Evaluation of Automatic Speech Recognition Systems

**Matheus Xavier Sampaio, Regis Pires Magalhães,
Ticiana Linhares Coelho da Silva, Lívia Almada Cruz,
Davi Romero de Vasconcelos, José Antônio Fernandes de Macêdo
and Marianna Gonçalves Fontenele Ferreira**

[1] Insight Data Science Lab - Universidade Federal do Ceará (UFC)
Fortaleza - CE - Brazil

{xavier,regis,ticianalc,livia,daviromero,jose.macedo,marianna}@insightlab.ufc.br

Abstract. *Automatic Speech Recognition (ASR) is an essential task for many applications like automatic caption generation for videos, voice search, voice commands for smart homes, and chatbots. Due to the increasing popularity of these applications and the advances in deep learning models for transcribing speech into text, this work aims to evaluate the performance of commercial solutions for ASR that use deep learning models, such as Facebook Wit.ai, Microsoft Azure Speech, and Google Cloud Speech-to-Text. The results demonstrate that the evaluated solutions slightly differ. However, Microsoft Azure Speech outperformed the other analyzed APIs.*

Resumo. *O Reconhecimento Automático de Fala (ASR) é uma tarefa essencial para muitos aplicativos, como geração automática de legendas para vídeos, pesquisa por voz, comandos de voz para casas inteligentes e chatbots. Devido à crescente popularidade desses aplicativos e aos avanços nos modelos de deep learning para transcrição de fala em texto, este trabalho tem como objetivo avaliar o desempenho de soluções comerciais para ASR que utilizam modelos de deep learning, como Facebook Wit.ai, Microsoft Azure Speech, e Google Cloud Speech-to-Text. Os resultados demonstram que as soluções avaliadas diferem ligeiramente. No entanto, o Microsoft Azure Speech superou as outras APIs analisadas.*

1. Introduction

Automatic Speech Recognition (ASR) techniques to transform speech-to-text [Reddy 1976] have gained increased importance in recent years and have applications in many problems, such as screen readers, automatic video and music captioning, etc [Graves et al. 2013]. One of those applications is chatbots, which gained popularity due to the adoption of messaging services and advances in Artificial Intelligence and Deep Learning.

This work proposes to evaluate the performance of the commercial APIs of ASR Facebook Wit.ai, Microsoft Azure Speech Services, and Google Cloud Speech-to-Text on a Portuguese dataset. The most used metric to evaluate ASR is the Word Error Rate (WER) [Këpuska and Bohouta 2017], however, it is limited to determine the rate of incorrect words in the transcription. In this work, we applied other NLP metrics to also validate if the models keep the original sentence structure and organization and if they generate transcriptions with similar vectorial representation.

Tabela 1. Comparison of models used by the ASR APIs

Paper	Architecture	Training Corpus	Test WER
Facebook AI Research [Baevski et al. 2020]	Encoder-Decoder built with fully connected CNN	1041 hours of audio combining the Wall Street Journal and Librispeech datasets in English	2.4% in the Wall Street Journal dataset
Microsoft AI and Research[Xiong et al. 2018]	CNN Encoder and BiLSTM Decoder	2000 hours of audio from the Switchboard dataset and 25 hours of audio from the CallHome dataset, both in English	5.1 % in the SwitchBoard dataset and 9.8 % in the CallHome dataset
Google AI[Chiu et al. 2018]	LAS with Multi-headed Attention	12500 hours of audio consisting of 15 million phrases taken from Google Voice Search in English	5.6% in Google Voice Search dataset

Related Works. Other works evaluated ASR models [Këpuska and Bohouta 2017, Filippidou and Moussiades 2020], but they aimed at evaluating performance in English and using only WER or WER in combination with the precision and recall of words [Filippidou and Moussiades 2020]. [de Lima and Da Costa-Abreu 2020] presents a survey of techniques and data sets for ASR in Portuguese. However, it does not offer an experimental evaluation of these techniques. The main contribution of this work is to offer a comparison of different ASR models according to different metrics. This may help data scientists to choose one of these available models.

2. ASR Models

Facebook AI Research developed Wav2vec 2.0 [Baevski et al. 2020] which applies the concept of unsupervised pre-training that learns a general representation of speech from unlabeled examples [Schneider et al. 2019]. The pre-trained models are fine-tuned for speech recognition by adding a projection layer at the top of the context network to predict one of the classes representing the vocabulary of the task.

The **Microsoft AI and Research** team proposed a speech recognition system composed of a combination of Convolutional Neural Networks (CNN) architectures Residual Network (ResNet) and Layer-wise Context Expansion with Attention (LACE), and Bi-directional Long Short Term Memory (Bi-LSTM) layers [Xiong et al. 2016, Xiong et al. 2018]. In addition, it adds language models based on LSTM at the word and character levels responsible for reclassifying the output at the end of the model.

The **Google AI** team proposed the Listen, Attend, and Spell (LAS) [Chan et al. 2016], that uses an Encoder-Decoder with Recurrent Neural Network (RNN). They improved LAS by adding a Multi-headed Attention layer, a new training metric based on the minimum rate of word errors, and the use of an external language model during inference [Chiu et al. 2018].

Table 1 presents a comparison between the architectures, training data, and test results of the models. The papers show the WER of the audio transcriptions on test sets selected for each work.

3. Experimental Setup

Datasets. The experiments use public and collaborative audio datasets in Portuguese, the Mozilla Common Voice[1] and the Voxforge[2] datasets. Mozilla Common Voice collaborators can record the audio and evaluate the available data quality. Voxforge is composed of sentences from audiobooks of the public domain. Table 2 presents the characteristics of the used datasets such as the number of recorded sentences, the size of vocabulary, the average time of audios, average length of sentences in terms of the number of characters, the original audio format, and frequency. We conduct the experiments on 10,000 sentences selected from these datasets, accounting for 11 hours and 40 minutes of audio. We chose all sentences with female or not informed voices and randomly selected from the remaining male voices to reduce the data imbalance, with the resulting distribution presented in Table 3.

Tabela 2. Characteristics of the datasets

Corpus	Number of sentences	Size of vocabulary	Average audio duration in seconds	Average sentence size in characters	Format and frequency of the original audio
Mozilla Common Voice	8014	8596	$4.43(\pm 1.42)$	$42.73(\pm 19.81)$	mp3 48KHz
Voxforge	4115	566	$3.66(\pm 1.19)$	$27.34(\pm 11.72)$	wav 48KHz

Tabela 3. Gender distribution of the voices in the datasets

Voice gender	Mozilla Common Voice	Voxforge
Male	4500 (75%)	2800 (70%)
Female	900 (15%)	200 (5%)
Not informed	600 (10%)	1000 (25%)

Metrics. The metrics used in this work assess the transcript quality by evaluating the difference between two sentences and their contexts. While the **WER** is the most used metric in ASR works, it is limited only to the transcription accuracy at the level of words. Using **BLEU** [Papineni et al. 2001] and **METEOR** [Banerjee and Lavie 2005], it is possible to evaluate whether the transcription maintains the context and organization of the sentence. BLEU is a metric originally proposed for neural machine translation that claims to be highly correlated with human assessment. METEOR was proposed to fix some limitations of BLEU, it is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. At the same time, the **Cosine Similarity** allows us to determine how close the two sentences are in a defined vector space. For the Cosine Similarity, we use the Word Embedding vectors produced in [Hartmann et al. 2017] by using the Word2Vec approach in both Continuous Bag of Words (CBOW) and Skip-Gram variations, with 50 dimensions.

[1] `https://commonvoice.mozilla.org/pt`
[2] `http://www.voxforge.org/pt`

4. Results

These experiments aim to evaluate the quality of Wit.ai, Azure Speech Services, and Google Cloud Speech to Text APIs transcriptions when applied to an extensive dataset in Portuguese. We have also tried to use the Amazon Transcribe service, but we decided not to use it because of the long transcription time for short audios compared to the other services. The metrics used in these experiments were WER, BLEU, METEOR, and Cosine Similarity. Tables 4 and 5 present a summary of the results of the experiments for the datasets.

The experiments showed a result for WER between 7.25% and 12.58%. Microsoft Azure Speech presented the best results for both the Mozilla Common Voice and Voxforge datasets, with 9.56% and 7.25%, respectively, which leads us to believe that this API has the least distortion between the original sentence and the transcribed text.

To obtain a more accurate picture of the APIs' ability to maintain sentence structure during transcriptions, we use BLEU and METEOR metrics. Scores between 0.83 and 0.90 for BLEU and 0.87 and 0.91 for METEOR show that the APIs can recognize the words in the sentences and maintain their organization to the original text. Finally, the cosine similarity values between 0.90 and 0.95 obtained for the Word Embeddings demonstrate that when evaluating the sentences taking them to a multidimensional vector space in which we abstract the word order, the texts produced by the APIs are pretty similar to the original sentences. The Skip-Gram variations obtained a marginally better result compared to CBOW. Again, Microsoft Azure Speech was superior to the other APIs.

Tabela 4. API results on Mozilla Common Voice Corpus

API	WER	BLEU	METEOR	Word2Vec CBOW	Word2Vec SKIP
Facebook Wit.ai	12.29%	0.831	0.881	0.911	0.914
Microsoft Azure Speech Services	**9.56%**	**0.871**	**0.909**	**0.927**	**0.929**
Google Cloud Text-to-Speech	12.58%	0.827	0.881	0.904	0.907

Tabela 5. API results on Voxforge Corpus

API	WER	BLEU	METEOR	Word2Vec CBOW	Word2Vec SKIP
Facebook Wit.ai	11.44%	0.856	0.873	0.919	0.920
Microsoft Azure Speech Services	**7.25%**	**0.900**	**0.906**	**0.946**	**0.947**
Google Cloud Text-to-Speech	10.49%	0.862	0.874	0.925	0.925

We also investigate the influence of voice gender on transcription quality. It is possible to observe that all APIs presented better results when recognizing male voices, with WER variations around 4% for Wit.ai and Azure Speech Services, and almost 6% for Google Cloud Speech-to-Text. This disparity in results by gender of voice is repeated in the other metrics, as shown in Figure 1.

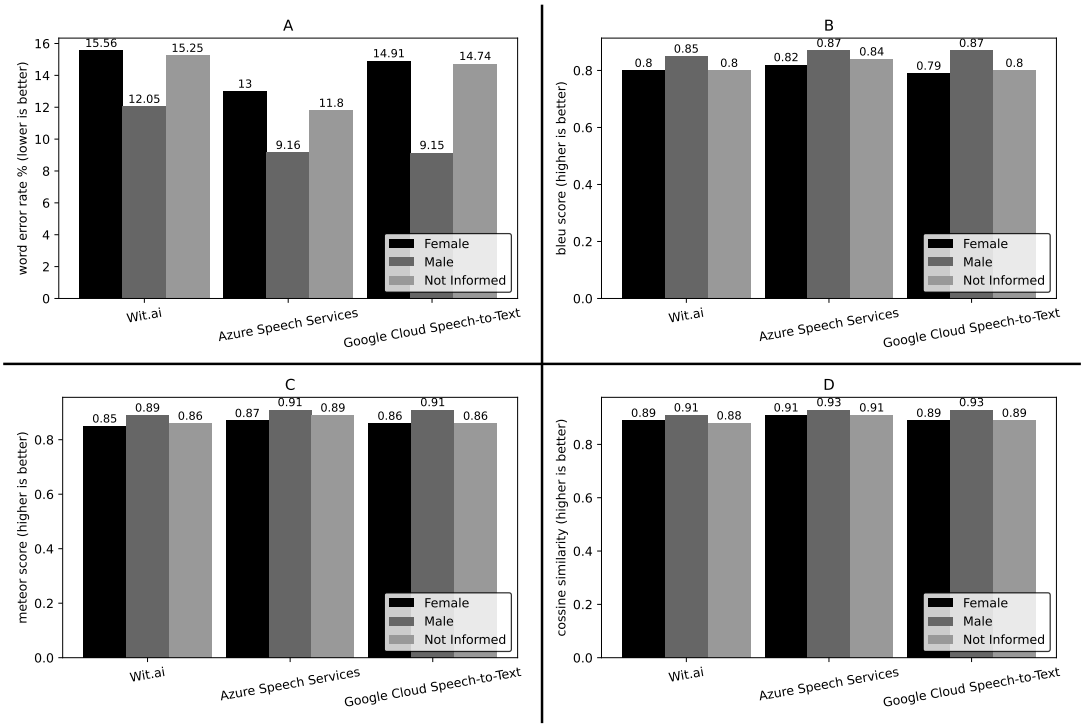

Figura 1. API WER results by Gender. A shows the WER percentages; B shows the BLEU Score; C shows the Meteor Score; D shows the Cosine Similarity

5. Conclusion

This work proposes an evaluation of speech recognition systems to be used to interact with chatbots through voice. In order to achieve this objective, we explored tools to execute the ASR tasks, obtained datasets, and studied metrics for carrying out tests and experiments.

After analyzing the techniques used by the APIs, we carried out experiments to assess the quality of speech transcription in Portuguese. We used WER, the primary metric for analyzing voice-to-text transcription, in addition to metrics that calculate the similarities between sentences, with text translation evaluation metrics BLEU and METEOR, and Cosine Similarity using Word Embeddings. For this, we used the datasets Mozilla Common Voice and Voxforge.

We can also observe the impact that the voice gender has on the accuracy of the transcriptions. The results showed similar performances between the tools in all metrics, with an advantage to Microsoft Azure Speech Services. The three APIs offer better WER values when transcribing male voices.

For future works, the comparison between the accent of different regions of Brazil could evaluate if it influences the quality of transcriptions. We also intend to train custom models using a diverse and localized dataset, using some open-source models, like Facebook's fairseq[3] and HuggingFace's Wav2Vec[4].

Acknowledgments. This work is partially supported by the FUNCAP projects 04772314/2020, 04772420/2020 and 04772551/2020.

[3]https://github.com/pytorch/fairseq

[4]https://huggingface.co/transformers/model_doc/wav2vec2.html

Referências

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020*, pages 12449–12460.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition.

Chiu, C.-C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., et al. (2018). State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE.

de Lima, T. A. and Da Costa-Abreu, M. (2020). A survey on automatic speech recognition systems for portuguese language and its variations. *Computer Speech & Language*, 62:101055.

Filippidou, F. and Moussiades, L. (2020). A benchmarking of ibm, google and wit automatic speech recognition systems. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 73–82. Springer.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Këpuska, V. and Bohouta, G. (2017). Comparing speech recognition systems (microsoft api, google api and cmu sphinx). *Int. J. Eng. Res. Appl*, 7(03):20–24.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Reddy, D. R. (1976). Speech recognition by machine: A review. *Proceedings of the IEEE*, 64(4):501–531.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech 2019*, pages 3465–3469.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (2016). Achieving human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., and Stolcke, A. (2018). The microsoft 2017 conversational speech recognition system. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5934–5938. IEEE.