UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Verônica Gesteira Souza

Explorando aplicações de inteligência artificial para a mixagem musical

Campinas

2024

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Verônica Gesteira Souza

Explorando aplicações de inteligência artificial para a mixagem musical

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestra em Engenharia Elétrica, na Área de Telecomunicações e Telemática.

Orientador: Prof. Dr. Bruno Sanches Masiero

Este exemplar corresponde à versão final da tese defendida pela aluna Verônica Gesteira Souza, e orientada pelo Prof. Dr. Bruno Sanches Masiero

Campinas

2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

So89e

Souza, Verônica Gesteira, 1998-
Explorando aplicações de inteligência artificial para a mixagem musical / Verônica Gesteira Souza. – Campinas, SP : [s.n.], 2024.

Orientador: Bruno Sanches Masiero.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Faculdade de Engenharia Elétrica e de Computação.

1. Mixagem (Som). 2. Sistemas inteligentes. 3. Inteligência artificial. I. Masiero, Bruno Sanches, 1981-. II. Universidade Estadual de Campinas (UNICAMP). Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações Complementares

Título em outro idioma: Exploring artificial intelligence applications to music mixing
Palavras-chave em inglês:
Mixing (Sound)
Intelligent systems
Artificial intelligence
Área de concentração: Telecomunicações e Telemática
Titulação: Mestra em Engenharia Elétrica
Banca examinadora:
Bruno Sanches Masiero [Orientador]
Maurício do Vale Madeira da Costa
Alexandre Virginelli Maiorino
Data de defesa: 05-07-2024
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0009-0009-9627-2679
- Currículo Lattes do autor: http://lattes.cnpq.br/6714985161771342

COMISSÃO JULGADORA - DISSERTAÇÃO DE MESTRADO

Candidata: Verônica Gesteira Souza RA: 187814

Data da Defesa: 05 de julho de 2024

Título da Tese: “Explorando aplicações de inteligência artificial para a mixagem musical”.

Prof. Dr. Bruno Sanches Masiero (Presidente)
Prof. Dr. Maurício do Vale Madeira da Costa
Prof. Dr. Alexandre Virginelli Maiorino

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

Dedico esse trabalho ao meu pai, Antônio, à minha mãe, Aspásia, e à minha irmã, Melina. Tudo por vocês.

Agradecimentos

Gostaria de agradecer primeiramente à minha mãe, Aspásia Basile Gesteira Souza, e ao meu pai, Antônio Carlos Gesteira Souza, meus maiores companheiros e incentivadores ao longo desses anos. Vocês, que já me dedicaram livros, incontáveis horas de trabalho e imensuráveis quantidades de cuidado, hoje eu dedico a vocês esta dissertação. Cada momento que compartilhamos juntos me dá a certeza de que ser filha de vocês é minha maior sorte e meu maior orgulho.

Agradeço imensamente à minha irmã, Melina Gesteira Souza, por ter aberto os caminhos para mim e por ter me ensinado tanto, seja sobre cálculo, seja sobre a vida. Nada disso teria acontecido sem você. Te amo tanto, irmã.

Aos meus queridos avós paternos, Dilermanda Gesteira Souza e Wilques Souza, por terem atravessado o Brasil, e aos meus avós maternos, Vasiliki Dimitrios (Kotapita) Stavrakas (em memória) e Georges Basile Stavrakas (em memória), por terem atravessado o Atlântico. Daria tudo para poder compartilhar esse momento com vocês em vida. Ser neta de gregos e baianos me ensinou coisas que a escola e a universidade jamais poderiam me ensinar. Aos demais familiares, perto ou longe, obrigada por terem cuidado de mim e por terem feito parte de tudo isso. Da zona leste para o mundo!

Ao meu orientador, Bruno, obrigada por aceitar me orientar e por ter me guiado ao longo de todo o processo de realização deste trabalho e escrita desta dissertação. Levarei cada ideia trocada para o resto da minha vida profissional e acadêmica. Obrigada por todo o apoio e, sobretudo, obrigada por tudo que você tem feito pela nossa Faculdade.

Às amizades feitas antes da universidade e ao longo desses oito anos de Unicamp: nada teria sentido sem vocês ao meu lado! Um agradecimento especial aos colegas que compõem as entidades estudantis da Elétrica por manterem pulsando nossa comunidade. Aos colegas que me acompanharam durante as gestões da empresa júnior e do centro acadêmico, e aos colegas da atlética e do ramo estudantil, obrigada por, mesmo cansados, nunca pararem de imaginar uma universidade melhor e uma engenharia melhor. Aos colegas pesquisadores e docentes dos laboratórios LAC, Nics e Dspcom: obrigada por cada ideia compartilhada e por serem uma inspiração acadêmica e pessoal.

Por último, minha eterna gratidão a todas as pessoas que me acompanharam de maneira direta ou indireta na minha trajetória musical e acadêmica. Por cada artista que me inspirou, por cada pessoa que me encorajou e por cada profissional que acreditou em mim e me abriu as portas. Não estaria aqui sem vocês.

Resumo

A mixagem multicanal é o processo no qual as várias faixas de uma peça musical são processadas e combinadas de maneira a atingir alguns objetivos técnicos e artísticos. Sendo essa uma tarefa de alta complexidade e subjetividade, nos últimos anos surgiram aplicações de inteligência artificial que buscaram automatizar por completo a mixagem. Este trabalho buscou realizar a revisão de dois dos modelos de inteligência artificial mais recentes propostos para essa tarefa e, posteriormente, medir seus resultados por meio de um teste subjetivo de preferência, onde as mixagens feitas pelos modelos foram confrontadas entre si e também por mixagens feitas por humanos. Nossos resultados apontaram que as mixagens humanas foram preferidas a maior parte das vezes pelos participantes, independentemente do nível de conhecimento em mixagem de cada sujeito. Quando as mixagens dos dois modelos foram confrontadas entre si, os resultados divergiram, ora apontando uma preferência por um e ora por outro modelo. Alguns possíveis pontos de melhoria e direções futuras foram apontados, bem como questões éticas sobre o desenvolvimento dessa classe de modelos.

Palavras-chaves: mixagem automática; sistemas inteligentes de mixagem.

Abstract

Multitrack mixing is the process in which the various tracks of a musical piece are processed and combined to achieve certain technical and artistic goals. As this is a task of high complexity and subjectivity, in recent years, artificial intelligence applications that seek to fully automate mixing have emerged. This study aimed to review two of the most recent artificial intelligence models proposed for this task and subsequently measure their results through a subjective preference test, where the mixes made by the models were compared with each other and also with mixes made by humans. Our results indicated that human mixes were preferred most of the time by the participants, regardless of each subject's level of mixing knowledge. When the mixes of the two models were compared with each other, the results diverged, sometimes indicating a preference for one model and sometimes for the other. Some possible areas of improvement and future directions were identified, as well as ethical issues regarding the development of this class of models.

Keywords: automatic mixing; Intelligent Mixing Systems.

"As ciências fornecem um entendimento de uma experiência universal, as artes são um entendimento universal de uma experiência pessoal...as duas são parte de nós e uma manifestação da mesma coisa."

(Mae Jemison)

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

No contexto da produção musical e da música gravada, a **mixagem multicanal** pode ser descrita como o processo no qual os vários elementos de uma peça musical, sejam estes gravados ou sintetizados, são devidamente processados e combinados em um arquivo final (*mixdown* ou *master track*), normalmente contendo um canal (mono), dois canais (estéreo) ou mais canais (formatos de áudio imersivo, por exemplo) (IZHAKI, 2017). Cabe ao profissional de mixagem garantir que os elementos sonoros de uma composição musical estejam ajustados, encaixados e destacados adequadamente de acordo com a **visão artística** estabelecida pelo artista (COLONEL; REISS, 2021). Por ser uma tarefa de grande relevância técnica e artística e ao mesmo tempo de complexidade elevada, nos últimos anos ocorreram muitos esforços no sentido de estudar a sua automatização (MOFFAT, 2021). Os primeiros estudos nessa área propunham abordagens baseadas em sistemas especialistas e em aprendizado de máquina; entretanto, na última década, as abordagens que têm se mostrado mais promissoras são baseadas em aprendizado profundo, enquanto abordagens generativas ainda necessitam de maior exploração (STEINMETZ *et al.*, 2022).

A mixagem de um conjunto de áudios multicanal envolve a manipulação de quatro aspectos sonoros: o tempo, a frequência, o volume (ou nível) e o espaço (ou ambientação) (IZHAKI, 2017). Tal manipulação é feita pelo emprego de uma série de ferramentas como equalizadores, *faders* de volume, compressores, *reverbs* e assim por diante (MAN *et al.*, 2019). Um ponto que torna o processo de mixar extremamente complexo é o fato de não existir apenas um resultado final correto, já que existem múltiplas mixagens para uma mesma música que são comercialmente viáveis (MAN *et al.*, 2015).

Avaliar um *mixdown*, portanto, não é uma tarefa simples e objetiva. Trabalhos recentes utilizam principalmente **métricas subjetivas** (testes de percepção) para avaliar e comparar diferentes mixagens para uma mesma música ou até mesmo para confrontar uma mixagem feita por um humano com uma feita por inteligência artificial (STEINMETZ *et al.*, 2022).

Atualmente as mixagens geradas pelos sistemas automáticos disponíveis tem resultados muito limitados quando comparados aos resultados produzidos por profissionais do áudio (STEINMETZ *et al.*, 2022). Este trabalho tem como objetivo principal, portanto, o de responder às seguintes perguntas de pesquisa: "Como as mixagens feitas por diferentes modelos de IA são avaliadas quando confrontadas entre si e com mixagens feitas por humanos? Há alguma diferença na avaliação na visão de profissionais, amado-

res e entusiastas?". Destas perguntas, temos a seguinte hipótese nula: "Não há diferença significativa na preferência entre mixagens feitas por humanos e por modelos de IA entre pessoas com diferentes níveis de conhecimento sobre mixagem".

De maneira a rejeitar ou confirmar a hipótese nula, será realizada uma revisão do atual estado da arte da mixagem automática, bem como de abordagens generativas utilizadas em outras tarefas de áudio. Posteriormente, os resultados obtidos por dois modelos de mixagem automática distintos serão comparados entre si e com os resultados gerados por humanos em um teste subjetivo.

Para expor o que foi desenvolvido, este trabalho está dividido em 6 capítulos. O primeiro é a introdução aqui contida. O segundo capítulo apresenta uma fundamentação acerca dos principais conceitos e ferramentas da mixagem musical, tendo em vista que não é um assunto de amplo conhecimento. No terceiro capítulo é feita uma extensa revisão bibliográfica acerca das tecnologias existentes para automatizar a tarefa da mixagem musical, bem como uma revisão sobre aplicações musicais de modelos generativos.

No quarto capítulo, a metodologia utilizada é apresentada, incluindo detalhes do treinamento dos modelos e do desenho do teste de percepção. No capítulo quinto são apresentados e discutidos os resultados do teste de percepção. Por fim, o trabalho é concluído no sexto capítulo e possíveis direções futuras são apontadas.

Uma observação final pertinente a ser feita sobre este trabalho é que, no melhor do nosso entendimento, existem poucos trabalhos sobre mixagem e sobre mixagem automática feitos em outra língua que não seja a língua inglesa. Isso implica no fato de que muitos termos não possuem tradução difundida e logo serão apresentados no idioma original.

2 Fundamentos

A mixagem, dentro do contexto da música gravada em estúdio, é um dos estágios finais na concepção de um fonograma, situando-se após a etapa de produção e antes da masterização, como ilustrado na Figura 2.1. Normalmente, um projeto musical recebe uma verba e é designado a um produtor ou produtora musical, que será responsável por conduzir toda a etapa de produção, que envolve a composição, o arranjo, a gravação e a edição. O(A) produtor(a) musical costuma ser responsável também por definir quem serão os profissionais envolvidos em cada etapa, desde compositores e músicos, até os profissionais de mixagem e masterização (SAVAGE, 2014).

Como resultado da etapa de produção, um arquivo multicanal contendo todos os elementos da peça musical em questão é gerado e enviado para o profissional responsável pela mixagem. Antes do processo de mixagem em si, é feita uma preparação onde os diferentes elementos podem ser agrupados. Um grupo de instrumentos constitui uma espécie de "canal", no qual algumas faixas podem ser enviadas a ele e podem ser processadas, além de individualmente, de maneira conjunta. Um exemplo tradicional é o agrupamento dos elementos da bateira (como bumbo, caixa, chimbal, etc.) em um único grupo, o qual chamamos de subgrupo ou *bus group*. Após a realização da mixagem, um arquivo com o áudio finalizado é gerado e enviado ao profissional da masterização. A masterização é o último passo na cadeia da produção musical, sendo a última oportunidade de potencializar aspectos estéticos (como timbre) e técnicos (como *loudness*, que será futuramente aprofundado) de uma música ou de um conjunto de músicas como um álbum, por exemplo (KATZ, 2003).

Como mencionado, não existem regras que determinem qual o resultado emergente correto ou incorreto da mixagem, no entanto, a aplicação de ferramentas é norteada por alguns conceitos técnicos e estéticos (DOBROWOHL *et al.*, 2019). Além de garantir a clareza de cada elemento musical presente, o processo é guiado também pela **visão artística** inicialmente estabelecida pelo produtor musical responsável pela música ou álbum e posteriormente refinada pela(o) engenheira(o) de mixagem. A elaboração da visão de mixagem é uma etapa crucial do processo como um todo, sendo normalmente estabele-

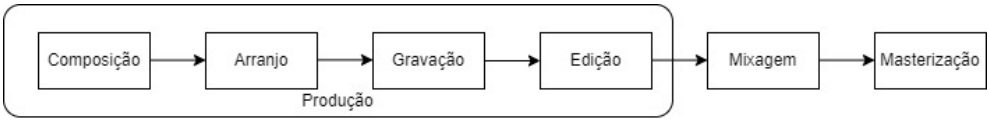

Figura 2.1 – Representação das etapas envolvidas na produção, mixagem e masterização de uma música.

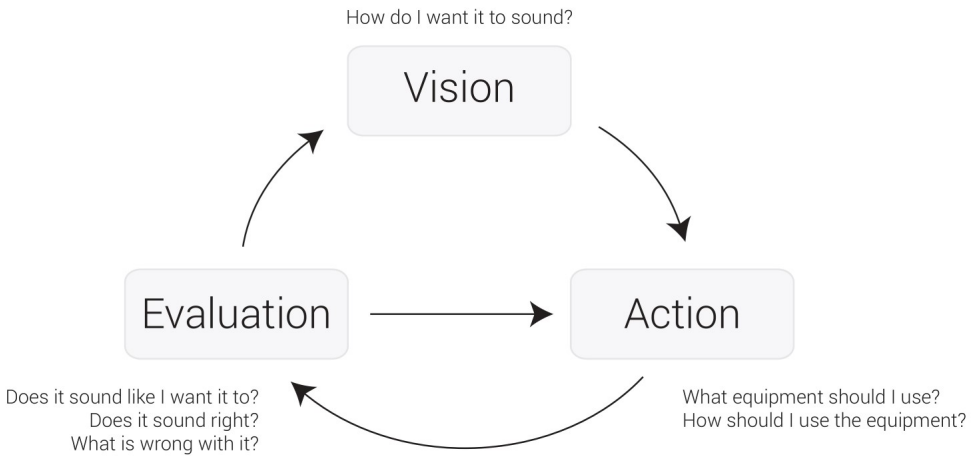

Figura 2.2 – Representação do processo de tomada de decisão da pessoa responsável pela mixagem, levando em consideração a visão artística estabelecida (IZHAKI, 2017).

cida utilizando uma combinação de práticas comumente aplicadas ao gênero ou estilo do produto a ser mixado e de referências observadas em outras produções similares. Tendo a visão estabelecida, a pessoa responsável pela mixagem normalmente avalia cada ação tomada ao longo da mixagem para garantir que a ação teve o resultado desejado. Uma representação desse processo iterativo pode ser observada na Figura 2.2.

Alguns gêneros e estilos musicais possuem particularidades próprias a determinadas linhas estéticas que são muito observadas no fazer artístico, bem como pelos usuários e consumidores finais. A pessoa responsável pela mixagem pode selecionar um fonograma, ou um conjunto de fonogramas, para referência na determinação da visão. O(A) artista ou produtor(a) musical também podem indicar essas referências. Usualmente a referência é do mesmo gênero musical e com uma instrumentação similar ao fonograma sendo mixado e também é uma obra que já foi ouvida extensivamente pela(o) engenheira(o) de mixagem (IZHAKI, 2017).

A seguir será apresentado um breve panorama histórico sobre a mixagem e também serão aprofundados alguns conceitos e ferramentas importantes para o melhor entendimento da função de um profissional da mixagem e da relevância deste trabalho como um todo.

2.1 Uma breve história da Mixagem

O início da história do áudio gravado é comumente atribuído ao ano de 1877 com o surgimento do fonógrafo, invenção de Thomas Edison, que consistia em um cilindro giratório capaz não apenas de armazenar, como também reproduzir informações sonoras. Como alternativa ao fonógrafo, Emile Berliner propôs em 1887 um sistema de gravação e

reprodução baseado em discos planos - o gramofone - que possuía uma série de vantagens em relação ao fonógrafo como maior espaço para armazenamento e a possibilidade de gerar cópias dos registros com facilidade, possibilitando assim o início da comercialização de gravações musicais. Outra contribuição importante para a história do áudio gravado foi dada pelo Dinamarquês Valdemar Poulsen, que inventou o primeiro sistema de gravação com o uso de magnetismo em 1898, o telegrafone, que funcionava através da atuação de um eletroímã que magnetiza um fio de aço de acordo com os sons captados (ARAÚJO, 2015).

Até então, todos os elementos de uma música eram gravados como um único produto final e o processo de gravação era realizado em estúdios sob a supervisão de um técnico de gravação que escolhia o posicionamento correto dos músicos em relação ao único captador sonoro e escolhia a sensibilidade do diafragma que iria atuar no captador (ARAÚJO, 2015). Posteriormente, com a possibilidade de empregar mais de um microfone na mesma gravação e de adicionar câmaras reverberantes durante a execução da música, as possibilidades de atuação do técnico de gravação começaram a expandir. Embora a figura do técnico de gravação pudesse se assemelhar à figura de um(a) engenheiro(a) de mixagem, a tarefa da mixagem como conhecemos hoje surgiu apenas no final da década de 1950, justamente com o advento da gravação multicanal.

Em 1955, a companhia estadunidense *Ampex* lança o primeiro gravador multicanal, com 8 canais de gravação. O engenheiro da *Ampex* Ross Snyder, insipirado pela técnica de *overdubbing* criada pelo guitarrista Lester William Polsfuss (*Les Paul*), foi o responsável por desenvolver o gravador multicanal com a tecnologia *Sel-Sync* (*Selective Synchronous* ou Sincronia Seletiva em português) que permitia que cada elemento fosse manipulado individualmente dentro de uma gravação (SNYDER, 2003). Ao longo dos anos 1960, as possibilidades artísticas inauguradas pela gravação multicanal chamaram a atenção de grandes artistas da época como os Beach Boys e os Beatles, o que impulsionou a evolução dos dispositivos de gravação no decorrer da década de 1970 e no início da década de 1980 (OWSINSKI, 2014). No final da década de 1980 e no início dos anos 1990 mesas digitais de áudio começam a se difundir pelos estúdios, bem como a criação de softwares que simulavam sistemas de gravação, as chamadas Estações Virtuais de Áudio (do inglês *Digital Audio Workstation* ou simplesmente *DAW*). Nesse período também se iniciou a comercialização dos primeiros computadores pessoais, o que acabou por baratear os custos de produção dos equipamentos de áudio e também popularizar o ato de gravar e registar sons (PHILLIPS, 2016).

Em 1996 a empresa alemã Steinberg lançou o software *Cubase VST*, a primeira *DAW* que possibilitava a inclusão de efeitos virtuais (*plugins*) de empresas terceiras, o que marcou o início de uma nova era para a história da música gravada. Entre o final dos anos

1990 e início dos anos 2000, novas empresas surgiram e a tecnologia das *DAWs* evoluiu de forma muito acelerada, de modo que em meados dos anos 2000 a utilização das *DAWs* havia se tornado o novo padrão do mercado fonográfico. As ferramentas digitais aos poucos foram agradando aos engenheiros e engenheiras de mixagem e tomando o espaço das ferramentas analógicas, de modo que a mixagem progressivamente foi deixando de ser um processo restrito ao ambiente do estúdio profissional com espaçosos e caros equipamentos analógicos e passasse a ocorrer quase exclusivamente dentro do computador (OWSINSKI, 2014).

2.2 Os objetivos e os domínios da Mixagem

De modo a delimitar a função do profissional da mixagem, Izhaki define quatro objetivos e quatro domínios principais para a tarefa da mixagem (IZHAKI, 2017). Os objetivos são: emoção, equilíbrio, definição e interesse. Enquanto os domínios são: o tempo, a frequência, o nível sonoro e o espaço. Esses componentes se relacionam entre si à medida em que os domínios nada mais são do que os aspectos que podem ser manipulados em uma música de modo a atingir os objetivos.

2.2.1 Os objetivos da Mixagem

2.2.1.1 Emoção

Garantir que o contexto emocional de uma música seja adequadamente refletido ao ouvinte final é um objetivo presente em todas as etapas da produção de um fonograma ilustradas na Figura 2.1. O profissional da mixagem poderá utilizar de sua criatividade e sensibilidade para que os elementos da peça musical se mantenham dentro dessa coerência emocional. Se a música em questão é um *jazz* suave, por exemplo, cuja visão artística seria a de transmitir uma sensação de tranquilidade ao ouvinte, não seria usual que os elementos dessa música fossem processados de uma maneira agressiva e que descaracterizassem os sons dos instrumentos. Por outro lado se a música em questão se trata de um *heavy metal*, espera-se que haja, por exemplo, a distorção de alguns dos elementos musicais (IZHAKI, 2017).

2.2.1.2 Equilíbrio

O equilíbrio (ou balanço) normalmente é buscado nos seguintes domínios: frequência, espaço e nível sonoro (IZHAKI, 2017). Mais a frente, cada domínio será aprofundado, mas em termos gerais é possível dizer que faz parte da mixagem garantir que cada elemento esteja representado de uma maneira coerente e que haja coesão entre os elementos. Um exemplo intuitivo de desequilíbrio seria se tivéssemos um instrumento com

um nível sonoro muito superior aos demais, o que provavelmente geraria estranhamento no ouvinte.

2.2.1.3 Definição

A Definição pode ser descrita como a capacidade de distinguir e reconhecer um som (IZHAKI, 2017). Em uma música normalmente haverá elementos que conflitam entre si no sentido de ocuparem regiões parecidas no espectro em frequência. Na maioria das vezes a mixagem tem como objetivo o de fazer com que as ideias musicais sejam traduzidas da melhor maneira possível. No caso da voz humana, por exemplo, é esperado que o ouvinte possa ouvir e compreender as palavras que estão sendo cantadas. O profissional da mixagem pode aplicar algumas ferramentas de modo a garantir a definição dos elementos. Nas sessões seguintes, será aprofundado como isso é feito.

2.2.1.4 Interesse

O Interesse pode ser definido como o objetivo de manter e potencializar o interesse presente na música desde a concepção da composição e do arranjo (IZHAKI, 2017). A pessoa responsável pela mixagem pode então, por exemplo, processar cada elemento de maneira distinta ao decorrer da música de modo a manter o ouvinte interessado.

2.2.2 Os domínios da Mixagem

2.2.2.1 Tempo

Existem efeitos que podem ser aplicados na mixagem que têm funções atreladas ao preenchimento do domínio tempo. Na próxima sessão, as ferramentas empregadas na mixagem serão aprofundadas, porém é possível citar efeitos como o *delay* e o *reverb*, que são capazes de alterar aspectos relativos ao tempo de um som.

2.2.2.2 Frequência

O domínio da frequência, restringido pela extensão audível ao ouvido humano (entre 20 Hz e 20 kHz), é um dos aspectos mais complexos para se lidar na mixagem, já que elementos podem conflitar entre si no âmbito da frequência, acarretando no fenômeno do **mascaramento**. O mascaramento em frequência é um fenômeno psicoacústico, que ocorre quando um som com maior nível sonoro (mascarador) prejudica a escuta de um outro som com menor nível sonoro (mascarado), quando ambos estão em regiões próximas em frequência (BOSI; GOLDBERG, 2002).

Esse fenômeno pode ser explicado pelo fato de que tanto o som mascarador quanto o som mascarado, por terem conteúdos similares em frequência, geram excitação

na mesma região da membrana basilar da cóclea, estrutura presente no ouvido humano responsável por transformar as oscilações mecânicas do som em sinais elétricos. Isso faz com que o som de maior nível sonoro se sobressaia em relação ao som de menor nível sonoro (BOSI; GOLDBERG, 2002).

Na literatura existem formas distintas de divisão do domínio da frequência, porém uma divisão possível é reparti-lo em quatro registros: graves (de 20 Hz até 250 Hz), médio-graves (de 250 Hz até 2 kHz), médio-agudos (de 2 kHz até 6kHz) e agudos (de 6kHz até 20 kHz) (IZHAKI, 2017). O espectro de frequências de cada elemento de mixagem (voz ou instrumento) ocupa um lugar na extensão do âmbito de frequências, ou seja, cada elemento tem uma contribuição espectral na mixagem como um todo. A ação de equilibrar a distribuição espectral dos elementos na mixagem é objetivamente uma ação de **balanço tonal**. Normalmente obter um balanço tonal adequado envolve identificar as regiões em frequência que estão em excesso ou em falta e manipular cada elemento de maneira a corrigir esses desequilíbrios.

Entre profissionais da mixagem, existem uma série de termos subjetivos que são utilizados para se referir a problemas de balanço tonal. Por exemplo, se há excesso de graves, pode-se dizer que a mixagem está com excesso de graves (*boomy*) e se há falta de graves, pode-se dizer que a mixagem está magra (*thin*). Também existem termos que são utilizados sem conotação positiva ou negativa, e sim realizam associações entre termos subjetivos com regiões em frequência e podem ajudar na comunicação entre profissionais da mixagem e outros profissionais da música. Os termos mais usuais podem ser encontrados na Figura 2.3.

2.2.2.3 Nível Sonoro

O domínio do Nível Sonoro, popularmente chamado de volume, pode ser manipulado de duas maneiras: absoluta e relativa. O nível sonoro absoluto é o nível da *master track* como um todo, enquanto o nível sonoro relativo seria a relação entre os níveis dos elementos entre si. A mixagem pouco se ocupa da manipulação absoluta do nível sonoro, isso é de maior responsabilidade do profissional da masterização. Já a manipulação relativa do nível dos elementos é uma das tarefas cruciais da mixagem (IZHAKI, 2017).

De modo a garantir a inteligibilidade dos elementos, ajustar os níveis relativos é primordial. Uma técnica muito comum entre profissionais da mixagem consiste em iniciar o processo da mixagem pela definição dos níveis de cada elemento, garantindo que haja equilíbrio. Normalmente o equilíbrio significa elevar o nível sonoro dos elementos mais importantes de uma música. Se estivermos ouvindo um solo de guitarra, por exemplo, não seria coerente que a guitarra estivesse com um nível sonoro muito menor em comparação aos demais elementos.

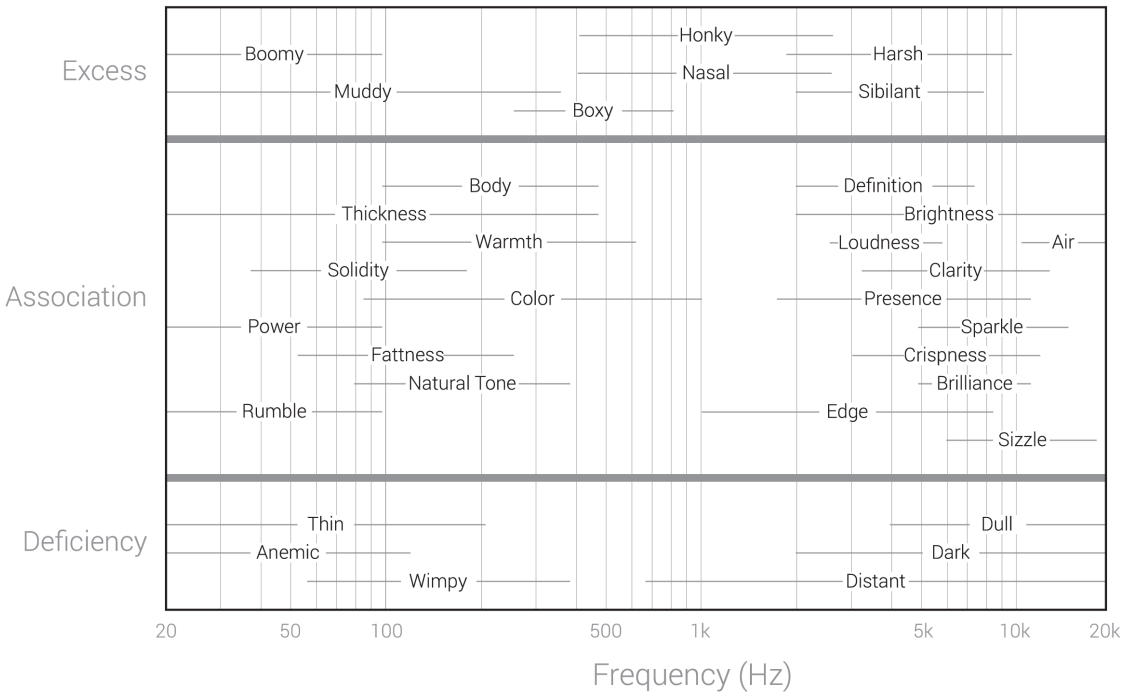

Figura 2.3 – Jargões utilizados por profissionais da mixagem de modo a expressar características de regiões em frequência, bem como excesso ou falta de algum conjunto de frequências. (IZHAKI, 2017).

Importa destacar que, para sinais digitais, o nível sonoro é representado em decibéis relativos à escala completa (do inglês *decibels related to full scale*), abreviado como dBFS. O decibel é uma unidade de medida relativa, que expressa sempre uma razão logarítmica em relação a um nível de referência. No caso do dBFS, o nível de referência é o valor máximo de nível sonoro permitido nos sistemas digitais, ou seja, zero dBFS representa o maior nível digital possível (REISS; MCPHERSON, 2014).

Outro tópico importante dentro do domínio do nível sonoro é a manipulação da **dinâmica**. A dinâmica resulta da gestão e manuseio do alcance dinâmico geral (utiliza-se normalmente o termo em inglês *dynamic range*), que pode ser definido como a diferença entre o nível mais fraco e o mais forte de um sinal (KATZ, 2003). Segundo Izhaki (2017), o elemento dinâmica pode ainda ser dividido 'entre macrodinâmica e microdinâmica. (...) Macrodinâmica está relacionada a variações de nível para eventos mais amplos que uma nota musical (...), enquanto microdinâmica está relacionada à variação de nível ocorrendo no detalhe, ou seja, no acontecimento de cada nota musical tocada(...).' Associamos microdinâmica ao envoltório dinâmico (volume x tempo) de cada som individualmente, normalmente dividido em quatro partes: ataque (*attack*), decaimento (*decay*), sustentação (*sustain*) e liberação ou extinção/desinência (*release*) (BURRED *et al.*, 2009).

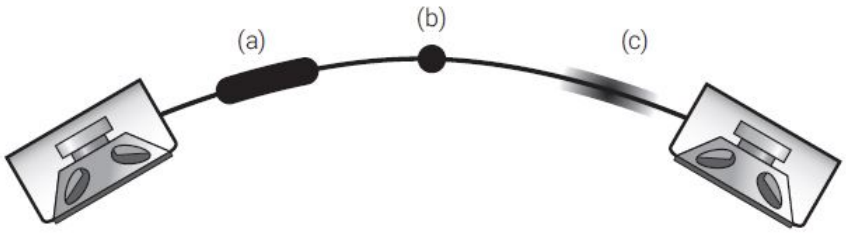

Figura 2.4 – Ilustração da imagem estéreo e das características com que os elementos musicais podem ser percebidos. (IZHAKI, 2017).

A maneira pela qual se procede a manipulação da microdinâmica e da macrodinâmica em uma mixagem, pode depender de alguns fatores técnicos para facilitar a inteligibilidade de um elemento, mas é principalmente vinculada ao gênero e ao estilo da música a ser mixada. Por exemplo, enquanto o uso de manipulação dinâmica é de suma importância no repertório pop ou hip-hop, ele é mínimo, se não raro no jazz ou na música de concerto (OWSINSKI, 2014).

2.2.2.4 Espaço

O domínio do espaço pode ser dividido em dois subdomínios: a imagem estéreo e a profundidade. A imagem estéreo, considerando que estamos tratando de áudio no formato estéreo, é o espaço imaginário formado entre os alto-falantes direito e esquerdo. Já a profundidade é a distância aparente que um elemento possuí em relação ao ouvinte.

A manipulação da imagem estéreo em uma mixagem pode ocorrer de quatro maneiras: localização, abertura, foco e distribuição. A localização diz respeito à posição de um elemento musical no eixo direita-esquerda. A abertura diz respeito ao espaço na imagem estéreo que um elemento ocupa. O foco diz respeito à capacidade de distinguir ou não a origem de um som dentro da imagem estéreo. Finalmente, a distribuição diz respeito a maneira como os elementos estão distribuídos na imagem estéreo. Na Figura 2.4 temos uma analogia visual para os conceitos presentes na imagem estéreo. Para um ouvinte posicionado no centro, entre os dois monitores de áudio, podemos dizer que o elemento (a) está localizado mais a esquerda do que os elementos (b) e (c) por exemplo. No quesito abertura, o elemento (a) está mais aberto do que (b). Já em termos de foco, o elemento (c) está mais desfocado do que (a) e (b). Por último, a distribuição diz respeito à maneira que (a), (b) e (c) estão posicionados na imagem estéreo.

Uma consideração importante a ser feita sobre o subdomínio estéreo é que o princípio do equilíbrio também está muito presente na sua manipulação. Balanço estéreo é o nome dado para a ação de balancear os níveis sonoros e o conteúdo em frequência dos canais direito e esquerdo de modo que não haja grandes diferenças entre eles. Quando há

muita discrepância entre o nível sonoro e o conteúdo em frequência dos canais direito e esquerdo o resultado pode gerar bastante desconforto ao ouvinte.

O outro subdomínio que compõe o domínio do espaço é a profundidade. A profundidade aqui é tratada em termos relativos, ou seja, podemos dizer que um elemento está mais próximo ou mais distante que outro. Existem alguns efeitos que podem ser aplicados para trazer, perceptivamente, um elemento mais para perto ou mais para longe do ouvinte, e que serão futuramente aprofundados, mas de maneira geral podemos dizer que o nível sonoro de um elemento em relação ao nível sonoro dos demais elementos costuma estar ligada à percepção da sua profundidade.

A manipulação da imagem estéreo e da profundidade, portanto, dão ao profissional da mixagem a possibilidade de posicionar cada elemento de uma peça musical no espaço. Esse posicionamento é muito importante para garantir que sejam cumpridos os quatro objetivos da mixagem - emoção, equilíbrio, definição e interesse.

2.3 As ferramentas da mixagem

Para atingir os objetivos da mixagem, são empregadas diversas ferramentas que manipulam os domínios apresentados na sessão anterior. Importa ressaltar que não existe uma ordem pré estabelecida para a utilização de cada uma dessas ferramentas, uma vez que isso fica a critério do profissional. A seguir serão apresentadas algumas das principais ferramentas utilizadas na mixagem.

2.3.1 Monitoramento

Monitoramento é o nome dado ao sistema de reprodução de áudio através do qual o profissional da mixagem realizará o processo de escutar criticamente o material sonoro e realizar as manipulações necessárias. Normalmente esse sistema consiste em um fone de ouvido ou em um conjunto de caixas de som (também chamados de monitores). Como cada circuito de amplificação possui uma resposta em frequência característica, o sistema de monitoramento influencia o conteúdo que estará sendo ouvido. No caso do uso de caixas de som, é importante pontuar que a característica acústica da sala na qual os monitores estão posicionados também influencia na audição (SENIOR, 2018).

2.3.2 Medidores de volume

Controlar o volume dos elementos ao longo da mixagem, como dito anteriormente, é uma tarefa importante. O uso de medidores de nível sonoro são de grande utilidade na medida em que oferecem uma referência visual do nível dos elementos na

mixagem. Os medidores mais utilizados são os medidores de pico e os de valor médio. Os medidores de pico mostram o valor de pico do sinal naquele instante (valor instantâneo) e tem como valor máximo em sistemas digitais o 0 dBFS, já os medidores de valor médio mostram uma média do nível do sinal em uma janela de tempo pré-definida (IZHAKI, 2017).

2.3.3 Panorama

Dentro do domínio do espaço, e mais precisamente no domínio da imagem estéreo, o posicionamento dos elementos no eixo horizontal (direita-esquerda) se dá pela manipulação do panorama. O ato de panoramizar, portanto, envolve controlar qual será a diferença de nível sonoro de um elemento entre os canais direito e esquerdo. Um sinal mono, possuí canais direito e esquerdo exatamente iguais e logo, como o nível sonoro que chega aos dois ouvidos é a mesma, o sinal aparenta estar vindo de uma fonte sonora centralizada. Já um sinal panoramizado para a esquerda possuí maior nível no canal esquerdo do que no direito e, portanto, aparenta estar vindo de uma fonte sonora posicionada à esquerda e vice-versa para um sinal panoramizado para a direita. Hoje em dia, as *DAWs* possuem um botão de panorama em cada canal, o que permite a manipulação do panorama de cada elemento individualmente de forma prática.

2.3.4 Equalizadores

Segundo Izhaki (2017), "distinguir frequências e dominar a manipulação delas é talvez o maior desafio a ser enfrentado em uma mixagem". O equalizador, mesmo não sendo a única forma de manipular frequências em uma mixagem, é a principal e mais importante ferramenta com esta finalidade. O princípio de correção de frequência está interligado à ação de diminuir ou acrescentar energia em determinadas bandas de frequência de um sinal, o que pode ser realizado através do uso de filtros. Equalizador, de uma maneira geral, é o nome dado para um conjunto de filtros. A seguir serão detalhados alguns dos filtros e equalizadores mais utilizados.

Os filtros mais elementares são os **filtros passa-alta (FPA)** e **filtros passa-baixa (FPB)**, representados respectivamente nas Figuras 2.5 e 2.6. O funcionamento desses filtros é bastante simples: a partir da inserção de uma frequência de referência (frequência de *cut-off*) as frequências acima serão atenuadas (FPB) ou mantidas (FPA) e as frequências abaixo serão mantidas (FPB) ou atenuadas (FPA). A taxa na qual essas frequências serão atenuadas dependerá de um outro parâmetro chamado de *slope* (taxa de inclinação em português), que é medida em dB/oitava e determina quão brusca será a atenuação a partir da frequência de *cut-off*. Uma taxa de inclinação de 6dB/oitava,

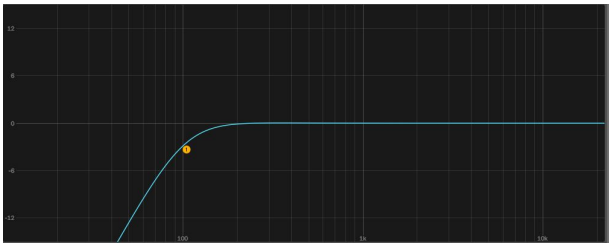

Figura 2.5 – Filtro passa-altas.

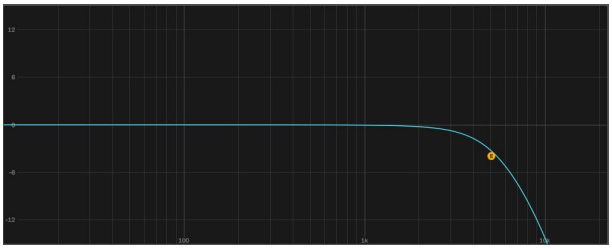

Figura 2.6 – Filtro passa-baixas.

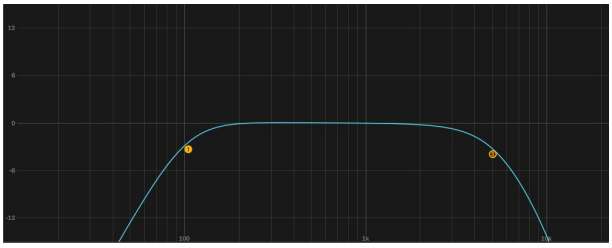

Figura 2.7 – Filtro passa-bandas.

por exemplo, significa que a partir da frequência de *cut-off* cada oitava sofrerá redução de 6 dB (IZHAKI, 2017). Um FPA e um FPB podem ser utilizados individualmente ou combinados de modo a formarem um **filtro passa-banda** (Figura 2.7, onde serão escolhidas duas frequências de *cut-off*.

Os filtros do tipo prateleira (do inglês *shelf*) possuem três parâmetros: frequência central, *slope* e ganho. Existem dois tipos de filtros prateleira: *low-shelf* e *high-shelf*. Os *low-shelf* (Figura 2.9) atenuam ou aumentam (dependendo do valor do ganho) as frequências menores que a frequência central e não alteram as frequências acima. Já o *high-shelf* (Figura 2.8) funciona de maneira inversa. O parâmetro *slope* funciona de forma similar aos filtros passa banda, mas no caso determinará a inclinação (em dB/oitava) com a qual o filtro irá atingir o ganho desejado (TRUAX, 1999).

Os filtros do tipo passa-banda/rejeita-banda são utilizados para agir em bandas de frequências de forma mais específica, podendo atenuá-las (rejeita-banda) ou aumentá-las (passa-banda). Os parâmetros a serem definidos são três: frequência central, ganho e fator de qualidade (largura da banda de frequência). Essa curva também é chamada de *bell*, pois seu formato se assemelha a um sino. Seu funcionamento é dado da seguinte forma: a frequência central determinada será a frequência de referência e o ganho será a quantidade em dB que esta frequência será atenuada ou amplificada. O fator de qualidade

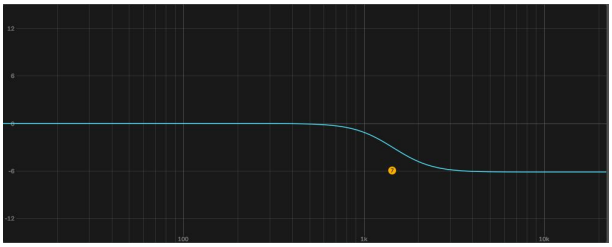

Figura 2.8 – Equalizador *high-shelf* com ganho negativo.

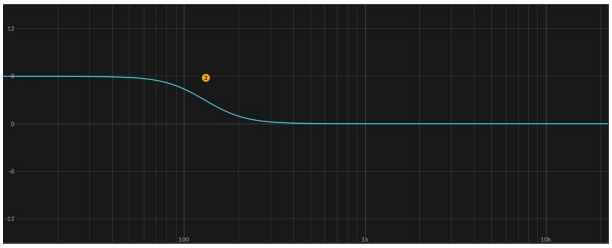

Figura 2.9 – Equalizador *low-shelf* com ganho positivo.

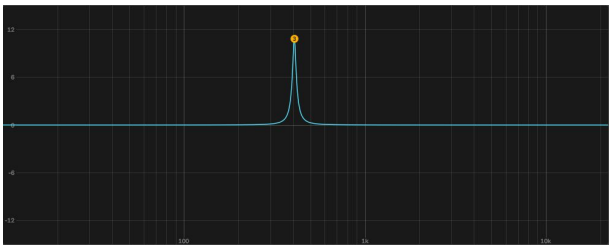

Figura 2.10 – Equalizador *bell* com ganho positivo e valor do fator Q elevado.

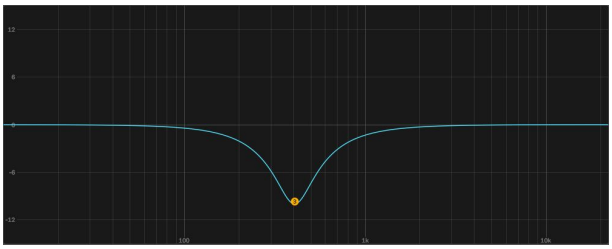

Figura 2.11 – Equalizador *bell* com ganho negativo e valor do fator Q mediano.

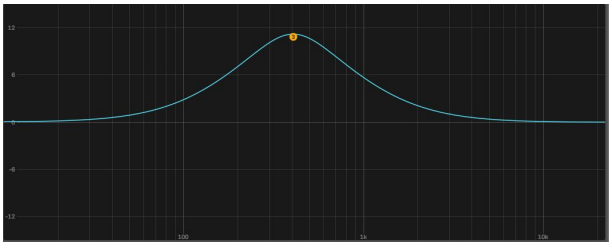

Figura 2.12 – Equalizador *bell* com ganho positivo e valor do fator Q baixo.

Q diz respeito a quão específica será sua ação, ou seja, quanto maior o valor de Q, mais estreita em termos de largura de banda, será a ação da curva e quanto menor o valor de Q, maior será a banda de frequência que sofrerá alteração. Nas Figuras 2.10, 2.11 e 2.12 temos, respectivamente, representados equalizadores com valor de Q elevado, mediano e baixo.

A equalização possui uma série de aplicações ao longo do processo da mixagem, sendo as principais:

- Manipulação tonal: lapidar as características sonoras e o timbre de cada instrumento.
- Separação e definição: diferenciar informações sonoras que estejam ocupando uma mesma posição no espectro de frequências de maneira a não se mascararem entre si, permitindo assim que o ouvinte possa distinguir os vários elementos sobrepostos.
- Ajuste de nível: caso seja necessário atenuar ou aumentar o volume de uma região específica de frequências.
- Remoção de conteúdo indesejado: atenuação de frequências ressonantes indesejadas, ruído ou sons capturados acidentalmente.

2.3.5 Compressores

O compressor é um dos dispositivos utilizados no controle da variação dinâmica e sua origem foi motivada pela necessidade de controlar automaticamente o nível de sinal para diferentes aplicações (IZHAKI, 2017). Embora sua função original seja a de controle dinâmico, o compressor também é utilizado como uma ferramenta para alterar características timbrísticas do som (MOORE *et al.*, 2016). De maneira geral, a operação de um compressor envolve os seguintes parâmetros: *threshold* (limiar), *ratio* (razão), *attack* (tempo de ataque), *release* (tempo de liberação/desinência), *gain* (ganho) e *knee* (joelho) (ELIASSON, 2019).

O parâmetro de *threshold* é o nível limiar de intensidade do sinal a partir do qual o compressor passará a atuar: cada vez que o sinal de entrada ultrapassa o nível de intensidade escolhido como limiar, o compressor inicia sua atuação para reduzir o volume. Essa redução é determinada pelo parâmetro *ratio*, que é a razão entre o nível do sinal de entrada e o nível do sinal de saída (ELIASSON, 2019). Em outras palavras, um ratio de 2:1 implica que para cada 2 dB que o sinal de entrada ultrapasse do valor de limiar, o sinal na saída do compressor tenderá a estar apenas 1 dB acima do valor do limiar (OWSINSKI, 2014). Na Figura 2.13 é possível ver o sinal resultante quando processado por diferentes razões de compressão. Quanto maior a razão de compressão, menor o alcance dinâmico do sinal.

A transição entre o sinal não comprimido e o início da compressão pode ser mais ou menos brusca, sendo determinada pelo parâmetro *knee*. Um *knee* suave irá partir da razão 1:1 (não compressão) para a razão selecionada de forma gradual, enquanto um *knee* agressivo irá fazer essa transição mais rapidamente (IZHAKI, 2017). O *knee*, no

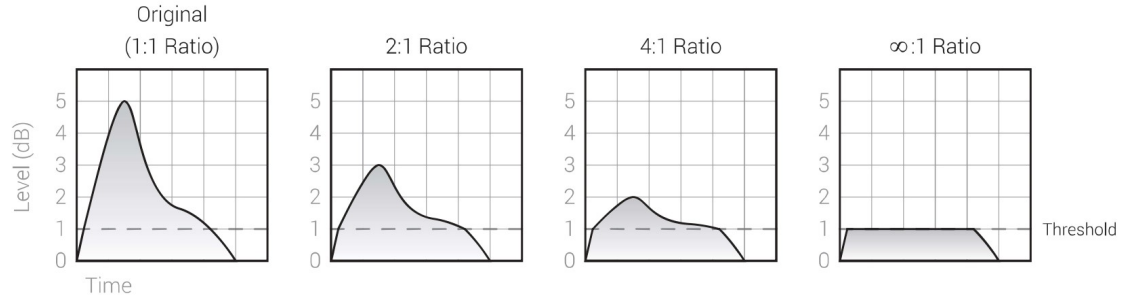

Figura 2.13 – Diferentes razões de compressão aplicadas ao mesmo sinal. Em todos os gráficos o valor selecionado de *threshold* é o mesmo. (IZHAKI, 2017).

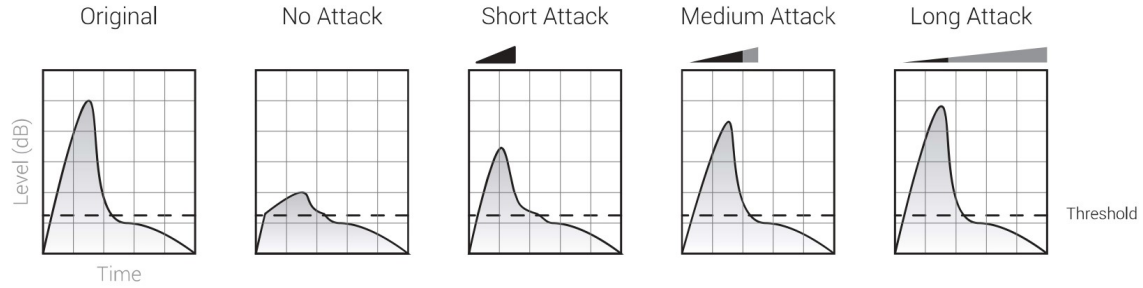

Figura 2.14 – Diferentes tempos de ataque aplicados ao mesmo sinal. Quanto mais duradouro o tempo de ataque, mais o início do sinal (ataque) original é mantido. Em todos os gráficos o valor selecionado de *threshold* é o mesmo e o tempo de liberação é zero. (IZHAKI, 2017).

entanto, não diz respeito à velocidade de atuação do compressor (com a qual o compressor inicia e termina sua atuação), sendo esta dada respectivamente pela escolha do tempo de ataque e do tempo de liberação. Por atuarem no domínio do tempo, os parâmetros de ataque e liberação influenciam diretamente a envoltória sonora do sinal que está sendo processado, como pode ser visto nas Figuras 2.14 e 2.15. Por fim, o parâmetro ganho oferece a possibilidade de aumentar ou reduzir o nível sonoro do sinal no estágio de saída do processamento de compressão.

Com o advento do processamento digital de áudio e o surgimento de compressores digitais, os parâmetros se tornaram muito mais customizáveis, uma vez que os modelos analógicos tinham limitações de recursos e valores pré-estabelecidos para cada parâmetro. Apesar de possuírem menos opções de controle que os compressores digitais, alguns modelos de compressores analógicos são até hoje cultuados pelo seu desempenho e por suas características sonoras (MOORE *et al.*, 2016). Existem quatro tipos de compressores analógicos: valvulados, FET (transistor de efeito de campo), ópticos e VCA (amplificador controlado por voltagem) (OWSINSKI, 2014).

Os primeiros modelos de compressores criados foram os valvulados, que atenuavam o sinal de saída através do uso de válvulas sensíveis ao sinal de entrada. O emprego de válvulas gera distorções harmônicas, mudando aspectos timbrísticos do som, o que faz

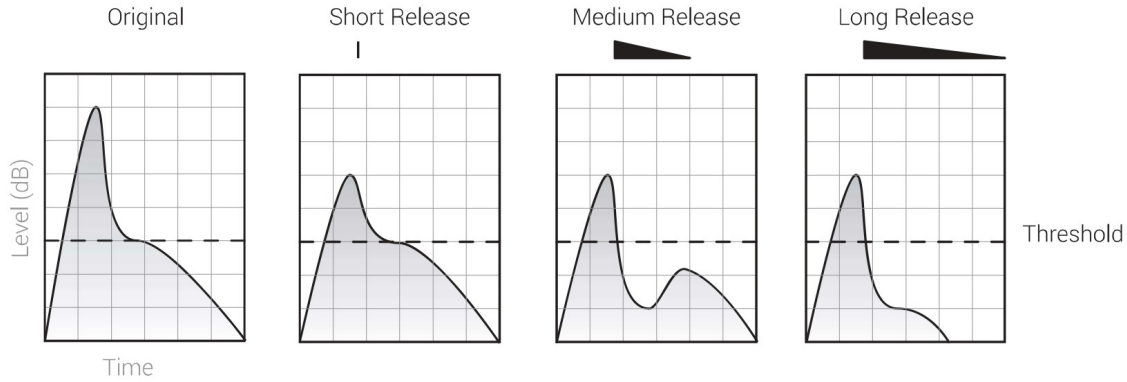

Figura 2.15 – Diferentes tempos de liberação aplicados ao mesmo sinal. Quanto mais curto o tempo de liberação, mais o final do sinal (decaimento) original é mantido. Quanto mais duradouro o tempo de liberação, mais tempo o compressor levará para deixar de atuar no sinal. Em todos os gráficos o valor selecionado de *threshold* é o mesmo e o tempo de ataque é zero. (IZHAKI, 2017).

dos compressores valvulados 'coloridos', termo utilizado para se referir a dispositivos que adicionam intensidade em bandas de frequências do sinal de forma não linear durante o processamento (MOORE *et al.*, 2016). Esse tipo de compressor não possui controle de razão e tem valores de ataque e liberação maiores, sendo, portanto, mais lentos (IZHAKI, 2017).

O circuito dos compressores ópticos é formado por uma fonte luminosa atrelada ao sinal de entrada e uma fotorresistência (OWSINSKI, 2014). A atenuação do sinal se dá por meio da interação entre a fonte luminosa e a fotorresistência, ou seja, quanto maior a intensidade do sinal de entrada no circuito, maior a intensidade da luz emitida pela fonte luminosa e, assim, maior a resistência do fotorresistor, reduzindo a intensidade do sinal (MOORE *et al.*, 2016). Os compressores ópticos possuem controle de razão, porém possuem tempo de resposta lento e produzem coloração ao sinal processado (SAVAGE, 2014).

Os compressores do tipo FET empregam transistores de efeito de campo que reduzem o sinal por meio do controle da sua resistência através da alteração da voltagem aplicada ao terminal *gate* (MOORE, 2017). Os compressores FET apresentam tempos de ataque e de liberação mais rápidos do que os valvulados e os ópticos, e permitem o controle de *ratio* (MOORE *et al.*, 2016).

Os compressores VCA são os mais recentes entre os compressores analógicos, tendo surgido na década de 1980 (OWSINSKI, 2014). Seu funcionamento ocorre através do controle de voltagem de um circuito de amplificação integrado. Caracterizam-se por ter tempos de resposta mais rápidos, e por ser muito mais 'transparente', termo utilizado

para ferramentas que não alteram as características em frequência do sinal (ou seja, não 'colorido') (MOORE *et al.*, 2016).

Além dos quatro tipos de compressores analógicos, existem, como mencionado, os compressores digitais, que tem seu funcionamento ditado por operações matemáticas, sendo altamente customizáveis e precisos (IZHAKI, 2017). Atualmente, além da vasta gama de compressores digitais disponíveis no mercado, os principais compressores analógicos foram emulados em *plugins*, trazendo toda a praticidade de uma ferramenta digital, mas buscando emular as características sonoras que por tantos anos foram apreciadas (OWSINSKI, 2014).

Com tantos tipos de ferramentas para compressão, cabe à(ao) engenheira(o) de mixagem escolher a mais adequada de acordo com as características sonoras do elemento a ser processado, a finalidade do tratamento de compressão e o contexto no qual o trabalho se insere.

O uso dos compressores é bastante vasto e dentre as suas várias utilizações, podemos destacar:

- Acentuar detalhes sonoros: comprimir o sinal para que algumas de suas nuances, que porventura sejam pouco perceptíveis, possam ficar mais evidentes (mais destacadas em intensidade).
- Balancear volume: diminuir a variação da amplitude dinâmica dos sinais de maneira a ter um maior controle do volume dos elementos durante a mixagem.
- Aumentar o volume percebido (*loudness*): comprimir um som permite manter o nível de pico do sinal e ao mesmo tempo aumentar o nível médio o que proporciona uma sensação maior de *loudness*.
- Alterar a envoltória dinâmica: ao mudar o nível do sinal ao longo do tempo, altera-se a envoltória dinâmica de cada som, podendo enfatizar ou atenuar seu transitório de ataque, por exemplo.
- Adicionar conteúdo frequencial: Como mencionado, alguns modelos analógicos de compressores possuem como característica a alteração do timbre do sinal processado, o que pode ser desejado em alguns contextos (IZHAKI, 2017).

2.3.6 *Limiters*

Outra ferramenta amplamente usada que manipula a amplitude dinâmica do sinal é o *Limiter* (limitador). O *limiter* pode ser considerado um tipo de compressor com uma razão de compressão bastante elevada (acima de 10:1), o que na prática significa

que uma vez selecionado o *threshold* o sinal de saída não ultrapassará esse valor, assim limitando o nível do sinal (OWSINSKI, 2014).

2.3.7 *Delay* (Atraso)

O *delay*, no contexto da mixagem, tem como função a de criar um efeito de eco. Os dois principais parâmetros em um *delay* são o tempo de atraso, que determina o tempo entre o sinal original e a primeira repetição, e o *feedback*, que determina o ganho no loop de realimentação, que vão decrescendo em intensidade (SENIOR, 2018). De modo geral os *delays* podem ser classificados em: curtos, com tempo de atraso entre 40-150 ms; médios, entre 150-400 ms e os longos, acima de 400 ms.

As possibilidades e situações de aplicação são vastas quando se trata do uso de *delays*, dependendo do estilo musical e do gosto pessoal da(o) engenheira(o) de mixagem. A funcionalidade do *delay* gira em torno da criação de uma ambientação acústica, adicionando sensação de espaço e profundidade; do posicionamento espacial um elemento de mixagem dentro da imagem sonora total; bem como de explorar ou preencher eventuais vazios, brechas ou silêncios (IZHAKI, 2017).

2.3.8 *Reverb* (Reverberação)

Para Pujahari (2017): 'A reverberação é (...) causada pela propagação e pela decorrente multiplicidade de reflexões (de modo difuso) dentro de um espaço estanque.' A diferença entre *delay* e reverberação é que o *delay* é a simples repetição do sinal original, já a reverberação leva em conta o comportamento do som no ambiente no qual está naturalmente ou artificialmente inserido. Os primeiros usos de reverberação artificial para fins musicais tinham por objetivo adicionar realismo às gravações feitas em estúdio. Porém, ao longo do tempo, o *reverb* se tornou também um recurso criativo capaz de simular respostas de salas existentes no mundo físico e também de salas imaginárias (SENIOR, 2018). Existem cinco principais tipos de simulação de reverberação: *hall*, sala, câmara, placa (*plate*), não-linear e mola.

Os *reverbs* de *hall*, sala e câmara são os tipos naturais, que simulam salas acústicas. Já os tipos placa, mola e não-linear são artificiais. O tipo *plate* é obtido através da vibração de uma placa metálica suspensa tendo nas extremidades transdutores elétricos de sinal sonoro, induzindo na placa o sinal direto e captando o retorno difuso. A reverberação de mola (*spring reverb)* funciona através da passagem do som pela extremidade de uma mola, que é capturada por um transdutor na outra extremidade. Já o *reverb* não-linear é um produto da era digital, permitindo uma customização completa dos parâmetros e

a simulação de respostas acústicas de salas que talvez jamais poderiam ser construídas (OWSINSKI, 2014).

Os parâmetros mais comuns de controle da reverberação são: quantidade de som direto em relação à quantidade de som reverberante, pré-delay, tempo de decaimento (tempo de reverberação – T60), tamanho da sala (*size*), densidade, difusão e equalização da reverberação.

O som direto é o próprio sinal original, de entrada, antes de ser reverberado. O pre-delay se refere ao tempo entre o sinal direto e a primeira reflexão. O tempo de decaimento é o tempo de reverberação, que é o tempo necessário para que as reflexões diminuam de 60 dB de nível em relação ao som direto. O tamanho da sala (*size*) diz respeito justamente ao espaço físico que está sendo simulado (ao volume da sala). Os parâmetros de densidade e difusão dizem respeito ao comportamento das reflexões em campo direto e em campo reverberante, e a equalização da reverberação uma simulação da absorção em função da frequência das ondas sonoras por parte dos materiais de revestimento da sala simulada (PUJAHARI, 2017).

A funcionalidade da reverberação em uma mixagem pode ter várias nuances, desde adequar o som à estética de um gênero musical específico, à criação de um espaço sonoro, até resolver problemas técnicos que podem surgir durante uma mixagem. Dentre as principais aplicações do *reverb* podemos citar: simular profundidade espacial, dar mais realismo à escuta de uma performance musical, adequar o espaço de escuta para uma determinada música ou produção sonora; mudar o timbre de um elemento; criar um efeito de dramatização; salientar ou destacar um elemento; posicionar um elemento na imagem sonora estereofônica; ou adicionar coesão a um elemento dentro do contexto da mixagem (IZHAKI, 2017).

2.3.9 Distorção

Vista muitas vezes como indesejável no processo de gravação, a distorção é uma ferramenta de muito valor na mixagem, permitindo a adição de conteúdo em termos de timbre a um som enquanto mantém as características do elemento sonoro sendo mixado (SAVAGE, 2014). A distorção era uma característica muito presente na era analógica do áudio, apesar de tecnicamente seu controle em produção ter sempre sido uma questão problemática. Porém quando aplicada de maneira coerente e dosada na medida certa, a distorção imprimia características timbrísticas expressivas e interessantes, o que fez com que seu uso fosse assimilado e mantido na era digital (SENIOR, 2018).

3 Sistemas Inteligentes de Mixagem

Tendo estabelecido um panorama histórico da mixagem, bem como introduzido alguns dos conceitos e ferramentas que compõem essa prática, explicita-se ainda mais a sua natureza complexa. A formação de um profissional da mixagem com habilidades de escuta refinadas, portanto, leva anos. Um problema, então, surgiu nas duas últimas décadas: apesar do advento do áudio digital ter tornado os meios de produção musical muito mais acessíveis, permitindo com que uma quantidade cada vez maior de músicas sejam gravadas e produzidas, a formação de bons profissionais da mixagem não acompanhou essa demanda (VANKA *et al.*, 2023b).

A busca por uma maior investigação sobre a complexidade da mixagem bem como a procura por suprir a falta de profissionais qualificados motivou tentativas de automatizar partes do processo ou até mesmo o processo da mixagem como um todo (MAN *et al.*, 2019). A primeira vez que o termo Mixagem Automática foi empregado dentro desse contexto foi em 2007 num estudo propondo um sistema baseado em regras para automatizar a panoramização dos elementos em uma mixagem ao vivo (GONZALEZ; REISS, 2007). Hoje em dia, esse campo de estudos é denominado tanto pelo termo Mixagem Automática quanto pelo termo Sistemas Inteligentes de Mixagem (Intelligent Mixing Systems) (MOFFAT, 2021).

As primeiras abordagens desse campo de estudo buscaram o desenvolvimento de sistemas especialistas, se valendo de conceitos da psicoacústica (MAN *et al.*, 2017). Mais recentemente, com o grande desenvolvimento do campo da Inteligência Artificial (IA), os estudos se voltaram para abordagens direcionadas para dados (*data-driven*) utilizando técnicas de aprendizado de máquina e aprendizado profundo (MOFFAT, 2021). Até onde sabemos, ainda não existem ferramentas para mixagem que utilizam Inteligência Artificial Generativa.

Dos estudos envolvendo Inteligência Artificial, alguns buscaram a automatização do processo como um todo, ou seja, sistemas nos quais o arquivo multicanal é fornecido e como saída se tem a faixa estéreo mixada (MOFFAT; SANDLER, 2019; MARTINEZ-RAMIREZ *et al.*, 2021a; STEINMETZ *et al.*, 2021; COLONEL; REISS, 2021; MARTÍNEZ-RAMÍREZ *et al.*, 2022; KOO *et al.*, 2023; KOSZEWSKI *et al.*, 2023). Já outros buscaram utilizar modelos de IA para modelar e automatizar o funcionamento de certas ferramentas ou de certos processos dentro da mixagem (MARTINEZ-RAMIREZ *et al.*, 2018; SHENG; FAZEKAS, 2019; NERCESSIAN, 2020; KUZNETSOV *et al.*, 2020; STEINMETZ; REISS, 2021).

3.1 Categorias de Usuários

Um aspecto que norteia o desenho e o desenvolvimento de Sistemas Automáticos de Mixagem é o usuário final. Antes do advento do áudio digital, o processo de produzir e mixar uma música era restrito aos estúdios musicais profissionais, logo as ferramentas de mixagem possuíam apenas um público-alvo: a(o) engenheira(o) de mixagem profissional. Hoje em dia, no entanto, qualquer pessoa com aceso à internet pode instalar uma estação digital de áudio (*DAW*) e pode acessar diferentes cursos e conteúdos sobre mixagem (VANKA *et al.*, 2023b).

Considerando esse cenário, atualmente podemos elencar três principais categorias de usuários de ferramentas tradicionais de mixagem e que, portanto, também são o público-alvo dos Sistemas Automáticos de Mixagem: amadores, entusiastas e profissionais (VANKA *et al.*, 2023a).

O(a) amador(a), dentro do contexto aqui abordado, é a pessoa que não tem experiência com mixagem, mas que geralmente tem alguma outra relação com a música, como instrumentista ou produtor(a) musical. Também pode se tratar de uma pessoa que acabou de ter contato com a mixagem, mas ainda não sabe o suficiente para executar a mixagem de uma música ou até mesmo de uma pessoa que é apenas consumidora de música.

Já a classe de entusiastas, comporta as pessoas que, graças à internet, possuem acesso à conteúdo extensivo sobre mixagem e podem até cobrar pelos serviços de mixagem, mas que ainda não são profissionais ou não pretendem se tornar profissionais. O(A) entusiasta pode conhecer os principais jargões da mixagem, bem como o funcionamento das principais ferramentas e pode até possuir certas habilidades de escuta crítica.

Por fim, a classe de profissionais agrupa as pessoas que estudaram extensivamente o assunto e que se dedicam profissionalmente a serviços de mixagem. Essas pessoas geralmente trabalham em estúdios caseiros (*home studios*) ou estúdios profissionais e possuem bons equipamentos de monitoração. Essas pessoas também possuem habilidades de escuta bastante refinadas e sabem utilizar as ferramentas de áudio com autonomia.

Uma vez detalhadas as três categorias de usuários, o desenvolvimento de estudos no campo da Mixagem Automática pode beneficiá-los de maneiras diferentes. Na próxima sessão serão abordados os graus de automatização que um Sistema Automático de Mixagem pode adotar e qual seria o público-alvo de cada um deles.

3.2 Graus de Automatização

No desenvolvimento de Sistemas Inteligentes de Mixagem, um fator a ser considerado é a interação computador-humano. Segundo Moffat (2021), são estabelecidos, portanto, quatro **graus de automatização**: automático, independente, recomendativo e exploratório.

No grau automático, o sistema recebe as faixas de áudio e é esperado que retorne a mixagem final, sem nenhuma ou com mínima intervenção humana. A pessoa operadora desse sistema pode interagir de forma mínima como, por exemplo, selecionando o gênero da música a ser mixada, porém não possuí controle sobre o produto final apresentado. Ferramentas com alto grau de automatização podem beneficiar usuários com pouca ou nenhuma experiência com mixagem.

No grau independente, a operadora do sistema se coloca na posição de supervisora, validando as ações tomadas pelo sistema. Uma interação possível nesse grau seria a de solicitar a automatização de uma tarefa, como por exemplo a equalização de um elemento musical, e então poder alterar o resultado emergente ou até mesmo desfazer o processamento feito pelo sistema de mixagem. O grau independente pode beneficiar principalmente pessoas com nível intermediário de conhecimento em mixagem, que podem não saber realizar ainda certos processamentos.

No grau recomendativo, o sistema analisa e interpreta em tempo real as decisões tomadas pela(o) engenheira(o) de mixagem e retorna sugestões de processamento do áudio. Sistemas com esse grau de automatização podem, por exemplo, sugerir algum processamento específico como recomendar o aumento no volume de determinados elementos. Cabe ao operador do sistema acatar ou não as sugestões do sistema recomendativo. Esse tipo de sistema pode beneficiar todo tipo de público. Enquanto amadores e entusiastas podem utilizar sistemas recomendativos como maneira de aprendizado, profissionais podem utilizá-los como maneira de validarem suas mixagens.

O grau exploratório tem como objetivo o de fornecer visualizações e informações que podem ajudar no processo de tomada de decisão do humano. Neste grau, o sistema não tem capacidade de processar o áudio, apenas de analisá-lo e compará-lo de acordo com referências pré-estabelecidas pelo operador. Ferramentas exploratórias podem auxiliar tanto amadores quanto entusiastas que ainda não são capazes de detectar problemas em suas mixagens apenas por suas habilidades de escuta.

Um ponto de destaque a ser feito sobre ferramentas de mixagem automática dos graus exploratório, independente e recomendativo é que já estão comercialmente disponíveis muitas ferramentas com esse grau de automatização que empregam algum tipo

de Inteligência Artificial no seu desenvolvimento. A empresa estadunidense *iZotope*[1], por exemplo, é muito conhecida por produzir ferramentas inteligentes para produção, mixagem e masterização, que são amplamente utilizadas por profissionais. No entanto, nenhuma dessas ferramentas possuí código aberto ao público, o que impede que se compreenda por completo o funcionamento delas.

Tendo em vista cada grau de automatização aqui descrito, o campo de estudo da Mixagem Automática, portanto, se debruça sobre o desenvolvimento de soluções para os diversos tipos de público, pensando nos principais problemas que cada público enfrenta e qual grau de automatização será mais pertinente. Neste trabalho iremos nos aprofundar nas soluções do tipo automático que utilizam Inteligência Artificial, e mais especificamente, nas que são capazes de realizar a tarefa da mixagem em sua totalidade, ou seja, que recebem as faixas de áudio e retornam o *mixdown*. Na próxima seção iremos formalizar matematicamente o funcionamento desse tipo de Sistema Inteligente de Mixagem e apresentar as duas abordagens possíveis de processamento de áudio.

3.3 Arquitetura

De maneira geral, um sistema automático de mixagem recebe um conjunto de **N** faixas sonoras $x_1, x_2, ..., x_N$, cada uma contendo **T** amostras. Essas faixas serão processadas e combinadas em um produto final Y, $Y \in \mathbb{R}^{2\times T}$ (para uma mixagem estéreo) (STEINMETZ *et al.*, 2022).

Segundo a literatura, há duas maneiras principais para tratar o problema da mixagem no âmbito da Inteligência Artificial: Transformação Direta e Estimativa de Parâmetros (STEINMETZ *et al.*, 2022) (MOFFAT, 2021). Um esquemático das duas abordagens pode ser observado na Figura 3.1 e a seguir cada abordagem será descrita.

3.3.1 Transformação Direta

Na transformação direta o processamento das faixas é feito diretamente no domínio do áudio ou então pela modificação do áudio em algum domínio reversível, como em uma transformada de Fourier de tempo curto ou em espectrogramas (MOFFAT, 2021). Nessa abordagem, não é utilizado um intermediário interpretável, diferentemente dos modelos baseados em estimativa de parâmetros.

Em termos de treinamento do modelo, na transformação direta é necessário ter acesso apenas às faixas musicais sem processamento e ao arquivo da mixagem final, não sendo necessário ter conhecimento sobre o processamento que foi aplicado. É, portanto,

[1] https://www.izotope.com/

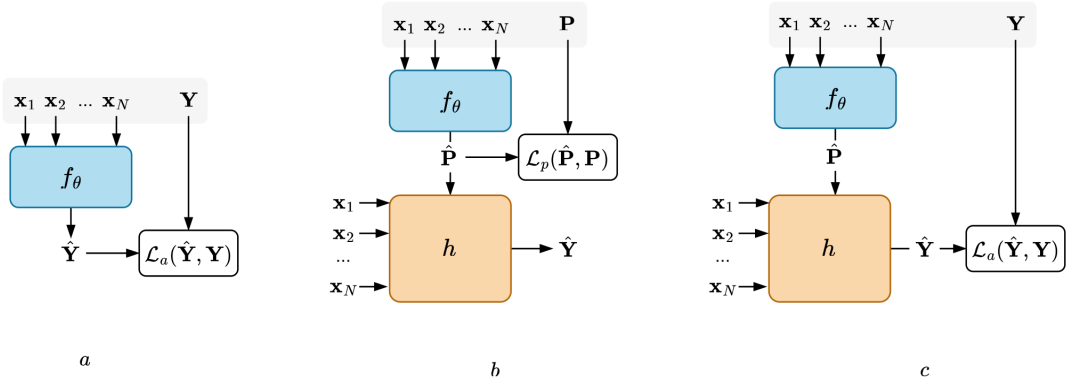

Figura 3.1 – Representação das abordagens: (a) transformação direta, (b) estimativa de parâmetros com parâmetros reais disponíveis e (c) estimativa de parâmetros sem parâmetros reais disponíveis. Fonte: (STEINMETZ *et al.*, 2022).

construído um modelo f_θ que tem como entrada o conjunto de N faixas e é treinado para produzir uma estimativa do produto final da mixagem $\hat{Y}$ que seja a mais próxima possível do Y correspondente. Isso envolve otimizar os parâmetros θ do modelo de acordo com uma função custo (*loss*) que mede a distância $\mathcal{L}(\hat{Y}, Y)$. Essa distância pode ser computada no domínio do tempo ou da frequência. Na subseção 3.5 serão aprofundadas as principais funções custo utilizadas para treinar modelos de áudio.

O ponto positivo desse método é que ele é flexível, uma vez que não estabelece premissas sobre cada mixagem, porém necessita de grandes quantidades de dados para um treino adequado e pode gerar transformações indesejadas no sinal de entrada (artefatos). Outros pontos negativos são a falta de interpretabilidade e controle, já que essa é uma abordagem de "caixa preta" e não permite o ajuste posterior do processamento aplicado (STEINMETZ *et al.*, 2022).

Entre as aplicações prévias utilizando transformação direta para a mixagem podemos citar a arquitetura *Mix-Wave-U-Net*, proposta por Martinez-Ramirez *et al.* (2021a) e revisada por Koszewski *et al.* (2023). Essa arquitetura é baseada na arquitetura *Wave-U-Net* proposta anteriormente por Stoller *et al.* (2018) para separação de fontes sonoras, com adaptações para incluir um arquivo de múltiplas faixas na entrada e uma saída estéreo.

A modelagem *Mix-Wave-U-Net* consiste em uma rede *encoder-decoder*, onde no codificador o sinal sofre uma redução de resolução e, em cada etapa de redução, informações são retiradas e compartilhadas com a etapa correspondente no decodificador, que consiste em uma série de etapas que aumentam a resolução do dado. Um ponto importante dessa implementação é que o modelo aceita um número fixo pré determinado de entradas. Mais detalhes da arquitetura estão representados na Figura 3.2.

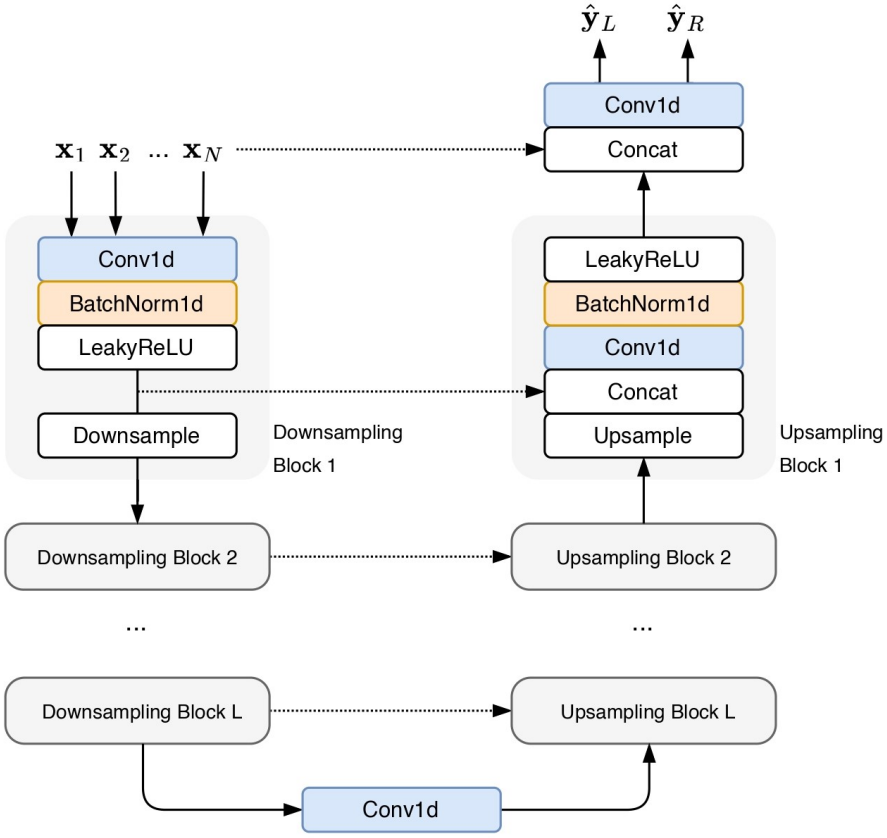

Figura 3.2 – Arquitetura proposta na abordagem Mix-Wave-U-Net. Fonte: (STEINMETZ *et al.*, 2022).

Outra aplicação para mixagem de um modelo de transformação direta foi proposta por Martínez-Ramírez *et al.* (2022) baseada no trabalho previamente desenvolvido para modelagem de efeitos de áudio (MARTINEZ-RAMIREZ *et al.*, 2020). A arquitetura proposta é dividida em três partes: um *front-end* adaptativo, um *mixer* no espaço latente e um *back-end* de síntese. Uma representação visual da arquitetura pode ser encontrada na Figura 3.3.

No *front-end* o áudio de entrada sofre convoluções no domínio do tempo, de modo a aprender uma representação latente e também gera uma conexão residual que será utilizada no *back-end* para facilitar a síntese do áudio final. O *mixer* do espaço latente é composto por uma Rede Convolucional Temporal (TCN) seguida de uma Rede Neural Recorrente Bidirecional (BLSTM), de modo a melhorar o aprendizado de dependências de longo-prazo. O *back-end* de síntese realiza a reconstituição do sinal modificado através de uma Rede de Compressão e Excitação (*Squeeze-and-Excitation Networks*) e uma convolução.

Uma outra contribuição proposta por Martínez-Ramírez et. al. nesse mesmo estudo foi um método para adaptar conjuntos de dados utilizados em problemas de sepa-

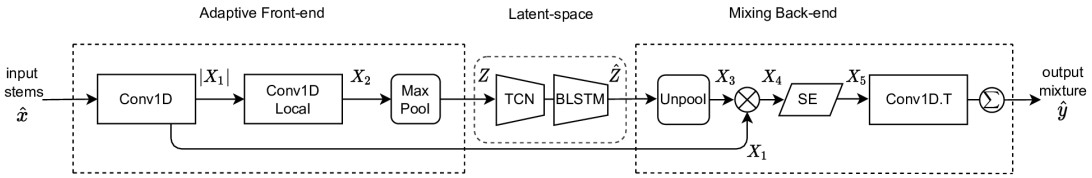

Figura 3.3 – Arquitetura proposta por Martínez-Ramírez et. al.Fonte: (MARTÍNEZ-RAMÍREZ *et al.*, 2022).

ração de fontes sonoras para serem reaproveitados no treinamento de modelos de mixagem automática. Dado que nos conjuntos de dados de problemas de separação de fontes as faixas com os elementos separados já estão processadas, foi proposto nesse estudo um método de pré-processamento dessas faixas de modo a normalizar algumas métricas sonoras (reverberação, equalização, alcance dinâmico, *loudness* e panoramização). Essa normalização nas faixas individuais, segundo o estudo, permite que um modelo de mixagem automática possa inferir o processamento empregado, já que a soma das faixas pré-processadas diferem da mixagem final.

3.3.2 Estimativa de Parâmetros

Nessa abordagem, é construído um modelo f_θ que estima os parâmetros de um console de mixagem pré-estabelecido. Esse console é descrito como uma função h que tem como argumentos o conjunto de faixas N e o conjunto de parâmetros P utilizados em cada efeito para cada faixa para produzir a faixa final $Y = h(x_1, x_2, ..., x_N, p_1, p_2, ..., p_N)$. Na maioria das configurações, é assumido que a cadeia de processamento para cada canal é a mesma e na mesma ordem, por exemplo o sinal passa por um *fader* de volume, um equalizador e um *reverb*, cada elemento com seus parâmetros individualizados para cada faixa, porém sem a possibilidade de adicionar mais efeitos. Dada essa limitação de assumir o mesmo processamento para cada faixa, esse método limita a expressividade do sistema se comparado ao método da transformação direta; no entanto, reduz o número de tarefas realizadas pelo modelo f e promove mais controle e interpretabilidade, já que é uma abordagem "caixa transparente" e permite o conhecimento dos parâmetros aplicados à cada efeito.

Essa abordagem possui duas formas de implementação, uma na qual os parâmetros de referência, ou seja, as configurações dos efeitos utilizados pela pessoa que realizou a mixagem são conhecidos e outra quando os parâmetros de referência não são conhecidos. No caso onde os os parâmetros utilizados para produzir a mixagem estão disponíveis, é possível projetar um modelo que aprenda a estimar o conjunto de parâmetros $\hat{P}$ através do calculo da distância entre os parâmetros de referência e preditos $\mathcal{L}(\hat{P}, P)$. Os principais empecilhos dessa abordagem são a falta de dados disponíveis e também a

falta de generalização do modelo. Para o outro caso, quando os parâmetros de referência não estão disponíveis e estão disponíveis apenas as trilhas antes do processamento e após o processamento, o treino do modelo envolve estimar os parâmetros $\hat{P}$ que mapeiam as faixas de entrada ao produto final da mixagem $\hat{Y}$ o aproximando de Y, minimizando portanto a função custo $\mathcal{L}(\hat{Y}, Y)$ (STEINMETZ *et al.*, 2022).

Uma implementação de sistema automático de mixagem utilizando a estratégia da estimativa de parâmetros denominada Console de Mixagem Diferenciável (*Differenciable Mixing Console)* foi proposta por Steinmetz *et al.* (2021). Nessa abordagem uma rede neural é empregada para estimar os parâmetros de um console de mixagem previamente estabelecido (Figura 3.4). A arquitetura consiste em três sub-sistemas, cada um implementado como uma rede neural separada: codificador (*encoder*), pós-processador e rede de transformação, descritos a seguir e ilustrados na Figura 3.5.

O codificador tem como objetivo extrair informações relevantes de cada faixa de entrada utilizando uma rede neural convolucional (*CNN*), produzindo um *embedding* para cada segundo de cada faixa sonora (*embedding* de entrada). *Embedding* é o nome dado para uma representação densa e de baixa dimensionalidade de um dado. Essas representações então são combinadas em um *embedding* de contexto que captura informações sobre todas as faixas, permitindo o acúmulo de informações sobre o conteúdo que está sendo mixado como um todo.

O pós-processador consiste em uma rede neural *perceptron* multicamadas (*MLP*) onde a entrada é a concatenação dos *embedding* de entrada com o *embedding* de contexto. Essas entradas são mapeadas para os parâmetros de controle de cada canal respectivo do console de mixagem. Por último, a rede de transformação recebe o áudio de entrada para cada faixa e os parâmetros preditos no pós-processador para manipular o áudio e criar o *mixdown final*. Os efeitos de áudio foram implementados como *neural proxies*, ou seja, foi utilizada uma rede neural, treinada para emular o funcionamento do console de mixagem proposto na Figura 3.4.

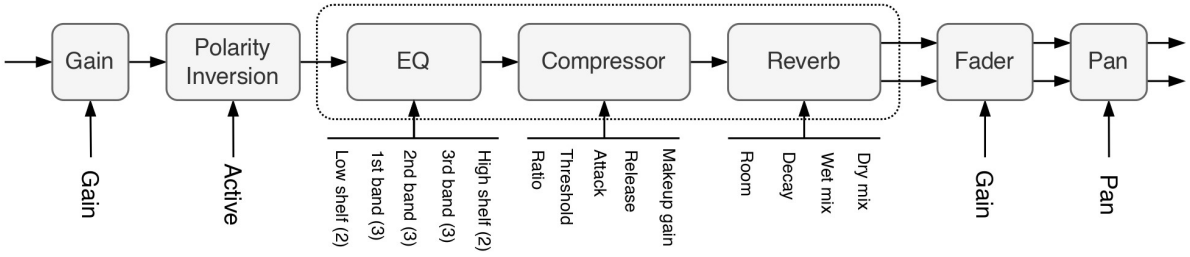

Figura 3.4 – Console de mixagem assumido nessa implementação como rede de transformação. Fonte: (STEINMETZ *et al.*, 2022).

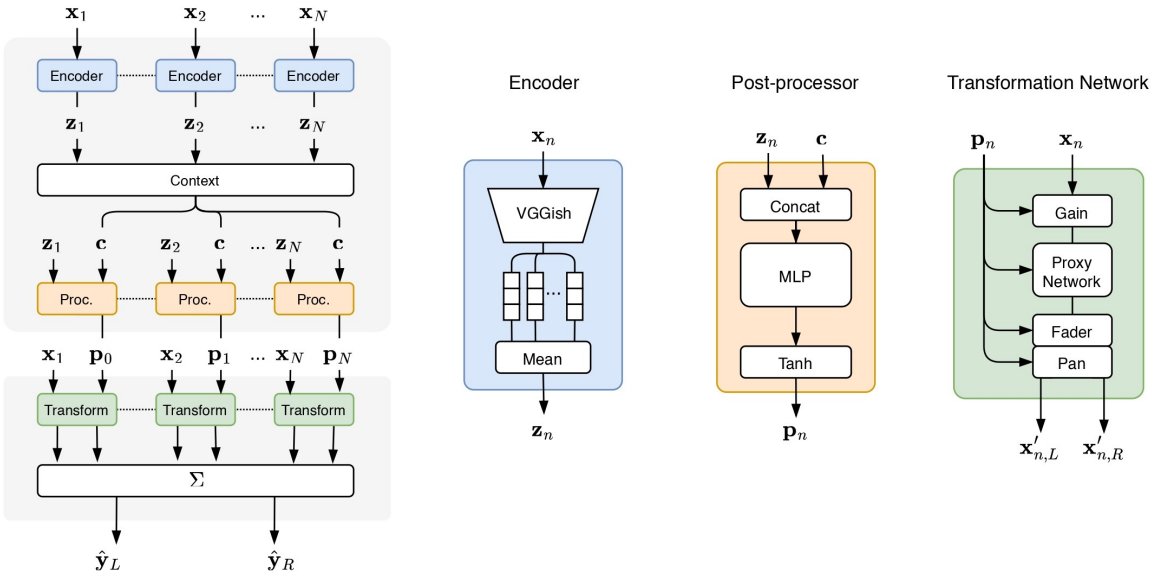

Figura 3.5 – Arquitetura dos subsistemas do DMC. Fonte: (STEINMETZ *et al.*, 2022).

3.4 Inteligência Artificial Generativa

Um modelo de Inteligência Artificial Generativo pode ser descrito, de forma ampla, como um modelo capaz de sintetizar um dado novo que seja estatisticamente provável de existir, dado o conjunto de dados no qual o modelo foi treinado (FOSTER, 2022). Mesmo que o campo de estudo da IA Generativa tenha se iniciado nos anos 1980, apenas em 2017, com a introdução da arquitetura *transformers*, que o enorme potencial tecnológico dessa área foi revelado e a aplicação de abordagens generativas foram estendidas para uma série de tarefas, incluindo atividades no domínio do áudio como conversão texto-fala, síntese sonora e realce de fala (BARNETT, 2023).

Em 2022, Steinmetz *et al.* (2022) propuseram a seguinte formulação para o problema da mixagem: um modelo generativo $p(Y|x_1, x_2, ..., x_N)$ pode ser construído de forma a capturar o processo de produção de uma mixagem dadas as faixas de entrada e suas respectivas mixagens. Embora tenha sido formulada, até o momento, ao melhor do nosso conhecimento, nenhuma abordagem generativa de grau automático para a tarefa da mixagem foi explicitamente desenvolvida. A seguir, serão apresentadas algumas classes de modelos generativos e algumas aplicações previamente propostas que podem ser adaptadas para a mixagem.

3.4.1 Autoencoder Variacional

A arquitetura do AutoEncoder Variacional (VAE) é uma derivação da arquitetura do Autoencoder, onde um encoder é responsável por reduzir a dimensionalidade do dado e, em seguida, um decoder tenta reconstituir o dado. A diferença entre o AutoEnco-

der tradicional e o variacional é que o VAE realiza a codificação entre o sinal original e sua representação em forma de uma distribuição probabilística, o que possibilita a geração de dados que são similares estatisticamente ao conjunto de dados de treinamento (FOSTER, 2022).

Uma aplicação relevante da arquitetura VAE previamente proposta no domínio do áudio inclui o modelo RAVE proposto por Caillon e Esling (2021). O objetivo do modelo RAVE é o de sintetizar sinais de áudio musicais, utilizando uma conjunção da arquitetura VAE com a arquitetura GAN, que será aprofundada na sessão 3.4.2. A seguir, o funcionamento dos AutoEncoders Variacionais será aprofundado, bem como a arquitetura RAVE.

Em termos matemáticos, um modelo generativo tem como objetivo o de mapear um conjunto de dados $x \in \mathbb{R}^{d_x}$ através da modelagem da sua distribuição $p(x)$. Podemos considerar que a geração de x está condicionada pelas variáveis latentes $z \in \mathbb{R}^{d_z}$, responsáveis pelas variações presentes em x. O modelo então é definido pela distribuição conjunta $p(x, z) = p(x|z)p(z)$, cuja solução não é trivial dada a grande complexidade do conjunto de dados. Os AutoEncoders Variacionais buscam solucionar esse problema pela introdução de um modelo de inferência $q_\phi(z|x)$, otimizado para minimizar a divergência de Kullback-Leibler, $\mathcal{D}_{KL}$, ou seja a diferença entre a distribuição verdadeira $p(z|x)$ e a distribuição inferida $q_\phi(z|x)$:

$$\Phi^* = \underset{\phi}{\operatorname{argmin}}\, \mathcal{D}_{KL}[q_\phi(z|x) \parallel p(z|x)], \tag{3.1}$$

que pode ser rearranjada para obter o objetivo final utilizado para treinar um modelo do tipo AutoEncoder Variacional, chamada de limite inferior da evidência (do inglês *Evidence Lower Bound (ELBO)*):

$$\mathcal{L}_{\phi,\theta}(x) = -\mathbb{E}_{q_\phi(z|x)}[log(p_\theta(x|z))] + \mathcal{D}_{KL}[q_\phi(z|x) \parallel p(z|x)]. \tag{3.2}$$

A equação em 3.2 minimiza a reconstituição do erro do modelo através da probabilidade do conjunto de dados quando fornecido um logaritmo latente $log(p_\theta(x|z))$, enquanto regulariza a distribuição posterior $q_\phi(z|x)$ para seguir uma distribuição previamente definida $p(z)$. Tanto q_ϕ quanto p_θ são distribuições posteriores parametrizadas por redes neurais, chamadas respectivamente de codificador e decodificador.

Na arquitetura RAVE, o codificador consiste na transformação do áudio em uma representação latente de 128 dimensões. Isso é atingido através de uma decomposição do sinal de entrada em 16 bandas de frequência, seguido de uma rede neural convolucional com função de ativação do tipo *Leaky ReLu*. Já o decodificador, inspirado na arquitetura *Wave-U-Net*, realiza o *upsampling* do sinal; porém, em vez de devolver diretamente a forma de onda, o sinal passa por três sub-redes diferentes. A primeira sub-rede é responsável por

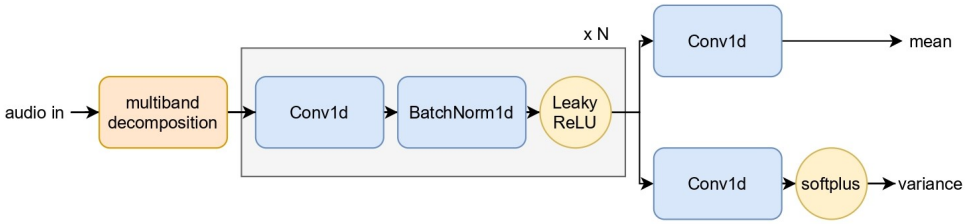

Figura 3.6 – Arquitetura do *Encoder* do modelo RAVE. Fonte: (CAILLON; ESLING, 2021).

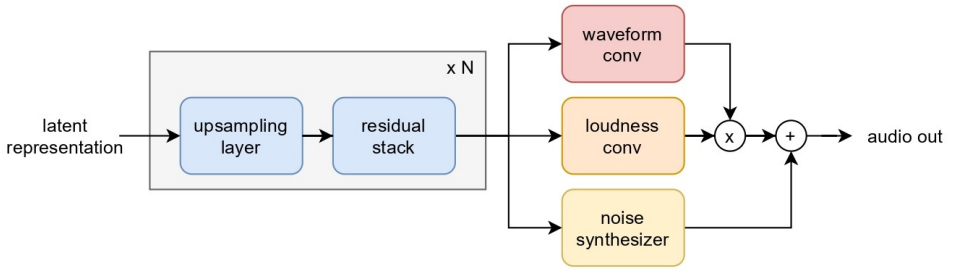

Figura 3.7 – Arquitetura do *Decoder* do modelo RAVE. Fonte: (CAILLON; ESLING, 2021).

gerar a forma de onda juntando todas as 16 bandas de frequência. Em seguida, o resultado da primeira sub-rede é multiplicado pela saída da segunda sub-rede, responsável por definir o nível de audibilidade (*loudness*) do sinal. Finalmente, a última sub-rede adiciona um pouco de ruído em frequência ao sinal. Na Figura 3.6 está um esquemático da arquitetura do codificador e na Figura 3.7 está um esquemático do decodificador.

O treino do modelo se dá em duas etapas: aprendizado de representação e *fine-tuning* adversarial. Essa maneira de treinamento ajuda o AutoEncoder Variacional a aprender uma representação que inclua atributos de alto-nível, em vez de incluir no espaço latente variações de baixo-nível que não são relevantes para a audição humana.

Na primeira etapa, de aprender a representação, tanto codificador quando decodificador são treinados com uma função de perda que mistura *ELBO* com a distância no domínio da frequência (distância espectral multibanda) entre os sinais reais e os sinais sintetizados. Na segunda etapa, apenas o decoder é treinado utilizando um objetivo adversarial (Redes Generativas Adversariais serão aprofundadas na sessão 3.4.2).

O modelo RAVE pode ser treinado para uma série de tarefas, uma delas é a transferência de timbre. Dados dois conjuntos de dados, o modelo consegue transformara gravação de um violino, por exemplo, em um sinal similar porém como se a mesma linha melódica estivesse sendo tocada por um outro instrumento, como uma guitarra. Dada essa natureza da arquitetura, algumas adaptações poderiam ser feitas de modo a permitir que o modelo gere uma mixagem, tendo um arquivo multicanal de faixas não processadas.

3.4.2 Redes Adversariais Generativas

Uma Rede Adversarial Generativa (GAN, do inglês *generative adversarial network*) é uma arquitetura baseada na interação entre duas redes neurais: a rede neural geradora (gerador) e a rede neural discriminadora (discriminador). O papel do gerador é o de, a partir de um ruído aleatório, gerar um dado que seja compatível com os dados presentes na base de dados, e o discriminador tem como objetivo identificar se o dado observado é realmente pertencente à base de dados ou se é um dado forjado pelo gerador. Nessa interação, a rede geradora fica progressivamente mais competente em sintetizar dados que sejam similares aos reais e a rede discriminatória fica melhor em reconhecer dados gerados (FOSTER, 2022). A arquitetura GAN foi primeiramente proposta por Goodfellow *et al.* (2014) e sua formulação será detalhada a seguir.

De modo a aprender a distribuição p_g do conjunto de dados x, é definida previamente uma distribuição p_z de um ruído variável z. Então um mapeamento para o espaço do conjunto de dados é representado por $G(z;\theta_g)$, onde G é uma função diferenciável com parâmetros θ_g. Uma segunda função $D(x;\theta_d)$ é definida, cuja saída é um escalar. $D(x)$, portanto, representa a probabilidade de x ser pertencente ao conjunto de dados em vez de ser pertencente à p_g. O treino de D é feito de maneira a maximizar a probabilidade de acertar o rótulo (real ou gerado) tanto nos dados do conjunto de treino, quanto nos dados gerados por G. Simultaneamente, G é treinada para minimizar $log(1 - D(G(z)))$. Temos portanto:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\ p_{data}(x)}[logD(x)] + \mathbb{E}_{z\ p_z(z)}[log(1 - D(G(z)))]. \tag{3.3}$$

Uma aplicação prévia de GAN para uma tarefa de áudio foi proposta por Chen *et al.* (2022) para realizar transições de música de maneira automática. No universo da discotecagem, a figura da(o) DJ realiza transições entre músicas com o uso de algumas ferramentas que também são utilizadas na mixagem musical multicanal. Importa diferenciar que, no contexto da discotecagem, a palavra mixagem também é aplicada para se referir ao ato de realizar a transição de uma música para a outra, porém essa prática não se assemelha à prática da mixagem multicanal que é centro deste trabalho.

No trabalho desenvolvido por Chen et. al. foi proposta a arquitetura DJtransGAN, onde, a partir de um par de músicas(x_1, x_2) e um conjunto de equalizadores e *faders* de volume, o modelo aplica efeitos de áudio em cada música e depois soma as duas músicas para obter a transição x_3. Para o desenvolvimento do modelo DJtransGAN, foi necessário um conjunto de dados de transições de músicas feitas por DJs humanos profissionais rotulados como dados verdadeiros, para assim utilizar a estratégia de treino adversarial.

O Gerador (G) foi construído numa arquitetura baseada no Console de Mixagem Diferencial proposta por Steinmetz *et al.* (2021) e previamente detalhada na Subseção

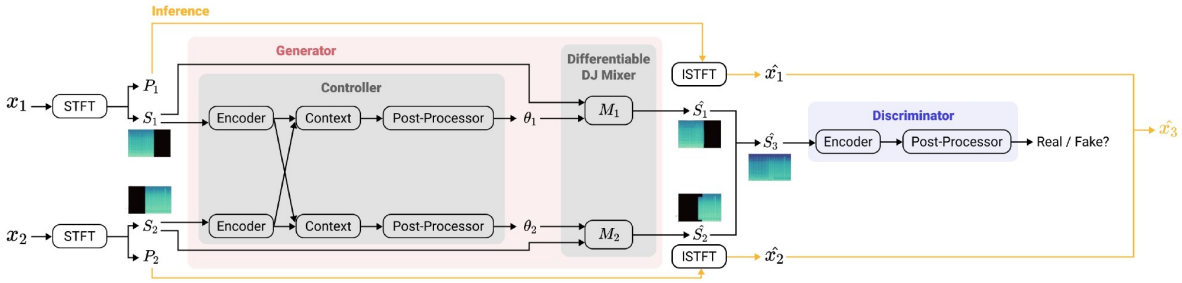

Figura 3.8 – Esquemático da arquitetura DJtransGAN. Fonte: (CHEN *et al.*, 2022).

3.3.2, onde o objetivo é encontrar os parâmetros θ_1, θ_2 que controlam os equalizadores e o *fader* de volume de cada música. A entrada do gerador recebe os pares de trechos de músicas pré estabelecidos (x_1, x_2) e realiza uma Transformada de Fourier de Tempo Curto nesse par para obter seus respectivos espectrogramas (S_1, S_2) e suas informações de fase (P_1, P_2).

A saída do gerador constitui nos espectrogramas devidamente processados pelos efeitos de áudio, $\hat{S}_1$ e $\hat{S}_2$ e, a partir das informações de fase P_1, P_2, os sinais de áudio são reconstituídos ($\hat{x}_1$ e $\hat{x}_2$). Os dois espectrogramas $\hat{S}_1$ e $\hat{S}_2$ então são somados em $\hat{S}_3$, que serve de entrada para o discriminador (D), que retorna o rótulo predito (real ou forjado). O treino do modelo então se dá por uma função de perda *min-max*, onde o discriminador tem como objetivo acertar a classe do sinal e o gerador tem como objetivo não ser detectado pelo discriminador. Mais detalhes da arquitetura podem ser vistos na Figura 3.8.

3.4.3 Modelos Autorregressivos

Modelos Autorregressivos são aqueles que tratam o processo de gerar dados como um problema sequencial, modelando a probabilidade condicional de um elemento dados os elementos anteriores (FOSTER, 2022). Essa família de modelos engloba arquiteturas bastante relevantes atualmente, em especial a arquitetura Transformers (VASWANI *et al.*, 2017) e o modelo GPT-3, que sustentou a primeira versão comercialmente lançada do *ChatGPT* (BROWN *et al.*, 2020).

Um modelo autoregressivo, portanto, é aquele que estima a densidade de probabilidade de um exemplo $X \in \mathcal{R}^M$ pela sua decomposição sequencial, utilizando a cadeia de probabilidade:

$$p(X) = \prod_{m=0}^{M-1} p(X_m | X_{<m}), \tag{3.4}$$

em que cada X_m é tipicamente um *token* (no universo do áudio podemos dizer que o *token* representa um grupo de amostras do áudio). Pela cadeia de probabilidade, temos que a

densidade da entrada X pode ser estimada através da estimativa sequencial da densidade condicional de todos os *tokens* que compõem X (HAWTHORNE *et al.*, 2022).

Dado que a tarefa de geração de música é uma tarefa sequencial, onde cada pedaço gerado depende do anterior para fazer sentido, existem algumas aplicações prévias de arquiteturas utilizando modelos autorregressivos para a síntese musical no domínio do áudio como o modelo *Jukebox* (DHARIWAL *et al.*, 2020), desenvolvido por trabalhadores da empresa *OpenAi* [2], capaz de gerar áudios em alta qualidade, a uma taxa de amostragem de 44,1 kHz.

O modelo *Jukebox* utiliza uma combinação das arquiteturas Autoencoder Variacional e Transformer. Enquanto a abordagem do autoencoder variacional é utilizada para aprender uma representação musical simplificada pela criação de *tokens*, a abordagem de transformers é utilizada para aprender as dependências temporais na música e assim aprimorar a geração sequencial. Quanto aos codificadores e decodificadores, o modelo prevê a codificação e decodificação em três resoluções diferentes: alto, médio e baixo nível.

O treinamento do modelo *Jukebox* se dá primeiro pelo treinamento não supervisionado, onde o modelo é treinado em um grande conjunto de músicas para aprender representações gerais de áudio. Em seguida, é feito um *fine-tuning* em conjuntos específicos de gênero, estilos e artistas. Para a geração de música é necessário que o usuário forneça uma descrição de áudio ou pedaço de áudio, o que condiciona a geração do modelo a determinados estilos musicais.

Uma observação pertinente é a de que modelos de Processamento de Linguagem Natural(PLN ou NLP, do inglês *Natural Language Processing*), que normalmente utilizam estruturas autoregressivas, podem ser treinados em conjuntos de dados específicos, de modo a extrair conhecimento em determinada área. A plataforma do *Chat GPT*, por exemplo, permite que usuários realizem esse processo, de modo a obter um assistente virtual especialista em um assunto. Existem alguns usuários que realizaram esse treinamento específico em textos relacionados à mixagem musical e disponibilizam essas versões online. Esses modelos, como são modelos de PLN, não são capazes de processar o áudio de maneira direta e nem de analisar o áudio sendo mixado, mas são capazes de dar conselhos gerais sobre mixagem.

3.4.4 Difusão

A classe de Modelos de Difusão tem seu funcionamento pautado em dois processos: aplicação de difusão e reversão de difusão. A aplicação de difusão consiste na

[2] https://openai.com/

introdução de ruído no conjunto de dados de treino de maneira gradual, até que os dados se tornem puro ruído Gaussiano. Já a reversão da difusão consiste na aplicação de um modelo capaz de reconhecer o ruído acrescentado e removê-lo (FOSTER, 2022).

Uma formulação possível para modelos de difusão é chamada de Modelos Probabilísticos de Remoção de Ruído de Difusão (DDPMs, do inglês *Denoising Diffusion Probabilistic Models*) (CROITORU *et al.*, 2023). Dado o conjunto de dados original x_0 com distribuição $p(x_0)$, podemos gerar as versões com ruído (corrompidas) $x_1, x_2, ...$ da seguinte maneira:

$$p(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta} \cdot x_{t-1}, \beta_t \cdot I\right), \forall t \in \{1, ..., T\}, \tag{3.5}$$

em que $\mathcal{N}(x; \mu, \sigma)$ representa a distribuição normal de média μ e covariância σ que produz x, T é o número de passos de difusão, $\beta_1, ..., \beta_T \in [0, 1)$ são hiperparâmetros representando a variância ao longo de cada passo e I é a matriz identidade contendo a mesma dimensão da entrada x_0. Dada a Equação 3.5, quando t pertence a uma distribuição uniforme, é possível formular a versão distorcida x_t partindo diretamente de x_0 tal que

$$p(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\hat{\beta}_t} \cdot x_0, (1-\hat{\beta}_t) \cdot I\right), \tag{3.6}$$

em que $\hat{\beta}_t = \prod_{i=0}^{t} \alpha_i$ e $\alpha_t = 1 - \beta_t$. De maneira geral, podemos normalizar uma amostra qualquer x de distribuição normal $x \sim \mathcal{N}(\mu, \sigma^2 I)$ subtraindo a média μ e dividindo pelo desvio padrão σ, resultando na amostra $z = \frac{x-\mu}{\sigma}$ de distribuição normal padrão $z \sim \mathcal{N}(0, I)$. Para obter x partindo de z, podemos realizar o inverso: multiplicar z por σ e somar μ. Dessa ideia podemos então amostrar x_t de $p(x_t|x_0)$ como

$$x_t = \sqrt{\hat{\beta}_t} \cdot x_0 + \sqrt{1-\hat{\beta}_t} \cdot z_t, \tag{3.7}$$

em que $z_t \sim \mathcal{N}(0, I)$. Dadas as Equações 3.5, 3.6 e 3.7, é possível gerar uma nova amostra de $p(x_0)$, partindo de uma amostra $x_T \sim \mathcal{N}(0, I)$. Podemos treinar uma rede neural que recebe como entrada um dado com ruído x_t e aprende a prever a média $\mu_\theta(x_t, t)$ e a covariância $\sum_\theta(x_t, t)$, utilizando o processo reverso: $p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu(x_t, t), \sum(x_t, t))$.

Recentemente, essa classe de modelos tem sido bastante utilizada em arquiteturas para síntese sonora (EVANS *et al.*, 2024) (MARIANI *et al.*, 2023). O trabalho desenvolvido por Mariani *et al.* (2023), especificamente, consistiu em um modelo baseado em difusão capaz de tanto separar quanto sintetizar fontes sonoras, o *Multi-Source Diffusion Model (MSDM)*. Para tal, o conjunto de dados consistiu em áudios multicanal de uma mesma composição que foram submetidos ao processo de difusão e reversão de difusão de maneira conjunta como ilustrado na Figura 3.9.

O modelo *MSDM* portanto, é capaz de realizar três tarefas distintas: síntese total, síntese parcial e separação de fontes. A síntese total consiste em gerar uma composição completa (com vários elementos sonoros), já a síntese parcial gera uma faixa sonora

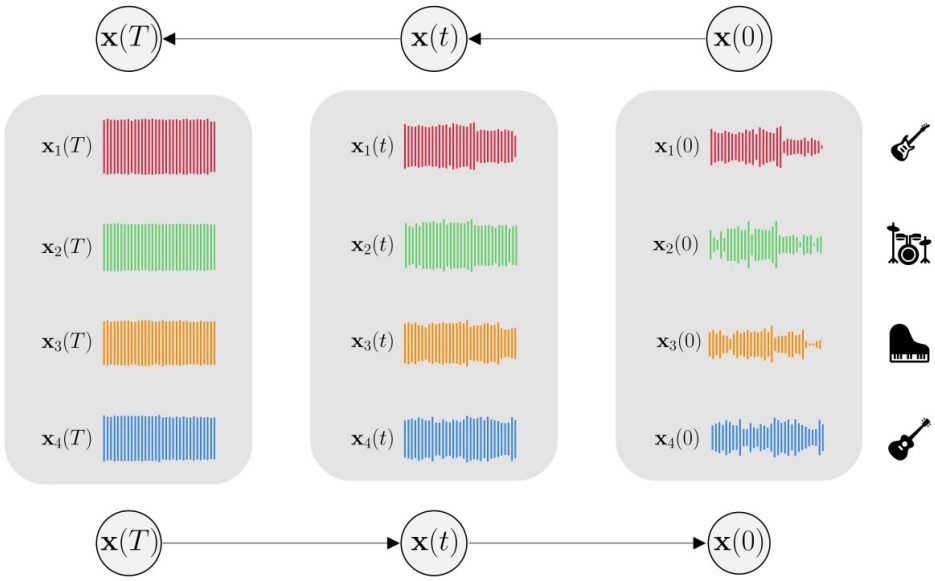

Figura 3.9 – Esquemático do método utilizado pelo modelo MSDM. À direita temos o conjunto de dados antes da inserção de ruído e para a esquerda temos os mesmos dados em diferentes passos de tempo t, onde o ruído foi acrescentado. Fonte: (MARIANI *et al.*, 2023).

de acompanhamento, dadas as demais (como compor uma linha melódica de baixo em cima de um sinal fornecido com uma linha melódica de piano, por exemplo). A separação de fontes consiste no problema clássico de, dado uma composição completa, extrair os elementos musicais individuais.

3.5 Função Custo

A Função Custo, também chamada de Função de Perda (*Loss Function*), pode ser definida, de maneira geral, como a função a ser minimizada durante o treinamento de um modelo. Para aplicações no universo do áudio, essa função pode ser computada tanto no domínio da frequência como no domínio do tempo. Especificando para o problema da mixagem, como descrito na sessão 3.3, dado uma função custo $\mathcal{L}$, podemos calcular a distância entre a mixagem esperada Y e a mixagem predita pelo modelo $\hat{Y}$, onde tanto Y quanto $\hat{Y} \in \mathbb{R}^{2\times T}$ (STEINMETZ *et al.*, 2022). Ressalta-se que a mixagem esperada corresponde à mixagem presente na base de dados, que é apenas uma das várias mixagens possíveis para uma mesma música. A seguir são enumeradas algumas funções usualmente aplicadas para o processamento de áudio.

1. **Erro Médio Absoluto (*MAE* ou ℓ_1)** - Dado que y é a mixagem esperada no domínio do tempo e $\hat{y}$ é a mixagem predita no domínio do tempo, e que ambas possuem N amostras e são faixas estéreo, podemos calcular ℓ_1 no domínio do tempo

da seguinte maneira:

$$\mathcal{L}_{\text{MAE}}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{2N} \sum_{i=1}^{2} \sum_{j=1}^{N} |\hat{\mathbf{y}}_{i,j} - \mathbf{y}_{i,j}|. \tag{3.8}$$

Também pode ser calculada no domínio da frequência, uma vez computadas as transformadas de Fourier de tempo curto (*STFT*): $\hat{Y}_i = |\text{STFT}(\hat{y}_i)|$ e $Y_i = |\text{STFT}(y_i)|$. Dado que i é o número de canais, F é o número de intervalos de frequência e K o número de janelas de tempo, temos:

$$\mathcal{L}_{\text{STFT}}^{\text{MAE}}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2KF} \sum_{i=1}^{2} \sum_{k=1}^{K} \sum_{f=1}^{F} \Big(|\text{STFT}(\hat{\mathbf{y}}_i)|[f,k] - |\text{STFT}(\mathbf{y}_i)|[f,k] \Big). \tag{3.9}$$

2. **Erro Quadrático Médio (*MSE* ou ℓ_2)** - De maneira similar ao Erro Médio Absoluto, podemos calcular tanto no domínio do tempo (Equação 3.10) quanto no domínio da frequência (Equação 3.11).

$$\mathcal{L}_{\text{MSE}}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{2N} \sum_{i=1}^{2} \sum_{j=1}^{N} |\hat{\mathbf{y}}_{i,j} - \mathbf{y}_{i,j}|^2. \tag{3.10}$$

$$\mathcal{L}_{\text{STFT}}^{\text{MSE}}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2KF} \sum_{i=1}^{2} \sum_{k=1}^{K} \sum_{f=1}^{F} \Big(|\text{STFT}(\hat{\mathbf{y}}_i)|[f,k] - |\text{STFT}(\mathbf{y}_i)|[f,k] \Big)^2. \tag{3.11}$$

3. **Transformada de Fourier de Tempo Curto Multi-Resolução (*MR-STFT Loss Function*** - Um problema originado no uso da Transformada de Fourier de Tempo Curto é que dependendo da escolha dos parâmetros a medida do erro pode ser diferente (STEINMETZ *et al.*, 2022). Uma forma de mitigar essa questão é escolher um conjunto de M parâmetros diferentes para a *STFT* e então tirar a média simples das funções custo (calculadas por meio da Equação 3.9 ou da Equação 3.11) resultantes de cada um desses parâmetros.

$$\mathcal{L}_{\text{MR-STFT}}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{M} \mathcal{L}_{\text{STFT}_m}(\hat{\mathbf{Y}}, \mathbf{Y}) \tag{3.12}$$

4. **Soma e Diferença (*Sum and Diference Loss Function*)** - No contexto da mixagem, o conteúdo de cada canal estéreo (direito e esquerdo) tem importância, logo uma função de perda que leve em consideração o posicionamento dos elementos na imagem estéreo é de grande relevância. A Função de Soma e Diferença proposta por Steinmetz *et al.* (2021) utiliza a soma e a diferença dos canais direto e esquerdo de modo a considerar a imagem estéreo no cálculo da função. Primeiro, os sinais são computados no domínio do tempo:

$$\mathbf{y}_{\text{sum}} = \mathbf{y}_{\text{left}} + \mathbf{y}_{\text{right}}, \quad \mathbf{y}_{\text{diff}} = \mathbf{y}_{\text{left}} - \mathbf{y}_{\text{right}}. \tag{3.13}$$

Isso ocorre tanto para a mixagem de referência (y) quanto para a mixagem estimada pelo modelo $\hat{y}$. Em seguida, é computada a função de perda *MR-STFT* descrita na Equação 3.12 para a soma dos canais da mixagem de referência em comparação com a soma dos canais da mixagem estimada e depois o mesmo procedimento é repetido para a diferença entre os canais, resultando assim na seguinte equação:

$$\mathcal{L}_{S/D}(\hat{\mathbf{Y}}, \mathbf{Y}) = \mathcal{L}_{\text{MR-STFT}}(\hat{\mathbf{Y}}_{\text{sum}}, \mathbf{Y}_{\text{sum}}) + \mathcal{L}_{\text{MR-STFT}}(\hat{\mathbf{Y}}_{\text{diff}}, \mathbf{Y}_{\text{diff}}) \tag{3.14}$$

Uma consideração importante a ser feita sobre as Funções de Perda para modelos de áudio é que dependendo do objetivo, mais do que aproximar a saída do modelo de uma referência, a percepção humana é o que norteia seu desenvolvimento. Tendo isso em vista, uma forma possível de levar a percepção humana em consideração é empregar conceitos da psicoacústica para dar mais peso à determinadas frequências durante o treinamento do modelo (WRIGHT; VÄLIMÄKI, 2020).

A curva A é a curva de isoaudibilidade, ou seja, mostra para cada frequência com qual intensidade um som deve ser reproduzido (em dB pascal) para que ele seja percebido com o mesmo volume de um tom em 1kHz. Uma ponderação A (*A-weighting filter*) pode ser aplicado à um sinal, de modo a representar a compensação da curva de isoaudibilidade. A resposta em frequência da ponderação A pode ser observada na Figura 3.10.

O trabalho realizado por Wright e Välimäki (2020) testou o uso de diferentes filtros durante o treinamento de um modelo que tinha como objetivo o de emular o funcionamento de um amplificador de guitarra. Os autores constataram, por meio de um teste de percepção, que o modelo treinado com a ponderação A melhorou de forma significativa a similaridade percebida entre o áudio objetivo e o áudio processado pelo modelo.

3.6 Discussão

Neste capítulo foram, portanto, apresentados aspectos relevantes acerca do campo de estudos dos Sistemas Inteligentes de Mixagem. Uma revisão bibliográfica sobre o atual estado da arte foi feita, bem como um levantamento de aplicações de modelos de Inteligência Artificial Generativa para tarefas musicais que se assemelham à mixagem. Dada a diferença entre as duas abordagens possíveis para sistemas de mixagem apresentadas, foram escolhidos dois modelos para compor este trabalho, cada um representando uma abordagem distinta. A arquitetura baseada em transformação direta selecionada foi a *Mix-Wave-U-Net* e a arquitetura baseada em estimativa de parâmetros selecionada foi a *Differenciable Mixing Console* (Console de Mixagem Diferenciável). Essa seleção se deu

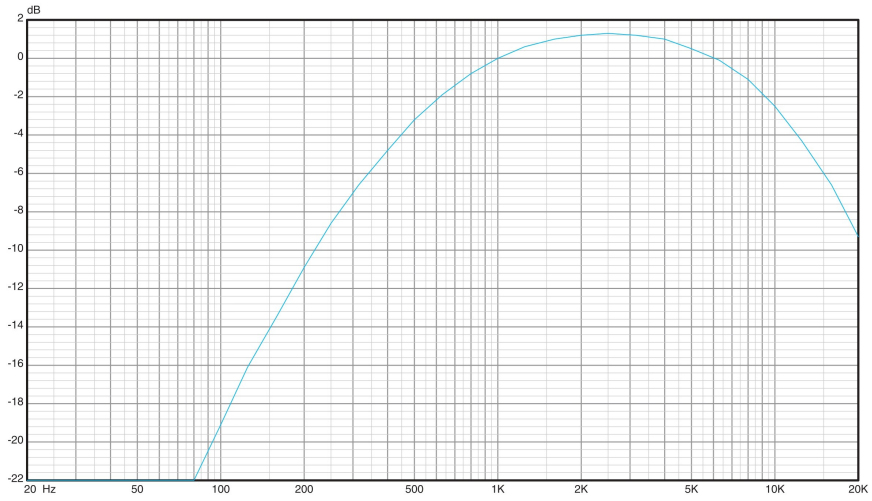

Figura 3.10 – Resposta da magnitude (dB) em frequência da ponderação A. Fonte: (DAWSON, 2005)

pelo fato de que ambas as implementações originais foram disponibilizadas na íntegra pelos desenvolvedores, o que facilita a compreensão e o desenvolvimento deste trabalho.

4 Metodologia

A metodologia adotada neste trabalho será detalhada a seguir. Dois modelos de Mixagem Automática foram implementados e treinados em linguagem *Python*, um pertencendo à abordagem de estimativa de parâmetros (Console de Mixagem Diferenciável) e outro pertencendo à abordagem de transformação direta (*Mix-Wave-U-Net*). Para possibilitar o treino, um conjunto de dados com 195 músicas foi preparado. Por fim, foi conduzido um teste subjetivo para avaliação das performances dos modelos quando comparados entre si e com mixagens feitas por humanos.

4.1 Banco de dados

Para a construção do conjunto de dados desse trabalho, foram utilizadas as músicas contidas no banco de dados MedleydB, que está dividido em dois volumes: V1 (BITTNER *et al.*, 2014) e V2 (BITTNER *et al.*, 2016). No total, os dois volumes do MedleydB somam 196 músicas e a obtenção dos arquivos foi realizada através de uma solicitação aos desenvolvedores. Uma das músicas do banco de dados veio corrompida, resultando em 195 músicas de gêneros variados.

O banco de dados disponibiliza as faixas dos elementos isolados sem processamento, bem como a mixagem estéreo respectiva de cada música. Como cada música tem um número distinto de elementos musicais, foi feito um pré processamento nos dados de modo a fixar o número de faixas de entrada dos modelos. O ato de fixar o número de entradas permite a utilização de todas as músicas da base dados, em vez de escolher um número máximo de faixas e descartar as músicas que contém um número superior de faixas.

O pré processamento consistiu em agrupar manualmente as faixas em quatro grupos: baixo, bateria, vocais e outros. Esse agrupamento segue a estrutura de outros *datasets* utilizados em trabalhos de modelos de mixagem automática na literatura (MARTÍNEZ-RAMÍREZ *et al.*, 2022; KOSZEWSKI *et al.*, 2023). O agrupamento simples consiste na soma das faixas dos elementos isolados de cada grupo e posterior normalização de cada grupo. Algumas das limitações dessa maneira de agrupamento serão discutidas no Capítulo 5.

As músicas então foram divididas de maneira aleatória em treino e validação, sendo 80% (156 músicas) utilizadas para treino e 20 % (39 músicas) utilizadas para validação. Não foi realizada validação cruzada (*k-Folding*).

4.2 Implementação

A implementação dos modelos *Mix-Wave-U-Net* e *Diferentiable Mixing Console* foram baseadas nas implementações disponibilizadas por Steinmetz *et al.* (2022). A seguir serão detalhados os modelos e as alterações feitas em cada uma das implementações.

4.2.1 Arquitetura de Transformação Direta - *Mix-Wave-U-Net*

A arquitetura *Mix-Wave-U-Net*, como detalhado na sessão 3.3.1, apresenta uma abordagem de transformação direta baseada em redes do tipo Encoder-Decoder. Mais detalhes sobre a arquitetura podem ser encontrados na Tabela 4.1. Aqui ressalta-se que foram escolhidos 8 canais de entrada (4 faixas estéreo) e o número de camadas escolhido foi de L = 10, ou seja, são 10 blocos de *downsampling* e 10 blocos de *upsampling*, de acordo com a literatura (MARTINEZ-RAMIREZ *et al.*, 2021b; KOSZEWSKI *et al.*, 2023). Foram mantidos os comprimentos dos *kernels* o uso da função de ativação *Parametric ReLu* em relação a implementação de Steinmetz *et al.* (2022), que serviu de base para esse trabalho.

Tabela 4.1 – Ordem das operações em cada um dos blocos que compõem o Codificador e o Decodificador da arquitetura *Mix-Wave-U-Net*

Bloco	Operação
Downsampling	Convolução1D BatchNormalization PReLu Convolução1D
Upsampling	Upsampling *Skip connection* Convolução1D BatchNormalization PReLu

4.2.2 Arquitetura de Estimativa de Parâmetros - DMC

Como abordado anteriormente na subseção 3.3.2, a arquitetura do Console Diferenciável de Mixagem utiliza uma abordagem de estimativa de parâmetros, onde uma interface interpretável é responsável por realizar o processamento das faixas de áudio. Em relação à implementação original, duas principais diferenças foram feitas no console: a primeira em relação aos efeitos implementados e a segunda no desenho do processamento do console. O trabalho original implementava os efeitos por meio de redes neurais que simulavam o funcionamento de efeitos de áudio (*neural proxies*) e o console proposto seguia o descrito na Figura 3.4. Neste trabalho, os efeitos de áudio utilizados foram os efeitos de áudio disponibilizados pela biblioteca *Pedalboard* (SOBOT, 2021), com exceção

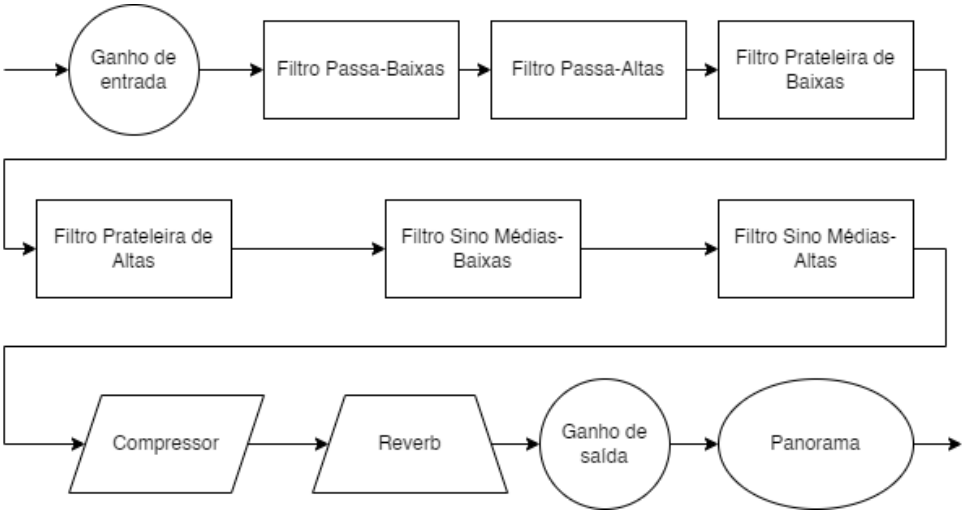

Figura 4.1 – Fluxo de processamento do sinal no console proposto.

do ganho de entrada e do panorama, que foram implementados seguindo a implementação original.

Quanto ao desenho do console, optou-se por realizar algumas alterações, em especial para simular as bandas de equalização de um console de mixagem real, o SSL-4000. Para cada um dos vinte e um parâmetros do console proposto, foi colocado um valor de mínimo e máximo possível. Mais detalhes sobre a ordem do processamento e os valores possíveis de cada parâmetro podem ser conferidos na Figura 4.1 e na Tabela 4.2. Para os filtros passa-altas e passa-baixas, o decaimento é fixo de 6 dB/oitava. Já para os filtros prateleira, o valor de Q é fixo em 0.7.

Destaca-se que o desenho do console pode ser feito de muitas formas distintas. O console aqui proposto buscou disponibilizar uma gama de funcionalidades ao modelo, mas ao mesmo tempo restringir os valores mínimos e máximos de alguns parâmetros, considerando que o modelo teria acesso apenas ao material agrupado para realizar a mixagem. O uso de reverberação, por exemplo, foi limitado, já que seu uso direto em um agrupamento não é muito usual. O mesmo raciocínio foi aplicado ao panorama.

4.2.3 Função Custo

Como detalhado na Seção 3.5, a Função Custo em aplicações de áudio pode ser calculada tanto no domínio do tempo quanto no domínio da frequência e também pode ponderar questões de psicoacústica ou não. Neste trabalho a Função Custo escolhida foi a Soma e Diferença, calculada no domínio da frequência, com a adição da ponderação A de modo a considerar a percepção humana. A implementação da Função Custo foi feita por meio da biblioteca *auraloss* (STEINMETZ; REISS, 2020).

Tabela 4.2 – Parâmetros presentes na implementação do Console Diferenciável de Mixagem.

Efeito	Parâmetro	Unidade	Mínimo	Máximo
Ganho de entrada	Ganho	dB	-48	12
Filtro passa altas	frequência de *cutoff*	Hz	0	350
Filtro passa baixas	frequência de *cutoff*	Hz	3000	22000
Filtro prateleira de altas	frequência de *cutoff* ganho	Hz dB	1500 -5	16000 5
Filtro prateleira de baixas	frequência de *cutoff* ganho	Hz dB	30 -5	450 5
Filtro sino médias-altas	frequência de *cutoff* ganho q	Hz dB	600 -5 0.5	7000 5 3
Filtro sino médias-baixas	frequência de *cutoff* ganho q	Hz dB	200 -5 0.5	2500 5 3
Compressor	*threshold* razão ataque *release*	dB ms ms	-5 1 1 100	5 20 30 4000
Reverberação	tamanho da sala *wet level*		0 0	0.5 0.33
Ganho de saída	ganho	dB	-48	12
Panorama	panorama		0.3	0.7

4.3 Treino

Para o treinamento dos modelos, foi utilizada uma GPU NVidia A100, através da plataforma Google Colab. Os hiperparâmetros utilizados foram os seguintes:

1. *Batch Size*: 16
2. *Learning Rate*: 3e-4
3. Otimizador: Adam
4. *Scheduler*: Cosseno

O modelo MWUN levou cerca de dois dias para treinar, enquanto o modelo DMC levou cerca de 12 horas para treinar, sendo feita a parada manual quando a função custo de validação apresentou estabilidade. Para cada modelo foi selecionada como melhor versão aquela cuja época apresentou menor valor de Função Custo de validação. Nas Figuras 4.2 e 4.3 é possível observar a progressão da função perda nas etapas de treino e validação para cada um dos modelos.

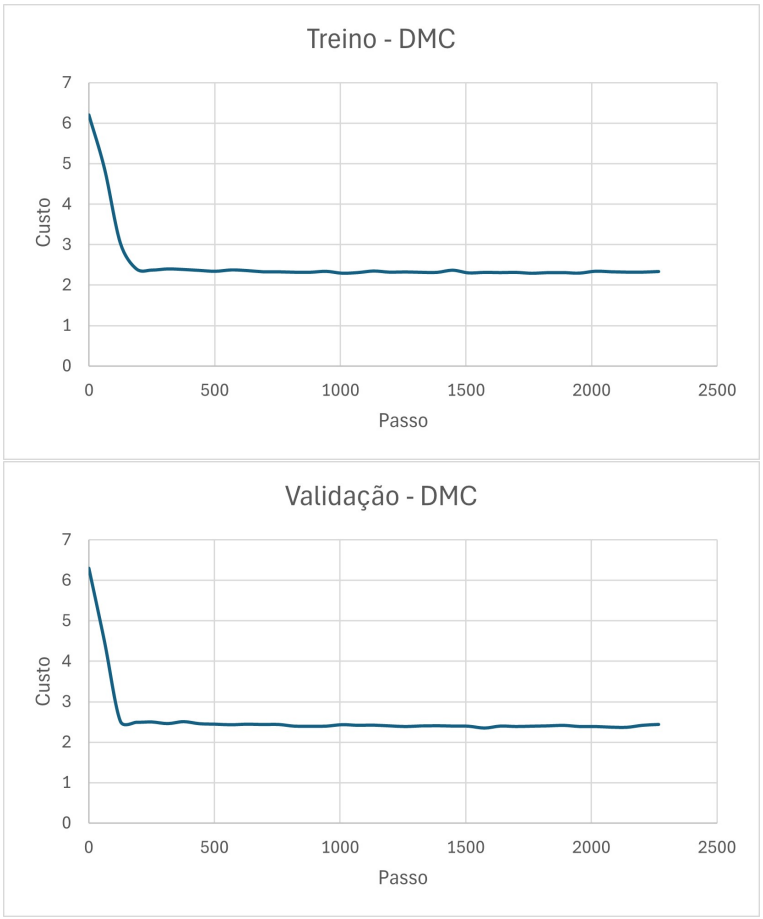

Figura 4.2 – Comportamento da função custo nos conjuntos de treino (em cima) e validação (embaixo) para o modelo Console de Mixagem Diferenciável.

4.4 Coleta de Mixagens

De modo a avaliar a performance dos modelos em diferentes contextos musicais, foram selecionadas três músicas diferentes disponibilizadas no conjunto de dados da *Cambridge Music Technology*, um conjunto de dados de projetos multicanal livres de direitos autorais feito para propósitos educacionais (SENIOR, 2018). As músicas selecionadas, juntamente com os respectivos gêneros podem ser conferidas na Tabela 4.3.

Tabela 4.3 – Músicas selecionadas como conjunto de teste para o modelo.

Música	Artista	Gênero
Summertime	**Bolz & Knecht**	**Instrumental**
Set Me Free	**Ghostly Beard**	**Jazz Pop**
King Of The Weekend	**Babe Grand**	**Techno**

4.5 Teste Subjetivo

Dada a natureza subjetiva e complexa da mixagem, a maneira mais usual de medir o desempenho de um Sistema Automático de Mixagem é através de testes subjetivos

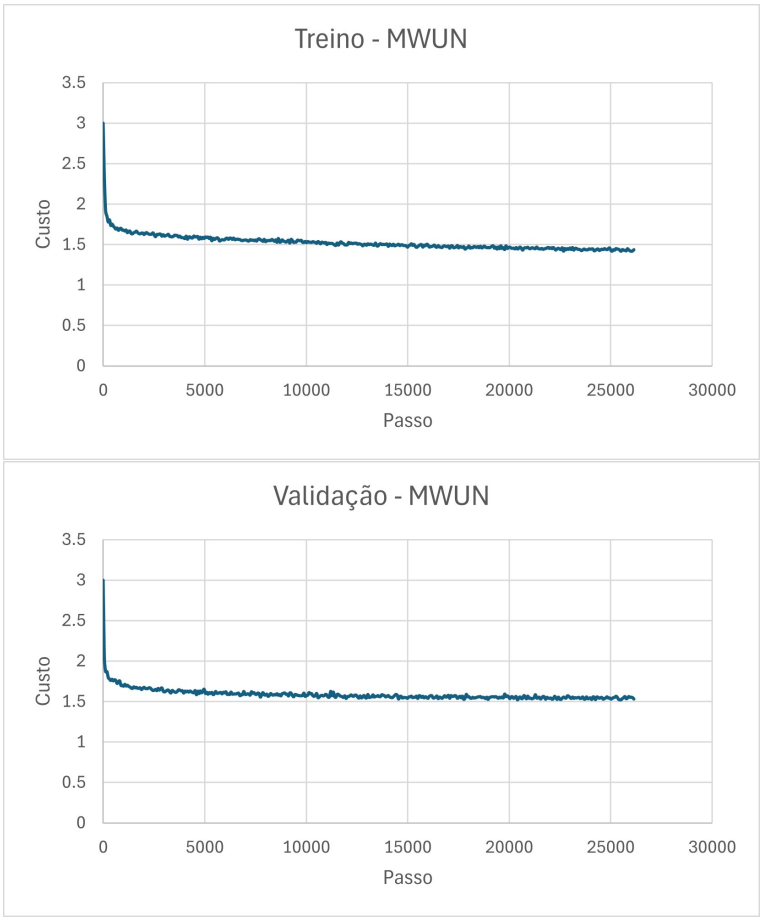

Figura 4.3 – Comportamento da função custo nos conjuntos de treino (em cima) e validação (embaixo) para o modelo *Mix-Wave-U-Net*.

de audição (MAN *et al.*, 2019). Para o desenho deste teste subjetivo, foram retomadas a pergunta de pesquisa e a hipótese nula (BECH; ZACHAROV, 2007).

As perguntas de pesquisa definidas foram as seguintes: "Como as mixagens feitas por diferentes modelos de IA são avaliadas quando confrontadas entre si e com mixagens feitas por humanos? Há alguma diferença na avaliação na visão de profissionais, amadores e entusiastas?". Dadas as perguntas, a hipótese nula que se estabelece é a seguinte: "Não há diferença significativa na preferência entre mixagens feitas por humanos e por modelos de IA entre pessoas com diferentes níveis de conhecimento sobre mixagem".

Com a pergunta e hipótese nula estabelecidas, foi optado por conduzir um Teste AB. Foram selecionados trechos de 29 segundos de cada música e para cada música foram apresentadas 3 mixagens diferentes (MWUN, DMC e humana). As mixagens de cada música foram então pareadas aleatoria e anonimamente, de modo a formar 9 comparações (3 para cada música). Todas as mixagens foram normalizadas em *loudness* de -23 LUFS e estavam sincronizadas, de maneira a permitir a audição das diferentes mixagens para uma mesma música de maneira sincronizada e possibilitar realizar um *loop* de determinado trecho.

Figura 4.4 – Foto da interface do teste AB. Abaixo estão os controles de *playback* e de *loop*.

O teste foi desenhado de maneira online através da plataforma especializada em testes auditivos *Go Listen* (BARRY *et al.*, 2021). Uma foto da interface do teste pode ser observada na Figura 4.4. Para coletar variáveis experimentais, foram acrescentadas algumas perguntas de controle no início do teste, bem como um áudio para ajuste de volume. As perguntas de controle visam a obtenção de algumas informações básicas sobre o voluntário que realizou o teste, bem como entender o sistema (monitores e/ou fones de ouvido) e o ambiente (silencioso/ruidoso, aberto/fechado, acusticamente tratado) no qual o teste foi realizado. As perguntas de controle que foram feitas são as seguintes:

1. Qual a sua faixa etária?
2. Como você descreveria a sua audição?
3. Quais gêneros musicais você costuma ouvir?
4. Como você descreveria seu conhecimento em mixagem musical?
5. Quantos anos você possui de experiência com mixagem?
6. Por qual sistema você está realizando esse teste?
7. Como você descreveria o seu ambiente atual?
8. Você está realizando este teste no seu sistema habitual de som, ou seja, onde você costuma ouvir música?

9. Saberia dizer o modelo dos fones ou monitores de áudio que está usando?

10. Se estiver utilizando monitores de áudio, a sua sala é acusticamente tratada?

Com exceção das perguntas 9 e 10, que eram de texto corrido, as demais eram perguntas cujas respostas deveriam ser selecionadas entre as opções fornecidas. No Apêndice A estão contidas as perguntas de controle e as repostas possíveis para cada pergunta, da maneira como foram apresentadas aos participantes do teste.

De acordo com as definições de amador, entusiasta e profissional aprofundadas na seção 3.1, foram definidas as metas de participação de no mínimo 24 amadores, 20 entusiastas e 20 profissionais, de acordo com a referência bibliográfica (BECH; ZACHAROV, 2007). Como o teste foi feito de maneira online, também colocamos como meta uma participação ampla de pelo menos 100 sujeitos no geral. O teste foi aprovado pelo Comitê de Ética em Pesquisa da Unicamp, no processo de número 67203223.0.0000.5404.

5 Análise e Discussão

Neste capítulo serão apresentados os resultados relativos ao teste subjetivo detalhado no capítulo anterior. Primeiro, será apresentado um panorama geral sobre o perfil das pessoas que responderam ao teste e os resultados quantitativos gerais. Em seguida serão apresentados os resultados quantitativos estratificados por nível de conhecimento do sujeito. Por fim, será realizada a discussão, apresentando alguns comentários qualitativos feitos.

Foram fornecidas ao total 9 comparações, 3 para cada música. A primeira música do teste é a música intitulada "*King of the Weekend*" do artista *Babe Grand*, pertencente ao gênero de música eletrônica *Techno*. A segunda música do teste consiste na música "*Summertime*" interpretada pela dupla *Bolz & Knecht* e corresponde a uma versão instrumental de uma música de mesmo nome composta originalmente por George Gershwin, em 1934, e regravada por diversos artistas ao longo do tempo. A terceira e última música se chama "*Set Me Free*" do cantor e compositor *Ghostly Beard* e corresponde a uma composição pop com influências de jazz.

Os arquivos podem ser acessados pelo seguinte link (<https://drive.google.com/drive/folders/1uxXekkTNhHwO1gseAzM7GZvYemHoItQm?usp=drive_link>). Ressaltando que no link os arquivos estão rotulados (mixagem humana, feita pelo modelo DMC ou feita pelo modelo MWUN), mas no teste os sujeitos não tiveram acesso aos rótulos, ou seja, foi realizado um teste cego.

5.1 Resultados Quantitativos Gerais

O teste subjetivo foi conduzido durante a segunda metade do mês de Maio de 2024 e foi divulgado de maneira online (e-mail e redes sociais). O teste contou ao todo com 125 respostas. Das 125 respostas obtidas, 32 foram dadas por profissionais da mixagem, 36 por entusiastas e 57 por amadores. Dados sobre a faixa etária e o sistema de reprodução dos sujeitos podem ser conferidos na Figura 5.1, onde é possível observar que a faixa etária mais comum entre os sujeitos é entre 25 e 34 anos e o sistema de reprodução mais utilizado foram fones de ouvido.

Os pareamentos das trilhas apresentadas foram feitos de maneira aleatória e sua ordem pode ser observada na Tabela 5.1. Nas demais tabelas os pareamentos seguirão sempre o mesmo ordenamento, diferente do que foi apresentado no teste, de maneira a facilitar a compreensão dos resultados.

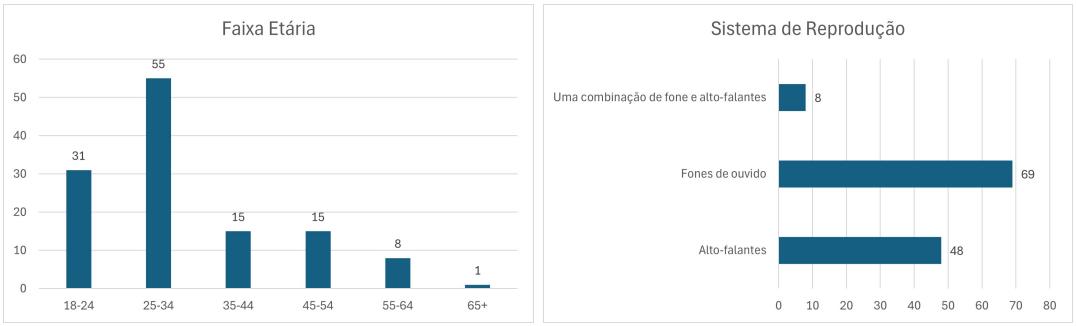

Figura 5.1 – Distribuição da faixa etária dos participantes (à esquerda) e contagem do sistema de reprodução utilizado (à direita).

Tabela 5.1 – Relação dos pareamentos feitos em cada música.

		A	B
Música 1 (Techno)	Par 1	DMC	Humana
	Par 2	DMC	MWUN
	Par 3	MWUN	Humana
Música 2 (Instrumental)	Par 1	MWUN	Humana
	Par 2	Humana	DMC
	Par 3	DMC	MWUN
Música 3 (Jazz)	Par 1	Humana	MWUN
	Par 2	MWUN	DMC
	Par 3	Humana	DMC

Na Tabela 5.2 estão compilados os resultados gerais de preferência para cada um dos pareamentos feitos. É possível notar nas três músicas que nos pareamentos cuja mixagem humana estava presente, ela foi preferida a maior parte das vezes (Música 1 - Par 1 e Par 3; Música 2 - Par 1 e Par 2; Música 3 - Par 1 e Par 3). Na primeira música, quando o pareamento foi entre as mixagens feitas pelos modelos (Par 2), a preferência se dividiu de maneira similar. Na segunda música (Par 3) houve uma leve preferência pela mixagem feita pelo modelo Console de Mixagem Diferenciável. Já na terceira música (Par 2) a mixagem do modelo *Mix-Wave-U-Net* foi preferida mais vezes nesse cruzamento.

5.2 Resultados Quantitativos Estratificados

5.2.1 Amadores

Dos 57 participantes que se declaram com pouco ou nenhum conhecimento sobre mixagem, 3 declararam ter um diagnóstico profissional de alguma condição referente à audição e por isso daqui em diante serão excluídos da análise, resultando em 54 respostas. Os resultados relativos a essa população podem ser encontrados na Tabela 5.3.

Tabela 5.2 – Resultados gerais do teste subjetivo, com a contagem e a porcentagem de participantes que selecionaram cada uma das mixagens de preferência por pareamento.

Música	Contagem			Porcentagem		
	DMC	MWUN	Humana	DMC	MWUN	Humana
Música 1	65	60		52	48	
	22		103	17,6		82,4
		31	94		24,8	75,2
Música 2	74	51		59,2	40,8	
	19		106	15,2		84,8
		15	110		12	88
Música 3	41	84		32,8	67,2	
	14		111	11,2		88,8
		20	105		16	84

Para esse conjunto de participantes, podemos observar que em todas as músicas a mixagem humana foi preferida sobre as demais mixagens. Quando as mixagens feitas pelos modelos foram diretamente confrontadas os resultados divergiram. Na primeira música, segundo pareamento, a preferência se dividiu em 51,85% para a mixagem do modelo DMC e 48,15% para a mixagem do modelo MWUN. Na segunda música, terceiro pareamento, a maioria dos participantes preferiu a mixagem do modelo DMC (59,26%) frente ao modelo MWUN (40,47%). Na terceira música, segundo pareamento, a preferência tendeu para a mixagem feita pelo modelo MWUN (66,67%) em vez do modelo DMC (33,33%).

Tabela 5.3 – Resultados do teste subjetivo dos participantes da categoria amador, com a contagem e a porcentagem de participantes que selecionaram cada uma das mixagens de preferência por pareamento.

Música	Contagem			Porcentagem		
	DMC	MWUN	Humana	DMC	MWUN	Humana
Música 1	28	26		51,85	48,15	
	12		42	22,22		77,78
		16	38		29,63	70,37
Música 2	32	22		59,26	40,74	
	13		41	24,07		75,93
		12	42		22,22	77,78
Música 3	18	36		33,33	66,67	
	11		43	20,37		79,63
		16	38		29,63	70,37

5.2.2 Entusiastas

Dos 36 participantes que se declaram com algum conhecimento sobre mixagem, nenhum declarou ter alguma condição diagnosticada por profissional. Os resultados relativos a essa população podem ser encontrados na Tabela 5.4.

Tabela 5.4 – Resultados do teste subjetivo dos participantes da categoria entusiasta, com a contagem e a porcentagem de participantes que selecionaram cada uma das mixagens de preferência por pareamento.

Música	Contagem			Porcentagem		
	DMC	MWUN	Humana	DMC	MWUN	Humana
Música 1	18	18		50	50	
	7		29	19,45		80,55
		6	30		16,67	83,33
Música 2	25	11		69,45	30,55	
	5		31	13,89		86,11
		2	34		5,56	94,44
Música 3	11	25		30,55	69,45	
	2		34	5,56		94,44
		2	34		5,56	94,44

Entre os entusiastas, a mixagem humana foi preferida em todos os pareamentos nos quais ela esteve presente. Nos pareamentos nos quais a mixagem humana não estava presente, houveram diferentes tendências. Na primeira música, a preferência dos participantes se dividiu exatamente na metade no segundo pareamento. Na segunda música, terceiro pareamento, 69,45% dos participantes preferiram a mixagem feita pelo modelo DMC enquanto 30,55% preferiu a mixagem feita pelo modelo MWUN. Já na terceira música, segundo pareamento, a maioria dos participantes (69,45%) preferiu a mixagem realizada pelo modelo MWUN, enquanto 30,55% preferiu a mixagem feita pelo modelo DMC.

5.2.3 Profissionais

Dos 32 participantes que se declaram profissionais da mixagem, 5 declararam ter um diagnóstico profissional de alguma condição referente à audição, resultando em 27 respostas de profissionais sem comprometimento da audição. Os resultados relativos a essa população podem ser encontrados na Tabela 5.5.

Entre profissionais, a preferência pela mixagem humana foi bastante evidente. Na segunda música, em especial, nos pareamentos nos quais a mixagem humana estava presente (primeiro par e segundo par), todos os participantes preferiram a mixagem humana. Quanto aos pareamentos nos quais a mixagem humana não estava presente, na primeira e segunda música a preferência ficou bem próxima dos 50% para a mixagem de cada modelo. Já na terceira música, segundo par, a mixagem do modelo MWUN foi preferida por 70,37% dos participantes enquanto 29,63% preferiu a mixagem do modelo DMC.

Tabela 5.5 – Resultados do teste subjetivo dos participantes da categoria profissional, com a contagem e a porcentagem de participantes que selecionaram cada uma das mixagens de preferência por pareamento.

Música	Contagem			Porcentagem		
	DMC	MWUN	Humana	DMC	MWUN	Humana
Música 1	14	13		51,85	48,15	
	1		26	3,71		96,29
		5	22		18,52	81,48
Música 2	13	14		48,15	51,85	
	0		27	0		100
		0	27		0	100
Música 3	8	19		29,63	70,37	
	0		27	0		100
		1	26		3,71	96,29

5.3 Análise de Variância

Foi realizada uma análise de variância de modo a testar a hipótese nula de que "Não há diferença significativa na preferência entre mixagens feitas por humanos e por modelos de IA entre pessoas com diferentes níveis de conhecimento sobre mixagem". Para tanto, as respostas dos três grupos (amadores, entusiastas e profissionais) foram combinadas, como mostrado na Tabela 5.2.

Os dados coletados são do tipo *categóricos*, no caso, *prefiro* versus *não prefiro* e, de acordo com Bech e Zacharov (2007), estes dados geralmente não possuem distribuição estatística normal, o que implica que os dados devem ser analisados com técnicas estatísticas não-paramétricas. Foi realizado um teste de Kruskal-Wallis para determinar se há um efeito do tipo de mixagem sobre a preferência dos voluntário. Os resultados indicam uma diferença significativa, $\chi^2(17) = 11{,}42$, $\mathbf{p = 0{,}003}$. Portanto, rejeitamos a hipótese nula e concluímos que há diferença significativa no nível de preferência entre as diferentes técnicas de mixagem.

As comparações *post-hoc*, feitas por meio do Teste de Tukey, indicaram que o ranqueamento médio da preferência da mixagem humana (3) é significativamente diferente (maior) do que as demais, enquanto a diferença entre o ranqueamento médio das preferências das mixagens realizadas pelos modelos DMC (1) e MWUN (2) não apresentam diferença significativa, como pode ser visto na Figura 5.2.

Também foram realizadas comparações *post-hoc* para os resultados estratificados entre Amadores, Entusiastas e Profissionais, representados, respectivamente, nas Figuras 5.3, 5.4 e 5.5. Esses resultados indicam que, nos diferentes estratos, o ranqueamento médio da preferência da mixagem humana (3) também é significativamente maior do que as demais.

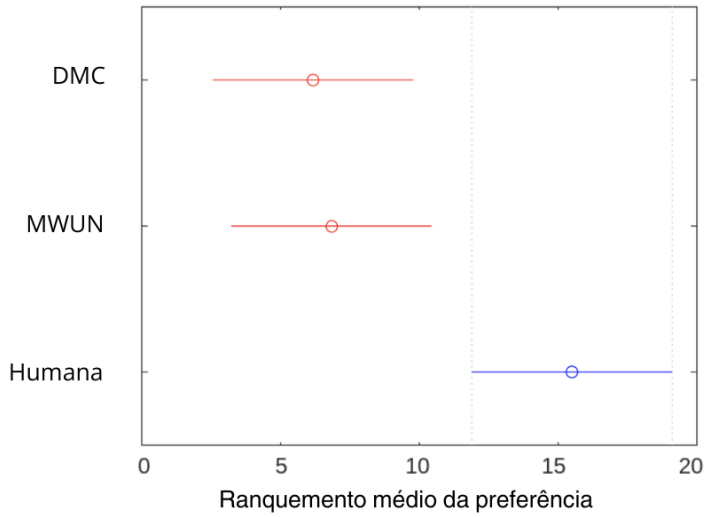

Figura 5.2 – Teste de Tukey realizado nos dados de preferência geral.

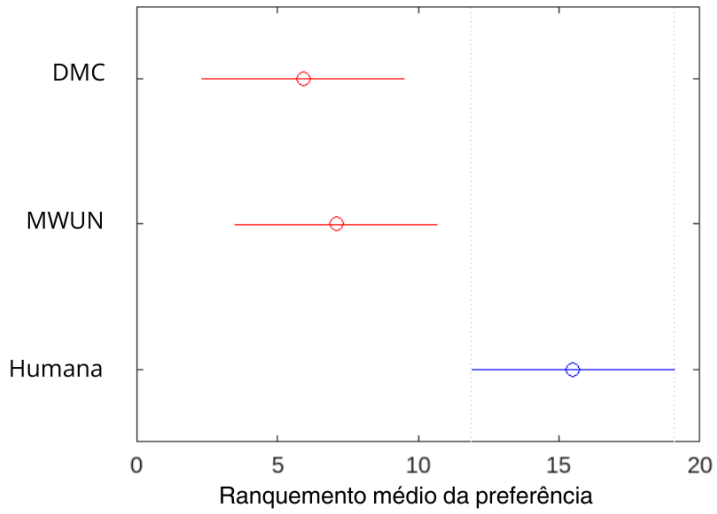

Figura 5.3 – Teste de Tukey realizado nos dados de preferência dos participantes amadores.

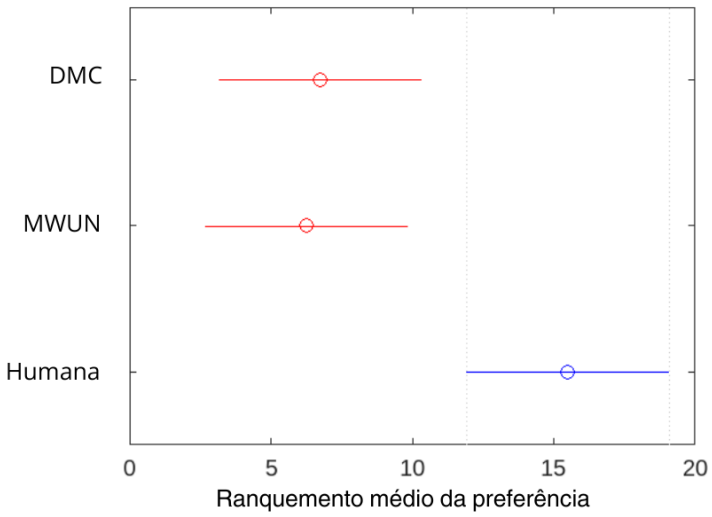

Figura 5.4 – Teste de Tukey realizado nos dados de preferência dos participantes entusiastas.

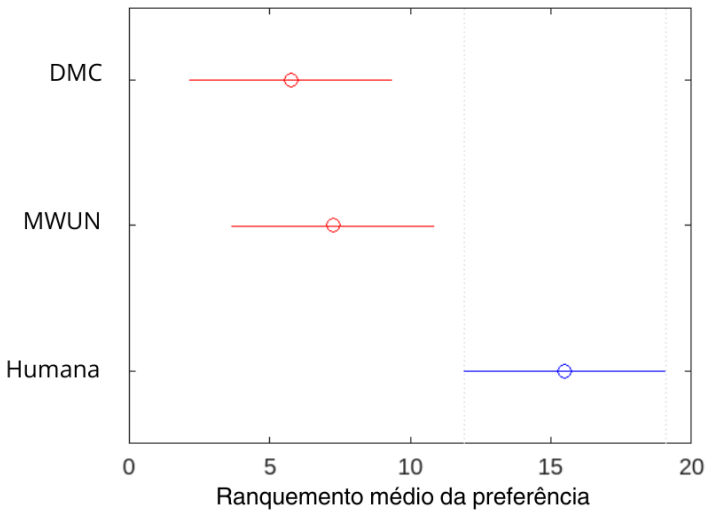

Figura 5.5 – Teste de Tukey realizado nos dados de preferência dos participantes profissionais.

5.4 Resultados Qualitativos

A seguir os resultados serão apresentados de maneira qualitativa com o suporte dos comentários feitos pelos participantes. Clicando sobre o título de cada pareamento é possível acessar as faixas de áudio que foram apresentadas no teste. Os comentários dos participantes na íntegra podem ser lidos no Apêndice B.

5.4.1 Primeira Música: Babe Grand - "King of the Weekend" (Techno)

5.4.1.1 Par 1 - DMC (A) x Humana (B)

Entre o grupo de profissionais, foi destacado que a mixagem B possuía graves mais pronunciados e uma melhor imagem estéreo. O único profissional que escolheu A pontuou que, na sua opinião, a mixagem B possuía graves demais. Os entusiastas que escolheram a mixagem A ressaltaram que gostaram mais da clareza e do uso de reverberação, enquanto os que escolheram B também ressaltaram os graves como ponto positivo. Já os amadores fizeram pontuações parecidas com os entusiastas, de que A possuiu uma ambientação mais interessante e B possuiu graves mais agradáveis.

5.4.1.2 Par 2 - DMC (A) x MWUN (B)

Neste pareamento a maioria dos comentários foi em torno da distribuição estéreo dos elementos. A maioria dos profissionais que escolheram A ressaltaram que acharam essa versão mais equilibrada em termos de imagem estéreo, enquanto os que escolheram B ressaltaram que a distribuição estéreo está mais interessante. Entre os entusiastas, os que escolheram A ressaltaram que a panoramização da opção B causou desconforto, enquanto os que escolheram B ressaltaram que a opção A falta espacialidade. Entre os amadores que escolheram a mixagem A foi ressaltado que a mixagem B parecia desbalanceada (o canal esquerdo mais alto que o direito) e também algumas pessoas perceberam um chiado na mixagem B. Já as que escolheram a B ressaltaram que gostaram da ambientação dos elementos nessa mixagem.

5.4.1.3 Par 3 - MWUN (A) x Humana (B)

Os profissionais que escolheram a mixagem B destacaram que os graves foram um ponto positivo, enquanto os profissionais que escolheram A ressaltaram que B está com sobra de graves. Essa percepção se estendeu de maneira geral para os grupos de entusiastas e de amadores. Algumas pessoas também comentaram sobre a presença de ruídos em A.

5.4.2 Segunda Música: Ghostly Beard - "Set Me Free" (Jazz Pop)

5.4.2.1 Par 1 - MWUN (A) x Humana (B)

Neste pareamento, o grupo de profissionais escolheu de maneira unânime a mixagem B, ressaltando que B possuí um palco sonoro melhor estruturado, enquanto A falta distribuição no palco sonoro. Também foi destacado que o balanço tonal de B agradou mais. No grupo de entusiastas, quem escolheu A não realizou comentários sobre sua escolha, enquanto quem escolheu B ressaltou que a equalização dos elementos está melhor nessa mixagem. Já os amadores que escolheram a mixagem A justificaram a preferência de elementos específicos como a voz ou os graves, enquanto os que escolheram B ressaltaram que essa mixagem permitia melhor identificação dos elementos.

5.4.2.2 Par 2 - Humana (A) x DMC (B)

Neste pareamento os profissionais também preferiram de maneira unanime a mixagem A, ressaltando que B parece ter desalinhamentos de fase e que o uso de reverberação causou estranhamento. Também foi comentado sobre o balanço tonal e a dinâmica de A estarem superiores. Entre os entusiastas, quem escolheu a versão A comentou que a inteligibilidade dos elementos está melhor e que os elementos estão mais equilibrados. Quem escolheu B comentou que a voz foi melhor processada e que os agudos estão mais agradáveis nessa versão. Entre os amadores quem escolheu a mixagem A destacou a clareza dos elementos nessa versão e que a mixagem B causou uma sensação de abafamento. Das pessoas que escolheram B, o grave foi pontuado como o aspecto que mais agradou.

5.4.2.3 Par3 - DMC (A) x MWUN (B)

Entre os profissionais, quem escolheu a mixagem A comentou que a equalização está melhor e que B possuí artefatos, quem escolheu B comentou sobre melhor espacialização e melhor equilíbrio de volumes Do grupo de entusiastas, 8 participantes de 36 comentaram que não gostaram de nenhuma das duas mixagens. Entretanto, os participantes que escolheram a mixagem A ressaltaram que o processamento de bateria está melhor e que B apresenta problemas no processamento dinâmico, que não está natural, enquanto quem escolheu B destacou melhor espacialidade. Entre o grupo de amadores quem escolheu a versão A ressaltou que pareceu mais equilibrada e que B pareceu um pouco estridente, enquanto quem escolheu B ressaltou que a mixagem está mais interessante e que os elementos aparentam estar mais definidos.

5.4.3 Terceira Música: Bolz & Knecht - "Summertime" (Instrumental)

5.4.3.1 Par 1 - Humana (A) x MWUN (B)

Entre o grupo de profissionais que selecionaram a mixagem A, a maioria pontuou que todos os aspectos de A estão melhores (dinâmica, espacialidade, balanço tonal etc.) e alguns profissionais comentaram que B não parece que foi mixada. O único participante que escolheu a mixagem B não comentou sua escolha. Entre os entusiastas, os que escolheram a mixagem A relataram melhor equilíbrio tonal, mais inteligibilidade e melhor representação dos graves. Já os que escolheram B não comentaram suas respostas. Entre os amadores que escolheram A, a maioria dos relatos destacou melhor inteligibilidade dos elementos e maior equilíbrio de volumes, enquanto quem escolheu B relatou que essa versão deu mais ênfase ao sax e que isso agradou.

5.4.3.2 Par 2 - MWUN (A) x DMC (B)

Entre os profissionais, o comentário de que nenhuma das duas mixagens agradou foi recorrente. Entretanto, os que selecionaram a mixagem A comentaram que a profundidade, o balanço tonal e o uso do panorama estão melhores nessa versão. Quem escolheu B comentou que o uso da reverberação está mais interessante nessa mixagem. Entre os entusiastas, também houve comentários de que nenhuma das mixagens agradou. Quem escolheu A justificou de maneira geral sua escolha citando que o uso de reverberação em B desagradou. Quem escolheu B comentou que a mixagem, mesmo com erros, estava mais interessante e agradável. Entre os amadores, quem escolheu A comentou que B parece abafada e distante enquanto A tem maior definição dos instrumentos. Quem escolheu B não comentou sobre aspectos específicos que agradaram, apenas que preferiram essa mixagem de uma maneira geral.

5.4.3.3 Par 3 - Humana (A) x DMC (B)

Neste pareamento o grupo de profissionais foi unânime em A. Houveram comentários de que a ambiência de A está melhor, bem como a imagem estéreo e a definição dos instrumentos agradou mais. Entre o grupo de entusiastas, quem escolheu A comentou melhor inteligibilidade e definição entre os elementos e pontuou que o uso de reverberação em B causou estranhamento, enquanto quem escolheu B comentou que achou a mixagem mais adequada ao gênero musical. Entre amadores, quem escolheu A comentou maior definição dos elementos nessa versão e houveram comentários de que a mixagem B parece abafada e com "eco". Quem escolheu B não fez comentários específicos sobre o que gostou nessa mixagem.

5.5 Discussão

As respostas para o teste subjetivo aqui desenhado mostraram algumas tendências acerca da preferência dos participantes. Quando comparadas a mixagem humana com a mixagem realizada pelo modelo Console de Mixagem Diferencial (DMC), a mixagem humana foi preferida nas três músicas e essa preferência foi percebida entre todas as categorias de participantes (amadores, entusiastas e profissionais). Quando a comparação se estabeleceu entre a mixagem feita pelo modelo DMC e pelo modelo *Mix-Wave-u-Net* (MWUN) a preferência dos participantes ficou bastante dividida na primeira música, pendeu para mixagem feita pelo modelo DMC na segunda música e pendeu para a mixagem feita pelo modelo MWUN na terceira música. Por fim, quando o pareamento comparou a mixagem realizada pelo modelo MWUN com a mixagem humana, a mixagem humana foi preferida na maioria das vezes nas três músicas entre todos os estratos de participantes.

Um ponto importante de ressaltar é que os modelos tiveram acesso à um material agrupado em quatro faixas, o mesmo agrupamento aplicado à base de dados detalhado no Capítulo 4. Já as pessoas que realizaram as mixagens humanas tiveram acesso irrestrito às faixas, possibilitando outros tipos de processamentos. Parte dos comentários sobre limitações da imagem estéreo nas mixagens realizadas pelos modelos podem estar associados ao fato de que o agrupamento dos elementos não permitia grandes alterações no posicionamento de cada elemento individual em relação ao que já foi executado durante a etapa de produção. No Capítulo 6 mais detalhes serão descritos sobre maneiras de equilibrar as possibilidades de cada mixagem.

Nas condições sob as quais o teste foi realizado, os dados obtidos, portanto, rejeitam a hipótese nula de que não haveria diferença significativa na preferência entre mixagens feitas por humanos e por modelos de IA entre pessoas com diferentes níveis de conhecimento sobre mixagem. Os resultados indicam que há diferença na preferência entre as diferentes mixagens de maneira geral e que essa diferença também é notada quando analisados os resultados separados por nível de conhecimento em mixagem dos participantes.

6 Conclusão e Direções Futuras

Neste trabalho, buscou-se realizar uma extensa revisão bibliográfica sobre o atual estado da arte no âmbito dos Sistemas Inteligentes de Mixagem, incluindo um panorama atualizado sobre as aplicações de modelos generativos a outras tarefas musicais. Foram revistos em profundidade dois modelos distintos capazes de automatizar a tarefa da mixagem multicanal de maneira completa: o Console Diferenciável de Mixagem (STEINMETZ *et al.*, 2021) e o *Mix-Wave-u-Net (MWUN)* (MARTINEZ-RAMIREZ *et al.*, 2021a). Esses modelos foram treinados em um conjunto de dados totalmente livre de direitos autorais e utilizando uma função custo que leva em consideração a percepção humana. Por fim, os resultados desses modelos foram avaliados através de um teste AB, desenhado de maneira a capturar a preferência de pessoas com diferentes níveis de conhecimento sobre mixagem.

Os resultados obtidos no teste subjetivo indicaram que os modelos de inteligência artificial aqui implementados apresentam resultados inferiores, em termos de preferência entre os participantes, quando comparados diretamente com mixagens feitas por humanos. Quando as mixagens realizadas pelos modelos foram confrontadas entre si, os resultados divergiram em cada música. Para a primeira música a preferência ficou praticamente dividida entre as duas mixagens, já para a segunda música o modelo Console de Mixagem Diferenciável foi ligeiramente preferido, enquanto para a terceira música o modelo *MWUN* foi preferido. É possível notar que, entre os grupos com diferentes níveis de conhecimento em mixagem (amadores, entusiastas e profissionais) as preferências se mantiveram em todos os pareamentos, o que pode indicar a relevância de se considerar a opinião do público em geral quando se avalia um Sistema Inteligente de Mixagem.

As principais limitações que apareceram nos comentários dos participantes com relação às mixagens feitas pelos modelos foram em relação ao domínio do espaço. Comentários sobre a distribuição dos elementos na imagem estéreo e o uso de reverberação foram bastante recorrentes entre os participantes. Esse estranhamento pode ter ocorrido, pois na mixagem, de uma maneira geral, alguns processamentos no domínio do espaço são raramente executados. A voz principal, por exemplo, costuma ficar posicionada exatamente no centro do palco sonoro e qualquer pequena alteração no seu panorama para a direita ou para a esquerda não é muito usual. Outro exemplo seria o uso de reverberação, que é um efeito que se aplica em elementos específicos dentro da mixagem. Isso indica que os modelos tiveram dificuldade ao aprender a manipulação do domínio do espaço. Outros comentários frequentes entre os participantes também versaram sobre uma má represen-

tação das frequências graves e sobre a presença de artefatos nas mixagens realizadas pelo modelo *Mix-Wave-u-Net*.

Um ponto de destaque levantado anteriormente por Martínez-Ramírez *et al.* (2022) é o de que em testes de percepção para avaliação da qualidade de mixagens, os sujeitos costumam ser mais críticos com o material sendo apresentado do que seriam em outros contextos. Isso abre a possibilidade de se explorar novos métodos de avaliação do desempenho de Sistemas Inteligentes de Mixagem.

Quanto ao desempenho limitado das mixagens realizadas pelos modelos é importante destacar que, embora a maioria dos modelos de mixagem automática propostos anteriormente utilizem uma base de dados onde os arquivos de entrada são resumidos a quatro faixas (baixo, voz, bateria e outros), isso retira significativamente o controle do modelo em relação ao tipo de processamento que pode ser executado em cada elemento musical. O pré-processamento realizado na base de dados nesse trabalho, portanto, pode ter prejudicado a performance dos modelos. Em desenvolvimentos futuros, seria interessante fornecer aos humanos que executarão a mixagem os mesmos materiais fornecidos aos modelos. Também seria interessante que houvesse uma documentação adequada do processo de mixagem humana, já que o humano pode utilizar o material fornecido de maneiras que os modelos não podem, como duplicar faixas, por exemplo.

Um outro ponto importante de limitação é a quantidade de dados disponíveis que estão publica e legalmente acessíveis para o desenvolvimento desse tipo de estudo. Embora existam estudos que mostram que dados feitos para outros tipos de problemas de áudio podem ser reaproveitados para o desenvolvimento de Sistemas Automáticos de Mixagem (MARTÍNEZ-RAMÍREZ *et al.*, 2022), muitos conjuntos de dados ainda são de uso privado ou suas músicas devem ser processadas para serem utilizadas. A ampliação dos dados disponibilizados legalmente para essa finalidade também pode possibilitar a exploração de modelos que possam realizar mixagens em formatos de áudio além do estéreo. Mesmo que o formato estéreo ainda seja predominante no consumo de música, cada vez mais se tem promovido formatos imersivos de áudio. Com bases de dados adequadas, o desenvolvimento de ferramentas inteligentes de mixagem voltadas para áudio imersivo seria um caminho potencial a ser explorado.

Como detalhado ao longo deste trabalho, uma direção futura natural para os Sistemas Automáticos de Mixagem é o uso de Inteligência Artificial Generativa. No entanto, algumas discussões éticas são bastante latentes, principalmente quando levado em consideração que o aumento de dados estruturados para a mixagem é uma das barreiras que freia a implementação dessa classe de modelos. Segundo Barnett (2023), existem poucos estudos sendo conduzidos por parte dos desenvolvedores de modelos generativos aplicados ao áudio acerca dos impactos negativos e implicações éticas que esse tipo de

tecnologia pode causar. Ainda segundo a autora, existem seis principais pontos de preocupação que devem ser considerados ao desenvolver modelos generativos aplicados à música, são eles: a perda de autoria; a rigidez criativa; a predominância de viés ocidental; a violação de direitos autorais; e a apropriação cultural.

Uma outra preocupação ética foi também levantada por Morreale (2021), onde foi feito um levantamento sobre como as grandes corporações podem se beneficiar de modelos de IA aplicados a música de maneira desproporcional, prejudicando pessoas cujo trabalho pertence a alguma das etapas da cadeia de produção musical. Uma vez que muitas pesquisas relativas a modelos de IA com aplicação musical, incluindo Sistemas Automáticos de Mixagem, estão sendo desenvolvidas com o apoio de grandes empresas do mercado da música, Morreale questiona as motivações por trás do desenvolvimento desses modelos. Dado que muitas dessas empresas tem no *streamming* de músicas uma das suas principais fontes de renda, o uso de modelos de IA poderia aumentar o número de músicas sendo produzidas, mixadas e masterizadas, sem a necessidade de remunerar compositores(ras), produtores(ras), engenheiros(ras) de mixagem e de masterização.

Ao longo da história da mixagem, a interação entre o tripé da opinião pública, da(o) engenheira(o) de mixagem e da tecnologia deu origem a vários movimentos estéticos, misturando referencias entre gêneros musicais diferentes e criando novas sonoridades e movimentos artísticos. Embora o campo de estudo dos Sistemas Inteligentes de Mixagem ainda esteja em seus anos iniciais, é possível observar as potenciais contribuições que a comunidade musical pode extrair do desenvolvimento dessa área: profissionais podem automatizar tarefas exaustivas, entusiastas podem aprender mais sobre mixagem e músicos amadores podem acessar mixagens de qualidade para suas músicas.

Com isso posto, existem considerações éticas e práticas que devem ser feitas acerca do desenvolvimento de modelos capazes de realizar mixagens musicais. Destaca-se a necessidade de ouvir todas as esferas envolvidas de modo a promover um desenvolvimento responsável dessas tecnologias, levando em consideração que existe muito trabalho humano contido em cada produção fonográfica. Por último, a construção de Sistemas Inteligentes de Mixagem não deve omitir o fato de que a engenharia de mixagem é uma complexa junção de técnica e arte, que tem desempenhado um papel primordial na forma como as pessoas experienciam o ato de ouvir música desde o início da história da música gravada.

Referências

ARAÚJO, D. V. G. *Uma breve história da mixagem: origem, técnicas, perpcepção e futuros avanços*. Tese (Doutorado) — [sn], 2015. Citado na página 20.

BARNETT, J. The ethical implications of generative audio models: A systematic literature review. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. [S.l.: s.n.], 2023. p. 146–161. Citado 2 vezes nas páginas 44 e 75.

BARRY, D.; ZHANG, Q.; SUN, P. W.; HINES, A. Go listen: An end-to-end online listening test platform. *Journal of Open Research Software*, v. 9, n. 1, 2021. Citado na página 61.

BECH, S.; ZACHAROV, N. *Perceptual audio evaluation-Theory, method and application*. [S.l.]: John Wiley & Sons, 2007. Citado 3 vezes nas páginas 60, 62 e 67.

BITTNER, R. M.; SALAMON, J.; TIERNEY, M.; MAUCH, M.; CANNAM, C.; BELLO, J. P. Medleydb: A multitrack dataset for annotation-intensive mir research. In: *ISMIR*. [S.l.: s.n.], 2014. v. 14, p. 155–160. Citado na página 55.

BITTNER, R. M.; WILKINS, J.; YIP, H.; BELLO, J. P. Medleydb 2.0: New data and a system for sustainable data collection. *ISMIR Late Breaking and Demo Papers*, p. 36, 2016. Citado na página 55.

BOSI, M.; GOLDBERG, R. E. *Introduction to digital audio coding and standards*. [S.l.]: Springer Science & Business Media, 2002. v. 721. Citado 2 vezes nas páginas 22 e 23.

BROWN, T.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A. *et al.* Language models are few-shot learners. *Advances in neural information processing systems*, v. 33, p. 1877–1901, 2020. Citado na página 48.

BURRED, J. J.; ROBEL, A.; SIKORA, T. Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE, v. 18, n. 3, p. 663–674, 2009. Citado na página 24.

CAILLON, A.; ESLING, P. Rave: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011*, 2021. Citado 3 vezes nas páginas 10, 45 e 46.

CHEN, B.-Y.; HSU, W.-H.; LIAO, W.-H.; RAMÍREZ, M. A. M.; MITSUFUJI, Y.; YANG, Y.-H. Automatic dj transitions with differentiable audio effects and generative adversarial networks. In: IEEE. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2022. p. 466–470. Citado 3 vezes nas páginas 10, 47 e 48.

COLONEL, J. T.; REISS, J. Reverse engineering of a recording mix with differentiable digital signal processing. *The Journal of the Acoustical Society of America*, AIP Publishing, v. 150, n. 1, p. 608–619, 2021. Citado 2 vezes nas páginas 16 e 36.

CROITORU, F.-A.; HONDRU, V.; IONESCU, R. T.; SHAH, M. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, 2023. Citado na página 50.

DAWSON, S. *A-Weighting, the saviour of many a poor piece of equipment | HiFi Writer Blog — hifi-writer.com*. 2005. <https://hifi-writer.com/wpblog/?p=1001>. [Accessed 13-06-2024]. Citado 2 vezes nas páginas 10 e 54.

DHARIWAL, P.; JUN, H.; PAYNE, C.; KIM, J. W.; RADFORD, A.; SUTSKEVER, I. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. Citado na página 49.

DOBROWOHL, F. A.; MILNE, A. J.; DEAN, R. T. Timbre preferences in the context of mixing music. *Applied Sciences*, MDPI, v. 9, n. 8, p. 1695, 2019. Citado na página 18.

ELIASSON, S. *Comparing Compressor Interface Designs: How do visual displays on digital compressors impact how audio engineers navigate an interface and the choices they make?* 2019. Citado na página 30.

EVANS, Z.; CARR, C.; TAYLOR, J.; HAWLEY, S. H.; PONS, J. Fast timing-conditioned latent audio diffusion. *arXiv preprint arXiv:2402.04825*, 2024. Citado na página 50.

FOSTER, D. *Generative deep learning*. [S.l.]: "O'Reilly Media, Inc.", 2022. Citado 5 vezes nas páginas 44, 45, 47, 48 e 50.

GONZALEZ, E. P.; REISS, J. D. Automatic mixing: live downmixing stereo panner. In: *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx'07)*. [S.l.: s.n.], 2007. p. 63–68. Citado na página 36.

GOODFELLOW, I.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative adversarial nets. *Advances in neural information processing systems*, v. 27, 2014. Citado na página 47.

HAWTHORNE, C.; JAEGLE, A.; CANGEA, C.; BORGEAUD, S.; NASH, C.; MALINOWSKI, M.; DIELEMAN, S.; VINYALS, O.; BOTVINICK, M.; SIMON, I. *et al.* General-purpose, long-context autoregressive modeling with perceiver ar. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2022. p. 8535–8558. Citado na página 49.

IZHAKI, R. *Mixing audio: concepts, practices, and tools*. [S.l.]: Routledge, 2017. Citado 16 vezes nas páginas 9, 16, 19, 21, 22, 23, 24, 25, 27, 28, 30, 31, 32, 33, 34 e 35.

KATZ, B. *Mastering audio: the art and the science*. [S.l.]: Butterworth-Heinemann, 2003. Citado 2 vezes nas páginas 18 e 24.

KOO, J.; MARTÍNEZ-RAMÍREZ, M. A.; LIAO, W.-H.; UHLICH, S.; LEE, K.; MITSUFUJI, Y. Music mixing style transfer: A contrastive learning approach to disentangle audio effects. In: IEEE. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2023. p. 1–5. Citado na página 36.

KOSZEWSKI, D.; GÖRNE, T.; KORVEL, G.; KOSTEK, B. Automatic music signal mixing system based on one-dimensional wave-u-net autoencoders. *EURASIP Journal on Audio, Speech, and Music Processing*, Springer, v. 2023, n. 1, p. 1, 2023. Citado 4 vezes nas páginas 36, 40, 55 e 56.

KUZNETSOV, B.; PARKER, J. D.; ESQUEDA, F. Differentiable iir filters for machine learning applications. In: *Proc. Int. Conf. Digital Audio Effects (eDAFx-20)*. [S.l.: s.n.], 2020. p. 297–303. Citado na página 36.

MAN, B. D.; BOERUM, M.; LEONARD, B.; KING, R.; MASSENBURG, G.; REISS, J. D. Perceptual evaluation of music mixing practices. In: AUDIO ENGINEERING SOCIETY. *Audio Engineering Society Convention 138*. [S.l.], 2015. Citado na página 16.

MAN, B. D.; REISS, J.; STABLES, R. Ten years of automatic mixing. 2017. Citado na página 36.

MAN, B. D.; STABLES, R.; REISS, J. D. *Intelligent Music Production*. [S.l.]: Focal Press, 2019. Citado 3 vezes nas páginas 16, 36 e 60.

MARIANI, G.; TALLINI, I.; POSTOLACHE, E.; MANCUSI, M.; COSMO, L.; RODOLÀ, E. Multi-source diffusion models for simultaneous music generation and separation. *arXiv preprint arXiv:2302.02257*, 2023. Citado 3 vezes nas páginas 10, 50 e 51.

MARTINEZ-RAMIREZ, M.; REISS, J. *et al.* End-to-end equalization with convolutional neural networks. 2018. Citado na página 36.

MARTINEZ-RAMIREZ, M.; STOLLER, D.; MOFFAT, D. A deep learning approach to intelligent drum mixing with the wave-u-net. In: AUDIO ENGINEERING SOCIETY. [S.l.], 2021. Citado 3 vezes nas páginas 36, 40 e 74.

MARTINEZ-RAMIREZ, M.; STOLLER, D.; MOFFAT, D. A deep learning approach to intelligent drum mixing with the wave-u-net. In: AUDIO ENGINEERING SOCIETY. [S.l.], 2021. Citado na página 56.

MARTINEZ-RAMIREZ, M. A.; BENETOS, E.; REISS, J. D. Deep learning for black-box modeling of audio effects. *Applied Sciences*, MDPI, v. 10, n. 2, p. 638, 2020. Citado na página 41.

MARTÍNEZ-RAMÍREZ, M. A.; LIAO, W.-H.; FABBRO, G.; UHLICH, S.; NAGASHIMA, C.; MITSUFUJI, Y. Automatic music mixing with deep learning and out-of-domain data. *arXiv preprint arXiv:2208.11428*, 2022. Citado 6 vezes nas páginas 10, 36, 41, 42, 55 e 75.

MOFFAT, D. Ai music mixing systems. *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity*, Springer, p. 345–375, 2021. Citado 4 vezes nas páginas 16, 36, 38 e 39.

MOFFAT, D.; SANDLER, M. Machine learning multitrack gain mixing of drums. In: AUDIO ENGINEERING SOCIETY. *Audio Engineering Society Convention 147*. [S.l.], 2019. Citado na página 36.

MOORE, A. *An investigation into non-linear sonic signatures with a focus on dynamic range compression and the 1176 fet compressor*. Tese (Doutorado) — University of Huddersfield, 2017. Citado na página 32.

MOORE, A.; TILL, R.; WAKEFIELD, J. P. An investigation into the sonic signature of three classic dynamic range compressors. 2016. Citado 4 vezes nas páginas 30, 31, 32 e 33.

MORREALE, F. Where does the buck stop? ethical and political issues with ai in music creation. 2021. Citado na página 76.

NERCESSIAN, S. Neural parametric equalizer matching using differentiable biquads. In: *Proc. Int. Conf. Digital Audio Effects (eDAFx-20)*. [S.l.: s.n.], 2020. p. 265–272. Citado na página 36.

OWSINSKI, B. *The mixing engineer's handbook*. [S.l.]: Course Technology, Cengage Learning, 2014. Citado 9 vezes nas páginas 20, 21, 25, 30, 31, 32, 33, 34 e 35.

PHILLIPS, E. P. M. Exploring potential of the mix: Historical milestones and expanded perspectives. In: *Mixing Music*. [S.l.]: Routledge, 2016. p. 28–43. Citado na página 20.

PUJAHARI, A. Towards automatic reverb addition for production oriented multi-track audio mixing. 2017. Citado 2 vezes nas páginas 34 e 35.

REISS, J. D.; MCPHERSON, A. *Audio effects: theory, implementation and application*. [S.l.]: CRC Press, 2014. Citado na página 24.

SAVAGE, S. *Mixing and mastering in the box: the guide to making great mixes and final masters on your computer*. [S.l.]: Oxford University Press, 2014. Citado 3 vezes nas páginas 18, 32 e 35.

SENIOR, M. *Mixing secrets for the small studio*. [S.l.]: Routledge, 2018. Citado 4 vezes nas páginas 26, 34, 35 e 59.

SHENG, D.; FAZEKAS, G. A feature learning siamese model for intelligent control of the dynamic range compressor. In: IEEE. *2019 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2019. p. 1–8. Citado na página 36.

SNYDER, R. H. Sel-sync and the"octopus": How came to be the first recorder to minimize successive copying in overdubs. *ARSC Journal.*, Association for Recorded Sound, v. 34, n. 2, p. 209–213, 2003. Citado na página 20.

SOBOT, P. *Pedalboard*. Zenodo, 2021. Disponível em: <https://doi.org/10.5281/zenodo.7817838>. Citado na página 56.

STEINMETZ, C. J.; PONS, J.; PASCUAL, S.; SERRÀ, J. Automatic multitrack mixing with a differentiable mixing console of neural audio effects. In: IEEE. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2021. p. 71–75. Citado 5 vezes nas páginas 36, 43, 47, 52 e 74.

STEINMETZ, C. J.; REISS, J. D. auraloss: Audio focused loss functions in PyTorch. In: *Digital Music Research Network One-day Workshop (DMRN+15)*. [S.l.: s.n.], 2020. Citado na página 57.

STEINMETZ, C. J.; REISS, J. D. Efficient neural networks for real-time modeling of analog dynamic range compression. *arXiv preprint arXiv:2102.06200*, 2021. Citado na página 36.

STEINMETZ, C. J.; VANKA, S. S.; RAMíREZ, M. A. M.; BROMHAM, G. *Deep Learning for Automatic Mixing*. ISMIR, 2022. Disponível em: <https://dl4am.github.io/tutorial>. Citado 11 vezes nas páginas 9, 10, 16, 39, 40, 41, 43, 44, 51, 52 e 56.

STOLLER, D.; EWERT, S.; DIXON, S. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*, 2018. Citado na página 40.

TRUAX, B. *Handbook of Acoustic Ecology*. [S.l.]: Simon Fraser University, 1999. Citado na página 28.

VANKA, S. S.; SAFI, M.; ROLLAND, J.-B.; FAZEKAS, G. Adoption of ai technology in the music mixing workflow: An investigation. *arXiv preprint arXiv:2304.03407*, 2023. Citado na página 37.

VANKA, S. S.; SAFI, M.; ROLLAND, J.-B.; FAZEKAS, G. The role of communication and reference songs in the mixing process: Insights from professional mix engineers. *arXiv preprint arXiv:2309.03404*, 2023. Citado 2 vezes nas páginas 36 e 37.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017. Citado na página 48.

WRIGHT, A.; VÄLIMÄKI, V. Perceptual loss function for neural modeling of audio systems. In: IEEE. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2020. p. 251–255. Citado na página 53.

APÊNDICE A – Teste subjetivo - Perguntas de controle

Nas próximas páginas deste anexo estão contidas as perguntas de controle realizadas aos participantes do teste subjetivo. Nota-se que algumas perguntas permitiam apenas a seleção de uma ou mais das respostas possíveis enquanto outras perguntas permitiam a escrita de texto.

Teste de Percepção - Mixagem

Olá!

Neste teste serão avaliadas diferentes mixagens para uma mesma música. Para cada música você escutará duas versões por vez e terá que escolher a sua preferida entre as duas.

Você poderá também escrever o que gostou ou não em cada mixagem, o que nos ajuda muito a compreender a sua percepção, mas não é obrigatório.

O tempo estimado de resposta dessa pesquisa é de 12 minutos.

Antes de começarmos, será apresentado o termo de consentimento livre e esclarecido e serão feitas algumas perguntas de controle.

Termo de Consentimento Livre e Esclarecido

No link abaixo você pode consultar os termos dessa pesquisa. Caso aceite, você poderá seguir para o teste. Lembrando que a pesquisa é anônima, suas respostas não estarão atreladas a nenhum tipo de identificação.

https://drive.google.com/file/d/1PI7hfarNiyqOV8PthZa9c1AhPxpEJSMI/view?usp=sharing

☐ Aceito os termos.

This question is required *

NEXT QUESTION

Idade

Qual a sua faixa etária?

- ○ 18-24
- ○ 25-34
- ○ 35-44
- ○ 45-54
- ○ 55-64
- ○ 65+

This question is required *

NEXT QUESTION

Audição

Como você descreveria a sua audição?

- ○ Minha audição é normal.
- ○ Eu tenho uma audição comprometida diagnosticada por um profissional.
- ○ Não sei responder

This question is required *

NEXT QUESTION

Preferência Musical

Quais gêneros musicais você costuma ouvir? Selecione todos que se apliquem

- [ ] Axé
- [ ] Blues
- [ ] Clássica
- [ ] Eletrônica/Dance/EDM
- [ ] Hip-Hop/Rap
- [ ] Instrumental
- [ ] Jazz
- [ ] K-Pop
- [ ] Metal
- [ ] MPB
- [ ] Pop
- [ ] R&B/Soul
- [ ] Reggae
- [ ] Rock
- [ ] Sertanejo
- [ ] Outros

This question is required *

NEXT QUESTION

Conhecimento em Mixagem

Como você descreveria seu conhecimento em mixagem musical?

- ( ) Eu conheço nada ou pouco sobre mixagem.
- ( ) Eu já estudei ou estudo sobre mixagem, mas não sou um profissional da área.
- ( ) Eu sou profissional da área.

This question is required *

NEXT QUESTION

Anos de Experiência

Quantos anos você possui de experiência com mixagem?

- ◯ Não possuo experiência.
- ◯ 1-2
- ◯ 2-5
- ◯ 5-10
- ◯ 10+

This question is required *

NEXT QUESTION

Sistema

Por qual sistema você está realizando esse teste?

- ◯ Alto-falantes
- ◯ Fones de ouvido
- ◯ Uma combinação de fone e alto-falantes

This question is required *

NEXT QUESTION

Sistema

Como você descreveria o seu ambiente atual?

- ◯ Estou em um lugar fechado e silencioso.
- ◯ Estou em um lugar fechado e barulhento.
- ◯ Estou em um lugar aberto e silencioso.
- ◯ Estou em um lugar aberto e barulhento.

This question is required *

NEXT QUESTION

Sistema

Você está realizando este teste no seu sistema habitual de som, ou seja, onde você costuma ouvir música?

- ◯ Sim
- ◯ Não
- ◯ Não sei dizer

This question is required *

NEXT QUESTION

Sistema

Saberia dizer o modelo dos fones ou monitores de áudio que está usando?

NEXT QUESTION

Sistema

Se estiver utilizando monitores de áudio, a sua sala é acusticamente tratada?

NEXT QUESTION

APÊNDICE B – Teste subjetivo - Comentários

A seguir serão apresentados os comentários dos participantes do teste subjetivo feitos em cada um dos pareamentos. Os comentários estão apresentados na íntegra, exatamente da forma como foram escritos, sem correções ortográficas.

B.0.1 Primeira Música: Babe Grand - “King of the Weekend” (Techno)

B.0.1.1 Par 1 - DMC (A) x Humana (B)

· "B,pra esse tipo de musica, os graves do B me agradam mais"

· "A,Curti mais a equalização da B (graves mais pronunciados), porém a presença de artefatos estranhos na B me fez escolher a A mesmo assim."

· "B,Prefiro a B pois está com mais graves, que são característica do estilo e mais bem equilibrados entre si os elementos, a mixagem A está com uma imagem stereo muito pequena, com os elementos que estão no mid soando muito embolados. "

· "B,Eu gostei mais da clareza da voz da opção B, mas a batida ficou muito baixa. Fiquei na dúvida."

· "B,Para esse estilo, a presença de sub graves faz mais sentido e confere profundidade à música. Nesse caso, a opção B se destacou."

· "A,Gostei mais da batida mais alta, acho que combinou mais com a sensação da música"

· "B,A audição de B é mais confortável, menos "dura"e destaca melhor os vocais"

· "B,A esta bem alto os agudos. e não fica tão claro o que esta sendo falado ou cantado"

· "B,muito mais graves e profundidade"

· "B,Melhor definição e sem coloração"

· "B,Escolhi B, pois é a mix onde a voz está mais presente, mas poderia ter um pouco menos de graves."

· "B,a mixagem B está mais equilibrada, seja no volume dos elementos, seja nas frequências: na A faltam os graves e sobram médios e agudos"

· "A,B tem Graves demais"

· "B,Opção B tem mais Clareza. As fontes sonoras são melhor localizaveis. O espaço é mais equilibrado e limpo. Há RVB em proporção adquada."

· "A,Mais claro"

· "B,melhor equilíbrio de freqs."

· "B,Nesta, os graves e a voz são mais proeminentes, além de ter menos som de ambiência, o que traz mais definição para o áudio."

· "B,Voz fica mais audível"

· "B,Espectro mais equilibrado, melhor imagem estéreo e melhor nivelamento entre vozes e instrumentos."

· "B,Prefiro a versão B, com graves mais acentuados"

· "B,A opção A é meio sem grave, som distante, reverb meio feio. A B tem um som bem mais vivo e cheio"

· "B,A mixagem B é infinitamente melhor que a A. poderia escrever um texto enorme sobre as diferenças. Mas a mixagem A está em Mono, o equilibrio de volumes entre os instrumentos é ruim e desbalanceado e os timbres dos instrumentos mal ajustados."

· "B,Subs interessantes"

· "B,Stereo e mais graves"

· "A,B está com mais graves porém perde definição. A tem mais definição"

· "B,Grave melhor colocado, estereo melhor tratado."

· "B,O som da mixagem B parece mais "tratado"e a música mais nítida"

· "B,Sinto que na B podemos perceber os elementos da musica melhor"

· "B,profundidade dos graves, pulsação da música e equilibrio de frequencias"

· "B,O som esta mais equilibrado, as vozes estão mais presentes e te mais energia nos graves, o que é importante para o estilo"

· "B,alto eh sempre melhor"

· "B,tem coisas que gosto mais na primeira, tem coisas que gosto mais da segunda. o reverbão geral e o synth mais em primeiro plano da primeira me aprazem, mas escolhi a segunda pela nitidez da voz [apesar que na primeira metade é legal a voz afundada no meio da mix, mas quando entra a parte cantada prefiro com certeza a mix B]"

· "B,B mais esquilibrada, grave mais presente e medio agudo mais controlado"

· "B,A música B parece mais agradavel, a A parece mais agressiva e metálica"

· "B,Me pareceu mais intimista. Gosto mais de sons intimistas"

· "B,Acho que para o gênero em questão, os kicks mais graves e bem definidos da faixa B são mais adequados."

· "B,B"

· "B,A é mais estridente e a voz é abafada, B é mais confortável, ouço a voz e os graves com mais clareza"

· "A,Áudio mais limpo, o B parece "abafado

· "B,Tem muito eco na primeira. Dá ate uns sustos. A faixa B é a melhor"

· "B,mais envolvente"

· "B,A opção A me trouxe uma sensação de desconforto logo de cara, um som que parece de certa forma distante, onde eu escuto o vocal e os beats, mas sem o grave, algo que foi totalmente diferente na música B, só o fato de conseguir sentir o grave já mudou muito a sensação ao escutar a faixa."

· "B,A faixa A nao tem separacao estereo e nao da pra entender nada"

· "B,Os médios estão mais equilibrados e o grave está do jeito que eu gosto. Mas o synth poderia ser mais alto"

· "B,Mix em stereo, soa muito mais ampla, com profundidade e ajuda a enriquecer o arranjo, pois destaca mais os detalhes. Além disso, o spectro sonoro está bem distribuído. Regiões graves no centro, vocais bem definidos e linhas melódicas instrumentais bem distribuídas."

· "A,A primeira parece ser mais limpa e clara enquanto que a segunda parece que está faltando algo na mixagem"

· "B,Achei levemente mais calma que a A"

· "B,Beat mais encorpado (grave), mais volume dos vocais, equalização mais grave em relação à opção A. Para mim, me pareceu mais confortável de ouvir a opção B, apesar de nenhuma das duas estar bem mixada."

· "B,"B"tem o baixo bem mais energético, dá ênfase nos pontos corretos"

· "A,A alternativa "A"apresenta um aspecto mais limpo e com uma batida mais compassada, trazendo mais levessa e agrado ao escutar"

· "A,A B me soa abafada"

· "B,Em A, os instrumentos e vocal parecem estar todos amontoados numa mesma faixa de frequência e vir de uma mesma fonte. Já em B os graves estão mais presentes, há uma melhor separação dos elementos e o vocal está muito mais claro."

· "B,A está com vocais escondidos e pouca profundidade, sem graves"

· "B,Graves mais ricos destacam os efeitos na voz"

· "B,A batida mais forte é mais dinâmica (tipo um beat atrás mais forte). A pessoa cantando também combina legal"

· "B,A 'A' soa tem muita reverberação e um balanço tonal estranho, com graves e médios recuados. A B soa mais natural, apesar de eu achar que ficou um pouco "escura"(pouco agudo)"

· "B,Graves definidos, som melhor distribuído e equilibrado, profundidade"

· "B,O grave na B é bem mais presente, sendo muito mais agradável prós meus ouvidos e mais próximo do padrão de música que estou acostumado a ouvir"

· "A,Baixo mais vivo"

· "B,Achei o grave do B mais seco e me agradou mais"

· "B,Pra mim soa mais equilibrado entre vocal e instrumental"

· "B,O outro fica muito agudo pra mim."

· "B,Batida da A parece estourada. Também dá pra distinguir melhor a voz na B."

· "A,Gostei mais da primeira musica porque tem um som mais aberto, dançante. Parece que está numa sala. Mas a outra também é legal."

· "B,opção B muito melhor pelos graves e por ser menos metálico."

· "A,A mixagem A dá um tom de Minimal/Industrial e Dark, enquanto a B vira mais uma música normal e sem originalidade"

· "B,Muitos agudos no A"

· "A,Parece que a música tá mais "viva

· "B,o vocal é mais alto e a batida mais abafada"

· "B,B possui mais nitidez no som, também no fone é possível ter percepção 3D de alguns efeitos. A caixa da percurssão no A parece uma lata de goiabada, som com muita compressão"

· "B,Eu gostei mais da B, porque a música de fundo está mais baixa, o que dá uma sensação mais agradável e suave em comparação com a A."

· "B,B parece mais suave"

· "B,A é muito alto"

· "B,Grave mais definido, tem mais punch. Vocais mais claros e com mais definição. A letra "A"tem um som de reverb de sala que incomoda um pouco, tem também uma estética de música 32 bits de video game, o que é legal, mas são propostas diferentes de mixagem e eu ainda prefiro a segunda opção."

· "B,Senti que a opção B tem batidas mais fortes e algo mais abafado, para mim combinou com a música, mas ambos ficaram bons, fui pelo meu gosto mesmo"

· "B,B da pra ouvir melhor a voz e o som parece mais "encorpado

· "B,Achei o som mais nítido"

B.0.1.2 Par 2 - DMC (A) x MWUN (B)

· "B,Som mais preenchido e especializado na B"

· "A,Prefiro a mixagem A pois está com a posição de paneamento dos elementos mais concisa."

· "B,Gostei mais da B, mas o som ficou desbalanceado, favorecendo o lado esquerdo do fone."

· "A,A imagem estéreo da opção B ficou desequilibrada, em minha opinião. Por isso, a opção A foi escolhida."

· "A,Não sei explicar, mas achei que ela tem menos picos mais altos que me incomodaram mais na B"

· "B,percebo melhor a espacialidade dos canais"

· "B,trás uma sensação de profundidade, o que não ouço tão presente na opção A"

· "A,a A parece mono e a B estéreo, mas A parece mais balanceada e B parece exceder alguns limites (clipping)"

· "A,Perdeu qualidade mas ainda audível"

· "A,O som da.mix B está horrível, marquei A, mas a mix que escolhi está sem voz presente"

· "A,Parece ser igual"

· "A,ambas as mixes parecem desequilibradas, todavia a B parece mais comprimida, perdendo um pouco da dinâmica"

· "B,A é muito mono - B demasiado estéreo, levemente fora de fase e com distorção passando do pont0"

· "A,Apesar de estar ainda com muita RVB a opção A é melhor que a B que está com ruídos adquiridos por processamento de sinal com excesso de dinâmica.erando distorção e saturação"

· "A,o palco sonoro da outra opção era muito confuso"

· "B,Nesta, pois o panorama é mais aberto."

· "B,Espacialização sonora mais interessante"

· "A,a B está com a imagem estéreo bem desequilibrada. os problemas da A continuam...."

· "B,Versão B, por conta da amplitude espacial"

· "A,Não gostei muito do som da A, mas prefiro em relação à B. A separação stereo da B tá meio desconfortável, pra mim incomoda um pouco"

· "B,Novamente a mixagem A está em mono e todos os timbre centrados numa região média, com equilíbrio de volumes ruins. A voz está baixa. A mix B está melhor distribuída espacialmente na imagem estereo o que abre espaço para que se ouça melhor cada elemento. Além disso ela explora melhor o espectro de frequências do grave ao agudo. A voz pode ser melhor compreendida."

· "A,A outra tem problemas de fase"

· "B,há mais clareza nos "vocais

· "A,B soa descentralizado?"

· "B,B está com estéreo mais espalhado."

· "A,Parece que tem algo em fase ou mal alinhado."

· "A,O áudio da mixagem A é mais claro que o da mixagem B, que parece estar "descentralizado

· "B,A musica parece estar mais espacializada, mesmo com um dos elementos mais alto ainda esta mais nitida"

· "A,apesar da B ter mais info stereo, não gostei mto da distribuição,"

· "B,Tem uma impressão 3D melhor, e a voz mais presente e mais energia nos graves"

· "A,fora de fase eh pior, embora mais alot"

· "B,regiao grave e medio grave mais presente e definida em B"

· "A,A parece estar debaixo d'agua mas achei que ficou interessante. A B soa mais robótica"

· "B,De alguma forma senti a segunda opção mais ritmada"

· "A,Na faixa B o som que foi jogado para a esquerda causa bastante distração e desconforto."

· "A,A"

· "A,Parece que a B tá mais chiada ou estourada (pode ser meu celular tbm, mas acho q não pq não tá no máximo, o volume do auto falante)"

· "B,Prefiro a faixa B, pois a A ainda me traz uma sensação abafada e talvez de estar ouvindo apenas o eco da voz e na B eu sinto que o vocal já está mais próximo mesmo"

· "B,A A eh muito mono e confusa. A B tem mais separacao e da pra endender melhor, apesar do nheconheco na orelha esquerda me incomodar um pouco"

· "B,A opção A é um pouco embolada. Não consigo focar direito em nenhum dos elementos da música. A opção B também me parece um pouco desequilibrada, mas conseguimos mais destaque na voz e uma melhor percepção do arranjo como um todo."

· "A,Marquei a A por ser obrigatório marcar alguma, mas não conseguiria escolher entre A e B, não consegui reconhecer alguma diferença marcante entre ambas"

· "B,Opção B pelos vocais estarem mais altos em volume e equilibrados com o instrumental em relação à opção A"

· "A,a B parece desbalanceada. levando mais de um som para um ouvido do que outro"

· "B,B, pois a voz fica mais clara. "A"parece som de fone de ouvido de 10 reais"

· "A,Me soa similar"

· "A,Em A tenho a impressão de que tudo está dentro de uma sala com bastante reverb. Em B esse efeito é menor e há maior separação das faixas, mas em contrapartida me parece que alguns elementos da música estão mais para a esquerda e talvez até oscilando um pouco o pan, o que me incomoda mais do que o reverb"

· "B,Espacialidade mais interessante"

· "A,A segunda ficou mais baixo uma parte legal da música, e deu muita visibilidade pra voz. Prefiro a outra que não fica camuflado o beat atras"

· "A,A 'B' tem uma diferença muito exagerada entre os canais L/R, que incomoda usando fone de ouvido"

· "A,A outra opção é muito barulhenta e tem muita variação. Essa opção é mais coesa e definida"

· "B,A B tem subgraves mais fortes, além de ter dado destaque pra voz se comparado com a A e a sensação da melhora do campo estéreo também é bem evidente"

· "B,Baixo parece mais vivo também"

· "A,Achei o som do A com menos chiado e isso me agradou mais"

· "A,Nesse trecho a qualidade parece um pouco melhor na A"

· "A,Não sei muito bem o porquê, mas acho esse mais confortável de ouvir"

· "B,Gostei mais da segunda música porque parece mais harmonica, a voz está mais audivel."

· "B,Tive muita dificuldade de perceber a diferença entre essas duas"

· "B,opção B melhor pelo estéreo."

· "B,Parece ter um efeito stereo meio 3D"

· "B,a voz esta mais clara"

· "B,B parece ser um som stereo enquanto A é mono"

· "A,foi difícil escolher a melhor, muito parecido"

· "B,Sinceramente, não consegui perceber a diferença."

· "A,A parece mais suave"

· "B,A Letra "B"tem um pan estranho mas interessante, da pra ouvir melhor os vocais e os elementos de percussão, no geral, soam melhor do que os da letra "A". A letra "A"também tem um som de reverb que eu não gosto e que deixa os pratos com um som que eu acho que não combina com a música."

· "B,A opção A ficou meio "embolada", não sei se gostei disso"

· "A,Gostei mais do grave na A"

B.0.1.3 Par 3 - MWUN (A) x Humana (B)

· "B,Prefiro a mixagem B pois mesmo com o grave estando exagerado demais ela soa mais dentro do "padrão"de balanço tonal para este estilo de música."

· "B,A voz ficou bem mais nítida."

· "B,A opção B se destacou pela presença dos graves e imagem estéreo mais equilibrada."

· "A,Na B parece que está faltando alguma coisa, alguma profundidade, parece mais 'fechada'"

· "B,caso similar ao primeiro par; presença de um grave mais interessante"

· "B,Áudio definido mas com pequena modulação/DELAY e FASER..."

· "B,Marquei B aleatoriamente, as duas estão horríveis, a menos quebessa distorção seja uma proposta de produção, mas duvido disso!!"

· "B,novamente, ambas as mixes parece desequilibradas, todavia a B parece conservar os elementos essenciais à música em melhor condição de volume e compressão"

· "A,B Muito Grave"

· "B,A opção A era a B da questão anterior e segue suja. A opção B aqui é mais limpa apesar de carecer de profundidade e de definição dos planos de mixagewm"

· "B,Kick e Baixo muito melhores"

· "B,dinâmica muito superior; a outra opção está sem graves"

· "B,Nesta, o grave é mais proeminente. Este panorama me agrada mais em relação ao da faixa anterior."

· "B,Espectro mais equilibrado (grave mais eficaz) , melhor imagem estéreo e melhor nivelamento entre vozes e instrumentos. A faixa A provavelmente tem alguns problemas de cancelamento de fase na imagem estéreo."

· "A,Apesar de gostar mais do bumbo na versão B, prefiro a versão A por conta dos médios e agudos, que sumiram na versão B"

· "B,Na A falta grave. O som da B é redondinho, achei ótimo"

· "B,A mixagem A prioriza a região de médias frequências especialmente no HiHat da música chamando mais atenção do que a voz, Ocupando um espaço importante na música. Com o passar do tempo essa faixa de médios cansa o ouvido. A mixagem B é mais equilibrada, possui graves com melhor definição, a voz se destaca tendo outro elementos ao seu redor."

· "B,Subs interessantes e mix mais limpa"

· "A,Os efeitos no "vocal"deram uma encorpada mais interessante"

· "B,A parece ter compressão exagerada no médio"

· "B,"melhor escolha artisita"em comparação com a outra, os volumes na A estao estranhos, parece que quem mixou "nao conhece o estilo"."

· "B,A escolhida se aproxima mais do tipo de mixagem que eu estou acostumado a ouvir"

· "B,B tem curva de EQ mais legal e pulsa mais"

· "B,Tem os medios mais controlados, o som fica um pouco mais fechado, porem é mais agradável de se ouvir"

· "B,fora de fase eh pior, e mais alto eh melhor"

· "B,a A está muito agudo e a B está com muuito filtro de agudo, foi difícil escolher, mas acho que pelo estilo a mais "escura"me parece que faz mais sentido."

· "B,B mais equilibrado em geral, porem A tem algo de interessante/competitivo na regiao aguda e em termos de espacialidade"

· "B,A B soa mais "redonda"menos agressiva"

· "B,Mais intimista, "abafado". Sensação melhor na minha opinião"

· "B,Acredito que os volumes dos instrumentos na faixa B estão mais adequados para a música."

· "B,B"

· "B,A B tem menos eco, então eu prefiro"

· "B,Ainda sinto a A abafada em relação a B mesmo"

· "B,Grave da B muito mais presente e estereo eh o que se espera de uma musica"

· "B,A mix soa mais agradável, apesar de desequilibrada. Muitos elementos escondidos, como as frases do sintetizador, que poderiam ter mais destaque, assim como alguns elementos de percussão. No entanto, a opção A tem bem mais defeitos, como a mix no geral soar embolada e 'entubada'. Por isso, prefiro a B."

· "B,O som metálico é robótico mais alto da A realmente incomodou muito meus ouvidos."

· "A,Prefiro a opção A pois a B está com bons graves mas agudos abafados."

· "A,alguns sons do B parecem muito abafados"

· "B,B, pois a ênfase é no baixo e os vocais continuam bem audíveis. "A"tem muito foco nos agudos."

· "B,Está a batida da escolha "B"possui uma marcação de passos mais destacada que chama atenção"

· "A,B soa abafado"

· "B,Em A não há uma boa separação dos elementos, que novamente parecem estar amontoados numa mesma faixa de frequência. Além disso os graves são pouco presentes e os hihats parecem estar com uns artefatos estranhos (tipo um efeito de gate, "tremendo"). Em B tudo está mais claro, sinto que a mix tá pendendo mais pro grave mas prefiro assim rs"

· "B,Consigo escutar tudo com mais clareza na B, imagem sonora melhor"

· "B,Snare do loop "a"muito alto, mascara outros elementos do trecho"

· "B,A tem um "zigzag"que incomoda atras, parece um ruido. b parece mais limpo"

· "B,Apesar da 'B' e praticamente mutar uma track, novamente o desequilíbrio L/R da 'A' me incomodou"

· "B,Grave melhor, stereo melhor, posição dos elementos, profundida"

· "B,A B é mais agradável novamente por causa dos graves, nela é possível ouvir o groove do baixo, coisa que não dá na A. Além disso na A os agudos estão muito evidentes, na segunda ainda acho que ficaram fechados demais, mas me agrada mais. Existe a diferença no synth ressaltado também, muito melhor quando só a frase está ressaltada e tem essa estética meio espacial, do que na versão A onde o synth destacado toca a todo momento e fica meio irritante junto com o destaque dos agudos"

· "B,Me enganei, quis chamar de baixo, mas acho que seja a marcaçao viva"

· "B,Achei o B menos chiado e com o grave mais seco isso me agradou mais"

· "B,Qualidade da voz parece melhor pra mim e mais equilibrado com o instrumental"

· "B,Gosto mais do baixo do B"

· "B,A parece que tem um chiado de fundo constante"

· "B,Gostei mais da musica B porque parece mais limpa, como se fosse menos informaçao no mesmo volume."

· "B,B com graves mais presentes."

· "B,Com o tempo, o B se torna mais confortável de ouvir com o tempo"

· "B,o som parece sem ruidos"

· "B,O contra-baixo não é perceptível no A, mas sim no B. Fora que há melhor separação dos instrumentos em B. Além do fato dos graves serem melhor ouvidos."

· "B,Gostei mais da B pelo mesmo motivo da música de fundo estar mais baixa."

· "B,A repetiçao mais baixa agradou mais"

· "A,B tem um toque meio música de balada que não gosto"

· "B,A letra "B"tem um grave melhor no kick, os elementos percussivos soam melhor num geral e os vocais também soam melhor (gostei mais do efeito oitavado da voz na letra "B"do que na da letra "A"). Na letra "A"o hit-hat conflita um pouco com os vocais, tendo um volume alto demais. Ainda sorbe a letra "A", na transição pros vocais aos 15s, o crash dos pratos não soou bem, parecendo um som de pc travando."

· "B,Adorei a voz na opção B"

· "B,Gostei da batida, grave mais encorpado e da pra ouvir melhor a voz"

B.0.2 Segunda Música: Ghostly Beard - "Set Me Free" (Jazz Pop)

B.0.2.1 Par 1 - MWUN (A) x Humana (B)

· "B,Prefiro a mixagem B pois está com os elementos mais bem equilibrados em volume e paneamento, a mixagem A soa muito mais mono e embolada."

· "B,Achei os sons muito mais definidos na parte B"

· "B,A opção B apresentou uma sonoridade bem mais interessante, com a bateria mais presente e um melhor equilíbrio tonal."

· "B,Na B parece que dá para ouvir muito mais coisa, gostei mais"

· "B,A versão A parece mais "chapada". Versão B parece ter mais camadas sonoras"

· "B,a voz da musica A deveria estar na musica B"

· "B,ouveše muito melhor a instrumentação"

· "B,Traduz uma sensação orgânica mesmo sendo digitalizado"

· "B,A muito mono e com truques de fase exagerados em certos momentos"

· "B,O espaço estereofonico esta muito mais claro e límpido em B. Os elementos mixados também são percebidos com mais precisão em B."

· "B,Mais clara"

· "B,palco sonoro muito superior, a outra opção soa péssima, só a voz está melhor"

· "B,Nesta, o panorama é mais aberto e tudo tem mais definição. O timbre da bateria é mais real, pois a na A é muito sintético, coisa que me desagrada."

· "B,Espectro mais equilibrado, melhor imagem estéreo e melhor nivelamento entre vozes e instrumentos."

· "B,Gosto mais da voz na versão A, com mais médio-agudos e reverb. Mas escolhi a versão B por conta dos graves e da clareza da bateria."

· "B,Sem comentários, a mixagem A parece a de um aluno que nunca mixou na vida e está mixando pela primeira vez. Brincadeiras à parte, apensar da mix A não estar em Mono a distribuição espacial dos elementos é central. Muitos elementos na mesma faixa de frequência brigando por espaço na mixagem comprometendo a clareza das frases musicais."

· "B,Pressão mais interessante"

· "B,Nesse tipo de ritmo os graves precisam ser mais evidentes"

· "B,A parece mal equalizado, bateria sem presença"

· "B,Escolhas artisticas das coisas e tratamento."

· "B,Mixagem B possui mais clareza e nitidez"

· "B,Na opção escolhida sinto os instrumentos ocupando espaços mais bem definidos e não "batendo tanto

· "B,profundidade e curva de EQ"

· "B,A Mixagem da letra A esta bem desequilibrado o balanço tonal, soa bem embolado o som. Sinto os graves e medio grave congestionado no centro"

· "B,mais alto e o balanco parece melhor"

· "B,acho os timbres dos instrumentos muito mais definidos na mix B [o reverb da A embola tudo]"

· "B,B parece ter mais emoção e movimento"

· "B,Na faixa B o estéreo é bem agradável. Voltando da B para a A, parece que ficou tudo muito encaixotado."

· "A,A"

· "A,Em A percebo os sons de maneira mais homogênea, em B eles se destacam muito uns em relação aos outros"

· "B,A A me parece menos natural (parece que tirou alguns dos harmonicos agudos)"

· "B,Eu sinto que na faixa A há algo em maior evidência abafando outras regiões do som e na B eu estou mais próximo dos Subwoofer, mas que ainda sinto muito bem todo o conjunto da música"

· "B,A eh praticamente mono e os graves me incomodam. A da B me lembra da mix de tracks parecidas tipo o album "Turbo"da banda "Dirty Loops

· "B,Mix muito mais cristalina e equilibrada."

· "A,A única diferença que senti é que a B é mais "alta"que A"

· "B,Opção B pelos sons estarem mais claros e com vocal mais equilibrado em relação aos instrumentos."

· "B,"A"parece mais abafado"

· "B,B é muito melhor equilibrada e mais gostosa de ouvir. "A"tem foco excessivo no teclado e apaga todo o resto."

· "B,Por conta do acompanhamento de percussão a opção "B"traz mais conexão e ritmo"

· "B,Em A as baterias (principalmente o snare) parecem longe demais. O vocal tem o um bom destaque mas os outros instrumentos não têm clareza. Também parecem existir artefatos parecidos com o efeito de gate. B está infinitamente melhor, todos os instrumentos estão bastante claros e sustentam muito bem o vocal. A bateria tá incrível e dá pra sentir muito o groove do baixo, que era quase inexistente em A. Na minha opinião é uma ótima mix! c:"

· "B,Loop b com som mais brilhante e reverb mais perceptivel"

· "B,B está muito mais limpo."

· "B,B soa muito mais 'cheia' e 'expansiva'"

· "B,Som mais vívido, nítido, distribuído e equilibrado"

· "B,Gostei das duas, mas pensando em mix e master escolho a segunda, o volume dela está mais alto, além da evidência de efeitos com acho na voz. A versão A está mais seca, parece uma versão ao vivo e a B parece versão estúdio"

· "A,Vocal"

· "B,Achei o som do B mais vivo e isso me agradou mais"

· "B,Instrumental mais equilibrado, os instrumentos entre si parecem equilibrados, e o vocal tá bem melhor"

· "B,Parece mais limpo"

· "B,Gostei mais da segunda música, parece um pouco mais animada, forte."

· "B,B com instrumentos mais presentes. A com vocal mais equilibrado. As duas são boas."

· "B,Combinou mais com o estilo de música ao vivo de bar"

· "A,Som muito mais limpo"

· "B,O som do B é mais cristalino e limpo, não tão comprimido, além de ter melhor percepção de cada instrumento"

· "B,Gostei mais da B, ironicamente, porque o instrumental está mais alto e a música dá mais emoção. Acho que no caso da música eletrônica anterior, a música não transmitia a mesma alegria que essa e, por isso, foi beneficiada a que tinha um volume da música de fundo menor."

· "B,Na versão B, dá pra ouvir uma gama maior de instrumentos."

· "B,B é mais pra cima"

· "B,A letra "A"tem um som mais abafado num geral, o baixo não tem definição, os metais não tem brilho e há elementos da música que só da pra escutar na versão da letra "B", bem como os efeitos de delay na voz do cantor. Na letra "A"também o shake (chocalho) esta com um volume desnecessariamente alto, sendo até mais alto que a própria bateria, algo que não acontece na letra "B". Entretanto gostei dos vocais "mais próximos"da letra "A". A letra "B"tem um som muito médio, principalmente no baixo e na bateria, parecendo que tem pouco grave na mix."

· "A,Essa foi difícil de escolher, fui na A pois achei a voz mais clean e gostei disso"

· "B,Da pra ouvir melhor e distinguir os instrumentos"

· "B,Achei mais "agudo"e as palavras com "eco"me agradaram."

B.0.2.2 Par 2 - Humana (A) x DMC (B)

· "A,Prefiro a mixagem A pois está com o stereo bem mais aberto, com o paneamento que me agrada mais e com um balanço entre os elementos melhor."

· "A,Achei o som B abafado, como se estivesse dando uma ênfase artifical nos graves. Entretanto, eu gostei muito mais do efeito (do que me parece ser um) reverb na voz da opção B."

· "A,Apesar de soar bem mais comprimida, o que considero relativamente ruim, a opção A me pareceu mais equilibrada e com o vocal melhor encaixado na mix."

· "A,A B parece distante, a voz parece ecoar"

· "A,A mesma coisa da anterior: A parece ter mais camadas e B parece mais "chapada

· "A,musica B parece captação apenas do ambiente com exceção da voz"

· "A,B tá muito flat"

· "A,Orgânico, transmite uma sensação de ambiente fechado sem receber"

· "A,O instrumento da A com o vocal de B"

· "A,Estéreo melhor trabalhado"

· "A,A opção A é a B da questão anterior. A Opção B aqui está com o espaço pobre, fechado. Parecer que o L-R se dirigiram quase ao redor do Centro. Falta definição na equalização geral."

· "A,Mais clara e brilhante que B"

· "A,tudo parece melhor nessa opção, menos a voz"

· "A,Nesta, pois o som da sala (ambiência) está equilibrado."

· "B,Suaviza os agudos, que estão sofríveis."

· "A,Espectro mais equilibrado, melhor imagem estéreo e melhor nivelamento entre vozes e instrumentos. A faixa A é a B da questão anterior."

· "A,Prefiro a versão A por conta dos graves e médios, que acentua a caixa e o bumbo. Mas gosto mais da voz B, com bastante reverb, pois creio que combina com o clima da música."

· "A,Acho que a B tem reverb demais, som distante"

· "A,Mix B tem pouca espacialidade, em Mono, muito reverb."

· "A,Mais pressão"

· "A,na A o vocal tá no volume mais confortável, sem estranheza"

· "A,B instrumentos desequilibrados, muito para trás no campo"

· "A,a B Parece ter coisas em fase, posicionamentos estranhos. Muito reverb no hh"

· "A,O áudio A dá mais detalhe para os instrumentos e integra-os de forma mais agradável à voz"

· "A,abertura stereo, efeitos, dinâmica"

· "A,A letra "A"alem de estar mais equilibrada esta estéreo e mais punch"

· "A,a eh mais alto"

· "A,a parte instrumental com certeza a A e a parte com voz com certeza a B"

· "A,B soa abafada"

· "A,Os metais ficaram muito escondidos na faixa B. Acho que eles são importantes na música e gostei mais de como soou na faixa A. O reverb na voz também está excessivo na faixa B."

· "A,Música mais clara, áudio mais "limpo

· "B,Por mais q a A me pareça mais clara, com mais agudos, eu consigo ouvir melhor o baixo na B então escolho a B (mesmo tendo eco na voz do solo, q eu não sou muito fã)."

· "A,A faixa B me passa a sensação mais forte de um reverb na música"

· "A,A mix está muito mais equilibrada."

· "A,Apesar da A incomodar um pouco minha audição, a falta de mudança na "tonalidade"não me cativa enquanto ouvinte(estou usando aspas porque não sei se este é o termo certo)"

· "A,A por estar mais claro, enquanto B parece estar com reverberação de surround acentuada."

· "A,B parece muito abafado"

· "A,A é bem melhor de se ouvir. B tem muito foco nos vocais e nos agudos, apagando todo o resto."

· "A,A A tá muito boa, instrumentos bem separados, claros e cobrindo todo o espaço. B novamente parece que tá numa sala, muito reverb! Parece também que todos os instrumentos estão no centro, o que não é legal"

· "A,A está melhor e límpido. B parece que tem ruido e está mais baixo."

· "A,equilíbrio tonal estranho na B"

· "A,Som mais claro e nítido, cristalino, bem distribuído Centro Stereo"

· "A,A versão A está mais clara, na B a sensação é de que a gravação foi feita em um auditório ou grande salão, os instrumentos estão longes e a voz evidente e mais

seca se comparada com a A, que parece uma versão estúdio mais próxima do que estou acostumado"

· "A,Parece mais viva"

· "A,Achei a voz da opção A mais natural/acústica melhor"

· "B,Achei que no A o som tava meio abafado, B tava soando melhor"

· "B,Acho o som mais confortável"

· "A,Gostei mais da primeira musica porque parece mais harmonica/dançante"

· "A,B parece distante e menos balanceada. Opção A com camadas mais agradáveis."

· "A,O B ficou muito apagado e sem graça para uma música que parece ser animada"

· "B,voz mais demarcada"

· "A,O som de A é mais nítido, enquanto B é meio abafado"

· "A,Gostei mais da A pelo mesmo motivo anterior: o instrumental está mais alto."

· "A,Idem. Na versão A, dá pra ouvir uma maior gama de intrumentos"

· "A,A letra "B"parece que foi gravada num banheiro."

· "A,A opção A valorizou os instrumentos melhor, na B achei que tiveram batidas fortes que abafaram os outros instrumentos"

· "A,É menos "abafado"e da pra distinguir melhor os instrumentos"

B.0.2.3 Par3 - DMC (A) x MWUN (B)

· "A,Tem que escolher uma das duas?"

· "B,O equilíbrio da mixagem A me agrada mais, porém ela está soando como estivesse sendo tocada dentro de uma sala com sonoridade bem reflexiva (garagem, banheiro etc), a mixagem da B me agrada mais pois está com o som mais presente. "

· "B,Achei o B mais definido e também como se fosse algo mais "aveludado", mais agradável aos ouvidos do que a opção A"

· "A,Não gostei de nenhuma das duas, mas opção A foi a melhor. A opção B estava distorcida e com algumas falhas estranhas no áudio."

· "B,A A parece mais harmoniosa"

· "A,A tá flat mas B parece apresentar uns problemas"

· "B,Muito bom, sem distorção da ouvir de boa"

· "A,A mais equilibrada apesar de muito mono em vários trechos, mas vocais muito secos e altos. B voz com ambiente errado"

· "B,A opção B aqui é muito mais clara que a Opção A (opção B da questão anterior). Porém há algumas "saturações"coincidentes com flutuações de compressão mal ajustada."

· "A,Bateria mais viva"

· "A,tudo melhor, e os artefatos do prato da outra opção são horríveis."

· "B,Apesar desta ser ruim, igualmente no primeiro par, é menos pior que a faixa A."

· "B,as duas faixas apresentam uma mixagem não ideal. A por falta de espectro grave "no meio"da imagem e B por desequilíbrio na região médio grave. Fico com a B."

· "A,Versão A, por conta dos graves e médio-agudos mais acentuados e do reverb na voz."

· "A,Difícil escolher, mas prefiro a A. A bateria ficou muito sumida na B"

· "B,As duas mixagens apresentam problemas, aliás poderia falar de problemas em TODAS as mixagens que ouvi. Mas entre as duas a mix B é a melhor pelas mesmas razões: espacialidade em estereo, equilíbrio de volumes, timbres melhores ajustados, melhor uso de efeitos espaciais."

· "B,A outra está em mono. Mas a que escolhi também não está tão boa."

· "A,Sinto o vocal muito limpo não fica legal, se ele estiver tão limpo tem que ser com o volume no nível instrumental."

· "B,B tem Caixa e bumbo da bateria sem presença, mas melhor equalização da voz"

· "B,Mais de meu agrado a escolha estetica....."

· "B,Na opção A eu sinto que falta espacialidade e os instrumentos se chocam um pouco"

· "B,B tá meio desequilibrada, mas a voz tá mais bonita na B"

· "A,Não gostei de nenhuma, porém a letra "A", mesmo tendo a sensação de estar congestionada no centro, prefiro ela do que a letra B"

· "A,b fora de fase, a melhor"

· "A,prefiro a A apesar do teclado estar bem mais apagado e não dar pra entender direito o solo"

· "A,A soa como se estivesse ao ar livre"

· "A,O instrumental na música A está bem melhor, porém a voz tem muito reverb. Escolha difícil."

· "A,Olha, a A parece ter muito mais camadas. A B é como se os instrumentos estivessem todos em cima de um único microfone."

· "B,A B me traz a sensação de que o cantor está colado do meu lado e com um som mais natural, na A me traz a sensação de algo mais artificial no vocal"

· "B,Os dois soam esquisito e nao me agradam. Escolhi a B pq eh o menos pior."

· "A,B está "abafado

· "A,Não gosto de nenhuma das duas. Mas a segunda opção tem problemas de compressão e saturação, que me soa mais desagradável."

· "B,Deixar a percussão da bateria em menor destaque em detrimento aos outros instrumentos me fez escolher a B, pois acredito que a sonoridade fica melhor desta maneira"

· "A,A maior clareza da bateria e o vocal com reverberação deixam mais confortável em relação à B, mas nenhuma das opções está perfeita."

· "A,B parece estar levando mais para o lado direito o som"

· "A,A é melhor, a B parece meio "estourada"e tem microvariações estranhas no volume às vezes (após 12s)."

· "A,A A tá ok. A B me aparenta estar com o panning meio esquisito, os instrumentos melódicos parecem mais pra esquerda. Também apresenta artefatos do tipo gate, dá pra perceber bastante quando toca o prato"

· "A,Voz e sintetizador mais "limpos"no loop a"

· "A,Gosto como da ênfase nos trompetes ou algo assim. b parece que é muito sim misturado"

· "B,B tem uma variação estranha de volume, e soa mais unidimensional"

· "A,Mais cristalino, menos variações de volume e stereo"

· "B,A sensação dos graves é melhor na B, o som me parece bem equilibrado. Na parece que sobra Reverb e os agudos das percs ficam mais evidentes do que eu acho adequado para uma versão estúdio"

· "A,Bateria muito mais presente"

· "B,Senti que a B tem a voz com menos eco e isso me agradou mais"

· "A,Achei que o B parecia meio estourado"

· "B,Achei mais balanceado o grave e o agudo nessa"

· "B,Gostei mais dessa musica porque parece mais forte"

· "A,Nenhuma agrada muito. Opção A parece melhor nos intrumentos, mas não no vocal"

· "A,parece mais equilibrado o A"

· "A,Em B os agudos são mais estridentes"

· "B,Não notei diferença entre as faixas."

· "B,A letra "A"parece que foi gravada também num banheiro, porém a letra "B"tem o problema de ser muito abafada também. Ambas não me agradam tanto, mas a que me incomodou menos foi a letra "B

· "B,A B me animou mais kkkkkk acho que a A ta meio calma demais, um pouco entediante"

· "B,Gostei mais da voz na B"

B.0.3 Terceira Música: Bolz & Knecht - "Summertime" (Instrumental)

B.0.3.1 Par 1 - Humana (A) x MWUN (B)

· "B,Gostei mais da B. Mais definido."

· "A,Opção A de longe a melhor. Possui equilíbrio tonal, instrumentos bem destacados, ênfase na percussão no violão... A opção B não possui graves e está levemente distorcida."

· "A,Mais equilibrada"

· "B,A é melhor acabada. B me dá a sensação que estou ouvindo "ao vivo

· "A,B flat e nasalado"

· "A,Maravilha sem comparação"

· "A,Mais presente"

· "A,Saxofone com ambiente e local correto na mix"

· "A,A opção A é a melhor. A B parece que traz tudo mais ao centro e destaca as primeiras reflexões tomadas na gravação do sax."

· "A,Mais clara"

· "A,Tudo soa melhor, mas a percussão da outra é um horror :)"

· "A,Nesta, pois na B parece que nem mixada foi de tão ruim que está."

· "A,A faixa B nem foi mixada."

· "A,Versão A, com menos agudos, o que deixa o sopro mais agradável aos meus ouvidos, e com mais graves, o que permite ouvir melhor a percussão."

· "A,O som da A é bem mais aberto"

· "A,Entre as duas mixagens a A é melhor. A mix B o saxofone... aquilo era um saxofone?"

· "A,A outra está em mono e horrível"

· "A,A, no solo o instrumento solista tem que ter a maior clareza possível"

· "A,Melhor uso das ferramentas (compressor, etc...)"

· "A,O som da mixagem B parece estar "preso", sem poder ressoar completamente como no A"

· "A,profundidade e efeitos"

· "A,balanco da a eh melhor"

· "A,o sax clean da B fica com o som horroroso. o reverb da A maquia um pouco"

· "A,equalizacao e compressao muito melhor em A"

· "A,A soa melhor, parece ter tido mais tratamento ou microfones melhores, som está mais nítido"

· "A,Na faixa B o saxofone está muito mais alto que o resto da instrumentação."

· "A,A B parece que puseram uma surdina no sax. :(O timbre fica estranhom"

· "A,O som da melodia do sax está mais limpa e próxima dos outros instrumentos do que na faixa B"

· "A,Na B o sax ta dentro de um banheiro e a percussao no violao nao tem impacto algum"

· "B,Na A da pra escutar melhor todos os instrumentos"

· "A,Só detecto mixagem na primeira opção. rs"

· "A,Gostei mais da primeira pois é possível ouvir melhor os demais instrumentos pra além do sax"

· "A,Novamente, a A é muita alta"

· "A,A pelos graves estarem mais presentes, embora faltem os agudos da opção B."

· "A,B parece uma gravação caseira. não sei explicar"

· "A,A é muito melhor. B dá ênfase apenas para o saxofone."

· "A,Opção "A"trás um som mais limpo e gostoso de se ouvir"

· "A,Em A tudo está muito claro, separação ótima dos instrumentos. Só sinto que tá um pouco mais pra esquerda do que eu gostaria, principalmente a percussão. Em B parece que o volume de alguns instrumentos estão variando e há pouca energia no registro grave. Também parece haver um pouco de ruído"

· "A,Reverb do loop "a"da mais destaque no saxofone"

· "A,A tambor ficou em evidência gostei mais"

· "A,A soa com mais espacialidade"

· "A,Som parece ter maior qualidade, sons mais equilibrados"

· "A,A primeira opção soa como uma versão pós tratamentos, a segunda parece a versão original da gravação"

· "B,Eu sou saxofonista, amo ouvir um tenor"

· "A,Achei o instrumento de sopro da opção A mais afinado e isso me agradou mais"

· "A,O B tinha mais ruído, no A tava mais limpo"

· "B,Gostei mais do som da B"

· "A,Essa primeira musica é mais agradavel de ouvir"

· "A,Opção A mais equilibrada"

· "A,Parece que o som está mais limpo"

· "A,elementos mais aparentes"

· "A,Melhor equilibrio entre o saxofone(lado esquerdo) e violão(lado direito. Em B o sax parece estar à frente do microfone enquanto o violão está atrás, o que prejudica sua percepção. Além do som em A ser mais nítido e com mais graves."

· "A,Gostei mais da A porque dá para escutar melhor o saxofone."

· "A,A B parece mais crua"

· "A,O B dá uma estourada no som."

· "A,A pareceu que tem menos coisas acontecendo"

· "A,Parece que a letra "B"é a gravação bruta da música, não tem brilho nenhum no sax e no violão. Os detalhes da percussão no violão não possuem a definição e o punch que tem na letra "A

· "A,O saxofone mistura melhora com os outros instrumentos, na B ele parece isolado"

B.0.3.2 Par 2 - MWUN (A) x DMC (B)

· "B,A: estridente demais, B: abafada demais"

· "A,mixagem A está com uma sonoridade muito melhor em questão de profundidade, paneamento, presença dos instrumentos principais, a B parece toda sendo tocada em um ambiente pequeno e vivo através de um speaker."

· "B,Embora eu ache que o B tenha ficado um pouco abafado, o A esconde muito a percurssão, que fica mais pronunciada no B, e ter esse acesso mais direto à percurssão torna para mim a experiência mais prazerosa."

· "B,Não gostei de nenhuma das opções. Mas achei a B melhor."

· "A,A B da impressão que o sax está meio longe, mas na A esta meio alto. Tive dificuldade de escolher nessa"

· "A,A é mais equilibrada entre solo e acompanhamento. B tem mais acompanhamento do que solo"

· "A,não gostei de nenhuma das 2; A é menos pior"

· "B,Low fi mas ainda sim é a melhor"

· "B,B com um pouco menos de efeito/sala/profundidade no sax"

· "A,Nenhuma das duas. As duas tem sérios problemas. A está distorcida, sem mix. B com reverb errado"

· "A,A versão A fica aqui mais clara e definida do que esta B, que está com muita RVB e com espaço confuso."

· "B,Parece ter um reverb no sax que dá mais volume"

· "B,o som está afundado, velado, parece um mic de sala. mas a outra tem um material agudo muito ríspido."

· "A,Nesta, pois mesmo parecendo que não foi mixada, parece que gravaram com microfones e a B parece que gravaram com um celular um pouco distante."

· "A,Nenhuma das 2 está bem mixada. A faixa muito "dry"e com uma frequência "chata"no sax (entre 2 e 4k provavelmente) e a B com ambiência exagerada. Fico com a A por ser mais clara musicalmente."

· "A,Versão A, por estar mais aberta e brilhante, ainda que eu perceba um pouco de excesso nos agudos."

· "A,O som no B ficou bem distante"

· "A,Duas mixagens ruins, mas a A é melhor definida em timbre, os instrumentos estão espacializados na imagem estéreo, timbres mais bem definidos"

· "B,As duas são péssimas, mas escolhi a menos ruim"

· "A,Esse efeito de Reverb deixou muito embolado, preferi a A"

· "A,B muito reverb"

· "B,Vou nesse por curtir a estetica das coisas mais afastadas"

· "A,Som da mixagem B parece mais distante que o som na primeira mixagem"

· "A,A opção B parece abafada"

· "B,tá melhor, mas tem umas ressonâncias no médio grave que não gostei... A tá seca demais..."

· "A,a eh mais alto"

· "A,no geral a A tá melhor, mas o som de sax tá horrososo também [acho que é a peça atrás do saxofone o problema], mas o violão tá muito melhor na A. o reverb da B afunda tudo pra dentro da água"

· "A,B parece que está abafado ou em um lugar com eco"

· "A,Reverb demais na faixa B, e o som ficou abafado."

· "A,Violão ficou mais evidente"

· "B,Olha, difícil essa, pq a A o timbre parece distorcido e a B, com o timbre mais natural, perde um pouco da rispidez do ataque das notas que tem no A. Msm assim, vou de B."

· "B,Não sei explicar, sinto que contradiz com o que falei nas outras, mas a faixa B me trouxe um maior conforto do que na faixa A, o sax parece um pouco mais afastado, mas isso combinou com a melodia passando um clima muito gostoso mesmo, na faixa A eu tenho muito a sensação de que o sax tá sendo gravado dentro de um estúdio, mas

que os outros instrumentos já foram gravados antes e o sax está apenas acompanhando por cima a faixa mesmo"

· "A,Nao gostei de nenhuma. A A eh mono demais, a percussao do violao nao tem impacto e o violao nao "brilha". A B tem um reverb nadaver."

· "A,Duas opções desequilibradas. Mas a primeira soa mais agradável, pois traz uma percepção melhor dos elementos que compõe o arranjo."

· "B,Acho que deixar o som do instrumento principal mais grave funcionou bem pra musica"

· "A,Impressão de a A estar tudo mais perto, enquanto a B está tudo mais longe."

· "A,B parece muito distante"

· "A,A é melhor, pois pelo menos dá ênfase no sax. B é abafada, dá a impressão de ter sido gravada fora da sala/longe dos instrumentos."

· "A,A apresenta um pouco de ruído, a percussão está pouco presente mas no geral está ok. Em B o clarinete (?) que me soa como o instrumento principal poderia estar mais presente e brilhante. Os instrumentos também estão pouco separados. Essa foi a mais difícil de escolher até agora, vou ficar com a A por que o clarinete (?) tá melhor pra mim."

· "B,Loop "a"parece mais próximo mas a ambiencia do loop "b"agradou mais"

· "A,B parece que tá ruidoso, baixo"

· "A,A outra opção parece que só tinha reverb. Essa o som tem mais nitidez"

· "B,Minha sensação é de que na B parece que eu estou ouvindo um show em um teatro, já na A me parece um ao vivo ou uma vídeo gravado ao vivo e postado no YouTube. Gosto mais do B"

· "A,Saxofone mais marcado"

· "B,Senti os instrumentos de sopro mais afinado/limpo na opção B e isso me agradou mais"

· "B,O A tinha mais ruído agr e o B tava mais limpo"

· "A,Fica mais limpo pra ouvir"

· "A,Gostei mais dessa primeira musica porque é como se ela estivesse mais próxima/forte"

· "B,Eu gosto mais quando tem os elementos de percussão mais destacados"

· "A,O B pareceu muito distante, bem mais sintético"

· "A,Som com menos reverb e som ambiente. Em B parece que é apenas um microfone captando o som em uma sala vazia, o que gera um eco"

· "A,Gostei mais da A, porque senti que o saxofone está mais abafado na B."

· "B,A parece mais gravada ao vivo"

· "B,Acho q o B combina melhor com o gênero musical"

· "A,B pareceu meio abafado"

· "B,As duas tem uma estética diferente por causa do tipo de sala. A letra "B"parece que foi gravada num metro e distante dos microfones, e tem um som abafado e sem brilho, mas é uma estética interessante ainda sim. A letra "A"é bem mais próxima e não é tão abafada, mas não tem a personalidade da letra "B

· "B,achei mais "gostosinho"de ouvir"

· "A,Na B o saxofone parece distante, no A ele fica mais limpo"

B.0.3.3 Par 3 - Humana (A) x DMC (B)

· "A,A soa bem mais profissional, B parece uma captação de uma apresentação ao vivo em um local sem tratamento acústico."

· "A,Acho que o B ficou muito puxado pro grave e a percurssão ficou meio abafada justamente na parte dos estalos da percurssão aos 12-13 segundos, que é um momento muito prazeroso dessa música e que merece ser evidenciado"

· "A,Gostei mais da opção A. A imagem estéreo está mais aberta, os instrumentos mais presentes e, ainda assim, equilibrados."

· "A,Gostei mais do final da A"

· "A,percebo a presença mais claramente do sax, violão e percussão que aparenta ser no próprio violão"

· "A,B tá flat"

· "A,Lindo, aparece que estou ouvindo acústico na minha frente"

· "A,B porem suavisando o ataque da percussão e acrescentando muito sutilmente um pouco de grave do instrumento grave da percussão"

· "A,A com reverb correto - B mono e com reverb errado"

· "A,Opção A é melhor. O Balanço tonal de A é correto o de B é anasalado. O espaço em A é mais aberto e rico. Em B é muito fechado no Centro e cria uma amalgama sonoro unico embrulhado."

· "A,Idem ao anterior"

· "A,essa opção tem mais presença, parece mais equilibrada, a outra não tem graves"

· "A,Nesta, pois na B parece que, além de não ser mixada, foi gravada com um celular um pouco distante."

· "A,a Faixa A tem uma boa mixagem (equilibrada e com boa imagem estéreo), a melhor entre todas até agora para esta música. A faixa B segue com os problemas de ambiência."

· "A,Versão A, mais aberta e equilibrada e sem exagero de reverb"

· "A,O som do B ficou meio distante, pequeno"

· "A,Pressão boa de escutar. A outra está em mono."

· "A,O estalo da percussão ficou muito legal em A, e deu um Molejo pro som"

· "A,B com muito médio e reverb no sax"

· "A,cara.... curto as duas, com suas esteticas diferentes kkkkk"

· "A,som do sax , profundidade"

· "A,a eh mais alto portanto melhor"

· "A,achei essa A a mais equilibrada das opções"

· "A,B soa abafado"

· "A,Na faixa B sinto que o saxofone está distante. Na faixa A tudo ficou bem claro e agradável de se ouvir."

· "A,A"

· "A,Ouço esse metal mais claro e mais estridente em A, ele se destaca mais"

· "A,A B tem muito eco. As notas acabam se sobrepondo"

· "A,Agora a faixa B traz muito eco, como se estivesse sendo gravado dentro de um banheiro, não que seja ruim, mas a faixa A trouxe uma melhor qualidade"

· "A,A soa como uma mix profissional. B ta num banheiro com reverb e mono."

· "A,Não gostei das duas"

· "A,Acredito que tenha algum problema de fase na primeira opção, mas ela soa como uma mix. É mais agradável. A segunda opção não soa como uma mix."

· "A,A qualidade do som parece melhor"

· "A,Mais ou menos a mesma justificativa da resposta anterior."

· "A,B parece gravado em um banheiro"

· "A,A é sem dúvidas a melhor. B é abafada e dá impressão de ter sido gravada longe dos instrumentos."

· "A,A me parece pender um pouco pra esquerda mais do que eu gostaria, mas no geral todos os instrumentos estão claros e com uma sonoridade limpa e agradável. B sofre do mesmo problema de estar tudo amontoado no mesmo espaço e faixa de frequência. Dá a impressão que as tracks foram gravadas mas não mixadas ainda."

· "A,Loop "b"soa sem graves"

· "A,A tem uma batida além do trompete curti mais"

· "A,Som tem maior qualidade, bem equilibrado, som agradável"

· "A,A versão A deixa o áudio muito mais evidente, pra frente, muito mais próximo do que de costume nas plataformas de streamming"

· "B,Achei som do instrumento de sopro da opção B mais suave e isso me agradou mais"

· "B,B tava mais limpo com pouco ruído"

· "A,Fica melhor para ouvir os médios e agudos eu acho"

· "A,Gostei mais da primeira musica, é um pouco mais forte. A outra também é lega, mas prefiro a primeira.."

· "A,O A está mais alto e o B parece que está abafado"

· "A,Mesmo justificativa da anterior. Parece que a captação dos instrumentos em A são individuais, ao passo que em B é uma captação única, em uma sala que promove eco."

· "A,Gostei mais da A, porque senti o saxofone abafado na B."

· "A,B parece apresentar um eco"

· "B,O B combina melhor com o gênero musical"

· "A,B pareceu meio abafado e baixo"

· "A,A letra "A"é bem mais clara, dando pra ouvir os respiros do sax e as nuances da percussão no violão e a letra "B"é mais abafada e não tem tanta personalidade mesmo parecendo também ser gravada num metro."

· "A,Como a música tem muito instrumento, acho que prefiro o A por conseguir ouvi-los melhor"

· "A,A batida fica mais evidente e o sax mais limpo"