Design for Inclusion
I. Garofolo and G. Bencini (Eds.)

doi:10.3233/SHTI230394

Towards Universally Designed Communication: Opportunities and Challenges in the Use of Automatic Speech Recognition Systems to Support Access, Understanding and Use of Information in Communicative Settings

Martina PUCCI[a,1]
[a] *Ca' Foscari University of Venice*
ORCiD ID: Martina Pucci https://orcid.org/0000-0002-6561-333X

Abstract. Unlike physical barriers, communication barriers do not have an easy solution: people speak or sign in different languages and may have wide-ranging proficiency levels in the languages they understand and produce. Universal Design (UD) principles in the domain of language and communication have guided the production of multimodal (audio, visual, written) information. For example, UD guidelines encourage websites to provide information in alternative formats (for example, a video with captions; a sign language version). The same UD for Learning principles apply in the classroom, and instructors are encouraged to prepare content to be presented multimodally, making use of increasingly available technology. In this chapter, I will address some of the opportunities and challenges offered by automatic speech recognition (ASR) systems. These systems have many strengths, and the most evident is the time they employ to convert speech sounds into a written form, faster than the time human transcribers need to perform the same process. These systems also present weaknesses, for example, a higher rate of errors when compared to human-generated transcriptions. It is essential to weigh the strengths and weaknesses of technology when choosing which device(s) to use in a universally designed environment to enhance access to information and communication. It is equally imperative to understand which tools are most appropriate for diverse populations. Therefore, researchers should continue investigating how people process information in a multimodal format, and how technology can be improved based on this knowledge and users' needs and feedback.

Keywords. automatic speech recognition, captions, transcriptions, technology, communication, communicative settings, universal design, universal design for learning, universal design and individual differences, multimodality.

[1] Martina Pucci, Department of Linguistics and Comparative Cultural Studies, Ca' Foscari University of Venice, Ca' Bembo, Fondamenta Tofetti – Dorsoduro 1075, Venice, Italy; E-mail: martina.pucci@unive.it.

1. Introduction

Access to information in critical domains for citizenship and well-being such as, for example, health, emergency information, individual rights and the law by any means of communication (e.g., television, the internet) are fundamental rights for all citizens. However, there are instances in which individual circumstances or characteristics hamper access to information or communication. Unlike physical barriers, communication barriers are difficult to overcome due to the heterogeneity of populations: factors such as literacy, proficiency, disability, can interfere with access to information if this is presented in ways that are unavailable or inaccessible to the end user [1]. Universal Design (UD) principles and Universal Design for Learning (UDL) guidelines have contributed to the matter by proposing strategies to overcome communication barriers depending on different contexts and speakers [2-4]. For example, these guidelines suggest the employment of *multimodality*, that is, the presentation of information in more than one modality (e.g., spoken + written modalities), also with the help of technological tools [5]. One of the devices that can help with the multimodal presentation of information is automatic speech recognition (ASR), a system that turns the speech signal into a written transcription [6].

The purpose of this essay is to discuss the role of ASR systems in facilitating access to information for diverse users. I will discuss the opportunities ASR systems offer and the current limitations of the technology. It will also provide a brief overview of the roles of UD principles and UDL guidelines in developing policies to increase the use of technology suited to diverse users. Finally, I will address some of the challenges developers and researchers have to face before implementing these systems in universally designed environments.

2. Universal Design and technology use

UD principles guide experts in the design of environments, products, and communication systems [2-4]. The ultimate goal is to ensure that users benefit from final designs without having to be further adapted to their needs. While this approach can contribute to the elimination of architectural barriers in public or private buildings, there is no easy solution to remove *communication barriers*. Different native languages between two or more speakers or various proficiency levels in a foreign language can hinder access and dissemination of information in communicative settings. In the same way, the modality with which information is conveyed is another frequent obstacle. One modality may be available to a large segment of the population, but not to others: for example, audio input for deaf and hard-of-hearing people or written input for blind people. If information is delivered in only one modality (e.g., spoken or written only), it could preclude access to crucial information for these individuals, potentially excluding them from communication. Luckily, technological advancements in the last decades have contributed to removing some of these barriers. Today, these tools can assist people in presenting information in more than one modality simultaneously (*multimodality*), such as combined spoken and written input. For these reasons, UD principles and UDL guidelines encourage designers and instructors to combine the use of various technological tools to deliver information multimodally, supporting users and communication [5].

Researchers have been studying the cognitive mechanisms underlying the processing of multimodal input for years. Specifically, research has highlighted how the presentation of information in multiple modalities (e.g., audio + written input) benefits language comprehension, vocabulary learning, and memory for content [7, 8]. For example, the simultaneous presentation of auditory input and written transcription help diverse students recover missing or incomplete information [9]. Another example concerns low-proficient speakers of a foreign language. In this case, written input can help these speakers in segmenting their interlocutor's speech stream while following a lecture [10].

With this in mind, governments and supranational organizations have developed policies and projects aimed at improving the use of technology and enhancing the inclusion of diverse users in various settings, especially in the educational one [11-13]. The higher education sector has begun to equip its buildings with more advanced technological tools and adopt UD and UDL guidelines to promote inclusion, but this process is still ongoing [14, 15]. Specifically, UDL guidelines recommend the use of technology to enhance individual autonomy and encourage the use of alternative learning strategies. These guidelines also prompt instructors to explore different methods of presenting the content of their lectures to provide easier access to information and improve communication, combining multimodality and technology use [5].

In the last few decades, developers have focused on building a system that institutions are starting to employ to present information multimodally and improve communication, that is, automatic speech recognition (ASR).

3. ASR: opportunities and current limitations

Automatic speech recognition is defined as "the process of converting a speech signal into a sequence of words (i.e., spoken words to text) by means of an algorithm implemented as a computer program" [6 - p. 394], with *words* defined as the "best-decoded sequence of linguistic units" [16 - p. 18]. A closer examination of the standard structure of an ASR system reveals that it replicates more simply some of the processes involved in human speech processing and language comprehension [16]. ASR systems are composed of five components:

- The *acoustic front-end* is devoted to speech signal analysis and feature (or parameter) extraction [6].
- The *acoustic model* is a list of statistical representations of the sounds (phones) of words [6].
- The *language model* is the module devoted to word identification. It clusters phones into words, helping the acoustic model disambiguate the phones in a chain. It also groups words based on the statistical probability of appearing together in a sequence [6].
- The *lexicon* is a list of words (with their phonological description) that interacts with the acoustic and language models. As part of the building process of the system, developers create and define not only this list of words but also the data contained in the acoustic and language models [6, 16].
- The *decoder* is an algorithm that searches for "the most likely word sequence *w* given the observation sequence *o*, and the acoustic-phonetic-language model" based on the target language [6].

While building ASR systems, developers provide sounds, words, and rules of a target language to the two models and the lexicon. This process (alongside the training stage) has the goal of improving the decoding processing of the speech signal and increasing the accuracy of transcriptions from speech to text. When given an input, the acoustic front-end analyzes the speech signal and extracts the relevant features to be processed. The decoder computes the words based on the extracted features and the data contained in the models and the lexicon. The output of this decoding stage is the written transcript of the hypothesized words in the speech signal analyzed by the system [6]. The most relevant outputs of ASR systems are transcriptions and automatic captions - that is, subtitles generated in real-time by an ASR system while a speaker is talking. It usually takes a few seconds for ASR to generate captions: this is due to the computing phase and printing of the outputs. These processes, therefore, cause brief temporal misalignments between spoken input and written output.

ASR is rapidly becoming one of the most relevant tools for multimodal access to information. This technology improves human-human interaction in settings where language barriers impede communication (for example, in the absence of interpreters or when they provide interpreting services only for one language) [9], providing multimodal access to information. These systems can be employed in different settings, such as work meetings, conferences, national and supranational institution sessions (e.g., national parliaments), and legal processes. Users can read whole transcripts after the end of such meetings, but they can also access information on a PC with the aid of real-time captions due to the fast speed of conversion of the speech signal into written transcripts, facilitating communication [17, 9]. The advantage of this method lies in the fact that machines transcribe speech inputs faster than humans. In this last case, the transcription process is a costly and time-consuming task that usually requires days to be carried out [18]: ASR systems thus help users accelerate this process by automating it.

Researchers have been investigating how diverse populations may use ASR outputs in various settings. For example, in educational settings, students may use transcriptions to support note-taking and revision of previous notes [1, 9, 19-22]. Regarding real-time captioning, researchers have investigated how captions affect comprehension during work meetings where communication is carried out in a foreign language (English in the majority of cases). In this setting, users can rely on captions to understand the information conveyed by spoken utterances in those circumstances where comprehension may have been impeded. For this reason, these users prefer that the ASR-generated captions align as closely as possible with the speech signal, while the accuracy of transcription is not deemed as relevant as the speed of text presentation [17, 23]. However, the accuracy of captions plays a relevant role in other settings, aiding access to information and supporting listening comprehension [10, 19, 23]. In the educational setting, for example, research on the benefits of captions on language comprehension in different populations, has mainly focused on human-made subtitles, while ASR-generated captions have generally received less attention. ASR has been the core of the Liberated Learning concept, a project that focuses on how these systems could "provide universal access to lecture material for students with diverse backgrounds" [1, 19-22]. Experimental studies conducted in this project and others have highlighted the benefits for students (e.g., an increase in word recognition and better comprehension of the content of lectures), but they have also sought to surface common problems with current ASR systems. The major problem (as already briefly mentioned) was the lack of accuracy in transcriptions [1, 9, 19-22].

Improving the accuracy of ASR systems has always been a challenge for developers, and this is due to various reasons. Speech variability is one of these factors that affect ASR systems' accuracy and general robustness (defined as "the system's ability to successfully deal with different aspects of variability in the speech signal" by Karpagavalli & Chandra, 2016 - p. 401). Speech variability is linked to:

- Characteristics of speakers (physical traits that affect voice structure);
- Sociolinguistic factors (regional or foreign accents);
- Spontaneous speech (speech rate, connected speech, disfluencies such as false starts, hesitations, etcetera);
- Emotions [6, 24, 25].

Each of these factors defines the uniqueness of each speaker. At the same time, they challenge the ability of ASR to accurately decode speech signals: these factors alter the spectrum of the speech signal, affecting its features and preventing the correct decoding of sounds, lowering the accuracy of transcriptions [24]. Additionally, signal degradation may be caused by external factors, including the structure of the ASR system itself, environmental noise, and the quality of the hardware that collects the speech signals [6, 24, 25]. It is clear that, if these systems are not accurate at transcribing speech signals, they cannot be reliable for a wide range of users, as this will hinder comprehension when they should be helping [9, 26].

4. ASR and technology: what’s next? Considerations for the future

Developers are facing many challenges to improving ASR [24, 25], and some of the problems are currently undermining the implementation of these systems as a primary service in various settings [9]. Feedback from users also highlights the need to increase transcription accuracy before considering the adoption of this technology to support communication and facilitate access to information in universally designed environments [1, 9, 20, 21].

For ASR to become fully implemented, a well-refined and ethically-approved user-centered approach will have to be incorporated into the process of technology development [27]. Specifically,

- Different users should be encouraged by developers throughout the development process of ASR systems to state their needs, share their doubts, and provide feedback.
- At the same time, developers need to enhance the accuracy of ASR systems by increasing the robustness of speech recognition models.
- Researchers should continue expanding their knowledge of the mechanisms underlying the processing of multimodal information by cooperating with persons with different needs.

Human beings are complex. We all have different characteristics and needs. Taking a user-centered approach to technology advancement implies that individuals from diverse populations work together with researchers and developers to (I) contribute to an increase in knowledge about the mechanisms behind human cognition and (II) express their views, requirements, and feedback during the development phase of devices.

All around the world, individuals should actively cooperate with academics by participating in research projects that aim at expanding their knowledge of the cognitive mechanisms involved in accessing and processing multimodal information. This refined knowledge should then be transferred to developers for technology development [28] since it will help them create or enhance systems that meet the characteristics of users. At the same time, developers should listen to diverse individuals and encourage them to share their requests, doubts, needs, and feedback during the entire development process of any technological tool. Research on the impact of ASR systems on comprehension where feedback from participants was collected has already demonstrated how imperative it is to listen to diverse users [9, 19, 26]. Additionally, national and supranational organizations should continue funding projects that integrate basic research with technology advancement, promoting user-centered approaches. It is critical to underline the importance of supporting these projects since progress in this kind of technology is linked to knowledge of the mechanisms behind human functioning [28].

There is another matter to remember when considering the implementation of ASR in communicative settings. Designers and instructors should remember that there does not exist a single device or system that guarantees the same degree of effective communication and access to information for diverse individuals. Therefore, they should regard ASR as one of the many tools that can potentially be employed in universally designed environments. As a large body of research has already highlighted, this is because users utilize technology based on their characteristics, needs, and strategies. For example, native speakers of a target language do not use ASR outputs the same way as second language learners [19]. To the same extent, some deaf and hard-of-hearing individuals may rely on ASR as support for the input received from sign language interpreters, while others will not [9, 26]. To promote access to information and communication, institutions should provide a range of technological tools for students, ask them which devices they prefer, helping them find the ones most suitable for all. Last, but not least, once ASR systems will have reached acceptable accuracy levels, local governments should also consider implementing ASR in public settings such as post offices, banks, hospitals, and municipal/national offices. ASR use will help improve communication between officials and citizens, facilitating access to information. The implementation of this technology in public and private settings also requires policies aimed at guaranteeing funding to buy these systems and providing support to users. Developers and researchers should continue conversations with supranational institutions, local governments, and the public to stress the importance of access to information and efficient communication.

Technology and ASR can be active players in the progress of society, ensuring equity in communication and access to information in universally designed environments where a service is provided to all.

5. Conclusions

Over the last few decades, advances in information and communication technology have widened access to information and communication for a wide range of users. Electronic devices, websites, and mobile apps have been developed following UD principles to improve access to information and facilitate communication for all via multimodality [3, 4]. At the educational level, instructors have been encouraged to follow UDL guidelines to create multimodal content with the same aim [2, 5]. ASR is one of the tools that can be used to present information multimodally, supporting equity in accessing information for diverse learners and improving communication [1, 9, 19-22]. However, this technology needs to be refined before being implemented in universally designed settings. In this essay, I focused on ASR technology, its opportunities, and current issues. ASR systems can be improved thanks to the synergistic work of developers, researchers, and users, combining basic research with technological advancement, and focusing on the characteristics of end users. Designers and instructors should be aware of the need to evaluate the use of a range of technological devices rather than choosing one solution for all. The choice should be made based on how these tools can benefit diverse users and students in the context in which they are planning to implement them.

Funding

Martina Pucci was supported by a doctoral student Research and Innovation scholarship (PON Ricerca & Innovazione 2014-2020 - CUP: H75F21002110005) funded by the Italian Ministry of Universities and Research (MUR) and the European Union (FSE REACT EU).

References

[1] Wald M, Bain K. Universal access to communication and learning: the role of automatic speech recognition, Universal Access in the Information Society. 2008;6(4):435–47.

[2] UDL: The UDL Guidelines [Internet]. 2018 [cited 2022 Dec]. Available from: https://udlguidelines.cast.org/.

[3] Definition and overview | Centre for Excellence in Universal Design. 2020 [cited 2022 Dec]. universaldesign.ie. Available from: https://universaldesign.ie/what-is-universal-design/definition-and-overview/.

[4] The 7 Principles | Centre for Excellence in Universal Design [Internet]. 2020 [cited 2022 Dec]. Available from: https://universaldesign.ie/what-is-universal-design/the-7-principles/.

[5] UDL: Offer alternatives for auditory information [Internet]. [cited 2022 Dec]. Available from: https://udlguidelines.cast.org/representation/perception/alternatives-auditory.

[6] Karpagavalli S, Chandra E. A Review on Automatic Speech Recognition Architecture and Approaches. International Journal of Signal Processing, Image Processing and Pattern Recognition. 2016;9(4):393–404.

[7] Gernsbacher MA. Video captions benefit everyone. Policy insights from the behavioral and brain sciences. 2015;2(1):195-202.

[8] Montero Perez M. Second or foreign language learning through watching audio-visual input and the role of on-screen text. Language Teaching. 2022;55:163-92.

[9] Butler J, Trager B, Behm B. Exploration of Automatic Speech Recognition for Deaf and Hard of Hearing Students in Higher Education Classes. In: The 21st International ACM SIGACCESS Conference on Computers and Accessibility [Internet]. Pittsburgh PA USA: ACM; 2019. p. 32–42. Available from: https://dl.acm.org/doi/10.1145/3308561.3353772.

[10] Venturini S, Vann MM, Pucci M, Bencini GM. Towards a More Inclusive Learning Environment: The Importance of Providing Captions That Are Suited to Learners' Language Proficiency in the UDL Classroom. In: Garofolo I, Bencini G, Arenghi A, editors. Proceedings of the Sixth International Conference on Universal Design: Transforming Our World through Universal Design for Human Development. Amsterdam, Berlin, Washington (DC): IOS Press; 2022. p. 533–40.

[11] ICT for Inclusion [Internet]. European Agency for Special Needs and Inclusive Education. [cited 2022 Dec]. Available from: https://www.european-agency.org/activities/ict4i.

[12] Digital Education Action Plan (2021-2027) | European Education Area [Internet]. [cited 2023 Jan 8]. Available from: https://education.ec.europa.eu/node/1518.

[13] Piano Nazionale Scuola Digitale – Scuoladigitale [Internet]. [cited 2022 Dec]. Available from: https://scuoladigitale.istruzione.it/pnsd/.

[14] Bencini GM, Garofolo I, Arenghi A. Implementing universal design and the ICF in higher education: Towards a model that achieves quality higher education for all. In: Craddock G, Doran C, McNutt L, Rice D, editors. Transforming Our World through Design, Diversity and Education: Proceedings of Universal Design and Higher Education in Transformation Congress 2018. Amsterdam, Berlin, Washington (DC): IOS Press; 2018. p. 464–72.

[15] Bencini G, Arenghi A, Garofolo I. Is My University Inclusive? Towards a Multi-Domain Instrument for Sustainable Environments in Higher Education. In: Verma, I, editor. Universal Design 2021: From Special to Mainstream Solutions. Amsterdam, Berlin, Washington (DC): IOS press; 2021. p. 137–43.

[16] Juang BH, Rabiner LR. Automatic speech recognition – a brief history of the technology development. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara. 2005;1:1-24.

[17] Shimogori N, Ikeda T, Tsuboi S. Automatically generated captions: will they help non-native speakers communicate in English? In: Proceedings of the 3rd ACM International Conference on Intercultural Collaboration, ICIC '10; 2010 Aug 19-20; New York, NY: Association for Computing Machinery; 2010. p. 79–86.

[18] Chan WS, Kruger JL, Doherty S. Comparing the impact of automatically generated and corrected subtitles on cognitive load and learning in a first-and second-language educational context. LANS – TTS. 2019;18:237-72.

[19] Ryba K, McIvor T, Shakir M, Paez D. Liberated Learning: Analysis of University Students' Perceptions and Experiences with Continuous Automated Speech Recognition. E-Journal of Instructional Science and Technology. 2006;9(1):1-19.

[20] Wald, M. Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. Interactive Technology and Smart Education. 2006.

[21] Wald M. An exploration of the potential of Automatic Speech Recognition to assist and enable receptive communication in higher education. ALT-J. 2006;14(1):9–20.

[22] Wald M. A research agenda for transforming pedagogy and enhancing inclusive learning through synchronised multimedia captioned using speech recognition. EdMedia+ Innovate Learning. 2007:4479–85.

[23] Cao X, Yamashita N, Ishida T. Effects of Automated Transcripts on Non-Native Speakers' Listening Comprehension. IEICE Trans Inf & Syst. 2018;E101.D(3):730–9.

[24] Benzeghiba M, De Mori R, Deroo O, Dupont S, Erbes T, Jouvet D, et al. Automatic speech recognition and speech variability: A review. Speech Communication. 2007 Oct 1;49(10):763–86.

[25] Alharbi S, Alrazgan M, Alrashed A, Alnomasi T, Almojel R, Alharbi R, et al. Automatic Speech Recognition: Systematic Literature Review. IEEE Access. 2021;9:131858–76.

[26] Berke L, Caulfield C, Huenerfauth M. Deaf and hard-of-hearing perspectives on imperfect automatic speech recognition for captioning one-on-one meetings. In: Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility. 2017. p. 155–64.

[27] Vredenburg K, Mao JY, Smith PW, Carey T. A survey of user-centered design practice. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2002. p. 471–8. (CHI '02). doi: https://doi.org/10.1145/503376.503460

[28] The long-term benefits of basic research for technology | Shaping Europe's digital future [Internet]. 2020 [cited 2022 Dec]. Available from: https://digital-strategy.ec.europa.eu/en/news/long-term-benefits-basic-research-technology.