Artigo Original

https://doi.org/10.1590/2317-6431-2025-3117pt

Audiology Communication Research
ISSN 2317-6431

Processamento de arquivos de áudio para composição de testes de reconhecimento de fala em português

Processing of audio files for the construction of speech recognition tests in portuguese

Kauana Knappmann[1,2], *Stephan Paul*[2,3]

RESUMO

Objetivo: desenvolver códigos computacionais para modificação automatizada de grandes quantidades de sentenças gravadas, que realizem modificações de formato, filtragens, simulem o processamento dos sinais sonoros em implantes cocleares e ajustem as médias quadráticas de amplitude, a fim de equalizar o volume sonoro percebido entre sentenças. **Métodos:** para os diferentes processamentos pretendidos, foram desenvolvidos códigos em Python, usando as interfaces Spyder e pacotes tais como o *pydub*, *soundfile*, *os* e *numpy*. Os códigos foram testados em dois conjuntos de arquivos de áudio gravados previamente em português brasileiro, nos formatos .MP3 e .WAV. **Resultados:** foram implementados códigos para modificação do formato dos arquivos, ajuste de *fade-in* e *fade-out*, filtragem de passa-alta, vocoderização opcional e ajuste das médias quadráticas das amplitudes. Os testes dos códigos desenvolvidos em dois conjuntos de sentenças disponíveis em .WAV e .MP3 na língua portuguesa demonstraram resultados consistentes com o esperado. **Conclusão:** desenvolveram-se códigos na linguagem Python para modificação de maneira automatizada de arquivos de áudio, disponíveis no *site* GitHub para adaptações e aprimoramentos por terceiros.

Palavras-chave: Audiologia; Processamento de sinais assistido por computador; Testes auditivos; Implante coclear; Simulação por computador

ABSTRACT

Purpose: To develop computational codes for the automated modification of large quantities of sentence recordings, capable of performing format modifications, filtering, simulating sound signal processing in cochlear implants, and adjusting root mean square amplitude to equalize perceived volume between sentences. **Methods:** Python codes were developed for the intended processes, using the Spyder interface and packages such as pydub, soundfile, os, and numpy. The codes were tested on two sets of previously recorded audio files in Brazilian Portuguese, in .MP3 and .WAV formats. **Results:** Codes were implemented for 1) file format modification, 2) fade-in and fade-out adjustment, 3) high-pass filtering, 4) optional vocoderization, and 5) adjustment of root mean square amplitude. Testing the developed codes on two sets of sentence recordings available in .WAV and .MP3 formats in Portuguese showed consistent results as expected. **Conclusion:** Python codes were developed for the automated modification of audio files, available on the GitHub website for further adaptations and improvements by third parties.

Keywords: Audiology; Signal processing computer-assisted; Hearing tests; Cochlear implantation; Computer simulation

Trabalho realizado no Laboratório de Vibrações e Acústica – LVA, Universidade Federal de Santa Catarina – UFSC – Florianópolis (SC), Brasil.
[1]Departamento de Fonoaudiologia, Universidade Federal de Santa Catarina – UFSC – Florianópolis (SC), Brasil.
[2]Laboratório de Vibrações e Acústica – LVA, Departamento de Engenharia Mecânica, Universidade Federal de Santa Catarina – UFSC – Florianópolis (SC), Brasil.
[3]Departamento de Engenharia Mecânica, Universidade Federal de Santa Catarina – UFSC – Florianópolis (SC), Brasil.
Conflito de interesses: A Cochlear® forneceu apoio financeiro para o desenvolvimento deste estudo apenas para a autora Kauana Knappmann.
Contribuição dos autores: KK e SP foram responsáveis pela concepção e delineamento do estudo; KK foi responsável pela execução e análise dos dados; SP foi responsável pela coordenação geral, revisão crítica e edição final do manuscrito.
Declaração de Disponibilidade de Dados:
Os dados de pesquisa não estão disponíveis.
Financiamento: Cochlear Latinoamerica S.A.
Autor correspondente: Kauana Knappmann. E-mail: kauanak21@gmail.com
Recebido: Novembro 12, 2025; **Aceito:** Abril 01, 2026
Editor-Chefe: Maria Cecilia Martinelli Iorio.
Editor Associado: Liliane Desgualdo Pereira.

INTRODUÇÃO

A perda auditiva compromete a comunicação dos indivíduos, a interação social e a qualidade de vida[1]. Estima-se que até o ano de 2050 aproximadamente 2,5 bilhões de pessoas apresentarão hipoacusia, das quais 700 milhões poderão necessitar de algum tipo de tratamento[2], o que, por sua vez, requer instrumentos clínicos capazes de avaliar o desempenho auditivo de forma funcional.

A avaliação de reconhecimento de fala assume um papel central na prática audiológica, permitindo estimar a capacidade do indivíduo de compreender a fala em situações próximas das demandas cotidianas, considerando aspectos espectrotemporais e linguísticos[3]. É aplicada em diferentes populações clínicas, incluindo indivíduos com perda auditiva de variados graus, usuários de aparelhos de amplificação sonora individual (AASI) e de candidatos e usuários de implante coclear (IC)[4].

Dependendo do público-alvo (adultos, crianças, usuários de AASI/IC ou não), há diferentes formas de avaliar o reconhecimento de fala, podendo envolver não somente variações nos procedimentos de avaliação e no material de fala (palavras isoladas ou sentenças), presença de ruído competitivo, número de falantes, balanceamento fonético e procedimentos de validação, mas também formas específicas de preparação das amostras de áudio da fala[5]. Por exemplo, no caso específico da avaliação do reconhecimento de fala de usuários de IC, deve-se considerar que o processamento de sinal acústico no sistema auditivo com IC difere da audição acústica natural, já que, no IC, o estímulo sonoro é convertido em sinais elétricos distribuídos em número limitado de canais, resultando em redução da resolução espectral e alterações temporais significativas[6]. Em razão da degradação intrínseca ao processamento elétrico, usuários de IC podem apresentar efeito teto quando expostos a materiais com alta previsibilidade linguística, limitando a sensibilidade do teste[7]. A degradação intrínseca pode ser simulada usando vocoder[8]. Em indivíduos com audição normal, a avaliação de reconhecimento de fala envolve principalmente o uso de ruídos competitivos, sem a necessidade de simular a degradação coclear[5].

No Brasil, de acordo com pesquisas, os principais testes utilizados incluem o *Hearing in Noise Test* (HINT), adaptado para a língua portuguesa, as Listas de Sentenças em Português e ainda as Listas de Sentenças do Centro de Pesquisas Audiológicas (CPA). Embora representem avanços significativos na padronização nacional, cada um desses instrumentos apresenta características próprias, com vantagens e limitações específicas[9].

O *Hearing in Noise Test* (HINT)[10] foi desenvolvido para avaliar a inteligibilidade de sentenças em presença de ruído competitivo, permitindo estimar limiares de reconhecimento sob condições controladas. Sua principal vantagem está na padronização do procedimento e na ampla aplicabilidade clínica. Entretanto, o custo do material é alto para o uso clínico e o teste foi validado inicialmente para pessoas com audição normal.

O teste Listas de Sentenças em Português[11] é amplamente utilizado na prática clínica brasileira e diversos estudos reconhecem sua aplicabilidade e relevância clínica. Entretanto, está disponível somente em formato *compact disc* e sua aplicação exige que o próprio fonoaudiólogo ajuste manualmente os níveis de fala e ruído durante o procedimento, o que pode introduzir variabilidade nas condições de teste.

As Listas de Sentenças do Centro de Pesquisas Audiológicas (CPA)[12] também são bastante usadas na prática clínica nacional, especialmente em serviços de IC, apresentando organização estruturada e uso consolidado. O material é composto por sentenças foneticamente balanceadas, mas conforme aponta a literatura, sua aplicação ocorre predominantemente de forma falada, não havendo uniformidade nas condutas de aplicação entre os diferentes serviços[9].

No cenário internacional, diversos instrumentos são utilizados na avaliação do reconhecimento de fala, incluindo o HINT em sua versão original, o *Bamford-Kowal-Bench Speech-in-Noise Test* (BKB-SIN), o *Quick Speech-in-Noise* (QuickSIN) e o *Words-in-Noise* (WIN), tanto para avaliação em indivíduos com audição normal, quanto com perda auditiva[13]. Ainda, para avaliação de indivíduos com perda auditiva severa a profunda e usuários de IC, foi desenvolvido o *AzBio Sentence Test*[14]. O teste utiliza múltiplos locutores com frases pouco previsíveis, organizadas em listas estatisticamente equivalentes. Estudos demonstram maior sensibilidade para discriminar níveis de desempenho em adultos implantados quando comparado ao HINT, reduzindo a ocorrência do efeito teto[15]. Testes já validados e utilizados em diversos países de línguas diferentes[16] são boas opções para adaptação ao português do Brasil, podendo oferecer avaliação precisa e confiável, preenchendo uma lacuna que existe nos testes validados no Brasil.

De qualquer modo, a validade e a confiabilidade dos testes de percepção de fala na prática audiológica dependem do rigor metodológico empregado na construção e padronização dos estímulos acústicos utilizados, fazendo com que a gravação e a manipulação desses estímulos no processo de desenvolvimento de testes de reconhecimento de fala sejam obrigatoriamente realizadas de forma criteriosa e padronizada em grandes números de arquivos de áudio[14,17]. Modificações compreendem, tipicamente, filtragem de frequências indesejadas, equalização de volume, remoção ou inserção de pausas, suavização de transições, entre outras[18]. Ainda, visando ao desenvolvimento de testes de reconhecimento de fala para usuários de implante coclear, pode ainda ser empregada a utilização de vocoder como ferramenta de simulação do processamento de sinal do dispositivo, replicando a transformação de sinal acústico em estimulação elétrica[19].

Tendo em vista a complexidade dos diversos processamentos necessários em grande número de arquivos de áudio, especialmente durante o processo de desenvolvimento de testes, fica evidente que as etapas de processamento não só precisam ser realizadas de forma criteriosa e padronizada, mas também de forma automatizada. Tal processamento é possível utilizando-se ferramentas de processamento digital de sinais, com a devida automatização de processos.

Diante dessa necessidade, e para dar suporte ao desenvolvimento de novos testes de reconhecimento de fala no Brasil, incluindo versões brasileiras de testes já consagrados, foi desenvolvido um conjunto de códigos e rotinas computacionais de processamento de sinais que automatizam diversas tarefas necessárias para a manipulação de material de áudio (sentenças) já gravadas. Embora criado inicialmente para a manipulação de gravações de sentenças de um teste específico, o conjunto de códigos pode ser aplicado em outras propostas de testes de reconhecimento de fala.

PRINCÍPIOS DE PROCESSAMENTO DIGITAL DE SINAIS VISANDO À PADRONIZAÇÃO DE ESTÍMULOS DE FALA

Com a transição da tecnologia analógica para o ambiente computacional e, portanto, digital, o Processamento Digital de Sinais (PDS) tornou-se uma ferramenta útil e indispensável para análise e manipulação de sinais dos mais variados tipos, incluindo sinais de áudio e sinais de fala[20]. Os conceitos fundamentais de PDS no contexto do presente trabalho são apresentados na sequência.

Formatos de áudio

Um formato de áudio digital é, essencialmente, o método ou contêiner utilizado para codificar e armazenar o sinal sonoro em formato digital, sendo que os diferentes formatos disponíveis determinam tanto a fidelidade dessa representação, como opções de múltiplos canais, espacialização, etc. Os formatos mais comuns para sinais de áudio são atualmente .WAV e MP3. Enquanto o formato .WAV (*waveform Audio File Format*) armazena a representação digital mais exata possível da onda sonora original, o .MP3 é um formato de compressão "com perdas" (*lossy*), que emprega algoritmos que descartam informações espectrais consideradas inaudíveis pelo sistema auditivo humano para reduzir o tamanho do arquivo[21].

Filtragem digital e o filtro passa-alta Butterworth

A filtragem digital é uma operação (matemática) que modifica o conteúdo de frequências (espectro) de um sinal, permitindo a passagem de determinadas faixas de frequência, enquanto atenua ou bloqueia outras. No contexto da manipulação de sinais de áudio, a filtragem pode ser utilizada para remoção de ruídos indesejados de baixa frequência, muito comumente presentes em gravações devido a vibrações estruturais, o impacto do fluxo de ar no microfone ou interferência de rede elétrica[22]. Dependendo da faixa de frequência a ser removida e da faixa de frequência a ser preservada, podem ser empregadas diferentes tipologias de filtros (filtros passa-baixa, passa-alta, rejeita-faixa ou atenua-faixa). Para remoção de conteúdo espectral de baixa frequência em sinais de áudio, utiliza-se um "filtro passa-alta", sendo o tipo mais comum o filtro de Butterworth. Caracterizado por resposta de frequência "maximamente plana" na banda de passagem, não cria distorções espectrais e, portanto, alterações do timbre na parte do espectro preservado. A transição entre a faixa de frequência rejeitada e a faixa de frequência preservada é governada pela "ordem" do filtro, sendo que filtros com ordens mais elevadas resultam em filtragens seletivas. Entretanto, um aumento na ordem do filtro eleva o custo computacional do processamento e pode gerar outros problemas, como o efeito de *ringing* (efeito residual audível no áudio), que ocorre devido a uma resposta ao impulso demasiadamente oscilatória. Portanto, a escolha da ordem exige um compromisso adequado entre a eficácia na remoção do ruído e a preservação das características temporais do estímulo[20].

Suavização de transientes - *Fade-in* e *Fade-out*

A gravação e a segmentação de sinais de fala frequentemente introduzem pequenas descontinuidades no domínio do tempo, especificamente nos pontos de corte nos quais a amplitude do sinal, ainda que muito pequena e quase inaudível, difere de zero. Durante a reprodução, essa transição não perfeita de um valor não nulo, ainda que muito pequeno, a zero, gera um transiente espectral de banda larga, audível como um clique indesejado. Para mitigar esse artefato, aplicam-se envoltórias de *fade-in* e *fade-out*, por meio da multiplicação das amostras nas extremidades do sinal por uma função suave de janelamento, que modula, de forma suave, a amplitude de zero até o valor nominal da onda (*fade-in*) e vice-versa (*fade-out*). Com janelas temporais da ordem de milissegundos, a técnica assegura uma transição suave sem artefatos audíveis[23].

Normalização de amplitude - ajuste RMS (*Root Mean Square* ou Valor Quadrático Médio) em dBFS (*Decibels Full Scale*)

Para que os itens (sentenças) de um teste de fala sejam clinicamente comparáveis, também precisam ser calibrados para proporcionarem a mesma sensação de volume sonoro ao ouvinte. Para sinais com espectros similares, o que se aplica a sinais de sentenças, a padronização pode ser realizada pela equalização do valor RMS da variável dependente, que reflete a energia média do sinal ao longo do tempo[24]. Tendo em vista o uso de formatos digitais .WAV ou .MP3, o ajuste do valor RMS é realizado em uma escala digital chamada dBFS[1]. Cabe notar que o nível de pressão sonora (SPL) de reprodução, em termos de dB SPL (nível de pressão sonora em decibéis), obtido na reprodução do sinal por meio de alto-falantes ou fones de ouvido dependerá do *software* e dos equipamentos utilizados para reprodução. Entretanto, assumindo linearidade e invariância no tempo desse sistema de reprodução, o fato de existirem arquivos sonoros com a mesma amplitude média quadrática expressa na escala dBFS contribui para assegurar que, na reprodução, o nível de pressão sonora de reprodução dB SPL também tenha consistência relativa entre os estímulos[17,20].

Vocoder

No contexto da simulação acústica da audição por implante coclear (IC), vocoders, como o AngelSim[25], operam fundamentalmente como modelos de degradação espectral estritamente fundamentados em etapas sucessivas de processamento digital de sinais. Inicialmente, o sinal de entrada é submetido a um banco, ou conjunto, de filtros passa-banda, partindo o espectro de frequências em um número restrito de canais de análise, o que mimetiza a resolução espectral discretizada imposta pela matriz de eletrodos intracocleares. Em cada canal analítico, a

[1] O termo dBFS não é usualmente traduzido para o português, sendo empregado na forma original (decibels relative to full scale). A tradução técnica mais precisa seria "decibels em relação ao fundo de escala (digital)". Em instrumentação, "fundo de escala" corresponde ao valor máximo que um sistema ou uma escala pode representar antes de ocorrer saturação (clipping). Diferentemente de escalas como dB SPL, o dBFS tem como referência o nível máximo digital, definido como 0 dBFS; todos os demais valores são expressos em números negativos (por exemplo, –18 dBFS).

envoltória temporal da amplitude é extraída — tipicamente através de retificação (meia onda ou onda completa), seguida de filtragem passa-baixa, emulando a taxa de estimulação pulsátil transmitida à interface eletroneural. Essa envoltória de baixa frequência é então utilizada para modular em amplitude um sinal portador sintetizado, que pode ser ruído de banda estreita ou tom puro. A superposição de todos os canais modulados sintetiza um estímulo acústico final, caracterizado pelo severo borramento espectral (*spectral smearing*) e pela restrição de pistas de estrutura fina (*fine structure*). Consequentemente, a preservação prioritária das pistas temporais grosseiras força o sistema auditivo central do ouvinte a recrutar os mesmos mecanismos neurocognitivos *top-down* de fechamento perceptual exigidos de um usuário de IC para a extração e a decodificação da fala[26].

MÉTODOS

Materiais utilizados

O material de áudio gravado em português brasileiro disponível para o trabalho foi composto por dois conjuntos distintos de sentenças, destinados a diferentes propostas de testes de reconhecimento de fala. Por terem sido gravados por terceiros, não houve acesso à documentação referente aos critérios linguísticos de construção das listas e não houve interação direta com os participantes das gravações, dispensando a necessidade de aprovação pelo Comitê de Ética em Pesquisa com Seres Humanos (CEPSH) da instituição e também da assinatura do Termo de Consentimento Livre e Esclarecido (TCLE). Embora tivessem origem e propósitos distintos, ambos os materiais apresentaram inconsistências técnicas típicas, como grandes variações de volume e ruído de fundo e, portanto, foram considerados excelentes materiais para constituírem objeto de tratamento pelas rotinas de processamento de sinal desenvolvidas neste trabalho.

O conjunto A de sentenças foi composto por 987 sentenças gravadas em um estúdio profissional por terceiros. O material foi gravado por quatro locutores (dois do gênero masculino e dois do gênero feminino). O primeiro falante masculino gravou 248 sentenças, enquanto o segundo gravou 231. Em relação às falantes femininas, a primeira registrou 257 sentenças e a segunda, 251. As gravações foram disponibilizadas para este estudo em formato .MP3, com taxa de amostragem de 44.100 Hz.

O conjunto B consistiu em 106 gravações de sentenças para a construção de uma avaliação da percepção de fala para crianças falantes do português brasileiro, pronunciadas por somente uma falante feminina. Essas gravações estavam disponíveis em formato .WAV, com taxa de amostragem de 44.100 Hz.

Processamento de sinal aplicado ao estudo

Considerando o grande número de arquivos existentes, as etapas de manipulação acústica foram realizadas por meio de rotinas computacionais automatizadas, na linguagem de programação Python (versão 3.10), linguagem de alto nível, utilizada para aplicações científicas e de processamento de sinais. O desenvolvimento foi conduzido no ambiente Spyder (distribuição Anaconda).

Utilizaram-se várias bibliotecas e pacotes para realização de funções específicas voltadas principalmente para modificação de arquivos de áudio. As bibliotecas são coleções de módulos que implementam várias funções e métodos pré-definidos, além de classes, para executar tarefas específicas de forma eficiente. Permitem o uso de códigos já existentes, simplificando processos[27].

A título de exemplo, as bibliotecas *soundfile* e *pydub* foram usadas para a manipulação apropriada dos arquivos de áudio, enquanto as bibliotecas *os* e *numpy* serviram para a gestão e a manipulação dos caminhos dos arquivos e *arrays* numéricos, respectivamente. Já a biblioteca *scipy* foi utilizada para funcionalidades estatísticas, auxiliando a análise e o processamento dos dados sonoros. Para facilitar o acesso de outros pesquisadores da área, o código-fonte foi disponibilizado integralmente em um repositório público da plataforma GitHub.

As etapas de processamento dos arquivos foram configuradas da seguinte forma:

1. Padronização de formato: de início, os arquivos .MP3 foram convertidos para o formato .WAV, garantindo que todos os arquivos tivessem as mesmas características iniciais em termos de formato de áudio;

2. Suavização temporal: aplicou-se uma rampa de *fade-in* e *fade-out* com duração de 1 ms cada e envoltórias de meia onda de cosseno, buscando eliminar artefatos audíveis na transição entre sentenças, na reprodução;

3. Filtragem passa-alta: para eliminar ruídos ambientais de baixa frequência que contaminaram as gravações, implementou-se um filtro do tipo Butterworth de 5ª ordem, com frequência de corte em 80 Hz. A frequência de corte de 80 Hz elimina os ruídos mais evidentes nas gravações, incluindo o ruído da rede elétrica na frequência de 60 Hz, e preserva, ao mesmo tempo, as frequências mais baixas da voz masculina (frequência fundamental perto de 110 Hz)[20,28];

4. Simulação de implante coclear (vocoder): para simular a percepção auditiva de usuários de implante coclear, utilizou-se o processamento *batch* de um vocoder (AngelSim™). No caso deste estudo, os parâmetros utilizados no AngelSim foram: onda portadora senoidal; 3 e 5 canais espectrais/3 e 5 canais estimulados; mapeamento de Greenwood[29]; banco de filtros de análise 200 Hz a 7000 Hz e inclinação em 24 dB; frequência de corte inferior das envoltórias 160 Hz, filtro passa-alta com *slope* de 24 dB/oitava. Os parâmetros de síntese mantiveram correspondência com os valores definidos na etapa de análise, para garantir coerência entre extração de envoltória e reconstrução do sinal;

5. Equalização de amplitude: as médias quadráticas de amplitude (*Root Mean Square* - RMS) de todas as sentenças foram normalizadas para um valor-alvo de -20 dBFS. O processo de ajuste dos valores RMS compreende várias etapas: carregamento dos arquivos, cálculo do RMS atual, definição do valor de RMS desejado, cálculo do fator de escala para ajustar o RMS ao valor desejado e aplicação da correção às amplitudes do sinal. Embora os dados de áudio já estivessem representados em ponto flutuante nesse estágio do fluxo de processamento, a conversão para ponto flutuante foi realizada quando necessário, especialmente em arquivos originalmente

codificados em formato *integer*, a fim de garantir maior precisão nas operações matemáticas envolvidas no ajuste do RMS. Após a normalização, foi inserido um intervalo padronizado de silêncio no início e no fim dos arquivos. Essa etapa foi realizada posteriormente ao ajuste do RMS para evitar interferência no nível normalizado, considerando que os silêncios originais, de duração variável, haviam sido previamente removidos.

Aferição acústica

Por fim, vários arquivos sonoros ajustados pelos códigos foram reproduzidos por um sistema de reprodução sonora com caixas de som GENELEC e tiveram seus níveis de pressão sonora aferidos por meio de um sistema de medição SQuadriga II, com microfone capacitivo GRAS 46AQ. O sistema de medição foi previamente calibrado com o calibrador Larson Davis CAL200.

RESULTADOS

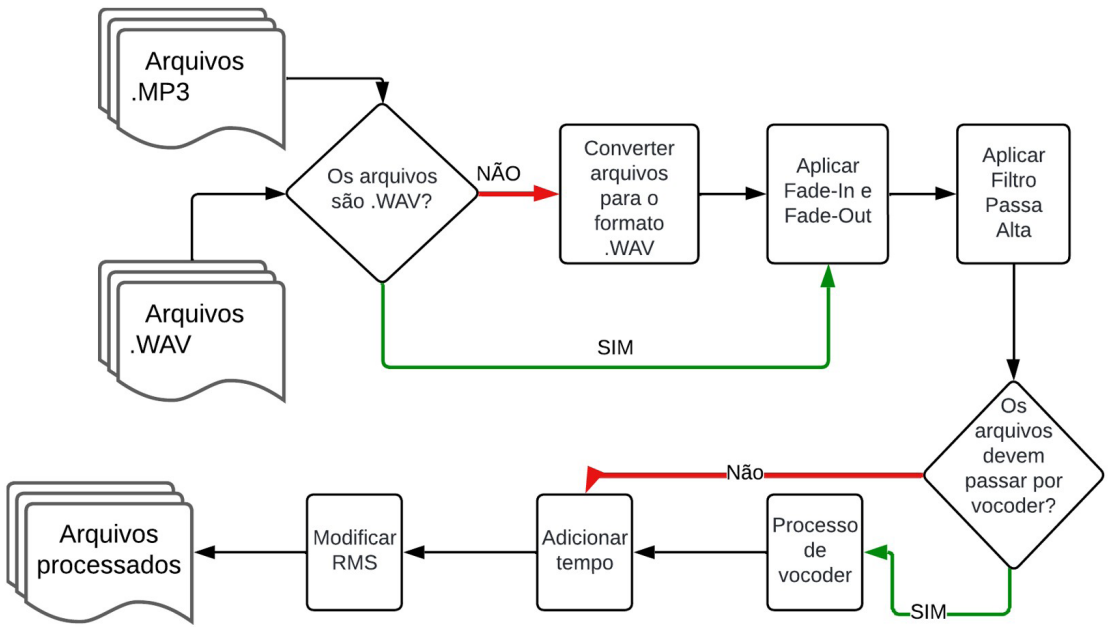
Todos os processamentos de sinais necessários, alcançados com os códigos implementados, para fins de validação, foram organizados em diferentes etapas, conforme mostra o esquemático da Figura 1. Algumas das etapas foram executadas de forma obrigatória, enquanto que outras etapas foram facultativas, como, por exemplo, a vocoderização, que só é necessária caso se queira simular o efeito de um implante coclear.

Arquivos de áudio em formato de áudio .MP3 foram convertidos, inicialmente, por meio de códigos para o formato .WAV, mesmo que as informações removidas durante a compressão .MP3 não pudessem ser recuperadas ao converter o arquivo para .WAV.

No que concerne à aplicação do *fade-in* e do *fade-out*, a Figura 2 painel A, representa o histórico no tempo de um arquivo-exemplo pertencente ao conjunto A, antes e depois do processamento, marcando a diferença entre a cor cinza para a sentença antes de qualquer modificação e a cor azul, que representa a sentença após a aplicação do *fade-in*, do *fade-out* e do filtro passa-alta. As diferenças observadas foram bastante relevantes, evidenciando que o *fade-in* e o *fade-out* atuam sobre as amplitudes do início e do fim do sinal de áudio.

O processo de filtragem por meio do filtro passa-alta pode ser evidenciado da melhor forma por meio da comparação dos espectrogramas de magnitude do sinal de áudio antes (Figura 3 painel A) e depois (Figura 3 painel B) da aplicação do filtro. Verifica-se, pelo espectrograma de magnitude da Figura 3 painel A, que o arquivo-exemplo original estava contaminado com energia de frequências baixas (<80 Hz), e que essa energia de baixa frequência esteve presente, inclusive, antes e depois do sinal da fala, demonstrando claramente que se tratava de ruído não relacionado ao sinal de fala. O item que contém a fala está indicado por uma caixa vermelha. Já o espectrograma da Figura 3 painel B, mostra que o filtro removeu de forma satisfatória a contaminação em baixa frequência, preservando as frequências do sinal de fala.

A vocoderização, opcional, foi realizada com diferentes configurações do vocoder, conforme descrito na seção "Processamento de sinal aplicado ao estudo". Criou modificações consideráveis no espectro e na definição temporal dos sons, aproximando-se da maneira que um implante coclear apresenta os sons ao usuário desse dispositivo. Na Figura 4 painel A e Figura 4 painel B apresenta-se, por meio do histórico no tempo, um sinal-exemplo e a sua modificação após o processamento do vocoder no arquivo. A Figura 3 painel C mostra o espectrograma do sinal de áudio depois da passagem pelo vocoder, com as modificações feitas, evidenciando as mudanças no sinal em comparação ao espectrograma apresentado na Figura 3 painel



Legenda: RMS = *Root Mean Square* **(ou Valor Quadrático Médio).**
Fonte: Elaborada pelo autor, 2023
Figura 1. Fluxograma do processo de ajuste de *Root Mean Square* em sinais de áudio

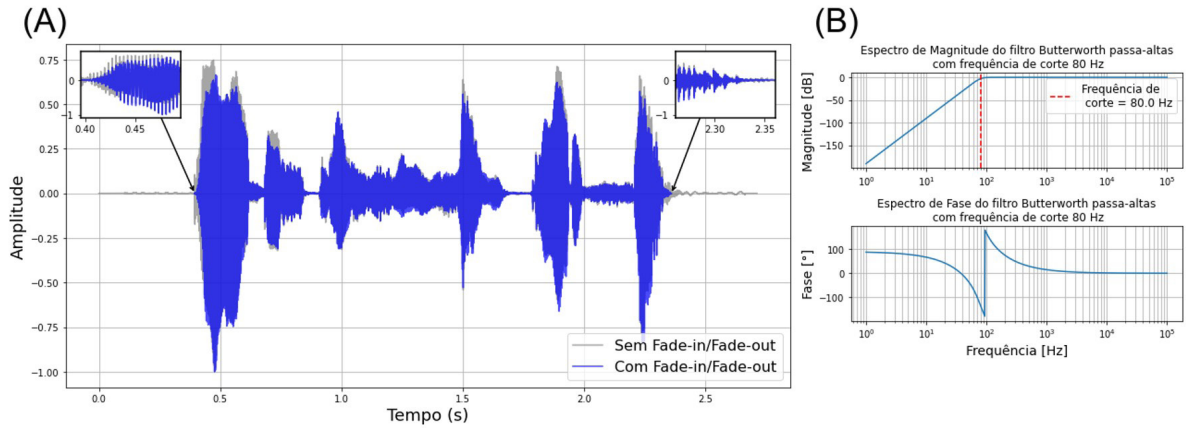


Figura 2. Pré-processamento dos sinais de áudio. (A) Histórico no tempo do sinal sonoro antes e depois da aplicação do *fade-in*/*fade-out*, remoção de silêncio e filtro passa-alta; (B) Espectro de magnitude e fase da função resposta em frequência do filtro passa-alta Butterworth, com frequência de corte em 80 Hz
Fonte: Elaborada pelo autor, 2023-2024

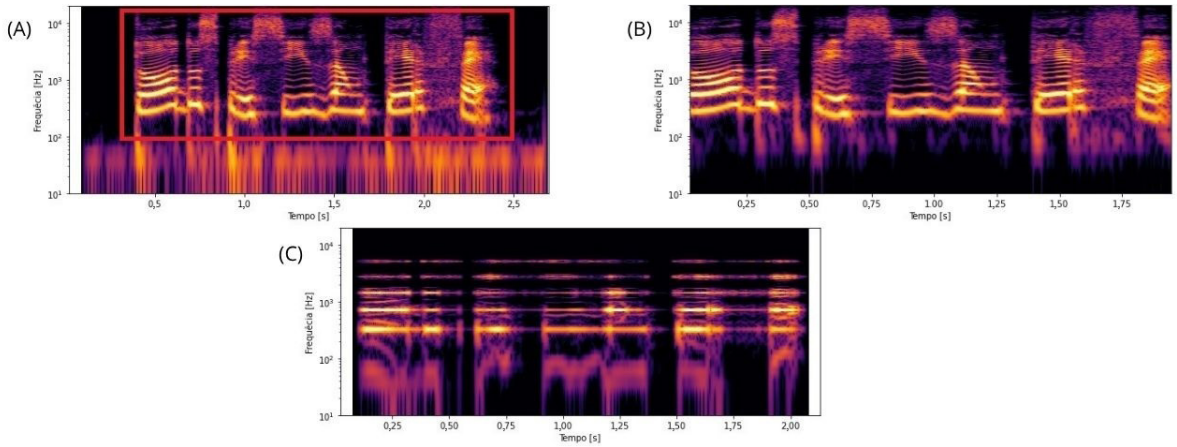


Figura 3. Espectrogramas dos sinais sonoros em diferentes etapas. (A) Espectrograma do sinal sonoro antes da aplicação do *fade-in*/*fade-out*, remoção de silêncio e filtro passa-alta; (B) Espectrograma do sinal sonoro após aplicação do *fade-in*/*fade-out*, remoção de silêncio e filtro passa-alta; (C) Espectrograma do sinal sonoro após processamento com vocoder tonal de cinco canais
Fonte: Elaborada pelo autor, 2023

Figura 4. Histórico temporal dos sinais de áudio. (A) Registro temporal de um arquivo de áudio antes do processo de vocoderização tonal de cinco canais; (B) Registro temporal do mesmo arquivo após o processo de vocoderização tonal de cinco canais
Fonte: Elaborada pelo autor, 2023

B, principalmente a perda de resolução em frequência típica em sinais processados por implantes cocleares.

Para avaliar o funcionamento correto do código que ajusta a média quadrática das amplitudes (ajuste RMS) para um nível alvo, a fim de garantir que todas as sentenças (arquivos de áudio) evocassem a mesma sensação de volume sonoro, foram processados os dois conjuntos de sentenças do português brasileiro disponíveis para teste dos códigos. O conjunto A foi processado tanto sem passar por vocoderização, como após passagem pelo processo de vocoderização. Já o conjunto B foi processado apenas sem vocoderização. A média quadrática de todos os conjuntos foi ajustada para um nível-alvo de -20 dBFS. A Figura 5 painel A representa, em azul, a distribuição das médias quadráticas das amplitudes do conjunto de 987 arquivos do conjunto A de sentenças antes do ajuste do RMS, por meio do código implementado. Verificou-se uma grande variação das médias quadráticas das amplitudes, o que causaria grandes diferenças na percepção do volume sonoro e possivelmente no reconhecimento das diferentes sentenças.

O ajuste das médias quadráticas das amplitudes das sentenças do conjunto A a, valor-alvo de -20 dBFS, sem a aplicação do vocoder, resultou em médias quadráticas das amplitudes com média e mediana igual a -20 dBFS, com desvio padrão de apenas 0,01 dBFS, conforme mostra a Figura 5 painel A na cor verde.

Já para o mesmo conjunto das sentenças A, após aplicação do vocoder, o ajuste automático para um nível-alvo de -20dBFS resultou em média e mediana das amplitudes ajustadas de -20,34 dBFS e -20,31 dBFS, respectivamente, com desvio padrão de 0,12 dBFS e diferença máxima de 0,8 dBFS, conforme se verifica na Figura 5 painel A, na cor cinza. Nota-se, portanto, uma variação um pouco maior em sentenças que passaram pelo vocoder, ainda que inferior a 1 dB.

O conjunto B de arquivos também foi processado por todos os códigos em Python, com exceção da etapa da vocoderização. Todas as sentenças desse conjunto também tiveram seus níveis de média quadrática de amplitudes ajustadas pelos códigos a um valor-alvo de -20 dBFS. A Figura 5 painel B representa um histograma dos níveis de média quadrática das amplitudes da pasta de 106 arquivos sonoros desse conjunto antes da aplicação do ajuste, na cor azul, evidenciando que os níveis médios quadráticos variaram entre -28 dBFS e -14 dBFS, ainda que a maioria dos arquivos tivesse níveis médios quadráticos entre -20 dBFS e -14 dBFS. A Figura 5 painel B, na cor verde, mostra os níveis médios quadráticos da mesma seleção de arquivos após a modificação dos níveis médios quadráticos a um valor-alvo de -20 dBFS. O histograma evidencia que o ajuste foi muito exato, com a média dos valores RMS em -20 dBFS, com mediana também de -20 dBFS e desvio padrão de 0,04 dBFS.

Ao finalizar todas as etapas de processamento, todos os arquivos apresentaram-se modificados. Foram salvos em formato.WAV com taxa de amostragem em 44.100 Hz e RMS muito próximos ou iguais a -20 dBFS.

Por fim, a verificação do volume sonoro das sentenças, quando reproduzidas em ambiente controlado, mostrou que os níveis de pressão sonora a um metro de distância da fonte sonora foram de entre 65,4 e 66,4 dB SPL, variação inferior ao JND (*Just Noticeable Difference*) aferido na prática para o nível de pressão sonora.

DISCUSSÃO

Os códigos de processamento de sinais de fala desenvolvidos em Python foram avaliados frente a diferentes aspectos de qualidade, fundamentais para a validação de testes de reconhecimento de fala. Verificou-se que a aplicação automatizada de *fade-in* e *fade-out,* exemplificada na Figura 2 painel A, e sugerida na literatura proporcionou transições suaves de sons nas extremidades do sinal sonoro, minimizando descontinuidades abruptas e audíveis, em todos os sinais processados[20,21].

A aplicação do filtro passa-alta Butterworth de 5ª ordem com frequência de corte de 80 Hz mostrou-se eficaz na remoção de componentes espectrais indesejados de baixa frequência, incluindo interferências associadas à rede elétrica. A escolha da frequência de corte fundamentou-se na distribuição espectral típica da fala, considerando que a frequência fundamental média da voz masculina situa-se em torno de 110 Hz, na frequência

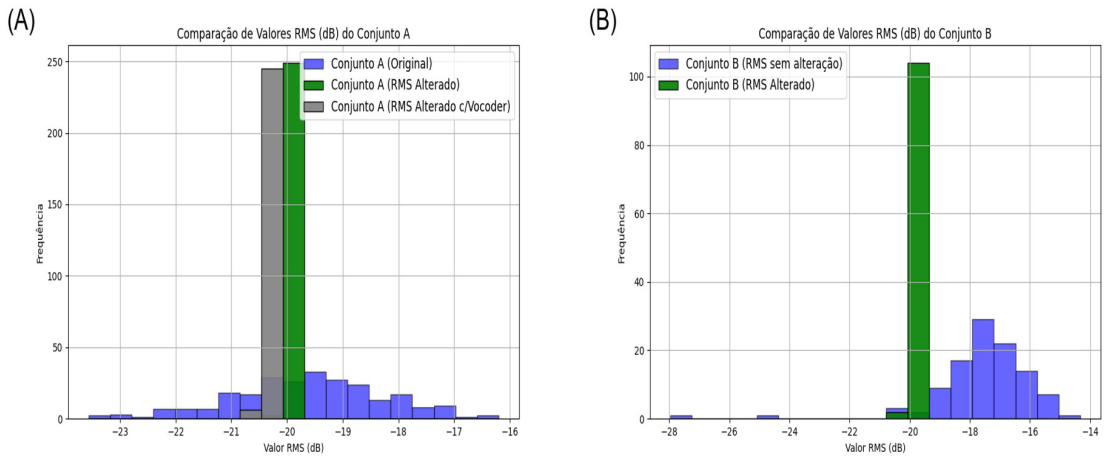


Figura 5. Distribuição dos valores *Root Mean Square*. (A) Histograma da distribuição de valores *Root Mean Square* do Conjunto A, antes e depois dos processos (ajuste de *Root Mean Square* e vocoder); (B) Histograma da distribuição dos níveis de *Root Mean Square* do Conjunto B, antes e depois dos processos

Legenda: RMS = *Root Mean Square* **(ou Valor Quadrático Médio); dB = decibel.**

Fonte: Elaborada pelo autor, 2023

da rede elétrica (60 Hz) e nas frequências de ruídos acústicos observados nas gravações originais, de modo que atenuações abaixo de 80 Hz removem os ruídos indesejados sem comprometer o conteúdo linguístico[20,22,29].

O ajuste de valores RMS para um valor-alvo (-20 dBFS) resultou em elevada homogeneidade dos valores RMS das sentenças, com variações inferiores a 1 dBFS, mesmo após a etapa da vocoderização. Considerando o limiar de discriminação de sensação de volume sonoro de 1 dB[30], os resultados indicaram controle efetivo da variável "nível de apresentação", fator potencialmente influente no desempenho em testes de reconhecimento de fala. O discreto aumento da variabilidade após a vocoderização era esperado, em razão da redistribuição espectral de energia promovida pelo processamento por canais, mas não ultrapassou limites clinicamente relevantes[10,24].

A vocoderização das sentenças, realizada com processamento *batch* em *software* externo (AngelSim) com três e cinco canais espectrais, produziu as alterações espectrotemporais compatíveis com a descrição clássica de simulação de implante coclear, em que o menor número de canais reduz a resolução espectral e aumenta a dependência de pistas temporais. Efeitos de borramento espectral e restrição da estrutura fina, observados nos gráficos apresentados, confirmaram o que era esperado para vocoders[19,26]. Essa simulação controlada permite investigar o impacto de diferentes configurações de processamento antes de aplicações clínicas.

A aferição acústica em campo livre demonstrou variação do nível de pressão sonora inferior a 1 dB entre estímulos, evidenciando que a padronização digital se refletiu de maneira consistente no ambiente físico[10,24].

Além do rigor técnico, destaca-se a viabilidade operacional do sistema, com processamento completo de 987 sentenças em, aproximadamente, 40 minutos. Essa automatização reduz erros humanos inerentes à edição manual de áudio e favorece reprodutibilidade metodológica. Futuramente, pretende-se desenvolver ainda uma interface gráfica (GUI) para facilitar o uso dos códigos por pesquisadores não familiarizados com a linguagem computacional Python, ampliando o acesso a ferramentas de processamento de fala baseadas em evidências.

CONCLUSÃO

Tendo em vista a necessidade de uma ferramenta computacional para rápida modificação automatizada de grandes quantidades de arquivos de áudio, em particular as gravações de sentenças, foram desenvolvidos códigos na linguagem computacional Python. Esses códigos não precisam ser utilizados exclusivamente da maneira apresentada neste estudo, mas podem ser adaptados para serem usados de acordo com as necessidades de cada aplicação.

Tais códigos estão disponíveis no *site* GitHub, com acesso aberto à comunidade científica e com a possibilidade de aprimoramento das técnicas apresentadas, nomeados de Codigos_Processamento_Sinal_de_Fala.

REFERÊNCIAS

1. Força MT, Cal RVR, Santos SR, Pereira LC, Boas GPV, Zell RGA, et al. Análise comparativa da percepção da perda auditiva com o resultado da audiometria em pacientes adultos e idosos do Hospital Bettina Ferro de Souza/PA. BJHR. 2020;3(6):17457-73. https://doi.org/10.34119/bjhrv3n6-162.
2. WHO: World Health Organization. World report on hearing [Internet]. Geneva: WHO; 2021 [citado em 2025 junho 25]. Disponível em: https://www.who.int/publications/i/item/9789240020481
3. Costa LD, Vaucher AVA, Pagliarin KC, Costa MJ. Teste de palavras no ruído: desenvolvimento, validação e valores de referência. CoDAS. 2024;36(3):e20230091. https://doi.org/10.1590/2317-1782/20242023091en. PMid:38836822.
4. Campos Salvato C, Araújo SRS, Muller R, Soares AD, Chiari BM. Correlação entre reconhecimento de fala, tempo de privação auditiva e tempo de uso de Implante Coclear em usuários com surdez pós-lingual. Distúrb Comun. 2020;32(3):396-405. https://doi.org/10.23925/2176-2724.2020v32i3p396-405.
5. Ferreira MC, Zamberlan-Amorim NE, Wolf AE, Reis ACMB. Influence of different types of noise on sentence recognition in normally hearing adults. Rev CEFAC. 2021;23(5):e2121. https://doi.org/10.1590/1982-0216/20212352121.
6. Wilson BS, Dorman MF. Cochlear implants: current designs and future possibilities. JRRD. 2008;45(5):695-730. https://doi.org/10.1682/JRRD.2007.10.0173. PMid:18816422.
7. Gifford RH, Shallop JK, Peterson AM. Speech recognition materials and ceiling effects: considerations for cochlear implant programs. Audiol Neurootol. 2008;13(3):193-205. https://doi.org/10.1159/000113510. PMid:18212519.
8. Johnson PA, McNamara DM, Ziarani AK. A novel VOCODER for cochlear implants. In: 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2008; Vancouver, BC. New York: IEEE; 2008. p. 4732-5.
9. Pinheiro MMC, Vieira MG, Vieira LM, Koerich I, Rosseto I, Lazzarotto-Volcão C, et al. Adaptação de listas de sentenças para avaliação da percepção da fala. CoDAS. 2022;34(1):e20200301. https://doi.org/10.1590/2317-1782/20202020301. PMid:35019063.
10. Sbompato AF, Corteletti LCBJ, Moret ADLM, Jacob RTDS. Hearing in Noise Test Brazil: standardization for young adults with normal hearing. Braz J Otorhinolaryngol. 2015;81(4):384-8. https://doi.org/10.1016/j.bjorl.2014.07.018. PMid:26130593.
11. Costa MJ. Listas de sentenças do português. Santa Maria: Pallotti; 1998. 48 p.
12. Valente SLO. Elaboração de listas de sentenças construídas na língua portuguesa [dissertação]. São Paulo: Pontifícia Universidade Católica; 1998.
13. Wilson RH, McArdle RA, Smith SL. An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss. J Speech Lang Hear Res. 2007;50(4):844-56. https://doi.org/10.1044/1092-4388(2007/059). PMid:17675590.
14. Spahr AJ, Dorman MF, Litvak LM, Van Wie S, Gifford RH, Loizou PC, et al. Development and validation of the AzBio sentence lists. Ear Hear. 2012;33(1):112-7. https://doi.org/10.1097/AUD.0b013e31822c2549. PMid:21829134.
15. Massa ST, Ruckenstein MJ. Comparing the performance plateau in adult cochlear implant patients using HINT and AzBio. Otol Neurotol. 2014;35(4):598-604. https://doi.org/10.1097/MAO.0000000000000264. PMid:24557031.
16. Jung C, Shah KV, Ferraro T, Bigelow DC, Ruckenstein MJ, Peng-Hwa T. Non-english validation of the azbio sentence test: a systematic review of current applications and early adoption patterns. Otol Neurotol.

2026;47(1):e8391. https://doi.org/10.1097/MAO.0000000000004663. PMid:41145396.

17. Nilsson M, Soli SD, Sullivan JA. Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. J Acoust Soc Am. 1994;95(2):1085-99. https://doi.org/10.1121/1.408469. PMid:8132902.
18. Upadhyay N, Karmakar A. A multi-band speech enhancement algorithm exploiting iterative processing for enhancement of single channel speech. J Signal Inf Process. 2013;4(2):197-211. https://doi.org/10.4236/jsip.2013.42027.
19. Karoui C, James C, Barone P, Bakhos D, Marx M, Macherey O. Searching for the sound of a cochlear implant: evaluation of different vocoder parameters by cochlear implant users with single-sided deafness. Trends Hear. 2019;23:2331216519866029. https://doi.org/10.1177/2331216519866029. PMid:31533581.
20. Välimäki V, Reiss J. All about audio equalization: solutions and frontiers. Appl Sci. 2016;6(5):129. https://doi.org/10.3390/app6050129.]
21. Raj VA, Dhas MDK. Analysis of audio signal using various transforms for enhanced audio processing. Int J Health Sci. 2022;6(2):8890-7.
22. Berglund B, Hassmén P, Job RFS. Sources and effects of low-frequency noise. J Acoust Soc Am. 1996;99(5):2985-3002. https://doi.org/10.1121/1.414863. PMid:8642114.
23. Lupşa-Tătaru L. Customizing audio fades with a view to real-time processing. Applied Computer Science. 2019;15(4):16-26. https://doi.org/10.35784/acs-2019-27.
24. Ricketts TA, Bentler RA. The effect of test signal type and bandwidth on the categorical scaling of loudness. J Acoust Soc Am. 1996;99(4):2281-7. https://doi.org/10.1121/1.415415. PMid:8730074.
25. Emily Shannon Fu Foundation. AngelSim (TigerCIS): Cochlear Implant and Hearing Loss Simulator [Internet]. 2013 [citado em 2025 junho 25]. Disponível em: http://angelsim.emilyfufoundation.org/
26. Cychosz M, Winn MB, Goupell MJ. How to vocode: uing channel vocoders for cochlear-implant research. J Acoust Soc Am. 2024;155(4):2407-37. https://doi.org/10.1121/10.0025274. PMid:38568143.
27. The Python Standard Library [Internet]. 2001 [citado em 2025 junho 25]. Disponível em: https://docs.python.org/3/library/index.html
28. Kent RD, Read C. Acoustic analysis of speech. 2nd ed. San Diego: Singular Publishing Group; 2002.
29. Greenwood DD. Auditory masking and the critical band. J Acoust Soc Am. 1961;33(4):484-502. https://doi.org/10.1121/1.1908699.
30. Forinash K, Christian W. Sound: an interactive ebook [Internet]. 2012 [citado em 2025 junho 25]. Disponível em: https://www.compadre.org/books/SoundBook