Intelligent Environments 2018
I. Chatzigiannakis et al. (Eds.)

doi:10.3233/978-1-61499-874-7-557

"Hey Siri, Do You Understand Me?": Virtual Assistants and Dysarthria

Fabio BALLATI [a], Fulvio CORNO [a] and Luigi DE RUSSIS [a,1]
[a] *Politecnico di Torino, Corso Duca degli Abruzzi, 24 Torino, Italy 10129*

Abstract. Voice-activated devices are becoming common place: people can use their voice to control smartphones, smart vacuum robots, and interact with their smart homes through virtual assistant devices like Amazon Echo or Google Home. The spread of such voice-controlled devices is possible thanks to the increasing capabilities of natural language processing, and generally have a positive impact on the device accessibility, e.g., for people with disabilities. However, a consequence of these devices embracing voice control is that people with dysarthria or other speech impairments may be unable to control their intelligent environments, at least with proficiency. This paper investigates to which extent people with dysarthria can use and be understood by the three most common virtual assistants, namely Siri, Google Assistant, and Amazon Alexa. Starting from the sentences in the TORGO database of dysarthric articulation, the differences between such assistants are investigated and discussed. Preliminary results show that the three virtual assistants have comparable performance, with an accuracy of the recognition in the range of 50-60%.

Keywords. dysarthria, conversational assistant, smart home, persons with disabilities, virtual assistants

1. Introduction

The last five years have seen the spread of voice-controlled smart environments, powered by virtual assistants like Siri or Amazon Alexa. Nowadays, people can speak to their smartphones, smart watches, smart homes, connected vacuum robots, and even smart cars, with the aims of setting alarms, controlling other devices in their smart environments, playing music, or requiring various types of information (e.g., the weather forecast). This spread of voice-controlled devices was possible thanks to the advances made in speech recognition, and was seen as a viable alternative to touch screens. Displays, in fact, are typically more expensive and bulky as components, and they are impossible to operate hands-free. By using speech as the primary input, virtual assistants can bypass or minimize the more "conventional" input methods (i.e., keyboard, mouse, and touch), thus making voice-controlled devices useful and accessible to people with disabilities. However, while persons with motor disabilities may benefit from these virtual assistants, those with cognitive, sensory, or speech disorders may be unable to fully use them. For example, Bigham et al. [1] demonstrated that the Google's speech recognition system

[1] Corresponding Author: Luigi De Russis, Politecnico di Torino, Corso Duca degli Abruzzi, 24 Torino, Italy 10129; E-mail: luigi.derussis@polito.it.

does not work well for people who are deaf and hard of hearing, and they expected that recognizing deaf speech will remain challenging for both automatic and human-powered approaches.

In this paper, we present an initial work on enabling people with speech impairments to access voice-controlled devices, adopted more and more in smart homes around the world. In particular, we focus on people with *dysarthria*, a motor speech disorder characterized by poor articulation of phonemes that makes it difficult to pronounce words. We investigated the interaction of people with dysarthria with three of the most used virtual assistants, included in several standalone and mobile devices: Apple's Siri, Google Assistant, and Amazon Alexa. With such users, at what point do virtual assistants become limited and demonstrate a low reliability? Could people with different degrees of dysarthria easily access and be understood by those voice-controlled devices? Even if the percentage of recognized speech is low, could virtual assistants leverage the context as extracted from the requests to provide a suitable answer?

To answer these questions and investigate the differences between the three virtual assistants, we extracted 17 appropriate sentences from the TORGO database of dysarthric articulation [2]. The database contains dysarthric speech samples from eight speakers with cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS) and was developed in a collaboration between the departments of Computer Science and Speech-Language Pathology at the University of Toronto and the Holland-Bloorview Kids Rehab hospital in Toronto, Canada. We, then, submitted each speech sample to every virtual assistant, separately, and analyzed the given answers. We were interested in the dysarthric *sentence comprehension* and in the *consistency of the answers* as indicators of the reliability of such assistants. With the former, we evaluate the accuracy of the speech-to-text recognition, while the latter is an indicator of the appropriateness of the responses provided by the assistants. Results show that the three virtual assistants have comparable performance for both sentence comprehension and consistency of the answers, with a correct transcription percentage of around 50-60%.

2. Related Work

Speech technology in general, and automatic speech recognition (ASR) in particular, is not new for people with disabilities. It was being used to increase accessibility in mainstream operating systems since decades, as an alternative method to control the computer or to compose document through dictation systems (e.g., Dragon [3]). For example, while speech recognition is available in Microsoft Windows since Vista, the latest version of the operating system include the Cortana digital assistant to help users set reminders, open apps, find information, and send emails and texts [4]. Similarly, speech recognition as an input to electronic assistive technology was investigated both in general and for dysarthria. Hawley [5] presents an 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 reports that, given adequate time, training, and support, commercial ASR systems for PCs 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 the home or other electronic devices is less reliable and more problematic, especially for dysarthric speech. To overcome this kind of issues, researchers investigated new methods and proposed dedicated ASR systems for dysarthria, e.g., by using ergodic hidden Markov models [6] or articulatory dynamic Bayes networks [7].

Specific HCI research in the domain of technology for people with speech impairments is, instead, still quite limited [8]. Sears et al. [9] offer an overview of HCI research for people with "significant speech and physical impairments", by focusing on communication aids. More recently, Derboven et al. [8] describe the design of ALADIN, a self-learning speech recognition system for people with physical disabilities, many of whom also have speech impairments. ALADIN is designed to allow users to use their own specific words and sentences, adapting itself to the speech characteristics of the user, and primarily targets smart homes.

Finally, a few works explore usability, reliability, and accessibility issues of virtual assistants. López et al. [10] present a usability evaluation of the most used speech-based virtual assistants (i.e., Alexa, Siri, Cortana, and Google Assistant) and highlight that there is still a lot to do to improve the usability and reliability of these systems. Bigham et al. [1], instead, focus on the issues that may arise from the usage of commercial virtual assistants by people who are deaf and hard of hearing. They propose two technical approaches for enabling deaf people to provide input to those assistants, i.e., human computation workflows for understanding speech and mobile interfaces that can be instructed to speak on the user's behalf. Similarly to the work of Bigham et al., we focus on the issues that may arise from the usage of virtual assistants, but we were specifically interested in dysarthric speech and in evaluating the current behaviors of these assistants.

3. Dysarthric Speech for Virtual Assistants

Dysarthric speech is the speech produced by people with dysarthria. Dysarthria is caused by disruptions in the neuro-motor interface, typically as a consequence of cerebral palsy or the Parkinson's disease. These disruptions distort motor commands to the vocal articulators, thus resulting in atypical and relatively unintelligible speech in most cases. Dysarthric speech may be characterized by a slurred, nasal-sounding or breathy speech, an excessively loud or quiet speech, problems speaking in a regular rhythm, with frequent hesitations, and monotone speech. As a consequence of these problems, a person with dysarthria may be difficult to understand and, in some cases, she may only be able to produce very short phrases, single words, or no intelligible speech at all. As a result, enabling modern ASR to effectively understand dysarthric speech is a major need, both for virtual assistants and for computers, since other physical impairments often associated with dysarthria can make other forms of input, such as keyboards or touch screens, especially difficult.

To begin exploring the issues of understanding dysarthric speech by contemporary virtual assistants, we extracted a corpus of sentences from the *TORGO database* [2]. The database is one of the few freely available[2] collections of dysarthric speech in English, and it consists of aligned acoustics and measured 3D articulatory features from speakers with either CP or ALS, all of them with various degree of dysarthria. The dataset contains

[2] for academic and non-profit purposes

both audio files and transcriptions of four types of sentences to control different abilities of dysarthric speakers: *non-words*, *short words*, *restricted sentences*, and *unrestricted sentences*. Non-words were used to control baseline abilities of dysarthric speakers, especially to gauge their articulatory control in the presence of plosives and prosody, like high-pitch and low-pitch vowels. Short words are useful for studying speech acoustics and for hypothetical command for accessible software. They include the repetitions of the English digits, and words like 'yes', 'no', 'up', 'down', 'back', 'select', etc. Restricted sentences are full and syntactically correct sentences, including "The Grandfather Passage[3]" and phoneme-rich sentences like "The quick brown fox jumps over the lazy dog." Finally, unrestricted sentences were natural descriptive text of 30 different images, to more accurately represent disfluencies and syntactic variation of natural speech.

To evaluate dysarthric speech with the three virtual assistants, we looked in the TORGO database for sentences similar to the commands used to control Alexa, Siri, and Google Assistant (e.g., [11]) in the home. We extracted 5 different sentences pronounced by 7 different speakers. Those sentences were extracted from all the types of sentences included in the dataset, excluding non-words. We, then, used these commands to evaluate the sentence comprehension and the answers consistency of the virtual assistants. The five unique sentences are:

1. Some hotels are available nearby
2. Please, open the window
3. Today's date
4. Start
5. Play

4. Evaluation

The evaluation of Siri, Alexa, and Google Assistant for dysarthric speech focuses on *sentences comprehension* and *consistency of the answers*.

For *sentence comprehension*, we evaluate the accuracy of the speech-to-text recognition process adopted by the virtual assistants. Both Siri and Google Assistant, in fact, provide the user with the transcription of the received command (if operated on devices with a display). The evaluation of the comprehension is, therefore, given by the similarity between the expected transcription (as provided in the TORGO database) and the transcribed output of the assistants, both in terms of the number of correctly recognized sentences and as the Word Error Rate (WER). Alexa, instead, does not provide a transcription of the request but it only gives a binary indication about the recognition of the input (i.e., it warns the user if it was not able to recognize the input speech). Therefore, we only qualitatively compared Alexa with the other assistants, for this criterion.

Consistency of the answers is an indicator of the appropriateness of the assistants' responses, given as the number and percentage of times that the three assistants provided appropriate responses to the user's queries. Even if the accuracy of the speech-to-text recognition process is low, in fact, virtual assistants may leverage the context or some specific recognized keywords to provide a suitable response.

[3] a public domain text frequently used to gather a speech sample that contains nearly all of the phonemes of American English.

4.1. Study Description

We extracted the 5 sentences reported in the previous section from the TORGO database, as pronounced by 7 persons with dysarthria (5 males and 2 females). Since the sentences were not available for all the users, we obtained the audio files of 17 sentences, overall. In details, the first sentence was pronounced by two people (F3 and M5), the second sentence by F4 only, the third by M5 only, the fourth sentence by all the users except M4, while the last sentence was pronounced by all the users. After the selection of the sentences, we ensured that all the speech samples were perfectly recognizable by a human listener, to avoid submitting to the assistants any sentence that even a person would not be able to understand. Tables 1, 2, 3 report all the 17 sentences with the details of the study.

The evaluation took place in a quiet room of our university. The speech samples were played on a laptop connected to an external speaker. Each sentence was played for each virtual assistant, separately, and the results of the operation (i.e., recognized request and related response) were noted down by the experimenter. The virtual assistants were run on dedicated devices: an iPhone 7 (iOS 11.2) was used for Siri, a Nexus 5X (Android 8.1) for Google Assistant, while Amazon Alexa was used through a browser-based interface (i.e., the Amazon Echo Simulator [12]). The Amazon Echo Simulator console, in particular, was useful to help overcome the absence of the requests transcription.

Before starting with the evaluation, we extracted from the TORGO database the same 5 sentences but pronounced by people without any speech impairment. We carefully and successfully checked that each virtual assistant recognized and transcribed those speech samples without any errors nor problems.

5. Results and Discussion

Overall, the three virtual assistants performed similarly with the 17 dysarthric speech samples for both *sentences comprehension* and *consistency of the answers*, as reported in Tables 1 (results with Siri), 2 (Google Assistants) and 3 (Amazon Alexa), and summarized in Table 4. For what concerns sentence comprehension, Siri was the only assistant that *tried* to recognize all 17 sentences, by transcribing something. The other two assistants indicated, instead, that they were not able to recognize anything for some speech samples. Google Assistant performed the worst, with 4 not recognized speech samples, while Amazon Alexa did not recognize 3 sentences, only. However, this difference between the virtual assistants disappeared when analyzing the **correct** transcriptions and the relative Word Error Rate (WER)[4]. Siri, in fact, was able to correctly recognize 8 speech samples, i.e., all but one of the sentences pronounced by F3, F4, M3, and M4, with an average WER of 69.41% and an overall recognition percentage of 47%. Google Assistant performed better with a recognition percentage of 58.82%: it was able to recognize 10 speech samples, with an average WER of 15.38%. Differently from Amazon Alexa and Siri, Google Assistant recognized a sentence from M2 and one from M5, indeed.

[4] WER is a commonly used performance measure for speech recognition systems that includes substitution errors (i.e., miss-recognition of one word for another), deletion errors (i.e., words missed by the recognition system), and insertions (i.e., words introduced into the text output by the recognition system). It can be greater than 100% when the transcription has more insertions than deletions.

User	Original Sentence	Siri Transcription	Correct?	WER	Siri Response	Appropriate?
M01	play	hey	No	100%	Hello	No
	start	go	No	100%	You were saying	No
M02	play	hello	No	100%	Hi there	No
	start	can i	No	200%	Interesting question	No
M03	play	play	Yes	0%	Ok... (play some music)	Yes
	start	start	Yes	0%	I'm not sure I understand	Yes
M04	play	play	Yes	0%	Ok... (play some music)	Yes
M05	some hotels are available nearby	resume route girl or a rare burger nearby	No	140%	Ok. Here's what I found nearby	Yes
	play	siri hi	No	200%	Hi, what can I do for you?	No
	start	no		No	100% Ok, I didn't think so	No
	today's date	do you do do	No	200%	This is about you, not me	No
F03	some hotels are available nearby	show hotel are available nearby	No	40%	I found quite a few hotels fairly close to you	Yes
	play	play	Yes	0%	Playing all songs, shuffled	Yes
	start	start	Yes	0%	Ok...	Yes
F04	please open the window quickly	please open the window quickly	Yes	0%	Hmm, I don't see anything connected, but I can help once you've set something up	Yes
	play	play	Yes	0%	Playing all songs, shuffled	Yes
	start	start	Yes	0%	I'm not sure I understand	Yes

Table 1. The full list of sentences per user with the responses provided by Siri. The correctness of the transcription, the Word Error Rate, and the appropriateness of the response are reported.

Finally, for what concerns the consistency of the answers, all three virtual assistants were consistent in their answers, e.g., Siri with the "play" sentences always executed some music, Google Assistant always proposed some games to play, while Amazon Alexa always replied with "what do you want to hear?". The appropriateness of the responses was similar for the three assistants, as they leveraged the context or some specific keywords to provide a suitable answer. In particular, Google Assistant tried to provide a pertinent answer when it recognized some words, e.g., it showed a link to the AroundMe app or some TripAdvisor pages when it recognized "hotels" or "nearby". A similar behavior was exhibited by Siri with the "nearby" word. However, Google Assistant performed slightly better than Siri (11 vs. 10 appropriate responses), while Alexa

User	Original Sentence	Assistant Transcription	Correct?	WER	Assistant Response	Appropriate?
M01	play	-	No	-	-	No
	start	dart	No	100%	Search "dart" on Google	No
M02	play	play	Yes	0%	You can play one of this games from Playstore	Yes
	start	-	No	-	-	No
M03	play	play	Yes	0%	You can play one of this games from Playstore	Yes
	start	start	Yes	0%	This came back from a search of "start" in the dictionary	Yes
M04	play	play	Yes	0%	You can play one of this games from Playstore	Yes
M05	some hotels are available nearby	hotels available	No	60%	Search on Google (hotels)	Yes
	play	-	No	-	-	No
	start	-	No	-	-	No
	today's date	today's date	Yes	0%	12/12/2017	Yes
F03	some hotels are available nearby	someone tells are available nearby	No	40%	Here to help (suggest the "Around Me" app)	Yes
	play	play	Yes	0%	You can play one of this games from Playstore	Yes
	start	start	Yes	0%	This came back from a search of "start" in the dictionary	Yes
F04	please open the window quickly	please open the window quickly	Yes	0%	Search on Google (window quickly)	No
	play	play	Yes	0%	You can play one of this games from Playstore	Yes
	start	start	Yes	0%	This came back from a search of "start" in the dictionary	Yes

Table 2. The full list of sentences per user with the responses provided by Google Assistant. The correctness of the transcription, the Word Error Rate, and the appropriateness of the response are reported.

performed the worst, by providing 7 suitable response, only. It should be notice that the Siri approach of trying of answering in any case provided a benefit in this case: it moved, in fact, from 8 correctly transcribed sentences to 10, while Google Assistant moved from 10 to 11.

User	**Original Sentence**	**Recognized by Alexa?**	**Alexa Response**	**Appropriate?**
M01	play	No	-	No
	start	Yes	-	No
M02	play	Yes	Hi there	No
	start	No	-	No
M03	play	Yes	What do you want to hear?	Yes
	start	Yes	What do you want to hear?	Yes
M04	play	Yes	What do you want to hear?	Yes
M05	some hotels are available nearby	Yes	I'm still learning about bird questions. Try asking "tell me a bird fact" to learn more about them.	No
	play	Yes	-	No
	start	Yes	-	No
	today's date	No	-	No
F03	some hotels are available nearby	Yes	To search for local businesses you need to enter your address in your settings	
	play	Yes	What do you want to hear?	Yes
	start	Yes	What do you want to hear?	Yes
F04	please open the window quickly	Yes	Window quickly doesn't support there	No
	play	Yes	What do you want to hear?	Yes
	start	Yes	What do you want to hear?	Yes

Table 3. The full list of sentences per user with the responses provided by Amazon Alexa. The appropriateness of the response is indicated in the last column.

	Sentence Comprehension			**Consistency of the Answers**	
	Recognized Sentences	*Correct Transcriptions (# and %)*	*WER (mean)*	*Appropriate Responses (#)*	*Appropriate Responses (%)*
Siri	**17**	8 (47%)	69.41%	10	58.82%
Google Assistant	13	**10 (58.82%)**	**15.38%**	**11**	**64.7%**
Amazon Alexa	14	-	-	7	41.17%

Table 4. Summary of the results according to the sentence comprehension and consistency of the answers criteria.

5.1. Discussion

Starting from the results about dysarthric speech comprehension and consistency of the answers, a similarity and a couple of differences emerge. The similarity is related to the almost equal level of recognition, with a percentage of correct transcriptions around the range of 50-60% (more precisely, 47-58.82%): a similar range was already found for contemporary ASR systems used by users with other speech impairments [1]. While slight differences in which sentences were recognized by the assistants exist, they seem not to be significant, at least with the limited data available in the TORGO database. The main difference between the virtual assistants is, instead, related to the provided answers. While Siri always tries to answer any request, even if it does not recognize any word, Amazon Alexa and Google Assistant use an opposite approach as they provide a response if they recognize some words, only. Such fallback mechanisms, however, are different for Amazon Alexa and Google Assistant: while the former may say "I'm still learning about bird questions", the latter starts a Google search with the recognized words. Another difference we noticed during the evaluation is that Amazon Alexa tries not to reply to single-word commands as it seems to prefer longer sentences.

5.2. Study Limitations

We would like to acknowledge that this evaluation presents two limitations, which emphasize the preliminary nature of this work. The first one is the relatively low number of speech samples present in the TORGO database that are suitable for virtual assistants. The second one resides in the choice of playing sentences from a speaker instead of by a human speaker. This was, obviously, inevitable since we chose to adopt the TORGO dataset for this work. While we do not have any evidence that this choice negatively impacted the results of the evaluation, involving human participants may improve the ecological validity of the results.

6. Conclusion and Future Work

Voice-controlled smart environments, powered by virtual assistants like Siri or Amazon Alexa, are becoming mainstream. The reliability of the intelligent environment they control, strongly depend on their capability of understanding the requests they receive, and of correctly acting on the environment.

In this paper, we have presented an initial investigation of the accessibility challenges presented by such virtual assistants for dysarthric speech. By using 5 different sentences pronounced by 7 diverse speakers with dysarthria, we evaluated the performances of the three most common virtual assistants, i.e., Siri, Google Assistant, and Amazon Alexa, according to two criteria: dysarthric sentence comprehension and consistency of the answers. Preliminary results show that the three assistants have comparable performance and similar behaviors for both criteria, with a recognition percentage of around 50-60%. Similar recognition values were already found for contemporary ASR systems when used by people with other speech impairments, e.g., deaf people.

Future work will include a more extensive evaluation, both in variety and in number of sentences pronounced by people with dysarthria. Moreover, we would like to better assess such virtual assistants by characterizing their usefulness for different degree of dysarthric speech (e.g., moderate vs. severe) as well as by employing speech-controlled devices like Google Home and Amazon Echo instead of smartphones and web-based interfaces to better assess their reliability in an intelligent environment. Finally, we will use the outcome of this evaluation as a starting point to improve the accessibility and the recognition capabilities of such assistants.

References

[1] J. P. Bigham, R. Kushalnagar, T.-H. K. Huang, J. P. Flores, S. Savage, On how deaf people might use speech to control devices, *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, 383–384, New York, USA, 2017.

[2] F. Rudzicz, A. K. Namasivayam, T. Wolff, The TORGO database of acoustic and articulatory speech from speakers with dysarthria, *Language Resources and Evaluation* **46**, 4 (2012), 523–541.

[3] Nuance Communications Inc., Dragon speech recognition, January 2018, `https://www.nuance.com/dragon.html` (last visited on April 14, 2018).

[4] Microsoft Inc., Windows accessibility, `https://www.microsoft.com/en-us/accessibility/windows` (last visited on April 14, 2018).

[5] M. S. Hawley, Speech recognition as an input to electronic assistive technology, *British Journal of Occupational Therapy* **65**, 1 (2012), 15–20.

[6] P. D. Polur, G. E. Miller, Investigation of an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals, *Medical Engineering & Physics* **28**, 8 (2006), 741–748.

[7] F. Rudzicz, Using articulatory likelihoods in the recognition of dysarthric speech, *Speech Communication* **54**, 3 (2012), 430–444.

[8] J. Derboven, J. Huyghe, D. De Grooff, Designing voice interaction for people with physical and speech impairments, *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, 217–226, New York, USA, 2014.

[9] A. Sears, M. Young, The human-computer interaction handbook, *Physical Disabilities and Computing Technologies: An Analysis of Impairments*, 482–503, L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 2003.

[10] G. López, L. Quesada, L. A. Guerrero, Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces, 241–250, Springer International Publishing, 2018.

[11] T. Martin, D. Priest, The complete list of Google Home commands so far, `https://www.cnet.com/how-to/google-home-complete-list-of-commands/` (last visited on April 14, 2018).

[12] iQuarius Media, Echosim.io community edition, `https://echosim.io` (last visited on April 14, 2018).