Ferramentas e recursos disponíveis para reconhecimento de fala em Português Brasileiro

Matheus N. S. M. de Lima
matheus.marques_96@hotmail.com
IFSC - Instituto Federal de Santa Catarina
Florianópolis, Santa Catarina, Brasil

Bruna H. Coelho
bhamess@icloud.com
IFSC - Instituto Federal de Santa Catarina
Florianópolis, Santa Catarina, Brasil

Fabrício Y. K. Takigawa
takigawa@ifsc.edu.br
IFSC - Instituto Federal de Santa Catarina
Florianópolis, Santa Catarina, Brail

RESUMO

Speech recognition allows natural communication between the humans and machines. With Industry 4.0 there is a great demand for systems that perform this task, since human-machine integrations are increasingly attractive. Currently, there are several tools and resources that perform this activity, with some companies providing their audio recognition services through the Application Programming Interface, such as Microsoft, Google, IBM and Wit. On the other hand, there are offline libraries and open source that can also be explored like Vosk. Each company has its business rule and its specificity, in this sense it is difficult to know which is the most interesting for each situation. Thus, a comparison was made between speech recognition services in terms of usability, limitation and precision. In the comparison, speech recognition performance metrics were used in a set of audios, using the programming language Python.

KEYWORDS

Conjunto de Dados, Transcrição de Áudio, Sistemas de Reconhecimento de Fala, Interface de Programação de Aplicações

1 INTRODUÇÃO

A necessidade por reconhecimento de fala é algo cada vez mais presente no cotidiano. O conceito da Internet das Coisas (IoT) e da Indústria 4.0 eleva cada vez mais a proximidade do ser humano com a máquina. Nesse sentido, a crescente necessidade por ferramentas e recursos acarretou na abertura de um mercado de serviços *online* que realizam a tarefa de reconhecimento de fala. Pode-se citar, como exemplos, as empresas Microsoft, Google, IBM e Wit. Entretanto, em paralelo a esse mercado, existem iniciativas de código aberto que também buscam desempenhar essa atividade de forma *offline*, como CMU Sphinx e Vosk [1].

Dentre essas diversas opções, seja de código aberto ou fechado, cada serviço pode possuir vantagens e desvantagens. Pode-se afirmar que a quantidade de opções disponíveis, dificulta a tomada de decisão do consumidor. Nesse sentido, neste artigo, inicialmente, foi realizado um levantamento das principais ferramentas e os recursos atuais disponíveis. E utilizando a linguagem de programação Python, foram realizados testes com um conjunto de áudios para analisar o desempenho de cada utilitário. A Figura 1 exibe o processo geral da execução e a metodologia utilizada para a análise dos serviços selecionados.

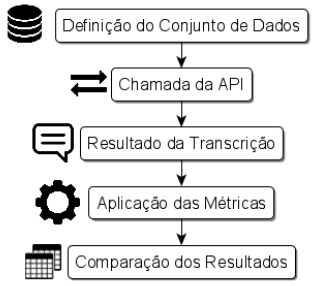

Figura 1: Processo geral da proposta de análise dos serviços de transcrição.

O conjunto de dados (*dataset*) é composto de trechos de cem áudios com duração inferior a vinte segundos. Os arquivos de áudios são gravações telefônicas de conversas entre operadores do setor elétrico. Nesse sentido, será evidenciado o desempenho geral dos serviços e, em paralelo, será avaliada a transcrição de palavras específicas do contexto técnico, como nome dos equipamentos, unidades de medidas, entre outros.

A principal forma de disponibilização dos serviços de transcrição de voz (*Speech to Text*) é por meio de Interface de Programação de Aplicações (*Application Programming Interface - API*), que se trata de um conjunto de tarefas que permite o envio de informações para um servidor, e a partir disso, obtém-se uma resposta (*request–response*). Com isso, as empresas prestadoras de serviço podem monitorar o uso de seus serviços e cobrar taxas caso for cabível.

O artigo está organizado da seguinte maneira: na Seção 2 estão ilustrados os requisitos necessários para o desenvolvimento da análise e quais os utilitários utilizados no artigo. Na Seção 3, os principais resultados são apresentados. Na Seção 4 são abordadas as conclusões; e por fim, na Seção 5, as considerações finais e os desenvolvimentos futuros são expostos.

2 MATERIAL E MÉTODOS

A principal ferramenta usada para a realização da transcrição do *dataset* foi a linguagem de programação Python. A utilização dessa ferramenta apresenta muitas vantagens, sendo as principais: *software* de código aberto (*open-source*), adaptação em diferentes sistemas operacionais, linguagem intuitiva e *user-friendly*, possibilidade de implementação de várias bibliotecas, manejamento de memória, operações em múltiplas *threads* e grande documentação básica [2].

Para a programação com Python, basicamente, é necessário ter definido três itens: o *Core*, o Ambiente de Desenvolvimento Integrado e as Bibliotecas. O *Core* é o interpretador da linguagem junto com algumas bibliotecas padrões. O Ambiente de Desenvolvimento Integrado (do inglês *Integrated Development Environment*) é a interface usada para o desenvolvimento, cita-se: Pyzo, PyCharm, Spider, Jupiter. E as bibliotecas são uma coleção de módulos de *script* que executam determinada atividade. No contexto atual, vale citar as seguintes bibliotecas e módulos:

- `numpy`: pacote para Python suportar *arrays* e matrizes multidimensionais.
- `glob`: módulo para encontrar todos os nomes de um determinado caminho.
- `pydub`: biblioteca para manipular áudio com uma interface simples e de alto nível.
- `auditok`: biblioteca para detecção de atividade de áudio.
- `matplotlib`: biblioteca abrangente para a visualização de dados.
- `pathlib`: biblioteca para manipular caminhos de sistema de arquivos de maneira independente.
- `SpeechRecognition`: biblioteca para realização de reconhecimento de fala, com suporte para diversos *engines* e APIs, *online* e *offline*
- `wave`: fornece uma interface para o formato de som WAV.
- `jiwer`: pacote Python para cálculo *Word Error Rate (WER)*, *Match Error Rate (MER)*, *Word Information Lost (WIL)* e *Word Information Preserved (WIP)* de uma transcrição.

Em relação ao formato dos áudios, são trechos de ligações (formato WAV) com um único canal e taxa de amostragem de 8.000 kHz, contendo um diálogo específico relacionado a operação do setor elétrico. Os serviços que foram selecionados para análise e comparação de resultados por meio da transcrição do *dataset* são:

- `Microsoft Azure`: reconhecimento de voz, permite a transcrição em tempo real e em lote de *streams* de áudio em texto [3].
- `Google Cloud`: usa tecnologia de reconhecimento de voz de código fechado com base em aprendizagem profunda [4].
- `Wit`: grátis até para uso comercial, suporta 130 línguas. Suporta as linguagens de programação Node, Python, e Ruby [5].
- `IBM Watson Speech to Text`: serviço aproveita o aprendizado de máquina para combinar o conhecimento da gramática, a estrutura do idioma e a composição de sinais de áudio e de voz e atualiza continuamente e refina sua transcrição à medida que recebe mais falas [6].
- `Vosk`: Suporta 17 línguas e dialetos, código-aberto e funciona *offline* [7].

Vale destacar que os testes foram realizados usando o plano gratuito disponibilizado pelas empresas citadas. No entanto, para algumas, existe a possibilidade de contratação de um plano pago com mais vantagens e menos limitações. A Tabela 1 mostra as limitações dos planos disponibilizados gratuitamente.

Tabela 1: Limitações dos planos gratuitos.

	Azure	IBM	Google	Wit	Vosk
Online/Pagos	Sim	Sim	Sim	Sim	Não
Limite/mês	300 min	500 min	60 min	-	∞
Limite/*request*	1 min	100 MB	1 min	20 seg	∞
Personalização	Sim	Sim	Não	Não	Sim

Com relação as métricas para análise das transcrições, adotou-se os métodos recomendados pela literatura, *Word Error Rate (WER)*, *Match Error Rate (MER)*, *Word Information Lost (WIL)*. [8]. A métrica mais comum é a *Word Error Rate (WER)*, que usa a soma de substituições (S), deleções (D), inserções (I) e divide pela soma de palavras da sentença correta (N) [9].

$$WER = \frac{S + D + I}{N} \tag{1}$$

Uma segunda métrica recomendada por [10] é a *Match Error Rate (MER)*, que é soma de S, D e I dividida pela soma dos acertos (H), S, D e I.

$$MER = \frac{S + D + I}{H + S + D + I} = 1 - \frac{H}{N} \tag{2}$$

E a terceira métrica recomendada por [8] é a *Word Information Lost (WIL)*, que é um menos os acertos ao quadrado (H*H) dividido pela multiplicação da quantidade de palavras na entrada (N_1) pela quantidade de palavras na saída (N_2).

$$WIL = 1 - (\frac{H}{N_1}\frac{H}{N_2}) \tag{3}$$

Para exemplificar o comportamento das métricas a Tabela 2 exibe o resultado em porcentagem dos critérios para determinadas entradas e saídas. Em que, X, Y e Z são palavras arbitrárias, I remete a uma inserção e D a uma deleção [8].

Tabela 2: Comparação WER, MER e WIL [8].

Entrada	Saída	H	S	D	I	%WER	%MER	%WIL
X	X	1	0	0	0	0	0	0
Xiii	XXYY	1	0	0	3	300	75	75
XYX	XZd	1	1	1	0	67	67	83
X	Y	0	1	0	0	100	100	100
Xi	YZ	0	1	0	1	200	100	100

Para a avaliação geral dos serviços foi considerada a média das métricas. Neste sentido, quanto menor o valor, melhor será a precisão geral do serviço. Adicionalmente às métricas, foi avaliado o desempenho dos serviços em relação a palavras-chaves relacionadas ao contexto do *dataset*. Foi verificado o desempenho da transcrição de dez palavras específicas e suas variações, estas palavras são termos técnicos comuns utilizados na operação do setor elétrico.

A Tabela 3 exibe as dez palavras-chaves selecionadas e a quantidade de ocorrências de cada uma no *dataset*.

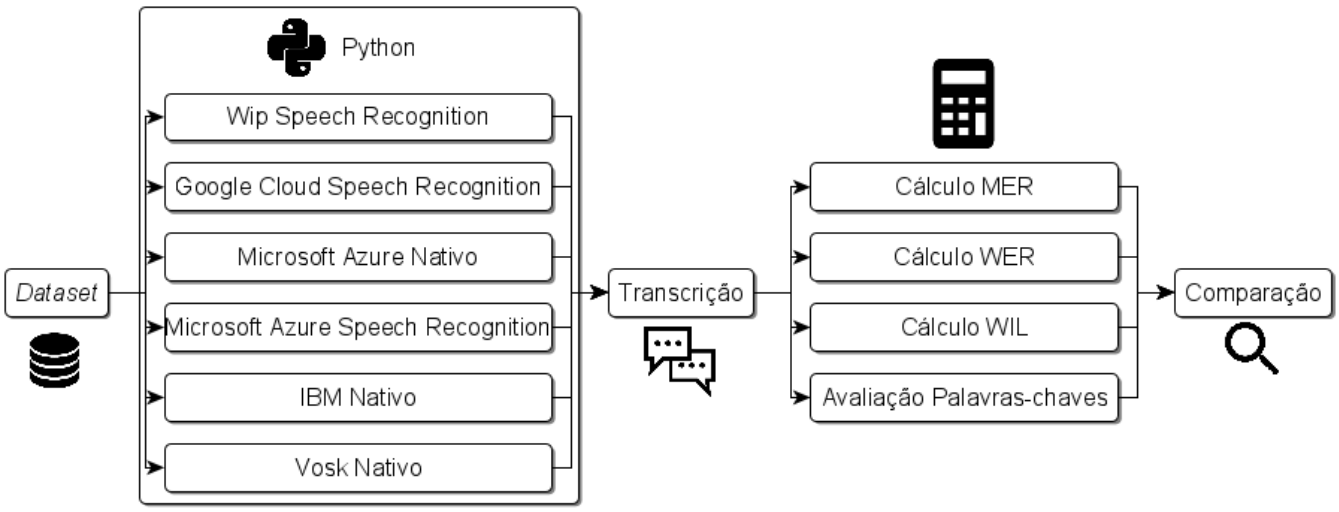

Figura 2: Processo de transcrição e análise.

Tabela 3: Listagem de palavras-chaves e suas ocorrências.

Ordem	Palavra-chave	nº de Ocorrências
i	kV	25
ii	MW	15
iv	disjuntor	6
v	tap	7
vi	tensão	15
vii	barra	10
viii	máquina	7
ix	usina	9
x	compensador	6

Tabela 4: Resultados das métricas WER, MER e WIL para cada serviço em porcentagem.

Serviço	Biblioteca	WER	MER	WIL	Total
Azure	SpeechRecognition	50	46	67	54
Azure	Nativamente	49	46	65	53
Google	SpeechRecognition	50	49	67	55
Wit	**SpeechRecognition**	**45**	**42**	**60**	**49**
Vosk	Nativamente	94	94	97	95
IBM	Nativamente	72	68	88	76

Para cada ocorrência de um termo específico foi verificado se ocorreu a transcrição correta. Nesse sentido, foi possível avaliar os serviços de forma geral e também de forma específica para o contexto proposto.

O processo adotado para a análise e comparação dos serviços é evidenciado na Figura 2. Conforme ilustrado, foi realizada a transcrição e análises usando 6 serviços:

- `Microsoft Azure` com a biblioteca Speech Recognition
- `Microsoft Azure` com a biblioteca nativa
- `Google Cloud` com a biblioteca Speech Recognition
- `Wit` com a biblioteca Speech Recognition
- `IBM Watson` com a biblioteca nativa
- `Vosk` com a biblioteca nativa

3 RESULTADOS

Por meio do uso da linguagem de programação Pyhon em conjunto com a IDE Spyder, foi realizada a transcrição do *dataset* e análise dos resultados de forma comparativa por meio das métricas propostas. A Tabela 4 sumariza os resultados das métricas para cada serviço testado.

A partir da Tabela 4 é possível constatar que o serviço que apresentou o melhor desempenho foi o Wit, seguido pela Microsoft Azure usando a biblioteca nativa da Microsoft. Vosk e IBM apresentaram os piores resultados, consecutivamente. A seguir é ilustrado um resumo do que foi observado dos serviços disponíveis, na ordem dos piores para os melhores, na métrica utilizada.

Vosk é uma biblioteca *offline* que funciona baseado em um modelo treinado. Para a execução das transcrições foi utilizado o modelo pt-BR disponibilizado no site gratuitamente. Esse modelo é relativamente pequeno e não foi treinado para executar transcrições de ligações telefônicas. Por isso a transcrição acabou não sendo efetiva (a biblioteca Vosk foi a única que não transcreveu todos os cem áudios do *dataset*, apenas 28 áudios), o que acarretou em uma perda significativa na sua avaliação por meio das métricas utilizadas (WER, MER e WIL).

Por outro lado, a IBM é um serviço consolidado no mercado. E, suspeita-se que o serviço gratuito não contempla um modelo voltado para transcrição de ligações telefônicas. Adicionalmente, o Google Cloud apresentou um desempenho mediano, mas que poderia ser melhorado com a contratação dos serviços específicos para ter acesso a modelos de transcrição de áudios de ligações telefônicas.

O serviço da Microsoft Azure foi testado de duas formas, com o uso da biblioteca nativa da empresa e com a biblioteca SpeechRecognition. Nesse sentido, foi possível comparar o desempenho do serviço usando as bibliotecas. A Tabela 4 indica que o uso da biblioteca nativa apresentou melhor desempenho. Logo, entende-se que a biblioteca SpeechRecognition deve realizar algum tipo de tratamento nos dados antes de realizar a requisição.

O Wit, por sua vez, foi o que apresentou o melhor desempenho dos serviços analisados. Pode-se afirmar que o Wit obtém boas respostas na transcrição de comandos rápidos. Entretanto, o serviço apresenta limitação na requisição (20 segundos de áudio por requisição) e não possibilita a personalização do serviço.

Em relação as palavras-chaves selecionadas para avaliação, foram encontradas no total 123 ocorrências no conjunto de dados transcritos. A Tabela 5 evidencia em porcentagem a quantidade de acertos de cada palavra para cada serviço. Note que quando o serviço tiver a sigla SR remete que a requisição foi feita usando a biblioteca SpeechRecognition e quando tiver a sigla N remete a biblioteca nativa do serviço.

Tabela 5: Avaliação geral de acertos de palavras-chaves em porcentagem.

Serviço	i	ii	iii	iv	v	vi	vii	viii	ix	x
Azure SR	24	**27**	83	50	**57**	0	**90**	**100**	**100**	67
Azure N	28	20	83	67	43	0	70	86	**100**	67
Google SR	**56**	0	74	**100**	29	**73**	70	71	78	67
Wit SR	44	20	**91**	**100**	14	67	90	**100**	**100**	**83**
Vosk N	0	0	0	0	0	0	0	0	0	0
IBM N	0	0	74	33	0	0	0	0	78	17

A Tabela 6 exibe o resultado total de acertos, em porcentagem, para cada serviço.

Tabela 6: Avaliação total de acertos de palavras-chaves em porcentagem.

Serviço	Total
Wit SR	**67**
Google SR	59
Azure SR	53
Azure N	50
IBM N	22
Vosk N	0

Mantendo as constatações realizadas anteriormente, o Wit apresentou os melhores valores de acerto das palavras-chave. Entretanto, em segundo lugar, para essa métrica, está o Google Cloud. Para as métricas anteriores (WER, MER e WILL) o segundo melhor serviço tinha sido o Azure usando a biblioteca nativa. Entretanto ao observar somente as palavras do contexto técnico (voltadas à operação do setor elétrico), o segundo melhor recurso é o da empresa Google.

Outra análise feita foi através de uma média do percentual de acertos para cada um dos dez termos técnicos selecionados para a avaliação das palavras-chaves. Pode-se observar num contexto geral quais palavras obtiveram maior número de acertos, e em contrapartida, quais palavras apresentaram maiores dificuldades para serem transcritas. A Tabela 7 exibe a porcentagem de acertos médio para cada uma das palavras-chaves.

Tabela 7: Porcentagem de acertos médio referente a cada palavra-chave.

Palavras-chaves	Média percentual
Usina	**76**
Máquina	71
Geração	68
Barra	58
Disjuntor	58
Compensador	50
Tap	33
Tensão	27
kV	26
MW	13

Comparando os resultados apresentados na Tabela 7, observa-se que as palavras com maiores dificuldades de transcrição foram: "MW", "kV"e "tensão", atingindo uma média percentual menor do que trinta porcento. As palavras com melhor desempenho não ultrapassaram os oitenta porcento, essa média pode ter sido prejudicada por conta dos serviços que apresentaram uma performance muito inferior.

Em termos gerais, deve-se levar em consideração que a análise está sendo feita a partir de áudios em gravações telefônicas. Portanto, trata-se de um ambiente não-controlado, isto é, um ambiente passível de interferências sonoras do local como vozes e ruídos. Na análise de palavras-chaves foram considerados termos técnicos, que talvez não sejam muito usuais na linguagem informal do português brasileiro, o que pode ter determinado resultados negativos se levar em consideração que os serviços utilizados não foram treinados para estas especificidades.

4 CONSIDERAÇÕES FINAIS

Todas as constatações realizadas são referentes a transcrição dos dados em questão, logo para análises mais abrangentes e gerais recomenta-se aumentar e diversificar o *dataset* explorado. O contexto atual se limita ao desempenho da transcrição dos serviços gratuitos em relação a comunicação verbal da operação do setor elétrico.[11].

Nos testes efetuados no artigo, pode-se observar que o Wit foi o que apresentou o melhor desempenho entre os serviços analisados. Entretanto, esse serviço é o que mais apresenta limitação por requisição, sendo que em cada requisição só pode ser enviado 20 segundos de áudio. Em relação a personalização dos modelos, as empresas Azure e IBM possibilitam a personalização e o Wit não. Apesar dessas limitações, deve-se levar em consideração que o objetivo do Wit é transcrever comandos rápidos e executar determinadas ações baseado na intenção do usuário.

Por outro lado, mesmo com desempenho insatisfatório, o Vosk é um serviço promissor e pode ser explorado com uma abordagem

diferente ao utilizado (modelo disponível no site). É possível usar um programa chamado Kaldi e aprimorar um modelo existente para obter resultados mais atraentes. Outra possibilidade é o uso do Vosk Server (*online*) e consumir o serviço por meio do uso de um *WebSocket* (tecnologia que permite a comunicação entre o cliente e um servidor) com um modelo voltado para chamadas telefônicas. Entretanto, até o momento, esse serviço não estava disponível para o público, podendo ser testado somente mediante contato com a empresa desenvolvedora.

REFERÊNCIAS

[1] Lawrence R. Rabiner B. H. Juang. Automatic speech recognition a brief history of the technology development, 2004.

[2] T. Viarbitskaya and A. Dobrucki. Audio processing with using python language science libraries. In *2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 350–354, 2018. doi: 10.23919/SPA.2018.8563430.

[3] Microsoft. Speech service documentation, 2020. URL https://docs.microsoft.com/en-us/azure/cognitive-services/Speech-Service/.

[4] Google. Speech-to-text, 2020. URL https://cloud.google.com/speech-to-text.

[5] Wit. Build natural language experiences, 2013. URL https://wit.ai.

[6] IBM. Speech to text, 2020. URL https://cloud.ibm.com/catalog/services/speech-to-text.

[7] Alpha Cephei. Vosk, 2020. URL https://alphacephei.com/vosk/.

[8] Phil Green Andrew C. Morris, Viktoria Maier. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *INTERSPEECH*, 2004.

[9] Alex Acero Xuedong Huang, Raj Reddy. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, 1nd. edition, 2001. ISBN 0-13-022616-5.

[10] Elias Pimenidis Ilias Maglogiannis, Lazaros Iliadis. benchmarking of ibm, google and wit automatic speech recognition systems. *Artificial Intelligence Applications and Innovations*, 2020. doi: 10.1007/978-3-030-49161-1_7.

[11] ONS. Comunicação verbal na operação, 2020. URL http://www.ons.org.br/%2FMPO%2FDocumento%20Normativo%2F4.%20Rotinas%20Operacionais%20-%20SM%2010.22%2F4.1.%20Rotinas%20Gerais%2F4.1.7.%20Relacionamento%20Operacional%2FRO-RO.BR.01_Rev.12.pdf.