Evaluating Google Speech-to-Text API's Performance for Romanian e-Learning Resources

Bogdan IANCU
The Bucharest University of Economic Studies, Romania
bogdan.iancu@ie.ase.ro

This paper presents a way of performing ASR on multimedia e-learning resources available in Romanian with the usage of the Google Cloud Speech-to-Text API. The material presents the history of ASR systems together with the main approaches used by the algorithms behind these systems. The cloud computing providers, that offer ASR solutions via SaaS, are analyzed as well. After performing a short literature review, the author focuses on applying the Google Cloud Speech-to-Text API on various video e-learning resources available online on YouTube. By doing this, the resources can be easily indexed and transformed into searchable materials. The WER score is used in order to measure the accuracy of the model and to compare it with similar works. The results are more than satisfying, thus the proposed model can be used as a method of automating the indexing of multimedia e-learning resources.
Keywords: *ASR, Speech-to-text, Romanian, WER, E-learning*

1 Introduction

Automatic Speech Recognition, abbreviated ASR, is not a new topic in the Computer Science field. In fact, the first steps were done in the '50s when a digit recognition system was developed by AT&T Bell Labs [1]. However, it was the 2000s when the speech recognition systems have successfully started to understand large vocabularies in uncontrolled environments. Table 1 summarizes the evolution of ASR systems. Most probably, in the following years, the humanity will face real time speech recognition systems that will achieve an accuracy similar to the one of a native speaker [2]. But, until then, let us see the existing ASR related algorithms.

Table 1. Growth of ASR Systems [2]

Year	Progress of ASR Systems
1952	Digit Recognizer
1976	1000 word connected recognizer with constrained grammar
1980	1000 word LSM recognizer (separate words w/o grammar)
1988	Phonetic typewriter
1993	Read texts (WSJ news)
1998	Broadcast news, telephone conversations
1998	Speech retrieval from broadcast news
2002	Rich transcription of meetings, Very Large Vocabulary, Limited Tasks, Controlled Environment
2004	Finnish online dictation, almost unlimited vocabulary based on morphemes
2006	Machine translation of broadcast speech
2008	Very Large Vocabulary, Limited Tasks, Arbitrary Environment
2009	Quick adaptation of synthesized voice by speech recognition (in a project where TKK participates in)

DOI: 10.12948/issn14531305/23.1.2019.02

Year	Progress of ASR Systems
2011	Unlimited Vocabulary, Unlimited Tasks, Many Languages, Multilingual Systems for Multimodal Speech Enabled Devices
Future Directions	Real time recognition with 100% accuracy, all words that are intelligibly spoken by any person, independent of vocabulary size, noise, speaker characteristics or accent

There are several approaches, in terms of algorithms, that were or are still used for performing ASR [2]:

- *template-based approaches* focus on matching unknown speech with pieces of already known information;
- *knowledge-based approaches*, where variations in speech are saved into the system and more complex rules are deduced by an inference engine;
- *neural network-based approaches* which use neural network AI algorithms to automatically detect speech based on training data;
- *dynamic time warping-based approaches* that focus on matching sequences of the same speech that vary in time or speed;
- *statistical-based approaches* that use automatic learning procedures on large amounts of training data; the most popular statistical algorithm is the Hidden Markov Model (abbreviated HMM) which is the current state-of-the-art.

Nevertheless, this paper scope is not to develop an ASR algorithm for Romanian. There are several attempts in this direction like [3] and [4], but because Romanian is categorized as an under-resourced language [5] it would be relatively hard to develop a complete ASR algorithm with a satisfying performance. Instead of that, we will focus in the next part of this chapter on the ASR algorithms available in cloud as part of the main cloud computing providers' portfolio. There are five major players in this field, each one of them with its own personal assistant: Google with the so called Google Assistant, Apple with Siri, Microsoft with Cortana, Amazon with Alexa and IBM with Watson. From these five players, Apple doesn't have a cloud service that allows the usage of the technology behind Siri and IBM doesn't have an assistant per se, Watson being used as a term with larger connotations. Below are presented the main cloud computing providers that offer ASR algorithms in the form of SaaS.

Amazon provides a cloud service called Amazon Transcribe that focuses on ASR. It was launched in 2018 and supports five languages in different dialects: English, Spanish, French, Italian and Portuguese [6]. Unfortunately, the support for Romanian was not announced yet and it is still questionable if it will be.

Microsoft offers a Speech to Text component in its Azure Cloud Service [7]. This one has support for six languages in different dialects (English, Chinese, French, German, Italian and Spanish), but again, Romanian is not among them.

IBM has integrated also in the Watson Cloud Services a speech to text algorithm. It supports the largest number of languages among the systems presented so far (Portuguese, French, German, Japanese, Korean, Mandarin Chinese, Modern Standard Arabic, Spanish, UK English and US English), but still no support for Romanian [8].

Google is the only cloud provider that supports Romanian among the languages available in the Cloud Speech-to-Text component. The component offers ASR support for around 120 languages and dialects [9] and it will be the one that we will focus on in the following chapters of this paper.

2 Related Work

This chapter focuses on the literature review in the field of ASR in general, ASR applied on Romanian, in particular, and Google Speech Recognition API alternatives. Therefore, authors in [5] do a survey of the under-resourced languages for speech recognition. Romanian

DOI: 10.12948/issn14531305/23.1.2019.02

is identified as being one of them, among other European languages like Croatian, Icelandic, Latvian, Lithuanian and Maltese. Another problem raised by this paper, that affects Romanian too, is the use of diacritics. Even if human readers can easily identify texts without diacritics, these texts cannot be used as training data for an ASR algorithm. The paper also proposes WER (Word Error Rate) as the main metric for evaluating ASR performance.
Paper [3] presents some approaches for performing ASR for Romanian language. The authors use a training dataset of 3300 phrases uttered by 11 speakers, 7 males and 4 females. They use the WRR (Word Recognition Rate) as main metric of their algorithm and obtain the highest recognition rate, 90.41%, by using cepstral analysis. This is one of the best results obtained so far by an ASR algorithm on Romanian and will be the point of reference for our approach.
In paper [2] are presented the main challenges related to ASR. A classification of the speech recognition algorithms is done, together with a presentation of the evolution of ASR systems (both topics are already presented in the introduction). Also, the results obtained in [3] are listed here as the main accomplishment for Romanian.
In [4] the authors try to enhance ASR for Romanian by using machine translated and web-based text corpora. They are using a training dataset of 644 phrases uttered by 21 speakers, or, in terms of time, more than 6.5 hours of speaking. The lowest WER obtained was 20.7% when the largest dataset (europarl + 9am.DRS2 + hotnews.DRS2) was used. We will not focus on this approach in the current paper, first of all because we are not developing an ASR algorithm per se and second of all because Google is already using web-based text corpora for training its speech-to-text algorithm. The result can be however usefully for comparison purposes.
Paper [10] measures the accuracy of the Google Web Speech API in recognizing and transcribing words spoken by Japanese English learners. The tests are performed on simple English phrases. The results present Google's API with a mean accuracy of 89.4 for native speakers and 65.7 for non-native speakers. No alternatives to Google's API are presented in this paper and because the authors do not specify if standard metrics (WER, WRR) are used for evaluating the algorithm's performance and what they mean by accuracy, it is relatively hard to compare this approach with similar ones.
In [11] the authors try to outperform the Google Speech Recognition API by using Sphinx, a group of open source speech recognition systems developed at Carnegie Mellon University. The corpus consists of 3000 sentences related to bus schedule information. Sphinx obtained a WER of 51.2% and outperformed Google by 3.3%. The results are arguably relevant because, first of all, the difference is not significant and, second off all, because Google does not train its algorithm to be domain-specific (Pittsburgh bus stations in this case). In the same time, a similar approach for Romanian could take a lot of development time, most of it being spent for gathering training data. Since the results obtained in this paper when it comes to compare the two systems' accuracy are more or less similar, we considered using the out-of-the-box Google algorithm for measuring ASR performance on Romanian e-learning resources. Future work can focus also on testing Sphinx's performance for Romanian in the conditions of a larger corpus of YouTube video than the one used by this paper.
To conclude this chapter, there are several experiments that tried to perform ASR for Romanian, but none of them for e-learning resources. The added value of this paper in comparison with similar works that are using Google Speech-to-Text API is, first of all, the fact that we are doing this on an under-resourced language and, second of all, the fact that we will try to outperform the existing attempts.

3 Google Speech Recognition API in the context of Romanian e-learning resources

Google Speech Recognition was publically launched in 2008 in the form of the Google Voice Search app for iPhone. With the usage of the large quantities of data stored on its

DOI: 10.12948/issn14531305/23.1.2019.02

servers and of the already existing machine learning algorithms, Google created the first large scale ASR system, being considered the pioneer of modern ASR algorithms [12].
The Cloud Speech-to-Text API (figure 1) was launched in April 2017 and has included since the beginning support for 90 languages, 30 being added afterwards in August 2017 [13]. Beside the basic ASR, the cloud service offers additional features like phrase hints, real-time streaming, language auto-detection, inappropriate content filtering, automatic punctuation, multichannel recognition and others [14].
"In ancient times having power meant having access to data, today having power means knowing what to ignore" said Yuval Noah Harari in his book "Homo Deus". We are leaving in a decade where information is everywhere. The problem that raises is how can one know if the e-learning resources that he or she is interested in are the proper ones. Especially if the resources are in video format. With YouTube gaining more and more grip, humanity has to deal with 300 hours of video uploaded to this platform every minute [15]. Among those are, of course, valuable resources, but how can one determine what videos to watch and what to ignore, which ones are valuable e-learning materials and which one are just cat videos. The problem exists in large deposits of e-learning data also. How can one determine fast if a video contains the information that he or she is interested in? Even if the subject of the video or the title suggests so, the content may not be the desired one.
One possible way of resolving this is to automatically transcribe the voice from multimedia e-learning materials. By doing this, classical text-based search algorithms can be applied to the newly obtained text resource. The search can be done in a statistical manner, based on keywords or by using Named Entity Recognition algorithms which can determine the taxonomy of the content within the video, together with different relations between the identified entities. From this point on, semantic approaches can be used to filter the data [16].

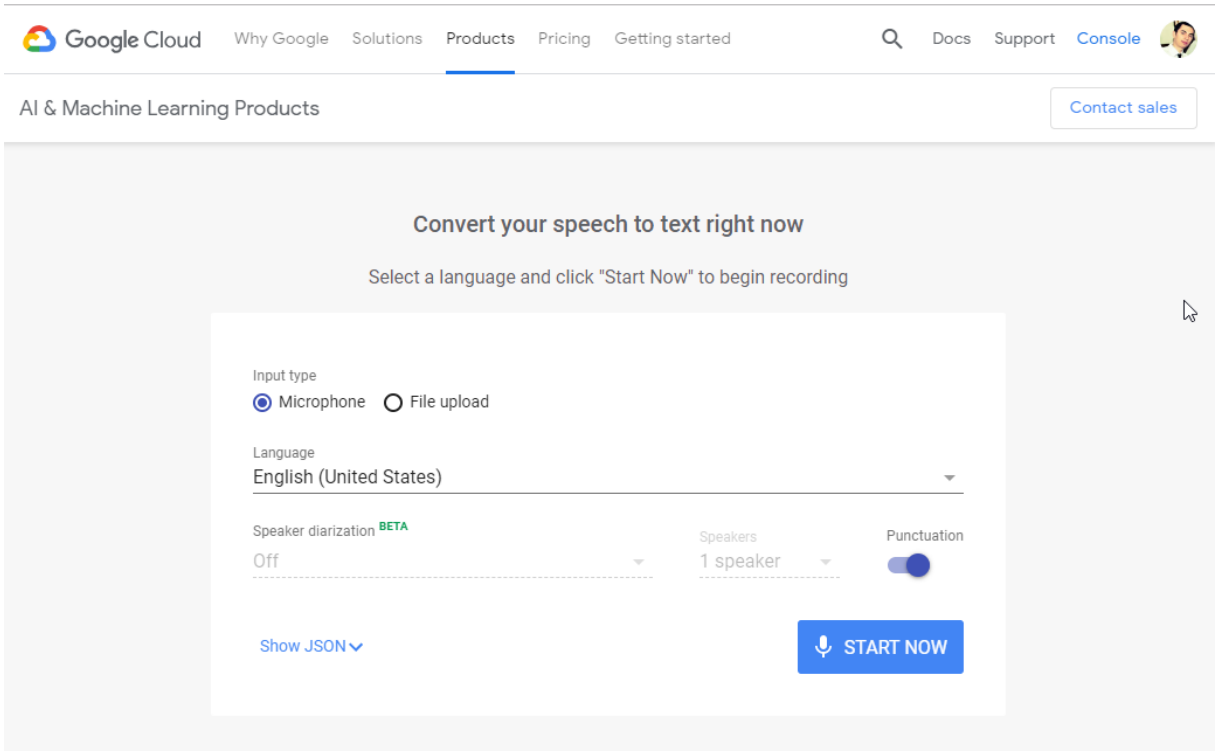

Fig. 1. The Google Cloud Speech-to-Text official website with the demo interface of the product in foreground [14]

The ASR algorithms can help in automating the task of transcribing the speech from video

DOI: 10.12948/issn14531305/23.1.2019.02

(tutorials, webinars, online lessons) or audio (e-books, podcasts) e-learning resources, thing that we will focus on in greater detail in the next chapter.

4 Evaluating Google Speech Recognition API's Performance for Romanian

In order to evaluate Google Cloud Speech-to-Text API's performance when it comes to Romanian e-learning resources, a corpus of YouTube videos was selected. The dataset is widely presented in table 2, but in large terms it consisted of 20 e-learning videos from the YouTube platform, 10 of them presented by male speakers (5 different speakers) and 10 of them by female speakers (5 different speakers). Because of the limitations imposed by the free version of the Google Cloud Speech-to-Text API, we considered and analyzed just the first minute of each video. The correct transcription for each video was manually created prior to the speech-to-text analysis, obtaining a corpus of 3100 words divided in 210 sentences.

Table 2. Testing dataset summary

Video no.	Sentences	Words	Main subject	Speaker
1	10	142	Entrepreneurship	male
2	18	140	Entrepreneurship	male
3	8	144	Programming	male
4	10	134	Programming	male
5	12	209	IT	male
6	18	195	Lifestyle	male
7	6	57	Auto	male
8	16	186	Media	male
9	18	109	Lifestyle	male
10	7	160	Lifestyle	male
11	7	151	Lifestyle	female
12	7	172	Personal development	female
13	6	151	Parenting	female
14	6	119	Lifestyle	female
15	7	167	Lifestyle	female
16	7	162	Lifestyle	female
17	13	217	Beauty	female
18	11	195	Lifestyle	female
19	9	134	Beauty	female
20	8	162	Lifestyle	female

The most widely present measurement of a speech recognition, that we also used to evaluate our proposed system's accuracy is Word Error Rate (WER) [2], computed as following:

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

where S is the number of substitution, D is the number of deletions, I is the number of insertions and N is the total number of words. The pseudocode for computing WER is presented in figure 2. For the training dataset, the implementation of the WER algorithm in Python from [17] was used.

DOI: 10.12948/issn14531305/23.1.2019.02

```
function WER(Reference r, Hypophysis h)
    int[|r| + 1][|h| + 1] D
    for (i = 0; i ≤ |r|; i++) do
        for (j = 0; j ≤ |h|; j++) do
            if i == 0 then
                D[0][j] ← j
            else if j == 0 then
                D[i][0] ← i
            end if
        end for
    end for

    for (i = 1; i ≤ |r|; i++) do
        for (j = 1; j ≤ |h|; j++) do
            if r[i − 1] == h[j − 1] then
                D[i][j] ← D[i − 1][j − 1]
            else
                sub ← D[i − 1][j − 1] + 1
                ins ← D[i][j − 1] + 1
                del ← D[i − 1][j] + 1
                D[i][j] ← min(sub, ins, del)
            end if
        end for
    end for

    return D[|r|][|h|]
end function
```

Fig. 2. The pseudocode for the calculation of WER with Levenshtein distance [17]

The WER value obtained for the whole corpus was 30.96% which places our approach under the results obtained in [3], which had a WER of just 9.59% (WER = 1 – WRR) and [4], which had a WER of 20.7%, but above the results obtained by the authors of [11] where WER was 54.5%. Having in consideration the fact that both approaches (ours and [11]) are using Google Speech Recognition API we can consider that our results are more than satisfying.

Another aspect that we must mention is that our results are influenced also by the quality of the videos (the quality of the audio recordings actually), by the person's accent and speech clarity, by the background noises and by other external factors. Most probably the better results of other algorithms were obtained in controlled environments by using high quality audio recordings. To emphasize this, we must say that one of the videos, where the female speaker had a clear voice and the recording was done by using a high quality microphone, obtained a WER of just 9.93% which is comparable with the results obtained by [3]. Table 3 and table 4 illustrate the results obtained per speaker gender and per video subject.

Table 3. Results grouped by speaker's gender

Speaker	Number of speakers	WER
male	5	38.41%
female	5	24.02%

DOI: 10.12948/issn14531305/23.1.2019.02

Table 4. Results grouped by video's subject

Subject	Number of videos	WER
Auto	1	22.81%
Beauty	2	40.43%
Entrepreneurship	2	30.85%
IT	1	28.71%
Lifestyle	9	28.76
Media	1	31.72%
Parenting	1	9.93%
Personal development	1	11.63%
Programming	2	57.19%

We can see from the tables above that the best results were obtained by female speakers and that the subject with the lowest WER was Parenting, even though just one video is not enough to validate this hypothesis. Another interesting fact is that the subjects with the highest WER were Programming and Beauty. This can be due to the fact that both domains are using a lot of English terms to describe some of the entities, tools or actions involved, terms that cannot be easily translated into Romanian, thus unidentifiable by a Romanian focused ASR system.

5 Conclusions and future work

The main objective of this research paper was to obtain an accurate enough ASR model that can process multimedia e-learning resources which contain Romanian speech. For achieving this, first the ASR term was defined, then a short history of speech recognition systems was presented. The main algorithms used to perform ASR, together with the cloud providers that offer ASR solutions in the form of SaaS were analyzed as well. After a short literature review which emphasized the added value of the current research, the paper shifted into applying the Google Cloud Speech-to-Text API on a corpus of various video e-learning resources available online on YouTube. The main reason behind this approach was to transform the multimedia resources into text-based resources with the scope of making them more searchable by classic keyword-based algorithms or by semantic solutions. Another side effect of this approach is that it can help the automation of subtitles generation for e-learning videos. A difficult process that is done in most of the cases (at least for Romanian) manually and that can help people with hearing problems.

Thus, a dataset of 3100 words divided in 210 sentences was used, uttered by 10 speakers (5 males and 5 females). In order to measure the accuracy of the model, the Word Error Rate indicator was computed. The Google Cloud Speech-to-Text API obtained a WER of 30.96% for the used dataset, which is a better result that the one obtained by similar works which are using Google Speech Recognition API, but worse than other solutions performed for Romanian in controlled environments with homogenous datasets. Even so, some videos obtained promising results, having an WER of just 9.93%, which gives us hope that by tuning the system properly and by using more qualitative audio recordings, the current model has the potential of obtaining better results.

Future work will focus on using a larger corpus with higher quality audio files with the scope of obtaining a WER under 10%. If this objective will not be met, we are also considering using the Sphinx's implementation in Python to check if it can outperform the Google's algorithm.

Another future objective is to use the results obtain by the ASR algorithm as input data for a Named Entity Recognition solution that will semantically index the multimedia e-learning resources of interest. By taking this approach, similar e-learning materials can be aggregated into a domain ontology and semantic searches can be made.

DOI: 10.12948/issn14531305/23.1.2019.02

Acknowledgement
This paper presents results obtained within the PN-III-P1-1.2-PCCDI-2017-0272 ATLAS project ("Hub inovativ pentru tehnologii avansate de securitate cibernetică / Innovative Hub for Advanced Cyber Security Technologies "), financed by UEFISCDI through the PN III –"Dezvoltarea sistemului national de cercetare-dezvoltare", PN-III-P1-1.2-PCCDI-2017-1 program.

References

[1] K. Davis, R. Biddulph and S. Balashek, "Automatic recognition of spoken digits," *The Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637-642, 1952.
[2] V. Radha and C. Vimala, "A review on speech recognition challenges and approaches," *World of Computer Science and Information Technology Journal (WCSIT)*, vol. 2, no. 1, pp. 1-7, 2012.
[3] C. O. Dumitru and G. Inge, "A comparative study of feature extraction methods applied to continuous speech recognition in romanian language," in *Proceedings ELMAR 2006*, Zadar, Croatia, 2006.
[4] H. Cucu, L. Besacier, C. Burileanu and A. Buzo, "Enhancing automatic speech recognition for Romanian by using machine translated and web-based text corpora," in *SPECOM 2011*, 2011.
[5] B. Laurent, B. Etienne, K. Alexey and S. Tanja, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85-100, 2014.
[6] "Amazon Transcribe now supports speech-to-text in French, Italian, and Brazilian Portuguese," Amazon, 20 December 2018. [Online]. Available: https://aws.amazon.com/about-aws/whats-new/2018/12/amazon-transcribe-now-supports-speech-to-text-in-french-italian-and-brazilian-portuguese/. [Accessed 19 January 2019].
[7] "Speech to Text," Microsoft, February 2019. [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/. [Accessed 19 January 2019].
[8] "IBM Cloud Docs / Speech to Text - Language Support," IBM, 7 February 2019. [Online]. Available: https://console.bluemix.net/docs/services/speech-to-text/index.html#languages. [Accessed 9 February 2019].
[9] "Cloud Speech-to-Text API - Language support," Google, 17 January 2019. [Online]. Available: https://cloud.google.com/speech-to-text/docs/languages. [Accessed 19 January 2019].
[10] T. Ashwell and J. R. Elam, "How Accurately Can the Google Web Speech API Recognize and Transcribe Japanese L2 English Learners' Oral Production?," *Jalt Call Journal*, vol. 13, no. 1, pp. 59-76, 2017.
[11] P. Lange and D. Suendermann-Oeft, "Tuning Sphinx to outperform Google's speech recognition API," in *Proc. of the ESSV 2014, Conference on Electronic Speech Signal Processing*, 2014.
[12] C. Boyd, "The Past, Present, and Future of Speech Recognition Technology," Medium, 10 January 2018. [Online]. Available: https://medium.com/swlh/the-past-present-and-future-of-speech-recognition-technology-cf13c179aaf. [Accessed 19 January 2019].
[13] "Cloud Speech-to-Text API - Release notes," Google, 20 February 2019. [Online]. Available: https://cloud.google.com/speech-to-text/docs/release-notes. [Accessed 22 January 2019].
[14] "Cloud Speech-to-Text," Google, February 2019. [Online]. Available: https://cloud.google.com/speech-to-text/. [Accessed 19 January 2019].
[15] "YouTube by the Numbers: Stats, Demographics & Fun Facts," Omnicore, 6 January 2019. [Online]. Available: https://www.omnicoreagency.com/youtube-statistics/. [Accessed 19 January 2019].
[16] I. Ivan, C. Brândaș and A. Zamfiroiu, "Audit Validation Using Ontologies,"

Informatica Economică, vol. 19, no. 2, pp. 25-33, 2015.

[17] M. Thoma, "Word Error Rate Calculation," 15 November 2013. [Online]. Available: https://martin-thoma.com/word-error-rate-calculation/. [Accessed 19 January 2019].

Bogdan IANCU has graduated The Faculty of Cybernetics, Statistics and Economic Informatics from The Bucharest University of Economic Studies in 2010. He has a master's degree in Economic Informatics (2012) and a PhD in Economic Informatics starting from 2015 in the field of Ontologies and eLearning. He is an Assistant Lecturer in The Department of Economic Informatics and Cybernetics from The Bucharest University of Economic Studies. His current research focuses on semantic technologies and ontologies innovations. Other fields of interest include machine learning, cybersecurity, mobile devices, embedded systems and IoT.

DOI: 10.12948/issn14531305/23.1.2019.02