Universidade Federal de Minas Gerais
Escola de Engenharia
Curso de Graduação em Engenharia Elétrica

Separação de Fontes de Áudio

Kayke Renan Campos Silva

Orientador: Prof. Adriano Vilela Barbosa

Belo Horizonte, Dezembro de 2023

Monografia

Separação de Fontes de Áudio

Monografia submetida à banca examinadora designada pelo Colegiado Didático do Curso de Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, como parte dos requisitos para aprovação na disciplina Trabalho de Conclusão de Curso.

Belo Horizonte, Dezembro de 2023

Resumo

Ainda a ser realizado!

Agradecimentos

Aqui vai o texto dos agradecimentos.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A separação de fontes de áudio, em inglês *Audio Source Separation (ASS)*, consiste no problema de simular a capacidade do sistema auditivo humano de extrair uma fonte de áudio de interesse dentre outras fontes misturadas e ruídos. Este problema é conhecido na literatura como o "efeito da festa de coquetel"(*the cocktail party effect*) ou "escuta seletiva"[1], descrito como um ambiente de festa com a presença de diversas pessoas conversando, possivelmente com música e outros ruídos onde um indivíduo é capaz de ignorar e suprimir as diversas fontes sonoras enquanto foca e extrai apenas aquela(s) que o interessa, seja uma conversa, música, instrumento(s), etc.

1.1 Motivação e Justificativa

A separação de fontes de áudio (ASS) é um caso especial do problema geral de separação de fontes na área de processamento de sinais que decorre do fato de que os sinais no mundo real geralmente vêm misturados com outros sinais e ruídos, especialmente no domínio eletromagnético pelo fenômeno de acoplamento que ocorre quando os sinais compartilham um mesmo canal de transmissão e/ou ambiente de geração, por exemplo os sinais biológicos captados por eletrodos, ondas eletromagnéticas em telecomunicações, etc.

Quando os diferentes sinais possuem bandas espectrais distintas, o problema de separação pode ser solucionado através da filtragem, como ocorre nos sistemas de telecomunicações através da modulação e demodulação, mas quando os espectros se sobrepõem, a solução é mais difícil por se tratar de um problema inverso [2]. Neste tipo de problema busca-se, a partir de observações de um fenômeno em um determinado domínio, inferir as causas ou os melhores modelos, geralmente em um outro domínio, que explicam e/ou aproximam os dados observados. Como exemplo, pode-se citar a descoberta de Netuno a partir de medições que indicavam desvios da previsão teórica da órbita de Urano.

A separação de fontes é um problema ainda em aberto e portanto diversas abordagens para abordá-lo foram e têm sido propostas, combinando heurísticas e diferentes métodos computacionais, assim como informações sobre os sinais e o processo de mistura quando possível. O interesse no problema de separação de fontes se justifica pelo fato de que os resultados podem ser aplicados a diversos outros problemas, como a separação de instrumentos para remixagem de músicas antigas que dispunham de poucos canais no processo de gravação, extração de voz e outras informações em gravações de áudio para uso em perícia criminal, separação de sinais biológicos em exames e pesquisas, cancelamento de ruídos em

fones de ouvidos e em microfones visando melhorar a qualidade das comunicações, dentre outras aplicações.

Este trabalho terá como foco o caso especial de separação de instrumentos musicais em gravações, tendo como inspiração os recentes resultados obtidos com a re-mixagem do disco *Revolver* da banda *The Beatles*. Na época da gravação do disco, apenas 4 canais de áudio estavam disponíveis, de forma que diversos instrumentos tiveram de ser misturados em cada trilha, tornando a separação dos instrumentos para re-mixagem um problema sem resultados satisfatórios através das técnicas clássicas, tornando-se possível apenas recentemente com a aplicação de sistemas de redes neurais profundas, como declarado pelo produtor Giles Martins, responsável pela re-mixagem da discografia da banda [3].

1.2 Objetivos do Projeto

O trabalho consiste em implementar e comparar os resultados da separação de 4 instrumentos (bateria/instrumentos per cursivos, vocais, guitarras e baixo) utilizando duas abordagens, uma mais clássica com o método computacional de fatoração de matrizes não-negativas, como descrito no estudo [4] e outra abordagem mais recente (informar sobre os períodos de tempo em que cada uma foi proposta e estudada) utilizando aprendizado de máquina a partir de redes neurais profundas com auxílio da biblioteca em Python Open-Unmix [5]. A escolha dos instrumentos se deu pela facilidade de encontrá-los em diversas músicas de gêneros diferentes, além do fato de separar em instrumentos que possuem afinação (vocais, guitarras e baixo) e que não possuem afinação (bateria e instrumentos percussivos)

A comparação dos resultados se dará a partir de um estudo com um número de pessoas ainda a ser definido, em que inicialmente será apresentada uma gravação musical com a mistura dos instrumentos e em seguida, de forma randomizada sem que o participante saiba qual método está sendo apresentado, para cada instrumento separado perguntar se alguma dos métodos teve um desempenho melhor do que o outro ou se apresentam a mesma qualidade.

Após esta etapa de seleção do melhor método para cada instrumento, será apresentado novamente para cada participante aquelas trilhas de cada instrumento que foram escolhidas por este como as melhores, sendo que no caso de empate será escolhido um dos métodos de forma aleatória, e em seguida será pedido para ordenar a qualidade de separação de cada instrumento do melhor para o pior ou se apresentam a mesma qualidade. A partir destes resultados, será possível decidir se algum método teve um desempenho estatisticamente melhor do que o outro e se dentro de cada método a qualidade de separação para cada instrumento varia ou não.

A opção pelo estudo com participantes se dá pelo fato do ser humano ser considerado o padrão ouro com relação à capacidade de separação de fontes de áudio e por não ter atualmente um modelo psicoacústico que consiga reproduzir de forma objetiva a habilidade humana. A opção pela comparação qualitativa entre os métodos vem da dificuldade de se atribuir números absolutos para a qualidade da separação que demandam métricas e métodos estatísticos mais complexos, portanto optou-se pela simplicidade além do fato de não exigir dos participantes alguma forma de conhecimento prévio da tarefa de separação dos sinais e a consequente medição da qualidade, bastando apenas sua percepção.

Capítulo 2

Metodologia e Revisão Bibliográfica

Neste capítulo, você deve apresentar uma breve revisão bibliográfica sobre as técnicas utilizadas para solução do problema.

Ao longo do tempo, diversas técnicas e abordagens foram desenvolvidas e aplicadas ao problema de ASS, levando a diferentes linhas de pesquisa em que todas têm em comum a aplicação em maior ou menor grau de conceitos e ferramentas de processamento de sinais, modelagem de fontes de áudio como fonte-filtro, álgebra linear, probabilidade e estatística [5]. Além disso, sinais de áudio multicanais (como áudio estéreo, surround 5.1, surround 7.1) demandam técnicas diferentes daquelas utilizadas em sinais que possuem apenas um canal (áudio mono), contribuindo ainda mais para a diversidade de linhas de pesquisa.

2.1 Blind Source Separation e Análise de Componentes Independentes

Historicamente, as primeiras abordagens para tratar o problema partiram de simplificações da teoria de sinais e sistemas dinâmicos lineares em que a partir de um número conhecido F de fontes e um número conhecido S de sensores (microfones), os sinais captados por cada microfone no domínio do tempo poderiam ser representados a partir da convolução entre o sinal de cada fonte e a resposta ao impulso do sistema composto pela agregação dos efeitos de propagação e interferência agindo entre cada fonte e o sinal captado pelo sensor (como a reverberação do ambiente de gravação, interferências de outras fontes, filtros dos circuitos eletrônicos dos sensores, etc.). Como pode-se notar, ainda que simplificações na forma da premissa de o número de fontes ser conhecido e o fato deste tipo de modelo de respostas ao impulso considerar que as fontes são pontuais e estacionárias [6], a identificação e modelagem do sistema dinâmico através das respostas ao impulso não é uma tarefa trivial e geralmente sequer possível, portanto, foram necessárias simplificações adicionais deste problema como a consideração da escassez de informação sobre as fontes e sobre o processo de mistura no que é conhecido na literatura como Blind Source Separation (BSS) [7] e [8], na qual a técnica de Análise de Componentes Independentes (Independent Component Analysis - ICA) é uma das mais exploradas até hoje.

A ICA faz uma série de suposições simplificadoras, dentre elas que as fontes são estatisticamente independentes, que a mistura pode ser modelada como uma simples combinação linear dos sinais da fonte (ou seja, a mistura é apenas instantânea e efeitos de convolução

não são modelados), além de que a solução do sistema linear formado pode ser obtida apenas quando há um número de microfones igual ou maior do que o número de fontes (S >= F). A formalização matemática da ICA assim como melhorias do método visando resolver as limitações de número de fontes, efeitos de convolução e não-linearidades do processo de mistura são exploradas em [9] dado que os sinais de áudio de forma geral não atendem às restrições impostas pela técnica, por exemplo o fato da maioria ser no formato mono ou estéreo e possuir mais do que duas fontes sonoras.

2.2 Máscaras Espectrais

Outra abordagem derivada diretamente da teoria de processamento de sinais consiste na filtragem do sinal misturado, isto é, realizar uma espécie de equalização na mistura atenuando ou amplificando determinadas frequências assumindo que diferentes fontes apresentam características e regiões espectrais diferentes, porém, como o sinal de áudio muda ao longo do tempo, é necessário que esta equalização também mude de forma correspondente, portanto, de forma a obter este efeito, utilizam-se máscaras espectrais que consistem em filtros cujas características são variantes no tempo [11], portanto, as máscaras espectrais operam tanto no domínio da frequência quanto no domínio do tempo simultaneamente, logo, sua aplicação se dá numa representação Tempo-Frequência do sinal como o espectrograma do sinal misturado. Uma breve explicação do espectrograma é feita no capítulo de fundamentos teóricos.

Trecho do capítulo de fundamentos teóricos explicando o espectrograma

O espectrograma é uma representação de como o espectro de frequências do sinal varia ao longo do tempo obtido através da transformada de Fourier (TF) de tempo curto (STFT), onde definem-se intervalos de tempo (janelas/bins) nos quais serão realizados a TF de forma sucessiva ao longo da duração do sinal. O espectrograma é obtido a partir da magnitude da TF para cada intervalo de tempo, como ilustrado na figura [REF.], onde o eixo vertical corresponde às frequências, o eixo horizontal ao tempo e a cor de cada ponto a intensidade/magnitude da TF naquela frequência naquela janela de tempo.

Existem diversas técnicas para obter máscaras espectrais de acordo com as características do sinal (como o número de canais) e premissas sobre as fontes misturadas (como o número de fontes, a distribuição espectral de cada uma, etc.) e a separação é obtida através da multiplicação matricial da máscara pela representação tempo-frequência do sinal. Para o caso em que há apenas um canal, a máscara espectral é geralmente representada como uma matriz de valores reais entre 0 e 1. Além disso, elas podem ser classificadas de acordo com os valores que podem assumir, como as máscaras binárias que são restritas apenas aos valores 0 e 1 (i.e., de forma discreta) e que são aplicadas em problemas de separação de fala [12], enquanto as máscaras que podem assumir qualquer valor real entre 0 e 1 são classificadas como soft-masks ou ratio-masks e apresentam desempenho superior às máscaras binárias como demonstrado nos estudos [13] e [14] em que a separação obtida apresenta uma melhor relação sinal-ruído (SNR). Já quando o sinal é multicanal, dificuldades adicionais como a diferença de tempo entre os sinais da mesma fonte em cada canal devem ser levadas em conta e como consequência as máscaras espectrais passam a ser representadas como matrizes de valores complexos de forma a modelar a diferença de fase entre os sinais.

2.3 Modelagem Fonte-Filtro

Outra abordagem envolve a modelagem das diferentes fontes de áudio a partir do modelo fonte-filtro constituído por um sinal de excitação (a fonte) que é alterado por um filtro que representa alguma característica do meio ou instrumento. A alteração do sinal pelo filtro se dá pela mudança do envelope espectral do sinal. Esta abordagem é muito utilizada na modelagem de voz como em [15], onde a fonte do sinal de excitação corresponde aos pulsos glotais gerados pelas cordas vocais e o filtro corresponde ao trato vocal.

Outra possibilidade é a modelagem tanto de melodias cantadas por vozes quanto de instrumentos afinados dado que em teoria musical as notas são determinadas por frequências específicas (chamadas de frequências fundamentais) e o timbre dos instrumentos é derivado do envelope espectral dos harmônicos que são múltiplos inteiros da frequência fundamental, logo, pode-se modelar a fonte como o sinal senoidal puro com a frequência fundamental e o timbre através do filtro. Além disso, esta modelagem é facilitada dado que geralmente apenas um conjunto reduzido de notas ocorrem em cada música por causa de conceitos de teoria musical como escala, tom, campo harmônico, etc.

O modelo fonte-filtro e as considerações sobre a característica harmônica de instrumentos e melodias cantadas necessitam de métodos de identificação da frequência fundamental assim como a sua correspondente variação ao longo do tempo para extração das melodias. Estes métodos em conjunto com o modelo fonte-filtro foram explorados em trabalhos como em [16] onde após realizar a identificação das frequências fundamentais ao longo da melodia através da STFT, realizou-se a sintetização das fontes através do modelo fonte-filtro de cada uma. Abordagem similar foi utilizada no trabalho [17] de forma a separar sinais musicais de voz acompanhada de piano. Porém os resultados obtidos pela sintetização a partir do modelo fonte-filtro e da harmonicidade sofrem do problema de soarem artificiais pela característica metalizada, como discutido em [6], decorrente das diferenças entre os sinais de excitação estimados para as fontes e os sinais reais, o que expõe a sensibilidade destes métodos à qualidade do modelo fonte-filtro e à qualidade de extração da melodia o que impede a aplicação em gravações que fogem das condições ideais que o método exige. Além disso, nem todos instrumentos podem ser modelados como harmônicos, como é o caso de instrumentos percussivos, demandando portanto novas abordagens.

2.4 Fatoração e Co-Fatoração de Matrizes Não-Negativas

De forma a atacar as limitações dos métodos harmônicos com modelo fonte-filtro, outras abordagens passaram a considerar a natureza repetitiva de sinais musicais, como as repetições rítmicas, as repetições de notas em escalas, as de acordes na progressão harmônica e as repetições de estruturas musicais como versos e refrões presentes em músicas populares. Essas repetições podem ser modeladas como um conjunto de estruturas básicas (features) que aparecem e são combinadas em determinados momentos, no caso de sinais musicais isso corresponde a considerar que o espectrograma da mistura pode ser decomposto em poucos espectrogramas bases que são ativados e combinados ao longo do tempo, sendo este o fundamento do algoritmo de fatoração de matrizes não-negativas (Non-negative Matrix Factorization - NMF) [18] em que uma matriz de valores não-negativos é decomposta como o produto entre outras duas matrizes também não negativas, sendo que o algoritmo é explicado com mais detalhes no capítulo de fundamentos teóricos. O algoritmo de NMF foi aplicado

em sinais musicais em [19] e [20], porém como o algoritmo não garante unicidade de soluções e por ser um caso especial de BSS, a interpretação da separação em instrumentos é prejudicada além do fato de que por ser um algoritmo de aproximação por ?low-rank?, ele não é adequado para vocais e instrumentos que apresentam menos repetições. De forma a atacar estes problemas, uma variação do algoritmo chamada de co-fatoração de matrizes não-negativas visa realizar a fatoração conjunta da matriz alvo (o espectrograma) e uma matriz contendo informação adicional como melodias, vocais e acompanhamento quando disponíveis de forma que a fatoração obtida apresenta componentes compartilhadas com as matrizes de informação, como mostrado pelos resultados em [21] e [22] para o caso de instrumentos percussivos e em [23] para vocais.

2.5 Redes Neurais Artificias

Atualmente as abordagens que têm produzido resultados mais promissores e superiores aos métodos já explorados são aquelas baseadas em aprendizado de máquina como redes neurais artificiais, dada a disponibilidade atual do grande volume de dados para treinamento. Como os métodos discutidos acima são baseados sempre em alguma espécie de modelagem que exigem determinadas suposições que quando não satisfeitas produzem resultados insatisfatórios, as abordagens por aprendizado de máquina se destacam pela ausência destas suposições, onde o modelo é obtido pelo treinamento a partir dos exemplos. A primeira aplicação das redes neurais em ASS se deu em [24] com o emprego de redes neurais recorrentes (Recurrent Neural Networks - RNN). Já em [25], redes neurais convolucionais (Convolutional Neural Networks - CNN) foram utilizadas para extração de vocais. Em [26], os autores aplicaram redes neurais feedforward (Feedforward Neural Networks - FNN) para separação de instrumentos e re-mixagem de músicas de Jazz.

Os trabalhos mais recentes buscam associar redes neurais profundas com técnicas dos modelos descritos neste capítulo, aumentando assim o desempenho destes algoritmos, como em [27] que na saída de uma RNN foi inserida uma camada extra de uma máscara espectral usada para gerar os espectros das fontes separadas e portanto o processo de aprendizado aproxima de forma indireta a máscara espectral ideal que produz a melhor separação, onde os resultados obtidos superaram os algoritmos estado da arte em separação de voz.

Os problemas associados com o emprego de redes neurais são o custo computacional no treinamento dos algoritmos e a necessidade de uma grande base de dados para treinamento, além da dificuldade de interpretação dos milhares de parâmetros dos modelos que estas redes geram [6].

2.6 Resumo do Capítulo

Neste capítulo foram apresentadas as diversas técnicas aplicadas ao problema de separação de fontes de áudio ao longo do tempo, chegando mais recentemente nas técnicas de aprendizado profundo com redes neurais que apresentam os melhores resultados em comparação às técnicas clássicas baseadas apenas nas ferramentas de processamento de sinais.

Capítulo 3

Resultados

Para a execução do projeto, algumas etapas de desenvolvimento tiveram de ser seguidas: familiarização com o sistema, estudo dos módulos envolvidos, leitura dos requisitos, elaboração de documento descrevendo todo o processo de implementação e relacionamento com os diversos módulos, implementação e testes.

3.1 Atividades do Projeto

3.2 Requisitos do Sistema

Para referenciar a Figura 3.1, veja arquivo .tex.

Aqui começa uma sub-seção.

3.3 Desenvolvimeto e Implementação

Aqui começa outra seção.

Para inserir a tabela abaixo, veja arquivo .tex.

Aqui você referencia a tabela: a Tabela 3.1 explicita os pontos mais relevantes na implementação do SRUC.

3.4 Testes

3.5 Resumo do Capítulo

Esse capítulo pode ser dividido em duas partes $f = ma$ blaba

$$f = ma \tag{3.1}$$

$$x = 2 \tag{3.2}$$

$$\tag{3.3}$$

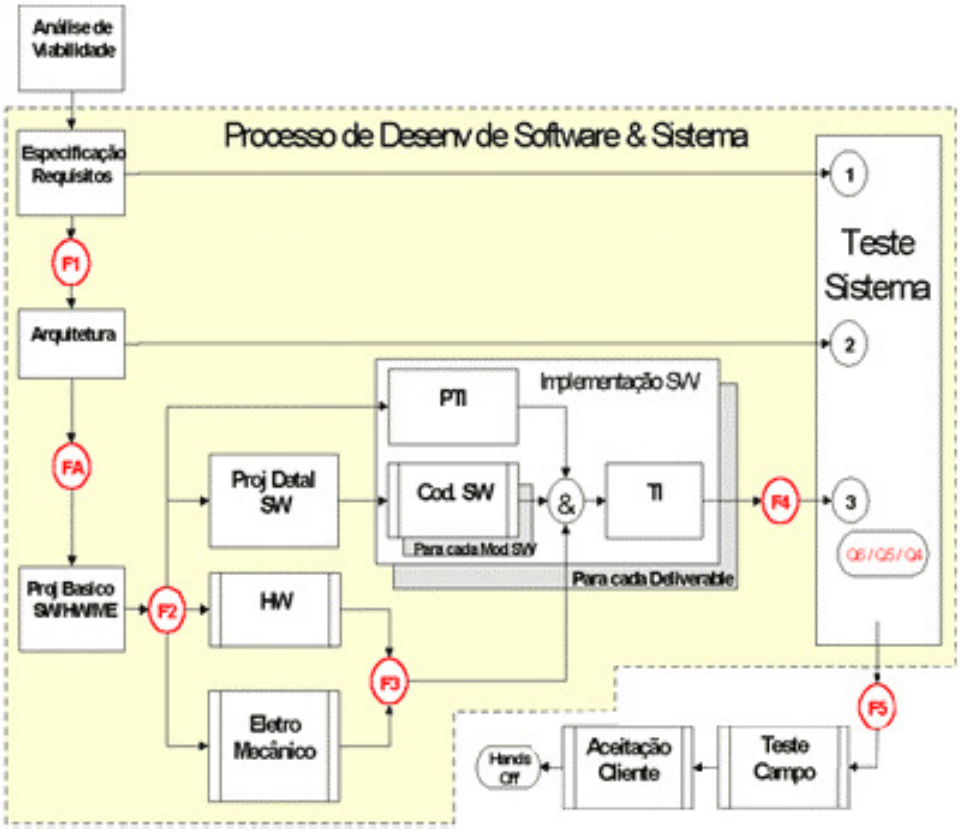

Figura 3.1: Ciclo de desenvolvimento de um projeto

1.Uso do serviço	Para o assinante rastrear uma chamada, ele deverá tirar o telefone do gancho, esperar pelo tom de discagem e então discar o código de acesso ao serviço.
2.Processamento do serviço	Caso o assinante tenha acesso ao serviço SRUC, ele deverá ouvir um anúncio, ao discar o código de acesso, explicando que o serviço SRUC foi acessado. Dessa forma, se os dados a serem rastreados forem suficientes, o sistema deverá fornecer uma mensagem de confirmação de serviço realizado
3. Ativação da última chamada recebida	A ativação do serviço somente será válida para a última chamada recebida.
4. Mais de uma ativação para a mesma chamada	Se o assinante tentar ativar o serviço para a mesma chamada ele deverá ouvir novamente o anúncio de serviço realizado, mas não irá gravar os dados novamente
5. Número privado do assinante A	O sistema deverá mostrar o número do assinante chamador mesmo que este não possa ser mostrado.
6. Chamadas intercentrais	Para que o serviço possa valer para chamadas intercentrais a central deverá utilizar a sinalização SS7, e o número do assinante A será obtido pela mensagem IAM.
7. Informações de um registro	Um *trace* do serviço deverá possuir os seguintes itens: Número do assinante A Hora da chamada recebida Data da chamada recebida Número do assinante B Hora da solicitação do serviço Data da solicitação do serviço Dados sobre rota para chamadas intercentrais
8. Tratamento para assinante sem serviço	Se um assinante discar o código de acesso ao serviço, a central deverá fornecer tratamento padrão de acesso negado.
9. Tipos de telefones	A central deve permitir que o assinante com o serviço possua tanto DTMF quando Dial Pulse
10. Comandos do sistema supervisório	O sistema supervisório conectado à central deverá disponibilizar um comando para que o operador possa descarregar o arquivo com os *traces* das chamadas para os diversos assinantes de uma central. Um comando para visualizar os *traces* também será necessário.

Tabela 3.1: Requisitos do Serviço SRUC

$$f = ma \tag{3.4}$$
$$x = 2 \tag{3.5}$$
$$\tag{3.6}$$

$$f = ma \tag{3.7}$$
$$x = 2$$
$$\tag{3.8}$$

Capítulo 4

Conclusões

4.1 Considerações Finais

Aqui vai o texto da conclusão.

4.2 Propostas de Continuidade

Referências Bibliográficas

[1] The cocktail party effect. https://en.wikipedia.org/wiki/Cocktail_party_effect. Acesso em 22 de Abril de 2023.

[2] Inverse problem. https://en.wikipedia.org/wiki/Inverse_problem. Acesso em 22 de Abril de 2023.

[3] Giles martin says that he could only remix the beatles' revolver album in stereo because of peter jackson's ai-powered audio separation technology: "it opened the door". https://www.musicradar.com/news/the-beatles-revolver-stereo-remix-ai-tech. Acesso em 22 de Abril de 2023.

[4] Alan Freihof Tygel. MÃ©todos de fatoraÃ§Ã£o de matrizes nÃ£o-negativas para separaÃ§Ã£o de sinais musicais. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Aug 2009.

[5] Z. Rafii, A. Liutkus, F. R. StÃ¶ter, S. I. Mimilakis, D. FitzGerald, and B. Pardo. An overview of lead and accompaniment separation in music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(8):1307–1335, Aug 2018.

[6] R. Gribonval E. Vincent, N. Bertin and F. Bimbot. From blind to guided audio source separation: How models and side information can improve the separation of sound. *IEEE Signal Processing Magazine*, 31(3):107–115, May 2014.

[7] P. Common and C. Jutten. *Handbook of blind source separation*. Academic Press, 2010.

[8] G. R. Naik and W. Wang. *Blind source separation*. Springer-Verlag Berlin Heidelberg, 2014.

[9] Karhunen J. Oja E. Hyvarinen, A. *Independent Component Analysis*. New York, John Wiley and Sons, 2001.