UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

JOÃO PEDRO MENDES DE OLIVEIRA

O uso de ferramentas computacionais na remoção de vazamentos sonoros em gravações de bateria

São João del-Rei
2024

JOÃO PEDRO MENDES DE OLIVEIRA

O uso de ferramentas computacionais na remoção de vazamentos sonoros em gravações de bateria

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação. Universidade Federal de São João del-Rei.

Flávio Luiz Schiavoni
Orientador

José Guilherme Allen Lima
Coorientador

Henrique Maia Lins Vaz
Membro externo

João Pedro Hallack Sansão
Membro interno

São João del-Rei
2024

Ficha catalográfica elaborada pela Divisão de Biblioteca (DIBIB)
e Núcleo de Tecnologia da Informação (NTINF) da UFSJ,
com os dados fornecidos pelo(a) autor(a)

O48u

Oliveira, João Pedro Mendes de.
O uso de ferramentas computacionais na remoção de vazamentos sonoros em gravações de bateria / João Pedro Mendes de Oliveira ; orientador Flávio Luiz Schiavoni; coorientador José Guilherme Allen Lima. - São João del-Rei, 2024.
155 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciência da Computação) -- Universidade Federal de São João del-Rei, 2024.

1. Engenharia de áudio. 2. Separação de fontes de áudio. 3. Separação de fontes de bateria. 4. Recuperação de informação musical.. I. Schiavoni, Flávio Luiz, orient. II. Lima, José Guilherme Allen, co-orient. III. Título.

Este trabalho é dedicado a Maria Mathilde Mendes Gotardelo, minha tia querida, educadora e professora aposentada da Universidade Federal de Juiz de Fora (UFJF). Um testemunho vivo de como a educação tem o poder de transformar o indivíduo e o mundo ao seu redor.

Através de seu exemplo como educadora, Mathilde influenciou e continua a influenciar, ainda hoje, aos 92 anos, gerações dentro da família Mendes — uma família amplamente composta por professores, mestres, pedagogos e profissionais licenciados. Por meio do trabalho dessas pessoas, diversas vidas foram transformadas ao longo dos anos, seja no magistério, no ensino público — federal, estadual e municipal —, no ensino privado ou em qualquer outra modalidade na qual tenham trabalhado. Tudo isso é fruto de uma semente plantada há muito tempo por Mathilde e, da mesma forma, este mestrado também é um reflexo dessa semente!

Mathilde é bacharel e licenciada em Letras Neo-Latinas pela Faculdade de Filosofia e Letras de Juiz de Fora, bacharel em Direito pela UFJF, mestre em Educação pela Universidade Federal Fluminense (UFF), e portadora de diversos títulos de pós-graduação Lato Sensu. Foi professora da Faculdade de Educação e do Curso de Especialização em Engenharia de Segurança do Trabalho, ambos na UFJF. Sua atuação enquanto docente na universidade gerou frutos que reverberam até hoje em nossa cidade, como o exemplo da atual prefeita de Juiz de Fora, que foi aluna de uma de suas turmas.

Além disso, Mathilde teve uma trajetória educacional significativa, atuando como diretora do Curso Normal do Instituto Metodista Granbery, professora na Escola Estadual Professor Sebastião Patruz de Souza e como professora de Educação Física no Colégio São José (atualmente Instituto Viana Jr), todos em Juiz de Fora. Também contribuiu para a fundação da Faculdade de Filosofia, Ciências e Letras nas cidades de Além Paraíba e Cataguases.

Desde muito cedo, tia Mathilde me ensinou a importância da interdisciplinaridade. Quando eu era criança, ainda com os sonhos de ser astronauta, arqueólogo ou biólogo — carreiras motivadas pela magia da infância — sempre encontrei nela apoio e incentivo. Meu primeiro livro de ciências, "Segredos do Mundo Animal", presente dado por ela, permanece guardado com muito carinho até hoje.

Durante o período pré-universitário, cogitei me aventurar por diversas áreas, como História, Música, Arquitetura, Engenharia e outras. Mesmo com as dúvidas na escolha, encontrei em tia Mathilde o apoio necessário em cada etapa dessa jornada. Afinal, para ela, o que importava não era o curso em si, mas o acesso ao conhecimento transformador por trás do ensino de cada área. Demorei, mas aprendi essa lição!

Quando arrependido da minha escolha acadêmica e decidido a redirecionar os meus estudos para o campo da música e das artes, recebi o apoio da tia Mathilde, que na época me presenteou com

parte de sua coleção de discos de vinil. As coleções "Grandes Compositores da Música Universal" e "The Wonderful World of Music" ainda estão guardadas comigo, com muito carinho e apreço. Se hoje utilizo o conhecimento adquirido em Engenharia e Computação para contribuir no avanço das pesquisas em Música, foi porque, nesse momento crucial, encontrei grande apoio e incentivo dela.

Por fim, esta dedicatória também é motivada pelo imenso carinho que tia Mathilde sempre dispensou à minha família. Graças à sua condição financeira, conquistada por meio de seu acesso aos estudos, ela ajudou meus avós, Josias e Gertrudes, a construírem seu lar e sua vida. Além disso, foi uma incentivadora fundamental da minha mãe, Christina, durante toda a construção de sua carreira na educação pública, seja por meio de conselhos ou de apoio material. Jamais esquecerei das memórias de infância, quando os carros novos adquiridos pela minha mãe se pareciam sempre com os carros antigos de tia Mathilde. Na época, eu não entendia. Hoje, eu compreendo!

Por todo o carinho e exemplo que transformaram a nossa família e que me permitiram chegar até aqui, é que dedico este trabalho com todo o meu carinho à minha querida tia Mathilde.

AGRADECIMENTOS

Certa vez, me deparei com uma frase de um pensador romano que me chamou muita atenção: “A gratidão é a mãe de todas as virtudes.” Desde então, esse pensamento de Cícero me acompanha. Na vida, é preciso aprender a ser grato! Ser grato é uma forma generosa de retribuir, com afeto, todo o esforço das pessoas que nos ajudam a ir além. Por isso, quero dedicar uma boa dose de energia a essa parte. Afinal, para chegar até aqui, contei com o auxílio de muitas mãos amigas!

Os primeiros a quem sempre devo agradecer são os meus pais, João e Christina. Agradeço não só pelo suporte e pelo apoio que recebi durante o mestrado, mas também pelo amor e pelo cuidado que me acompanham por toda a vida. Sou imensamente grato a eles por me ensinarem a enxergar o valor da educação. Se hoje sou quem sou, é porque me inspirei no exemplo deles. A eles, que ouviram os meus primeiros acordes e hoje me veem atuando profissionalmente como músico; que me ensinaram as minhas primeiras palavras e agora me veem conquistando um título de pós-graduação: todo o meu amor e minha eterna gratidão! (Amo vocês! <3)

Um outro agradecimento muitíssimo especial vai para o meu querido orientador, Flávio. Quando cogitei cursar uma pós-graduação, em meio a um universo de dúvidas, fui acolhido e incentivado por ele — antes mesmo de ele me conhecer! Conviver com o Flávio durante todo esse tempo foi uma verdadeira dádiva. As coisas que aprendi com ele ao longo dessa trajetória me ajudaram a me tornar não só um pesquisador melhor, mas também um ser humano melhor! O Flávio é o exemplo de educador que eu quero ser um dia: alguém que reconhece que, por trás de todo trabalho, há um ser humano trabalhando. Se não fosse por essas qualidades e pelo olhar acolhedor dele, eu provavelmente teria desistido no meio do caminho. Encerro esse ciclo com muito mais do que um orientador; ganhei um amigo com quem sei que posso contar para o que der e vier! Obrigado Flávio!

Mudar para São João del-Rei foi um desafio que eu nunca tinha encarado antes. Sair da casa dos meus pais e assumir novas responsabilidades foi algo totalmente novo para mim. Mas a verdade é que eu nunca passei por nada sozinho! Durante toda essa jornada, pude contar com a companhia de amigos valiosos que estiveram ao meu lado. Sou profundamente grato ao Vini, Carneiro, Isadora, Andressa, Matanza, Rodolfo, Emanuel, Emerson, Rebeca, Bradoki, Trotta, Ítalo, Karen, Yago, Carioca, Bianca, Jota, Lellis, Pedro, Naty, Sid, Evaldo, Paulo, Davi, Tony, Rômulo, Samuel, Vitin, Felipe, Frank, Alberto, Monica, Montse, Max, Maria, Uchoa, Andreza, Alice, Sofia, Ana, Amanda, Keth, João Lara, Bolin, Cleisson, Sick, Aretha, Luan, Matheus, Josi, Leo, LB, Polyana, Lucas, Hugo, Gustavo, André, Rato (*in memoriam*), Bebel, Pri, Christian, Giovane, Caroline e todos os outros que, porventura, eu tenha esquecido de mencionar! Essa jornada não teria sido tão incrível sem vocês!

Agradeço também aos companheiros de trabalho do laboratório ALICE. Conviver no ALICE expandiu meus horizontes e me fez enxergar a magia por trás da interdisciplinaridade.

Lá, tive a oportunidade de trabalhar ao lado de pessoas com as mais variadas formações — computação, música, moda, dança, teatro, cerâmica e engenharia — e aprender um pouco de cada coisa com cada um. Tenho certeza de que essa experiência expandiu meu olhar e me fez enxergar o mundo sob uma nova perspectiva. Devo isso a vocês!

Não posso deixar de fazer um agradecimento muito especial aos queridos mestres com os quais tive a oportunidade de conviver em sala de aula. As disciplinas que cursei e os estágios de docência que realizei contribuíram significativamente para a minha formação profissional. Nesse sentido, agradeço ao Flávio, Carol, Edimilson, Diego, Milene e Leo que, por meio de seu trabalho, contribuíram tanto para a minha trajetória.

Para finalizar esta sessão de agradecimentos, gostaria de registrar minha gratidão a algumas pessoas que não se encaixam nas categorias anteriores, mas que foram extremamente importantes durante este período. Agradeço ao Ari e à Leidiane, que gentilmente me alugaram o imóvel que chamei de lar por muitos meses, e à Jaquiele, secretária do programa, que me auxiliou inúmeras vezes em diferentes situações, sempre com muita prestatividade e boa vontade. Por fim, mas não menos importante, agradeço aos professores Missionário José, Henrique Vaz e João Sansão, que aceitaram compor a banca de defesa da dissertação, mesmo em circunstâncias tão adversas. As contribuições de vocês foram fundamentais para a construção deste trabalho.

"Se o problema tem solução, não esquente a cabeça, porque tem solução.
Se o problema não tem solução, não esquente a cabeça, porque não tem solução."
(Provérbio tibetano)

RESUMO

A produção musical é uma área intrinsecamente ligada à tecnologia. As diferentes formas de registrar performances musicais ao longo dos anos influenciaram diretamente a maneira como as músicas são criadas. Com o avanço tecnológico das últimas décadas, o processo de produção musical passou por adaptações contínuas até atingir o formato atual, caracterizado por etapas que abrangem uma série de tarefas técnicas e estéticas. O advento das tecnologias digitais revolucionou a indústria musical, trazendo facilidades e novas possibilidades. Com o computador como centro da produção musical, tornou-se possível o desenvolvimento de diversas técnicas e ferramentas inovadoras para o processamento de áudio. Muitas dessas técnicas têm base em métodos computacionais tradicionais, abrangendo tanto o processamento de sinais quanto o uso da inteligência artificial. Isso permitiu que questões antes impossíveis de serem abordadas no ambiente analógico agora pudessem ser resolvidas no ambiente digital. Uma dessas questões diz respeito aos vazamentos sonoros em gravações de bateria. Embora esses vazamentos não representem necessariamente um problema na produção musical, a separação dessas interferências pode facilitar o processamento sonoro, trazendo mais dinamismo e eficiência para profissionais da área de áudio em contextos de produção contemporânea. Nesse sentido, este trabalho conecta os campos da música, do áudio e da computação ao investigar a viabilidade do uso de ferramentas de separação de fontes de áudio em situações reais de gravação de bateria. Dentro desse campo, uma subcategoria recente, conhecida como separação de fontes de bateria, tem se destacado na literatura, proporcionando avanços significativos para a solução dessas questões. Neste estudo, uma das principais ferramentas de separação de peças de bateria, o LarsNet, é aplicada a situações práticas e reais de gravação de bateria, permitindo avaliar o desempenho do modelo e sua aplicabilidade no cotidiano de profissionais da área de áudio. A ferramenta é testada em duas gravações de baterias executadas e captadas em contextos distintos, oferecendo uma análise comparativa de suas capacidades. Ao final do trabalho, os resultados são apresentados e discutidos, fornecendo uma visão detalhada da capacidade da ferramenta em lidar com os desafios apresentados e sua viabilidade no uso cotidiano por produtores musicais e engenheiros de áudio.

Palavras-chaves: engenharia de áudio, separação de fontes de áudio, separação de fontes de bateria, recuperação de informação musical.

ABSTRACT

Title: The use of computational tools in the removal of sound spill in drum recordings.

Music production is an area inherently connected to technology. The various methods of recording musical performances over the years have directly influenced the way music is created. With the technological advancements of recent decades, the music production process has undergone continuous adaptations until reaching its current form, characterized by stages that encompass a series of technical and aesthetic tasks. The advent of digital technologies has revolutionized the music industry, bringing conveniences and new possibilities. With the computer as the center of music production, it has become possible to develop a range of innovative techniques and tools for audio processing. Many of these techniques are based on traditional computational methods, encompassing both signal processing and the use of artificial intelligence. This has made it possible to address issues that were previously unfeasible in the analog environment but are now resolvable in the digital domain. One such issue concerns sound spill in drum recordings. Although these leaks do not necessarily pose a problem in music production, separating these interferences can facilitate sound processing, bringing more dynamism and efficiency to audio professionals in contemporary production contexts. In this regard, this work bridges the fields of music, audio, and computing by investigating the feasibility of using audio source separation tools in real-world drum recording situations. Within this field, a recent subcategory known as drum source separation has gained prominence in the literature, providing significant advances toward solving these issues. In this study, one of the leading tools for drum piece separation, LarsNet, is applied to practical and real-world drum recording scenarios, enabling the evaluation of its performance and applicability in the daily routines of audio professionals. The tool is tested on two drum recordings performed and captured in distinct contexts, offering a comparative analysis of its capabilities. At the conclusion of the study, the results are presented and discussed, providing a detailed view of the tool's ability to address the presented challenges and its feasibility for everyday use by music producers and audio engineers.

Keywords: audio engineering, audio source separation, drum source separation, music information retrieval.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

A/D	Analógico para digital
ASS	Audio Source Separation
CNN	*Convolutional neural network*
D/A	Digital para analógico
DAW	*Digital audio workstation*
DNN	*Deep neural network*
DSS	Drum Source Separation
FASST	Flexible Audio Source Separation Toolbox
IHC	Interface Humano-Computador
IMP	Intelligent Music Production
ISMIR	International Society for Music Information Retrieval
ISTFS	Inverse Short-Time Fourier Transform
MIDI	Musical Instrument Digital Interface
MIR	Music Information Retrieval
MSS	Music Source Separation
NIME	New Interfaces for Musical Expression
NMF	Non-Negative Matrix Factorization
SiSEC	Signal Separation Evaluation Campaign
STFT	Short-Time Fourier Transform
TCC	Trabalho de Conclusão de Curso
UFSJ	Universidade Federal de São João del-Rei

SUMMARY

1 INTRODUÇÃO

A interdisciplinaridade é uma prática que contribui significativamente para o desenvolvimento científico. A união de esforços entre diferentes áreas do conhecimento permite que uma disciplina preencha lacunas que outras não conseguem suprir. Dessa forma, trabalhos com abordagens interdisciplinares promovem encontros e discussões que integram pesquisadores com habilidades complementares, convergindo para objetivos comuns.

Na ciência da computação, a computação aplicada é um terreno fértil para o desenvolvimento de trabalhos interdisciplinares. Ao aplicar métodos e ferramentas computacionais em diferentes contextos, os desenvolvedores precisam compreender os desafios específicos de cada área. Essa compreensão é essencial para garantir que as ferramentas criadas sejam adequadas e eficazes na solução dos problemas propostos.

Já no campo da música, a computação desempenha um papel fundamental. Tecnologias de gravação e reprodução musical moldam a forma como a música é criada, produzida e consumida. Na era digital, essas duas áreas estão cada vez mais próximas, impulsionando inovações e trazendo contribuições mútuas. O tema deste trabalho insere-se nesse contexto, buscando aproximar alguns desafios da produção musical de técnicas e métodos computacionais em desenvolvimento.

Este estudo aborda a utilização de algoritmos de separação de fontes de áudio como uma possibilidade de remoção de vazamentos sonoros em gravações de bateria. A escolha desse tema está diretamente relacionada à prática profissional do autor, que encontrou na computação uma oportunidade de adquirir novos conhecimentos aplicáveis ao seu dia a dia. A seção a seguir descreve as motivações que impulsionaram a escrita e o desenvolvimento do tema.

1.1 Motivação pessoal e profissional

Nesta seção, motivado pelo meu orientador, permito-me abrir mão da formalidade da narrativa em terceira pessoa para compartilhar um pouco da minha história e da motivação pessoal por trás da escrita deste trabalho. A paixão pela música é algo que carrego comigo desde a infância. Em 2003, com apenas sete anos de idade, fui introduzido a esse universo encantador por meio da minha primeira aula de teclado. Desde então, essa paixão tornou-se minha companheira ao longo da minha trajetória de vida.

Durante esses 21 anos dedicados ao estudo da música, meu envolvimento com essa forma de arte tornou-se tão profundo que passou a ser impossível separá-la da minha vida profissional e acadêmica. No entanto, em 2013, ao escolher o curso superior que seguiria, optei pela Engenharia. Essa decisão foi extremamente difícil, pois envolveu uma série de fatores pessoais e familiares. Em 2015, fui aprovado e comecei o curso de Bacharelado em Engenharia Mecatrônica no Instituto Federal do Sudeste de Minas Gerais (IF Sudeste MG).

Em 2018, já no meu quarto ano do curso de Engenharia, percebi que não tinha interesse em seguir a carreira tradicional de engenheiro. No entanto, o conhecimento adquirido ao longo da graduação, especialmente nas áreas de eletrônica e processamento de sinais, era valioso demais para ser deixado de lado. Motivado por isso, decidi buscar na interdisciplinaridade entre música e tecnologia uma nova razão para concluir o curso. Nesse período, tive a oportunidade de colaborar com meu orientador de graduação, Prof. Dr. Rodrigo Arruda, desenvolvendo projetos e pesquisas que integravam esses dois campos. Essa parceria culminou no desenvolvimento do meu Trabalho de Conclusão de Curso (TCC), intitulado "Análise comparativa entre um compressor de áudio analógico e seus simuladores digitais" (OLIVEIRA, 2021). Nesse trabalho, utilizei técnicas de processamento de sinais digitais para estudar compressores de áudio, ferramentas essenciais no contexto da produção musical e da engenharia de áudio.

Em 2019, inspirado pela gravação de um projeto musical da banda que eu participava na época, decidi me aprofundar nos estudos para atuar profissionalmente como produtor musical e engenheiro de áudio. Percebi que a combinação do conhecimento técnico adquirido na faculdade de Engenharia aliada à bagagem musical que já possuía me proporcionava uma base sólida para me tornar um profissional capacitado nessa área. A partir de 2020, iniciei minha trajetória profissional nesse segmento, contribuindo para o lançamento de projetos musicais de diversos artistas da minha cidade natal, Juiz de Fora, e fortalecendo minha atuação no mercado musical local.

Em 2021, concluí minha graduação em Engenharia Mecatrônica, em meio ao contexto desafiador do auge da pandemia de COVID-19. Aproveitando esse momento de transição, decidi buscar uma pós-graduação na área de Engenharia de Áudio, com o intuito de aprofundar meus conhecimentos e unir minhas paixões por tecnologia e música. No entanto, durante minha pesquisa, constatei que esse tema específico ainda não é oferecido no Brasil como curso de graduação ou pós-graduação. Apesar de existirem opções como Engenharia Acústica e Produção Fonográfica, ambas se mostraram distantes dos objetivos que eu desejava alcançar para minha formação acadêmica e desenvolvimento profissional.

Buscando uma forma de dar continuidade aos meus estudos, entrei em contato com o Prof. Dr. Marcelo Wanderley, renomado pesquisador na área de Tecnologia Musical da Universidade McGill, em Montreal, Canadá. Através desse contato, fui apresentado a diversos pesquisadores cujas linhas de pesquisa dialogavam diretamente com os temas que eu desejava explorar. Esses profissionais atuavam em programas de pós-graduação variados, abrangendo desde faculdades de música até áreas como engenharia e ciência da computação. Após muitas trocas de *e-mail* e reflexões, tomei a decisão de prosseguir meus estudos sob a orientação do Prof. Dr. Flávio Schiavoni na Universidade Federal de São João del-Rei (UFSJ).

Ingressar em uma pós-graduação em Ciência da Computação representou uma oportunidade única para explorar novos horizontes do conhecimento. Durante o curso, tive acesso a disciplinas que abrangiam uma ampla gama de áreas, como Computação Musical, Aprendizado

de Máquina, Ciência de Dados, Realidade Virtual, Teoria de Linguagens, Redes de Computadores e Bancos de Dados. Contudo, o caminho até encontrar uma linha de pesquisa que realmente alinhasse meus objetivos acadêmicos e profissionais exigiu tempo e reflexão.

Durante o período regulamentar do programa, explorei dois temas distintos para minha dissertação. Apesar de ter realizado publicações relacionadas a ambos, ainda sentia a necessidade de encontrar um tema que se conectasse de forma mais direta à minha prática profissional e às minhas aspirações acadêmicas. Próximo da conclusão do curso, decidi mudar o foco pela segunda vez, concentrando meus esforços em um terceiro tema, que é o abordado nesse trabalho. Embora o tempo disponível para aprofundar os resultados tenha sido limitado, acredito que a temática abordada possui uma relevância significativa para os campos da produção musical e da engenharia de áudio contemporâneas. Essa pesquisa marca o início de um estudo que pretendo expandir e aprofundar durante o doutorado.

Refletindo sobre minha motivação profissional, noto que a tecnologia tem conquistado um papel cada vez mais significativo nos processos de produção musical. Como discutido no Capítulo 2, os processos modernos de produção musical englobam diversas etapas e procedimentos que podem ser amplamente beneficiados pelo desenvolvimento tecnológico, impulsionado pelos avanços na computação. No entanto, ainda percebo lacunas que permanecem sem solução. Um exemplo notável é a aplicação de ferramentas tecnológicas voltadas para a separação de vazamentos em gravações, uma questão recorrente e desafiadora nesse contexto.

Embora o vazamento sonoro em gravações não seja necessariamente considerado um problema, ele pode limitar algumas possibilidades de edição e correção do material gravado. Isso ocorre porque o vazamento interfere diretamente na separação precisa dos elementos sonoros. Em um cenário influenciado pela dinâmica da indústria cultural, os processos de criação musical estão se tornando cada vez mais rápidos. Essa aceleração busca atender à crescente demanda por uma produção constante de conteúdo. Nesse contexto, o desenvolvimento de ferramentas que simplifiquem e otimizem essas etapas, proporcionando edições e manipulações de áudio mais rápidas, pode se tornar um recurso extremamente valioso.

Ao explorar possibilidades de implementações tecnológicas computacionais aplicadas à produção musical e à engenharia de áudio, identifiquei uma lacuna significativa na área de soluções para a separação de vazamentos em gravações de bateria. Notei que este tema é pouco explorado na literatura, com uma quantidade limitada de estudos publicados. Além disso, a separação sonora entre peças de bateria também apresenta uma carência de contribuições significativas. Por isso, um dos principais objetivos deste trabalho é integrar essas duas áreas, fomentando uma discussão sobre como elas podem se complementar e contribuir mutuamente para o avanço uma da outra.

1.2 Objetivos

De acordo com esse contexto, os objetivos deste trabalho são definidos da seguinte forma:

- Objetivo geral:
  - Utilizar a ferramenta de separação de fontes sonoras de peças de bateria, LarsNet, em duas situações reais e distintas de gravação de bateria e avaliar os resultados obtidos sob a perspectiva da produção musical.

- Objetivos específicos:
  - Traçar um histórico do desenvolvimento das tecnologias de gravação, destacando como as inovações tecnológicas influenciaram os processos de produção musical.
  - Descrever a importância do desenvolvimento das tecnologias de áudio digital para a produção musical moderna.
  - Ilustrar as principais etapas de um processo de produção musical e os desafios envolvidos em cada uma.
  - Caracterizar a bateria enquanto instrumento, detalhando suas particularidades.
  - Apresentar um breve histórico sobre o desenvolvimento da área de separação de fontes de áudio e sua convergência com esforços voltados para a separação de peças de bateria.
  - Introduzir ao leitor as duas gravações utilizadas como casos de estudo neste trabalho.
  - Discutir os resultados da aplicação do LarsNet nos sinais selecionados.
  - Sugerir, com base nos resultados dos experimentos, possíveis melhorias para a aplicação dos métodos avaliados em tarefas cotidianas da produção musical.

1.3 Organização do trabalho e recomendações de leitura

O trabalho é organizado em 6 capítulos, incluindo esta introdução. A estrutura foi elaborada para apresentar ao leitor os conceitos fundamentais necessários para as discussões desenvolvidas ao longo do texto. Por se tratar de um estudo interdisciplinar, abrangendo temas das áreas de música e computação, a organização busca fornecer as informações essenciais para a compreensão do conteúdo, independentemente da área de atuação do leitor.

No Capítulo 2, o leitor é introduzido ao universo da música, da produção musical e da engenharia de áudio. Esse capítulo apresenta, de forma breve, o desenvolvimento histórico das tecnologias de registro de áudio e suas contribuições para os processos modernos de

produção. Além disso, descreve as diversas etapas que compõem esse processo e suas respectivas atribuições.

O Capítulo 3 foca no principal objeto de estudo deste trabalho: a bateria. Ele detalha a composição desse instrumento e suas particularidades, abordando como ela é tratada dentro da produção musical, especialmente em relação ao vazamento sonoro entre suas peças durante as gravações. Também são apresentados métodos e tecnologias tradicionais utilizados nesse contexto.

Após a ambientação musical, o Capítulo 4 introduz os aspectos computacionais na discussão. Uma breve contextualização é feita sobre como a computação se relaciona com pesquisas em tecnologia musical, culminando nos algoritmos de separação de fontes sonoras. Os principais algoritmos são apresentados e contextualizados, facilitando o entendimento de seu uso em cenários musicais.

O Capítulo 5 apresenta a metodologia dos experimentos realizados neste trabalho. São detalhados os dois objetos de prova utilizados para avaliar os desempenhos dos métodos computacionais selecionados, além das metodologias de avaliação adotadas no experimento. Após isso, os resultados dos experimentos são apresentados e discutidos. Essas discussões permitem avaliar as limitações e os acertos da ferramenta empregada. Para auxiliar nessa análise, o capítulo conta com uma série de recursos multimídia disponíveis.

Por fim, o Capítulo 6 conclui o trabalho. São sugeridos pontos de melhoria com base nos resultados obtidos e propostas de trabalhos futuros, além de novas linhas de pesquisa que exploram a interseção entre computação e produção musical.

Por se tratar de um trabalho que envolve questões relacionadas ao áudio, foram utilizados recursos multimídia no formato PDF para ilustrar exemplos e resultados. Dessa forma, toda vez que o leitor encontrar uma caixa de texto verde, esta estará acompanhada de um arquivo de áudio[1]. Recomenda-se que a leitura seja realizada com um bom sistema de monitoramento de áudio ou com fones de ouvido para uma melhor experiência.

[1] Caso o recurso não funcione, os áudios também estarão disponibilizados no seguinte endereço: <https://alice.ufsj.edu.br/audios_oliveira2024/>

2 MÚSICA, PRODUÇÃO MUSICAL E ENGENHARIA DE ÁUDIO

A música é uma forma de manifestação artística (CANUDO, 1911) e, por isso, é frequentemente associada ao campo das ciências humanas. No entanto, sua aplicabilidade vai além dessa esfera, permitindo uma ampla gama de trabalhos interdisciplinares. Por exemplo, a música pode ser utilizada em pesquisas na área da saúde, consolidando-se também como um objeto de estudo valioso para as ciências biológicas (BATALHA et al., 2022).

Diferente de outras formas de artes, como as artes plásticas, a música é uma manifestação artística de natureza impermanente, o que significa que a prática musical, por si só, não é suficiente para garantir seu registro. Quando não gravada, tende a se perder com o passar do tempo. Embora tradições orais e sistemas de notação auxiliem em sua preservação, eles apenas fornecem pistas sobre como uma performance era realizada. Historicamente, a cultura musical de muitos povos se perdeu devido à ausência de registros adequados, uma consequência direta do caráter efêmero da música (TARUSKIN, 2010).

Diante desse desafio, a humanidade buscou maneiras de registrar performances musicais à medida que os avanços tecnológicos tornaram possíveis gravações sonoras. Desde meados do século XIX, já existiam indícios de produção tecnológica voltadas para essa finalidade (MILEHAM, 2009). Esse processo evoluiu através de diferentes etapas, passando por dispositivos mecânicos e eletromagnéticos até chegar à gravação digital, que é a forma mais utilizada atualmente (MACEDO, 2007).

O uso de tecnologias mecânicas, eletromagnéticas e digitais no campo musical estabelece uma relação interdisciplinar entre a música, enquanto forma de arte, e áreas das ciências exatas, como a ciência da computação e as engenharias (BROWN, 1995). Essa integração permite o desenvolvimento de ferramentas e técnicas para o registro, processamento e reprodução de som, ampliando as possibilidades de criação (WILMERING et al., 2020) e preservação musical. As interações entre esses campos têm sido fundamentais para a transformação das práticas musicais e a ampliação do acesso à música em formatos cada vez mais diversificados.

O desenvolvimento tecnológico impulsionou o surgimento de um mercado específico para a música, conhecido como indústria fonográfica. A mercantilização da música e os diferentes formatos de reprodução sonora disponíveis contribuíram para a padronização das formas de realizar e registrar performances musicais. Esse processo de criação e registro de uma obra musical é denominado produção musical (COSTA; CATALAN, 2019).

Apesar das padronizações trazidas pelo mercado fonográfico, a produção musical permanece um processo bastante livre no campo artístico, permitindo que diferentes artistas concretizem suas obras por diversos caminhos. Assim, não há regras rígidas para a produção musical. No entanto, algumas etapas são comuns a diferentes processos, seja por exigências

técnicas relacionadas às tecnologias disponíveis ou por questões mercadológicas (PAIXÃO, 2013).

Antes de aprofundar nas etapas comuns aos processos de produção musical, é interessante diferenciar dois conceitos: produção musical e engenharia de áudio, ou engenharia de som. Conforme discutido anteriormente, há uma relação estreita entre a criação musical contemporânea e as tecnologias que possibilitam o registro dessas criações. Por esse motivo, é comum que os conceitos de produção musical e engenharia de áudio sejam confundidos.

A diferenciação entre produção musical e engenharia de áudio se dá principalmente pelas suas áreas de enfoque e aplicação prática. A engenharia de áudio concentra-se nos aspectos técnicos e científicos do som, englobando desde a captura e criação até a manipulação e reprodução sonora. Esse campo inclui conhecimentos em acústica, eletrônica, física do som e computação aplicada ao áudio. Por isso, o engenheiro de áudio é o responsável por ajustar e implementar as configurações técnicas que garantem a qualidade sonora do produto.

A produção musical, por outro lado, envolve um papel mais artístico e gerencial. O produtor musical trabalha diretamente com o artista, orientando o processo criativo, avaliando os resultados e tomando decisões importantes para garantir que a obra atinja seus objetivos estéticos e mercadológicos. Além de conhecimentos artísticos, é essencial que o produtor possua habilidades interpessoais, já que ele coordena a equipe envolvida em todo o processo de produção musical (MACEDO, 2007).

Assim, produção musical e engenharia de áudio caminham juntas, complementando-se. Enquanto o produtor define a direção criativa e estratégica do projeto, o engenheiro de áudio executa as soluções técnicas necessárias para materializar essa visão, utilizando seu domínio sobre as tecnologias e equipamentos de áudio. Essa parceria permite que as criações musicais ganhem forma, unindo a visão artística com a excelência técnica (MACEDO, 2007).

A seguir, serão discutidas algumas etapas que são comuns à maioria dos processos de produção musical. Considera-se que esses processos têm como objetivo final a criação de um fonograma adequado para o mercado atual e para os formatos de reprodução disponíveis. Essas etapas abrangem tanto aspectos artísticos quanto técnicos, assegurando que a obra musical seja criada, refinada e finalizada de acordo com as exigências de qualidade sonora e com as tendências de consumo, visando alcançar um público mais amplo nos diversos meios de distribuição e plataformas digitais.

2.1 Composição

A criação musical frequentemente nasce de um conceito ou de um conjunto de inspirações que refletem a individualidade e a visão artística de seu criador. O primeiro passo nesse processo é estruturar e desenvolver esses elementos. Esse estágio, conhecido como composição,

é o ponto de partida para transformar abstrações em uma ideia musical mais concreta.

Nesse momento, o compositor traduz suas concepções iniciais em componentes musicais tangíveis, como letras, melodias, harmonias e narrativas. Esse processo exige não apenas criatividade, mas também habilidades técnicas para transformar emoções e intenções em uma linguagem sonora articulada. O produtor musical, nesse contexto, desempenha um papel essencial. Ele atua como um colaborador estratégico, auxiliando na organização e no refinamento das ideias do compositor, garantindo que o projeto alcance sua máxima expressão artística e técnica.

A composição, por sua natureza criativa, é profundamente influenciada pelas vivências e referências únicas de cada indivíduo. Essa característica confere à prática uma flexibilidade singular, permitindo que cada artista desenvolva seu próprio método de trabalho, livre de regras ou padrões preestabelecidos. Essa liberdade resulta em obras diversificadas, que refletem de maneira autêntica as experiências e intenções do compositor.

Embora seja uma atividade essencialmente artística, a composição frequentemente incorpora tecnologias que potencializam o processo criativo. Muitos compositores recorrem a ferramentas de gravação de áudio para registrar, organizar e explorar suas ideias musicais. Essas tecnologias desempenham um papel crucial na produção contemporânea, ampliando as possibilidades. Além de registrar performances, elas permitem capturar, experimentar e armazenar conceitos em evolução, facilitando revisões e aprimoramentos ao longo do processo.

2.2 Arranjo

Outra etapa essencial no processo de produção musical é o arranjo, que sucede a composição. Diferentemente do processo inicial de criação, o arranjo é o momento em que as ideias concebidas começam a ganhar contornos mais definidos, aproximando-se do formato final da obra. Essa etapa, assim como a composição, é marcada por um alto grau de criatividade e expressão artística, o que contribui para a singularidade de cada produção musical.

Durante o arranjo, elementos fundamentais da música, como ritmo, harmonia, melodia e textura, são estruturados e refinados com precisão. Esse trabalho detalhado não apenas define a identidade sonora da canção, mas também a prepara para ser executada ou registrada em sua forma definitiva. É uma fase crucial, onde as escolhas feitas impactam diretamente a estética e a funcionalidade da obra.

O responsável por liderar essa etapa é o arranjador. Ele desempenha um papel essencial na definição de aspectos técnicos e criativos, como a seleção dos instrumentos que comporão o arranjo, a criação de melodias complementares, a implementação de possíveis reharmonizações e a organização de elementos rítmicos. Para realizar essas tarefas, o arranjador combina sensibilidade artística, profundo conhecimento técnico e habilidade em interpretar e concretizar a

visão do compositor.

A atuação do produtor musical também é indispensável nesse momento. Ele atua como um elo entre a visão artística do projeto e as limitações práticas envolvidas. Por exemplo, se o artista pretende incorporar uma orquestra ao arranjo, cabe ao produtor avaliar a viabilidade dessa proposta. Caso seja viável, o produtor organiza o orçamento, contrata os profissionais necessários e gerencia os recursos para garantir a realização dessa ideia. Além disso, o produtor colabora na definição de aspectos estéticos e garante que o arranjo esteja alinhado com os objetivos gerais do projeto.

A tecnologia desempenha um papel significativo no processo de arranjo. Além de viabilizar o registro e o armazenamento das ideias, ela permite a pré-visualização das propostas sonoras antes mesmo da gravação final. Essa funcionalidade oferece maior flexibilidade ao processo criativo, tornando-o mais dinâmico e eficiente. Esse avanço é possível graças a ferramentas conhecidas como instrumentos virtuais.

Os instrumentos virtuais são programas de computador que têm como objetivo simular o som de instrumentos reais. Essa simulação pode ser feita por meio da reprodução de amostras gravadas desses instrumentos, conhecidas como *samples*, ou através de técnicas de síntese sonora, como a síntese subtrativa, a síntese de distorção de fase e a síntese de modulação em frequência, também conhecida como síntese FM (MASSEY; NOYES; SHKLAIR, 1987).

O controle desses instrumentos virtuais pode ser realizado de maneira eficiente por meio de um protocolo conhecido como Musical Instrument Digital Interface (MIDI) (MAZZOLA et al., 2018). Esse protocolo foi desenvolvido para facilitar a comunicação entre instrumentos eletrônicos, computadores e outros dispositivos musicais. Em vez de transmitir áudio, o MIDI envia dados que descrevem ações musicais, como a execução de notas, controle de volume e alterações de timbre. Esses dados são transmitidos entre dispositivos, permitindo que, por exemplo, um sintetizador reproduza sons a partir das instruções recebidas (RUSS, 2012a).

O protocolo MIDI codifica as ações realizadas pelo músico, como pressionar teclas ou manipular controles, transformando-as em comandos digitais que podem ser reproduzidos. Além disso, permite o armazenamento dessas instruções em arquivos digitais padronizados, conhecidos como arquivos MIDI, para uso e edição futuros (MAZZOLA et al., 2018). Esses comandos registram nuances importantes da performance, como o momento em que as notas foram tocadas, sua duração e a intensidade aplicada em cada uma (GILLICK et al., 2019). Os dispositivos que utilizam esse protocolo para controlar uma performance são chamados de controladores MIDI (RUSS, 2012b).

Para ilustrar a importância dessas tecnologias no processo de produção musical, pode-se retomar o exemplo de um arranjo com orquestra. Gravar uma orquestra real pode ser um processo extremamente caro e complexo, devido ao número de músicos e profissionais envolvidos, além dos altos custos de produção (MÉNDEZ, 2022). Um instrumento virtual que

simula uma orquestra permite ao arranjador testar essa estética sonora antes de decidir pela gravação com uma orquestra real (RUSS, 2012c). Em muitos casos, os instrumentos virtuais são utilizados como uma alternativa viável para substituir elementos que sejam financeiramente ou logisticamente inviáveis em uma produção (MCGUIRE; KAPLAN; KAPLAN, 2005).

A criação e o uso de instrumentos virtuais exemplificam novamente a relação interdisciplinar entre a música e a ciência da computação. Por serem programas de computador, o desenvolvimento desses instrumentos envolve conceitos e áreas como Interface Humano-Computador (IHC) (HSU et al., 2013), Computação Musical (YUN; CHA, 2013) e Novas Interfaces para Expressão Musical, em inglês New Interfaces for Musical Expression (NIME) (ROCHA; TEIXEIRA; SCHIAVONI, 2019).

Embora composição e arranjo sejam diferenciados neste trabalho, alguns autores os integram em um processo único, conhecido como pré-produção (PAIXÃO, 2013). Independentemente da abordagem, o principal objetivo dessas etapas é assegurar que, ao final, o artista e os músicos estejam devidamente preparados para registrar suas ideias na fase de gravação.

2.3 Captação/gravação

Após a finalização das etapas iniciais do processo de criação musical, torna-se necessário dar vida às ideias desenvolvidas. Esse passo envolve a execução prática das composições. A partir dessa execução, ocorre o registro sonoro em formato de áudio, conhecido como gravação ou captação.

As gravações geralmente são realizadas em ambientes conhecidos como estúdios de gravação. Os estúdios com viés profissional são projetados com uma maior preocupação, tanto em termos de acústica quanto de arquitetura, para que as gravações ocorram com a melhor qualidade possível. Além disso, esses esforços também visam garantir conforto e bem-estar auditivo para a equipe envolvida (HUBER; RUNSTEIN, 2018j). Ele são equipados com a infraestrutura necessária para converter o som dos instrumentos em dados de áudio. Com o avanço da tecnologia de gravação digital, muitos artistas também passaram a realizar a etapa de gravação em ambientes domésticos, conhecidos como *home studios* (PRAS; GUASTAVINO; LAVOIE, 2013). A Figura 1 exemplifica o tipo de estrutura encontrada em um estúdio de gravação profissional, enquanto a Figura 2 mostra um exemplo de *home studio*.

É nessa etapa que a engenharia de áudio assume um papel central no processo de produção musical. As funções do produtor musical e do engenheiro de áudio se tornam claramente delimitadas, conforme discutido anteriormente. Além disso, essa fase marca o ponto em que muitos conceitos de arte e tecnologia se interligam, reforçando a interdisciplinaridade que é essencial para o desenvolvimento deste trabalho. Decisões técnicas têm implicações artísticas e, da mesma forma, decisões artísticas influenciam os aspectos técnicos. Trata-se de uma etapa em que arte e tecnologia caminham lado a lado.

Figura 1 – Estúdio de gravação profissional.

Fonte: adaptado de (TUCCONI, 2018) e (TEN, 2013).

Figura 2 – *Home studio.*

Fonte: adaptado de (SMITH, 2010) e (YAKARTEPE, 2013).

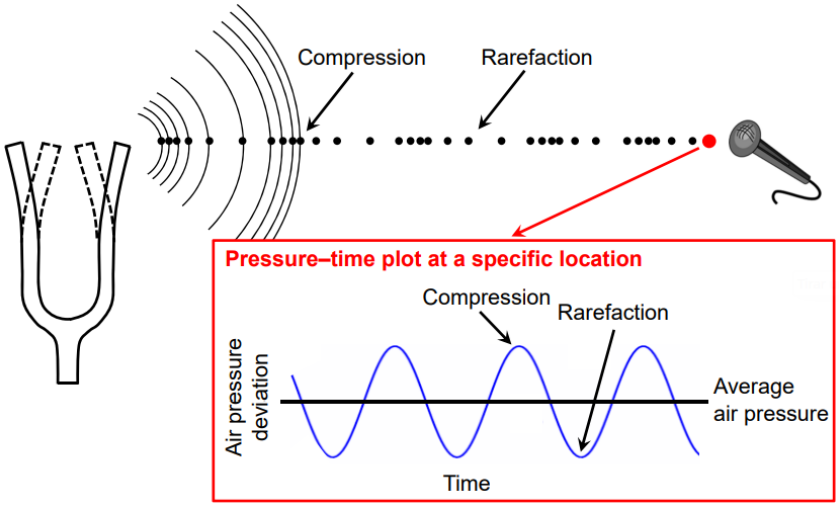
O processo de gravação pode ser realizado de diversas maneiras. No entanto, apesar dessa pluralidade, é possível dividi-lo em duas categorias principais: gravação em sistemas analógicos (HUBER; RUNSTEIN, 2018a) e gravação em sistemas digitais (HUBER; RUNSTEIN, 2018b). Essa classificação é baseada no tipo de tecnologia de registro utilizada para capturar as performances executadas. Para compreender com clareza as diferenças entre as duas formas de captação sonora, é fundamental entender alguns conceitos básicos.

Som e áudio

O primeiro conceito a ser apresentado é o de som. Ele pode ser definido como um fenômeno físico caracterizado por uma onda mecânica longitudinal que se propaga através do ar ou de outro fluido elástico. Essa onda é gerada por um objeto vibrante, que provoca

deslocamentos e oscilações nas moléculas do fluido (BRANDÃO, 2018). Essas perturbações criam regiões de compressão e rarefação em sua estrutura, que podem ser captadas por sistemas auditivos, como o ouvido humano, ou por dispositivos eletrônicos, como microfones (MÜLLER, 2021b). A Figura 3 ilustra este processo.

Figura 3 – Ilustração do fenômeno sonoro.



Fonte: (MÜLLER, 2021b).

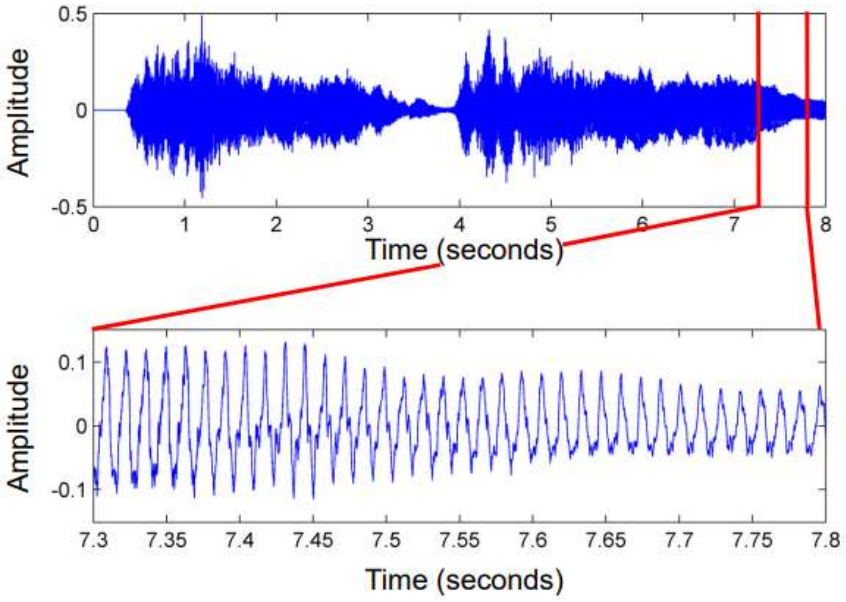
Quando se estuda os fenômenos sonoros, é importante destacar que a capacidade auditiva humana é limitada a um intervalo que varia, aproximadamente, entre 20 Hz e 20 kHz. Ondas com frequências inferiores a 20 Hz são chamadas de infrassom, enquanto aquelas com frequências superiores a 20 kHz são denominadas ultrassom. Embora essas ondas sejam fundamentais em outras aplicações, elas são irrelevantes para o estudo do som audível (BRANDÃO, 2018).

Quando um som, dentro do limiar da audição humana, é reproduzido, transmitido ou recebido na forma de sinal elétrico, ele passa a ser denominado áudio. Nesse contexto, o sinal de áudio pode ser definido como uma representação do som. Um sinal de áudio contém todas as informações necessárias para a reprodução sonora, incluindo aspectos temporais, dinâmicos e tonais (MÜLLER, 2021b).

Uma forma bastante comum de representar sinais de áudio é por meio de um modelo gráfico conhecido como forma de onda, ou *waveform* em inglês. Ele consiste em uma representação cartesiana que traduz a variação da pressão do ar de um som, relacionando-a à amplitude do sinal de áudio ao longo do tempo (MÜLLER, 2021b). A Figura 4 ilustra tal representação.

O sinal de áudio analógico é aquele que ocorre no mundo real (MÜLLER, 2021a). Ele pode ser encontrado na forma de sinais elétricos, como na saída de um microfone, ou na forma de ondas magnéticas, como no caso das informações registradas em uma fita de gravação de áudio (HUBER; RUNSTEIN, 2018a). Dessa forma, os sinais de áudio analógico podem ser definidos como representações em tempo contínuo (MÜLLER, 2021a), ou seja, funções cuja variável

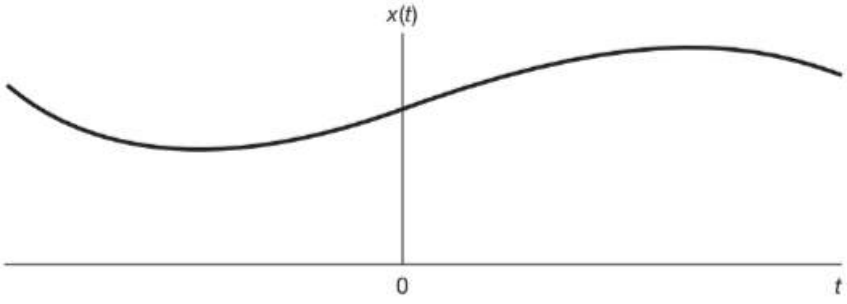
Figura 4 – Representação de um sinal de áudio em forma de onda.



Fonte: (MÜLLER, 2021b).

independente é contínua (OPPENHEIM; WILLSKY; NAWAB, 2010). A Figura 5 apresenta um exemplo da representação de um sinal contínuo.

Figura 5 – Representação de um sinal contínuo.



Fonte: (OPPENHEIM; WILLSKY; NAWAB, 2010).

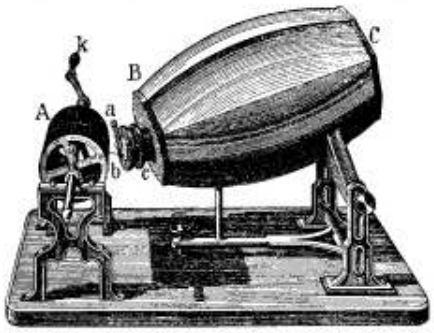
Gravação do som

Antes do advento do computador como ferramenta de gravação, todo o processo era realizado de forma totalmente analógica. Os primeiros sistemas de gravação eram bastante rudimentares e apresentavam várias limitações. Abaixo estão algumas das primeiras tentativas de registro sonoro:

- **Fonautógrafo**: desenvolvido por Édouard-Léon Scott de Martinville em meados da década de 1850, o fonautógrafo registrava visualmente as ondas sonoras em cilindros cobertos de fuligem. No entanto, essa tecnologia não permitia a reprodução do som

gravado, limitando-se apenas à representação visual das ondas sonoras (MILEHAM, 2009). A Figura 6 ilustra um fonautógrafo.

Figura 6 – Ilustração de um fonautógrafo.



Phonautograph.

BC, barrel with opening at *C*; *c*, brass tube with membrane and style at *b*, and movable piece *a*, by which the position of the nodal points can be regulated; *k*, handle to turn cylinder (*A*) covered with lampblacked paper.

Fonte: (CENTURY DICTIONARY, 1891).

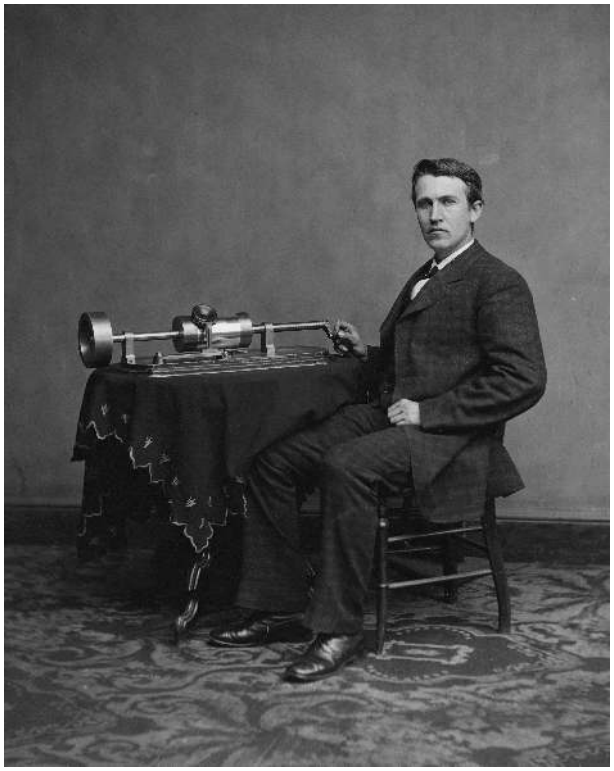
- **Fonógrafo**: introduzido por Thomas Edison em 1877, o fonógrafo teve sua criação inspirada no paleofone de Charles Cros (PRAS; GUASTAVINO; LAVOIE, 2013). Foi o primeiro dispositivo capaz de gravar e reproduzir som. Utilizando uma agulha para gravar padrões em folhas de estanho, o fonógrafo permitiu o registro e a posterior reprodução do som, representando um avanço significativo na tecnologia de gravação (MILEHAM, 2009). A Figura 7 mostra Edison ao lado de um fonógrafo.

- **Grafofone**: desenvolvido na década de 1880 por Chichester Bell e Charles Tainter, o grafofone trouxe melhorias na qualidade sonora ao usar cilindros de cera mais duráveis, substituindo as folhas de estanho do fonógrafo. Embora tenha melhorado a durabilidade e a clareza das gravações, ainda apresentava limitações em termos de capacidade e resistência (MILEHAM, 2009). A Figura 8 ilustra um grafofone.

- **Gramofone**: introduzido por Emile Berliner em 1887, o gramofone inovou ao substituir os cilindros por discos planos. Essa mudança facilitou a duplicação das gravações, tornando o processo mais eficiente e permitindo a distribuição comercial em larga escala. O gramofone revolucionou a indústria da música, tornando as gravações acessíveis a um público muito maior (MILEHAM, 2009). A Figura 9 apresenta Emile Berliner ao lado de um gramofone.

Apesar de todas essas tentativas, o processo de gravação de áudio em ambiente analógico, como o conhecemos hoje, só se consolidou com o advento dos magnetófonos. O primeiro dispositivo dessa natureza que se tem registro foi o telegrafone, desenvolvido por Valdemar Poulsen na década de 1930. Essa tecnologia permitia a gravação magnética de um sinal de áudio utilizando um fio metálico como meio de gravação. Futuramente, esse meio viria a ser

Figura 7 – Thomas Edison ao lado de um fonógrafo.

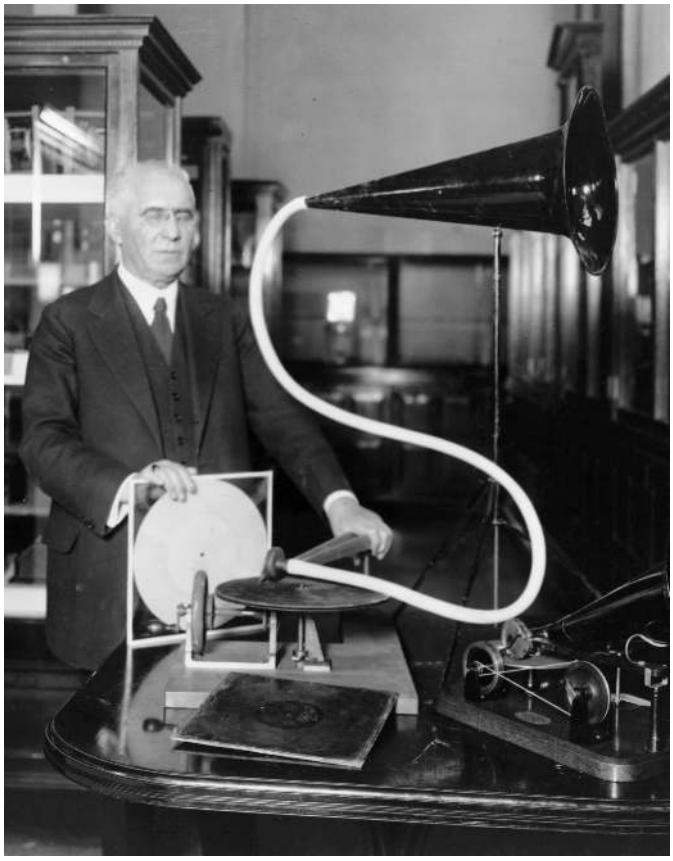


Fonte: (HANDY, 1878).

Figura 8 – Ilustração de um grafofone.

Fonte: (MAISON DE LA BONNE PRESSE, 1901).

Figura 9 – Emile Berliner ao lado de um gramofone.



Fonte: (AUTOR DESCONHECIDO, entre 1910 e 1929).

substituído por fitas magnéticas, proporcionando uma melhora na qualidade sonora registrada (MILEHAM, 2009). Embora apresentasse algumas desvantagens, como a fragilidade da fita e o tamanho volumoso dos equipamentos, esse tipo de sistema permaneceu como padrão nos estúdios de gravação analógicos e continua sendo utilizado até hoje em produções que optam por esse formato. A Figura 10 apresenta exemplares de gravadores de fita.

Figura 10 – Exemplos de gravadores de fita.

Fonte: adaptado de (WILMUT, 1878) e (WILMUT, 1969).

O surgimento da gravação em fita magnética trouxe uma série de avanços significativos no processo de gravação. Diferentemente dos formatos anteriores, esse novo método permitia a edição da performance executada. Com ele, era possível gravar uma mesma performance várias vezes e selecionar os melhores trechos de cada uma, criando uma faixa híbrida composta por diferentes execuções (MACEDO, 2007). Isso se tornava viável porque a fita magnética podia ser cortada e colada, permitindo ajustes e edições conforme as necessidades específicas do projeto (PRAS; GUASTAVINO; LAVOIE, 2013).

Uma segunda vantagem proporcionada pelo uso de fitas magnéticas no registro de áudio analógico foi o surgimento de uma prática conhecida como *overdubbing* (PRAS; GUASTAVINO; LAVOIE, 2013). Com o avanço das tecnologias de gravação, tornou-se possível adicionar uma nova faixa enquanto se ouve o conteúdo já gravado, sem sobrescrevê-lo (MACEDO, 2007). Essa inovação permitiu que partes mal executadas em uma performance pudessem ser substituídas por novas gravações, sem a necessidade de cortar a fita magnética.

Outra grande transformação no processo de gravação veio com o surgimento dos gravadores multipista, após a década de 1960. Essa tecnologia revolucionária permitiu que cada instrumento fosse registrado em uma faixa de áudio independente. Embora pareça uma mudança simples, essa inovação alterou profundamente o modo como as produções eram realizadas. Com os instrumentos gravados separadamente, as possibilidades de edição, *overdubbing* e manipulação de áudio aumentaram significativamente (PRAS; GUASTAVINO; LAVOIE, 2013). Além disso, a capacidade de registrar os instrumentos de forma isolada impactou diretamente as etapas após a gravação, como a mixagem e a masterização, proporcionando maior controle e precisão no tratamento do áudio (MACEDO, 2007). A Figura 11 apresenta um gravador multipista.

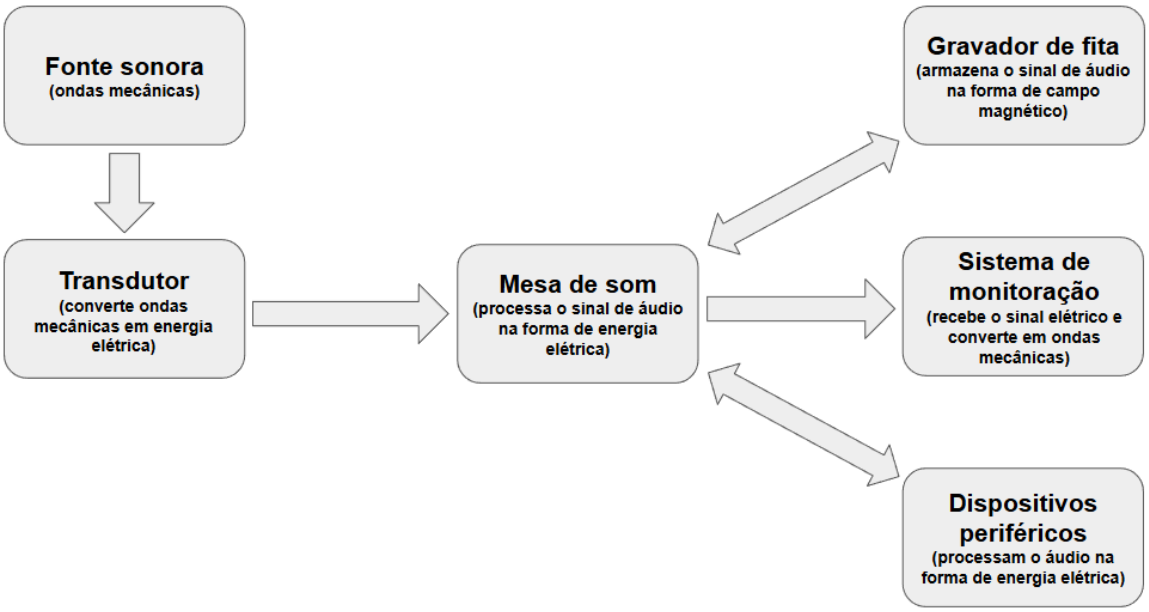
O fluxo de sinal básico em uma gravação em sistema analógico ocorre da seguinte maneira. Primeiro, o som emitido pela fonte sonora é convertido em sinais elétricos por um dispositivo transdutor, geralmente um microfone ou captador. Esses sinais são então enviados por cabos, do transdutor até um equipamento conhecido como mesa de som. A mesa de som é um aparelho de processamento de áudio que permite rotear o sinal para os demais equipamentos do estúdio. Nesse processo, a máquina de fita desempenha o papel de armazenar as performances capturadas em fitas magnéticas (HUBER; RUNSTEIN, 2018a). A Figura 12 representa um esquema de gravação em sistema analógico.

Embora seja uma forma de gravação mais antiga, a produção musical em gravadores de fita ainda é bastante utilizada atualmente. Grande parte da motivação para trabalhar com sistemas analógicos está relacionada a questões estéticas. O uso de dispositivos físicos com uma ampla variedade de circuitos eletrônicos faz com que a sonoridade obtida por esse processo seja única. Além da característica sonora específica de um gravador de rolo, muitos dos equipamentos utilizados nesses sistemas empregam válvulas eletrônicas em seus circuitos, o que adiciona uma distorção harmônica particular ao som.

Figura 11 – Exemplo do um gravador de fita multipista.

Fonte: (CALYSO, 2014).

Figura 12 – Exemplo do fluxo de sinal em uma gravação analógica.



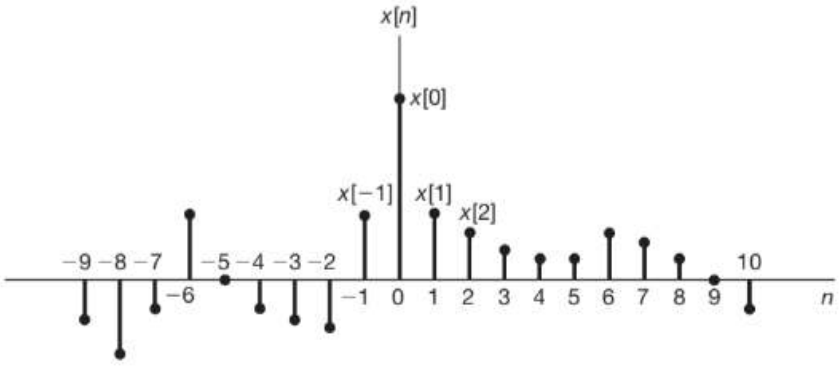
Fonte: acervo do autor.

Gravação Digital

O advento do computador como ferramenta de registro de áudio transformou profundamente o processo de produção musical (PRAS; GUASTAVINO; LAVOIE, 2013). Com sua popularização e o aumento do poder computacional, o computador tornou-se o centro da produção musical moderna (WILMERING et al., 2020). Isso se deve à sua capacidade de executar tarefas de forma eficiente, como gravação, edição, mixagem e masterização de áudio, funções que antes exigiam equipamentos físicos dedicados. Com essa tecnologia, processos que eram complexos e demorados, tornaram-se muito mais rápidos e simples, permitindo ajustes precisos com o uso de interfaces intuitivas.

Devido à sua arquitetura de funcionamento, os computadores não operam com sinais analógicos, mas sim com sinais digitais. Isso ocorre porque os computadores são limitados a registrar uma quantidade finita de valores (MÜLLER, 2021a). Dessa forma, os sinais digitais são representados como sinais de tempo discreto, nos quais a variável independente assume valores discretos (OPPENHEIM; WILLSKY; NAWAB, 2010). A Figura 13 apresenta uma representação desse tipo de sinal.

Figura 13 – Representação de um sinal discreto.



Fonte: (OPPENHEIM; WILLSKY; NAWAB, 2010).

Sinais de áudio analógico podem ser convertidos em sinais de áudio digital por meio de uma prática denominada digitalização. Esse processo transforma sinais de tempo contínuo em sinais de tempo discreto. A digitalização de sinais de áudio é dividida em dois sub-processos: amostragem e quantização (MÜLLER, 2021a).

O processo de amostragem tem a função de converter os valores contínuos de um sinal em valores discretos. Uma maneira comum de realizar essa tarefa é por meio de um método conhecido como amostragem equidistante (MÜLLER, 2021a), que pode ser representado por

$$x(n) := f(nT) \tag{2.1}$$

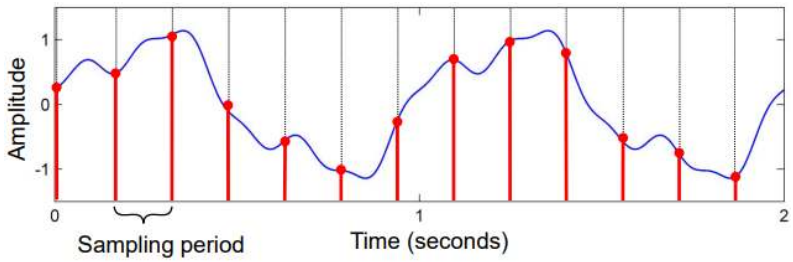
onde $x(n)$ é a amostra do sinal analógico f em um instante de tempo $t = nT$, e T é o período de amostragem. O inverso de T é denominado taxa ou frequência de amostragem. Dessa forma,

a frequência de amostragem é definida por

$$F_s := \frac{1}{T} \tag{2.2}$$

onde F_s representa a frequência de amostragem, geralmente expressa em Hz (MÜLLER, 2021a). A Figura 14 ilustra o processo de amostragem de um sinal analógico utilizando 13 amostras e uma frequência de amostragem de 6,5 Hz.

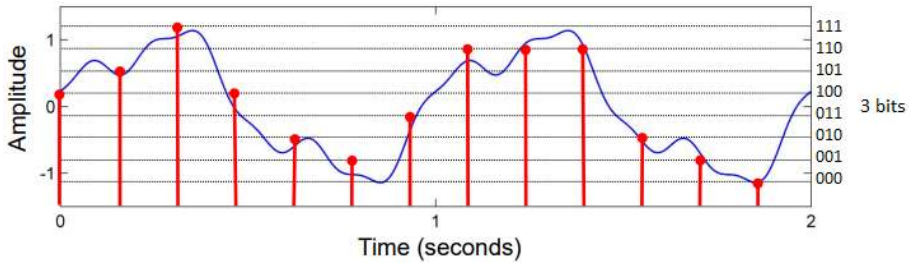
Figura 14 – Representação do processo de amostragem de um sinal analógico utilizando 13 amostras e frequência de amostragem igual a 6,5 Hz.



Fonte: (MÜLLER, 2021a).

O processo de quantização tem como função representar os valores contínuos de amplitude de um sinal de áudio analógico em valores discretos. Como mencionado anteriormente, os computadores são limitados quanto à quantidade de valores que podem registrar. Por esse motivo, a quantização é fundamental para permitir o armazenamento e a manipulação de sinais de áudio no domínio digital. No caso de um sinal de áudio elétrico, esse processo consiste em registrar o nível de tensão do sinal em *bits* (MÜLLER, 2021a). O valor registrado em *bits* é conhecido como profundidade de *bits*, ou resolução do sinal de áudio. Quanto maior a profundidade de *bits*, maior será a resolução e a qualidade do sinal de áudio digitalizado (HUBER; RUNSTEIN, 2018b). A Figura 15 ilustra o processo de quantização do sinal amostrado na Figura 14.

Figura 15 – Representação do processo de quantização de um sinal analógico utilizando uma resolução de 3 *bits*.



Fonte: adaptado de (MÜLLER, 2021a).

O processo de conversão de um sinal de áudio analógico em digital é comumente referido como conversão analógico-digital (A/D). De maneira semelhante, o processo inverso, que converte um sinal digital em analógico, é denominado conversão digital-analógico (D/A).

Ambos os processos são de extrema importância no campo da engenharia de áudio, pois permitem a transição entre os domínios analógico e digital.

A manipulação de áudio em sistemas digitais é realizada em uma interface conhecida como estação de trabalho de áudio digital, ou *digital audio workstation* (DAW), em inglês. Essas ferramentas são essenciais para as atividades de produção musical e engenharia de som realizadas diretamente no computador, um processo comumente denominado *in the box*. Os *softwares* DAW permitem a gravação, edição e processamento de arquivos de áudio, sendo amplamente utilizados na indústria. Existem diversas empresas que produzem DAW no mercado, cada uma voltada para diferentes perfis de usuários e necessidades. Apesar dessa variedade, a maioria das DAW compartilha duas interfaces gráficas principais: a janela de edição e a janela de mixagem (HUBER; RUNSTEIN, 2018c)

A janela de edição é a interface que exibe ao usuário a visualização gráfica das formas de onda dos sinais de áudio gravados. Essa representação gráfica permite visualizar a amplitude do som ao longo do tempo, facilitando a compreensão e manipulação dos dados de áudio. Devido a essa funcionalidade, a janela de edição é amplamente utilizada para realizar operações como cópias, recortes e colagens no sinal de áudio, justificando seu nome (HUBER; RUNSTEIN, 2018c). A Figura 16 ilustra as janelas de edição de diversas DAWs.

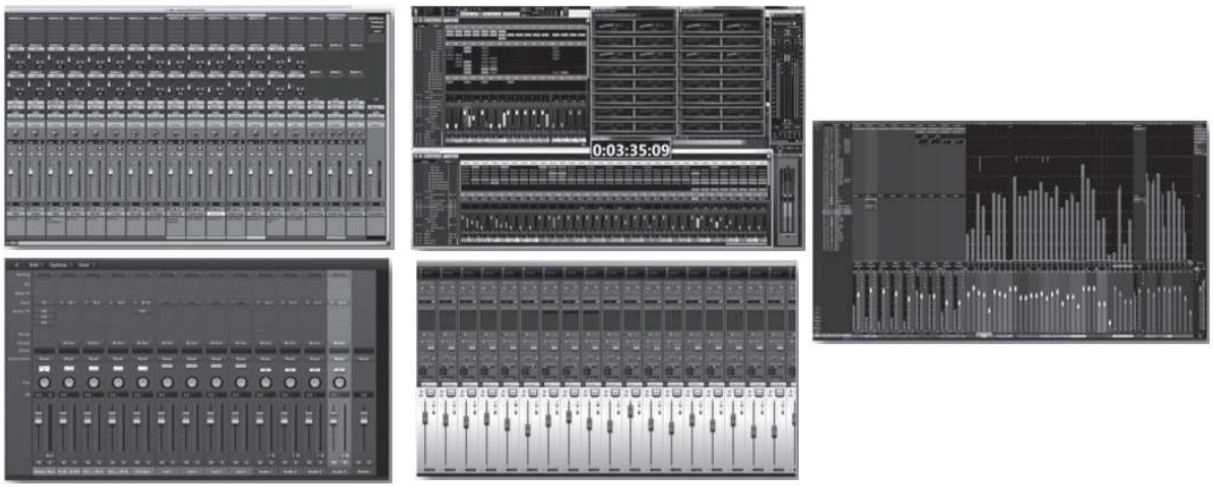
Figura 16 – Janela de edição das DAW Pro Tools (superior esquerda), Cubase (superior direita), Logic (inferior esquerda), Reaper (inferior direita) e Studio One (direita central).

Fonte: adaptado de (HUBER; RUNSTEIN, 2018c).

A janela de mixagem, por sua vez, é a interface que apresenta ao usuário os canais de áudio da DAW, com o objetivo de simular os canais de uma mesa de som analógica. Também chamada de *mixer*, essa janela oferece acesso a todas as funções de mixagem e masterização da DAW. Entre as funcionalidades disponíveis estão os controles de volume de saída, geralmente representados por potenciômetros deslizantes, conhecidos como *faders*, e os controles de panorama, representados por potenciômetros radiais. Graças a essas representações gráficas,

a janela de mixagem é utilizada para tarefas como o nivelamento de volumes entre faixas, a modificação da posição dos elementos na imagem estéreo e a inserção de processamentos em cada faixa (SAVAGE, 2014a). Alguns desses conceitos serão desenvolvidos em mais detalhes nas seções seguintes. A Figura 17 ilustra as janelas de mixagem de diversas DAW.

Figura 17 – Janela de mixagem das DAW Pro Tools (superior esquerda), Cubase (superior direita), Logic (inferior esquerda), Reaper (inferior direita) e Studio One (direita central).



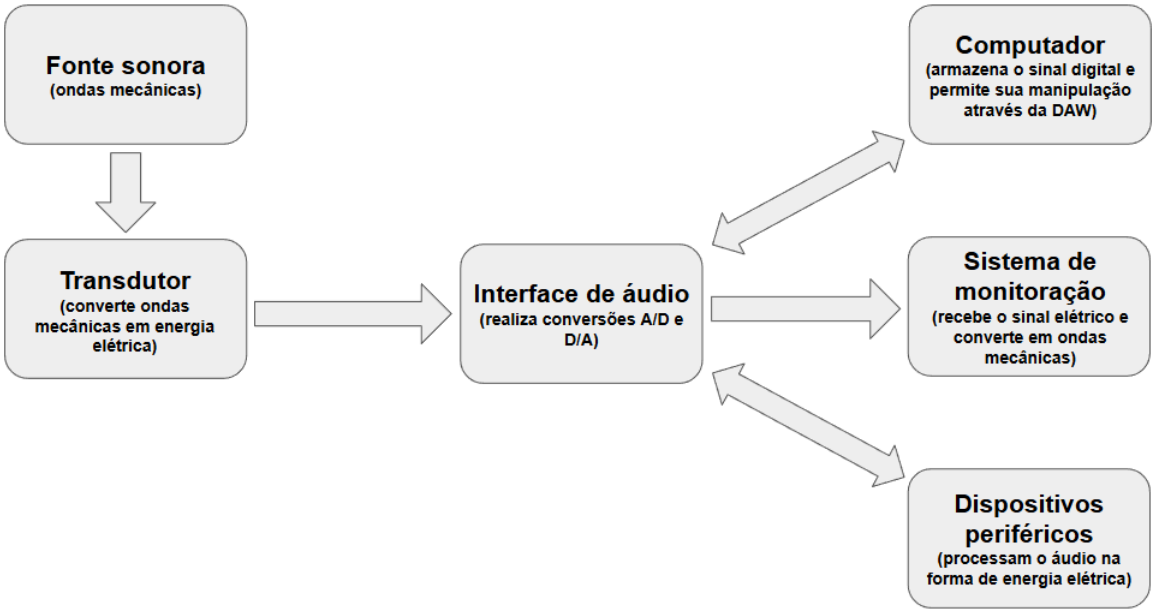
Fonte: adaptado de (HUBER; RUNSTEIN, 2018c).

Uma importante característica das DAW é a capacidade de utilizar aplicações remotas conhecidas como *plugin*. Essas aplicações podem ser descritas como processadores de sinal. Entre os tipos de processamento realizados por *plugin*, destacam-se a equalização, compressão e reverberação (SAVAGE, 2014a), que serão abordados em seções posteriores. Além disso, os *plugin* também permitem a integração de instrumentos virtuais dentro do ambiente de uma DAW, ampliando as possibilidades de criação e produção musical (HUBER; RUNSTEIN, 2018c).

Ao comparar o fluxo de sinal entre uma gravação analógica e uma digital, é possível identificar tanto semelhanças quanto diferenças. Em ambos os processos, a etapa de transdução sonora é a mesma, permitindo o uso dos mesmos microfones. No entanto, a principal diferença surge após a transdução.

No sistema analógico, o sinal permanece no domínio analógico durante todo o percurso, enquanto no digital, ele é convertido em dados discretos por um dispositivo chamado interface de áudio. Essas interfaces se conectam ao computador por meio de tecnologias como USB, FireWire ou Thunderbolt. No caso da gravação analógica, as informações são armazenadas em fitas magnéticas, já na digital, os dados são gravados na memória do computador. O roteamento do áudio, que no processo analógico é gerenciado pela mesa de som, no digital é realizado pela combinação da interface de áudio e da DAW. A Figura 18 apresenta um esquema de gravação em sistema digital.

Figura 18 – Exemplo do fluxo de sinal em uma gravação digital.



Fonte: acervo do autor.

Embora as ilustrações apresentadas forneçam uma visão geral, é importante destacar que variações na montagem e nas configurações dos equipamentos podem ocorrer entre diferentes processos de gravação. O mais relevante é que, ao término da captação, independente do método utilizado, as ideias concebidas durante a pré-produção tenham sido executadas pelos músicos e devidamente registradas em formato de áudio. No entanto, esse registro, por si só, não é suficiente para atender aos requisitos mercadológicos exigidos na produção de um fonograma. Por isso, o material gravado é encaminhado para a fase seguinte, conhecida como mixagem.

2.4 Mixagem

As etapas que sucedem a gravação são conhecidas como pós-produção. Durante essa fase, ocorrem processos essenciais como a mixagem e a masterização (PAIXÃO, 2013). Ambos são fundamentais para adaptar o material gravado aos diversos formatos disponibilizados pelos meios de reprodução. Essas etapas são importantes, pois influenciarão diretamente na qualidade da experiência de escuta do ouvinte.

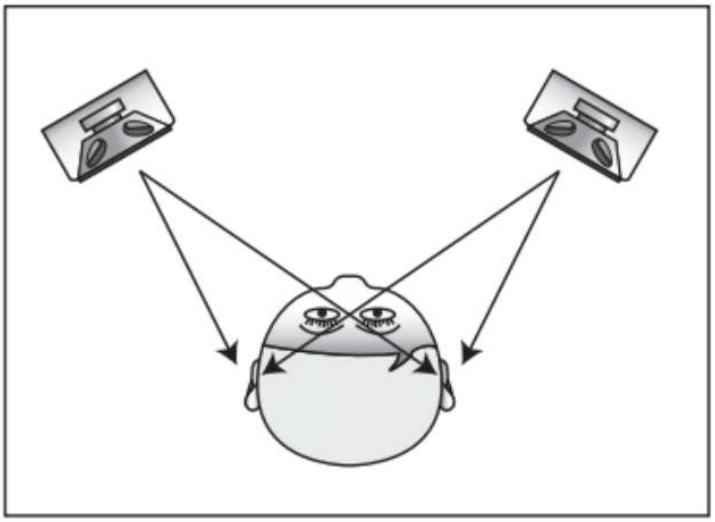
Definir o processo de mixagem pode ser desafiador, pois envolve tanto aspectos técnicos quanto artísticos, que são altamente personalizados de acordo com cada música. Em uma abordagem técnica, a mixagem pode ser vista como um processo de ajuste das faixas gravadas para garantir clareza sonora e inteligibilidade (CARVALHO; PEREIRA, 2017). Por outro lado, ela vai além da técnica ao buscar realçar a emoção da obra, favorecendo a performance, a musicalidade e executando ideias criativas que potencializam o impacto artístico da canção (IZHAKI, 2017i).

Uma metáfora comum para descrever o processo de mixagem é a da culinária. Assim como os ingredientes de uma receita precisam ser preparados e combinados, o material gravado passa por um processo de tratamento e organização. A mixagem envolve transformar esses elementos separados em um conjunto harmonioso, assim como o preparo dos ingredientes resulta em uma prato saboroso. Dessa forma, o material bruto gravado é trabalhado até se transformar em uma faixa final, pronta para ser apreciada.

Panorama e imagem estéreo

O formato de reprodução da faixa mixada dependerá do objetivo final, pois diferentes mídias utilizam diferentes formatos de som. Por exemplo, produções cinematográficas frequentemente usam o sistema *surround* 5.1, que possui seis canais de áudio. Já a maioria das produções musicais utiliza o estéreo, que trabalha com dois canais. Assim, o resultado final de uma mixagem musical é, geralmente, um arquivo de dois canais de áudio (SAVAGE, 2014b). A Figura 19 ilustra um sistema sonoro em estéreo.

Figura 19 – Exemplos de um sistema sonoro em estéreo.



Fonte: (IZHAKI, 2017k).

O estéreo trabalha com duas fontes sonoras, uma à esquerda e outra à direita do ouvinte, criando o chamado panorama ou imagem estéreo. A percepção de posicionamento dos instrumentos nesse espaço sonoro surge da diferença de volume entre as duas fontes (IZHAKI, 2017h). Além do volume, o cérebro também processa diferenças de tempo e frequências entre os sons captados por cada ouvido para determinar a localização espacial dos objetos sonoros (IZHAKI, 2017k). A igualdade entre as duas fontes sonoras gera a sensação de que o som está centralizado no panorama estéreo (HUBER; RUNSTEIN, 2018i).

Esse recurso é amplamente utilizado em produções musicais para gerar diferentes sensações. Alguns efeitos processam o áudio de maneira distinta entre as duas fontes sonoras,

criando a sensação de movimento dentro da música (SENIOR, 2018). Além disso, o uso do panorama estéreo facilita o posicionamento de elementos na mixagem, evitando conflitos de frequência entre eles e permitindo uma melhor organização espacial e clareza (SAVAGE, 2014a).

O posicionamento de uma faixa de áudio no panorama, dentro de uma DAW, pode ser determinado pelo uso de potenciômetros digitais. Esses potenciômetros utilizam escalas que mostram a distribuição de um som entre os canais direito e esquerdo. Quando um elemento está 100% à direita, o volume no canal esquerdo é nulo, e vice-versa (IZHAKI, 2017k). A Figura 20 exemplifica como o potenciômetro controla o panorama na DAW, permitindo um posicionamento preciso dos elementos da mixagem entre os dois canais.

Figura 20 – Exemplos de potenciômetros digitais utilizados para deslocar objetos no panorama estéreo.

Fonte: (IZHAKI, 2017k).

Equalização

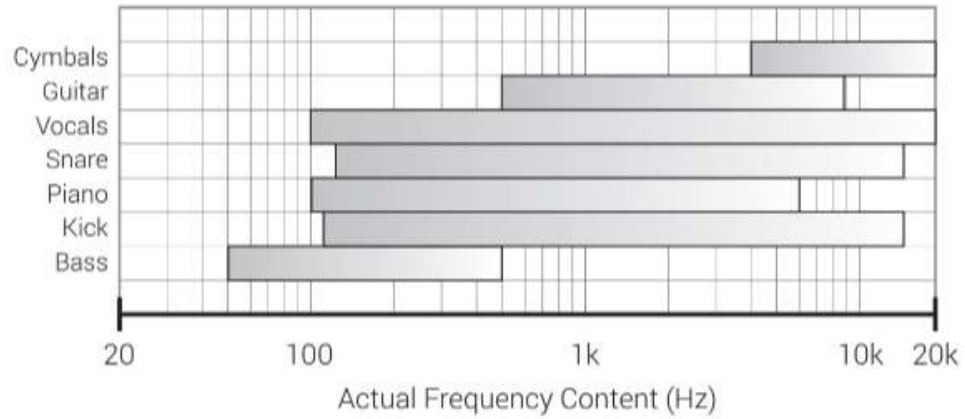
A sensação espacial dentro de uma mixagem pode ser expandida para além do eixo horizontal, alcançando também o eixo vertical. Enquanto o posicionamento horizontal é determinado pela diferença de volume de um elemento entre as fontes sonoras, a sensação de posicionamento no eixo vertical, muitas vezes, pode ser trabalhada ao manipular o espectro de frequências do material sonoro. Alguns autores da área de produção musical afirmam que sons mais agudos tendem a soar mais elevados e sons graves mais baixos, graças a efeitos psicoacústicos e características morfológicas do corpo humano (GIBSON, 2019b).

Os autores que defendem esse argumento partem do princípio de que as ondas sonoras de baixa frequência têm um comprimento de onda maior e se propagam pelo chão, gerando a sensação de que os graves vêm de baixo. As ondas de alta frequência, por sua vez, seriam mais direcionais e chegariam diretamente aos ouvidos, criando a percepção de que os agudos

vêm de cima. Além disso, eles também afirmam que o corpo humano possui duas cavidades ressonantes: a caixa torácica, que por ser maior, ressoa com frequências mais graves, e a cabeça, que por ser menor, ressoa com frequências mais agudas. A localização das duas cavidades no corpo ajudaria a reforçar essa percepção vertical (GIBSON, 2019b).

O processamento utilizado para alterar a amplitude das frequências de um elemento dentro de uma mixagem, e que consequentemente alteraria a sua percepção vertical, é a equalização (GIBSON, 2019b). A análise das frequências em uma produção musical é crucial, já que cada instrumento apresenta uma distribuição única, influenciando diretamente seu timbre. Por exemplo, instrumentos como o contrabaixo concentram-se em frequências graves, enquanto pratos de bateria possuem mais conteúdo nas frequências agudas. Já o piano pode abranger uma ampla faixa, desde os graves até os agudos (IZHAKI, 2017f). A Figura 21 compara as faixas de frequências de diversos instrumentos como bateria, violão, voz humana e piano, destacando suas diferentes características sonoras.

Figura 21 – Distribuição das frequências de diversos instrumentos através do espectro de frequência sonoro.



Fonte: adaptado de (IZHAKI, 2017f).

Como é possível perceber na Figura 21, algumas regiões de frequência são ocupadas por diversos instrumentos. Esse tipo de fenômeno não é desejável dentro de um processo de mixagem, pois pode resultar em problemas como a falta de inteligibilidade dos detalhes de cada instrumento. A sobreposição de frequências pode ocultar características importantes de um sinal de áudio ou até fazer com que a música soe confusa ou sobrecarregada. Por esse motivo, o processo de equalização é tão importante na mixagem (IZHAKI, 2017f).

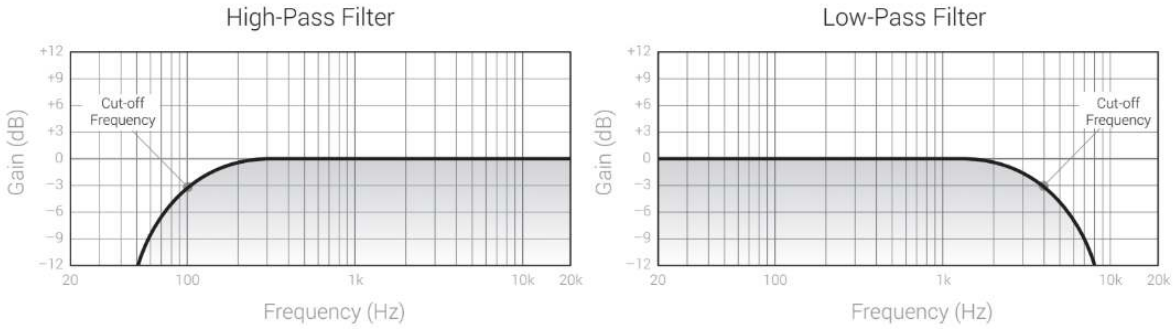
A equalização é realizada através de um dispositivo chamado equalizador, que permite atenuar ou amplificar faixas específicas de frequências dentro do espectro de um sinal de áudio. Esse processo ajuda a moldar o conteúdo sonoro, equilibrando o espectro de frequências e realçando os aspectos desejados dos timbres dos instrumentos. Com isso, é possível aprimorar a definição dos elementos musicais, contribuindo para uma mixagem mais clara e precisa, na qual cada instrumento tem seu espaço bem definido na produção final (IZHAKI, 2017f). O

equalizador pode ser encontrado tanto em formato analógico, como circuitos e dispositivos, quanto em digital, como um *plugin* (HUBER; RUNSTEIN, 2018d).

Em termos de eletrônica e processamento de sinais, os equalizadores podem ser considerados como filtros que modificam o espectro de frequência de um sinal de áudio. Eles são classificados conforme o tipo de processamento realizado (IZHAKI, 2017f). Entre as categorias de equalizadores, destacam-se:

- **Filtros de passagem**: são filtros que operam a partir de uma frequência de corte, permitindo que as frequências acima ou abaixo dela permaneçam inalteradas, enquanto atenuam completamente as demais. Sua implementação eletrônica é relativamente simples, bastando a combinação de um capacitor com um resistor. Filtros que preservam frequências graves são chamados de passa-baixas, enquanto os que preservam frequências agudas são conhecidos como passa-altas (IZHAKI, 2017f). A Figura 22 exemplifica o funcionamento desses filtros.

Figura 22 – Exemplo de filtros de passagem.



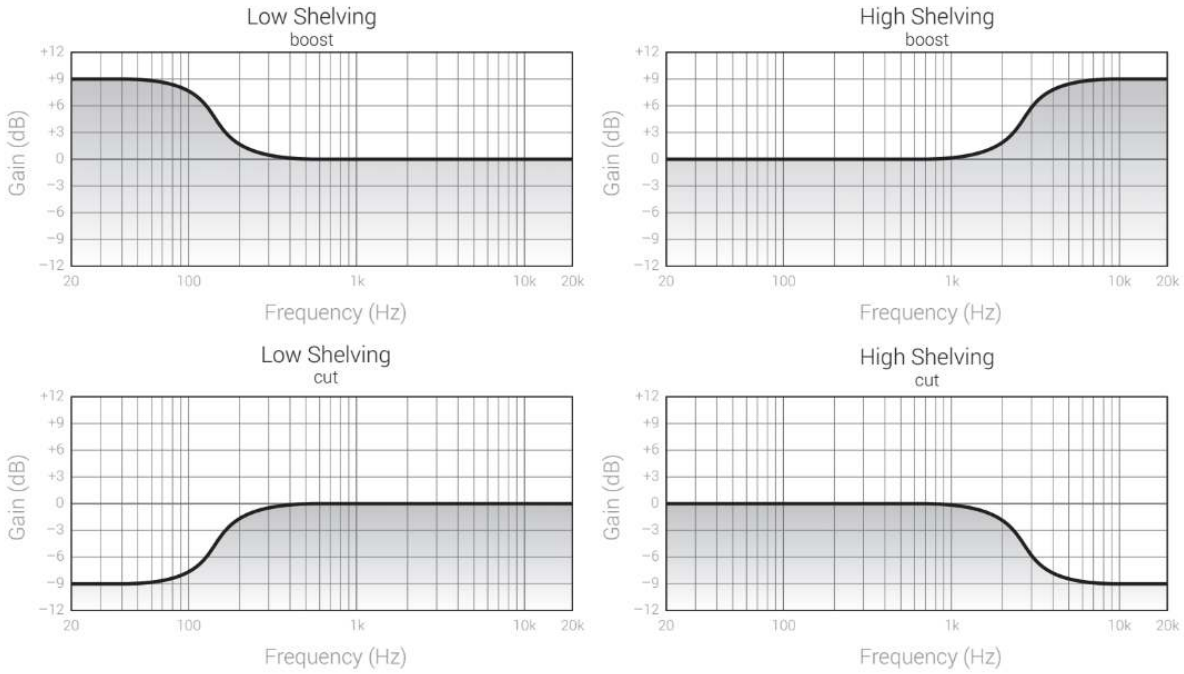
Fonte: (IZHAKI, 2017f).

- **Filtros de prateleira**: são chamados assim devido ao formato de sua curva de resposta, que pode lembrar uma prateleira. Esses filtros também operam com base em uma frequência de referência. No entanto, diferentemente dos filtros de passagem, que eliminam as frequências fora da faixa de corte, os filtros de prateleira permitem tanto amplificar quanto atenuar as frequências além da frequência de referência (IZHAKI, 2017f). A Figura 23 exemplifica o funcionamento desse tipo de filtro.

- **Filtros paramétricos**: também são conhecidos como filtros de pico (HUBER; RUNSTEIN, 2018d), filtros de banda ou filtros de sino. Eles atuam sobre uma faixa de frequências específica, permitindo sua amplificação ou atenuação. Alguns desses filtros também possibilitam o ajuste da largura da banda afetada por meio de um parâmetro chamado

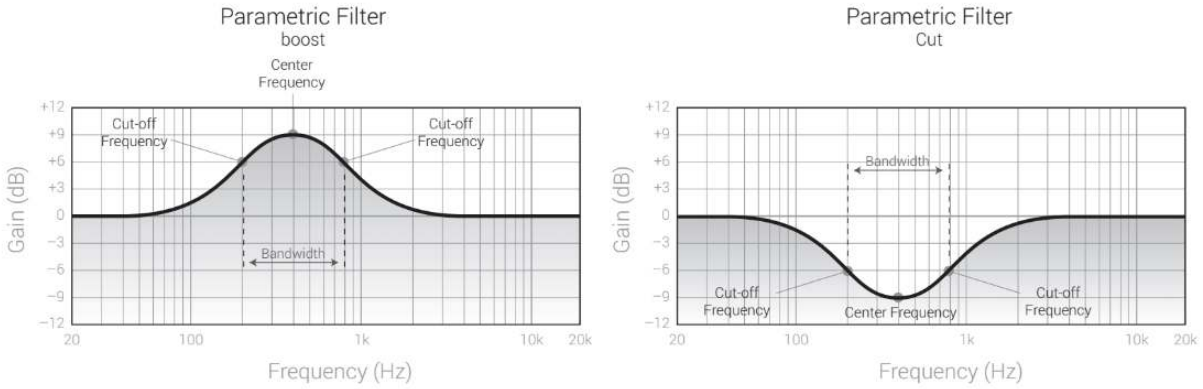
fator de qualidade[1], ou Q. O nome desse tipo de filtro se deve ao seu uso em equalizadores paramétricos, onde foram inicialmente introduzidos (IZHAKI, 2017f). A Figura 24 ilustra a sua ação.

Figura 23 – Exemplo de filtros de prateleira.



Fonte: (IZHAKI, 2017f).

Figura 24 – Exemplo de filtros paramétricos.



Fonte: (IZHAKI, 2017f).

O tema da equalização e equalizadores é vasto, oferecendo uma enorme gama de possibilidades. Embora seja possível expandir esse assunto por muitos parágrafos, isso poderia

[1] Apesar do nome utilizar a palavra qualidade, o termo neste contexto não carrega juízo de valor, mas representa a característica da razão de abertura da banda.

tornar a leitura deste trabalho muito extensa e cansativa. O que realmente importa é que o leitor compreenda os princípios básicos dessa técnica e como ela ajuda a resolver questões fundamentais no processo de mixagem, garantindo que os elementos sonoros tenham seu espaço definido e contribuam para uma mixagem equilibrada e clara.

Processadores de dinâmica

Além do espectro de frequências, outra característica relevante para a mixagem é a dinâmica da performance. No contexto de mixagem, a dinâmica se refere à variação da amplitude do sinal ao longo do tempo. Ela pode ser analisada de duas maneiras: macrodinâmica e microdinâmica. A macrodinâmica lida com as variações de nível ao longo de uma faixa completa, enquanto a microdinâmica analisa as variações de volume de cada nota individualmente durante a execução do instrumento (IZHAKI, 2017g).

Dentro do contexto da microdinâmica, é possível introduzir o conceito de envelope dinâmico, que se refere às variações de amplitude que ocorrem dentro de cada nota ou batida. Esse aspecto é fundamental para a construção do timbre de um instrumento, pois ao lado do espectro de frequências, ele influenciará diretamente na percepção sonora e no caráter de cada som (IZHAKI, 2017g).

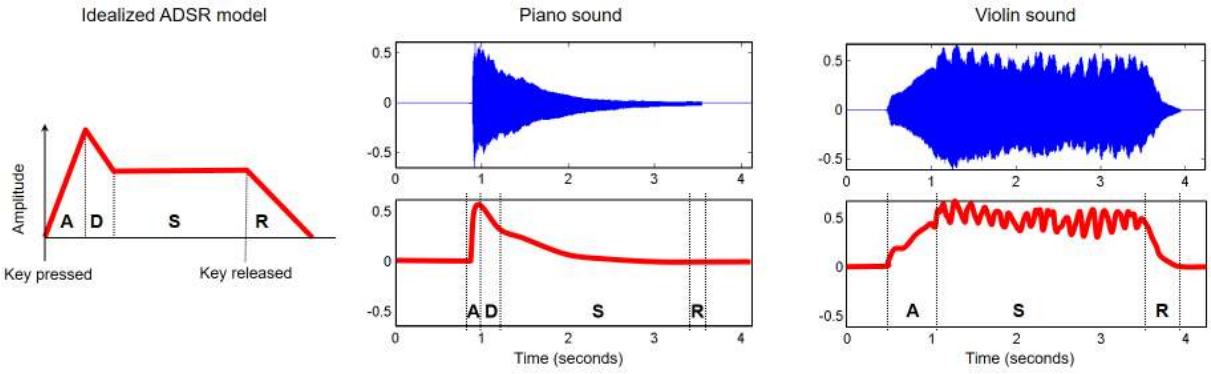
O envelope dinâmico de um sinal pode ser analisado a partir de um sistema classificatório conhecido como ADSR (HUBER; RUNSTEIN, 2018i). Este sistema é composto por quatro estágios:

- ***Attack*:** diz respeito ao tempo que o som leva para atingir o seu valor de amplitude máximo, a partir do momento em que é tocado (HUBER; RUNSTEIN, 2018i).
- ***Decay*:** considera a velocidade de decaimento da amplitude do som após atingir o pico de valor do sinal por intermédio do *attack* (HUBER; RUNSTEIN, 2018i).
- ***Sustain*:** trata-se da duração do tempo de sustentação de uma nota após o *attack* e o *decay* (HUBER; RUNSTEIN, 2018i).
- ***Release*:** refere-se ao tempo que a nota continua soando no instrumento, após o *sustain*, até se extinguir completamente (HUBER; RUNSTEIN, 2018i).

A Figura 25 ilustra o envelope dinâmico de um sinal, onde cada etapa é destacada com os algarismos correspondentes. Na imagem, é possível ver a aplicação desse conceito comparando uma nota de piano, que tem um *attack* rápido e um *release* curto, com uma nota de violino, cujo *attack* é mais gradual e o *sustain* mais prolongado, destacando as diferenças nos envelopes dinâmicos entre esses dois instrumentos.

Quando o sinal de áudio de uma nota contém um trecho curto e com amplitude elevada, semelhante a um ruído, ele é chamado de transiente. Na prática, os transientes geralmente

Figura 25 – Representação das componentes que compõem o envelope dinâmico.



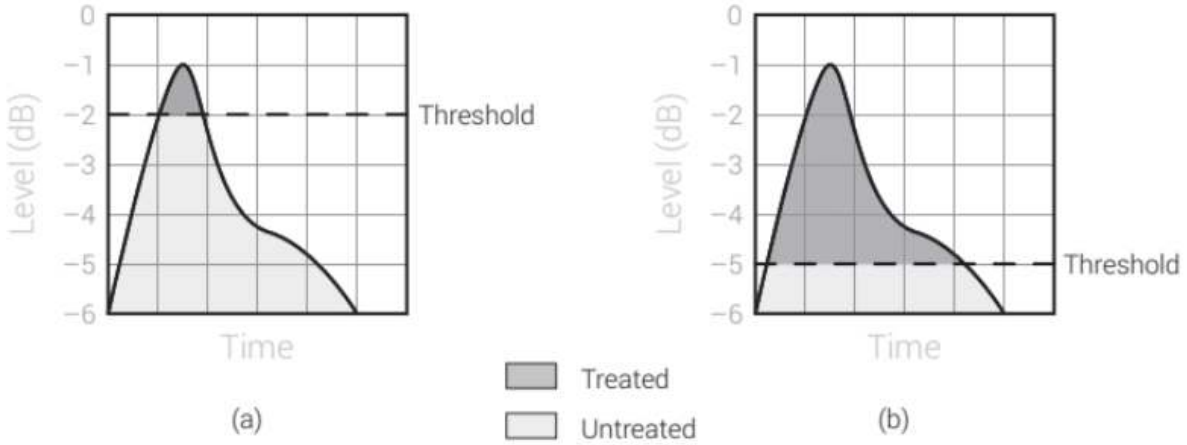
Fonte: (MÜLLER, 2021b).

representam o momento em que ocorre a interação física com o instrumento, como o martelo de um piano atingindo as cordas, o toque de um dedo nas cordas de um violão ou uma baqueta batendo em uma peça de bateria. Esses transientes são cruciais para definir o caráter e a articulação do som gerado (MÜLLER, 2021b).

No processo de mixagem, algumas ferramentas são empregadas para manipular e ajustar a dinâmica de um sinal de áudio. As principais ferramentas para essa função são os compressores e os expansores, também conhecidos como *expanders*. Essas ferramentas permitem controlar a amplitude do sinal, ajudando a comprimir variações extremas ou expandir a faixa dinâmica (IZHAKI, 2017g). Apesar das diferenças entre compressores e expansores, ambos compartilham alguns parâmetros fundamentais que são cruciais para o controle dinâmico. Entre eles, destacam-se:

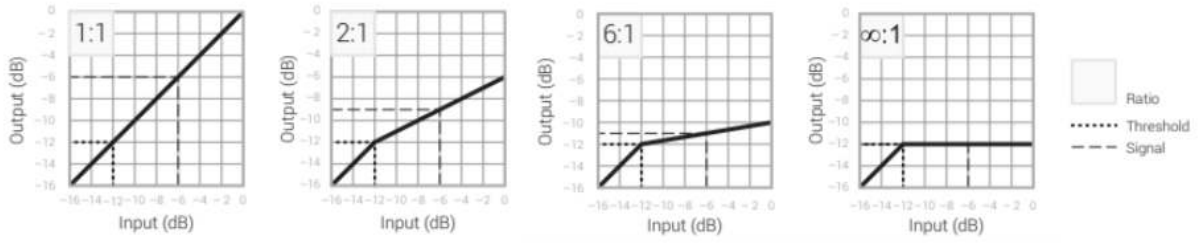
- ***Gain*:** esse parâmetro controla a atenuação ou amplificação do nível do sinal que entra no processador, ajustando a intensidade do som (IZHAKI, 2017b).
- ***Threshold*:** esse parâmetro define um valor de referência que, quando o sinal de áudio ultrapassa, ativa o funcionamento do dispositivo. A Figura 26 ilustra o seu funcionamento em um compressor, mostrando a parcela do sinal que será comprimida (IZHAKI, 2017b).
- ***Ratio***: se o *threshold* determina a partir de qual ponto o processador atuará, o *ratio* define a intensidade desse processamento. Ele é representado por uma relação numérica entre a amplitude de entrada e a de saída. Em um compressor, por exemplo, um *ratio* de 2:1 significa que, para cada 2 dB que o sinal excede o *threshold*, ele será reduzido para 1 dB na saída (IZHAKI, 2017b). A Figura 27 ilustra essa relação em termos de amplitude de entrada e saída, enquanto a Figura 28 mostra o efeito ao longo do tempo.
- ***Attack*:** o *attack* é um parâmetro temporal que define a rapidez com que o processamento começa a atuar após o sinal ultrapassar o valor do *threshold*. Ele determina quanto tempo

Figura 26 – Exemplificação do parâmetro *threshold* em um compressor.



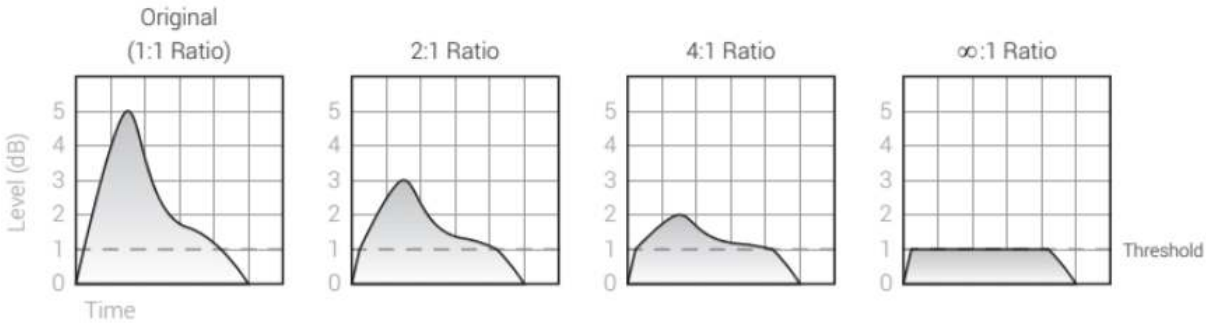
Fonte: (IZHAKI, 2017b).

Figura 27 – Exemplificação do parâmetro *ratio* através da relação das amplitudes de um sinal na entrada e na saída do compressor.



Fonte: (IZHAKI, 2017b).

Figura 28 – Exemplificação do parâmetro *ratio* através de sinais representados no tempo.

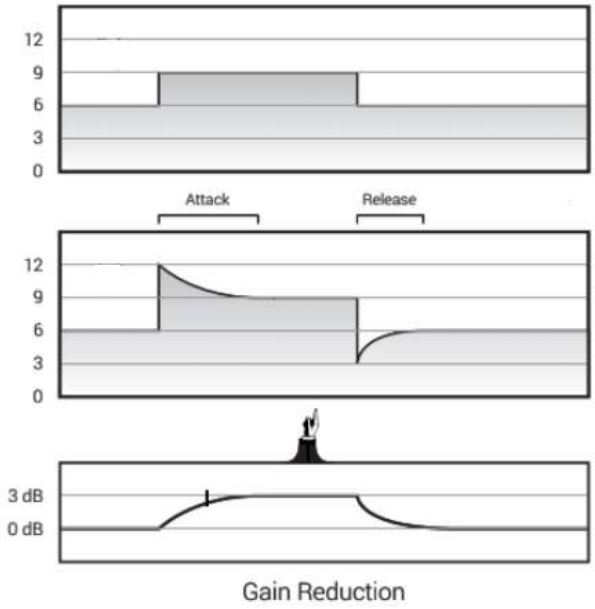


Fonte: (IZHAKI, 2017b).

levará para que a taxa de processamento atinja a proporção configurada pelo *ratio*. Um *attack* rápido inicia a atuação imediatamente, enquanto um *attack* mais lento permite que parte do transiente inicial do sinal seja preservada antes da atuação completa (IZHAKI, 2017b).

- ***Release*:** de maneira semelhante ao *attack*, o *release* também é um parâmetro temporal. No entanto, ao invés de determinar a rapidez com que o processamento começa, o *release* define por quanto tempo o processamento continuará a atuar, mesmo após o sinal original retornar para abaixo do valor do *threshold*. Isso influencia o tempo de recuperação do sinal à sua amplitude original (IZHAKI, 2017b). A Figura 29 demonstra como os parâmetros de *attack* e *release* atuam sobre um compressor.

Figura 29 – Exemplo de como o *attack* e o *release* de um compressor atuam sobre um sinal.

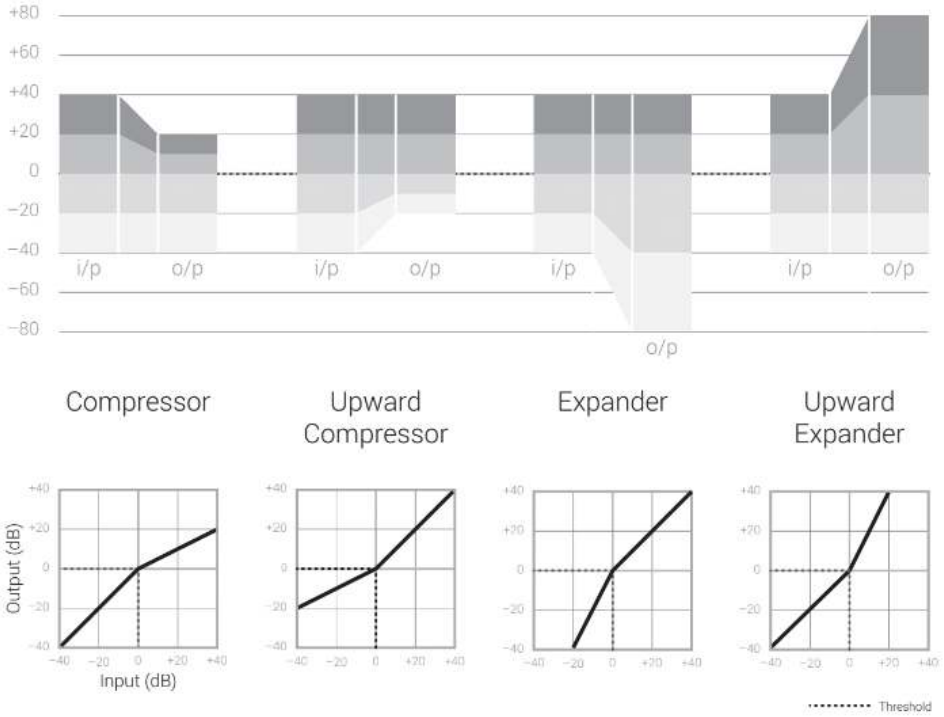


Fonte: (IZHAKI, 2017b).

A principal diferença entre um compressor e um *expander* está em como cada dispositivo processa o sinal em relação ao *threshold*. O compressor reduz a amplitude da porção do sinal que ultrapassa o *threshold*, enquanto o *expander* atua na parte do sinal abaixo desse valor, diminuindo sua amplitude. Além desses dois, existem também o *upward compressor*, que aumenta o nível do sinal abaixo do *threshold*, e o *upward expander*, que amplifica os sinais acima do *threshold* (IZHAKI, 2017g). A Figura 30 exemplifica o funcionamento desses dispositivos.

Quando um compressor tradicional é configurado com um *ratio* de valor muito alto, ele passa a ser chamado de limitador ou *limiter*, pois limita o sinal a um valor máximo de amplitude, impedindo que ele o ultrapasse. Da mesma forma, um *expander* com *ratio* muito alto é chamado de *gate*, já que reduz praticamente todo o sinal abaixo do *threshold*, funcionando

Figura 30 – Forma de atuação dos diferentes processadores de dinâmica.



Fonte: (IZHAKI, 2017g).

como uma espécie de portão que se fecha para os sinais mais fracos (IZHAKI, 2017g). A Figura 31 exemplifica o funcionamento desses dois dispositivos.

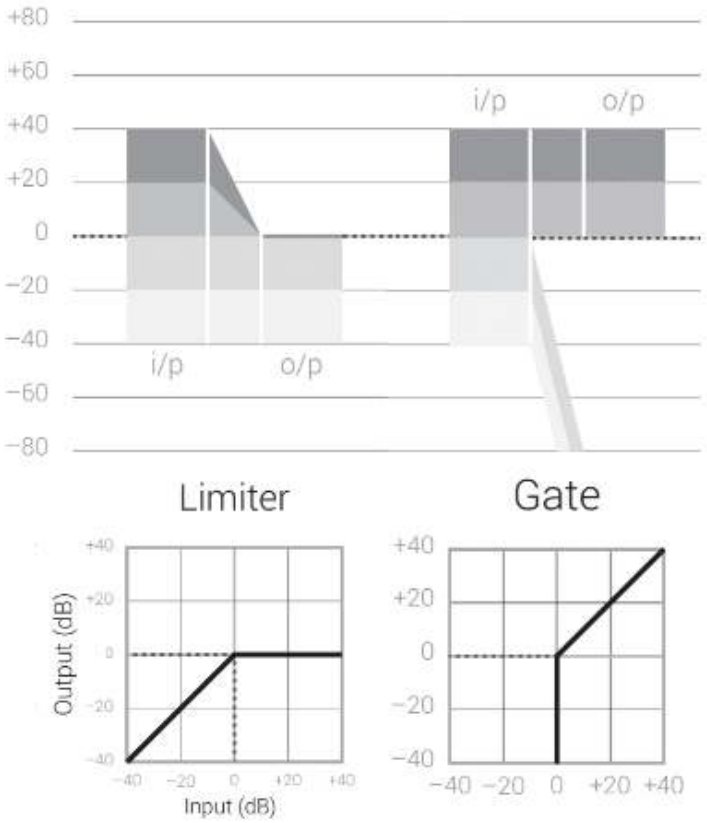
O uso de compressores na mixagem vai além das tarefas relacionadas ao controle de dinâmica. Isso ocorre porque os circuitos dos compressores físicos podem ser projetados com diferentes métodos e componentes, como válvulas, transistores e sensores ópticos. Essas variações construtivas conferem aos compressores sonoridades e características únicas. Assim como os equalizadores, os compressores podem ser encontrados tanto em dispositivos físicos quanto em formatos digitais, como *plugin* (IZHAKI, 2017b).

Reverberação

Um terceiro tipo de processamento fundamental na mixagem é a adição de efeitos de reverberação às faixas de áudio. A reverberação pode ser explicada como o som que persiste em um espaço após a emissão sonora ter sido interrompida. Um exemplo desse fenômeno é o prolongamento do som de uma palma executada em uma sala vazia. As ondas sonoras são refletidas nas superfícies ao redor e retornam ao ouvinte após múltiplas reflexões, criando o *reverb* (IZHAKI, 2017m).

A estética sonora da reverberação é influenciada por diversos fatores. Um desses fatores está relacionado ao formato e ao material das superfícies que refletem o som, impactando suas características harmônicas ao reforçar ou atenuar certas frequências (WILMERING et al., 2020).

Figura 31 – Forma de atuação do *limiter* e do *gate*.



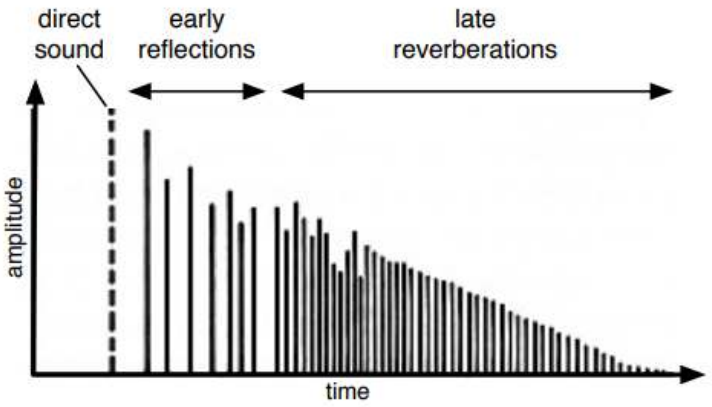
Fonte: adaptado de (IZHAKI, 2017g).

A propagação do som e sua absorção por materiais também afetam o decaimento da amplitude (IZHAKI, 2017m). O fenômeno da reverberação pode ser dividido em três estágios principais: o som direto, as reflexões iniciais e as reflexões tardias, cada uma contribuindo para o caráter espacial do som (WILMERING et al., 2020).

- **Som direto**: refere-se ao som emitido pela fonte sonora que chega primeiro aos ouvidos do ouvinte, sem interferências de reflexões ou reverberações. Ele é o som puro e não faz parte do *reverb*. Esse som viaja pela menor distância entre a fonte e o ouvinte, trafegando em linha reta, sem interagir com superfícies refletoras (IZHAKI, 2017m). Por conta disso, ele é percebido de forma clara e definida, sendo essencial para a percepção da localização espacial da fonte sonora.
- **Reflexões iniciais**: são as primeiras a alcançar o ouvinte após o som direto, influenciando significativamente a percepção do tamanho do ambiente e o timbre do *reverb* (WILMERING et al., 2020). Elas ocorrem quando o sinal é refletido por uma ou duas superfícies e chegam ao ouvinte em até 100 ms após o som direto, dependendo das dimensões do ambiente (IZHAKI, 2017m).
- **Reflexões tardias**: são compostas por diversas reflexões difusas que ocorrem após as

reflexões iniciais. A diferença temporal entre essas reflexões não é facilmente percebida pelo ouvido humano, mas pode ser calculada por meio de métodos estatísticos (WILMERING et al., 2020). Essas reflexões também são conhecidas como "cauda do *reverb*" e, por se encontrarem com um número maior de superfícies, sofrem maior absorção, o que leva ao decaimento da amplitude sonora (IZHAKI, 2017m). A Figura 32 ilustra o fenômeno relacionando a amplitude do som com o tempo decorrido.

Figura 32 – Representação das reflexões de uma reverberação.



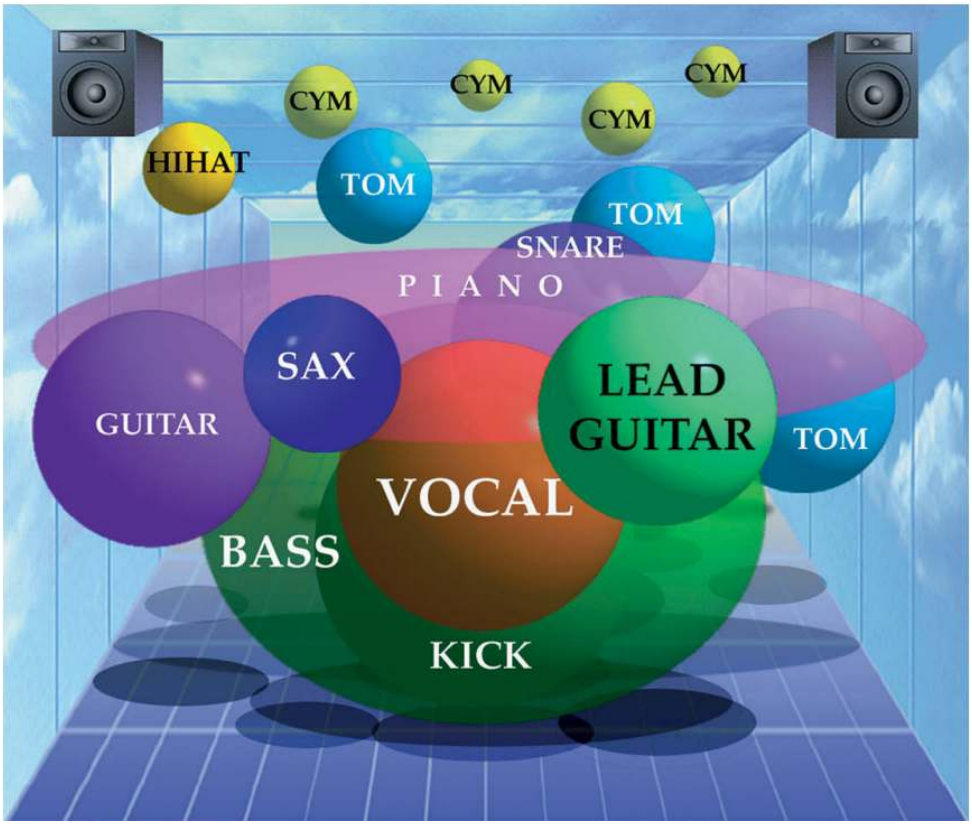
Fonte: (WILMERING et al., 2020).

Existem vários métodos que podem gerar *reverb* além das reflexões acústicas naturais do ambiente. Os diferentes métodos de produzir a reverberação irão gerar efeitos com timbres diferentes. Um *reverb* pode ser criado, por exemplo, em uma câmara de eco (*reverb chamber*), através do uso de molas (*spring reverb*), de chapas metálicas (*plate reverb*) ou de emulações digitais (*reverb* convolucional) (IZHAKI, 2017m).

Na produção musical, o *reverb* é utilizado por diversas razões. Primeiramente, ele ajuda a criar uma sensação mais natural ao som, imitando as reflexões de um ambiente real, como no caso de uma guitarra captada sem o uso de microfones. Além disso, o *reverb* pode homogeneizar a mixagem ao adicionar uma coesão auditiva entre diferentes elementos processados. Ele também pode criar profundidade, oferecendo ao ouvinte uma percepção espacial, tornando a música mais imersiva e dimensional (IZHAKI, 2017m).

A combinação de efeitos de *reverb*, ajustes de dinâmica e equalização completam a tridimensionalidade de uma mixagem ao recriar uma sensação de profundidade. Isso porque o som de objetos que estão mais distante soam de forma diferente dos mais próximos. Através da combinação dessas ferramentas é possível emular alguns desses efeitos (GIBSON, 2019b). Essa abordagem tri-dimensional contribui para o valor artístico da mixagem. A Figura 33 demonstra como diferentes elementos sonoros podem utilizar esses recursos para criar imagens sonoras dentro de uma música.

Figura 33 – Distribuição dos elementos sonoros em uma mixagem.



Fonte: (GIBSON, 2019a).

Apesar da relevância dos processamentos de compressão, equalização e reverberação para a mixagem, o processo vai muito além dessas ferramentas. Uma ampla variedade de efeitos, técnicas e processamentos está disponível para explorar e trabalhar diferentes estéticas musicais. Entre os efeitos mais conhecidos, destacam-se o *delay* (IZHAKI, 2017c), a distorção (IZHAKI, 2017d), e os efeitos de modulação (IZHAKI, 2017j) como *chorus*, *flanger*, *phaser* e *tremolo*.

Como mencionado no início desta seção, a tarefa de explicar o processo de mixagem pode ser complexa devido à subjetividade e individualidade presentes em cada mixagem. Compreender os principais tipos de processamento aplicados nesse contexto é uma maneira eficiente de capturar a essência do processo. Os áudios a seguir exemplificam essa diferença, comparando a mesma música antes e depois de passar pelo processo de mixagem, oferecendo uma visão prática das melhorias e ajustes feitos ao longo dessa fase crucial.

Trecho musical antes da mixagem

Trecho musical depois da mixagem

2.5 Masterização

A masterização é a etapa final dentro do processo de produção musical em relação ao processamento de áudio. É um estágio é crucial, pois refina ainda mais todo o trabalho realizado na mixagem (IZHAKI, 2017l). Enquanto a mixagem otimiza individualmente cada faixa, a masterização visa equilibrar as faixas para garantir que o álbum soe coeso. Entre os aspectos trabalhados estão o volume, o equilíbrio de frequências, a faixa dinâmica e os efeitos, ajustando cada um para criar um produto final harmônico e consistente (SAVAGE, 2014c).

Outra importante função da masterização dentro do processo de produção musical é servir como uma segunda avaliação do trabalho de pós-produção. Os engenheiros de masterização geralmente possuem vasta experiência (HUBER; RUNSTEIN, 2018f) e operam em ambientes acústicos otimizados para esse processo (IZHAKI, 2017l). Isso permite, por exemplo, que eles possam identificar e corrigir eventuais problemas que tenham passado despercebidos durante a mixagem (SAVAGE, 2014c).

Por ser a última etapa de edição de áudio antes da distribuição, a masterização também prepara o material para atender às exigências técnicas de diferentes mídias de reprodução sonora. Por exemplo, ao masterizar para vinil, é essencial ajustar as frequências graves e as fases entre os canais direito e esquerdo devido às limitações físicas do formato (IZHAKI, 2017l). Já mídias como CD e DVD lidam com diferentes taxas de amostragem, com o CD utilizando 44,1 kHz e o DVD, 48 kHz (HUBER; RUNSTEIN, 2018h).

A masterização encerra o processo de pós-produção do ponto de vista técnico, mas depois disso, o material deve ser distribuído e vendido, abrindo novas etapas que envolvem aspectos como *marketing*, gerenciamento de vendas, lançamento, campanhas, e outros temas administrativos (HUBER; RUNSTEIN, 2018e). No entanto, como esses tópicos não são de grande relevância para o contexto deste trabalho, eles não serão abordados aqui.

3 A BATERIA

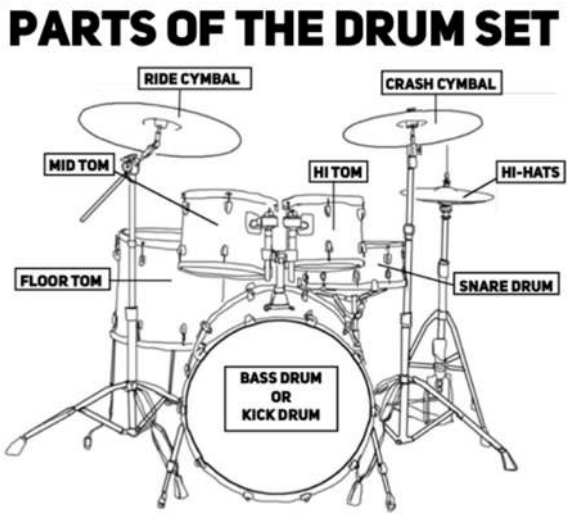
Para compreender a ambientação do trabalho é preciso conhecer seu principal objeto de estudo: a bateria. Trata-se de um instrumento musical amplamente utilizado na atualidade. Classificada como instrumento de percussão (LEE, 2019), a bateria é essencial na música popular contemporânea, sendo uma das principais responsáveis por definir o andamento rítmico de uma canção (ABE; MURAKAMI; MIURA, 2012).

3.1 Partes de uma bateria

Apesar de ser frequentemente vista como um único instrumento, a bateria é, na realidade, um conjunto formado por diferentes instrumentos de percussão. Esse agrupamento começou a se consolidar no final do século XIX, quando, por razões econômicas e logísticas, surgiu a prática de tocar múltiplos instrumentos ao mesmo tempo (HARTIGAN, 1995). Além disso, os componentes de uma bateria moderna carregam influências de várias culturas, refletindo a diversidade de suas origens (JOHNSON, 2021).

As diferentes combinações dos instrumentos que compõe uma bateria recebem o nome de *kits* de bateria. Embora não exista uma regra fixa para a configuração de um *kit*, a maioria dos modelos amplamente utilizados é formada principalmente por tambores e pratos percussivos. A Figura 34 apresenta um exemplo de *kit* de bateria tradicional, que inclui bumbo, caixa, tons e surdos, além de um chimbal e diferentes tipos de pratos (ZHANG; CALLISON-BURCH, 2023).

Figura 34 – Exemplo de um *kit* de bateria tradicional.

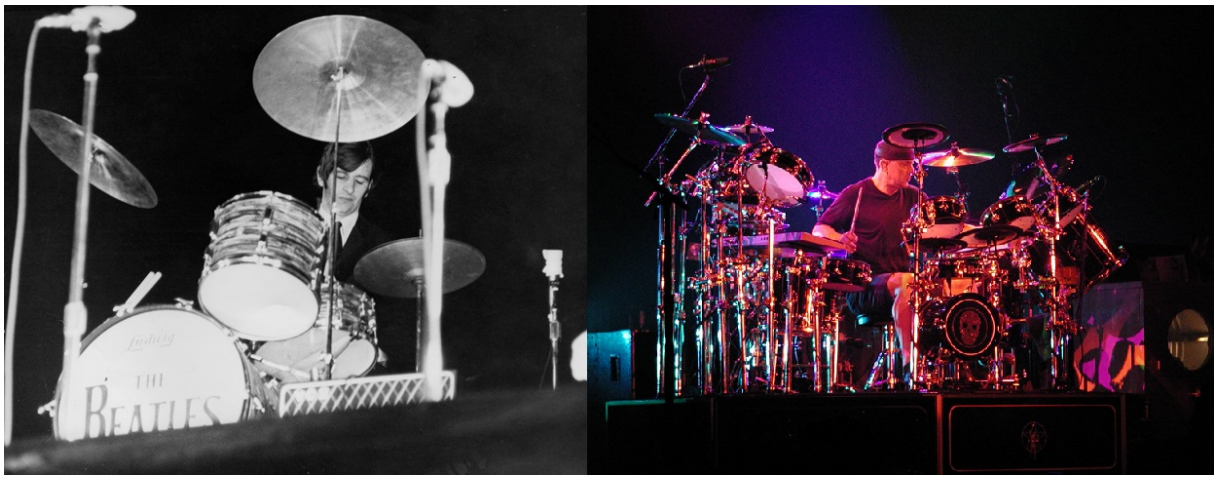


Fonte: (ZHANG; CALLISON-BURCH, 2023).

A quantidade de peças em um *kit* de bateria é determinada por diversos fatores, que vão desde o gosto pessoal do baterista até questões de técnica, estética e o gênero musical que está sendo tocado. Para ilustrar essa diferença, podemos citar Ringo Starr, baterista da

lendária banda The Beatles, que é famoso por utilizar *kits* de bateria extremamente pequenos e simples, em comparação com Neil Peart, baterista do trio de *rock* progressivo Rush, conhecido por seus *kits* extremamente amplos e complexos, repletos de elementos para sua execução altamente técnica e elaborada. A Figura 35 destaca essa diferença, comparando visualmente os dois bateristas e seus respectivos *kits* lado a lado.

Figura 35 – Comparação entre os *kits* de bateria de Ringo Starr e Neil Peart.



Fonte: (BARCHARD, 1964), (FRANGI, 2004).

As subseções a seguir apresentam as peças mais comuns encontradas nos *kits* de bateria.

3.2 Bumbo

O bumbo é o maior tambor de uma bateria. Devido à sua construção, ele é responsável por produzir o um dos sons mais fortes e graves de todo o *kit*. Por essa razão, também é chamado de *bass drum* ou *kick*, em inglês. A Figura 36 apresenta um exemplo de um bumbo de bateria.

Devido ao seu tamanho, o bumbo é apoiado no chão, sustentado por dois suportes, com peles posicionadas verticalmente. Por essa configuração, ele é geralmente tocado com os pés, utilizando um batedor feito de material macio, como feltro ou borracha. O batedor é acionado por um pedal, cujo mecanismo inclui mola, eixo, correntes e engrenagens para movimentar o batedor em direção à pele do bumbo, gerando o som grave característico. A Figura 37 ilustra o mecanismo de pedal usado para tocar o bumbo.

A característica sonora forte e marcante do bumbo faz com que ele seja amplamente utilizado para marcar o tempo forte do compasso em uma música. Ele desempenha um papel fundamental na construção rítmica de diversos gêneros musicais, fornecendo uma base sólida. O áudio a seguir exemplifica a sonoridade típica de um bumbo de bateria.

Figura 36 – Exemplo de um bumbo de uma bateria.

Fonte: acervo do autor.

Figura 37 – Mecanismo de pedal utilizado para tocar o bumbo.

Fonte: acervo do autor.

Som do bumbo

De modo geral, os *kits* de bateria possuem apenas um bumbo. Contudo, em estilos específicos como o *heavy metal*, alguns bateristas optam por incluir mais de um bumbo em seu *kit*. Isso ocorre porque, nesses estilos, as linhas de bumbo frequentemente apresentam muitas notas rápidas, permitindo que o baterista as execute utilizando ambas as pernas. A Figura 38 ilustra o *kit* do baterista brasileiro Eloy Casagrande, que adota essa configuração.

Figura 38 – Eloy Casagrande utilizando dois bumbos em seu *kit* de bateria.

Fonte: (RASCHKA, 2014).

Outro recurso amplamente utilizado nesse contexto é o mecanismo conhecido como pedal duplo. Esse tipo especial de pedal conta com dois batedores, controlados individualmente por dois pedais distintos. Assim, os bateristas conseguem executar uma grande quantidade de notas rápidas utilizando apenas um único bumbo no *kit*. A Figura 39 apresenta um exemplo de pedal duplo.

3.3 Caixa

Assim como o bumbo, a caixa (ou *snare*), também é um tambor especial dentro de um *kit* de bateria. O que a diferencia de um tambor comum é a esteira de fios metálicos instalada em sua pele inferior[1]. Esse mecanismo confere à caixa uma sonoridade única, distinta dos outros tambores do *kit*. A Figura 40 ilustra uma caixa e como a esteira é instalada nela.

[1] As peles superiores dos tambores, que recebem o impacto das baquetas, são popularmente conhecidas como peles batedeiras. Por outro lado, as peles inferiores são chamadas de peles de resposta.

Figura 39 – Exemplo de um pedal duplo.

Fonte: acervo do autor.

Figura 40 – Exemplo de uma caixa de uma bateria.

Fonte: acervo do autor.

Ao contrário do bumbo, que é tocado com os pés, a caixa é tocada com as mãos. Para isso, os bateristas geralmente utilizam um par de baquetas. Essas baquetas são instrumentos essenciais para o controle e a precisão da execução rítmica na caixa, permitindo uma vasta gama de dinâmicas e articulações no som. Elas podem ser fabricadas com diferentes materiais, como madeira ou fibra, e vêm em uma variedade de tamanhos, espessuras e formatos de ponteira, adaptando-se a diversas técnicas e estilos de execução. A Figura 41 ilustra diferentes tipos de baqueta.

Figura 41 – Exemplos de baquetas utilizadas por bateristas.

Fonte: acervo do autor.

Devido às suas características sonoras, a caixa, juntamente com o bumbo, desempenha um papel central na levada rítmica de uma música. Existem diversas técnicas para tocar a caixa, o que contribui para sua expressividade. Entre as mais comuns, destacam-se: tocar atacando apenas a pele superior, tocar atacando apenas o aro, tocar atacando pele e aro simultaneamente (técnica conhecida como *rimshot*) e tocar com a esteira afrouxada, criando uma variedade de timbres e efeitos sonoros. Os áudios a seguir demonstram a diversidade de possibilidades sonoras da caixa.

Som da caixa (pele)

Som da caixa (aro)

Som da caixa (*rimshot*)

Som da caixa (esteira afrouxada)

3.4 Tons e surdos

Os demais tambores que compõem uma bateria são conhecidos como tons e surdos. Apesar da diferença nos nomes, eles apresentam construções, funções e sonoridades bastante semelhantes. A principal distinção está em suas dimensões e na forma de apoio: enquanto os tons são menores e suspensos no *kit*, os surdos possuem dimensões maiores e são apoiados no chão com suportes. Por essa razão, os surdos também são conhecidos como *floor tom* e produzem uma sonoridade mais grave em comparação aos tons tradicionais. A Figura 42 mostra exemplos de tons, e a Figura 43, de surdos.

Figura 42 – Exemplos de tons.

Fonte: acervo do autor.

Figura 43 – Exemplo de surdos.

Fonte: acervo do autor.

Um *kit* de bateria pode ser equipado com uma quantidade variável de tons e surdos. Geralmente, essas peças são usadas para realizar transições rítmicas na música, popularmente conhecidas como viradas. Os áudios a seguir contêm gravações do som de tons e surdos, permitindo observar as diferenças de timbre entre eles.

Som do tom

Som do surdo

3.5 Chimbal

Assim como o bumbo e a caixa são exemplos de tambores com funções específicas em um *kit* de bateria, o chimbal é um exemplo de prato com função e sonoridade singulares. Também conhecido como *hi-hat*, ele é composto por dois pratos de igual dimensão, montados um sobre o outro com as bordas alinhadas. O prato inferior permanece fixo em um suporte, enquanto o superior é conectado a um mecanismo acionado por pedal. Esse mecanismo permite unir ou separar os pratos, criando diferentes variações sonoras. A Figura 44 ilustra o mecanismo de funcionamento do chimbal.

Figura 44 – Exemplo de um chimbal e seu mecanismo.

Fonte: acervo do autor.

O chimbal pode ser tocado de várias maneiras. Uma delas é pressionando o pedal, que movimenta o prato superior em direção ao inferior, produzindo um som específico ao se encontrarem. Outra forma é atacar o prato superior com as baquetas. Isso pode ser feito com os pratos encostados, conhecido como chimbal fechado, ou com eles separados, na configuração

de chimbal aberto. A versatilidade do chimbal em diferentes configurações, permite criação de variações rítmicas e sonoras.

Juntamente com o bumbo e a caixa, o chimbal é frequentemente utilizado na marcação rítmica principal das músicas. Ele costuma ser empregado para marcar os tempos de um compasso e, em alguns casos, também as subdivisões do tempo. Os áudios a seguir exemplificam diferentes configurações sonoras que podem ser extraídas de um chimbal.

Som do chimbal aberto

Som do chimbal fechado

Som do chimbal acionado por pedal

3.6 O prato de condução

O prato de condução, também conhecido como *ride*, embora se pareça com outros pratos, desempenha uma função específica na execução da bateria. Normalmente, é o maior e mais pesado prato do *kit*, o que o torna ideal para marcar o ritmo de maneira contínua, assim como o chimbal. Por isso, o *ride* é frequentemente usado como uma alternativa ao chimbal, oferecendo variações sonoras que enriquecem a levada rítmica da música em diferentes momentos. A Figura 45 mostra um exemplo de um prato de condução.

A maioria dos pratos de uma bateria possui uma geometria específica, dividida em área de ataque, área de condução e cúpula (WERNER, 2015). A área de ataque é a parte externa, que gera sons mais brilhantes quando tocada com baquetas. A área de condução, localizada mais ao centro, produz um som mais definido e controlado, ideal para marcar o ritmo. Já a cúpula é o ressalto central do prato, que gera um som mais concentrado e metálico. Essas áreas proporcionam uma diversidade de timbres e variações sonoras. A Figura 46 apresenta o perfil mecânico de um prato de bateria.

Os áudios a seguir exemplificam as diferenças sonoras de uma condução rítmica usando diferentes regiões de um mesmo prato de condução, tocadas com uma baqueta de madeira.

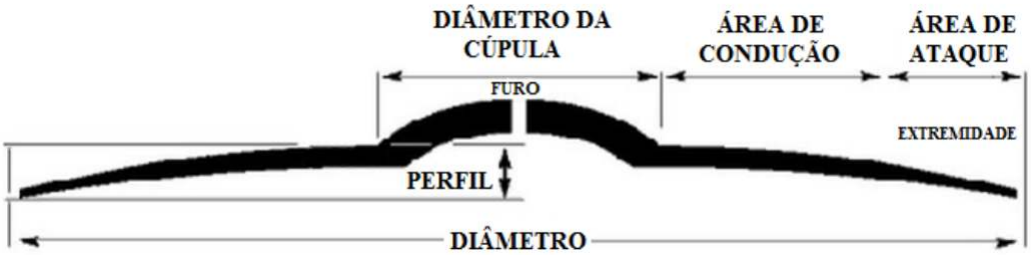
Prato de condução (extremidade)

Prato de condução (cúpula)

Figura 45 – Exemplo de um prato de condução.

Fonte: acervo do autor.

Figura 46 – Perfil mecânico de um prato de bateria.



Fonte: (WERNER, 2015).

3.7 Demais pratos

Além dos pratos mencionados até agora, um *kit* de bateria pode incluir uma grande variedade de outros pratos. Eles se diferenciam por características de construção, como tamanho, diâmetro, espessura, formato, processo de fabricação (BORATTO et al., 2021) e tipo de liga metálica (BORATTO et al., 2024). Cada uma dessas propriedades influencia diretamente a sonoridade de cada prato, resultando em timbres e respostas únicas que podem ser utilizados para enriquecer a dinâmica e a textura sonora em diferentes estilos musicais.

Os pratos de bateria são tradicionalmente organizados em categorias distintas, de acordo com suas características sonoras, físicas e a maneira como são utilizados em diferentes contextos musicais (BORATTO, 2022). Essa classificação é fundamental para entender as funções de cada prato dentro de um *kit* de bateria e como eles contribuem para a estética e dinâmica de uma música. Entre os exemplos mais conhecidos dessa categorização, destacam-se:

- Pratos de ataque (ou *crash*): recebem esse nome porque são usados em transições rítmicas que exigem uma resposta rápida e acentuada, caracterizada por um decaimento sonoro curto. Geralmente, são fabricados com diâmetros que variam entre 14 e 18 polegadas. Costumam ser tocados utilizando uma baqueta que atinge a extremidade do prato, produzindo um som explosivo e breve (WERNER, 2015) (BORATTO, 2022).

- *Splash*: geometricamente, os *splash* se assemelham aos *crash*, mas possuem diâmetros muito menores. Essa característica proporciona aos *splash* uma curta duração sonora, o que os torna ideais para momentos que demandam acentuações pontuais e rápidas na música (BORATTO, 2022).

- *China*: diferente dos *crash* e *splash*, os pratos *china* possuem uma geometria distinta, com bordas inclinadas para cima e uma curvatura mais acentuada. Essa forma confere ao *china* uma sonoridade única, com um timbre agressivo e explosivo, sendo ideal para destacar acentuações na música (BORATTO, 2022).

- *Stack*: o *stack* é uma configuração formada pelo empilhamento de dois ou mais pratos, permitindo uma ampla gama de possibilidades sonoras, dependendo de como os pratos são combinados. Embora essa montagem ofereça diversidade, o *stack* mantém uma característica sonora específica, com um ataque seco e curta duração. Devido a essa particularidade, é frequentemente utilizado para criar efeitos e texturas distintas em passagens musicais (BORATTO, 2022).

A Figura 47 apresenta os diferentes tipos de pratos mencionados anteriormente, oferecendo uma visão clara de suas características físicas e formatos distintos. Nos áudios a seguir, é possível ouvir as sonoridades de um *crash*, um *splash*, um *china* e um *stack*, permitindo perceber de forma prática as diferenças sonoras entre eles.

Figura 47 – Diferentes tipos de pratos de bateria. Da esquerda para a direita: *crash*, *splash*, *china* e *stack*.

Fonte: acervo do autor.

3.8 Outras possibilidades

Embora os elementos descritos até agora sejam amplamente utilizados na montagem de um *kit* de bateria, nada impede que outros instrumentos percussivos sejam incorporados. Isso cria uma combinação praticamente infinita de possibilidades, considerando a vasta gama de elementos percussivos disponíveis.

A escolha desses elementos adicionais depende do gosto pessoal do baterista, suas influências musicais e culturais, e do gênero musical que está sendo tocado. Alguns exemplos de outros instrumentos de percussão que podem ser integrados a um *kit* de bateria são: *cowbells*, platinelas, tamborins, chocalhos, sinos, *pads* eletrônicos, entre outros.

Além disso, nada impede que os elementos mencionados anteriormente sejam tocados por métodos não convencionais, também conhecidos como técnicas estendidas. Essas técnicas podem envolver uma ampla variedade de possibilidades, que vão desde o uso alternativo de acessórios convencionais, como a própria baqueta, até o emprego de acessórios não convencionais, como arcos de violino, as mãos, escovas de plástico, entre outros (COUTURIER; DAIGLE, 2022). Essas abordagens criativas permitem explorar novos timbres e texturas, ampliando as possibilidades sonoras do *kit*.

3.9 A bateria na produção musical

A bateria desempenha um papel crucial na produção musical moderna. Em formações de banda amplamente utilizadas, ela costuma ser o principal e, muitas vezes, o único instrumento percussivo. O ritmo definido pela bateria tem o poder de transformar completamente a intenção de uma música. Além disso, as viradas de bateria criam transições e momentos dinâmicos, enquanto sua levada frequentemente serve como um elemento central na identificação do gênero musical.

Dada a sua importância, a bateria assume um papel essencial nos processos de produção musical. Assim, ela se relaciona diretamente com as diversas etapas dessa produção, demandando a criação de métodos, técnicas e ferramentas específicas para explorar as possibilidades que o instrumento oferece. As subseções a seguir discutem como a bateria costuma ser abordada em duas etapas fundamentais da produção musical: gravação e mixagem

3.9.1 A bateria na gravação

Gravar uma bateria pode ser uma tarefa desafiadora, pois, conforme mencionado anteriormente, ela é composta por diversos elementos de percussão diferentes. Além disso, não existe um método único e correto para captar o som da bateria. A gravação pode ser realizada com o uso de um único microfone ou de dezenas de microfones, sem uma regra predeterminada. A decisão sobre o método ideal fica a critério do produtor musical e do engenheiro de áudio.

Ainda que não exista uma maneira única ou correta de captar o som da bateria, métodos que utilizam microfones direcionados a cada peça individualmente podem oferecer vantagens significativas, como maior controle sobre o volume de cada elemento, a possibilidade de aplicar processamento específico para cada peça e o posicionamento individual das peças dentro de um panorama estereofônico. Por essa razão, é comum que a muitas das produções musicais utilizem vários microfones na gravação da bateria. Nesse trabalho, será adotado um método de gravação amplamente utilizado em estúdios profissionais, que visa captar cada peça da bateria de forma individualizada (HUBER; RUNSTEIN, 2018g). Esse método inclui:

- 1 microfone direcionado à pele de resposta do bumbo;
- 2 microfones para captar a caixa: um direcionado à pele batedeira, e o outro direcionado à pele de resposta, onde se encontra a esteira[2];
- 1 microfone direcionado para o prato superior do chimbal, que se movimenta através do mecanismo acionado pelo pedal;
- 1 microfone apontado para a pele batedeira de cada tom e surdo do instrumento;

[2] A presença de dois microfones na caixa se justifica pelas diferenças de timbre entre o som gerado pelo toque da baqueta na pele batedeira e o som produzido pela esteira. Ambos são elementos importantes na sonoridade da peça e, sempre que possível, é comum utilizar dois microfones para captá-los individualmente.

- 2 microfones posicionados acima da bateria[3].

O sinal de cada microfone é registrado em uma faixa separada de um sistema multipista, permitindo que cada elemento da bateria seja processado individualmente durante a mixagem. A Figura 48 ilustra detalhadamente o método de microfonação mencionado.

Figura 48 – Método de microfonação de bateria.

Fonte: acervo do autor.

3.9.2 A bateria na mixagem

O processo de gravar individualmente o som das peças de uma bateria oferece diversas possibilidades criativas na mixagem. Com os sinais dos microfones registrados em diferentes pistas, é possível controlar o volume de cada peça separadamente. Isso permite ajustar o som do instrumento às necessidades estéticas de diferentes estilos musicais. Por exemplo, em uma estética sonora que exija maior destaque para o ritmo marcado pelo bumbo, o volume dessa peça pode ser aumentado em relação às outras.

Outra vantagem da captação multipista está relacionada ao panorama da música. O modelo de reprodução sonora predominante na atualidade é o estéreo, no qual os sons são reproduzidos por duas fontes sonoras, uma posicionada à esquerda e outra à direita. Essa configuração cria, de forma psicoacústica, a sensação de posicionamento dos elementos dentro da música. Com as peças registradas separadamente, é possível explorar esse recurso de maneira criativa. Ao aumentar o volume de um tambor na fonte sonora esquerda em relação à direita, pode-se deslocá-lo para a esquerda no panorama musical.

Além dessas vantagens, a gravação multipista permite o processamento individual de cada peça da bateria, o que é essencial para várias estéticas sonoras. Muitos estilos musicais, por exemplo, se caracterizam pela adição de efeitos, como *delays* e *reverbs*, nas caixas das baterias.

[3] Esses microfones são conhecidos como *overheads* e são responsáveis por capturar o som da bateria de modo geral, trazendo também a informação sonora dos pratos.

Outra técnica amplamente utilizada no ambiente digital é o *triggering*, que detecta o som de uma peça, como o bumbo, e adiciona automaticamente um som amostrado de outro bumbo ao sinal gravado (IZHAKI, 2017e).

3.10 O vazamento sonoro entre as peças da bateria

A bateria é um instrumento que gera um elevado nível de pressão sonora. Durante a gravação, esse volume intenso pode fazer com que o som de uma peça seja captado pelos microfones destinados a outras peças. Esse fenômeno, no qual sons de outras partes da bateria são registrados por um microfone específico, é conhecido como vazamento.

Embora os vazamentos nem sempre representem um problema eles podem, em algumas situações, gerar complicações na produção musical. Uma dessas questões está relacionada à correlação de fases entre os áudios captados. Por exemplo, o som da caixa pode ser registrado não apenas pelo microfone dedicado a ela, mas também por microfones próximos, como os do chimbal, dos tons e dos *overheads*. Como se trata de uma única fonte sonora captada por transdutores posicionados em diferentes locais, a soma desses sinais pode provocar alterações no espectro de frequências da caixa. Esse fenômeno é conhecido como *comb filter*.

Em relação aos exemplos mencionados na seção anterior, o vazamento também pode ser enxergado como um empecilho na aplicação de alguns processamentos. No exemplo da adição dos efeitos em uma caixa, se o sinal dela possuir vazamento de chimbal, ele também será processado, o que pode não ser o desejado. Além do mais, a presença de um vazamento alto pode atrapalhar na detecção das peças durante a aplicação de um *trigger*.

Por fim, os vazamentos também podem apresentar desafios durante a edição de tempo de uma performance. Como os sons de uma peça estão presentes nas faixas de outras, é necessário criar um agrupamento de todas as faixas da bateria. Dessa forma, qualquer alteração feita em um elemento será replicada em todos os outros, garantindo a integridade do conjunto. No entanto, isso impede a edição isolada de uma peça quando outra está sendo tocada simultaneamente. Por exemplo, se a caixa foi tocada fora do tempo, mas ao mesmo tempo o chimbal foi acionado, a faixa da caixa não poderá ser editada individualmente sem que o som do chimbal também seja alterado.

3.11 Métodos de separação sonora tradicionais

Ao longo do tempo, produtores e engenheiros de áudio têm buscado soluções para remover ou, ao menos, minimizar os impactos causados pelos vazamentos. Para isso, desenvolveram métodos e técnicas utilizando ferramentas tradicionais amplamente empregadas no campo do áudio. Nas subseções seguintes, serão explorados alguns desses métodos.

3.11.1 Equalização

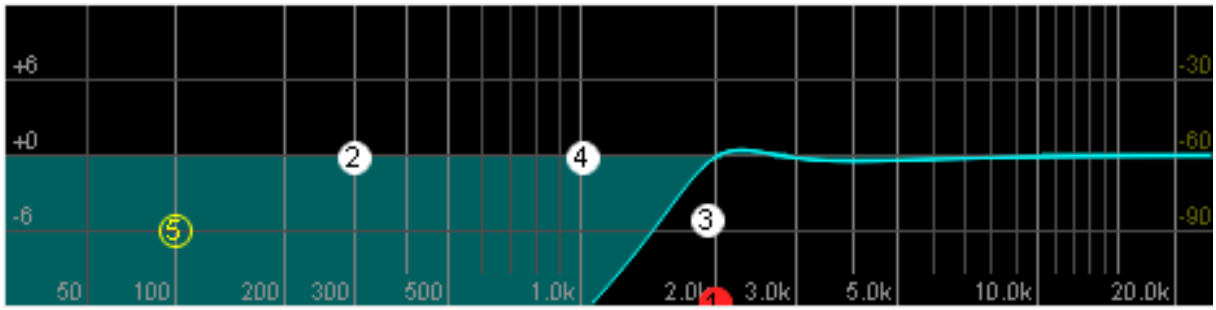
A equalização é um dos métodos frequentemente utilizados para minimizar os efeitos dos vazamentos. Esse método baseia-se no fato de que a informação sonora principal de cada peça da bateria geralmente está concentrada em regiões específicas do espectro de frequências. Assim, ao atenuar as frequências fundamentais de uma peça indesejada, é possível reduzir sua presença dentro de uma faixa, ajudando a mascarar o vazamento.

No exemplo abaixo, pode-se ouvir o som captado por um microfone direcionado a um prato de condução durante uma gravação. No áudio original, é possível perceber o vazamento de bumbo, caixa e tambores nesse microfone. Como o prato de condução possui uma sonoridade característica na região aguda do espectro de frequências, foi utilizado um filtro passa-altas para atenuar as frequências graves e mascarar os vazamentos. O resultado dessa técnica pode ser conferido nos áudios apresentados a seguir e o processamento realizado é mostrado na Figura 49.

Prato de condução antes da equalização

Prato de condução depois da equalização

Figura 49 – Equalização aplicada no microfone do prato de condução.



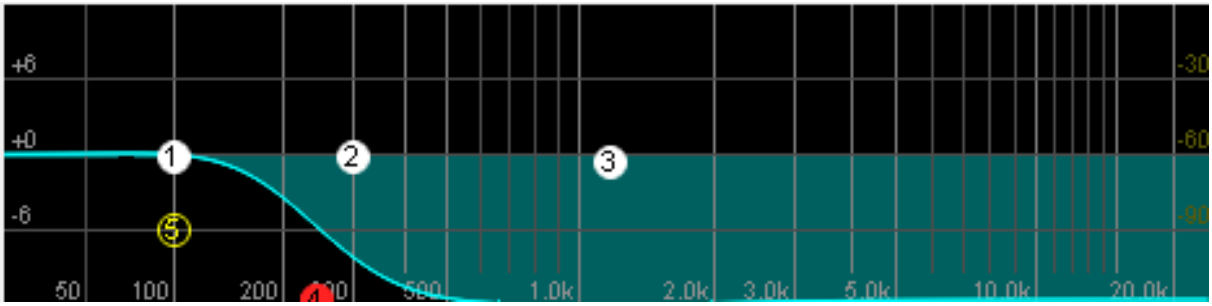
Fonte: acervo do autor.

Em um segundo exemplo, foi adotada a abordagem inversa. Utilizou-se a gravação do microfone direcionado ao bumbo da mesma performance. Assim como no caso anterior, é possível identificar vazamentos de caixa, tambores e pratos no microfone do bumbo. Nesse caso, considerando que o bumbo concentra suas informações sonoras mais relevantes na região grave do espectro, foi aplicado um filtro passa-baixas para atenuar as frequências mais altas e reduzir os efeitos dos vazamentos. O resultado dessa técnica pode ser ouvido nos áudios apresentados abaixo e o processamento realizado é mostrado na Figura 50.

Bumbo antes da equalização

Bumbo depois da equalização

Figura 50 – Equalização aplicada no microfone do bumbo.



Fonte: acervo do autor.

Embora seja uma técnica útil, a equalização não resolve completamente o problema dos vazamentos. No exemplo do prato, a equalização conseguiu tornar os vazamentos quase imperceptíveis. No entanto, isso ocorreu à custa de prejuízos significativos em seu timbre, já que a atenuação dos vazamentos resultou na perda de muitas das propriedades sonoras das frequências médias, que são importantes para a identidade sonora do instrumento.

No caso do bumbo, embora a equalização tenha conseguido atenuar os vazamentos, eles ainda permanecem perceptíveis após o processamento. Além disso, houve prejuízos consideráveis ao timbre do bumbo, que ficou limitado apenas às frequências graves, comprometendo sua sonoridade completa e característica.

A equalização pode resultar em algumas consequências indesejadas. Isso ocorre porque, ao atenuar uma faixa de frequência para minimizar um vazamento, tanto a informação da peça original quanto a da peça vazada são igualmente reduzidas. Esse efeito pode se tornar especialmente problemático em situações em que as peças operam em regiões de frequência próximas ou sobrepostas, dificultando a separação adequada dos sons.

3.11.2 *Gates* e *expanders*

Os *gates* e *expanders* são frequentemente utilizados em situações onde a amplitude do vazamento é significativamente menor em relação ao sinal da peça principal. Isso porque essas ferramentas operam com base na detecção do sinal por meio do parâmetro de *threshold*. Quando o vazamento possui uma amplitude elevada, a aplicação desses processadores pode gerar grandes prejuízos ao sinal principal, comprometendo a qualidade do som desejado.

Um exemplo clássico do uso de *gate* nesse contexto é na separação de vazamentos em microfones de bumbo. Por estar localizado próximo ao chão e mais distante das demais peças, o microfone do bumbo geralmente registra vazamentos com menor amplitude. No exemplo abaixo, foi aplicado um *gate* para remover os vazamentos em uma gravação de bumbo, destacando apenas o som principal.

Nesse caso, é possível perceber que a ferramenta foi muito mais eficaz na remoção do vazamento, garantindo que o som principal do bumbo se destacasse. Em termos de espectro de frequências, o sinal do bumbo não sofreu perdas significativas. No entanto, em termos de dinâmica, houve uma redução perceptível no *attack* e no *release* do bumbo devido à ação do *gate*. Isso resultou em um som mais "seco" e menos natural, alterando parcialmente a característica original do instrumento.

Outro aspecto a ser destacado é que, embora a ferramenta tenha se mostrado eficaz na redução do vazamento, ela não conseguiu eliminá-lo completamente. Isso fica evidente em algumas batidas, onde ainda é possível ouvir uma sobra do som de prato junto com o bumbo. Esse fato ressalta uma limitação desse método: a dificuldade de remover vazamentos que ocorrem simultaneamente à peça original. Ajustar o *threshold* para evitar capturar esse vazamento implicaria na exclusão de outras partes em que o bumbo foi tocado com menor intensidade, comprometendo ainda mais o resultado final.

3.11.3 Automações de volume

A maioria das DAWs disponíveis no mercado oferece um recurso conhecido como automação (IZHAKI, 2017a). Essa funcionalidade permite ajustar diferentes parâmetros da DAW para que eles alterem seus valores ao longo da performance, de maneira programada. Entre os diversos parâmetros que podem ser automatizados, destaca-se o controle do volume, que possibilita mudanças dinâmicas e precisas na mixagem.

A automação de volume pode ser uma ferramenta eficaz para reduzir os efeitos dos vazamentos em uma performance. Para isso, basta ajustar o volume nas partes em que o vazamento ocorre, reduzindo sua intensidade. Na maioria das DAWs, esse recurso é configurado por meio de um controle gráfico que acompanha a *waveform* do sinal, permitindo alterações precisas e visuais ao longo da faixa.

Para exemplificar o funcionamento dessa técnica, foi utilizada uma gravação de caixa, captada por um microfone posicionado na pele superior do instrumento. No áudio captado, é possível identificar vazamentos de diversas outras peças, como o bumbo, os tons e o prato de condução. Foi aplicada uma automação de volume com o objetivo de minimizar esses vazamentos. Os áudios abaixo apresentam o sinal original e o resultado após a aplicação do procedimento. A Figura 51 ilustra em detalhes a automação realizada, alinhada à *waveform* do sinal.

Caixa antes da automação

Caixa depois da automação

Figura 51 – Automação de volume aplicada no microfone da caixa.

Fonte: acervo do autor.

Embora seja semelhante ao método que utiliza *gates* e *expanders*, a automação pode oferecer um resultado mais preciso e menos suscetível a erros. Isso ocorre porque *gates* e *expanders* são ferramentas de atuação automática que, se não forem ajustadas corretamente, podem apresentar falhas. Um exemplo comum é em performances onde as peças são tocadas com sutileza, caso em que a amplitude do som principal pode se aproximar da dos vazamentos, resultando na eliminação do som desejado junto com o vazamento.

A automação de volume, por ser um processo realizado manualmente, consegue evitar esse tipo de problema. Isso ocorre porque o operador pode distinguir entre uma performance

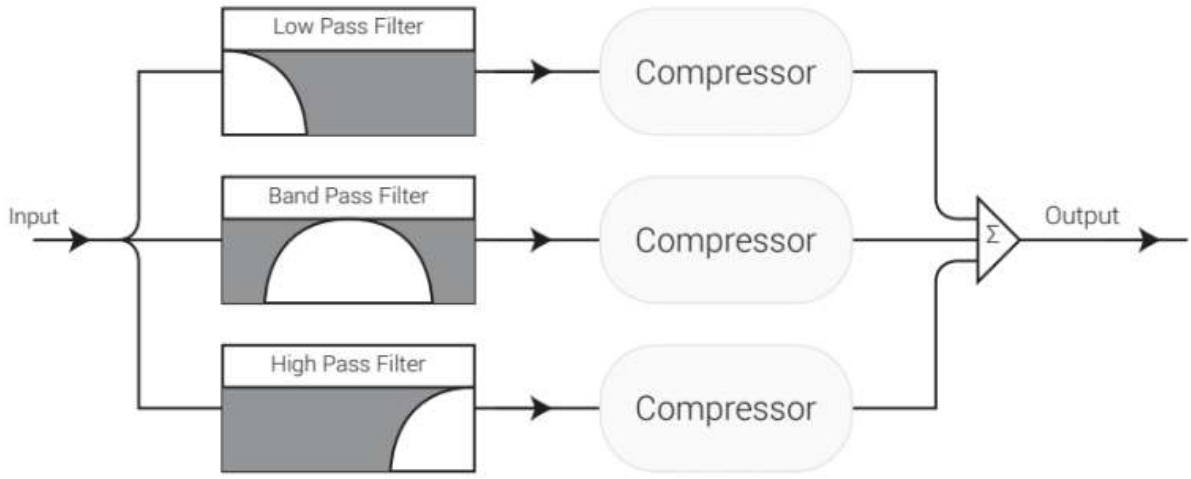
sutil da peça e um vazamento, ajustando o volume de forma mais precisa. Em contrapartida, a automação é significativamente mais trabalhosa do que o uso de processadores automáticos. Isso porque cada ajuste deve ser configurado manualmente na DAW, o que pode demandar muito tempo, especialmente em trabalhos mais longos.

Outro ponto de limitação do método de automação é que ele enfrenta as mesmas restrições do método anterior. Quando sinais são executados simultaneamente, torna-se impossível remover o vazamento sem causar prejuízos significativos ao sinal original. No exemplo apresentado, isso fica evidente no final da performance, durante a virada, quando alguns tons e batidas no prato de condução ocorrem ao mesmo tempo que a caixa, tornando inviável separá-los adequadamente.

3.11.4 Compressão multibanda

A compressão multibanda é um tipo especial de compressão. Nela, a redução de margem dinâmica é aplicada à uma determinada faixa de frequência do sinal, e não ao espectro por completo. Para isso, o compressor multibanda separa o sinal de em entrada em diferentes bandas, permitindo que cada uma seja comprimida individualmente (IZHAKI, 2017b). A Figura 52 ilustra o funcionamento desse dispositivo.

Figura 52 – Funcionamento de um compressor multibanda.



Fonte: adaptado de (IZHAKI, 2017b).

No contexto de separação de vazamentos, a compressão multibanda é utilizada de forma similar ao equalizador, pois seu princípio de aplicação também considera as regiões do espectro onde os instrumentos possuem informações mais relevantes. No exemplo abaixo, o áudio contendo a gravação do prato de condução, apresentado anteriormente na seção sobre equalização, foi processado novamente, desta vez com o uso de compressão multibanda. Os resultados podem ser ouvidos nos áudios a seguir.

Prato de condução antes da compressão multibanda

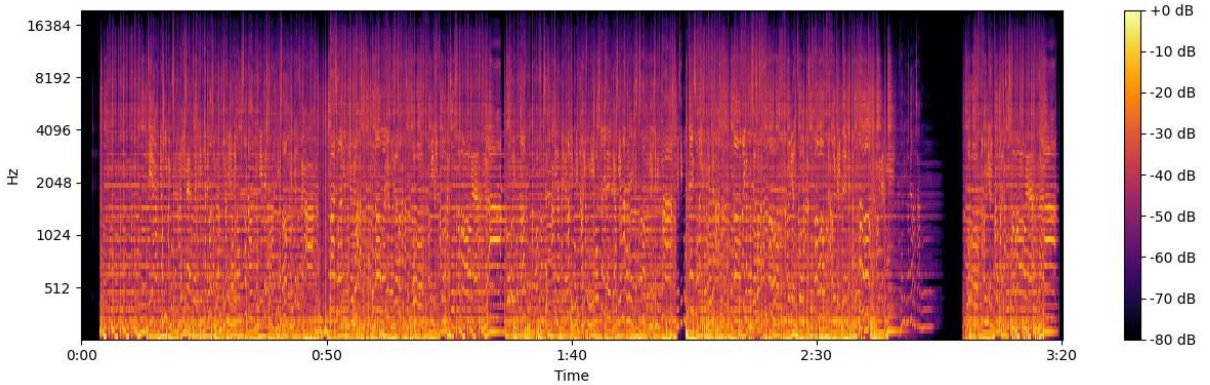
Prato de condução depois da compressão multibanda

Esse método difere do equalizador, pois não elimina completamente a informação da banda comprimida, o que traz tanto vantagens quanto desvantagens ao processo. A principal vantagem é que, ao preservar parte da banda, ele mantém informações importantes para o sinal principal. Por outro lado, isso impede que os vazamentos sejam completamente eliminados. Assim, essa ferramenta é mais adequada para situações em que se busca suavizar o vazamento de sinais, mas sem comprometer significativamente o conteúdo essencial do sinal principal.

3.11.5 Editor de áudio espectral

O último método abordado nesta seção é o uso de ferramentas conhecidas como editores de áudio espectrais. O funcionamento dessas ferramentas é baseado em uma representação essencial do sinal de áudio: o espectrograma. Essa representação combina três das propriedades mais importantes de um sinal (amplitude, tempo e frequência) em um único gráfico. Nas representações convencionais, esses atributos são dispostos da seguinte maneira: o tempo é representado no eixo das abscissas do plano cartesiano, a frequência ocupa o eixo das ordenadas, e a amplitude é indicada por meio de uma escala de cores na imagem. A Figura 53 ilustra um exemplo de espectrograma.

Figura 53 – Representação de uma música através de um espectrograma.



Fonte: acervo do autor.

Os editores de áudio espectrais aproveitam a representação do áudio como imagem para realizar operações de processamento e edição. A maioria dos *softwares* que oferecem edição espectral inclui diversas ferramentas, como pincéis e caixas de seleção, permitindo que a edição

do áudio seja feita diretamente pela manipulação da imagem gerada pelo espectrograma. Isso possibilita ajustes precisos e detalhados no áudio com base em sua visualização gráfica.

Para exemplificar essa técnica, foi utilizado um trecho curto de uma virada captado pelo microfone da caixa. Durante as batidas, é possível ouvir o som de um tom. Foi aplicada uma edição de áudio espectral para remover o som desse tom, e o resultado pode ser conferido nos áudios apresentados a seguir.

Caixa antes da edição espectral

Caixa depois da edição espectral

Essa técnica oferece uma vantagem significativa em relação às demais apresentadas, pois permite a separação de vazamentos que ocorrem simultaneamente com a peça principal, sem causar perdas perceptíveis no som desejado. No entanto, para obter bons resultados, é necessário que o operador possua certa habilidade com a ferramenta. Além disso, por ser um processo detalhado e minucioso, exige muito tempo para remover vazamentos de faixas de áudio longas. Por essa razão, essa abordagem é mais comumente empregada para eliminar vazamentos curtos e pontuais, como o som de metrônomos ou ruídos externos.

4 SEPARAÇÃO AUTOMÁTICA DE FONTES DE ÁUDIO

As ciências da computação abrangem diversas disciplinas aplicadas em contextos variados. Ferramentas e conceitos computacionais são cada vez mais empregados para solucionar problemas em diferentes áreas. No cenário da música, produção musical e engenharia de áudio contemporâneos, a computação surge como uma importante aliada, viabilizando uma ampla gama de tarefas e impulsionando transformações. Como resultado dessa interseção, várias linhas de pesquisa que integram música e tecnologia são desenvolvidas, trazendo inovações e novas contribuições.

Um exemplo inicial de como a computação pode ser aplicada no contexto musical é a linha de pesquisa conhecida como Novas Interfaces para Expressão Musical, ou New Interfaces for Musical Expression (NIME). Aqui, a computação atua como uma ponte entre diferentes interfaces de controle e módulos de síntese sonora, estabelecendo essa conexão por meio de estratégias de mapeamento. O objetivo desse tipo de pesquisa é possibilitar a criação de novos instrumentos musicais digitais, oferecendo suporte adicional para que artistas aprimorem suas performances (MEDEIROS et al., 2014).

Outro exemplo relevante dentro do assunto é a linha de pesquisa conhecida como Produção Musical Inteligente, ou Intelligent Music Production (IMP). Nela, a inteligência artificial e outras técnicas computacionais são aplicadas à produção musical. Por meio da IMP, é possível desenvolver ferramentas avançadas que colaboram diretamente com engenheiros de áudio em tarefas complexas, como a mixagem e a masterização. Essas ferramentas não apenas oferecem recomendações, mas também introduzem automações que otimizam o fluxo de trabalho desses profissionais. Com algoritmos baseados em dados, esses sistemas mapeiam as características do áudio e realizam ajustes baseados no conhecimento acumulado de especialistas, garantindo produções mais eficientes e com qualidade sonora aprimorada (MOFFAT; SANDLER, 2019).

A computação e a música também se encontram nos desafios relacionados à prática musical conhecida como música distribuída. Essa forma de performance musical envolve a utilização de redes de computadores para conectar músicos, permitindo a realização de apresentações colaborativas a distância, um tema intimamente ligado à computação. Nesse contexto, diversas ferramentas de distribuição de sinais e protocolos de rede têm sido desenvolvidos para apoiar essa prática. Essa área de pesquisa é particularmente relevante em um mundo globalizado, possibilitando a execução de performances mesmo entre músicos localizados em diferentes partes do mundo (SCHIAVONI; BIANCHI; QUEIROZ, 2014).

Uma quarta área de interesse em pesquisa que relaciona computação e áudio é a conhecida como Recuperação de Informação Musical, ou Music Information Retrieval (MIR). Esse campo de estudo é focado na extração de características musicais significativas a partir de diferentes fontes, como o sinal de áudio, representações simbólicas e fontes contextuais. Dentro de MIR, exploram-se temas como a classificação de músicas, recomendação, reconhecimento

de padrões, modelagem de usuários e desenvolvimento de interfaces que facilitam o acesso a grandes coleções de música. Esses estudos envolvem métodos para descrever, categorizar e organizar o conteúdo musical, além de compreender o contexto em que a música é consumida e apreciada (SCHEDL et al., 2014).

Muitas das ferramentas desenvolvidas no estudo de MIR são projetadas para analisar diferentes tipos de informações presentes em um sinal musical. Essas informações podem ser classificadas de acordo com sua natureza (DOWNIE, 2003):

- **Temporal:** corresponde à duração dos eventos musicais, incluindo indicadores de tempo, duração das notas, pausas e acentuações. Assim, a faceta temporal reflete elementos rítmicos em um arquivo musical (DOWNIE, 2003).
- **Harmônica:** abrange as características polifônicas de um áudio musical, surgindo quando duas ou mais notas soam simultaneamente, formando harmonias (DOWNIE, 2003).
- **Timbral:** diz respeito às qualidades que distinguem auditivamente uma mesma nota tocada em diferentes instrumentos. Assim, essa faceta está relacionada à "coloração" sonora específica de cada fonte sonora (DOWNIE, 2003).
- **Editorial:** representa as instruções de execução de uma peça musical, como o modo específico de tocar uma nota. Inclui elementos como dedilhados, ornamentos, dinâmicas e articulações. Como essas instruções influenciam diretamente o som produzido, a distinção entre as facetas editorial e timbral pode ser sutil (DOWNIE, 2003).
- **Textual:** refere-se basicamente à letra das músicas, desempenhando um papel importante na diferenciação de canções com o mesmo conteúdo melódico, mas letras distintas, como em traduções e paródias (DOWNIE, 2003).
- **Bibliográfica:** engloba os metadados de um arquivo de áudio, como título, compositores, arranjadores e data de publicação. É a única faceta que não deriva diretamente do conteúdo musical em si (DOWNIE, 2003).

A análise dessas informações por meio das tecnologias de MIR possibilita a execução de diversas tarefas, como a criação de sistemas de recomendação personalizados, a identificação de músicas a partir de pequenos trechos, a geração automática de *playlists* e a realização de análises aprofundadas para entender tendências culturais e estilísticas na música (SCHEDL et al., 2014).

Embora o termo *music* faça parte do nome, os estudos em MIR não se limitam exclusivamente ao campo musical. As técnicas desenvolvidas pela comunidade têm aplicação em diversas áreas, incluindo a saúde. A análise sonora através de MIR pode, por exemplo, ser utilizada na detecção de doenças, analisando sons como a fala (YILDIRIM et al., 2021) e a tosse (MOUAWAD;

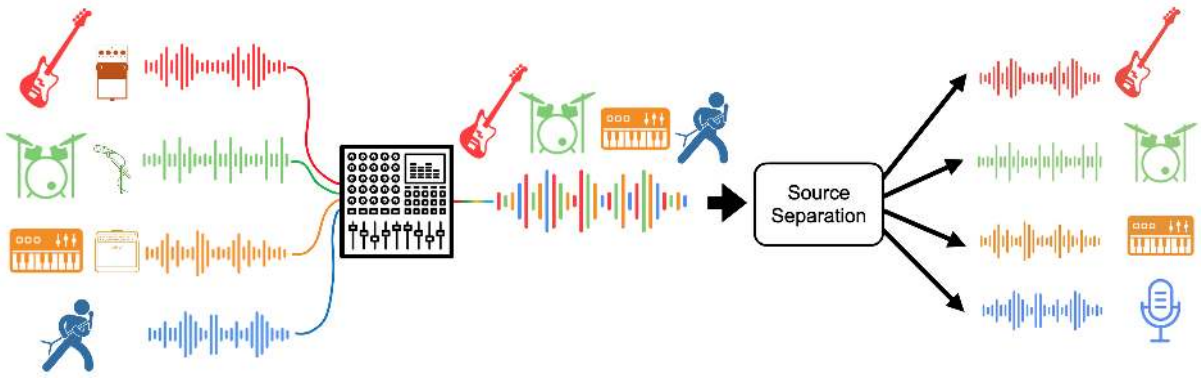
DUBNOV; DUBNOV, 2021) de um paciente. Essa gama de aplicações vai além, abrangendo áreas como a fabricação mecânica (BORATTO, 2022), a agricultura (ALI; DHANARAJ; NAYYAR, 2023) e muitos outros campos (BORATTO et al., 2022).

4.1 Audio Source Separation

Dentre os diversos temas abordados nas pesquisas em MIR, destaca-se um de importância central para este trabalho: a Separação de Fontes de Áudio, conhecida como Audio Source Separation (ASS). Essa área de estudo busca desenvolver métodos para isolar sons específicos em gravações contendo misturas complexas de elementos sonoros, como instrumentos, vozes e ruídos sobrepostos (MANILOW; SEETHARMAN; SALAMON, 2020). Quando aplicado à separação de fontes em sinais de áudio musicais, ela é denominada Separação de Fontes Musicais, ou Music Source Separation (MSS) (CANO et al., 2018).

Conforme discutido no Capítulo 2, durante o processo de mixagem, diferentes faixas de uma gravação multipista são combinadas em um único arquivo de áudio. Esse arquivo final é o que é comercializado e, consequentemente, disponibilizado ao grande público. Assim, o consumidor geralmente não tem acesso às faixas individuais de uma música. Nesse contexto, os algoritmos de ASS surgem como uma tentativa de realizar o processo inverso, buscando separar as diferentes faixas que foram integradas durante a mixagem. A Figura 54 ilustra esse processo.

Figura 54 – Relação entre mixagem e a separação de fontes sonoras.



Fonte: adaptado de (MANILOW; SEETHARMAN; SALAMON, 2020).

A separação de diferentes fontes sonoras é uma técnica versátil com uma ampla gama de aplicações. Esse processo pode ser utilizado em tarefas cotidianas, como remover vocais para criar uma versão de karaokê de uma música ou isolar um instrumento específico em uma performance para fins de estudo musical. Além disso, a separação de fontes se mostra útil em contextos não musicais envolvendo fala, como na separação das vozes de duas pessoas

em um arquivo de áudio ou na remoção de ruídos indesejados de uma gravação (MANILOW; SEETHARMAN; SALAMON, 2020).

Essa técnica também apoia o desenvolvimento de outras áreas de pesquisa (MANILOW; SEETHARMAN; SALAMON, 2020). Entre elas estão a transcrição musical automática (PLUMBLEY et al., 2002) (MANILOW; SEETHARAMAN; PARDO, 2020), a sincronização entre letra e música (FUJIHARA et al., 2006), a detecção de instrumentos musicais (HEITTOLA; KLAPURI; VIRTANEN, 2009), o reconhecimento automático de letras em canções (MESAROS; VIRTANEN, 2010), a identificação automática de vocais (WENINGER; WÖLLMER; SCHULLER, 2011) (HU; LIU, 2015) (SHARMA; DAS; LI, 2019), a detecção de atividade vocal (STOLLER; EWERT; DIXON, 2018), a identificação de frequências fundamentais (JANSSON et al., 2019) e a compreensão das previsões de modelos de áudio de caixa-preta (HAUNSCHMID; MANILOW; WIDMER, 2020b) (HAUNSCHMID; MANILOW; WIDMER, 2020a).

4.2 O uso de ferramentas de ASS como uma ferramenta para a remoção de vazamentos

O vazamento sonoro entre diferentes microfones em gravações de bateria não é uma questão recente. Além de ser comum e recorrente, é um dos desafios mais antigos enfrentados nos processos de produção musical. Como discutido no Capítulo 3, diversas abordagens e ferramentas foram desenvolvidas ao longo dos anos com o objetivo de solucionar ou minimizar esse fenômeno. Entretanto, os métodos tradicionais frequentemente apresentam limitações que comprometem a qualidade do resultado final. Além disso, muitos deles demandam tempo e habilidades específicas para sua execução.

Com o avanço dos algoritmos de ASS, essa tecnologia tem se mostrado uma alternativa promissora para superar as limitações das ferramentas convencionais. Diversos fatores evidenciam o potencial dos algoritmos de ASS como uma solução eficaz para a remoção de vazamentos em gravações de bateria. Um exemplo é o sucesso já alcançado na separação de outros elementos sonoros, como fala e performances vocais. Além disso, ferramentas baseadas em ASS já são utilizadas na prática cotidiana de engenheiros de áudio e produtores musicais para lidar com vazamentos em outros contextos, reforçando sua aplicabilidade nesse cenário.

Um exemplo prático da aplicação de ferramentas baseadas em ASS para a remoção de vazamentos ocorre em gravações realizadas ao vivo. Nesse contexto, os músicos geralmente compartilham o mesmo ambiente, como um palco ou uma sala de estúdio. Apesar da experiência do engenheiro de áudio responsável pela captação, frequentemente é inevitável a ocorrência de vazamentos entre os microfones dos diferentes instrumentos. Isso, especialmente em locais onde estão presentes fontes sonoras de alta intensidade, como baterias e guitarras amplificadas em equipamentos de alta potência. Esses vazamentos podem comprometer a qualidade do material gravado, dificultando processos subsequentes de edição e mixagem. Ferramentas de

ASS, nesse cenário, oferecem uma solução eficiente para mitigar essa questão, permitindo um isolamento mais preciso dos instrumentos.

Os áudios a seguir ilustram um exemplo real em que uma ferramenta de ASS foi utilizada para remover vazamentos em um contexto de gravação ao vivo. Nessa gravação, fones de ouvido foram utilizados como recurso de monitoração, eliminando a necessidade de caixas de som, o que evitou seu vazamento nos microfones. Instrumentos elétricos e digitais, como guitarra, baixo e teclado, foram gravados diretamente na interface de áudio, com monitoramento realizado pelas vias de fones.

No entanto, a bateria e os vocais representaram um desafio, pois não poderiam ser captados de forma silenciosa. Como o som da bateria é significativamente mais alto que o da voz, um vazamento expressivo da bateria no microfone vocal era inevitável. No primeiro áudio, é possível ouvir o som captado pelo microfone da voz de forma bruta, sem qualquer processamento. Já no segundo áudio, apresenta-se o resultado da aplicação de uma ferramenta[1] baseada em ASS, que removeu os vazamentos da bateria, permitindo um isolamento mais eficiente do vocal.

Antes do uso do algoritmo de ASS

Depois do uso do algoritmo de ASS

A análise auditiva dos áudios apresentados evidencia que a ferramenta de ASS utilizada conseguiu separar eficazmente o vazamento, removendo o som da bateria sem causar prejuízos significativos à performance vocal. Com isso, foi possível processar e aprimorar a gravação vocal utilizando *softwares* de afinação tonal e adicionando efeitos como *reverb* e *delay*. Sem a remoção do vazamento, tais processamentos seriam inviáveis, já que o som da bateria captado como vazamento também seria afetado, causando alterações indesejáveis ao som geral da bateria quando as faixas fossem somadas.

O uso de ferramentas de ASS nesse contexto foi viabilizado pelos avanços nas pesquisas dessa área. No entanto, apesar de os algoritmos modernos apresentarem resultados satisfatórios em termos de separação, eles ainda enfrentam limitações significativas quanto aos tipos de separação que podem realizar. Isso ocorre porque, historicamente, essas tecnologias foram desenvolvidas para identificar e separar sonoridades em classes pré-definidas, como vocais, bateria, baixo e uma categoria genérica denominada “outros”, que agrupa elementos que não se enquadram nas demais classes. Algumas ferramentas específicas conseguem ainda identificar instrumentos adicionais, como piano e violão (MEZZA et al., 2024b). Nos áudios abaixo, é possível ouvir a separação de diferentes instrumentos utilizando essa tecnologia. Na

[1] A ferramenta utilizada nestes exemplos foi a Moises AI. <https://moises.ai/>

performance apresentada, a música foi separada em vocais, baixo, bateria, violões e órgão, sendo este último agrupado na categoria “outros”.

4.3 Contextualização das ferramentas de ASS

A separação de fontes em arquivos musicais é um desafio que tem intrigado pesquisadores há décadas. A ausência de uma fórmula matemática simples para realizar essa separação de forma eficiente é uma das razões que tornam essa tarefa tão complexa e desafiadora. Além disso, fatores como a pluralidade sonora dos diversos instrumentos utilizados nas gravações e das diferentes técnicas de gravação, mixagem e masterização, fazem com que seja necessário explorar conhecimentos sobre a forma como a produção foi realizada para permitir boas separações (STÖTER et al., 2019).

Muitas das ferramentas desenvolvidas para superar os desafios da separação de fontes sonoras utilizam técnicas de processamento de sinal ou aprendizado de máquina, como redes neurais (LI et al., 2024). Em ambos os casos, é essencial a disponibilidade de dados para treinar e avaliar os métodos. No entanto, o acesso às faixas originais e isoladas das gravações musicais tem sido historicamente limitado. Isso ocorre porque a maioria das músicas comerciais está protegida por direitos autorais, dificultando a disponibilização dessas faixas pela indústria musical (STÖTER et al., 2019).

Com o avanço das tecnologias digitais, a produção musical migrou do ambiente analógico para o computador, ampliando as possibilidades na área. Essa transformação viabilizou

a criação de bases de dados específicas para a separação de fontes sonoras. Recursos como performances MIDI e instrumentos virtuais tem sido fundamentais para a construção de bases relevantes para esse propósito (MEZZA et al., 2024b).

Outro fator que contribuiu para o avanço na criação de bases de dados foi o uso de licenças como a Creative Commons por artistas e produtores musicais. Essas licenças possibilitam o compartilhamento gratuito das faixas de gravação, facilitando a criação de novas bases de dados. Como resultado, essas bases têm sido amplamente utilizadas no desenvolvimento de métodos de ASS baseados em aprendizado de máquina, o que tem contribuído significativamente para o aprimoramento das pesquisas na área (STÖTER et al., 2019)

O desenvolvimento da comunidade de pesquisa também teve um papel fundamental nos avanços alcançados nas tarefas de ASS, complementando a expansão das bases de dados. Iniciativas e eventos têm sido organizados para promover o compartilhamento de conhecimento e resultados entre pesquisadores da área. Um exemplo importante é a Signal Separation Evaluation Campaign (SiSEC), que se dedica a avaliar os progressos em ASS por meio de comparações sistemáticas entre algoritmos, utilizando tanto métricas objetivas quanto avaliações subjetivas baseadas na percepção de usuários (WARD et al., 2018).

A mentalidade dos pesquisadores em disponibilizar seus códigos e resultados de forma aberta também tem sido crucial para os avanços na área. Essa prática é essencial, pois promove a reprodutibilidade, permite que novas pesquisas aproveitem os avanços já alcançados e estimula melhorias nos resultados existentes. Além disso, muitos autores se preocupam em desenvolver suas ferramentas utilizando plataformas e *frameworks* acessíveis e amplamente adotados pela comunidade científica (STÖTER et al., 2019).

Algumas iniciativas, como o SigSep[2], desempenham um papel essencial na introdução de novos pesquisadores ao campo da ASS. Esse ambiente oferece acesso a bases de dados gratuitas, ferramentas de código aberto para separação de fontes sonoras e uma variedade de materiais didáticos e tutoriais voltados para o ensino de conceitos relacionados à área (SIGSEP, 2019). Além disso, eventos focados em MIR, como as conferências promovidas pela International Society for Music Information Retrieval (ISMIR), frequentemente disponibilizam treinamentos específicos sobre ASS, produzindo recursos valiosos para pesquisadores iniciantes e interessados no tema (MANILOW; SEETHARMAN; SALAMON, 2020).

4.4 Tecnologias que baseiam ferramentas de ASS

Conforme mencionado anteriormente, as ferramentas de ASS são, em sua maioria, fundamentadas em técnicas de processamento de sinais ou métodos de aprendizado de máquina (LI et al., 2024). Ambas as áreas são amplamente pesquisadas e aplicadas no campo da ciência

[2] <https://sigsep.github.io/>. Acesso em 29/11/2024.

da computação. Nas subseções a seguir, serão apresentados alguns dos principais métodos que embasam essas ferramentas.

4.4.1 Fatoração de Matrizes Não-Negativas

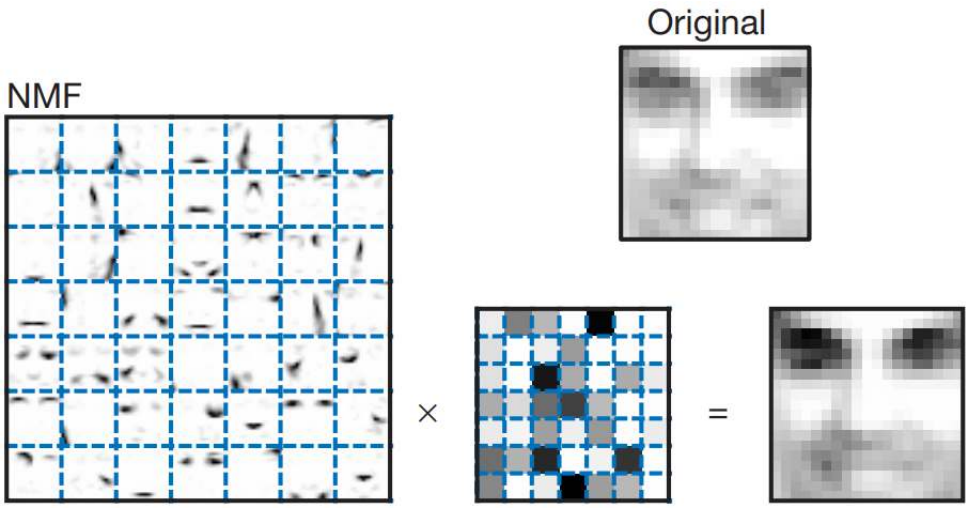
A Fatoração de Matrizes Não-Negativas, conhecida em inglês como Non-Negative Matrix Factorization (NMF), é uma técnica de processamento de sinais utilizada para decompor dados complexos, como imagens, textos ou sons. Essa decomposição transforma os dados em componentes mais simples, facilitando sua interpretabilidade. Assim, algoritmos baseados em NMF conseguem aprender por meio da análise das diferentes partes que compõem os dados estudados (LEE; SEUNG, 1999).

Em NMF, os dados analisados são representados por uma matriz V, que é aproximada por meio da fatoração em duas matrizes menores, W e H, de acordo com a relação

$$V \approx WH \tag{4.1}$$

onde todos os valores em W e H são não-negativos. Essa restrição permite a utilização de diversas imagens como base para representar o objeto estudado (LEE; SEUNG, 1999). A Figura 55 apresenta um exemplo de processamento de imagem utilizando o método de NMF.

Figura 55 – Aplicação do método de NMF em processamento de imagens.



Fonte: adaptado de (LEE; SEUNG, 1999).

Esse método apresentou sucesso significativo em diversas aplicações de áudio, especialmente em tarefas como separação de fontes e transcrição musical. Sua capacidade de decompor dados em componentes distintos e interpretáveis o torna uma ferramenta indispensável para o processamento de sinais complexos, como aqueles presentes em gravações vocais e musicais (LAROCHE et al., 2015).

No contexto de sinais de áudio, a matriz V é comumente utilizada para representar o espectrograma de magnitude dos dados de entrada. A matriz W, por sua vez, funciona como

um dicionário de padrões característicos, como espectros de instrumentos, enquanto H contém os coeficientes que descrevem as variações temporais desses padrões (LAROCHE et al., 2015).

A NMF é especialmente eficaz na separação de fontes harmônicas e percussivas devido às diferenças estruturais evidentes em seus espectrogramas. Isso permite ao método diferenciar essas fontes ao analisar suas partes constituintes. Esses fatores combinados explicam a ampla utilização da NMF em contextos de ASS (LAROCHE et al., 2015).

4.4.2 Redes Neurais

As redes neurais são ferramentas essenciais no campo da inteligência artificial. Inspiradas no funcionamento do cérebro humano, essas redes utilizam modelos matemáticos para simular o modo como sistemas biológicos realizam tarefas. Devido a essa inspiração, elas são amplamente aplicadas em problemas complexos, onde as relações entre variáveis não são claramente definidas (FLECK et al., 2016).

O principal diferencial das redes neurais está em sua capacidade de aprendizado por meio de exemplos e na generalização das informações adquiridas. Isso possibilita a construção de modelos não lineares organizados em estruturas compostas por camadas interligadas. Cada camada contém unidades de processamento, conhecidas como neurônios, que realizam cálculos matemáticos específicos (FLECK et al., 2016).

Essas redes são formadas por três tipos de camadas: entrada, intermediárias (ou ocultas) e saída. A camada de entrada recebe os dados que serão processados pelas camadas intermediárias, onde os neurônios aplicam pesos às informações recebidas. O resultado final desse processamento é entregue pela camada de saída (FLECK et al., 2016). Essa organização é ilustrada na Figura 56.

O valor calculado por um neurônio pode ser expresso matematicamente da seguinte forma

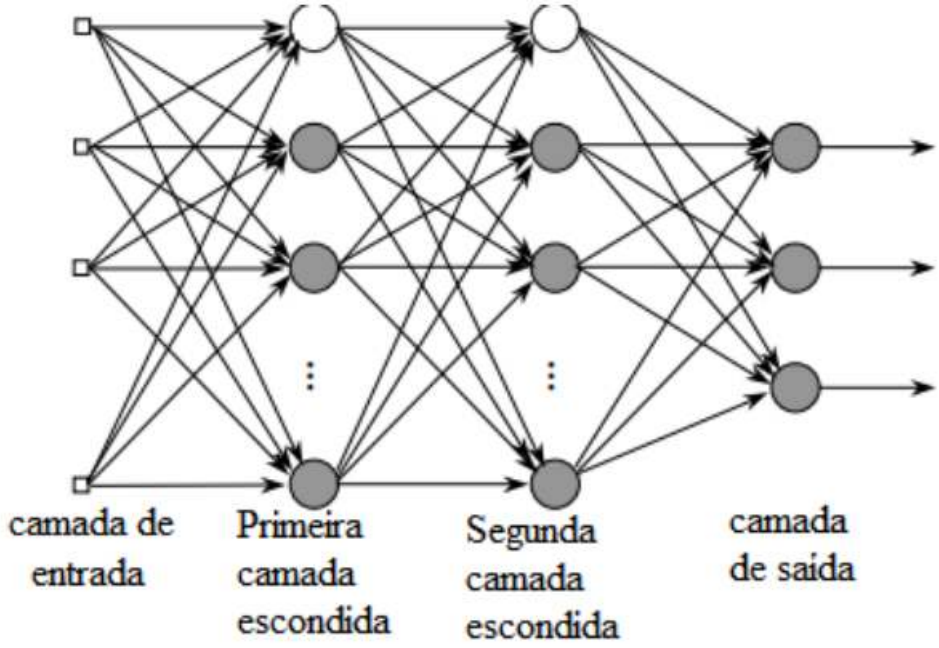
$$y_k = AF\left(\sum_{i=1}^{n}(y_i w_{ki}) + b_k\right) \tag{4.2}$$

onde y_k é a saída do neurônio, AF representa a função de ativação, y_i são os valores de entrada, w_{ki} são os pesos associados, e b_k é o *bias*. Nesse contexto, a função de ativação regula a amplitude da saída, enquanto o *bias* ajusta a entrada para a função de ativação (FLECK et al., 2016).

Os valores de saída da rede são avaliados de acordo com o tipo de aprendizado utilizado. No aprendizado supervisionado, esses valores são comparados com o esperado utilizando uma função de erro. Já no aprendizado não supervisionado, a rede avalia os resultados com base em métricas como correlação e redundância nos dados de entrada (FLECK et al., 2016).

Quando as redes neurais são aplicados em contextos de aprendizado profundo, elas recebem o nome de redes neurais profundas, em inglês *deep neural networks* (DNN). A principal diferença entre esses dois tipos de redes está no número de camadas: enquanto redes

Figura 56 – Exemplo de uma rede neural com duas camadas intermediárias.



Fonte: (FLECK et al., 2016).

tradicionais podem ter poucas camadas, as redes profundas possuem pelo menos quatro. Esse abordagem permite a extração de características em cada camada. Dessa forma, elas combinam características de baixo nível em representações de alto nível, possibilitando uma compreensão mais detalhada das características dos dados (YI et al., 2016).

Em comparação com as redes neurais comuns, as DNN se destacam por sua capacidade superior de capturar características relevantes e modelar relações complexas entre variáveis. Isso as torna ideais para resolver problemas com grande grau de complexidade e estrutura não-linear (YI et al., 2016). Entretanto, para alcançar esse nível de desempenho, as redes neurais profundas dependem de um volume substancial de dados durante o treinamento (RATNAPARKHI; PILLI; JOSHI, 2016), sendo esse um dos principais desafios em sua aplicação.

Uma abordagem alternativa baseada em redes neurais tradicionais, que possui grande relevância para este trabalho, é a rede neural convolucional, ou *convolutional neural network* (CNN). Assim como as redes tradicionais, as CNN também são inspiradas em estruturas biológicas. Contudo, seu funcionamento é baseado na região cortical dos mamíferos. Essas estruturas são compostas por pequenas áreas de células sensíveis a regiões específicas do campo visual. Por esse motivo, as CNN são amplamente aplicadas em tarefas que envolvem dados no formato de imagens (SAKIB et al., 2019).

As arquiteturas das CNNs são organizadas em várias camadas, cada uma desempenhando funções específicas. Essas camadas podem ser classificadas em quatro tipos principais: convolucionais, de *pooling*, totalmente conectadas e de perda (BAI; LI, 2023).

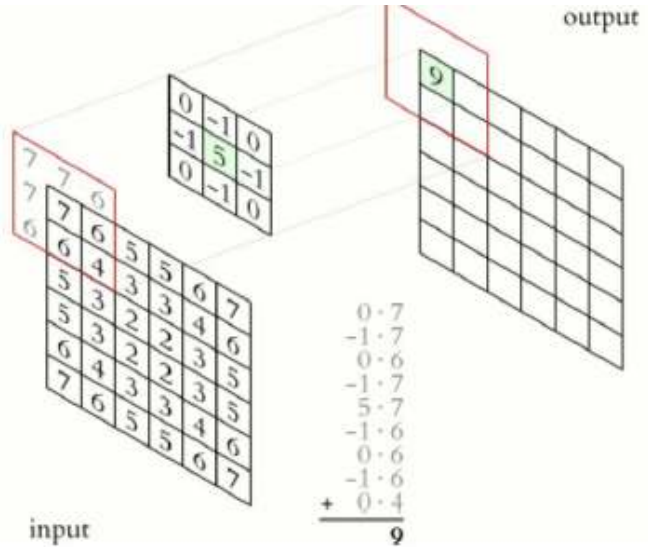
- **Camadas convolucionais**: essas camadas processam os dados de entrada por meio de filtros convolucionais, compostos por uma série de *kernels*. Esses filtros atuam sobre pequenas áreas dos dados, realizando cálculos matemáticos, como a multiplicação dos valores da entrada pelos valores do *kernel*. O resultado é um tensor de saída reduzido em altura e largura, mas que preserva as características mais relevantes, como padrões ou texturas, permitindo que a rede interprete os dados de maneira mais eficiente (BAI; LI, 2023). A Figura 57 ilustra o funcionamento desse tipo de filtro.

- **Camadas de *pooling***: têm a função de diminuir a quantidade de dados a serem processados, gerando uma versão de menor resolução da entrada enquanto preservam as informações essenciais. Esse processo, que não requer treinamento, organiza os dados gerados pelas camadas anteriores, facilitando seu processamento pelas camadas subsequentes e melhorando a eficiência da rede durante o treinamento e a execução (BAI; LI, 2023).

- **Camadas totalmente conectadas**: são camadas em que cada neurônio estabelece conexões com todos os neurônios das camadas anterior e posterior. Elas são responsáveis por realizar a classificação final do modelo e utilizam funções de ativação baseadas em transformações lineares aplicadas aos dados de entrada, incorporando um valor de *bias* para ajustar os resultados (BAI; LI, 2023).

- **Camadas de perda**: avaliam os resultados gerados pela rede, penalizando-a quando as saídas finais diferem dos valores esperados. Essas camadas utilizam funções de perda específicas para cada tipo de tarefa, contribuindo para a otimização de diferentes sistemas. Assim, sua presença ajuda a evitar a geração de resultados incorretos (BAI; LI, 2023).

Figura 57 – Representação da atuação de um filtro convolucional sobre uma imagem.

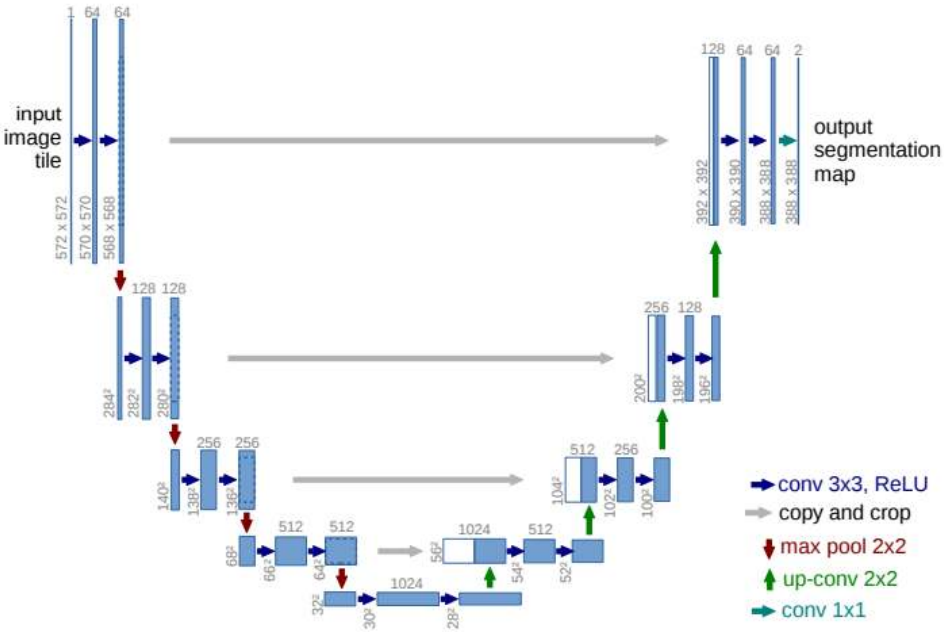


Fonte: (BAI; LI, 2023).

Entre as arquiteturas de CNN desenvolvidas, a U-Net merece destaque pela sua relevância neste trabalho. Essa rede foi concebida como uma solução robusta para problemas de segmentação de imagens biomédicas. Sua concepção foi impulsionada por uma necessidade específica: alcançar segmentações precisas em contextos onde a disponibilidade de dados anotados é limitada, uma dificuldade comum na área biomédica (RONNEBERGER; FISCHER; BROX, 2015).

A U-Net foi desenhada com o propósito de combinar a captura de contexto global e a recuperação precisa de detalhes locais. Essa abordagem é refletida em sua arquitetura em forma de "U", composta por um caminho contratante e um caminho expansivo. O caminho contratante reduz progressivamente a resolução espacial, extraindo características globais por meio de camadas de convolução e *pooling*. Já o caminho expansivo reverte esse processo, utilizando operações de *upsampling* e concatenação de mapas de características para reconstruir detalhes com precisão. Esse *design* simétrico possibilita que a U-Net segmente imagens complexas de maneira rápida e eficiente, mesmo em cenários com limitação na quantidade de dados (RONNEBERGER; FISCHER; BROX, 2015). A Figura 58 apresenta uma representação gráfica do modelo U-Net.

Figura 58 – Representação gráfica da arquitetura do U-Net.



Fonte: (RONNEBERGER; FISCHER; BROX, 2015).

Além disso, a implementação da U-Net inovou ao utilizar uma estratégia de aumento de dados baseada técnicas de deformações elásticas. Isso permitiu que o modelo adquirisse robustez frente a variações comuns em imagens biomédicas. Tal combinação de arquitetura eficiente e aumento de dados transformou a U-Net em uma ferramenta amplamente reconhecida e adaptada para diversas áreas (RONNEBERGER; FISCHER; BROX, 2015).

No campo da ASS, a U-Net encontrou uma nova aplicação. Ao ser implementada neste contexto, a rede demonstrou um desempenho notável ao lidar com representações espectrográficas de sinais de áudio. Dessa forma, ela foi adaptada para lidar com esses desafios, consolidando-se como uma peça-chave nas ferramentas modernas de separação de fontes. Atualmente esse modelo de rede neural é a base para o desenvolvimento de importantes ferramentas voltadas para o DSS, como LarsNet (MEZZA et al., 2024b).

4.5 Ferramentas importantes para a história da MSS

A história do desenvolvimento de ferramentas de separação de fontes sonoras remonta ao ano de 2011. Neste ano foi publicado o openBliSSART (WENINGER; LEHMANN; SCHULLER, 2011), considerado como o primeiro *software* de separação de fontes disponível publicamente (STÖTER et al., 2019). O openBliSSART[3] é uma ferramenta de código aberto desenvolvida para a separação cega de fontes sonoras com foco em flexibilidade e modularidade. Desenvolvida em C++ , utiliza o método de NMF e algumas de suas variantes. Sua arquitetura permite tanto separações supervisionadas quanto não supervisionadas, sendo útil para aplicação em separação de instrumentos musicais, melhoria de fala em ambientes ruidosos e outras tarefas de processamento de áudio (WENINGER; LEHMANN; SCHULLER, 2011).

A partir do openBliSSART, diversas outras ferramentas focadas em ASS foram desenvolvidas, incluindo a Flexible Audio Source Separation Toolbox (FASST)[4] (SALAÜN et al., 2014), a Unwist[5] (ROMA et al., 2016) e a Nussl[6] (MANILOW; SEETHARAMAN; PARDO, 2018). No entanto, algumas dessas ferramentas mais antigas enfrentam desafios relacionados à falta de atualizações regulares. A Unwist, por exemplo, não é atualizada desde 2017 (STÖTER et al., 2019), enquanto a Nussl apresentou problemas de atualização de domínio durante sua tentativa de instalação pelo autor, durante a época da escrita desse trabalho.

Embora essas ferramentas tenham desempenhado um papel histórico importante e contribuído significativamente para a construção do estado da arte em ASS, nenhuma delas conseguiu apresentar desempenhos notavelmente superiores em relação aos resultados publicados até 2015 (UHLICH; GIRON; MITSUFUJI, 2015) (STÖTER et al., 2019). No entanto, a partir de 2019, uma nova geração de ferramentas de ASS começou a emergir na literatura, introduzindo abordagens inovadoras e avanços significativos. A seguir, são apresentadas algumas dessas ferramentas, destacando suas contribuições e relevância no contexto atual da pesquisa.

[3] <https://openblissart.github.io/openBliSSART/>. Acesso em 29/11/2024.
[4] <https://gitlab.inria.fr/bass-db/fasst>. Acesso em 29/11/2024.
[5] <https://github.com/IoSR-Surrey/untwist>. Acesso em 29/11/2024.
[6] <https://github.com/nussl/nussl>. Acesso em 29/11/2024.

4.5.1 Open-Unmix

O Open-Unmix se destacou como uma ferramenta fundamental na evolução dos sistemas de ASS. Desenvolvido com base em DNN, o *software* foi idealizado para ser uma referência na área, figurando entre as primeiras ferramentas de código aberto a atingir resultados de alta qualidade. Além disso, suporta *frameworks* amplamente reconhecidos no aprendizado profundo e oferece um modelo pré-treinado que facilita o uso por artistas e pesquisadores interessados em explorar suas funcionalidades (STÖTER et al., 2019).

Uma característica distintiva do Open-Unmix é sua integração em um ecossistema aberto voltado para a separação de fontes musicais. Esse ecossistema abrange não apenas o *software* em si, mas também bases de dados como MUSDB18 (RAFII et al., 2017) e MUSDB18-HQ (RAFII et al., 2019), ferramentas de avaliação como o museval (STÖTER; LIUTKUS, 2019), além de um ambiente colaborativo gerido pela comunidade SigSep (SIGSEP, 2019), que incentiva pesquisas reprodutíveis e colaborativas. Dessa forma, o Open-Unmix desempenha um papel crucial no avanço das pesquisas em separação musical, promovendo a inovação dentro da comunidade científica (STÖTER et al., 2019).

Os desenvolvedores do Open-Unmix tinham dois objetivos principais em mente ao conceber a ferramenta: alcançar alto desempenho nos resultados e oferecer uma solução de fácil compreensão. De acordo com os idealizadores, etapas como pré e pós-processamento apresentam desafios técnicos significativos na separação de fontes sonoras. O Open-Unmix foi projetado para mitigar esses obstáculos por meio do compartilhamento de conhecimento técnico, além de facilitar a experimentação com novas representações e arquiteturas, impulsionando avanços futuros no campo (STÖTER et al., 2019).

Inicialmente, a implementação do Open-Unmix foi realizada em PyTorch, escolhido por sua simplicidade e modularidade. Posteriormente, o modelo foi adaptado para NNabla, com planos de expansão para TensorFlow. No entanto, os modelos pré-treinados não foram disponibilizados nessas versões alternativas, uma vez que seria inviável garantir resultados idênticos entre diferentes *frameworks*. Por isso, o modelo de referência em PyTorch foi mantido como padrão para comparações em pesquisas acadêmicas (STÖTER et al., 2019).

Com o objetivo de facilitar o aprendizado em processamento de áudio, os desenvolvedores do Open-Unmix adotaram estratégias que aprimoram a usabilidade da ferramenta. Componentes como pré e pós-processamento, carregamento de dados, treinamento e o modelo em si foram projetados de forma modular, permitindo fácil atualização ou substituição. Essa estrutura torna o *software* independente e altamente adaptável às diferentes necessidades dos usuários. O Open-Unmix adota uma abordagem acessível, permitindo que os usuários iniciem o treinamento rapidamente com uma base de dados baixada automaticamente, promovendo um aprendizado mais prático e eficiente (STÖTER et al., 2019).

O Open-Unmix encontra-se disponível gratuitamente em seu repositório no GitHub[7].

4.5.2 Demucs

O Demucs é um modelo de separação de fontes musicais baseado em aprendizado profundo. Ele foi projetado para lidar com quatro fontes principais: vocais, bateria, baixo e outros acompanhamentos. Ao contrário dos métodos baseados em espectrogramas, que criam máscaras sobre o espectrograma da mixagem e reutilizam a fase do sinal de entrada, o modelo opera diretamente no domínio da forma de onda. Essa abordagem elimina a necessidade de passos adicionais de síntese que podem introduzir artefatos e possibilita um treinamento de ponta a ponta, apresentando desempenho superior em comparação com métodos anteriores também baseados em formas de onda (DÉFOSSEZ et al., 2019).

A arquitetura do Demucs combina elementos da rede Wave-U-Net (uma variação da U-Net) com técnicas de síntese de áudio, integrando um codificador convolucional, uma camada LSTM bidirecional e um decodificador convolucional conectados por *skip connections*. Diferentemente de outros métodos, utiliza convoluções transpostas em vez de interpolação linear, permitindo um número maior de canais, o que contribui para seu desempenho superior. Além disso, emprega uma técnica equivalente ao uso de taxas de aprendizado específicas por camada, o que acelera a redução da perda de treinamento e melhora a convergência (DÉFOSSEZ et al., 2019).

Para lidar com a escassez de dados disponíveis, o modelo adota um esquema de aumento de dados semissupervisionado. Neste esquema, trechos de áudio em que pelo menos uma fonte está ausente são extraídos e remixados com exemplos supervisionados para criar novas combinações de treinamento. Essa abordagem utiliza 2000 músicas não rotuladas para reforçar o conjunto supervisionado, ajudando a reduzir a tendência ao *overfitting* observada em experimentos iniciais com os dados disponíveis (DÉFOSSEZ et al., 2019).

O modelo trouxe avanços significativos na separação de fontes musicais, reduzindo a diferença de desempenho entre métodos baseados em formas de onda e espectrogramas. Ele demonstrou resultados competitivos com os melhores algoritmos baseados em espectrogramas da época, alcançando um destaque especial na separação de linhas de baixo. Além disso, continuou a se beneficiar do uso de dados não rotulados, mesmo quando treinado com exemplos adicionais supervisionados (DÉFOSSEZ et al., 2019).

O código e os modelos pré-treinados estão disponíveis em seu repositório no GitHub[8] (DÉFOSSEZ et al., 2019).

[7] <https://github.com/sigsep/open-unmix-paper-joss>. Acessado em 29/11/2024.
[8] <https://github.com/facebookresearch/demucs>. Acessado em 29/11/2024.

4.5.3 Spleeter

O Spleeter é uma ferramenta amplamente utilizada no estudo de algoritmos de ASS. Desenvolvido com modelos pré-treinados, o Spleeter foi projetado para ser de fácil usabilidade, permitindo que arquivos de áudio sejam separados em múltiplos componentes com a execução de um único comando. Baseado na biblioteca TensorFlow, ele se destaca por sua alta eficiência em desempenho e velocidade, características que o tornam uma solução robusta para pesquisas e aplicações práticas na área. Além disso, o Spleeter pode processar grandes volumes de áudio com rapidez, sendo capaz de separar até 100 segundos de áudio estéreo em menos de 1 segundo, o que o torna ideal para tarefas que envolvem grandes conjuntos de dados (HENNEQUIN et al., 2020).

Os modelos pré-treinados disponibilizados pela ferramenta oferecem diferentes modalidades de separação de fontes sonoras (HENNEQUIN et al., 2020), sendo elas:

- Vocais e instrumental (2 faixas);
- Vocais, bateria, baixo e outros (4 faixas);
- Vocais, bateria, baixo, piano e outros (5 faixas).

Embora o Spleeter disponibilize modelos pré-treinados de alta qualidade, ele também oferece a flexibilidade de treinamento de novos modelos ou o ajuste fino dos modelos existentes por meio do TensorFlow. Para isso, é necessário que o usuário disponha de uma base de dados adequada contendo as fontes separadas. Essa flexibilidade torna o Spleeter uma ferramenta valiosa no estudo de ASS (HENNEQUIN et al., 2020).

O Spleeter foi desenvolvido com o objetivo de contribuir para a comunidade de MIR, oferecendo uma ferramenta eficiente para diversas tarefas. De acordo com seus autores, ele pode ser utilizado em atividades como sincronização e transcrição de letras a partir de áudio, transcrição de aspectos instrumentais como melodias, harmonias e ritmos, identificação de cantores, classificação com base em características musicais, como humor e gênero, extração de melodias vocais e identificação de covers. Seus resultados, na época de sua publicação, posicionaram-no como um dos melhores modelos de separação em 4 faixas disponíveis publicamente (HENNEQUIN et al., 2020).

De acordo com seus autores, o Spleeter foi o primeiro algoritmo na literatura a alcançar uma separação eficaz de piano como uma faixa distinta, consolidando-se como um marco na pesquisa de separação de fontes sonoras. Embora seja uma tecnologia de código aberto, a divulgação de sua base de dados foi limitada devido a questões de direitos autorais. Assim, a disponibilização de modelos pré-treinados tornou-se a principal forma de contribuição concreta da ferramenta para a comunidade acadêmica (HENNEQUIN et al., 2020).

A arquitetura do Spleeter é baseada nas redes neurais convolucionais do tipo U-Net, amplamente reconhecidas por sua eficiência em tarefas de ASS (JANSSON et al., 2017). Essas redes utilizam uma estrutura composta por codificadores e decodificadores conectados por meio de *skip connections*, permitindo a preservação de informações relevantes ao longo do processamento. No caso do Spleeter, cada U-Net é composta por 12 camadas, sendo 6 dedicadas à codificação e 6 à decodificação (HENNEQUIN et al., 2020).

No treinamento do Spleeter, foi utilizada uma norma L1 para calcular as diferenças entre os espectrogramas mascarados do áudio de entrada e os espectrogramas das fontes isoladas. O modelo foi treinado com um conjunto de dados interno da plataforma Deezer, conhecido como Bean *dataset*. Para ajustar os parâmetros da rede durante o treinamento, foi utilizado o algoritmo Adam (KINGMA; BA, 2017). O treinamento foi realizado ao longo de uma semana utilizando uma única GPU. O modelo apresentou desempenho competitivo em relação ao estado da arte da época (HENNEQUIN et al., 2020).

No Spleeter, a separação de áudio é realizada utilizando espectrogramas estimados a partir de máscaras específicas. Dois métodos principais são empregados para a criação dessas máscaras: as máscaras suaves e a filtragem Wiener multicanal (LIUTKUS; STöTER, 2019). Enquanto as máscaras suaves ajustam diretamente as intensidades do espectrograma original, a filtragem Wiener multicanal é utilizada para aprimorar a qualidade da separação, reduzindo ruídos e interferências (HENNEQUIN et al., 2020).

O Spleeter está disponível gratuitamente em seu repositório oficial no GitHub[9], onde podem ser encontrados tanto o código-fonte da ferramenta quanto os modelos pré-treinados. Além disso, ele é distribuído como um pacote Python independente e como uma receita para contêineres Docker autônomos, permitindo sua utilização em diversas plataformas (HENNEQUIN et al., 2020).

4.5.4 Meta-TasNet

O Meta-TasNet é uma ferramenta de ASS baseada em aprendizado profundo, inspirada no modelo Conv-TasNet (LUO; MESGARANI, 2019). O Conv-TasNet foi originalmente desenvolvido para separação de fala e utiliza redes neurais convolucionais para processar os dados no domínio do tempo. Dessa forma, o Meta-TasNet adapta essa arquitetura para a separação de fontes musicais (SAMUEL; GANESHAN; NARADOWSKY, 2020).

Adotando uma abordagem de meta-aprendizado, o modelo aprimora a separação de fontes musicais. Em vez de criar redes separadas para cada fonte, ele emprega uma rede geradora que calcula os parâmetros dos extratores. Com essa abordagem, é possível compartilhar os parâmetros entre os modelos de separação, tornando os extratores mais eficazes, com menos parâmetros e desempenho superior, em comparação com os modelos tradicionais que exigem

[9] <https://github.com/deezer/spleeter>. Acessado em 29/11/2024.

redes distintas para cada tipo de fonte (SAMUEL; GANESHAN; NARADOWSKY, 2020).

O Meta-TasNet adota uma arquitetura hierárquica, onde um gerador ajusta os parâmetros de um submodelo de mascaramento, conhecido como extrator. Esse ajuste é feito de forma específica para cada instrumento, levando em conta as relações entre diferentes instrumentos, o que contribui para uma separação mais eficaz. Com essa abordagem de meta-aprendizado, o modelo se torna mais compacto e rápido, mantendo a precisão na separação (SAMUEL; GANESHAN; NARADOWSKY, 2020).

O modelo foi avaliado usando o conjunto de dados MUSDB18 (RAFII et al., 2017), e os resultados demonstraram alto desempenho, especialmente nas tarefas de separação de vocais, bateria, baixo e outros instrumentos. Para aprimorar ainda mais a separação, a arquitetura foi modificada com a implementação de um modelo multi-estágio, no qual a resolução do áudio aumenta progressivamente a cada estágio. Além disso, ao reduzir a quantidade de parâmetros necessários para a separação, o Meta-TasNet mostrou-se eficiente, tornando-se ideal para ser implementado em dispositivos com recursos limitados, como dispositivos móveis e sistemas embarcados (SAMUEL; GANESHAN; NARADOWSKY, 2020).

O Meta-TasNet, bem como seus modelos pré-treinados, estão disponíveis publicamente no GitHub[10], permitindo que pesquisadores e desenvolvedores explorem sua eficácia em outras tarefas de separação de fontes musicais.

4.5.5 CrossNet-UMX

O CrossNet-UMX é uma arquitetura aprimorada para separação de fontes musicais baseada no Open-Unmix (STÖTER et al., 2019). O modelo propõe duas inovações principais para melhorar a performance da separação: a perda multi-domínio e a perda de combinação, além de uma modificação na arquitetura da rede. A perda multi-domínio permite que o modelo aproveite tanto a representação no domínio da frequência quanto no domínio do tempo, considerando as diferenças em ambos os domínios para melhorar a qualidade da separação. Para implementar a perda multi-domínio, camadas STFT ou *inverse* STFT (ISFT) são adicionadas à rede durante o treinamento, permitindo que o modelo aprenda as relações temporais e espectrais dos sinais (SAWATA et al., 2021).

A perda de combinação atua no modelo ao combinar máscaras de saída, o que permite calcular um número maior de funções de perda. Esse processo contribui para a redução do vazamento de instrumentos para outras fontes, pois considera a influência mútua entre elas, melhorando a precisão na separação (SAWATA et al., 2021).

Além das novas funções de perda, o CrossNet-UMX altera a arquitetura do Open-Unmix, possibilitando que as redes de extração de fontes distintas compartilhem informações. Esse compartilhamento permite que o modelo aprenda a colaborar entre as redes, aprimorando a

[10] <https://github.com/pfnet-research/meta-tasnet>. Acessado em 29/11/2024.

separação das fontes. Como resultado, a performance do modelo é significativamente melhorada ao considerar as interações entre as fontes (SAWATA et al., 2021).

Os experimentos realizados com o conjunto de dados MUSDB18 (RAFII et al., 2017) demonstraram que o CrossNet-UMX supera o Open-Unmix em termos de qualidade. O modelo apresentou melhorias significativas em diversas métricas, como a redução de artefatos e o aprimoramento na separação de fontes. Esses resultados validam as modificações propostas e comprovam a eficácia da arquitetura CrossNet-UMX em sistemas de separação de fontes musicais (SAWATA et al., 2021).

O CrossNet-UMX e seus modelos pré-treinados estão disponíveis publicamente no GitHub[11].

4.6 Avaliação de algoritmos de ASS

A avaliação dos resultados de separação de fontes é um grande desafio na área de ASS. Um dos principais fatores que torna essa tarefa tão complexa é a ampla variedade de aplicações em que os algoritmos de ASS podem ser utilizados, fazendo com que o desempenho esteja diretamente relacionado ao contexto específico da aplicação (VINCENT; GRIBONVAL; FEVOTTE, 2006). Por isso, metodologias tradicionalmente usadas para avaliar separações em ASS podem não ser adequadas para avaliar separações realizadas em peças de bateria, devido às diferenças nas abordagens e nos objetivos de cada técnica.

Outro fator que torna a avaliação de separações em ASS desafiadora é o fato de que, durante o processo de mixagem, as faixas originais gravadas passam por diversos processamentos, como compressão, equalização, reverberação e adição de efeitos, conforme discutido no Capítulo 2. Esses tratamentos transformam significativamente o som original, de modo que, por exemplo, o vocal presente no arquivo final da música difere em vários aspectos do vocal gravado originalmente, apresentando alterações em dinâmica, frequência e fase. Por isso, comparar uma faixa separada com sua gravação original pode ser uma tarefa complexa, mesmo quando a separação do elemento em uma música mixada é tecnicamente bem-sucedida, pois a faixa separada pode não corresponder exatamente à gravação original.

Como uma tentativa de superar esses desafios, algumas soluções foram propostas em eventos como o SiSec. Um exemplo é a criação e utilização de bases comuns para a avaliação de algoritmos, como o MUSDB18 (RAFII et al., 2017). Desde 2018, o MUSDB18 tem sido amplamente utilizado para avaliar e medir o desempenho de algoritmos de ASS tradicionais. A base é composta por 150 faixas completas, em estéreo, totalizando cerca de 10 horas de áudio. Os sinais estão divididos em quatro categorias: baixo, bateria, vocais e outros, abrangendo uma variedade de gêneros musicais, como jazz, eletrônico, metal, entre outros (WARD et al., 2018).

[11] <https://github.com/sony/ai-research-code/tree/master/x-umx>. Acessado em 29/11/2024.

Embora a padronização de bases de dados para a avaliação de ferramentas de ASS seja um passo importante, é igualmente essencial a implementação de métricas que possibilitem a quantificação dos resultados. Na literatura, algumas métricas relevantes incluem:

- ***Source-to-distortion ratio* (SDR)**: o SDR, como o próprio nome sugere, é uma métrica que avalia o nível de semelhança entre o sinal separado e o arquivo original. Em outras palavras, ele mede a distorção introduzida durante o processo de separação de fontes. Essa métrica é calculada pela equação

$$SDR := 10\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}||^2} \tag{4.3}$$

  onde s_{target} representa a parte da fonte verdadeira corretamente identificada na separação, e_{interf} é o erro devido à interferência de outras fontes, e_{noise} corresponde aos ruídos, como os captados por microfones ou adicionados no processo de separação e e_{artif} refere-se aos artefatos criados pelo algoritmo durante a separação (VINCENT; GRIBONVAL; FEVOTTE, 2006).

- ***Source-to-interferences ratio* (SIR)**: é uma métrica baseada no mesmo princípio do SDR. No entanto, ao invés de avaliar todos os erros gerados durante o processo de separação, ele se concentra exclusivamente nos erros causados pela interferência de outras fontes na estimativa. Essa métrica pode ser expressa pela equação

$$SIR := 10\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{interf}}||^2} \tag{4.4}$$

  onde s_{target} representa a parte da fonte verdadeira corretamente identificada na separação e e_{interf} é o erro devido à interferência de outras fontes (VINCENT; GRIBONVAL; FEVOTTE, 2006).

- ***Source-to-noise ratio* (SNR)**: é uma métrica que avalia os erros causados especificamente pelo ruído no processo de separação de fontes. Essa métrica é representada pela equação

$$SNR := 10\log_{10}\frac{||s_{\text{target}} + e_{\text{interf}}||^2}{||e_{\text{noise}}||^2} \tag{4.5}$$

  onde s_{target} representa a parte da fonte verdadeira corretamente identificada na separação, e_{interf} é o erro devido à interferência de outras fontes e e_{noise} corresponde aos ruídos (VINCENT; GRIBONVAL; FEVOTTE, 2006).

- ***Source-to-artifacts ratio* (SAR)**: é a métrica que avalia os erros causados pelos artefatos gerados pelo algoritmo durante o processo de separação de fontes. Essa métrica é representada pela equação

$$SAR := 10\log_{10}\frac{||s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}}||^2}{||e_{\text{artif}}||^2} \tag{4.6}$$

onde s_{target} representa a parte da fonte verdadeira corretamente identificada na separação, e_{interf} é o erro devido à interferência de outras fontes, e_{noise} corresponde aos ruídos, como os captados por microfones ou adicionados no processo de separação e e_{artif} refere-se aos artefatos criados pelo algoritmo durante a separação (VINCENT; GRIBONVAL; FEVOTTE, 2006).

- ***Image-to-spatial distortion ratio* (ISR)**: é a métrica que representa a preservação da imagem espacial da fonte verdadeira após a reconstrução do sinal (MALI; MAHAJAN, 2024). É descrita pela fórmula

$$ISR_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t s_{ij}^{\text{img}}(t)^2}{\sum_{i=1}^{I}\sum_t e_{ij}^{\text{spat}}(t)^2} \tag{4.7}$$

  onde $s_{ij}^{\text{img}}(t)$ representa a imagem espacial verdadeira e $e_{ij}^{\text{spat}}(t)^2$ representa a distorção espacial (VINCENT et al., 2007).

Por meio desse ecossistema, que integra bases de dados de referência e métricas avaliativas, a comunidade científica tem impulsionado o desenvolvimento de aplicações voltadas para a avaliação de ferramentas de ASS. Essas aplicações estão disponíveis na forma de *frameworks*, implementados em diferentes ambientes, como MATLAB e Python (WARD et al., 2018).

4.7 Drum Source Separation

Conforme apresentado na seção anterior, embora as ferramentas de ASS sejam capazes de realizar diferentes tipos de separação, seus modelos geralmente são treinados com base no exemplo clássico de divisão entre vocais, baixo, bateria e outros elementos musicais. Apesar de essa abordagem proporcionar resultados consistentes na extração da bateria, ela a trata como uma única faixa. Desse modo, todas as peças que compõem a bateria são interpretadas como um único elemento sonoro (MEZZA et al., 2024b).

Recentemente, uma nova linha de pesquisa denominada Drum Source Separation (DSS) tem emergido na literatura, visando a separação individual dos componentes de uma bateria. Essa abordagem representa um avanço significativo no processo inverso à mixagem, permitindo aplicações criativas e contribuindo para estudos musicais (MEZZA et al., 2024b). Além disso, o DSS auxilia em tarefas de pesquisa, como a transcrição automática de linhas de bateria (**??**).

O desenvolvimento da área de DSS enfrenta atualmente diversos desafios, sendo um dos principais a escassez de trabalhos dedicados ao tema. Em razão disso, a comunidade que se dedica ao avanço dessa tecnologia ainda está focada em resolver problemas fundamentais. Um dos esforços mais destacados é a criação de bases de dados amplas que contenham gravações de peças de bateria isoladas, essenciais para o treinamento de modelos. Essa necessidade é

especialmente relevante, dado que a maioria dos algoritmos utilizados atualmente para realizar a separação sonora se baseia em modelos de aprendizado profundo (MEZZA et al., 2024b).

Outro desafio abordado pela comunidade é a melhoria dos resultados utilizando ferramentas já existentes. Alguns autores têm investigado métodos para aprimorar ainda mais a acurácia das ferramentas empregadas na separação de peças de bateria. Um exemplo é o uso de descritores estatístico sobre frequência e tempo, como o centroide espectral, a dispersão espectral e o fluxo espectral. Estes, combinados com métodos já estabelecidos, demonstraram otimizações significativas nos resultados apresentados na literatura (LI et al., 2024).

Um terceiro viés de pesquisa consiste na comparação e avaliação de métodos já existentes. Por meio de processos como o *benchmarking*, diversos autores se dedicam a analisar o desempenho das principais ferramentas descritas na literatura, aplicadas em tarefas de DSS. Esse tipo de comparação é essencial para compreender aspectos específicos do funcionamento dos métodos, como a identificação da eficiência de abordagens que operam com dados em domínios híbridos (MEZZA et al., 2024a).

Para viabilizar pesquisas e o treinamento de modelos para a separação de peças de bateria, é fundamental ter acesso a bases de dados específicas de gravações de performances de bateria. Graças aos avanços em áreas como transcrição automática de bateria e ao crescente interesse de pesquisadores na área de DSS, diversas bases de dados já estão disponíveis gratuitamente. Entre elas, destacam-se:

- **ENST-Drums**: o ENST-Drums[12] é um banco de dados audiovisual anotado, desenvolvido para estudos de transcrição musical multimodal e análise automática de cena e gesto. Três bateristas gravaram oito faixas de áudio, com performances filmadas por duas câmeras. Parte do banco foi disponibilizada ao público, enquanto o restante foi retido pela instituição promotora. O material gravado totaliza 75 minutos, com variações na montagem das baterias e nos artefatos utilizados. A base abrange gêneros como bossa, *disco*, afro, *reggae*, *jazz*, *swing*, salsa, cha-cha, oriental, *rock*, *blues*, metal, *hard rock*, valsa, *funk* e *country*. As gravações foram realizadas com oito microfones, em taxa de amostragem de 44,1 kHz e resolução de 16 *bits* (GILLET; RICHARD, 2006).

- **RBMA13**: o RBMA13[13] é um conjunto de dados que contém 30 faixas do *sampler* 2013 Red Bull Music Academy Various Assets. Embora esteja associado a uma marca, os dados foram disponibilizados gratuitamente, o que justifica sua inclusão nesta lista. As faixas possuem duração média de 3 minutos e 50 segundos, totalizando aproximadamente 1 hora e 43 minutos de gravações. Além disso, elas abrangem uma variedade de gêneros musicais (VOGL et al., 2017).

[12] <https://perso.telecom-paristech.fr/grichard/ENST-drums/>. Acessado em 29/11/2024.
[13] <http://ifs.tuwien.ac.at/~vogl/datasets/>. Acessado em 29/11/2024.

- **IDMT-SMT-Drums**: o IDMT-SMT-Drums[14] (WEBER et al., 2023) é uma base de dados desenvolvida para estudos de transcrição automática de bateria e ASS. Composta por 608 arquivos no formato WAV, a base totaliza uma duração de 2167 horas, com gravações em taxa de amostragem de 44,1 kHz, resolução de 16 *bits* e formato monofônico. Inclui 104 gravações de bateria, registrando *loops* que abrangem bumbo, caixa e chimbal. Combina tanto performances reais quanto sintetizadas (DITTMAR; GÄRTNER, 2014).

- **MDB Drums**: o MDB Drums[15] é uma base de dados desenvolvida para pesquisas de transcrição automática de bateria, criada a partir do subconjunto MusicDelta da base de dados MedleyDB. Ele contém 23 faixas com duração média de 54 segundos, totalizando 20,7 minutos de gravações de bateria separadas por faixas. Além disso, a base inclui arquivos multipista, contabilizando 7994 *onsets*, divididos em 6 classes: bumbo (1539), caixa (2654), chimbal (2639), tons (90), pratos (1002) e outros elementos percussivos (70). Esses *onsets* destacam diferentes técnicas de execução e formatos das peças, abrangendo gêneros como *rock*, *country*, *disco*, *reggae* e *jazz*. Todas as faixas consistem em gravações reais, realizadas por músicos em baterias verdadeiras (SOUTHALL et al., 2017).

- **TMIDT**: o trabalho intitulado "Towards Multi-Instrument Drum Transcription" (TMIDT)[16] (VOGL; WIDMER; KNEES, 2018) criou e disponibilizou uma base de dados de peças de bateria com o objetivo de ampliar o volume de dados disponíveis na área e buscar uma distribuição equilibrada entre as classes de instrumentos. Para isso, foram utilizadas trilhas MIDI de músicas populares ocidentais. O banco de dados é composto por 4197 trilhas com uma duração média de 3 minutos e 41 segundos cada, totalizando aproximadamente 259 horas de gravações. As performances incluem uma ampla variedade de peças, como bumbo, caixa, palmas, diferentes tons, chimbal, tamborim, prato de condução, *cowbell* e diversos tipos de pratos. Variações de performance também foram incorporadas nesse contexto (VOGL; WIDMER; KNEES, 2018).

- **Groove MIDI Dataset (GMD)**: o GMD[17] é uma base de dados composta por 13,6 horas de performances de bateria, gravadas por 10 bateristas, sendo 5 profissionais e 5 amadores. As gravações foram realizadas em uma bateria eletrônica, resultando em 1.150 arquivos no formato MIDI. O banco de dados inclui informações sobre peças como bumbo, caixa, três tons, chimbal, um prato de ataque e um de condução. Apesar do formato MIDI, a bateria eletrônica utilizada permitiu o registro de diferentes técnicas de execução, como descrito no Capítulo 3. As performances variaram em termos de tempo e gênero musical (GILLICK et al., 2019).

[14] <https://www.idmt.fraunhofer.de/en/publications/datasets/drums.html>. Acessado em 29/11/2024.
[15] <https://github.com/CarlSouthall/MDBDrums>. Acessado em 29/11/2024.
[16] <http://ifs.tuwien.ac.at/~vogl/dafx2018/>. Acessado em 29/11/2024.
[17] <https://magenta.tensorflow.org/datasets/groove>. Acessado em 29/11/2024.

- **E-GMD**: o Expanded Groove MIDI Dataset (E-GMD)[18] é um conjunto de dados criado com o objetivo de expandir o GMD (GILLICK et al., 2019). Além dos arquivos já presentes no GMD, o E-GMD inclui gravações adicionais de 43 *kits* de bateria, abrangendo desde sons eletrônicos até acústicos. As gravações foram realizadas em uma bateria eletrônica, com os novos *kits* registrados a uma frequência de amostragem de 44,1 kHz e resolução de 24 *bits*. Ao todo, o E-GMD contém 1059 performances MIDI, resultando em 45537 arquivos e um total de 444,5 horas de gravação (CALLENDER; HAWTHORNE; ENGEL, 2020).

- **Crash Cymbal Sounds**: a Crash Cymbal Sounds[19] é uma base de dados desenvolvida para o estudo de pratos de bateria, com foco específico na classificação de diferentes pratos *crash* com base em suas composições químicas. Diferentemente das bases motivadas por pesquisas em transcrição automática ou separação de fontes de áudio (DSS), essa base foi criada para identificar as características sonoras de pratos feitos com ligas de bronze distintas. Ela contém registros de quatro pratos *crash*, totalizando 276 amostras sonoras, cada uma com 21 segundos de duração, resultando em 96,6 minutos de gravações (BORATTO, 2022).

- **StemGMD**: o StemGMD[20] é uma base de dados significativa para a área de separação de fontes de áudio (DSS), desenvolvida com base nas bases GMD (GILLICK et al., 2019) e E-GMD (CALLENDER; HAWTHORNE; ENGEL, 2020). Ela foi criada para abordar um problema recorrente na área: a falta de bases de dados amplas que disponibilizassem gravações de bateria com faixas isoladas. O StemGMD utilizou o material em MIDI fornecido pelo GMD e pelo E-GMD para gerar uma base adequada ao treinamento de DSS, reduzindo as diversas classes dessas bases em 9 novas categorias: bumbo, caixa, tom agudo, tom médio-grave, surdo, chimbal aberto, chimbal fechado, prato de ataque e prato de condução. Os arquivos de áudio foram produzidos utilizando 10 *kits* de bateria diferentes e registrados com taxa de amostragem de 44,1 kHz e resolução de 16 *bits*. O StemGMD contém 136 horas de performances de bateria, distribuídas em 103500 arquivos, totalizando 1224 horas de áudio (MEZZA et al., 2024b).

Apesar de muitas ferramentas tradicionais de separação de fontes de áudio, como as mencionadas na Seção 4.5, poderem ser treinadas para realizar separações generalistas e apresentarem bons resultados na separação de bateria, isso não garante que sejam as mais adequadas para o contexto de separação de peças individuais. Esse tipo de separação impõe desafios específicos que diferem dos exemplos clássicos de ASS, como a separação entre vocal, baixo, bateria e outros instrumentos. Esses desafios decorrem das características únicas das peças de bateria e da complexidade de suas interações no contexto musical.

[18] <https://magenta.tensorflow.org/datasets/e-gmd>. Acessado em 29/11/2024.
[19] <https://data.mendeley.com/datasets/9tytvdxd24/1>. Acessado em 29/11/2024.
[20] <https://zenodo.org/records/7860223>. Acessado em 29/11/2024.

Antes de entrar em detalhes sobre os algoritmos que têm sido utilizados nas pesquisas de DSS, é importante realizar algumas contextualizações sobre práticas que estão se consolidando nesse campo. Assim como nas pesquisas tradicionais de ASS, que utilizam classes como vocais, baixo, bateria e outros como base, as pesquisas em DSS também têm sugerido suas próprias categorias. As principais classes adotadas são: bumbo, caixa, tons (abrangendo tanto tons quanto surdos), chimbal e pratos (incluindo todos os pratos, exceto o chimbal) (MEZZA et al., 2024a).

Por ser um tema recente na literatura, houve poucas tentativas de uso, criação e adaptação de algoritmos voltados para essa tarefa. Como consequência, torna-se desafiador determinar um estado da arte no que diz respeito aos algoritmos mais eficientes para tarefas de DSS. Algumas ferramentas que foram utilizadas em *benchmarking* (MEZZA et al., 2024a), para avaliar a separação de peças de bateria incluem:

- **MDX23C** : o MDX23C (FABBRO et al., 202) é um modelo baseado em uma rede neural do tipo U-Net, na qual as camadas tradicionais de convolução são substituídas por blocos especializados que combinam operações no domínio do tempo e da frequência. Nesses blocos, as convoluções aplicadas ao sinal analisam as características nas dimensões de tempo e frequência, seguidas de camadas totalmente conectadas que ajustam os parâmetros do modelo para cada quadro temporal, garantindo um processamento mais preciso. Além disso, o modelo utiliza uma abordagem em que as partes real e imaginária do espectrograma complexo são tratadas como canais separados, o que permite ao sistema capturar informações mais ricas sobre o sinal de áudio. Outra característica importante é a técnica de subdivisão das frequências em bandas menores, que são reorganizadas e processadas separadamente, permitindo que o modelo aprenda padrões distintos para cada faixa de frequência. Por fim, o MDX23C combina funções de perda que avaliam a qualidade da separação tanto no domínio do tempo quanto em múltiplas resoluções do espectrograma, o que melhora significativamente seu desempenho na tarefa de separação de fontes musicais, proporcionando resultados robustos e precisos em diferentes contextos de áudio (MEZZA et al., 2024a).
- **HT-Demucs**: o Hybrid Transformer Demucs (ROUARD; MASSA; DéFOSSEZ, 2022), ou HT-Demucs, é uma das evoluções do Demucs, descrito anteriormente. Sua arquitetura é baseada em duas U-Nets paralelas: uma operando no domínio do tempo (através de convoluções temporais) e a outra no domínio do espectrograma (com convoluções sobre as frequências). A inovação do HT-Demucs está na substituição das camadas centrais de convolução por um codificador *transformer* entre domínios, que utiliza autoatenção para capturar relações locais dentro de um domínio e atenção cruzada para integrar informações entre os domínios. Essa abordagem permite que o modelo combine características temporais detalhadas com o contexto espectral, aumentando a precisão na separação de fontes. O HT-Demucs também implementa subdivisão de bandas em versões avançadas,

permitindo o processamento de sub-bandas específicas do sinal de áudio. Ao ser treinado com perdas no domínio do tempo e otimizado com dados adicionais, o HT-Demucs mostrou excelente desempenho em tarefas de demixagem de *kits* de bateria (MEZZA et al., 2024a).

- **BS-RoFormer**: O Band-Split RoFormer (LU et al., 2023), ou BS-RoFormer, é um modelo avançado que utiliza uma abordagem híbrida baseada em redes *transformers* e na divisão do espectrograma em bandas de frequência. Inicialmente, o espectrograma é segmentado em bandas de frequência não sobrepostas, que são processadas individualmente por camadas dedicadas de redes neurais. Em seguida, o modelo aplica duas formas distintas de atenção: a primeira realiza autoatenção ao longo do tempo, analisando as relações temporais no sinal, enquanto a segunda aplica atenção nas bandas de frequência, capturando as dependências espectrais. A arquitetura é aprimorada com o uso de posicionamento rotacional para consultas e chaves, uma técnica que ajusta a posição das informações com base em sua localização temporal ou espectral, aumentando a capacidade do modelo de identificar dependências de longo alcance no sinal. Por fim, o Band-Split RoFormer utiliza máscaras complexas preditivas para o espectrograma de entrada, permitindo reconstruir com precisão o sinal original no domínio do ISTFT (MEZZA et al., 2024a).

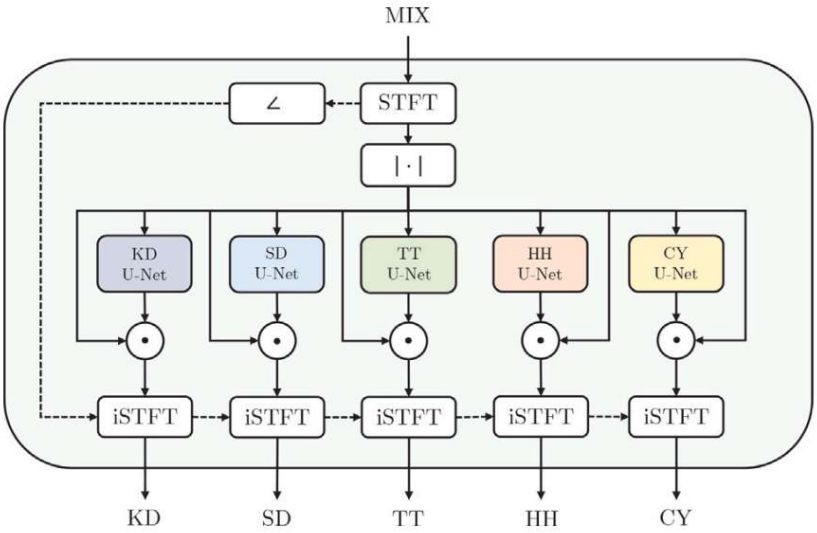
Apesar do bom desempenho das ferramentas mencionadas, a LarsNet (MEZZA et al., 2024b) foi a escolhida para a realização deste trabalho. Essa ferramenta foi desenvolvida especificamente para a separação de peças de bateria, categorizando os dados em cinco classes: bumbo, caixa, tons, chimbal e pratos. A classe tons agrupa todos os tambores que não sejam o bumbo ou a caixa, enquanto a classe pratos inclui os demais pratos que não correspondem ao chimbal. Além disso, a LarsNet disponibiliza modelos pré-treinados, o que facilita sua implementação (MEZZA et al., 2024b).

O desenvolvimento do LarsNet foi inspirado em ferramentas anteriores, como o Spleeter (HENNEQUIN et al., 2020), o Hybrid Demucs (uma versão atualizada do Demucs que estende sua funcionalidade para análise em múltiplos domínios) (DéFOSSEZ, 2022) e modelos baseados em U-Nets (JANSSON et al., 2017). Sua arquitetura é composta por cinco U-Nets paralelas, cada uma representando uma classe do modelo. Essas U-Nets operam com os sinais tanto no domínio do tempo quanto no domínio da frequência (MEZZA et al., 2024b). A Figura 59 ilustra o modelo de arquitetura do LarsNet.

Conforme ilustrado, nessa arquitetura cada U-Net utiliza uma porção da magnitude da STFT de uma mixagem em estéreo de bateria. Ela também gera uma máscara específica para cada um dos elementos a serem separados. Ao final do processo, o sinal no domínio do tempo é reconstruído aplicando a ISTFT ao produto entre a STFT complexa da mistura e a máscara gerada (MEZZA et al., 2024b).

No modelo, cada U-Net é composta por 13 camadas convolucionais que processam os

Figura 59 – Arquitetura do LarsNet.



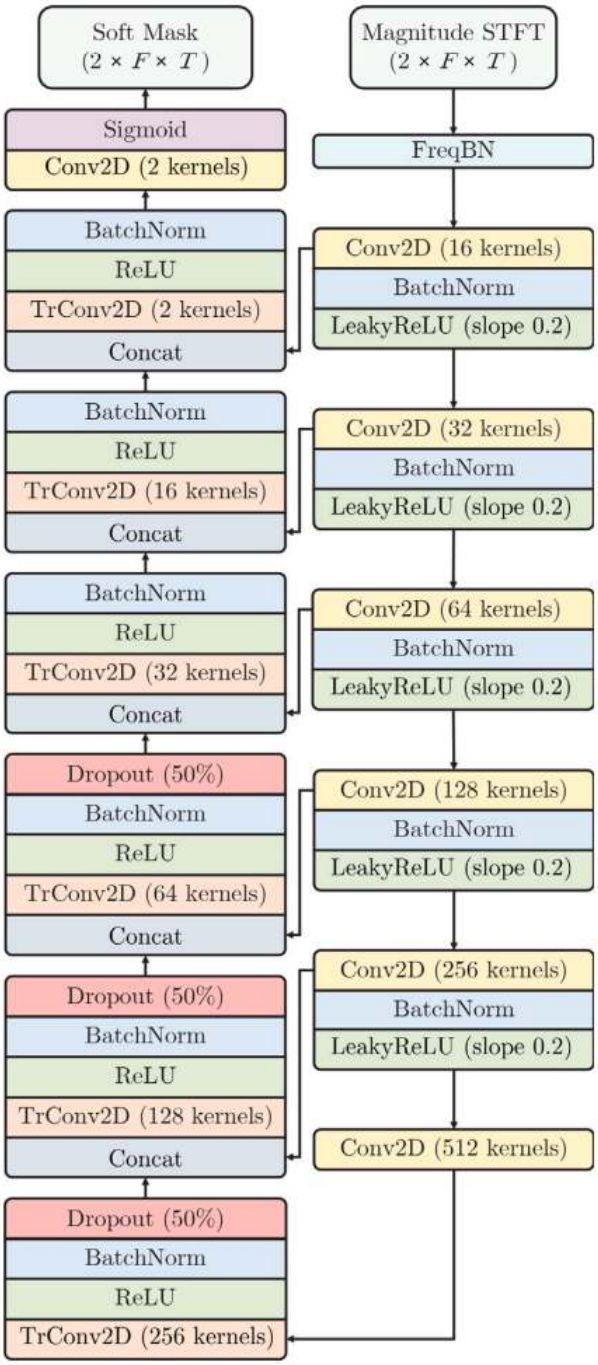
Fonte: (MEZZA et al., 2024b).

dados utilizando *kernels* de tamanho 5x5, com *stride* de 2x2 e *padding* de 2x2. A única exceção é a última camada do decodificador, que emprega *kernels* de 4x4, dilatação de 2x2 e *padding* de 3x3. Esse decodificador produz uma máscara suave usando uma função de ativação sigmoidal, que é ajustada com preenchimento de zeros para recuperar o tamanho original na dimensão temporal. Os blocos de tamanho T são então concatenados, e o sinal no domínio do tempo é reconstruído aplicando a ISTFT ao espectro ajustado (MEZZA et al., 2024b). A Figura 60 ilustra essa estrutura.

O desenvolvimento do LarsNet foi realizado em conjunto com a criação da base de dados StemGMD, previamente mencionada neste trabalho. Essa base foi utilizada no treinamento do modelo pré-treinado que acompanha a ferramenta. Durante esse processo, diversas técnicas de aumento de dados foram aplicadas para ampliar a variabilidade do conjunto e melhorar a capacidade de generalização do modelo (MEZZA et al., 2024b). As técnicas utilizadas incluem:

- **Aumento por troca de *kits***: consistiu em criar novas faixas ao combinar faixas de diferentes *kits* com o mesmo padrão, selecionados de forma aleatória.
- **Aumento por duplicação**: gerou novas faixas somando os sinais de um mesmo padrão provenientes de dois *kits* distintos.
- **Aumento por mudança de afinação**: introduziu novas variações ao ajustar a afinação de diversas peças em um valor aleatório dentro do intervalo de 3 semitons.
- **Aumento por saturação**: aplicou efeitos de compressão e saturação não-lineares em faixas existentes, criando novas variações a partir do processamento.

Figura 60 – Arquitetura de cada U-Net.

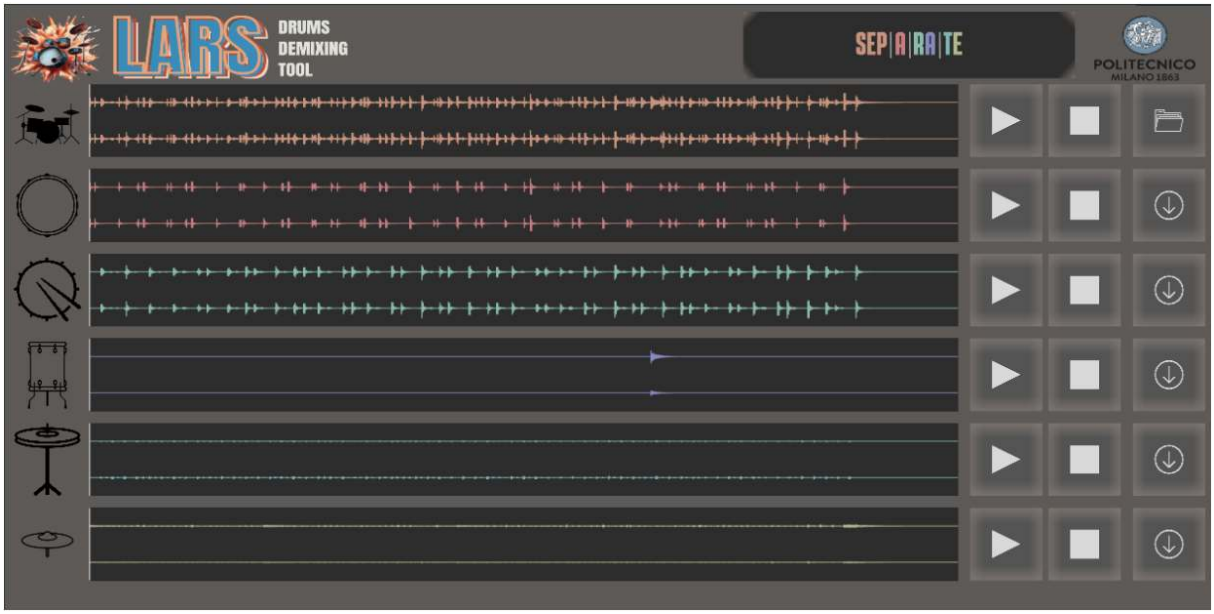


Fonte: (MEZZA et al., 2024b).

- **Aumento por troca de canais**: alterou a disposição dos canais direito e esquerdo de algumas faixas de áudio, gerando novas versões.
- **Aumento por remixagem**: criou novas faixas ao ajustar o ganho de diversas pistas, multiplicando-as por valores escalares dentro de um intervalo definido.

Uma característica importante dessa ferramenta, especialmente para este trabalho, é o fato de seus desenvolvedores terem trabalhado na adaptação de sua funcionalidade para o formato de *plugin*. Esse *plugin*, de código aberto, foi denominado LARS e pode ser utilizado diretamente em uma DAW. Além disso, ele foi projetado com uma interface gráfica de usuário simples e intuitiva (MEZZA et al., 2023), conforme mostrado na Figura 61.

Figura 61 – Interface gráfica de usuário do *plugin* LARS.



Fonte: (MEZZA et al., 2023).

Esse aspecto é fundamental para sua aplicação prática no contexto da produção musical, considerando que a maioria dos usuários que necessitam realizar separações de vazamentos são profissionais da área de áudio, e não de computação. Dessa forma, a adaptação do algoritmo para uso em DAWs possibilita que, mesmo sem conhecimento técnico em computação ou programação, esses usuários possam utilizar a ferramenta de forma eficiente.

5 EXPERIMENTO E RESULTADOS

Conforme discutido no Capítulo 1, este trabalho busca estabelecer uma conexão entre o desenvolvimento de ferramentas computacionais e os desafios encontrados no processo de produção musical. Para atingir esse objetivo, é crucial que os métodos computacionais sejam capazes de lidar com a realidade prática dos estúdios de gravação e estejam alinhados às necessidades dos engenheiros de áudio. Nesse contexto, o experimento proposto utilizou duas gravações reais de bateria, realizadas em cenários distintos, para avaliar se ferramentas de DSS, no estágio atual de desenvolvimento, são viáveis como solução prática para a questão dos vazamento das peças.

A ideia central do experimento é bastante direta: utilizar um algoritmo de DSS para eliminar vazamentos em gravações reais de bateria. Contudo, apesar de parecer uma tarefa simples, é necessário observar alguns critérios importantes. O principal deles é que, em um contexto de produção musical profissional, o áudio resultante da separação não pode conter artefatos perceptíveis. Por exemplo, ao separar o som da caixa de um vazamento de chimbal, qualquer degradação muito evidente na qualidade sonora pode tornar a separação inviável, sendo preferível, nesse caso, não realizar o processo ou optar por métodos tradicionais de redução de vazamentos.

O segundo critério importante é que muitos dos algoritmos projetados para tarefas de DSS foram treinados utilizando a base de dados StemGMD (MEZZA et al., 2024a), composta por sons gerados a partir de instrumentos virtuais. Por conta disso, um modelo treinado exclusivamente com essa base pode ter desempenho limitado em cenários que envolvem gravações de baterias reais. Esse desafio é ainda mais evidente quando as gravações são realizadas em condições adversas, como o uso de baterias de baixa qualidade ou em ambientes sem tratamento acústico apropriado.

O terceiro ponto relevante para este experimento está relacionado à capacidade do algoritmo em lidar com abordagens sonoras que não estão contempladas nas amostras da StemGMD. Essa limitação ocorre porque a base utiliza sons fornecidos pelos *kits* de bateria virtual empregados em sua criação. Um exemplo claro disso é a gravação da caixa: enquanto nos *samples* da base StemGMD o som da caixa e da esteira já aparecem combinados, em gravações reais esses sons são captados separadamente. Outro exemplo envolve os pratos, já que a base trabalha apenas com dois tipos, o prato de ataque e o de condução. Em contrapartida, situações reais frequentemente incluem outros tipos de pratos, como *chinas* e *splashes*, que não estão representados na base, podendo comprometer o desempenho do modelo nesses casos.

Por fim, uma questão interessante a ser considerada envolve a natureza sonora dos vazamentos. Embora o som de um vazamento carregue características da sonoridade da peça que o originou, ele representa uma informação distinta. Há diferenças significativas entre o som direto da peça e o seu vazamento, como intensidade, fase, espectro de frequência,

volume e dinâmica. Essas distinções levantam um ponto crucial: como um algoritmo de DSS se comportaria em situações onde algumas peças fossem captadas apenas como vazamentos, sem a presença de uma fonte direta clara? Essa situação pode testar os limites da capacidade do modelo em diferenciar e separar fontes em condições menos ideais.

A ferramenta selecionada para a execução do experimento neste trabalho foi o LarsNet (MEZZA et al., 2024b). Essa escolha foi fundamentada no foco das pesquisas que deram origem à ferramenta, alinhadas ao objetivo deste estudo. Embora existam outras ferramentas que apresentaram resultados superiores em testes de *benchmarking* (MEZZA et al., 2024a), o LarsNet foi a única, até o momento da realização desta pesquisa, a disponibilizar modelos pré-treinados. Essa característica foi crucial para viabilizar o desenvolvimento do trabalho dentro do prazo limitado disponível.

Outro aspecto determinante na escolha desse algoritmo foi sua adaptação para o formato de *plugin* (MEZZA et al., 2023). Considerando o caráter interdisciplinar deste trabalho, espera-se que o público-alvo inclua tanto pessoas com conhecimento em computação quanto profissionais das áreas de música e produção. Nesse contexto, a disponibilização do algoritmo como *plugin* facilita a reprodutibilidade do experimento, permitindo que qualquer usuário, mesmo sem conhecimentos avançados em computação, possa replicar a experimentação diretamente em uma DAW.

Para avaliar o desempenho da ferramenta, foram escolhidas duas gravações feitas em contextos distintos. A primeira ocorreu em um estúdio profissional, com equipamentos de alta qualidade e controle acústico adequado. A segunda, por sua vez, foi capturada em um ambiente ao vivo, sem qualquer tipo de tratamento acústico. Essas gravações representam diferentes cenários de captura, evidenciando os desafios relacionados aos vazamentos sonoros entre as peças da bateria. As próximas seções descrevem em detalhes o processo de coleta desses dados.

5.1 Gravação 1 - Bateria captada em estúdio de gravação profissional

A primeira gravação utilizada neste trabalho foi realizada por um baterista profissional. O responsável pela execução foi João Cordeiro, músico natural de Juiz de Fora (MG), com mais de 20 anos de carreira e ampla experiência em estúdios de gravação. Seu portfólio[1] inclui colaborações com renomados artistas brasileiros, como Emmerson Nogueira e Milton Nascimento.

A captação sonora foi realizada em um estúdio projetado para gravações de alta qualidade, com um ambiente acusticamente tratado e controlado. O *kit* de bateria utilizado foi o *Recording Custom*, fabricado pela Yamaha, conforme ilustrado na Figura 62. Embora a imagem

[1] <https://open.spotify.com/playlist/5Joan3jexHzavFQjJfr45t?si=afc9250cd88448fa>. Acessado em 29/11/2024.

mostre a bateria com cinco tons, apenas os três tons originais do *kit* foram gravados. Essa decisão teve como objetivo criar uma base de dados mais alinhada aos tipos de *kits* comumente utilizados em contextos de gravação, excluindo extensões ou peças adicionais. As especificações detalhadas da configuração da bateria estão descritas na Tabela 1.

Figura 62 – *Kit* de bateria Yamaha Recording Custom utilizada no experimento.

Fonte: imagem gentilmente cedida por João Cordeiro.

A gravação seguiu um modelo muito similar ao descrito na Subseção 3.9.1, diferenciando-se apenas pela adição de um microfone dedicado exclusivamente à captura do som do prato de condução. Os microfones utilizados eram de alta qualidade e amplamente reconhecidos no mercado por seu excelente desempenho em estúdios profissionais. A Tabela 2 apresenta a metodologia empregada na captação e descreve os microfones utilizados nesse processo.

No total, 10 microfones foram utilizados durante a gravação. A conversão A/D do áudio foi feita por meio de uma interface de áudio Focusrite, modelo Scarlett 18i20, que suporta 8 canais de entrada. Para expandir sua capacidade para 10 canais, foi conectado um expansor Presonus, modelo Digimax D8, via cabo óptico. A gravação foi realizada com uma frequência de amostragem de 44,1 kHz e resolução de 24 *bits*. A Tabela 3 detalha a configuração das conexões dos microfones aos canais desses dispositivos.

Tabela 1 – Configuração do primeiro *kit* utilizado no experimento.

Nome da peça	Dimensão	Marca	Modelo
Bumbo	22”	Yamaha	Recording Custom
Caixa	14” x 6”	Yamaha	Recording Custom
Tom	10”	Yamaha	Recording Custom
Tom	12”	Yamaha	Recording Custom
Surdo	14”	Yamaha	Recording Custom
Chimbal	14”	Bosphorus	Versa
Crash	16”	Zildjian	K Custom Dark
Crash	15”	Zildjian	A
Ride	22”	Bosphorus	Wide Ride
Splash	10”	Wuhan	10

Fonte: informações cedidas por João Cordeiro.

Tabela 2 – Método de microfonação utilizado no primeiro *kit* do experimento.

Marca do microfone	Modelo	Peça direcionada
Sennheiser	e602 II	Bumbo
Audix	i5	Caixa (pele superior)
Shure	SM57	Caixa (esteira)
Sennheiser	MD 421	Tom de 10”
Sennheiser	MD 421	Tom de 12”
Audix	D6	Surdo de 14”
Rode	M5	Prato superior do chimbal
Rode	M5	Prato de condução
Neumann	KM 184	Pratos (lado esquerdo)
Neumann	KM 184	Pratos (lado direito)

Fonte: informações cedidas por João Cordeiro.

Tabela 3 – Conexões dos microfones no sistema de gravação do primeiro experimento.

Microfone	Peça	Equipamento
Sennheiser e602 II	Bumbo	Focusrite Scarlett 18i20
Audix i5	Caixa (pele superior)	Focusrite Scarlett 18i20
Shure SM57	Caixa (esteira)	Presonus Digimax D8
Sennheiser MD 421	Tom de 10”	Presonus Digimax D8
Sennheiser MD 421	Tom de 12”	Presonus Digimax D8
Audix D6	Surdo de 14”	Presonus Digimax D8
Rode M5	Chimbal	Focusrite Scarlett 18i20
Rode M5	Prato de condução	Focusrite Scarlett 18i20
Neumann KM 184	Pratos (lado esquerdo)	Focusrite Scarlett 18i20
Neumann KM 184	Pratos (lado direito)	Focusrite Scarlett 18i20

Fonte: informações cedidas por João Cordeiro.

A performance selecionada para análise foi uma execução no estilo *jazz*, escolhida devido à sua capacidade de explorar de maneira abrangente as diferentes peças do *kit* de bateria.

Esse gênero musical é reconhecido por sua complexidade rítmica, riqueza em técnicas e ampla variação de dinâmicas sonoras, proporcionando um campo de análise robusto e diversificado. Essas características fazem com que a gravação seja especialmente adequada para testar a eficácia da ferramenta, desafiando sua capacidade de lidar com situações acústicas variadas e tecnicamente exigentes. O áudio abaixo apresenta a gravação utilizada no experimento.

Gravação 1

5.2 Gravação 2 - Bateria captada em gravação ao vivo

A segunda captação de bateria do trabalho foi realizada em um contexto de gravação ao vivo, durante uma sessão para o projeto Jelly Grooves[2]. Esse projeto musical tem como objetivo prestar homenagem a artistas consagrados da música internacional, por meio de releituras de seus maiores sucessos. Para a gravação, foi utilizado um *kit* de bateria Tama Starclassic Maple, ilustrado na Figura 63. Assim como na primeira captação, nem todas as peças do *kit* foram utilizadas. A Tabela 4 apresenta uma descrição detalhada das peças empregadas na sessão de gravação.

Tabela 4 – Configuração do segundo *kit* utilizado no experimento.

Nome da peça	Dimensão	Marca	Modelo
Bumbo	20”	Tama	Starclassic Maple
Caixa	14” x 7”	Taye	Studio Maple Java Burst
Tom	10”	Tama	Starclassic Maple
Surdo	14”	Tama	Starclassic Maple
Surdo	16”	Tama	Starclassic Maple
Chimbal	13”	Zildjian	I *hihat*
Crash	14”	Zildjian	I *crash*
Crash	16”	Sabian	B8X
Ride	21”	Zildjian	A *Sweet Ride*
Splash	10”	Sabian	XSR
China	18”	Sabian	XS20

Fonte: registros do autor.

Diferente da primeira captação, a segunda gravação foi realizada na sala da casa de um dos integrantes do projeto, um ambiente desprovido de qualquer tratamento acústico. O método de microfonação empregado foi semelhante ao utilizado no experimento anterior, permitindo uma futura comparação direta dos resultados peça por peça. A Tabela 5 apresenta uma descrição detalhada dos microfones utilizados nessa gravação.

Assim como no experimento anterior, foi utilizada uma interface de áudio Focusrite, modelo Scarlett 18i20, juntamente com um expansor de canais Phonic, modelo Firefly ADA 8,

[2] <https://www.youtube.com/@JellyGrooves>. Acessado em 29/11/2024.

Figura 63 – *Kit* de bateria Tama Starclassic utilizada no experimento.

Fonte: acervo do autor.

Tabela 5 – Método de microfonação utilizado no segundo *kit* do experimento.

Marca do microfone	Modelo	Tipo de transdução	Peça direcionada
AKG	D112 MKII II	Dinâmico	Bumbo
Shure	SM57	Dinâmico	Caixa (pele superior)
Shure	SM57	Dinâmico	Caixa (esteira)
Sennheiser	E 604	Dinâmico	Tom de 10”
Sennheiser	E 604	Dinâmico	Surdo de 14”
Sennheiser	e602 II	Dinâmico	Surdo de 16”
Sennheiser	E 614	Condensador	Prato superior do chimbal
Arcano	BD7	Condensador	Prato de condução
Shure	SM81	Condensador	Pratos (lado esquerdo)
Shure	SM81	Condensador	Pratos (lado direito)

Fonte: registros do autor.

conectado via cabo óptico. A gravação foi realizada com uma taxa de amostragem de 48 kHz e resolução de 24 *bits*. Contudo, diferente da primeira gravação, as conexões dos microfones no sistema de gravação não foram registradas nesta sessão. O resultado sonoro da gravação pode ser escutado no áudio abaixo.

Gravação 2

5.3 Metodologia de avaliação

Embora esforços tenham sido feitos para avaliar sistemas de ASS, as particularidades e desafios inerentes à DSS exigem a definição de parâmetros específicos. Bases de dados tradicionais, como o MUSDB18, não são adequadas para validar modelos de DSS, pois não abordam de forma detalhada as características exclusivas dos sons de bateria. Em resposta a essa limitação, estudos recentes no campo de DSS (MEZZA et al., 2024a) têm utilizado uma seção específica da base StemGMD, denominada StemGMD Eval Session, para medir a eficácia dos algoritmos. Essa seção inclui 10 *kits* de bateria selecionados a partir do conjunto principal (MEZZA et al., 2024b).

Outro estudo relevante no campo de DSS (LI et al., 2024) adotou uma abordagem distinta ao treinar e avaliar seus modelos, construindo uma base de dados própria denominada SYN5-Drums. Essa base foi criada utilizando amostras de uma biblioteca de som comercial, o que provavelmente justificou a decisão dos autores de não disponibilizá-la publicamente. Além da SYN5-Drums, os pesquisadores integraram o conjunto público IDMT-SMT-Drums para complementar seus experimentos. A validação dos algoritmos foi conduzida exclusivamente com dados da base elaborada especificamente para o trabalho.

Apesar das divergências nos dados utilizados para validação, os dois trabalhos compartilham pontos em comum significativos. Primeiramente, ambos adotaram a métrica SDR para avaliar os resultados, indicando que essa métrica é amplamente aceita e eficaz para medir o desempenho em tarefas de separação em DSS. Além disso, os dois estudos incluíram modelos baseados na arquitetura U-Net em suas análises, reforçando sua eficiência em tarefas de separação de fontes de bateria.

Mesmo com o desenvolvimento das metodologias e das métricas descritas, o experimento deste trabalho enfrenta um novo desafio. Isso porque, nesse caso, o trabalho foi desenvolvido pensando na aplicação das tecnologias de DSS em um contexto real de produção musical. Isso envolve uma série de questões que dificultam uma avaliação sistemática e quantitativa das tarefas propostas.

O primeiro ponto é que, neste trabalho, o objetivo não é avaliar o desempenho da ferramenta de forma generalista, mas sim sua eficácia em uma tarefa específica: a remoção de vazamentos. Assim, o foco da avaliação não está no modelo em si, mas nos resultados que ele produz ao processar os sinais. Nesse contexto, o desempenho geral da ferramenta enquanto modelo é irrelevante se ela não for eficiente na remoção de vazamentos, pois, nesse caso, ela não atenderia aos propósitos desta pesquisa.

Outro ponto importante é que o campo de atuação da pesquisa abrange gravações de baterias reais. Por mais vastas e úteis que bases de dados como o StemGMD sejam para pesquisas em DSS, elas ainda são bases sintéticas, construídas a partir de performances MIDI. Em uma gravação real, surgem novos elementos, que vão desde os vazamentos entre as peças

até a ressonância harmônica entre elas. Em bases de dados baseadas em *kits* amostrados, esses elementos frequentemente não são considerados. No entanto, eles estão presentes em contextos reais e precisam ser levados em conta para garantir a aplicabilidade prática dos resultados.

Acerca das métricas utilizadas, como o SDR, é importante notar que, para sua aplicação, é necessário comparar os dados resultantes da separação com as faixas originais isoladas do instrumento. No entanto, esse tipo de avaliação supervisionada enfrenta grandes desafios em um contexto de gravações reais. Isso ocorre porque é impossível para um baterista executar a mesma performance exatamente da mesma forma duas vezes. Mesmo que ele grave um ritmo de caixa isoladamente e, em seguida, reproduza o mesmo ritmo junto com as demais peças, haverá variações inevitáveis de tempo, timbre, intensidade, entre outros aspectos. Essas diferenças inviabilizam, em grande medida, o uso de métricas tradicionais nesse tipo de cenário.

As limitações do contexto em que a pesquisa se insere levam a metodologia de avaliação de resultados a adotar um enfoque qualitativo. Isso ocorre porque a inviabilidade de utilizar métricas quantitativas exige uma análise comparativa baseada na percepção auditiva. Embora esse tipo de avaliação não seja tão comum no campo da computação, em contextos de música e produção musical ele é fundamental para validar os resultados, já que a escuta desempenha um papel essencial e insubstituível nessas áreas.

Apesar dessa limitação na abordagem, é possível utilizar recursos interativos para melhorar a compreensão dos resultados obtidos. Um exemplo é a inclusão de materiais multimídia, como arquivos de áudio, prática já adotada em seções anteriores. Esses recursos facilitam a escuta, permitindo uma análise mais direta e intuitiva dos resultados, o que contribui significativamente para o entendimento das discussões apresentadas nas seções de resultados.

Outro recurso importante para auxiliar na avaliação dos resultados é o uso de espectrogramas. Por representarem diferentes aspectos do sinal de áudio, como amplitude, tempo e espectro de frequências de forma simultânea, eles oferecem um vislumbre gráfico que complementa a análise auditiva dos resultados. Essa abordagem visual facilita a identificação de padrões, diferenças e características específicas que poderiam passar despercebidas apenas pela escuta.

5.4 Avaliação de resultados

A análise dos resultados deste experimento foi estruturada em torno das cinco categorias de separação definidas pelo LarsNet: bumbo, caixa, tons, chimbal e pratos. Para isso, os 10 microfones empregados nas gravações foram agrupados conforme as peças do *kit* de bateria que pretendiam captar, sendo alocados às respectivas categorias. A Tabela 6 apresenta a classificação detalhada dos microfones para cada uma dessas categorias

Com base na classificação apresentada, cada seção deste capítulo analisará o desempenho

Tabela 6 – Classificação de cada microfone de acordo com as classes do LarsNet.

Peça alvo do microfone	Classe
Bumbo	Bumbo
Caixa (superior)	Caixa
Caixa (esteira)	Caixa
Tom/surdo menor	Tom
Tom/surdo médio	Tom
Tom/surdo maior	Tom
Chimbal	Chimbal
Prato de condução	Pratos
Overheads	Pratos

Fonte: registros do autor.

do algoritmo em relação aos microfones associados a cada classe. Em cada seção a seguir, serão discutidos tanto os resultados para os microfones utilizados no experimento realizado em estúdio quanto para os empregados na gravação ao vivo, permitindo uma avaliação comparativa entre os dois cenários.

5.5 Resultados para a classe bumbo

As diferenças entre os sinais de bumbo dos dois *kits* podem ser observadas nas Figuras 64 e 65, respectivamente. Em ambos os casos, a frequência fundamental desses instrumentos está abaixo de 500 Hz. Ao compará-los, nota-se que o bumbo do *kit* 1 possui uma característica sonora mais grave, enquanto o do *kit* 2 apresenta maior conteúdo sonoro em frequências agudas. O som de ambos pode ser ouvido nos áudios disponibilizados abaixo.

Microfone do bumbo - *Kit* 1 (estúdio)

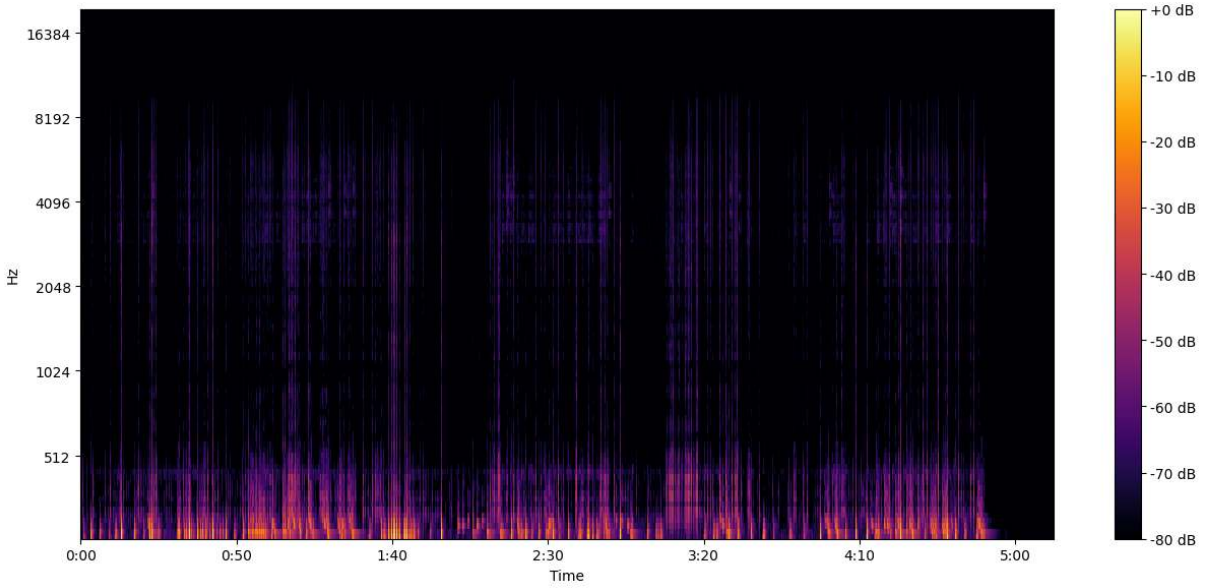
Microfone do bumbo - *Kit* 2 (ao vivo)

Os sinais foram processados pelo algoritmo LarsNet e classificados em cinco categorias distintas, conforme definido pelo algoritmo. Os resultantes da separação na classe bumbo podem ser ouvidos a seguir.

Microfone do bumbo - *Kit* 1 (estúdio) - SEPARADO

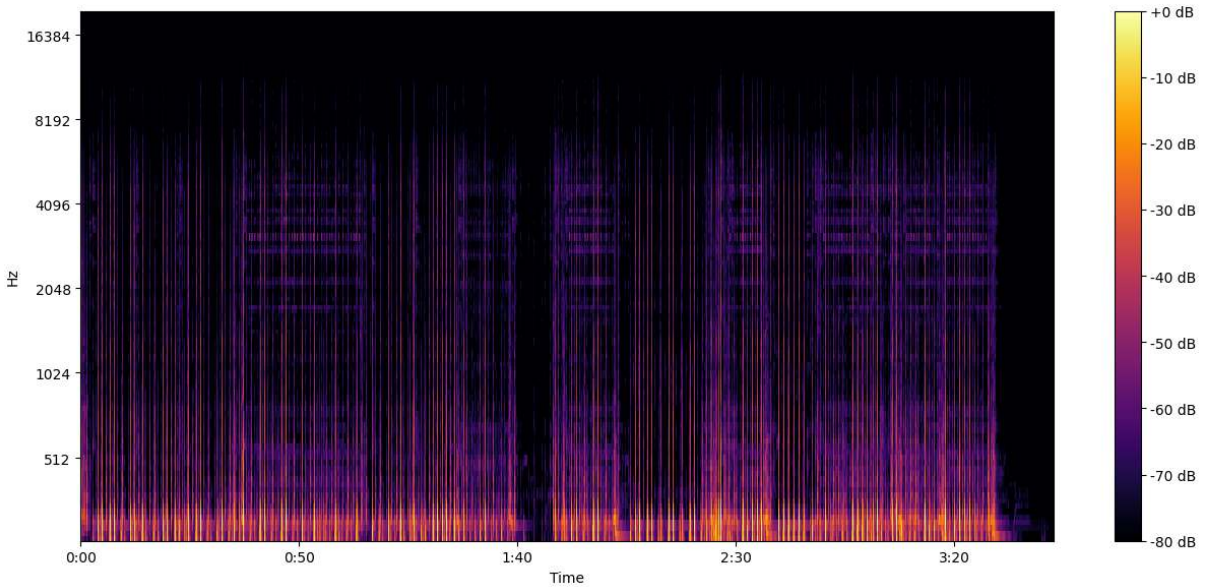
Microfone do bumbo - *Kit* 2 (ao vivo) - SEPARADO

Figura 64 – Microfone do bumbo do *kit* 1 (estúdio) antes da separação.



Fonte: acervo do autor.

Figura 65 – Microfone do bumbo do *kit* 2 (ao vivo) antes da separação.



Fonte: acervo do autor.

No *kit* 1, o algoritmo realizou a remoção de vazamento de forma satisfatória. Ao escutar o áudio, é possível perceber que restaram poucos resquícios de vazamento, predominantemente em frequências agudas. No entanto, esses resquícios apresentam uma amplitude muito baixa quando comparados ao sinal do bumbo. Isso evidencia que a separação, nesse contexto, é ideal para cenários em que se deseja realizar transcrições e *triggering* da bateria.

Apesar do bom desempenho na remoção de vazamentos no *kit* 1, ocorreram algumas alterações significativas no timbre da peça. De maneira geral, o bumbo perdeu características sonoras nas regiões mais agudas, resultando em um som mais "abafado". Além disso, foi possível observar que a intensidade com que as peças são tocadas influencia o timbre após a separação. Isso fica evidente no tempo 0:33, onde uma série de bumbos é tocada com intensidade progressivamente aumentada. Observa-se que, nos momentos em que os bumbos são tocados mais suavemente, o timbre após a separação se torna mais grave, enquanto os bumbos mais fortes mantêm frequências agudas. Outro exemplo que reforça esse fato ocorre a partir do tempo 1:10, onde um bumbo forte e dois fracos são tocados em sequência.

Uma última característica observada no resultado da separação do bumbo do *kit* 1 é que, em momentos onde o bumbo é tocado várias vezes em um curto intervalo de tempo, houve maior incidência de vazamento residual. Isso é evidente no tempo 1:43, onde o bumbo é tocado 11 vezes consecutivas. Nesse trecho, é possível ouvir não apenas resíduos de vazamentos do prato, mas também um vazamento de caixa próximo ao tempo 1:45.

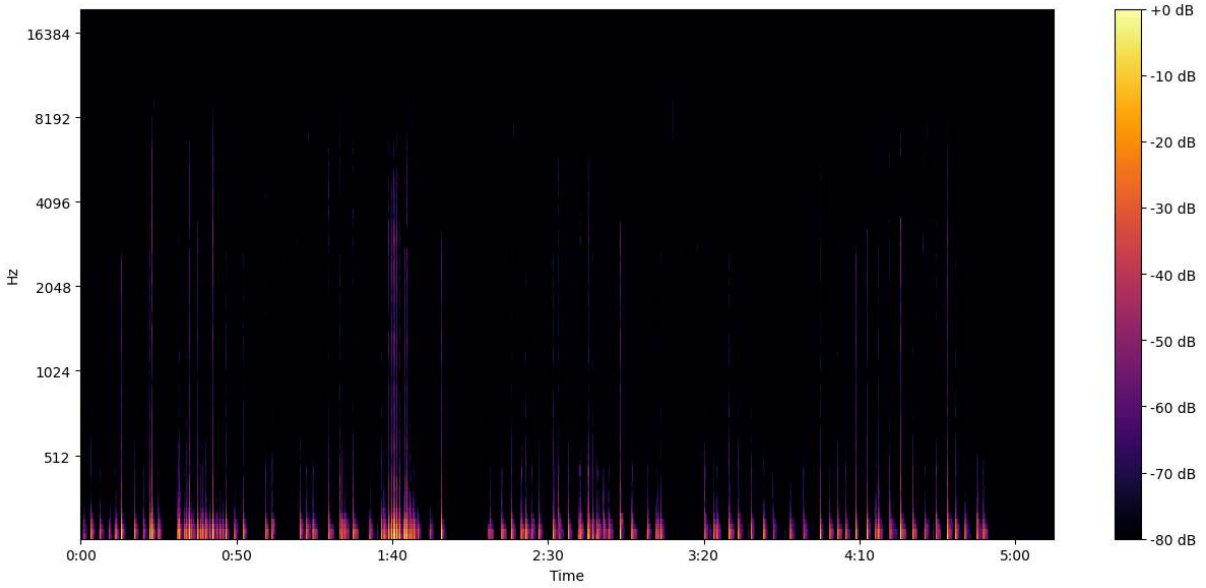
O resultado da separação do bumbo do *kit* 2 também se mostrou satisfatório. Houve poucas ocorrências de vazamento, sendo a maioria delas relacionada a pratos tocados simultaneamente ao bumbo. O primeiro bumbo do áudio, tocado em 0:02, pode ser usado como exemplo desse tipo de vazamento. Esses resultados sugerem que o algoritmo pode estar interpretando algumas características do som dos pratos como parte do som do bumbo.

Um resultado interessante da segunda performance é que ela preservou mais fielmente o timbre original da peça analisada. É possível observar que, nessa gravação, o bumbo foi tocado de forma mais intensa e com menos variações de dinâmica. Isso reforça a ideia de que a preservação das frequências que compõem sua sonoridade após a separação está relacionada à intensidade com que ele foi tocado. Para apoiar o processo de análise, os espectrogramas dos sinais separados estão apresentados nas Figuras 66 e 67.

5.6 Resultados para a classe caixa

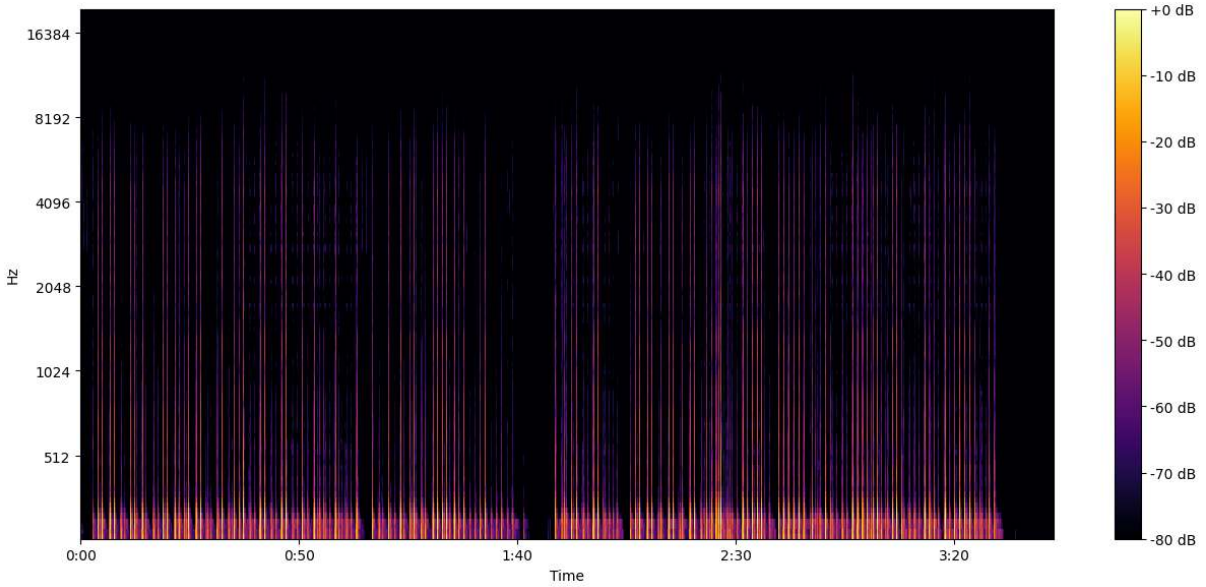
Nessa classe do algoritmo, foram realizadas separações utilizando dois microfones distintos: um apontado para a parte superior da peça (denominado aqui como "caixa") e outro direcionado para a parte inferior (chamado de "esteira"). Apesar de captarem o mesmo instrumento, ambos apresentam características sonoras distintas. Os sinais de áudio originais captados por ambos os microfones podem ser ouvidos nos áudios abaixo. As Figuras 68 e 69

Figura 66 – Microfone do bumbo do *kit* 1 (estúdio) depois da separação.



Fonte: acervo do autor.

Figura 67 – Microfone do bumbo do *kit* 2 (ao vivo) depois da separação.



Fonte: acervo do autor.

apresentam os espectrogramas das caixas dos *kits* 1 e 2, respectivamente, enquanto as Figuras 70 e 71 mostram os espectrogramas das esteiras dos *kits* 1 e 2, respectivamente.

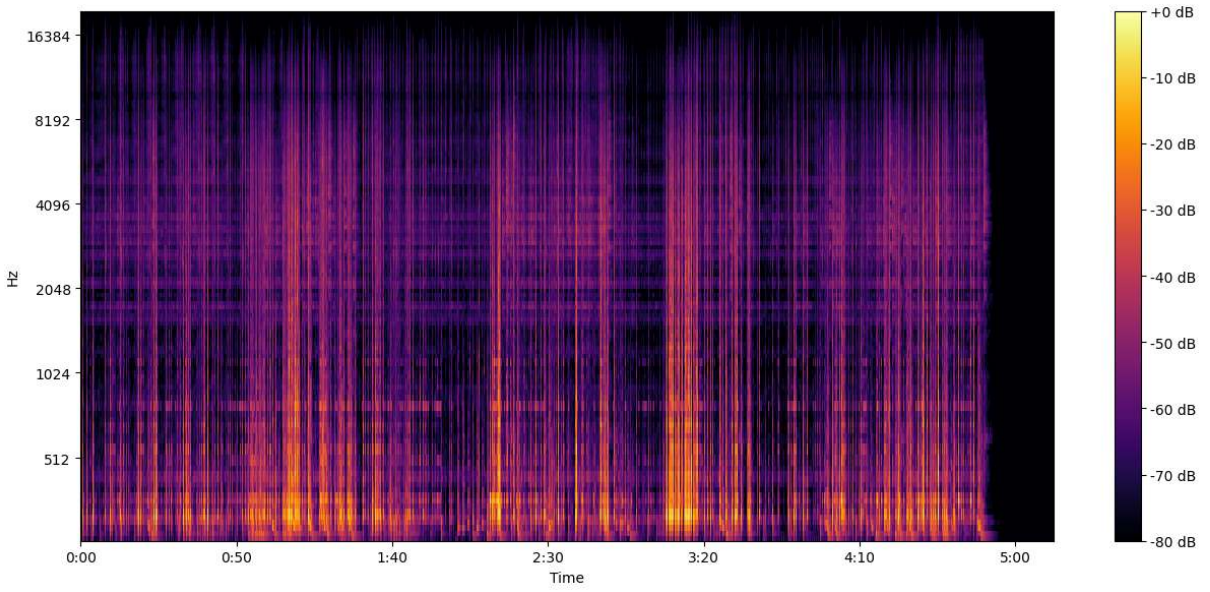
Microfone da caixa - *Kit* 1 (estúdio)

Microfone da caixa - *Kit* 2 (ao vivo)

Microfone da esteira - *Kit* 1 (estúdio)

Microfone da esteira - *Kit* 2 (ao vivo)

Figura 68 – Microfone da caixa do *kit* 1 (estúdio) antes da separação.

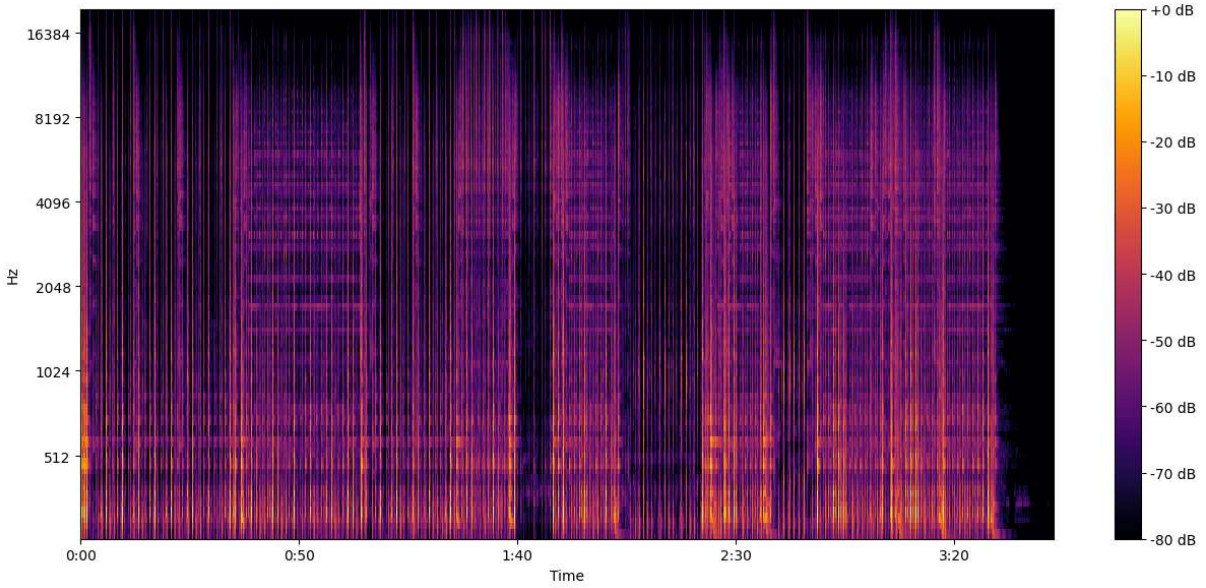


Fonte: acervo do autor.

A caixa apresenta diferenças significativas entre as duas performances. Na do *kit* 1, o baterista utilizou uma grande variedade de técnicas, que abrangem desde toques diretos, *himshot* até o uso do rebote da pele na baqueta para criar efeitos. Além disso, foi empregada uma ampla gama de dinâmicas, variando de toques extremamente suaves a golpes mais agressivos.

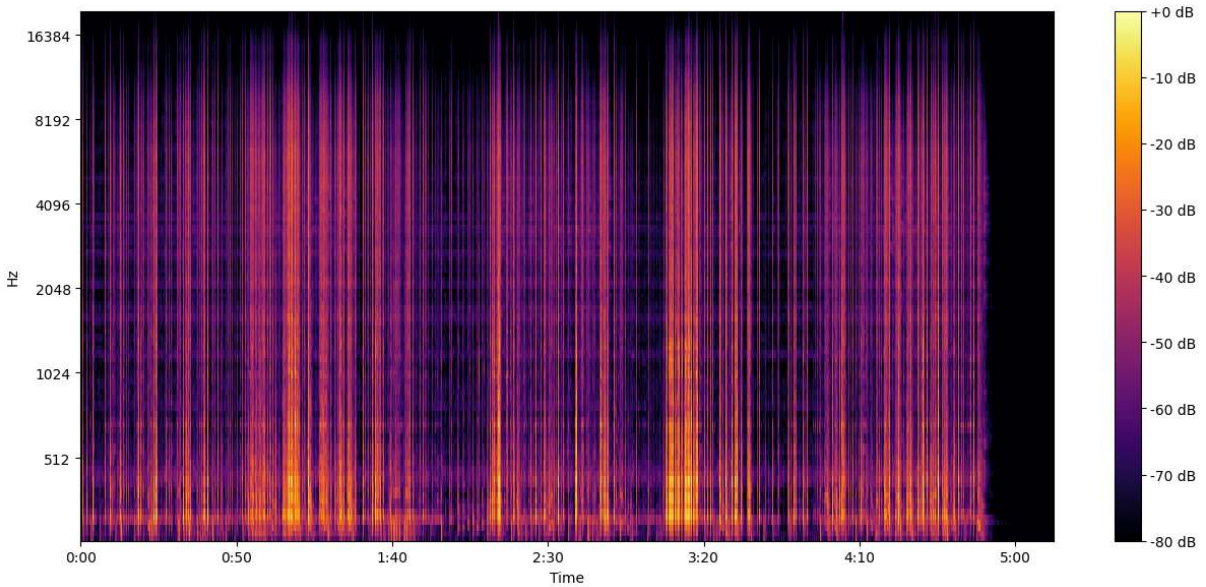
Já na performance do *kit* 2, a caixa foi tocada de maneira mais simples, com maior intensidade e menor variação dinâmica. Em ambas as performances, a técnica de tocar o aro foi utilizada: no *kit* 1, ela ocorre entre 2:00 e 2:09, enquanto no *kit* 2 é aplicada nos intervalos de

Figura 69 – Microfone da caixa do *kit* 2 (ao vivo) antes da separação.



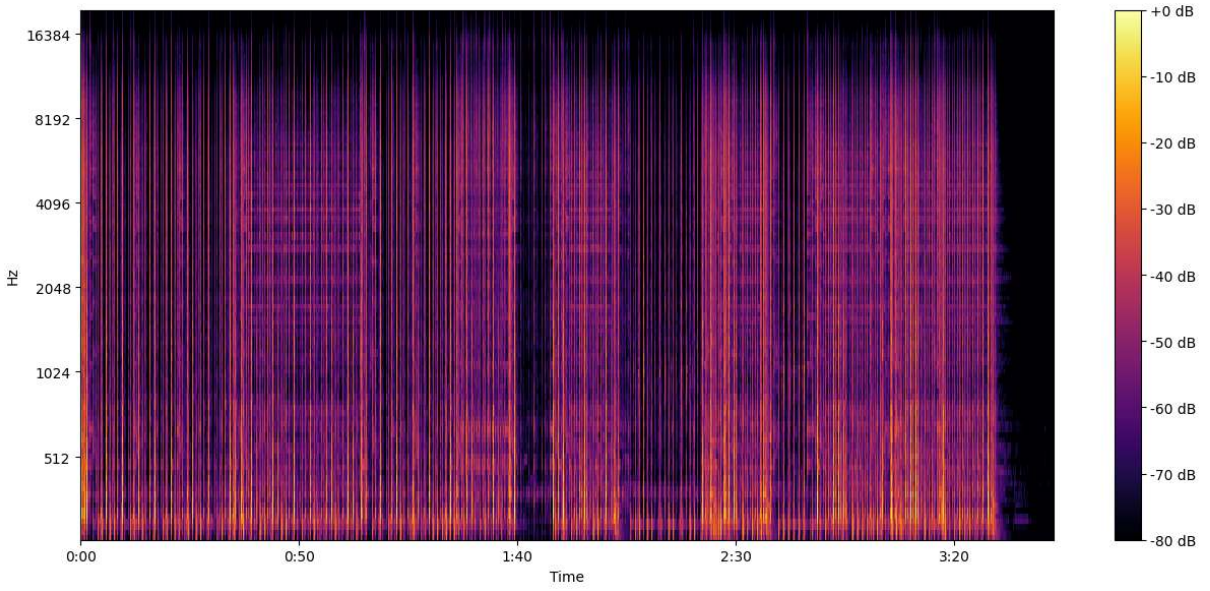
Fonte: acervo do autor.

Figura 70 – Microfone da esteira do *kit* 1 (estúdio) antes da separação.



Fonte: acervo do autor.

Figura 71 – Microfone da esteira do *kit* 2 (ao vivo) antes da separação.



Fonte: acervo do autor.

2:06 a 2:20 e de 2:40 a 2:46. Os resultados da separação, após o processamento pelo LarsNet, podem ser ouvidos nos áudios abaixo.

Microfone da caixa - *Kit* 1 (estúdio) - SEPARADO

Microfone da caixa - *Kit* 2 (ao vivo) - SEPARADO

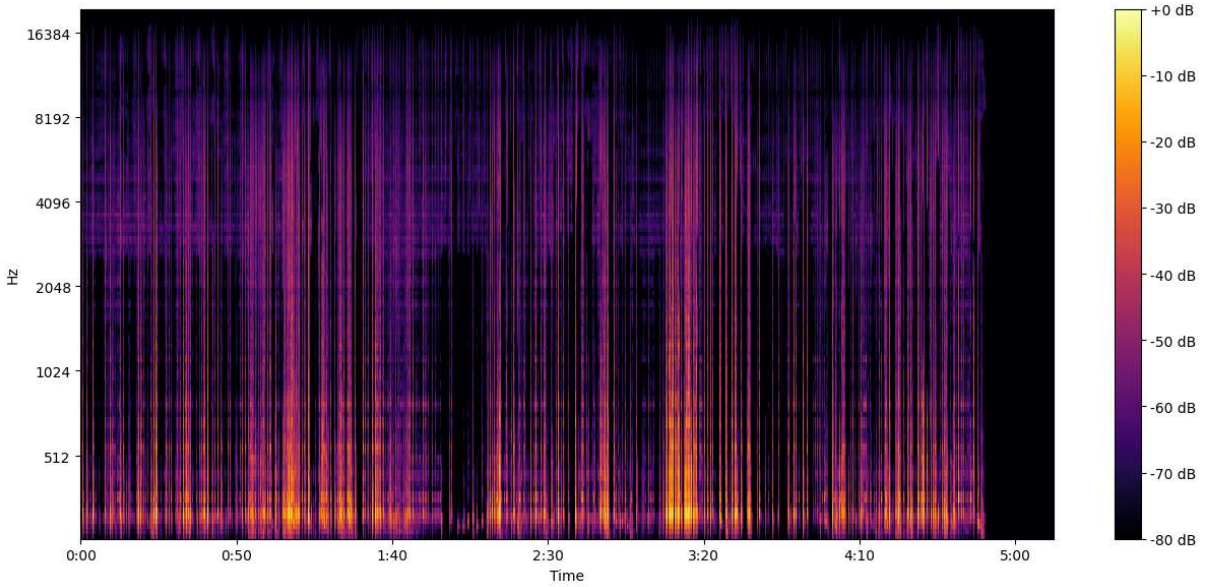
Ao contrário das performances do bumbo, que apresentaram resultados relativamente positivos, a separação das caixas apresentou diversos problemas. Na separação da performance do *kit* 1, é possível notar a ocorrência de vazamentos significativos, principalmente do prato de condução, que já é perceptível logo no início da faixa. Além disso, em trechos onde a caixa é tocada repetidamente em um curto período de tempo, como entre 0:53 e 1:00, observa-se uma grande sobra de outros elementos.

De forma geral, a caixa manteve uma similaridade de timbre com o som original na maior parte do tempo. No entanto, durante a separação, em trechos onde é utilizada a técnica *rimshot*, o timbre da caixa é alterado, perdendo frequências agudas, como na passagem de 4:29 a 4:32. Por outro lado, o som do aro foi completamente perdido durante a separação.

Na separação do *kit* 2, os resultados também não foram satisfatórios. Logo nas primeiras batidas, é possível notar alterações no timbre da caixa. Esse fenômeno se repete em diversos momentos durante a faixa, sendo especialmente perceptível entre 3:05 e 3:08. Além disso,

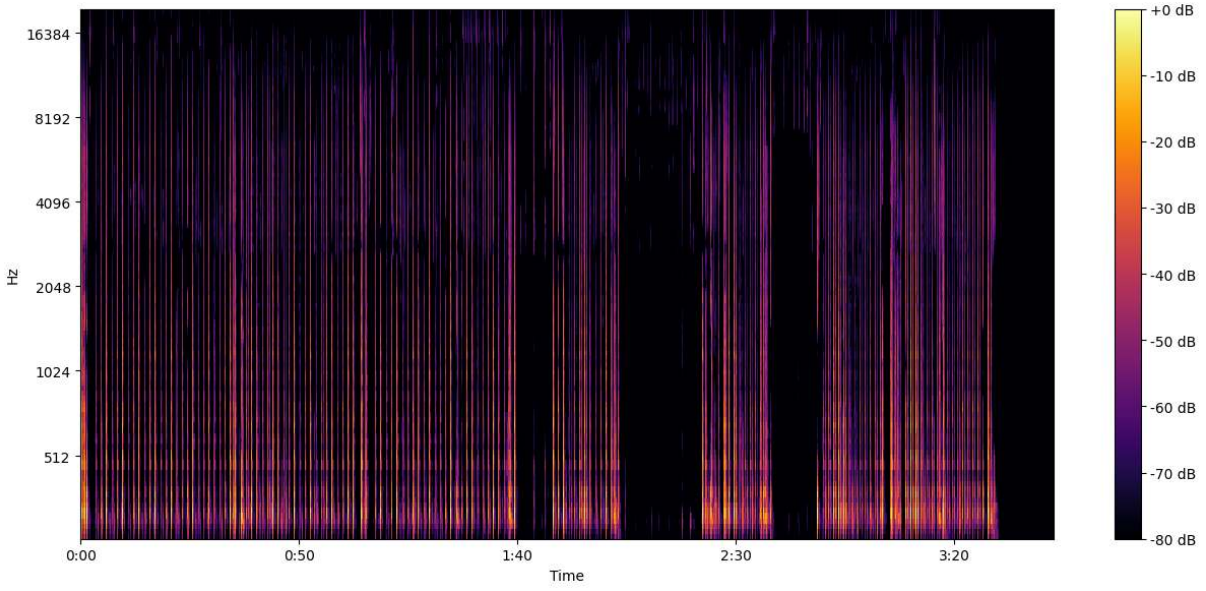
ao longo de toda a extensão do áudio, vazamentos de diferentes peças podem ser ouvidos, como o bumbo em 0:03, o chimbal em 1:28 e os tons em 2:02. Assim como no *kit* 1, o aro foi completamente removido nesta performance. As Figuras 72 e 73 apresentam os espectrogramas dos resultados da separação. Principalmente no espectrograma do segundo *kit*, é possível visualizar espaços vazios no espectro, onde deveria estar o som do aro.

Figura 72 – Microfone da caixa do *kit* 1 (estúdio) depois da separação.



Fonte: acervo do autor.

Figura 73 – Microfone da caixa do *kit* 2 (ao vivo) depois da separação.



Fonte: acervo do autor.

Após realizar a separação no microfone da caixa, o LarsNet foi utilizado para separar o som da mesma performance, registrando a caixa do ponto de vista da esteira. Os resultados podem ser ouvidos nos áudios abaixo

Microfone da esteira - *Kit* 1 (estúdio) - SEPARADO

Microfone da esteira - *Kit* 2 (ao vivo) - SEPARADO

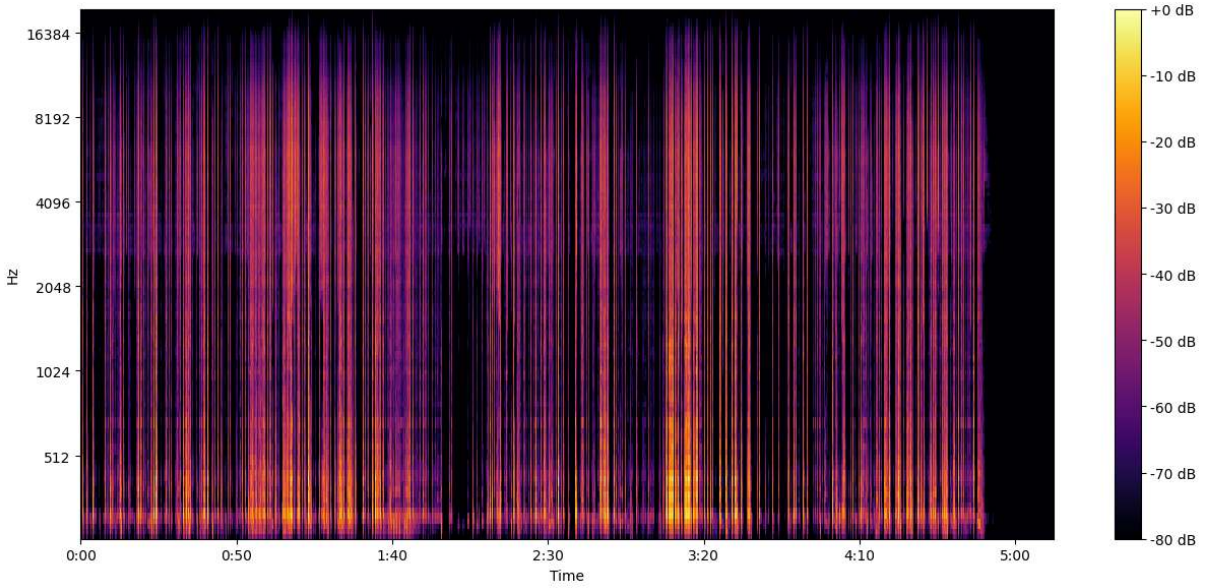
O desempenho da separação nos microfones da esteira não diferiu significativamente das separações realizadas nos microfones da caixa. Todos os problemas e exemplos mencionados para a separação da caixa do *kit* 1 também se aplicam à separação de sua esteira. As diferenças mais notáveis estão relacionadas à intensidade dos vazamentos de cada peça. Isso ocorre porque o microfone da esteira está posicionado mais distante das peças superiores da bateria e mais próximo ao bumbo. Dessa forma, vazamentos de tambores e pratos ainda ocorrem, mas com menor intensidade. Por outro lado, os vazamentos do bumbo tornam-se mais perceptíveis, como no trecho de 1:40. Apesar dessas diferenças, o algoritmo apresentou limitações no processo de separação. Nesse exemplo, o som do aro é audível, mas apenas como leves resquícios sonoros.

No *kit* 2, os resultados também não diferiram muito. Os vazamentos permaneceram persistentes e perceptíveis durante toda a performance. Apesar disso, é possível afirmar que houve uma leve melhora na preservação do timbre da caixa. Isso provavelmente se deve ao fato de que o microfone da esteira está mais afastado do aro, o que reduziu a presença do som do aro no áudio captado durante a execução de técnicas como o *rimshot*. Isso pode indicar uma possível influência dessa componente na preservação do timbre em casos de separação de caixa. Assim como no *kit* 1, o som do aro foi preservado apenas como resquícios sonoros. As Figuras 74 e 75 apresentam os espectrogramas dessa separação e auxiliam na análise visual dos resultados.

5.7 Resultados para a classe tons

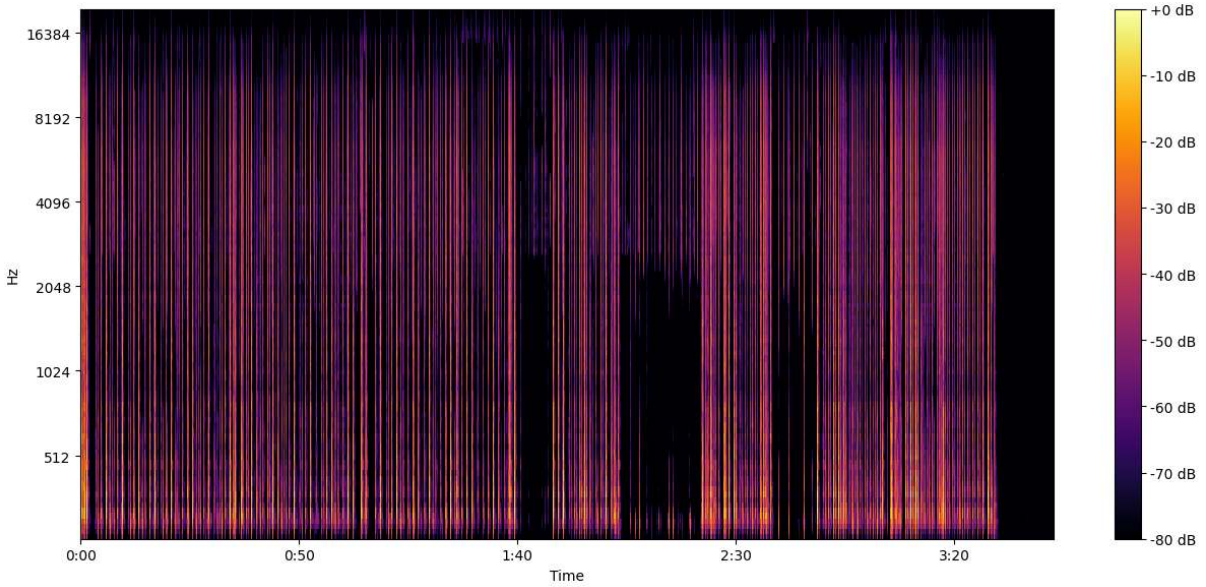
Nesta seção, para facilitar a análise dos resultados, não será feita a diferenciação entre tons e surdos. Isso porque, enquanto o *kit* 1 utilizou dois tons e um surdo, o *kit* 2 empregou um tom e dois surdos, o que poderia gerar confusão na análise. Assim, todos os tambores serão denominados como tom e numerados de 1 a 3, sendo o 1 o mais agudo e o 3 o mais grave. As Figuras 76, 77 e 78 apresentam os espectrogramas dos tons 1, 2 e 3 do *kit* 1, respectivamente. Já as Figuras 79, 80 e 81 mostram os espectrogramas dos tons 1, 2 e 3 do *kit* 2. Os áudios abaixo apresentam a gravação original de cada tambor.

Figura 74 – Microfone da esteira do *kit* 1 (estúdio) depois da separação.



Fonte: acervo do autor.

Figura 75 – Microfone da esteira do *kit* 2 (ao vivo) depois da separação.



Fonte: acervo do autor.

Microfone do tom 1 - *Kit* 1 (estúdio)

Microfone do tom 2 - *Kit* 1 (estúdio)

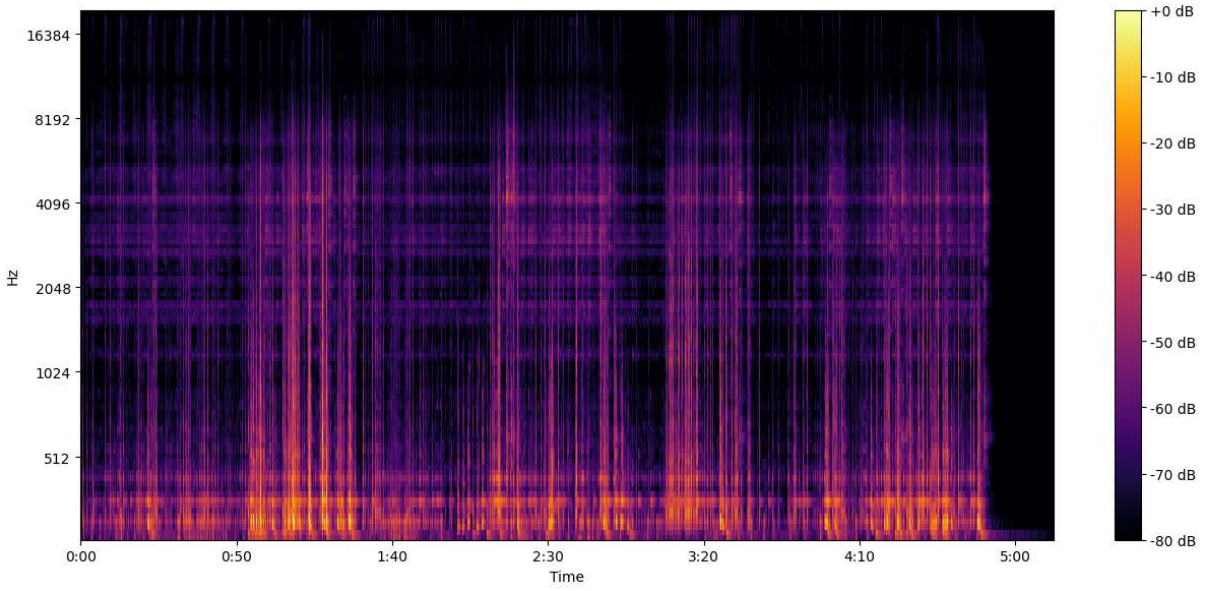
Microfone do tom 3 - *Kit* 1 (estúdio)

Microfone do tom 1 - *Kit* 2 (ao vivo)

Microfone do tom 2 - *Kit* 2 (ao vivo)

Microfone do tom 3 - *Kit* 2 (ao vivo)

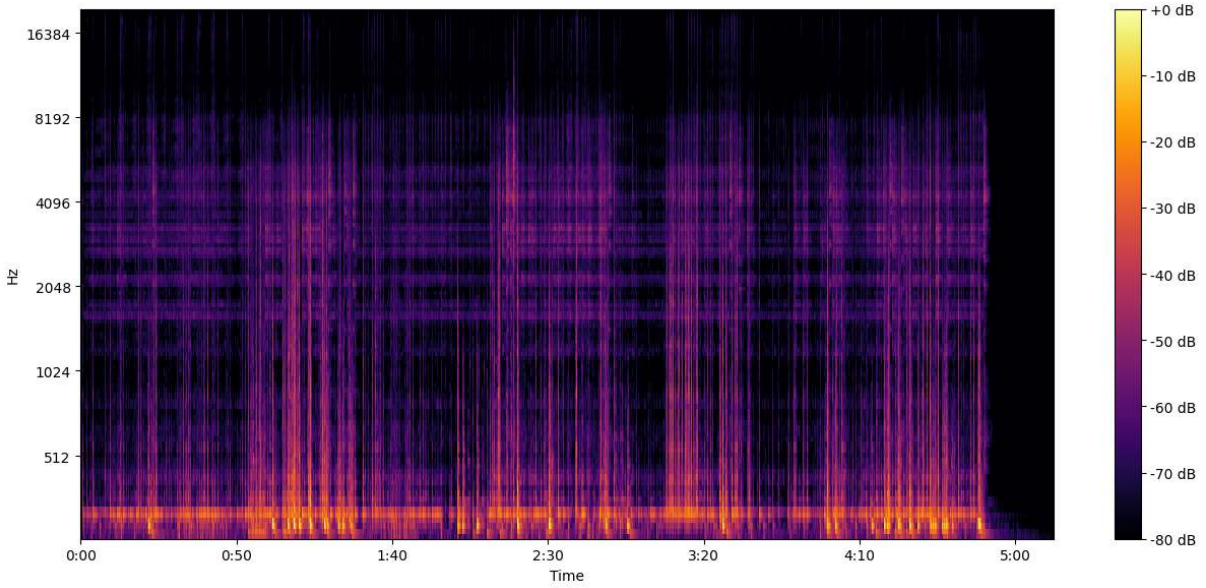
Figura 76 – Microfone do *tom* 1 do *kit* 1 (estúdio) antes da separação.



Fonte: acervo do autor.

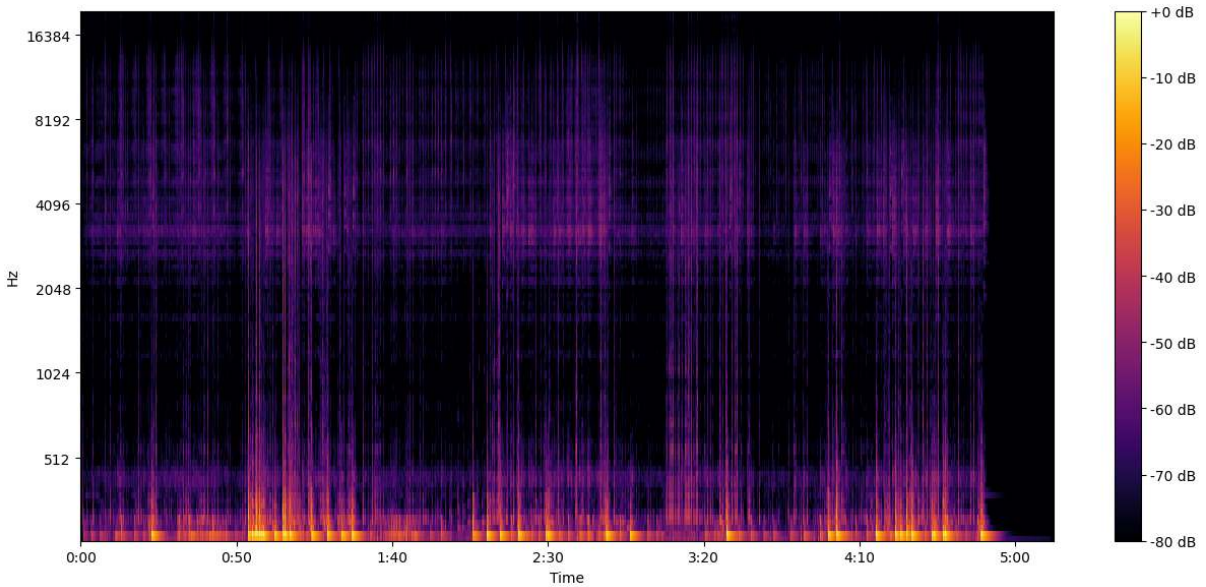
O uso dos tons varia entre as duas performances. Na primeira, por se tratar de um improviso musical livre, o baterista utiliza os tons com maior frequência e por períodos mais longos. Já na segunda gravação, onde a bateria acompanha uma banda, os tons são utilizados de forma mais pontual, atuando principalmente como apoio durante as viradas.

Figura 77 – Microfone do *tom* 2 do *kit* 1 (estúdio) antes da separação.



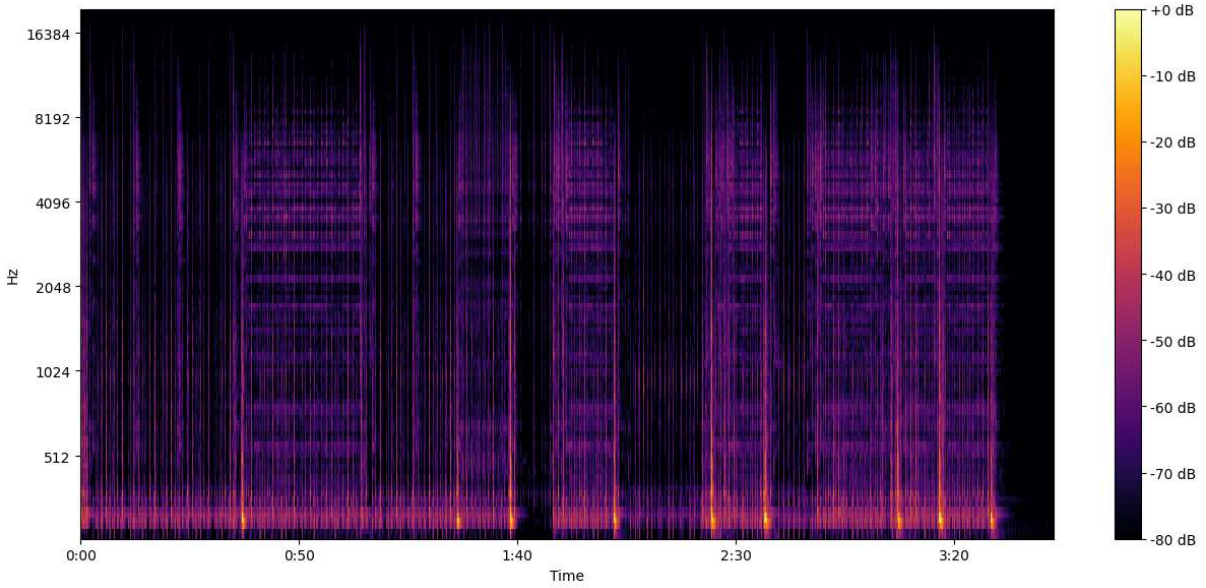
Fonte: acervo do autor.

Figura 78 – Microfone do *tom* 3 do *kit* 1 (estúdio) antes da separação.



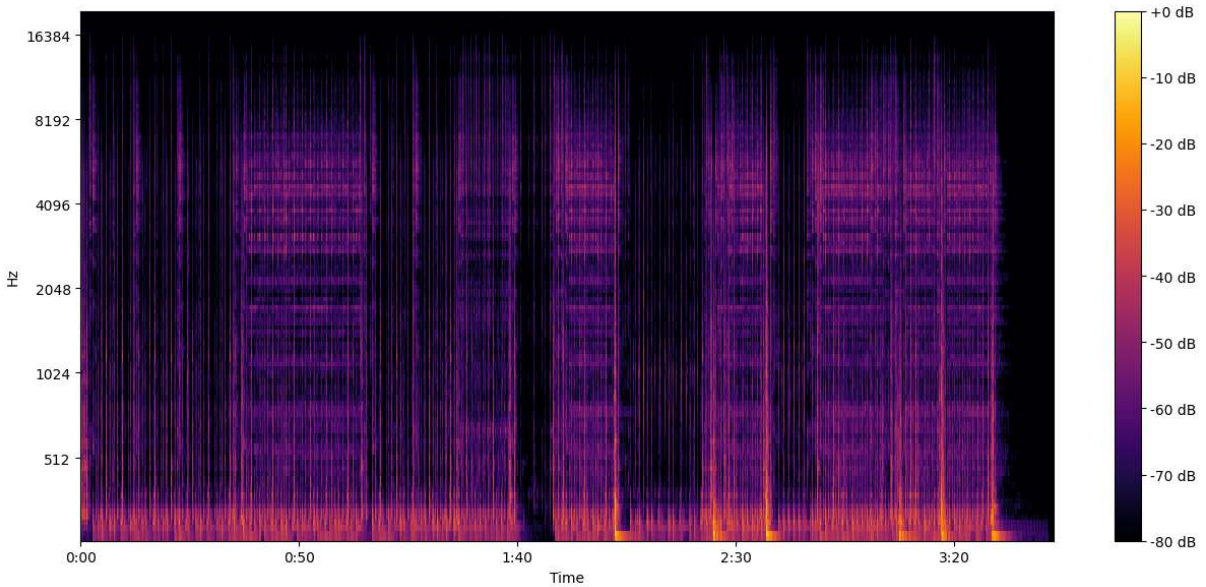
Fonte: acervo do autor.

Figura 79 – Microfone do *tom* 1 do *kit* 2 (estúdio) antes da separação.



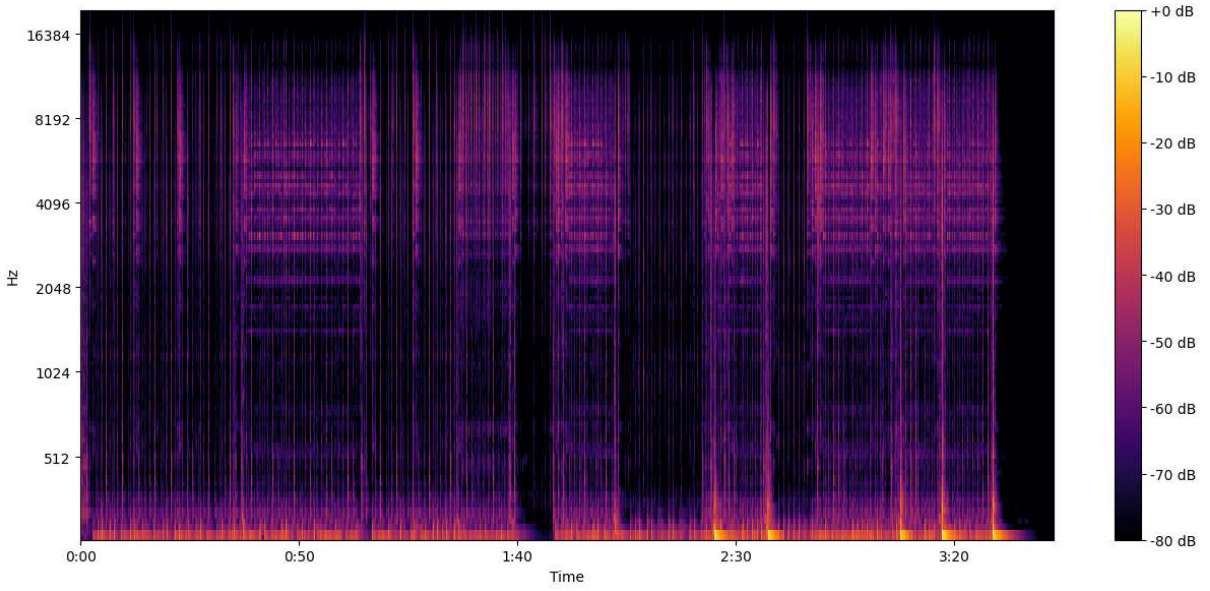
Fonte: acervo do autor.

Figura 80 – Microfone do *tom* 2 do *kit* 2 (estúdio) antes da separação.



Fonte: acervo do autor.

Figura 81 – Microfone do *tom* 3 do *kit* 2 (estúdio) antes da separação.



Fonte: acervo do autor.

Um fator interessante a ser observado é que, em ambas as gravações, é possível ouvir uma nota musical quando um dos tons é tocado. Essa nota é consequência da ressonância natural da própria peça, que gera essa sonoridade. Em bancos de dados amostrados, essa característica muitas vezes é removida por meio de processamentos e técnicas específicas, não estando presente em muitas amostras. No entanto, em contextos de gravações reais, trata-se de um fenômeno comum e natural do próprio instrumento. O resultado da separação do primeiro *kit* pode ser conferido nos áudios abaixo.

Microfone do tom 1 - *Kit* 1 (estúdio) - SEPARADO

Microfone do tom 2 - *Kit* 1 (estúdio) - SEPARADO

Microfone do tom 3 - *Kit* 1 (estúdio) - SEPARADO

Em termos de identificação da peça e remoção de vazamento, o resultado se mostrou o mais satisfatório do trabalho até o momento. O algoritmo foi preciso em identificar os momentos em que os três tons são tocados e separá-los corretamente. Durante os trechos em que as peças não estão sendo tocadas, não há nenhum registro sonoro presente na faixa de áudio. Entretanto, especialmente na faixa do tom 3, é possível ouvir vazamentos de outros elementos durante a execução do tom. A passagem a partir de 2:10 ilustra esse fenômeno.

Apesar do bom desempenho na identificação e separação dos tons, houve uma perda significativa no timbre do instrumento. Como mostrado nas Figuras 82, 83 e 84, a maior parte das frequências acima de 3 kHz foi eliminada. Além disso, percebe-se que, em alguns momentos, como em 1:19 da faixa do tom 2, há a introdução de artefatos no som da peça.

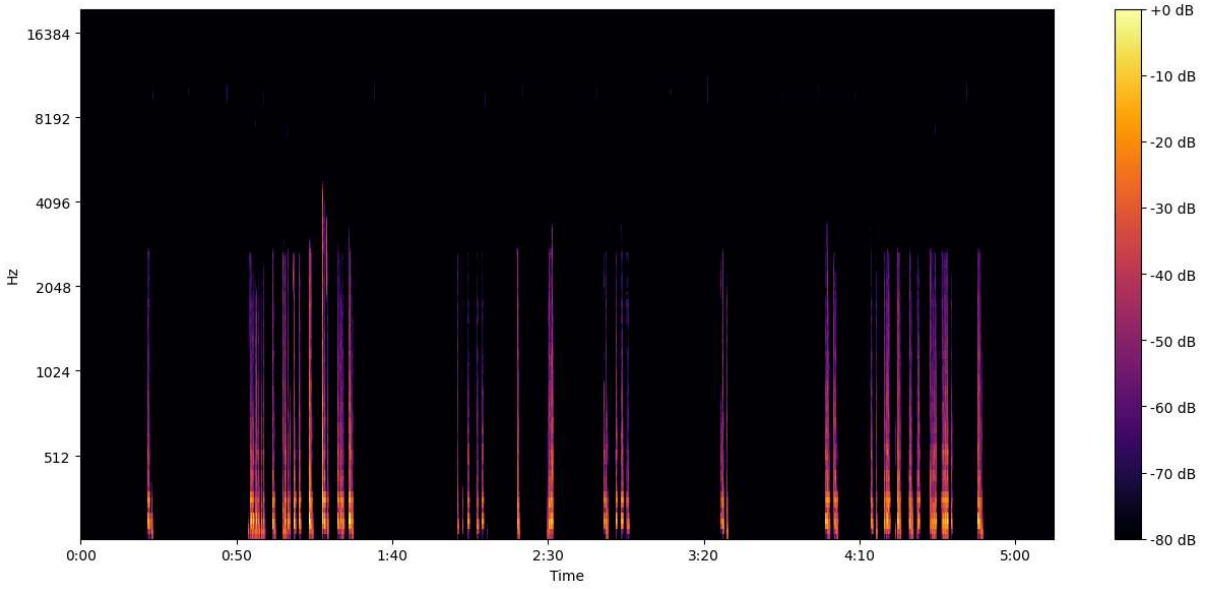
Os resultados dos sons separados para o *kit* 2 são apresentados nos áudios abaixo.

Microfone do tom 1 - *Kit* 2 (ao vivo) - SEPARADO

Microfone do tom 2 - *Kit* 2 (ao vivo) - SEPARADO

Microfone do tom 3 - *Kit* 2 (ao vivo) - SEPARADO

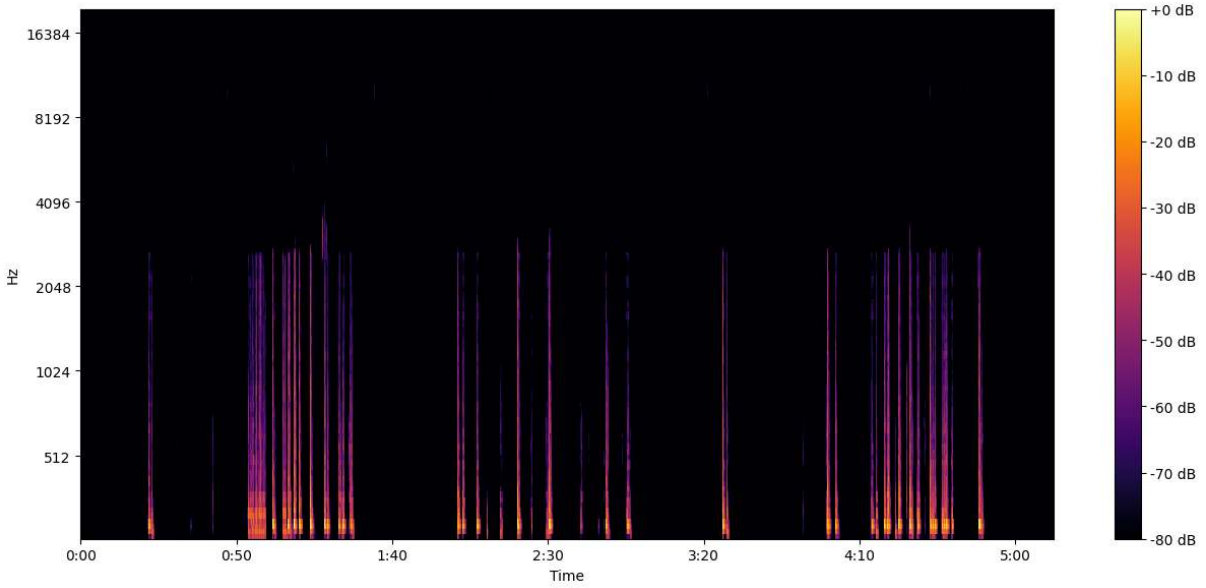
Figura 82 – Microfone do *tom* 1 do *kit* 1 (estúdio) depois da separação.



Fonte: acervo do autor.

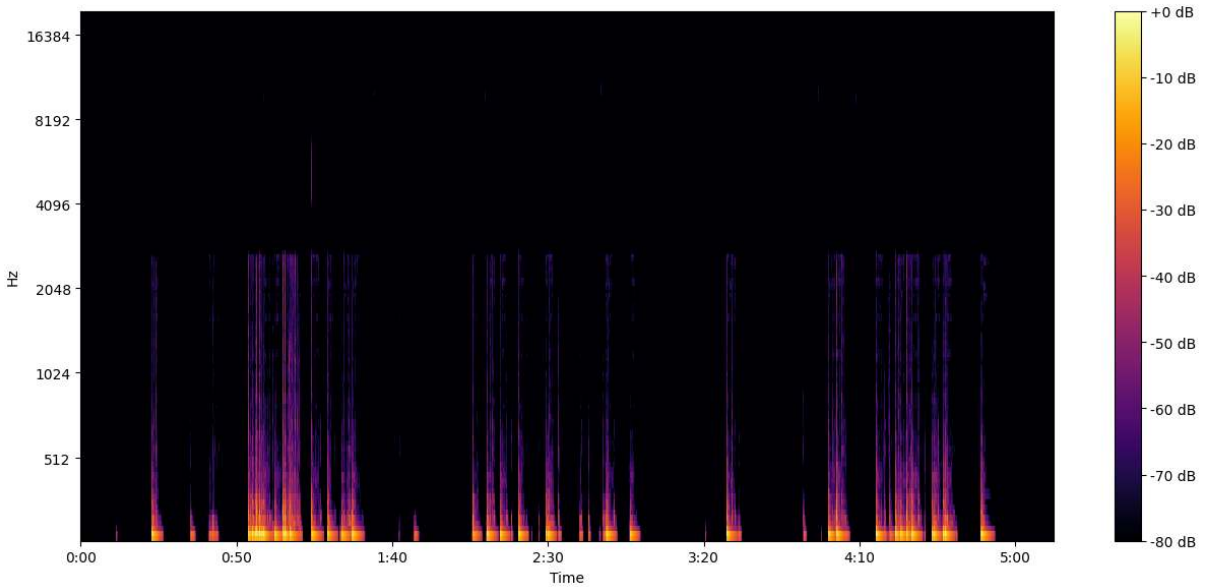
Assim como no *kit* 1, a identificação das peças no *kit* 2 também apresentou um resultado satisfatório no que diz respeito à identificação dos tons e à remoção de vazamentos. No entanto, essa separação apresentou tanto pontos positivos quanto negativos em comparação à primeira. O ponto positivo é que, como pode ser observado nas Figuras 85, 86 e 87, algumas peças conseguiram preservar informações nas regiões agudas, resultando em um timbre mais próximo ao original. O ponto negativo, por outro lado, é que na separação do tom 1, o som perdeu

Figura 83 – Microfone do *tom* 2 do *kit* 1 (estúdio) depois da separação.



Fonte: acervo do autor.

Figura 84 – Microfone do *tom* 3 do *kit* 1 (estúdio) depois da separação.

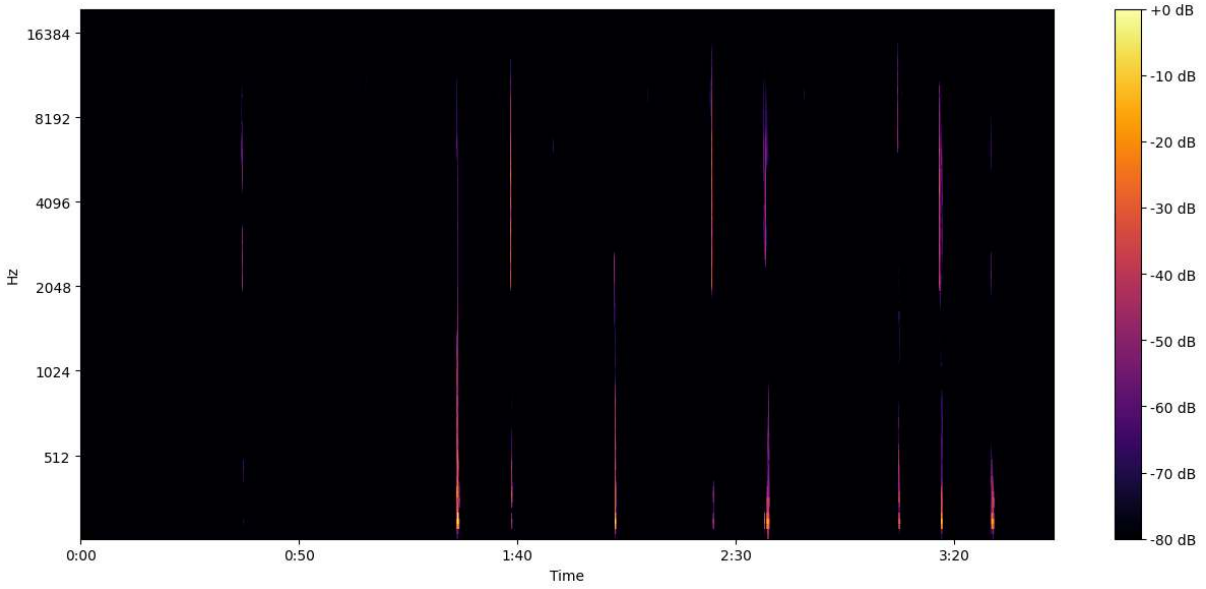


Fonte: acervo do autor.

informações essenciais para a caracterização da peça, transformando-se em um ruído que não corresponde à sua sonoridade original.

Apesar do bom desempenho geral, é pouco provável que um produtor musical utilize esses sinais resultantes da ferramenta em um projeto devido às alterações timbrísticas. Além disso, uma prática comum no processamento de sinais dos tons é o recorte manual dos momentos em que são tocados. Como essas peças são tocadas principalmente em viradas e não fazem parte da execução rítmica contínua, torna-se mais eficiente editá-las manualmente do que recorrer a uma ferramenta de DSS que pode introduzir perdas no timbre.

Figura 85 – Microfone do *tom* 1 do *kit* 2 (estúdio) depois da separação.



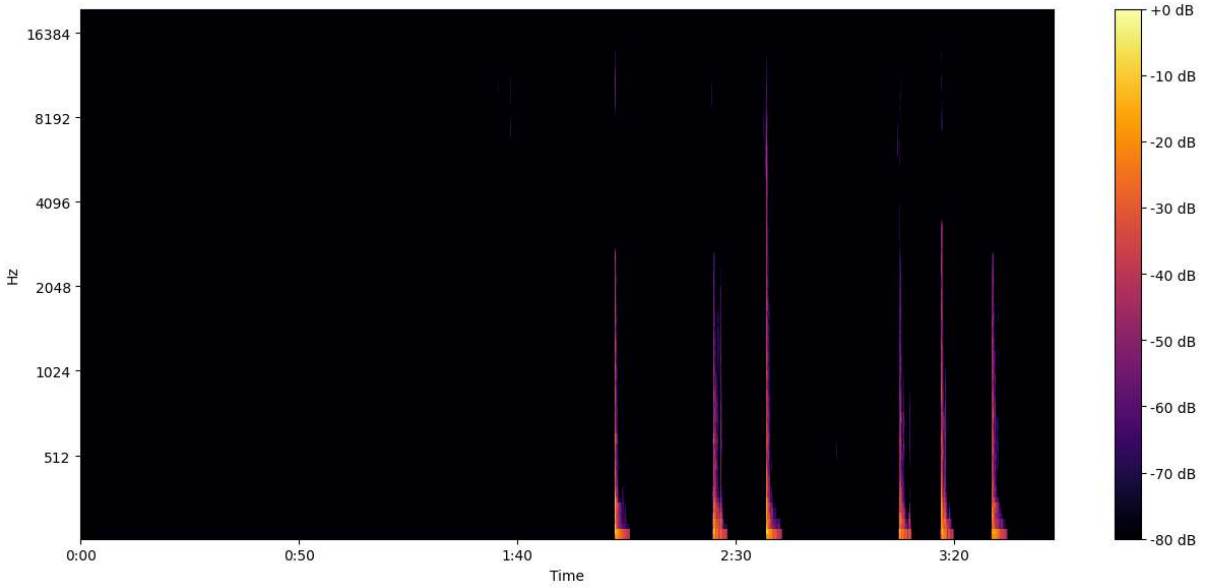
Fonte: acervo do autor.

5.8 Resultados para a classe chimbal

Conforme discutido no Capítulo 3, o chimbal desempenha um papel fundamental na construção do ritmo de uma música. Além disso, é um instrumento que oferece uma ampla gama de possibilidades de execução. Por essa razão, o LarsNet dedica uma classe específica para a sua separação. Nos áudios a seguir, é possível ouvir a captação do microfone do chimbal nas duas performances analisadas, enquanto as Figuras 88 e 89 apresentam os espectrogramas correspondentes.

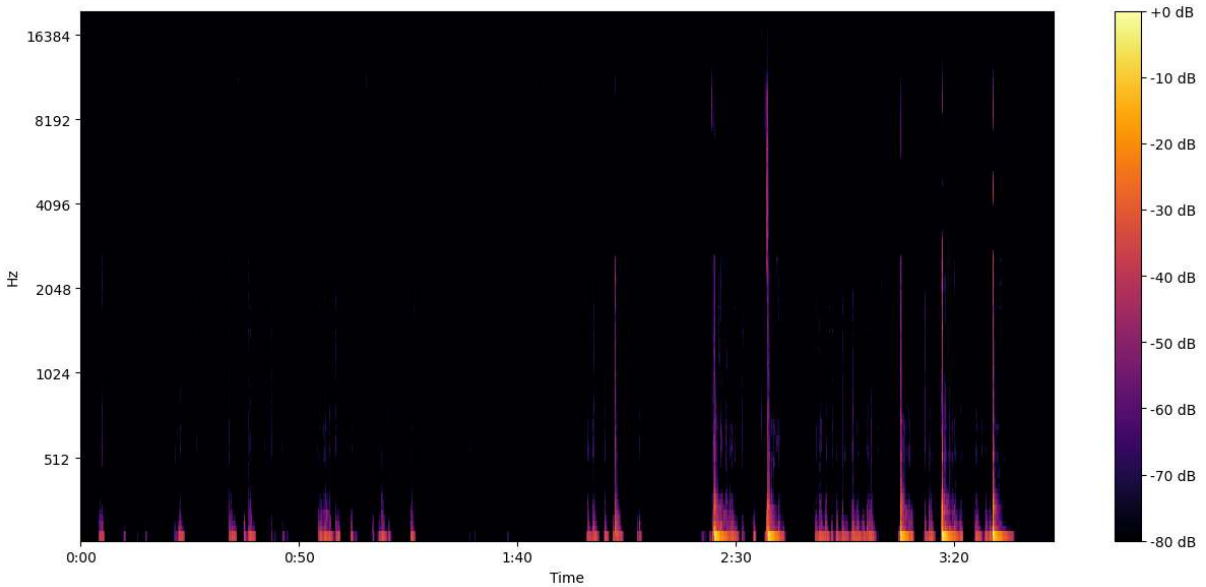
Microfone do chimbal - *Kit* 1 (estúdio)

Figura 86 – Microfone do *tom* 2 do *kit* 2 (estúdio) depois da separação.



Fonte: acervo do autor.

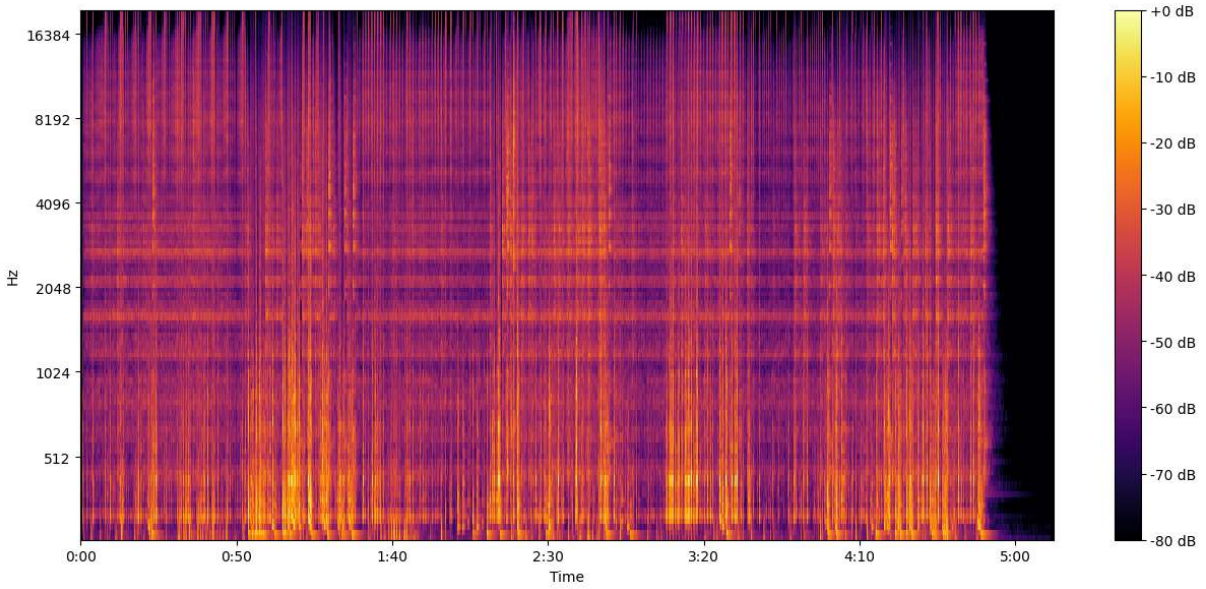
Figura 87 – Microfone do *tom* 3 do *kit* 2 (estúdio) depois da separação.



Fonte: acervo do autor.

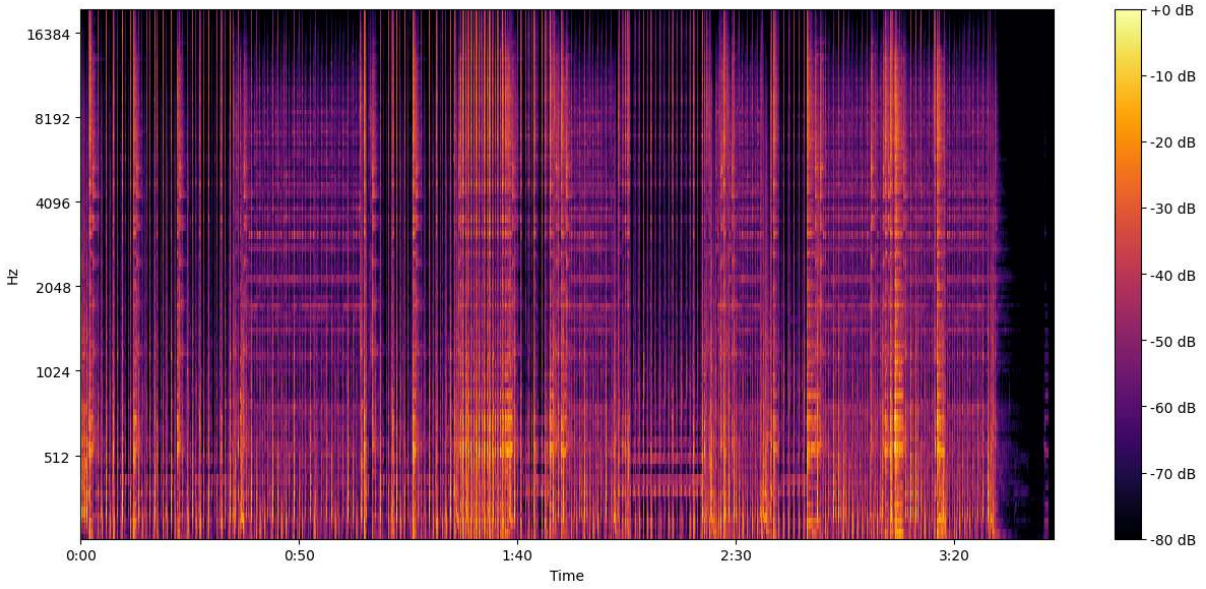
Microfone do chimbal - *Kit* 2 (ao vivo)

Figura 88 – Microfone do chimbal do *kit* 1 (estúdio) antes da separação.



Fonte: acervo do autor.

Figura 89 – Microfone do chimbal do *kit* 2 (ao vivo) antes da separação.



Fonte: acervo do autor.

É possível perceber que a execução do chimbal difere significativamente entre as duas performances. Na do *kit* 1, devido à estética característica do *jazz*, o chimbal é tocado

majoritariamente por meio do acionamento do pedal, enquanto a baqueta é utilizada para golpear outras peças da bateria. Já na segunda gravação, o chimbal é tocado predominantemente com baquetas, alternando entre os modos aberto e fechado. Os áudios abaixo apresentam os resultados da separação de ambos os sinais processados pelo LarsNet.

Microfone do chimbal - *Kit* 1 (estúdio) - SEPARADO

Microfone do chimbal - *Kit* 2 (ao vivo) - SEPARADO

Na separação do *kit* 1, é possível perceber que o algoritmo conseguiu identificar o chimbal em boa parte da performance. No entanto, uma série de outros elementos também foi erroneamente classificada como chimbal, principalmente os pratos. Como a performance faz uso constante do prato de condução e de outros pratos, em muitos momentos suas sonoridades foram identificadas como pertencentes ao chimbal. Esse problema não se restringe apenas aos pratos, sendo possível escutar resquícios de outras peças, como a caixa. O trecho entre 0:07 e 0:13 ilustra essa questão com clareza.

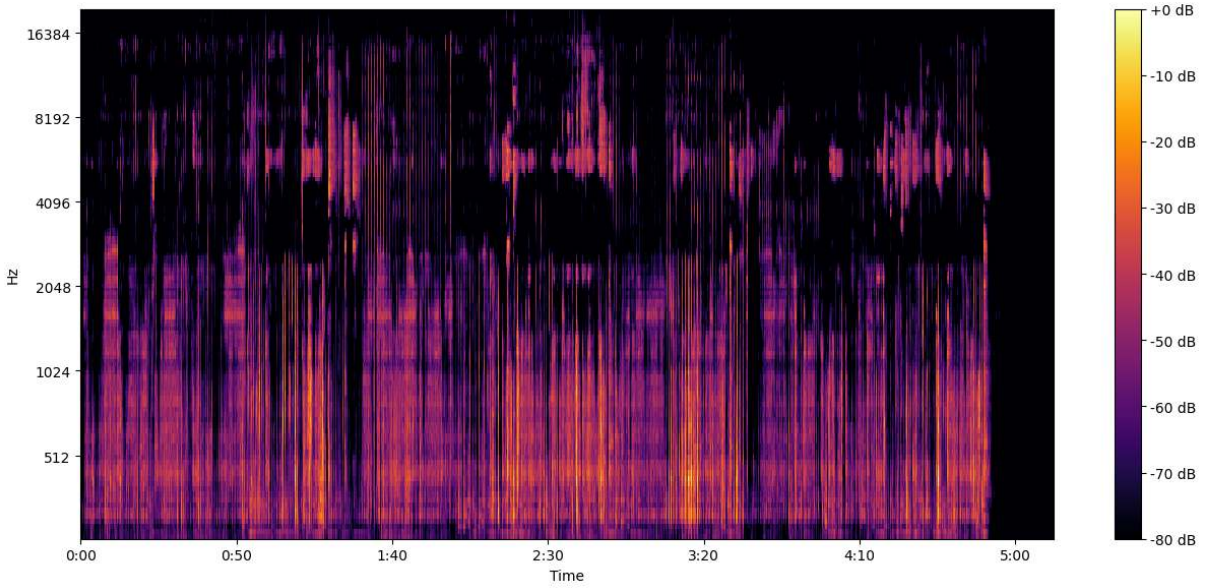
Além da questão do vazamento, é possível observar que houveram alterações no timbre do instrumento. No trecho de 0:14 a 0:19 é possível escutar que algumas execuções de chimbal diferem das outras em termos de sonoridade. A Figura 90 apresenta o espectrograma resultante da separação. Através dele é possível observar como as frequências são distribuídas ao longo da performance.

Os mesmos problemas encontrados na separação do chimbal no *kit* 1 também são observados no *kit* 2. No entanto, de modo geral, o algoritmo foi mais preciso na identificação do chimbal, resultando em uma faixa com menos vazamentos. Isso pode ser atribuído à forma como o chimbal foi tocado em ambas as performances. Enquanto na primeira performance o chimbal foi tocado de maneira mais suave e discreta, na segunda ele foi executado com mais intensidade e atacado diretamente com a baqueta. Esse resultado sugere que as nuances da performance podem influenciar diretamente na qualidade da separação, sendo necessário considerá-las durante o processo de estudo e análise. A Figura 91 apresenta o espectrograma da separação do chimbal no *kit* 2, proporcionando uma melhor visualização dos resultados.

5.9 Resultados para a classe pratos

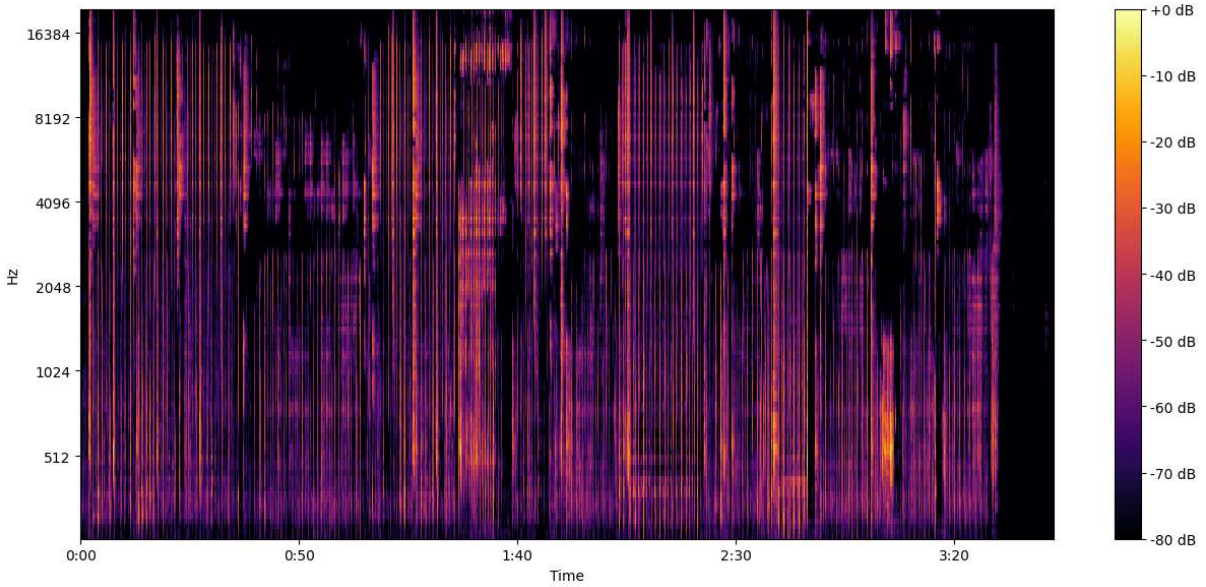
Por fim, a última classe a ser analisada é a classe de pratos. Na ferramenta, essa categoria engloba todos os pratos que compõem um *kit*, com exceção do chimbal. Nessa etapa, serão trabalhados três microfones: os dois *overheads* e o microfone do prato de condução.

Figura 90 – Microfone do chimbal do *kit* 1 (estúdio) depois da separação.



Fonte: acervo do autor.

Figura 91 – Microfone do chimbal do *kit* 2 (ao vivo) depois da separação.



Fonte: acervo do autor.

Os *overheads* representam um caso particular na gravação de bateria. Embora sejam posicionados acima dos pratos para captá-los, esses microfones captam a performance geral de todas as peças da bateria. Dessa forma, seria impreciso afirmar que os *overheads* têm como única intenção captar os pratos. Entretanto, como nos exemplos utilizados as demais peças possuem microfones dedicados, os *overheads* foram utilizados para testar a separação dos pratos pelo algoritmo. Por trabalharem em conjunto, os sinais de ambos os microfones foram transformados em uma única faixa estéreo.

Já em relação à captação do prato de condução, vale destacar que essa não é uma prática comum em todas as sessões de gravação de bateria. Frequentemente, o som desse prato é captado junto com os demais pelos *overheads*. No entanto, por se tratar de um prato de grande importância na condução rítmica das músicas, muitos engenheiros de áudio optam por utilizar um microfone dedicado, como foi o caso das gravações utilizadas neste trabalho.

Os áudios abaixo apresentam ambas as gravações, dos *overs* e do prato de condução. Enquanto isso, as Figuras 92 e 93 ilustram os espectrogramas dos *overheads* dos *kits* 1 e 2, respectivamente. Já as Figuras 94 e 95 mostram os espectrogramas dos pratos de condução dos *kits* 1 e 2, respectivamente.

Microfones dos *overheads* - *Kit* 1 (estúdio)

Microfones dos *overheads* - *Kit* 2 (ao vivo)

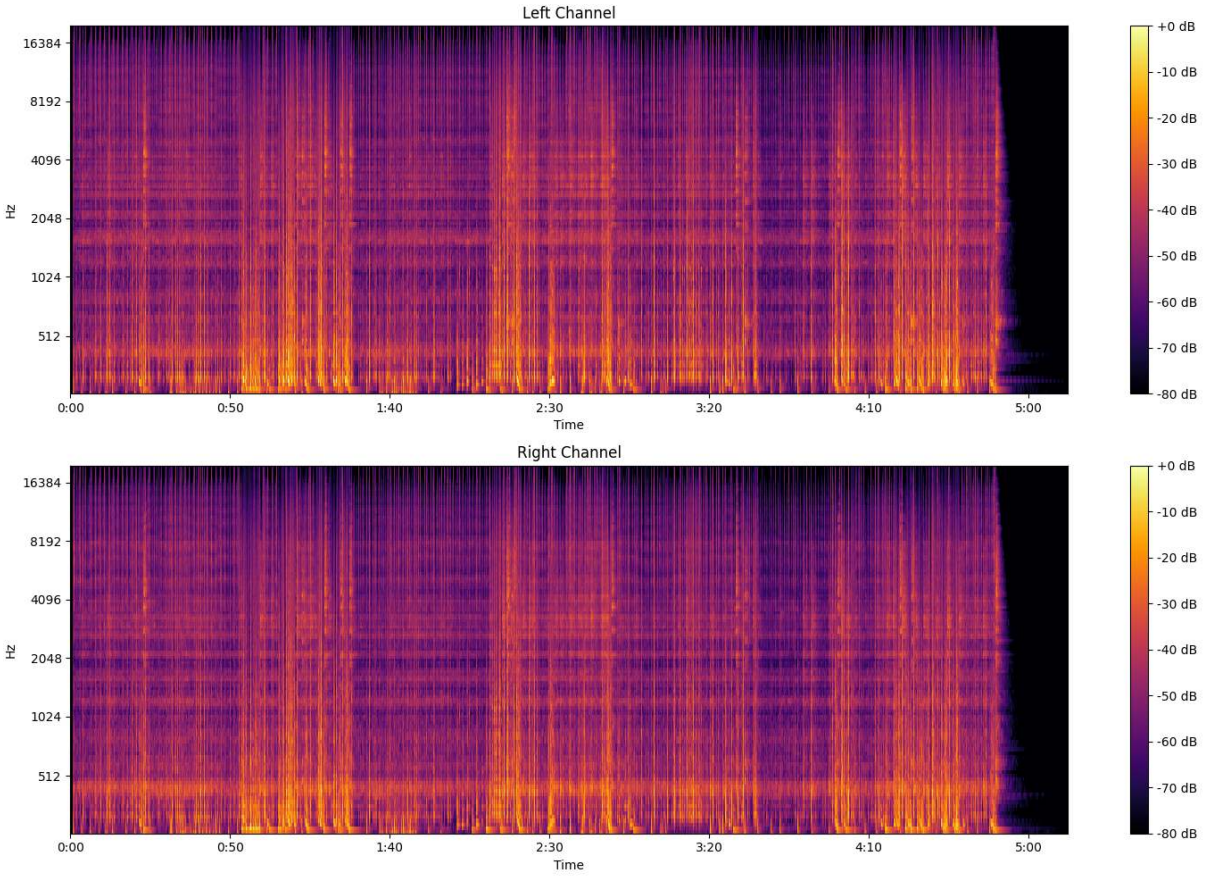
Microfone do prato de condução - *Kit* 1 (estúdio)

Microfone do prato de condução - *Kit* 2 (ao vivo)

Ao comparar ambas as performances, é possível perceber diferenças no uso dos pratos. Na gravação em estúdio, o prato de condução desempenha um papel central, conduzindo o ritmo do *jazz* com toques predominantemente suaves ao longo de quase toda a faixa. Os demais pratos, por sua vez, são utilizados pontualmente para ataques esporádicos. Já na gravação ao vivo, o prato de condução é usado de forma mais pontual, servindo como uma alternativa de marcação de tempo para o chimbal, enquanto os demais pratos são utilizados principalmente para compor as viradas.

O primeiro exemplo a ser analisado é o dos microfones *overheads*. Os áudios, com os pratos separados em estéreo, são apresentados abaixo. As Figuras 96 e 97 exibem os espectrogramas das performances em estúdio e ao vivo, respectivamente.

Figura 92 – Microfones dos *overs* do *kit* 1 (estúdio) antes da separação.



Fonte: acervo do autor.

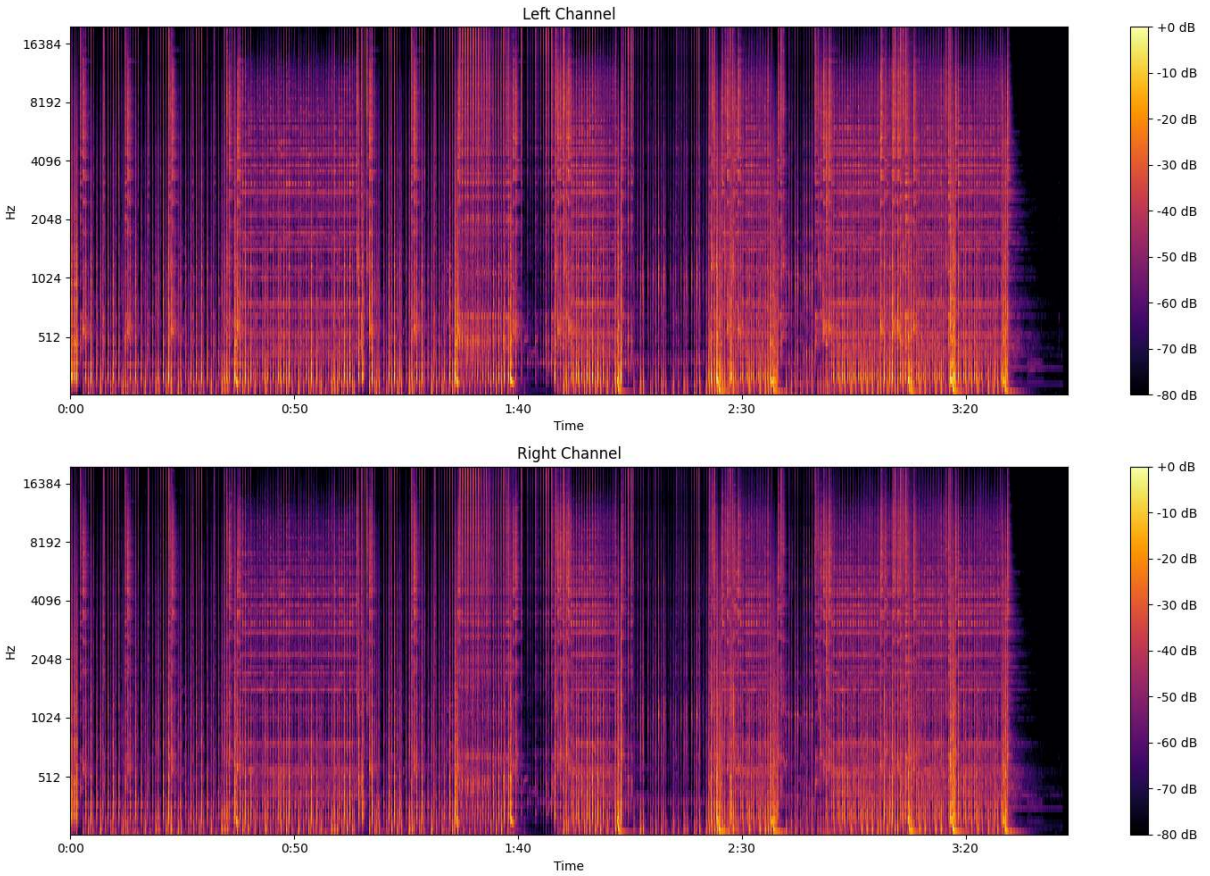
Microfones dos *overheads* - *Kit* 1 (estúdio) - SEPARADO

Microfones dos *overheads* - *Kit* 2 (ao vivo) - SEPARADO

A separação dos pratos nos microfones *overheads* do *kit* 1 apresentou um dos melhores desempenhos registrados neste trabalho. Apesar da performance complexa, com diversas nuances de toque e variações rítmicas, os pratos foram separados dos vazamentos de forma satisfatória. Apenas um pequeno vazamento residual de outras peças pode ser ouvido durante a execução, mas com amplitude muito baixa.

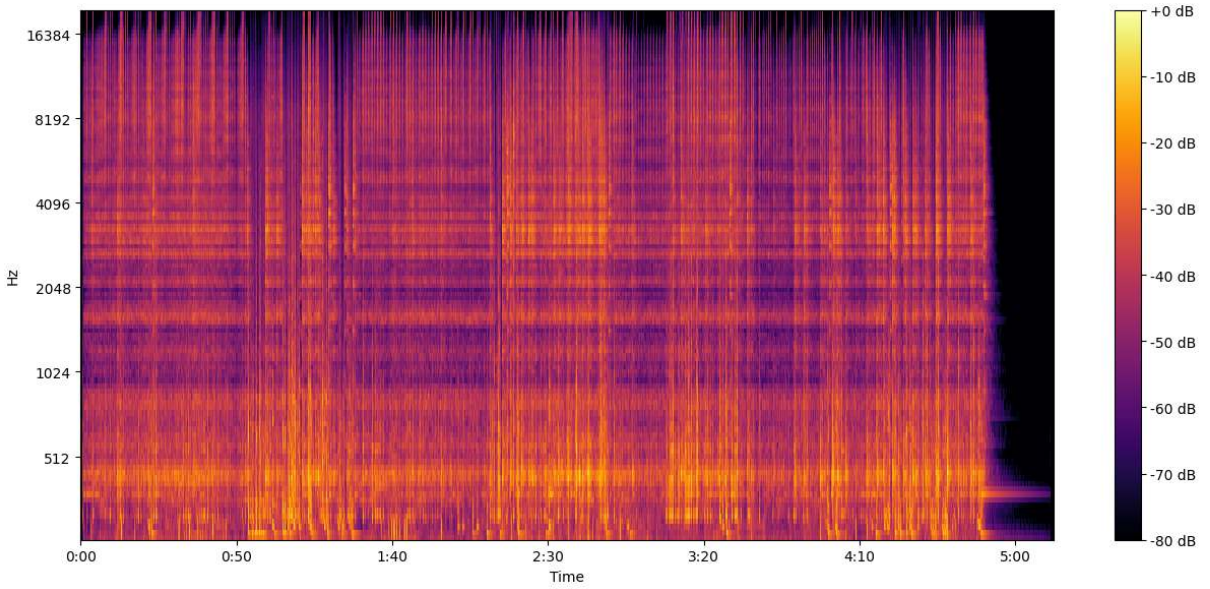
Outro fator relevante nesse resultado foi a capacidade do algoritmo de preservar a qualidade dos timbres após a separação. Esse é um grande feito, considerando que os microfones *overheads* estão entre os que mais sofrem com vazamentos quando o objetivo é a captação dos pratos. Em alguns poucos trechos da performance, é possível notar distorções sonoras, como em 2:10; no entanto, esses casos foram exceções.

Figura 93 – Microfones dos *overs* do *kit* 2 (ao vivo) antes da separação.



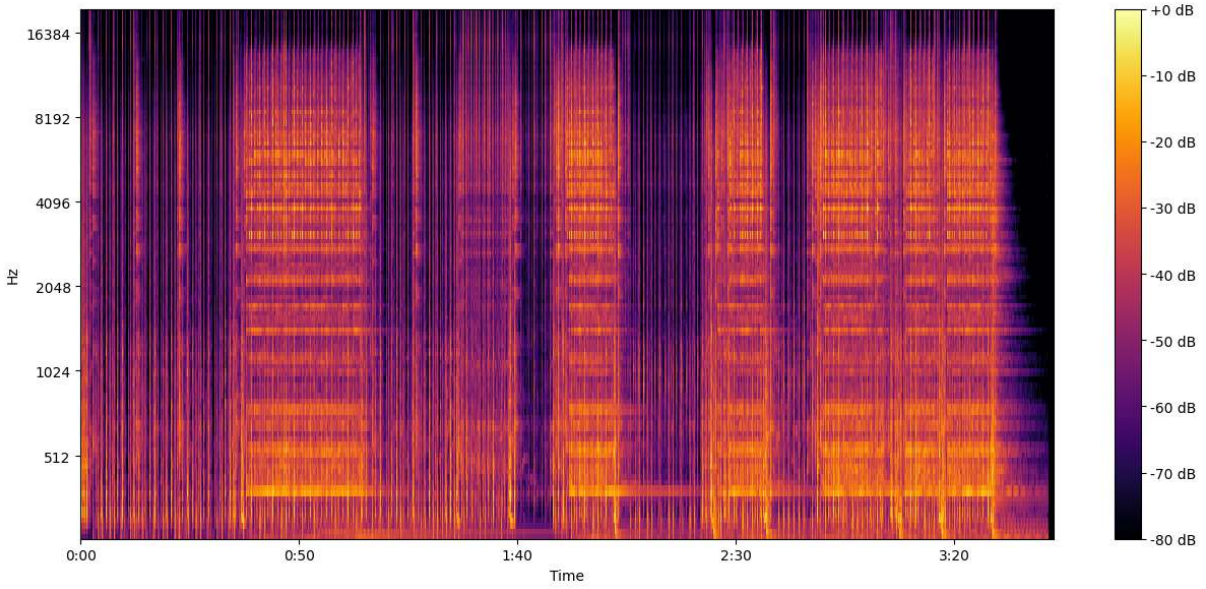
Fonte: acervo do autor.

Figura 94 – Microfone do prato de condução do *kit* 1 (estúdio) antes da separação.



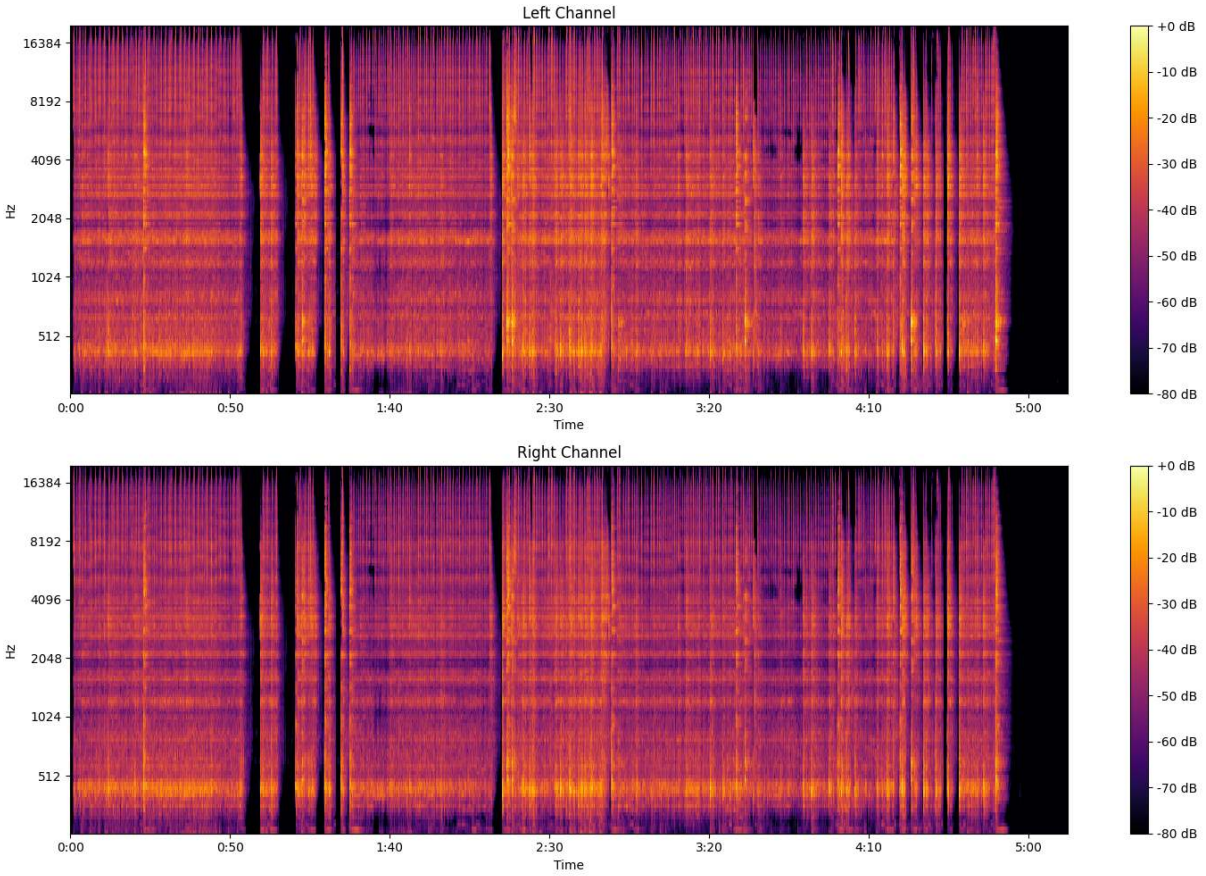
Fonte: acervo do autor.

Figura 95 – Microfone do prato de condução do *kit* 2 (ao vivo) antes da separação.



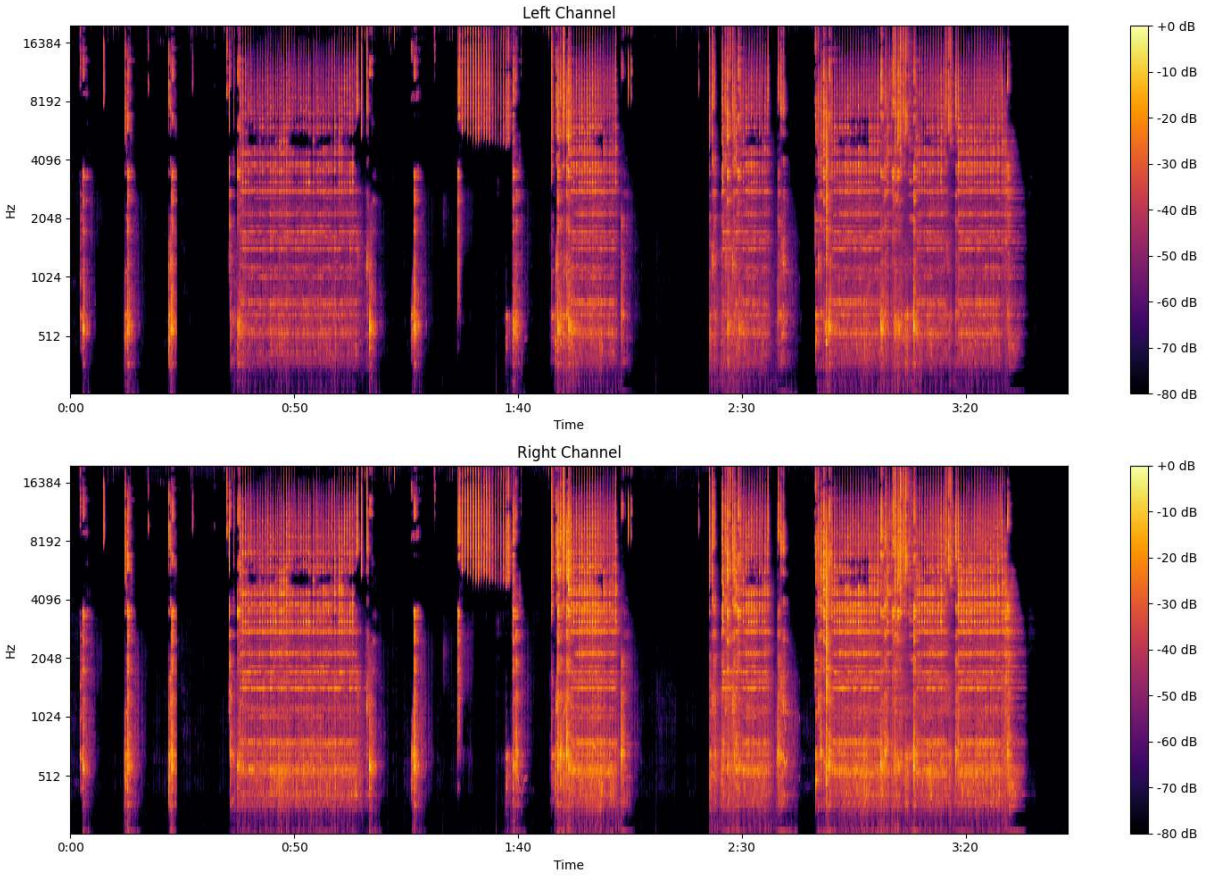
Fonte: acervo do autor.

Figura 96 – Microfones dos *overs* do *kit* 1 (estúdio) depois da separação.



Fonte: acervo do autor.

Figura 97 – Microfones dos *overs* do *kit* 2 (ao vivo) depois da separação.



Fonte: acervo do autor.

Apesar do bom desempenho no *kit* 1, o algoritmo não apresentou resultados tão satisfatórios para o *kit* 2. Primeiramente, houve uma alteração perceptível no timbre dos pratos, especialmente nos pratos que não são de condução. Além disso, no trecho entre 1:26 e 1:38, o algoritmo identificou o som do chimbal como pertencente a esta classe.

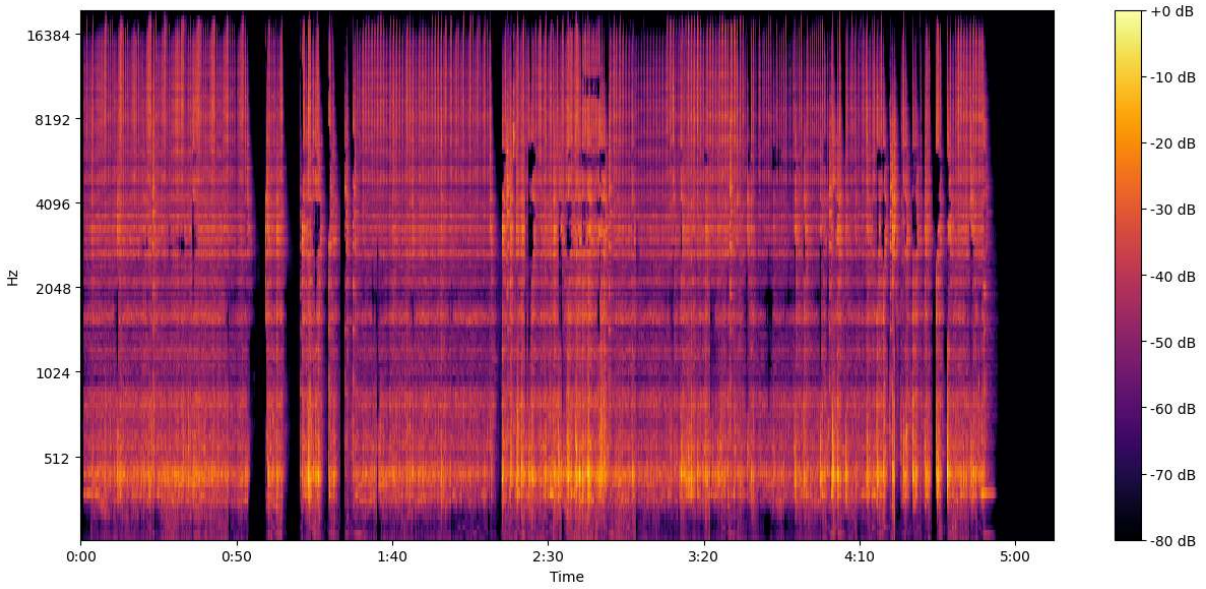
Os mesmos experimentos foram realizados com o microfone dedicado ao prato de condução. Os resultados sonoros podem ser ouvidos nos áudios abaixo. As Figuras 98 e 99 mostram, respectivamente, os resultados para os *kits* 1 e 2.

Microfone do prato de condução - *Kit* 1 (estúdio) - SEPARADO

Microfone do prato de condução - *Kit* 2 (ao vivo) - SEPARADO

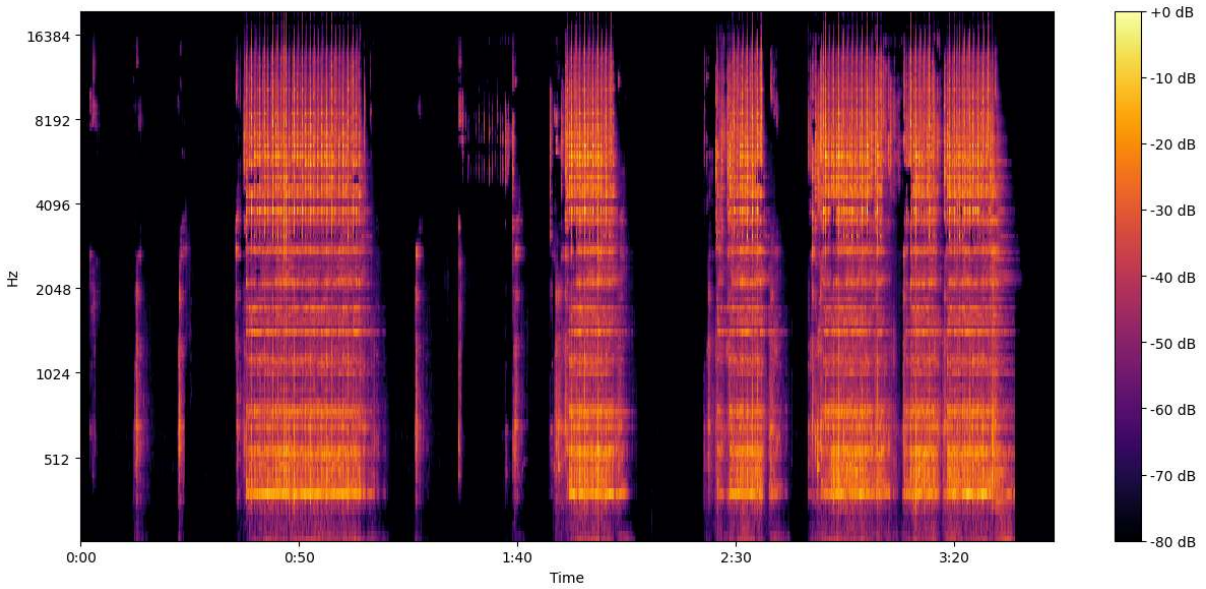
Os resultados obtidos com o microfone do prato de condução no *kit* 1 foram bastante semelhantes aos encontrados nos *overheads* do mesmo *kit*. O algoritmo foi capaz de identificar as

Figura 98 – Microfone do prato de condução do *kit* 1 (estúdio) depois da separação.



Fonte: acervo do autor.

Figura 99 – Microfone do prato de condução do *kit* 2 (ao vivo) depois da separação.



Fonte: acervo do autor.

nuances da performance, preservando adequadamente o timbre e apresentando baixa incidência de vazamentos. No entanto, nesse caso, as distorções no som foram mais recorrentes. Exemplos dessas distorções podem ser ouvidos nos trechos de 0:35, 0:59, 1:10, 1:19 e 2:23.

No *kit* 2, houve algumas melhorias em relação aos resultados obtidos com os *overheads*. A primeira foi a ausência de detecção equivocada do chimbal. A segunda é que o timbre do prato de condução foi preservado em um nível satisfatório. Apesar dessas melhorias, a maioria dos outros pratos apresentou perda significativa de informações sonoras acima de 4 kHz, como demonstrado na Figura 99, o que alterou significativamente o seu timbre.

5.10 Comparação com os métodos tradicionais

Para entender a usabilidade de uma ferramenta de DSS na remoção de vazamentos, é necessário comparar os resultados obtidos com os métodos tradicionais descritos na Seção 3.11. De modo geral, o LarsNet levou, em média, 35 segundos para processar cada faixa da primeira gravação e 50 segundos para cada faixa da segunda gravação. Esse desempenho representa um ponto positivo para ferramentas desse tipo, pois o tempo necessário para um operador ajustar métodos tradicionais costuma ser significativamente maior. Além disso, abordagens que envolvem operações manuais, como automação de volume e edição espectral, poderiam levar minutos ou até horas, mesmo quando realizadas por um profissional experiente.

Alguns dos resultados obtidos podem ser comparados diretamente com os de ferramentas tradicionais. A separação do bumbo é um exemplo disso. Em muitos casos, o uso do *gate* é a técnica aplicada para reduzir os vazamentos em seu microfone. No entanto, essa ferramenta atua com base em um *threshold*, o que pode resultar no silenciamento indesejado de partes onde o som do bumbo possui amplitude mais baixa, como o *attack* e o *release*. Por outro lado, a separação automática mantém essas informações, mas não preserva o timbre com a mesma fidelidade, como demonstram os resultados apresentados. Nesse contexto, a escolha da ferramenta mais adequada dependerá dos objetivos específicos do usuário.

Em algumas situações, a escolha da ferramenta mais adequada pode ser influenciada pela performance musical executada. Um exemplo disso é o caso dos microfones da caixa. Conforme apresentado, no *kit* 1, a caixa é tocada de forma suave, com o uso de diversas técnicas, enquanto outras peças, como o chimbal e o prato de condução, são tocadas ao fundo. Esse cenário representa uma separação bastante complexa de ser realizada por métodos tradicionais e reflete as particularidades de um gênero musical específico. O aprimoramento da qualidade das ferramentas de DSS as tornaria ideais para lidar com esse tipo de contexto.

Por outro lado, no *kit* 2, a caixa é tocada com maior intensidade, apresentando uma amplitude que a destaca em relação às demais peças. Além disso, durante a maior parte da performance, a única peça tocada simultaneamente à caixa é o chimbal, o que facilita sua separação dos outros elementos. Nesse caso, a utilização de uma técnica como a automação de

volume poderia gerar um resultado mais satisfatório, visto que os algoritmos de DSS não apenas alteraram significativamente o som do instrumento, mas também eliminaram as performances com o uso do aro e permitiram a permanência de alguns vazamentos.

Outro exemplo que possibilita uma comparação com os métodos tradicionais é o caso dos tons. Em grande parte dos ritmos da música popular, essa peça é utilizada poucas vezes ao longo da música, sendo geralmente restrita às viradas. Nesse tipo de performance, pelo fato de os tons serem tocados de forma pontual, métodos tradicionais de separação, mesmo que demandem mais tempo para serem implementados, podem ser uma alternativa viável em relação às ferramentas de DSS, considerando a baixa frequência de execuções desses elementos na música.

Dentre os resultados obtidos no trabalho, a separação dos pratos se destaca como um exemplo em que a ferramenta de DSS apresentou um desempenho que, possivelmente, eliminaria a necessidade de utilizar outros métodos de separação. Isso se deve ao fato de que a ferramenta foi capaz de realizar uma tarefa complexa para métodos tradicionais, em pouco tempo e com uma preservação satisfatória do timbre. Além disso, essa separação foi aplicada nos *overheads*, que são os microfones mais suscetíveis a capturar vazamentos de outras peças durante a gravação. Esses resultados evidenciam o potencial promissor dessas ferramentas, especialmente se alcançarem separações de qualidade ainda mais elevada.

Por fim, conclui-se que a escolha entre o uso de uma ferramenta de DSS e um método tradicional está diretamente relacionada à qualidade da separação oferecida pela ferramenta de DSS. Em muitos casos, é preferível preservar a qualidade sonora, mesmo que isso demande mais trabalho, do que optar por uma solução rápida que comprometa o timbre ou mantenha vazamentos indesejados. Assim, o aprimoramento da qualidade na separação é um fator essencial para a consolidação dessas ferramentas no contexto da produção musical, especialmente considerando sua velocidade em comparação aos métodos tradicionais.

6 CONCLUSÃO

Este trabalho oferece uma contribuição relevante para o campo da computação aplicada, especialmente no desenvolvimento de algoritmos de DSS. A aplicação da ferramenta em situações de gravações reais permitiu identificar desafios e limitações em cenários práticos. Essas constatações fornecem subsídios importantes para que pesquisadores possam aprimorar esses algoritmos no futuro.

Conforme discutido no Capítulo 5, diversos fatores interferem na qualidade da separação realizada. Entre esses fatores, destacam-se o estilo da performance, a intensidade com que a peça foi tocada, a técnica utilizada e as outras peças tocadas simultaneamente. Todos esses elementos influenciam diretamente no desempenho da separação, evidenciando a importância de estudá-los para o desenvolvimento de modelos mais robustos voltados para gravações reais.

Este trabalho introduz na literatura a discussão sobre o uso de modelos de DSS em situações reais no cotidiano de profissionais da produção musical e engenharia de áudio. A integração entre os avanços tecnológicos no campo da computação e o cenário de produção artística pode gerar benefícios significativos. Esses avanços têm o potencial de influenciar diretamente a maneira como esses profissionais trabalham e executam suas produções.

A remoção de vazamentos em gravações, abordada neste trabalho, pode ser considerada uma nova área de aplicação dentro dos estudos de DSS. Essa abordagem apresenta desafios específicos, como a ausência de métodos quantitativos para avaliação dos resultados, uma vez que gravações reais não permitem a aplicação de métricas baseadas em dados supervisionados. Como trabalho futuro, propõe-se o desenvolvimento de novas formas de avaliação voltadas para esse tipo de tarefa.

Um dos pontos fundamentais discutidos neste trabalho é que gravações de bateria reais apresentam diferenças significativas em comparação aos *kits* amostrados de instrumentos virtuais. Isso evidencia a necessidade de considerar essas particularidades no desenvolvimento de modelos de DSS. Uma proposta interessante para trabalhos futuros é a ampliação das bases de dados existentes, incorporando sons reais de gravações.

A criação de bases de dados com gravações reais de bateria, em tamanho adequado para treinar DNN, representa um desafio significativo. No entanto, por meio da interdisciplinaridade, é possível desenvolver soluções que facilitem esse processo. Uma proposta interessante seria a criação de um sistema mecatrônico capaz de tocar automaticamente as peças da bateria com base em informações extraídas de arquivos MIDI, permitindo a captura de sons reais de forma automatizada, sem demandar horas de execução de um músico.

Além das gravações com peças reais, dados contendo amostras de vazamentos também podem ser registrados e utilizados para treinar modelos. Isso porque, apesar de constarem de uma mesma fonte sonora, os vazamentos são capturados por microfones diferentes, com

amplitude, fase e espectros diferentes. Treinar modelos com esse tipo de dados pode ser relevante para avanços nas tarefas de separação de fontes sonoras.

Outro ponto evidenciado neste trabalho é que o estilo musical executado influencia diretamente na qualidade da separação proposta. Isso se torna evidente ao analisar os resultados das separações em gravações de estilos distintos. Como cada gênero musical possui suas particularidades, diversificar as performances presentes nas bases de treinamento pode ser um esforço promissor, com potencial para melhorar significativamente o desempenho das ferramentas de DSS.

Acerca dos resultados obtidos com a implementação do algoritmo LarsNet, foram observados tanto resultados promissores quanto resultados insatisfatórios. Devido à quantidade de particularidades e peculiaridades presentes em cada bateria, uma solução interessante seria dividir o modelo em mais classes, permitindo uma identificação sonora mais precisa de determinados elementos. Um exemplo claro é o som do aro da caixa, que possui grande importância em diversos gêneros musicais, mas não é corretamente identificado pelo algoritmo como parte da caixa.

Por fim, este trabalho se dedicou a organizar informações relevantes para essa discussão, abordando os processos que envolvem a produção musical, a descrição detalhada da bateria e seu papel na produção musical e na engenharia de áudio, além de uma breve introdução e contextualização sobre as áreas de ASS e DSS. Dessa forma, o estudo oferece uma base completa para interessados nesses temas, apresentando-os de maneira acessível e com uma abordagem simplificada.

REFERÊNCIAS

ABE, Y.; MURAKAMI, Y.; MIURA, M. Automatic arrangement for the bass guitar in popular music using principle component analysis. **Acoustical Science and Technology**, Acoustical Society of Japan, v. 33, n. 4, p. 229–238, 2012. ISSN 1347-5177. Disponível em: <https://www.jstage.jst.go.jp/article/ast/33/4/33_E1101/_article/-char/ja/>.

ALI, M. A.; DHANARAJ, R. K.; NAYYAR, A. A high performance-oriented ai-enabled iot-based pest detection system using sound analytics in large agricultural field. **Microprocessors and Microsystems**, Elsevier, v. 103, p. 104946, novembro 2023. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0141933123001904>.

AUTOR DESCONHECIDO. **German-American engineer Emile Berliner (1851-1929) with the model of the first phonograph machine which he invented.** entre 1910 e 1929. Domínio público, via Wikimedia Commons. Acessado em 29 de novembro de 2024. Disponível em: <https://commons.wikimedia.org/wiki/File:Emile_Berliner_with_phonograph_(cropped_portrait).jpg>.

BAI, M.; LI, M. A presentation of structures and applications of convolutional neural networks. **Highlights in Science, Engineering and Technology**, Darcy & Roy Press, v. 61, p. 180–187, julho 2023. ISSN 2791-0210. Disponível em: <https://drpress.org/ojs/index.php/HSET/article/view/10291>.

BARCHARD, V. **Ringo Starr performing with the Beatles at the Gator Bowl in Jacksonville, 1964**. 1964. Domínio público, via Wikimedia Commons. Acessado em 29 de novembro de 2024. Disponível em: <https://commons.wikimedia.org/wiki/File:Ringo_Starr_drumming.jpg>.

BATALHA, J. C. R. et al. Music therapy and its effects in the hospital environment. **Research, Society and Development**, CDRR Editors, v. 11, n. 6, p. e12411626747, abril 2022. ISSN 2525-3409. Disponível em: <https://rsdjournal.org/index.php/rsd/article/view/26747>.

BORATTO, T. H. et al. Data-driven cymbal bronze alloy identification via evolutionary machine learning with automatic feature selection. **Journal of Intelligent Manufacturing**, Springer, v. 35, p. 257–273, janeiro 2024. Disponível em: <https://link.springer.com/article/10.1007/s10845-022-02047-3>.

BORATTO, T. H. d. A. **Aprendizado de máquina para a classificação automática de pratos de bateria conforme a proporção de estanho presente em suas ligas de bronze**. 104 p. Dissertação (Mestrado) — Universidade Federal de Juiz de Fora, Juiz de Fora, 2022. Disponível em: <https://repositorio.ufjf.br/jspui/handle/ufjf/15052>.

BORATTO, T. H. d. A. et al. Análise dos efeitos de dois diferentes métodos de fabricação em pratos de bateria. In: **Proceedings of the 17th Brazilian Symposium on Computer Music**. Curitiba, PR, Brasil: Associação Brasileira de Engenharia e Ciências Mecânicas, 2021. p. 1–8. Disponível em: <https://www.researchgate.net/publication/352266969>.

BORATTO, T. H. de A. et al. Recuperação de informação musical como ferramenta para análise sonora: uma inicialização. In: **Anais do FIA 2020/22 | XXIX Sobrac**. Florianópolis, SC, Brasil: ABRAC, 2022. p. 1–10. Disponível em: <https://alice.ufsj.edu.br//papers/2022Boratto.pdf>.

BRANDÃO, E. Fundamentos. In: _____. **Acústica de salas: projeto e modelagem**. 1. ed. São Paulo, Brasil: Blucher, 2018. cap. 1, p. 57–112. ISBN 978-85-212-1006-1.

BROWN, A. Digital technology and the study of music. **International Journal of Music Education**, Sage Publications, os-25, n. 1, p. 14–19, maio 1995. ISSN 1744-795X. Disponível em: <https://journals.sagepub.com/doi/10.1177/025576149502500102>.

CALLENDER, L.; HAWTHORNE, C.; ENGEL, J. **Improving Perceptual Quality of Drum Transcription with the Expanded Groove MIDI Dataset**. 2020. Disponível em: <https://arxiv.org/abs/2004.00188>.

CALYSO, S. **Otari 2"machine**. 2014. CC BY-SA 3.0, via Wikimedia Commons. Acessado em 29 de novembro de 2024. Disponível em: <https://commons.wikimedia.org/wiki/File:Otari_MX-80.JPG>.

CANO, E. et al. Musical source separation: An introduction. **IEEE Signal Processing Magazine**, IEEE, v. 36, n. 1, p. 31–40, janeiro 2018. ISSN 1558-0792. Disponível em: <https://ieeexplore.ieee.org/document/8588410>.

CANUDO, R. **Manifesto of the Seven Arts**. [S.l.: s.n.], 1911.

CARVALHO, L. R.; PEREIRA, A. T. C. Mixagem de som na interface. In: **Anais do 16° USIHC – Congresso Internacional de Ergonomia e Usabilidade de Interfaces Humano Computador**. São Paulo, Brasil: Blucher, 2017. v. 3, n. 11, p. 2156 – 2164. ISSN 2318-6968. Disponível em: <www.proceedings.blucher.com.br/article-details/mixagem-de-som-na-interface-25881>.

CENTURY DICTIONARY. **Phonoautograph**. 1891. Domínio público, via Wikimedia Commons. Acessado em 29 de novembro de 2024. Disponível em: <https://commons.wikimedia.org/wiki/File:Phonautograph-cent2.png>.

COSTA, A.; CATALAN, L. B. O emergir da música popular e suas interfaces com a indústria fonográfica. **Caderno CRH**, Universidade Federal da Bahia, v. 32, n. 87, p. 517–535, dezembro 2019. ISSN 1983-8239. Disponível em: <https://periodicos.ufba.br/index.php/crh/article/view/32241>.

COUTURIER, G.; DAIGLE, M. **Masque de Fer: Extended Drum Kit Techniques**. 2022. Timbre and Orchestration Resource. Acessado em 07 de outubro de 2024. Disponível em: <https://www.timbreandorchestrationresource.org/project-reports/masque-de-fer>.

DITTMAR, C.; GÄRTNER, D. Real-time transcription and separation of drum recordings based on nmf decomposition. In: **DAFx**. Erlangen, Alemanha: [s.n.], 2014. v. 17, p. 187–194. Disponível em: <https://www.dafx14.fau.de/proceedings.html>.

DOWNIE, J. S. Music information retrieval. In: _____. **Annual review of information science and technology**. Medford, Estados Unidos: Information Today Books, 2003. v. 37, p. 295–340. Disponível em: <https://www.music.mcgill.ca/~ich/classes/mumt611_06/downie_mir_arist37.pdf>.

DÉFOSSEZ, A. et al. **Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed**. 2019. Disponível em: <https://arxiv.org/abs/1909.01174>.

DéFOSSEZ, A. **Hybrid Spectrogram and Waveform Source Separation**. 2022. Disponível em: <https://arxiv.org/abs/2111.03600>.

FABBRO, G. et al. The sound demixing challenge 2023 – music demixing track. **Transactions of the International Society for Music Information Retrieval**, ISMIR, v. 7, n. 1, p. 63–84, abril 202. ISSN 2514-3298. Disponível em: <https://transactions.ismir.net/articles/10.5334/tismir.171>.

FLECK, L. et al. Redes neurais artificiais: princípios básicos. **Revista Eletrônica Científica Inovação e Tecnologia**, Universidade Tecnológica Federal do Paraná, Medianeira, PR, Brasil, v. 7, n. 15, p. 47–57, janeiro/junho 2016. ISSN 2175-1846. Disponível em: <https://periodicos.utfpr.edu.br/recit/article/view/4330>.

FRANGI, E. **Neil Peart in concert with Rush. Milan, Italy (September 21, 2004)**. 2004. Domínio público, via Wikimedia Commons. Acessado em 29 de novembro de 2024. Disponível em: <https://commons.wikimedia.org/wiki/File:Neil-Peart.jpg>.

FUJIHARA, H. et al. Automatic synchronization between lyrics and music cd recordings based on viterbi alignment of segregated vocal signals. In: **Eighth IEEE International Symposium on Multimedia (ISM'06)**. São Diego, Estados Unidos: IEEE, 2006. p. 257–264. ISBN 0-7695-2746-9. Disponível em: <https://ieeexplore.ieee.org/abstract/document/4061176>.

GIBSON, D. The best of the colors visuals. In: _____. **The art of mixing: a visual guide to recording, engineering, and production**. 3. ed. Nova York, Estados Unidos e Abingdon, Inglaterra: Routledge, 2019. p. 5–17. ISBN 978–1–135–125221–8.

GIBSON, D. Visual representations of "imaging". In: _____. **The art of mixing: a visual guide to recording, engineering, and production**. 3. ed. Nova York, Estados Unidos e abingdon, Inglaterra: Routledge, 2019. cap. 2, p. 37–62. ISBN 978–1–135–125221–8.

GILLET, O.; RICHARD, G. Enst-drums: an extensive audio-visual database for drum signals processing. In: **International Society for Music Information Retrieval Conference (ISMIR)**. [s.n.], 2006. Disponível em: <https://archives.ismir.net/ismir2006/paper/000027.pdf>.

GILLICK, J. et al. Learning to groove with inverse sequence transformations. In: **International conference on machine learning**. Long Beach, Estados Unidos: PMLR, 2019. p. 2269–2279. ISSN 2640-3498. Disponível em: <https://proceedings.mlr.press/v97/gillick19a/gillick19a.pdf>.

HANDY, L. C. **Thomas Edison and his early phonograph. Cropped from Library of Congress copy.** 1878. Domínio público, via Wikimedia Commons. Acessado em 29 de novembro de 2024. Disponível em: <https://commons.wikimedia.org/wiki/File:Edison_and_phonograph_edit1.jpg>.

HARTIGAN, R. The heritage of the drumset. **African American Review**, JSTOR, v. 29, n. 2, p. 234–236, 1995. Disponível em: <https://www.jstor.org/stable/3042298>.

HAUNSCHMID, V.; MANILOW, E.; WIDMER, G. **audioLIME: Listenable Explanations Using Source Separation**. 2020. Disponível em: <https://arxiv.org/abs/2008.00582>.

HAUNSCHMID, V.; MANILOW, E.; WIDMER, G. **Towards Musically Meaningful Explanations Using Source Separation**. 2020. Disponível em: <https://arxiv.org/abs/2009.02051>.

HEITTOLA, T.; KLAPURI, A.; VIRTANEN, T. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In: **10th International Society for Music Information Retrieval Conference**. Kobe, Japão: ISMIR, 2009. p. 327–332. Disponível em: <https://ismir2009.ismir.net/proceedings/OS3-2.pdf>.

HENNEQUIN, R. et al. Spleeter: a fast and efficient music source separation tool with pre-trained models. **Journal of Open Source Software**, v. 5, n. 50, p. 2154, 2020. Disponível em: <https://www.theoj.org/joss-papers/joss.02154/10.21105.joss.02154.pdf>.

HSU, M.-H. et al. Spider king: Virtual musical instruments based on microsoft kinect. In: **2013 International Joint Conference on Awareness Science and Technology & Ubi-Media Computing (iCAST 2013 & UMEDIA 2013)**. Aizu-Wakamatsu, Japão: IEEE, 2013. p. 707–712. ISBN 978-1-4799-2364-9. Disponível em: <https://ieeexplore.ieee.org/document/6765529>.

HU, Y.; LIU, G. Separation of singing voice using nonnegative matrix partial co-factorization for singer identification. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, IEEE, v. 23, n. 4, p. 643–653, janeiro 2015. ISSN 2329-9304. Disponível em: <https://ieeexplore.ieee.org/document/7021947>.

HUBER, D. M.; RUNSTEIN, R. E. The analog tape recorder. In: ____. **Modern recording techniques**. 9. ed. Nova York, Estados Unidos e Londres, Inglaterra: Routledge, 2018. cap. 5, p. 175–193. ISBN 978–1–315–66695–2.

HUBER, D. M.; RUNSTEIN, R. E. Digital audio technology. In: ____. **Modern recording techniques**. 9. ed. Nova York, Estados Unidos e Londres, Inglaterra: Routledge, 2018. cap. 6, p. 195–217. ISBN 978–1–315–66695–2.

HUBER, D. M.; RUNSTEIN, R. E. The digital audio workstation. In: ____. **Modern recording techniques**. 9. ed. Nova York, Estados Unidos e Londres, Inglaterra: Routledge, 2018. cap. 7, p. 219–264. ISBN 978–1–315–66695–2.

HUBER, D. M.; RUNSTEIN, R. E. Effects processing. In: ____. **Modern recording techniques**. 9. ed. Nova York, Estados Unidos e Londres, Inglaterra: Routledge, 2018. cap. 15, p. 403–441. ISBN 978–1–315–66695–2.

HUBER, D. M.; RUNSTEIN, R. E. Introduction. In: ____. **Modern recording techniques**. 9. ed. Nova York, Estados Unidos e Londres, Inglaterra: Routledge, 2018. cap. 1, p. 1–41. ISBN 978–1–315–66695–2.

HUBER, D. M.; RUNSTEIN, R. E. Mastering. In: ____. **Modern recording techniques**. 9. ed. Nova York, Estados Unidos e Londres, Inglaterra: Routledge, 2018. cap. 20, p. 533–552. ISBN 978–1–315–66695–2.

HUBER, D. M.; RUNSTEIN, R. E. Microphones: Design and applications. In: ____. **Modern recording techniques**. 9. ed. Nova York, Estados Unidos e Londres, Inglaterra: Routledge, 2018. cap. 4, p. 105–173. ISBN 978–1–315–66695–2.

HUBER, D. M.; RUNSTEIN, R. E. Multimedia and the web. In: ____. **Modern recording techniques**. 9. ed. Nova York, Estados Unidos e Londres, Inglaterra: Routledge, 2018. cap. 11, p. 349–370. ISBN 978–1–315–66695–2.

HUBER, D. M.; RUNSTEIN, R. E. Sound and hearing. In: ____. **Modern recording techniques**. 9. ed. Nova York, Estados Unidos e Londres, Inglaterra: Routledge, 2018. cap. 2, p. 43–74. ISBN 978–1–315–66695–2.

HUBER, D. M.; RUNSTEIN, R. E. Studio acoustics and design. In: ____. **Modern recording techniques**. 9. ed. Nova York, Estados Unidos e Londres, Inglaterra: Routledge, 2018. cap. 3, p. 75–103. ISBN 978–1–315–66695–2.

IZHAKI, R. Automation. In: ____. **Mixing audio: concepts, practices, and tools**. 3. ed. Nova York, Estados Unidos e Abingdon, Inglaterra: Routledge, 2017. cap. 28, p. 474–481. ISBN 978-1-315-71694-7.

IZHAKI, R. Compressors. In: ____. **Mixing audio: concepts, practices, and tools**. 3. ed. Nova York, Estados Unidos e Abingdon, Inglaterra: Routledge, 2017. cap. 17, p. 274–336. ISBN 978-1-315-71694-7.

IZHAKI, R. Delays. In: ____. **Mixing audio: concepts, practices, and tools**. 3. ed. Nova York, Estados Unidos e Abingdon, Inglaterra: Routledge, 2017. cap. 22, p. 381–397. ISBN 978-1-315-71694-7.

IZHAKI, R. Distortion. In: ____. **Mixing audio: concepts, practices, and tools**. 3. ed. Nova York, Estados Unidos e Abingdon, Inglaterra: Routledge, 2017. cap. 25, p. 451–459. ISBN 978-1-315-71694-7.

IZHAKI, R. Drum triggering. In: ____. **Mixing audio: concepts, practices, and tools**. 3. ed. Nova York, Estados Unidos e Abingdon, Inglaterra: Routledge, 2017. cap. 26, p. 460–464. ISBN 978-1-315-71694-7.

IZHAKI, R. Equalizers. In: ____. **Mixing audio: concepts, practices, and tools**. 3. ed. Nova York, Estados Unidos e Abingdon, Inglaterra: Routledge, 2017. cap. 15, p. 210–265. ISBN 978-1-315-71694-7.

IZHAKI, R. Introduction to dynamic range processors. In: ____. **Mixing audio: concepts, practices, and tools**. 3. ed. Nova York, Estados Unidos e Abingdon, Inglaterra: Routledge, 2017. cap. 16, p. 266–273. ISBN 978-1-315-71694-7.

IZHAKI, R. Mixing domains and objectives. In: ____. **Mixing audio: concepts, practices, and tools**. 3. ed. Nova York, Estados Unidos e Abingdon, Inglaterra: Routledge, 2017. cap. 7, p. 66–81. ISBN 978-1-315-71694-7.

IZHAKI, R. Music and mixing. In: ____. **Mixing audio: concepts, practices, and tools**. 3. ed. Nova York, Estados Unidos e Abingdon, Inglaterra: Routledge, 2017. cap. 1, p. 7–13. ISBN 978-1-315-71694-7.

IZHAKI, R. Other modulation tools. In: ____. **Mixing audio: concepts, practices, and tools**. 3. ed. Nova York, Estados Unidos e Abingdon, Inglaterra: Routledge, 2017. cap. 23, p. 398–405. ISBN 978-1-315-71694-7.

IZHAKI, R. Panning. In: ____. **Mixing audio: concepts, practices, and tools**. 3. ed. Nova York, Estados Unidos e Abingdon, Inglaterra: Routledge, 2017. cap. 14, p. 190–209. ISBN 978-1-315-71694-7.

IZHAKI, R. Related issues. In: ____. **Mixing audio: concepts, practices, and tools**. 3. ed. Nova York, Estados Unidos e Abingdon, Inglaterra: Routledge, 2017. cap. 5, p. 45–52. ISBN 978-1-315-71694-7.

IZHAKI, R. Reverbs. In: ____. **Mixing audio: concepts, practices, and tools**. 3. ed. Nova York, Estados Unidos e Abingdon, Inglaterra: Routledge, 2017. cap. 24, p. 406–450. ISBN 978-1-315-71694-7.

JANSSON, A. et al. Joint singing voice separation and f0 estimation with deep u-net architectures. In: **27th European Signal Processing Conference (EUSIPCO)**. IEEE, 2019. p. 1–5. ISBN 978-9-0827-9703-9. ISSN 2076-1465. Disponível em: <https://ieeexplore.ieee.org/document/8902550>.

JANSSON, A. et al. Singing voice separation with deep u-net convolutional networks. In: **Proceedings of the 18th ISMIR Conference**. Suzhou, China: ISMIR, 2017. v. 18, p. 745–751. Disponível em: <https://archives.ismir.net/ismir2017/paper/000171.pdf>.

JOHNSON, H. Chinese toms in the making of the drum kit: Localization and exoticism. **Journal of Popular Music Education**, Intellect Ltd, v. 5, n. 2, p. 227–242, julho 2021. ISSN 2397-673X. Disponível em: <https://intellectdiscover.com/content/journals/10.1386/jpme_00059_1>.

KINGMA, D. P.; BA, J. **Adam: A Method for Stochastic Optimization**. 2017. Disponível em: <https://arxiv.org/abs/1412.6980>.

LAROCHE, C. et al. A structured nonnegative matrix factorization for source separation. In: **2015 23rd European Signal Processing Conference (EUSIPCO)**. Nice, França: IEEE, 2015. p. 2033–2037. ISBN 978-0-9928-6263-3. ISSN 2076-1465. Disponível em: <https://ieeexplore.ieee.org/document/7362741>.

LEE, D. Hornbostel-sachs classification of musical instruments. **Knowledge Organization**, Nomos Verlagsgesellschaft, v. 47, n. 1, p. 72–91, 2019. Disponível em: <https://openaccess.city.ac.uk/id/eprint/22554/>.

LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. **Nature**, Nature Publishing Group, Londres, Reino Unido, v. 401, n. 6755, p. 788–791, outubro 1999. ISSN 1476-4687. Disponível em: <https://www.nature.com/articles/44565>.

LI, S. et al. Improving drum source separation with temporal-frequency statistical descriptors. In: **2024 IEEE International Conference on Multimedia and Expo (ICME)**. Niagara Falls, Canada: IEEE, 2024. p. 1–6. ISBN 979-8-3503-9015-5. Disponível em: <https://www.computer.org/csdl/proceedings-article/icme/2024/10688211/20F09cS5nb2>.

LIUTKUS, A.; STöTER, F.-R. **sigsep/norbert: First official Norbert release**. Zenodo, 2019. Disponível em: <https://doi.org/10.5281/zenodo.3269749>.

LU, W.-T. et al. **Music Source Separation with Band-Split RoPE Transformer**. 2023. Disponível em: <https://arxiv.org/abs/2309.02612>.

LUO, Y.; MESGARANI, N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 27, n. 8, p. 1256–1266, agosto 2019. ISSN 2329-9290. Disponível em: <https://dl.acm.org/doi/10.1109/TASLP.2019.2915167>.

MACEDO, F. A. B. O processo de produção musical na indústria fonográfica: questões técnicas e musicais envolvidas no processo de produção musical em estúdio. **Revista eletrônica de musicologia**, Universidade Federal do Paraná, XI, p. 1–7, setembro 2007. ISSN 1415-952X. Disponível em: <http://www.rem.ufpr.br/_REM/REMv11/12/12-macedo-gravacao.html>.

MAISON DE LA BONNE PRESSE. **A Columbia type AT Graphophone, shown in a 1901 catalog published by the Maison de la Bonne Presse.** 1901. Domínio público, via Wikimedia Commons. Acessado em 29 de novembro de 2024. Disponível em: <https://commons.wikimedia.org/wiki/File:Graphophone1901.jpg>.

MALI, S. G.; MAHAJAN, S. P. Blind sound source separation by combining the convolutional neural network and degree separator. **Traitement du Signal**, International Information and Engineering Technology Association, v. 41, n. 3, p. 1429–1439, junho 2024. ISSN 1958-5608. Disponível em: <https://www.iieta.org/journals/ts/paper/10.18280/ts.410331>.

MANILOW, E.; SEETHARAMAN, P.; PARDO, B. The northwestern university source separation library. In: **Proceedings of the 19th ISMIR Conference**. Paris, França: ISMIR, 2018. p. 297–305. Disponível em: <https://pseeth.github.io/public/papers/manilow_seetharaman_ismir18.pdf>.

MANILOW, E.; SEETHARAMAN, P.; PARDO, B. Simultaneous separation and transcription of mixtures with multiple polyphonic and percussive instruments. In: **ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. Barcelona, Espanha: IEEE, 2020. p. 771–775. ISBN 978-1-5090-6631-5. ISSN 2379-190X. Disponível em: <https://ieeexplore.ieee.org/document/9054340>.

MANILOW, E.; SEETHARMAN, P.; SALAMON, J. **Open Source Tools & Data for Music Source Separation**. https://source-separation.github.io/tutorial, 2020. Disponível em: <https://source-separation.github.io/tutorial>.

MASSEY, H.; NOYES, A.; SHKLAIR, D. **A Synthesist's Guide to Acoustic Instruments**. Nova York, Londes e Sydney: Amsco Publications, 1987. ISBN 0.7119.1124.X.

MAZZOLA, G. et al. A short history of midi. In: _____. **Basic Music Technology: An Introduction**. Springer International Publishing, 2018. p. 115–116. ISBN 978-3-030-00982-3, 978-3-030-00982-3. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-030-00982-3_11>.

MCGUIRE, J.; KAPLAN, J.; KAPLAN, K. Virtual orchestras: Engineering innovation and musicians collide. In: **2005 Annual Conference**. Portland, Oregon: ASEE Conferences, 2005. p. 10.1453.1 – 10.1453.7. ISSN 2153-5965. Https://peer.asee.org/15163. Disponível em: <https://peer.asee.org/15163>.

MEDEIROS, R. et al. Challenges in designing new interfaces for musical expression. In: **Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience: Third International Conference, DUXU 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part I 3**. Springer, 2014. p. 643–652. ISBN 978-3-319-07667-6. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-319-07668-3_62>.

MESAROS, A.; VIRTANEN, T. Automatic recognition of lyrics in singing. **EURASIP Journal on Audio, Speech, and Music Processing**, Springer, p. 1–11, fevereiro 2010. ISSN 1687-4722. Disponível em: <https://asmp-eurasipjournals.springeropen.com/articles/10.1155/2010/546047>.

MEZZA, A. I. et al. Benchmarking music demixing models for deep drum source separation. In: **2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)**. Erlangen, Alemanha: IEEE, 2024. p. 1–6. Disponível em: <https://ieeexplore.ieee.org/document/10704147>.

MEZZA, A. I. et al. Toward deep drum source separation. **Pattern Recognition Letters**, Elsevier, v. 183, p. 86–91, julho 2024. ISSN 0167-8655. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167865524001351>.

MEZZA, A. I. et al. Lars: An open-source vst3 plug-in for deep drums demixing with pretrained models. In: **Ismir 2023 Hybrid Conference**. Milão, Itália: ISMIR, 2023. Disponível em: <https://ismir2023program.ismir.net/lbd_349.html>.

MILEHAM, R. Sounds of history [audio recording]. **Engineering & Technology**, Institution of Engineering and Technology, v. 4, n. 18, p. 22–24, outubro 2009. ISSN 1750-9637. Disponível em: <https://digital-library.theiet.org/content/journals/10.1049/et.2009.1802>.

MOFFAT, D.; SANDLER, M. B. Approaches in intelligent music production. **Arts**, MDPI, v. 8, n. 4, p. 125, setembro 2019. ISSN 2076-0752. Disponível em: <https://www.mdpi.com/2076-0752/8/4/125>.

MOUAWAD, P.; DUBNOV, T.; DUBNOV, S. Robust detection of covid-19 in cough sounds: using recurrence dynamics and variable markov model. **SN Computer Science**, Springer, v. 2, n. 1, p. 34, janeiro 2021. Disponível em: <https://link.springer.com/article/10.1007/s42979-020-00422-6>.

MÉNDEZ, P. E. La transformación digital en la producción musical de bandas sonoras en españa. **Cuadernos de Investigación Musical**, Universidad de Castilla-La Mancha, n. 15, p. 171–185, maio 2022. ISSN 2530-6847. Disponível em: <https://revista.uclm.es/index.php/cuadernosdeinvestigacionmusical/article/view/2863>.

MÜLLER, M. Fourier analysis of signals. In: _____. **Fundamentals of music processing: Using Python and Jupyter notebooks**. 2. ed. Erlangen, Alemanha: Springer, 2021. cap. 2, p. 39–117. ISBN 978-3-030-69808-9.

MÜLLER, M. Music representations. In: _____. **Fundamentals of music processing: Using Python and Jupyter notebooks**. 2. ed. Erlangen, Alemanha: Springer, 2021. cap. 1, p. 1–38. ISBN 978-3-030-69808-9.

OLIVEIRA, J. P. M. **Análise comparativa entre um compressor de áudio analógico e seus simuladores digitais**. Monografia (Trabalho de Conclusão de Curso), Juiz de Fora, 2021. Disponível em: <https://www.researchgate.net/publication/363694427_Analise_comparativa_entre_um_compressor_de_audio_analogico_e_seus_simuladores_digitais>.

OPPENHEIM, A. V.; WILLSKY, A. S.; NAWAB, S. H. A transformada de fourier de tempo contínuo. In: _____. **Sinais e sistemas**. 2. ed. São Paulo, Brasil: Pearson Education do Brasil, 2010. cap. 4, p. 165–206. ISBN 978-85-4301-380-0.

PAIXÃO, L. F. da. **A indústria fonográfica como mediadora entre a música e a sociedade**. 104 p. Dissertação (Mestrado) — Universidade Federal do Paraná, Curitiba, 2013. Disponível em: <https://acervodigital.ufpr.br/handle/1884/30351>.

PLUMBLEY, M. D. et al. Automatic music transcription and audio source separation. **Cybernetics and Systems**, Taylor & Francis, v. 33, n. 6, p. 603–627, 2002. Disponível em: <https://doi.org/10.1080/01969720290040777>.

PRAS, A.; GUASTAVINO, C.; LAVOIE, M. The impact of technological advances on recording studio practices. **Journal of the American Society for Information Science and Technology**, Wiley Online Library, v. 64, n. 3, p. 612–626, janeiro 2013. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/asi.22840>.

RAFII, Z. et al. **MUSDB18 - a corpus for music separation**. Zenodo, 2017. Disponível em: <https://zenodo.org/records/1117372>.

RAFII, Z. et al. **MUSDB18-HQ - an uncompressed version of MUSDB18**. Zenodo, 2019. Disponível em: <https://zenodo.org/records/3338373>.

RASCHKA, A. **Sepultura at Vainstream Rockfest 2014.** 2014. CC BY-SA 3.0, via Wikimedia Commons. Acessado em 29 de novembro de 2024. Disponível em: <https://commons.wikimedia.org/wiki/File:2014-07-05_Vainstream_Sepultura_Eloy_Casagrande_02.jpg>.

RATNAPARKHI, A. A.; PILLI, E.; JOSHI, R. C. Survey of scaling platforms for deep neural networks. In: **2016 International Conference on Emerging Trends in Communication Technologies (ETCT)**. Dehradun, Índia: IEEE, 2016. p. 1–6. ISBN 978-1-5090-4505-1. Disponível em: <https://ieeexplore.ieee.org/document/7882969>.

ROCHA, G. L.; TEIXEIRA, J.; SCHIAVONI, F. Ha dou ken music: Mapping a joysticks as a musical controller. In: SCHIAVONI, F. et al. (Ed.). **Proceedings of the 17th Brazilian Symposium on Computer Music**. São João del-Rei, MG, Brasil: Sociedade Brasileira de Computação, 2019. p. 69–75. ISBN 978-85-7669-490-8. ISSN 2175-6759. Disponível em: <https://compmus.ime.usp.br/sbcm/2019/assets/proceedings.pdf>.

ROMA, G. et al. Untwist: A new toolbox for audio source separation. In: **Late-Breaking Demo Session of the 17th International Society for Music Information Retrieval Conference**. Nova York, Estados Unidos: ISMIR, 2016. p. 7–11. Disponível em: <https://www.researchgate.net/publication/308054917>.

RONNEBERGER, O.; FISCHER, P.; BROX, T. **U-Net: Convolutional Networks for Biomedical Image Segmentation**. 2015. Disponível em: <https://arxiv.org/abs/1505.04597>.

ROUARD, S.; MASSA, F.; DéFOSSEZ, A. **Hybrid Transformers for Music Source Separation**. 2022. Disponível em: <https://arxiv.org/abs/2211.08553>.

RUSS, M. Background. In: _____. **Sound synthesis and sampling**. 3. ed. Oxford, Reino Unido e Burlington, Estados Unidos: Elsevier, 2012. cap. 1. ISBN 978-0-240-52105-3.

RUSS, M. Controllers. In: _____. **Sound synthesis and sampling**. 3. ed. Oxford, Reino Unido e Burlington, Estados Unidos: Elsevier, 2012. cap. 8. ISBN 978-0-240-52105-3.

RUSS, M. Sound-making techniques. In: _____. **Sound synthesis and sampling**. 3. ed. Oxford, Reino Unido e Burlington, Estados Unidos: Elsevier, 2012. cap. 7. ISBN 978-0-240-52105-3.

SAKIB, S. et al. An overview of convolutional neural network: Its architecture and applications. **Preprints**, MDPI, fevereiro 2019. Disponível em: <https://www.preprints.org/manuscript/201811.0546>.

SALAÜN, Y. et al. The flexible audio source separation toolbox version 2.0. In: **ICASSP**. Florença, Itália: Hal Open Science, 2014. p. https://inria.hal.science/hal–00957412v1/. Disponível em: <https://inria.hal.science/hal-00957412v1/>.

SAMUEL, D.; GANESHAN, A.; NARADOWSKY, J. Meta-learning extractors for music source separation. In: **ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. Barcelona, Espanha: IEEE, 2020. v. 18, p. 816–820. ISBN 978-1-5090-6631-5. ISSN 2379-190X. Disponível em: <https://ieeexplore.ieee.org/document/9053513>.

SAVAGE, S. Building a mix: The concepts and tools in detail. In: _____. **Mixing and mastering in the box: the guide to making great mixes and final masters on your computer**. Oxford, Inglaterra e Nova York, Estados Unidos: Oxford University Press, 2014. cap. 4, p. 63–122. ISBN 978-0-19-992930-6.

SAVAGE, S. Delivering mixes: Formats, mix types, and multiple mixes. In: _____. **Mixing and mastering in the box: the guide to making great mixes and final masters on your computer**. Oxford, Inglaterra e Nova York, Estados Unidos: Oxford University Press, 2014. cap. 8, p. 186–194. ISBN 978-0-19-992930-6.

SAVAGE, S. Mixing and mastering. In: _____. **Mixing and mastering in the box: the guide to making great mixes and final masters on your computer**. Oxford, Inglaterra e Nova York, Estados Unidos: Oxford University Press, 2014. p. 1–4. ISBN 978-0-19-992930-6.

SAWATA, R. et al. All for one and one for all: Improving music separation by bridging networks. In: **ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. Toronto, Canadá: IEEE, 2021. p. 51–55. ISBN 978-1-7281-7605-5. ISSN 2379-190X. Disponível em: <https://ieeexplore.ieee.org/document/9414044>.

SCHEDL, M. et al. **Music information retrieval: Recent developments and applications**. [S.l.]: Now Foundations and Trends, 2014. 152 p. ISBN 978-1-6019-8807-2.

SCHIAVONI, F. L.; BIANCHI, A. J.; QUEIROZ, M. Ferramentas livres para distribuição de áudio em rede. **Cadernos de Informática**, Universidade Federal do Rio Grande do Sul, v. 8, n. 2, p. 1–10, março 2014. ISSN 1519-132X. Disponível em: <https://seer.ufrgs.br/index.php/cadernosdeinformatica/article/view/v8n2p01-10>.

SENIOR, M. Stereo enhancements. In: _____. **Mixing secrets for the small studio**. Oxford, Inglaterra e Nova York, Estados Unidos: Routledge, 2018. cap. 18, p. 305–320. ISBN 978-0-19-992930-6.

SHARMA, B.; DAS, R. K.; LI, H. On the importance of audio-source separation for singer identification in polyphonic music. In: **Interspeech**. Graz, Áustria: ISCA, 2019. p. 2020–2024. ISSN 2958-1796. Disponível em: <https://www.isca-archive.org/interspeech_2019/sharma19d_interspeech.html>.

SIGSEP. **SigSep - Open Resources for Music Source Separation**. 2019. Acesso em 29/11/2024. Disponível em: <https://sigsep.github.io/>.

SMITH, A. **Marsha & Adam's Humble Studio**. 2010. CC BY-SA 2.0, via Wikimedia Commons. Acessado em 29 de novembro de 2024. Disponível em: <https://commons.wikimedia.org/wiki/File:Marsha_%26_Adam%27s_Humble_Studio.jpg>.

SOUTHALL, C. et al. Mdb drums: An annotated subset of medleydb for automatic drum transcription. In: **International Society for Music Information Retrieval Conference**. Suzhou, China: [s.n.], 2017. Disponível em: <https://www.open-access.bcu.ac.uk/6179/1/Southall2017a.pdf>.

STOLLER, D.; EWERT, S.; DIXON, S. Jointly detecting and separating singing voice: A multi-task approach. In: **International Conference on Latent Variable Analysis and Signal Separation**. Springer, 2018. v. 1, p. 329–339. ISBN 978-3-319-93764-9. ISSN 1611-3349. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-319-93764-9_31>.

STÖTER, F.-R. et al. Open-unmix-a reference implementation for music source separation. **Journal of Open Source Software**, v. 4, n. 41, p. 1667, setembro 2019. ISSN 2475-9066. Disponível em: <https://joss.theoj.org/papers/10.21105/joss.01667>.

STÖTER, F.-R.; LIUTKUS, A. **museval 0.3.0**. Zenodo, 2019. Disponível em: <https://zenodo.org/records/3376621>.

TARUSKIN, R. The curtain goes up. In: **The Oxford History of Western Music**. Oxford University Press, 2010. cap. 1. Disponível em: <https://www.oxfordwesternmusic.com/view/Volume1/actrade-9780195384819-div1-001011.xml>.

TEN, J. **Studio Main Floor at Crescente Studio**. 2013. CC BY-SA 3.0, via Wikimedia Commons. Acessado em 29 de novembro de 2024. Disponível em: <https://commons.wikimedia.org/wiki/File:Studio_Main_Floor_at_Crescente_Studio.jpg>.

TUCCONI, M. **ARC Studio Kontrollraum**. 2018. CC BY-SA 4.0, via Wikimedia Commons. Acessado em 29 de novembro de 2024. Disponível em: <https://commons.wikimedia.org/wiki/File:ARC_Studio_Control_Room_(side_view_2).jpg>.

UHLICH, S.; GIRON, F.; MITSUFUJI, Y. Deep neural network based instrument extraction from music. In: **2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. South Brisbane, Austrália: IEEE, 2015. p. 2135–2139. ISBN 978-1-4673-6997-8. ISSN 2379-190X. Disponível em: <https://ieeexplore.ieee.org/document/7178348>.

VINCENT, E.; GRIBONVAL, R.; FEVOTTE, C. Performance measurement in blind audio source separation. **IEEE Transactions on Audio, Speech, and Language Processing**, v. 14, n. 4, p. 1462–1469, 2006. ISSN 1558-7924. Disponível em: <https://ieeexplore.ieee.org/document/1643671>.

VINCENT, E. et al. First stereo audio source separation evaluation campaign: data, algorithms and results. In: DAVIES, M. E. et al. (Ed.). **International Conference on Independent Component Analysis and Signal Separation**. Berlin, Heidelberg: Springer, 2007. p. 552–559. ISBN 978-3-540-74494-8. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-540-74494-8_69>.

VOGL, R. et al. Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks. In: **Proceedings of the 18th ISMIR Conference**. Suzhou, China: ISMIR, 2017. v. 18, p. 150–157. Disponível em: <https://archives.ismir.net/ismir2017/paper/000123.pdf>.

VOGL, R.; WIDMER, G.; KNEES, P. **Towards multi-instrument drum transcription**. 2018. Disponível em: <https://arxiv.org/abs/1806.06676>.

WARD, D. et al. Sisec 2018: State of the art in musical audio source separation-subjective selection of the best algorithm. In: **WIMP: Workshop on Intelligent Music Production**. Huddersfield, Reino Unido: Hal Open Science, 2018. p. hal–01945362f. Disponível em: <https://inria.hal.science/hal-01945362>.

WEBER, F. et al. **IDMT-SMT-Drums Dataset**. Zenodo, 2023. Disponível em: <https://doi.org/10.5281/zenodo.7544164>.

WENINGER, F.; LEHMANN, A.; SCHULLER, B. Openblissart: Design and evaluation of a research toolkit for blind source separation in audio recognition tasks. In: **2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. Praga, República Tcheca: IEEE, 2011. p. 1625–1628. ISBN 978-1-4577-0537-3. ISSN 1520-6149. Disponível em: <https://ieeexplore.ieee.org/abstract/document/5946809>.

WENINGER, F.; WÖLLMER, M.; SCHULLER, B. Automatic assessment of singer traits in popular music: Gender, age, height and race. In: **12th International Society for Music Information Retrieval Conference**. Miami, Estados Unidos: ISMIR, 2011. p. 37–42. Disponível em: <https://www.researchgate.net/publication/224929629_Automatic_Assessment_of_Singer_Traits_in_Popular_Music_Gender_Age_Height_and_Race>.

WERNER, K. **Caracterização de aspectos do timbre de pratos de percussão através de análises psicoacústicas**. 201 p. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, Florianópolis, 2015. Disponível em: <https://repositorio.ufsc.br/handle/123456789/169649>.

WILMERING, T. et al. A history of audio effects. **Applied Sciences**, MDPI, v. 10, n. 3, p. 791, janeiro 2020. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/10/3/791>.

WILMUT, R. **BBC recording room in 1962 with EMI BTR2 recorders.** 1878. CC BY-SA 3.0, via Wikimedia Commons. Acessado em 29 de novembro de 2024. Disponível em: <https://commons.wikimedia.org/wiki/File:BTR2_1961-11-12.jpg>.

WILMUT, R. **Early model Studer professional tape recorder, 1969.** 1969. CC BY-SA 3.0, via Wikimedia Commons. Acessado em 29 de novembro de 2024. Disponível em: <https://commons.wikimedia.org/wiki/File:Studer1969.jpg>.

YAKARTEPE, M. G. **Udu drum hang afokse afoxe Yamaha Klarnet Klasik Akustik Gitar Klavye Yamaha or700 Baglama şan solfej Saks Kromatik Mızıka Kanuni trompet Turumpet çelo ukulele**. 2013. CC BY-SA 3.0, via Wikimedia Commons. Acessado em 29 de novembro de 2024. Disponível em: <https://commons.wikimedia.org/wiki/File:Yakartepe%27s_home_studio_1.jpg>.

YI, H. et al. A study on deep neural networks framework. In: **2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)**. IEEE, 2016. p. 1519–1522. ISBN 978-1-4673-9613-4. Disponível em: <https://ieeexplore.ieee.org/document/7867471>.

YILDIRIM, K. et al. Diagnosis of parkinson's disease with acoustic sounds by rule based model. In: **Trends in Data Engineering Methods for Intelligent Systems: Proceedings of the International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2020)**. Springer, 2021. p. 59–75. ISBN 978-3-030-79357-9. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-030-79357-9_7>.

YUN, Y.; CHA, S.-H. Designing virtual instruments for computer music. **International Journal of Multimedia and Ubiquitous Engineering**, Global Vision Press, v. 8, n. 5, p. 173–178, 2013. ISSN 1975-0080. Disponível em: <http://dx.doi.org/10.14257/ijmue.2013.8.5.16>.

ZHANG, L.; CALLISON-BURCH, C. **Language Models are Drummers: Drum Composition with Natural Language Pre-Training**. 2023. Disponível em: <https://arxiv.org/abs/2301.01162>.