# Análise de ferramentas para reconhecimento de fala em ambientes infantis

**Ana Cláudia Croys Felthes[1], Tiago Alexandre Dócusse[1]**
[1]Instituto Federal de São Paulo Campus Barretos
Palavras-chave: Processamento de áudio; Aplicativos

## Introdução

Crianças com necessidades especiais podem ter dificuldade na comunicação, seja por motivos neurológicos ou sensoriais (REED, 2018). Para estas crianças, é comum utilizar uma prancha de comunicação contendo cartões informativos, uma técnica da Comunicação Aumentativa e Assistiva (CAA) (ASSISTIVA, 2023), normalmente estas pranchas são físicas e possuem cartões físicos, sendo exibidos às crianças à medida em que a comunicação ocorre.

A utilização de técnicas de CAA vem sido bastante estudadas recentemente, e com o avanço das tecnologias, cada vez mais é possível desenvolver sistemas para o auxílio de pessoas com essa necessidade. Citamos, como exemplos recentes de trabalhos desenvolvidos na área, (AMERY et al., 2022), (FRIZELLE e LYONS, 2022), (LUO et al., 2022) e (TEGLER et al., 2020).

Um exemplo de utilização da prancha de cartões é quanto um instrutor exibe um cartão para uma determinação ação (por exemplo, beber água), de forma a reforçar a iteração entre a ação realizada e o desenho presente no cartão. Apesar de fácil utilização, pode apresentar uma certa demora na escolha do cartão correto, caso haja um grande número de cartões a serem pesquisados. Dessa forma, com o advento de *smartphones*, *tablets* e televisores com recursos como execução de aplicativos e acesso à *internet*, é possível utilizar aplicativos com esta capacidade, que possa exibir o cartão desejado através da fala de um utilizador, sem precisar procurar manualmente este cartão.

A fala é capturada através do som, composto por uma onda eletromagnética que pode ser captada por um instrumento e digitalizada, de forma a ser processada computacionalmente (HAYKIN, 2002). Dessa forma, é possível utilizar ferramentas computacionais para identificar as palavras ditas por uma pessoa, e realizar ações com base nesse resultado.

Atualmente, existem diversas ferramentas com interfaces de programação de aplicação - *Application Programming Interface* (API) - disponíveis para a síntese da fala, processo no qual as palavras contidas em um arquivo de áudio são extraídas para posterior utilização. Podemos destacar algumas, como a Google Speech-to-Text (GOOGLE, 2023), Microsoft Speech to Text (MICROSOFT, 2023), Whisper Speech to Text (OPENAI, 2023) e Amazon Transcribe (AMAZON WEB SERVICES, 2023).

Essas ferramentas são executadas em um servidor externo, possuindo APIs com ações públicas que podem ser acessadas por dispositivos, como aplicativos em um *smartphone* ou um sítio de *internet*. Assim é possível, para um aplicativo, capturar o áudio de um utilizador e usar a API de uma ferramenta para identificar as palavras ditas, utilizando esse resultado na busca do cartão desejado.

No entanto, ambientes com crianças nem sempre são silenciosos. Elas podem estar assistindo televisão, podem haver outras crianças conversando no mesmo ambiente, adultos, etc. Dessa forma, a capacidade de reconhecimento das palavras ditas pode ser prejudicada, sendo necessário avaliar as ferramentas disponíveis de forma a identificar qual seria a melhor adaptada para este tipo de ambiente.

## Objetivos

O objetivo geral deste projeto é avaliar recursos de reconhecimento de fala de ferramentas que possuem APIs disponíveis na *internet*, tanto em relação à sua capacidade de identificação de palavras em ambientes comuns a crianças, quanto o tempo de resposta para que esta identificação ocorra, de forma a ser viável para ser utilizada no dia a dia da comunicação alternativa de uma pessoa que possui necessidade deste recurso.

## Material e Métodos

Foram avaliadas duas ferramentas para que pudessem ter o áudio testado: Google Speech-to-Text (GOOGLE, 2023), paga, e Whisper Speech to Text (OPENAI, 2023), gratuita. Foram criados, como exemplo, três áudios de comandos que são utilizados em situações de apresentação de cartões para as crianças no dia a dia, como por exemplo: "Rafael, você quer beber água?". Esses áudios foram capturados sem ruídos aparentes. Para verificar a capacidade das ferramentas de transcrever o texto em situações do dia a dia, foram capturados áudios de ruídos de fundo comuns, como ventilador ligado, televisão ligada com música de fundo tocando e crianças brincando, sendo então adicionados aos áudios originais. No total, 21 áudios foram gerados, sendo 7 para cada áudio original gravado, cada um contendo combinações de ruídos diferentes. Todos eles foram gravados a uma taxa de amostragem de 44.100Hz e exportados utilizando a codificação *Pulse-Code Modulation* (PCM) de 16 *bits*.

## Resultados e Discussão

Os áudios gerados foram enviados às APIs das ferramentas através da *internet*, sendo calculado o tempo de resposta de cada ferramenta e comparado o texto transcrito com o áudio originalmente capturado. Na Tabela 1 a seguir é exibido um exemplo dos dados referentes à frase "Rafael, você quer beber água?", com diferentes ruídos de fundo adicionados, de acordo com a coluna "Nome": em A, o ruído de fundo é referente a um ventilador; em B, uma música; em C, crianças brincando; em D, ventilador e música combinados; em E, ventilador e crianças brincando combinados; em F, música e crianças brincando combinados; e em G, ventilador, música e crianças brincando combinados. Ainda nesta tabela, é exibida a relação sinal-ruído – *Signal to Noise Ratio* (SNR) – calculada em decibéis, bem como a diferença de energia do sinal com o ruído adicionado para o sinal original, e a diferença de volume do sinal original para o sinal com o ruído adicionado, também em decibéis.

Tabela 1 – Dados referentes aos áudios original e com ruídos adicionados para a frase "Rafael, você quer beber água?"

| Nome | SNR (dB) | Energia (%) | Dif. (dB) |
|---|---|---|---|
| A | 47,97 | 99,30 | 20,80 |
| B | 48,02 | 99,15 | 20,82 |
| C | 18,49 | 86,29 | 8,03 |
| D | 41,08 | 98,47 | 17,81 |
| E | 18,02 | 85,81 | 7,82 |
| F | 18,02 | 85,69 | 7,82 |
| G | 17,57 | 85,23 | 7,62 |

Fonte: criação dos autores

Na Tabela 2 a seguir é exibido um resumo dos resultados do processamento dos 21 arquivos de áudios analisados pelas duas ferramentas, sendo exibido o tempo médio de processamento, em segundos, bem como a porcentagem de acerto na transcrição do áudio enviado.

Tabela 2 – Resumo dos resultados obtidos

| Ferramenta | Tempo médio (s) | Acertos |
|---|---|---|
| Google Speech-to-Text | 4,03 | 80,95 % |
| Whisper Speech to Text | 16,37 | 100,00 % |

Fonte: criação dos autores

A análise da Tabela 2 nos permite visualizar que, em relação à quantidade de acertos, a ferramenta Whisper Speech to Text apresentou melhores resultados que a ferramenta Google Speech-to-Text, que apresentou erros na transcrição de apenas um dos três áudios gerados, quando o ruído das crianças brincando foi adicionado (casos C, E, F e G exidos na Tabela 1). Nas outras situações, ambas as ferramentas obtiveram transcrição correta em relação ao texto falado no áudio. Já em relação ao tempo de processamento, a ferramenta Google Speech-to-Text se mostrou superior à ferramenta Whisper Speech to Text, porém, ambas demoraram um tempo considerável para apresentar uma resposta, algo que não é desejável pois pode frustrar o usuário enquanto ele aguarda o processamento do áudio enviado. Vale ressaltar que, apesar de ser uma ferramenta paga, foi utilizada a versão gratuita de testes da ferramenta Google Speech-to-Text, sendo que não podemos afirmar se os resultados obtidos pela versão

paga podem ser diferentes, pois não foram feitos testes com ela. Também é importante ressaltar que ambas as ferramentas são executadas em um servidor na *internet*, e não é conhecida a carga de processamento do servidor no momento da requisição realizada, o que pode influenciar o tempo de resposta em comparação a um servidor dedicado apenas ao processamento das requisições de uma única aplicação.

## Conclusões

O trabalho apresentado mostrou que a utilização de ferramentas *online* para reconhecimento de fala é viável do ponto de vista de acerto na transcrição da fala, porém, o tempo para seu processamento é alto, o que pode inviabilizar sua utilização. Mais estudos devem ser realizados, com outras ferramentas e também com ferramentas dedicadas apenas a uma aplicação, para verificar melhorias que possam tornar a sua utilização viável.

## Agradecimentos

Agradeço a instituição de ensino onde faço minha graduação por tantas oportunidades de aprendizado.

## Referências Bibliográficas

AMAZON WEB SERVICES. Amazon Transcribe - Conversão de fala para texto. Disponível em: <https://aws.amazon.com/pt/transcribe/>. Acesso em: 2 fev 2023.

AMERY R., et al. Designing augmentative and alternative communication systems with Aboriginal Australians: vocabulary representation, layout, and access. **Augmentative and Alternative Communication**, v. 38, n. 4, p. 221-235, 2022.

ASSISTIVA. Comunicação alternativa. Disponível em: <https://www.assistiva.com.br/ca.html>. Acesso em: 2 fev 2023.

FRIZELLE, P. LYONS, C. The development of a core key word signing vocabulary (Lámh) to facilitate communication with children with down syndrome in the first year of mainstream primary school in Ireland. **Augmentative and Alternative Communication**, v. 38, n. 1, p. 53-66, 2022.

GOOGLE. Speech-to-Text: reconhecimento de fala automático. Disponível em: <https://cloud.google.com/speech-to-text?hl=pt-br>. Acesso em: 2 fev 2023.

HAYKIN, S.; VEEN, B. V. **Sinais e sistemas**. Porto Alegre: BOOKMAN, 2002.

LUO F., et al. Working with children with cortical visual impairment who use augmentative and alternative communication: implications for improving current practice. **Augmentative and Alternative Communication**, v. 38, n. 2, p. 91-105, 2022.

MICROSOFT. Speech to Text - Audio to Texto Translation. Disponível em: <https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text/>. Acesso em: 2 fev 2023.

OPENAI. Whisper: Speech Recognition System. Disponível em: <https://openai.com/research/whisper>. Acesso em: 4 set 2023.

REED, V. A. **An introdocution to children with language disorders**. 5a. ed. Pearson, 2018.

TEGLER H., et al. Creating a response space in multiparty classroom settings for students using eye-gaze accessed speech-generating devices. **Augmentative and Alternative Communication**, v. 36, n. 4, p. 203-213, 2020.