On the Impact of Dysarthric Speech on Contemporary ASR Cloud Platforms

Journal of Reliable Intelligent Environments manuscript No.
(will be inserted by the editor)

On the Impact of Dysarthric Speech on Contemporary ASR Cloud Platforms

Luigi De Russis · Fulvio Corno

Received: date / Accepted: date

Abstract The spread of voice-driven devices has a positive impact for people with disabilities in smart environments, since such devices allow them to perform a series of daily activities that were difficult or impossible before. As a result, their quality of life and autonomy increase. However, the speech recognition technology employed in such devices becomes limited with people having communication disorders, like dysarthria. People with dysarthria may be unable to control their smart environments, at least with the needed proficiency; this problem may negatively affect the perceived reliability of the entire environment. By exploiting the TORGO database of speech samples pronounced by people with dysarthria, this paper compares the accuracy of the dysarthric speech recognition as achieved by three speech recognition cloud platforms, namely IBM Watson Speech-to-Text, Google Cloud Speech, and Microsoft Azure Bing Speech. Such services, indeed, are used in many virtual assistants deployed in smart environments, such as Google Home. The goal is to investigate whether such cloud platforms are usable to recognize dysarthric speech, and to understand which of them is the most suitable for people with dysarthria. Results suggest that the three platforms have comparable performance in recognizing dysarthric speech, and that the accuracy of the recognition is related to the speech intelligibility of the person. Overall, the platforms are limited when the dysarthric speech intelligibility is low (80-90% of word error rate), while they improve up to reach a word error rate of 15-25% for people without abnormality in their speech intelligibility.

Keywords Automatic speech recognition · Speech-to-text · Dysarthria · Accessibility · Comparison · Cloud Platform

L. De Russis, F. Corno
Politecnico di Torino
Dipartimento di Automatica e Informatica
Corso Duca degli Abruzzi, 24
10129 Torino, Italy
E-mail: fulvio.corno, luigi.derussis@polito.it

1 Introduction

Speech recognition technology entered the public life rather recently, with launch events from tech giants making worldwide headlines. Voice-driven interfaces are, therefore, becoming commonplace: people can use their voice to control their smart home or their in-car systems. Such devices, mostly powered by Automatic Speech Recognition (ASR) cloud platforms like Google Cloud Speech, have a positive impact for people with disabilities [18,4,15]. Through such devices, people with disabilities can indeed perform a series of activities that were difficult or impossible before, e.g., controlling their connected vacuum robots, setting alarms, turning on and off lights, playing music, etc.

However, the ASR technology employed in contemporary voice-driven devices becomes limited with users having moderate to severe speech disorders like *dysarthria*. Dysarthria is a motor speech disorder resulting from neurological injury of the motor component of the motor-speech system, characterized by poor articulation of phonemes. Problems in word articulation impact the performance of ASR, with a consequent negative impact on the perceived reliability of the entire smart environment in which voice-driven devices are used.

By exploiting the TORGO [20] database of speech samples pronounced by people with dysarthia, this paper compares the accuracy of the dysarthric speech

recognition as achieved by three popular ASR platforms, i.e., IBM Watson Speech-to-Text, Google Cloud Speech, and Microsoft Azure Bing Speech. We focused on such cloud platforms due to their global availability, their wide usage, and the fact that their algorithms are continuously maintained and updated. To have a common baseline to conduct the comparison, the three platforms were also evaluated with the same TORGO speech samples, but pronounced by people without any speech impairments. Such platforms are used in different voice-based devices and interfaces, like Google Home ans the in-car "Ask Mercedes".

TORGO, in fact, is a well-know dataset of dysarthric speech, developed as a result of collaboration between the University of Toronto's departments of Computer Science and Speech Language Pathology, and the Holland-Bloorview Kids Rehab hospital in Toronto, Canada. It contains dysarthric speech samples, the corresponding *original textual sentences*, and documentation from 8 speakers (5 males, 3 females) with cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS), all of them with dysarthria. The database also contains an evaluation of speech intelligibility for the eight participants, according to the "Intelligibility Severity Rating" section of the Frenchay Dysarthria Assessment [5]. In addition, it contains speech samples and the corresponding original textual sentences from a control group of 7 speakers (4 males, 3 females) without any speech impairment.

The goal of this paper is two-fold: to investigate whether such cloud platforms are usable to recognize dysarthric speech, and to understand which of them is the most suitable for people with dysarthria. To do this, we rely on the transcribed sentences provided by the ASR platform. ASR cloud platforms, indeed, process speech samples and produce a *default transcribed sentence* and a set of *transcription alternatives* (also called alternatives). Alternatives transcriptions are variations of the default sentences, presented in no particular order. To evaluate the accuracy of the dysarthric speech recognition by the three ASR platforms, we computed the word error rate (WER) by comparing each *default transcribed sentence* with the *original text sentence* (provided by TORGO). In addition, to understand which cloud platform is the most suitable for people with dysarthria, we further analyzed the results of the best ASR platforms to get some additional insights about the mistakes in the *default transcribed sentences* and to check whether a *transcription alternative* is better than the *default transcribed sentence*.

Results of the comparison suggest that the performances in recognizing dysarthric speech are comparable among the three platforms. Moreover, the accuracy of the recognition is strictly related to speech intelligibility of persons with dysarthria, for all the three ASR platforms. In particular, the ASR platforms present limited results when the dysarthric speech intelligibility is severely distorted, with a WER in the range of 55-75% (S.D. around 10-20%). When the speech intelligibility has no particular abnormalities, instead, the average WER is in the range of 15-25% (S.D. around 20%), not so different from the average WER of the control group (5%). Finally, results show that, in 60% of the cases, the *default transcribed sentence* is not the best transcription for the original text sentence, and that Google Cloud Speech is currently the most suitable platform for handling dysarthrich speech.

2 Related Works

This work provides an overview on the issues that may arise from the usage of voice-driven interfaces when they need to handle dysarthric speech. To do so, it contributes with an evaluation of the behavior and reliability of popular Automatic Speech Recognition (ASR) systems.

Despite speech technology in general, and ASR in particular, are not new for people with disabilities, specific research in the domain of technology for people with speech impairments is still quite limited. Speech technology and ASR have been used to increase accessibility in mainstream operating systems since decades, as an alternative method to compose documents through dictation systems, to control computers and smartphones. Similarly, speech recognition as an input to electronic assistive technology was investigated both in general and for dysarthria.

In 2002, Hawley [8] presented an early overview, based on a literature review and clinical observations, upon the suitability and performance of speech recognition for computer access by people with disabilities, including people with dysarthria. He reported that, given adequate time, training, and support, commercial ASR systems for computers are often appropriate for people with no, mild, or moderate speech impairments. People with dysarthria achieve lower recognition rates, but speech recognition can be still a useful input method for some individuals. Conversely, Hawley discovers that speech as a mean of controlling electronic devices such as smartphones and appliances is more troublesome, especially for dysarthric speech.

To overcome this kind of issues, researchers investigated several new methods, datasets, and proposed dedicated dysarthric speech recognition systems. Rudzicz et al. [20], for instance, describes the acquisition and the composition of TORGO, a database of dysarthric speech in terms of aligned acoustics and articulatory

data, mainly from individuals whose speech impediments were caused by cerebral palsy or amyotrophic lateral sclerosis.

Rudzicz [19], starting from the TORGO database and leveraging the task-dynamics theory, also proposes a new method for acoustic-to-articulatory inversion for dysarthria, which estimates positions of the vocal tract given acoustics using a nonlinear Hammerstein system. The approach uses adaptive kernel canonical correlation analysis and is found to be significantly more accurate than mixture density networks, at or above the 95% level of confidence for most vocal tract variables for dysarthric speech. In addition, he introduces a new method for ASR in which acoustic-based hypotheses are re-evaluated according to the likelihoods of their articulatory realizations in task-dynamics.

Kim et al. [14, 13] investigate dysarthric speech recognition using Kullback-Leibler divergence-based hidden Markov models. In the model, the emission probability of state is modeled by a categorical distribution using phoneme posterior probabilities from a deep neural network, and therefore, it can effectively capture the phonetic variation of dysarthric speech. Through an experimental evaluation on a database of several hundred words, they show that the proposed approach provides substantial improvement over the conventional Gaussian mixture model and deep neural network based speech recognition systems.

More recently, Joy et al. [10] adopted the TORGO database to explore multiple ways to improve Gaussian mixture model and deep neural network (DNN) based hidden Markov model (HMM) ASR systems. Their work shows significant improvements over the previous attempts in building such ASR systems with TORGO. In their work, they trained speaker-specific acoustic models by tuning various acoustic model parameters, using speaker normalized cepstral features and building complex DNN-HMM models with dropout and sequence-discrimination strategies. The DNN-HMM models for severe and severe-moderate dysarthric speakers were further improved by leveraging specific information from dysarthric speech to DNN models trained on audio files from both dysarthric and normal speech, using generalized distillation framework.

Yu et al. [21] presents an initial attempt to develop an ASR system for the Universal Access Speech (UA-Speech) database [12]. A range of deep neural network (DNN) acoustic models and their more advanced variants based on time delayed neural networks (TDNNs) and long short-term memory recurrent neural networks (LSTM-RNNs) were developed. Speaker adaptation by learning hidden unit contributions (LHUC) was used. The authors further built a semi-supervised complementary auto-encoder system, to improve the bottleneck feature extraction. Two out-of-domain ASR systems separately trained on broadcast news and switchboard data were cross domain adapted to the UA-Speech data and used in system combination. The final combined system gave an overall word accuracy of 69.4% on a 16-speaker test set.

While several efforts were oriented in developing novel, yet dedicated, ASR systems for people with different degrees of dysarthria, only a few works explore accessibility issues of virtual assistants and voice-driven devices when they need to handle sentences pronounced by people with speech impairments. Glasser et al. [6] focus on the issues that may arise from the usage of two virtual assistants by people who are deaf and hard of hearing. Bigham et al. [3], instead, propose two technical approaches for enabling deaf people to provide input to voice-driven devices, i.e., human computation workflows for understanding speech and mobile interfaces that can be instructed to speak on the user's behalf. Ballati et al. investigate the interaction of dysarthric speech data with three widely used virtual assistants, included in several standalone and mobile devices (Apple's Siri, Google Assistant, and Amazon Alexa), both in English [2] and in Italian [1].

Similar to the work of Glasser et al. and Ballati et al., we focus on the issues that may arise from the usage of voice-driven assistants, but we are specifically interested in dysarthric speech and in the evaluation of the behavior and reliability of the contemporary and popular ASR cloud platforms that often empower voice-driven devices, namely Google's, IBM's, and Microsoft's services. In addition, we would also understand which of them (if any) could be the most suitable platform to be used by people with dysarthria.

3 Background and Problem Statement

3.1 The TORGO Database

To study the behavior of the three ASR cloud platforms we need to obtain an appropriate number of sentences pronounced by people with dysarthria. A few datasets about dysarthric speech were produced by the research community, with the most notable being the TORGO database of dysarthric articulation [20], UA-Speech database of spastic dysarthria [12], and the Nemours database of dysarthric speech [16]. Despite all those datasets were used to improve or create new ASR models able to tackle dysarthric speech, for the purpose of this paper we decided to adopt the TORGO database. We looked for an available dataset with full sentences, indeed, to have samples as ecologically valid

as possible, and with a set of related speech samples pronounced by people without any speech impairments; UA-Speech does not include sentences (but only words) nor non-dysarthric speech samples, and we were not able to get the Nemours database. Moreover, the TORGO database is the newest one and its speech samples were collected within a professional setting.

The TORGO dataset is the result of a collaboration between the departments of Computer Science and Speech-Language Pathology, both at the University of Toronto and the Holland-Bloorview Kids Rehab Hospital in Toronto, Canada. It includes a large number of sentences, with data collected between 2008 and 2010. It contains approximately 23 hours of English speech samples, transcripts, and documentation from 8 speakers (5 males, 3 females) with cerebral palsy or amyotrophic lateral sclerosis (ALS), and from 7 speakers (4 males, 3 females) from a non-dysarthric control group. Both cerebral palsy and ALS speakers were affected by dysarthria with disruptions of motor commands of the vocal articulators, with an atypical and relatively unintelligible speech in most cases [11].

The speech intelligibility of the 8 speakers with dysarthria ranges from "no abnormalities" (for three of them, 2 females and 1 male) to "severely distorted" (the remaining speakers, 4 males and 1 female).

Sentences in the TORGO database consist of non-words, short words, restricted sentences, and unrestricted sentences. Non-words were used to control baseline abilities of speakers with dysarthria, especially to gauge their articulatory control in the presence of plosives and prosody, like high-pitch and low-pitch vowels. The short words (e.g., 'yes', 'no', 'select', 'increase', ...) are useful for studying speech acoustics without the need for word boundary detection. The restricted sentences include 162 sentences from the sentence intelligibility section of "Assessment of intelligibility of dysarthric speech" [5] and 460 sentences derived from the TIMIT database [22]. The unrestricted sentences were elicited by asking participants to spontaneously describe thirty images in interesting situations taken randomly from Webber Photo Cards - Story Starters, originally designed to prompt students to tell or write a story [20]. Each speaker was recorded while reading different sentences, only partially shared among them. For the purpose of this paper, to perform an appropriate comparison, we only considered the sentences shared among speakers, i.e., 38 sentences for male speakers and 13 for females.

3.2 Cloud Platforms

To perform a comparison to recognize dysarthric speech, we selected three of the most used cloud ASR platforms: IBM Watson Speech-to-Text [9], Google Cloud Speech [7], and Microsoft Azure Bing Speech [17]. We choose such cloud platforms due to their global availability, the fact that they are constantly and directly maintained and updated, and their wide usage also in voice-based devices.

To transcribe the human voice accurately, these services leverage machine intelligence to combine information about grammar and language structure with knowledge of the composition of the audio signal. Those cloud services continuously return and retroactively update a transcription as more speech is heard. For all three platforms, the service interfaces share many common input features for transcribing speech-to-text, such as supported audio formats, languages and models. The platforms also support various output features like speaker labels, keyword spotting, maximum alternatives and interim results, word alternatives, word or sentence confidence, word timestamp, profanity filtering, and smart formatting. These features are exploited in the analysis to look for the most accurately transcribed sentence.

3.3 ASR for Dysarthric Speech

To better understand which features a sentence pronounced by a person with dysarthria exhibit, we present here three examples (a, b, c). In each example, the first sentence is the *original text sentences* as in the TORGO database. The second one is the *default transcribed sentence* obtained by IBM Watson Speech-to-Text, starting from the audio file produced by one of the five males with dysarthria present in the dataset.

a.1 A long flowing beard clings to his chin.
a.2 A long flowing gear things to his chin.
b.1 You wished to know all about my grandfather.
b.2 You wish to know all about nine.
c.1 She had your dark suit in greasy wash water all year.
c.2 She had your dark suit an greasy wash water all re a.

By comparing the *original text sentences* ("1") with the *default transcribed sentence* ("2") for each of the three examples, we can notice that the second sentences are not the correct transcription of the first one. In these cases, the person with dysarthria could not fully benefit from using an ASR platform.

In general, three problems may arise when an ASR platform does not recognize a sentence correctly. First,

some wrong words may be present into the *default transcribed sentence* (e.g., as in "a.2"). Second, some of the words present in the *original text sentences* can be missing in the *default transcribed sentence* (e.g., "b.2"). Finally, the *default transcribed sentence* may present more words than the *original text sentences* (e.g., "c.2").

Wrong words, *missing words*, and words in excess are the three typical error types in dysarthric speech recognition. They may happen separately or together, according to the speech intelligibility of the speaker (e.g., "c.2" exhibits both words in excess and wrong words).

3.3.1 Research Questions

The following research questions guided our work towards the goal of investigating whether and to which extent the three ASR cloud platforms are usable to successfully recognize dysarthric speech:

RQ1 Are ASR platforms suitable for recognizing dysarthric speech? What is the attained recognition rate?

RQ2 What kinds of transcription errors are more frequent, in case of imperfect/partial recognition?

RQ3 Can transcription alternatives (as provided by ASR platforms) be used to improve the overall recognition result?

To answer RQ1, we will investigate the accuracy in transcription of the three ASR platform by computing the WER for the default transcribed sentence of the TORGO speech samples. For RQ2, instead, we will analyze the transcription error for each default transcribed sentence, to define the most common type of errors that arise for the transcription of dysathric speech. Finally, to answer RQ3, we will start from the *transcription alternatives* to check whether one of them would be the best transcription for the original sentence.

4 Methodology

The analysis about the accuracy of ASR platforms with dysarthric speech has as the initial input the speech samples and the relative *original text sentences* from the TORGO database, and as the final output the analysis of the *transcription alternatives* from ASR platforms. Figure 1 shows the four phases of the analysis described in this paper. First, we selected the common *original text sentences* and related speech sample from the TORGO database, separately for males and females, to perform a balanced comparison. In the second phase, each speech sample was fed to each ASR platform, and the resulting *default transcribed sentence* and the set of *transcription alternatives* is saved for each speech sample. In the third phase, we compute the WER for all the transcribed sentences to answer RQ1. The last phase analyzed the sentences from the best platforms (emerged from the previous phase), to answer both RQ3 and RQ4.

4.1 Phase 1 - TORGO Sentences Selection

First of all, we identified the common *original text sentences* and the related speech samples separately for males and females from the TORGO database. In fact, each speaker from TORGO database was recorded while reading different sentences, only partially shared among them. To perform a balanced and fair comparison, we selected the common sentences. To keep the number of sentences as high as possible, we needed to consider sentences pronounced by male speakers separate from the ones from females. The output of the first phase is 13 common *original text sentences* with related speech samples for the three females, and 38 *original text sentences* with related speech samples for the five males. For the control group, we selected the same sentences.

4.2 Phase 2 - Automatic Speech Recognition

In the second phase, we submit each speech sample identified in the Phase 1 to every ASR platform, separately, to analyze the speech samples related to the common *original text sentences*. Each ASR platform recognized all speech samples from each speaker. The output of the automatic speech recognition process is a list of thirty *transcription alternatives* (plus the default transcribed sentence) and the related level of confidence, for each platforms and for each speaker.

4.3 Phase 3 - Suitability Analysis

The third phase analyzes the *default transcribed sentence* in terms of its "suitability". In this phase, we analyzed the sentences from speakers belonging to both the "no abnormal" and the "severely distorted" intelligibility speech categories. The goal of this suitability analysis is to evaluate the accuracy in transcription of the *default transcribed sentence*, for each speaker, thus answering RQ1. We computed the WER between the *original text sentences* and the *default transcribed sentence*, as provided by each platform. The WER is defined as $WER = (S + I + D)/(S + D + C)$, where I is the number of word insertions, D the number of word deletions, S the occurrence of word substitutions, while C is the number of correctly transcribed words.

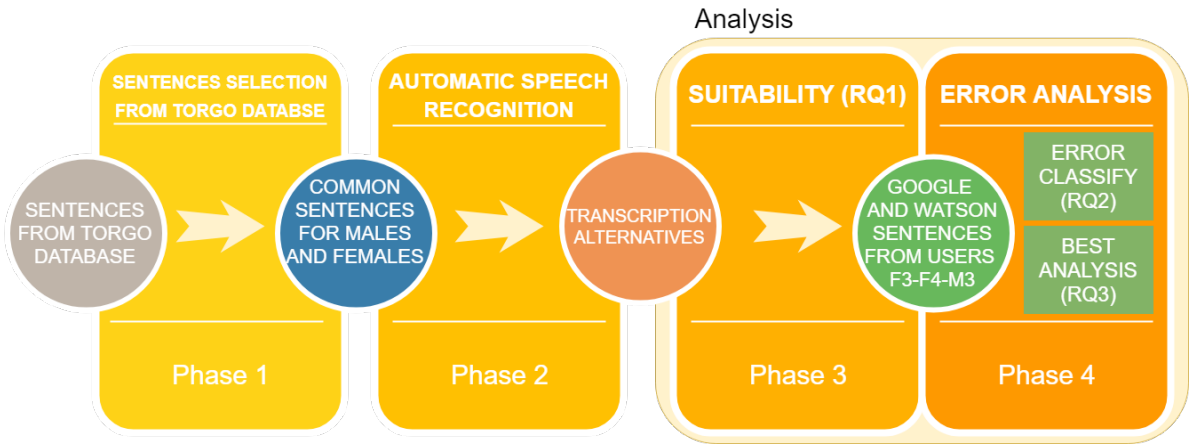

Fig. 1 The phases followed to perform the analysis.

4.4 Phase 4 - Error Analysis

The error analysis consisted in two different evaluation: *error classification* and *best analysis*. The best platforms from the quantitative analysis were selected as a reference for this phase.

4.4.1 Error Classification

For the error classification, we classified the mistakes that are present in each of the default transcribed sentence to define the most common typology of error which stem from the ASR process of dysarthric speech. The "error classification" step allowed us to answer RQ2.

4.4.2 Best Analysis

The goal of the "best analysis" is to find out a *transcription alternative* better than the *default transcribed sentence*. To do so, we computed the WER between the *original text sentences* and each *transcription alternative* from the list of 30 alternatives. Afterwards, we selected the *best transcription alternative* for each ASR platform.

The *best transcription alternative* is the alternative with the smallest WER, i.e., the best transcription result among all the *alternatives*. The identification of the *best transcription alternative* is possible due to the knowledge of the *original text sentences*. We also kept into account the *position* of the selected *best transcription alternative* into the list of the 30 *alternatives*. This step allowed us to answer RQ3.

Table 1 WER from the suitability analysis

Speaker	**WER**		
	Google	**Microsoft**	**IBM**
All dysarthric users	59.81%	62.94%	67.35%
- "No abnormalities"	16.11%	23.16%	14.89%
- "Severely distorted"	78.21%	78.59%	89.08%
Control group	3.95%	6.94%	5.26%

Table 2 Correctly transcribed sentences (default transcription)

Speaker	**Correctly transcribed sentences**		
	Google	**Microsoft**	**IBM**
All dysarthric users	15.28 (35)	9.17 (21)	14.85 (34)
- "No abnormalities"	51.56 (33)	31.25 (20)	53.15 (34)
- "Severely distorted"	1.21(2)	0.61 (1)	0.00 (0)
Control group	69.15	43.02	62.99

5 Results: Suitability Analysis

For what concerns *Phase 3 - Suitability Analysis*, Table 1 shows the accuracy in transcription, in terms of WER, evaluated for all the users and the two speech intelligibility categories. In addition, in the bottom of Table 1, we show the results for the control group. Table 2 shows, instead, the quantity of correctly transcribed sentences, evaluated for users in the same way as we did in table one.

Considering all the analyzed sentences from all dysarthric speakers, the average WER for Google is slightly lower than the average WER for Microsoft and IBM (WER: 59.81% for Google vs. WER: 62.94% for Microsoft vs. WER: 67.35% for IBM, SD around 35% for each ASR platform). Instead, the number of correctly transcribed sentences is generally low for all three ASR

platforms, with the platforms provided by Google and IBM that show similar values, while is clearly lower for Microsoft's platform.

By looking at the three users evaluated as "no abnormalities" in speech intelligibility, results confirm the positive behavior for Google, close to the performance of IBM Watson Speech-to-Text (WER around 15% for both ASR platforms). Microsoft cloud platform, instead, is the worst among the three ASR platform, with a WER of 23.16%. Almost all the correctly transcribed sentences for the three ASR platforms belong to speakers with this speech intelligibility.

The average WER for the remaining five speakers, i.e., those who exhibit a "Severely distorted" speech intelligibility, grows strongly, in the range between 75% and 90%. Result for those speakers substantiate the previous data for Google. In fact, the WER for Google is around 80%. The WER obtained for IBM is around 90%, while Microsoft's ASR platform exhibits a good WER at 80%. For these speakers, the number of correctly transcribed sentences is unfortunately close to zero.

Finally, the average WER for the control group are clearly low and similar for all ASR platforms (Table 1), around 5%. For what concerns the number of correctly transcribed sentences in the control group, Google's and IBM's platforms presents much better performances than Microsoft's (Table 2).

5.1 Discussion

The suitability analysis explores the accuracy of the speech samples transcriptions for all the speakers across the two speech intelligibility categories. From this analysis emerges that, overall, Google Cloud Speech has the best performance in terms of WER (59,81%, first row of Table 1), strictly followed by Microsoft Azure Bing Speech (62,94%). In addition, the analysis points out that the behavior of the three ASR platforms is strictly related to the speech intelligibility, with a level of accuracy for people with a mild level of disarthria ("No abnormalities" row in both Tables 1 and 2) slightly higher than the control group.

For the speakers who have a "severely distorted" speech intelligibility, instead, the average WER is high: this suggests that, at the moment, the use of ASR cloud platforms is not suitable or advisable, to avoid misbehavior or deficiency in the perceived reliability of a voice-driven device powered by such a technology. In this case, we should acknowledge that the best results are provided by Google's and Microsoft's platforms (around 78% for both platforms, see Table 1).

Finally, for what specifically concerns speakers with a "no abnormalities" speech intelligibility, results for Google's and IBM's platforms are similar (16.11% vs. 14.89%, respectively) and sharply better compared with the result of the other dysarthric speakers. For these people, Google Cloud Speech and IBM Watson Speech-to-Text recognize correctly half of the sentences (33 vs. 34, respectively, as shown in Table 2). For the "severely distorted" speakers, conversely, all the platforms only recognize 0-2 sentences at most (Table 2).

Nevertheless, to answer RQ1, people with mild speech impairments cannot fully exploit such voice-driven devices which only use Google Cloud or IBM Watson, at least with the same proficiency of people without any speech impairment. Indeed, the average WER for speakers in the control group is around 5% vs. the WER of around 15% obtained for people with a mild dysarthria.

6 Results: Error Analysis

The best platforms from the quantitative analysis (i.e., Google Cloud Speech and IBM Watson Speech-to-Text) were selected as a reference for the error analysis. We also considered speakers with the "no abnormalities" speech intelligibility, since no further analysis can be done upon the results of the other dysarthric speakers.

6.1 Error Classification

We analyzed the *default transcribed sentences* from Google Cloud Speech and IBM Watson Speech-to-Text with a WER different from 100%, by only considering speakers with a "no abnormalities" speech intelligibility. To answer RQ2, we find out 5 main typologies of mistakes:

One Wrong Word: in the transcribed sentence, there is a single incorrectly transcribed word.

Wrong Words: in the transcribed sentence, there are two or more incorrectly transcribed words.

Missing word(s): in the transcribed sentence, there are one or more words for which the transcription is missing.

Split word(s): in the transcribed sentence, two or more words are the transcription of a single word from the *original text sentence* .

Multiple mistakes: in the transcribed sentence, there are multiple occurrences of the previous errors.

Starting from this error typologies, we classified all the transcribed sentences. Table 3 shows how many transcribed sentences occur for each typologies of errors. From the analyzed *default transcribed sentences*,

Table 3 Occurrences for error typologies

Error typology	Error occurrences	
	Google	**IBM**
One Wrong Word	12	13
Wrong Words	6	1
Missing word(s)	1	2
Split word(s)	1	3
Multiple mistakes	7	9
Correctly transcribed	34	33

most of the errors are about incorrect words. In fact, 20 sentences out of 27 (for Google) and 14 sentences out of 28 (for IBM) have "One Wrong Word" or "Wrong Words" types of mistake. The error typology "Missing Word" and "Word split" were not particularly common. A quite big number of sentences falls in the "Multiple mistakes" type.

6.1.1 Discussion

The main typology of error is "One Wrong Word" which is the only error that arise in around 50% of the analyzed sentences for both ASR platforms. Iin these sentences, often the error is related to articles or prepositions: this could entail a low impact on the semantic meaning of the sentence and an opportunity to tackle as future work.

The other main source of error is "Multiple mistakes'. In this case, for each sentence there are more than one error, belonging to different error typologies. For this reason, sentences classified as "Multiple mistakes" are strongly different from the original sample.

6.2 Best Analysis

Table 4 shows, for Google's and IBM's platforms separately, the comparison between *default transcribed sentences* and *best transcription alternatives*. By analyzing the *best transcription alternatives*, the average WER decreases sharply for both platforms. The WER for the two ASR platforms is lower than 10% (close to 5% for Google Cloud Speech). Moreover, results in terms of *correctly transcribed sentences* grows for IBM (Δ + 15%) and increases even more for Google(Δ + 25%).

Finally, Table 5 shows, in parenthesis, the number of correctly transcribed sentences emerging from the "best analysis" for each of the error typologies previously defined. By considering one of the most common type of error, i.e., the "One Wrong Word", both ASR platforms have the *best transcription alternatives* as the correct transcription in almost all the cases. By looking at the other main source of errors, i.e., "Multiple mistakes", Table 5 shows that Google's and IBM's platforms do not have the correct transcription among their *transcription alternatives*, in almost all cases. For sentences in the category "Wrong Words", quite numerous for Google Cloud Speech, the ASR platform has the correct transcription in just one case (out of 6).

We conclude that we can positively answer to RQ3, since in most cases the best transcription is already provided by each ASR platform. However, such a positive answer is strongly dependent from our knowledge of the "original sentence", which is typically not available to ASR platforms.

6.2.1 Discussion

In the "best analysis", we looked for the *best transcription* among the various *transcription alternatives* provided by Google's and IBM's platform. We would like to define whether and which of them is the best transcription of the original text sentence. This was possible thanks to the knowledge of the *original text sentence* (provided by TORGO).

Result in terms of WER and correctly transcribed sentences significantly improve the good results of both platforms. In particular, the WER for Google Cloud Speech is around 5%, a value close to the WER of control group and declared for people without speech impairments by the all the three ASR platforms.

By considering the improved performance obtained after selecting the best transcription, we discovered that the set of transcription alternatives should be leveraged to improve the recognition accuracy. This means that, for ASR platforms, it is already possible to retrieve a better transcription, without modifying the underline models and methods adopted for speech recognition. However, we should highlight that we exploited the knowledge of the *original text sentence* provided by TORGO, which is usually unknown for an ASR platform.

7 Limitations

Our study exhibits some limitations. First of all, we assumed that voice-driven devices (and virtual assistants) *only* rely on ASR cloud platform for the speech analysis and comprehension. This seems confirmed by previous studies both with dysarthric speech (e.g., [2,1]) and with other voice impairments (e.g., [3,6]). However, further investigations are needed: voice-driven devices can leverage from other information like the context or the overall conversation to partially tackle some errors, thus being able to provide a correct interpretation and appropriate actions. For sure, virtual assistants should

Table 4 Results for speakers with"no abnormalities" speech intelligibility

	Google			IBM		
	Default	**Best**	Δ	**Default**	**Best**	Δ
% correctly transcribed sentences	51.56	77.05	+25.49	53.15	68.85	+15.70
Residual WER (%)	16.11	5.32	−10.79	14.89	8.87	−6.02

Table 5 Occurrences for error typologies within the "best analysis"

Error typology	**Error occurrences**	
	Google	**IBM**
One Wrong Word	12 (9)	13 (11)
Wrong Words	6 (1)	1 (0)
Missing word(s)	1 (1)	2 (0)
Split word(s)	1 (1)	3 (2)
Multiple mistakes	7 (1)	9 (2)

exploit other information to improve the overall recognition of the speech.

Eventually, the study leverages on a single dataset (i.e., TORGO), not designed for virtual assistants and with speech samples collected in a professional setting. Despite TORGO proved to be a valid option for this study, different datasets with speech samples collected in a more ecological way could further enhance the behavior of ASR platforms (and/or voice-driven assistants) with dysarthric speech.

8 Conclusion

Voice-activated device, powered by automatic speech recognition platforms like Google Cloud Speech, have now become common. However, the usability of such devices and the perceived reliability of the resulting operations is strictly related to their capability of accurately recognize speech, and to correctly understand its meaning.

In this paper, we studied an accessibility challenges presented by automatic speech recognition platforms when they have to manage dysarthric speech. By using different sentences pronounced by 8 diverse speakers with dysarthria, we evaluated the performances of the three most common automatic speech recognition cloud platforms, namely, IBM Watson Speech-to-Text, Google Cloud Speech, and Microsoft Azure Bing Speech. We performed two analyses: a suitability and an error analysis.

Results show that the three cloud platforms have different behavior. In terms of word error rate and by considering all the dysarthric speakers, Google Cloud Speech has the best results among all the platforms, with an average WER of 59.81% (WER: 62.94% for Microsoft's and WER: 67.35% for IBM's platforms). Nevertheless, for dysarthric speech, the Google platform does not reaches the WER of speakers without any speech impairment (i.e., 4.9%). Comparable performance between people with dysarthria and people without speech impairments can only be obtained by considering a *best transcription alternative*, computed starting from the set of sentences provided by Google Cloud Speech. This is possible thanks to apriori knowledge of the *original text sentence*, only. Lastly, we discover that the most common mistakes in transcription of dysarthric speech with "no abnormalities" are about incorrect transcription of one or more words. Often, these mistakes do not prevent the understanding of the sentences.

Future work will include the study for an algorithm able to improve the selection among the transcription alternatives, which should not be based on apriori knowledge of the original text sentence.

Acknowledgements The authors would like to thanks Fabio Ballati for his contribution to the data analysis and for the software implementation to interact with each ASR cloud platform.

References

1. Ballati, F., Corno, F., De Russis, L.: Assessing virtual assistant capabilities with italian dysarthric speech. In: Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '18, pp. 93–101. ACM, New York, NY, USA (2018). DOI 10.1145/3234695.3236354
2. Ballati, F., Corno, F., De Russis, L.: "hey siri, do you understand me?": Virtual assistants and dysarthria. In: Intelligent Environments 2018: Workshop Proceedings of the 14th International Conference on Intelligent Environments, pp. 557–566. IOS Press (2018). DOI 10.3233/978-1-61499-874-7-557
3. Bigham, J.P., Kushalnagar, R., Huang, T.H.K., Flores, J.P., Savage, S.: On how deaf people might use speech to control devices. In: Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS'17. ACM Press (2017). DOI 10.1145/3132525.3134821
4. DeRosier, R., Farber, R.S.: Speech recognition software as an assistive device: a pilot study of user satisfaction and psychosocial impact. Work **25**(2), 125–134 (2005)
5. Enderby, P.: Frenchay dysarthria assessment. International Journal of Language & Communication Disorders **15**(3), 165–173 (1980). DOI 10.3109/13682828009112541

6. Glasser, A.T., Kushalnagar, K.R., Kushalnagar, R.S.: Feasibility of using automatic speech recognition with voices of deaf and hard-of-hearing individuals. In: Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS'17. ACM Press (2017). DOI 10.1145/3132525.3134819
7. Google: Cloud speech-to-text. https://cloud.google.com/speech-to-text/ (2018)
8. Hawley, M.S.: Speech recognition as an input to electronic assistive technology. British Journal of Occupational Therapy **65**(1), 15–20 (2002). DOI 10.1177/030802260206500104
9. IBM: Watson speech to text. https://www.ibm.com/cloud/watson-speech-to-text (2018)
10. Joy, N.M., Umesh, S.: Improving acoustic models in torgo dysarthric speech database. IEEE Transactions on Neural Systems and Rehabilitation Engineering **26**(3), 637–645 (2018). DOI 10.1109/TNSRE.2018.2802914
11. Kent, R.D.: Research on speech motor control and its disorders: a review and prospective. Journal of communication disorders **33**, 391–427; quiz 428 (2000)
12. Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., Frame, S.: Dysarthric speech database for universal access research. In: Interspeech, pp. 1741–1744 (2008)
13. Kim, M., Kim, Y., Yoo, J., Wang, J., Kim, H.: Regularized speaker adaptation of kl-hmm for dysarthric speech recognition. IEEE Transactions on Neural Systems and Rehabilitation Engineering **25**(9), 1581–1591 (2017). DOI 10.1109/TNSRE.2017.2681691
14. Kim, M., Wang, J., Kim, H.: Dysarthric speech recognition using kullback-leibler divergence-based hidden markov model. In: Interspeech 2016, pp. 2671–2675 (2016). DOI 10.21437/Interspeech.2016-776
15. Koester, H.H.: Usage, performance, and satisfaction outcomes for experienced users of automatic speech recognition. The Journal of Rehabilitation Research and Development **41**(5), 739 (2004). DOI 10.1682/jrrd.2003.07.0106
16. Menendez-Pidal, X., Polikoff, J.B., Peters, S.M., Leonzio, J.E., Bunnell, H.T.: The nemours database of dysarthric speech. In: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, vol. 3, pp. 1962–1965 (1996). DOI 10.1109/ICSLP.1996.608020
17. Microsoft: Bing speech. https://azure.microsoft.com/en-us/services/cognitive-services/speech/ (2018)
18. Pradhan, A., Mehta, K., Findlater, L.: "accessibility came by accident": Use of voice-controlled intelligent personal assistants by people with disabilities. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, pp. 459:1–459:13. ACM, New York, NY, USA (2018). DOI 10.1145/3173574.3174033
19. Rudzicz, F.: Using articulatory likelihoods in the recognition of dysarthric speech. Speech Communication **54**(3), 430 – 444 (2012). DOI 10.1016/j.specom.2011.10.006
20. Rudzicz, F., Namasivayam, A.K., Wolff, T.: The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Language Resources and Evaluation **46**(4), 523–541 (2012). DOI 10.1007/s10579-011-9145-0
21. Yu, J., Xie, X., Liu, S., Hu, S., Lam, M.W.Y., Wu, X., Wong, K.H., Liu, X., Meng, H.: Development of the cuhk dysarthric speech recognition system for the ua speech corpus. In: Proc. Interspeech 2018, pp. 2938–2942 (2018). DOI 10.21437/Interspeech.2018-1541
22. Zue, V., Seneff, S., Glass, J.: Speech database development at MIT: Timit and beyond. Speech Communication **9**(4), 351–356 (1990). DOI 10.1016/0167-6393(90)90010-7