Text-to-speech vs. human voiced audio descriptions: a reception study in films dubbed into Catalan

Anna Fernández-Torné, Universitat Autònoma de Barcelona
Anna Matamala, Universitat Autònoma de Barcelona

ABSTRACT

This article presents an experiment that aims to determine whether blind and visually impaired people would accept the implementation of text-to-speech in the audio description of dubbed feature films in the Catalan context. A user study was conducted with 67 blind and partially sighted people who assessed two synthetic voices when applied to audio description, as compared to two natural voices. All of the voices had been previously selected in a preliminary test. The analysis of the data (both quantitative and qualitative) concludes that most participants accept Catalan text-to-speech audio description as an alternative solution to the standard human-voiced audio description. However, natural voices obtain statistically higher scores than synthetic voices and are still the preferred solution.

KEYWORDS

Accessibility, audio description, audiovisual translation, text-to-speech, speech synthesis, Catalan language, blind, visually impaired.

1. Introduction

Accessibility has become a major concern in society in recent decades, and laws are being enforced to guarantee disabled people's rights. Present legislation states that sensorial accessibility to audiovisual content should be provided: theatres, museums, TV broadcasters and web designers, among others, are endeavouring to make their content accessible to persons with disabilities and to comply with regulations.

For users who are blind or visually impaired, audio description (AD) allows access to visual information (images) appearing on screen, which they would otherwise miss. Audio description can be defined as an inter-semiotic translation in which the visuals are transferred into words that are received aurally by end users (Orero 2007; Orero and Matamala 2007). In films these oral descriptions are inserted in the silent gaps in the dialogue, i.e. when characters are not talking, and create a coherent whole with the film dialogues and soundtrack (Braun 2011). However, because creating and voicing an audio description is a time-consuming and costly process, this access service is not as widely available as one might expect. This is especially striking in social media environments, but also in other traditional broadcasting contexts.

In view of the need for wider availability of audio described audiovisual products, research on technological processes which fully or partially automate the audio description workflow is considered relevant, from a

scientific, social and economic point of view. Within this general framework, this article aims to present the results of research in which Catalan audio description using text-to-speech (TTS) software was assessed, and compared to standard human-voiced audio descriptions. Our final aim was to find out whether TTS AD in Catalan would be accepted by blind and visually impaired patrons as an alternative solution and to compare the scores attributed to both natural and artificial voices on key aspects. The project's novelty lies in the language under analysis (Catalan). In addition, the methodological approach is also new in comparison with existing text-to-speech audio description tests: on the one hand, it provides a detailed analysis of many features instead of asking about general opinions or perceptions; secondly, it assesses text-to-speech audio description against human-voiced audio description instead of evaluating it in isolation, as further explained in section 3.2.

The article presents, first of all, a review of related work, focussing on text-to-speech audio description but also widening the scope to present other text-to-speech applications in audiovisual translation and media accessibility (section 2). Methodological aspects are detailed in section 3, and results are discussed in section 4. Conclusions and possibilities for further research close the article.

2.   Text-to-speech audio description: an overview

Blind and visually impaired people use text-to-speech in many contexts, and its usefulness has already been proved in different domains. Cryer and Home (2008) analyse the use of synthetic speech technology by blind and partially sighted people. Inspired by Freitas and Kouroupetroglou (2008), they list the many areas in which speech technologies can be used: mobility aids (for instance, GPS navigation devices), educational tools (talking dictionaries, audio textbooks), entertainment (audio subtitles (AST), speaking electronic programming guides) and communication (screen reading software on computers). Speech synthesis seems to offer quicker access to information (Llisterri *et al.* 1993) and guarantees independence of the user (González García 2004), among other aspects. Cryer and Home (2008) point out two relevant research results of their overview of text-to-speech usage by blind and partially sighted people: firstly, the direct impact of each user experience on the acceptance of synthetic speech, as people gradually get used to synthetic voices, and, secondly, the impact of the naturalness and the context where the artificial voice is being used.

As for text-to-speech audio description (TTS AD), it has been researched within a project developed at the University of Warsaw, Poland, aiming to assess its feasibility and its reception among visually impaired people. Szarkowska (2011: 144) states that "instead of recording a human voice reading out the AD script, TTS AD can be read by speech synthesis

software". This guarantees the cost-effectiveness of the AD production in comparison with traditional methods of AD production.

The project analysed the application of TTS AD in several types of audiovisual products:

- in a monolingual feature film in Polish (Szarkowska 2011), where the artificial voice tested was Ewa (female voice), by Ivo Software;
- in a dubbed educational TV series for children (Walczak and Szarkowska 2010), where the artificial voice tested was Ewa (female voice), by Ivo Software;
- in a foreign fiction film, with voice-over (Szarkowska and Jankowska 2012), where the artificial voice tested was Krzysztof (male), by Loquendo;
- in a non-fiction film, with audio subtitling (Mączyńska 2011), where the artificial voices tested were Zosia (female voice) for the AD, and Krzysztof (male voice) for the AST, both by Loquendo;
- in a dubbed feature film (Drożdż-Kubik 2011), where the artificial voice tested was Ewa (female voice), by Ivo Software.

The number of participants ranged from 17 in Drożdż-Kubik (2011) to 76 in Walczak and Szarkowska (2010). The conclusions for each study were as follows: Szarkowska (2011) and Szarkowska and Jankowska (2012) stated that most respondents accepted TTS AD both as an interim solution and as a permanent option; Walczak and Szarkowska (2010) emphasised that most participants enjoyed the voice used in the test, and Mączyńska (2011) and Drożdż-Kubik (2011) explained that a majority of respondents found TTS AD acceptable, although it was not the preferred solution. Hence all five studies showed that most viewers accept TTS in AD.

On a similar note, and inspired by Chapdelaine and Gagnon's work (2009) on an accessible website platform for rendering different levels of audio description (as far as quantity and quality of AD is concerned) on demand, Kobayashi *et al.* (2009: 249) describe a "technique to use synthesized speech to add AD to online videos on any websites." The three steps of their project include determining whether or not synthesised voice can compete with real voices, designing a text-based format to describe the AD scripts, and developing authoring software. Step one is thoroughly explained in Kobayashi *et al.* (2010): 115 visually-impaired adult participants took part in an informal survey in Japan where three kinds of voice were tested (human, standard TTS, and prototype TTS). This first experiment was followed by an in-depth interview session with three participants. The study continued in the US, where 236 participants completed a survey, followed by an additional in-depth interview session with eight participants. A follow-up study with 24 participants closed the research. It included additional variables such as long vs short stimuli, expressive TTS technology vs standard TTS, expert vs novice descriptions, and standard vs extended descriptions. All in all, this broad study showed

that synthesised descriptions are generally accepted, especially for relatively short videos and informational content.

With their more experimental approach, Encelle *et al.* (2011) present an exploratory work on video accessibility for the blind and visually impaired with "audio enrichments composed of speech synthesis and earcons (i.e. nonverbal audio messages)" (123). Their study with 21 blind volunteers show that earcons associated with speech synthesis are useful for understanding set-related information, i.e. enriching videos with the use of earcons to complement speech synthesis helps convey visual information.

Moving from academia to industry, the firm Swiss TXT is already planning to offer audio description in which text-to-speech technologies are implemented (Caruso 2012). A web-based editor for transforming text into speech which can be used for audio description has also been developed by Mieskes and Martínez (2011). The editor contains features which allow the speaking rate and pitch to be set, as well as phonetic tuning functionalities. The described scenario would allow a user to upload an existing description or create a new one, upload the corresponding movie and synthesise the descriptions. Similarly, Oncins *et al.* (2013) have developed the Universal Accessibility System, a multi-language and multi-system mobile application to make live performing arts accessible. The system is designed to offer automatic AD through speech synthesis as well as other features (subtitling, spoken subtitles, an emergency pack, etc.).

Research on text-to-speech in audiovisual translation (AVT) goes beyond audio description and is especially relevant in a strongly related transfer mode: audio subtitling or spoken subtitles, where a synthetic voice is used to automatically read aloud the subtitles and make them accessible not only to blind and visually-impaired people, but also to people with reading difficulties. This service has been implemented in television broadcasts in countries such as the Netherlands (Verboom *et al.* 2002) and Sweden (De Jong 2006), where two digital boxes are needed to make it work. To expand the availability of spoken subtitles and avoid the need for a special decoder, a user-based device for reading aloud subtitles (Subpal) has been proposed by Nielsen and Bothe (2007), and a free and open-source tool has been developed by Derbring, Ljunglöf and Olsson (2012) within the SubTTS project (Derbring, Ljunglöf and Olsson 2010).

Finally, it is worth mentioning that, focusing exclusively on the language under analysis in this research, Alías, Iriondo and Socoró (2011) present the state of speech synthesis implementation in Catalonia which includes the most relevant companies, research centres and products relating to Catalan synthetic voice generation, and carry out field work to map the actual usage of text-to-speech in Catalan audiovisual media. In a specific

section of their article devoted to blind and visually impaired users, they point out that most of them think text-to-speech could be used in AD as long as more natural and expressive voices can be developed, although no specific quantitative data are given.

3.   Methodological aspects: materials and method

This section describes the participants involved in the current experiment, the voices used, the film and clip selection process, the evaluation questionnaires drafted, the actual development of the test, and the statistical methods used.

3.1. Participants

Since it was "impossible to map 'the population' from which a random sample" was to be taken (Bryman 2012: 416), an a priori generic purpose sampling strategy was adopted. Such a strategy implied the establishment of certain criteria for selecting participants at the outset of the research. A total of 67 persons participated in the test (55% female, 45% male). The mean age was 52, with ages ranging from 21 years old to 85 years old. Thirty-three participants (49%) were 50 or younger, the others being older (51%). A more detailed distribution of the participants is shown in Table 1:

	Women	Men	Total
From 20 to 34	5	4	9 (13%)
From 35 to 44	5	7	12 (18%)
From 45 to 54	7	8	15 (22%)
From 55 to 64	9	6	15 (22%)
From 65 to 74	9	4	13 (19%)
More than 80	2	1	3 (4%)
TOTAL	37 (55%)	30 (45%)	67

Table 1. Participants' distribution based on sex and age.

The age range was not limited to account for the whole spectrum of the adult population to which AD is offered. Additionally, acceptance of synthetic voices is often linked to their usage, and limiting the age range to younger or older participants would have probably had an effect on the results.

Using the World Health Organisation's classification of visual impairments (2013), 51% of the participants described their disability as blindness, whereas 49% declared it to be low vision, with visual impairment being from birth in 30 cases (45%).

As far as the participants' educational background is concerned, 51% reached at least first degree university level (Bachelor's degree or equivalent), whilst 24% did not reach the first stage of secondary school. 46% reported being unemployed, whilst 13% declared to be employed in clerical posts.

3.2. Voice selection

It was decided that a total of four voices (a male and a female artificial and a male and a female natural voice) would be included in the experiment to avoid any gender bias. In order to select them, a pre-test with 20 participants was carried out, as described in Matamala, Fernández-Torné and Ortiz-Boix (2013). Ten synthetic voices (see Table 2) and ten natural voices (both professional and non-professional voice artists selected by the Catalan School of Dubbing ECAD) were assessed by the participants (see Fernández-Torné and Matamala 2013 for further details on the methodological aspects of the pre-test).

Female	Male
Laia by Acapela	Jan by FestCat
Anna by iSpeech	Teo by FestCat
Meritxell by Verbio	Oriol by Verbio
Montserrat by Loquendo	Jordi by Loquendo
Ona by FestCAt	Pep by FestCat

Table 2. Artificial voices.

This pre-test allowed us to select the voices to be used in the experiment, namely a professional voice talent (a female natural voice), a non-professional but trained voice talent (a male natural voice), Laia by Acapela (a female synthetic voice), and Oriol by Verbio (a male synthetic voice).

3.3. Film and clip selection

The voices were tested in an audio described film excerpt. Various factors influenced the film selection process: first of all, this experiment is part of a wider project in which other technologies such as machine translation are to be tested in the English-Catalan language pair (Fernández-Torné, Matamala and Ortiz-Boix 2012). Therefore, a film which had already been audio described in Catalan (for the TTS AD experiments) and that had also been audio described in English (for the machine translation tests) was required. A dubbed fiction feature film or a children's animation film were the only options, as these were the only dubbed audiovisual products that were audio described in Catalan at the time the experiment took place.

Children's animation films were disregarded as our intended target audience were adults; hence a dubbed fiction film had to be selected.

Secondly, defining the specific genre was also considered relevant, since in TTS evaluation studies in other fields such as audiobooks, the text type has been shown to have a significant influence on the results. For instance, Hinterleitner *et al.* (2011) have proven that seven out of the 11 rating scales used in their study were influenced by the type of text when assessing the quality of the same synthetic voice. Our final decision was not to favour any particular film genre, and a film belonging to a "miscellaneous" category according to Salway *et al.*'s (2004) classification was chosen.

Finally, from a more practical point of view, it was considered that the availability of the English original script, the English AD script, the Catalan dubbed script, and the Catalan AD script would speed up the research process, and *Closer* (2004, directed by Mike Nichols) was selected. However, to limit the duration of the experiment, it was decided to carry out the experiment using short clips rather than the whole film, unlike the five studies within the TTS project developed at the University of Warsaw and the Jagiellonian University of Krakow (Szarkowska and Jankowska 2012).

As far as the clip selection was concerned, it was decided that two different clips, one clip for female voices and another one for male voices, would be chosen to minimis e fatigue and the impact of a learning effect on the subjects. Additionally, an in-depth analysis of the film, of the AD script and of the individual AD units was performed, in order to obtain two comparable clips in terms of content (neutral in both cases, with no potentially distracting and/or offensive content), length (3 minutes in clip 1 vs 3 minutes and 6 seconds in clip 2), intervening characters (Anna and Dan in both clips), background music (the same opera for both), and AD density (571 characters vs 537 characters respectively). Clips were randomly assigned a voice gender, either masculine or feminine, for the audio description.

3.4. Evaluation questionnaires

For the human assessment of synthetic voices, the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T) recommends using a Mean Opinion Score (MOS) test, by which listeners are asked to rate several systems taking into account various items (ITU Recommendation P.85 1994), hence this was our chosen approach. The items to be included in our questionnaire were selected after a thorough comparison of various tests in text-to-speech evaluation. These are:

- ITU Recommendation P.85 (1994), which includes seven 5-point scales and one 2-point (yes-no) scale;
- Viswanathan and Viswanathan (2005), who propose 11 items to be assessed on a 5-point scale;
- Cryer, Home and Morley Wilkins (2010), who suggest twelve 5-point scales; and
- Hinterleitner *et al.* (2011), who put forward an evaluation protocol for the assessment of TTS in audiobook reading tasks, concluding that eight scales out of the eleven they tested should be kept, with a continuous 7-point rating scale.

It was finally decided to limit the number of items and to focus on issues directly linked to end-user reception rather than on the intelligibility dimension, since intelligibility was taken for granted in the selected voices and was deemed more relevant for system performance testing. The final list of items included in our questionnaire is listed next, in the same order as they were presented to participants when given the instructions. Participants assessed each item on a 5-point scale.

Overall impression: a global score, the general opinion participants have of the voice of the audio description.

Accentuation: this score assesses whether the stress is put on the right syllable.

Pronunciation: measures to what extent words are correctly uttered according to Catalan phonetics.

Speech pauses: evaluates whether the voice stops when needed between sentence components and between sentences.

Intonation: assesses whether the pitch curve accurately represents the sentence type (whether it is a question, an exclamation or a declarative sentence).

Naturalness: in synthetic voices, this item assesses to which extent the voice resembles a human voice; in natural voices, it is related to the degree the human voice is forced and dramatised.

Pleasantness: conveys to what extent the listener finds the voice pleasant.

Listening effort: involves subjectively assessing whether listening to the voice for a long period of time would be tiring or tedious.

Acceptance: is used to indicate whether the voice is deemed adequate to voice audio descriptions.

It must be stressed that a careful translation into Catalan of each of the previous items, validated by a professional translator and tested in a pilot test, was carried out. It was also decided that heading descriptors were not to be used in the real test since the choice of an oral delivery mode for the test instead of a written one made the use of headings before a question quite awkward and it actually did not enhance comprehension. Therefore, participants were directly asked the questions and read aloud the 5 possible answers to each question preceded by their corresponding score: from least positive (1) to most positive (5) (see Annex 1 for the back translation into English of the actual Catalan questionnaire).

Regarding the order of the items, the overall impression and acceptance items were kept in the first and last positions respectively following the other tests. A logical order was proposed for the remaining scales, from more specific questions to broader ones: word-centered questions (accentuation and pronunciation), phrase-centered questions (speech pauses and intonation), voice-centered questions (naturalness and pleasantness) and a global question (listening effort).

As well as the questionnaire, a post-questionnaire was included, inspired by the works of Walczak (2010), Mączyńska (2011), Chmiel and Mazur (2012) and Pazos (2012). Its aim was to gather information on the participant demographics and to get more subjective information on personal preferences and usage of audio described audiovisual products and TTS applications in devices and/or computers. As in previous studies in the field (Walczak 2010, Mączyńska 2011), such questions were included in a post-questionnaire rather than in a pre-questionnaire. This decision was motivated by our wish to be as tactful as possible, trying not to ask potentially sensitive questions at the beginning of the test. The post-questionnaire, translated from Catalan into English, is included in Annex 2.

3.5. Procedure

Participants did the experiment on a one to one basis in a sound proof booth, following approved ethical procedures. Listening conditions were controlled: the stimuli were played with VLC Media Player and presented through professional headphones, Beats mixr by Dr. Dre. All participants were volunteers and listened to all stimuli, following a within-subjects design.

The experimental session was initially tested in a pilot test which was developed as follows: participants were given an overview of the project and the actual experiment, and were required to sign a Participant Information Sheet and Consent Form previously approved by the University Ethical Committee. They were instructed to assess each AD voice independently, and a thorough explanation of the nine items for

which they were to give ratings was provided (see previous section). A warm-up task using a voice that was not included in the actual experiment was also carried out.

The main experiment then started, and participants were asked to listen to the four voices, replicating always the same pattern: audio stimulus reproduction, 5-second pause, questions 1 to 9 read aloud by the researcher, oral reply by the participant that was written down by the researcher, and a final 3-second pause. The listening order of the voices was randomised across participants, always presenting the synthetic voices first to avoid a negative impact on the TTS system evaluation, as suggested by van Santen (1993), and Viswanathan and Viswanathan (2005: 62). This part of the experiment lasted 22 minutes and 36 seconds, and the test finished with the post-questionnaire, which was read aloud by the researcher, who would again write down the answers in the corresponding form.

3.6. Statistical methods

For the eight items to be considered (accentuation, pronunciation, speech pauses, intonation, naturalness, pleasantness, listening effort, acceptance) descriptive statistics (mean, median, standard deviation, minimum, maximum and percentiles) were calculated. Figures 2 and 3 display the means and the medians for all the items. A multinomial model was established for each item under analysis as the dependent variable and the type of voice as the independent variable. However, some of the items had very low frequencies in some of the categories, so they were recategorised as a binary outcome (scores 1, 2, and 3 were grouped under the category "low score", whereas scores 4 and 5 were grouped under the category "high score"). Then logistic regression models were used to assess the probability of obtaining a high score.

Overall impression was also analysed using a multinomial model, taking into account the voice, the gender, the age (categorised in under and above 50 to balance groups) and the disability type as independent variables.
All results were obtained using SAS, v 9.2 (SAS Institute Inc, USA). For the decisions, significance level was fixed at 0.05.
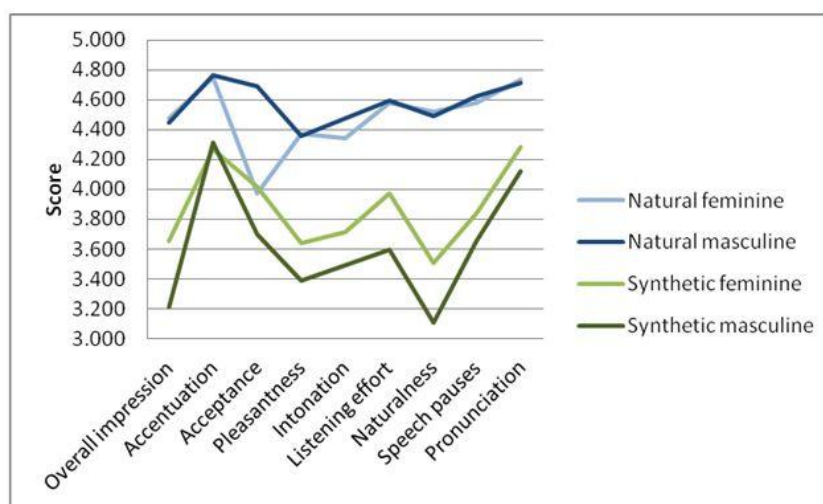
4.   Results and discussion

From the mean scores of the items (Figure 2) we notice that the natural male voice obtains higher scores in:
- accentuation (4.761, stdev=0.495),
- acceptance (4.687, stdev=0.583),
- intonation (4.478, stdev=0.682),
- listening effort (4.597, stdev=0.605), and

- speech pauses (4.627, stdev=0.624).
The natural feminine voice, by contrast, obtains higher scores in
- pleasantness (4.373, stdev=0.671),
- naturalness (4.522, stdev=0.725),
- overall impression (4.478, stdev=0.725), and
- pronunciation (4.731, stdev=0.479).

The lowest scores for natural voices are related to the female voice acceptance (3.970, stdev=0.244) and intonation (4.343, stdev=0.708). However, for the purposes of our study, what is especially interesting is not which voice gets higher scores on what items, but to observe that the results for the synthetic voices is close to that of natural voices, and that all the scores of the synthetic voices are above 3.1, reaching 4.313 in the accentuation of the synthetic male voice and 4.284 for the pronunciation of the feminine synthetic voice.
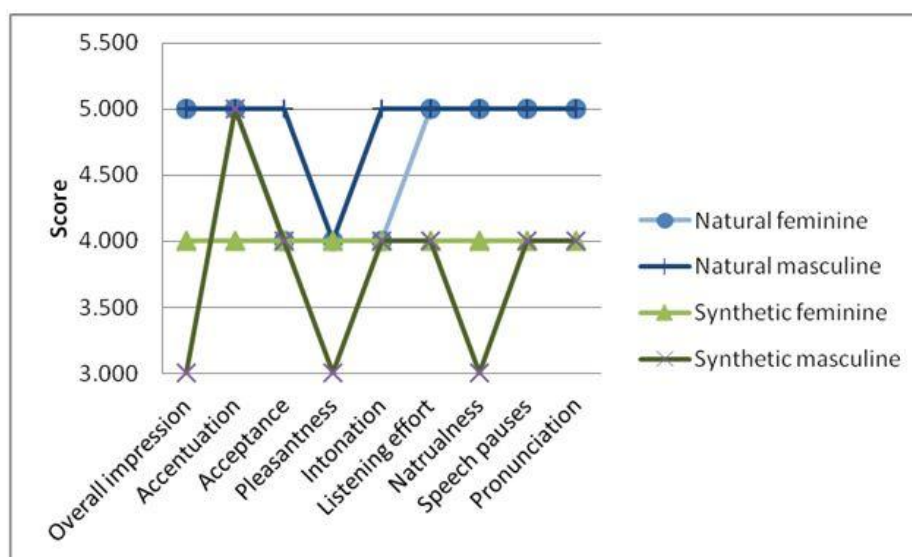

Figure 1. Mean scores of all scales for all voices.

However, since all items were collected as scores between 1 and 5, the medians (see Figure 3) may be more robust than the means. It must be stressed that all median scores are between 3.0 and 5.0. Both male and female natural voices obtain 5.0 in accentuation, listening effort, naturalness, pronunciation, speech pauses, and overall impression, and 4.0 in pleasantness. However, in acceptance and intonation the male natural voice gets higher scores (5.0 vs 4.0). This shows how the more subjective aspects which relate to end users' preferences (for instance, acceptance) present greater variation, whilst standard features that a professional describer masters (e.g. accentuation and pronunciation) are more stable. It also shows how even a natural voice may not get the highest mark in terms of pleasantness or intonation.

In as far as artificial voices are concerned, the female voice obtains 4.0 in all items under analysis, whilst the male artificial voice ranges from 3.0 (pleasantness, naturalness, overall impression) to 5.0 (accentuation), with

most items rated 4.0 on a 5-point scale (acceptance, intonation, listening effort, speech pauses, pronunciation). Again, what is especially relevant is the fact that all items are assessed above 3.0 and that in some items the median scores are the same for some natural and artificial voices. This is the case of accentuation (same scores for both natural voices and the male artificial voice), acceptance (same scores for the female natural voice and both artificial voices), pleasantness (same scores for natural voices and the female artificial voice), and intonation (same scores for the natural female voice and both artificial voices).

Figure 2. Median scores of all scales for all voices.

An analysis taking into account the ordinal characteristic of the items is based on the multinomial or logistic models. Statistically significant differences between the synthetic voices and their natural counterparts were found in all items under analysis. In all cases the natural voices were considered to be better than the artificial ones (see Annex 3 for further details).

When comparing the two artificial voices, the analysis shows that the synthetic feminine voice was more accepted, required less effort, was considered to be more natural and obtained a better score in the overall impression than the synthetic masculine one. As for the rest of the items (accentuation, pleasantness, intonation, speech pauses and pronunciation), no statistically significant differences were found.
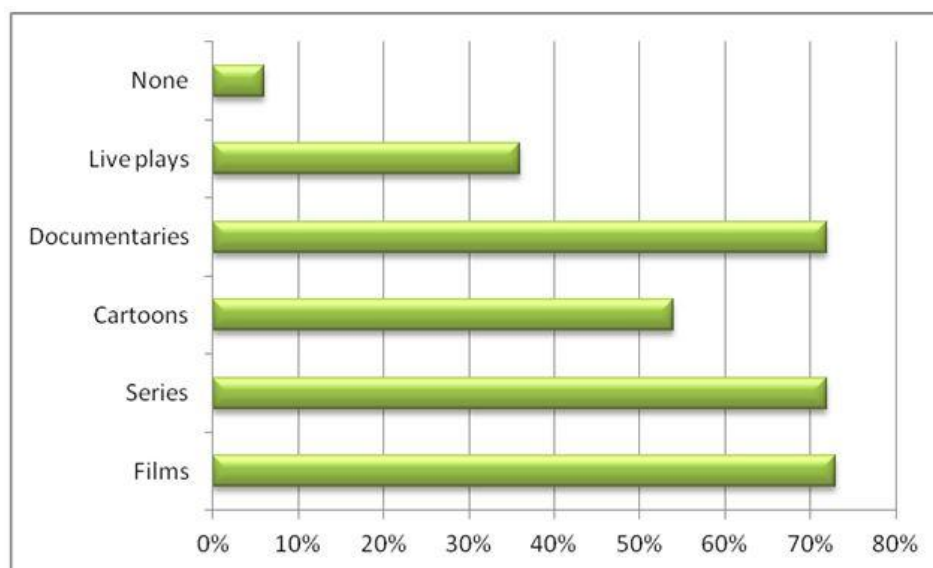
Focusing on the overall impression, the multinomial model allows us to conclude that women (OR=1.67, IC=(0.96,2.90)) and the group below 50 (OR=1.89, IC=(1.09,3.30)) gave statistically significant higher scores than men and people older than 50, respectively. No statistically significant differences were found related to the disability type.

To complement previous statistical analyses, the post-questionnaire provides qualitative data that will be discussed next. When asked about their preferences in terms of a male or a female AD voice, 72% declared they did not have any preferences, with only 16% stating that it depends on the audiovisual product. The reasons for preferring either a female or a male voice in such cases were the topic (in 7 out of the 11 cases, that is 64%) and the characters (in 4 instances, that is 36%).

When asked about their preferences regarding a human or a synthetic voice, 81% of the informants stated that they preferred a human voice to read the AD, 1% declared that they preferred a synthetic voice, 3% said that it depended on the audiovisual product, and 15% declared they did not have any specific preferences as long as the artificial voice sounded natural enough and was not tiring. It must be noted, for example, that in the case of the synthetic voices tested, the naturalness mean scores were 3.507 for the female voice and 3.104 for the male one, and the listening effort mean scores were 3.836 and 3.657 respectively, which are quite strong results in a 5-point scale. It must be also stressed that 51 informants (76%) said they normally use electronic devices with synthetic voice applications on a daily basis.

When explicitly asked about the TTS AD as an alternative solution to human voiced audio description, 94% of participants responded positively. Twenty-two participants, i.e. 33%, stated that the main reason for accepting TTS AD as an alternative solution was that it would definitely increase the amount of audio described audiovisual products. Eight out of these 22 participants explained that it would reduce both the costs and time for creating such products. Nine participants (13%) stated that it could be an alternative solution because the quality of synthetic voices was already good enough. On the other hand, 10 informants (15%) stated that synthetically voiced AD was better than no AD at all, while 9 respondents (13%) argued it should only be an alternative, not the usual situation.

When questioned about specific kinds of audiovisual products, the preferences varied slightly, as shown in Figure 4: most of the participants agreed on applying TTS AD in documentaries (48 respondents), series (48 respondents) and films (49 respondents); not so many people agreed on applying it to cartoons (36 respondents) and even less informants were willing to implement it in live plays (24 respondents), with 4 participants being against implementing it at all.

Figure 3. Audiovisual products that could be used with TTS AD.

Finally, a question about their opinion after listening to the four voices included in the experiment showed a preference for the masculine natural voice (42%) and the feminine natural voice (38%), although 14% said they preferred the feminine synthetic voice and 6% selected the male synthetic voice. These qualitative data match with the results obtained both in the descriptive and inferential statistics, which actually graded voices in the same order: the natural masculine voice was the one which obtained better mean scores, closely followed by the natural feminine, then the synthetic feminine and finally the synthetic masculine.

5.   Conclusions

This article has presented a first analysis of text-to-speech audio description in Catalan, as compared to human-voiced audio descriptions, using both male and female voices. Participants have assessed the voices taking into account various items (overall impression, accentuation, pronunciation, speech pauses, intonation, naturalness, pleasantness, listening effort, and acceptance), providing data of both a quantitative and a qualitative nature.

Results show that natural voices in our experiment have statistically higher scores than synthetic voices. They also show that the synthetic feminine voice has higher mean scores than the synthetic masculine voice in all items but accentuation. This proves that the preferential choice of blind and partially sighted persons is the audio description voiced by a human, rather than by a speech synthesis system. This does not mean, though, that TTS AD is not accepted by end users, as shown by the fact that 94% of the participants consider TTS an alternative acceptable solution, and 20% of the respondents actually state that their preferred voice from the four under analysis is a synthetic one. Moreover, it is

particularly relevant that no mean score of any of the items goes under 3.1 on a 5-point scale. As an example, the acceptance item's lowest score is a 3.7 (for the synthetic masculine voice) and the overall impression item's lowest score is a 3.2 (also for the synthetic masculine voice).

This experiment follows previous research on TTS AD carried out in Poland and Japan but it is the first of its kind in Catalan. However, it also has its own limitations. First of all, since the study used a non-probability sampling approach (Bryman 2012: 418), the results cannot be generalised to the whole Catalan blind and partially sighted population. Another of its setbacks is the length of the clips: it remains to be seen whether the results would remain the same in longer productions and in various genres. It would also be highly interesting to see whether reception varies in productions originally shot in Catalan and in dubbed productions, since the language and the sound conditions are different. Another topic worth researching would not only be the perceived quality based on a list of previously selected items, but also the engagement of the audience, in line with Fryer and Freeman's research (2013). Finally, it would also be worth researching end users behaviour if given the possibility of tuning their own AD preferences, at least as far as voice, voice gender and volume are concerned, in line with Walczak and Szarkowska's approach (2012).

All in all, it is our hope that this type of research will allow us to find new ways of increasing access to culture and entertainment for the blind and visually impaired, both on traditional and new media. We are convinced that speech technologies but also other language and visual processing technologies will play a key role and will open a myriad of research possibilities.

Acknowledgements

Bibliography

- **Alías, Francesc, Ignasi Iriondo, and Joan Claudi Socoró** (2011). "Aplicació de tècniques de generació automàtica de la parla en producció audiovisual." *Quaderns del CAC*, 37(1), 105-114. http://www.cac.cat/pfw_files/cma/recerca/quaderns_cac/Q37_Alias_etal.pdf (consulted 30.07.2014)

- **Braun, Sabine** (2011). **"**Creating coherence in Audio Description." *Meta: Journal des Traducteurs / Meta: Translator's Journal*, 56(3), 645-662. http://www.erudit.org/revue/meta/2011/v56/n3/1008338ar.pdf (consulted 30.07.2014)

- **Bryman, Alan** (2012). *Social Research Methods*. Oxford: Oxford University Press.

- **Caruso, Beatrice** (2012) "Audio Description Using Speech Synthesis." Languages and the Media. 9th International Conference on Language Transfer in Audiovisual Media. Conference Catalogue. Berlin, Germany: ICWE, 59-60.

- **Chapdelaine, Claude and Langis Gagnon** (2009). "Accessible Videodescription On-Demand." *ASSETS '09. Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. New York: ACM, 221-222.

- **Chmiel, Agnieska and Iwona Mazur** (2012). "AD reception research: Some methodological considerations." Elisa Perego (ed.) (2012). *Emerging Topics in Translation: Audio Description*. Trieste: EUT Edizioni Università di Trieste, 57-80.

- **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, Sarah Home and Sarah Morley Wilkins** (2010). *Synthetic speech evaluation protocol*. Technical report #7. Birmingham: RNIB Centre for Accessible Information (CAI).

- **De Jong, Frans** (2006). "Access Services for Digital Television." Ricardo Perez-Amat and Álvaro Pérez-Ugena (eds) (2006). *Sociedad, integración y televisión en España*. Madrid, Spain: Laberinto Comunicación, 331-344.

- **Derbring, Sandra, Peter Ljunglöf and Maria Olsson** (2009). "SubTTS: Light-weight automatic reading of subtitles." Kristiina Jokinen and Eckhard Bick (eds) (2009). *Nodalida'09: Proceedings of the 17th Nordic Conference of Computational Linguistics*. NEALT Proceedings Series, vol. 4. Odense, Denmark: Northern European Association for Language Technology.

- **Drożdż-Kubik, Justyna** (2011). "Harry Potter i Kamień Filozoficzny słowem malowany – czyli badanie odbioru filmu z audiodeskrypcją z syntezą mowy." MA Thesis. Jagiellonian University.

- **Encelle, Benoît, Magali Ollagnier-Beldame, Stéphanie Pouchot and Yannick Prié** (2011). "Annotation-based video enrichment for blind people: A pilot study on the use of earcons and speech synthesis." *ASSETS '11: Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*. New York, USA: ACM, 123-130.

- **Fernández-Torné, Anna, Anna Matamala and Carla Ortiz-Boix** (2012). "Technology for accessibility in multilingual settings: the way forward in AD?" Paper presented at *The translation and reception of multilingual films* (University of Montpellier 3, 15-16 June 2012). http://ddd.uab.cat/record/117160 (consulted 30.07.2014)

- **Fernández-Torné, Anna and Anna Matamala** (2013). "Methodological considerations for the evaluation of TTS AD's acceptance in the Catalan context." Paper presented at *ARSAD (Advanced Research Seminar).* (Autonomous University of Barcelona, 13-14 March 2013). http://ddd.uab.cat/record/117078 (consulted 30.07.2014)

- **Freitas, Diamantino and Georgios Kouroupetroglou** (2008). "Speech technologies for blind and low vision persons." *Technology and Disability* 20, 135-156.

- **Fryer, Louise and Jonathan Freeman** (2013). "Visual impairment and presence: measuring the effect of audio description." *Proceedings of the 2013 Inputs-Outputs Conference: An Interdisciplinary Conference on Engagement in HCI and Performance*. New York, USA: ACM, article n° 4.

- **González García, Luis** (2004). "Assessment of text reading comprehension by Spanish-speaking blind persons." *British Journal of Visual Impairment* 22 (1), 4-12.

- **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).

- **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. http://www.itu.int/rec/T-REC-P.85-199406-I/en (consulted 23.07.2014).

- **Kobayashi, Masatomo, Kentarou Fukuda, Hironobu Takagi, and Chieko Asakawa** (2009). "Providing synthesized audio description for online videos." *ASSETS '09: Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility*. New York, USA: ACM, 249-250.

- **Kobayashi, Masatomo, Trisha O'Connell, Bryan Gould, Hironobu Takagi, and Chieko Asakawa** (2010). "Are Synthesized Video Descriptions Acceptable?" *ASSETS '10: Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility*. New York, USA: ACM, 163-170.

- **Ljunglöf, Peter, Sandra Derbring and Maria Olsson** (2012). "A free and open-source tool that reads movie subtitles aloud." *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies*. Montreal, Canada: Association for Computational Linguistics (ACL), 1-4.

- **Llisterri, Joaquim, Natividad Fernàndez, Francesc Gudayol, Juan José Poyatos and Josep Martí** (1993). "Testing users' acceptance of Ciber232, a test to speech system used by blind people." Granström, B., Hunnicutt, S., and Spense, K.E. (eds) (1993) *Speech and Language Technology for Disabled Persons. Proceeding of an ESCA Workshop*. Stockholm, Sweden. 203-206.

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

- **Matamala, Anna, Anna Fernández-Torné and Carla Ortiz-Boix** (2013). "Enhancing sensorial and linguistic accessibility: further developments in the TECNACC and ALST projects." Paper presented at the *5th International Conference Media for All. Audiovisual Translation: Expanding Borders*. (Dubrovnik, Croatia, 25-27 September 2013). http://ddd.uab.cat/record/116868 (consulted 30.07.2014)

- **Mieskes, Margot and Juan Martínez Pérez** (2011). "A Web-based Editor for Audio-titling using Synthetic Speech." *3rd International Symposium on Live Subtitling with Speech Recognition*. Antwerp http://www.respeaking.net/Antwerp%202011/Webbased_editor.pdf (consulted 30.07.2014)

- **Nielsen, Simon and Hans-Heinrich Bothe** (2007). "SUBPAL: A Device for Reading Aloud Subtitles from Television and Cinema." Marion A. Hersh and James Ohene-Djan (eds) (2008). *Proceedings of the Conference & Workshop on Assistive Technologies for People with Vision & Hearing Impairments Assistive Technology for All Ages CVHI 2007*. CEUR Workshop Proceedings, vol. 415. http://ceur-ws.org/Vol-415/paper17.pdf (consulted 30.07.2014)

- **Oncins, Estella, Oscar Lopes, Pilar Orero, Javier Serrano and Jordi Carrabina** (2013). "All together now: a multi-language and multi-system mobile application to make living performing arts accessible." *The Journal of Specialised Translation* 20, 147-164.

- **Orero, Pilar** (2007). "Sampling audio description in Europe." Jorge Díaz Cintas, Pilar Orero and Aline Remael (eds) (2007). *Media for All. Subtitling for the Deaf, Audio Description, and Sign Language*. Amsterdam/New York: Rodopi, 111-125.

- **Orero, Pilar and Anna Matamala** (2007). "Accessible opera: overcoming linguistic and sensorial barriers." *Perspectives. Studies in Translatology* 15(4), 262-277. https://ddd.uab.cat/record/117149 (consulted 30.07.2014)

- **Pazos, Patricia** (2012). *Audiosubtitulación: una posible solución para la accesibilidad a los medios audiovisuals.* MA thesis. Autonomous University of Barcelona.

- **Salway, Andrew, Elia Tomadaki and Andrew Vassiliou** (2004). *Building and analysing a corpus of AD scripts. TIWO Television in Words. Report on Workpackage 2*. Surrey, UK: University of Surrey.

- **Szarkowska, Agnieska** (2011). "Text-to-speech audio description: towards wider availability of AD." *The Journal of Specialised Translation* 15, 142-162.

- **Szarkowska, Agnieska 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, Italy: EUT Edizioni Università di Trieste, 81-98.

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

- **Verboom, Maarten, David Crombie, Evelien Dijk and Mildred Theunisz** (2002). "Spoken Subtitles: Making Subtitled TV Programmes Accessible." Klaus Miesenberger, Joachim Klaus and Wolfgang L. Zagler (eds) (2002). *Proceedings of Computers Helping People with Special Needs, 8th International Conference, ICCHP 2002.* Berlin-Heidelberg, Germany: Springer-Verlag, 295-302.

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

- **Walczak, Agnieska and Agnieska Szarkowska** (2012). "Text-to-speech audio description of educational materials for visually impaired children." Silvia Bruti and Elena Di Giovanni (eds) (2012). *Audio Visual Translation across Europe: An Ever-Changing Landscape*. Berna/Berlin: Peter Lang, 209-234.

Website

World Health Organisation (2013) Fact Sheet no. 282. http://www.who.int/mediacentre/factsheets/fs282/en/ (consulted 17.07.2014).

Biography

Anna Fernández-Torné holds an MA in Audiovisual Translation (UAB), MA in Language Consultancy in the Media (UAB) and a European MA in Audiovisual Translation (Parma University). Her research centres on audio description and technologies, both text-to-speech and machine translation. She has been a freelance translator since 2004, specialising in audiovisual translation, and lectures at the MA in Audiovisual Translation at Universitat Autònoma de Barcelona. anna.torne@gmail.com

Anna Matamala, holds a PhD in Applied Linguistics (UPF) and is a tenured lecturer at the UAB. She has published in international refereed journals such as *Meta, Perspectives, Babel, VIAL* and *Linguistica Antverpiensia*, among others. She is the author of a book on interjections and lexicography (IEC, 2005), co-author of a book on voice-over (Peter Lang, 2010), and co-editor of three volumes on AVT and media accessibility. anna.matamala@uab.cat

Annex 1. Questionnaire

How would you describe the quality of the voice you have just heard?
1. Bad
2. Regular
3. Neutral
4. Good
5. Excellent

Did you detect anomalies in terms of the accentuation of words?
1. Yes, al lot of them
2. Yes, many
3. Yes, some
4. Yes, but only a few
5. No, none

Did you notice anomalies in terms of pronunciation?
1. Yes, al lot of them
2. Yes, many
3. Yes, some
4. Yes, but only a few
5. No, none

Do you think the voice makes pauses when it is needed?
1. No, never
2. No, almost never
3. Yes, normally
4. Yes, almost always
5. Yes, always

How would you rate the intonation of sentences?
1. Very bad
2. Bad
3. Good
4. Quite good
5. Very good

How would you define the degree of naturalness of the voice?
1. Very unnatural
2. Unnatural
3. Natural
4. Quite natural
5. Very natural

To what extent do you deem this voice to be pleasant?
1. Very unpleasant
2. Unpleasant

3. Neutral
4. Pleasant
5. Very pleasant

Do you think listening to this voice for a long time would be tiring?
1. Yes, a lot
2. Yes, quite a lot
3. Yes, a little bit
4. No, not much
5. No, not at all

Do you think this voice could be used for voicing audio descriptions?
1. No, never
2. No, almost never
3. Yes, in some cases
4. Yes, in many cases
5. Yes, always

Annex 2. Post-questionnaire

*Mandatory field

- Identifier *

Enter your initials (first name initial, first surname initial and second surname initial) followed by your age. Do not leave any blank space in between.

- Age*

- Sex*

Male / Female

- Level of studies reached*

Lower than first stage of secondary school

Secondary education, first stage

Secondary education, second stage

Advanced vocational education

First cycle university education (diploma, degree, engineering or graduate studies)

Second cycle university education (master, postgraduate or doctoral studies)

- In case you have reached university education, please specify.

- Occupation*

Public administration management and management of companies with 10 or more wage earners.

Management of companies with less than 10 wage earners

Management of companies without wage earners

Professions associated with 2nd and 3rd cycle university degrees and the like

Professions associated with a 1st cycle university degree and the like

Support technicians and professionals

Administrative type employees

Catering services workers and personal services workers

Protection and security service workers

Retail workers and the like

Workers skilled in agriculture and fishing

Skilled construction workers, except machinery operators

Skilled workers in the extractive industry, metallurgy, construction of machinery and related trades.

Skilled workers from the graphic arts, textile and tailoring, elaboration of food, cabinetmakers, craftspersons and other similar industries

Fixed machinery and industrial installation operators; fitters and assemblers.

Mobile machinery drivers and operators

Unskilled workers in the service sector (except transports)

Agriculture, fishing, construction, manufacturing industries and transport labourers.

Armed forces

Unemployed for longer than one year

Unemployed, seeking a first job

- Profession in your own words
- Kind of visual impairment according to WHO*
 Blindness / Low vision
- How long have you been visually impaired for? *
From birth / For less than 1 year / For between 1 and 10 years / For between 11 and 20 years / For more than 20 years any
- Have you ever seen an audio described product (films, series, theatre plays, etc.)?*
Yes / No
- In case you have, which kind of products? (You can tick more than one answer)
Films / Series / Cartoons / Theatre plays / Opera plays
- How often do you use audio described products?*
At least once a day / At least once a week / At least once a month / Never / Other
- Do you prefer the AD to be read by*
A man / A woman / It depends on the audiovisual product / I don't care
- If it depends on the audiovisual product, what does it depend on exactly?
- You prefer the AD to be read by*
A human voice / An artificial voice / It depends on the audiovisual product / I don't care
- If it depends on the audiovisual product, what does it depend on exactly?
- Do you use electronic devices with synthetic voice applications, such as mobile phones or computers?*
Yes / No
- How often do you use them?*
At least once a day / At least once a week / At least once a month / Never
- Have you ever used audio described products with synthetic voice?*
Yes / No
- Do you think it is an alternative solution to human voiced audio description?*
Yes / No
- Why do you think so?*
- What kind of products would you use with synthetic voiced AD? (You can tick more than one answer)*
Films / Series / Cartoons / Documentaries / Live plays / None
- Which voice, from the 4 voices you have just heard, did you like the most?*
The masculine synthetic voice / The masculine natural voice / The feminine synthetic voice / The feminine natural voice
- Would you be able to rank them in order, from the one you like the most to the one you liked the least?
- Other comments

Annex 3. Odds ratio (OR) tables

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.3878	0.2041	0.7367
Synthetic masculine vs natural masculine	0.06400	0.03085	0.1328
Synthetic masculine vs natural feminine	0.05216	0.02484	0.1095
Synthetic feminine vs natural masculine	0.1650	0.08252	0.3300
Synthetic feminine vs natural feminine	0.1345	0.06652	0.2719
Natural masculine vs natural feminine	0.8150	0.4144	1.6029

Overall impression

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	1.1856	0.5856	2.4005
Synthetic masculine vs natural masculine	0.1604	0.06801	0.3783
Synthetic masculine vs natural feminine	0.1774	0.07617	0.4129
Synthetic feminine vs natural masculine	0.1353	0.05743	0.3187
Synthetic feminine vs natural feminine	0.1496	0.06433	0.3478
Natural masculine vs natural feminine	1.1058	0.4334	2.8212

Accentuation

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.6593	0.3323	1.3081
Synthetic masculine vs natural masculine	0.1128	0.05074	0.2506
Synthetic masculine vs natural feminine	0.1000	0.04445	0.2251
Synthetic feminine vs natural masculine	0.1710	0.07730	0.3784
Synthetic feminine vs natural feminine	0.1517	0.06774	0.3398
Natural masculine vs natural feminine	0.8871	0.3765	2.0898

Pronunciation

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.6419	0.3409	1.2088
Synthetic masculine vs natural masculine	0.08431	0.04036	0.1761
Synthetic masculine vs natural feminine	0.1045	0.05101	0.2142
Synthetic feminine vs natural masculine	0.1313	0.06367	0.2710
Synthetic feminine vs natural feminine	0.1628	0.08043	0.3297
Natural masculine vs natural feminine	1.2397	0.5845	2.6297

Speech pauses

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.6922	0.3463	1.3835
Synthetic masculine vs natural masculine	0.1176	0.04655	0.2973
Synthetic masculine vs natural feminine	0.1568	0.06638	0.3701
Synthetic feminine vs natural masculine	0.1700	0.06693	0.4316
Synthetic feminine vs natural feminine	0.2265	0.09541	0.5375
Natural masculine vs natural feminine	1.3324	0.4610	3.8508

Intonation

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.4920	0.2652	0.9126
Synthetic masculine vs natural masculine	0.07831	0.03867	0.1586
Synthetic masculine vs natural feminine	0.07904	0.03905	0.1600
Synthetic feminine vs natural masculine	0.1592	0.08044	0.3150
Synthetic feminine vs natural feminine	0.1607	0.08124	0.3177
Natural masculine vs natural feminine	1.0092	0.4981	2.0447

Naturalness

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.5685	0.2815	1.1482
Synthetic masculine vs natural masculine	0.07245	0.02695	0.1948
Synthetic masculine vs natural feminine	0.08628	0.03367	0.2211
Synthetic feminine vs natural masculine	0.1274	0.04734	0.3431
Synthetic feminine vs natural feminine	0.1518	0.05913	0.3895
Natural masculine vs natural feminine	1.1909	0.3705	3.8281

Pleasantness

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.4473	0.2363	0.8468
Synthetic masculine vs natural masculine	0.1148	0.05619	0.2347
Synthetic masculine vs natural feminine	0.1205	0.05921	0.2453
Synthetic feminine vs natural masculine	0.2568	0.1272	0.5181
Synthetic feminine vs natural feminine	0.2695	0.1340	0.5417
Natural masculine vs natural feminine	1.0494	0.5012	2.1973

Listening effort

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.3832	0.1814	0.8096
Synthetic masculine vs natural masculine	0.07689	0.02486	0.2378
Synthetic masculine vs natural feminine	0.01831	0.002365	0.1417
Synthetic feminine vs natural masculine	0.2007	0.06254	0.6438
Synthetic feminine vs natural feminine	0.04778	0.006047	0.3775
Natural masculine vs natural feminine	0.2381	0.02552	2.2217

Acceptance