Text-to-speech voice-over?
A study on user preferences in the voicing of wildlife documentaries

Anna Matamala, Carla Ortiz-Boix

Abstract
In many countries the television broadcast of wildlife documentaries is nowadays translated from English and voiced by professional voice talents in the target language. This article discusses an alternative scenario in which text-to-speech would be used for the voicing of what is called "text-to-speech voice-over". It reports the results of an experiment in which a group of volunteers assessed excerpts voiced by synthetic voices as compared to excerpts voiced by human voices. Although human voices receive globally better assessments, results leave the door open to future research in the field.

Keywords
Audiovisual translation, voice-over, wildlife documentaries, speech synthesis, technologies

Introduction

Automatization (or semi-automatization) is increasingly present in our society. Research efforts in implementing new technologies in various processes are being made, and the field of translation and interpreting is no exception. In the area of audiovisual translation, however, research has traditionally lagged behind compared to other translation modalities, and has focused almost exclusively on the machine translation of subtitles (Volk 2008, De Sousa et al. 2011, Del Pozo 2014).

Taking into account this situation, the ALST project was launched (Matamala 2016). Although limited in its funding and scope, it aimed to research the implementation of speech recognition, machine translation and text-to-speech in two audiovisual transfer modes which share a key feature: their oral delivery. On the one hand, audio description (Maszerowska et al. 2014) was chosen as an instance of sensorial accessibility; on the other, voice-over (Franco et al. 2010) was selected as an instance of linguistic accessibility.

Within the ALST project research has been carried out on: a) the implementation of speech recognition in audio description (Delgado et al. 2015) and in voice-over (Matamala et al. 2017); b) the implementation of machine translation in audio description (Fernández-Torné and Matamala 2016), and in wildlife documentaries to be voiced-over (Ortiz-Boix and Matamala 2015, 2017), and also on c) the application of text-to-speech in audio description (Fernández-Torné and Matamala 2015). This article presents the last piece of research carried out within the project, and focuses on text-to-speech (TTS) in the voice-over (VO) of wildlife documentaries.

Voice-over is a pre-recorded audiovisual transfer mode mainly used in non-fictional content in which a translating voice is superimposed on the original voice, which can still be heard underneath. The translation usually fits in a limited space, beginning some words after the original utterance starts and finishing some words before the original utterance finishes (Matamala, forthcoming). It is also used for fictional genres in some Eastern European countries, presenting slightly different features. Traditional lip-synch dubbing constraints

(Chaume 2012, Matamala 2010) do not apply to voice-over, which often coexists with off-screen dubbing in non-fictional content. Off-screen dubbing refers to an audiovisual transfer mode in which the original words are totally substituted by a translated version, so that the original speech cannot be heard. It is usually applied when the speaker – generally a narrator in non-fictional content – is off-screen.

Voice-over does not generally involve any automatization processes. However, it is worth stressing that "automatic voice-over" was considered by the Strategic Research Agenda for Multilingual Europe 2020 as an open challenge in creative contents and creative works:

> open challenges are the automatic production of sign-language translation and dubbing. Especially automatic dubbing will be a hard task since it requires the interpretation of the intonation in the source language, the generation of the adequate intonation in the target language, and finally lip synchronization. An easier method would be automatic voice-over. In 2020 we will see a wide use of automatic subtitling and first successful examples of automatic voice over for a few languages (Rehm and Uszkoreit 2012:38)

Automatic voice-over would probably involve a machine translation (with post-editing) of the original content, followed by a text-to-speech voicing of this output. This latter aspect, namely "text-to-speech voice-over" (TTS VO), will be approached here. This paper presents the results of an experiment designed to gather users' opinions on wildlife documentary excerpts translated from English into Spanish and voiced by human and synthetic voices.

The article begins with a summary of text-to-speech research, focusing exclusively on the field of audiovisual translation. It then summarizes methodological aspects, and presents and discusses the results. Conclusions and further research avenues are presented at the end of this article which is exploratory in nature.

Text-to-speech systems in audiovisual translation

Research on text-to-speech systems in audiovisual translation has focused mainly on its use in audio description and audio subtitling. A project developed at the University of Warsaw, Poland, assessed the feasibility of text-to-speech audio description (TTS AD) and its reception among blind and visually impaired audiences, reducing costs and increasing accessibility. The project applied TTS AD to a monolingual feature film in Polish (Szarkowska 2011), to a dubbed educational TV series for children (Walczak and Szarkowska 2010), to a foreign fiction film with voice-over (Szarkowska and Jankowska 2012), to a non-fiction film with audio subtitling (Mączyńska 2011), and to a dubbed feature film (Drożdż-Kubik 2011). The majority of respondents found TTS AD acceptable, but not the preferred solution.

Similarly, Kobayashi et al. (2010) report on the results of an informal survey in both Japan and the USA with 115 and 236 visually-impaired adult participants, respectively, followed by in-depth interview sessions with three participants in the first case and eight in the second. Three types of voices were tested (human, standard TTS, and prototype TTS), and results show that synthesized audio descriptions are generally accepted, especially for relatively short videos and informational content.

Fernández-Torné and Matamala (2015) carried out similar research in which both synthetic and human voices were compared in Catalan audio descriptions. After a pre-test in which the "best" male/female human/synthetic voices were selected in a sample of 20 participants, 67 blind and visually-impaired volunteers took part in the main experiment. Four voices (male/female, human/synthetic) were assessed using a questionnaire inspired by ITU (1994), Viswanathan and Viswanathan (2005), Hinterleitner et al. (2011) and Cryer et al. (2010). Results show that natural voices have statistically higher scores than synthetic ones. However, 94% of the participants consider TTS AD to be an alternative acceptable solution to human audio description. Additionally, it is worth mentioning that no mean score of any of the items under analysis went under 3.1 on a 5-point Likert scale.

Apart from audio description, synthetic voices are extensively found in audio subtitling, where they are used to automatically read aloud subtitles and make them accessible not only to blind and visually impaired audiences but also to users with reading difficulties. This service, also called spoken subtitles, has been implemented in the Netherlands (Verboom et al. 2002), Denmark (Thrane 2013) and Sweden (De Jung 2006). Although audio subtitling can be delivered by a human voice, especially in combination with audio description (Braun and Orero 2010, Benecke 2012, Remael 2012), a synthetic voice is generally used when implemented independently from audio description in live content. Thrane (2013) looks in more detail into this rather unknown modality, describing the various productions systems and reporting on various experiments carried out with a sample of 16 adults. Her aim was to find out the main barriers adults find when using spoken subtitles, to elucidate whether different genres (news, documentaries, and fiction) imply different difficulties, and to get feed-back from users. Her results indicate that the main barriers found by users in spoken subtitles are related to synchronization issues, pronunciation, the presence of multiple voices, speed and split sentences, and that spoken subtitles receive a poorer user evaluation in fiction than in news and documentaries.

Beyond the field of audiovisual translation, text-to-speech systems applied to audiovisual content, not always including a translation process, have also been researched in various projects but they are beyond the scope of this paper (see, for example, Alías et al. 2011).

Methodological aspects

This section describes the participants' profiles, materials used, test development, and analysis.

Participants

Sixteen participants, aged 21-29 years old (mean age= 26), took part in the experiment. They were all Spanish native-speaker volunteers, both undergraduate (4) and graduate students (12). None reported having any uncorrected vision or hearing impairments, and none had previously seen the excerpts under analysis. All of them reported watching a maximum of one wildlife documentary per month.

In terms of audiovisual transfer modes, they reported the watching habits included in Table 1, which shows a prevalence of subtitling (75% replied "frequently" or "quite frequently"), which is in line with a growing tendency in younger generations to use subtitled

content (Matamala et al. 2017), contrary to what used to be standard practice in Spain, a traditionally dubbing country.

	Very frequently	Frequently	Quiet frequently	Occasionally	Rarely	Very Rarely	Never
Dubbing	12.5%	25%		25%	12.5%		25%
Subtitling		50%	25%	12.5%	12.5%		
VO		12.5%	12.5%	25%	25%		25%

Table 1 *Watching habits of participants*

In terms of audiovisual transfer mode preferences in voiced-over documentaries, there is a high variability among participants. To the statement "I'd rather watch voiced-over documentaries than subtitled documentaries", 12.5% strongly agreed and the same percentage strongly disagreed; 12.5% agreed and 12.5% somewhat agreed with the statement, whilst 25% neither agreed nor disagreed, and 25% disagreed. When the comparison is with dubbing ("I'd rather watch voiced-over documentaries than dubbed documentaries"), the response was less variable but showed opposing trends: 50% strongly agreed, whilst 25% strongly disagreed and 25% disagreed.

Materials

The materials used were two self-contained video excerpts in English of a 7-minute wildlife documentary film entitled *Must Watch: a Lioness Adopts a Baby Antelope*, currently available on Youtube as an independent video (https://www.youtube.com/watch?v=mZw-1BfHFKM). These excerpts are part of the episode *Odd Couples* from the series *Unlikely Animal Friends* (National Geographic 2009). They are both similar in terms of length (1:41 minutes versus 1:52 minutes), number of words (283 versus 287), speakers (the same two experts and a narrator appear in both), and segments of speech (8 versus 9). They both feature a male narrator and two experts, a male and a female, talking to the camera.

For each clip two versions were created in Spanish: one version with only human voices in the target language, and one version with only synthetic voices in the target language. The text-to-speech system used was developed by Verbio and the voices chosen were "Laura" for the female expert, "Carlos" for the male narrator, and "Javier" for the male expert. Human voices were selected by a professional dubbing studio.

Questionnaire design

A pre-questionnaire gathered data about the participants' profiles, including information about their age, mother tongue, educational level, vision or hearing impairments, watching habits, and preferences regarding audiovisual transfer modes.

A first post-questionnaire (PQ1) was developed for first-time viewings. It included five open comprehension questions that could be answered with short replies. Correct replies were given 1 point, incorrect replies were given 0 points, and partially correct answers scored 0.5 points, totaling a maximum of 5 points. This first post-questionnaire also aimed to gather opinions from participants in terms of self-reported interest, engagement, and enjoyment. Participants had to report their level of agreement on a 7-point Likert scale with the following statements:

1. The excerpt was interesting.
2. I will look into more information about the unusual couple presented in the documentary.
3. I lost the notion of time while I was watching the excerpt
4. I followed the excerpt actively.
5. I paid more attention to the excerpt than to my own thoughts.
6. I enjoyed watching the excerpt.
7. If the documentary were to be voiced with these voices, I would not watch it.

Next, they were asked to rate the voices heard on a 7-point Likert scale in terms of quality, naturalness, and comprehensibility. Despite the existence of established questionnaires for assessing synthetic voices in which extensive lists of items are evaluated (see Fernández-Torné and Matamala 2015 for an overview), a shorter version focusing on only those three items was prioritized.

Then, they were asked about preferences. A first question asked whether they had liked all voices equally and, when a negative answer was given, they were requested to order the voices according to their preferences. A second question asked if they thought all voices were human (the possible answers being "yes", "no", and "I don't know") and, if the reply was negative, participants were asked to indicate which one(s) they thought were human. They were also asked if they thought synthetic voices could be used to voice documentaries. In both cases the answers to be chosen were "yes", "no", and "I don't know", and there was an open field to explain their choice. Overall, the first post-questionnaire included questions on the clip (comprehension; self-reported interest, engagement, and enjoyment) and on the voices (quality, naturalness, comprehensibility, preferences, voice identification).

A second post-questionnaire (PQ2) was also developed for second-time viewings. It replicated the questions in PQ1 with two exceptions: first of all, it excluded comprehension questions as it was considered a second viewing would definitely increase understanding. Secondly, it included additional questions regarding participants' preferences after watching both clips. In this regard, they were asked to indicate their preferred version for each excerpt, without knowing which one used human or synthetic voices. Finally, they were explicitly asked whether, according to them, both versions could be broadcast on television.

Test development and analysis

Participants were received in a computer lab individually and, after filling in information and consent forms approved by UAB Ethics Committee, the pre-questionnaire was administered. Information about the audiovisual content context was given to them and they were requested to watch one excerpt. PQ1 was then administered to them. Next, they watched the same excerpt with different voices, and PQ2 was given to them. The same procedure was followed for the second excerpt. The order of the excerpts, and type of voices, were randomized and balanced across participants (excerpt 1/excerpt 2, human/synthetic voices), who did not know which versions were watching.

Results and discussion

Results are presented differentiating between aspects dealing first with the clips and then with questions addressing directly the voices.

Understanding and enjoying the audiovisual excerpts

Concerning comprehension, which was assessed only in the first viewing, Table 2 presents the data, where 5 would be 100% comprehension. Results show that human voices are slightly better understood, especially in the first clip. One can also observe that the second clip is not understood as well as the first one.

	E1-H	E1-S	E2-H	E2-S
Median	4.75	3.8	3.2	3
Mean	4.63	3.38	3.38	3.25

Table 2 *Comprehension levels (E= excerpt, H=human, S=synthetic)*

However, it is worth stressing that two participants had a very poor comprehension of E1-S, which impacted negatively in the results. It remains to be seen in future experiments with bigger samples whether the fact that human voices are better understood than synthetic voices was indeed caused by the usage of text-to-speech systems or was more related to the participants' profile and/or the excerpt characteristics.

When participants were asked directly about their interest in the clips, results show that human-voiced excerpts were considered slightly more interesting than their TTS counterparts in excerpt 1 but not in excerpt 2, where the synthetic version got slightly better assessments. However, when the question about interest was not so directly formulated, they tended to show a very low interest in general. Still, no significant differences between the TTS and the human-voiced documentaries were found. Table 3 summarizes the results, and highlights the median values in bold. When the median is between two nominal values, both are highlighted.

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somew hat agree	Agree	Strongly agree
"The excerpt was interesting"							
E1-H	0%	0%	12.5%	25%	**12.5%**	**25%**	25%
E1-S	12.5%	12.5%	12.5%	**25%**	12.5%	12.5%	12.5%
E2-H	0%	0%	25%	0%	**25%**	**25%**	25%
E2-S	0%	0%	0%	28.6%	14.3%	**42.9%**	14.3%
"I will look into more information about the unusual couple presented in the documentary"							
E1-H	37.5%	**12.5%**	**12.5%**	12.5%	12.5%	12.5%	12.5%
E1-S	37.5%	**12.5%**	**12.5%**	12.5%	12.5%	12.5%	12.5%
E2-H	25%	**25%**	**12.5%**	12.5%	0%	0%	25%
E2-S	14.3%	**42.9%**	28.6%	0%	14.3%	0%	0%

Table 3 *Interest in excerpts*

When asked about their engagement with the content, results (Table 4) showed that almost no differences were found depending on the voices used. A small difference in terms of excerpts

can be observed, because the second one obtained slightly better results, but the voice selection did not seem to impact on the results.

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
"I lost the notion of time while I was watching the excerpt"							
E1-H	0%	0%	12.5%	**37.5%**	**25%**	25%	0%
E1-S	0%	12.5%	25%	**37.5%**	12.5%	12.5%	0%
E2-H	0%	25%	0%	**25%**	**12.5%**	12.5%	25%
E2-S	0%	0%	0%	42.9%	**14.3%**	28.6%	14.3%
"I followed the excerpt actively"							
E1-H	0%	0%	25%	0%	**25%**	**25%**	25%
E1-S	0%	0%	12.5%	25%	**25%**	25%	12.5%
E2-H	0%	0%	12.5%	0%	**37.5%**	**25%**	12.5%
E2-S	0%	0%	0%	14.3%	28.6%	**28.6 %**	28.6%
"I paid more attention to the excerpt than to my own thoughts"							
E1-H	0%	0%	25%	12.5%	**25%**	12.5%	25%
E1-S	0%	0%	25%	0%	**50%**	12.5%	12.5%
E2-H	0%	0%	12.5%	0%	**50%**	12.5%	25%
E2-S	0%	0%	0%	14.3%	28.6%	**28.6 %**	28.6%

Table 4 *Self-reported engagement and attention*

It must be noted that one clear limitation of this experiment is that excerpts are short, hence it can be difficult for users to engage with the content and lose the notion of time, as asked for instance in the first question. However, the conditions are equal for both human-voiced and TTS-voiced documentaries, and what interests us is not the engagement felt but the comparison between the conditions under analysis.

Self-reported enjoyment was another aspect under analysis, and it was assessed through one statement (see Table 5). Results in this case show lower values for TTS documentaries. Whilst they "somewhat agree" or "agree" with the statement stating they have enjoyed watching the excerpt in the human-voiced versions, they "somewhat disagree" or remain neutral in the TTS versions.

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
"I enjoyed watching the excerpt"							
E1-H	0%	0%	12.5%	12.5%	**25%**	**37.5%**	12.5%
E1-S	12.5%	25%	**25%**	12.5%	12.5%	12.5%	12.5%
E2-H	0%	12.5%	0%	25%	**37.5%**	12.5%	12.5%
E2-S	14.3%	0%	0%	**42.9%**	28.6%	14.3%	0%

Table 5 *Self-reported enjoyment*

Overall, one can observe slightly better values for the human voices, especially in terms of comprehension and enjoyment, but results are not consistent across excerpts and are not strikingly different.

Opinions about the voices

When explicitly asked whether they would be willing to watch the whole documentary with these voices, results show a considerable difference between TTS documentaries and human-voiced ones: while they tend to say that they would not watch it with TTS voices (50% "strongly agree" in the first excerpt and 42.9% "somewhat agree" in the second), the opposite trend is shown in the excerpts voiced by human professionals, as seen in Table 6.

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
"If the whole documentary was voiced with these voices, I would not watch it"							
E1-H	12.5%	**50%**	25%	12.5%	0%	0%	0%
E1-S	12.5%	25%	0%	**12.5%**	0%	0%	**50%**
E2-H	25%	**37.5%**	0%	37.5%	0%	0%	0%
E2-S	14.3%	14.3%	14.3%	0%	**42.9%**	0%	14.3%

Table 6 *Voice acceptance values*

In trying to find an explanation to these replies, a set of questions was directly concerned with the voices, differentiating between the female expert voice, the narrator (also a male) and the male expert voice. Issues under analysis dealt with overall quality, naturalness, and comprehensibility of each voice. Tables with the results are presented as an annex to make the article more readable, but are discussed next.

Regarding the quality of the voices, results show that natural voices are generally considered "good" or "pretty good", whilst synthetic voices received worse assessments. However, there were remarkable differences among the synthetic voices: while the expert female voice was not assessed positively, with a median between "bad" or "pretty bad", both the narrator and the male expert voice were assessed as "pretty good" in the second excerpt.

Concerning the naturalness of the voices, human voices were assessed quite positively. However, although being natural voices, they did not reach the best possible mark in median values, but were generally assessed as "good" or "pretty good". On the contrary, synthetic voices were assessed with lower marks but with interesting differences: for instance, the same female voice was assessed with a median between "very bad" and "bad" in one excerpt, but it was only considered "pretty bad" in excerpt 2. Similarly, the narrator's voice was considered also "bad" in excerpt 1 but "neither bad nor good" in excerpt 2. The same trend is observed in the male expert's voice, where the voice was considered "pretty bad" in excerpt 1, but "pretty good" in excerpt two. When looking at the excerpts trying to find a reason for this divergence, we cannot see any relevant difference that could explain the results obtained. Apart from this pattern related to the excerpts, it can also be observed that the naturalness of the artificial voices is assessed differently, with the female voice receiving low values and the male voices obtaining higher values across excerpts.

In terms of comprehensibility, no median values were below average: human voices were assessed mainly as “good” or “very good”, while synthetic voices presented greater variability but still high median values.

Overall, it seems that human voices are better assessed, but synthetic voices get acceptable evaluations in terms of comprehensibility. In terms of naturalness and overall quality, the values are lower, with variability among voices. Indeed, when asked whether they liked all voices equally, it comes as no surprise that the reply was “no” in 100% of the participants after watching the first excerpt with synthetic voices, and 71.4% after watching the second one. When asked to order their synthetic voices according to their preferences, results show a preference for the voice of the narrator in clip 1, which was chosen as the preferred voice by 62.5%, and for both the narrator and the male expert voice in clip 2 (40% narrator, 40% male voice). This shows that among synthetic voices there are uneven qualities and their selection can have a direct impact on the results.

When asked whether they liked all voices equally in the human-voiced excerpts, 25% of participants replied “no” in both excerpts, with diverging results in terms of their preferred voices. This indicates that, even in human voices of the same professional standard, there are differences in terms of preferences.

When asked whether they believed all voices were human, it is surprising to observe that the reply was not as straightforward as one might expect. When the clip contained only human voices, most participants gave the correct answer (62.5%), but some still expressed doubts (12.5% in clip 1 and 37.5% in clip 2) and others directly got it wrong (25% in clip 1). When the excerpt only contained synthetic voices, 100% correctly identified it in the second excerpt, but replies were multiple for the first one: although the vast majority (75%) rightly identified the voices as non-human, 12.5% thought they were human, and 12.5% expressed their uncertainty. A more detailed analysis per voice shows that the female artificial voice is the one that is more clearly identified as synthetic (85.7% of participants in the first excerpts and 71.4% in the second), or human (100% in clip 1), which is in line with preferences expressed before by participants. Conversely, the male voices under analysis generated more doubts.

When specifically asked whether they think TTS voices could be used in documentaries, opinions are quite divided and seem to be influenced by the clip: in clip 1, 37.5% think TTS voices could be used, 43.75% think they could not be used, and the rest do not know. In clip 2, the percentages are the opposite: 46.45% think TTS voices could be used, 40.2% think they could not be used, and the rest do not know.

An open question gathered qualitative replies that, although not numerically significant, shed more light on participants’ views. For instance, one participant thought synthetic voices impair comprehension whilst another one stated that they are uncomfortable to listen to. Another added that they sound “robotic” and therefore the viewers detach themselves from the documentary. Another participant simply stated that quality would be worse with TTS VO, an opinion not shared by another informant, who thought that in this way comprehension could be improved, or another, who stated that it highly depends on the synthetic voice. In this regard, this informant liked the narrator’s voice but did not like the female voice, a trend that seems to be shared by other participants according to the data presented before. Another volunteer, after listening to synthetic voices without knowing they were synthetic, stated that if the ones they had just heard were synthetic, they could be used, showing the potential of this technology.

Finally, participants were asked which version they preferred and all but one participant when watching excerpt 2 selected the human version. However, when asked if they believed both versions could be broadcast on television, results are not so direct: 50% gave a positive reply and 50% gave a negative reply in excerpt one, while the percentage of positive replies rose to 57.1% for the second excerpt. This gives an average of 53.11% of participants thinking this could be broadcast, which indirectly hints at its acceptability for television.

Conclusions

To sum up, this article has presented the results of an exploratory experiment in which excerpts of wildlife documentaries voiced by human professionals and by synthetic voices were compared using various parameters, in order to assess whether what we have termed TTS VO could be broadcast on television.

Overall, human voices get better values, especially in terms of comprehension and self-reported enjoyment, but in other parameters such as self-reported engagement differences are almost non-existent. It seems that our participants would not be willing to watch a whole documentary voiced with speech synthesis, but even so more than half of them consider that the excerpts could be broadcast on television. And, surprisingly enough, some of them do not correctly identify text-to-speech voices as such.

Our analysis also shows that comprehensibility, tightly linked to intelligibility, is no longer an issue of TTS but naturalness and overall quality may be. Indeed, naturalness, tightly linked to emotions, is a relevant area of research in speech synthesis and one that is expected to have an impact in the field in the short-term. It is also worth mentioning that the three artificial voices used in the experiment are also assessed differently, which shows the impact of voice selection on the assessment. Undoubtedly, there is also an issue of personal preferences, which implies that even human voices, which are "natural" per se, are not considered natural enough by some of the participants.

Another aspect observed in our experiment is that, even in balanced excerpts, which have been presented to participants in a random order, there are differences. Extrapolating from this fact the high variability in terms of speakers, contents, and registers of wildlife documentaries, one could expect to get diverging feed-back from users depending on the content features, making it a challenge to select the type of content that would be more suitable for TTS.

Finally, when participants are asked whether TTS VO could be used in documentaries, results show opposing views among participants, with an average of 42% in favour, 42% against, and 16% who simply do not know.

Our research has provided an innovative approach to the topic of voice-over and the translation of documentaries, where studies on automatization are scarce. However, many questions remain open after this research. It is obvious that human voices are preferred but the reaction towards TTS voices is not one of total rejection. Future advances in speech synthesis may yield better results and open new possibilities in the field. Additionally, this could be seen as a solution in environments where human voicing is not possible. Beyond traditional broadcast television, many other platforms provide audiovisual content, created by both professionals and amateurs. Further research with wider samples of participants, other

languages, voices and content, and also longer excerpts which allow for a thorough statistical analysis should be carried out to provide more conclusive results.

Acknowledgements

This research was supported by the Spanish Ministerio de Economía y Competitividad funds (reference code FFI-2012-31024) and by Catalan government funds (2014SGR0027, 2017 resolution pending). We would like to thank the volunteers that took part in the experiment. Special thanks are due to Verbio for providing the voices used in the experiment free of charge. We are also grateful to ECAD (Escola Catalana de Doblatge) for recording the stimuli with high professional standards of quality.

References

ALÍAS, Francesc, IRIONDO, Ignasi, SOCORÓ, Joan Claudi. 2011. Aplicació de tècniques de generació automàtica de la parla en producció audiovisual. In *Quaderns del CAC*, vol. 37, no. 1, pp. 105-114.

BENECKE, Bernd. 2012. Audio description and audio subtitling in a dubbing country: Case studies. In PEREGO, Elisa (ed.) *Emerging Topics in translation: audio description*. Trieste: EUT, 2012, pp. 99-104.

BRAUN, Sabine, ORERO, Pilar. 2010. Audio description with audio subtitling – an emergent modality of audiovisual localization. In: *Perspectives. Studies in* Translatology, vol. 18, no. 3, pp. 173-188.

CHAUME, Frederic. 2012. *Audiovisual translation: Dubbing*. Manchester: St. Jerome, 2012.

CRYER, Heather, HOME, Sarah, MORLEY WILKINS, Sarah. 2010. *Synthetic speech evaluation protocol*. *Technical report #7*. Birmingham: RNIB Centre for Accessible Information (CAI), 2010.

DE JONG, Frans. 2006. Access devices for digital television. In PÉREZ-AMAT, Ricardo, PÉREZ-UGENA, Álvaro (eds) *Sociedad, integración y televisión en España*. Madrid, Spain: Laberinto Comunicación, 2006, pp. 331-344.

DE SOUSA, Sheila C., AZIZ, Wilker, SPECIA, Lucia. 2011. Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. *Proceedigs of the International Conference Recent Advances in Natural Language Processing*, 2011, pp. 97-103.

DELGADO, Héctor, MATAMALA, Anna, SERRANO, Javier. 2015. Speaker diarization and speech recognition in the semi-automatization of audio description: an exploratory study on future possibilities. In *Cadernos de Traduçao*, vol. 35, no. 2, pp. 308-324.

DEL POZO, Arantza. 2014. *SUMAT final report* [online]. 2014 [cit. 2017-12-19]. Available at: <http://www.sumat-project.eu/uploads/2014/07/D1-5_Final-Report-June-2014.pdf>

DROZDZ-KUBIK, Justyna. 2011. Harry Potter i Kamień Filozoficzny słowem malowany – czyli badanie odbioru filmu z audiodeskrypcją z syntezą mowy. MA Thesis. Krakow: Jagiellonian University, 2011.

FERNÁNDEZ-TORNÉ, Anna, MATAMALA, Anna. 2015. Text-to-Speech vs Human Voiced Audio Descriptions: A Reception Study in Films Dubbed into Catalan. In *The Journal of Specialised Translation* [online]. 2015, vol. 24 [cit. 2017-12-19], pp. 61-88. Available at: <http://www.jostrans.org/issue24/art_fernandez.pdf>.

FERNÁNDEZ-TORNÉ, Anna, MATAMALA, Anna. 2016. Machine translation in audio description? Comparing creation, translation and post-editing efforts. In *SKASE. Journal of translation and interpretation*, vol. 9, no. 1, pp. 64-87.

FRANCO, Eliana, MATAMALA, Anna, ORERO, Pilar. 2010. *Voice-over translation: an overview*. Bern: Peter Lang, 2010.

HINTERLEINER, Florian, NEITZEL, Georgina, MÖLLER, Sebastian, NORRENBROCK, Christoph. 2011. An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks. In *Proceedings of the Blizzard Challenge Workshop*. International Speech Communication Association (ISCA), 2011.

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, Switzerland: ITU [cit. 2017-12-19]. Available at: http: //www.itu.int/rec/T-REC-P.85-199406-I/en

KOBAYASHI, Masatomo, O'CONNELL, Trisha, GOULD, Bryan, TAKAGI, Hironobu, ASAKAWA, Chieko. 2010. Are synthesised video descriptions acceptable? In *ASSETS '10: Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility*. New York, USA: ACM, 2010, pp. 163-170.

MAZYNSKA, Magdalena. 2011. *TTS AD with audio subtitling to a non-fiction film. A case study based on* La Soufriere *by Werner Herzog*. MA Thesis. Warsaw: University of Warsaw, 2011.

MASZEROWSKA, Anna, MATAMALA, Anna, ORERO, Pilar (eds). 2014. *Audio description. New perspectives illustrated*. Amsterdam: Benjamin, 2014.

MATAMALA, Anna. 2010. Translations for dubbing as dynamic texts: strategies in film synchronisation. In *Babel*, vol. 56, no. 2, pp. 101-118.

MATAMALA, Anna. 2016. The ALST project: technologies for audio description. In MATAMALA, Anna, ORERO, Pilar (eds). *Researching audio description. New approaches*. London: Palgrave Macmillan, 2016, pp. 269-284.

MATAMALA, Anna. forthcoming. Voice-over: practice, research, and future possibilities. In PÉREZ-GONZÁLEZ, Luis (ed.) *The Routledge Handbook of Audiovisual Translation*. London: Routledge.

MATAMALA, Anna, PEREGO, Elisa, BOTTIROLLI, Sara. 2017. Dubbing versus subtitling yet again? An empirical study on user comprehension and preferences in Spain. In *Babel*, vol. 63, no. 3, pp. 423-441.

MATAMALA, Anna, ROMERO-FRESCO, Pablo, DANILUK, Lukasz. 2017. The use of respeaking for the transcription of non-fictional genres: an exploratory study. In *Intralinea*, vol. 19.

ORTIZ-BOIX, Carla, MATAMALA, Anna. 2015. Post-editing wildlife documentary films: a new possible scenario? In *The Journal of Specialised Translation*, vol. 26, pp. 187-210.

ORTIZ-BOIX, Carla, MATAMALA, Anna. 2017. Assessing the quality of post-edited wildlife documentaries. In *Perspectives. Studies in Translatology*, vol. 25, no. 4, pp. 571-593.

REHM, George, and USZKOREIT, Hans (eds). 2012. *Strategic Research Agenda for Multilingual Europe*. Berlin: Springer, 2012.

REMAEL, Aline 2012. Audio description with audio subtitling for Dutch multilingual films: manipulating textual cohesion on different levels. In *Meta*, vol. 57, no. 2, pp. 385-407.

SZARKOWSKA, Agnieszka. 2011. Text-to-speech audio description: towards wider availability of AD. In *The Journal of Specialised Translation*, vol. 15, pp. 142-162.

SZARKOWSKA, Agnieszka, JANKOWSKA, Anna. 2012. Text-to-speech audio description of voice-over films. A case study of audio described *Volver* in Polish. In PEREGO, Elisa (ed.) *Emerging topics in translation: audio description*. Trieste: EUT, 2012, pp. 81-98.

THRANE, Lisbeth Kvistholm. 2013. *Text-to-speech on Digital TV An Exploratory Study of Spoken Subtitles on DR1Syn*. MA Thesis. Copenhaguen: University of Copenhagen, 2013.

VERBOOM, Maarten, CROMBIE, David, DIJK, Evelien, THEUNISZ, Mildred. 2002. Spoken subtitles: making subtitled TV programmes accessible. In MIESENBERGER, Klaus, KLAUS, Joachim, ZAGLER, Wolfgang L. (eds) *Proceedings of Computers Helping People with Special Needs, 8th International Conference, ICCHP 2002*. Berlin-Heidelberg, Germany: Springer-Verlag, 2002, pp. 295-302.

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

VOLK, Martin. 2008. The automatic translation of film subtitles. A machine translation success story. In *Journal for Language Technology and Computational Linguistics*, vol. 23, no. 2, pp. 113-125

WALCZAK, Agnieszka, SZARKOWSKA, Agnieszka. 2012. Text-to-speech audio description of educational materials for visually impaired children. In BRUTI, Silvia, DI GIOVANNI, Elena (eds). *Audio Visual Translation across Europe: An Ever-Changing Landscape*. Bern/Berlin: Peter Lang, 2012, pp. 209-234.

Annex

	Very bad	Bad	Pretty bad	Neither bard nor good	Pretty good	Good	Very good
Female voice							
E1-H	0%	0%	0%	0%	37.5%	**37.5%**	25%

E1-S	37.5%	**12.5%**	**12.5%**	0%	25%	12.5%	0%
E2-H	0%	0%	0%	0%	25%	**37.5%**	37.5%
E2-S	14.3%	14.3%	**28.6%**	14.3%	28.6%	0%	0%
Narrator							
E1-H	0%	0%	0%	12.5%	**37.5%**	**37.5%**	12.5%
E1-S	37.5%	0%	**25%**	0%	25%	12.5%	0%
E2-H	0%	0%	12.5%	0%	0%	**62.5%**	25%
E2-S	0%	28.6%	14.3%	0%	**42.9%**	14.3%	0%
Male voice							
E1-H	0%	0%	0%	25%	**25%**	**37.5%**	12.5%
E1-S	37.5%	0%	**12.5%**	**25%**	12.5%	12.5%	0%
E2-H	0%	0%	12.5%	0%	12.5%	**50%**	25%
E2-S	0%	28.6%	14.3%	0%	**42.9%**	14.3%	0%

Table 7 *Overall quality of the voices*

	Very bad	Bad	Pretty bad	Neither bad nor good	Pretty good	Good	Very good
Female voice							
E1-H	0%	0%	0%	0%	37.5%	**50%**	12.5%
E1-S	**50%**	**12.5%**	0%	25%	0%	12.5%	0%
E2-H	0%	0%	12.5%	12.5%	12.5%	**37.5%**	25%
E2-S	42.9%	0%	**28.6%**	0%	28.6%	0%	0%
Narrator							
E1-H	0%	0%	0%	12.5%	**62.5%**	25%	0%
E1-S	37.5%	**25%**	0%	12.5%	25%	0%	0%
E2-H	0%	0%	25%	0%	12.5%	**37.5%**	25%
E2-S	0%	37.5%	0%	**25%**	25%	12.5%	0%
Male voice							
E1-H	0%	0%	0%	37.5%	**37.5%**	25%	0%
E1-S	25%	12.5%	**12.5%**	**37.5%**	0%	12.5%	0%
E2-H	0%	0%	12.5%	12.5%	12.5%	**37.5%**	25%
E2-S	0%	0%	0%	28.6%	**42.9%**	14.3%	14.3%

Table 8 *Naturalness of the voices*

	Very bad	Bad	Pretty bad	Neither bad nor good	Pretty good	Good	Very good
Female voice							
E1-H	0%	0%	0%	0%	25%	**25%**	**50%**
E1-S	37.5%	0%	0%	**12.5%**	**0%**	**37.5%**	12.5%
E2-H	0%	12.5%	0%	0%	0%	**37.5%**	**50%**
E2-S	0%	0%	14.3%	14.3%	14.3%	0%	**57.1%**
Narrator							

E1-H	0%	0%	0%	0%	25%	**25%**	**50%**
E1-S	0%	12.5%	0%	**12.5%**	**25%**	12.5%	12.5%
E2-H	0%	0%	0%	12.5%	0%	**37.5%**	**50%**
E2-S	0%	0%	0%	14.3%	28.6%	0%	**57.1%**
Male voice							
E1-H	0%	0%	0%	0%	25%	**25%**	**50%**
E1-S	12.5%	0%	0%	**37.5%**	**12.5%**	25%	12.5%
E2-H	0%	0%	0%	12.5%	0%	**37.5%**	**50%**
E2-S	0%	0%	0%	28.6%	14.3%	0%	**57.1%**

Table 9 *Comprehensibility of the voices*

Corresponding author: Anna Matamala
Universitat Autònoma de Barcelona
Edifici K-1002
08193 Bellaterra
e-mail: anna.matamala@uab.cat

In SKASE Journal of Translation and Interpretation [online]. 2018, vol. 11, no. 1 [cit. 2018-21-07]. Available online <http://www.skase.sk/Volumes/JTI14/pdf_doc/02.pdf>. ISSN 1336-7811