T(ime) T(o) S(tart) synthesising audio description in China? Results of a reception study

Irene Tor-Carroggio, Autonomous University of Barcelona

ABSTRACT

Text-to-speech audio description has proven to be an accepted method to increase the output of this access service in several languages, such as English, Japanese, Polish, Spanish and Catalan – at least as an interim solution until more audio description is available with human narrators. In the light of China's need to make audio description more widely available, we propose text-to-speech audio description as a means to provide this access service both faster and more economically. This article describes and analyses the results of a reception study carried out in China to test the acceptance of text-to-speech audio description in comparison with standard human-voiced audio description. The study sample consisted of forty participants and was carried out using clips from a Chinese historical movie. With the exception of comprehension, the results show that natural voices score statistically higher than synthetic voices, which suggests that those with sight loss prefer human-voiced audio description. Yet, it cannot be stated that text-to-speech audio description is not accepted: more than half of the study participants not only see this as an interim alternative, but also as a welcome permanent solution if it means more audio described movies.

KEYWORDS

Audio-visual translation, media accessibility, audio description, text-to-speech, speech synthesis, reception study.

1. Introduction

China is the country with the highest number of people with disabilities (Wu and Xie 2015: 21). It not only numbers among the countries with the oldest populations, along with Japan, the United Kingdom and the USA (Wang *et al*. 2014: 76), but also its elderly population is expected to rise to 30% by 2050 (Wang *et al*. 2014: 76). Countries facing ageing societies need to tackle a plethora of pressing challenges, including how to accommodate people with disabilities because disability is closely related to age (Peng *et al.* 2010). Although audio description (AD) is still in its infancy in China, especially as an object of academic study, it is of paramount importance to make it more widely available and broaden its scope, as it is mainly restricted to films (Tor-Carroggio and Casas-Tost, *forthcoming*); this would not only benefit people with sight loss, but also sighted audiences (Ramos 2015). Various low-cost proposals have been put forward to extend this type of intersemiotic translation and tested with reception studies, such as AD translation (Jankowska 2015; Jankowska *et al*. 2017) and text-to-speech (TTS) AD (Szarkowska 2011). The latter is, for example, widely used in theatres that offer AD in Spain (Hermosa, *forthcoming*). According to Szarkowska (2011: 145), TTS AD:

> […] offers unequalled cost-effectiveness in terms of AD production in comparison with conventional methods of producing audio description as it does not require the

> recording of the AD script (for pre-recorded AD) nor does it incur any human labour costs for the reading out of the AD script (for live AD).

Nonetheless, and unlike languages such as Catalan (Fernández-Torné and Matamala 2015) and Japanese (Kobayashi *et al.* 2010), TTS AD has never been formally tested in Chinese, which means that it cannot yet be considered a potential solution to expedite the production of AD and make it more cost-effective. In fact, TTS AD in Chinese may produce different results from previous research in terms of reception, because TTS is constantly being improved and has to meet different challenges compared to other languages. Wang *et al*. (2006) list some of these challenges, the most salient examples being the large number of commonly used Chinese characters (around 20,000 excluding ancient documents), phonological changes occurring in tonal languages, words not being formally segmented and the polyphony of around 600 characters. Consequently, and in order to study how Chinese AD users respond to the way TTS tackles these challenges, a study was deemed necessary to bridge this research gap. Since TTS AD had already been evaluated in other languages and contexts, a replication study was considered to be the most suitable methodological approach.

Replication is crucial for scientific progress and, as Olalla-Soler (2019) argues, the scarcity of available literature on this issue in Translation and Interpreting Studies begs the need for more replication studies. Contrasted hypotheses should be tested many times to counteract the high probability of producing false research results (Ioannidis 2005), and in the case of this study, a constructive replication has been performed, i.e., the reliability of the effect detected in previous studies has been assessed by modifying a limited number of aspects of the experimental design (Olalla-Soler 2019).

Hence, the objective of this article is twofold. First, it attempts to evaluate TTS AD in Chinese by comparing it with standard human-voiced AD regarding key features. Second, it also attempts to analyse and demonstrate whether TTS AD would be accepted in China as either an interim solution until there is more AD available with human narrators, a permanent solution, or both. This paper is divided into four sections. The first reviews previous studies on TTS, particularly TTS AD, in order to situate this research initiative within the current academic context. The second describes the methodology employed for this quasi experimental research. The third presents and discusses the results obtained and the final section sets out the study conclusions, its limitations and future perspectives to explore.

2. Previous research on TTS

TTS has proven to be an effective support instrument for people with sight loss (Cryer and Home 2008). Its applications are quite diverse and range from global positioning systems (GPS) to educational, communication and

entertainment tools (Cryer and Home 2008: 5). Some studies suggest that how acceptable these voices are depends on how much experience users have with them (Szarkowska and Jankowska 2012), and that not only is it possible to get used to them (Hjelmquist *et al*. 1990 in Cryer and Home 2008), but also in some cases users have stated they prefer these types of voices. The reasons given are their lack of expressiveness, because this feature allows them to concentrate more on the content, it grants them more autonomy and guarantees more confidentiality (Llisterri *et al*. 1993). Although in most cases users prefer natural voices (Szarkowska 2011; Fernández-Torné and Matamala 2015), they are also aware that products with artificial voices expedite access to information, since they do not need to wait for someone to voice the content for them. This was clearly shown in the study by Thompson *et al*. (1999 in Cryer and Home 2008), whose objective was to discover how those with sight loss preferred to receive tax information. Stevens *et al*. (2005) noted that the key factor behind accepting synthetic voices lies in how natural they sound. In short, these previous studies suggest that users' subjective evaluations of artificial voices are based on how much experience they have with them, the context in which they are used and the voice features.

Cryer and Home (2009) conducted a study on the acceptance of TTS in audio books. This study is important as it is the only one which collected very valuable qualitative information that can be extrapolated to other immersive fields such as movies. In their case, the opinions reported by users were quite varied: some saw the potential of the idea, while others considered it to be an aberration and stressed its shortcomings. However, most participants came to the conclusion that not all types of books were suited to being read with artificial voices: while likely applications were instructional, educational or non-fiction books, a human narrator was preferred for fiction and leisure reading. Cryer and Home (2009: 26) also reported that the acceptance of TTS audio books is very likely to be experience-dependent, since TTS is better accepted among people who have been previously exposed to it. Yet, they also admitted that this is not always the case. Finally, despite the diverse opinions, there were a few participants in their study who stated they would accept this solution if it meant more access to information.

2.1. Previous research on TTS in AVT

TTS has been applied to various modalities of AVT, namely voice-over, audio subtitling (AST) and AD. Regarding the former, Matamala and Ortiz-Boix (2018) compared TTS with human voices in a wildlife documentary with a sample of 16 participants and two clips that lasted around two minutes. Using questionnaires, they elicited participants' opinions in terms of self-reported interest, engagement and enjoyment. Participants were also asked to rate the quality, naturalness and comprehensibility of the voices they heard, as well as to respond to questions on comprehension and preferences. Although human voices were rated higher in general, there

were no differences in their self-reported engagement. However, a clear positive opinion about using artificial voices as a possible substitute for human voices for this type of translation was not observed, although more than half of their respondents stated excerpts voiced with TTS would be acceptable in a TV broadcast.

TTS AST has also been tested and is actually a reality in many countries such as Spain (Rovira-Esteva and Tor-Carroggio 2018), the Netherlands, Denmark and Sweden (Iturregui-Gallardo 2019). While some authors focused on providing improved technical solutions to deliver TTS AST (Derbring *et al.* 2009), Thrane (2013) evaluated the difficulties encountered by 16 TTS AST consumers when watching the news, documentaries and fiction content. She concluded that some of the most problematic areas were synchronisation, pronunciation, the presence of multiple voices and speed. She also discovered that TTS ASTs were considered more acceptable for non-fiction content such as the news. A similar conclusion was reached by Kobayashi *et al.* (2010), who found that TTS AD was generally accepted, especially in short informative videos. Kobayashi *et al.* (2010) assessed the acceptance and effectiveness of TTS AD in various genres through surveys and in-depth interview sessions in both Japan and the US. In the survey distributed in Japan, users ranked the voices heard (human, prototype TTS and standard TTS), whilst during the interviews they were asked to evaluate their experience using enjoyability and intelligibility as evaluation criteria. Although the study carried out in the US shared the same objectives, it was slightly different. The survey was designed to test comprehension and comfort, as well as effectiveness and preferences. The test ended with in-depth interviews to explore the characteristics of TTS AD.

Kobayashi *et al*.'s (2010) study has not been the only one putting TTS AD to the test. Previous studies also explored TTS AD acceptance in Polish through questionnaires using a wide range of materials: an educational animated series (Walczak 2010), a Polish feature film (Szarkowska 2011), a foreign film dubbed into Polish (Drożdż-Kubik 2011), a foreign movie with voice-over (Szarkowska and Jankowska 2012) and a documentary with AST (Mączyńska and Szarkowska 2011). Despite the variety of genres evaluated, the conclusions were similar in all cases: TTS AD is accepted in most cases as an interim solution, and sometimes even as a permanent solution. Nevertheless, TTS AD seems to be more acceptable with non-fiction content. It is worth noting that Szarkowska and Jankowska (2012: 86) pointed out quality (especially intelligibility and naturalness) as playing a crucial role in user comprehension and attitudes towards TTS. Also, and compared to viewers with low vision, Szarkowska and Jankowska (2012) detected that blind participants are more supportive of TTS. This finding could be explained by blind viewers being more dependent on AD and thus wanting more audio described films regardless of the voice.

Besides Polish, TTS AD has also been tested in other European languages, such as Catalan (Fernández-Torné and Matamala 2015). Fernández-Torné

and Matamala (2015) tested TTS AD acceptance in a dubbed feature film. Their reception study was conducted with 67 users, who assessed two synthetic voices applied to AD, as well as two natural voices. Participants were administered a questionnaire to rate voices taking into account various end user reception-related items, and to answer questions about their personal preferences. The conclusion reached is in line with those of the Polish project: most participants accepted Catalan TTS AD as an alternative solution to the human-voiced AD. However, natural voices outperformed the artificial ones tested and were still the preferred solution.

These studies focused mainly on the acceptance of TTS AD, but other aspects have also been investigated, such as the effect of TTS AD both on emotion and presence (Fryer and Freeman 2014; Walczak and Fryer 2017) using the emotion elicitation scale and the ITC-Sense of presence inventory, as well as preference questions. Results indicate that higher levels of presence are obtained for AD delivered by a human voice, especially for drama. Likewise, AD delivered by a human voice also enhances emotion.

Finally, although no academic study of TTS AD in Chinese can be reported, the Chinese company Shanghai Gaozhi Keji Ltd. (上海高智科技公司) had already produced a few films with TTS AD back in 2014. According to this company, the production of an entire film usually took them around three days and they used a free online TTS to voice the scripts (the voice can be listened at https://www.xfyun.cn/services/online_tts). The main problem detected at the time was related to the sentence rhythm, especially in names of people and places that are not well-known. This happened before the current group of radio presenters that volunteer to voice ADs in Shanghai was created (Tor-Carroggio and Casas-Tost, *forthcoming*). Yet, this TTS AD was never tested (formally or informally) with users and was rapidly substituted for human-voiced versions.

3. Methodology

The stimuli used in our study came from a fiction product kindly provided by the Shanghainese AD production centre Sound of Light (Guangying zhi Sheng, 光影之声). This non-profit organisation describes 50 movies every year and each AD script is supervised by an AD user. We worked from Tor-Carroggio's study (2020) on Chinese AD user movies preferences, which concluded that informants (N = 52), who were mostly elderly people from Shanghai, preferred historical movies when asked which type they would like this experiment to be carried out with. This genre ranked first and second, respectively, in the two questionnaires presented. This finding sheds some light on the study participants' preferences, since those recruited for this experiment were expected to have a similar demographic profile. This situation was due to the author having a limited number of user sources, which obviously limits the representativeness of our sample. Consequently, the Chinese historical movie *Our time will come* (Mingyue Ji

Shi You, 明月几时有), directed by Ann Hui and released in 2017, was chosen. This movie revolves around the resistance movement during Japan's occupation of Hong Kong in the 1940s.

The film and its AD script were analysed in depth in order to select two clips that were as comparable as possible in terms of content (food is mentioned in both), length (almost 5 minutes), intervening characters (the film's two leading actresses appear in both), background music (almost no music or striking sounds are heard), and AD density (around 600 characters each). This information is summarised in Table 1. Both clips start describing what Fang Mu (one of the main characters' mother) is doing and coincide in mentioning items such as a mirror and rice. Both clips include people's names, as this is one of the elements that can be problematic for TTS to resolve. Also, the AD briefly overlaps with the dialogue in both cases. Yet, the spots were wisely chosen by the Chinese audio describers because no relevant information is missed, as the Sound of Light guidelines recommend (Tor-Carroggio and Vercauteren, *forthcoming*). Finally, in these two clips, some of the AD units were rather long. We thought having longer units would allow users to listen to the voice for a longer time and, therefore, get more used to it. This was deemed important as TTS acceptance appears to be dependent on how much users are exposed to it (Szarkowska and Jankowska 2012).

The movie was already described in Chinese with a male human voice, consequently, the gender of the artificial voice was chosen to match the original clip (male as well). The male voice was synthesised using the Chinese software Ke Da Yuyin Ku 4.0 (科大语音酷 4.0) as it is available for free. Since in one of the clips only female characters intervened, having a male voice was considered to be a good choice to be able to clearly differentiate the AD voice. Regarding the audio mix, the researcher was assisted by the company Shanghai Gaozhi Keji Ltd, which has experience in AD mixing. At this point, it must be pointed out that, although TTS AD can be read directly by speech synthesis software, recording it was seen as the safest option taking into account the number of individual homes the researcher had to visit and the various problems that could arise.

Only one minor change had to be applied to the original AD script so that the TTS would have a better final effect. The retroflex sound *er* (儿), a characteristic of Chinese northern dialects which could be not pronounced in all the cases in this sample without affecting comprehension, was deleted because the TTS would not join its sound with the character just before it, which is how it should be read.

Clip	Human voice	Artificial voice	Characteristics
Clip A	Male	Male	4'56'' min 556 Chinese characters

Clip B	Male	Male	4’37’’ min 592 Chinese characters

Table 1. Clips selected and their characteristics

A questionnaire to rate the voice in each clip was designed. So, each participant answered two questionnaires, which were administered orally by the researcher. These questionnaires included comprehension questions, as recommended by Chmiel and Mazur (2012), and a list of nine parameters that had to be evaluated after listening to each clip:

- Voice naturalness
- Voice pleasantness
- Speech pauses
- Ease of listening
- Comprehension
- Pronunciation
- Intonation
- Acceptance
- Overall impression.

These items were selected mainly from the International Telecommunications Union (ITU) Recommendation P.85 (1994), which defines a testing method for evaluating the subjective quality of synthetic speech. Other authors were consulted to add other suitable items to be evaluated, such as Hinterleitner *et al.* (2011) and Viswanathan and Viswanathan (2005), following the research designed by Fernández-Torné and Matamala (2015).

Although the objective (i.e. technical) acoustic evaluation of synthetic voices has proven to be useful “to measure where voices differ from a human utterance” (Cryer *and* Home 2010: 6), a more subjective approach involving user testing was taken for one main reason: users are those who would ultimately use TTS AD, although they “differ in both ability and opinion” (Cryer and Home 2010: 7). Also, “[s]ubjective user testing is more useful for someone considering the voice for use in a product or service, to find out whether users are happy with the voice” (Cryer and Home 2010: 4). The most common test for measuring opinions in this regard —and the one recommended by the ITU— is the Mean Opinion Score (MOS). This test involves participants listening to synthetic speech and rating the voices on a simple 5-point categorical scale; scores are then averaged across the group. In our case, though, the scale was changed to 0-10, which is allowed by the ITU (1994: 1). Other studies have also used scales with higher granularity because, theoretically, they can result in smaller standard deviations of MOS (Streijl *et al.* 2016). In our case, this was done mainly to avoid statistical ties as much as possible.

The experiment was approved by our university's Ethics Committee in June 2019 and piloted with three users before formally conducting it. In each experiment the informant was first read the information sheet and the consent form out loud. Their expressed consent to participate in the study was recorded on audio. During the experiment each participant watched two clips: one described with TTS and another one described with a human voice. Table 2 presents the listening order of the voices for participants, which followed a Latin square. This order was repeated with all the users recruited. Artificial voices (*) were always presented first to avoid a negative impact on their evaluation, as did Kobayashi *et al.* (2010) and Fernández-Torné and Matamala (2015). The latter based their decision on van Santen (1993), and Viswanathan and Viswanathan (2005: 62).

User	Clip 1	Clip 2
01	A*	B
02	B*	A

Table 2. Listening order of the clips

A post-questionnaire was drafted based on Szarkowska and Jankowska (2012) and Fernández-Torné and Matamala (2015). It was aimed at gathering more qualitative information and the participants' demographic data. Given that all our participants suffered from sight loss because they are the main target group AD caters for, there were no questions relating to their disability. This was decided in accordance with the capabilities approach suggested by Mitra (2006), which essentially explains how disability may be triggered by three different factors: the individual's personal characteristics (e.g., impairment, age, race, gender), the individual's resources and the individual's environment (social, economic, political). Consequently, it was believed that disability can be disentangled from being closely related to a physical impairment, so questions regarding disability were deemed unnecessary.

All the questionnaires were translated into Chinese by a professional Chinese translator, whose work was checked by another Chinese translator. The translators were required to use rather simple language because we expected that many users would be elderly people with little or no education at all.

Users were recruited through personal contacts and with the kind help of the Shanghai Guide Dogs' Club and the Shanghai Association of Persons with Disabilities. The only requirement was to have watched at least one audio described movie. All the tests took place in Shanghai, more specifically in the users' homes, in community centres and in East China Normal University. The session in East China Normal University was followed by a focus group with ten users that shared their views with us on TTS AD and on AD in China in general.

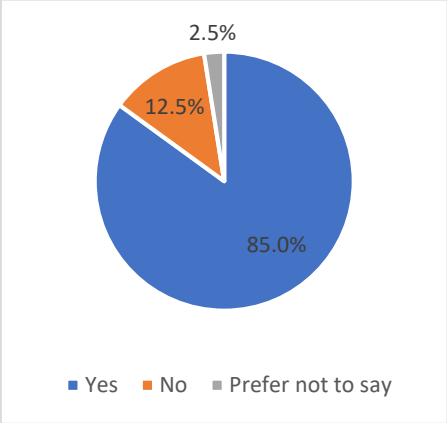
Finally, the quantitative results were analysed statistically using SPSS (version 25). Non-parametric tests were run (i.e. Wilcoxon signed-rank test and Spearman's rank correlation coefficient) due to the limited size of our sample. The p-value threshold for the declaration of statistical significance was set at 0.05.

4. Results and discussion

This section presents the results yielded by our study. It begins with a presentation of the demographic characteristics of our sample and then moves on to present both the quantitative and qualitative data gathered.

4.1. Description of the sample

Forty AD users took part in the test, 52.5% of whom were men, while 45% were women (one person did not answer this question). As expected, most of the participants were quite elderly: average 58.64 and median 61 (Min=30, Max=78). Regarding their educational background, the majority of the participants had completed secondary education or had attended a vocational school (72.5%), close to a fifth had attended university (17.5%), one person had only completed primary education and another had not received any formal education at all (two participants declined to answer this question). As for their experience with TTS, 85% of the participants claimed to have used it before (Figure 1), mainly in their mobile phones and computers, and more than half of the participants reported using TTS often or very often (Figure 2).



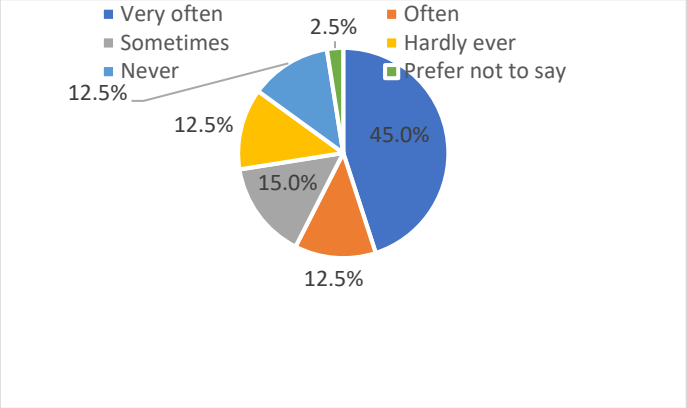


Figure 1. Experience with TTS **Figure 2. TTS frequency of use**

All the participants had experience with AD at least once. More than half of the AD users recruited reported using AD in movies often or very often (67.5%) whenever it is available, which is not always the case. This has to be framed within the Shanghainese context: despite having other opportunities at different venues, only one live audio described movie is available in cinemas every month, screened in specific cinemas but just for one day (Tor-Carroggio and Casas-Tost, *forthcoming*).

4.2. Comprehension and quality of the voices

After watching each of the two clips, the users were asked to answer three comprehension questions. The average number of correct answers was 1.42 for the clip audio described with TTS and 1.75 for the clip audio described with a human voice. Yet, no statistical differences were detected ($Z=-1.62$, $p=0.10$, with 14 ties). These rather low scores might be explained by different factors, such as some of the users being quite aged, which could mean that their cognitive faculties are less sharp. Also, the number of correct answers being higher in the case of the clip voiced with a human voice can be regarded as an effect of the order in which they were presented. Since the TTS version was always presented first, participants may have understood the clip better the second time or been prepared for the comprehension questions. It was decided not to tell them in advance what kind of questions they would need to answer because we intended to simulate a real situation in which users would not be paying excessive attention to every single detail for fear of a test. Furthermore, almost half of the tests were carried out in the users' homes, which were not always quiet environments. Choosing their homes as the testing venue was the only option possible in many cases, since it was quite inconvenient for many of them to leave their homes. However, this situation was seen as an opportunity to see users' response to a more ecological environment, which in the end is where AD TTS would most likely be used were it to be implemented. After carrying out a Mann-Whitney U test, no statistical differences were spotted in terms of correct answers to the comprehension questions between those who watched the clips at home and the rest who did not ($Z=-0.28$ and $p=0.77$ for the TTS clips, and $Z=-0.23$ and $p=0.81$ for the clips described with a human voice).

Nonetheless, and in order to clarify which factors had had an influence on the informants' comprehension, we decided to recruit a younger sample to watch the same clips in similar conditions. Fifty sighted third-year undergraduate students agreed to participate. They were BA Translation and Interpreting students from the Hangzhou Foreign Languages Institute, and were blindfolded during the experiment, which took place in one of their classrooms. In their case, the average number of correct answers was 1.80 for the clip audio described with TTS and 2.10 for the clip audio described with a human voice. Although it might seem that the students understood the contents of the clip better when described with a human voice, no statistical differences were detected ($Z=-1.81$, $p=0.07$, with 17 ties). These slightly better results might be linked to the students' age (they were all in their early twenties), to their education background and/or to the testing venue environment (which was quieter). Yet, no statistical differences were detected between the students and the AD users either for the first clip ($Z=-1.92$, $p=0.054$) or for the second one ($Z=-1.62$, $p=0.10$). This shows that, in this second round, the parameter 'comprehension' did not obtain statistically better results with younger and better-educated respondents in

a quieter environment. Yet, it is also true that it was the first time for our participants to listen to an AD, so maybe they found it hard to cope with the initial excitement and so much aural input. In any case, this gives some food for thought regarding the cognitive load AD users can cope with, regardless of their age and education background. In fact, when the questionnaire was first piloted with a user who works as an AD reviewer, this person also got some comprehension answers wrong and underlined the need to review what aspects were important to include in an AD script.

It should be pointed out, however, that although sighted students participated in the study and answered the exact same questions as the users with sight loss, this article mainly focuses on the data obtained from the visually impaired users, since, as Walczak (2010: 39) notes, they are the major beneficiaries of AD. Mendoza and Matamala (2019) reached a similar conclusion in Spain: unlike subtitles, people consuming AD tend to be visually impaired. In fact, these researchers found that not even the professionals of this service usually watch audio described products. Given that AD is still a novelty in China, and also that the students who participated had no previous experience with AD, their participation in the study was purely anecdotal and also used for them to experience the possibilities AD experimental research can offer. Therefore, the following data refers exclusively to the participants with sight loss.

As for the quality of the voices, the average score for each parameter tested was as follows (Table 3):

Parameter	MOS TTS	MOS Human voice
Voice naturalness	8.35	9.60
Voice pleasantness	8.20	9.18
Intonation	7.65	9.25
Pronunciation	9.10	9.60
Speech pauses	8.77	9.38
Ease of listening	9.20	9.63
Comprehension	9.80	9.85
Acceptance	9.13	9.68
Overall impression	8.57	9.33

Table 3. Descriptive results

Although the average number of correct answers for comprehension questions was rather low, most users thought they had understood everything. In fact, the parameter that was best assessed in both cases was ‘comprehension’. Yet, what is especially relevant is the fact that all items are assessed above 6, which is the minimum score to pass any test in China and, thus, our study’s threshold. All the parameters related to the human voice were above nine and scored better than those related to the TTS, however, some parameters from the latter also scored above nine (pronunciation, ease of listening, comprehension and acceptance). The

parameter that leaves most room for improvement is intonation, but that comes as no surprise since other studies (such as Walczak 2010) had already warned about unnatural intonation being one TTS's main drawbacks. Similarly, Fryer and Freeman (2014) also concluded prosody is a critical component of AD content in terms of presence and emotion elicitation, and Kobayashi *et al.* (2010: 167) had already reported users complaining about TTS being "less comfortable due to their 'flat' intonation." The importance of this parameter cannot be underestimated since, for example, Ramos (2015: 88) claimed that intonation clearly influences the emotional impact of texts.

Table 4 shows that statistical differences were detected in almost all the parameters under study. Thus, it can be stated that the quality of the human voice is regarded as better than that of the artificial one, except for the parameter 'comprehension,' in which no statistical differences were spotted.

Parameter	**Wilcoxon test results**
Voice naturalness	Z=-3.29 *p*<0.01 ties=10
Voice pleasantness	Z=-2.73 *p*<0.01 ties=11
Intonation	Z=-3.51 *p*<0.01 ties=11
Pronunciation	Z=-2.29 *p*=0.02 ties=21
Speech pauses	Z=-2.40 *p*=0.01 ties=21
Ease of listening	Z=-1.99 *p*=0.05 ties=22
Comprehension	Z=-1.42 *p*=0.15 ties=32
Acceptance	Z=-2.04 *p*=0.04 ties=20
Overall impression	Z=-2.60 *p*<0.01 ties=15

Table 4. Wilcoxon test results

It was deemed necessary to check whether there were any correlations among the parameters tested for the TTS. Statistically significant correlations were found in all cases except for two pairs ('speech pauses' and 'comprehension,' and 'comprehension' and 'acceptance'). The statistically significant correlation coefficients were above 0.5.

4.3. Preferences

Apart from quantitative data, our participants also provided some qualitative comments regarding their preferences when consuming AD, as well as the idea of applying TTS to AD. All the comments gathered are reproduced in this section. At this point it should be mentioned that users were generally open to any ideas that could eventually allow them to watch movies along with a sighted audience, for example, using earphones.

When asked about which of the two voices heard they had liked better, half of the participants selected the human voice, although some users were not very satisfied with it because they thought it had an accent and that it seemed to be reciting rather than describing. The comment regarding the voice talent's accent surprised us, since the majority of volunteers currently audio describing in Shanghai are professional radio presenters who are trained to speak perfect Standard Chinese. Future studies should not take this for granted and make sure the voice reflects the standard accent. Yet, 27.5% of those who responded had no clear preference and had liked both, while 20.0% preferred the artificial voice (Figure 3). In fact, the artificial voice was unexpectedly praised for being clear and some of those who liked it better claimed that it was very smooth and did not have an accent. Szarkowska and Jankowska (2012: 85) also report many users liking standard accent voices.

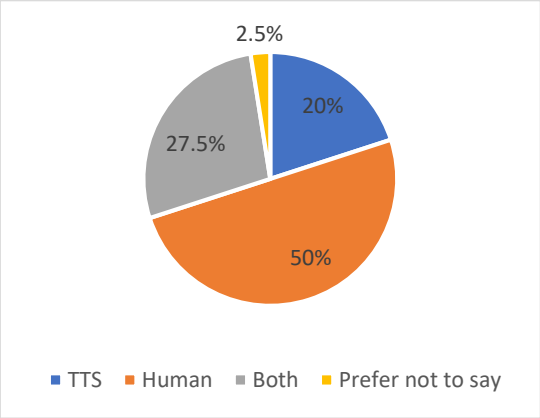


Figure 3. Voice preferred

Nonetheless, when asked what type of voice they would like to listen to in audio described films, over 50% chose human voices (57.5%), whereas only one user selected TTS (2.5%). Despite this clear difference in proportion, 15% of those who responded said they did not care and almost a quarter of the participants (22.5%) claimed that it depended on the movie.

Those who stated that it depended on the type of movie considered documentaries to be the most suitable genre for TTS AD. This coincides with Kobayashi *et al*. (2010: 170), who concluded that TTS AD is more "suitable for informational videos where understanding is the critical factor," and Fernández-Torné and Matamala's findings (2015). Similarly, Walczak and Fryer (2017: 77) found that the users recruited preferred to watch dramas with a human voice and were more sceptical about doing so with TTS. For the documentary, "out of 36 participants, 31 (86%) were eager to watch the documentary also with TTS AD" (Walczak and Fryer 2017: 77). Similarly, Walczak and Fryer (2017) came to the conclusion that presence rates, as well as levels of interest and confusion were comparable for documentaries audio described by a human voice and TTS. Yet, it should be noted that documentaries are not the products that benefit most from AD (Walczak 2010: 43). Our participants also considered historical and martial arts films to be appropriate for TTS AD. There was also one user who underlined AD was an art and the reason why TTS AD is not suitable for all kinds of movies. Nevertheless, users felt being able to understand the AD was more important than the voice used.

Although there was a clear preference for human voices (both quantitatively and qualitatively), all participants except for one stated that they would accept TTS AD as an interim solution until there is more AD available with human narrators (Figure 4) and almost 63% of the users would even accept it as a permanent solution (Figure 5).

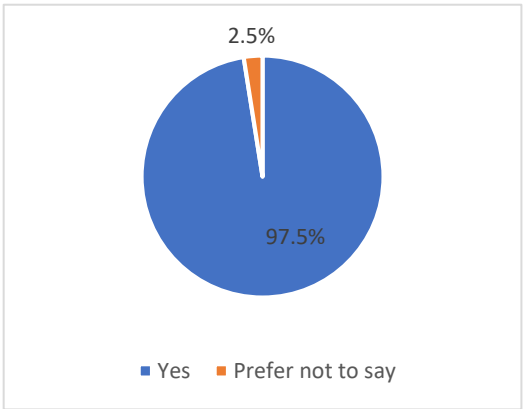


Figure 4. Interim acceptance

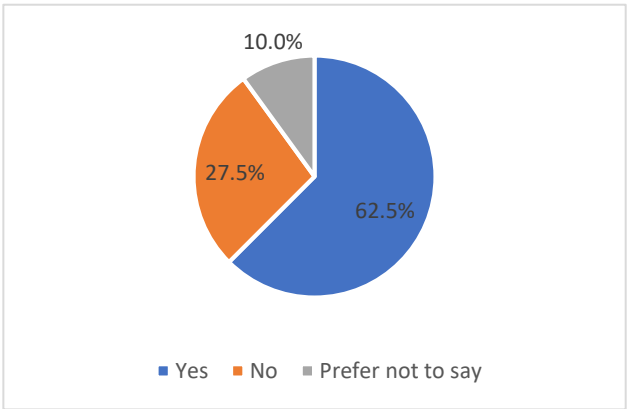


Figure 5. Permanent acceptance

As Table 5 illustrates, these numbers are very similar to those obtained in previous studies.

Study	Language and audio-visual stimuli	Acceptance of TTS as an interim solution	Acceptance of TTS as a permanent solution
Szarkowska (2011)	Polish/ Polish feature film	95%	58%
Szakowska and Jankowska (2012)	Polish/ Foreign film with Polish voice-over	95%	70%
Fernández-Torné and Matamala (2015)	Catalan/ Miscellaneous film	94% believe TTS AD can be an alternative solution to human-voiced AD	
Present study	Chinese/ Historical film	97.5%	62.5%

Table 5. Comparison with similar studies

The acceptance of TTS AD in the Chinese context cannot be divorced from the scarcity of audio described movies, which means that users do not have enough movies to choose from. Many of our participants complained about not being able to choose what audio described movies to watch and also about the scantiness of newer and foreign films. Szarkowska and Jankowska (2012: 84) had already pointed out that people with sight loss want to watch foreign films just like the sighted audience. Therefore, they were open to solutions that could increase the availability of audio described movies. Nonetheless, not everybody who accepted TTS AD did so because the alternative was fewer ADs or none at all. Some of our participants had difficulty in differentiating the artificial voice from the human one, and some did not even realise that the first clip had been described with an artificial voice. Also, some participants stated that having real voices for movie AD did not necessarily guarantee AD excellence. In fact, some users confessed checking who the AD voicer is in the monthly live AD sessions in cinemas to see whether they like the volunteer in charge of voicing AD script or not. Moreover, some participants mentioned that it is sometimes difficult to know when the actors talk and when the describer describes, since the voices might be very similar. These issues could be solved by using TTS AD, but also by making a better choice of human voices or a better mix of the original soundtrack with the AD.

Two users also mentioned that TTS is constantly improving, so they were already sure TTS AD could substitute traditional AD in the near future. In fact, some of them were aware of new TTS systems cloning human voices, which will make it necessary to replicate this study to validate our current findings. These two same users also raised an interesting point: they were

willing to pay for AD if that meant having more choices and having it online so that they could watch audio described movies at home. They also highlighted the fact of not liking having to feel grateful to the volunteers all the time and having to restrain themselves from pointing out aspects they do not like about the AD currently being delivered. Notwithstanding, the participants mainly complained that TTS AD was too monotonous and did not convey any emotions. But again, this was seen as a less relevant problem in comparison with not having access to more audio described materials.

Finally, it was deemed necessary to see whether the TTS acceptance item was correlated to either the participants' frequency of use of TTS or to how often they consume AD. No statistically significant correlations were found in such cases (r=0.06, p=0.72/ r=0.08, p=0.61). Yet, the users who participated in the focus groups mentioned them not having difficulties in accepting AD TTS for them being very used to artificial voices.

5. Conclusion

This article has presented the results of a reception study carried out in China to test the acceptance of TTS AD in comparison with conventional AD. The results show that natural voices have statistically higher scores than synthetic voices, which proves that people with sight loss prefer AD voiced by a human. Yet, this does not mean that TTS AD is not accepted, since it is viewed as both an interim alternative and even as a permanent solution by more than half of the participants if that means more access to audio described movies. We have also confirmed that prosody is still a pending subject for TTS, compared to other parameters. Therefore, if Chinese TTS companies were to suggest their voices for AD, they should improve them in this regard. Furthermore, the fact that our results confirm previous studies can be attributed to AD's situation in China being equivalent to that of, for example, Poland's ten years ago. Our participants gave the impression they would accept anything that could increase AD availability.

Although it is not the first time that TTS AD has been tested with users, our study has incorporated a series of features that make it innovative. To begin with, TTS AD had never been tested in Chinese. In fact, no AD-related tests had ever been conducted in Mainland China, so our study hopes to encourage other researchers, especially in the Chinese academic world, to engage in media accessibility research and, more specifically, in user testing. Moreover, the clips used in the study are slightly longer than those used by Fernández-Torné and Matamala (2015). This was something Fernández-Torné and Matamala suggested as a way of taking TTS AD research a step further. Fernández-Torné and Matamala (2015) also suggested testing it with different genres, which this study has also contributed to. In addition, this study followed a user-centric design and the genre was selected by users in a pre-study. Furthermore, comprehension questions were added before asking the users' opinion as regards the voices used. This was done

to assess user comprehension, since it is usually taken for granted but is crucial and has priority over technological innovations. We found that the use of TTS AD did not have a statistically significant impact on comprehension.

Obviously, this research has its limitations and we would like to note seven observations. First, only one synthetic voice is tested, which is also only compared to one natural voice. Furthermore, these two voices were not selected by AD users, so it is uncertain whether voices chosen by them would have yielded different results. Future studies should put to the test other voices that have previously been validated by users. Second, given the size and the socioeconomical disparities in China, our results cannot be extrapolated to the whole of China. Third, the size of our sample, which is not random, leaves room for improvement in future studies. Ramos (2015) acknowledges the difficulty of recruiting volunteers with sight loss. According to Ramos (2015: 87), "[t]his justifies the fact that reception studies in audio-visual translation usually work with smaller samples." Fourth, although the clips were slightly longer than those used by Fernández-Torné and Matamala (2015), they are still short and, therefore, studies with complete movies audio described with TTS should be conducted in the future. Fifth, Fernández-Torné and Matamala (2015) also suggest testing for engagement, which would have been useful in this case as well. Sixth, we are aware that by following the design by Fernández-Torné and Matamala (2015), the order of the clips was not sufficiently randomised (the synthesised speech was always presented first). The order effects could be further explored but right now our choice could also be seen as a limitation. Seventh, since we did not ask about the type of disability, it would have made more sense to include more potential AD users, such as people with mental impairments.

Nevertheless, it is hoped that this study has provided an empirical basis from which to initiate further debate and contributions to this area. Despite incorporating these innovative aspects, and given AD's dependence on volunteers, it would be interesting to investigate other ways to produce AD in a less time-consuming way in Chinese. An example of this could be AD translation, which has already proven to be successful in some language pairs, such as English and Polish (Jankowska 2015), while Szarkowska (2011) also suggested exploring AD templates. These two ideas could be particularly useful in Shanghai, since voicers outnumber AD scriptwriters. Also, it would be helpful to study how TTS can be used to provide more didactic tools for Chinese children with sight loss and how TTS affects emotion and presence rates.

Acknowledgements

This paper has been funded by the EasyTV project (GA761999), the Confucius Institute and the Department of Translation and Interpreting and East Asian Studies of the Autonomous University of Barcelona (UAB). The

author is member of the TransMedia Catalonia research group (2017SGR113) and is currently enrolled in the PhD Programme in Translation and Intercultural Studies at UAB. The author would like to thank the users who participated in the study and the reviewers for their enriching and helpful suggestions.

References

- **Chmiel, Agnieska and Iwona Mazur** (2012). "AD reception research: some methodological considerations." Elisa Perego (ed.) (2012). *Emerging topics in translation: Audio description*. Trieste: EUT, 57-74.

- **Cryer, Heather and Sarah Home** (2008). "Exploring the use of synthetic speech by blind and partially sighted people. Literature review #2." Birmingham: RNIB Centre for Accessible Information (CAI).

- **Cryer, Heather and Sarah Home** (2009). "User attitudes towards synthetic speech for Talking Books." Birmingham: RNIB Centre for Accessible Information (CAI).

- **Cryer, Heather and Sarah Home** (2010). *Review of methods for evaluating synthetic speech*. Technical report #8. Birmingham: RNIB Centre for Accessible Information (CAI).

- **Derbring, Sandra, Peter Ljunglöf and Maria Olsson** (2009). "TS: Light-weight automatic reading of subtitles." Kristiina Jokinen and Eckhard Bick (eds) (2009). *NODALIDA 2009 Conference Proceedings*. Odense: Northern European Association for Language Technology 272–274.

- **Drożdż-Kubik, Jan** (2011). *Harry Potter i Kamień Filozoficzny słowem malowany – czyli badanie odbioru filmu z audiodeskrypcją z syntezą mowy*. Masters dissertation. Jagiellonian University.

- **Fernández-Torné, Anna and Anna Matamala** (2015). "Text-to-speech vs. human voiced audio descriptions: a reception study in films dubbed into Catalan." *The Journal of Specialised Translation* 24, 61-88.

- **Fryer, Louise and Jonathan Freeman** (2014). "Can you feel what I'm saying? The impact of verbal information on emotion elicitation and presence in people with a visual impairment." Anna Felnhofer and Oswald D. Kothgassner (eds) (2014). *Proceedings of the International Society for Presence Research*. Wien: Facultas, 99-107.

- **Hermosa, Irene** (*forthcoming*). "Delivery Approaches in Audio Description for the Scenic Arts." To appear in *Parallèles*.

- **Hinterleitner, Florian *et al.*** (2011). "An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks." *Proceedings of the Blizzard Challenge Workshop*. International Speech Communication Association (ISCA). Florence.

- **Hjelmquist, Erland, Bengt Jansson and Gunilla Torell** (1990). "Computer-oriented technology for blind readers." *Journal of Visual Impairment and Blindness* 17, 210-215.

- **Ioannidis, John PA** (2005). "Why most published research findings are false." *PLoS Medicine* 2(8), 696–701.

- **ITU**-T Recommendation P.85 (1994). *Telephone transmission quality subjective opinion tests. A method for subjective performance assessment of the quality of speech voice*

output devices. Geneva: ITU. http: //www.itu.int/rec/T-REC-P.85-199406-I/en (consulted 7.11.2019).

- **Iturregui-Gallardo, Gonzalo** (2019). *Audio Subtitling: Strategies for Voicing Subtitles and Their Effect on Film Enjoyment and Emotion*. PhD thesis. Autonomous University of Barcelona.

- **Jankowska, Anna** (2015). *Translating Audio Description Scripts: Translation as a New Strategy of Creating Audio Description*. Berna: Peter Lang.

- **Jankowska, Anna, Michał Milc and Louise Fryer** (2017). "Translating audio description scripts… into English." *SKASE Journal of Translation and Interpretation* 10(2), 2-16.

- **Kobayashi, Masatomo *et al.*** (2010). "Are Synthesized Video Descriptions Acceptable?" *ASSETS '10: Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility*. New York: ACM, 163-170.

- **Llisterri, Joaquim *et al.*** (1993). "Testing users' acceptance of Ciber232, a text to speech system used by blind persons." Björn Granström, Sheri Hunnicut and Karl-Erik Spens (eds) (1993). *Speech and Language Technology for Disabled Persons. Proceedings of an ESCA Workshop*. Stockholm: KTH-ESCA, 203-206.

- **Matamala, Anna and Carla Ortiz-Boix** (2018). "Text-to-speech voice-over? A study on user preferences in the voicing of wildlife documentaries." *SKASE Journal of Translation and Interpretation* 1, 24-38.

- **Mączyńska, Magdalena and Agnieszka Szarkowska** (2011). "Text-to-speech audio description with audio subtitling to a non-fiction film "La Soufriere" by Werner Herzog." Paper presented at the *Media for All Conference. Audiovisual Translation: Taking stock* (Imperial College London, 28 June 2011).

- **Mendoza, Nuria and Anna Matamala** (2019). "Panorama de la enseñanza de la audiodescripción en España: resultados de un cuestionario." *MonTI. Monografías de Traducción e Interpretación* 11, 155-185.

- **Mitra, Sophie** (2006). "The Capability Approach and Disability." *Journal of Disability Policy Studies* 16(4), 236-247.

- **Olalla-Soler, Christian** (2019). "Practices and attitudes toward replication in empirical translation and interpreting studies." *Target* 32(1), 3-36.

- **Peng, Xiaoxia *et al.*** (2010). "Ageing, the urban-rural gap and disability trends: 19 years of experience in China - 1987 to 2006." *PLoS ONE* 5(8), e12129.

- **Ramos, Marina** (2015). "The emotional experience of films: does audio description make a difference?" *The Translator* 21(1), 68-94.

- **Rovira-Esteva, Sara and Irene Tor-Carroggio** (2018). "Serveis d'accessibilitat sensorial a les televisions que emeten en català: situació actual i propostes de futur." *Quaderns del CAC* 44(21), 71-80.

- **Stevens, Catherine *et al.*** (2005). "On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference." *Computer Speech and Language* 19, 129-146.

- **Streijl, Robert C., Stefan Winkler and David S. Hands** (2016). "Mean Opinion Score (MOS) revisited: Methods and applications, limitations and alternatives." *Multimedia Systems* 22(2): 213–227.

- **Szarkowska, Agnieszka** (2011). "Text-to-speech audio description: towards wider availability of AD." *The Journal of Specialised Translation* 15, 142-163.

- **Szarkowska, Agnieszka and Anna Jankowska** (2012). "Text-to-speech audio description of voice-over films. A case study of audio described *Volver* in Polish." Elisa Perego (ed.) (2012). *Emerging Topics in Translation: Audio description*. Trieste: EUT, 81-98.

- **Thompson, Leanne, Christopher Reeves and Kate Masters** (1999). "In the balance: making financial information accessible." *British Journal of Visual Impairment* 17(2), 65-70.

- **Thrane, Lisbeth** (2013). *Text-to-speech on Digital TV An Exploratory Study of Spoken Subtitles on DR1Syn*. Masters dissertation. University of Copenhagen.

- **Tor-Carroggio, Irene and Helena Casas-Tost** (*forthcoming*). "Who Is Currently Audio Describing in China? A Study of Chinese Audio Describer Profiles." To appear in *MonTI* in 2020.

- **Tor-Carroggio, Irene** (2020). "The customer is always right: study on Chinese persons with sight loss' opinion on their experience with audio description." *Disability & Society*. Doi: 10.1080/09687599.2020.1713727

- **Tor-Carroggio, Irene and Gert Vercauteren** (*forthcoming*). "When East Meets West: A Comparison of Audio Description Guidelines in China and Europe" To appear in *HIKMA* in 2020.

- **van Santen, Jan P.H.** (1993). "Perceptual experiments for diagnostic testing of text-to-speech systems." *Computer Speech & Language* 7(1), 49–100.

- **Viswanathan, Mahesh and Madhubalan Viswanathan** (2005). "Measuring speech quality for text-to-speech systems development and assessment of a modified mean opinion score (MOS) scale." *Computer Speech and Language* 19, 55-83.

- **Walczak, Agnieszka** (2010). *Audio description for children. A case study of text-to speech audio description of educational animation series Once Upon a Time... Life*. Masters dissertation. University of Warsaw.

- **Walczak, Agnieszka and Louise Fryer** (2018). "Vocal delivery of audio description by genre: measuring users' presence." *Perspectives* 26(1), 69-83.

- **Wang, Ren-Hua *et al*.** (2006). "Mandarin text-to-speech synthesis." Chin-Hui Lee *et al*. (eds) (2006). *Advances in Chinese Spoken Languages Processing.* Singapore: World Scientific, 99-124.

- **Wang, Hua *et al*.** (2014). "Zhongguo Renkou Laolinghua Shehui Fazhan yu Yingdui Celüe." *Zhongguo Shehui Yixue Zazhi* 31(2), 75-77.

- **Wu, Zongyi and Zhenzhen Xie** (2015). "Zhongguo shizhang koushu yingxiang fuwu de fazhan xianzhuang yu dalu tuiguang." *Xinwen Yanjiu Shikan* 6(10), 20-22.

Biography

Irene Tor-Carroggio is a Ph.D. student in Translation and Intercultural Studies at the Universitat Autònoma de Barcelona (UAB) and is also a member of the research group TransMedia Catalonia (2017SGR113). She holds a B.A. in Translation and Interpretation from the UAB (2013) and also an M.A. in International Business from Shanghai University of Finance and Economics (2017). She is part of the EU-funded project EasyTV, http://easytvproject.eu.

Email: irene.tor@uab.cat