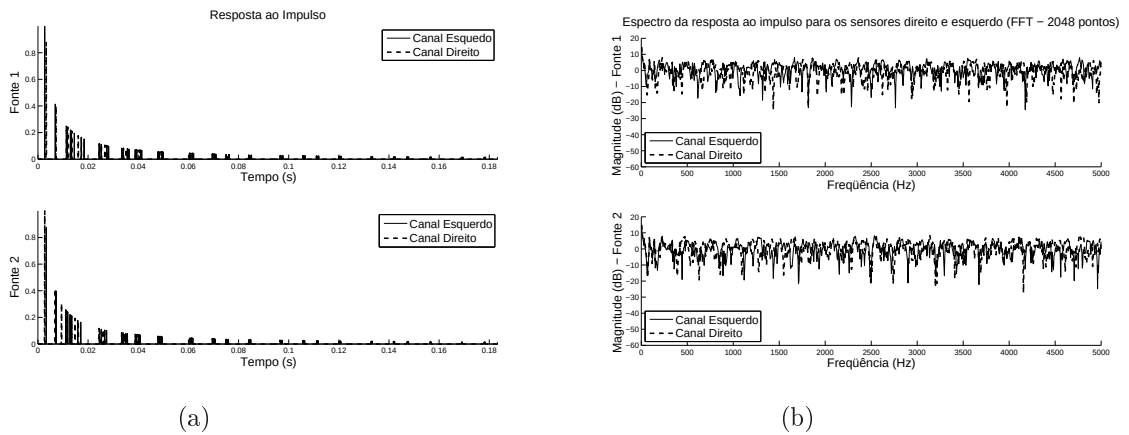


Figura 4.9: Respostas ao impulso simuladas considerando reflexão em apenas quatro paredes do ambiente (Ax1, Ax2, Ay1 e Ay2) com coeficiente de absorção igual a 0.01. (a) Respostas ao impulso simuladas. (b) Espectros das respostas ao impulso simuladas.

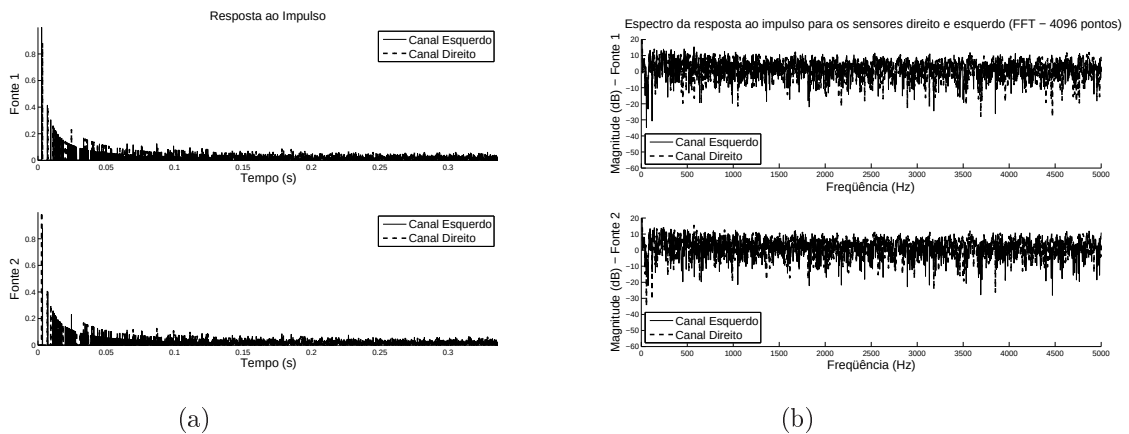


Figura 4.10: Respostas ao impulso simuladas considerando reflexão em cinco paredes do ambiente (Ax1, Ax2, Ay1, Ay2 e Az1) com coeficiente de absorção igual a 0.01. (a) Respostas ao impulso simuladas. (b) Espectros das respostas ao impulso simuladas.

No segundo caso analisado, ao aplicar a máscara ideal sobre os sinais misturados e considerando diversos tempos de reverberação pode-se observar que, para o caso do ambiente anecóico, a relação sinal-distorção média obtida para os sinais fonte foi de 9 *dB* e para tempo de reverberação menor que 300 *ms* esta medida caiu para 8 *dB*. Para tempo de reverberação igual a 3,9 segundos a relação sinal-distorção reduziu para 3 *dB*, mas mesmo com esta redução, os sinais recuperados ainda se mantiveram intelegíveis.

Na terceira situação a máscara binária obtida para o ambiente anecóico foi utilizada diretamente sobre cada sinal fonte convoluído com a resposta ao impulso do ambiente em relação à posição da fonte e sensor. Neste caso pode-se verificar que, até aproximadamente o tempo de 300 *ms*, a máscara ideal praticamente não sofre modificações, uma vez que os resultados obtidos foram os mesmos que os encontrados na primeira situação. Porém, para

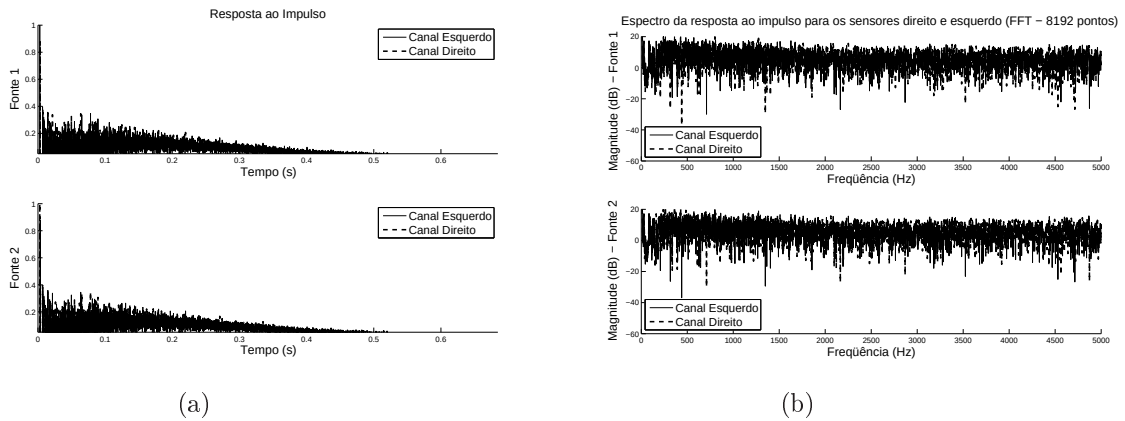


Figura 4.11: Respostas ao impulso obtidas considerando reflexão em todas as paredes, com baixo coeficiente de absorção das paredes (cerca de 0.08). (a) Respostas ao impulso simuladas. (b) Espectros das respostas ao impulso simuladas.

Reflexões	Tempo de Reverberação (ms)
Uma parede	33
Duas paredes	40
Três paredes	182
Quatro paredes	290
Cinco paredes	980
Seis paredes	1750
Dimensões dobradas	3900

Tabela 4.1: Variação do tempo de reverberação a medida que se aumenta a quantidade de paredes com reflexão no ambiente simulado.

tempos maiores que 300 ms , os resultados encontrados sofrem uma redução maior que quando comparados com os resultados encontrados no primeiro caso, ou seja, a máscara ideal não é mais a mesma que foi obtida para o caso do ambiente anecóico.

Na quarta situação a máscara binária obtida para o ambiente anecóico foi utilizada sobre os sinais misturados e considerado também diversos tempos de reverberação. Neste caso, pode-se verificar também que até aproximadamente o tempo de 300 ms a máscara ideal praticamente não sofre modificações, uma vez que os resultados obtidos foram os mesmos que os encontrados na segunda situação proposta. Porém, para tempos maiores que 300 ms os resultados encontrados também sofrem uma redução maior quando comparados com os resultados encontrados no segundo caso.

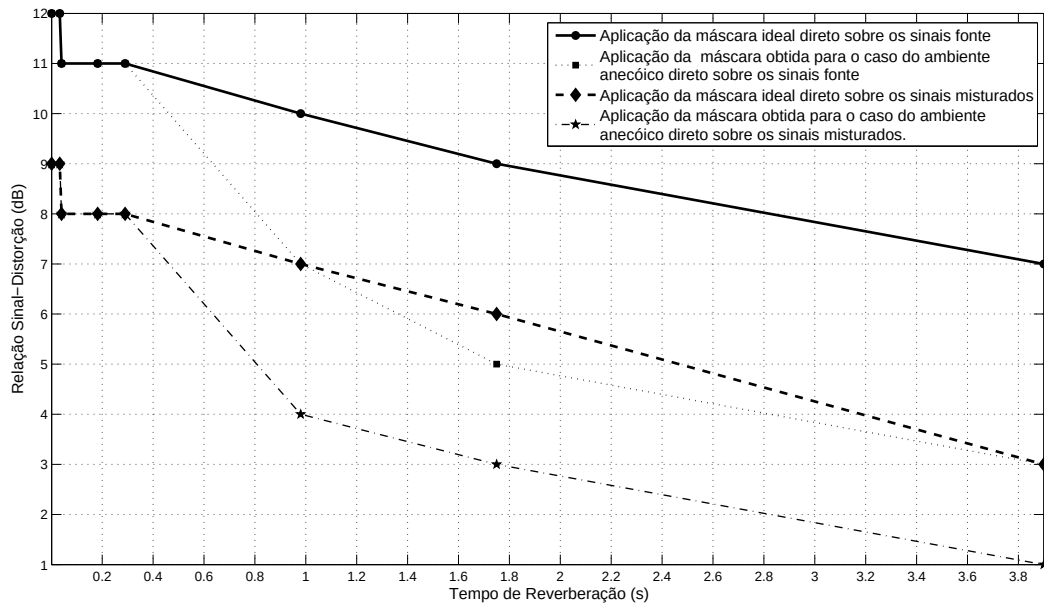


Figura 4.12: Relação sinal-distorção dos sinais recuperados ao aplicarmos uma máscara binária para tempos de reverberação variados. A qualidade dos sinais recuperados foi avaliada para as seguintes situações: i) aplicação da máscara ideal direto sobre os sinais fonte; ii) aplicação da máscara ideal direto sobre os sinais misturados; iii) aplicação da máscara obtida para o caso do ambiente anecóico direto sobre os sinais fonte e iv) aplicação da máscara obtida para o caso do ambiente anecóico direto sobre os sinais misturados.

4.3 Gravações em ambientes reais com reverberação

Para análise em condições reais do efeito da reverberação na utilização de máscaras ideais para a separação dos sinais fonte, foram feitas gravações em dois ambientes distintos, com tempos de reverberação iguais a aproximadamente 100 e 1100 *ms*, respectivamente. As posições das fontes e sensores, bem como as dimensões dos ambientes são mostradas na Figura 4.14. Os sinais fonte utilizados nas gravações foram gravados a uma taxa de amostragem de 44.100 *kHz* e depois foram reamostrados a uma taxa de 10.000 *kHz*, com tempo de duração de cada sinal em torno de 45 segundos. Foram utilizados os seguintes sinais durante as gravações: i) voz masculina (s1); ii) voz feminina (s2); iii) música - *trecho da 5ª sinfonia de Beethoven* (s3) e iv) ruído branco (s4) ¹. As máscaras ideais foram extraídas a partir das gravações realizadas separadamente para cada sinal fonte. As gravações foram realizadas da seguinte forma: primeiramente, foram gravados os sinais fonte separados (para permitir a extração das máscaras ideais) e, posteriormente, os sinais foram reproduzidos de forma a obter misturas 2x2 utilizando os sinais fonte.

A resposta ao impulso de um ambiente pode ser medida através da gravação de um

¹Ruído branco gerado aleatoriamente com uma distribuição uniforme e reproduzido através de alto-falantes de alta fidelidade.

signal (impulso) com curta duração e que tenha espectro de potência plano. A utilização de impulsos sonoros, tais como: explosão de pólvora ou bater de palmas não são muito confiáveis por não apresentarem um espectro plano em toda a faixa de frequência. Uma outra forma para se obter a resposta ao impulso de um ambiente é a utilização de um pulso estendido no tempo (*Time-Stretched Pulse - TSP*) (Aoshima, 1981). Este sinal possui um espectro de frequência plano. O sinal *TSP* utilizado (no domínio da frequência) é dado pela expressão abaixo (Suzuki, Asano, Kim, e Sone, 1995):

$$T(k) = \begin{cases} \exp(jpk^2), & 0 \leq k < N/2, \\ 1, & k = N/2, \\ T^*(N - k), & N/2 < k < N, \end{cases} \quad (4.3)$$

onde o símbolo p determina a extensão do sinal. A função de transferência inversa pode ser obtida através da expressão: $T^{-1}(k) = 1/T(k)$, conforme mostrado em (Suzuki et al., 1995). O sinal *TSP* utilizado nas gravações para obtenção da resposta ao impulso dos ambientes é mostrado na Figura 4.13.

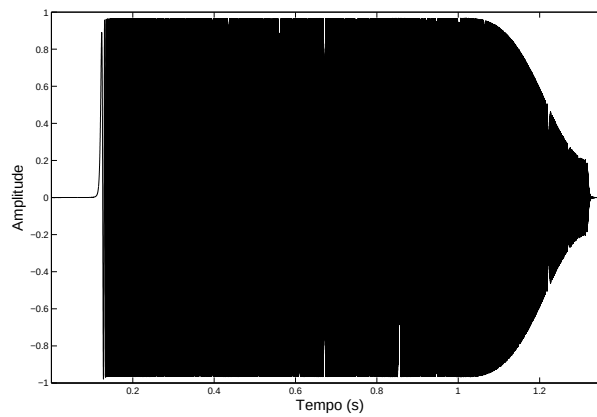


Figura 4.13: Pulso estendido no tempo (*TSP*) utilizado para as medições das respostas ao impulso dos ambientes.

A resposta ao impulso pode ser medida através dos seguintes passos. Primeiramente, calcula-se a *DFT* do sinal gravado após o toque do sinal de *TSP*. Posteriormente multiplica-se o resultado da *DFT* pelo sinal $T^{-1}(k)$. Os cálculos acima são baseados numa *DFT* de comprimento igual a N (Suzuki et al., 1995). Esta operação apresenta algumas deficiências, tais como: a função de transferência da caixa de som e do microfone utilizados estão sendo integrados juntamente com a função de transferência do ambiente, além da dificuldade em sincronizar o sinal original com a gravação do sinal reverberado. Porém, para os objetivos deste trabalho, podemos considerar as respostas ao impulso obtidas como sendo uma boa aproximação. As respostas ao impulso medidas para os dois ambientes mencionados são mostradas na Figura 4.15. O tempo de reverberação (T_{60}) foi

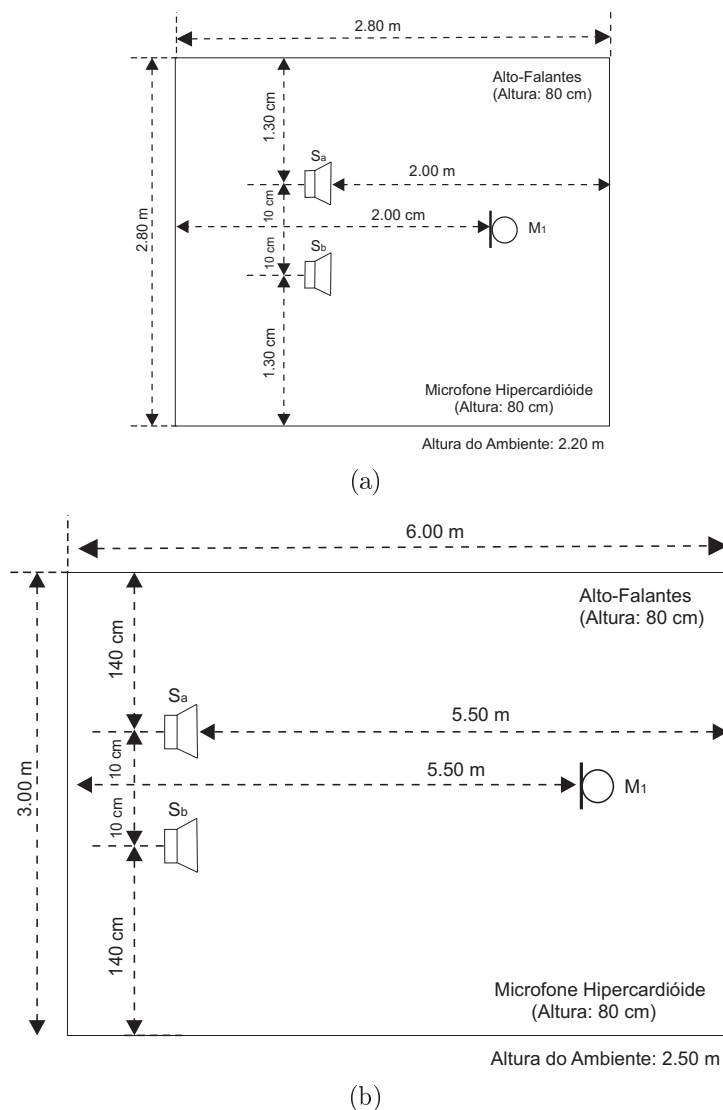


Figura 4.14: Ambientes onde foram realizadas as gravações dos sinais fonte. (a) Ambiente 1: estúdio vazio com pouca reverberação (as paredes são revestidas de carpete). (b) Ambiente 2: sala vazia com reverberação (paredes revestidas de cimento e tinta).

calculado de acordo com o método proposto por Schroeder (1965), através do decaimento da energia do sinal (resposta ao impulso) até o nível de -60 dB .

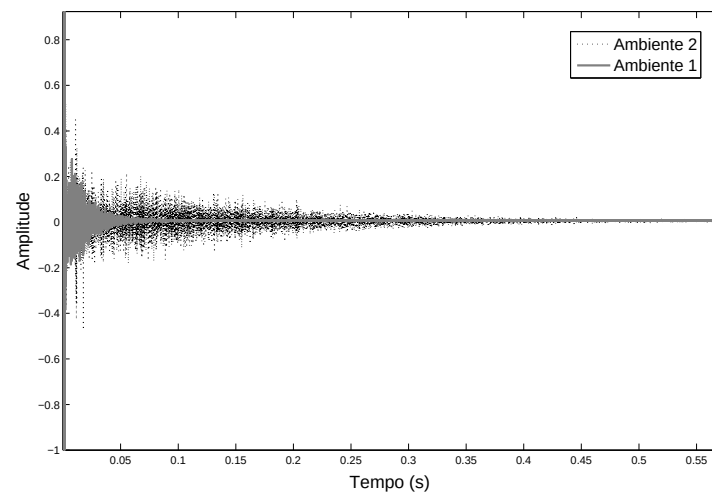


Figura 4.15: Figura comparativa contendo as duas respostas ao impulso obtidas para os ambientes (mencionados na Figura 4.14) com tempos de reverberação iguais a 100 ms (Ambiente 1) e 1100 ms (Ambiente 2).

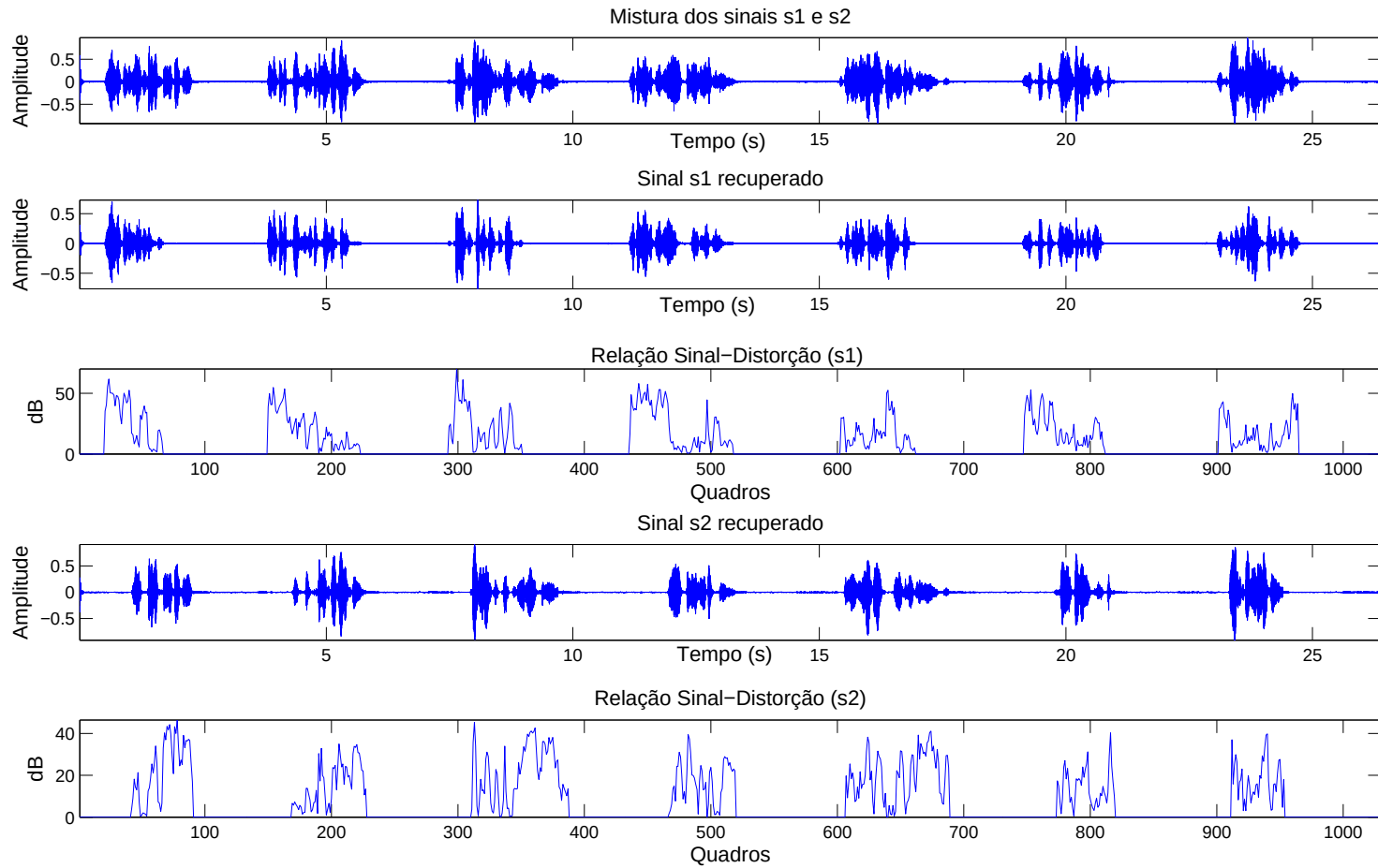


Figura 4.16: Gravações de dois sinais de voz $s1$ e $s2$ (uma voz masculina e uma voz feminina) realizadas no ambiente 1 com tempo de reverberação igual a 100 ms . De cima para baixo as figuras correspondem, respectivamente: i) mistura dos sinais de voz $s1$ e $s2$; ii) sinal $s1$ recuperado ao aplicarmos a máscara ideal sobre a mistura; iii) relação sinal-distorção do sinal $s1$ recuperado em relação ao sinal $s1$ gravado separadamente ($\overline{SDR} = 9\text{ dB}$); iv) sinal $s2$ recuperado ao aplicarmos a máscara ideal sobre a mistura; v) relação sinal-distorção do sinal $s2$ recuperado em relação ao sinal $s2$ gravado separadamente ($\overline{SDR} = 8\text{ dB}$).

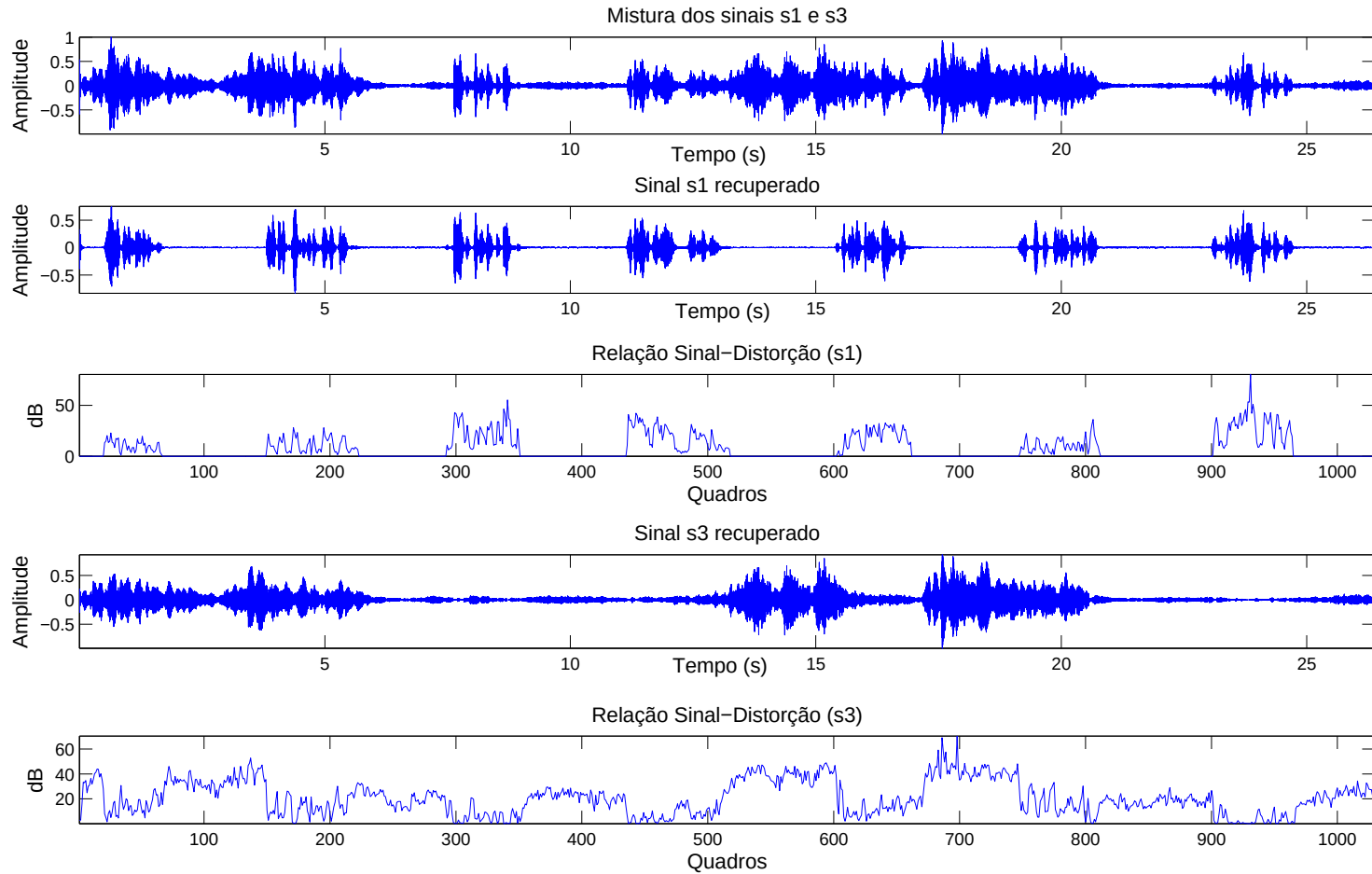


Figura 4.17: Gravações de dois sinais $s1$ (voz masculina) e $s3$ (música) realizadas no ambiente 1 com tempo de reverberação igual a 100 ms. De cima para baixo as figuras correspondem, respectivamente: i) mistura dos sinais de voz $s1$ e música $s3$; ii) sinal $s1$ recuperado ao aplicarmos a máscara ideal sobre a mistura; iii) relação sinal-distorção do sinal $s1$ recuperado em relação ao sinal $s1$ gravado separadamente ($\overline{SDR} = 8$ dB); iv) sinal $s3$ recuperado ao aplicarmos a máscara ideal sobre a mistura; v) relação sinal-distorção do sinal $s3$ recuperado em relação ao sinal $s3$ gravado separadamente ($\overline{SDR} = 21$ dB).

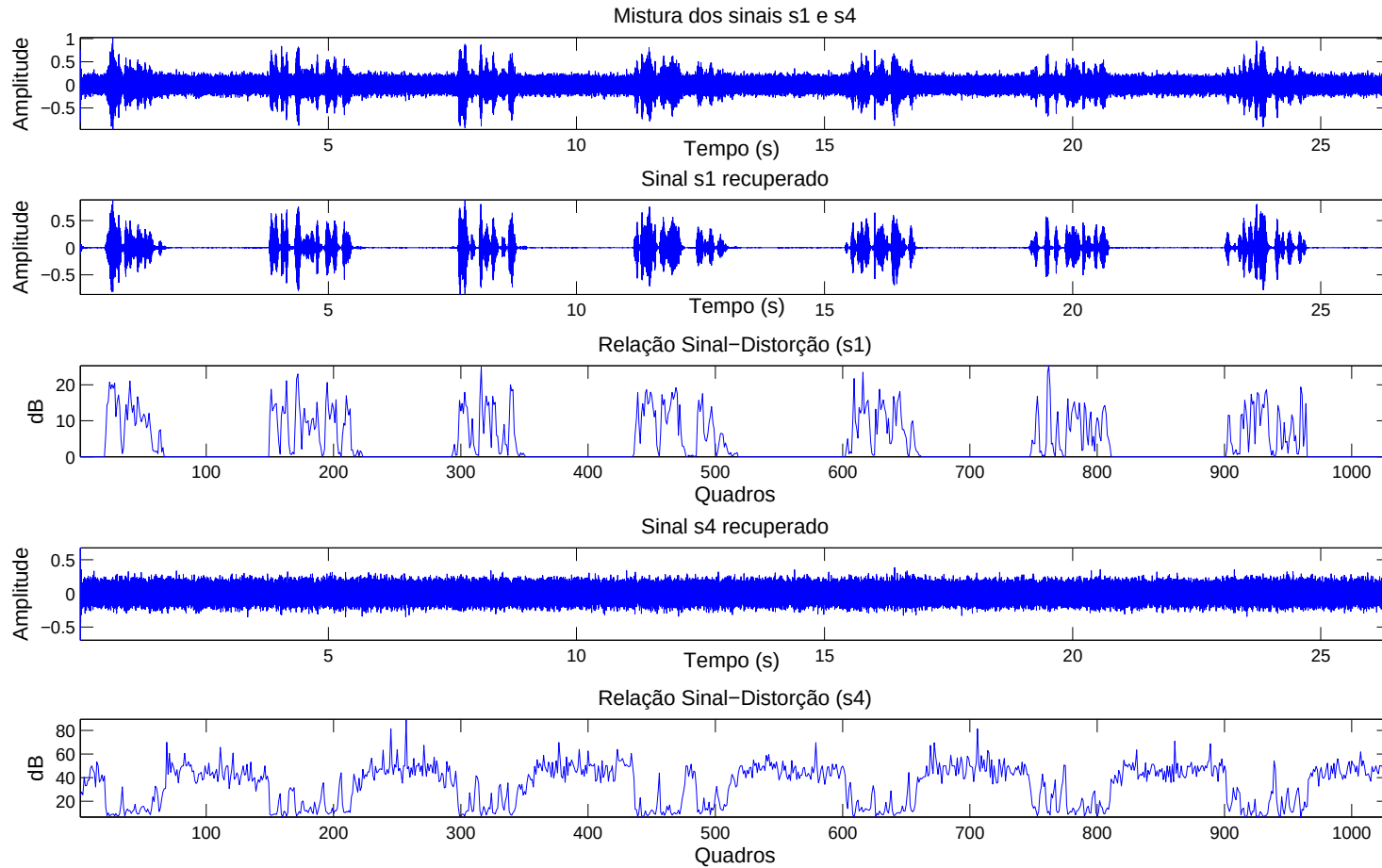


Figura 4.18: Gravações de dois sinais s_1 (voz masculina) e s_4 (ruído branco) realizadas no ambiente 1 com tempo de reverberação igual a 100 ms. De cima para baixo as figuras correspondem, respectivamente: i) mistura dos sinais de voz s_1 e o ruído branco s_4 ; ii) sinal s_1 recuperado ao aplicarmos a máscara ideal sobre a mistura; iii) relação sinal-distorção do sinal s_1 recuperado em relação ao sinal s_1 gravado separadamente ($\overline{SDR} = 4$ dB); iv) sinal s_4 recuperado ao aplicarmos a máscara ideal sobre a mistura; v) relação sinal-distorção do sinal s_4 recuperado em relação ao sinal s_4 gravado separadamente ($\overline{SDR} = 34$ dB).

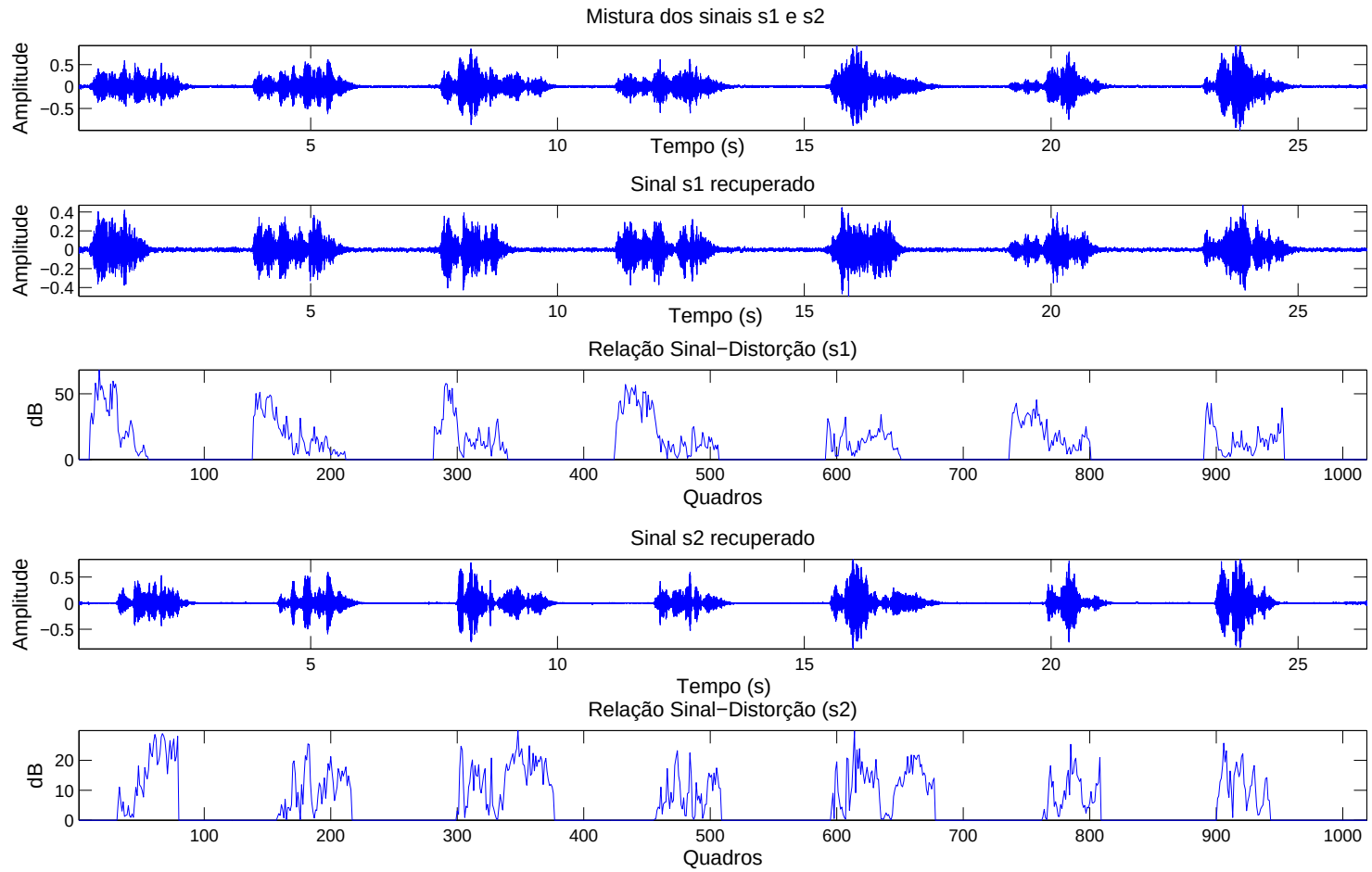


Figura 4.19: Gravações de dois sinais de voz s_1 e s_2 (uma voz masculina e uma voz feminina) realizadas no ambiente 2 com tempo de reverberação igual a 1100 ms. De cima para baixo as figuras correspondem, respectivamente: i) mistura dos sinais de voz s_1 e s_2 ; ii) sinal s_1 recuperado ao aplicarmos a máscara ideal sobre a mistura; iii) relação sinal-distorção do sinal s_1 recuperado em relação ao sinal s_1 gravado separadamente ($\overline{SDR} = 8$ dB); iv) sinal s_2 recuperado ao aplicarmos a máscara ideal sobre a mistura; v) relação sinal-distorção do sinal s_2 recuperado em relação ao sinal s_2 gravado separadamente ($\overline{SDR} = 4$ dB).

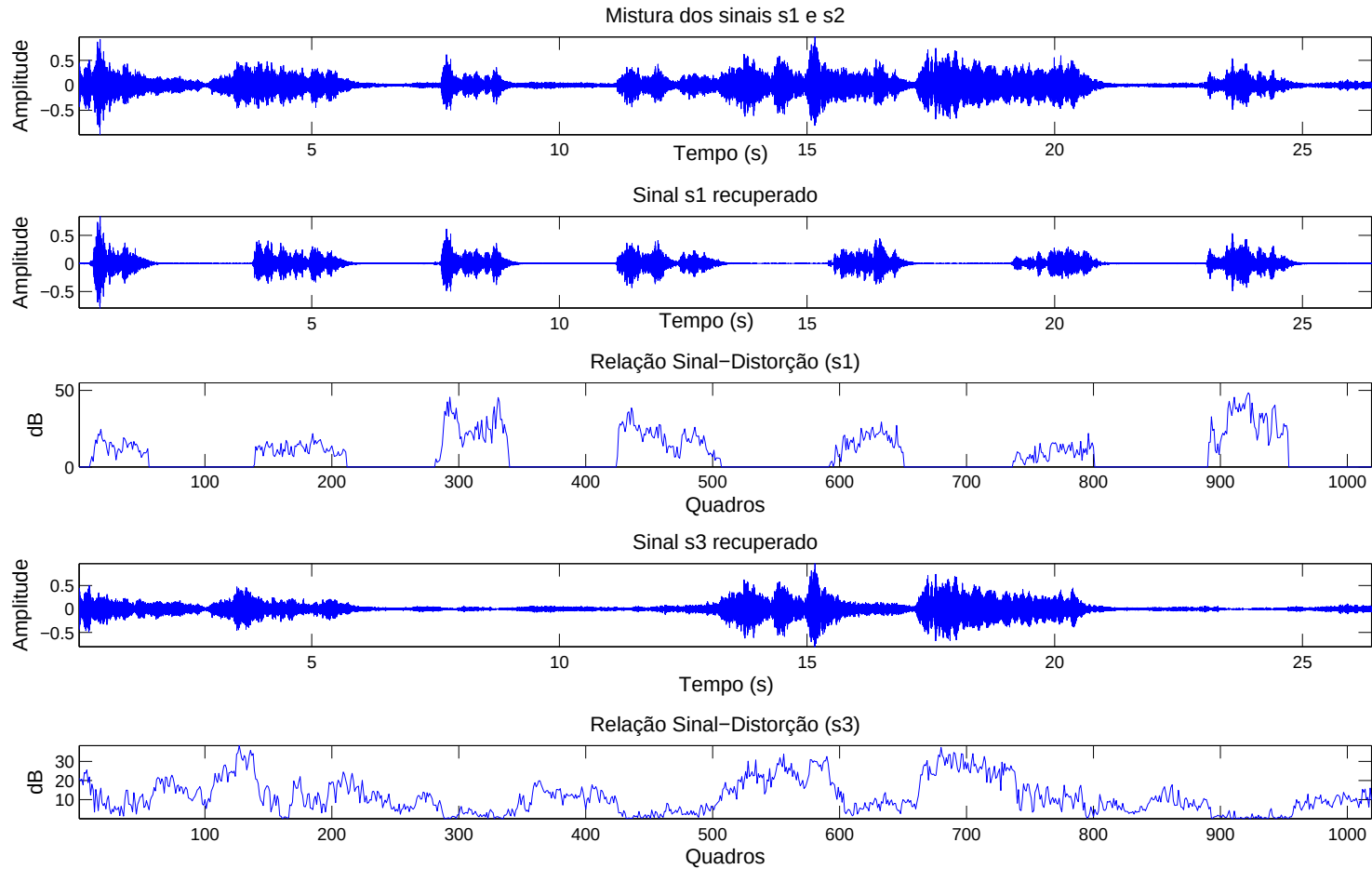


Figura 4.20: Gravações de dois sinais $s1$ (voz masculina) e $s3$ (música) realizadas no ambiente 2 com tempo de reverberação igual a 1100 ms . De cima para baixo as figuras correspondem, respectivamente: i) mistura dos sinais de voz $s1$ e música $s3$; ii) sinal $s1$ recuperado ao aplicarmos a máscara ideal sobre a mistura; iii) relação sinal-distorção do sinal $s1$ recuperado em relação ao sinal $s1$ gravado separadamente ($\overline{SDR} = 8\text{ dB}$); iv) sinal $s3$ recuperado ao aplicarmos a máscara ideal sobre a mistura ; v) relação sinal-distorção do sinal $s3$ recuperado em relação ao sinal $s3$ gravado separadamente ($\overline{SDR} = 11\text{ dB}$).

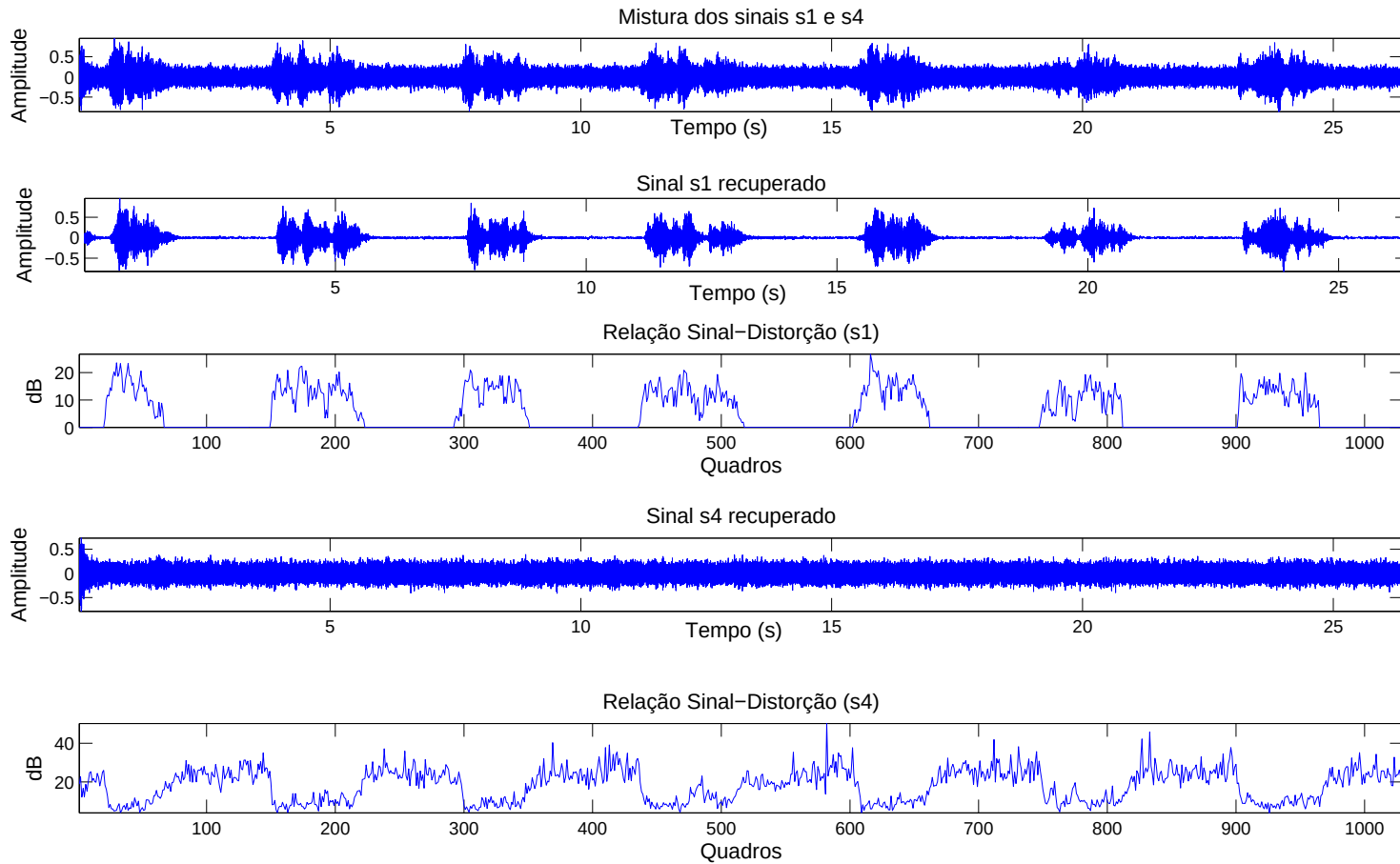


Figura 4.21: Gravações de dois sinais s_1 (voz masculina) e s_4 (ruído branco) realizadas no ambiente 2 com tempo de reverberação igual a 1100 ms. De cima para baixo as figuras correspondem, respectivamente: i) mistura dos sinais de voz s_1 e o ruído branco s_4 ; ii) sinal s_1 recuperado ao aplicarmos a máscara ideal sobre a mistura; iii) relação sinal-distorção do sinal s_1 recuperado em relação ao sinal s_1 gravado separadamente ($\overline{SDR} = 4$ dB); iv) sinal s_4 recuperado ao aplicarmos a máscara ideal sobre a mistura; v) relação sinal-distorção do sinal s_4 recuperado em relação ao sinal s_4 gravado separadamente ($\overline{SDR} = 18$ dB).

Na Tabela 4.2 pode-se observar os valores da relação sinal-distorção (SDR) obtidos para as misturas (2x2) de sinais gravados em ambientes com tempos de reverberação diferentes: 100 e 1100 ms , conforme mencionado anteriormente. Ao aplicar a máscara ideal sobre uma mistura contendo dois sinais de voz (uma voz masculina e outra feminina) a SDR média relativa aos sinais diminui aproximadamente cerca de 3 dB , para o caso do ambiente com tempo de reverberação igual a 1100 ms . Para o caso do sinal s_2 pode-se verificar que a relação sinal-distorção diminui de 8 dB (ambiente com $T_{60}= 100$ ms) para 4 dB (ambiente com $T_{60}= 1100$ ms). Embora no ambiente com maior tempo de reverberação a relação sinal-distorção obtida seja inferior, verifica-se que o sinal separado ainda continua inteligível. Nas outras duas situações propostas observa-se também que o desempenho médio ($\overline{SDR}$) diminui no ambiente com tempo de reverberação igual a 1100 ms : cerca de 5 dB para a gravação (mistura) dos sinais s_1 (voz masculina) e s_3 (música) e aproximadamente 8 dB para a gravação (mistura) dos sinais s_1 (voz masculina) e s_4 (ruído branco). Os sinais relativos as misturas e os sinais separados recuperados para cada uma das misturas e ambientes descritos na Tabela 4.2 são mostrados nas figuras 4.16, 4.17, 4.17, 4.19, 4.20 e 4.21.

Tabela 4.2: Aplicação da máscara ideal sobre uma mistura de sinais gravados em ambientes com tempos de reverberação iguais a 100 e 1100 ms .

<i>Ambiente 1: Tempo de reverberação igual a 100 ms</i>		$\hat{s}_a$	$\hat{s}_b$	Média
Gravação (mistura) dos sinais s_1 (voz masculina) e s_2 (voz feminina)	$SDR(dB)$	9	8	9
Gravação (mistura) dos sinais s_1 (voz masculina) e s_3 (música)	$SDR(dB)$	8	21	15
Gravação (mistura) dos sinais s_1 (voz masculina) e s_4 (ruído branco)	$SDR(dB)$	4	34	19
<i>Ambiente 2: Tempo de reverberação igual a 1100 ms</i>		$\hat{s}_a$	$\hat{s}_b$	Média
Gravação (mistura) dos sinais s_1 (voz masculina) e s_2 (voz feminina)	$SDR(dB)$	8	4	6
Gravação (mistura) dos sinais s_1 (voz masculina) e s_3 (música)	$SDR(dB)$	8	11	10
Gravação (mistura) dos sinais s_1 (voz masculina) e s_4 (ruído branco)	$SDR(dB)$	4	18	11

4.4 Influência do tempo de reverberação no mascaramento espectral

À medida em que aumenta-se o tempo de reverberação, a separação de sinais fonte via mascaramento espectral tem sua eficiência diminuída devido ao espalhamento das frequências ao longo do tempo. A separação de sinais fonte através da aplicação de máscaras ideais considera que os sinais sejam esparsos. Com o aumento do tempo de reverberação e, consequentemente, o aumento do espalhamento das frequências ao longo do tempo, pode-se observar que a consideração de esparsidade dos sinais fonte fica comprometida em ambientes com tempo de reverberação elevados.

As figuras 4.22 e 4.23 mostram os espectrogramas e as curvas de decaimento de energia de um sinal *TSP* obtidos considerando a simulação de um ambiente anecóico e de ambientes reverberantes. Observa-se através dos espectrogramas obtidos que, à medida que se aumenta o tempo de reverberação, ocorre um espalhamento uniforme de todas as frequências ao longo do tempo. A figura 4.24 mostra o espectrograma de um sinal *TSP* obtido para o caso de gravações reais em ambientes com tempos de reverberação iguais a 100 *ms* e 1100 *ms*, respectivamente. Para obtenção do tempo de reverberação (T_{60}) foi triplicado o tempo de decaimento do nível do sinal de -5 *dB* até o nível de -25 *dB*. Nota-se neste caso que o espalhamento das frequências ao longo do tempo não acontece de forma uniforme como nos resultados obtidos através de simulações. O espalhamento das frequências ao longo do tempo, de acordo com a frequência observada, ocorre com intervalos de tempo diferentes. Este fato mostra que o processo usado para simular as respostas ao impulso não descreve exatamente o comportamento de ambientes reais.

4.5 Testes de percepção

Nesta parte do trabalho, são apresentados os resultados dos testes realizados para avaliação da percepção de inteligibilidade de sinais de voz recuperados utilizando técnicas de mascaramento espectral. Os testes foram divididos em duas etapas: primeiramente analisou-se a inteligibilidade dos sinais recuperados utilizando o algoritmo *DUET* e uma versão simplificada do mesmo proposta no capítulo 3; e posteriormente avaliou-se a inteligibilidade dos sinais recuperados através da aplicação da máscara ideal sobre a mistura de dois sinais de voz considerando diversos tempos de reverberação.

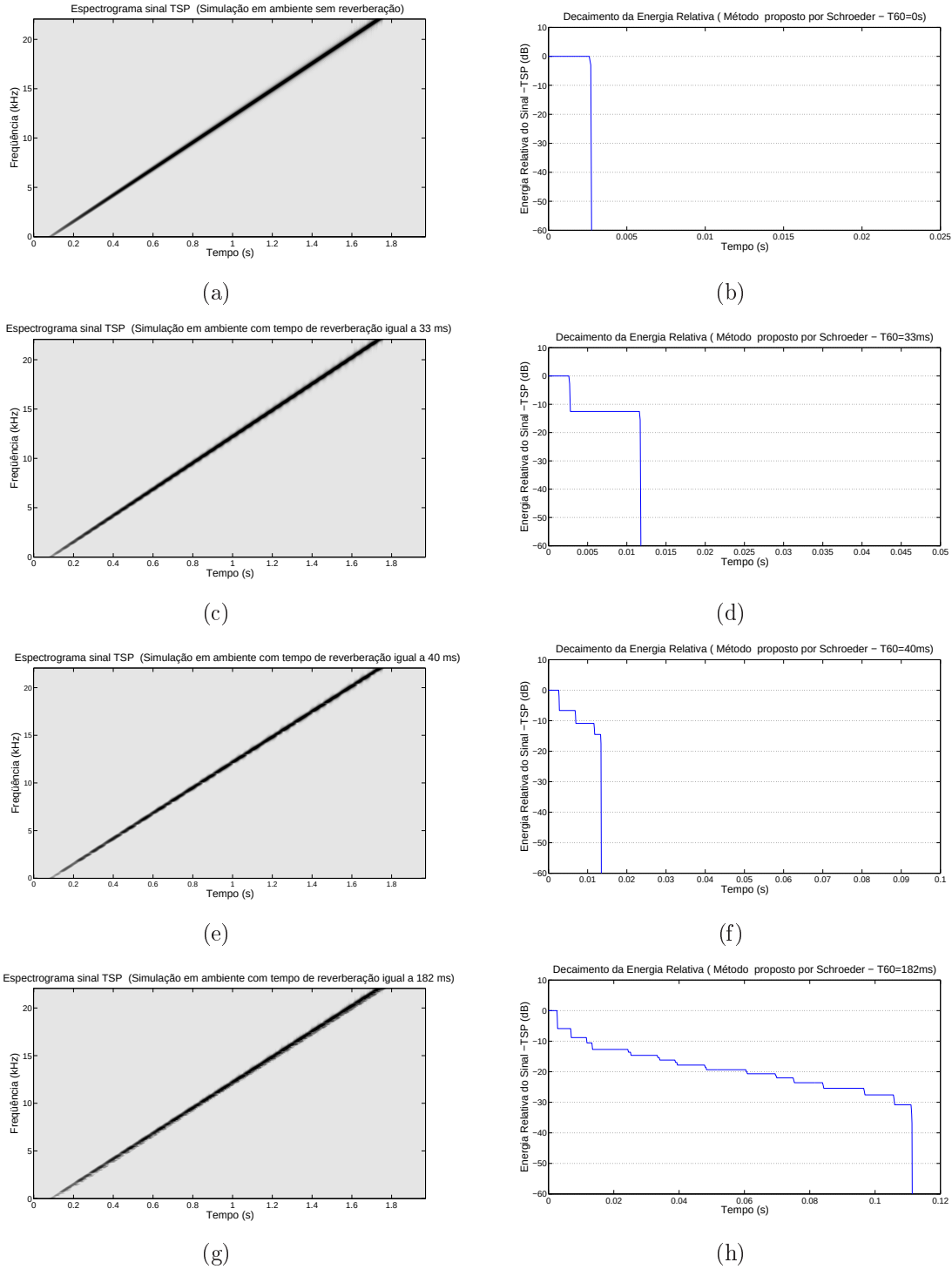


Figura 4.22: Espectrogramas e curvas de decaimento da energia obtidas para um sinal *TSP*. O tempo de reverberação (T_{60}) foi especificado de acordo com o tempo de decaimento do nível do sinal de -5 dB até -25 dB . (a) Espectrograma obtido para o caso de simulação de um ambiente anecóico. (b) Curva de decaimento de energia obtido para o caso de simulação de um ambiente anecóico. (c) Espectrograma obtido considerando a simulação de um ambiente com tempo de reverberação igual a 33 ms . (d) Curva de decaimento de energia considerando a simulação de um ambiente com tempo de reverberação igual a 33 ms . (e) Espectrograma obtido considerando a simulação de um ambiente com tempo de reverberação igual a 40 ms . (f) Curva de decaimento de energia considerando a simulação de um ambiente com tempo de reverberação igual a 40 ms . (g) Espectrograma obtido considerando a simulação de um ambiente com tempo de reverberação igual a 182 ms . (h) Curva de decaimento de energia considerando a simulação de um ambiente com tempo de reverberação igual a 182 ms .

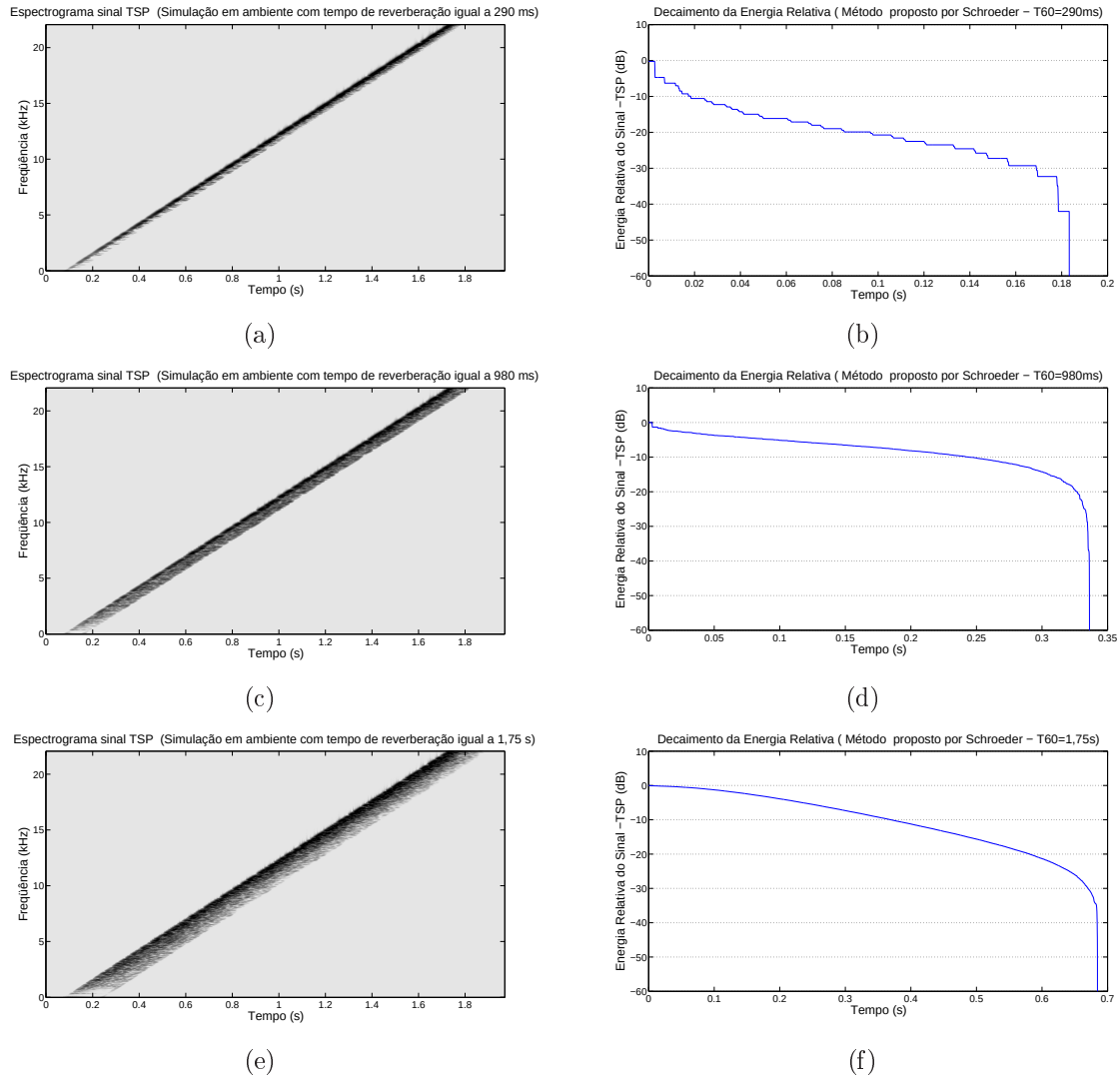
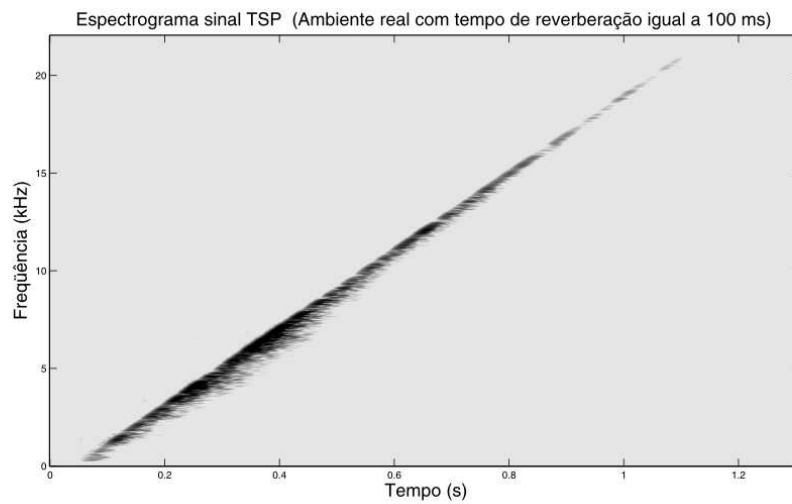
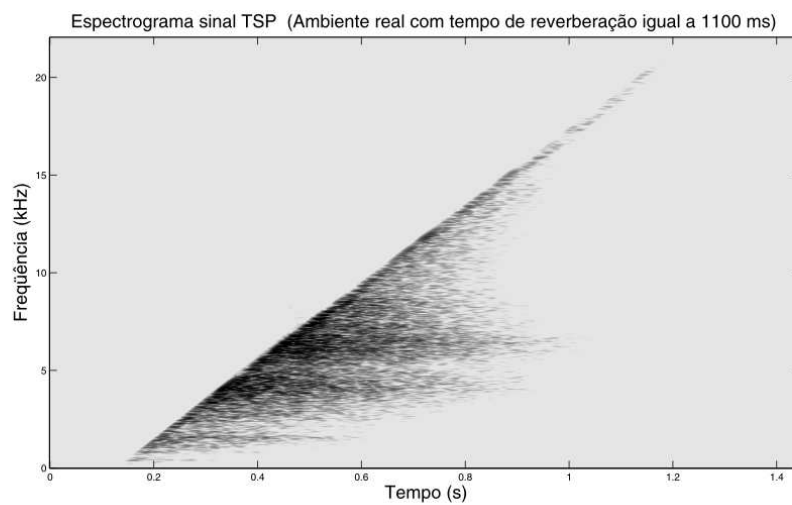


Figura 4.23: Espectrogramas e curvas de decaimento da energia obtidas para um sinal *TSP*. (a) Espectrograma obtido considerando a simulação de um ambiente com tempo de reverberação igual a 290 ms. (b) Curva de decaimento de energia considerando a simulação de um ambiente com tempo de reverberação igual a 290 ms. (c) Espectrograma obtido considerando a simulação de um ambiente com tempo de reverberação igual a 980 ms. (d) Curva de decaimento de energia considerando a simulação de um ambiente com tempo de reverberação igual a 980 ms. (e) Espectrograma obtido considerando a simulação de um ambiente com tempo de reverberação igual a 1750 ms. (f) Curva de decaimento de energia considerando a simulação de um ambiente com tempo de reverberação igual a 1750 ms.



(a)



(b)

Figura 4.24: Espectrogramas de um sinal *TSP* obtido através de gravações em ambientes reais. (a) Espectrograma obtido para um ambiente com tempo de reverberação igual a 100 ms. (b) Espectrograma obtido para um ambiente com tempo de reverberação igual a 1100 ms.

4.6 Testes subjetivos para avaliação de desempenho do algoritmo *DUET* e da versão simplificada proposta

Este teste tem como objetivo avaliar a inteligibilidade dos sinais recuperados através da aplicação de uma máscara binária estimada com base nos parâmetros de amplitude e atraso relativos, para o caso do algoritmo *DUET*, e avaliar a inteligibilidade também para o caso de se obter a máscara binária utilizando apenas a informação de amplitude relativa. A base de dados utilizada para estes testes é composta por dez sentenças. As sentenças de voz correspondem a sinais de voz separados pelos dois métodos mencionados e foram obtidas para os casos de misturas instantâneas, contendo atrasos, convoluídas artificialmente, gravações reais em ambiente aberto e gravações em ambiente com reverberação, sendo as mesmas misturas e sinais recuperados descritos nas seções 3.10, 3.11, 3.12, 3.13 e 3.14. O teste foi realizado por um total de 10 participantes, sendo entregue um questionário para preenchimento durante o teste com as devidas instruções sobre como se proceder durante o experimento. Os testes foram realizados individualmente e através da utilização de um fone de ouvido, cada participante ouvia sentença por sentença, avaliava a inteligibilidade do sinal e anotava a frase percebida no questionário entregue. A Tabela 4.3 mostra as frases utilizadas, bem como a sequência em que foram reproduzidas nos testes. As respostas dos questionários foram transferidas para tabelas que tinham a finalidade de indicar a quantidade de palavras corretas em cada frase reproduzida. A inteligibilidade das frases foi expressa pelo percentual de acertos definido pela razão entre a quantidade de palavras corretamente percebidas e a quantidade de palavras presentes nas frases. Os resultados obtidos neste experimento são apresentados na Figura 4.25. Ao verificarmos os resultados obtidos observar-se que os algoritmos *DUET* e a versão simplificada proposta apresentam desempenho similares em termos de inteligibilidade. Apenas em uma situação específica, na mistura contendo atrasos, observou-se um desempenho do método baseado apenas na amplitude relativa significativamente superior ao desempenho do *DUET*, mas no geral o desempenho dos algoritmos são bem similares. Outro ponto a ser destacado refere-se a inteligibilidade observada nas gravações reais, realizadas em campo aberto (ambiente com pouca reverberação) e em uma sala reverberante (tempo de reverberação igual a 1100 ms). No ambiente reverberante a taxa de acertos foi aproximadamente da ordem de 30 %, porém em ambiente aberto os algoritmos apresentaram boa inteligibilidade na separação dos sinais, obtendo uma taxa de acertos da ordem de 78 %.

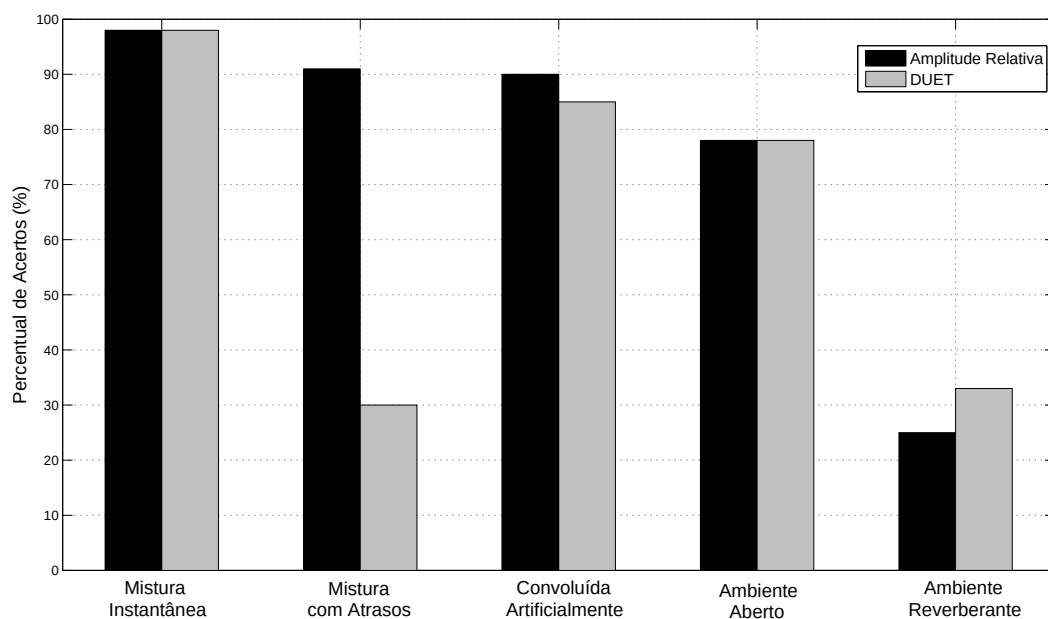


Figura 4.25: Resultado do teste subjetivo individual de avaliação de inteligibilidade de sinais de voz separados, a partir de vários tipos de misturas (instantâneas, com atrasos, convoluídas artificialmente, gravações reais em ambiente aberto e gravações em ambiente com reverberação), utilizando o algoritmo *DUET* e uma versão simplificada proposta baseada somente na informação de amplitude relativa. Nos testes foram utilizados dez avaliadores. A inteligibilidade das frases foi expressa pelo percentual de acertos definido pela razão entre a quantidade de palavras corretamente percebidas e a quantidade de palavras presentes nas frases, conforme mostrado na Tabela 4.3.

Tabela 4.3: Sequência de frases reproduzidas durante o primeiro experimento para avaliação de inteligibilidade de sinais de voz recuperados utilizando técnicas de mascaramento espectral.

Ordem	Mistura	Algoritmo	Sentença	Quantidade Palavras
1	Instantânea	Amplitude Relativa	Tem um programa extremamente interessante.	5
2	Instantânea	<i>DUET</i>	Duas toneladas de remédio serão distribuídas.	6
3	Instantânea	<i>DUET</i>	Tem um programa extremamente interessante.	5
4	Instantânea	Amplitude Relativa	Duas toneladas de remédio serão distribuídas.	6
5	Atraso	<i>DUET</i>	Circula pela fábrica inteira.	4
6	Atraso	Amplitude Relativa	Estados atingidos pelos temporais.	4
7	Atraso	Amplitude Relativa	Circula pela fábrica inteira.	4
8	Atraso	<i>DUET</i>	Estados atingidos pelos temporais.	4
9	Convoluída Artificialmente	<i>DUET</i>	Conhece todo processo de fabricação de cerveja.	7
10	Convoluída Artificialmente	Amplitude Relativa	Nesta terça-feira o presidente lula e quatro ministros.	8
11	Convoluída Artificialmente	Amplitude Relativa	Conhece todo processo de fabricação de cerveja.	7
12	Convoluída Artificialmente	<i>DUET</i>	Nesta terça-feira o presidente lula e quatro ministros.	8
13	Ambiente Aberto	Amplitude Relativa	Os temporais são um fenômeno fora de época.	8
14	Ambiente Aberto	<i>DUET</i>	E ao final da visita você entra em um salão.	10
15	Ambiente Aberto	<i>DUET</i>	Os temporais são um fenômeno fora de época.	8
16	Ambiente Aberto	Amplitude Relativa	E ao final da visita você entra em um salão.	10
17	Ambiente Reverberante	Amplitude Relativa	Na saída eu comprei esta camisa que estou usando.	9
18	Ambiente Reverberante	<i>DUET</i>	Um dos ministros que vai acompanhar o presidente.	8
19	Ambiente Reverberante	<i>DUET</i>	Na saída eu comprei esta camisa que estou usando.	9
20	Ambiente Reverberante	Amplitude Relativa	Um dos ministros que vai acompanhar o presidente.	8

4.7 Testes subjetivos para avaliação da inteligibilidade de sinais de voz recuperados através da aplicação da máscara ideal sobre a mistura de dois sinais de voz considerando diversos tempos de reverberação

A inteligibilidade dos sinais de voz separados através de uma máscara binária espectral ideal foi também avaliada para diversos tempos de reverberação. A base de dados utilizada para estes testes é composta por dezesseis sentenças. As sentenças de voz correspondem a sinais de voz separados através da aplicação da máscara ideal sobre misturas convoluídas considerando diversos tempos de reverberação (anecóico, 33 *ms*, 40 *ms*, 182 *ms*, 290 *ms*, 980 *ms*, 1750 *ms* e 3900 *ms*), os sinais separados e as misturas convoluídas são os mesmos descritos na seção 4.2. O teste foi realizado por um total também de 10 participantes, sendo entregue um questionário para preenchimento durante o teste com as devidas instruções sobre como se proceder durante o experimento. Novamente, os testes foram realizados individualmente e através da utilização de um fone de ouvido, cada participante ouvia sentença por sentença, avaliava a inteligibilidade do sinal e anotava a frase percebida no formulário entregue. A Tabela 4.4 mostra as frases utilizadas e a sequência em que foram reproduzidas durante a realização dos testes. As repostas dos questionários foram transferidas para tabelas que tinham a finalidade de indicar a quantidade de palavras corretas em cada frase reproduzida. A inteligibilidade das frases foi expressa pelo percentual de acertos definido pela razão entre a quantidade de palavras corretamente percebidas e a quantidade de palavras presentes nas frases. Os resultados obtidos mostram que a medida em que se aumenta o tempo de reverberação a inteligibilidade do sinal reduz, conforme mostrado na Figura 4.26. Analisando os resultados obtidos pode-se constatar que foram muito similares aos obtidos quando realizamos os mesmos testes anteriormente, porém utilizando medidas objetivas (relação sinal-distorção), conforme mostra a Figura 4.12. Os testes subjetivos apresentam uma alta taxa de acertos (superior a 90 %) para sinais recuperados a partir de misturas com tempo de reverberação menores que 300 *ms*.

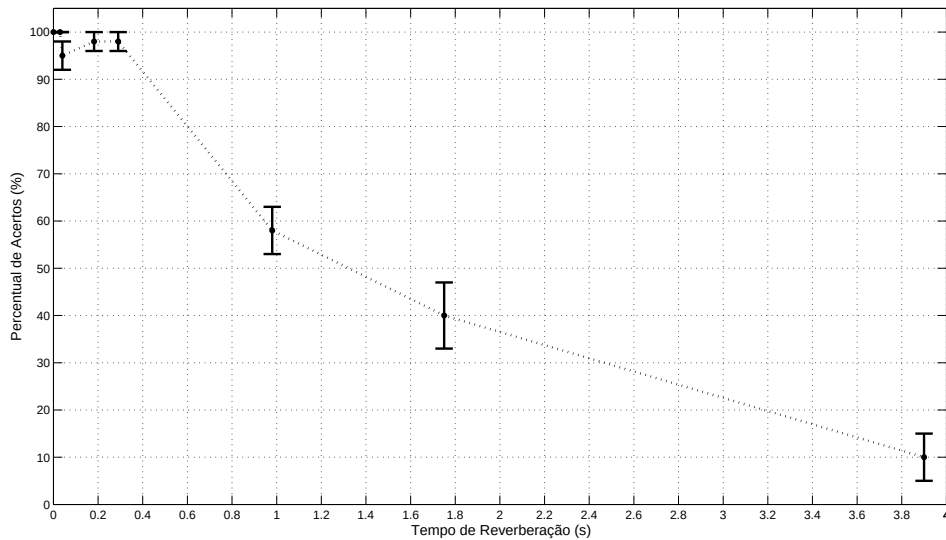


Figura 4.26: Resultado do teste subjetivo individual de avaliação de inteligibilidade de sinais de voz separados através de uma máscara binária espectral ideal considerando diversos tempos de reverberação. Nos testes foram utilizados dez avaliadores. A inteligibilidade das frases foi expressa pelo percentual de acertos definido pela razão entre a quantidade de palavras corretamente percebidas e a quantidade de palavras presentes nas frases, conforme mostrado na Tabela 4.4. A curva pontilhada apresenta as percentagens médias de acertos dos avaliadores referentes as sentenças reproduzidas de acordo com o tempo de reverberação. As retas cruzando os pontos apresentam o desvio padrão dos resultados obtidos em relação a cada tempo de reverberação analisado.

4.8 Algoritmos propostos para separação de fontes usando apenas uma mistura

Para o caso de sinais gravados em ambientes reais, ou seja, misturas contendo reverberação, o desempenho dos algoritmos de separação cega de sinais está longe do ideal. No caso mais extremo, onde temos apenas uma mistura e duas ou mais fontes, temos um problema ainda em aberto para o caso de sinais complexos como são os sinais de voz. Neste caso, simples considerações estatísticas relacionadas aos sinais fonte não são suficientes para a separação. A maioria dos algoritmos propostos para a separação de fontes utilizando apenas uma mistura se baseiam na estimação de máscaras ideais e em um banco de dados contendo informações utilizadas para treinamento do sistema de separação. Visando a implementação de um sistema de separação de fontes utilizando apenas uma mistura e que não necessite de treinamento, uma série de testes para a estimação de máscaras ideais com base em medidas de não-gaussianidade e na medida de energia é apresentada nas seções seguintes e suas limitações discutidas.

Tabela 4.4: Sequência de frases reproduzidas durante o segundo experimento para avaliação de inteligibilidade de sinais de voz recuperados utilizando uma máscara binária ideal a partir de misturas convoluídas com diferentes tempos de reverberação.

Ordem	Reverberação	Sentença	Quantidade Palavras
1	Anecóico	Leila tem um lindo jardim.	4
2	Anecóico	Nosso telefone quebrou.	3
3	980 <i>ms</i>	A justiça é a única vencedora.	6
4	40 <i>ms</i>	A casa foi vendida sem pressa.	6
5	290 <i>ms</i>	Recebi nosso amigo para almoçar.	5
6	1750 <i>ms</i>	Isto se resolverá de forma tranquila.	6
7	3900 <i>ms</i>	Sei que atingiremos nosso objetivo.	5
8	182 <i>ms</i>	Trabalhando com união rende muito mais.	6
9	33 <i>ms</i>	O analfabetismo é a vergonha do país.	7
10	980 <i>ms</i>	Hoje dormirei bem.	3
11	40 <i>ms</i>	Queremos discutir o orçamento.	4
12	290 <i>ms</i>	Zé vai mais rápido.	4
13	1750 <i>ms</i>	João deu pouco dinheiro.	4
14	3900 <i>ms</i>	Ainda são seis horas.	4
15	182 <i>ms</i>	Ela tem muita fome.	4
16	33 <i>ms</i>	Desculpa se magoei o velho.	5

4.9 Medidas de não-gaussianidade

Muitos algoritmos de *ICA* se baseiam em medidas de não-gaussianidade. O teorema do limite central assume que a distribuição da soma de duas variáveis aleatórias independentes tendem a apresentar uma distribuição mais gaussiana que a distribuição dos dois sinais separados. Desta forma, pode-se utilizar a maximização de medidas de não-gaussianidade para a recuperação de sinais misturados.

4.9.1 Curtose

Uma medida clássica de não-gaussianidade é a curtose que pode ser definida como:

$$Curtose(y) = E\{y^4\} - 3(E\{y^2\})^2. \quad (4.4)$$

O valor de curtose de uma variável pode ser positivo ou negativo. As variáveis aleatórias que apresentam valores negativos são chamadas de sub-gaussianas, e aquelas com valores positivos são chamadas de super-gaussianas, conforme ilustrado na Figura 4.27.

Uma medida típica de não-gaussianidade é o valor absoluto da curtose. O quadrado

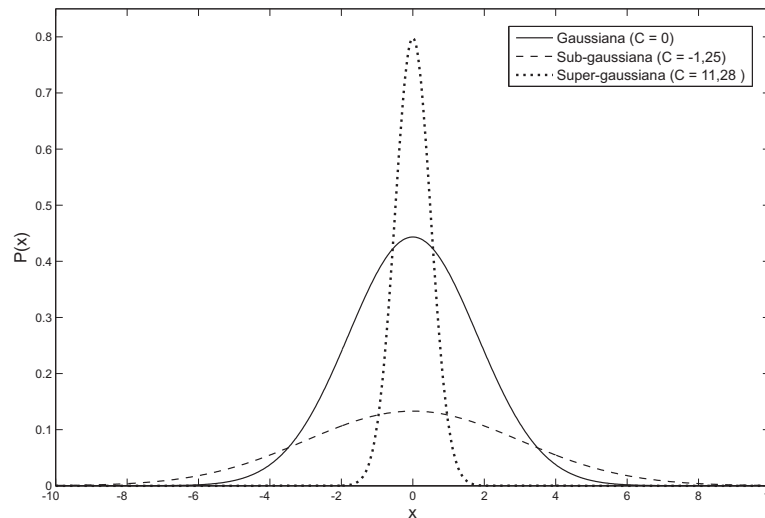


Figura 4.27: Função de densidade de probabilidade gaussiana, super-gaussiana e sub-gaussiana. Pode-se verificar que a distribuição gaussiana apresenta valor de curtose igual a zero, a distribuição super-gaussiana apresenta valor igual a 11,28 e a distribuição sub-gaussiana valor igual a -1,25.

da curtose também pode ser usado. Neste caso, a curtose apresenta valores iguais a zero para uma variável gaussiana e maior que zero para a maioria das variáveis aleatórias não-gaussianas. Existem variáveis aleatórias, com distribuição não-gaussiana, que apresentam curtose iguais a zero mas podem ser consideradas um caso raro (Hyvarinen e Oja, 2000).

Entretanto a curtose apresenta alguns problemas quando seus valores são estimados a partir de dados amostrados. O principal problema deve-se ao fato da curtose ser uma medida muito sensível a pontos fora do padrão (erros de amostragem). O valor estimado sobre poucos pontos de observação pode se basear em erros de amostragem ou em observações irrelevantes. Desta forma, a medida de curtose não é uma medida robusta de não-gaussianidade. Uma medida mais robusta que a curtose é a negentropia (Hyvarinen e Oja, 2000).

4.9.2 Negentropia

A medida de negentropia é baseada na teoria da informação. A entropia de uma variável aleatória pode ser interpretada como o grau de informação relacionado a uma variável observada. Quanto mais aleatória ou maior a dificuldade de predição de uma determinada variável, maior será a sua entropia. A entropia de uma variável aleatória discreta (Y) pode ser definida como:

$$Entropia(Y) = - \sum_i P(Y = a_i) \log P(Y = a_i), \quad (4.5)$$

onde a_i corresponde aos valores possíveis de Y . Para análise da entropia de um sinal contínuo, temos que a entropia diferencial (*Entropia*) de uma variável aleatória y com densidade $f(y)$ é definida como:

$$Entropia(y) = - \int_i f(y) \log f(y) dy. \quad (4.6)$$

Um resultado fundamental da teoria da informação diz que uma variável gaussiana tem maior entropia entre todas as variáveis aleatórias com mesma variância. Isto significa que pode-se utilizar a entropia como medida de não-gaussianidade. A entropia apresenta um valor pequeno para distribuições que são claramente concentradas em um determinado valor, ou seja para distribuições super-gaussianas. Visando obter uma medida de não-gaussianidade que seja zero para uma variável aleatória e apresente sempre valores positivos é utilizada uma versão ligeiramente modificada da definição de entropia diferencial chamada de negentropia. A negentropia (J) é definida como:

$$J(y) = Entropia(y_{gauss}) - Entropia(y), \quad (4.7)$$

onde y_{gauss} é uma variável aleatória gaussiana com a mesma matriz de covariância que y . Devido às propriedades mencionadas anteriormente a negentropia assume sempre valor positivo e valor igual a zero se y apresenta uma distribuição gaussiana. A negentropia é um bom estimador de não-gaussianidade. O problema ao utilizar a medida de negentropia é o fato de ser uma medida de difícil cálculo computacional. A estimação da negentropia usando a definição acima requer a estimação da função de densidade de probabilidade do sinal (*pdf*). Desta forma, existe a necessidade de introduzirmos algumas aproximações no cálculo da negentropia (Hyvarinen e Oja, 2000).

O método clássico para aproximação da negentropia utiliza momentos de alta ordem:

$$J(y) \approx \frac{1}{12} E\{y^3\}^2 + \frac{1}{48} Curt(y)^2, \quad (4.8)$$

onde a variável aleatória y é considerada como tendo média zero e variância unitária. Entretanto esta aproximação também não apresenta robustez. Para evitar estes problemas novas aproximações são propostas (Hyvarinen e Oja, 2000). Uma aproximação baseada no princípio da máxima entropia é dada através da seguinte expressão:

$$J(y) \approx \sum_{i=1}^P k_i [E\{G_i(y)\} - E\{G_i(v)\}]^2, \quad (4.9)$$

onde k_i é uma constante positiva e v é uma variável gaussiana com média zero e variância unitária. A variável y também apresenta média zero e variância unitária e G_i corresponde a funções não-quadráticas. No caso de utilizarmos apenas uma função não-quadrática G

a aproximação pode ser definida como:

$$J(y) \propto [E\{G(y)\} - E\{G(v)\}]^2, \quad (4.10)$$

para praticamente qualquer função não-quadrática G . Um ponto importante consiste na escolha da função G . Escolhendo uma função G que não cresça muito rápido obtém-se um estimador robusto. As seguintes escolhas de funções G apresentam bom desempenho (Hyvarinen e Oja, 2000):

$$G_1(u) = \frac{1}{a_1} \log(\cosh(a_1 u)), \quad (4.11)$$

$$G_2(u) = -\exp(-u^2/2), \quad (4.12)$$

onde $1 \leq a_1 \leq 2$ é uma constante apropriada. A função G_1 representa uma função de contraste que geralmente proporciona bons resultados. Quando os sinais são independentes e com uma distribuição super-gaussiana, ou quando a robustez é um fator muito importante, a função G_2 fornece uma aproximação melhor segundo Hyvarinen (1999). Porém, neste trabalho, foi utilizada a função G_1 que apresentou melhores resultados. Desta maneira, obtém-se uma aproximação da negentropia que fornece uma boa relação de compromisso entre as duas medidas clássicas de não-gaussianidade: curtose e negentropia. Esta aproximação é simples, possui baixo custo de processamento e apresenta robustez.

4.10 Análise das medidas de curtose e negentropia para quadros de sinais de voz separados e misturados.

Nesta seção serão analisados os valores de curtose e negentropia obtidos para cada quadro de dois sinais de voz separados e também para o caso da mistura dos sinais (os sinais foram somados). Os sinais de voz foram amostrados a uma taxa de 10000 Hz e foram utilizados quadros com 512 amostras (sobreposição de 50%) e um total de 240 quadros.

A curtose assume valor zero quando a variável aleatória apresenta uma distribuição gaussiana, valores negativos para uma distribuição sub-gaussiana e valores positivos para uma distribuição super-gaussiana. No caso de sinais de voz (que apresentam uma distribuição laplaciana), pode-se dizer que, maximizando a curtose, maximiza-se a independência das variáveis. A Figura 4.28(a) mostra os valores de curtose obtidos para os quadros de dois sinais de voz separados e para a mistura (soma) dos mesmos. Para a análise dos valores obtidos considerou-se o valor da curtose ao quadrado, afim de evitar

valores negativos. Desta forma, ao analisar os valores encontrados, verifica-se que 45,0% dos quadros da soma dos sinais apresentam valores de curtose ao quadrado maiores que os valores máximos obtidos entre os quadros dos sinais separados. Observa-se também que apenas 10,4% dos quadros analisados da soma dos sinais apresentam valores de curtose ao quadrado menores que os valores mínimos encontrados entre os quadros dos sinais separados.

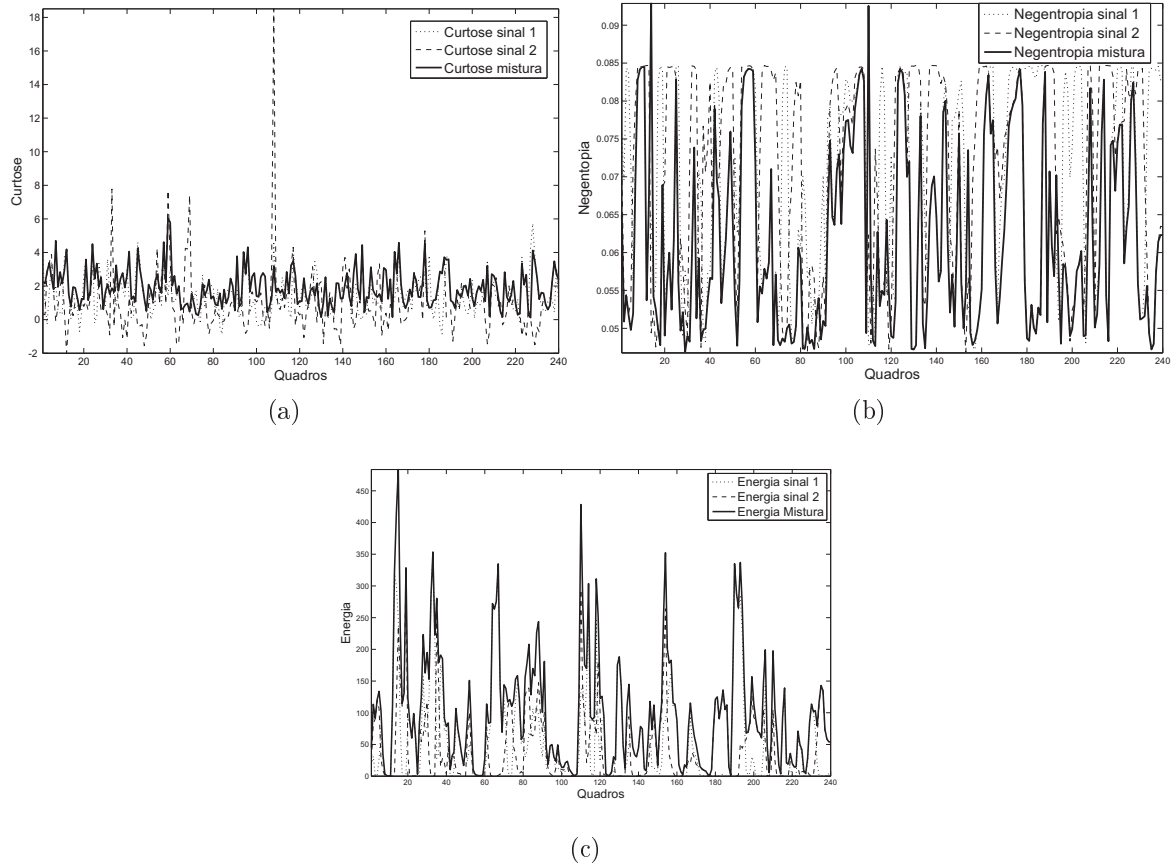


Figura 4.28: Valores de curtose, negentropia e energia obtidos para os quadros de dois sinais de voz separados e para a mistura (soma) dos mesmos. (a) Valores de Curtose. Para a análise dos valores obtidos considerou-se o valor da curtose ao quadrado, afim de evitar valores negativos. Pode-se verificar neste caso, que 45% dos quadros analisados da soma dos sinais apresentam valores de curtose ao quadrado maiores que os valores máximos obtidos entre os quadros dos sinais separados. E observamos também que apenas 10,4% dos quadros analisados da soma dos sinais apresentam valores de curtose ao quadrado menores que os valores de mínimo encontrados entre os quadros dos sinais separados. (b) Valores de Negentropia. Ao verificar os resultados percebe-se que cerca de 85% dos quadros analisados da soma dos sinais apresentam valores de negentropia menores que os valores de mínimo encontrados entre os quadros dos sinais separados. Observa-se também que aproximadamente 3,8% dos quadros analisados da soma dos sinais apresentam valores de negentropia maiores que os valores máximos obtidos entre os quadros dos sinais separados. (c) Valores de Energia. Como era de se esperar 100% dos quadros analisados da soma dos sinais apresentam valores de energia maiores que os valores de máximo encontrados entre os quadros dos sinais separados.

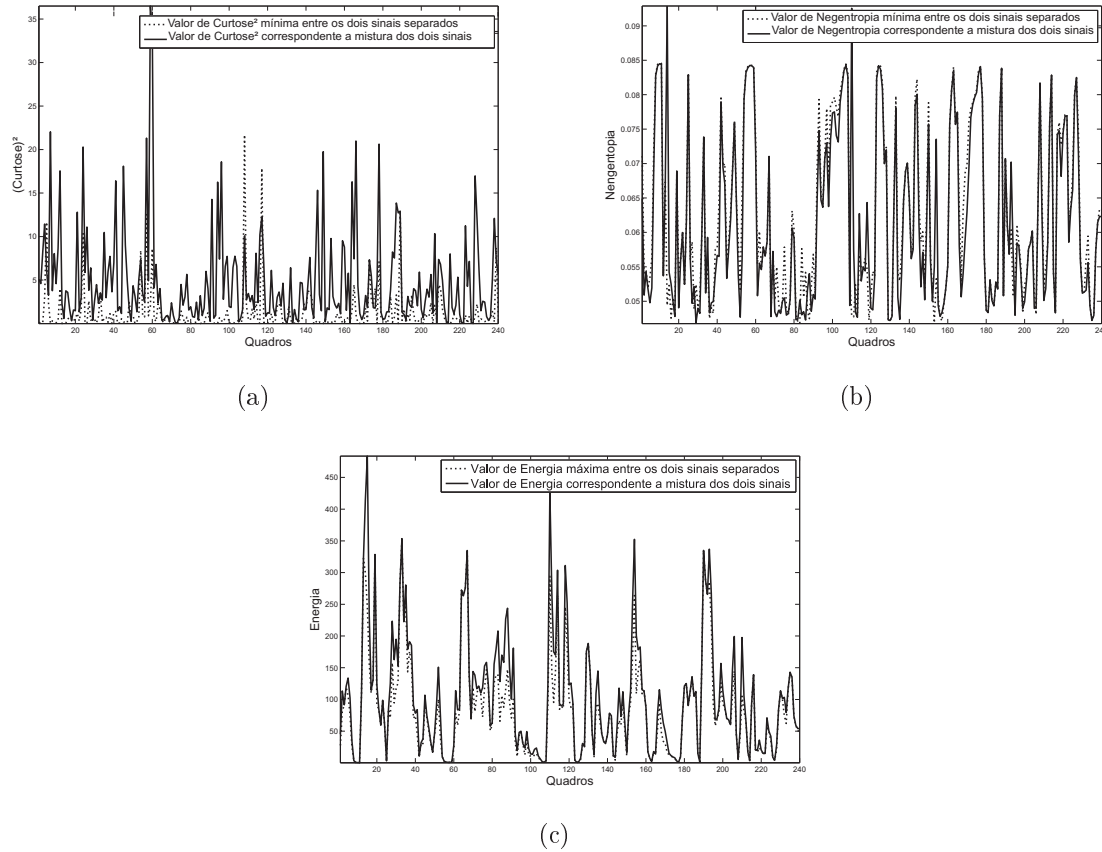


Figura 4.29: As curvas tracejadas correspondem aos menores valores do quadrado da curtose e da negentropia, e aos maiores valores de energia entre os quadros dos dois sinais de voz separados. As curvas contínuas correspondem aos valores do quadrado da curtose, negentropia e energia para a mistura (soma) dos dois sinais. (a) Valores relativos ao quadrado da curtose. O coeficiente de correlação linear entre a curva contendo os valores mínimos de curtose ao quadrado relativos aos quadros dos sinais separados e a curva contendo o valores de curtose ao quadrado dos quadros da mistura dos sinais é de 0,22. (b) Valores de negentropia. O coeficiente de correlação linear entre a curva contendo os valores mínimos de negentropia relativos aos quadros dos sinais separados e a curva contendo os valores de negentropia dos quadros da mistura dos sinais é de 0,91. (c) Valores de energia. O coeficiente de correlação linear entre a curva contendo o valores máximos de energia relativos aos quadros dos sinais separados e a curva contendo os valores de energia dos quadros da mistura dos sinais é de 0,95.

As variáveis gaussianas apresentam a maior entropia entre todas as variáveis aleatórias de igual variância. Desta maneira, maximizar a negentropia significa minimizar a informação mútua ao analisar a independência entre variáveis. Ao verificar os resultados, mostrados na Figura 4.28(b), percebe-se que cerca de 85,0% dos quadros analisados da soma dos sinais apresentam valores de negentropia menores que os valores mínimos encontrados entre os quadros dos sinais separados. Observa-se ainda que aproximadamente 3,8% dos quadros analisados da soma dos sinais apresentam valores de negentropia maiores que os valores máximos obtidos entre os quadros dos sinais separados, o que sugere que, no caso de poucos dados amostrados, a medida de negentropia é uma medida de não-gaussianidade mais robusta que a curtose. Os valores de curtose, negentropia e energia dos sinais separados e da mistura são mostrados na Figura 4.28.

Pode-se observar também que a medida de negentropia mínima entre os quadros dos sinais de voz separados é aproximadamente igual à medida de negentropia dos quadros da mistura (soma) dos dois sinais, conforme ilustrado na Figura 4.29(b). O mesmo já não ocorre com a medida de curtose. O coeficiente de correlação linear entre a curva contendo os valores mínimos dos quadros dos sinais separados e a curva contendo os valores dos quadros da mistura dos sinais é de 0,91 para os valores de negentropia e de 0,22 para os valores do quadrado da curtose. O coeficiente de correlação linear entre a curva contendo os valores máximos de energia dos quadros dos sinais separados e a curva contendo os valores de energia dos quadros da mistura dos sinais é de 0,95.

Observa-se também que a diferença entre o valor mínimo da negentropia entre os quadros dos sinais separados e o valor da negentropia de cada quadro da mistura (soma) dos dois sinais aumenta à medida que introduz-se erros na máscara ideal, conforme mostrado na Figura 4.30. Isto sugere que, quando se usa uma máscara ideal ou próxima do ideal o valor da diferença encontrado é próximo de um valor mínimo.

Um dos objetivos do trabalho é explorar as medidas de não-gaussianidade e de energia para obtenção das máscaras ideais. A maioria dos sistemas de separação cega de fontes que utiliza apenas uma mistura para a separação dos sinais necessita de dados para treinamento. Desta forma, seria muito útil a construção de um sistema de separação de fontes que utilize uma única mistura e que não necessite de treinamento. Para isto, verificou-se a viabilidade de estimar a máscara ideal a partir de um conjunto de 256 máscaras, onde foram sendo inseridos erros (*bit a bit*) a partir da máscara ideal, e com base em medidas de não-gaussianidade e de energia. Na Tabela 4.5 pode-se verificar as porcentagens relativas à relação sinal-distorção para os quadros dos sinais de voz estimados. Na primeira linha da tabela, ao aplicar a máscara ideal sobre a soma de dois sinais de voz e considerar o valor máximo de *SDR* obtido entre os dois sinais, observa-se que nenhum quadro de voz, em comparação com o quadro do sinal original, apresentou medida de *SDR* abaixo de 2

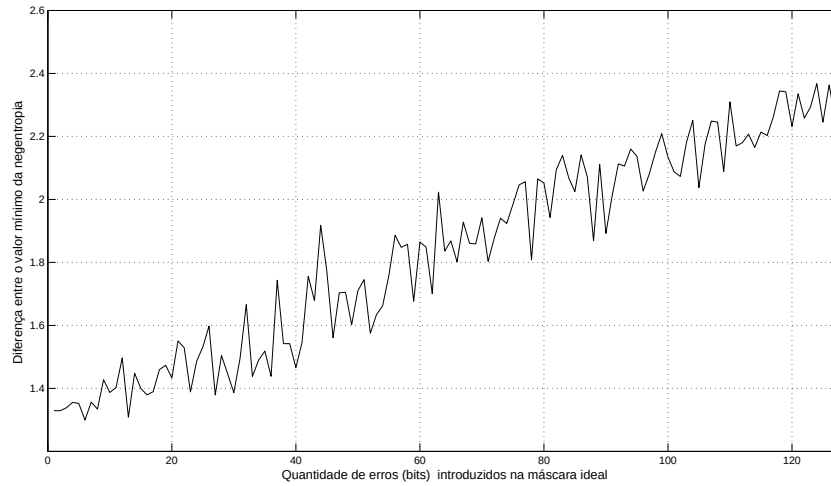


Figura 4.30: Diferença entre o valor mínimo da negentropia de cada quadro entre os sinais separados e o valor da negentropia de cada quadro da mistura (soma) dos dois sinais.

dB e que cerca de 91,7% dos quadros analisados apresentam relação sinal-distorção acima de 6 dB . Nos quatro itens restantes da tabela verificou-se o desempenho obtido durante a estimação das máscaras ideais através dos seguintes critérios utilizados, respectivamente: i) maximização da medida de negentropia para cada quadro dos sinais separados; ii) minimização da diferença entre o valor mínimo da negentropia de cada quadro entre os sinais separados e o valor da negentropia de cada quadro da mistura (soma) dos dois sinais; iii) maximização da medida de curtose ao quadrado para cada quadro dos sinais separados e iv) maximização da energia para cada quadro dos sinais separados.

Tabela 4.5: Percentual da relação sinal-distorção referente aos quadros dos sinais de voz estimados em comparação com cada quadro do sinal original.

SDR dB (%)	<0	0 a 2	2 a 4	4 a 6	6 a 8	8 a 10	10 a 12	>12
SDR máxima	0,0	0,0	0,8	7,5	13,8	16,6	13,8	47,5
Máxima Negentropia	17,9	60,8	16,3	4,2	0,4	0,4	0,0	0,0
Diferença Negentropia	5,4	10,8	10,8	12,9	15,4	11,3	9,6	23,8
Máxima Curtose	4,6	13,3	15,4	18,8	18,8	8,3	7,9	12,9
Máxima Energia	5,8	13,8	13,3	10,4	9,6	6,3	7,0	33,8

Ao maximizar a medida de negentropia verifica-se que 99,2% dos quadros apresentam uma relação sinal-distorção menor que 6 dB e apenas 0,8% dos quadros apresentam uma medida de SDR maior que 6 dB . Ao utilizar o critério de busca do valor mínimo da diferença de negentropia para estimação das máscaras ideais, verifica-se que 39,9% dos quadros apresentam uma relação sinal-distorção menor que 6 dB e cerca de 60,1% dos

quadros apresentam uma medida de SDR maior que 6 dB . Já ao maximizar a medida de curtose verifica-se que 52,1% dos quadros apresentam uma relação sinal-distorção menor que 6 dB e aproximadamente 47,9% dos quadros apresentam uma medida de SDR maior que 6 dB . Ao maximizarmos a medida de energia observa-se que 43,3% dos quadros apresentam uma relação sinal-distorção menor que 6 dB e 56,7% dos quadros apresentam uma medida de SDR maior que 6 dB . Os valores encontrados acima, correspondem ao valor máximo de SDR estimado para apenas uma das fontes e ainda não permitem a separação dos sinais fonte.

4.11 Considerações finais

Neste capítulo foram apresentadas várias análises realizadas com o objetivo de verificar a viabilidade e limitações de técnicas de separação via mascaramento espectral de acordo com o tempo de reverberação. Ao utilizar a máscara ideal na separação de sinais obtemos uma relação sinal-distorção da ordem de 12 dB , porém o sinal ainda mantém boa inteligibilidade. A medida que afasta-se da máscara ideal, ou a medida em que inserimos erros na máscara pode-se verificar que a relação sinal-distorção diminui. Ao analisar a queda de desempenho do processo de separação em função da distância entre a máscara ideal e a máscara efetivamente usada, verificou-se uma queda de desempenho de 3 dB quando aproximadamente 10 % dos bits são invertidos. Verificou-se também que à medida que aumenta-se o tempo de reverberação a eficiência das técnicas de separação via mascaramento espectral diminui. Isto ocorre devido a um espalhamento das frequências ao longo do tempo prejudicando a separação via máscaras ideais. Ao analisar misturas (obtidas através de simulações) com tempos de reverberação menores que 300 ms percebe-se que o resultado é o mesmo obtido para o caso de misturas anecóicas. Para tempos de reverberação da ordem de 4 s , a relação sinal-distorção diminui bastante, cerca de 8 dB , mas o sinal ainda permanece inteligível. Nos testes de percepção realizados observou-se um desempenho similar entre os algoritmos *DUET* e a versão simplificada baseada na informação apenas de amplitude relativa. Nas gravações reais realizadas em ambiente aberto (ambiente com pouca reverberação), os testes de percepção indicam alta inteligibilidade dos sinais recuperados pelos algoritmos, com uma taxa de acertos da ordem de 78%. Os testes subjetivos também apresentaram boa inteligibilidade dos sinais recuperados utilizando máscaras ideais, com taxas de acertos superiores a 90%, para misturas com tempo de reverberação menores que 300 ms . O caso de apenas uma mistura e duas ou mais fontes, corresponde a um problema ainda em aberto. A partir das análises apresentadas, verificou-se a viabilidade de estimar a máscara ideal a partir de um conjunto de 256 máscaras, onde foram sendo inseridos erros (bit a bit) a partir da máscara ideal, com

base em medidas de não-gaussianidade (curtose e negentropia) e na medida de energia. Atualmente, os algoritmos propostos para a separação cega de sinais que usam apenas uma mistura utilizam-se de um banco de dados para o treinamento do sistema. Visando a implementação de um sistema que utilize apenas uma mistura e não necessite de treinamento, foram realizadas análises preliminares para tentar estimar a máscara ideal. O ser humano a partir de uma única mistura consegue separar as fontes sonoras. Desta forma, pode-se considerar possível a separação de fontes a partir de uma única mistura. Porém, este problema ainda não tem uma solução conhecida.

Capítulo 5

Conclusão

Existe na literatura uma série de algoritmos de separação cega de fontes para misturas instantâneas capazes de obter uma separação quase perfeita (a menos de um fator de escala e permutação das fontes). Porém, para o caso de gravações reais de sinais de voz, em ambientes onde exista reverberação, os algoritmos de separação cega ainda apresentam um desempenho limitado.

Uma das situações mais complicadas envolvendo problemas de separação cega de fontes corresponde aos casos sub-determinados, ou seja, situações onde existem mais fontes que sensores. Muitos progressos têm sido alcançados assumindo considerações a respeito da esparsidade dos sinais fonte. Em uma mistura de sinais de voz, pode-se considerar que somente um sinal está ativo em um mesmo instante de tempo e frequência. Desta maneira, pode-se separar os sinais fonte através da aplicação de uma máscara ideal binária que privilegie o sinal com maior energia em um determinado instante de tempo e frequência.

Considerando que exista uma máscara ideal capaz de separar qualquer mistura de sinais de voz, o problema consiste em especificar um método capaz de estimar tais máscaras. Yilmaz e Rickard (2004a) propuseram um algoritmo (*DUET*) de separação de fontes utilizando duas misturas para a estimação das máscaras com base em dois parâmetros denominados: amplitude e atraso relativos. O algoritmo *DUET* foi desenvolvido considerando misturas sem reverberação, contendo somente atrasos. Porém, quando aplicado a misturas reverberadas não é capaz de realizar a separação dos sinais. Este estudo realizou uma análise detalhada do desempenho do algoritmo *DUET*, em diversas condições, através de simulações e experimentos. Uma conclusão interessante é que resultados similares são atingidos quando a informação de amplitude relativa é usada com ou sem a adição da informação de atraso relativo. Nos testes subjetivos realizados pode-se constatar também que o desempenho dos dois algoritmos são similares, em termos de inteligibilidade do sinal de voz.

Ao final da análise, surge o problema de determinar até que ponto a queda de de-

sempenho do algoritmo *DUET* é causada pela sua incapacidade de encontrar a máscara ideal ou pela incapacidade da máscara ideal separar misturas de sinais reverberados. Para responder a esta questão, verificou-se que a separação cega de fontes através da aplicação de máscaras ideais provoca uma distorção dos sinais separados, mas a qualidade sonora e a inteligibilidade dos sinais são perfeitamente aceitáveis em algumas situações. Ao aplicar a máscara ideal diretamente sobre os sinais fonte separados (originais), a relação sinal-distorção média obtida é de 12 *dB*. Quando aplica-se a máscara em uma mistura convoluída artificialmente, com tempo de reverberação menor que 300 *ms*, a $\overline{SDR}$ média cai para 8 *dB*, mas ainda se mantém inteligível. Outro ponto observado é o fato de que, para tempos de reverberação menores que 300 *ms*, os resultados são praticamente os mesmos obtidos para misturas anecóicas. Nos testes subjetivos realizados verifica-se boa inteligibilidade dos sinais recuperados utilizando máscaras ideais, com taxas de acertos superiores a 90%, para misturas com tempo de reverberação menores que 300 *ms*. Ao aplicar a máscara ideal a uma mistura (artificial) de dois sinais de voz com tempo de reverberação longo (3,9 *s*), verifica-se que a $\overline{SDR}$ obtida é igual a 3 *dB*. Para análise do efeito da reverberação na utilização de máscaras ideais na separação dos sinais fonte em condições reais, foram feitas gravações em ambientes com tempos de reverberação iguais a 100 *ms* e 1100 *ms*, obtendo-se relações sinal-distorção médias de 12 *dB* e 8 *dB*, respectivamente. À medida em que aumenta-se o tempo de reverberação, a separação de sinais fonte via mascaramento espectral tem sua eficiência diminuída devido ao espalhamento das frequências ao longo do tempo, reduzindo sua esparsidade.

Outro ponto observado, refere-se aos espectrogramas de sinais *TSP*. Verifica-se que, no caso de simulações, com o aumento do tempo de reverberação, ocorre um espalhamento uniforme de todas as frequências ao longo do tempo. Ao observar os espectrogramas dos mesmos sinais *TSP* para o caso de gravações reais em ambientes com tempos de reverberação iguais a 100 *ms* e 1100 *ms*, respectivamente, verifica-se que o espalhamento das frequências ao longo do tempo não acontece de forma uniforme como nos resultados obtidos através de simulações. Este fato mostra que o processo de simulação de reverberação usado neste estudo não descreve exatamente o comportamento de ambientes reais.

Dado que encontrar máscaras ideais com base em apenas uma mistura é ainda um problema em aberto, analisou-se a queda de desempenho do processo de separação em função da distância entre a máscara ideal e a máscara efetivamente usada. Os resultados encontrados indicam uma queda de desempenho de 3 *dB* quando aproximadamente 10% dos bits da máscara ideal estão invertidos.

Atualmente, os algoritmos propostos para a separação cega de fontes utilizando apenas uma mistura baseiam-se na estimação de máscaras ideais binárias através de um banco de dados contendo informações para treinamento do sistema. Visando a implementação de

um sistema de separação cega de sinais que utilize apenas uma mistura e não necessite de treinamento, foram realizadas análises preliminares para encontrar a máscara ideal com base na negentropia, na curtose e na energia dos sinais separados. A solução para este problema, entretando, permanece como um desafio.

Referências Bibliográficas

- Aichner, R., Buchner, H., Yan, F., e Kellermann, W. Real-time convolutive blind source separation based on a broadband approach. *Proceedings IEEE International Symposium on Independent Component Analysis and Blind Source Separation (ICA)*, páginas 840–848, 2004.
- Aires, F., Chedin, A., e Nadal, J. P. Independent component analysis of multivariate time series. *Application to the tropical SST variability, J. Geophys. Res.*, 105(D13):17–437, 2000.
- Amari, S. Differential-geometrical methods in statistics. *Lecture Notes in Statistics*, 28, 1985.
- Amari, S., Cichocki, A., e Yang, H. H. A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems 8*, 1996.
- Amari, S., Cichocki, A., e Yang, H. H. Multichannel blind deconvolution and equalization using the natural gradient. *IEEE International Workshop on Wireless Communication*, páginas 101–104, 1997.
- Anemuller, J. e Kollmeier, B. Adaptive separation of acoustic sources for anechoic conditions: a constrained frequency domain approach. *Speech Commun.*, 39(1-2):79–95, 2003.
- Aoshima, N. Computer-generated pulse signal applied for sound measurement. *J. Acoust. Soc. Am.*, 69(5):1484–1488, 1981.
- Araki, S., Makino, S., Aichner, R., Nishikawa, T., e Saruwatari, H. Subband based blind source separation with appropriate processing for each frequency band. *In Proc. Int. Conf. ICA and BSS*, páginas 499–504, abril de 2003a.
- Araki, S., Makino, S., Sawada, H., e Mukai, R. Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ica. *Fifth International Conference on Independent Component Analysis and Blind Signal Separation*, páginas 898–905, 2004.

- Araki, S., Mukai, R., Makino, S., Nishikawa, T., e Saruwatari, H. The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Transactions on Speech and Audio Processing*, 11(2):109–116, março de 2003b.
- Bach, F. e Jordan, M. Blind one-microphone speech separation: A spectral learning approach. *Neural Information Processing Systems (NIPS)*, 2004.
- Back, A. D. e Tsoi, A. C. Blind deconvolution of signals using a complex recurrent network. *Proceedings of the 1994 IEEE Workshop, Neural Networks for Signal Processing*, páginas 565–574, 1994.
- Barros, A. K., Rutkowski, T., Mansour, A., Itakura, F., e Ohnishi, N. Estimation of speech embedded in a reverberant environment with multiple sources of noise. *IEEE Transactions on Neural Networks*, 13(4):888–893, 2002.
- Barros, A.K., Mansour, A., e Ohnishi, N. Removing the artifacts from eletrocardiographic signals using independent component analysis. *Neurocomputing*, 22(1-3):173–186, 1998.
- Bell, A. J. e Sejnowski, T. J. An information maximisation approach to blind signal separation. *Neural Computation*, 7(6):1129–1159, 1995.
- Belouchrani, A. e Cardoso, J. F. Maximum likelihood source separation for discrete sources. *In Proc. EUSIPCO*, páginas 768–771, 1994.
- Belouchrani, A., Meraim, K. Abed, Cardoso, J. F., e Moulines, E. A blind source separation technique based on second order statistics. *IEEE Trans. on Signal Proc.*, 45(2): 434–444, fevereiro de 1997.
- Benveniste, A., Goursat, M., e Ruget, G. Robust identification of a nonminimum phase system: Blind adjustment of a linear equalizer in data communications. *IEEE Transactions on Automatic Control*, 25(3), 1980.
- Blin, A., Araki, S., e Makino, S. Underdetermined blind separation of convolutive mixtures of speech using time-frequency mask and mixing matrix estimation. *IEICE Trans Fundamentals*, (E88-A):1693–1700, julho de 2005.
- Bradaric, I., Petropulu, A.P., e Diamantaras, K.I. On blind identifiability of fir-mimo systems with cyclostationary inputs using second order statistics. *IEEE Transactions on Signal Processing*, 51(2):434–441, fevereiro de 2003.
- Bregman, A. S. Auditory scene analysis: the perceptual organization of sound. *Cambridge, MIT Press*, página 773, 1990.

- Brown, G.J. e Wang, D. L. *Speech Enhancement*, chapter Separation of speech by computational auditory scene analysis, páginas 371–402. Springer, New York, 2005.
- Buchner, H., Aichner, R., e Kellermann, W. Blind source separation for convolutive mixtures exploiting nongaussianity, nonwhiteness, and nonstationarity. *8th International Workshop on Acoustic Echo and Noise Control (IWAENC)*, páginas 275–278, setembro de 2003.
- Buchner, H., Aichner, R., e Kellermann, W. Trinicon-based blind system identification with application to multiple-source localization and separation. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, páginas 889–892, 2004.
- Campbell, D. R., Palomäki, K. J., e Brown, G. J. A matlab simulation of shoebox room acoustics for use in research and teaching. Technical report, 2007.
- Capdevielle, V., Serviere, Ch., e Lacoume, J. L. Blind separation of wide-band sources in the frequency domain. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 3:2080–2083, 1995.
- Cardoso, J. F. Source separation using higher order moments. *In Proc. ICASSP*, páginas 2109–2112, 1989.
- Cardoso, J. F. e Laheld, B. Equivariant adaptive source separation. *IEEE Trans. Signal Processing*, 44(12):3017–3030, 1996.
- Cardoso, J.F. e Souloumiac, A. Blind beamforming for non-gaussian signals. *IEE Proceedings F: Radar and Signal Processing*, 140:362–370, dezembro de 1993.
- Carvalho, C. R. Análise em componentes independentes baseada em um método geométrico-probabilístico. Dissertação de mestrado, Universidade Federal de Minas Gerais, 2000.
- Chan, Dominic C.B. *Blind signal separation*. Tese de doutorado, Department of Engineering, University of Cambridge, janeiro de 1997.
- Charkani, N. e Deville, Y. Optimization of the asymptotic performance of time-domain convolutive source separation algorithms. In *ESANN*, 1997.
- Chazan, D., Medan, Y., e Shvadron, U. Noise cancellation for hearing aids. *IEEE Trans. Acoust. Speech, Signal Processing*, 36(11):1697–1705, novembro de 1988.
- Cherry, e Collin, E. Recognition of speech with one and with two ears. *Journal of the Acoustical Society of America*, 24:975–979, 1953.

- Choi, S. e Cichocki, A. Adaptive blind separation of speech signals: Cocktail party problem. *In International Conference on Speech Processing*, páginas 617–622, 1997.
- Choi, S. e Cichocki, A. A hybrid learning approach to blind deconvolution of mimo systems. *In IEEE Signal Processing Workshop on Higher-order Statistics*, páginas 292–295, 1999.
- Choi, S., Cichocki, A., Park, Hyung-Min, e Lee, Soo-Young. Blind source separation and independent component analysis: a review. *Neural Information Processing - Letters and Reviews*, 6(1), 2005.
- Cichocki, A. e Unbehauen, R. Robust neural networks with on-line learning for blind identification and blind separation of sources. *IEEE Trans. Circuit and Systems-I*, CAS-43:894–906, 1996.
- Comon, P. Independent component analysis, a new concept? *Signal Processing, Elsevier*, 36:287–314, 1994.
- Comon, P. e Rota, L. Blind separation of independent sources from convolutive mixtures. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E86-A(3):542–549, 2003.
- Cristescu, R., Ristaniemi, T., Joutsensalo, J., e Karhunen, J. Blind separation of convolved mixtures for cdma systems. *Proceedings of the X European Signal Processing Conference (EUSIP 2000)*, páginas 619–622, 2000.
- Darmois, G. Analyse generale des liaisons stochastiques. *Rev. Inst. Internat. Stat.*, 21: 2–8, 1953.
- Delfosse, N. e Loubaton, P. Adaptive blind separation of independent sources: A deflation approach. *Signal Processing*, 45:59–83, 1995.
- Dinç, A. e Bar-Ness, Y. Bootstrap: A fast adaptive signal separator. *Proc. ICASSP*, páginas 325–328, 1992.
- Douglas, S. C. Blind source separation and independent component analysis: a cross roads of tools and ideas. *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japão, abril de 2003.
- Feder, M., Weinstein, E., e Oppenheim, A.V. Multi-channel signal separation by decorrelation. *IEEE Trans. Speech and Audio Processing*, 1(4):405–413, outubro de 1992.

- Fiori, S. Overview of independent component analysis technique with an application to synthetic aperture radar (sar) imagery processing. *Neural Networks for Analysis of Complex Scientific Data: Astronomy, Geology and Geophysics*, outubro de 2002.
- Fiori, S. e Celledoni, E. Neural learning by geometric integration of reduced 'rigid-body' equations. *Journal of Computational and Applied Mathematics archive*, 172(2):247–269, 2004.
- Furukawa, M., Hioka, Y., Ema, T., e Hamada, N. Introducing new mechanism in the learning process of fdica-based speech separation. *International Workshop on Acoustic Echo and Noise Control (IWANENC2003)*, páginas 291–294, 2003.
- Gandhi, M. A. e Hasegawa-Johnson, M. A. Source separation using particle filters. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, III: 2449–2452, 2004.
- Gerven, S. Van e Compernelle, D. Van. Signal separation by symmetric adaptive decorrelation: stability, convergence and uniqueness. *IEEE Transaction on Signal Processing*, 43(7):1602–1612, 1995.
- Girolami, M. e Fyfe, C. Extraction of independent signal sources using a deflationary exploratory projection pursuit network with lateral inhibition. *IEEE Proceedings - Vision, Image and Signal Processing*, 14(5):299–306, 1997.
- Godfrey, R. An information theory approach to deconvolution. *Stanford exploration project*, 15:157–182, 1978.
- Gorokhov, A. e Loubaton, Ph. Subspace based techniques for second order blind separation of convolutive mixtures with temporally correlated sources. *IEEE Transactions on Circuits and Systems*, 44:813–820, setembro de 1997.
- Guddeti, R. R. e Guddeti, B. Perceptually motivated blind source separation of convolutive mixtures. *In IEEE Conference on Acoustics, Speech, and Signal Processing*, 5: 273–276, março de 2005.
- Hashemi, S. e Hammond, J. K. Ill-conditioning of non-minimum phase systems. *EUSIPCO - European Signal Processing Conference*, 1996.
- Haykin, S. e Veen, B. Van. *Signal and Systems*. John Wiley and Sons, Inc., 1999. ISBN 0-471-13820-7.

- Hornillo-Mellado, S., Puntonet, C.G., Martín-Clemente, R., e Rodríguez-Álvarez, M. Characterization of the sources in convolutive mixtures: A cumulant-based approach. In *ICA*, páginas 586–593, 2004.
- Hérault, J. e Jutten, C. Space or time adaptive signal processing by neural network models. *AIP Conference Proceedings*, 151:206–211, 1986.
- Hu, G. e Wang, D. L. Speech segregation based on pitch tracking and amplitude modulation. in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, páginas 79–82, 2001.
- Hu, G. e Wang, D. L. *Topics in Acoustic Echo and Noise Control*, chapter An auditory scene analysis approach to monaural speech segregation, páginas 485–515. Springer, 2006.
- Hu, G. e Wang, D. L. Auditory segmentation based on onset and offset analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):396–405, fevereiro de 2007.
- Hyvarinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634, 1999.
- Hyvarinen, A. e Oja, E. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, outubro de 1997.
- Hyvarinen, A. e Oja, E. Independent component analysis by general non-linear hebbian-like rules. *Signal Processing*, 64(3):301–313, 1998.
- Hyvarinen, A. e Oja, E. Independent component analysis: algorithms and applications. *Neural Networks, Elsevier*, (13):411–430, março de 2000.
- Icart, S. e Gautier, R. "blind separation of convolutive mixtures using second and fourth order moments". *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 5:3018–3021, 1996.
- Ikeda, S. e Murata, N. A method of ica in time-frequency domain. *International conference on independent component analysis and signal separation*, páginas 365–371, 1999.
- Jang, G. J. e Lee, T. W. A probabilistic approach to single channel blind signal separation. *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, páginas 1173–1180, 2002.

- Jourjine, A., Rickard, S., e Yilmaz, O. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. *In IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP2000)*, 5:2985–2988, junho de 2000.
- Jutten, C. e Héroult, J. Blind separation of sources, part i: an adaptive algorithm based on neuromimetic architecture. *Signal Processing, Elsevier*, 24(1):1–10, julho de 1991.
- Jutten, C., Thi, L. Nguyen, Dijkstra, E., Vittoz, E., e Caelen, J. Blind separation of sources: An algorithm for separation of convolutive mixtures. *Proceedings of the International Signal Processing Workshop in Higher Order Statistics*, páginas 273–276, 1991.
- Karhunen, J., Hyvarinen, A., Vigario, R., Hurri, J., e Oja, E. Applications of neural blind separation to signal and image processing. *Proc. International Conference on Acoustics, Speech and Signal Processing (IEEEICASSP)*, 1:131–134, 1997.
- Kiviluoto, K. e Oja, E. Independent component analysis for parallel financial time series. *Proceedings of the ICONIP-1998*, páginas 895–898, 1998.
- Knezevic, D. *Blind Source Separation for Signal Processing Applications*. Tese de doutorado, School of Electrical, Electronic and Computer Engineering, University of Western Australia, novembro de 2004.
- Lambert, R. e Nikias, C. Polynomial matrix whitening and application to the blind deconvolution problem, 1995. URL citeseer.ist.psu.edu/lambert95polynomial.html.
- Lambert, R. H. A new method for source separation. *In Proceedings, IEEE Conference on Acoustics, Speech, and Signal Processing*, páginas 2116–2119, 1995.
- Lambert, R. H. *Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures*. Tese de doutorado, University of Southern California, 1996.
- Lee, B., Goudeseune, C., e Hasegawa-Johnson, M. A. Open-loop multi-channel inversion of room impulse response. Technical report, Electrical and Computer Engineering, University of Illinois at Urbana, Champaign, 2003.
- Lee, J. H., Jung, H. Y., Lee, T. W., e Lee, S. Y. Speech feature extraction using independent component analysis. *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3:1631–1634, 2000.
- Lee, T. W. Examples of blind source separation of recorded speech and music signals, 1998.

- Lee, T. W. e Jang, G. J. The statistical structures of male and female speech signals. *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001.
- Lindgren, U. e Veen, A. J. Van Der. Source separation based on second order statistics: an algebraic approach. *In Proc. of IEEE SP Workshop on SSAP, Corfu, Grécia*, junho de 1996.
- Lindgren, U. A. e Broman, H. Source separation using a criterion based on second-order statistics. *IEEE Transaction on Signal Processing*, 46(7):1837–1850, 1998.
- Makeig, S., Bell, A.J., Jung, T.-P., e Sejnowski, T. Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems*, 8 (145-151), 1998.
- Mansour, A., Barros, A.K., e Ohnishi, N. Blind separation of sources: methods, assumptions and applications. *Transactions on Fundamentals of Electronics, Communications and Computer Sciences, IEICE*, E83-A:1498–1512, 2000.
- Mansour, A., Jutten, C., e Loubaton, P. Subspace method for blind separation of sources and for a convolutive mixture model. *Signal Processing VIII, Theories and Applications, Elsevier*, páginas 2081–2084, setembro de 1996.
- McKeown, M., Makeig, S., Brown, S., Jung, T.-P., Kindermann, S., Bell, A.J., Iragui, V., e Sejnowski, T. Blind separation of functional magnetic resonance imaging (fmri) data. *Human Brain Mapping*, 6(5-6):368–372, 1998.
- Murata, N., Ikeda, S., e Ziehe, A. An approach to blind source separation based structure of speech signals. *Neurocomputing*, 41(1-4):1–24, agosto de 2001.
- Neely, S. T. e Allen, J. B. Invertibility of a room impulse response. *J. Acoust. Soc. Am.*, 66(1):165–169, 1979.
- Nix, J., Kleinschmidt, M., e Hohmann, V. Computational auditory scene analysis by using statistics of high-dimensional speech dynamics and sound source direction. *Proc. of Eurospeech*,, páginas 1441–1444, 2003.
- O’Grady, P. D., Pearlmutter, B. A., e Rickard, S. T. Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology (IJIST)*, páginas 18–33, 2005a.

- O'Grady, P. D., Pearlmutter, B. A., e Rickard, Scott T. Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology, Blind Source Separation and Deconvolution in Imaging and Image Processing*, abril de 2005b.
- Olsson, R. K. e Hansen, L. K. A harmonic excitation state-space approach to blind separation of speech. In *NIPS*, 2004a.
- Olsson, R. K. e Hansen, L. K. Probabilistic blind deconvolution of non-stationary sources. In *12th European Signal Processing Conference*, páginas 1697–1700, setembro de 2004b.
- Parra, L. e Spence, C. Convolutional blind source separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, páginas 320–327, maio de 2000.
- Parra, L., Spence, C., e Vries, B. De. Convolutional blind source separation based on multiple decorrelation, 1998. URL citeseer.ist.psu.edu/parra98convolutive.html.
- Pedersen, M. S., Larsen, J., Kjems, U., e Parra, L. C. A survey of convolutional blind source separation methods. In *Springer Handbook of Speech (to appear)*. Springer Press, setembro de 2007. URL <http://www2.imm.dtu.dk/pubdb/p.php?4924>.
- Prasad, R. K., Saruwatari, H., e Shikano, K. Problems in blind separation of convolutional speech mixture by negentropy maximization. *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, páginas 287–290, 2003.
- Puntonet, C.G., Prieto, A., Jutten, C., Rodriguez-Alvarez, M., e Ortega, J. Separation of sources: A geometry-based procedure for reconstruction of n-valued signals. *Signal Processing*, 46:267–284, 1995.
- Raki, S. A., Makino, S., Sawada, H., e Mukai, R. Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, III:81–84, 2005.
- Rao, C. R. Characterisation of the distribution of random variables in linear structural relations. *Sankhya, Series A*, 28:251–260, 1966.
- Reddy, A. M. e Raj, B. A minimum mean squared error estimator for single channel speaker separation. páginas 2445–2448, 2004a.
- Reddy, A. M. e Raj, B. Soft mask estimation for single channel speaker separation. 2004b.
- Rickard, S., Melia, T., e Fearon, C. Desprits - histogram based blind source separation of more sources than sensors using subspace methods. *IEEE Workshop on this paper*

- appears in: *Applications of Signal Processing to Audio and Acoustics*, páginas 5–8, 2005.
- Rickard, S. e Yilmaz, O. On the approximate w-disjoint orthogonality of speech. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3:3049–3052, 2002.
- Rindel, J. H. The use of computer modeling in room acoustics. *Journal of Vobroengineering*, 3(4):41–72, 2000.
- Roman, N. e Wang, D. L. Binaural sound segregation for multisource reverberant environments. in *Proc. IEEE ICASSP*, 2:373–376, 2004.
- Roman, N., Wang, D. L., e Brown, G. J. Speech segregation based on sound localization. *J. Acousit. Soc. Am.*, 114:2236–2252, 2003.
- Roweis, S. T. One microphone source separation. In *Neural Information Processing Systems (NIPS)*, páginas 793–799, 2000. URL citeseer.ist.psu.edu/roweis00one.html.
- Rutkowski, T., Cichocki, A., e Barros, A. K. Speech enhancement from interfering sounds using casa techniques and blind source separation. In: *3rd International Conference on Independent Component Analysis and Source Separation*, 2001.
- Sanchez, V. D. Frontiers of research in bss/ica. *Neuro Computing, Elsevier*, (49):7–23, outubro de 2002.
- Sato, Y. A method of self-recovering equalization for multilevel amplitude-modulation system. *IEEE Trans. on Communications*, 23:679–682, 1975.
- Savioja, L. *Modeling techniques for virtual acoustics*. Tese de doutorado, Helsinki University of Technology, agosto de 2000.
- Schobben, D. W. E e Sommen, P. C. W. On the indeterminacies of convolutive blind signal separation based on second order statistics. *Proceedings of the Fifth International Symposium on Signal Processing and Its Applications (ISSPA)*, 1:215–218, 1999.
- Schroeder, M.R. New method of measuring reverberation time. *J. Acoust. Soc. Am.*, 1965.
- Sharnsunder, S. e Giannakis, G. Multichannel blind signal separation and reconstruction. *IEEE Transaction on Speech and Audio Processing*, 5(6):515–528, novembro de 1997.

- Shashanka, M.V.S., Raj, B., e Smaragdis, P. Sparse overcomplete decomposition for single channel speaker separation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2:641–644, abril de 2007.
- Smaragdis, P. J. Information theoretic approaches to source separation. Dissertação de metrado, Massachusetts Institute of Technology, junho de 1997.
- Suzuki, Y., Asano, F., Kim, H.-Y., e Sone, T. An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses. *J. Acoust. Soc. Am.*, 97(2):1119–1123, 1995.
- Thawonmas, R., Cichocki, A., e Amari, S. A cascade neural network for blind signal extraction without spurious equilibria. *IEICE Trans. Fundamentals*, E81-A(9):1833–1846, 1998.
- Thomas, J., Deville, Y., e Hosseini, S. Time-domain fast fixed-point algorithms for convolutive ica. *IEEE Signal Processing Letters*, 13(4):228–231, abril de 2006.
- Thurstone, L. L. Multiple factor analysis. *Psychol. Rev.*, 38:406–427, 1931.
- Tordini, F. e Piazza, F. A semi-blind approach to the separation of real world speech mixtures. *Proc. of IEEE Int. Joint Conf. on Neural Networks (IJCNN2002)*, 2:1293–1298, 2002.
- Torkkola, K. Blind separation of convolved sources based on information maximization. *IEEE Workshop on Neural Networks for Signal Processing, Kyoto*, setembro de 1996.
- Torkkola, K. Blind separation of radio signals in fading channels. *Adv. Neural Inform. Process. Systems*, 1998.
- Vigario, R., Sarela, J., Jousmaki, V., e Oja, E. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. *Adv. Neural Inform. Process. Systems*, 10:229–235, 1998.
- Vincent, E., Gribonval, R., e F'evotte, C. Performance measurement in blind audio source separation. *IEEE Transactions on Speech and Audio Processing*, 14(4):1462–1469, 2006.
- Wang, W., Chambers, J. A., e Sanei, S. A joint diagonalization method for convolutive blind separation of nonstationary sources in the frequency domain. *ICA2003*, páginas 939–944, 2003a.
- Wang, W., Jafari, M. G., Sanei, S., e Chambers, J. A. Blind separation of convolutive mixtures of cyclostationary sources using an extended natural gradient method. *IEEE*

- International Symposium on Signal Processing and its Applications*, páginas 93–96, 2003b.
- Weintraub, M. The grasp sound separation system. *In IEEE Conference on Acoustics, Speech, and Signal Processing*, 9:68–72, 1984.
- Weiss, R. e Ellis, D. Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking. *Proc. Workshop on Statistical and Perceptual Audition (SAPA)*, 31-36, outubro de 2006.
- Westner, A. e Bove, V. M. Blind separation of real world audio signals using overdetermined mixtures. *In First International Conference on Independent Component Analysis*, 1999.
- Wiggins, R. A. Minimum entropy deconvolution. *Geoexploration*, páginas 21–35, 1978.
- Winter, S., Kellermann, W., Sawada, H., e Makino, S. Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1 - norm minimization. *Journal on Advances in Signal Processing (EURASIP)*, 2007.
- Winter, S., Sawada, H., Araki, S., e Makino, S. Overcomplete bss for convolutive mixtures based on hierarchical clustering. *Proceedings of International Workshop on Independent Component Analysis and Blind Signal Separation (ICA '04)*, páginas 652–660, setembro de 2004.
- Yang, H. H. e Amari, S. Adaptive online learning algorithms for blind separation: Maximum entropy and minimum mutual information. *Neural Computation*, 9:1457–1482, 1997.
- Yilmaz, O. e Rickard, S. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, julho de 2004a.
- Yilmaz, O. e Rickard, S. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, julho de 2004b.
- Yin, B. e Sommen, P.C.W. Adaptive blind signal separation using a new simplified mixing model. *Proc. ProRISC 99, 10th Annual Workshop on Circuits, Systems and Signal Processing, ISBN 90-73461-18-9, ed. J.P. Veen*, páginas 601–606, 1999.
- Yin, B. e Sommen, P.C.W. A new convolutive blind signal separation algorithm based on second order statistics using a simplified mixing model. *Proc. Eusipco 2000, European Signal Processing Conference*, páginas 2049–2053, 2000.