The cocktail party effect. An inclusive vision of conversational interactions

Isabella Loddo & Dario Martini

Published online: 06 Sep 2017.

Submit your article to this journal ↗

Article views: 112

View related articles ↗

View Crossmark data ↗

The cocktail party effect. An inclusive vision of conversational interactions

Isabella Loddo[a*], Dario Martini[b]

[a]Scuola di dottorato | Università Iuav di Venezia
 NTT Data Italia
[b]Università degli Studi della Repubblica di San Marino
*Corresponding author e-mail: isabella.loddo@gmail.com

Abstract: Studying the behaviour of disabled users can provide data for designing inclusive technologies for everyone. The focus of this paper is the field of inclusive design in conversational interaction.
Starting from the experience of a blind subject using screen readers in a professional scenario and setting them to a very high speech rate, we have investigated the evolution of speech based interactions from the perspective of the visually impaired and compared them to the current conversational interfaces.
The peculiar interactions set by the visually impaired motivate questions about inclusive design: How can we design conversational interactions for all? What can we learn from blind people using fast synthetic speech to browse digital products?
In this paper we have shown different strategies for increasing the usability of screen readers: speech synthesis and compression, natural vs artificial sound, multiple concurrent speech tracks. Our aim is to match them with an inclusive design approach in order to envision the future of conversational interaction.

Keywords: Conversational Interfaces, Screen readers, Accessibility, Synthetic Speech, Inclusive Design

1. Introduction

The vast majority of accessibility guidelines have been formulated at two separated levels: on one side, there are generic design principles; on the other, there are low level norms, referring to specific details, under specific platforms. Usually, they do not refer to specific scientific findings (Casali, 1995) and are referred to specific categories, among the users, in terms of adaptation, and in specific parts of the text (Bergman and Johnson, 1995). Stephanidis et al. introduced a process-oriented approach, coherent with user-centred design principles (Norman and Draper, 1986) and inclusive design, with unified norms, aiming to address the needs of disabled users, as a natural adaptation of rules of wider application. (Stephanidis et al., 1998)

The Design for all approach in HCI reflects an approach aiming to recognize and respect the widest range of abilities, necessities and preferences, in the development of digital products and environments. This way, it promotes a design perspective avoiding "special features" to respond to

special needs: instead, it aims to make possible a broad acceptability and adaptability of each proposed design solution.

Designing solutions that can be used by the disabled provides useful tools for designing technologies for everyone.

In this paper, we analyse the work performed with a visually impaired user, her behaviour in performing the analyses and using the test tools themselves.

2. User research: the case of an expert blind user

In order to evaluate the role of assistive technologies in the emotional experience of HCI users, we will take into account the work experience conducted with EB, a 38 year old blind girl, working as an IT professional. EB is not a common user, considering her high level skills and computer literacy.

EB is an expert user, open to experimentation.

Having the opportunity to work together with her for a fin-tech accessibility evaluation project, we managed to conduct an empirical, observation based study on her way of using assistive technologies, looking for the expression of affective qualities.

2.1 The user and the context

Working in an IT company environment, EB has a predefined activity schedule and needs to perform each task in order to proceed with the following one. Interacting by necessity, she is highly sequential and focused on her duties.

Having her chair in an open space office, she often seeks concentration by choosing a task specific room to work without distractions. Most of the time, she wears headphones, to reduce environmental noise and prevent her personal data from being unintentionally disclosed.

As a professional, she has chosen advanced assistive technologies, allowing her not to be bonded by her desk: her everyday tools include a laptop with a JAWS license and a portable Braille display (including a keyboard, that she mostly does not use), an iPhone (that she often uses to make calls or send text messages with Siri) and an Apple Watch, mostly used to keep track of time.

She is also using and testing programmable push buttons, but they still do not belong to her essential tool set.

2.2 Methodology

The methodology employed to deal with the blind user is not based on a quantitative analysis of the interaction quality in performing a specific task, but on the search for problematic conditions in her user experience and personal adaptations – intentional or not – to overcome an operational issue. Gathering qualitative data on alternative interaction patterns, we are looking for design hints that might have potential applications for all the users.

In the following experiment, EB has been using a screen-reader to perform a comparative web accessibility evaluation for a fin-tech company, simulating common personal finance operations.

She has verified the website compliance with the current W3C rules and tried to perform the same tasks with 2 competitors' websites, to compare their performance.

In addition to the necessary equipment, her hands and face have been captured with a double video shooting, in order to monitor her expressions while performing the evaluation and quantify the time spent in each task.

This test methodology has been set to combine the accessibility evaluation task itself with a further investigation on the user satisfaction and its emotional expression.

2.3 The experiment

The first phase of the experiment is aimed to evaluate which information and functions are accessible by the screen-reader and which are not. The lack of accessibility in certain parts of the website provokes emotional responses, influencing the first-look judgement of the service. Although every user has an emotional reaction towards his experiences in using websites, the ones who are not visually impaired can easily find hints or shortcuts to achieve their goals, whereas the blind can only rely on speech or Braille. They have no alternatives and can easily feel frustrated: often, the improvements in accessibility are slow and there has been almost no evolution in the mainstream assistive technologies in the last 20 years.

The following phase is based on the video shootings: one camera frames the subject, while the other records the screen. Combining the shootings in a simultaneous view allowed us to qualitatively evaluate the user satisfaction levels, in a highly informative recording. At a basic level, the video has the function to give an explanation on the visually impaired: how they explore a web page and how they interact with web contents. This is very valuable to the client (fin-tech company web developers), who may not have a direct experience of visually impaired people using their product.

Additionally, we were able to recognize the best and worst features of the website in terms of user satisfaction, being conveyed by facial expressions such as smiling, frowning, blushing, noticeable emotions such as joy, satisfaction, surprise, disappointment, anger, and behaviours such as relaxation, loss of interest or nervousness.

By reviewing and post-producing the clips, we have marked the most meaningful sections of the footage, with graphic emphasis and appropriate subtitles to synthetic speech.

The website accessibility compliance has been reported with comparative tables, based on the W3C guidelines. Each item has been graded, according to its compliance with the norms: the first column shows the rule id, the second shows its brief description, the third lists the solutions that could be undertaken in order to comply with the rule. In short, each item has been marked with a coloured cell, reporting whether the norm has been respected (green), not respected (red), or would not apply (grey). By reading the overall report, the tables offer a quick understanding of the accessibility and provides the client with clear information on the accessibility of their website. To get to consider desirable alternatives, EB tests other websites, providing the same functionality, with highly variable results. Through summarizing charts, the website accessibility is benchmarked both individually and in comparison with the competitors.

2.4 Empirical findings

As previously reported, the add of video shootings provides a considerable amount of additional data concerning accessibility and user satisfaction. Certain elements, in the website, result so problematic that the user reacts with visible anger and disappointment. *Captchas*, for instance, are often perceived like a wall, where the navigation stops. They often provide no clue to be read with a screen-reader, just like raster text elements or multimedia items without alternative captions.

Finding such items recalls a negative experience, affecting the overall satisfaction and the user's preference towards a website.
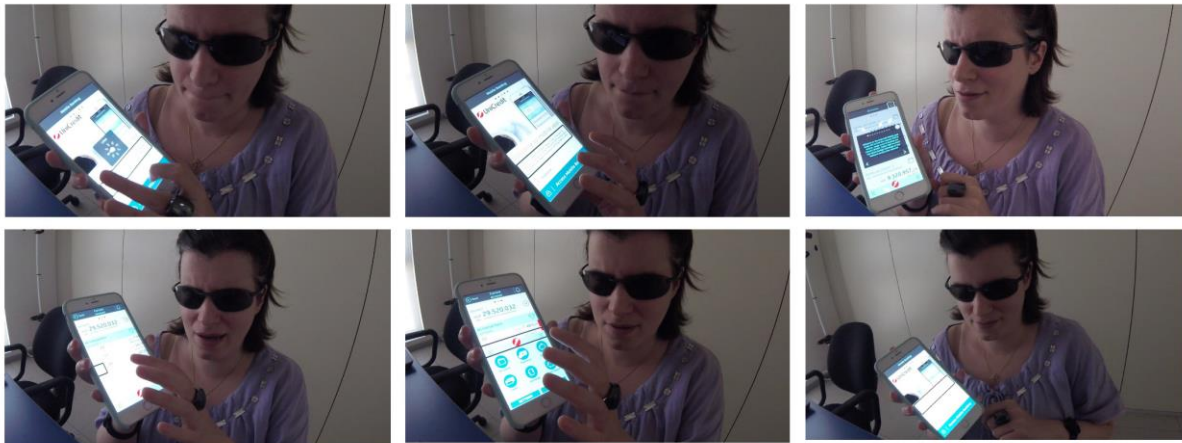

Figure 2. The subject showing emotional changes while performing the accessibility tests

Even when the navigation is successful in terms of mere W3C compliance, the user experience can be quite negative. Some tasks are feasible, but take a long time when compared to the graphical equivalent: this wait can be frustrating and can be perceived as a disabling factor. No one deserves a low efficiency interaction, even when "it works".
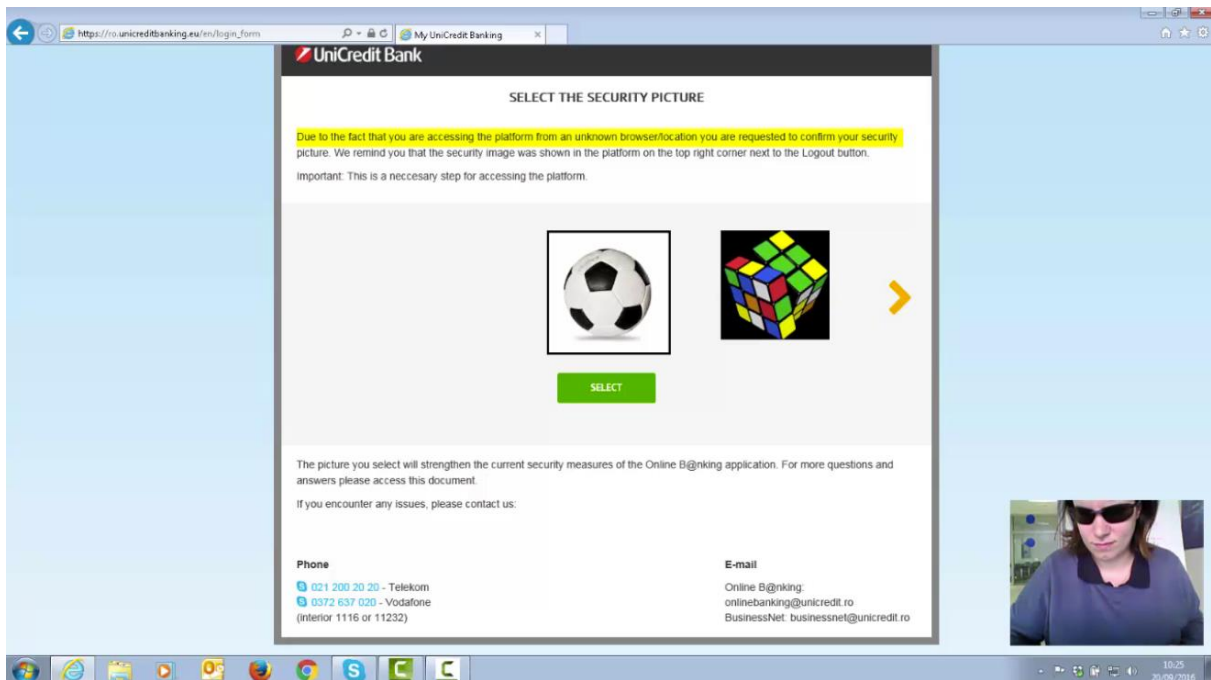

Figure 3. The subject being asked to choose between visual icons

Time is a crucial factor for the user's satisfaction also when the screen-reader itself is concerned: the user EB takes the speech ratio at a very high level, so that it feels unintelligible for those unfamiliar with such technology. Even at this speed, we often find the user to be frustrated with the waits, lags and overall length of the interaction.

As far as the interaction process is concerned, we see that EB is often looking for shortcuts, alternative paths to perform a task as fast as with the GUI. She is often speeding up navigation and searching for quicker ways to interact, closing the gap that she perceives.

2.5 The affective gap

The charted output of the analysis is capable of showing W3C compliance, but it lacks the ability of conveying the emotional impact of accessibility pitfalls: the rules are set from the developer's perspective, not from the user experience's point of view. Some rules, when broken, cause a worse dissatisfaction. In the case of the visually impaired, we find this gap to be even more uncomfortable. Disabled users have a greater dependence on technology, so unsuccessful technologies provoke more serious limitations and a worse emotional effect.

For the sake of improving user satisfaction, in a design for all approach, the emotional responses need to be evaluated on par with the accessibility guidelines.

3. Accessibility of speech-based interaction

First referred to elderly and disabled users, the issue of accessibility, with the higher expansion of the world wide web in the second half of the '90s, came to a broader meaning.

Under the term accessibility, researchers began to consider the need of any user, no matter what her preferences and abilities are, to have access to information and, in general, to access any function, in any context where it has been expected. To meet accessibility criteria, the functions have to be effective, efficient and satisfactory.

The differences among users have to be considered in terms of distinctive properties, individual cultural, in terms of nature and purpose of the tasks, of technological platform and devices used to access information.

Visually impaired people often use screen-readers, as do some people with low vision. A screen-reader scans websites for text content and converts it into spoken words, using a synthesized voice. They are the main alternative to Braille displays, that render text into a Braille tangible string. Actually, Braille displays are much less diffused, due to their higher price, and many users, especially in low income countries, use only screen-readers. (McCarthy et al., 2013)

Compared to the average user, the visually impaired tend to run screen-readers at an almost double speed. Being faster allows them to be more independent.

3.1 User adoption and preferences

Most visually impaired users tend to choose their first screen-reader according to the quality of the voice. They become more concerned with responsiveness and app support after the first experience, when they get to use it more frequently and in-depth (McCarthy et al., 2013).

In the earliest phase, most users quickly adapt to the chosen software. Afterwards, they largely become unwiling to change, even when they foresee potential improvements.
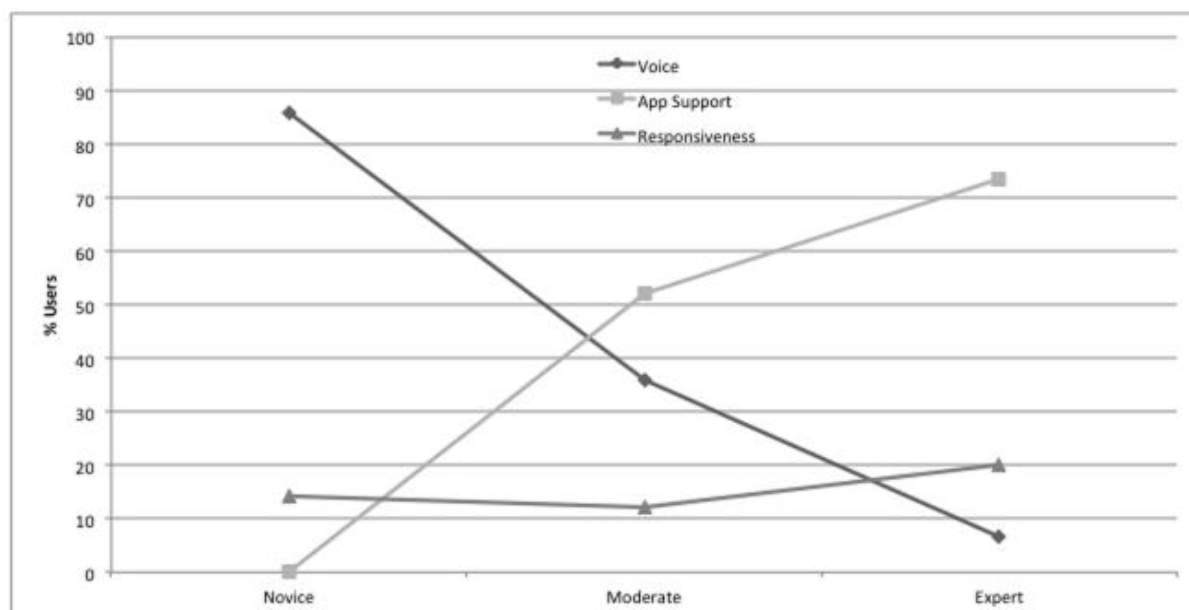
Figure 1. Percent of users choosing either TTS quality, responsiveness, or application support when asked "What is of primary importance in deciding your preference for one screen reader software over another?" (n = 48)

Although text-to-speech technologies are highly diffused among the visually impaired, little work has been done to analyse and compare the major tools and the approaches they are built upon (Stent et al., 2011). To date, the most popular solutions on the market have been benchmarked only partially, showing no clear performance differences.

In the case of early-stage blindness, the users are willing to adapt to their screen-readers more than they expect the software to be adjusted on their preferences (McCarthy et al., 2013). To achieve a faster interaction, they usually prefer information intelligibility over speech naturalness, but they mainly choose the software that is more familiar to their listening. (Stent et al.)

Actually, this preference is not relative to the visually impaired only. Many motor impaired prefer intelligible speech as well. An eminent case is the one of Stephen Hawking, using a custom speech synthesiser, that has no voice variation and puts influence in speech just by using punctuation.

In general, some authors state that disabled users are mainly focused on their tasks and do not pay attention to secondary elements. (Damper, 1984)

The most popular software, by the numbers, is JAWS, on Windows machines. In India, over 90% of the users prefer JAWS, even when they have a NVDA license, despite the 57% uses a pirate license. Recently, the expansion of conversational interfaces among consumer applications has brought to a wide diffusion of new tools, both proprietary and free / open-source.

3.2 Optimising listening speeds

The average user, at medium speech rates, finds the most natural voices to be the most intelligible. At very high speech rates, both the average and the visually impaired find synthetic speech to be more intelligible, even though it feels less natural.

Visually impaired users tend to change their preference towards synthetic speech, as it stands as a tool, rather than a human speaker. They also tend to prefer higher speech rates, with a general increase of their preferred rate as they become more experienced. Arguably, it is not just a change in preferences, it's the development of a new ability.

At very fast speech rates, speech does not resemble human dialogue anymore. Rather, it can be seen as a technique, depicting an auditory environment where contents are briefly shown by their first sounds and browsed at a glance, resembling a visual-like experience.

Though it is vastly known that the visually impaired use screen-readers at a much higher than normal rate, the first benchmarks have been performed after 2000. In 2003, Asakawa et al. have been rating the highest intelligible speech rate (over 80% of word comprehension) at about 500 wpm, about 1.6x the non-impaired maximum rate.

While novice users show an immediate increase of fast speech comprehension, expert users reach speech rates that are 2.5-2.8x faster than the reference rate and set their optimal rate much higher than the average user, both in subjective and objective evaluations.

Subsequent tests have been involving larger samples, other languages, to achieve a more granular comprehension of the relationship between visual impairment and text-to-speech systems.

Among expert users, the only measured difference is given by age: elderly users tend to decrease their peak intelligible speech rate, no matter if they are suffering from hearing loss.

Fast speech rates can be achieved in multiple ways: basically, the software can either increase the word count, without compressing each sound, or perform a time compression, linear or weighted.

Time compression algorithms can either manipulate speech by variating parameters for each phoneme ("formant" approach) or group speech units ("concatenative" approach).

Concatenative tools provide a more natural sound, closer to human speech, which is preferred by novice and non-impaired users. Formant tools may sound more mechanic, but are better in preserving the intelligibility of the phonemes and are generally preferred by expert users.

Every approach can benefit from further enhancements, such as tone tuning, stereo imaging, prosody and word emphasis modulation, to mark the role of each phoneme in a sentence.

There is almost no evidence on which of these enhancements can benefit intelligibility: most of them are left to personal preferences.

Since the first studies conducted by Asakawa et al., the ability of the visually impaired to be significantly faster in processing speech has been pointed out as a crucial design opportunity to develop inclusive HCI systems. Most of the business screen-readers, at the time, peaked far below the actual users' potential, determining frustration, stress and loss of productivity. By changing the speech rates according to the user's ability, experience and mood, the overall experience can be vastly improved.

A decade after, the improvements in machine learning and body sensing let us foresee speech-based interactions where the machine speech is dynamically adjusted to the user and the context.

Whereas the studies conducted to date state the benefits of a faster single speech, other enhancements still need to be tested.

3.3 The cocktail party effect

In the last few years, the use of concurrent speech is being developed as a new strategy for increasing speech rate in screen-readers, beyond the limits of single voice time compression.

The concept is based upon the so called *cocktail party effect* (Cherry, 1953), describing the human ability to understand relevant information from background conversations. The cocktail party effect

involves an incremented information bandwidth. The bandwidth saturates over 3 concurrent conversations, meaning that the average person can still pick information from any source, but no more than 4 concurrently.

We can assume that people with a high hearing ability can especially benefit from this effect and when hearing is the primary sensory channel. Most of the visually impaired fit well in this case, hence the interest in testing concurrent speech as an alternative approach for relevant scanning, the listening process involved, for example, in searching information on a web page. (Guerreiro et al., 2015)

In relevant scanning, the user does not need to understand each word in a specific sequence: rather, he aims to find the desired information, when recalled.

In 2015, Guerreiro and Gonçalves have measured the speed increase, quantifying the optimal rate at 1.75X, versus the 1.6X achievable with a time compression of a single speech. Besides the mere speed proportion, the speech intelligibility rate is more stable among the users, envisioning a more balanced and accessible approach. Though this technique has already proven to be promising, it is yet to be implemented in consumer applications and many aspects still need to be clarified: for instance, when would it be better to use 3 voices instead of 2? Are there beneficial sound effects, such as voice spatialization, channel separation or vocal tone arrangement?

4. Inclusion in speech-based interaction

Though the applications of synthetic speech are rapidly expanding in HCI (Drahota et al., 2008, Robinson and El Kaliouby, 2009) their use is held back by the lack of "human touch" and emotional interaction with users. (Mitchell and Xu, 2015)

If on one hand the idea of natural and human-likeness is broadly associated to an added value for interactions, several examples demonstrate that a human-like similarity does not imply a higher level of user satisfaction.

4.1 Human-likeness and affective interaction

The current speech-based interaction metrics are meant to increase the efficiency, either in terms of performance, or in terms of user satisfaction.

Though this criteria corresponds to a more user-centred approach and generally, the user satisfaction is better when the performance improves, it does not identify an absolute term of comparison and does not define the properties of a satisfying and "human-like" dialogue.

What are the distinctive properties of a human-like interaction with a computational system? Can we integrate a human-like design with different aesthetics, while keeping the intended interaction obvious and intuitive? What are the actual benefits of a human-like dialogue? Is it compatible with a design for all approach?

According to Edlund et al., speech-based interactions are largely perceived as metaphoric systems, where speech takes a precise purpose in an implicit mental model. (Edlund et al.)

Talking to computers is not common enough to be familiar per se. Rather, it works when placed in a mental model. Interface models, historically, are divided into tool-like and anthropomorphic models. (Qvarfordt, 2004)

The latter can be divided again in a human-human and a human-artificial model. In the second model, there is no need to enhance the naturalness of speech, because the speech is linked to a

machine interface. It performs a task, just like a displayed text or a keyboard input, providing an alternative to existing tools.

In the human-like model, the computer is proposed as an interlocutor, with conversational abilities. Though the user is aware that he is dealing with an artificial system, he engages a dialogue as if it was a human: not an alternative to computer peripherals, but a person to talk to.

The distinction between the two models, obviously, is not always explicit. In many current cases, the user is supposed to engage a human-like conversation, but later, he switches to a machine-like model, when the interface fails in providing a believable human behaviour.

Far from being applicable to any kind of interaction, the use of a human-like metaphor is useful when searching for data, when performing booking, ordering or payment tasks, in user assistance and troubleshooting, or in text input, through dictation.

In positive scenarios, the choice of talking to the system can be quicker, more secure (through step-by-step validation) and pleasurable. It can be also employed for communication tasks involving multiple users, mediating an interpersonal dialogue or simultaneously talking to several concurrent users.

The human-human and machine-human models are also distinguished by a different grade of dependence: while the human-like is expected to stand at the same level of the user, the machine listens, executes and obeys.

The ubiquitous use of personal assistants, though being aimed to define a virtual interlocutor, will necessarily be bonded to wait, obedience and subordination behaviours. The assistant would, therefore, be placed in an in-between metaphor, between the human and the machine: an android-like interface.

These examples convey the complexity of identifying a human-like metaphor that can be both persistent and believable.

5. Conclusions and future work

An inclusive approach to user experience design requires the participation of visually impaired users to test and develop each project. By observing a blind subject using a screen-reader, we noticed that her use of assistive technologies is very different from the regular. Screen-readers are especially set at high speech-rates among blind users, to facilitate relevant scanning tasks. Faster rates help them to achieve a quicker interaction with digital products so it has been investigated since 2003 to improve listening performance. Several techniques have been tested, among them time compression, "formant" synthesis and concurrent speech, taking advantage of the so called "cocktail party effect". These techniques lead to artificial sounding voice, aiming to a tool-like mental model of speech-based interaction. The strategies targeted to visually impaired users could benefit everyone, while the interaction quality of intelligent and natural conversation should be extended to the visually impaired. We can propose concurrent conversations tracks to provide a faster interaction for everyone. In order to develop inclusive conversational interactions we propose to extend the inclusion of the visually impaired in the design process and to apply a design for all methodology as a new paradigm for conversational user interfaces.

References

Asakawa, C., Takagi, H., Ino, S., and Ifukube, T. (2003). Maximum listening speeds for the blind. *Proceedings of the International Community for Auditory Display* (ICAD), (pp. 276–279).

Bänziger, T., Hosoya, G., & Scherer, K. R. (2015). Path Models of Vocal Emotion Communication. *PLoS ONE*, *10*(9), e0136675. http://doi.org/10.1371/journal.pone.0136675

Bergman, E., & Johnson, E. (1995). Towards Accessible Human-Computer Interaction. In Nielsen, J. (Ed.), *Advances in Human-Computer Interaction*, 5. New Jersey: Ablex Publishing Corporation.

Casali, S. P. (1995). A physical skills-based strategy for choosing an appropriate interface method. In Edwards, A. D. N. (Ed.), *Extra-ordinary Human Computer Interaction: Interfaces for people with disabilities*, (pp. 315-342). Cambridge University Press.

Cherry, E. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*.

Damper, R. (1984). Voice-input aids for the physically handicapped. Int. J. Man–Machine Stud., 21, 541-553.

Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication*, 50, (pp. 630-645).

Fernandez, R., Schlangen, D., & Lucht, T. (2007). Push-to-talk ain't always bad! comparing different interactivity settings in task-oriened dialogue. *Proceedings of DECALOG, the 11th International Workshop on the Semantics and Pragmatic of Dialogue (SemDial07)*. Trento: Italy

Funakoshi, K., Nakano, M., Kobayashi, K., Komatsu, T., & Yamada, S. (2010). Non-humanlike spoken dialogue: a design perspective. *Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 176–184. The University of Tokyo.

Guerreiro, J. & Gonçalves, D. (2015). Faster Text-to-Speeches: Enhancing Blind People's Information Scanning with Faster Concurrent Speech. *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '15)*, 3-11. New York: ACM.

Janse, E. (2004). Word perception in fast speech: artificially time compressed vs. naturally produced fast speech. *Speech Communication*, 42, (pp. 155–173).

Komatsu, T., & Yamada, S. (2010). Effects of adaptation gap on user's variation of impressions of artificial agents. *Proceedings WMSCI*.

Krauss R., & Bricker, P. (1967). Effects of transmission delay and access delay on the efficiency of verbal communication. *Journal of the aCoustic Society of America*, 41, 286–292.

Lee E.-J. (2010). The more humanlike, the better? How speech type and users' cognitive style affect social responses to computers. Comput. Hum. Behav. 26, 665–672.

McCarthy, T., Pal, J., & Cutrell, E. (2013). The "voice" has it: screen reader adoption and switching behavior among vision impaired persons in India. *Assistive Technology: The Official Journal of RESNA*, *25*(4), 222–229.

McTear, M., Callejas, Z., Griol, D. (2016). Emotion, Affect, and Personality. In McTear, M., Callejas, Z., Griol, D. (Eds.), *The Conversational Interface. Talking to Smart Devices* (pp. 309-328). Basel: Springer Verlag.

McTear, M., Callejas, Z., Griol, D. (2016). Affective Conversational Interfaces. In McTear, M., Callejas, Z., Griol, D. (Eds.), *The Conversational Interface. Talking to Smart Devices* (pp. 329-355). Basel: Springer Verlag.

Mitchell, R. L. C., & Xu, Y. (2015). What is the Value of Embedding Artificial Emotional Prosody in Human–Computer Interactions? Implications for Theory and Design in Psychological Science. *Frontiers in Psychology*, *6*, 1750. http://doi.org/10.3389/fpsyg.2015.01750

Moosand, A., & Trouvain, J. (2007). Comprehension of ultra-fast speech blind vs. "normally hearing" persons. *Proceedings of the 16th International Congress of Phonetic Sciences*.

Mori, M. (1970). The Uncanny Valley. Energy, 7(4), 33-35.

Norman, A., D., & Draper, W., S. (Eds.), (1986). User-centred system design: New perspectives in Human Computer Interaction. Hillsdale, NJ: LEA

Papadopoulos, K., Katemidou, E., Koutsoklenis, A., & Mouratidou, E. (2010). Differences among sighted individuals and individuals with visual impairments in word intelligibility. *Augmentative and Alternative Communication*, 26(4), 278–288.

Qvarfordt, P. (2004). Eyes on multimodal interaction. Doctoral dissertation, Linköping University.

Stephanidis, C., Akoumianakis, D., Sfyrakis, M., & Paramythis, A. (1998). Universal accessibility in HCI: Process-oriented design guidelines and tool requirements. *Proceedings of the 4th ERCIM workshop on user interfaces for all*, 19-21.

Stephanidis, C., Akoumianakis, D., Sfyrakis, M., & Paramythis, A. (1998, October). Universal accessibility in HCI: Process-oriented design guidelines and tool requirements. In *Proceedings of the 4th ERCIM Workshop on User Interfaces for all, Stockholm, Sweden* (pp. 19-21).

Story, M. F. (1998). Maximising Usability: The Principles of Universal Design. *Assistive Technology*, 10 (1), 4-12.

About the Authors:

Isabella Loddo's work concerns user experience and inclusion. She is Interaction Designer at NTT DATA and PhD Candidate at Iuav University of Venice, with a thesis project addressing the human relationship with technology, making more people feel comfortable with digital things.

Dario Martini Ph.D. is a product designer and design researcher. His interests include design for emergencies, aero-space design, robotics, bio-medical research and prosthetics. Since 2015, he is working as teaching assistant at the University of San Marino.