A Framework for Speech Recognition Benchmarking

Franck Dernoncourt, Trung Bui, Walter Chang

Adobe Research, USA

{dernonco,bui,wachang}@adobe.com

Abstract

Over the past few years, the number of APIs for automated speech recognition (ASR) has significantly increased. It is often time-consuming to evaluate how the performance of these ASR systems compare with each other, and against newly proposed algorithms. In this paper, we present a lightweight, open source[1] framework that allows users to easily benchmark ASR APIs on the corpora of their choice. The framework currently supports 7 ASR APIs and is easily extendable to more APIs.

Index Terms: speech recognition, benchmark

1. Introduction

The performance of automated speech recognition (ASR) systems has drastically improved over the past few years, to the point that some studies report performance results that equal or outperform humans [1, 2, 3, 4]. These systems allow users to interact with machines by voice, and be more efficient than when typing [5], for example. As a result, the use of ASR is becoming increasingly commonplace and the number of ASR APIs has significantly increased.

These ASR APIs are used by three categories of users: researchers, developers, and end-users. Researchers may use these APIs to obtain performance baselines for their new ASR algorithms. Developers and end-users want to select the API that satisfies their requirements (e.g., in terms of accuracy, language, latency, privacy, customization, or price).

In this paper, we present a lightweight framework that allows these three categories of users to easily benchmark ASR APIs on the corpora of their choice.

2. The ASR Benchmark Framework

2.1. Overview

The framework is written in Python 3, and runs on Linux, macOS, and Microsoft Windows. It currently supports the following ASR APIs: Google Speech Recognition [6], Google Cloud Speech API [7], Houndify API [8], IBM Speech-to-Text [9], Microsoft Bing Speech-to-Text [10], Speechmatics [11], Wit.ai [12]. The framework is easily extendable to more APIs.

The required format for corpora is a list of pairs of speech files and gold transcriptions. The framework comes with an example corpus as well as scripts to convert well-known speech corpora into this format. Speech files may be FLAC, Ogg, MP3, or WAV files.

Figure 1 presents an overview of the system. Listing 1 gives an overview of the configuration file.

[1] https://github.com/Franck-Dernoncourt/ASR_benchmark

```
[general]
data_folder              =  ../example_dataset
transcribe               = true
asr_systems              = google,ibm
overwrite_transcriptions = false
evaluate_transcriptions  = true
speech_file_type         = wav
delay_between_transcript = 0
speech_language          = en-US
transcription_encoding   = UTF-8

[credentials]
bing_key            = [removed]
google_credentials  = [removed]
houndify_client_id  = [removed]
houndify_client_key = [removed]
ibm_username        = [removed]
ibm_password        = [removed]
speechmatics_id     = [removed]
speechmatics_token  = [removed]
wit_ai_key          = [removed]
```

Listing 1: *Configuration file used to define a benchmark in the framework. This is the only file the user has to modify. The* `dataset_folder` *defines the location of the folder,* `transcribe` *indicates whether the speech files should be transcribed,* `asr_systems` *lists which ASR API(s) should be called,* `overwrite_transcriptions` *specifies whether a speech file that has already been transcribed should be transcribed again, and* `evaluate_transcriptions` *indicates whether the framework should compute performance metrics once the predicted transcriptions have been collected.*

2.2. Performance Metrics

The framework is provided with a performance assessment script that computes ASR metrics comparing the predicted transcriptions with the reference transcriptions. Tables 1 and 2 present some performance metrics of several ASR APIs on the publicly and freely available Common Voice [13] and LibriSpeech [14] corpora, as well as two internal corpora (Adobe Stock, which corresponds to short image search queries, and Image Edit Requests, which as its name indicates corresponds to short oral requests to edit an image). The corpus LibriSpeech is divided into two subsets: LibriSpeech-clean, and LibriSpeech-other. The former contains "clean" speech while the latter contains "more challenging" speech. For all these corpora, we only use the official test sets.

We wish to emphasize that the results we present do not aim at ranking existing ASR APIs, since the performance may be affected by whether the corpus was used as part of the training set. Also, different APIs may differ on how well they handle languages other than English, speaker accents, background noise, etc. Instead, the results we present aim at demonstrating the use of the benchmarking framework.

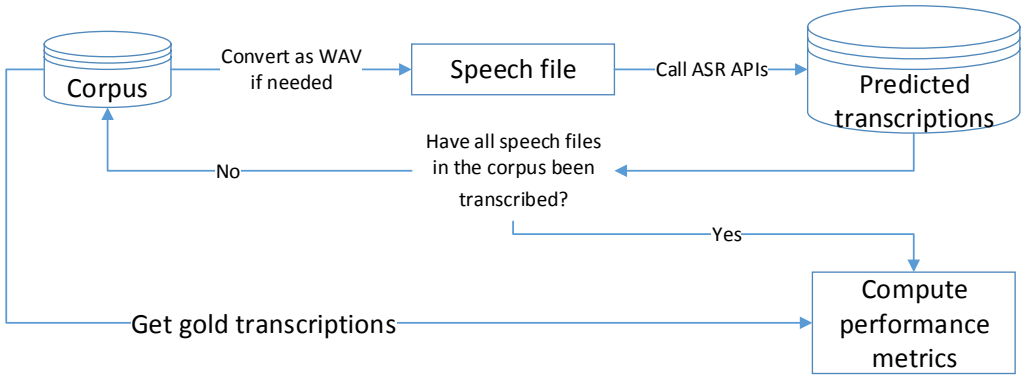

Figure 1: *Overview of the ASR benchmarking framework. First, the user has to provides a corpus that contains speech files with their reference (gold) transcriptions. The framework then converts each file to WAV format if needed, and calls the ASR APIs. When all speech files have been transcribed, the framework computes a set of performance metrics (e.g., word error rate) by comparing the predicted transcriptions with the gold transcriptions.*

Table 1: *Benchmark results presenting the word error rates expressed in percentage for several ASR APIs on the following 5 corpora: AS = Adobe Stock (4:28:05, 3184); CV = Common Voice (total length: 4:58:32, divided into 3995 speech files); IER = Image Edit Requests (2:29:09, 1289); LS-c = LibriSpeech clean (1:53:37, 870); LS-o = LibriSpeech other (5:20:29, 2939). Please refer to the GitHub repository (see footnote on page 1) for the most up-to-date and comprehensive benchmarks.*

API	CV	AS	IER	LS-c	LS-o
Google	23.2	24.2	16.6	12.1	28.8
Google Cloud	23.3	26.3	18.3	12.3	27.3
IBM	21.8	47.6	24.0	9.8	25.3
Microsoft	29.1	28.1	23.1	18.8	35.9
Speechmatics	19.1	38.4	21.4	7.3	19.4
Wit.ai	35.6	54.2	37.4	19.2	41.7
Human				5.8	12.7

Table 2: *Number of insertions, deletions, and substitutions when computing the word error rate on the predicted transcriptions for the LibriSpeech clean test corpus, which contains 18,533 tokens.*

API	Deletions	Insertions	Substitutions
Google	330	246	1614
Google Cloud	243	303	1741
IBM	166	269	1386
Microsoft	517	366	2595
Speechmatics	165	171	1018
Wit.ai	518	439	2604

3. Conclusion and Future Work

In this article we have presented a framework to benchmark ASR APIs. The framework is lightweight and easy to use: we hope it will make it more convenient for developers, end-users, and researchers to decide which ASR API to use for their needs and quickly compute some baseline performance for existing or new ASR corpora.

4. References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[3] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos *et al.*, "End to end speech recognition in english and mandarin," 2016.

[4] E. Edwards, W. Salloum, G. P. Finley, J. Fone, G. Cardiff, M. Miller, and D. Suendermann-Oeft, "Medical speech recognition: reaching parity with humans," in *International Conference on Speech and Computer*. Springer, 2017, pp. 512–524.

[5] S. Ruan, J. O. Wobbrock, K. Liou, A. Ng, and J. Landay, "Speech is 3x faster than typing for english and mandarin text entry on mobile devices," *arXiv preprint arXiv:1608.07323*, 2016.

[6] "Google Chrome's Speech API," https://www.chromium.org/developers/how-tos/api-keys, Accessed: March 3, 2018.

[7] "Google Cloud Speech API," https://cloud.google.com/speech, Accessed: March 3, 2018.

[8] "Houndify API," https://houndify.com, Accessed: March 3, 2018.

[9] "IBM Speech-to-Text," http://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/speech-to-text.html, Accessed: March 3, 2018.

[10] "Microsoft Bing Speech-to-Text API," https://www.microsoft.com/cognitive-services/en-us/speech-api, Accessed: March 3, 2018.

[11] "Speechmatics API," https://speechmatics.com, Accessed: March 3, 2018.

[12] "Wit.ai," https://wit.ai, Accessed: March 3, 2018.

[13] "Common Voice," https://voice.mozilla.org, Accessed: March 3, 2018.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.