O ESPECTADOR COMO CO-MIXADOR: INTELIGÊNCIA ARTIFICIAL, INTELIGIBILIDADE E A CUSTOMIZAÇÃO DA ESCUTA CINEMATOGRÁFICA[1]

Rodrigo Carreiro[2]
Débora Opolski[3]

Resumo: Este artigo analisa o impacto da inteligência artificial na reconfiguração da cadeia sonora cinematográfica, da produção à recepção. Partindo da tradição verbocêntrica do cinema, discute-se a atual crise de legibilidade vocal, intensificada por escolhas estéticas e pela compressão de áudio nas plataformas de streaming. Argumenta-se que a indústria responde a esse cenário por meio de algoritmos de machine learning que aprofundam a tendência da customização da escuta, transferindo ao espectador parte do controle da mixagem. Como estudo de caso, examina-se a ferramenta Dialogue Boost, da Amazon Prime Video. A partir de análises espectrográficas das séries Maravilhosa Sra. Maisel e Fallout, demonstra-se que o recurso atua pela atenuação seletiva de sons não vocais. Os resultados indicam o reforço de um modo telefônico de representação vocal, privilegiando a clareza semântica em detrimento da fidelidade acústica e tensionando a ética da preservação da obra audiovisual.

Palavras-Chave: Inteligência Artificial; Sound design; Inteligibilidade da Voz; Customização da Escuta; Análise Espectral.

[1] Esta pesquisa contou com fomento do CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico.

[2] Professor titular da Universidade Federal de Pernambuco (UFPE), onde atua no Programa de Pós-Graduação em Comunicação e no Bacharelado em Cinema e Audiovisual. Cursou Mestrado e Doutorado em Comunicação na própria UFPE, com pós-doutorado na Universidade Federal Fluminense (UFF). É bolsista de Produtividade em Pesquisa do CNPq desde 2020. Pesquisa *sound design*, escuta, processos criativos cinematográficos e gêneros fílmicos populares, especialmente horror e *thriller*. Autor dos livros A Linguagem do Cinema (Editora da UFPE, 2021), O som no cinema de horror (Editora da UFPR, 2023) e *Cinema de Horror: Uma Introdução* (2025, com Laura Cánepa), entre outros. Lidera o LAPIS (Laboratório de Pesquisa de Imagem e Som), grupo certificado pelo CNPq. E-mail: rcarreiro@gmail.com.
ORCID: https://orcid.org/0000-0003-3087-9557.
Currículo Lattes: http://lattes.cnpq.br/0263456536361820.

[3] Professora do Instituto Federal do Paraná (IFPR) e do Mestrado em Cinema e Artes do Vídeo da UNESPAR. Doutora em comunicação e Linguagens (Cinema e Audiovisual) pela Universidade Tuiuti do Paraná (UTP). Bolsista Capes/Fulbright para estágio de doutorado na University of Southern California (USC), School of cinematic Arts (2015- 2016). Mestre em Música (Teoria e Criação) pela Universidade Federal do Paraná (UFPR). Graduação em Música (Produção Sonora) também pela UFPR. Autora dos livros *Introdução ao Desenho de Som* (Editora da UFPB, 2013) e *Edição de Diálogos no Cinema* (Editora da UFPR, 2021). Integrante dos grupos de pesquisa Cinecriare Cinema: Criação e Reflexão, da UNESPAR e LAPIS, da UFPE. E-mail: deboraopolski@gmail.com. ORCID: https://orcid.org/0000-0002-7784-3626.
Currículo Lattes: http://lattes.cnpq.br/5694754137339356.

THE SPECTATOR AS CO-MIXER: ARTIFICIAL INTELLIGENCE, INTELLIGIBILITY, AND THE CUSTOMIZATION OF CINEMATIC LISTENING

Abstract: This article analyzes the impact of artificial intelligence on the reconfiguration of the cinematic sound chain, from production to reception. Drawing from cinema's verbocentric tradition (Chion, 2008), it discusses the current crisis of vocal intelligibility, intensified by aesthetic choices and audio compression on streaming platforms. It argues that the industry responds to this scenario through machine learning algorithms that deepen the existing trend of customized listening, transferring part of the mixing control to the viewer. As a case study, Amazon Prime Video's Dialogue Boost tool is examined. Based on spectrographic analyses of the series The Marvelous Mrs. Maisel and Fallout, the study demonstrates that the feature operates through the selective attenuation of non-vocal sounds. The results indicate the reinforcement of a "telephonic mode" of vocal representation, prioritizing semantic clarity over acoustic fidelity, and straining the ethics of audiovisual work preservation.

Keywords: Artificial Intelligence; Sound Design; Speech Intelligibility; Listening Customization; Spectral Analysis.

Introdução

A terceira década do século XXI ficará marcada, na historiografia das artes e da tecnologia, como o momento da irrupção massiva da inteligência artificial (IA) no cotidiano criativo. Ferramentas baseadas em modelos generativos, como ChatGPT para textos e Midjourney para imagens, provocaram um sismo nas estruturas tradicionais de autoria, originalidade e produção estética. No entanto, enquanto os debates acadêmicos e midiáticos se concentram majoritariamente na visualidade sintética e na automação da escrita, o espectro sonoro das mídias audiovisuais testemunha a radicalização de tendências pré-existentes de customização – um movimento talvez mais insidioso, pela invisibilidade inerente, que aprofunda a individualização da experiência espectatorial.

Se, por quase um século, a evolução tecnológica do cinema foi pautada por uma busca pela fidelidade acústica, o advento de algoritmos de machine learning e redes neurais profundas (Deep Neural Networks, ou DNN) radicaliza a tendência de plasticidade do material sonoro, elevando a níveis semânticos o que antes era restrito a manipulações puramente matemáticas. Diferente dos processadores digitais de sinal (DSP) convencionais, que operam sob lógicas matemáticas fixas de equalização e compressão, a IA vai além de processar o som; ela o interpreta semanticamente e o reconstrói (De Man et al., 2014).

Essa capacidade de agência maquínica sobre a matéria sonora reconfigura a cadeia produtiva e a experiência espectatorial. Na produção, testemunhamos a ressurreição vocal de atores via clonagem sintética e a separação de instrumentos musicais em gravações originais mono, desafiando ontologias de performance e arquivo (Pritchard, 2022). É, contudo, na ponta final dessa cadeia – na recepção doméstica – que este artigo localiza seu objeto de estudo, discutindo e analisando tecnologias que, instaladas em plataformas de transmissão e dispositivos de reprodução audiovisuais, prometem equacionar uma questão persistente e paradoxal do cinema contemporâneo: a crise da inteligibilidade dos diálogos.

A tradição verbocêntrica do cinema (Chion, 2008) tem enfrentado, nas últimas duas décadas, um momento de tensão. Observa-se, globalmente, um aumento exponencial nas reclamações do público sobre a dificuldade de compreender o que é

falado, em filmes e séries (Pearson, 2022). Foi em resposta a essa demanda que surgiram ferramentas como o Dialogue Boost, implementado pela Amazon Prime Video em 2023 (Amazon Staff, 2023). Este recurso, baseado em IA, promete isolar e realçar a voz humana, permitindo ao usuário alterar o equilíbrio da mixagem original, elevando a intensidade dos diálogos sobre os demais elementos sonoros.

Este artigo propõe investigar como a inteligência artificial vem desafiando a autoridade sonora cinematográfica, ao aprofundar uma tendência de alteração dos parâmetros de mixagem, do console do sound designer (e do diretor do filme) para o controle remoto do usuário. A hipótese central é que estamos aprofundando um modelo de customização da escuta, em que o espectador assume o papel de co-mixador, priorizando a eficácia comunicativa – o modo telefônico de apresentar diálogos, conforme descrito por James Lastra (2012) – em detrimento da complexidade estética sonora da obra original, o que pode tornar sacrificar parte do potencial expressivo e estético de efeitos sonoros e música.

Para sustentar essa investigação, adotamos uma abordagem metodológica híbrida, que combina revisão bibliográfica narrativa com análise fílmica sonora. Na vertente teórica, revisitamos autores fundamentais dos estudos de cinema e de som, como David Bordwell (1985), Michel Chion (2008) e Jonathan Sterne (2012), para contextualizar a centralidade histórica do diálogo em Hollywood e as tensões entre fidelidade e inteligibilidade. Na vertente analítica, realizamos um estudo de caso da ferramenta Dialogue Boost aplicada às séries A Maravilhosa Sra, Maisel (2017-2023) e Fallout (2024-). Para superar a subjetividade inerente à percepção auditiva, utilizamos a metodologia de análise espectral visual, proposta por Carreiro e Opolski (2022).

O artigo encontra-se estruturado em seis seções – cinco das quais subsequentes a esta introdução –, desenhando um arco que vai da produção industrial à recepção individual. A segunda seção mapeia o estado da arte das tecnologias generativas e de separação de fontes (demixing), detalhando casos emblemáticos, como a restauração do documentário The Beatles: Get Back (Peter Jackson, 2021) e a síntese vocal em The Mandalorian (2019-2023). A terceira seção recua historicamente para explicar por que os diálogos sempre foram a âncora narrativa do cinema clássico e analisa as causas multifatoriais do fenômeno contemporâneo de diálogos considerados inaudíveis.

A quarta seção discute a evolução tecnológica do som digital doméstico, das predefinições de equalização em Smart TVs à lógica da mídia baseada em objetos, argumentando que a personalização individual vem se consolidando como norma de consumo. A quinta seção apresenta o núcleo empírico, com o estudo de caso da ferramenta Dialogue Boost, da Amazon Prime Video. Por fim, as considerações finais sintetizam implicações éticas e estéticas de um modelo que radicaliza uma tendência pré-existente de customização da escuta, apontando para um futuro onde a integridade da mixagem se dissolve, em favor de uma experiência algoritmicamente mediada.

IA na cadeia produtiva do som

Antes de chegar aos algoritmos que operam em Smart TVs e plataformas de streaming, a inteligência artificial vem somando novas ferramentas a etapas cruciais da produção cinematográfica. A literatura técnica recente, bem como os anais de conferências da Audio Engineering Society (AES), apontam para mudanças nos protocolos contemporâneos de tratamento do som. Deixamos a era do Processamento Digital de Sinais (DSP), baseado em lógica matemática fixa e previsível, para adentrar a era do processamento neural, onde o computador apreende as características espectrais do som através de inferência estatística (De Man et al., 2019). Em 2026, essa transformação já impacta ao menos três eixos da cadeia produtiva: a síntese vocal, a separação de fontes e a mixagem assistida.

No campo da gravação, a mudança toma forma através principalmente da síntese neural de voz (Neural Text-to-Speech ou Neural TTS) e da transferência de estilo (Voice Conversion). No passado, o uso de sintetizadores e samples podiam fazer essa tarefa, mas com frequência soavam artificiais ou robóticos, ao manipular e editar fonemas pré-gravados. Os novos modelos de deep learning conseguem modelar a prosódia, a respiração e a intenção dramática humana (Respeecher, 2021).

Um caso paradigmático dessa aplicação é o rejuvenescimento de vozes de personagens icônicos. Na série The Mandalorian (2020) e em O Livro de Boba Fett (2021), a LucasFilm enfrentou o desafio de trazer de volta o personagem Luke Skywalker com a aparência e a voz de 20 anos de idade. A solução não envolveu apenas

rejuvenescimento visual (de-aging), mas reconstituição sonora completa. A documentação técnica da empresa ucraniana Respeecher (2021), responsável pela tecnologia, informa que uma rede neural foi treinada com horas de gravações de arquivo de Mark Hamill – incluindo radionovelas, entrevistas e filmes das décadas de 1970 e 1980. O algoritmo 'aprendeu' a identidade vocal do ator jovem e aplicou o modelo sobre a performance de um dublê no set, gerando uma voz sintética indistinguível da original.

Quase ao mesmo tempo, a tecnologia desenvolvida pela empresa Sonantic permitiu que o ator Val Kilmer, após perder a capacidade de falar devido a um câncer na garganta, tivesse a voz recriada digitalmente para o filme Top Gun: Maverick (Joseph Kosinski, 2022). O processo, descrito por Mello-Klein (2022), envolveu o treinamento da IA com arquivos de áudio antigos do ator para criar um 'motor de voz' capaz de ler novas falas com a textura e a prosódia características de Kilmer. Tais avanços levantam questões éticas complexas sobre a necrofilia digital e autoria de performance. Esses tópicos se tornaram centrais nas greves do sindicato SAG-AFTRA em 2023, que debateram os limites das réplicas digitais e a substituição do trabalho humano em tarefas de Automated Dialogue Replacement (ADR), mas escapam aos objetivos deste artigo.

No eixo de edição e restauração, o paradigma está sendo modificado pela possibilidade de alterar elementos da mixagem final de forma independente dos agrupamentos de faixas pré-mixadas. Nesse ponto, é preciso explicar que sons pertencentes às três categorias principais da trilha sonora (diálogos, efeitos sonoros e vozes) são agrupados em faixas, chamadas stems, e entregues para reprodução nos cinemas e outras mídias, na mixagem final. Antes do advento da nova tecnologia, chamada Sound Separation, era impossível a qualquer pessoa que não fosse o mixador do filme alterar um elemento isolado de um desses stems; agora, não mais. Redes neurais treinadas para reconhecer padrões espectrais específicos derrubaram essa ideia, permitindo uma espécie de 'desmixagem' (do inglês demixing) de faixas consolidadas.

O documentário The Beatles: Get Back (2021), dirigido por Peter Jackson, tem sido considerado o marco zero dessa técnica no mainstream. A equipe de som, liderada por Giles Martin e Emile de la Rey, deparou-se com gravações de 1969 captadas em

mono, onde as vozes dos músicos estavam pouco legíveis, mascaradas pelo som intenso das guitarras e baterias. Para solucionar o problema, a equipe desenvolveu um software denominado MAL (Machine Audio Learning).

Segundo Martin (2025), o sistema foi treinado para diferenciar, em nível granular, o espectro sonoro de uma guitarra do som da voz humana ou de uma peça de bateria. Isso permitiu à equipe isolar detalhes de execução musical, tornando os arranjos mais discerníveis. A tecnologia de separação de fontes não ficou restrita a superproduções; ferramentas comerciais como o módulo Music Rebalance do software iZotope RX e a ferramenta Enhance Speech da Adobe democratizaram o acesso a algoritmos que limpam ruídos extremos e isolam vozes com um clique. O que antes exigia horas de edição manual espectral agora é resolvido por inferência probabilística, às vezes em questão de segundos.

Por fim, na etapa de mixagem, a IA atua cada vez mais como 'copiloto'. Ferramentas de mercado, como os plugins iZotope Neutron e smart:EQ (Sonible), utilizam aprendizado de máquina para analisar a trilha sonora e identificar conflitos de frequência em tempo real (De Man, 2019). Esses sistemas mapeiam os pontos em que instrumentos concorrem pelo mesmo espaço espectral – por exemplo, o mascaramento da voz pela trilha musical – e realizam ajustes de equalização para garantir a clareza das palavras.

Esses sistemas de mixagem assistida prometem eficiência técnica, liberando o sound designer para decisões criativas. Contudo, eles estão longe de serem unanimidade, pois críticos apontam para o risco de uma homogeneização estética (Moffat, Sandler, 2019). Se os algoritmos são treinados com bases de dados de mixagens consideradas dominantes e tradicionais pela indústria hollywoodiana, a aplicação massiva dessas ferramentas pode levar a uma padronização da sonoridade, em nível global. David Moffat e Mark Sandler (2019) apontam que idiossincrasias sonoras, texturas ruidosas e outras opções estéticas e estilísticas podem ser eliminadas em prol de uma limpeza algorítmica.

É neste cenário de manipulação algorítmica profunda – onde a voz pode ser clonada, efeitos sonoros alterados ou eliminados, e o equilíbrio espectral automatizado – que surge a demanda pela possibilidade de aplicar todo esse poder de processamento na ponta final da cadeia: a reprodução na casa do espectador. A migração dessas

tecnologias do estúdio para a sala de estar não deve ser vista como uma ruptura, mas como a automação sofisticada de práticas de ajuste que o usuário já ensaiava, com equalizadores e outras ferramentas mais simples. No atual momento, a tecnologia vai do estúdio para a sala de estar, na tentativa de resolver um problema antigo e persistente, que discutiremos a seguir: a dificuldade de entender o que os atores dizem.

A construção técnica e a materialidade da escuta

A crise da inteligibilidade e a resposta industrial via customização algorítmica não devem ser vistas apenas como problemas de engenharia, mas como sintomas de uma disputa epistemológica sobre a natureza da escuta cinematográfica. Ao analisarmos a operação do Dialogue Boost, que privilegia a clareza semântica em detrimento da textura ambiental, deparamo-nos com o que Lisa Coulthard (2013) discute sobre a escuta somática, argumentando que certos sons no cinema – como o impacto de um soco ou o uso de zumbidos e drones no limiar mais grave da audição humana, entre 20 e 120 Hz – não servem apenas para informar, mas para afetar fisicamente o corpo do espectador. Ao 'higienizar' a trilha para destacar a voz, a inteligência artificial remove essa camada de impacto físico que confere peso e presença ao mundo diegético, transformando uma experiência que deveria ser sentida no corpo numa mera transmissão de dados verbais para o intelecto.

Essa desmaterialização vai de encontro ao que propõe Miguel Mera (2016). Ao explorar a dimensão háptica do som e da música, Mera sustenta que o cinema contemporâneo busca uma tangibilidade, uma relação de devir na qual o som toca a pele da audiência e cria uma imersão sensorial completa. A intervenção subtrativa da IA, contudo, opera o desmonte dessa arquitetura sensível. Ao minimizar perceptualmente os ambientes ao fundo para isolar o primeiro plano (a voz), o algoritmo esvazia a densidade háptica da obra, que Mera e outros pesquisadores (Kulezic-Wilson, 2019; Carreiro, 2025) descrevem como característica importante do sound design cinematográfico contemporâneo, restando apenas o logos desencarnado, flutuando em um vácuo acústico artificial que nega a complexidade espacial elaborada pela direção original.

Neste contexto de interatividade, estamos diante de um debate complexo a respeito de um cenário de espectatorialidade que reestrutura toda a lógica de criação e recepção de uma obra. Afinal, a possibilidade de reconfigurar o som através de ajustes individuais também ocasiona uma reorganização dos regimes de escuta audiovisual, porque geram possibilidades múltiplas de dramaturgias sonoras e audiovisuais. Quando o audioespectador passa a construir uma hierarquia perceptiva própria, individual, a escuta deixa de ser guiada pela dramaturgia sonora proposta pela instância criativa. Porém, na medida em que diálogos, efeitos e música são criados para interagirem (entre si e com a imagem), e serem interpretados em um todo complexo, o audioespectador altera o impacto dramatúrgico da obra, as funções narrativas, afetivas e emocionais estruturantes, sempre que opta por modificar as relações entre os diversos sons componentes da trilha sonora integrada.

Para fundamentar essa transformação, é imperativo recorrer à crítica seminal de Jonathan Sterne (2012) sobre a construção cultural dos sentidos. Na introdução de The Sound Studies Reader, Sterne (2012, p. 9) desmonta o que pode ser traduzido como "litania audiovisual"4: a crença romântica de que a audição seria naturalmente imersiva, esférica, oposta a uma visão direcional; a ideia de que o som acessa nosso interior mais profundo, enquanto a visão permanece na superfícies; a noção de que o som vem até nós e nos envolve, queiramos ou não, enquanto a visão pode ser dirigida apenas para o que desejamos olhar. O autor demonstra que essas características são moldadas tecnologicamente.

O Dialogue Boost do Prime Video comprova essa tese na prática: de certo modo, ele treina o espectador para uma escuta cirurgicamente direcional e, de certa forma, desencarnada, sem peso. Ao aplicarmos a lente de Sterne (2012) ao nosso objeto, torna-se evidente que a IA opera a etapa final dessa fabricação sensorial. Ao permitir eliminar a natureza esférica (o ambiente, a imersão háptica de Miguel Mera) para garantir a direcionalidade da mensagem, a tecnologia de inteligência artificial consolida um modelo de escuta instrumental e pouco sensível.

Se a inteligência artificial oferece hoje ferramentas para 'consertar' o som, é preciso entender primeiramente o que se considera 'quebrado'. A história do cinema

[4] Do original "audiovisual litany" (Sterne, 2012, p. 9).

sonoro não é apenas uma cronologia de invenções técnicas, mas a história da consolidação de uma hierarquia estética onde a voz humana ocupa o topo da pirâmide. Desde a introdução do som sincrônico em 1927, o diálogo estabeleceu-se como o fio condutor da narrativa clássica, subordinando a música e os efeitos sonoros a uma função de suporte ou ambiência (Bordwell, 1985), com exceções pontuais quase sempre oriundas do cinema de gênero, em especial o horror (Hutchings, 2004, p. 128; Carreiro, Cánepa, 2025, p. 97). A transição do cinema sem som sincronizado para o sonoro revelou, desde o início, uma preferência ontológica. Os primeiros filmes não foram rotulados de 'sonorizados', mas de talkies (ou seja, 'falados'). O jargão popular, usado pela imprensa da época, denunciava a expectativa do público e da indústria: o som servia, primordialmente, para veicular a fala, as vozes dos atores. Efeitos sonoros e música eram secundários.

Segundo David Bordwell (1985), no modelo clássico de Hollywood, a voz tornou-se tão central para a banda sonora quanto a figura humana o é para o enquadramento visual. O teórico argumenta que a voz individualiza o personagem, explicita as motivações psicológicas e garante a causalidade narrativa. Sem a compreensão clara do diálogo, a corrente de eventos que constitui o roteiro se rompe. Por consequência, tecnologias de gravação, edição, mixagem e reprodução cinematográficas foram desenvolvidas para proteger a inteligibilidade vocal. Em 1936, a RCA introduziu o microfone unidirecional, cuja cápsula com seletividade espacial permitia aos técnicos isolar a voz dos atores, rejeitando os ruídos do set de filmagem (Carreiro, 2018). Em 1938, a indústria adotou um protocolo chamado Academy Curve, com um padrão homogêneo de equalização para salas de cinema. Esse protocolo atenuava drasticamente as frequências agudas (acima de 8.000 Hz) e graves (abaixo de 1.000 Hz), ressaltando a faixa média (entre 1.000 Hz e 4.000 Hz) – justamente a região espectral onde residem os formantes da fala humana e a articulação das consoantes (Florian, 2002).

Michel Chion (2008) traz para os estudos de som essa tendência, nomeando-a com o termo verbocentrismo. Para o teórico francês, o cinema é uma arte verbocêntrica não apenas por convenção, mas por imperativo psicoacústico. O aparelho cognitivo humano, argumenta Chion (2008), possui uma tendência inata de isolar vozes em meio a uma paisagem sonora complexa e buscar nelas um significado semântico. Quando

assistimos a um filme, não ouvimos todos os sons com a mesma atenção; nossa escuta é hierarquizada, flutuando em busca da palavra, que organiza o sentido da cena.

Apesar dessa hegemonia histórica, a relação entre a voz e o espaço sonoro nunca foi isenta de tensões. James Lastra (2012) oferece uma contribuição fundamental para entendermos o atual cenário de crise, ao propor uma distinção entre dois modos de representação sonora: o fonográfico e o telefônico. O modo fonográfico, segundo ele, almeja fidelidade perceptual (Lastra, 2012, p. 248); ou seja, busca reproduzir o evento sonoro com todas as características espaciais e acústicas originais, captadas no espaço físico onde ele ocorreu. Se dois personagens conversam em uma fábrica ruidosa, uma gravação fonográfica fiel capturaria o barulho das máquinas competindo com as vozes, a reverberação do galpão e a distância física dos atores em relação ao microfone. O resultado é um realismo acústico imersivo, porém, muitas vezes, sujo e difícil de decifrar.

Em oposição, o modo telefônico prioriza a inteligibilidade e a transmissão da mensagem, em detrimento da verossimilhança espacial. Assim como o telefone, que elimina frequências desnecessárias para garantir que a mensagem chegue clara ao outro lado da linha, o cinema clássico prioriza a clareza da voz. Lastra (2012) observa que, na maioria dos filmes, mesmo se um personagem estiver distante da câmera ou em um ambiente barulhento, ouvimos sua voz com uma clareza e proximidade impossíveis na vida real. O som é limpo, seco e frontal, garantindo que nenhuma frase importante do roteiro seja perdida pelo espectador.

Essa dicotomia explica a natureza das ferramentas de IA contemporâneas. Ao minimizarem sonoridades para destacar a voz, recursos como o Dialogue Boost operam uma ‘telefonização’ mais radical da trilha sonora. Muitas vezes, essas tecnologias sacrificam a complexidade fonográfica (o ambiente, a música, a textura do mundo) para garantir a eficiência telefônica (a entrega da mensagem verbal). De todo modo, se o modelo telefônico reinou durante o século XX, as primeiras décadas do século XXI trouxeram uma ruptura que resultou na atual crise da inteligibilidade da voz. A partir dos anos 2000, e intensificando-se de 2010 em diante, cinéfilos e críticos começaram a notar um fenômeno desconcertante: acumulavam-se, em redes sociais e fóruns de Internet, reclamações de audioespectadores de que os diálogos estavam se tornando inaudíveis (Thornton, 2022).

Essa crise explodiu na cultura pop após o lançamento de Batman: O Cavaleiro das Trevas Ressurge (The Dark Knight Rises, 2012), de Christopher Nolan. A voz do vilão Bane (Tom Hardy), abafada por uma máscara e processada digitalmente, gerou uma onda de reclamações, obrigando o diretor a remixar o áudio antes do lançamento massivo. Nolan tornou-se pivô DA polêmica, com filmes subsequentes como Interestelar (Interstellar, 2014) e Tenet (2020) sendo alvo de críticas ferozes por mixagens que enterravam os diálogos sob camadas de música e efeitos sonoros.

O problema não é autoral, mas sistêmico, pois questões técnicas que envolvem o trabalho com áudio digital podem causar percepções desconfortáveis para o audioespectador, que aprendeu a ouvir os diálogos sempre de forma preponderante, quase sempre mais altos do que efeitos e música. Filmes e séries de diferentes realizadores e países de origens, vêm sendo objeto de reclamações parecidas. O seriado de guerra britânico SS-GB (2017), a série de aventura da Marvel O justiceiro (2017-2019) e o drama político brasileiro O mecanismo (2018) são três de muitas obras criticadas pelo público, a partir da mesma premissa comum: os diálogos estariam quase na mesma intensidade de música e efeitos sonoros. Mais de 100 pessoas foram ao Twitter reclamar desse problema, no dia da exibição do primeiro episódio de SS-GB, em 2017; o debate chegou à Câmara dos Lordes (o Congresso Nacional da Inglaterra), com parlamentares questionando a BBC por isso (Ward, 2017). José Padilha e Selton Mello, diretor e ator do seriado brasileiro O mecanismo, abordaram o mesmo tema em entrevistas, algum tempo depois, apontando um suposto erro de mixagem (Brito, 2019).

Em uma reportagem publicada no site especializado SlashFilm, Ben Pearson (2022) entrevistou editores, mixadores de som e sound designers sobre o fenômeno, revelando que existe consenso entre os profissionais de que o fenômeno é real. Essa unanimidade se estende ao reconhecimento de que não existe uma causa isolada, mas uma complexa interação de fatores que se agravaram nos últimos 20 anos. Dentre esses fatores, destaca-se a escolha estética naturalista adotada por diretores como Nolan, Alejandro Iñárritu e Alfonso Cuarón, que buscam um audiovisual mais verossímil, incorporando erros, incompreensões e murmúrios típicos da vida cotidiana. O hábito de consumir conteúdo audiovisual em dispositivos como smartphones e notebooks, com reprodução sonora de baixa qualidade; o uso maciço de microfones de lapela (que sacrifica frequências entre 2.000 e 4.000 Hz, essenciais para a clareza de consoantes, e

gera artefatos que degradam a qualidade acústica); e a compressão severa de áudio para plataformas de streaming (que diminui a dinâmica sonora, aproximando as distâncias entre os níveis de intensidades dos diálogos, da música e dos efeitos), tirando a clareza das transientes da fala (Pearson, 2022; Thornton, 2022) são outros fatores.

A customização da escuta

A resposta da indústria tecnológica à crise da inteligibilidade acelerou um movimento pré-existente de fragmentação e personalização. Historicamente, o ouvinte de música já dispunha de ferramentas para intervir na obra (como equalizadores gráficos e seletores de loudness); o que vemos agora é a transposição dessa cultura da intervenção para o domínio da narrativa audiovisual, agora mediada por metadados e algoritmos inteligentes. Nesse ponto, cabe observar que a individualização da experiência do usuário não é exatamente uma novidade, mas ocorria de forma mais limitada há poucos anos. Um exemplo pode ser encontrado por ocasião da digitalização do sinal televisivo, consolidada globalmente entre meados dos anos 2000 e a década de 2010 (no Brasil, o desligamento do sinal analógico iniciou-se em 2015). A mudança de tecnologia alterava a relação do espectador com o aparelho receptor; o televisor deixava de ser um mero exibidor de sinais de broadcasting para tornar-se um computador, capaz de processar metadados e oferecer ao usuário um menu de intervenções estéticas sobre a obra audiovisual.

No período analógico, a intervenção do audioespectador sobre o som se limitava ao ajuste de intensidade geral (volume) e, em aparelhos de maior fidelidade, a controles de equalização (frequências graves, médias e agudas). Com a massificação das Smart TVs, os fabricantes começaram a implementar perfis de processamento de áudio predefinidos (presets), desenhados para compensar as deficiências acústicas dos alto-falantes cada vez mais finos e posicionados na parte traseira dos aparelhos.

Os presets de áudio, em especial os modos ligados ao reforço dos diálogos ou da voz, funcionam aplicando uma curva de equalização em forma de sino, que acentua as frequências entre 1.000 Hz e 4.000 Hz, ao mesmo tempo em que aplicam compressão dinâmica para reduzir a discrepância entre sons fracos e fortes. Embora funcionais, essas

ferramentas terminam por 'achatar' a mixagem final, o que limita sua eficácia: ao aumentar a faixa de frequência onde ficam os transientes, muitas vezes distorcem a música ou alteram a timbragem geral. A verdadeira revolução na customização exigiria uma mudança na própria estrutura do arquivo de áudio, migrando do paradigma de canais para o de objetos sonoros.

Nesse ponto, é necessária uma explicação mais técnica. A tecnologia digital mais tradicional de som, adotada no começo da década de 1990, é baseada em canais (Channel-Based Audio), seja estéreo (2.0) ou surround (5.1, 7.1). Nesse modelo, o mixador distribui os efeitos sonoros e músicas por cada alto-falante disponível, direcionando a voz para o canal central. Uma vez renderizado o arquivo final, essas decisões são imutáveis, como uma fotografia impressa onde não se pode mais mover os personagens (Carreiro, 2018).

Desde o lançamento do Dolby Atmos, em 2012, e de concorrentes como o Auro 3D, a tecnologia de reprodução sonora mudou, passando à Mídia Baseada em Objetos (Object-Based Media), que, embora tecnicamente disruptiva, não concretiza uma ruptura ontológica ou epistemológica no modelo de escuta audiovisual dominante, mas sim a maturação tecnológica da tendência de individualização da escuta. Embora parta de uma mixagem tradicional em 7.1, a Mídia Baseada em Objetos trata cada elemento sonoro – um diálogo, um efeito sonoro – como objeto de áudio independente, que possui metadados e pode viajar entre os alto-falantes em qualquer direção (BBC Research & Development, 2023).

Esses metadados contêm instruções sobre onde o som deve ser posicionado no espaço 3D, qual sua intensidade relativa e como ele deve se comportar. O arquivo final não é uma mixagem pronta, mas um pacote de dados que será renderizado apenas no momento da exibição, pelo processador de áudio da casa do audioespectador. Isso permite uma flexibilidade inédita: se o diálogo é um objeto separado da música, o usuário final pode, teoricamente, alterar a intensidade apenas do diálogo, sem afetar o resto da trilha.

Neste sentido, tecnicamente, é preciso que haja uma reconfiguração dos objetivos criativos da etapa da mixagem. No novo contexto de customização da escuta, a função de um mixador pode vir a ser, mais do que criar um produto finalizado (como propõe a pós-produção de som, atualmente), mas um conjunto de camadas sonoras

reconfiguráveis em tempo real, num processo contínuo, ao gosto do audioespectador. No contexto da exibição em uma sala escura, lutava-se para controlar a escuta, tornando-a coletiva; o público ouvia a obra finalizada, de preferência em ambiente controlado, e com dispositivos adequados. No contexto contemporâneo da exibição predominantemente individual, a mixagem audiovisual deve criar um esquema de interação entre arquivos sonoros que permita múltiplas configurações adaptáveis de escuta. O mixador, então, passa a ter a função de estabelecer relações entre camadas e objetos sonoros, definindo o que pode ser alterado e o que deve ser preservado para que a dramaturgia sonora se sustente, deixando como resultado não uma mixagem final fechada, mas um campo de possibilidades de escutas sonoras.

Pesquisas apontam que o público está interessado em experimentar. A emissora pública britânica BBC foi pioneira em testar as implicações de acessibilidade dessa tecnologia. Ward (2017) relata uma série de experimentos realizados com a tecnologia de objetos para resolver as queixas de inteligibilidade de sua audiência. Em um teste com transmissões de partidas de futebol, a emissora permitiu que os telespectadores ajustassem, via controle remoto, o balanço entre o áudio da narração e o som da torcida no estádio. Os resultados foram contundentes. Segundo a pesquisa, 77% dos participantes aprovaram a funcionalidade e manifestaram o desejo de que o recurso fosse implementado definitivamente (Ward, 2017). A BBC expandiu os testes para a dramaturgia, lançando um episódio da série Doctor Who com áudio binaural e opções de mixagem personalizável.

Esses experimentos demonstraram que o público deseja atuar como co-mixador. A passividade espectatorial, presumida pelo modelo de cinema clássico, dá lugar a uma interatividade funcional. O espectador contemporâneo, consumindo conteúdo em ambientes adversos – no ônibus ou metrô com fones de ouvido, na cozinha com o barulho da geladeira, ou na sala com crianças brincando –, necessita adaptar o produto audiovisual a realidades acústicas distintas. A transição do consumo audiovisual para as plataformas de streaming acelerou ainda mais a adoção da lógica baseada em objetos, ainda que de forma híbrida. Plataformas como Netflix, Amazon Prime Video e Disney+ precisam entregar conteúdo para um ecossistema fragmentado de dispositivos (smartphones, tablets, soundbars, home theaters). Metadados

acoplados aos arquivos de áudio instruem o aplicativo sobre como guiar os objetos sonoros para cada situação (International Telecommunication Union, 2017).

Foi neste contexto que ferramentas que utilizam tecnologias de inteligência artificial, como o Dialogue Boost da Amazon, emergiram: não como uma anomalia, mas como a concretização comercial da noção de Mídia Baseada em Objetos, agora devidamente assistida por inteligência artificial. Se a tecnologia OBM fornece a estrutura (a separação dos elementos), a IA fornece a automação (a identificação e o isolamento de cada componente sonoro). Ao oferecer ao usuário opções como o Dialogue Boost, a Prime Video está, na prática, entregando uma interface simplificada de uma mesa de mixagem. O espectador não precisa entender de decibéis ou frequências; ele apenas seleciona o nível de conforto desejado. Contudo, como veremos na análise a seguir, a maneira como a IA executa esse comando revela estratégias de processamento que desafiam a integridade estética da obra original.

Estudos de caso: A Maravilhosa Sra. Maisel e Fallout

Para investigar empiricamente como a inteligência artificial intervém na estrutura da mixagem cinematográfica, no caso do Dialogue Boost, buscamos séries de destaque da plataforma, dos últimos três anos, que disponibilizam a ferramenta para o usuário. Selecionamos como objeto de estudo a série Maravilhosa Sra. Maisel (The Marvelous Mrs. Maisel, Amy Sherman-Palladino, 2017-2023), uma produção original da Amazon Prime Video. A escolha justifica-se por três critérios: primeiramente, trata-se de uma produção de orçamento generoso, cuja mixagem original segue padrões rigorosos de qualidade técnica, eliminando variáveis como captação de som direto defeituosa. Em segundo lugar, a série caracteriza-se por diálogos rápidos e sobrepostos, típicos do gênero cômico, o que torna a inteligibilidade um desafio intrínseco à estética escolhida. Por fim, a produção foi uma das pioneiras a disponibilizar o recurso Dialogue Boost, em abril de 2023 (Amazon Staff, 2023).

A ferramenta Dialogue Boost oferecia, ao usuário, em 2023, três níveis de intervenção: Low, Medium e High. Em 2025, ela disponibiliza apenas o Medium e o High. Para fins de contraste analítico, optamos por realizar uma análise comparativa entre a

trilha sonora original (selecionada como padrão pela plataforma) com a versão Dialogue Boost: High, que aplica o nível máximo de processamento algorítmico. A análise recaiu sobre os três primeiros episódios da quinta temporada.

Como sabemos, a análise de áudio esbarra com frequência na subjetividade da escuta dos analistas. Para superar essa limitação e conferir maior rigor científico à investigação, desenvolvemos uma metodologia de análise espectral (Carreiro, Opolski, 2022) e a utilizamos no estudo de caso. Esta abordagem propõe a transposição de dados sonoros para o domínio visual, através de espectrogramas tridimensionais. Este procedimento permite a observação objetiva de parâmetros físicos do som, que poderiam passar despercebidos numa audição menos especializada.

Vale a pena mencionar, ainda, que em mudanças de intensidade muito sutis, os gráficos espectrais podem não apresentar diferenças visíveis. Nesse sentido, reforçamos que essa abordagem metodológica constitui uma opção plausível para a análise de trilhas sonoras, mas sempre como apoio secundário à percepção sonora. A escuta atenta continua a ser o sentido primário a desencadear a análise. Contudo, é preciso lembrar que a escuta é um ato individual que, embora possa ser treinado e desenvolvido, está sujeito a idiossincrasias de natureza fisiológica e cognitiva, que o individualizam. O que queremos dizer é que, via de regra, a percepção visual do gráfico espectral não é uma metodologia que substitui a percepção sonora, mas a complementa.

O protocolo técnico consistiu na extração digital do áudio estéreo diretamente da plataforma de streaming. Utilizou-se a placa de som virtual Soundflower para rotear o áudio do navegador para o software de gravação Audacity, garantindo a integridade do sinal sem conversões analógicas. Os arquivos resultantes, no formato WAV sem compressão, foram importados em seguida para o software de restauração e edição espectral iZotope RX.

O iZotope RX gera gráficos tridimensionais, nos quais o eixo horizontal representa o tempo, o vertical representa as frequências (de 20 Hz a 20.000 Hz, mesma faixa da audição humana), e a intensidade é representada pela cor: tons esbranquiçados e alaranjados indicam alta energia sonora (volume mais alto), enquanto o fundo preto representa o silêncio; portanto, quanto mais clara a cor laranja, maior a intensidade do som. Essa ferramenta funciona como uma espécie de microscópio sonoro, permitindo

dissecar o som gravado em todas as camadas constituintes. As imagens que ilustram esse artigo, oriundas dos espectrogramas, foram geradas pelo iZotope RX.

O primeiro passo da análise foi a medição da intensidade média (Loudness Integrated) de cada obra, utilizando o padrão internacional ITU-R BS.1770-4, que regula níveis de áudio para broadcasting e streaming (International Telecommunication Union, 2017).

Figuras 1 e 2 - Episódio completo de A Maravilhosa Sra. Maisel, sem e com Dialogue Boost

Fonte: iZotone RX (acervo dos autores)

A expectativa inicial era de que a versão com Dialogue Boost apresentasse níveis de intensidade superiores, dada a promessa de reforço do diálogo. No entanto, a medição do primeiro episódio revelou um dado curioso. A trilha sonora original apresentou um Loudness Integrated de -21.4 LKFS5. Este valor está dentro dos padrões de medição utilizados pela plataforma Amazon Prime Video, cuja intensidade sonora média exigida aos mixadores é de -24 LKFS, com uma margem de 2 a 3 LKFS para mais e para menos (Amazon, 2016).

A surpresa maior veio depois: a versão de áudio extraída com a ferramenta Dialogue Boost: High ligada registrou -35.6 LKFS. Esse valor representa uma queda abrupta de 14.2 dB na intensidade sonora. Segundo Holman (2005, p. 213), uma variação de 10 dB é percebida psicoacusticamente pelo ouvido humano como uma redução de 50% da intensidade. Portanto, o audioespectador que ativa o recurso de

5 A intensidade da pressão sonora de produções audiovisuais para streaming é regulamentada internacionalmente pela norma técnica ITU-R BS.1170, acordada em 2015. Esta norma foi criada pela International Telecommunication Union (Sindicato Internacional de Telecomunicações, em tradução livre), estabelecendo o LKFS (Loudness K-eighted relative to Full Scale) como unidade de medida padrão a ser utilizada para a normalização do áudio, em sistemas de transmissão televisivas e streaming de música e vídeo (INTERNATIONAL, 2017). A normalização é a técnica capaz de igualar a intensidade dos sons ouvidos, tanto de uma produção para a outra dentro da mesma plataforma, quanto em plataformas diferentes.

reforço do diálogo no primeiro episódio recebe, ao contrário do que se poderia esperar, um som com menos intensidade. Essa discrepância implicou em algumas consequências para o trabalho de análise.

A primeira, e mais óbvia, residiu na falta de normalização no algoritmo de implementação da ferramenta. As duas faixas de áudio têm, entre si, uma discrepância superior às margens de 2 a 3 LKFS aceitos pela Amazon Prime Video. A partir desta primeira experiência, decidimos verificar se outros episódios tinham a mesma característica. Extraímos a trilha sonora dos episódios 2 e 3 da mesma série e descobrimos que, neles, ambas as versões analisadas (original e com Dialogue Boost: High) estão com loudness integrated alterado, variando entre -32 e -34 LKFS. Em outras palavras: embora possuam entre si níveis de intensidade semelhantes, as duas versões de áudio também estão com intensidade 50% abaixo da marca estabelecida pela Amazon para programas exibidos na plataforma.

Intrigados com essas informações, decidimos efetuar uma comparação com níveis de loudness integrated de outras produções originais da Amazon Prime Video. Extraímos as trilhas de áudio dos filmes O Som do Silêncio (Sound of Metal, Darius Marder, 2019), no qual o loudness integrated atinge média de -22,5 LKFS. Em Coda – No Ritmo do Coração (Sian Heder, 2021), o loudness integrated médio é de -21,4 LKFS. Essas medições nos deram a certeza de que a Prime Video segue o padrão da norma técnica internacional; por motivo desconhecido, contudo, na série Maravilhosa Senhora Maisel, a plataforma permitiu um padrão de loudness diferenciado. De toda forma, nos episódios 2 e 3, a diferença entre as versões original e com Dialogue Boost: High existe, mas é pequena (loudness integrated variando entre -32 e -34 LKFS). Esse fato possibilita uma percepção mais clara das alterações impressas à trilha sonora pela ferramenta, e por consequência, a análise comparativa entre as duas versões pôde ser feita de modo mais preciso.

A principal anomalia detectada – a intensidade geral cerca de 50% mais baixa na trilha geral de áudio de todos os episódios analisados – evidenciou a estratégia subtrativa da ferramenta de inteligência artificial da Amazon Prime Video: ao tentar isolar o diálogo, o algoritmo parece ter descartado tanta informação sonora (efeitos sonoros e música, sobretudo) que a energia média dos arquivos despencou.

Em seguida, para tentar compreender como o Dialogue Boost opera em nível microestrutural, isolamos cenas específicas. A análise visual dos espectrogramas confirmou que a inteligibilidade prometida pela Amazon não estava ocorrendo pela amplificação da voz, mas pela redução de intensidade sonora de ambientes, efeitos sonoros e músicas. A ferramenta de inteligência artificial podia isolar diálogos, mas ao invés de elevar a intensidade dos diálogos, fazia o contrário – reduzia os demais sons.

Um teste ainda mais complexo para a IA ocorreu aos 28 minutos do terceiro episódio. Na cena, dois personagens conversam em um bar, com música ao vivo. Eles falam sobre outros personagens, por sua vez sentados em mesas dispostas entre os protagonistas e o palco. As falas dos protagonistas possuem intensidade constante; os diálogos dos outros atores são evidenciados apenas quando viram assunto da conversa principal. Nesses momentos, a câmera focaliza as pessoas que são objeto da conversa, e as vozes delas são apresentadas com mais intensidade, para trazer ao som a perspectiva mais aproximada da edição de imagens, e ao mesmo tempo reforçar a presença no recinto.

Figuras 3 e 4: Cena da conversa no bar, sem e com Dialogue Boost

Fonte: iZotone RX (acervo dos autores)

Ocorre então um reforço sensorial: a imagem fornece elementos para ilustrar a informação semântica, porque o movimento de câmera e os cortes da montagem sublinham a história que está sendo verbalizada. Além do diálogo principal e das falas dos personagens que são o foco da conversa, temos outros dois elementos sonoros importantes para a cena: a música e o vozerio. Ambos parecem, em um primeiro momento, estar apenas compondo o ambiente e dando-lhe verossimilhança, mas ao final da cena, se articulam com o enredo, pois a banda para de tocar, reagindo a uma

ação mal educada da plateia. Neste momento, o vozerio se manifesta, mostrando contrariedade.

Nesta cena, o desafio algorítmico era maior: tanto a música cantada quanto o vozerio da multidão contêm sonoridades vocais humanas. Será que a IA seria capaz de distinguir a voz do protagonista do canto do vocalista, ou do vozerio dos figurantes? Para aumentar nossa dúvida, trata-se de uma cena sonoramente complexa. A observação do espectro mostra menos cor alaranjada na versão com Dialogue Boost; portanto, há menos intensidade sonora. Direcionando o foco do olhar para os segundos iniciais, conseguimos perceber com clareza a ausência relativa de frequências agudas e de seus devidos harmônicos. Uma visão aguçada pode identificar inúmeras outras diferenças.

A música e o vozerio tiveram a intensidade bastante minimizada na versão com Dialogue Boost. Em outras palavras: a ferramenta desenvolvida pela Amazon Prime Video não deu ênfase a essas vozes; agrupou-as junto com os demais sons do ambiente e minimizou as intensidades, para que apenas as falas personificadas, proferidas por personagens narrativamente relevantes, fossem enfatizadas. A partir desse dado, podemos concluir que o Dialogue Boost não identifica a qualidade de um som vocal e trabalha especificamente com ele. A ferramenta é programada para identificar os stems6 de sons, minimizando os stems de efeitos sonoros e música.

Além disso, o deslocamento do olhar analítico para os 18 segundos finais da cena comprova visualmente a nossa afirmação: o Dialogue Boost, neste caso, adiciona inteligibilidade não por enfatizar o diálogo, mas sim por minimizar a intensidade das outras sonoridades. As demais informações sonoras, como vozerio e música, permanecem, só que em menor intensidade, como podemos perceber observando o tom mais claro da cor laranja e os picos de intensidade da forma de onda (a linha horizontal azul no meio dos gráficos). A proposta estética da cena foi alterada; a música tocada pela banda e o burburinho da plateia foram suavizados, o que tornou a sequência menos tensa.

[6] *Stem* é o nome dado ao conjunto de sons que compreendem cada um dos 3 principais grupos da mixagem final: diálogos, efeitos sonoros e música. Na prática o áudio mixado é dividido em stems para que possa ser dublado em outros idiomas, mantendo a concepção do design sonoro original.

Nos exemplos analisados, o algoritmo parece operar através de uma hierarquização de foco: ele detecta os padrões de fala que estão em primeiro plano, e classifica todo o resto como ruído. Na versão original, a música e o walla (o vozerio) preenchem os espaços entre as falas dos protagonistas, criando uma mise-en-scène sonora contínua. Na versão com Boost, esses elementos de fundo são comprimidos, criando um efeito de pumping (oscilação de intensidade) perceptível. A ferramenta de IA não funciona como amplificador, mas como equalizador dinâmico subtrativo.

Além de analisar o uso do Dialogue Boost em uma das séries pioneiras a disponibilizar a ferramenta para o usuário, decidimos repetir a análise em outra produção mais recente, para verificar se o padrão continuava o mesmo. Após realizar uma apreciação auditiva de obras lançadas entre 2024 e 2025, identificamos que a resposta é sim: o padrão de uso da ferramenta – reduzir a intensidade de música e efeitos sonoros, mantendo o nível dos diálogos definidos pela mixagem original – é idêntico.

Ao procurar por obras realizadas nos últimos dois anos que contêm o Dialogue Boost, percebemos que são inúmeras as cenas em que a atuação da ferramenta é auditivamente perceptível. Em vez de atuar como recurso de ajuste discreto, a ferramenta frequentemente reorganiza de maneira abrupta a hierarquia das camadas sonoras, comprometendo (ao menos a ouvidos treinados) a reconstrução do espaço acústico. A título de exemplo, gostaríamos de citar quatro momentos de produções realizadas em 2024 e 2025.

De 2024, destacamos o episódio inicial da primeira temporada da série Fallout (Graham Wagner e Geneva Robertson-Dworet, 2024-). Numa cena de luta que ocorre aos 26 minutos, a supressão da música deixa um vazio sonoro peculiar, esvaziando a tensão dramática que tradicionalmente se constrói entre música, efeitos e diálogo. Outro exemplo bastante claro vem da série Expatriadas (Expats, Lulu Wang, 2023-2024), também do episódio inaugural da primeira temporada. Há um momento em que a canção entoada por Margaret, em resposta à música diegética, evidencia uma descontinuidade perceptiva entre a voz e o ambiente musical em que ela se encontra, aos 47 minutos.

Do ano de 2025, destacamos um momento oriundo do episódio inicial da segunda temporada de Fallout. Durante uma conversa aos 49 minutos, a explosão

(literal) da cabeça de um personagem tem seu potencial expressivo bastante reduzido pela atenuação brusca e grande dos efeitos sonoros, enfraquecendo o impacto sensorial do acontecimento. Por fim, o primeiro episódio da temporada inicial da animação The Mighty Nein (Critical Role Productions, 2025-), possui várias cenas nas quais a ausência quase absoluta de música e efeitos sonoros produz um esvaziamento artificial e súbito do espaço acústico, no qual o diálogo se apresenta como uma entidade sonora isolada, flutuando em um vácuo acústico.

Para efeito de comparação mais detalhada e aprofundada, feita com o uso do método de análise espectral, elegemos entre os quatro exemplos citados a cena da explosão da cabeça no episódio inicial da segunda temporada da série Fallout. Ao fazer a medição da intensidade média (Loudness Integrated) dos episódios, utilizando os mesmos critérios aplicados anteriormente, percebemos que a trilha sonora original está com Loudness Integrated de -23.3 LKFS. Já a versão da trilha sonora extraída com a ferramenta Dialogue Boost: High habilitada está com -31.2 LKFS. Como se pode notar, trata-se de padrão idêntico ao verificado em Maravilhosa Sra. Maisel.

Figuras 5 e 6 - Episódio da primeira temporada de Fallout, sem e com Dialogue Boost

Fonte: iZotope RX (acervo dos autores)

A visualização comparada da onda sonora no centro das duas imagens espectrais, que exibem o conteúdo sonoro do episódio completo, permite perceber a diferença significativa de intensidade com o Dialogue Boost ativado (Figura 5) e desativado (Figura 6). Para complementar nossa análise, aplicamos um zoom e visualizamos de maneira detalhada os gráficos espectrais do momento da explosão da cabeça, aos 48’35. Na cena, um trabalhador submetido à implantação de um chip cerebral que permite o controle de suas ações pela empresa desperta e troca breves

palavras com Lucy. A cabeça dele explode no exato momento em que profere, pela terceira vez (e aos gritos) a frase “Go home!”.

Acompanhando as figuras a partir dos diálogos transcritos, é possível notar que as informações de fala estão presentes em ambas as representações; contudo, na segunda imagem (Figura 8), verifica-se a acentuada redução da informação sonora dos efeitos sonoros, em tonalidade laranja-amarelada. A supressão da maior parte dos sons da explosão altera de maneira decisiva a experiência auditiva da cena, produzindo versões cujo impacto sensorial e expressivo se mostra radicalmente distinto.

Figuras 7 e 8 - Cena da conversa, sem e com Dialogue Boost

Fonte: iZotope RX (acervo dos autores)

Nossos achados nesta experiência analítica confirmam, materialmente, a teoria de James Lastra (2012) sobre o privilégio do modo telefônico no áudio de produções audiovisuais. O Dialogue Boost transforma a rica complexidade sonora em uma transmissão limpa e funcional, embora ainda verossímil e eficiente. A inteligibilidade é priorizada a ponto de descaracterizar – pelo menos no mais intenso dos modos de aplicação do Dialogue Boost – a narrativa espacial. Assim, o audioespectador ganha a palavra, mas perde o mundo em volta.

Considerações finais

A investigação conduzida ao longo deste artigo, culminando na análise espectral da ferramenta de inteligência artificial denominada Dialogue Boost, aponta para uma transformação estrutural na ontologia do som cinematográfico. Se, no século

XX, a mixagem final de um filme era encarada como uma estrutura fixa e autenticada pela visão do diretor, o século XXI, mediado por algoritmos de inteligência artificial, parece reconfigurar a obra audiovisual como um banco de dados fluido, passível de rearticulação em tempo real pelo usuário final. Isso tem consequências positivas e negativas.

O estudo de caso da série Maravilhosa Sra. Maisel revelou uma dissonância entre o discurso de marketing e a realidade técnica. A Amazon rotulou sua ferramenta com a promessa de um 'impulso' (boost), sugerindo acréscimo de energia. Contudo, os dados espectrais analisados demonstram que a IA opera pela subtração. Para garantir a inteligibilidade em um ecossistema de dispositivos precários (smartphones, tablets) e ambientes ruidosos, o algoritmo sacrifica parte da complexidade ambiental da obra.

Essa constatação valida a tensão teórica proposta por James Lastra (2012), sugerindo o Dialogue Boost como uma ferramenta de mecanização do modo telefônico de escuta cinemática. Ele prioriza a mensagem verbal, ao mesmo tempo em que relativiza o realismo fonográfico. Ao suavizar o som do metrô ou da banda no bar para destacar a fala, a ferramenta modifica a diegese fílmica e prioriza a compreensão semântica sobre a riqueza acústica; a palavra sobre a sensação.

Talvez a implicação mais significativa deste fenômeno seja o deslocamento da autoria. Mais do que inaugurar uma prática, ferramentas como o Dialogue Boost e as tecnologias de mídia baseadas em objetos aprofundam uma tendência já consolidada de individualização da experiência espectatorial no consumo doméstico. Assim como os sistemas de reprodução musical há décadas oferecem opções de equalização que permitem ao ouvinte intervir na massa sonora, essas novas ferramentas elevam o espectador ao status de co-mixador. Se, historicamente, a indústria operou sob um modelo de 'tamanho único' – no qual a mixagem final aprovada pelo diretor era a única via de acesso à obra –, a tecnologia contemporânea torna essa mixagem mais líquida, permitindo que os parâmetros originais sejam reconfigurados conforme a conveniência de quem ouve.

Tudo isso sugere um dilema ético sobre a integridade da obra de arte. Se um diretor como Christopher Nolan mixa intencionalmente um diálogo para ser mais sentido do que ouvido, até que ponto é aceitável que uma textura sônica caótica de guerra – como em Dunkirk (2014) – seja suavizada de acordo com o conforto do

espectador? E, no sentido contrário, em que medida a liquidez da mixagem transforma cada indivíduo diante de uma tela em coautor daquilo que vê? A fronteira entre customização (ajuste de conforto) e intervenção editorial (mudança de sentido) torna-se nebulosa.

Ao delegar ao algoritmo e ao usuário o poder de alterar o balanço espectral, a indústria permite que a intenção do autor se torne adaptável, de acordo com as preferências e a relação de experiência que o usuário deseja. O filme deixa de ser um discurso fechado para se tornar um serviço ajustável. Nesse sentido, a emergência da IA na ponta do consumo da cadeia produtiva do audiovisual sugere que o papel do mixador e do sound designer pode estar mudando. Em vez de esculpir realidades sonoras de universos diegéticos fechados, esses profissionais passam a trabalhar como arquitetos de possibilidades sonoras. Mais do que 'melhorar' o som de filmes e séries, a inteligência artificial pode estar reescrevendo o contrato social entre criadores e público, dentro da lógica de que o audioespectador possui uma liberdade estruturada, pois ele altera o que for permitido por realizadores e distribuidores.

Referências

AMAZON Original Spec and Delivery. **Gearspace**, 31 mar. 2016. Disponível em: https://gearspace.com/board/post-production-forum/1077559-amazon-original-spec-delivery.html. Acesso em: 25 nov. 2025.

AMAZON STAFF. Prime Video launches a new accessibility feature that makes it easier to hear dialogue in your favorite movies and series. **About Amazon**, 18 abr. 2023. Disponível em: https://www.aboutamazon.com/news/entertainment/prime-video-dialogue-boost. Acesso em: 24 nov. 2025.

BBC RESEARCH & DEVELOPMENT. Object-based media. **BBC**, 2023. Disponível em: https://www.bbc.co.uk/rd/object-based-media#work. Acesso em: 24 nov. 2025.

BORDWELL, David. The introduction of sound. In: BORDWELL, David; THOMPSON, Kristin; STEIGER, Janet (org.). **The classical Hollywood cinema**: film style & mode of production to 1960. New York: Columbia University Press, 1985. p. 298-308.

BRITO, Carlo. 'O mecanismo' volta e Selton Mello explica sussurros: 'Eu não falo baixo, as pessoas falam alto'. **G1**, 10 maio 2019. Disponível em: https://g1.globo.com/pop-arte/noticia/2019/05/10/o-mecanismo-volta-e-selton-mello-explica-sussurros-eu-nao-falo-baixo-as-pessoas-falam-alto.ghtml. Acesso em: 24 nov. 2025.

CARREIRO, Rodrigo. A história do som dos filmes. In: CARREIRO, Rodrigo (org.). **O som do filme**: uma introdução. Curitiba: Editora da UFPR, 2018. p. 35-86.

CARREIRO, Rodrigo. O papel da respiração no cinema de horror. **E-Compós**, v. 23, 2020.

CARREIRO, Rodrigo; OPOLSKI, Débora. O espectro do som como ferramenta de análise fílmica. **PÓS: Revista do Programa de Pós-Graduação em Artes da EBA/UFMG**, v. 12, n. 24, p. 388–414, 2022.

CARREIRO, Rodrigo. Construindo um cinema sensorial: música, voz e efeitos sonoros além de fronteiras. **Revista Música**, v. 25, n. 1, p. 377–401, 2025.

CARREIRO, Rodrigo; CÁNEPA, Laura Loguercio. **Cinema de Horror:** Uma Introdução. São Paulo: Gênio Editorial, 2025.

CHION, Michel. **A audiovisão**. Lisboa: Texto e Grafia, 2008.

COULTHARD, Lisa. Dirty sound: Haptic noise in new extremism. In: VERNALLIS, Carol; HERZOG, Amy; RICHARDSON, John (orgs.). **The Oxford Handbook of Sound and Image in Digital Media**. New York: Oxford University Press, p. 115-126, 2013.

DE MAN, Brecht et al. The mix evaluation dataset. In: **137th Audio Engineering Society Convention**. Los Angeles: AES, 2014. Disponível em: https://www.aes.org/e-lib/browse.cfm?elib=17478. Acesso em: 24 nov. 2025.

FLORIAN, Brian. Learning from history: Cinema sound and EQ curves. **Secrets of Home Theater and High Fidelity**, v. 9, n. 2, 2002. Disponível em: https://hometheaterhifi.com/volume_9_2/feature-article-curves-6-2002.html. Acesso em: 24 nov. 2025.

HOLMAN, Tomlinson. **Sound for film and television**. London: Routledge, 2005.

HUTCHINGS, Peter. **The Horror Film**. Edimburgh: Person Education Limited, 2004.

INTERNATIONAL TELECOMMUNICATION UNION**. Recommendation ITU-R BS.1770-4**: Algorithms to measure audio programme loudness and true-peak audio level. Genebra: ITU, 2017. Disponível em: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1770-4-201510-I!!PDF-E.pdf. Acesso em: 24 nov. 2025.

KULEZIC-WILSON, Danijela. **Sound design is the new score**: theory, aesthetics, and erotics of the integrated soundtrack. Oxford: Oxford University Press, 2019.

LASTRA, James. Fidelity versus intelligibility. In: STERNE, Jonathan (org.). **The Sound Studies Reader**. New York: Routledge, 2012. p. 248-254.

MERA, Miguel. Materializing film music. In: COOKE, M.; FORD, F. (Eds.). **The Cambridge companion to film music**. Cambridge: Cambridge University Press, p. 157-172, 2016.

MOFFAT, David; SANDLER, Mark. Approaches in Intelligent Music Production. In: **Arts**. v. 8, n. 4, p. 125, 2019. Disponível em: https://www.mdpi.com/2076-0752/8/4/125. Acesso em: 24 nov. 2025.

PEARSON, Ben. Here's why movie dialogue has gotten more difficult to understand (and three ways to fix it). **SlashFilm**, 22 ago. 2022. Disponível em: https://www.slashfilm.com/673162/heres-why-movie-dialogue-has-gotten-more-difficult-to-understand-and-three-ways-to-fix-it/. Acesso em: 24 nov. 2025.

PRITCHARD, Huw. The Beatles: Get Back - The Technical Story. **Sound on Sound**, fev. 2022. Disponível em: https://www.soundonsound.com/techniques/beatles-get-back. Acesso em: 24 nov. 2025.

RESPEECHER. Recreating the Legendary Luke Skywalker Voice for Disney's The Mandalorian. **Respeecher Case Studies**, 23 ago. 2021. Disponível em: https://www.respeecher.com/case-studies/legendary-luke-skywalker-voice-mandalorian. Acesso em: 24 nov. 2025.

STERNE, Jonathan. The Audio-Visual Litany. In: STERNE, Jonathan (org.). **The Sound Studies Reader**. New York: Routledge, p. 3-12, 2012.

THORNTON, Mike. TV subtitle usage up to 80% – what is going wrong with dialogue mixes?. **Production Expert**, 17 nov. 2022. Disponível em: https://www.pro-tools-expert.com/production-expert-1/tv-subtitle-usage-up-to-80-what-is-going-wrong-with-dialogue-mixes. Acesso em: 24 nov. 2025.

WARD, Lauren. Speak up! Why some TV dialogue is so hard to understand. **The Conversation**, 27 abr. 2017. Disponível em: https://theconversation.com/speak-up-why-some-tv-dialogue-is-so-hard-to-understand-75423. Acesso em: 24 nov. 2025.

Recebido em 30/01/2026
Aceito em 07/07/2026