INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Text-to-speech for individuals with vision loss: a user study

Monika Podsiadło, Shweta Chahar

Google
{mpodsiadlo, shwetachahar}@google.com

Abstract

Individuals with vision loss use text-to-speech (TTS) for most of their interaction with devices, and rely on the quality of synthetic voices to a much larger extent than any other user group. A significant amount of local synthesis requests for Google TTS comes from TalkBack, the Android screenreader, making it our top client and making the visually-impaired users the heaviest consumers of the technology. Despite this, very little attention has been devoted to optimizing TTS voices for this user group and the feedback on TTS voices from the blind has been traditionally less-favourable. We present the findings from a TTS user experience study conducted by Google with visually-impaired screen reader users. The study comprised 14 focus groups and evaluated a total of 95 candidate voices with 90 participants across 3 countries. The study uncovered the distinctitve usage patterns of this user group, which point to different TTS requirements and voice preferences from those of sighted users.

Index Terms: speech synthesis, human-computer interaction, speech perception

1. Introduction

Text-to-speech (TTS) to a casual user is an added modality when consuming digital content. However, to certain user groups text-to-speech is a crucial technology that enables access and provides functionality that is necessary rather than additional. For example, TTS can provide replacement voices for individuals with vocal disabilities [1] and voice interfaces for individuals with motor disabilities. Text-to-speech is also of critical importance to individuals with vision loss. Most visually impaired users consume digital content using screen readers, for example JAWS, Android's TalkBack or iOS's VoiceOver. Text-to-speech is a vital component of screen reading software that vocalizes the content of the screen to the user. Unlike sighted users who rely on other modalities, visually-impaired users rely on TTS for all or most of their interactions with devices, depending on the degree of their sight loss. In addition, a number of studies into voice perception [2], [3] suggest superior auditory perception in blind individuals as opposed to sighted people, making them more sensitive to the quality of TTS voices.

Given how important high-quality TTS is to individuals with vision loss, it is surprising how little prior research there is into the needs of this particular user group. A small pilot study investigated the effects of speaking rate on intelligibility of synthesised speech among blind individuals [4]. Concurrent TTS channels were also explored as an enhancement to browsing strategies of individuals with vision loss [5]. Such a solution takes advantage of the increased intelligibility of simultaneous speech channels among the blind and could potentially increase user productivity when using a screen reader. To our knowledge, however, there has been no user studies conducted to address specifically text-to-speech preferences of visually-impaired individuals.

On the other hand, research into TTS evaluation has a well-established tradition. The annual Blizzard Challenge [6] provides a comprehensive platform for in-depth testing of synthetic voices built on common datasets. There are well-established methods for assessing TTS intelligibility [7], including at increased speech rates [8]. There is also a body of research into the analysis of various TTS evaluation methods [9] as well as into how voice characteristics affect the acceptability of TTS among general population users [10].

With our research we combined the two traditions and structured a sizeable study employing standard TTS listening tests, but recruting solely participants who were daily screen reader users and had some form of vision loss. We also organized a series of focus groups in which we discussed text-to-speech in a more conversational and interactive setting with a small number of participants. This paper presents our findings about the needs, preferences and opinions regarding text-to-speech technology among individuals with visual impairments.

The paper is organized as follows. Section 2 details the design of the user study. Section 3 presents the findings from the focus group discussions. Section 4 presents the speech evaluation experiments, Section 5 discusses the data we collected, and finally, Section 6 concludes the paper.

2. The study

In late 2014 and 2015 Google organized a series of UX studies with visually-impaired users of text-to-speech. The objective of the studies was to establish whether heavy users of text-to-speech who rely on screen readers have different preferences and needs to other users of TTS, and what these preferences are. We initially conducted a study in England with 50 native speakers of British English. We later on conducted follow-on studies in the US and in Spain, recruiting 20 further participants in each country.

2.1. Participants

A total of 90 persons participated in the study. All participants were native speakers of the language used in experiments. 71% of the participants had complete vision loss, with the remaining participants retaining some vision. The onset of blindness varied, with 41% of participants congenitally blind. All participants were daily users of text-to-speech. The participants' age ranged from 18 to 81, with 55% of participants aged between 18 and 39. We had equal gender representation. The participants were compensated for their participation in the study.

http://dx.doi.org/10.21437/Interspeech.2016-1376

2.2. Evaluation

In designing the study our aim was twofold: to test a range of speech samples from candidate voices, as well as to collect user feedback on a variety of general issues surrounding TTS and voice interfaces. We therefore divided the study into two sections: a series of focus groups and a series of individual listening tests. In total, we conducted 14 focus groups and 90 individual test sessions in 3 countries.

2.3. Stimuli

To produce the speech samples used as stimuli in the study, we recorded professional voiceover artists in controlled studio conditions. We recorded 50 sentences from each artist. The sentences represented a variety of domains including, for example, newspaper text, weather reports, common device navigation strings and screen reader navigation strings. The study in England evaluated 40 voice samples, and in the US and Spain, 25 samples each. We also included samples from existing Google TTS databases from each locale: two British English voices, one American English voice and one Spanish voice. In crafting the stimuli, we selected voices that could be equally distributed along the dimensions of gender, voice age and reading style, as shown in Table 1.

Table 1: *Voice features matrix.*

Age	Gender	Reading Style
25-35 (Y)	Male (M)	Conversational (C)
40+ (O)	Female (F)	Narrative (N)

We defined the narrative reading style as more formal and less expressive, similar to that of a newsreader. We defined the conversational reading style as more informal, familiar and expressive, similar to an animated conversation between friends.

We will refer to particular voices by their indexed feature combination, for example FYC will be used to refer to a voice that is female, younger, and conversational in reading style.

3. Focus groups

3.1. Methodology

The focus groups were used to collect feedback from the participants in a more open-ended and conversational fashion. They were conducted in groups of six to eight participants, a moderator and an observer. Each focus group followed the same protocol. In the first half of the focus group, we played a range of TTS samples and asked participants to discuss the voice. We asked about perceived naturalness and pleasantness of the voices, their suitability for conducting particular tasks (such as reading a book, writing an email or filling a work spreadsheet), and which features stand out as particularly good or bad. In the second half of the focus groups, we discussed specific features like text normalization, acoustic characteristics of voices, latency, voice interface design, and what guides the voice choice.

Each focus group covered the questions included in the protocol. We only report points that were discussed by all groups.

3.2. Findings

Analyzing the recordings from focus groups, it became clear that the participants' feedback on TTS should be viewed in the context of unique usage patterns of individuals with visual impairments. We thus grouped our findings into two subsections: analysis of usage patterns and TTS feature requests.

3.2.1. Analysis of TTS usage patterns

The three factors that the focus group discussions pointed to as affecting TTS usage are: 1. Purpose of the action, 2. Type of environment, 3. Length of experience with TTS.

The purpose of action can be divided into two types: leisure-oriented mode and the factual mode. The leisure-oriented mode includes book or news reading and any kind of interaction conducted as a leisure activity. Here, the discussion strongly suggested that the most important factor is the intelligibility of the voice. This includes prosody that accurately expresses text events and thus facilitates intelligibility, increases comprehension and minimizes cognitive load. Participants also stressed that particular acoustic quality of the voice or its gender is unimportant, as long as the voice sounds pleasant, especially when listened to over longer periods of time. However, text-to-speech is still not seen as good enough to compete with audiobooks: all participants stated that they only use TTS for book reading if a recorded audiobook is not available.

The factual mode includes device navigation, work tasks, messaging or any activity where the focus is on accomplishing a task. Here, the participants pointed to efficiency as the most important factor in selecting text-to-speech. It is thus imperative that the voice is intelligble at high speech rates and the latency is minimized. The acoustic quality of the voice is seen as insignificant, and unnatural, robotic-sounding voices are acceptable as long as the efficiency can be maintained.

The second factor affecting TTS preferences is the length of TTS experience. Users new to the technology take time to learn and understand the voice. They prefer voices that sound more natural and human-like, and they listen to them at speech rates closer to natural. Experienced TTS users, on the other hand, get used to new voices quickly and prioritize its functionality over naturalness.

The final factor affecting TTS preferences is the environment. Here users differentiated between public and private. Public environment includes streets, offices, restaurants, public transport – any environment where the user is surrounded by others and privacy is a concern. In such an environment some users prefer to use headphones for privacy while others find this unsafe and prefer to use the phone speaker. When using the phone speaker, users keep the device close to their ear and adjust the volume to maximize intelligibility in the noisy environment while maintaining privacy. Private environment includes home or car, where the user does not have privacy concerns. Most users listen to TTS through the phone speaker and adjust volume to maximize intelligibility without any additional constraints.

3.2.2. TTS feature requests

In this section we summarize feedback regarding text-to-speech voice features that contribute towards a better user experience.

Intelligibility was strongly suggested as the most important feature of TTS voices. Clarity of pronunciation, consistency and intelligibility, also at high speech rates, were voted as infinitely more important than likeability, naturalness or conversational quality of the voice.

Latency followed as the second most important consideration. Participants pointed out how, especially if you rely on TTS feedback when typing, the milliseconds lost to latency very quickly add up to hours and days of lost productivity over a

working year. Similarly, even small improvements make a difference and are noticed by the user. We tried to quantify acceptable levels of latency, however the feedback we received was that anything other than instant will be noticed by the user and will affect their experience negatively.

While effiency and general utility of the voice dominated the discussions, the participants expressed strong opinions about the acoustic characteristics of voices as well. Participants were unanimous in stating that the most important quality in this respect is how the voice makes you feel when you are interacting with it, rather than its specific characteristics like gender or perceived age. A good TTS voice should sound helpful and relaxing, not increasing the user's stress levels and not adding to the frustrations of their daily lives. Voices that sound harsh, commanding or opinionated were particularly disliked.

Naturalness was seen as a less important feature than intelligibility but still very desirable as long as it does not increase latency or decrease intelligibility. The context in which the participants discussed naturalness was always that of intelligibility: the more natural the voice, the less cognitive effort is required to listen to it.

Focus groups also discussed expressiveness of TTS at length. Our participants recognized that emotional expression can be useful for some domains, such as book reading, however they preferred more neutral and less expressive voices. On the one hand, too much emotional expression is an unnecessary component when interacting with TTS in the factual mode. On the other, in the leisure mode users also prefer neutral emotional expression to allow them to project their own emotions into the reading material. This is akin to silent reading. Another point raised here was that any attempts at prosody or emotional expression that are less than fully successful are seen as disruptive and worse than consistently neutral expression, even if it is less appropriate for emotionally-charged contexts.

4. Speech sample evaluation

The second part of the study consisted in speech evaluation tests performed by each participant individually. In this section, the study followed closely standard listening test protocols for evaluating speech synthesis, for example those used in the Blizzard Challenge ([11]). Each participant spent 60 minutes completing a Mean Opinion Score (MOS) and an AB comparison tests of voices, as well as a verbalization preference test and an intelligebility test at different speaking rates. In this paper, we report the results from the first two tests: MOS and AB comparison.

4.1. Methodology

All listening tests were conducted in a quiet lab and the speech samples were played through laptop speakers. We decided not to use headphones to mimic the most common usage scenario more closely. Each listening test evaluated the same set of voices.

First, we conducted the MOS evaluation test. The participants were played a voice sample of three sentences. They were then asked how likeable they find the voice on a scale from 1 to 5, with half-points allowed. This was then repeated for between 9 and 14 unique voices per participant. Each participant therefore judged a randomly selected subset of the total evaluation set. Each voice in the evaluation set received ratings from 10 users.

In producing the final voice ranking, we tried to control for two types of biases: the differences between the evaluation subsets presented to each user (for example, if a particular set happened to be biased towards high-quality voices, or voices of particular gender) and the differences in individual user ratings (for example, if some users were more critical or more generous than others). We therefore performed two types of normalization. First, we computed z-scores for each user rating as

$$z = \frac{MOS - \mu}{\sigma} \tag{1}$$

where μ is the mean of all the ratings given by the user and σ is their standard deviation.

Second, for each of the voices under evaluation, we computed the population mean of all the ratings given by different users to this voice and their standard deviation. We then used the z-scores computed in (1) and calculate the absolute grading as

$$A_g = \mu_p + (z \cdot \sigma_p) \tag{2}$$

where μ_p and σ_p are the population mean and standard deviation for that voice, respectively.

We use the absolute grading in (2) as the normalized score for each voice in each user subset. We then computed the final MOS for each voice as the mean of the absolute gradings received from different users. The mean absolute grading is used to generate MOS ranking of all the voices.

Second, we conducted the AB voice comparison test. The participants were played a pair of natural speech sentences spoken by two voices one after the other. The sentence spoken by each voice in the pair was the same. The participants were then asked which voice of the pair they preferred as their screen reader voice. If requested, a participant could hear the pair of voices one more time before deciding. The test was repeated for 51 voice pairs for participants in the study in England, and 49 for the participants in the study in Spain and the US. Our goal was to produce a stack ranking of voices representing their likeability.

4.2. Results

We analyzed the correlation of the results of both tasks and found that the top-scoring voices from the MOS listening test were also placed near the top of the ranking in the AB test.

Figure 1 presentes how the different voice types from Table 1 were distributed among the 20 top voices in the evaluation.

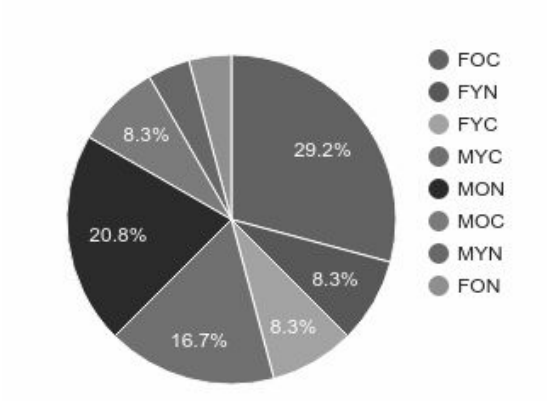

Figure 1: *Distribution of voice types in the top 20.*

Figure 2 presentes how the different voice types from Table 1 were distributed among the 20 worst voices.

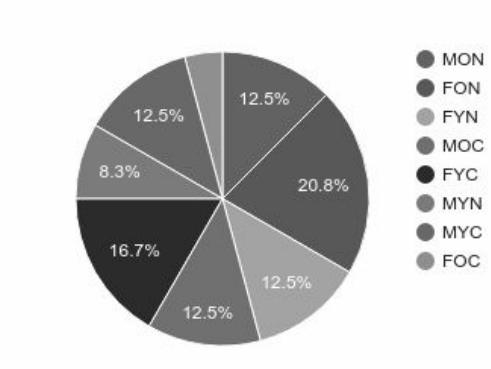

Figure 2: *Distribution of voice types in the bottom 20.*

Overall, the FOC and MON voice types were preferred and together comprised 60% of the top 20 set. Voices that sound more mature were overwhelmingly preferred by the participants–only 2 youger voices made it to the top 20. We also noted that the more formal narrative style of speech was preferred over the more informal, conversational style. MOC, FYC and FON voices were the least preferred. There was no signifficant difference between gender preferences among the voices, though among voices with conversational speaking style, female voices were preferred over male ones. There was no significant correlation between the participants' gender and their preferred voice gender. We also have not observed any other correlation between participants' demographic information and their voice preferences.

5. Discussion

The findings we presented from the focus group discussions suggest there is a variety of distinct use cases that a good TTS voice needs to cover to be acceptable. It is clear that it is extremely hard to satisfy all the user requirements with a single voice. On the one hand, extremely natural and human-sounding voices are necessary for an acceptable user experience in the leisure mode. Natural prosody without artificial artifacts also facilitates understanding by minimizing cognitive load. Such voices are also a good non-threatening introduction to text-to-speech for users who only recently acuired the visual impairement and started using the technology. On the other hand, more experienced users are accustomed to synthetic voices and are less sensitive to artificial artifacts in speech. They find consistently robotic speech to be preferable to speech that attempts to sound human-like but sometimes fails. More importantly, voices that can be highly intelligable at very fast speaking rates are vital for maintaining user productivity in professional contexts and for factual tasks. Right now, this typically means less expressive and more robotic-sounding voices.

The listening experiments in our study suggest what types of voices specifically stand a chance of satisfying these needs. We've observed a clear pattern of voice characteristics shared by the voices that were preferred the most by the participants. The top-scoring voices can all be described as calm, very consistent and with clear pronunciation. There is little emotional expression present and the voices sounds friendly but detached. The prosody is always natural, intonation not forced to give a particular effect. The reading style is fluent and the voices are pleasant, soft and attractive but in a way that does not command attention.

On the other hand, there were clear commonalities between the voices in the bottom of the ranking as well. Here we have voices that had non-standard accents (e.g. Irish in the English evaluation set) and affected prosody. They are all very expressive, sound highly dramatized and not neutral. The reading style in each case is emotionally charged and the voice's attitude towards the message is clearly audible. The voices stand out as distinctive and capture listener's attention easily. They can all be described as character voices emulating a radio DJ or a child's storyteller. Our participants commented that such voices would not even be desirable for domains such as book reading, where expressiveness can be seen as an asset. These preferences are also in line with the focus groups discussions.

Interestingly, while older voices are preferred for both genders, female voices scored higher if their reading style was informal/conversational. Conversely, male voices had better preference rates if their reading style was formal/narrative. Among the least favoured voices the reverse was true–female voices were penalized if their reading style was narrative, while male voices were penalized if their reading style was conversational. It would appear that a more authoritative tone is still more acceptable in male voices and a more approachable and familiar tone is preferred in female voices.

Crucially, we did not observe any signifficant differences between participant groups in the three countries. The feedback we collected in England was consistent with that collected in the US and Spain: users wanted minimal latency and high intelligibility at as fast speaking rates as possible. In addition, for a truly optimal experience in leisure more, highly natural voices that can compete with recorded content were desired.

6. Conclusions

In this paper we reported the findings from a user study investigating text-to-speech requirements of visually impaired users of screen readers. We presented an analysis of scenarios under which text-to-speech voices are used by this user group and how they translate into particular feature requests. We have shown that usage patterns of individuals with vision loss are distinct from those of other user groups, and result in markedly different requirements regarding synthetic voices. We have also presented the results of perceptual experiments showing how different voice qualities affect preference and likeability. We have concluded that lower-pitched, mature-sounding voices of either gender with neutral emotional expression are strongly preferred by visually impaired users. They are seen as minimizing listening fatique and increasing overall intelligibility. In addition, different voices are required for different applications depending on which voice features are prioritized in a given usage scenario.

7. References

[1] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.

[2] F. Gougoux, P. Belin, P. Voss, F. Lepore, M. Lassonde, and R. Zatorre, "Voice perception in blind persons: a functional magnetic resonance imaging study," *Neuropsychologia*, vol. 47, no. 13, pp. 2967–2974, 2009.

[3] K. Hugdahl, M. Ek, F. Takio, T. Rintee, J. Tuomainen, C. Haarala, and H. Hamalainen, "Blind individuals show enhanced perceptual and attentional sensitivity for identification of speech sounds," *Cognitive Brain Research*, vol. 19, no. 1, pp. 28–32, 2004.

[4] A. Stent, A. Syrdal, and T. Mishra, "On the intelligibility of fast synthesized speech for individuals with early-onset blindness," in *Proceedings of ASSETS'11*, 2011, pp. 211–218.

[5] J. Guerreiro and D. Goncalves, "Text-to-speeches: evaluating the perception of concurrent speech by blind people," in *Proceedings of ASSETS'14*, 2014, pp. 169–176.

[6] S. King and V. Karaiskos, "The Blizzard Challenge 2013," in *Blizzard Challenge Workshop, Barcelona, Spain*, 2013.

[7] C. Benoit, M. Grice, and V. Hazan, "The SUS: test a method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," *Speech Communication*, vol. 18, no. 4, pp. 381–392, 1996.

[8] E. Janse, S. Noteboom, and H. Quene, "Word-level intelligibility of time-compressed speech: Prosodic and segmental factors," *Speech Communication*, vol. 41, no. 1, pp. 287–301, 2003.

[9] T. Bunnel and J. Liley, "Analysis of methods for assessing TTS intelligibility," in *Proceedings of Sixth ISCA Workshop on Speech Synthesis, Bonn, Germany*, 2007.

[10] C. Stevens, N. Lees, J. Vonwiller, and D. Burnham, "On-line experimental methods to evaluate text-to-speech synthesis: Effects of voice gender and signal quality on intelligibility, naturalness and preference," *Computer Speech and Language*, vol. 19, no. 1, pp. 129–146, 2005.

[11] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proceedings of Sixth ISCA Workshop on Speech Synthesis, Bonn, Germany*, 2007.