Avaliação do Desempenho de Ferramentas de Transcrição de Áudio em Português para Análise de Dados da Web

Jonatas Santos[1], **Marcelo M. R. Araujo**[1], **Josemar Caetano**[1], **Yago Santos**[2]
Julio C. S. Reis[2], **Ana P. C. Silva**[1], **Jussara M. Almeida**[1]

[1]Universidade Federal de Minas Gerais (UFMG) – Brasil

[2]Universidade Federal de Viçosa (UFV) – Brasil

{jonatashds,mmra}@ufmg.br,josemarcaetano@dcc.ufmg.br,yago.santos@ufv.br,

jreis@ufv.br,{ana.coutosilva,jussara}@dcc.ufmg.br

Abstract. *In this work, we present an evaluation of the performance of Portuguese audio transcription tools used for Web data analysis. For this purpose, we explored a publicly accessible dataset, and performed analyses based on two main dimensions: (total) number of failures and accuracy. Our results present interesting findings that may be useful to guide researchers in choosing audio transcription methods for studies focused on the Portuguese language.*

Resumo. *Neste trabalho apresentamos uma avaliação do desempenho de ferramentas de transcrição de áudio em Português utilizados para análise de dados da Web. Para isso, exploramos um conjunto de dados de acesso público, e realizamos análises a partir de duas dimensões principais: número (total) de falhas e precisão. Nossos resultados apresentam descobertas interessantes que podem ser úteis para guiar pesquisadores na escolha de métodos de transcrição de áudio para estudos focados no idioma Português.*

1. Introdução

A popularização do uso da Internet impulsionou a produção e o compartilhamento de grandes volumes de dados em diferentes tipos de mídias em diversas plataformas da Web [Verma et al. 2016], incluindo as sociais. Se anteriormente a difusão de informação era majoritariamente realizada em formato texto (e.g., uma mensagem em uma rede social ou um comentário em um *blog*), atualmente vídeos e áudios passaram a ser ferramentas de comunicação preferenciais em muitos cenários [Resende et al. 2019]. Há de se ressaltar que conteúdos compartilhados nestes tipos de mídias tendem a ter um alcance maior de audiência, permitindo uma comunicação rápida entre pessoas e favorecendo a inclusão de pessoas com deficiências auditiva ou visual [McCleary and Viotti 2007], ou ainda, com dificuldades de escrita ou no uso de tecnologia.

Em especial, o compartilhamento de conteúdo em *áudio* tem se tornado muito comum na Web, principalmente com o surgimento de aplicativos de troca de mensagens instantâneas como o WhatAapp, Telegram e Viber. Porém, este aumento de conteúdo neste tipo de mídia trouxe desafios para uma multitude de serviços que exploram propriedades do conteúdo para prover diversas funcionalidades tais como identificação de tópico de discussão [Qiang et al. 2020], recomendação de conteúdo de interesse [Lops et al. 2019], detecção de conteúdo tóxico e discurso de ódio [Saha et al. 2021], ou ainda, desinformação [Resende et al. 2019]. Isto porque estes trabalhos exploram

fundamentalmente propriedades do conteúdo textual. Logo, adaptações para processar o conteúdo em áudio diretamente não escalam para o volume cada vez maior de conteúdo disponível, já que algoritmos de processamento de áudio tendem a ser mais complexos e a requerer maior capacidade de processamento [Rumsey 2012]. Assim, a *transcrição do áudio* se torna uma importante etapa para viabilizar a execução destes serviços que podem, então, atuar sobre o conteúdo textual da transcrição [Riggs and Knobloch-Westerwick 2022].

De forma geral, transcrições de áudio podem ser feitas manualmente ou automaticamente, por meio de ferramentas computacionais. Transcrições manuais tendem a ser mais precisas, mas muito mais dispendiosas já que requerem tempo e pessoas fluentes no idioma em que o áudio foi produzido, além de experiência relacionada ao contexto. As transcrições automáticas, por sua vez, tendem a ser mais rápidas, podendo assim escalar para um volume muito maior de dados. Porém, a qualidade da transcrição automática pode variar bastante, dependendo da ferramenta utilizada. Neste contexto, apesar da existência de várias ferramentas de transcrição de áudios, incluindo algumas de acesso gratuito, a literatura relacionada disponível foca primariamente em tarefas complementares à transcrição como a construção de conjuntos de dados [Kolobov et al. 2021], com destaque para outros idiomas, principalmente Inglês [Wang et al. 2020]. Existe, portanto, uma lacuna no que tange a avaliação sistemática e criteriosa do desempenho de diferentes ferramentas de transcrição, em especial para áudios em Português, considerando diferentes cenários e métricas de avaliação, que podem ser úteis para análise de dados da Web.

Assim, visando preencher esta lacuna de pesquisa, este trabalho apresenta uma avaliação do desempenho de 4 ferramentas de transcrição de áudios para Português a partir de duas dimensões: *(i) número de falhas*, que é o número de áudios que a ferramenta não foi capaz de transcrever fornecendo assim, uma análise de cobertura das mesmas e, *(ii) precisão*, que reflete a qualidade da transcrição realizada (i.e., similaridade entre o texto real – *baseline* – e o texto transcrito fornecido pela ferramenta). As ferramentas analisadas (i.e., Google, IBM Watson, Vosk, Microsoft), foram selecionadas a partir de uma busca ampla na Web por ferramentas populares de transcrição com suporte para áudios em Português e uso gratuito. Elas foram avaliadas em um repositório de dados da Web com acesso público, que foi selecionado por conter narrações curtas relacionadas a assuntos variados, representando, de forma satisfatória, características de áudios disseminados em diversos sistemas sociais, como WhatsApp e Telegram.

Em suma, nossos resultados evidenciam a sensibilidade dessas ferramentas ao mostrar que elas se comportam de maneiras distintas considerando as diferentes dimensões avaliadas (i.e., número de falhas e precisão). Em outras palavras, ferramentas que apresentam bom desempenho em relação ao número de falhas, não são precisas, em todos os cenários, para a transcrição dos aúdios, o que reforça a importância deste estudo. Esperamos que os resultados obtidos possam ser úteis para fomentar esforços que explorem conteúdo disseminado na Web no formato de áudio, auxiliando pesquisadores na escolha da ferramenta para transcrição do mesmo.

Na próxima seção discutimos brevemente trabalhos relacionados. Em seguida, na Seção 3, apresentamos a metodologia proposta para este estudo. Os resultados experimentais são apresentados e discutidos na Seção 4. Por fim, na Seção 5 apresentamos as considerações finais.

2. Trabalhos Relacionados

A transcrição de áudio é uma tarefa extremamente relevante que tem sido explorada em diversos contextos com propósitos distintos como mediar a comunicação entre usuários [Kolobov et al. 2021], incluindo o processo de conversação com pessoas surdas onde áudios extraídos de vídeos são transcritos para a Língua Brasileira de Sinais (i.e., Libras) [Guerra et al. 2020], e suportar a análise do discurso dos usuários (em áudio) a partir de diferentes plataformas, como WhatsApp [Maros et al. 2020], com objetivo de caracterizar propriedades (e.g., linguísticas, etc) bem como a dinâmica de propagação do conteúdo disseminado especificamente neste tipo de mídia (i.e., áudio).

De forma geral, no que tange ao processo de escolha das ferramentas exploradas nos trabalhos anteriores, percebemos que não existe uma metodologia sistemática e ou análises experimentais bem definidas que suportem e/ou justifiquem a ferramenta de transcrição de áudio selecionada. Comumente, a seleção dessas ferramentas consiste em um processo realizado com base em critérios distintos e não claros envolvendo aspectos subjetivos como facilidade de uso. Além disso, embora existam esforços anteriores que compararam o desempenho de ferramentas de transcrição de áudios em diferentes cenários, até onde sabemos, eles são limitados, por exemplo, a outros idiomas (e.g., Turco, Espanhol, etc [Kolobov et al. 2021]) e ferramentas específicos [Filippidou and Moussiades 2020], o que reforça a necessidade de realização de estudos focados investigação do desempenho de ferramentas de transcrição de áudios especificamente para o idiomas onde eles são precários, como o Português.

3. Metodologia

Nesta seção, apresentamos detalhes relativos à metodologia proposta para avaliação das ferramentas de transcrição de áudio em Português analisadas neste trabalho. Primeiramente, descrevemos os critérios utilizados para a seleção das ferramentas a serem analisadas. Em seguida, descrevemos a base de dados explorada, e por fim, apresentamos as métricas utilizadas para a avaliação de desempenho das ferramentas analisadas bem como detalhes relativos à configuração experimental.

Ferramentas para Transcrição de Áudio. Para seleção das ferramentas analisadas neste trabalho, primeiramente, realizamos uma pesquisa em artigos científicos [Sampaio et al. 2021, Kolobov et al. 2021] e técnicos[1] em busca das ferramentas mais utilizadas para a tarefa de transcrição automática de áudios. Em seguida, escolhemos as ferramentas mais populares com suporte ao Português brasileiro. Ao final desta etapa, selecionamos 4 (quatro) ferramentas[2]: Google Cloud Speech API (Google), IBM Watson Speech-to-Text (IBM Watson), Vosk API (Vosk), e Azure Microsoft Speech-to-Text (Microsoft). Neste contexto, é importante mencionar que filtramos apenas ferramentas de licença gratuita (i.e., total ou parcial). Exceto para a ferramenta Vosk, que é totalmente gratuita, para as demais (i.e., parcialmente gratuitas) foram geradas múltiplas chaves de acesso como estratégia para ampliação do número de requisições.

Base de Dados. Para uma análise sistemática do desempenho de ferramentas de transcrição de áudio precisamos de uma base de dados curada, com áudios em Português.

[1] https://www.voicegain.ai/post/speech-to-text-benchmark-june-2021

[2] https://cloud.google.com/speech-to-text?hl=pt-br, https://www.ibm.com/cloud/watson-text-tAo-speech, https://github.com/alphacep/vosk-api, https://azure.microsoft.com/en-us/services/cognitive-services/speech-services/

Assim, selecionamos a base de dados *"Common Voice Corpus - Mozilla Foundation"*[3], versão 7.0, que contém narrações de sentenças curtas de assuntos variados, como leitura de frases de livros, falas de personagens de filmes, descrição de objetos, dentre outros. Desta forma, utilizamos esta base de dados para considerar áudios com vocabulários diversificados, não especializados (i.e. somente sobre um assunto específico) e coloquiais, que são representativos da diversidade de áudios disseminados na Web, e comumente encontrados em plataformas sociais como WhatsApp e Telegram. No total, a base de dados analisada é composta por uma amostra de 723 áudios com duração de até 15 segundos.

Métricas. Para avaliação do desempenho das abordagens selecionadas consideramos duas dimensões: *(i) número (total) de falhas*, que podem ocorrer no processo de transcrição e, *(ii) precisão*, que mede a similaridade entre o texto transcrito automaticamente e o texto original extraído do áudio, via *baseline* (i.e., transcrição exata de cada arquivo de áudio), fornecida pela própria base de dados. Para *(i)*, calculamos a quantidade de falhas que cada ferramenta apresentou ao transcrever áudios. Se uma das ferramentas selecionadas apresentar erro(s) durante o processo de transcrição (e.g., a ferramenta não reconheceu o foi dito e retornou uma transcrição – resposta – vazia), contabilizamos como 1 falha. Já em *(ii)* mensuramos a qualidade da transcrição de cada ferramenta considerando duas estratégias amplamente exploradas neste contexto [Ratcliff and Metzener 1988, Sampaio et al. 2021]: (1) algoritmo *Gestalt Pattern Matching* (GPM) e (2) métrica *Word Error Rate* (WER). Em suma, enquanto a primeira estratégia considera o quanto o resultado da transcrição preserva a organização das sentenças originais dos áudios, a segunda estratégia considera a taxa de erros de palavras transcritas. Em outras palavras, para GPM, um valor mais alto indica uma maior similaridade em relação ao texto base, e para WER, um valor menor, representa menor erro relacionado a transcrição.

Configuração Experimental. Para garantir a equidade na comparação do número de falhas e precisão das ferramentas selecionadas, dividimos os áudios de cada base de dados em subconjuntos cuja duração estão nas seguintes faixas: 0 – 5 segundos, 5 – 10 segundos, e 10 – 15 segundos, com 300, 300 e 123 áudios em cada faixa, respectivamente. O objetivo é permitir uma comparação mais justa ao analisar áudios com tempo de duração similar.

Na seção a seguir apresentamos os resultados experimentais obtidos para cada uma das dimensões analisadas (i.e., número de falhas e precisão), considerando cada um dos subconjuntos da base de dados original, conforme supracitado.

4. Resultados Experimentais

A Tabela 1 apresenta o número (total) de falhas que cada uma das ferramentas analisados. Primeiramente, observamos que a ferramenta Vosk foi a única que não apresentou falhas durante a transcrição de todos os 723 áudios da base de dados selecionada. Isso significa que para todos os áudios ela foi capaz de fornecer um resultado de transcrição (diferente de vazio). Por outro lado, a ferramenta Microsoft apresentou o maior número de falhas de transcrição em todos os intervalos de tempo analisados, totalizando 61 falhas ($\approx 8\%$ do total da base). Note ainda que, a partir da análise das ferramentas Google e IBM Watson, é possível perceber uma tendência: quanto o maior o tempo de duração do áudio, menor o número de falhas (e.g., casos em que a resposta de retorno da ferramenta é vazia).

[3]https://commonvoice.mozilla.org/pt

Tabela 1. Número de falhas para cada ferramenta analisada.

Ferramenta	0 – 5 segundos	5 – 10 segundos	10 – 15 segundos	#Total Falhas
Google	10	1	1	12
IBM Watson	3	0	0	3
Vosk	0	0	0	0
Microsoft	19	32	10	61

Os resultados médios de precisão, obtidos a partir da aplicação do algoritmo GPM e da métrica WER são apresentados, respectivamente, nas Tabelas 2 e 3. A análise de precisão realizada a partir do algoritmo GPM, nos revela uma série de resultados interessantes. Primeiramente, podemos notar que a ferramenta Google apresentou a maior precisão média e menores desvios em todos os intervalos de tempo considerados. Depois, as ferramentas IBM Watson e Microsoft apresentam resultados sobrepostos (i.e., empatados), exceto para áudios mais longos (i.e., 10 – 15 segundos), onde a IBM Watson apresenta melhores resultados em comparação a ferramenta Microsoft. Por fim, a ferramenta Vosk, apresentou os menores valores de precisão e maiores desvios, respectivamente.

Por outro lado, quando analisamos os resultados de precisão utilizando a métrica WER (ver Tabela 3) observamos algumas particulariades. De forma geral, as ferramentas Vosk e Google, apresentaram os menores valores de precisão média (i.e., taxa de erros) para a base de dados analisada com exceção de áudios com menor duração (i.e., 0 – 5 segundos), onde os resultados apresentados pela Vosk são mais promissores em comparação a ferramenta Google. É interessante notar que para este mesmo conjunto (i.e., áudios com menor duração), os resultados apresentados pela ferramenta IBM Watson são satisfatórios, e pioram a medida que o tempo de duração do áudio aumenta. A ferramenta Microsoft, por sua vez, apresentou os piores resultados gerais em termos de WER, incluindo as maiores variações.

5. Considerações Finais

Este trabalho apresentou uma investigação do desempenho de 4 ferramentas de transcrição de áudio com suporte para o idioma Português para análise de dados da Web. Neste contexto, a comparação, que explorou um conjunto de dados de acesso público, foi realizada a partir de duas dimensões principais: (i) número (total) de falhas e (ii) precisão. Nossos resultados apresentam descobertas interessantes. Em termos de falhas, a ferramenta Vosk se mostrou a mais robusta, enquanto em termos de precisão, os resultados foram mais sensíveis, por exemplo, ao tempo de duração dos áudios. Esperamos que este estudo seja útil para guiar pesquisadores na escolha de métodos de transcrição de áudio no idioma Português para análise de dados oriundos de diferentes plataformas (e.g., sociais). Como trabalhos futuros prentendemos investigar o desempenho das ferramentas em novas bases de dados, a partir de novas dimensões (e.g., tempo) e eventualmente expandir o volume de ferramentas avaliadas.

Tabela 2. Precisão média da transcrição de áudios realizada pelas ferramentas utilizando o algoritmo GPM.

Ferramenta	0 – 5 segundos	5 – 10 segundos	10 – 15 segundos
Google	0,92 ($\pm$0,16)	0,94 ($\pm$0,09)	0,94 ($\pm$0,13)
IBM Watson	0,85 ($\pm$0,21)	0,89 ($\pm$0,13)	0,91 ($\pm$0,13)
Vosk	0,75 ($\pm$0,28)	0,80 ($\pm$0,21)	0,88 ($\pm$0,17)
Microsoft	0,86 ($\pm$0,26)	0,87 ($\pm$0,26)	0,84 ($\pm$0,29)

Tabela 3. Precisão média da transcrição de áudios realizada pelas ferramentas utilizando a métrica WER.

Ferramenta	0 – 5 segundos	5 – 10 segundos	10 – 15 segundos
Google	0,26 (±0,15)	0,16 (±0,22)	0,14 (±0,20)
IBM Watson	0,16 (±0,16)	0,20 (±0,14)	0,26 (±0,18)
Vosk	0,10 (±0,05)	0,16 (±0,07)	0,13 (±0,07)
Microsoft	0,22 (±0,43)	0,18 (±0,32)	0,23 (±0,35)

Agradecimentos. CNPq, FAPEMIG e Ministério Público de Minas Gerais (MPMG).

Referências

Filippidou, F. and Moussiades, L. (2020). A benchmarking of ibm, google and wit automatic speech recognition systems. In *IFIP Int'l Conference AIAI*.

Guerra, P. A. C., Silveira, S. R., Bertolini, C., Parreira, F. J., and Ulbricht, V. R. (2020). Aplicativo mobile para avaliar a acessibilidade de objetos de aprendizagem utilizando um sistema especialista. *Revista Educação Especial*, 33:1–26.

Kolobov, R., Okhapkina, O., Omelchishina, O., Platunov, A., Bedyakin, R., Moshkin, V., Menshikov, D., and Mikhaylovskiy, N. (2021). Mediaspeech: Multilanguage asr benchmark and dataset. *arXiv preprint arXiv:2103.16193*.

Lops, P., Jannach, D., Musto, C., Bogers, T., and Koolen, M. (2019). Trends in content-based recommendation. *User Modeling and User-Adapted Interaction*, 29(2):239–249.

Maros, A., Almeida, J., Benevenuto, F., and Vasconcelos, M. (2020). Analyzing the use of audio messages in whatsapp groups. In *Proc. of the WWW*.

McCleary, L. and Viotti, E. (2007). Transcrição de dados de uma língua sinalizada: um estudo piloto da transcrição de narrativas na língua de sinais brasileira (lsb).

Qiang, J., Qian, Z., Li, Y., Yuan, Y., and Wu, X. (2020). Short text topic modeling techniques, applications, and performance: a survey. *IEEE TKDE*.

Ratcliff, J. W. and Metzener, D. E. (1988). Pattern-matching-the gestalt approach. *Dr Dobbs Journal*, 13(7):46.

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019). (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *Proc. of the WWW*.

Riggs, E. E. and Knobloch-Westerwick, S. (2022). Beyond the text: Testing narrative persuasion mechanisms with audio messages. *Mass Communication and Society*.

Rumsey, F. (2012). *Spatial audio*. Routledge.

Saha, P., Mathew, B., Garimella, K., and Mukherjee, A. (2021). "short is the road that leads from fear to hate": Fear speech in indian whatsapp groups. In *Proc. of the WWW*.

Sampaio, M. X., Magalhães, R. P., da Silva, T. L. C., Cruz, L. A., de Vasconcelos, D. R., de Macêdo, J. A. F., and Ferreira, M. G. F. (2021). Evaluation of automatic speech recognition systems. In *Proc. of the SBBD*.

Verma, J. P., Agrawal, S., Patel, B., and Patel, A. (2016). Big data analytics: Challenges and applications for text, audio, video, and social media data. *IJSCAI*, 5(1):41–51.

Wang, Y., Luan, H., Yuan, J., Wang, B., and Lin, H. (2020). Laix corpus of chinese learner english: Towards a benchmark for l2 english asr. In *Proc. of the INTERSPEECH*.