AutoIPA: development of an online platform to facilitate the use of automated phonetic transcription using Artificial Intelligence

Guilherme Brizzi[1], **Ana Lilian Alfonso Toledo**[1], **Felipe Crivellaro Minuzzi**[2]

[1]Curso de Ciência da Computação, Universidade Federal de Santa Maria (UFSM) - Santa Maria, RS - Brasil

[2]Departamento de Matemática, Universidade Federal de Santa Maria (UFSM) - Santa Maria, RS - Brasil

gbrizzi@inf.ufsm.br, analilianat2014@gmail.com, felipe.minuzzi@ufsm.br

Abstract. *Phonetic transcription is a useful tool for describing the variability inherent in language. However, this technique remains largely inaccessible to the general public. Against this backdrop, AutoIPA (autoipa.org) was developed: an AI-powered web platform designed to simplify access to and use of phonetic transcription in the International Phonetic Alphabet (IPA). In this vein, various pre-existing models were studied and evaluated to enable the creation of an application that automates phonetic transcription. Initially, a scarcity of labeled datasets was noted, which constrains the performance of current models. Moreover, it was observed that these tools are still not very accessible.*

Resumo. *A transcrição fonética é uma ferramenta útil para descrever a variedade inerente à linguagem. Contudo, essa técnica ainda é pouco acessível para o público geral. A partir desse paradigma, desenvolveu-se o AutoIPA (autoipa.org): uma plataforma web baseada em IA para facilitar o acesso e a utilização da transcrição fonética para o Alfabeto Fonético Internacional (IPA). Nessa toada, foram estudados e avaliados diferentes modelos pré-existentes para viabilizar o desenvolvimento de uma aplicação que automatiza a transcrição fonética. Preliminarmente, notou-se a escassez de bases de dados rotuladas, o que restringe a performance dos modelos existentes. Ademais, observou-se que as ferramentas ainda são pouco acessíveis.*

1. Introdução

A diversidade da linguagem humana é notável: estima-se que há, no mundo, em torno de sete mil línguas sendo faladas e, dentre os falantes de cada língua, há incontáveis diferentes sotaques, dialetos, sistemas de escrita, jargões e gírias [Ethnologue 2025]. Cada interlocutor molda a língua à sua maneira, criando sua forma de expressão própria.

Mesmo no Brasil, um país em que quase toda a população fala a mesma língua, nota-se uma grande variação nas maneiras de se expressar quando comparamos diferentes regiões, origem étnicas, socioeconômicas e geográficas. Essa variação, além de diferenças em vocabulário e gramática, se manifesta fortemente na fonética, ou seja, simplificando, na forma dos sotaques da língua [Bisol 2005].

É possível que um usuário da língua diferencie a origem de falantes, por mais que as palavras faladas sejam as mesmas. Essa diferenciação é feita pela análise - ainda que inconscientemente - da fonética do falante.

A fonética, um dos principais componentes da linguagem, se refere ao ramo da linguística que estuda a produção e a percepção de sons, além dos mecanismos que as possibilitam. [O'Grady 2005] Uma maneira de analisar com maior precisão a fonética é representá-la por meio da escrita: ou seja, realizando-se uma transcrição fonética.

De modo geral, a transcrição é o processo que recebe um texto em um meio, como a voz, e o passa para outro, como texto escrito no alfabeto latino, tal qual o deste artigo. [Peterson 2015]

Assim, a transcrição fonética é o processo de representar a fala de modo escrito. Porém, sistemas de escrita tradicionais, como o alfabeto latino, o alfabeto árabe ou *kanji* japoneses não são representações fiéis da fonética, ao passo que suas representações textuais não correspondem aos sons emitidos por um determinado ouvinte. Um exemplo simples para demonstrar isso no português é a palavra "jogo", que, mesmo escrita *ipsis litteris*, pode receber diferentes sons e significados, conforme seu uso, vide tabela 1.

Tabela 1. Diferença de pronúncia e significado da palavra "jogo"

Pronúncia	**Classe/significado**	**Exemplo**
Jogo com **"o" fechado**, "ô"	Substantivo: partida	"O jogo começou agora."
Jogo com **"o" aberto**, "ó"	Verbo: forma de jogar	"Eu jogo bola todo dia."

Percebe-se, então, que o alfabeto latino convencional não é uma forma viável para a representação fidedigna do som, já que possui ambiguidades em que um símbolo pode representar múltiplos sons. Em outros sistemas de escrita, é comum que a representação gráfica tenha ainda menor relação com a fonética e, em alguns casos, nenhuma relação.

A partir desse paradigma, linguistas criaram o Alfabeto Fonético Internacional (em inglês, *IPA - International Phonetic Alphabet*). O IPA visa a ser uma forma neutra de representação que representa puramente os sons emitidos, sem ambiguidade. O IPA é construído a partir do alfabeto latino com a adição de uma série de símbolos especiais. O exemplo abaixo mostra um exemplo de transcrição fonética com IPA. [Battisti 2014]

Tabela 2. Transcrição fonética da palavra "cidade" em diferentes variantes linguísticas do RS

Palavra	**Variante linguística**	**Transcrição**
Cidade	Porto Alegre, RS	/s i d a dʒi/
	Uruguaiana, RS	/s i d a d e/

A associação entre fonemas produzidos na fala e símbolos do IPA é geralmente disposta em tabelas, como mostrado na Figura 1 [International Phonetic Association 2020].

Mesmo para os leitores com alguma familiaridade, é inegável que absorver todo o conteúdo não é uma tarefa trivial. Aprender a realizar transcrição fonética é um trabalho árduo, especializado e complexo.

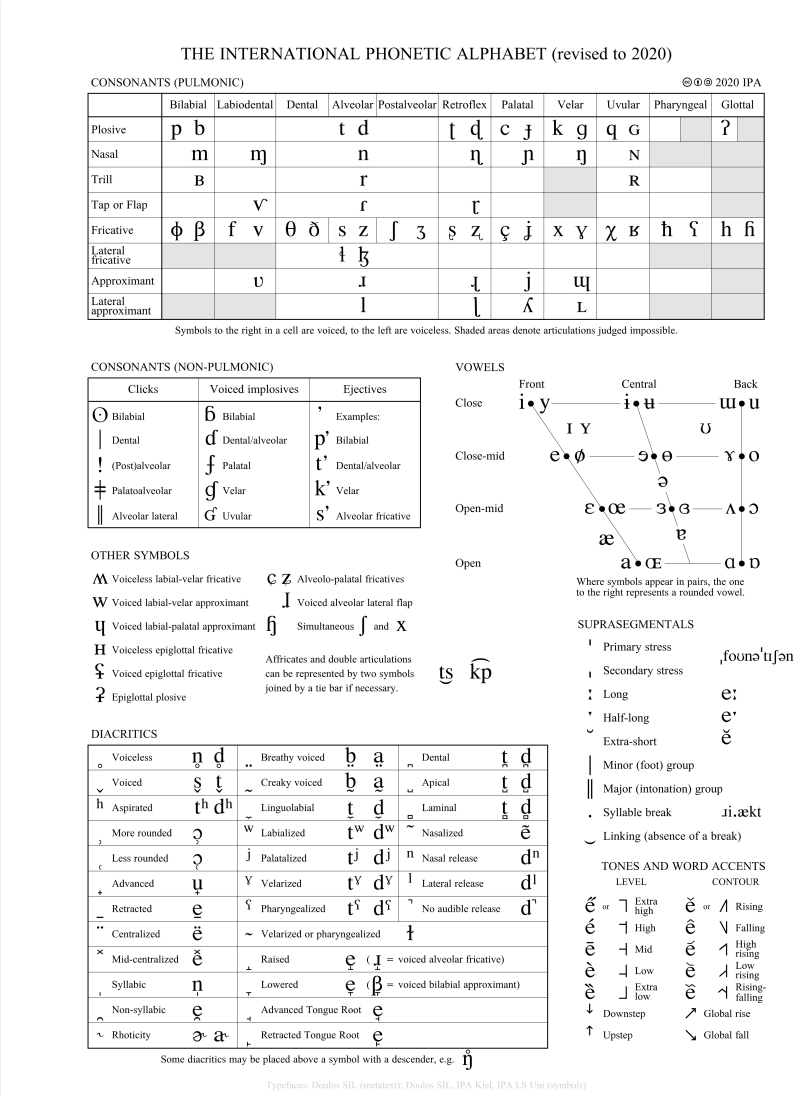

Figura 1. A tabela fonética do IPA

Nesse sentido, viu-se a necessidade de desenvolver ferramentas computacionais para a facilitação da transcrição fonética. Com os avanços recentes no desenvolvimento de Inteligência Artificial (IA), possibilitados pelo crescimento da capacidade de processamento e pelo aprimoramento das arquiteturas de redes neurais, possibilitou-se que tarefas complexas, sem abordagens procedurais, fossem realizadas automaticamente por computadores.

Partindo deste contexto, desenvolveu-se o ***AutoIPA***, uma plataforma movida à IA que facilita a realização, aprendizado e correção de transcrição fonética. O AutoIPA permite ao usuário gravar um áudio e ver a transcrição fonética correspondente no IPA.

Nas seções a seguir, explorar-se-á o funcionamento da plataforma e todos aspectos relativos ao seu desenvolvimento e impacto.

2. Motivação e objetivos

A representação acurada da fala em texto, por meio da transcrição fonética, possibilita uma análise mais facilitada da fonética. Sob posse desses dados em formato textual, dispensa-se ouvir o áudio a cada vez para extrair seu conteúdo.

Como o IPA é uma forma precisa de representação, é bastante útil para o ensino de línguas, a desambiguação fonética de termos desconhecidos e o estudo de variações linguísticas. A título de exemplo, no ensino de línguas, o conhecimento do IPA pode ser usado por um estudante para verificação da pronúncia e subsequente adequação desta à variante desejada [Atkielski 2005].

Entende-se, pois, que a popularização desse processo seria benefício. Contudo, como dito anteriormente, a transcrição é hoje tarefa especializada e pouco conhecida, sendo restrita aos cursos de Letras e às pessoas que tenham estudado o IPA.

Dessa forma, viu-se a oportunidade de auxiliar na disseminação da transcrição fonética e de sua compreensão através da criação de uma ferramenta de Inteligência Artificial. Com uma ferramenta assim, basta ao usuário falar em seu celular para ver os fonemas que compõem o trecho gravado, de forma simples e rápida.

Assim, objetivou-se desenvolver uma plataforma web que tornasse acessível a transcrição fonética partindo de qualquer idioma com destino no Alfabeto Fonético Internacional. Subsidiariamente, buscou-se desenvolver, melhorar e avaliar os modelos de IA que realizam essa transcrição.

3. Panorama Tecnológico

Com o avanço da Inteligência Artificial, várias tarefas intransponíveis com programação procedural tradicional se tornaram solucionáveis com abordagens de Aprendizado de Máquina. Isso se destaca, sobretudo, no campo de processamento de linguagem natural - área que, devido a sua inerente complexidade, teve avanços significativos apenas mais recentemente com o aumento do poder computacional e desenvolvimento de novas arquiteturas de redes neurais - como os transformadores. [Vaswani et al. 2017]

Mesmo com esses avanços, são poucos os modelos de IA existentes para realização de transcrição fonética. Ademais, não existe qualquer plataforma aberta e simplificada para a utilização de tais modelos - o que os torna restritos ao público com domínio técnico da programação e computação.

O estado da arte em processamento da voz humana é o modelo Wav2Vec2 do Facebook (atualmente, Meta). Esse modelo é baseado em uma arquitetura de transformadores e na ideia de aprendizado de máquina auto-supervisionado (*self-supervised learning*) em que o modelo foi treinado sobre áudio puro sem qualquer rotulação dos dados. Como resultado dessa fase de treinamento, obtém-se o modelo base do Wav2Vec2, que então passa por uma fase de *fine-tuning* em que ele é ajustado por meio de uma nova etapa de treinamento em que os rótulos desejados são adicionados [Baevski et al. 2020].

Boa parte dos modelos baseados no Wav2Vec2 são modelos de transcrição convencional de voz - isto é, voz para texto padrão em uma determinada língua, como fala em inglês para texto em inglês no alfabeto latino. Porém, esse modelo base está apto e é utilizado para uma ampla gama de tarefas: na plataforma de colaboração em IA Hug-

ging Face, há modelos baseados no Wav2Vec2 que fazem reconhecimento de emoções [Ravanelli et al. 2021] e gênero do interlocutor [Alefiury 2024], por exemplo.

Esse modelo também foi usado como base para a construção de modelos de transcrição fonética. Considerando que é um modelo base genérico, tendo uma base de dados rotulada com as transcrições fonéticas correspondentes no IPA, é possível realizar o *fine-tuning* do modelo para atuar nesta tarefa.

Contudo, diferentemente de bases de dados texto-voz[1] - que possuem quantidades enormes de dados amplamente disponíveis -, as bases de dados texto-IPA são mais escassas, o que diminui a qualidade dos modelos existentes. Na testagem preliminar realizada, alguns dos modelos encontrados apresentaram resultados inconsistentes e instáveis. Por exemplo, o modelo Allosaurus [Li et al. 2020] faz transcrições com acurácia aceitável em inglês, mas é ruim, à primeira vista, diferenciando fonemas da língua portuguesa.

Os modelos baseados no Wav2Vec2 apresentaram melhor performance. Os dois modelos usados para o desenvolvimento da plataforma AutoIPA - objeto deste artigo - foram *fine-tunes* originários dessa base. Na Seção 5, analisar-se-á a performance desses dois modelos.

- facebook/wav2vec2-lv-60-espeak-cv-ft [Facebook 2025a]
- facebook/wav2vec2-xlsr-53-espeak-cv-ft [Facebook 2025b]

A partir disso, evidenciou-se que ainda persistem desafios para o desenvolvimento de modelos capazes de transcrição fonética com desempenho de qualidade e generalização para múltiplos idiomas. Ademais, verificou-se que, mesmo os modelos existentes, são inacessíveis para o público leigo e representam uma barreira técnica ainda alta para adoção.

4. Metodologia e Implementação do Sistema

O AutoIPA consiste em uma aplicação web (autoipa.org)[2] que realiza a gravação de um trecho de voz (ou recebe um arquivo contendo uma gravação), envia para um servidor de *backend* em que é feito o processamento do arquivo por meio da rede neural, e envia a resposta com a predição da transcrição fonética de volta para o usuário. Dessa forma, do ponto de vista do usuário final, não há qualquer barreira técnica para o uso da plataforma. Essa escolha de simplicidade é um contraste com o panorama atual, em que é necessário configurar ambientes de desenvolvimento de software para a mera execução de um modelo.

Sob a ótica das tecnologias utilizadas para o funcionamento do AutoIPA, utilizou-se para desenvolver a interface de *frontend* do usuário as linguagens convencionais da Web - HTML, CSS e JavaScript, sem uso de *frameworks* adicionais. As Figuras 2 e 3 mostram a interface do AutoIPA.

Para o *backend*, optou-se pela utilização de um servidor em Python, utilizando a biblioteca Flask para o gerenciamento das requisições. Essa escolha foi feita por dois motivos: primeiramente, porque a flexibilidade da linguagem permite a utilização de várias

[1]Refere-se aqui a "texto" no sentido estrito de texto convencional, escrito usando o sistema de escrita da respectiva língua e não um sistema "neutro", como o Alfabeto Fonético Internacional.

[2]Alternativamente, disponível em autoipa.brizzigui.com.

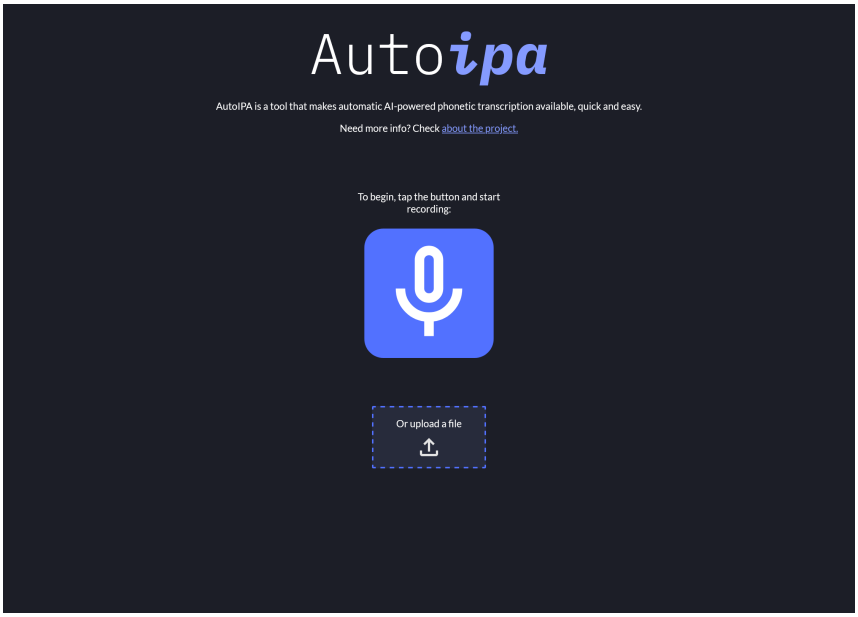

Figura 2. Captura de tela da página principal do AutoIPA.

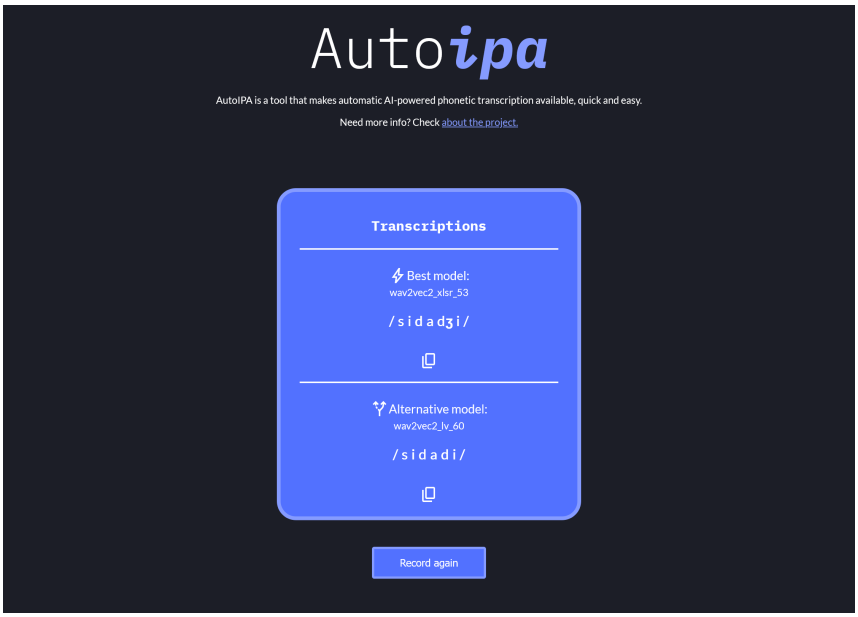

Figura 3. Captura de tela da página de resultados do AutoIPA.

bibliotecas auxiliares para pré-processamento de áudio e arquivos, suplementarmente, pois facilitou a conexão direta do servidor de *backend* com a execução dos modelos de IA, visto que a execução desses é feita a partir de um script em Python utilizando as bibliotecas de carregamento do Hugging Face (Transformers), as quais trabalham em cima do PyTorch para inferência [Wolf et al. 2020].

Quanto aos modelos de transcrição, utilizaram-se os *fine-tunes* baseados no Wav2Vec2, como mencionado anteriormente. Esse modelo base usa a arquitetura de transformadores, a qual é uma estrutura mais recente de rede neural, que analisa a entrada por partes e pode utilizar as partes "anteriores" para a inferência da entrada atual.

Devido a essa capacidade, a arquitetura de transformadores se tornou o estado da arte e é utilizada em aplicações em que a análise de contexto é importante - como em modelos de linguagem larga, por exemplo. Os modelos *wav2vec2-lv-60-espeak-cv-ft* e *wav2vec2-xlsr-53-espeak-cv-ft* usam base de dados adicionais - neste caso, o *Common Voice* - para especializar o modelo base genérico e torná-lo no modelo específico para transcrição fonética.

5. Avaliação e Resultados

Realizou-se uma comparação do desempenho dos modelos pesquisados para a criação do AutoIPA usando uma base de dados independente, não relacionada com o treinamento dos respectivos modelos, o *bookbot/ljspeech_phonemes* [Bookbot 2022]. Essa base de dados contém 13100 amostras de áudio em língua inglesa e suas respectivas transcrições fonéticas.

Avaliou-se a performance a partir da métrica de Taxa de Erro de Caracteres (*Character Error Rate* (CER)) [K et al. 2024], conforme os resultados da tabela 3.

Tabela 3. Taxa média de erro por caractere (CER) e desvio padrão (SD) para os modelos de IA

Modelo	CER médio	Desvio padrão (SD)
`xlsr_53`	0.1405	$1{,}78 \times 10^{-9}$
`lv_60`	0.1314	0.0476

Nota-se que ambos os modelos são similares, com o *wav2vec2-lv-60-espeak-cv-ft* tendo taxa de erros levemente menor, mas desvio-padrão maior em contraponto ao *wav2vec2-xlsr-53-espeak-cv-ft*. Esse resultado é esperado visto que ambos os modelos têm arquitetura e base de dados similares.

Nesse viés, vale ressaltar que a taxa de erro por caractere (CER) de 14.05% e 13.14% ainda são relativamente altas, mormente tendo em vista o fato de que trabalhos reconhecidos estabelecem 5% como uma taxa máxima aceitável de erro por caractere, em transcrição convencional [Pratap et al. 2024]. As taxas obtidas nesta avaliação são maiores e reforçam que os modelos existentes ainda são pouco robustos. Outrossim, destaca-se que a performance em inglês desses modelos tende a ser melhor que em outras línguas, as quais sofrem mais com a escassez de dados rotulados - e, dessarte, com mais erros. Portanto, é esperado que, em testes com bases de dados em outros idiomas, haja uma performance ainda mais reduzida.

Quanto ao impacto do desenvolvimento da plataforma, conquanto os resultados obtidos até o presente momento são indicativos de um potencial promissor, por se tratar de uma fase inicial, ainda é prematuro realizar generalizações ou se falar em resultados definitivos.

6. Discussão e Limitações

Evidenciou-se que os modelos atuais de transcrição fonética ainda são limitados, trazem resultados com pouca consistência e são de baixa acessibilidade. Quando comparados aos seus correlatos mais próximos, os modelos de transcrição textual convencional, é inegável que eles ainda estão subdesenvolvidos.

Contudo, percebe-se que, mesmo na escassez de dados, já foi possível utilizar modelos generalizados como base para criar sistemas de transcrição fonética funcionais. Dessa forma, possibilitou-se também a criação de uma plataforma aberta na web que tornasse essas ferramentas acessíveis - o AutoIPA, objeto deste artigo.

Nesse sentido, é pertinente reiterar a utilidade da transcrição fonética como ferramenta para uso geral: ela se mostra útil para aprimorar o ensino de idiomas, permitindo que alunos e professores identifiquem com precisão os sons e nuances da fala, o que facilita tanto a aquisição de uma nova língua quanto a correção da pronúncia [Atkielski 2005]. Ela também pode contribuir para o trabalho de fonoaudiólogos, em que essas transcrições são usadas para diagnosticar e tratar distúrbios da comunicação, além de promover a inclusão de pessoas com dificuldades de processamento auditivo e de leitura [Bates et al. 2024]. Em contextos de pesquisa linguística e preservação cultural, a ferramenta auxilia na documentação e estudo de dialetos e variações regionais, garantindo um acesso mais democrático e informado ao conhecimento da diversidade linguística da população [Bhaskararao 2004].

Compreende-se, pois, que a transcrição fonética não deve ser uma técnica nichada e restrita, mas sim algo de acesso e uso facilitados.

7. Conclusões e Trabalhos Futuros

Conclui-se que o desenvolvimento de ferramentas de transcrição fonética ainda é recente e limitado, mas seu potencial é amplo e benéfico. Há desafios, como a escassez de dados rotulados e a dificuldade na criação de novos dados, devido à barreira técnica da tarefa. Contudo, os fortes avanços das arquiteturas de redes neurais e dos *frameworks* de modelos pré-treinados permitiram resultados mesmo com um volume de dados rotulados mais reduzidos. Ademais, a criação da plataforma AutoIPA para acesso público a essas ferramentas é um dos passos para a constante melhoria do panorama atual.

A partir disso, tem-se a intenção de dar-se continuidade ao trabalho de desenvolvimento de ferramentas computacionais de IA voltadas para a linguística. Nesse campo, vê-se a oportunidade de realizar uma pesquisa voltada para mapeamento de sotaques automatizado e também o desenvolvimento de modelos de predição geográfica a partir da voz.

De forma geral, evidenciou-se que utilizar a tecnologia para melhor compreender e estudar a diversidade linguística é uma abordagem adequada e que deve ser mais explorada.

Referências

Alefiury (2024). wav2vec2-large-xlsr-53-gender-recognition-librispeech. Acesso em: maio 2025. Disponível em: `https://huggingface.co/alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech`.

Atkielski, A. (2005). Phonetic transcription can be a useful tool for teaching or correcting pronunciation in the esl/efl classroom. *Using Phonetic Transcription in Class*. p. 1–12.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. p. 1–5.

Bates, S., Watson, J., Heselwood, B., and Howard, S. (2024). Phonetic transcription in clinical practice. In Ball, M. J., Müller, N., and Spencer, E., editors, *The Handbook of Clinical Linguistics*. John Wiley & Sons. p. 471–489.

Battisti, E. (2014). Palatalização de t e d. In Bisol, L. and Battisti, E., editors, *O português falado no Rio Grande do Sul*. EDIPUCRS, Porto Alegre. p. 105–120.

Bhaskararao, P. (2004). Phonetic documentation of endangered languages: Creating a knowledge- base containing sound recording, transcription and analysis. *Acoustical Science and Technology*, 25(4). p. 219-226.

Bisol, L. (2005). *Introdução a estudos de fonologia do português brasileiro*. EDIPUCRS.

Bookbot (2022). Ljspeech phonemes dataset. `https://huggingface.co/datasets/bookbot/ljspeech_phonemes`. Acesso em: maio de 2025.

Ethnologue (2025). Ethnologue: Languages of the world. Disponível em: `https://www.ethnologue.com/`. Acesso em: abril de 2025.

Facebook (2025a). facebook/wav2vec2-lv-60-espeak-cv-ft: A fine-tuned model for speech recognition on commonvoice. Disponível em: `https://huggingface.co/facebook/wav2vec2-lv-60-espeak-cv-ft`. Acesso em: abril de 2025.

Facebook (2025b). facebook/wav2vec2-xlsr-53-espeak-cv-ft: A fine-tuned model for speech recognition on commonvoice. Disponível em: `https://huggingface.co/facebook/wav2vec2-xlsr-53-espeak-cv-ft`. Acesso em: abril de 2025.

International Phonetic Association (2020). The international phonetic alphabet (revised to 2020). `https://www.internationalphoneticassociation.org/IPAcharts/IPA_chart_orig/pdfs/IPA_Kiel_2020_full.pdf`. Official IPA chart rendered in the TeX TIPA Roman font.

K, T. D., James, J., Gopinath, D. P., and K, M. A. (2024). Advocating character error rate for multilingual asr evaluation. Acesso em: maio de 2025.

Li, X., Dalmia, S., Li, J., Littell, P., Lee, M., Yao, J., Anastasopoulos, A., Mortensen, D., Neubig, G., Black, A., and Metze, F. (2020). Universal phone recognition with a multilingual allophone system. In *2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona.

O'Grady, W. (2005). *Contemporary Linguistics: An Introduction*. Bedford/St. Martin's, 5th edition. p. 15.

Peterson, D. J. (2015). *The Art of Language Invention: From Horse-Lords to Dark Elves to Sand Worms, the Words Behind World-Building*. p. 18–23.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., et al. (2024). Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y. (2021). Speechbrain: A general-purpose speech toolkit. Acesso em: maio 2025. Disponível em: `https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP`.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Łukasz Kaiser, and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. p. 5998–6008.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages p. 38–45, Online. Association for Computational Linguistics.