

Evaluation of Audio Transcription Services

WP6 - *Automated media analysis*

1 Introduction

Nowadays, a range of automatic audio transcription services are offered by providers such as Google or Amazon. Most of these services are cloud-based and support transcription for a fixed set of languages, with the exception of CMU Sphinx – a language independent transcription framework that can be used with any language for which a language model can be provided. In contrast cloud-based services work out of the box and allow the user to offload non-negligible computational loads to an external device. We evaluate the suitability of four transcription services on a small dataset spanning 13 languages.

2 Dataset

The key concept audio handles the automated transcription of 14 interviews convering 13 languages. These interviews are available in either audio or video form. Two samples contain Russian language, the first one being purely russian (labeled *Russian(1)* in Figures 1, 2 and 3) and the other consisting of a Russian/Ukrainian mix (labeled *Russian(2)*). The second is considered to be Russian since transcription tests show considerably worse results if the origin language is set to Ukrainian. The full set of evaluated languages is as follows: English, Spanish, French, Italian, Russian, Dutch, Hebrew, Greek, Polish, Hungarian, Serbian, Slovenian and Czech. Interviews are between 1.5 and 5 hours long and are mostly audio encoded in the **mp3** format. Four samples (Italian, Hungarian, Serbian and Polish) are video that is encoded in the **mpg** format. In order to minimize the risk that personal information of interviewees is exposed to cloud-based services the first and last 15 minutes of each interview are discarded during preprocessing. Ground truth data is generated by professional human transcriptionists and is available in the **pdf** format. It is converted into plain text to facilitate further automated processing. This conversion step may introduce a small number of artificial errors since inconsistently formatted additional information such as timestamps and comments need to be removed.

3 Service Overview

This evaluation tests the suitability of the four services *Google Cloud Speech-to-Text*¹, *Azure Speech-to-Text*², *Amazon Transcribe*³ and *CMU Sphinx*⁴. All transcriptions are

¹<https://cloud.google.com/speech-to-text/>

²<https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/>

³<https://aws.amazon.com/de/transcribe/>

⁴<https://cmusphinx.github.io/>

obtained through the respective Python APIs. Utility, ease of use and to a lesser extent transcription speed may differ if APIs for other programming languages are used instead.

Google Cloud Speech-to-Text: With a total of 120 languages and variants, and support for all of the languages considered here, Google Cloud Speech-to-Text is the best option in terms of availability. It further distinguishes itself with an easy to use API and fast transcription. By default data logging is disabled and its cost is set to 0.024 \$/min . The data logging option allows Google to record audio data sent to Cloud Speech-to-Text and reduces the cost to 0.016 \$/min. For audio clips longer than one minute in length the API only allows asynchronous requests which require the file to be uploaded to Google Cloud Storage. Shorter audio clips can be transcribed through a synchronous request without explicit use of cloud storage. This service specifically discourages the use of lossy audio compression as used in the `mp3` format.

Azure Speech-to-Text: The transcription service offered by Microsoft Azure supports 39 languages and variants in total and 7 of the evaluated languages. Its cost is set to 0.0167 \$/min. Both synchronous and asynchronous transcriptions are available for audio clips of arbitrary length. The documentation does not specify whether data is kept after the transcription process has finished or if it is used for training purposes. While transcription is slightly slower, this service allows the user to run up to 20 transcription jobs in parallel.

Amazon Transcribe: Amazon Transcribe offers 31 languages and variants with support for 7 of the evaluated languages at a cost of 0.024 \$/min. This service is tightly integrated into the AWS ecosystem, a transcription job can only be started on data that has been uploaded to the Amazon S3 cloud storage service. Although transcription speed compared to other services is slow, Amazon Transcribe allows the user to start up to 100 concurrent transcription jobs. Any voice inputs processed by the service may be stored and used by AWS for the training of speech recognition models.

CMU Sphinx: The offline software CMU Sphinx is in principle language independent, it can transcribe any language for which a language model is available. Models for 15 languages (7 in the evaluated language set) have been collected by the developers and can be downloaded for free⁵. Transcription speed varies to a great extent between languages. The number of parallel transcription jobs is limited by the available computing power.

Other services such as *AssemblyAI* and *IBM Watson* have been initially considered but disregarded due to lacking language support. An overview of costs, transcription speed and summarized language support is given in Table 1 while Table 2 shows language availability in detail.

⁵<https://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/>

Service	Cost	Languages	Transcription Speed
Google Cloud Speech-to-Text	0.0240 \$/min	13/13	5.0 min/min up to 1 parallel job
Azure Speech-to-Text	0.0167 \$/min	7/13	2.0 min/min up to 20 parallel jobs
Amazon Transcribe	0.0240 \$/min	7/13	0.5 min/min up to 100 parallel jobs
CMU Sphinx	free (offline)	7/13	0.4 – 4.2 min/min measured on local device

Table 1: Overview of cost, language support and transcription speed (ratio of audio duration to transcription duration) of the compared services.





















































Language	Google Cloud STT	Azure STT	Amazon Transcribe	CMU Sphinx
English	 Yes	 Yes	 Yes	 Yes
Spanish	 Yes	 Yes	 Yes	 Yes
French	 Yes	 Yes	 Yes	 Yes
Italian	 Yes	 Yes	 Yes	 Yes
Russian	 Yes	 Yes	 Yes	 Yes
Dutch	 Yes	 Yes	 Yes	 Yes
Hebrew	 Yes	 No	 Yes	 No
Greek	 Yes	 No	 No	 Yes
Polish	 Yes	 Yes	 No	 No
Hungarian	 Yes	 No	 No	 No
Serbian	 Yes	 No	 No	 No
Slovenian	 Yes	 No	 No	 No
Czech	 Yes	 No	 No	 No

Table 2: Detailed language availability for each service.

4 Metrics

The evaluation of transcription quality is a non-trivial matter. Metrics such as Word Error Rate can only be computed if two given word sequences are not too dissimilar, while tf-idf fails whenever a word sequence cannot be separated into discrete sentences. Some metrics measure similarity on a word-by-word basis while others make statistical assertions based on the respective vocabularies. Metrics corresponding to both approaches have been used to ensure that at least one measurement can be made for each transcribed document and transcription service.

Word Error Rate

The standard measure [1, 2, 4] in speech recognition evaluation is *Word Error Rate*, or WER for short. In general it is not guaranteed that the length of a recognized word sequence is identical to its correct reference sequence. For this reason it is necessary to perform a string alignment step that attempts to line up identical words occurring in both sequences. Any word for which no counterpart is found falls into one of three categories:

- **Substitution:** A word that has been recognized in place of a different word in the reference sequence.
- **Deletion:** A word that occurs in the reference sequence but is missing in the recognized sequence.
- **Insertion:** A word that occurs in the recognized sequences but does not exist in the reference sequence.

The Word Error Rate is then defined as:

$$\text{WER} = \frac{S + D + I}{N} \quad (1)$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the total number of words in the reference sequence. In general a lower Word Error Rate indicates better transcription accuracy. The string alignment step may fail if the recognized word sequence diverges too far from the reference sequence. Other metrics are considered in order to retain comparability in such cases, however we consider WER to be the most meaningful metric.

tf-idf and Term Frequency

The tf-idf metric is a statistical quantity that indicates how important a word is to a single document in a collection of documents. It is frequently used for document classification and the evaluation of keyword detection algorithms [5, 3]. The name tf-idf is composed of abbreviations for the terms Term Frequency (tf) and Inverse Document Frequency (idf). If the number of occurrences of the term t in the document D is denoted $\#(t, D)$ then the Term Frequency is defined as

$$\text{tf}(t, D) = \frac{\#(t, D)}{\max_{t' \in D} \#(t', D)}. \quad (2)$$

It is the normalized frequency of t in D . The main advantage of this metric is that it can be computed for any two non-empty word sequences.

The inverse document frequency measures the amount of information obtained by the occurrence of a word in a document in relation to the entire corpus. For a term t and a corpus \mathcal{C} it is defined as

$$\text{idf}(t, \mathcal{C}) = -\log \frac{|\{D \in \mathcal{C} : t \in D\}|}{|\mathcal{C}|}. \quad (3)$$

The combination of Term Frequency and Inverse Document Frequency yields tf-idf:

$$\text{tf-idf}(t, D, \mathcal{C}) = \text{tf}(t, D) \text{idf}(t, \mathcal{C}). \quad (4)$$

For the purposes of this evaluation the term “document” refers to a single sentence in a word sequence and “corpus” refers to the entire word sequence. In direct consequence a tf-idf score can only be computed if the transcription service provides punctuation symbols.

Both tf-idf and Term Frequency are transformations that allow the conversion of word sequences into real valued vectors. After enumerating the total set of words that occur in two (or more) word sequences, the i -th entry of such a vector is calculated as the tf-idf score (or Term Frequency) of the i -th word. The similarity score between two vectors \mathbf{a}, \mathbf{b} is then calculated as their cosine similarity:

$$\text{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}. \quad (5)$$

5 Evaluation

The Word Error Rate metric is considered to be the default measure in automated speech recognition. In contrast to Term Frequency and tf-idf it is based on transcription success on a word-by-word basis. We consider transcription service rankings based on this metric to be the most meaningful. Evaluation results based on WER are shown in Figure 1 and Table 3. Excluding two languages, English and French, the service Google Cloud Speech-to-Text delivers the best results while scores for the two exceptions are still competitive.

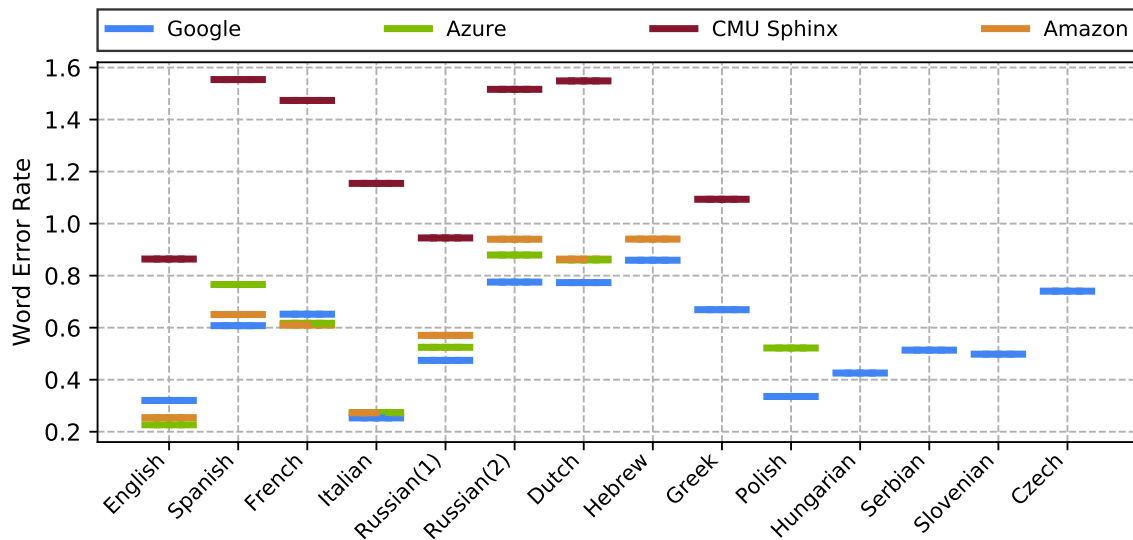


Figure 1: The word error rate achieved by services across all languages (lower is better).

Language	Google Cloud STT	Azure STT	Amazon Transcribe	CMU Sphinx
English	0.320	0.227	0.254	0.864
Spanish	0.608	0.766	0.650	1.554
French	0.652	0.617	0.610	1.473
Italian	0.253	0.274	0.273	1.155
Russian(1)	0.474	0.524	0.570	0.945
Russian(2)	0.775	0.879	0.940	1.516
Dutch	0.773	0.860	0.864	1.549
Hebrew	0.859	–	0.941	–
Greek	0.669	–	–	1.094
Polish	0.335	0.522	–	–
Hungarian	0.426	–	–	–
Serbian	0.514	–	–	–
Slovenian	0.498	–	–	–
Czech	0.740	–	–	–

Table 3: The word error rate achieved by services across all languages (lower is better).

In contrast to WER, the metrics Term Frequency and tf-idf merely measure the similarity of term distributions between two word sequences. Such a score would remain unchanged if, for example, the words in all sentences of a word sequence were reversed. Results based on tf-idf are shown in Figure 2 and Table 4 while results based on Term Frequency are shown in Figure 3 and Table 5. For tf-idf the ranking between services is less clear, with Google Cloud Speech-to-Text, Azure Speech-to-Text and Amazon Transcribe scoring generally similar and each having the highest score for more than one language. For some service/language pairs this metric could not be computed due to missing separation of the transcribed text into individual sentences as this feature was not always offered.

The measurements in Figure 3 are more spread out with Azure Speech-to-Text and Amazon Transcribe mostly scoring the highest, however common words such as “a” or “the” are assigned a disproportionately high term frequency score despite their low importance to a documents content. For this reason a score calculated solely based on Term Frequencies has limited informative value.

6 Conclusion

We compared the suitability of four automated transcription services on a dataset containing 13 languages. Appropriate metrics have been identified and it has been determined that, if applicable, the Word Error Rate measure give the best reflection of transcription quality. Based on quantitative results using this metric as well as due to its superior language support compared to other services in the lineup we determined that Google Cloud Speech-to-Text is best suited for the automated transcription of audio data related to the VHH project.

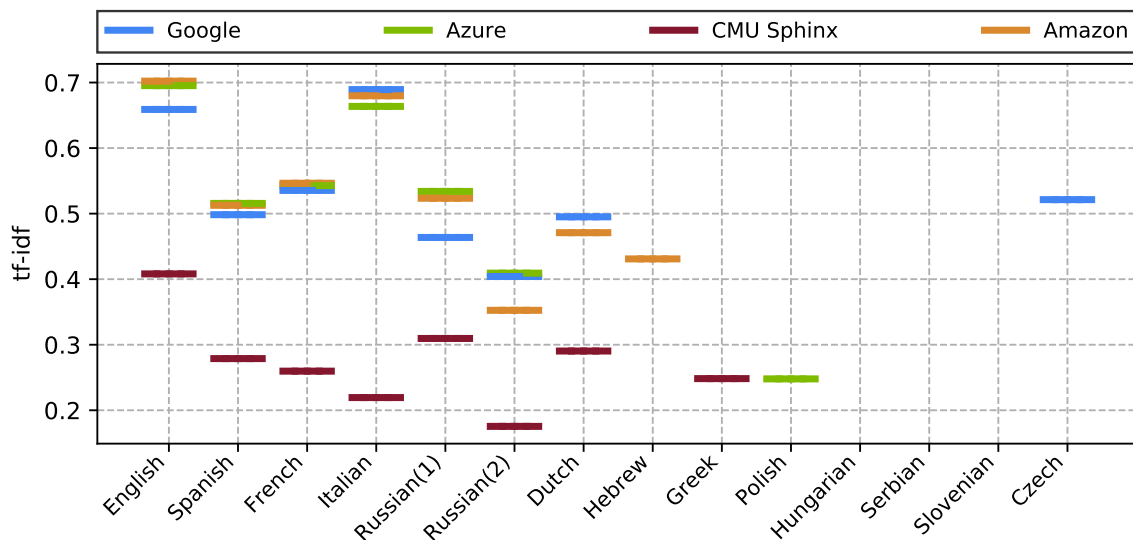


Figure 2: The tf-idf scores achieved by services across all languages (higher is better).

Language	Google Cloud STT	Azure STT	Amazon Transcribe	CMU Sphinx
English	0.659	0.695	0.702	0.408
Spanish	0.498	0.516	0.513	0.279
French	0.535	0.543	0.546	0.260
Italian	0.689	0.664	0.680	0.219
Russian(1)	0.464	0.534	0.523	0.310
Russian(2)	0.404	0.409	0.353	0.175
Dutch	0.495	—	0.471	0.290
Hebrew	—	—	0.431	—
Greek	—	—	—	0.248
Polish	—	0.248	—	—
Hungarian	—	—	—	—
Serbian	—	—	—	—
Slovenian	—	—	—	—
Czech	0.521	—	—	—

Table 4: The tf-idf scores achieved by services across all languages (higher is better).

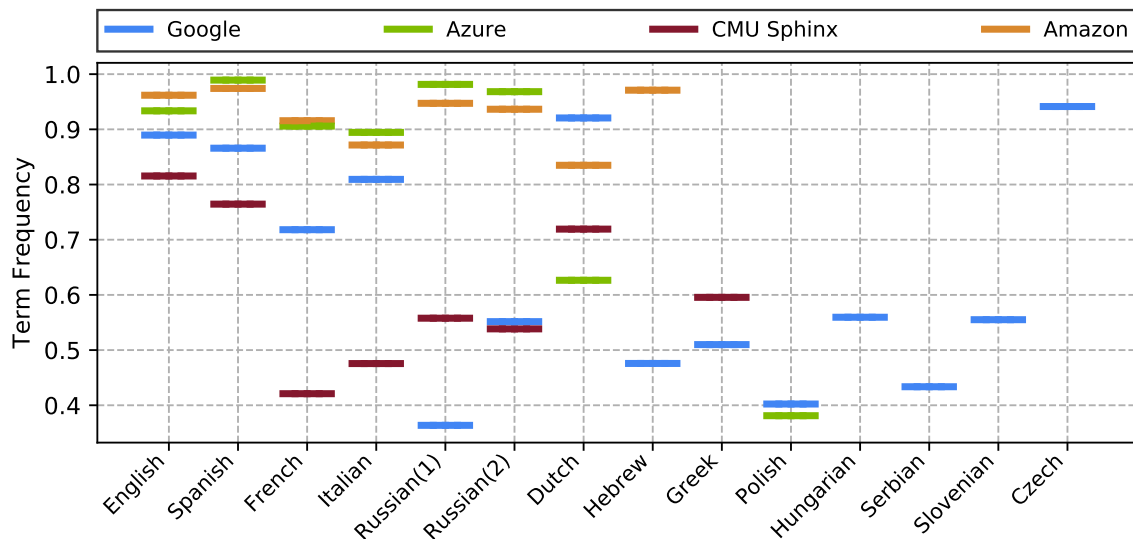


Figure 3: The Term Frequency scores achieved by services across all languages (higher is better).

Language	Google Cloud STT	Azure STT	Amazon Transcribe	CMU Sphinx
English	0.890	0.934	0.962	0.815
Spanish	0.866	0.989	0.974	0.764
French	0.718	0.906	0.916	0.421
Italian	0.809	0.894	0.872	0.476
Russian(1)	0.364	0.981	0.947	0.558
Russian(2)	0.552	0.968	0.936	0.538
Dutch	0.921	0.626	0.835	0.719
Hebrew	0.476	–	0.971	–
Greek	0.510	–	–	0.596
Polish	0.402	0.381	–	–
Hungarian	0.560	–	–	–
Serbian	0.434	–	–	–
Slovenian	0.555	–	–	–
Czech	0.941	–	–	–

Table 5: The Term Frequency scores achieved by services across all languages (higher is better).

References

- [1] Laila Dybkjaer, Holmer Hemsén, and Wolfgang Minker. *Evaluation of Text and Speech Systems*. 1st. Springer Publishing Company, Incorporated, 2007. ISBN: 1402058152.
- [2] María González et al. “An Illustrated Methodology for Evaluating ASR Systems”. In: *Adaptive Multimedia Retrieval. Large-Scale Multimedia Retrieval and Evaluation*. Ed. by Marcin Detyniecki et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 33–42. ISBN: 978-3-642-37425-8.
- [3] A. A. Hakim et al. “Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach”. In: *2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE)*. Oct. 2014, pp. 1–4. DOI: 10.1109/ICITEED.2014.7007894.
- [4] Youngja Park et al. “An empirical analysis of word error rate and keyword error rate”. In: *INTERSPEECH*. 2008.
- [5] Shahzad Qaiser and Ramsha Ali. “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents”. In: *International Journal of Computer Applications* 181 (July 2018). DOI: 10.5120/ijca2018917395.