UNIVERSIDADE FEDERAL DO PAMPA

OESLEY RODRIGUES MACHADO

BAHOKÊ: UMA ABORDAGEM DE APRENDIZADO PROFUNDO PARA SEPARAÇÃO CEGA DE FONTES E GERAÇÃO DE TRILHAS DE KARAOKÊ

Bagé
2025

OESLEY RODRIGUES MACHADO

BAHOKÊ: UMA ABORDAGEM DE APRENDIZADO PROFUNDO PARA SEPARAÇÃO CEGA DE FONTES E GERAÇÃO DE TRILHAS DE KARAOKÊ

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Engenharia de Computação como requisito parcial para a obtenção do grau de Bacharel em Engenharia de Computação.

Orientador: Gerson Alberto Leiria Nunes

Bagé
2025

Ficha catalográfica elaborada automaticamente com os dados fornecidos pelo(a) autor(a) através do Módulo de Biblioteca do Sistema GURI (Gestão Unificada de Recursos Institucionais).

M149b Machado, Oesley Rodrigues

Bahokê: Uma Abordagem de Aprendizado Profundo para Separação Cega de Fontes e Geração de Trilhas de Karaokê / Oesley Rodrigues Machado.

80 f.: il.

Orientador: Gerson Alberto Leiria Nunes

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal do Pampa, Engenharia de Computação, 2025.

1. Sistema de Karaokê, Processamento de Áudio, Separação Cega de Fontes, BSS, Separação de Fontes Musicais, Tecnologia de Karaokê, Aprendizado Profundo . I. Título.

OESLEY RODRIGUES MACHADO

BAHOKÊ: UMA ABORDAGEM DE APRENDIZADO PROFUNDO PARA SEPARAÇÃO CEGA DE FONTES E GERAÇÃO DE TRILHAS DE KARAOKÊ

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia de Computação como requisito parcial para a obtenção do grau de Bacharel em Engenharia de Computação.

Dissertação defendida e aprovada em: 16 de dezembro de 2025.

Banca examinadora:

_____

Prof. Dr. Gerson Leiria Nunes

Orientador

(UNIPAMPA)

_____

Prof. Dr. Bruno Silveira Neves

(UNIPAMPA)

_____

Prof. Dr. Sandro da Silva Camargo

(UNIPAMPA)

Assinado eletronicamente por **SANDRO DA SILVA CAMARGO**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 18/12/2025, às 21:13, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **GERSON ALBERTO LEIRIA NUNES**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 19/12/2025, às 16:09, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **BRUNO SILVEIRA NEVES**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 19/12/2025, às 16:21, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Dedico este trabalho ao meu pai, à minha mãe, à minha parceira de altas peripécias e a todos aqueles que persistem resilientes em busca de um futuro melhor.

AGRADECIMENTO

Agradeço a Deus por me permitir concluir este trabalho e a todos aqueles que, de forma direta ou indireta, colaboraram na elaboração do mesmo.

Agradeço profundamente aos meus pais, seu Odilon e dona Neusa, que não mediram esforços para me permitir concluir esta jornada e me tornar um engenheiro de computação. Meu muito obrigado.

Agradeço especialmente à minha noiva, Lourdes Maria Blanco Picanço Severo, que, além do incentivo à elaboração deste trabalho, contribuiu de outras formas. Meu muito obrigado.

Agradeço imensamente ao meu orientador, Prof. Dr. Gerson Alberto Leiria Nunes, pela sua disponibilidade, comprometimento, apoio e incentivo a este trabalho. Meu muito obrigado.

Agradeço ao professor e amigo Daniel Brum da Silva por participar deste trabalho atuando como *stakeholder* (parte interessada) do *software* desenvolvido. Meu muito obrigado.

Agradeço a todos que colaboraram com suas respostas na pesquisa realizada para o desenvolvimento deste trabalho, em especial familiares, amigos e professores do Instituto Artístico Carlos Gomes de Dom Pedrito/RS. Meu muito obrigado.

Agradeço aos meus colegas, docentes, à equipe diretiva e aos colaboradores da Unipampa, Campus Bagé, por fazerem parte desta trajetória. Meu muito obrigado.

“O computador é burro, na medida em que lhe falta a capacidade de hesitar.”
— Byul Chul-Han

RESUMO

Esta pesquisa propõe um sistema de karaokê inovador que utiliza técnicas avançadas de processamento de áudio, especificamente a Separação Cega de Fontes (BSS), para aprimorar a experiência do usuário. A principal contribuição é a resolução de uma lacuna na tecnologia atual de karaokê, pois ao permitir a extração direta de faixas instrumentais da música original, elimina a necessidade de recriar manualmente as partes instrumentais, um processo tradicionalmente demorado e complexo. Nesse sentido, a pesquisa também se aprofunda na aplicação avançada da separação de fontes de áudio, investigando e selecionando algoritmos BSS adequados para extrair faixas instrumentais da música popular brasileira ou regional. Desse modo, neste estudo, a qualidade das faixas geradas pelo sistema BSS é avaliada em comparação com as demais já existentes, considerando critérios como qualidade de separação, artefatos sonoros e eficiência computacional. Os resultados da avaliação de usuários validaram a eficácia da solução proposta, visto que o sistema Bahokê demonstrou alta usabilidade e aceitação, com $91,7\%$ dos participantes manifestando a intenção de uso diário/semanal. Esta validação empírica comprova que a aplicação de BSS e processamento de áudio de ponta não só supera os sistemas tradicionais em eficiência, mas também satisfaz a experiência do usuário, estabelecendo um novo padrão para ferramentas de prática musical. Em suma, este trabalho propõe e descreve um sistema de karaokê sofisticado que oferece separação vocal de alta qualidade, superando os sistemas tradicionais por meio da aplicação de processamento de áudio de ponta e redes convolucionais.

Palavras-chave: Sistema de Karaokê, Processamento de Áudio, Separação Cega de Fontes, BSS, Separação de Fontes Musicais, Tecnologia de Karaokê, Aprendizado Profundo .

ABSTRACT

This research proposes an innovative karaoke system that utilizes advanced audio processing techniques, specifically Blind Source Separation (BSS), to enhance the user experience. The main contribution is solving a gap in current karaoke technology, because by allowing the direct extraction of instrumental tracks from the original song, it eliminates the need to manually recreate the instrumental parts, a traditionally time-consuming and complex process. In this sense, the research also delves into the advanced application of audio source separation, investigating and selecting suitable BSS algorithms for extracting instrumental tracks from Brazilian or regional popular music. Thus, in this study, the quality of the tracks generated by the BSS system is evaluated in comparison with existing tracks, considering criteria such as separation quality, sound artifacts, and computational efficiency. User evaluation results validated the effectiveness of the proposed solution, as the Bahokê system demonstrated high usability and acceptance, with $91,7\%$ of participants expressing an intention to use it daily/weekly. This empirical validation proves that the application of BSS and cutting-edge audio processing not only surpasses traditional systems in efficiency but also satisfies the user experience, establishing a new standard for music practice tools. In summary, this work proposes and describes a sophisticated karaoke system that offers high-quality vocal separation, outperforming traditional systems through the application of state-of-the-art audio processing and convolutional networks.

Keywords: Karaoke System, Audio Processing, Blind Source Separation, BSS, Music Source Separation, Karaoke Technology, Deep Learning .

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
ADC	*Analog-to-Digital Converter*
AI	*Artificial Intelligence*
ASR	*Automatic Speech Recognition*
BSRNN	*Band-split Recurrent Neural Network*
BSS	*Blind Source Separation*
CAGR	*Compound Annual Growth Rate*
CPU	*Central Processing Unit*
CSS	*Cascading Style Sheets*
DAC	*Digital-to-Analog Converter*
DOM	*Document Object Model*
EEG	Eletroencefalograma
FAVENI	Faculdade de Venda Nova do Imigrante
GPU	*Graphic Processing Unit*
GUI	*Graphic User Interface*
HTML	*Hypertext Markup Language*
IC	Intervalo de Confiança
ICA	*Independent Component Analysis*
IRM	*Ideal Ratio Mask*
JS	*JavaScript*
JSON	*JavaScript Object Notation*
LSTM	*Long Short-Term Memory*
MEG	Magnetoencefalografia
MOS	*Mean Opinion Score*

MSS	*Music Source Separation*
NMF	*Non-negative Matrix Factorization*
PCM	*Pulse-Code Modulation*
ReLU	*Rectified Linear Unit*
SAR	*Source-to-Artifact Ratio*
SCA	*Sparse Component Analysis*
SDR	*Signal-to-Distortion Ratio*
SIR	*Source-to-Interference Ratio*
STFT	*Short-Time Fourier Transform*
UFPel	Universidade Federal de Pelotas
UI	*User Interface*
UNIPAMPA	Universidade Federal do Pampa
WER	*Word Error Rate*

SUMÁRIO

1 INTRODUÇÃO

Um autêntico clássico nunca se torna obsoleto; os karaokês ficaram famosos no Brasil na década de 1990; porém, continuam atraindo admiradores e conquistando novos locais nos estados brasileiros. Para dimensionar o entretenimento fornecido, quantifica-se que, em um único lugar, como a feira de São Cristóvão, no Rio de Janeiro, existem 70 karaokês. No coração das tradições nordestinas, entre lojas de artesanato e restaurantes típicos, existem mais de 70 bares, cujo principal atrativo é a música ao vivo, atraindo um público diversificado de todas as idades e movimentando as noites, madrugadas e os finais de semana (PINNA; ROCHA, 2025).

1.1 História

Segundo Zhou e Tarocco (2013), a ideia predominante é que o karaokê surgiu no Japão, devido à sua localização geográfica e seu criador ser Daisuke Inoue (músico e inventor japonês) que, infelizmente, não patenteou sua ideia em 1971, sendo ela registrada por Roberto del Rosario (empresário filipino) em 1975 como Sistema *Sing-Along* (cante-junto). O termo é uma combinação de duas palavras japonesas: *kara*, derivada de *karappo* (vazio), e *oke*, um acrônimo de *okesutura* (orquestra). No entanto, Hosokawa e Mitsui (2005) explicam que a palavra original *karaoke* em japonês não se traduz em "orquestra vazia", mas sim em "a orquestra na gravação está desprovida de vocais", aludindo tanto à máquina de karaokê quanto ao ato de cantar.

Importado para o Brasil em 1996 por um empresário coreano, o videokê, um dispositivo que combina karaokê e vídeo, foi desenvolvido pela *RAF Eletronics* buscando promover a habilidade musical dos seus usuários, além de impulsionar a movimentação de bares e outros locais de entretenimento. Similarmente, pode ser uma maneira de reunir toda a família em casa à noite ou nos fins de semana. O funcionamento ocorre a partir da seleção de uma canção contida em um cartucho, cuja letra e melodia são liberadas simultaneamente ao pressionar o botão de *play*, permitindo que o usuário cante e acompanhe a exibição da letra na tela. Quando a canção termina, o dispositivo avalia a performance e atribui uma pontuação ao usuário (BARELLI, 1996).

A avaliação da performance do cantor pode ser feita de várias formas, o método difere conforme o fabricante do aparelho, porém, na maioria das situações, os dispositivos apenas consideram o ritmo da música cantada. Em outras palavras, se o usuário cantar

ou falar as palavras com precisão, isto é, dentro do tempo estipulado pelo software, será bem pontuado. Na maioria dos casos, um algoritmo é responsável por identificar as sílabas da canção e validar se foram entoadas corretamente pelo cantor, contando o número de acertos para determinar a pontuação. Contudo, existem modelos que utilizam métodos mais avançados, buscando espelhar a maneira como os humanos avaliam uma performance. Além do ritmo, uma avaliação do volume da voz durante a canção e uma comparação da melodia são realizadas para confirmar se o cantor acertou o tom da canção (ESTRANHO, 2011).

1.2 Problema de pesquisa

Contextualizando, a separação cega de fontes (BSS, do inglês *Blind Source Separation*) é um campo do processamento digital de sinais que visa isolar fontes de áudio individuais a partir de uma mixagem complexa, sem informações prévias detalhadas sobre as fontes originais ou sobre como elas foram combinadas. Em uma gravação de uma música em que a voz do cantor e os instrumentos estão todos misturados em um único arquivo, a tecnologia BSS busca extrair, de forma autônoma, a voz dos instrumentos desta gravação (FU, 2025).

No contexto do karaokê, o potencial do BSS é vasto e revolucionário, uma vez que, tradicionalmente, a criação de faixas de karaokê envolve a reprodução ou recriação da parte instrumental da música, um processo que é demorado e nem sempre resulta em uma representação fiel da gravação original. Com o BSS, torna-se possível extrair a faixa instrumental diretamente da música original, eliminando a necessidade de recriação manual.

Considerando todas as observações anteriores, surge a seguinte pergunta: É possível desenvolver um sistema computacional que, com o uso de técnicas de BSS para extração de faixas instrumentais, se compare aos métodos tradicionais de reprodução/recriação em termos de fidelidade sonora, eficiência do processo e experiência do usuário?

1.3 Importância e motivação da pesquisa

A pesquisa nessa área é relevante devido ao crescente interesse em soluções tecnológicas que otimizem a produção de conteúdo musical e melhorem a experiência do usuário, uma vez que ao descobrir o potencial e as limitações do BSS no contexto específico do karaokê pode-se fornecer dados valiosos para diversos ramos do entretenimento. Desse modo, observa-se no relatório da Technavio (2025), que o mercado global de karaokê está passando por um crescimento significativo, devendo aumentar em US$ 442 milhões de 2025 a 2029, a uma Taxa de Crescimento Anual Composta (CAGR, do inglês *Compound Annual Growth Rate*) de 4,3%. Essa expansão é impulsionada pela crescente popularidade de atividades de lazer relacionadas à música e sistemas avançados de mixagem de música, embora os desafios incluam maior adoção de instrumentos musicais virtuais e *software* de produção.

1.4 Objetivos

Este estudo tem como objetivo geral desenvolver um sistema computacional de karaokê que utilize técnicas de separação de fontes sonoras para gerar faixas instrumentais de alta qualidade.

Como objetivos específicos, procurou-se:

- Investigar e selecionar o algoritmo de separação de fontes sonoras adequado para a extração de faixas instrumentais de músicas, considerando critérios como qualidade da separação, artefatos sonoros introduzidos e eficiência computacional;
- Pesquisar a viabilidade e selecionar um algoritmo capaz de gerar a letra da canção com base no áudio extraído do vocal;
- Desenvolver uma interface intuitiva focada em proporcionar uma experiência agradável e acessível a cantores e entusiastas de todos os níveis;
- Avaliar a qualidade das faixas instrumentais geradas pelo sistema utilizado técnicas de BSS em comparação com faixas instrumentais criadas por métodos tradicionais (reprodução ou recriação manual);
- Investigar a percepção dos usuários em relação à intuitividade e utilidade da ferramenta, bem como o impacto gerado em sua experiência;
- Analisar o potencial do sistema desenvolvido como ferramenta de apoio ao

treinamento vocal amador, identificando seus pontos fortes, limitações e possíveis melhorias futuras.

1.5 Metodologia

Demo (1985) caracteriza a metodologia como uma preocupação instrumental, a qual se dedica aos métodos e processos para realizar a ciência, funcionando como um orientador ou um conjunto de instrumentos e caminhos. O objetivo principal da ciência é abordar a realidade de maneira teórica e prática, e a metodologia fornece os diferentes caminhos para atingir esse objetivo.

Conforme o trabalho de Prodanov e Freitas (2013), na etapa inicial de um estudo, a pesquisa exploratória é tida como fundamental e, portanto, é a abordagem escolhida para a fase de diagnóstico do problema. Isso implica em tornar mais simples a definição do tema, a determinação de objetivos e a elaboração de hipóteses, ou até mesmo a descoberta de novos enfoques. As atividades habituais envolvem a pesquisa de referências bibliográficas, entrevistas com especialistas no assunto e análise de casos exemplares para ampliar a compreensão.

Em termos de procedimento técnico, recorreu-se à pesquisa experimental, que, de acordo com Prodanov e Freitas (2013), é uma metodologia de investigação eficaz, voltada para a identificação de causas e efeitos. Nela, um objeto de estudo é estabelecido, e o pesquisador escolhe com cuidado as variáveis que podem afetá-lo. O cerne consiste em recriar as condições de um fenômeno em um ambiente controlado, o que possibilita uma observação detalhada de como uma variável influencia o objeto. Para esse fim, costuma-se usar locais adequados, equipamentos e instrumentos de precisão, a fim de mostrar de forma clara como e porquê um fato acontece.

Realizada a pesquisa experimental, buscou-se trabalhos correlatos por meio de uma seleção criteriosa, a qual concentrou-se em estudos que abordavam temas semelhantes ao do trabalho desenvolvido, visando identificar e analisar pesquisas pertinentes para o contexto em questão. De acordo com Neiva e Silva (2016), há algumas etapas a serem seguidas para efetuar uma revisão sistemática:

1. A definição das questões de pesquisa, tendo como principal questão o desenvolvimento de um sistema computacional de karaokê que utilize técnicas de separação de fontes sonoras.

2. A definição das palavras-chave, as quais incluem: *blind*, *music*, *audio*, *vocal*, *source separation*, *techniques*, *separation in music*, *deep learning*, *automatic speech recognition*, *speech-to-text*, *BSS* e *karaoke*.
3. A definição das *strings* de busca, que envolve a separação das palavras-chave, a concatenação de seus sinônimos usando o conector *OR*, a organização em grupos de sinônimos e a conexão com *AND*. A tabela 1 exibe as *strings* de busca.

Tabela 1 – *Strings* de busca

(("BSS" OR "BLIND SOURCE SEPARATION") AND ("MUSIC SOURCE SEPARATION" OR "AUDIO SOURCE SEPARATION") AND ("DEEP LEARNING SOURCE SEPARATION") AND ("TECHNIQUES FOR VOCAL SEPARATION IN MUSIC" OR "TECHNIQUES FOR VOCAL SEPARATION IN AUDIO") AND "KARAOKE")

Fonte: Autor (2025)

4. A definição das bases de busca, a qual varia de acordo com a área da revisão sistemática. Considerando as principais bases na área de estudo, a revisão inclui as seguintes bases: *Google Scholar*[1], Sociedade Brasileira de Computação[2], *Springer*[3] e *IEEE Xplore* [4].
5. O refinamento das *strings*, após os primeiros testes das *strings* de busca nas bases escolhidas, não houve a necessidade de realizar muitos refinamentos, entretanto, realizou-se ajustes como: *(("deep learning techniques for vocal separation in music") AND ("blind source separation deep learning techniques") AND ("music source separation deep learning techniques"))*, para assim encontrar mais trabalhos correlatos.
6. A execução das *strings* de busca, conforme foram feitas as buscas, para cada base, mudanças acerca dos termos foram necessárias, devido a não existir em específico o desenvolvimento de um sistema para karaokê.
7. O armazenamento dos resultados encontrados, todos salvos em formatos compatíveis com as ferramentas utilizadas para consulta posteriori.
8. A partir desta etapa em diante, as definições dos critérios de seleção de trabalhos

[1] https://scholar.google.com/
[2] https://www.sbc.org.br/
[3] https://link.springer.com/
[4] https://ieeexplore.ieee.org/Xplore/home.jsp

são consolidadas, culminando nos estudos detalhados no Capítulo 3.

Abrangida uma gama significativa de conhecimento nas fases iniciais, a seguir é descrito como a aplicação se desenvolverá:

- Fonte dos dados: as músicas a serem utilizadas tanto para aprimorar quanto para utilizar modelos prontos serão um conjunto do próprio autor, em que a maioria será de contexto local, ou seja, músicas do gênero gaúcho e nativistas do estado do Rio Grande do Sul. Os critérios da seleção das músicas que fazem parte do conjunto foram selecionados em relação à qualidade de áudio e complexidade de instrumentação, além de músicas famosas no meio tradicionalista;
- Modelos de separação: a metodologia adotada inclui uma etapa rigorosa de avaliação comparativa de modelos. Serão investigadas e selecionadas as arquiteturas de *Deep Learning* de Separação Cega de Fontes que demonstrem maior pertinência e promessa de desempenho conforme o estado da arte e os estudos de revisão bibliográfica realizados;
- Validações dos modelos de separação: medição quantitativa da capacidade de separação das fontes (voz e acompanhamento) utilizando métricas padrão da área de Processamento de Sinais, como a SDR, a SIR, a SAR e a avaliação da fidelidade perceptual do áudio resultante, crucial para a aplicação em karaokê. Esta avaliação será complementada por testes de audição ou métricas que correlacionam melhor com a percepção humana, resultando na escolha do modelo a ser implementado no sistema de karaokê;
- Validação dos modelos de reconhecimento automático de fala (ASR), focada em duas dimensões: precisão textual e exatidão temporal. A precisão textual é quantificada pela taxa de erro de palavras (WER, do inglês *Word Error Rate*), comparando a transcrição do modelo com o gabarito de referência. A exatidão temporal, crucial para a sincronização visual do karaokê, será avaliada pela percepção do alinhamento (subjetiva) e pela eficácia da heurística de estruturação da letra, resultando na escolha do modelo ASR ideal para a implementação no sistema, equilibrando precisão técnica e usabilidade.

1.6 Organização do texto

Além do capítulo da introdução, onde é apresentado o tema geral da pesquisa e a contextualização dentro do campo de estudo, os próximos capítulos seguem como tal:

- Referencial Teórico: Fornece uma base teórica que sustenta a pesquisa. Apresentando e discutindo os principais conceitos, teorias e autores pertinentes para o tema estudado;
- Trabalhos Correlatos: Demonstra e analisa pesquisas e estudos anteriores que possuem alguma relação direta com o tema pesquisado;
- Desenvolvimento: Detalha como a pesquisa foi conduzida, incluindo a metodologia utilizada, os materiais e métodos empregados, e os procedimentos de coleta e análise de dados;
- Considerações Finais e Trabalhos Futuros: Sintetiza os principais achados da pesquisa, destacando as contribuições do estudo e sugerindo direções para pesquisas futuras.

2 REFERENCIAL TEÓRICO

O presente referencial teórico visa fornecer a fundamentação conceitual e as bases metodológicas que sustentam a pesquisa sobre a separação de fontes sonoras, com foco específico em suas aplicações para aprimorar a experiência de karaokê e as técnicas de treinamento vocal. Nesse contexto, esta seção explorará os fundamentos do processamento de sinais de áudio, abrangendo a natureza do som, sua representação digital e as técnicas de análise no domínio do tempo e da frequência. Em seguida, serão apresentados os conceitos e os desafios inerentes à separação de fontes sonoras, desde as abordagens clássicas até as mais recentes inovações.

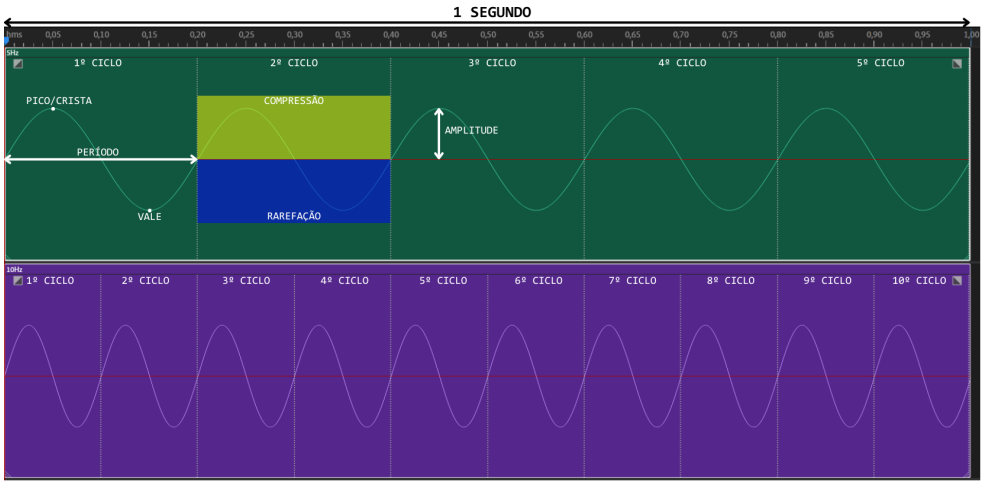
2.1 Som

O som é denominado como a vibração proveniente de uma fonte sonora, a qual pode compreender-se por: o diafragma de um transdutor acústico, uma corrente de ar que varia (como a voz), a explosão de algo inflamável, o choque entre dois corpos, elementos estruturais de instrumentos acústicos, entre outros, variando conforme a interpretação de cada indivíduo (GOMES; LABRADA, 2011). Tal vibração causa uma alternação de efeitos de compressão e rarefação que se espalham para além da origem como uma onda sonora. Quando a onda atinge o interior do ouvido humano, especificamente o tímpano, ela entra em sintonia ao vibrar com a superfície do tímpano, onde as terminações nervosas do corpo humano captam essas vibrações e as conectam ao cérebro humano, o qual as interpreta como sons (DODGE; JERSE, 1985). Logo, identifica-se que a natureza da vibração produzida por um corpo vibrante é única, fato que permite ao cérebro humano distinguir diferentes tipos de sons. Os gases (um exemplo comum é o ar), os sólidos e os líquidos podem ser elencados como meios transmissores do som, ou seja, todos transmitem o som, com maior ou menor eficiência. É importante notar que, na ausência de ar, como no espaço, a propagação dessas ondas sonoras é impossível e, portanto, o som não existe (MILETTO et al., 2004).

Segundo Miletto et al. (2004), o som é a vibração do ar, ou seja, variações na pressão do ar, que percebemos através dos nossos sentidos auditivos. Se tal pressão varia com um padrão repetitivo, dizemos que o som tem uma forma de onda periódica. Caso não haja um padrão, esse som é chamado de ruído. A forma como um som varia a pressão do ar ao longo do tempo é denominada onda sonora. Na figura 1, tal fenômeno é

demonstrado com suas peculiaridades e funções.

Figura 1 – Onda sonora, suas características e funções



$f = \frac{c}{seg}$

f = frequência
c = ciclos
seg = segundo

$C = 331 + 0,6 \times \Delta T$

C = velocidade do som em metros por segundo (m/seg)
331 = velocidade do som a zero graus Celsius
0,6 ≈ aumento de temperatura em um grau
ΔT = incremento ou diminuição da temperatura em graus Celsius

$L = \frac{C}{f}$

L = comprimento de onda em metros
C = velocidade de propagação da onda em m/seg
f = frequência da oscilação em Hertz

Fonte: Autor (2025), fórmulas de: Gomes e Labrada (2011)

O som em sua composição conta com os seguintes elementos:

- Altura: altura tonal tem sua base física na frequência fundamental da onda sonora. Por exemplo, em pianos, a mudança da altura tonal espelha-se diretamente na frequência das ondas sonoras geradas pelas suas teclas, com sons graves à esquerda e agudos à direita. A percepção subjetiva da altura tonal, que indica o quão alto ou baixo um som é, está relacionada ao número de ciclos de uma onda periódica por segundo, conhecido como frequência, assim, um ciclo dividido por um segundo resulta em um *Hertz* (Hz), que é a unidade internacional de medida da frequência. Sons agudos possuem uma frequência mais alta, enquanto os sons graves têm uma frequência mais baixa. A conexão entre frequência e altura tonal é logarítmica; o dobro de uma suposta frequência eleva seu som em uma oitava. Normalmente, o ouvido humano detecta frequências entre 20 Hz e 20.000 Hz, evidenciando a faixa espectral correspondente à percepção da altura tonal musical (MILETTO et al., 2004);
- Intensidade: a intensidade sonora, ou volume, em pianos é diretamente proporcional à força da atuação sobre a tecla. Uma percussão vigorosa resulta em maior amplitude da onda sonora, caracterizando um som forte. Inversamente, um toque suave gera ondas de menor amplitude, produzindo um som fraco. A

análise osciloscópica revela que a amplitude da onda, e não sua frequência, é o determinante físico do volume. Termos como "alto" e "baixo" referem-se precisamente à altura tonal (frequência), enquanto "forte" e "fraco" descrevem a intensidade (amplitude). Portanto, no domínio acústico, um som de "alta" amplitude corresponde a um som forte, e um som de "baixa" amplitude a um som fraco, distinguindo-se da terminologia coloquial.

Devido à lei logarítmica que rege a percepção auditiva de distintos níveis de intensidade sonora, o método de medição também é logarítmico. O decibel (dB) corresponde a um décimo de um Bel, o que indica uma relação entre dois valores e, portanto, não é absoluto (CAPEL, 1994);

- Timbre: a assinatura espectral de instrumentos musicais, ou seja, o timbre, é o causador da distinção sonora entre um piano e um violão, mesmo com altura tonal e volume idênticos, reside no timbre, uma propriedade acústica adicional. O timbre confere a cada instrumento sua identidade sonora única, resultante da complexa composição espectral da onda sonora. Analisando-se em um osciloscópio, sons com timbres distintos exibem formas de onda características. Ondas com contornos suaves tendem a produzir timbres aveludados, enquanto formas de onda mais angulares e ricas em harmônicos de alta frequência geram timbres mais brilhantes e penetrantes. O timbre, portanto, é determinado pela combinação e intensidade das diversas frequências constituintes do som fundamental, diferenciando a qualidade sonora dos instrumentos (MILETTO et al., 2004);
- Duração: de forma sucinta, é o tempo de emissão da onda sonora. Na notação musical, a duração é fundamental para o ritmo. A percepção da altura tonal requer uma duração mínima do estímulo sonoro, intrinsecamente ligada à sua frequência. O sistema auditivo necessita de um número suficiente de ciclos para identificar o tom. Frequências mais baixas, com períodos mais longos, demandam maior duração para a percepção tonal. Por exemplo, um tom de 100 Hz exige cerca de 40 ms, enquanto um de 1000 Hz pode ser percebido em aproximadamente 13 ms, evidenciando a relação inversa entre frequência e tempo mínimo de integração para a codificação da altura tonal (DODGE; JERSE, 1985).

2.2 Áudio digital

De acordo com Gomes e Labrada (2011), o áudio pode ser considerado como um conjunto de técnicas usadas no registro, reprodução e transmissão do som, ou como o som em si, que é convertido em outra forma de energia ou registrado através de algum princípio físico, químico ou elétrico.

Conforme Talbot-Smith (2001), os microfones são instrumentos cruciais em qualquer processo de gravação sonora e entender suas características é essencial. Existem duas características fundamentais em todo microfone. Uma é o sistema transdutor responsável pelo processo ao qual as ondas sonoras se transformam em sinais elétricos. Outra característica é a resposta polar, ou seja, a forma como o microfone reage a sons provenientes de diversas direções. Ao atingir um microfone, uma onda sonora gera uma voltagem que muda conforme o padrão da onda, logo, para que o computador possa processar o som, esse sinal analógico é transformado em uma sequência numérica, utilizando um conversor analógico-digital (ADC, do inglês *Analog-to-Digital Converter*). Em contrapartida, há o conversor digital-analógico (DAC, do inglês *Digital-to-Analog Converter*), para realizar o inverso, de converter números em tensões. Dispondo de tais informações, chegamos ao conceito de amostragem (em inglês *sampling*).

2.2.1 Amostragem

Segundo Miletto et al. (2004), amostragem é o processo de coletar um número específico de amostras (n) de uma forma de onda específica ($f(t)$) em um período de tempo específico (t). O procedimento de amostragem consiste em identificar n pontos de amplitude da onda e representá-los através de números que correspondam às respectivas amplitudes. No campo da engenharia, vale ressaltar que a técnica de converter um sinal analógico em uma sequência numérica é denominada modulação por código de pulso (PCM, do inglês *Pulse-Code Modulation*).

2.2.2 Taxa de amostragem

A taxa de amostragem é conhecida como intervalo de amostragem ou período de amostragem, compreendendo-se como o intervalo de tempo entre as amostras. O seu

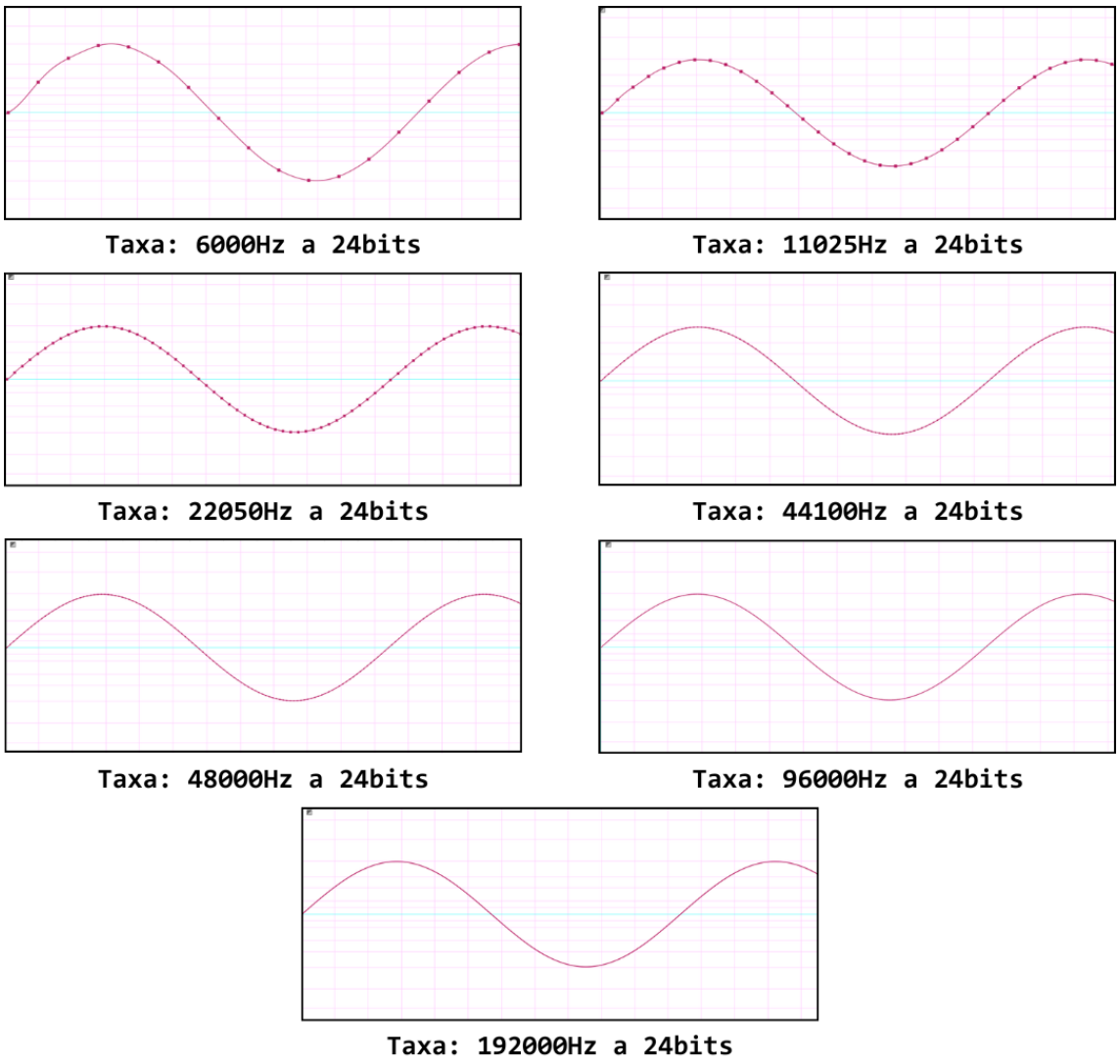
oposto, a quantidade de vezes que o sinal é amostrado em cada segmento, é conhecido como taxa de amostragem ou frequência de amostragem (em inglês *sampling rate*), sendo expressa em hertz (amostras por segundo). Dodge e Jerse (1985) explicam que o teorema de Nyquist-Shannon estabelece que, para representar digitalmente um sinal analógico sem perda de informação, a taxa de amostragem deve ser estritamente maior que o dobro da sua frequência máxima ($\frac{f}{2}$). A amostra que ultrapassa certos limites pode ocasionar o *aliasing*, uma distorção que cria frequências espúrias no sinal digital. Portanto, a precisão da representação de um fenômeno através da amostragem digital não depende de uma taxa infinita, mas sim de atender a critérios mínimos definidos pelo teorema de Nyquist.

Os aparelhos de som de alta fidelidade são concebidos para alcançar o limite máximo da audição humana, algo em torno de 20 kHz. Há diversas taxas de amostragem utilizadas para áudio, selecionadas por motivos técnicos e históricos. A frequência mais usual de amostragem para áudio digital é de 48 kHz, o que representa 41,67% da frequência amostrada (DODGE; JERSE, 1985). Observa-se a figura 2, onde, conforme a taxa de amostragem aumenta, são quase imperceptíveis os pontos na senoide com as maiores taxas.

2.2.3 Quantização

Um outro fator que pode impactar a fidelidade de um sinal digital depende de quão bem conseguimos medir o valor numérico de cada amostra. Esse estágio, chamado de quantização, consiste em mapear os valores infinitos do sinal original em um conjunto finito de números digitais. A maioria dos conversores de dados mede a resolução em bits, que é o tamanho binário do dado utilizado para representar cada amostra do sinal digital, portanto a resolução dos conversores de dados, que convertem sinais digitais em analógicos e vice-versa, é um fator crucial para o desempenho do sistema. De acordo com Dodge e Jerse (1985), quando há uma conversão, afirma-se que o sinal analógico está quantizado, já que sua representação digital só pode ser feita com uma resolução específica. Para uma representação mais detalhada das amostras, é preciso um número maior de bits por amostra, o que é denominado resolução do áudio digitalizado.

Figura 2 – Diferentes taxas de amostragem para a nota lá fundamental (440Hz)



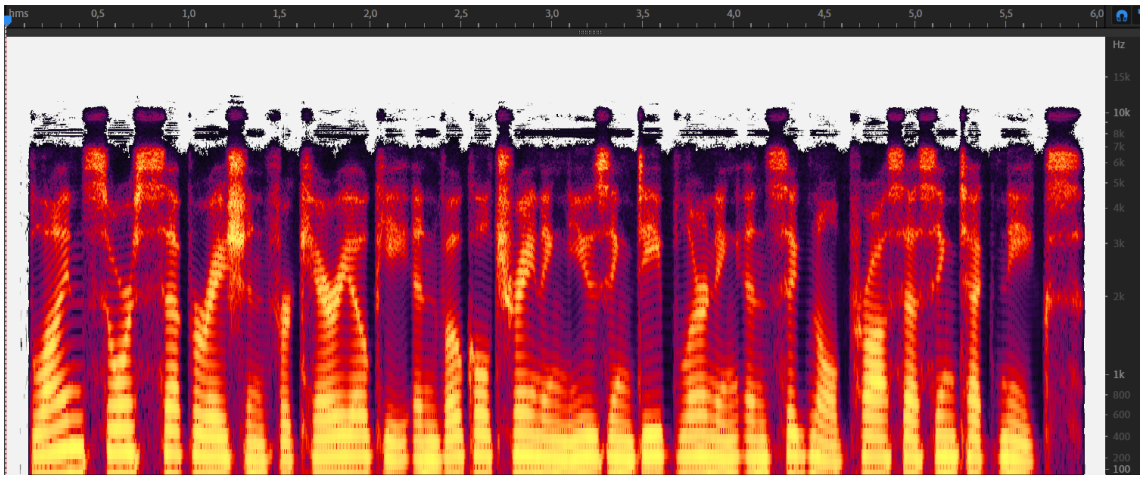
Fonte: Autor (2025)

2.2.4 Espectrograma

De acordo com Lampert e O'Keefe (2010) um espectrograma é uma ilustração visual que mostra como a energia acústica é distribuída entre as frequências e ao longo do tempo, permitindo, não apenas, observar se há mais ou menos energia em 10 Hz em comparação com 100 Hz, mas também identificar como os níveis de energia mudam ao longo do tempo. Verifica-se que a linha horizontal de um espectrograma geralmente indica o tempo, enquanto o eixo vertical mostra os passos discretos de frequência. Nesse viés, a quantidade de potência detectada é representada pela intensidade em cada ponto de tempo-frequência. A figura 3 representa um exemplo de espectrograma. No exemplo, as cores representam a amplitude (ou energia) de uma frequência específica em um

determinado momento, com tons de roxo indicando amplitudes baixas e cores mais brilhantes, até o amarelo radiante, correspondendo a amplitudes cada vez mais fortes. As cores podem variar de acordo com o *software* utilizado para visualização.

Figura 3 – Exemplo de um espectrograma



Fonte: Autor (2025)

2.2.5 Fatores de qualidade do áudio digital

Há fatores fundamentais que influenciam a qualidade do som digitalizado, como:

- Quanto maior for a taxa de amostragem, melhor será o detalhamento e o som ficará mais fidedigno;
- Quanto maior for a quantidade de bits, melhor será a resolução da amostra, isto é, mais números poderão ser representados;
- Qualidade do *hardware* diretamente envolvido no processo de captação, filtragem e amplificação do som, ou seja, a integridade física e eletrônica dos componentes que lidam com o som antes e durante a digitalização (MILETTO et al., 2004).

Após a introdução dos fatores, analogamente, pode-se realizar o seguinte cálculo: para um sinal de áudio que dura 1 segundo, utilizando uma frequência de amostragem igual a 44,1kHz (a mais empregada em gravação de CDs), quer dizer que o sinal terá 44100 amostras, empregando uma resolução de 16 bits (valor comum entre captações de som). Realiza-se a multiplicação entre os valores 44100 e 16, resultando que em 1 segundo há 705600 bits ou 88200 bytes ($\frac{705600}{8}$), ou seja, em 1 minuto, é ocupado em torno de 5MB de armazenamento.

2.2.6 Formatos de áudio digital

Diferentemente dos formatos analógicos profissionais e domésticos (semiprofissionais), os formatos digitais requerem a utilização de um decodificador-codificador para se integrarem, uma vez que cada formato tem uma linguagem distinta. Segundo Miletto et al. (2004), os formatos de arquivo mais comuns para a manipulação de áudio incluem:

- Wave, da Microsoft (extensão de arquivo .wav), é tipo áudio digitalizado sem compressão, baseado na codificação PCM, o que, dependendo da qualidade definida para gravação, acarreta na geração de um arquivo de áudio grande, a ponto de ocupar uma quantidade considerável de armazenamento;
- AIFF, da Apple (extensão de arquivo .aiff ou .aif), o AIFF possui as mesmas considerações do formato WAVE;
- MPEG Layer 3 (extensão .mp3), devido a questões de armazenamento, surgiram algoritmos para comprimir o áudio, sendo alguns desses algoritmos muito eficientes, como o caso do MPEG Layer 3, que alcança taxas de compressão de até 12:1 (1/12), mantendo uma qualidade admissível para o áudio, porém a perdas;
- Windows Media Audio, da Microsoft (extensão .wma), é uma opção de áudio digitalizado comprimido, para também abarcar problemas de espaço de armazenamento, entretanto, há remoção de certas frequências e harmônicos nesse formato, porém é mantida uma certa fidelidade ao original;
- Free Lossless Audio Codec (extensão .flac), um fato sobre o FLAC é que, ao contrário do MP3 (MPEG Layer 3), o FLAC comprime o tamanho do arquivo (geralmente para 50% a 70% do original), entretanto a descompressão resulta em uma cópia bit a bit idêntica ao áudio original.

Essa diversidade de formatos trouxe consigo uma vasta gama de *softwares* para tratá-los, desde editores de áudio profissionais como o Adobe Audition[5] e o Audacity®[6], que oferecem ferramentas avançadas de manipulação sonora, como equalização, mixagem e masterização, e para tarefas mais simples, como conversão de formatos ou cortes rápidos, há opções *on-line* e *softwares* mais intuitivos, como o Wavosaur[7]. A escolha do *software* ideal fica a cargo das necessidades do usuário e do nível de edição desejado.

[5] https://www.adobe.com/br/products/audition.html
[6] https://www.audacityteam.org/
[7] https://www.wavosaur.com/

2.3 Separação cega de fontes (BSS)

O trabalho de Fu (2025) ressalta que o processamento digital de sinais é um dos componentes essenciais da era digital e da informação, que surgiu com o rápido avanço das tecnologias de computadores, redes e comunicação. Na separação cega de fontes, o foco principal do processamento é a tecnologia que recupera ou separa o sinal da fonte do sinal misto observado em situações em que as informações prévias do canal de transmissão e do sinal da fonte são desconhecidas, tratando-se de um desafio clássico e complexo no processamento de sinais. Ainda, visualiza-se que processamento de sinais de fala, processamento de imagens, detecção de problemas mecânicos, processamento de sinais biológicos, processamento de sinais de comunicação e outras áreas fazem uso extensivo de tecnologias de separação cega de fontes.

O exemplo descrito por Xu (2019), em que, em um coquetel animado, múltiplas fontes sonoras criam uma cacofonia captada por vários microfones como sinais mistos. No entanto, apesar das dificuldades apresentadas, o cérebro humano consegue se concentrar facilmente em uma única voz em meio ao ruído. Em meio a um ambiente ruidoso, a Separação Cega de Fontes (BSS) resolve esse desafio, embora, sem conhecimento prévio das fontes sonoras ou da posição dos microfones, os algoritmos de BSS visem destrinchar esses sinais mistos, isolando a fala desejada. Essa área oferece técnicas poderosas para extrair componentes sonoros individuais de gravações complexas do mundo real, refletindo nossa notável capacidade auditiva em ambientes ruidosos. Tendo como base o número de microfones (sinais observados) em comparação com as fontes sonoras, a BSS é categorizada como:

- Sobredeterminado: quando o número de sinais observados é maior que o número de sinais de origem;
- Determinado: quando o número de sinais observados é igual ao número de sinais de fonte;
- Subdeterminado: quando o número de sinais observados é menor que o número de sinais de origem.

A teoria básica utilizada para o problema de sobredeterminado e determinado permanece a mesma, sendo solucionável através da Análise de Componentes Independentes (ICA, do inglês *Independent Component Analysis*), levando em conta a independência estatística entre os sinais de origem. Atendendo à natureza dispersa do sinal de entrada, a Análise de

Componentes Esparsos (SCA, do inglês *Sparse Component Analysis*) é empregada para solucionar a questão do subdeterminado.

2.3.1 Desenvolvimento da separação cega de fontes

A técnica de separação cega de fontes é uma técnica eficaz de processamento de sinais, sugerida no final dos anos 1980, mais especificamente em 1986, quando Hérault e Jutten apresentaram o algoritmo H-J (um modelo de rede neural de *feedback* com um algoritmo de aprendizado baseado em Hebb) para distinguir dois sinais provenientes de fontes independentes e mistas durante uma Conferência sobre Redes Neurais para Computação realizada nos Estados Unidos. O algoritmo H-J é capaz de distinguir dois sinais provenientes de fontes estatisticamente distintas e mistas através do método de pesquisa. Este estudo inaugurou um novo capítulo no campo do processamento de sinais e, desde então, o desafio BSS tem despertado grande interesse entre os pesquisadores. Como resultado de redes neurais artificiais, análise estatística de sinais e teoria da informação, a BSS emergiu como um tema relevante em pesquisa e inovação em diversos campos, particularmente nas ciências biomédicas, comunicação de sinais de fala, processamento de imagens, ciências do solo, econometria e análise de dados textuais (YU; HU; XU, 2014).

2.3.2 Análise de componentes independentes (ICA)

A análise de componentes independentes (ICA) destaca-se como a técnica de BSS mais difundida, fundamentando-se na suposição de que os sinais de origem são estatisticamente independentes, implicando que a informação proveniente de uma fonte não oferece subsídios acerca da outra. Ademais, a ICA presume que as fontes são não-gaussianas, uma vez que a independência estatística de variáveis gaussianas só resulta em sua incorrelação (e não na completa separação). A ICA procura realizar uma conversão linear dos sinais observados que aumente a não-gaussianidade dos componentes resultantes, o que normalmente se traduz em sua independência estatística. Muito empregada em processamento de áudio, como na separação de vozes, como no então exemplo citado "festa de coquetel", processamento de imagens e na neurociência para examinar dados de EEG/MEG, pode apresentar ambiguidades de permutação (a

disposição das fontes individuais pode ser aleatória) e escala (a extensão das fontes pode ser aleatória). Além disso, pode não funcionar adequadamente se as fontes forem gaussianas ou se a mistura for não-linear ou subdeterminada (menos sensores do que fontes) (YU; HU; XU, 2014).

2.3.3 Fatoração de matrizes não negativas (NMF)

Diferentemente da ICA, a NMF se fundamenta na ideia de que todos os dados e componentes de origem são não-negativos. Isso é válido em diversas situações reais, onde os sinais indicam quantidades físicas que não podem ser negativas, como a intensidade da luz, a frequência das palavras, a atividade muscular. A NMF divide uma matriz de dados não-negativa V em duas matrizes não-negativas, W (também conhecida como matriz de base ou componentes) e H (também conhecida como matriz de ativação ou pesos), de forma que $V = WH$. Os "componentes" subjacentes são representados nas colunas W, enquanto as colunas H mostram a quantidade desses componentes em cada observação. Utilizada em processamento de áudio (por exemplo, distinção entre música e voz, avaliação do timbre), processamento de texto (identificação de tópicos), visão computacional (identificação de rostos, extração de traços) e quimiometria. É preocupante a não uniformidade da solução, já que diversos pares de W e H podem resultar na mesma aproximação. A seleção do número de componentes também representa um obstáculo (YU; HU; XU, 2014).

2.3.4 Análise de componentes esparsos (SCA)

A SCA é especialmente eficaz no problema de BSS subdeterminado (em que a quantidade de fontes supera a quantidade de sensores), a suposição principal é que os sinais de origem são dispersos, o que implica que, em algum domínio (tempo ou frequência), a maior parte de seus valores é zero ou muito próximo disso, a SCA investiga essas fontes dispersas. No domínio da frequência, através da Transformada de Fourier de Curto Prazo (STFT, do inglês *Short-Time Fourier Transform*), os pontos de dados que representam um único pico de origem se alinham em torno das direções da matriz de mistura. Ao reunir esses pontos, podemos calcular a matriz de mistura e, posteriormente, as fontes. Eficaz em situações onde as fontes são naturalmente dispersas, como no áudio

(onde diversos instrumentos ou vozes podem apresentar picos de energia em frequências diferentes em momentos diferentes) ou no processamento de imagens para eliminação de ruído. A efetividade da SCA é fundamentalmente determinada pela dispersão das fontes. Se as fontes forem muito dispersas, a separação pode ser complicada. A exatidão na estimativa da matriz de mistura também é um elemento crucial (YU; HU; XU, 2014).

2.3.5 Medição e validação da qualidade dos sinais separados

De acordo com Lepcha et al. (2023) a pontuação média de opinião (MOS, do inglês *Mean Opinion Score*) é comumente empregada na técnica de avaliação qualitativa, que envolve avaliadores humanos solicitados a atribuir uma pontuação de qualidade perceptual. As classificações costumam variar de 1 (ruim) a 5 (excelente), assim, a média dessas notas é usada para calcular a MOS final. O teste MOS possui algumas deficiências, entre elas: variações nos critérios de avaliação. Porém, o teste MOS é uma técnica confiável para avaliar a qualidade perceptual.

O trabalho de Vincent, Gribonval e Fevotte (2006) descreve a Relação Sinal-Distorção (SDR, do inglês *Signal-to-Distortion Ratio*) como uma métrica utilizada para quantificar a qualidade de um sinal em comparação com sua versão distorcida ou ruidosa. A fórmula base para calcular o SDR é:

$$SDR = 10 \cdot log_{10}(\frac{E_s}{E_d})$$

Onde:

- E_s: é a energia do sinal original (ou de referência);
- E_d: é a energia da distorção ou ruído presente no sinal degradado. Geralmente calculada como a diferença entre o sinal original e o sinal estimado (com distorção).

Em aplicações como a desta pesquisa, e separação de fontes de áudio em geral, onde se tenta isolar uma fonte específica (voz, instrumento, etc.) de uma mistura, a distorção no SDR é decomposta em três componentes principais:

1. Interferência (e_{interf}): Componentes de outras fontes que vazaram para o sinal estimado;
2. Ruído (e_{noise}): Ruído ambiente ou aleatório;
3. Artefatos (e_{artif}): Novos componentes gerados pelo algoritmo de processamento que não estavam presentes no sinal original (por exemplo, eco ou reverberação

indesejado).

A fórmula fica da seguinte maneira:

$$SDR = 10 \cdot log_{10}(\frac{E_s}{e_{interf}+e_{noise}+e_{artif}})$$

O SDR é expresso em decibéis (dB), portanto, um SDR mais alto indica uma melhor qualidade do sinal, ou seja, menos distorção em relação ao sinal original.

2.4 Aprendizado profundo (*deep learning*)

Segundo Shinde e Shah (2018), a inteligência artificial (IA) diz respeito à capacidade de tornar as máquinas tão instruídas quanto o cérebro humano. Na área de Ciência da Computação, IA refere-se ao estudo de agentes inteligentes: qualquer aparelho que consiga perceber seu ambiente e executar ações que aumentem suas chances de alcançar seus objetivos de maneira bem-sucedida. A capacidade de aprender é um elemento essencial das máquinas, assim, o aprendizado de máquina acaba sendo um ramo da Inteligência Artificial. Desde a década de 1950, os cientistas da computação têm-se empenhado no campo do aprendizado de máquina. Nas últimas décadas, grandes investimentos foram feitos no progresso do aprendizado de máquina. Isso resulta em expectativas mais elevadas em relação às máquinas. O aprendizado profundo representa um esforço nesse sentido e cresce devido às Unidades de Processamento Gráfico (GPU, do inglês *Graphic Processing Unit*) potentes, *hardware* acessível e avanços em pesquisa de aprendizado de máquina e processamento de sinais. O termo aprendizagem profunda se refere a redes neurais artificiais de grande profundidade, e profundo é a expressão usada para designar as diversas camadas em uma rede neural. A rede profunda tem várias camadas ocultas, ao passo que uma rede superficial tem apenas uma. O aprendizado profundo é uma parte específica do aprendizado de máquinas, o qual trata-se de uma rede neural composta por diversas camadas e parâmetros. A grande parte das técnicas de aprendizado profundo emprega arquiteturas de rede neural. Em suma, o aprendizado profundo emprega uma sequência de várias camadas de unidades de processamento não-linear para identificar e modificar características. As camadas mais baixas, próximas à entrada de dados, assimilam características básicas, enquanto as camadas mais altas assimilam características mais complexas, oriundas das características das camadas mais baixas. A arquitetura cria uma representação poderosa e hierarquizada de atributos. Significa que o aprendizado profundo é apropriado para examinar e obter conhecimento

valioso tanto de grandes volumes de dados quanto de informações obtidas de diversas fontes.

As metodologias comuns de aprendizado profundo incluem Autocodificação (AE), Rede de Crenças Profundas (DBN), Rede Neural Convolucional (CNN), Rede Neural Recorrente (RNN), Rede Neural Recursiva e Aprendizado de Reforço Profundo Direto (PRD). Dentre as citadas, destacamos a CNN e a RNN por abordarem, de forma complementar, as duas principais características dos dados de áudio: o domínio da frequência e o domínio do tempo.

2.4.1 Rede neural convolucional (CNN)

De acordo com LeCun (1989) são um tipo específico de rede neural para o processamento de dados, com uma topologia conhecida, parecida com uma grade. Os dados de séries temporais, que podem ser vistos como uma grade unidimensional que coleta amostras em intervalos de tempo regulares, e os dados de imagem, que podem ser vistos como uma grade bidimensional de *pixels*[8], são exemplos disso. As redes convolucionais demonstraram grande sucesso em aplicações práticas. A denominação "rede neural convolucional" sugere que a rede utiliza um procedimento matemático conhecido como convolução. Diferente das redes neurais densas, as CNNs utilizam camadas convolucionais para extrair características locais de forma hierárquica. Em aplicações de BSS, essa arquitetura é fundamental para analisar a correlação entre amostras vizinhas no tempo e na frequência. Ao aplicar filtros que deslizam sobre o sinal, a rede consegue mapear dependências estruturais e extrair máscaras que permitem separar as fontes originais a partir de uma mistura complexa, preservando a integridade das características de cada sinal.

De um modo geral, as CNNs são concebidas para tratar dados que se apresentam em múltiplas matrizes, como um espectrograma, que é uma representação visual do som organizada em uma matriz de duas dimensões. Nessa estrutura, os eixos representam o tempo e a frequência, enquanto a intensidade de cada *pixel* indica a magnitude da energia sonora naquele instante, permitindo que a CNN identifique padrões específicos de cada fonte para realizar a separação. Numerosas formas de informação são representadas

[8]Do inglês *picture element* (elemento de imagem), é a menor unidade de uma imagem digital.

por matrizes múltiplas: 1D[9] para sinais e sequências, incluindo linguagem; 2D[10] para imagens ou espectrogramas de áudio; e 3D[11] para vídeos com estereoscopia ou imagens volumétricas. As CNNs baseiam-se em quatro conceitos fundamentais que tiram proveito das características dos sinais naturais: conexões locais, pesos compartilhados, agrupamento e a utilização de múltiplas camadas (LECUN; BENGIO; HINTON, 2015).

2.4.2 Rede neural recorrente (RNN)

Rumelhart, Hinton e Williams (1986) consideram a RNN como um grupo de redes neurais destinadas ao processamento de dados sequenciais. Da mesma forma que uma rede convolucional é um tipo de rede neural dedicada ao processamento de uma sequência de valores x, semelhante a uma imagem, uma rede neural recorrente é uma rede neural focada no processamento de uma sequência de valores $x_1, ..., x_n$. Da mesma forma que as redes convolucionais escalam eficientemente para processar sinais com alta resolução temporal ou espectrogramas extensos, as redes recorrentes são projetadas para lidar com sequências de áudio muito mais longas do que seria viável para redes neurais sem especialização sequencial. Enquanto as CNNs focam na extração de padrões locais, as redes recorrentes possuem a capacidade intrínseca de processar fluxos sonoros de duração variável, mantendo a memória de eventos passados para compreender a continuidade do sinal.

A retropropagação adquire um papel relevante tanto no âmbito das RNNs quanto das CNNs. De maneira simplificada, ela possibilita que a rede ajuste seus pesos internos para minimizar a diferença entre suas previsões e os resultados esperados. Suponha que a rede tenha feito uma previsão, mas tenha cometido um erro; a retropropagação determina como cada conexão da rede contribuiu para esse erro e, em seguida, ajusta essas conexões na direção correta para que, na próxima vez, a rede cometa menos erros. As RNNs são um tipo de rede neural especialmente eficiente para tarefas que requerem sequências de dados, como reconhecimento da fala humana ou realizar traduções. Isso ocorre porque, ao contrário de outras redes, elas têm a capacidade de lembrar eventos anteriores na sequência. As RNNs processam sequências elemento por elemento, preservando um vetor de estados em suas unidades ocultas, que guarda o histórico da sequência. Durante o

[9] Unidimensional; Uma dimensão; Uma linha.
[10] Bidimensional; Duas dimensões; Um plano.
[11] Tridimensional; Três dimensões; Um volume.

treinamento, as saídas ocultas em diferentes momentos são consideradas camadas de uma rede neural profunda, o que possibilita ao algoritmo de retropropagação aprender padrões sequenciais mais complexos (LECUN; BENGIO; HINTON, 2015).

2.5 Reconhecimento automático de fala (ASR)

A transcrição da fala em texto pode ser conhecida como reconhecimento automático de fala, sendo um campo estudado na área de reconhecimento de fala. Um exemplo que podemos citar é um *software* específico que recebe uma onda sonora e produz uma sequência de caracteres ou códigos de identificação de palavras que representam os termos falados na gravação. Em concordância com Zhang et al. (2021), no âmbito do reconhecimento automático de fala (ASR, do inglês *Automatic Speech Recognition*), o desafio é que há uma quantidade muito maior de quadros de áudio (o som é geralmente amostrado a 8kHz ou 16kHz) do que de texto. Isso significa que não há uma correspondência direta entre áudio e texto, uma vez que milhares de amostras podem equivaler a uma única palavra falada. Esses são problemas de aprendizado de sequência a sequência em que a saída é significativamente mais curta do que a entrada. Sendo assim, o aprendizado profundo é um elemento fundamental dos sistemas atuais de reconhecimento de fala utilizados por grandes corporações, como Microsoft, IBM e Google (HINTON et al., 2012).

2.5.1 Taxa de erro de palavras (WER)

De acordo com Park, Chen e Hain (2024), a taxa de erro de palavras (WER, do inglês *Word Error Rate*) é uma métrica utilizada para medir a qualidade das transcrições geradas por sistemas de Reconhecimento Automático de Fala (ASR) e para calcular essa taxa (WER), é necessário ter dados que foram transcritos manualmente para avaliar o desempenho dos sistemas de reconhecimento automático de fala (ASR).

A fórmula padrão do cálculo é:

$$WER = \frac{S+D+I}{N}$$

O sistema compara a transcrição palavra por palavra e identifica três tipos de erros:

- Substituição (S): quando uma palavra é trocada por outra (ex.: o modelo ouviu

"casa"em vez de "caça");

- Deleção (*D*): quando o modelo pula (passa por) uma palavra que existia no áudio original;
- Inserção (*I*): quando o modelo adiciona uma palavra que nunca foi dita no áudio;
- *N*: é o número total de palavras no texto de referência original.

O resultado geralmente é expresso em valor decimal e, quanto mais próximo de 0, mais perfeita é a transcrição. Um valor 0 significa que a transcrição é idêntica à referência.

2.5.2 Transformer

Vaswani et al. (2017), descrevem que Transformer são arquiteturas de rede neural que utilizam mecanismos de auto-atenção para processar informações de forma global. A maioria dos modelos que lidavam com sequências (como frases) usavam Redes Neurais Recorrentes e suas variações, que processavam dados sequencialmente (palavra por palavra). Isso tornava o treinamento lento e dificultava a captura de dependências de longo alcance. O Transformer resolveu isso com seu principal componente: o Mecanismo de Auto-Atenção (*Self-Attention*). De um modo geral, em vez de ler palavra por palavra (sequencialmente), o Transformer lê a sequência inteira de uma vez, prestando atenção em como cada palavra se relaciona com todas as outras palavras na frase.

Nos sistemas ASR modernos, a arquitetura Transformer é adaptada para mapear diretamente o áudio de entrada para a sequência de texto (caracteres ou palavras) de saída, eliminando a necessidade de múltiplos componentes separados. Por exemplo, tenha-se na entrada um sinal de áudio convertido em uma sequência de características acústicas (como espectrogramas). A arquitetura padrão *Encoder-Decoder* do Transformer é utilizada. O *Encoder* processa as características acústicas, usando o mecanismo de auto-atenção para entender as relações de longo alcance no áudio (por exemplo, como o som de uma vogal no início da fala se relaciona com o final da frase). O *Decoder* gera a sequência de texto (transcrição) com base na representação do *Encoder*, prestando atenção nas partes relevantes do áudio a cada caractere ou palavra que é produzida.

2.6 Modelos computacionais

O trabalho de Shi (2021) descreve que modelos computacionais traduzem problemas em linguagem matemática para simular e validar características da inteligência. Eles contribuem para a compreensão da estrutura funcional de fenômenos cognitivos específicos. Existem duas abordagens básicas para a modelagem cognitiva. A primeira se concentra nas funções mentais abstratas de uma mente perspicaz e emprega símbolos; a segunda segue as propriedades neurais e associativas do cérebro humano, recebendo o nome de subsimbólica, abrangendo modelos conexionistas e de redes neurais.

2.6.1 Exemplos de modelos para separação de vocais

Estes modelos geralmente utilizam redes neurais profundas para decompor uma mixagem em componentes instrumentais e vocais. Exemplos como:

- *Demucs*[12]: É um dos modelos de estado da arte mais conhecidos, utilizando arquiteturas de rede neural profunda. É capaz de separar vocais, bateria, baixo e outros instrumentos. Muitos serviços comerciais utilizam versões otimizadas deste modelo;
- *U-Net*[13]: Embora não sejam algoritmos independentes, muitas implementações de separação de fontes usam a arquitetura *U-Net* (originalmente para segmentação de imagens) adaptada para o domínio de tempo-frequência de áudio (espectrogramas);
- *Spleeter*[14]: Outro modelo popular baseado em aprendizado profundo que oferece separação rápida e de alta qualidade em diferentes configurações (vocal/instrumental; vocal/bateria/baixo/piano/outros;).

2.6.2 Exemplos de modelos para transcrição de voz

A conversão do áudio da fala em texto emprega modelos acústicos e modelos de linguagem para mapear fonemas a palavras. Exemplificando:

[12]https://arxiv.org/pdf/2211.08553
[13]https://arxiv.org/pdf/1806.03185
[14]https://joss.theoj.org/papers/10.21105/joss.02154

- *DeepSpeech*[15]: Um dos primeiros modelos open-source de ASR baseados em aprendizado profundo. Utiliza redes neurais recorrentes. Apesar de talvez não ser o mais preciso no momento, representa um marco significativo e foi empregado para experimentação com dados em português;
- *Whisper*[16]: Um modelo de reconhecimento de fala de aplicação ampla. Ele foi treinado com um vasto conjunto de dados de áudio variados e também atua como um modelo multitarefa, capaz de realizar reconhecimento de fala em vários idiomas, tradução de fala e identificação de idioma.
- *Wav2vec*[17]: Um modelo ousado que utiliza pré-treinamento auto-supervisionado em grandes volumes de áudio não rotulado, seguido de ajuste fino em dados rotulados. Isso o torna extremamente eficiente e adaptável a novos idiomas e domínios com menos dados rotulados. Existem versões pré-treinadas em múltiplos idiomas (como o XLSR-53) que incluem português;
- *Conformer*[18] e *Transformer*[19]: Arquiteturas de rede neural que combinam os pontos fortes das CNNs e *Transformers* para processar sequências de áudio, resultando em alta precisão e eficiência. Muitas das APIs de ASR comerciais como *Google Cloud Speech-to-Text, Amazon Transcribe*, utilizam variantes desses modelos.

[15]https://arxiv.org/pdf/1412.5567
[16]https://arxiv.org/pdf/2212.04356
[17]https://arxiv.org/pdf/1904.01038
[18]https://arxiv.org/pdf/2005.08100
[19]https://aclanthology.org/2020.emnlp-demos.6.pdf

3 TRABALHOS CORRELATOS

Este capítulo apresenta os estudos selecionados que formam parte da base desta pesquisa. A escolha desses trabalhos foi guiada por um rigoroso conjunto de critérios, cujas definições foram consolidadas a partir da etapa 8 do processo metodológico. Será descrito aqui cada um dos estudos, explicando sua relevância e como eles contribuem para o desenvolvimento e a fundamentação do presente trabalho. A análise desses estudos é crucial para estabelecer o contexto da pesquisa e identificar lacunas ou oportunidades que serão abordadas.

3.1 *Open-Unmix - A Reference Implementation for Music Source Separation*

O trabalho de Stöter et al. (2019) oferece várias contribuições importantes para o domínio da separação de fontes musicais, sendo os pontos principais:

- Implementação de ponta: o *Open-Unmix* oferece uma implementação de referência de código aberto que produz resultados de ponta na separação de fontes musicais. Não existia uma implementação anterior, portanto, esse trabalho preenche uma lacuna e possibilita que os pesquisadores comparem seus métodos com um padrão de alto desempenho;
- Estimulação da pesquisa: A implementação foi concebida para agilizar a pesquisa acadêmica, fornecendo implementações compatíveis com os *frameworks* de aprendizado profundo mais utilizados como: *PyTorch*[20], *Keras*[21], *NNabla*[22] e *TensorFlow*[23]. Essa flexibilidade proporciona que pesquisadores reproduzam resultados com facilidade, fomentando um ambiente de pesquisa colaborativa;
- Modelos pré-treinados: O artigo apresenta modelos pré-treinados para a separação de instrumentos disponíveis para usuários finais e artistas. Oportunizando que pessoas sem um profundo conhecimento técnico experimentem a separação de fontes, expandindo, dessa forma, a base de usuários e as possíveis utilizações da tecnologia;
- Foco na usabilidade: O *Open-Unmix* visa equilibrar alto desempenho com

[20]https://pytorch.org/
[21]https://keras.io/
[22]https://nnabla.org/
[23]https://www.tensorflow.org/

facilidade de compreensão. Essa escolha de design garante que o sistema não seja apenas eficaz, mas também compreensível, tornando-se um recurso valioso para futuras pesquisas e desenvolvimento na área. Os autores reconhecem que diversos pesquisadores se depararam com dificuldades em atividades de pré e pós-processamento devido à ausência de conhecimento compartilhado sobre o domínio. O *Open-Unmix* foi desenvolvido para mitigar essas questões e ainda atua como um elemento fundamental de um ecossistema aberto para separação musical, englobando conjuntos de dados abertos, ferramentas de *software* e técnicas de avaliação. O objetivo deste ecossistema é incentivar pesquisas reprodutíveis e apoiar progressos futuros no campo.

As contribuições para esta pesquisa são muito significativas, pois, além de disponibilizar *datasets*[24] com faixas musicais para testes e treinos, o seu repositório é bem documentado, auxiliando ao máximo pesquisadores, engenheiros de áudio e artistas com seus modelos prontos para o uso.

3.2 *Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation*

Os autores Stoller, Ewert e Dixon (2018) sugerem uma nova metodologia que atua diretamente no domínio do tempo, ao invés de se basear no espectro de magnitude. Esse método possibilita a modelagem de informações de fase, o que é fundamental para alcançar resultados de separação de alta qualidade. Ao não adotar transformações espectrais fixas, o modelo consegue captar com mais precisão as sutilezas dos sinais de áudio. O *Wave-U-Net* utiliza uma estratégia multiescala para a extração de características, redefinindo os mapas de características. Com essa técnica, o modelo é capaz de calcular e combinar características em várias escalas de tempo, o que melhora sua habilidade de separar fontes de áudio de forma mais eficiente. O artigo introduz uma modificação da arquitetura *U-Net*[25] projetada especificamente para sinais de áudio unidimensionais no domínio do tempo. Essa arquitetura foi desenvolvida para gerenciar de maneira eficiente as correlações temporais de longo alcance, o que é fundamental para a separação de áudio de alta qualidade. A arquitetura possui uma camada de saída que assegura a aditividade

[24] https://sigsep.github.io/datasets/

[25] De acordo com Ronneberger, Fischer e Brox (2015), é uma arquitetura de rede neural convolucional (CNN) desenvolvida especialmente para a segmentação de imagens, com uma estrutura em formato de U.

das fontes, permitindo que as fontes de áudio separadas sejam combinadas com exatidão. Esse recurso é fundamental para preservar a qualidade dos sinais de áudio durante a separação.

A principal contribuição desta arquitetura é sua capacidade de lidar com longas sequências temporais, permitindo captar nuances da performance vocal que métodos tradicionais ignoram. Em mixagens de áudio complexas, essa robustez no processamento do tempo e do fraseado resulta em uma separação de fontes mais precisa e fidedigna.

3.3 *Music Source Separation in the Waveform Domain*

A pesquisa idealizada por Défossez et al. (2021) apresenta o *Demucs*, um modelo que opera diretamente sobre a forma de onda, empregando uma arquitetura U-Net integrada a uma LSTM[26] bidirecional. Essa estrutura foi desenvolvida especificamente para aprimorar a separação de fontes musicais, superando o estado da arte em qualidade de áudio e precisão. Os autores utilizam o conjunto de dados *MUSDB18*[27], referencial padrão na literatura da área, cuja relevância advém da oferta de faixas multicanal isoladas (*stems*). Isso permite uma avaliação objetiva do desempenho em uma ampla diversidade de gêneros e texturas sonoras; como exemplo, o *Demucs* alcança, neste *dataset*, uma média de 6,3 dB na relação sinal-distorção (SDR). Com a inclusão de dados de treinamento extras, o modelo atinge 6,8 dB, ultrapassando inclusive o oráculo da Máscara de Razão Ideal (IRM) para a fonte de baixos, o que constitui um avanço importante no campo. Além disso, o artigo aborda progressos na quantização, possibilitando a compactação do modelo para 120 MB sem comprometer a precisão, fator fundamental para a viabilidade de implementação em aplicações do mundo real.

As contribuições que pode-se destacar do artigo são que o *Demucs* supera modelos existentes, em termos de qualidade de áudio e naturalidade. Essa melhoria pode aprimorar a experiência auditiva para usuários de karaokê, proporcionando faixas vocais mais nítidas e melhor separação da música de fundo, o que talvez seja crucial para o treinamento vocal.

[26] Tipo de rede neural recorrente (RNN)
[27] https://zenodo.org/records/1117372

3.4 *Hybrid Spectrogram and Waveform Source Separation*

O artigo de Défossez (2022) propõe um modelo híbrido que atua nos domínios da forma de onda e do espectrograma. Isso possibilita que o modelo identifique o domínio mais apropriado para cada fonte, integrando de maneira eficaz os pontos fortes de ambas as estratégias para aprimorar o desempenho em tarefas de separação de fontes. Sendo assim, há uma ampliação da arquitetura *Demucs* original, fundamentada em uma estrutura U-Net, para incorporar duas ramificações paralelas: uma voltada para o processamento temporal (onda) e outra para o processamento espectral (espectrograma). Uma melhoria de 1,4 dB no SDR em todas as fontes do conjunto de dados *MUSDB18*, com uma relação sinal-distorção (SDR) de 7,32 dB.

Para esta pesquisa, a abordagem híbrida pode ser adaptada a vários gêneros musicais, permitindo uma plataforma de karaokê versátil que pode atender às diferentes preferências do usuário. Essa adaptabilidade pode aumentar o engajamento e a satisfação do usuário em aplicativos de treinamento vocal, pois os usuários podem praticar com uma ampla variedade de estilos musicais. Outro ponto positivo são os avanços do modelo híbrido que podem permitir recursos de processamento em tempo real, o que, para aplicativos de karaokê, pode ser essencial.

3.5 *Hybrid Transformers for Music Source Separation*

Os pesquisadores Rouard, Massa e Défossez (2022) apresentam o *HT Demucs* (*Hybrid Transformer Demucs*), que é um modelo híbrido que combina processamento temporal e espectral. Ele substitui as camadas mais internas da arquitetura *Hybrid Demucs* existente por um *Transformer Encoder* de vários domínios, permitindo a autoatenção em um domínio e a atenção cruzada entre domínios. Essa inovação visa alavancar informações contextuais de longo alcance, que são cruciais para uma separação eficaz de fontes na música. O *HT Demucs* demonstra desempenho aprimorado em relação ao seu antecessor, o *Hybrid Demucs*. Quando treinado com dados adicionais (800 músicas de treinamento extras), o *HT Demucs* supera o *Hybrid Demucs* em 0,45 dB na relação sinal-distorção (SDR), mostrando a eficácia da nova arquitetura em aprimorar as tarefas de separação de fontes musicais.

O modelo *Hybrid Transformer Demucs* apresenta métodos inovadores para separação de fontes de música, particularmente por meio do uso de codificadores de

transformadores de vários domínios. Isso pode aumentar a capacidade de isolar os vocais das faixas instrumentais, fornecendo um áudio mais claro para fins de treinamento e desempenho. Entretanto, para esta pesquisa, o mesmo contribui de forma semelhante ao seu antecessor.

3.6 *Music Source Separation with Band-split RNN*

Os estudos de Luo e Yu (2022) sugerem uma nova arquitetura denominada BSRNN (Rede Neural Recorrente de Banda Dividida, do inglês *Band-split Recurrent Neural Network*), desenvolvida especificamente para a separação de fontes musicais. Este modelo segmenta explicitamente o espectrograma de valor complexo da mistura de entrada em diversas sub-bandas com larguras de banda distintas, o que possibilita um processamento mais eficiente dos sinais musicais. A ideia é realizar um processamento intercalado em nível de banda e nível de sequência usando redes neurais recorrentes. Essa abordagem ajuda a capturar as dependências intra-banda e a ordem sequencial dos sinais de áudio, o que é crucial para uma separação efetiva da fonte. O design leva em consideração o conhecimento a priori sobre as características da fonte alvo. Isso permite a otimização dos hiperparâmetros do modelo com base no tipo específico de instrumento musical que está sendo separado, o que é uma nova abordagem no campo. Os resultados indicam que o modelo proposto não só se destaca em desempenho, mas também se beneficia do estágio de ajuste fino semissupervisionado.

Os resultados experimentais do BSRNN (Band-split RNN) demonstram um desempenho consistente em diversos gêneros musicais. Essa robustez é particularmente benéfica para sistemas de karaokê, que demandam eficácia em uma ampla variedade de estilos e arranjos sonoros.

3.7 *Spleeter: A fast and efficient music source separation tool with pre-trained models*

Spleeter, uma ferramenta projetada por Hennequin et al. (2020) para separação de fontes musicais, enfatiza a facilidade de uso, alto desempenho de separação e velocidade, tornando-se acessível para vários usuários, incluindo pesquisadores e produtores musicais. O *Spleeter* oferece modelos pré-treinados que possibilitam aos usuários separar fontes de música sem a necessidade de grandes volumes de dados

de treinamento ou recursos computacionais. Esse recurso diminui consideravelmente o obstáculo para usuários que talvez não possuam a experiência ou os meios para treinar seus próprios modelos. Os modelos desenvolvidos demonstram desempenho competitivo em relação a sistemas de última geração no conjunto de dados *MUSDB18*. Notavelmente, o artigo relata que as métricas de desempenho do *Spleeter*, como a relação sinal-distorção (SDR), são comparáveis às dos sistemas existentes, como o *Open-Unmix* e o *Demucs*, embora o *Spleeter* não tenha sido treinado com dados do *MUSDB18*.

A disponibilidade de modelos pré-treinados no *Spleeter* permite uma implementação ágil em sistemas de karaokê. No contexto desta pesquisa, o modelo foi considerado por possibilitar a integração direta à aplicação sem a necessidade de treinamentos extensos ou recursos computacionais robustos, o que facilitaria o desenvolvimento inicial do protótipo. Outro benefício analisado foi a sua natureza de código aberto, com o código-fonte e os modelos disponibilizados publicamente em repositórios.

3.8 *MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation*

A perspectiva de Takahashi, Goswami e Mitsufuji (2018) apresenta uma estrutura que integra redes de memória de longo prazo (LSTM) à arquitetura *MMDenseNet*. Essa integração possibilita a modelagem eficaz de dependências temporais de longo prazo em sinais de áudio, otimizando o desempenho em tarefas de separação de fontes. Resultados experimentais desse trabalho demonstram que a arquitetura *MMDenseLSTM* supera tanto a *MMDenseNet* original quanto modelos LSTM autônomos. O método proposto por eles apresenta uma melhoria média de 0,2 dB na relação sinal-distorção (SDR) em comparação à *MMDenseNet*, validando sua eficácia. Além disso, a arquitetura foi concebida para operar com um número reduzido de parâmetros, diminuindo o tempo de processamento em relação a métodos de combinação convencionais.

A eficácia do modelo *MMDenseLSTM*, com seus parâmetros e tempo de processamento diminuídos, faz com que seja apropriado para uso em tempo real, como em sistemas de karaokês. Todavia, conforme explorado em outros trabalhos, não é o mais adequado para implementar nesta pesquisa.

3.9 *D3Net: Densely connected multidilated DenseNet for music source separation*

Os escritos de Takahashi e Mitsufuji (2021) introduzem a arquitetura *D3Net*, que emprega blocos aninhados de dilatação densa. Esse design inovador possibilita a aplicação eficaz de diversos fatores de dilatação na mesma camada, melhorando a habilidade do modelo de processar sinais de áudio em várias resoluções ao mesmo tempo. Ao incorporar a convolução multidilatada na arquitetura *DenseNet*, o *D3Net* consegue reduzir eficazmente os problemas de aliasing que podem surgir quando as convoluções dilatadas são aplicadas diretamente. Esse progresso é importante para preservar a qualidade do processamento do sinal de áudio. Resultados experimentais no conjunto de dados *MUSDB18* indicam que o *D3Net* atinge um desempenho de ponta, apresentando uma relação sinal-distorção (SDR) média de 6,01 dB. Esse resultado demonstra a eficácia das técnicas sugeridas em aplicações práticas para a separação de fontes musicais.

A combinação de técnicas de aprendizado profundo, conforme demonstrado no *D3Net*, com métodos tradicionais de processamento de sinal pode levar a soluções mais robustas para separação cega de fontes. Essa integração pode melhorar a precisão e a confiabilidade do isolamento vocal em vários ambientes de áudio, tornando-o adequado para um sistema de karaokê, como o desta pesquisa.

3.10 Considerações acerca dos trabalhos correlatos

Ao analisar os trabalhos correlatos, é evidente que a maioria das publicações aborda o problema sob a ótica de arquiteturas como Redes Neurais Convolucionais (CNNs), Redes Neurais Recorrentes (RNNs), notadamente LSTMs e GRUs, e, mais recentemente, *Transformers*. A aplicação dessas arquiteturas varia desde a estimação de máscaras espectrais até a modelagem direta de formas de onda. Modelos como *U-Net* e suas variações, que se destacam pela sua eficácia na segmentação de imagens, foram adaptados com sucesso para o domínio do espectrograma, tornando-se uma escolha popular para a separação em tempo-frequência.

Apesar da ubiquidade do *Deep Learning*, os trabalhos se diferenciam em aspectos como o âmbito aplicado (espectrogramas, formas da onda, híbridos), as funções de perda empregadas e as estratégias de aumento de dados. Além disso, há um esforço contínuo em otimizar o equilíbrio entre a qualidade da separação e a complexidade computacional dos modelos, visando aplicações em tempo real e em dispositivos com recursos limitados.

A consistência na adoção do *Deep Learning* sugere um consenso sobre sua superioridade em comparação a métodos tradicionais, impulsionando a pesquisa para aprimoramentos arquitetônicos e metodológicos dentro deste paradigma. Na tabela 2 é apresentada uma análise, especificando o âmbito aplicado, a relação sinal-distorção geral e o conjunto de dados utilizado em seu desenvolvimento.

Tabela 2 – Análise e contraste dos trabalhos correlatos

Trabalho	**Abordagem**	**SDR Geral**	***Dataset***
Stöter et al. (2019)	espectrograma	5,3 dB	*MUSDB18*
Stoller, Ewert e Dixon (2018)	forma da onda	3,2 dB	*MUSDB18*
Défossez et al. (2021)	forma da onda	6,3 dB	*MUSDB18*
Défossez (2022)	híbrido	7,7 dB	*MUSDB18*
Rouard, Massa e Défossez (2022)	híbrido	9,0 dB	Próprio
Luo e Yu (2022)	espectrograma	8,2 dB	*MUSDB18*
Hennequin et al. (2020)	espectrograma	5,9 dB	Próprio
Takahashi, Goswami e Mitsufuji (2018)	espectrograma	6,0 dB	Próprio
Takahashi e Mitsufuji (2021)	espectrograma	6,7 dB	Próprio

Fonte: Adaptação de Rouard, Massa e Défossez (2022)

Em síntese, a análise dos trabalhos correlatos reforça que a abordagem via *Deep Learning* é o caminho mais promissor para a separação de voz e instrumental de uma música, devido a diversas arquiteturas e otimizações já consolidadas. Este trabalho, ao aproveitar as melhores práticas e modelos existentes, não só visa aprimorar a separação de voz em cenários específicos, mas também adicionar um valor significativo ao transcrever automaticamente as letras das músicas. Essa integração é o pilar fundamental do Bahokê: um sistema que unifica a decomposição do sinal de áudio (separando a voz dos instrumentos) à geração de conteúdo textual sincronizado, proporcionando uma experiência de usuário sem precedentes.

4 DESENVOLVIMENTO

Este capítulo descreve o processo de criação do Bahokê, um *software* de karaokê, discutindo a estrutura do sistema e as fases de implementação que o transformam em uma ferramenta sólida e interativa. Desenvolver um aplicativo de karaokê vai além de apenas reproduzir áudio e letras; requer a integração eficaz dos módulos de processamento de áudio, interface do usuário e gerenciamento de arquivos.

A discussão inicia com a definição dos requisitos do programa, englobando funcionalidades essenciais como reprodução de faixas musicais, exibição sincronizada de letras, controle de volume, e demais características pertinentes ao funcionamento de um karaokê. Em seguida, é apresentada a arquitetura geral do sistema, delineando os principais componentes e suas interações, como o módulo de separação de fontes de áudio (para isolar vocais e instrumentais), o gerenciador de letras e a interface gráfica do usuário (GUI, do inglês *Graphical User Interface*).

O capítulo continua com a investigação das tecnologias e ferramentas utilizadas na implementação, abordando os detalhes das bibliotecas de processamento de áudio usadas na manipulação de sinais, dos *frameworks* de desenvolvimento para a interface gráfica do usuário e das alternativas de linguagem de programação. A implementação de recursos essenciais, como algoritmos para sincronização de letras com a música. Por último, são abordadas as técnicas de teste e otimização utilizadas para assegurar a estabilidade, performance e facilidade de uso do *software*, resultando em um aplicativo de karaokê completo e eficiente.

4.1 Estrutura e planejamento do projeto

Para criar um *software* de karaokê eficaz e agradável para o usuário, é fundamental estabelecer um conjunto claro de requisitos. Esses são classificados em categorias funcionais e não funcionais, assegurando que o aplicativo não só realize as funções desejadas, mas também proporcione uma experiência de uso satisfatória (SOMMERVILLE, 2011). Para isso, efetuou-se uma entrevista com o professor Daniel Brum da Silva, que iniciou seus estudos em música ainda na infância, nas cidades fronteiriças de Jaguarão/RS (Brasil) e Rio Branco (Uruguai), posteriormente ganhando experiência ao tocar em bandas de baile e acompanhar cantores. Sua carreira musical evoluiu para o serviço militar, onde atuou como músico e mestre da banda do 12º

Regimento de Cavalaria Mecanizada do Exército Brasileiro. Após cursar o bacharelado em Música na UFPel (não concluído), dedicou-se ao ensino como instrutor de música no Instituto Artístico Carlos Gomes, em Dom Pedrito/RS, até 2015. Atualmente, é tecnólogo em Gestão Pública e exerce o cargo de Chefe da Secretaria Administrativa da UNIPAMPA Campus Dom Pedrito, conciliando suas funções com a formação em Licenciatura em Música (UniCesumar), Especialização em Música (FAVENI) e pesquisas na área de luteria (a arte e o ofício da construção, restauração e conserto de instrumentos musicais de corda, como violões, guitarras e outros). Com base em sua experiência e conhecimento musical, solicitou-se que ele desempenhasse o papel de *stakeholder* para analisar tais requisitos que deveriam compor o Bahokê.

4.1.1 Requisitos funcionais

Estes requisitos descrevem o que o *software* deve fazer:

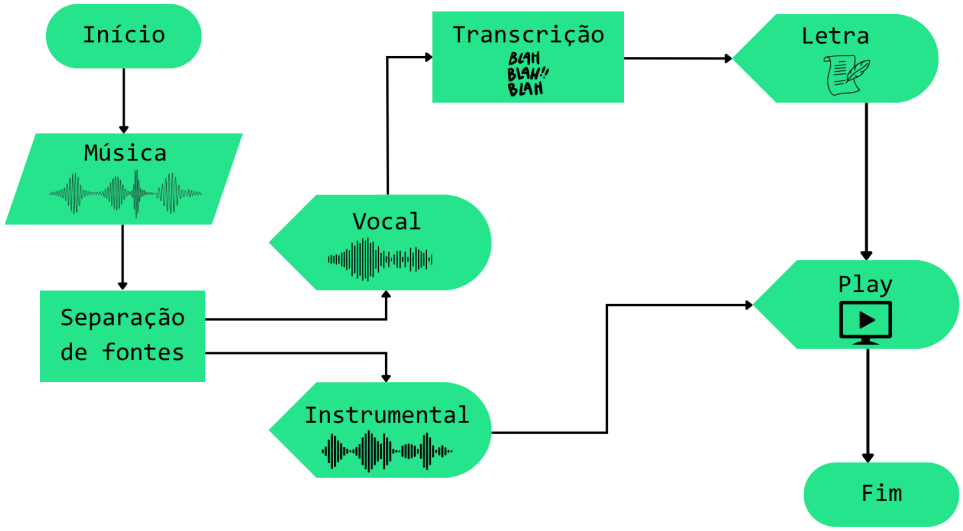
- Biblioteca musical: possibilidade de gerir as faixas musicais, como adicionar mais faixas ao conjunto;
- Entrada de áudio: o sistema deve ser capaz de aceitar faixas de áudio em formatos comuns como: MP3, WAV e FLAC;
- Separação de fontes musicais (vocal/instrumental): o *software* precisa ser capaz de reduzir ou remover a trilha vocal de uma música, deixando apenas a instrumental;
- Exibição sincronizada da letra: a letra da música em execução deverá ser exibida em tempo real, sincronizada com a reprodução da faixa instrumental. Idealmente, a linha atual ou a palavra sendo cantada deve ser destacada;
- Controle de reprodução: funções básicas de *player*, como tocar, pausar, parar, controle de volume e controle do cursor de momento (alterar o ponto de reprodução do áudio);
- Ajuste de tom (*pitch*) e de velocidade de reprodução do áudio (andamento): viabilidade de o usuário ser capaz de ajustar o tom da música para se adequar ao seu registro vocal e o andamento para facilitar o seu acompanhamento;
- Salvar configurações: eventualmente o usuário poderá salvar suas preferências de reprodução.

4.1.2 Requisitos não funcionais

Estes requisitos descrevem como o *software* deve ser:

- Usabilidade: a interface do usuário deve ser intuitiva, fácil de navegar e visualmente agradável, mesmo para usuários sem experiência técnica;
- Desempenho: o processamento deve ser realizado com baixa latência para uma experiência fluida em tempo real;
- Compatibilidade: deve ser compatível com sistemas operacionais amplamente utilizados e, idealmente, funcionar em diferentes configurações de *hardware*;
- Escalabilidade: a arquitetura deve permitir a adição de novas funcionalidades no futuro, como suporte a mais formatos de arquivo ou recursos avançados como análise vocal.

Figura 4 – Fluxograma do projeto



Fonte: Autor (2025)

4.2 Design da arquitetura do sistema

Nesta seção, será abordada a decisão das tecnologias específicas para cada módulo do projeto e sua estrutura básica, incluindo seus componentes, suas funcionalidades e as

interações entre eles. Para garantir que todos os requisitos do aplicativo sejam atendidos com sucesso, é essencial seguir um fluxograma claramente definido, como ilustrado na figura 4. Além de organizar as etapas do desenvolvimento de maneira lógica e sequencial, esse tipo de representação visual ressalta as interdependências entre as diversas fases e módulos do sistema. Seguir um fluxograma assegura que todos os requisitos, tanto funcionais quanto não funcionais, sejam tratados no momento adequado, reduzindo retrabalhos e otimizando a utilização de recursos.

4.3 Música

O arquivo de entrada representa o sinal sonoro bruto em sua forma completa. No Bahokê, este sinal é fornecido obrigatoriamente em formato MP3. Esta escolha foi definida por se tratar de um protótipo, no qual optou-se por não adotar outros formatos de áudio no momento a fim de reduzir a complexidade do desenvolvimento e garantir a estabilidade do fluxo de processamento. Ao padronizar a entrada, o sistema assegura uma manipulação mais coerente dos arquivos, servindo como a base necessária para as etapas subsequentes.

4.4 Separação de fontes

A etapa de Separação Cega de Fontes é crucial para isolar o sinal do vocal do conteúdo instrumental, garantindo uma entrada limpa para o processo seguinte de Reconhecimento Automático de Fala e para a geração da trilha de acompanhamento (*backing track/playback*). A seleção dos modelos para realizar a separação de fontes musicais, foi orientada por um critério duplo que equilibra a performance algorítmica com a viabilidade operacional e de licenciamento para um sistema *open-source* como o Bahokê. Com base nesses critérios, os modelos selecionados foram:

- Open-Unmix: Representa a arquitetura clássica baseada em U-Net com foco na modularidade e na separação limpa de *stems* (fontes).
- Spleeter: Notável por sua eficiência e velocidade, permitindo um *trade-off* favorável entre desempenho e tempo de processamento.
- Demucs (2022): Considerado o estado da arte, com arquiteturas avançadas que alcançam consistentemente as melhores métricas de qualidade (SDR), servindo

como *benchmark* de precisão.

Para mitigar a contaminação cruzada entre as faixas e otimizar a qualidade das fontes isoladas, realizou-se uma análise comparativa utilizando os três modelos, que representam diferentes abordagens arquiteturais na separação de fontes musicais.

Na tabela 3, observa-se tal comparação; ressalta-se que na coluna de tempo de duas fontes do *Open-Unmix* encontra-se vazia devido ao modelo não fornecer esse tipo de separação.

Tabela 3 – Análise dos modelos testados

Modelo	**Domínio**	***Dataset***	**SDR**	**Tempo 2 fontes**	**Tempo 4 fontes**
Open-Unmix	espectrograma	Autor	5,3 dB	—	13,71 min
Spleeter	espectrograma	Autor	5,9 dB	0,58 min	1,15 min
Demucs	híbrido	Autor	7,7 dB	3,99 min	4,22 min

Fonte: Autor (2025)

A avaliação comparativa foi conduzida em um conjunto de dados específico proveniente da biblioteca fonográfica particular do autor do texto. O tempo de processamento registrado reflete a etapa de separação de fontes em músicas com uma duração média de quatro minutos, estabelecendo uma base temporal para a inferência. Ressalta-se que, em tais testes, a inferência dos modelos foi executada exclusivamente em Unidade Central de Processamento (CPU), não sendo empregados recursos de aceleração por placas gráficas (GPUs).

A avaliação do desempenho de cada modelo foi baseada em métricas objetivas de qualidade de separação (SDR), além de uma análise perceptiva. Essa experimentação preliminar permitiu a seleção da arquitetura que minimiza o vazamento e a distorção no sinal vocal, garantindo a maior fidelidade possível para a etapa de transcrição e a melhor qualidade do áudio instrumental. O modelo selecionado para o fluxo principal foi o Demucs.

Os testes de validação revelaram que o sistema apresenta uma modularidade que o qualifica não apenas como uma plataforma de entretenimento (karaokê), mas também como um recurso de treinamento musical. Essa capacidade é explorada por meio da funcionalidade inerente à Separação Cega de Fontes, que permite a supressão seletiva de uma das fontes isoladas. Consequentemente, o usuário pode silenciar o sinal de um instrumento específico, por exemplo, a bateria, e praticar a execução do instrumento

em tempo real sobre a base instrumental remanescente, estabelecendo um ambiente de imersão (*play-along*) para o aprendizado instrumental.

4.5 Vocal e instrumental

O processo de separação de fontes resulta na geração de duas fontes essenciais: o fluxo vocal, que contém exclusivamente a voz do cantor, é crucialmente encaminhado para a etapa de transcrição, garantindo a maior fidelidade na conversão ASR; concomitantemente, o fluxo instrumental – o qual pode ser decomposto em múltiplas fontes (ex.: bateria, contrabaixo, demais instrumentos), formando a trilha de acompanhamento (karaokê base), que será utilizada diretamente na fase final do sistema.

4.6 Transcrição

Esta seção detalha a metodologia de transcrição empregada no estudo, a qual é importante para a subsequente quantificação do desempenho dos sistemas de reconhecimento automático de fala (ASR).

O corpus desta avaliação consistiu em 30 segmentos musicais, utilizados em sua duração integral. Para cada segmento, foi estabelecida uma transcrição de referência (*Ground Truth*) manual e verificada, servindo como padrão ouro para a mensuração do erro.

A avaliação da performance foi realizada por meio da aplicação de dois modelos distintos de ASR, o modelo Vosk[28] designado como A e o modelo WhisperX[29] como B. A seleção destes modelos foi estritamente orientada por um requisito de funcionalidade essencial para a aplicação final desta pesquisa: a capacidade de gerar uma saída de transcrição que incluísse *timestamps* (carimbos de tempo) para cada palavra reconhecida. A inclusão dos *timestamps* de palavra é um recurso fundamental para a sincronização precisa da letra com o áudio, viabilizando a correta exibição das letras (legenda) em formato de karaokê. Essa funcionalidade de alinhamento temporal é crucial para a usabilidade e a eficácia da aplicação, sendo o critério determinante para a escolha e a avaliação dos Modelos A e B em detrimento de outros sistemas de ASR. Portanto, os Modelos A e B foram selecionados não apenas por sua relevância no estado da arte, mas

[28] https://alphacephei.com/vosk/
[29] https://arxiv.org/pdf/2303.00747

primordialmente por sua capacidade de gerar marcações temporais em nível de palavra (*Word-Level Timestamps*). Essa funcionalidade permite que o modelo identifique com precisão o instante exato em que cada palavra começa a ser dita e o momento em que sua pronúncia é finalizada. Para o Bahokê, esse nível de detalhamento é indispensável, pois possibilita a sincronização perfeita entre o áudio processado e a exibição visual das letras, garantindo que o texto seja destacado conforme o progresso da performance vocal.

O cálculo do WER foi implementado utilizando a biblioteca jiwer (*Jitsi Word Error Rate*). A escolha desta ferramenta deve-se à sua eficiência e robustez, uma vez que utiliza implementações otimizadas em C++ para o cálculo da distância de edição, garantindo a acurácia e a velocidade necessárias para o processamento do corpus. A biblioteca foi utilizada para aplicar etapas de pré-processamento de texto (como remoção de pontuação e padronização para minúsculas) antes do cálculo, assegurando que a comparação entre referências e hipóteses fosse linguisticamente justa e representativa da precisão léxica dos modelos. A utilização do jiwer assegura que os resultados apresentados são baseados em uma metodologia computacionalmente validada e amplamente aceita pela comunidade de pesquisa em ASR.

Para cada um dos 30 segmentos, foram coletadas as transcrições do Modelo A e do Modelo B, esta etapa culminou na obtenção das métricas primárias de *Word Error Rate* (WER), que foram subsequentemente submetidas a uma análise estatística descritiva, incluindo média, mediana, variância e desvio-padrão, visando determinar a precisão e a consistência de cada modelo. Para a execução da Análise Estatística Descritiva, foi empregada a biblioteca NumPy (*Numerical Python*), que é a ferramenta fundamental na linguagem *Python* para a computação científica e operações com matrizes multidimensionais. A biblioteca NumPy foi especificamente utilizada para o cálculo das seguintes métricas a partir dos 30 valores individuais de WER de cada modelo:

- Medidas de tendência central:
  - Média aritmética (μ): Para indicar o desempenho médio geral do modelo no corpus.
  - Mediana: Para identificar o valor central do WER, sendo menos sensível a *outliers* (erros extremamente altos ou baixos em poucas amostras).
- Medidas de dispersão:
  - Variância (σ^2): Para quantificar a dispersão dos resultados de WER em torno da média.

- Desvio-padrão (σ): Para fornecer uma medida da variabilidade na mesma unidade do WER. O desvio-padrão é essencial, pois sua magnitude indica a consistência e robustez do modelo; um valor mais baixo sugere que o modelo mantém uma performance estável em todas as amostras.

A utilização do NumPy garantiu a eficiência computacional e a padronização estatística na sumarização dos resultados, fornecendo os parâmetros necessários para a comparação rigorosa da acurácia e da estabilidade entre o Modelo A e o Modelo B. Na tabela 4, contempla-se os cálculos realizados.

Tabela 4 – Análise da taxa de erro de palavras

Modelo	**Global**	**Média**	**Mediana**	**Variância**	**Desvio-padrão**
A	0,7852	0,8026	0,7656	0,0333	0,1824
B	0,3526	0,3562	0,3292	0,0457	0,2138

Fonte: Autor (2025)

A taxa de erro de palavras (WER) mede a proporção de erros de transcrição, quanto menor o valor, melhor a precisão do modelo, sendo assim observa-se a superioridade do Modelo B, que demonstrou uma precisão significativamente superior ao Modelo A, com um WER Global de 0,3526. Logo denota-se que, no total do seu corpus de estudo de 30 amostras, o Modelo B cometeu erros em 35,26% das palavras, enquanto o Modelo A errou em 78,52% das palavras. Portanto, o Modelo B é o sistema de ASR mais eficaz para a transcrição dos dados analisados.

As métricas de dispersão (desvio-padrão e variância) medem o quão espalhados estão os resultados de WER individual em torno da média, ou seja, quão consistente é o modelo em diferentes amostras. O Modelo B possui um desvio-padrão ligeiramente maior (0,2138) do que o Modelo A (0,1824), esse fenômeno sugere que a performance do Modelo B é levemente menos consistente que a do Modelo A, em outras palavras, embora o Modelo B seja muito mais preciso na média, sua taxa de erro (WER) tende a variar mais drasticamente entre as amostras mais fáceis e as mais difíceis do corpus. A diferença entre a Média e a Mediana indica a distribuição dos erros, para o Modelo A, a Média (0,8026) é significativamente maior que a Mediana (0,7656), assim insinua que a distribuição do WER é assimétrica positiva, indicando que a cauda de resultados inclui um pequeno número de amostras com WER excepcionalmente alto (próximo de 1,0), puxando a Média para cima. Para o Modelo B, a Média (0,3562) também é maior que a Mediana (0,3292),

confirmando uma assimetria positiva, o que é um achado comum e indica que, apesar de o Modelo B ser bom na maioria das amostras, ele falha consideravelmente em alguns poucos trechos difíceis.

Essa análise comparativa do WER revelou que o Modelo B é o sistema de reconhecimento automático de fala de melhor desempenho, atingindo um WER Global de 0,3526 (35,26%), representando uma redução de mais de 40 pontos percentuais na taxa de erro em relação ao Modelo A (WER Global: 0,7852). Embora o Modelo B apresente um desempenho superior, ele exibe um desvio-padrão marginalmente maior (0,2138 contra 0,1824 do Modelo A), esta descoberta, em conjunto com a Média ser superior à Mediana em ambos os casos, indica que a performance do Modelo B, apesar de ser mais precisa, é levemente menos consistente entre as 30 amostras, logo implica que o Modelo B é altamente eficaz em trechos de áudio de baixa dificuldade, mas é mais suscetível a grandes picos de erro em amostras com alta complexidade acústica ou linguística, justificando possíveis investigações futuras sobre a robustez do Modelo B sob condições adversas.

Para concluir esta análise realizou-se o cálculo do Intervalo de Confiança (IC), uma ferramenta fundamental na estatística inferencial, utilizada para estimar um parâmetro populacional desconhecido (como a verdadeira média do WER) a partir de uma amostra finita (os 30 segmentos musicais). Diferentemente da média amostral ($\bar{x}$), que fornece apenas uma estimativa pontual, o IC fornece uma faixa de valores dentro da qual o parâmetro populacional tem uma alta probabilidade de se encontrar.

No contexto da avaliação de ASR, o IC serve a dois propósitos:

- Precisão da estimativa: a largura do intervalo (a diferença entre os limites superior e inferior) reflete a precisão da estimativa. Um intervalo mais estreito sugere maior confiança de que a Média Amostral está próxima da Média Populacional.
- Significância estatística: o IC permite testar a diferença entre os modelos. Se os Intervalos de Confiança de dois modelos não se sobrepuserem, pode-se concluir que existe uma diferença estatisticamente significativa em seus desempenhos médios de WER, no nível de confiança de 95%.

Estabelecer um nível de confiança de 95% justifica que, se o processo de amostragem e cálculo fosse repetido inúmeras vezes, esperaríamos que 95% dos intervalos construídos contivessem a verdadeira média populacional do WER. O cálculo do IC foi realizado com base na Distribuição *t* de *Student* (devido ao tamanho da amostra ser $n = 30$), utilizando o desvio-padrão da amostra e o Erro Padrão da Média (SEM, do

inglês *Standard Error of the Mean*). Na tabela 5, descreve-se os resultados dos cálculos realizados.

Tabela 5 – Resultados obtidos no cálculo do IC

Modelo	**Média WER**	**Desvio-padrão**	**Limite Inferior**	**Limite Superior**	**Amplitude**
A	$0,8026$	$0,1824$	$0,7345$	$0,8707$	$0,1362$
B	$0,3562$	$0,2138$	$0,2764$	$0,4360$	$0,1596$

Fonte: Autor (2025)

No Modelo A, o WER populacional real é estimado em estar entre $0,7345$ e $0,8707$ (ou $73,45\%$ e $80,07\%$ de erro) com 95% de confiança, já no Modelo B, o WER populacional real é estimado em estar entre $0,2764$ e $0,4360$ ($27,64\%$ e $43,60\%$ de erro) com 95% de confiança. Um limite superior acima de $1,0$ (100%) é comum quando a média é muito elevada e o desvio-padrão é grande, o que indicaria que, em algumas amostras, o número de erros (substituições, deleções e inserções) poderia ser maior que a quantidade total de palavras de referência.

A amplitude do Intervalo de Confiança (a largura do intervalo) reflete a precisão da estimativa, sendo diretamente influenciada pelo desvio-padrão. O Modelo A apresentou um IC ligeiramente mais estreito ($0,1362$) do que o Modelo B ($0,1596$), o que, a princípio, indica que a estimativa da média do Modelo A é marginalmente mais precisa; no entanto, o desvio-padrão do Modelo B ($0,2138$) é maior que o desvio-padrão do Modelo A ($0,1824$). Essa combinação indica que, embora o Modelo B seja mais preciso na média, a sua taxa de erro ($0,3562$) varia mais ($\pm 0,2138$) entre as músicas, resultando em uma maior incerteza na sua estimativa populacional (IC mais largo).

O principal achado estatístico desta análise é a não sobreposição dos Intervalos de Confiança. O limite superior do Modelo B é $0,4360$ e o limite inferior do Modelo A é $0,7345$. Como o limite superior do Modelo B ($0,4360$) é menor que o limite inferior do Modelo A ($0,7345$), há uma diferença estatisticamente significativa entre os dois modelos com um nível de confiança de 95%. A análise inferencial reforçou as conclusões descritivas. A não sobreposição dos Intervalos de Confiança de 95% confirma que a superioridade de desempenho do Modelo B (IC: $[0,2764, 0,4360]$) sobre o Modelo A (IC: $[0,7345, 0,8707]$) é estatisticamente significativa. Enquanto o Modelo B apresenta maior variabilidade em suas estimativas (maior desvio-padrão), ele estabelece inequivocamente um nível de precisão de transcrição que é fundamentalmente superior ao do Modelo A

para este corpus.

4.7 Letra

O objetivo desta etapa foi desenvolver um sistema de exibição de letras de música sincronizadas (karaokê) dentro de uma interface gráfica de usuário baseada na biblioteca PyQt5 UI[30]. O desafio técnico central residiu na necessidade de alta performance de renderização para manter a fluidez visual. A abordagem inicial de utilizar o elemento de interação *QTextEdit* para carregar o conteúdo das letras via função em um intervalo de tempo muito curto (aproximadamente 30 vezes por segundo, 33*ms*) resultou em um gargalo de processamento. A re-renderização completa da árvore de elementos HTML a cada ciclo do temporizador sobrecarregava o *thread* principal da GUI, manifestando-se como travamento da interface.

Para mitigar o problema de desempenho, a arquitetura foi migrada para uma solução de *Web Rendering* otimizada, utilizando o elemento de interação *QWebEngineView*. A otimização reside na separação das responsabilidades entre o *backend* que gera comandos *JavaScript* (JS) para a UI e o *frontend* que executa comandos JS para manipular o DOM (*Document Object Model*) e aplicar estilos CSS (*Cascading Style Sheets*). O princípio fundamental é: em vez de recalcular e renderizar toda a estrutura HTML, o sistema apenas envia comandos para trocar a classe CSS do elemento correspondente, uma operação que o *browser engine* executa de forma instantânea e não-bloqueante.

O recurso de *smooth scroll* automático e progressivo exigiu uma reestruturação do HTML injetado e da lógica de quebra de linha. O sistema foi configurado para agrupar palavras em segmentos (linhas) com base em um critério heurístico: o início de uma nova linha é detectado quando a primeira letra da palavra é maiúscula, após a palavra inicial da música. O controle de *scroll* é implementado no método atualizar karaokê (disparado pelo temporizador).

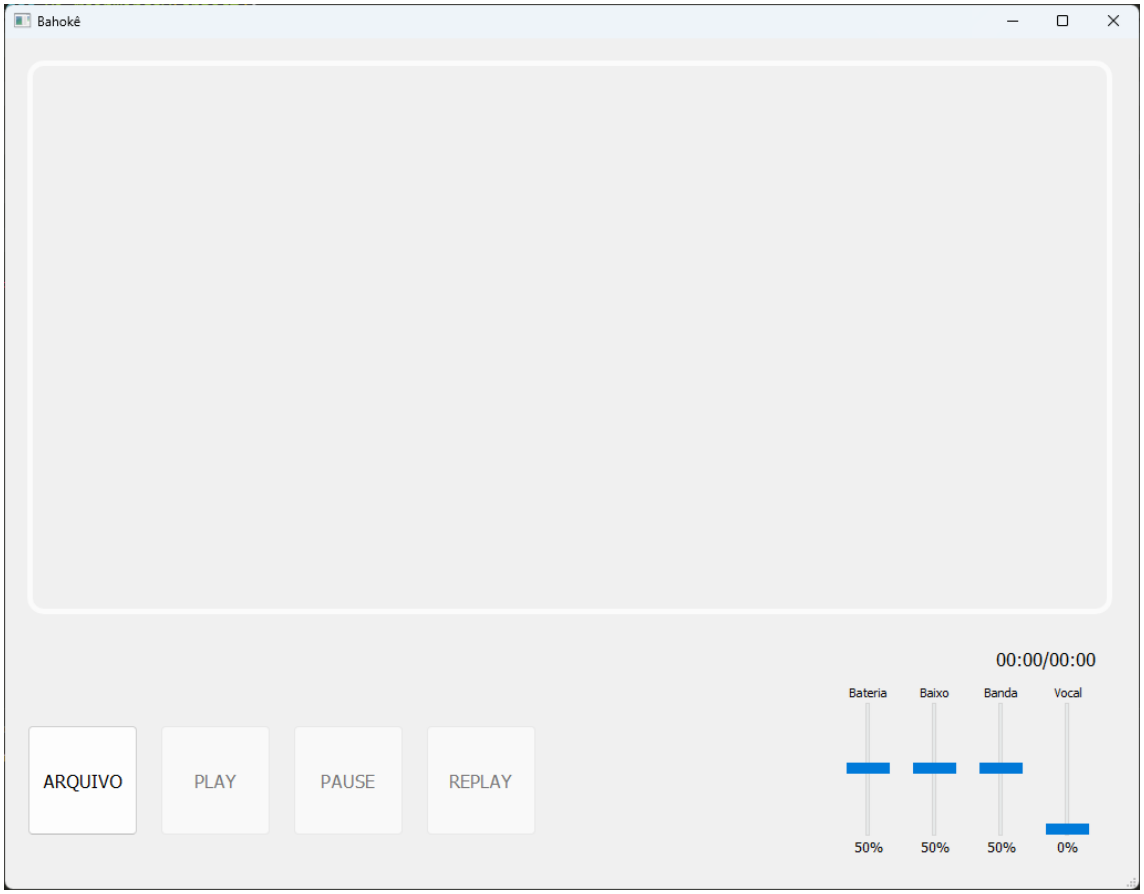
Assim, este design modular e a delegação de tarefas de renderização ao motor *web* garantem a sincronização precisa e a alta fluidez do efeito de karaokê, eliminando os problemas de travamento inerentes às abordagens baseadas na repintura recursiva do *QTextEdit*.

[30]https://pypi.org/project/PyQt5/

4.8 *Play*

Nesta arquitetura, o *Demucs* e o *WhisperX* foram integrados como motores de processamento para a separação de faixas e transcrição, respectivamente. Contudo, o diferencial de engenharia do Bahokê reside na implementação da etapa *Play*, que representa a fase de execução e consumo da saída do sistema. Nesta fase, o esforço de desenvolvimento concentrou-se em garantir que a sincronização temporal, o desempenho da GUI e o *feedback* visual em tempo real operassem de forma harmônica. Atuando como o módulo de renderização final, essa etapa integra as faixas de áudio isoladas aos metadados temporais da letra, convertendo saídas brutas de processamento em uma experiência de usuário estável, fluida e responsiva.

Figura 5 – Interface do Bahokê



Fonte: Autor (2025)

4.8.1 Inicialização e composição de mídia

Após escolher a música, iniciando a ação no botão nomeado arquivo, são realizadas as etapas iniciais de separação de fontes e a transcrição; logo após é então habilitado o botão *play* e, ao ser clicado, a execução inicia-se com a inicialização simultânea de quatro canais de áudio distintos. Todos os canais podem ser opcionalmente silenciados ou ter seus volumes ajustados (via deslizante de controle da faixa, como visto na 5) para permitir a prática vocal ou instrumental do usuário:

- Canal vocal: o arquivo de áudio *vocals.wav* é carregado em um *player* destinado aos vocais.
- Canal banda: o arquivo de áudio *other.wav* é carregado em um *player* secundário.
- Canal contrabaixo: o arquivo de áudio *bass.wav* é carregado em um *player* para a faixa de contrabaixo.
- Canal bateria: o arquivo de áudio *drums.wav* é carregado em seu *player* preparado.

A reprodução é iniciada através do método implementado no botão *play* da interface, garantindo que ambos os canais de áudio comecem no mesmo tempo zero, estabelecendo o relógio mestre do sistema.

4.8.2 Sincronização e módulo de controle temporal

O coração da etapa *Play* é o mecanismo de sincronização visual, que é acionado por um relógio, ajustado para um intervalo de 33*ms* (aproximadamente 30 quadros por segundo). A função central, é invocada a cada marcação do relógio e realiza duas operações críticas:

1. Coloração de palavras: é o mecanismo de *feedback* visual em tempo real, onde o sistema consulta continuamente a posição de reprodução do áudio (o relógio mestre) e a compara com os carimbos de tempo de início e fim de cada palavra da letra (metadados obtidos da Transcrição); ao detectar que a posição do áudio caiu no intervalo temporal de uma palavra específica, é emitido um comando *JavaScript* otimizado que altera a classe CSS do elemento HTML correspondente, colorindo a palavra de forma instantânea para indicar que ela está sendo cantada naquele exato momento.

2. Rolagem progressiva (*Scroll*): é o mecanismo de navegação visual que garante que a linha de letra atualmente em reprodução permaneça visível no centro da área de exibição. O sistema rastreia o tempo de áudio e o compara com os carimbos de tempo de fim de cada segmento (linha). Ao detectar a transição para um novo segmento, um comando *JavaScript* é injetado no motor de renderização. Este comando aciona o método de *scroll* no elemento da nova linha, configurando-o com o comportamento suave (*smooth*) e o alinhamento centralizado para uma transição visual fluida e focada.

4.8.3 Otimização de desempenho

A performance da etapa *Play* é mantida pelo princípio de mínima manipulação do DOM (árvore de objetos do html): o *backend* se concentra apenas no cálculo temporal e na emissão de comandos de texto, enquanto o pesado trabalho de renderização e animação é descarregado para o *thread* assíncrono do *QWebEngineView*, garantindo que a GUI permaneça responsiva e sem travamento durante toda a reprodução.

4.9 Requisitos para a execução do Bahokê

A execução robusta e eficiente do sistema Bahokê exige o cumprimento de requisitos mínimos de *hardware* e *software* para acomodar as demandas computacionais dos modelos e garantir a fluidez da interface gráfica.

4.9.1 Requisitos de ambiente de execução (*software*)

O sistema é construído sobre a arquitetura *Python*[31], sendo primariamente direcionado ao ambiente *Windows* devido a ser o ambiente acessível pelo autor, mas é potencialmente multiplataforma com a configuração correta:

- Sistema operacional: *Microsoft Windows* 10/11 (recomendado) ou distribuições *Linux* com suporte a *codecs* de mídia;
- Linguagem de programação: *Python* versão 3.12;

[31] https://www.python.org/

– FFmpeg[32]: uma estrutura de multimídia, com a capacidade de decodificar, codificar, entre outras funções, formatos de mídias antigos e modernos;
– Principais dependências[33] em *Python*:

  - Demucs: modelo para separação de fontes musicais, capaz de isolar vocais do restante do acompanhamento. No desenvolvimento do Bahokê foi utilizado o modelo *htdemucs*;
  - WhisperX: modelo de reconhecimento automático de fala (ASR). Gera transcrições de alta precisão com carimbos de tempo, marcando o início e fim da duração da palavra falada. No desenvolvimento do Bahokê foi utilizado o modelo *medium*;
  - PyQt5: para criar interfaces gráficas de usuário (GUIs) para aplicações *desktop*;
  - QtMultimedia: necessário para o funcionamento do *QMediaPlayer*;
  - QtWebEngine: Necessário para o *QWebEngineView*, motor de renderização das letras;
  - Pydub: para a manipulação de arquivos de áudio de forma simples e de alto nível;
  - Pathlib: oferece uma metodologia orientada a objetos para gerenciar caminhos do sistema de arquivos;
  - Unicodedata: biblioteca utilizada para a normalização de caracteres, garantindo que nomes de arquivos contendo caracteres especiais ou acentuações sejam processados corretamente pelo sistema;
  - Re: empregada em conjunto com a Unicodedata para a sanitização de caracteres, enquanto a Unicodedata normaliza os caracteres, a biblioteca Re aplica padrões de busca e substituição para remover caracteres não alfanuméricos, garantindo a integridade dos nomes de arquivos no sistema;
  - Sys: provê acesso a parâmetros e funções específicos do sistema;
  - Módulo JSON: responsável pelo armazenamento das informações relacionadas às transcrições das letras de forma estruturada.

Durante o desenvolvimento houve vários conflitos de dependências; assim sendo,

[32] https://www.ffmpeg.org/
[33] São pacotes, bibliotecas ou módulos de *software* dos quais um projeto depende para funcionar corretamente.

são listadas algumas bibliotecas e sua respectiva versão utilizada no projeto:

- Demucs: 4.0.1
- NumPy: 1.26.4
- ONNXRuntime: 1.17.3
- Pydub: 0.25.1
- PyQt5: 5.15.10
- PyQtWebEngine: 5.15.7
- Torch: 2.8.0
- TorchAudio: 2.8.0
- WhisperX: 3.7.4

4.9.2 Requisitos computacionais (*hardware*)

Devido à natureza do processamento de modelos de aprendizado de máquina em tempo real e em lote, a execução do Bahokê se divide em requisitos de processamento pesado (*backend*) e execução (*frontend*).

Processamento pesado (Separação e Transcrição):

- Processador (CPU): 4 cores físicos (ex.: Intel Core i5 de 7ª Geração ou equivalente AMD);
- Memória RAM: 8 GB;
- Placa de vídeo (GPU): nenhuma (pode rodar em CPU);
- Armazenamento: 5 GB de espaço livre para modelos e resultados.

Execução e interface gráfica (Reprodução):

- Processador (CPU): capacidade de processamento suficiente para manter o relógio do sistemas (*QTimer*) de 33*ms* estável;
- Memória RAM: 4 GB de memória livre para o ambiente de execução *Python* e o *QWebEngineView*.

5 ANÁLISE DE RESULTADOS E VALIDAÇÃO

Este capítulo apresenta a análise detalhada dos dados coletados a partir da avaliação qualitativa do Bahokê, com foco nas respostas e no *feedback* dos usuários. O grupo de participantes é composto, em sua maioria, por docentes do Instituto Artístico Carlos Gomes (IACG) de Dom Pedrito/RS, um ambiente especializado onde são ministradas aulas práticas de instrumentos musicais e disciplinas teóricas de música. A alta especialização dos participantes permite uma avaliação aprofundada da usabilidade, precisão técnica e relevância pedagógica das funcionalidades do sistema, como a separação de fontes e a sincronização temporal. O questionário aplicado encontra-se disponível no apêndice B.

O capítulo será estruturado para abordar os seguintes pontos:

- Apresentação e discussão dos resultados: serão apresentados os achados-chave da pesquisa, incluindo a identificação dos principais temas e padrões de opinião emergentes nos depoimentos dos usuários.
- Análise qualitativa detalhada: exploração aprofundada da experiência do usuário, focando em pontos fortes e pontos de melhoria conforme percebidos por aqueles que utilizaram o sistema.
- Validação da solução: utilização dos resultados coletados para validar ou refutar as hipóteses iniciais do projeto. Esta etapa confirma se o Bahoke é eficaz, utilizável e se resolve o problema proposto.

5.1 Informações de perfil

Para delinear o universo de análise, a pesquisa começou por coletar informações de perfil dos 12 participantes. A amostra revelou ser predominantemente adulta (83,3% entre 25 e 49 anos) e com forte inclinação para a prática musical, sendo 41,6% dos participantes identificados como Professores de Música ou com profissões relacionadas. Essa composição da amostra é crucial, pois indica que o *feedback* qualitativo posterior será fornecido por usuários com um alto nível de exigência e familiaridade com os conceitos de separação de trilhas e aprendizado musical.

5.2 Usabilidade e fluxo de trabalho

Os resultados são excepcionalmente positivos, validando o requisito de facilidade de uso e performance, como os recursos centrais do Bahokê também demonstram alta validação, conforme visto na tabela 6.

Tabela 6 – Resultados selecionados do questionário

Conceito	Média	Distribuição
O Bahokê é fácil de usar	4,92	11 de 12 deram nota 5
O Bahokê é rápido	4,75	9 de 12 deram nota 5
Qualidade da separação de fontes	4,92	11 de 12 deram nota 5
Clareza e impacto visual da coloração	4,75	9 de 12 deram nota 5
Precisão da transcrição/alinhamento temporal	4,58	9 de 12 deram nota 5

Fonte: Autor (2025)

As respostas abertas revelam três temas principais sobre a utilidade e a performance do sistema:

1. Essencialidade do Controle de Mixagem
   - O controle de volume de fontes (mixagem) é unanimemente considerado muito útil e essencial pelos usuários.
     - Finalidade principal (prática musical): a maioria dos usuários mencionou o uso para treinar a voz (baixar o vocal para treinar a voz) e para aulas de canto (muito útil para o ensino);
     - Flexibilidade e Customização: a relevância desta funcionalidade reside na autonomia concedida ao usuário para personalizar sua experiência de uso. Através do controle individual de ganho (volume) para cada trilha isolada, o sistema permite tanto o ajuste de uma mixagem personalizada para a sessão de karaokê quanto o isolamento técnico para o estudo musical. Essa versatilidade possibilita que o músico silencie um instrumento específico para praticar o *play-along* ou, inversamente, isole uma faixa para análise detalhada da execução original;
     - Valor agregado: é visto como prático e com potencial além do karaokê, citando o uso para produção musical ou para fins didáticos.
2. Mecanismo de Rolagem: Alto Foco na Prática

  - O mecanismo de rolagem progressivo (foco de linha) também foi unanimemente validado como um recurso de alto valor.
    - Apoio ao foco: o recurso cumpre sua função primária, sendo descrito como: ajudou a não se perder na hora que está cantando;
    - Eficiência pedagógica: foi explicitamente reconhecido como muito útil para o ensino, eficiente e dinâmico;
    - Confirmação de precisão: as respostas indiretamente confirmam a alta precisão da sincronização, já que a rolagem progressiva só é útil se estiver bem alinhada.

3. Performance Sólida com Ponto de Atenção
  - Os resultados de performance indicam uma execução robusta, mas identificam um ponto fraco na qualidade do conteúdo.
    - Performance (*lag/freezing*): a performance é validada. A maioria reporta não como resposta ou nenhum problema de atraso, lentidão ou travamento (média 4,75);
    - Ponto de atenção (qualidade da transcrição): a única referência negativa foi: "Não. Somente um erro na transcrição da música." Esse *insight* é crucial, pois, embora a performance do sistema seja elevada, a qualidade do conteúdo (precisão da letra) requer atenção, justificando assim as duas notas 3.

5.3 Impacto, aceitação e validação final

Estabelecida a eficácia e a alta satisfação dos usuários em relação à usabilidade e aos recursos centrais, esta seção objetiva mensurar o valor agregado e a intenção de uso do sistema, que são fatores cruciais para a validação do produto. Para tal, são analisadas as percepções dos participantes sobre as maiores vantagens competitivas e a frequência de utilização esperada, o que permite validar ou refutar as hipóteses iniciais do projeto sobre a relevância e a demanda da solução no contexto musical.

Sobre a questão relacionada à interface do Bahokê ser intuitiva, como média obteve-se 4,75 (9 de 12 deram nota 5), o que confirma a alta usabilidade já verificada. A interface é percebida como fácil de usar e fácil de entender. Acerca da frequência

de uso proposta, 91,7% dos participantes utilizariam o Bahokê Diariamente (41,7%) ou Semanalmente (50%), este é o dado mais importante para a validação da aceitação. A alta intenção de uso diário/semanal demonstra que o Bahokê atende a uma necessidade real no fluxo de trabalho/prática musical dos usuários, observando que a sua rejeição foi de 0%.

As respostas sobre as maiores vantagens convergem para dois temas principais: praticidade (tudo em um) e versatilidade (ensino e entretenimento). Em relação à praticidade, citamos comentários como: "Maior eficácia, tudo em um só", "Praticidade para tirar as músicas". O valor reside em não precisar de múltiplos aplicativos para separar áudio e sincronizar letra. A versatilidade e dupla função são vistas tanto para entretenimento/karaokê quanto para uso pedagógico/estudo. Com foco na audição musical, a separação permite: escutar o instrumento que preferir ou aprender os tons musicais, validando o impacto do recurso de áudio isolado.

No que se refere às funcionalidades essenciais, as respostas sobre o que mais sentiria falta se não existisse confirmam que o módulo de karaokê/transcrição e a mixagem/separação são o cerne do valor do Bahokê. A maior parte das respostas citou a mixagem do volume, a separação de fontes/vocal e banda e a transcrição da letra como funcionalidades insubstituíveis, assim validando que a proposta central do Bahokê (a integração dessas duas tecnologias) é o seu diferencial competitivo.

Quanto a oportunidades de melhoria, podem-se elencar os seguintes tópicos:

- Edição e customização: troca de tom (essencial para vocalistas e instrumentistas); ajuste de andamento; uso de cifras; mudança de cores e tamanho da fonte/*layout*; tais funcionalidades elevariam o Bahokê a um nível profissional;
- Integração/exportação: exportar a mixagem customizada; exportar partitura; salvar a mixagem (configurações); o que salienta a necessidade de uma funcionalidade para tirar o conteúdo de dentro do sistema;
- Navegação: uma linha do tempo que possa avançar ou retroceder o andamento da execução do áudio; em relação a usabilidade é um detalhe a ser adicionado.
- Instrumentos específicos: mais opções específicas de separação (ex.: acordeon, piano, violão); aponta a alta exigência do público músico.

5.4 Conclusões quanto às validações

O objetivo deste capítulo foi analisar os resultados da avaliação do sistema Bahokê para validar o cumprimento dos requisitos de projeto, usabilidade e aceitação. A partir da análise dos dados quantitativos e da codificação dos comentários qualitativos, as principais conclusões são:

- O Bahokê demonstrou alta taxa de sucesso na validação de seus requisitos centrais. O sistema é percebido como altamente utilizável, com médias de satisfação em usabilidade e performance ultrapassando $4,75$ em $5,0$. A proposta de valor, que é a integração da separação de fontes com o karaokê sincronizado, foi validada, sendo o controle de mixagem e a transcrição os recursos considerados essenciais pelos usuários.
- A intenção de uso revelou um forte indicador de aceitação e potencial de mercado, com $91,7\%$ dos participantes afirmando que utilizariam o Bahokê diariamente ou semanalmente. A versatilidade do aplicativo para entretenimento (karaokê) e uso didático (estudo musical) atende a uma necessidade clara e dupla do público-alvo, composto majoritariamente por músicos e professores.
- O principal ponto de atenção e oportunidade de melhoria reside na qualidade do conteúdo e nas funcionalidades avançadas de customização. A crítica mais recorrente, apesar do sistema ser robusto em performance, refere-se a erros pontuais na transcrição das letras. Além disso, o alto nível de exigência da amostra (músicos) demanda a inclusão de recursos de troca de tom, ajuste de andamento e exportação da mixagem para uso profissional e didático.

Em suma, a avaliação comprova a eficácia e a aceitação do Bahokê como uma ferramenta inovadora para a prática e aprendizado musical, estabelecendo uma base sólida para a continuidade do projeto mediante a incorporação dos itens de customização e aprimoramento da precisão do conteúdo.

6 CONCLUSÕES FINAIS E TRABALHOS FUTUROS

O presente trabalho atingiu seu objetivo principal ao propor e desenvolver o sistema Bahokê, que integra de maneira eficaz a separação de trilhas de áudio à exibição de letras sincronizadas. A solução mostrou-se comparável aos métodos tradicionais de reprodução musical, superando-os em termos de eficiência processual e apresentando uma elevada aceitação na experiência do usuário.

A eficiência da solução foi validada pela otimização do fluxo de trabalho, superando a complexidade e o tempo demandados pelos métodos convencionais de produção de karaokê. Esse desempenho é corroborado pela alta aceitação do Bahokê, que registrou $91,7\%$ de intenção de uso, e pela sua usabilidade intuitiva, avaliada com média de 4,92 em 5,0. Tais indicadores comprovam que o sistema atinge os objetivos propostos com menor esforço cognitivo e operacional por parte do usuário.

Em relação à fidelidade sonora, o módulo de separação de fontes obteve uma avaliação média de $4,92$ em $5,0$, validando a qualidade da extração do áudio para fins de prática musical. Apesar da validação positiva, o estudo identificou áreas críticas para o aprimoramento contínuo, as quais se tornam a agenda de trabalho futuro. O desafio central reside na precisão da transcrição das letras, que exige maior rigor algorítmico, especialmente em variações de andamento. Além disso, para atender às expectativas do público profissional, é indispensável a implementação de controles avançados de áudio, como ajuste de tom (transposição), alteração de andamento (tempo) e a funcionalidade de exportação da mixagem em formato de áudio.

Conclui-se que o Bahokê é uma solução tecnológica eficaz e relevante no domínio musical, apresentando um grande potencial para o mercado e servindo de base para futuras pesquisas no desenvolvimento de ferramentas inteligentes de auxílio à performance e ao aprendizado, confirmando a viabilidade de sistemas baseados em BSS quando o foco está na experiência e na praticidade do usuário.

REFERÊNCIAS

BAIN, M. et al. Whisperx: Time-accurate speech transcription of long-form audio. **INTERSPEECH 2023**, 2023.

BARELLI, S. **Videokê une diversão e negócios no mesmo local**. [S.l.]: Folha de S.Paulo, 1996. <https://www1.folha.uol.com.br/fsp/1996/1/28/tudo/6.html>. Acessado em 08 de abril de 2025.

CAPEL, V. **Audio and Hi-Fi Engineer's Pocket Book**. 3. ed. [S.l.]: Newnes, 1994.

COMON, P. Independent component analysis, a new concept? **Signal processing**, Elsevier, v. 36, n. 3, p. 287–314, 1994.

DEMO, P. **Introdução à metodologia da ciência SP: Atlas**. [S.l.]: Atlas, 1985.

DODGE, C.; JERSE, T. A. **Computer music: synthesis, composition, and performance**. [S.l.]: Macmillan Library Reference, 1985.

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

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

DéFOSSEZ, A. et al. Music source separation in the waveform domain. **ArXiv**, abs/2111.03603, 2021. Disponível em: https://hal.science/hal-02379796v2.

ESTRANHO, R. M. **Como o aparelho de videokê é capaz de dar notas às pessoas?** 2011. <https://super.abril.com.br/mundo-estranho/como-o-aparelho-de-videoke-e-capaz-de-dar-notas-as-pessoas/>. Acessado em 08 de abril de 2025.

FU, W. **Principle and Application of Blind Source Separation Technology**. [S.l.]: Springer, 2025.

GOMES, L. C.; LABRADA, J. **Áudio digital: livro didático**. 3. ed. Palhoça: UnisulVirtual, 2011. 269 p. ISBN 9788578172213.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

HANNUN, A. Y. The history of speech recognition to the year 2030. **CoRR**, abs/2108.00084, 2021. Disponível em: https://arxiv.org/abs/2108.00084.

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

HINTON, G. et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. **IEEE Signal Processing Magazine**, v. 29, n. 6, p. 82–97, 2012.

HOSOKAWA, S.; MITSUI, T. The genesis of karaoke. In: **Karaoke Around the World**. [S.l.]: Routledge, 2005. p. 41–54.

JÚNIOR, P. D. de T. **Separaçao automática de instrumentos de percussao brasileira a partir de mistura pré-gravada**. Tese (Doutorado) — MS thesis, Federal University of Amazonas, Manaus, Brazil, 2016.

KIM, M. et al. **KUIELab-MDX-Net: A Two-Stream Neural Network for Music Demixing**. 2021. Disponível em: https://arxiv.org/abs/2111.12203.

KONG, Q. et al. Decoupling magnitude and phase estimation with deep resunet for music source separation. **CoRR**, abs/2109.05418, 2021. Disponível em: https://arxiv.org/abs/2109.05418.

LAMPERT, T. A.; O'KEEFE, S. E. A survey of spectrogram track detection algorithms. **Applied acoustics**, Elsevier, v. 71, n. 2, p. 87–100, 2010.

LECUN, Y. **Generalization and Network Design Strategies**. [S.l.], 1989.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, Macmillan Publishers Limited, v. 521, p. 436–444, 2015.

LEPCHA, D. C. et al. Image super-resolution: A comprehensive review, recent trends, challenges and applications. **Information Fusion**, Elsevier, v. 91, p. 230–260, 2023.

LUO, Y.; MESGARANI, N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Institute of Electrical and Electronics Engineers (IEEE), v. 27, n. 8, p. 1256–1266, ago. 2019. ISSN 2329-9304. Disponível em: http://dx.doi.org/10.1109/TASLP.2019.2915167.

LUO, Y.; YU, J. **Music Source Separation with Band-split RNN**. 2022. Disponível em: https://arxiv.org/abs/2209.15174.

MAKINO, S. **Audio source separation**. [S.l.]: Springer, 2018. v. 433.

MILETTO, E. M. et al. Introdução à computação musical. In: SN. **IV Congresso Brasileiro de Computação**. [S.l.], 2004.

NAIK, G. R.; WANG, W. et al. Blind source separation. **Berlin: Springer**, Springer, v. 10, p. 978–3, 2014.

NASCIMENTO, J. R. **Separação de Fontes Sonoras Auxiliada por Deep Learning**. Dissertação (Dissertação de Mestrado) — Programa de Engenharia de Sistemas e Computação (PESC), Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia (COPPE), Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Dezembro 2021. Disponível em: https://www.cos.ufrj.br/uploadfile/publicacao/3033.pdf.

NEIVA, F. W.; SILVA, R. L. d. S. d. Revisão sistemática da literatura em ciência da computação – um guia prático. **RelaTeDCC**, Universidade Federal de Juiz de Fora, v. 001, 2016. Guia Prático.

NG, A. **Machine Learning Yearning: Technical Strategy for AI Engineers, In the Era of Deep Learning**. [S.l.]: deeplearning.ai, 2018.

PARK, C.; CHEN, M.; HAIN, T. **Automatic Speech Recognition System-Independent Word Error Rate Estimation**. 2024. Disponível em: https://arxiv.org/abs/2404.16743.

PINNA, J.; ROCHA, R. **Karaokê: uma onda que não sai de moda - Novas casas reforçam roteiro para soltar a voz no Rio**. [S.l.]: O Globo, 2025. <https://oglobo.globo.com/rioshow/eventos/guia/karaoke-uma-onda-que-nao-sai-de-moda>. Acessado em 04 de abril de 2025.

PRODANOV, C. C.; FREITAS, E. C. D. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico-2ª Edição**. [S.l.]: Editora Feevale, 2013.

ROMANO, M. R.; ATTUX, R. Um estudo sobre separação cega de fontes e análise de componentes independentes. **XXXV SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES E PROCESSAMENTO DE SINAIS – SBrT2017**, 2017.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. **CoRR**, abs/1505.04597, 2015. Disponível em: http://arxiv.org/abs/1505.04597.

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

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, Nature Publishing Group UK London, v. 323, n. 6088, p. 533–536, 1986.

SARUWATARI, H. et al. Speech enhancement based on blind source separation in car environments. In: IEEE. **21st International Conference on Data Engineering Workshops (ICDEW'05)**. [S.l.], 2005. p. 1205–1205.

SAWATA, R. et al. **All for One and One for All: Improving Music Separation by Bridging Networks**. 2021. Disponível em: https://arxiv.org/abs/2010.04228.

SHI, Z. **Intelligence science: Leading the age of intelligence**. [S.l.]: Elsevier, 2021.

SHINDE, P. P.; SHAH, S. A review of machine learning and deep learning applications. In: **2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)**. [S.l.: s.n.], 2018. p. 1–6.

SOMMERVILLE, I. **Engenharia de Software**. 9. ed. São Paulo: Pearson, 2011.

STOLLER, D.; EWERT, S.; DIXON, S. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. **CoRR**, abs/1806.03185, 2018. Disponível em: http://arxiv.org/abs/1806.03185.

STÖTER, F.-R. et al. Open-unmix - a reference implementation for music source separation. **Journal of Open Source Software**, The Open Journal, v. 4, n. 41, p. 1667, 2019. Disponível em: https://doi.org/10.21105/joss.01667.

TAKAHASHI, N.; GOSWAMI, N.; MITSUFUJI, Y. **MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation**. 2018. Disponível em: https://arxiv.org/abs/1805.02410.

TAKAHASHI, N.; MITSUFUJI, Y. **D3Net: Densely connected multidilated DenseNet for music source separation**. 2021. Disponível em: https://arxiv.org/abs/2010.01733.

TALBOT-SMITH, M. **Sound Engineering Explained**: Second edition. [S.l.]: Focal Press, 2001.

TECHNAVIO. **Karaoke Market - Industry Analysis, Trends, Market Size, and Forecast 2025-2029**. 2025. <https://www.technavio.com/report/karaoke-market-industry-analysis>. Acessado em 11 de abril de 2025.

VASWANI, A. et al. Attention is all you need. **CoRR**, abs/1706.03762, 2017. Disponível em: http://arxiv.org/abs/1706.03762.

VINCENT, E. et al. From blind to guided audio source separation: How models and side information can improve the separation of sound. **IEEE Signal Processing Magazine**, Institute of Electrical and Electronics Engineers, v. 31, n. 3, p. 107–115, maio 2014. Disponível em: https://inria.hal.science/hal-00922378.

VINCENT, E.; GRIBONVAL, R.; FEVOTTE, C. Performance measurement in blind audio source separation. **IEEE Transactions on Audio, Speech, and Language Processing**, v. 14, n. 4, p. 1462–1469, 2006.

VINCENT, E. et al. Blind audio source separation. **Queen Mary, University of London, Tech Report C4DM-TR-05-01**, 2005.

VINCENT, E.; VIRTANEN, T.; GANNOT, S. **Audio source separation and speech enhancement**. [S.l.]: John Wiley & Sons, 2018.

WEBSTER, M. B.; LEE, J. Blind source separation of single-channel mixtures via multi-encoder autoencoders. 2024. Disponível em: https://arxiv.org/abs/2309.07138.

WYSE, L. **Audio Spectrogram Representations for Processing with Convolutional Neural Networks**. 2017. Disponível em: https://arxiv.org/abs/1706.09559.

XIN, J. et al. Blind source separation and speech enhancement. **Mathematical Modeling and Signal Processing in Speech and Hearing Sciences**, Springer, p. 141–188, 2014.

XU, J. Application of blind source separation in sound source separation. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2019. v. 1345, n. 3, p. 032006.

YU, X.; HU, D.; XU, J. **Blind Source Separation Theory and Applications**. [S.l.]: Science Press, 2014.

ZHANG, A. et al. **Dive into Deep Learning**. [S.l.]: Cambridge University Press, 2021. Release 0.17.1, Dec 12, 2021.

ZHANG, L.; WANG, S.; LIU, B. Deep learning for sentiment analysis : A survey. **CoRR**, abs/1801.07883, 2018. Disponível em: http://arxiv.org/abs/1801.07883.

ZHOU, X.; TAROCCO, F. **Karaoke: The global phenomenon**. [S.l.]: Reaktion Books, 2013.

APÊNDICE A – TABELA DE TAXA DE ERRO DE PALAVRA

Tabela 7 – Dados complementares do cálculo de WER

Composição Musical	Tam. (min)	*A(min)	**B(min)	WER A	WER B
Abaralhando A Barbela	04:21	03:37	05:04	0,7943	0,3354
Amaro, Norico e Léco	04:03	03:35	05:33	1,2882	1,1354
Andejo	03:17	03:12	02:55	0,9625	0,4917
Batendo Água	03:44	03:37	03:32	0,7425	0,2052
Bugio do Fole Solto	03:39	05:26	03:23	0,9916	0,6878
Cadela Baia	03:22	01:52	02:44	0,7078	0,3735
Castração a Pealo	03:55	01:15	04:27	0,6667	0,2523
China Atrevida	03:42	02:06	03:08	0,6106	0,2056
Cinco e Meia da Manhã	04:12	02:03	04:03	0,7668	0,3738
De Relancina	04:21	02:29	03:39	0,5885	0,1728
Do Fundo da Grota	04:59	02:43	06:04	0,8365	0,5142
Eu, Mais Ela e o Tostado	04:00	06:54	03:49	0,9498	0,4702
Gritos de Liberdade	03:27	02:21	02:55	0,9000	0,3200
Hino Ao Rio Grande	03:06	00:44	02:37	0,8523	0,2282
Lanceiros Negros	03:50	02:31	04:20	0,8884	0,6736
Laço Armado	04:23	01:06	03:53	0,8209	0,2196
Milonga Abaixo de Mau Tempo	05:20	02:54	04:30	0,6538	0,3491
Nego Betão	04:25	03:49	04:22	0,9045	0,3701
No Chaquaio do Pandeiro	03:50	01:47	06:04	0,7368	0,2632
No Quadro do Coração	03:53	01:23	02:56	0,7644	0,0345
Nos Varzedos da Fronteira	03:46	02:54	03:38	1.3434	0,6566
Palanque do Passado, Esteio do Futuro	04:22	02:48	04:21	0,7154	0,1355
Pra Bailar de Cola Atada	04:11	02:55	04:23	0,6576	0,3333
Querência Amada	03:47	00:47	03:51	0,5541	0,3514
Ritual de Morte e Manada	04:00	02:33	03:08	0,9369	0,2973
Te Chamo Prenda	03:58	01:17	03:52	0,7110	0,2624
Tempo Bueno	03:16	00:44	02:28	0,6450	0,3250
Timbre de Galo	03:46	03:03	02:39	0,7583	0,2701
Tua Imagem Num Poema	04:09	00:50	03:56	0,5407	0,0447
Vanera do Peão Ajustado	04:16	02:19	02:36	0,7885	0,3348

Fonte: Autor (2025)

* Tempo de Transcrição A (min)

** Tempo de Transcrição B (min)

APÊNDICE B – QUESTIONÁRIO DE AVALIAÇÃO QUALITATIVA DO BAHOKÊ

Título da Pesquisa: Bahokê: Uma Abordagem de Aprendizado Profundo para Separação Cega de Fontes e Geração de Trilhas de Karaokê

Pesquisador Responsável: Oesley Rodrigues Machado

Orientador: Gerson Alberto Leiria Nunes

Contato do Pesquisador: oesleymachado.aluno@unipampa.edu.br

I. Introdução e Objetivo

O objetivo desta entrevista é coletar dados qualitativos sobre a usabilidade, a aceitação e o impacto das funcionalidades do sistema Bahokê, que combina a separação de trilhas de áudio (vocal, instrumental) com a exibição de letras sincronizadas (karaokê). Sua participação nos ajudará a entender como a tecnologia pode auxiliar na prática musical e no aprendizado de letras.

II. Procedimentos e Participação

Sua participação consiste em: Responder a perguntas demográficas básicas. Testar as funcionalidades do sistema (separação de fontes e karaokê sincronizado). Compartilhar sua opinião sobre a precisão da sincronização, a fluidez do sistema e a utilidade dos recursos de áudio isolados. A entrevista terá uma duração aproximada de 15 minutos.

III. Riscos e Benefícios

Riscos: Não há riscos físicos, morais, legais ou psicológicos conhecidos associados a esta entrevista. Benefícios: Sua contribuição é fundamental para o aprimoramento de ferramentas tecnológicas aplicadas à música e ao aprendizado.

IV. Confidencialidade e Privacidade (LGPD)

Garantimos o sigilo e a confidencialidade dos dados pessoais e das respostas fornecidas. Anonimato: Os dados serão tratados de forma agregada. Seus dados de contato e nome completo serão utilizados apenas para fins de registro interno da pesquisa e não serão associados às suas respostas ou citações em publicações, a menos que você explicitamente autorize a divulgação na Seção V. Uso dos Dados: Os dados coletados serão utilizados exclusivamente para fins de pesquisa, incluindo a análise, a elaboração de relatórios e a apresentação em trabalhos acadêmicos ou conferências.

V. Consentimento do Participante

Eu, _____, declaro que li e compreendi as informações contidas neste termo, tive a oportunidade de fazer perguntas e todas foram respondidas satisfatoriamente. Compreendo que minha participação é voluntária e posso me recusar a responder qualquer pergunta ou desistir de participar a qualquer momento, sem qualquer prejuízo ou ônus.

Declaro meu consentimento em participar desta pesquisa. Sim ( ) Não ( )

Autorização de divulgação das citações.

( ) Autorizo o uso das minhas citações e opiniões para fins de análise e pesquisa, desde que o meu nome real não seja divulgado e a citação seja atribuída a um código de participante.
( ) Autorizo a inclusão da minha citação na pesquisa, identificando apenas meu perfil (Ex: profissão), mas mantendo meu nome real confidencial.
( ) Autorizo, além do perfil, a divulgação do meu nome real junto com as citações e o conteúdo desta entrevista.
( ) Não autorizo a utilização de nenhuma citação ou opinião textual minha na pesquisa.

Em qual faixa etária você se encontra?

( ) 8 a 13 anos ( ) 14 a 17 anos ( ) 18 a 24 anos ( ) 25 a 34 anos
( ) 35 a 49 anos ( ) 50 a 65 anos ( ) 66 anos ou mais ( ) Prefiro não responder

Qual a sua profissão?

Instrumentos musicais que toca?

( ) Vocalista ( ) Violão/Guitarra ( ) Piano ( ) Acordeon/Gaita/Sanfona
( ) Sopros/Metais ( ) Contrabaixo ( ) Percussão ( ) Bateria
( ) Multi-instrumentista ( ) Nenhum ( ) Outro:

A escala de 1 (Ruim/Insatisfatório) a 5 (Excelente/Totalmente Satisfatório) é uma Escala de Concordância/Satisfação onde os valores representam níveis graduais de qualidade ou satisfação com o recurso avaliado.

O Bahokê é fácil de usar? (1) (2) (3) (4) (5)

Como você avalia a qualidade da separação de fontes (Vocal, Instrumental, etc.) para o uso em estudo/prática? (1) (2) (3) (4) (5)

O controle de volume das fontes (mixagem) é útil? Como você o utilizaria?

A precisão da transcrição da letra e o alinhamento temporal entre a palavra colorida e o áudio foram satisfatórios? (1) (2) (3) (4) (5)

Como você avalia o mecanismo de rolagem progressivo (foco de linha)? Ele ajudou a manter seu foco durante a prática?

Qual a clareza e o impacto visual da coloração de palavras? (Cor, Borda, Tamanho da Fonte) (1) (2) (3) (4) (5)

Você notou algum atraso, lentidão (lag) ou travamento (freezing) ao iniciar, parar ou durante o karaokê?

O Bahokê é rápido? (1) (2) (3) (4) (5)

A interface do Bahokê é intuitiva? (1) (2) (3) (4) (5)

Se o Bahokê estivesse disponível, qual seria a frequência com que você o utilizaria para aprender novas músicas ou praticar?
( ) Diariamente ( ) Semanalmente ( ) Raramente ( ) Nunca

Na pergunta anterior se respondeu à opção "nunca". Poderia informar o porquê?

Quais são as maiores vantagens de ter a separação de fontes integrada com a sincronização de letras em um único aplicativo?

Qual funcionalidade você sentiu falta no Bahokê?

Qual funcionalidade você mais sentiria falta se o sistema não a tivesse?

Deixe um comentário geral sobre sua experiência com o Bahokê e qualquer sugestão final.

APÊNDICE C – COMENTÁRIOS E SUGESTÕES FINAIS DA AVALIAÇÃO QUALITATIVA

A seguir, estão listados os comentários gerais e as sugestões finais fornecidos pelos participantes da pesquisa. Estes comentários complementam a análise qualitativa do Capítulo 5, detalhando a percepção individual sobre a experiência com o sistema Bahokê.

- "Ótima ferramenta tanto para entretenimento como para fins de didática em ensino musical."
- "Bahokê tem uma interface extremamente simples (que é um ponto positivo) e tem boas opções para criar um karaokê pessoal como o usuário desejar. Também vejo que seria de bastante utilidade em produções audiovisuais, como edição de áudio/vídeo."
- "Acho que incrementaria grandiosamente o programa se houvesse mais opções específicas de separação de instrumentos, como uma opção de mixagem específica para piano, violão, etc. Uma opção de exportar a mixagem customizada do usuário em formato de áudio (como .mp3) seria essencial para produção audiovisual. Como a ideia do Bahokê é ser um karaokê pessoal, acredito que casaria bem se tivesse opção de customizar o layout do programa, como o tamanho das legendas, a cor, a fonte, a cor de fundo."
- "Bahokê me surpreendeu de forma positiva e desejo o melhor para esse projeto."
- "Um aplicativo com muito potencial para ser explorado e aperfeiçoado."
- "Ótima ferramenta para uso de entretenimento, podendo assim ser transformada como função pedagógica no ensino de música, havendo algumas funcionalidades a mais, como: ajuste de andamento, mudança de tonalidade, uso de cifras, etc."
- "Muito interessante, gostei bastante!"
- "Foi uma experiência rápida, mas surgiram muitas ideias para trabalhar em aula a partir do Bahokê e acredito ser um projeto muito útil para quem trabalha no meio musical."
- "Achei realmente interessante e prático de ser usado, tanto na função de um karaokê como num separador de pistas para ouvir melhor o som de um determinado instrumento."
- "Que seja apresentado ao maior número de pessoas."
- "Achei ótimo!"

- "Ótima ideia e com grandes perspectivas."
- "Excelente! O projeto está bem estruturado, construído para músicos e público em geral, inclusive é uma excelente aposta para o entretenimento de eventos. Além disso, achei interessante trazer para o contexto nativista, uma área pouco explorada e abordada. Parabéns ao autor e ao orientador, excelente trabalho!"
- "Parabéns pela criação de uma ferramenta tão útil para o aprendizado musical e que, ao mesmo tempo, diverte o usuário com suas funcionalidades difíceis de serem encontradas juntas num mesmo aplicativo."