electronics

MDPI

Article

Four-Features Evaluation of Text to Speech Systems for Three Social Robots

Fernando Alonso Martin *, María Malfaz, Álvaro Castro-González, José Carlos Castillo and Miguel Ángel Salichs

Department of Robotics, University Carlos III of Madrid, Avda de la Universidad, 30, 28911 Leganés (Madrid), Spain; mmalfaz@ing.uc3m.es (M.M.); acgonzal@ing.uc3m.es (A.C.-G.); jocastil@ing.uc3m.es (J.C.C.); salichs@ing.uc3m.es (M.Á.S.)

* Correspondence: famartin@ing.uc3m.es

Received: 23 December 2019; Accepted: 22 January 2020; Published: 5 February 2020

Abstract: The success of social robotics is directly linked to their ability of interacting with people. Humans possess verbal and non-verbal communication skills, and, therefore, both are essential for social robots to get a natural human–robot interaction. This work focuses on the first of them since the majority of social robots implement an interaction system endowed with verbal capacities. In order to do this implementation, we must equip social robots with an artificial voice system. In robotics, a Text to Speech (TTS) system is the most common speech synthesizer technique. The performance of a speech synthesizer is mainly evaluated by its similarity to the human voice in relation to its intelligibility and expressiveness. In this paper, we present a comparative study of eight off-the-shelf TTS systems used in social robots. In order to carry out the study, 125 participants evaluated the performance of the following TTS systems: *Google*, *Microsoft*, *Ivona*, *Loquendo*, *Espeak*, *Pico*, *AT&T*, and *Nuance*. The evaluation was performed after observing videos where a social robot communicates verbally using one TTS system. The participants completed a questionnaire to rate each TTS system in relation to four features: *intelligibility*, *expressiveness*, *artificiality*, and *suitability*. In this study, four research questions were posed to determine whether it is possible to present a ranking of TTS systems in relation to each evaluated feature, or, on the contrary, there are no significant differences between them. Our study shows that participants found differences between the TTS systems evaluated in terms of intelligibility, expressiveness, and artificiality. The experiments also indicated that there was a relationship between the physical appearance of the robots (embodiment) and the suitability of TTS systems.

Keywords: text to speech systems; user studies; speech-based; accessibility technologies; natural language generation

1. Introduction

Social robots are intended to "live" around humans to help and/or entertain them. In this regard, the speech is probably the richest and the preferred way for humans to communicate, making the software that allows the robot to generate an artificial voice a crucial element during human–robot interaction. These systems, commonly known as Text To Speech (TTS) systems, can convert text to artificial voice.

There are several definitions of what a TTS system is. Van Bezooijen defines it as a system that 'allows the generation of novel (oral) messages, either from scratch (i.e., entirely by rule) or by recombining shorter pre-stored units' [1]. On the other hand, Handley uses the following definition: 'Speech synthesis systems, or speech synthesizers, are computer programs which automatically generate speech, i.e., systems which enable the computer to "*talk*" or "*speak*" to the user' [2]. Since the

beginning of the 1990s, several authors have analyzed the technological foundations of text to voice conversion using a computer [3]. Klatt presented one of the first analyses of the TTS systems available in the late 1980s [4].

Nowadays, there are many proposals of commercial TTS systems with different features and performance. Therefore, it seems that a comparative study of them would be quite useful for robotics researchers, among others. In this paper, we present a comparative study of the following TTS systems: *Ivona*, *Nuance*, *Google*, *Microsoft*, *AT&T*, *Espeak*, *Pico*, and *Loquendo*.

Since this work is motivated by our own need to decide which TTS system to select for our robots, each TTS has been configured and executed in three social robots, *Mbot*, *Mini*, and *Maggie*, the robots of the Social Robotics group of the RoboticsLab (University Carlos III of Madrid).

This comparative study is focused on the evaluation of four features: the first one is the degree of *intelligibility* of the generated speech. In this sense, the user must evaluate if the robot communicates with clarity and if the sentence is well understood. The second one is the *expressiveness* of the generated voice. That is, if the voice sounds monotonous or not: if it is able to emphasize certain words, to make pauses, to change the speech speed, etc. The third feature is related to the *artificiality* of the voice that is, if it is perceived as more or less robotic (in the sense of less human-like). Finally, we evaluate the *suitability* of the generated voice for the robot, i.e., if the voice fits with the external appearance of the robot. These features will translate into four research questions in Section 3.4, that is, the hypotheses that this study is intended to verify.

This paper is structured as follows. Section 2 offers a review of the TTS systems found in the literature, considering the platforms in which they are integrated. Next, in Section 3, there is described the procedure that has been followed. Then, Section 4 presents the results obtained for each evaluated feature. Finally, in Section 5, the authors present their conclusions and a discussion of this study, as well as the limitations and lessons learned.

2. Related Work

In this section, the most important TTS systems that are currently available are presented. Moreover, we also highlight those systems that are used in social robots and other electronic devices. After that, we review the previous literature of comparative studies of different TTS systems and the features that are evaluated.

2.1. Relevant TTS Systems in the Market

There are several TTS systems that are currently available for use, but in this section we are going to describe some of them. There are several references where the most relevant TTS systems are listed [5,6]. Some of them are the following:

- *Mbrola* is an open source artificial voice generation system which allows, at a low level, a great degree of control over the synthesized speech. In this sense, the user can configure various parameters to get precise prosodic control [7].
- *Loquendo TTS* synthesizes a human-like voice in multiple languages. It became very popular on Internet platforms, like Youtube, since the community employed it to generate tutorials and parodies.
- *Pico* is a TTS system developed by SVOX that currently is installed by default on most Android devices (at least until the 4.2 version). Note that SVOX and Loquendo were acquired by Nuance in 2011.
- *Nuance Real Speak* is the flagship product, regarding speech synthesis system, of the Nuance Communications company. It allows generating voices in several languages and it is the official voice of the virtual assistant of Apple, Siri.

- *Festival* is a general multilingual speech synthesis system originally developed at the Centre for Speech Technology Research (University of Edinburgh). It is distributed under a free software license similar to BSD.
- *Ivona*: This is the TTS system developed by the Amazon company. It is widely used in Amazon devices, such as the Kindle electronic reader.
- *Google*: This is the voice system developed by Google and is used in its applications, web services, and in its virtual assistant 'Google Now'. The generated voice is different in each language (supports over 80 languages and dialects).
- *Microsoft*: This is the voice system of the Microsoft company, and it is used in its services, applications, operative systems, and in its virtual assistant 'Cortana'.
- *AT&T*: This is the system developed by the AT&T company. It generates speech in eight different languages, and it is used in its call centres.
- *Verbio*: This is the system developed by the Spanish company Verbio. It is mainly used by call-centre services of companies and public institutions.

2.1.1. The TTS Systems Used in Social Robots and Other Electronic Devices

In social robotics, not all the robots communicate using synthesized speech. In fact, some of them do not use any sound to express themselves, such as the robot Keepon [8]. On the other hand, there is another group of social robots that express their internal state, but just using non-verbal sounds. This is the case of *Paro* [9], a baby seal robot that plays sounds similar to the ones that a real baby seal emits. Another example of a social robot with non-verbal communication skills is the robotic dog *Aibo* [10], developed by Sony. Using some pre-generated sounds, *Aibo* can share with the user some internal states such as 'happy', 'sad', etc. In addition, it can also communicate different events such as a user detection, an internal failure, etc.

Finally, there is another group of social robots that has verbal communication skills. For example, the *Nao* robot [11], developed by Aldebaran, is a humanoid robot that is able to interact with people in a natural way by voice and gestures. This little humanoid (it is about 58 cm tall) has become one of the research platforms most used by the robotic community for HRI (Human–Robot Interaction). This robot uses the TTS web service of *Nuance*. The same company has recently developed a new social robot called *Pepper* (more details in [12]). This is another humanoid robot, but bigger than *Nao* (about 120 cm tall). Pepper, like any other social robot, has been created to interact with humans using natural modes of interaction such as voice, gestures, nonverbal sounds, touch, etc. Moreover, Pepper includes another additional input and output channel: a tablet placed on its chest. As Nao, Pepper also uses Nuance as TTS.

Another social robot with verbal communication skills is *iCub* [13,14]. Its appearance is similar to a two-year old children (about 1 m tall), and it is used as a research platform to test learning algorithms, cognitive skills, and artificial intelligence algorithms. In this case, the TTS system used is *Acapela* [15]. The robot *Jibo* is another kind of social robot that has been recently developed. The company's founder, Cynthia Breazeal, describes *Jibo* 'as the result of R2D2 and Siri having a baby,' that is, it is a robot that is endowed with verbal and non-verbal communication skills. *Jibo* supports text-to-speech markup; this allows selecting which parts of the synthesized text should be given emphasis, or how unusual words or names should be pronounced. This feature is particularly important given the emphasis on the robot having its own specific personality. The *Yamaha Vocaloid Humanoid Robot* uses *Vocaloid* [16–18]. This system is different because it is used for talking and singing. Therefore, it allows creating very realistic artificial singing voices.

In relation to electronic devices, nowadays, there is a new generation of 'intelligent' devices (smartphones, tablets, smartspeakers, smartwatches...) that are equipped with a virtual assistants. The user interacts with this assistant by voice. In the case of the *Apple Inc.* devices, *iPad*, *iPhone*, *Watch*, etc., they have *Siri* [19], which uses *Nuance* for voice recognition and speech synthesis tasks. Android-based devices have another assistant, called *Google Assistant* [20], which uses the *Google TTS*

technology. Moreover, the *Amazon* devices, such as the *Kindle* [21] or *Echo* (https://www.amazon.es/Amazon-Echo-Altavoz-Inteligente-Alexa), use "under the hood" *Ivona*. Finally, there are other electronic devices, developed by *Microsoft*, which are equipped with the voice assistant *Cortana* [22] that uses *Microsoft TTS*.

2.2. Previous Comparative Studies of TTS Systems

The literature offers few comparative studies of the performance of TTS systems. In addition, the related references are not very recent, so perhaps some of the evaluated TTS systems are now discontinued. These studies have aimed to apply these results to improve the Interactive Voice Response applications (IVRs) that are used on "call centres". Nevertheless, in this section, we briefly present these papers.

In 2006, Roehling presented a comparative table of 12 TTS systems: *BabTTS*, *Natural Voices*, *DECtalk*, *Naxpres*, *Loquendo*, *Mulan*, *Speech SDK*, *RealSpeak*, *Festival*, *Gnuspeech*, *OpenMary*, and *ProSynth* [23]. This analysis studied the different features of the TTS systems needed for synthesizing expressive speech, considering pitch, duration, loudness, and voice quality. They concluded that *OpenMary* was the best solution to endow the robot *B21r* with an affective speech.

More recently, in 2009, Handley [2] presented a study focusing on the requirements of a TTS system to be used in Computer-Assisted Language Learning applications. These systems have control over the characteristics of the generated speech, that is, different styles (formal or familiar), different communicative rhythms (the speech rate), different tones of voice (timbre), and different ways of expression (interrogative, enunciative, imperative, exhortative, and exclamative). This study analyzed four TTS systems: *AT&T* , *Nuance Vocalizer* (the predecessor software of Nuance NaturalSpeak), *eLite*, and *Acapela BrightSpeech* [15]. The participants listened to the different systems using a PC and, after that, they completed an online questionnaire to assess their adequacy, acceptability, and intelligibility. Then, the average scores obtained for each TTS system and for each feature were presented. On average, the top-rated TTS system was *Acapela BrightSpeech*.

More recent papers focus on comparative studies of TTS systems for non-Latin languages, such as Arabic and Hindi. Research in this area has so far been mainly confined to English and other European languages (Spanish, German, French, Italian, etc.). For the Arabic [24] and Indian languages [25], such tools are still in their infancy, and the TTS systems developed are mainly used to help visually impaired people.

2.3. Evaluated Features of the TTS Systems

In order to determine the performance of a TTS system, it is necessary to define its characteristics and the way to evaluate them. In this sense, some authors have stated a formal definition of the features of a TTS system. According to Francis [26], the most important features of a TTS system are *intelligibility* and *naturalness*. First, he defines *intelligibility* as the ease of users' understanding the speech generated by humans or machines. Then, he defines a *natural conversation* (naturalness) as a speech that sounds as if it had been produced by a native speaker.

Handley [27] suggests that the quality of speech generated by a TTS system during Human–Computer Interaction (HCI) should be as *comprehensible*, *natural*, and *accurate* as possible. Later, as previously presented, Handley, in [2], presents a comparative study of different TTS systems and evaluates them in relation to the following features:

1. *Adequacy*: 'is the speech adequate for use as a reading machine (in comparison with other media)?'
2. *Acceptability*: 'is the speech acceptable for use as a reading machine (when is not possible to use other media)?'
3. *Comprehensibility*: 'is the message easy to understand?'
4. *Intelligibility*: 'are the individual phonemes/sounds and words easy to recognize (and discriminate one from another)?'

5. *Choice of pronunciation*: 'is the pronunciation correct?'
6. *Precision of phonemes*: 'was the articulation of the phonemes/sounds precise?'
7. *Appropriateness of prosody*: 'was the prosody (music) of the utterance appropriate?'
8. *Naturalness of phonemes*: 'do the phonemes/sounds sound natural/human?'
9. *Naturalness of prosody*: 'does the prosody (music) sound natural/human?'
10. *Expressiveness*: 'was the emotion expressed well?'
11. *Appropriateness of register*: 'was the register appropriate?'

On the other hand, the International Telecommunication Union (ITU-T), in 1994, set a questionnaire to evaluate TTS systems in voice applications (call centres) [28]. This questionnaire used the *Mean Opinion Score* (MOS) [29], and the evaluated features were the following:

1. *Sound quality acceptance*: related to the quality of the sound. This requires a yes or no answer.
2. *Listening effort*: related to the effort required to understand the message.
3. *Comprehension problems*: related to the difficulties to understand certain words.
4. *Articulation*: related to the question about if the sounds were distinguishable.
5. *Pronunciation*: related to the possible anomalies detected in pronunciation.
6. *Speaking rate*: related to the average speed of delivery.
7. *Pleasantness*: related to the pleasantness of the voice.
8. *Overall impression*.

The user evaluates each feature using a score from 1 to 5 (five-point Likert Scale), 5 being the most positive (except for sound quality acceptance, which required a yes/no answer). Other studies have been carried out using this MOS scale, or a modified version. This is the case of the research presented by Viswanathan in [30]. He uses an extended version of the MOS scale, and he concludes that the most important features to evaluate a TTS system are *intelligibility* and *naturalness*. Each of these concepts, according to that author, includes other features of a TTS system. That is, *naturalness* includes: *naturalness*, *ease of listening*, *pleasantness*, and *audio flow*; on the other hand, *intelligibility* includes: *listening effort*, *pronunciation*, *comprehension*, *articulation*, and *speaking rate*. More recently, in 2014, King [31] performs a review of the improvements obtained in the TTS technologies during the last decade. Again, he claims that the evaluations of *naturalness* and *intelligibility* are the main evaluation criteria for determining the quality of the speech synthesis. For social robots, Alonso [32] defines the *naturalness* of the generated speech as its degree of similarity with that emitted by a human, while the *intelligibility* is defined as the ease of the user's understanding the message generated by the robot. For that author, these two features are the most important ones during HRI.

3. Experiment

As already stated, in this paper, we present a comparative study of the performance of several TTS systems to be used in social robots. In order to carry out this study, some of the TTS systems that are currently available, described in Section 2.1, were integrated in our social robots, introduced in this section, particularly in Section 3.2. By means of a questionnaire, the participants evaluated them by rating their features.

3.1. The Compared Text-To-Speech Systems

The social robots used to carry out this comparative study have an interaction system known as the "Robotic Dialog System", or just RDS, presented in [33]. The RDS gives to these robots the capacity to interact with humans, especially using multimodal speech dialogs. In this study, we have implemented and used the component called 'Text-To-Speech'. This component allows our social robots to communicate with the users using different kinds of voice, language, volume, etc. In addition, it integrates the eight TTS systems that are analyzed in this paper:

1. *AT&T*
2. *Google*
3. *Ivona*
4. *Microsoft*
5. *Nuance*
6. *Loquendo (v7.7)*
7. *Espeak (v1.48)*
8. *Pico (v2018)*.

The first five TTS systems require an Internet connection (since they use web services), while the last three do not require a persistent connection.

We have selected these eight TTS systems based on three main requirements: (i) the system should be used in different domains, paying special attention to developments integrated by the robotics research community; (ii) the software should be open source or, at least, it should offer a trial version, and (iii) it should support the Spanish language with acceptable technical support. Thus, *Festival* was not selected since it does not offer robust Spanish support, and *Verbio* was discarded since it does not offer a trial version. It should be noted that the selected TTS systems (except for *Loquendo*) cannot be customized, that is, they offer just one version. In the case of *Loquendo*, we use its default speech in order to make all results in this study comparable.

3.2. The Social Robots

For the embodiment comparative, we have used our three social robots, built at the RoboticsLab from Universidad Carlos III of Madrid (Spain). These robots are: *Maggie* [34,35] (Figure 1), *Mini* [36] (Figure 2), and *Mbot* [37] (Figure 3).

Figure 1. Maggie.

Figure 2. Mini.

Figure 3. Mbot.

The robots integrate a dialog mechanism to enable natural HRI. For this reason, selecting the most adequate TTS system is crucial to enhance the user experience. Apart from the dialog system, the robots include high-quality speakers, microphones, and sound cards. The first robot, *Maggie*, is able to move through the environment to interact with people. The robot was originally designed as a generic research platform to test interaction mechanisms to improve the HRI experience. Maggie can communicate through sounds, gestures, and a touch-screen mounted in its chest. The robot has a rigid plastic shell and is 1.40 m tall. *Mini* is a desktop version of *Maggie*, also developed by the RoboticsLab,

that acts as a companion for elderly people. In contrast to Maggie, Minnie is shorter, (just 55 cm) and is covered in a plush-like soft fabric and integrates the same HRI capabilities as Maggie, with an external tablet to enhance interaction. Finally, *Mbot* is another mobile platform developed in the EU project MOnarCH [38]. The robot is 1.15 m tall, that is, like the height of an 8–11 year-old child, since this social platform was designed to interact with children at the pediatric ward of the Portuguese Oncology Institute in Lisbon (Portugal). Similarly to Maggie, Mbot's shell is of a rigid material, carbon fibre.

3.3. Procedure

As seen in Section 2, in order to determine the performance of a TTS system, several authors have proposed different sets of characteristics to be evaluated. In the present paper, considering these references, especially the ones presented by Handley [2] and Viswanathan [30], the performance of each TTS system is determined by the evaluation, using questionnaires, of the following features:

- **Intelligibility**: 'Can you clearly understand the voice of this robot?'
- **Expressiveness**: 'How do you perceive this robot's voice: monotonous or very expressive?'
- **Artificiality**: 'Do you think that this is a robotic voice?'
- **Suitability**: 'Do you think that this voice is suitable for this robot?'

Each of these questions have been rated using a Likert 5-point scale. In the case of *expressiveness*, the ranking varies between 'very monotonous' (1) and 'very expressive' (5). For the other features, a lower number of points corresponds to 'Not at all' while the maximum one is for 'Yes, absolutely.'

As can be observed, in addition to *intelligibility* (known as *comprehensibility* by Handley) and *naturalness* (also known as *expressiveness* by Handley), considering our target scenarios, TTS systems in Human–Robot interaction, and more specifically in social robotics, we have included two other important features: *artificiality*, related to the metallic/robotic sound of the voice, and *suitability*, related to the perception that the user has of whether the voice suits the robot considering its external appearance. The evaluation of these characteristics, as also stated in [32], is important in this kind of comparative study.

The questionnaires were created using the web tool 'Google Forms' [39]. The first page of the questionnaire is an introductory page where the user has to read some instructions about how to fill it in, and to answer some personal questions: age, gender, and educational level (university or non-university studies). The main part of the questionnaire is divided into eight pages, each one associated with a TTS system. The order of the pages was randomized when the forms were created. Every page shows a short video where the robot is talking using a specific TTS system. The robot says the following sentence in Spanish: 'This is the robot *X* and this is a test sentence to evaluate the TTS system *Y*.' After hearing this sentence, the user must score the four questions, and then the next page appears, showing the same robot using a different TTS system.

These questionnaires were distributed publicly for a month through the Internet using social networks in order to try to obtain the maximum diffusion. Each user was only allowed to fill out one questionnaire, so the user evaluated the performance of the eight TTS systems for just one robot. This assignment was made by the researchers, so the user did not know about the existence of the other robots, trying to balance the number of participants per robot/questionnaire type.

3.4. Research Questions

These questionnaires had two goals. The first one was to verify the following questions:

1. RQ1: are all TTS systems equally well understood?
2. RQ2: do all TTS systems have the same expressiveness?
3. RQ3: are all TTS systems equally perceived as robotic?
4. RQ4: are all TTS systems equally suitable for each robot?

In case the results confirm these RQs, then the second goal was to rank the TTS systems considering the features evaluated.

3.5. Participants

For this study, we obtained 125 questionnaires in all (for the three robots). The distribution among the robots is the following: 44 questionnaires for *Maggie* (35.2%), 42 for *Mini* (33.6%), and 39 for *Mbot* (31.2%).

Regarding their age, participants were grouped into three categories: 17–30 years, with 33 participants (26.4%); 31–40 years, with 86 participants (68.8%); and more than 41 years, with six participants (4.8%). Most of the participants were males (94 participants, which means 75.2% of the participants) and just 31 participants were females (24.8%). Finally, regarding the educational level, 24 participants (19.2%) say that they have carried out only non-university studies (just primary or secondary), while the majority of the participants (101, 80.8%) declare that they have carried out university studies (a bachelor's degree, masters, or PhD).

4. Results

This section introduces a thorough analysis of the questionnaires, grouping the results regarding the research questions presented in Section 3.4. The software used in the statistical analysis of the results was *IBM SPSS* [40].

In our analysis, we considered the scores given to each TTS system, our independent variables, considering all the research questions (features), our dependent measures: the mean and the standard deviation values were calculated and are presented in the next sections. We also had to prove that the differences between the mean values were significant for each TTS in relation to each dependent measure using one-way repeated measures ANOVA. After proving a statistically significant result from the above analyses, we could select which TTS systems differ from one another. This information was provided in the Pairwise Comparison tables, presented in Appendix A.

4.1. Intelligibility: Are All TTS Systems Equally Well Understood?

This first feature evaluates if the voice is clearly understood. Considering the results of the multivariate test, Wilks' Lambda (WL), there are significant differences between the TTS systems, $WL = 0.101, F(7, 118) = 149.89, p < 0.001$. In Figure A1 (see the Appendix A), the pairwise comparison table is presented.

Therefore, we can say that the answer to RQ1 is that not all the TTS systems are equally well understood. This answer allows ranking the TTS systems by representing the results in the order in which the TTS system with the highest mean value is situated first (at the left of the figure) and the one with the lowest mean value appears at the last position (at the right of the figure); see Figure 4. The ranking shows that, in terms of the intelligibility, the best-synthesized voice corresponds to *Google*. *Ivona* TTS also receives a good score. In fact, there is no significant difference with Google: $p = 0.228$. We can identify a second group significantly different from the previous ones. This is composed by *Loquendo*, *Nuance*, *Microsoft*, and *Pico*. The study shows that the intelligibility of *AT&T* and *Espeak* is noticeably worse.

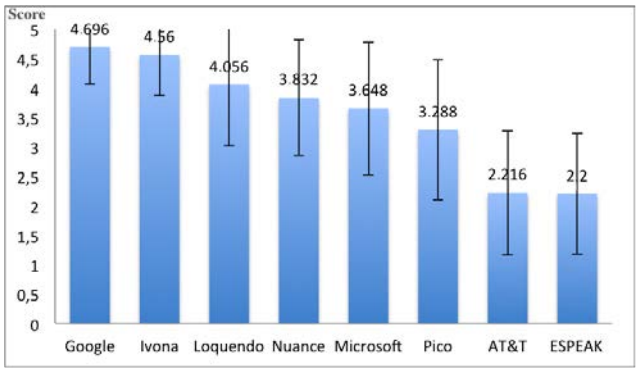

Figure 4. Ranking of TTS systems for *Intelligibility*. The vertical axis represents the mean score obtained on the questionnaires, being 5 the maximum. The error bars represent the standard deviation.

4.2. Expressiveness: Do All TTS Systems Have the Same Expressiveness?

This feature expresses how monotonous or expressive users perceive the synthetic voice generated by the TTS system. Again, we analyze the results provided by the ANOVA test. In this case, $WL = 0.25, F(7, 118) = 49.56, p < 0.001$, which means that the different TTS systems differ in *expressiveness*. For this reason, we can say that not all TTS systems have the same *expressiveness* (RQ2). The pairwise comparison table is presented in Figure A2; see Appendix A.

As in the previous feature, we can use the means and standard deviation to rank the systems evaluated regarding their expressiveness.

Again, *Google* TTS stands out, being perceived as the most expressive system, ($p < 0.05$) (see Figure 5). After *Google*, we find *Loquendo*, *Ivona*, *Microsoft*, and *Nuance* with no significant differences, $p = 1$, among them in terms of expressiveness. *Pico*, *AT&T*, and *Espeak* are perceived as the least expressive systems.

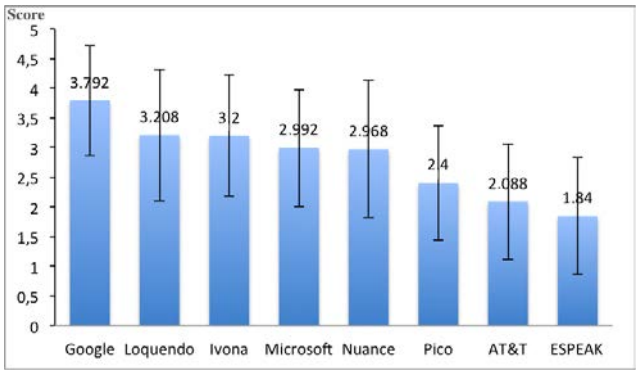

Figure 5. Ranking of TTS systems for *Expressiveness*. The vertical axis represents the mean score obtained on the questionnaires, 5 being the maximum. The error bars represent the standard deviation.

4.3. Artificiality: Are All TTS Systems Equally Perceived as Robotic?

Considering artificiality, the aim is to analyze how "robotic" the participants perceive the robot's voice. By robotics, we consider how not human-like or metallic the voice sounds. The results from the multivariate test, Wilks' Lambda, show significant differences between the TTS systems, $WL = 0.37, F(7, 118) = 28.17, p < 0.001$.

Given these results, the answer to the RQ3 is that not all the TTS systems are equally perceived as "robotic". Figure A3 (Appendix A) presents the pairwise comparison table for this feature.

The results show that *Espeak* was perceived as the most artificial TTS system, with a significant difference with respect to the other systems evaluated. Figure 6 shows the ranking regarding Artificiality where, after *Espeak*, the systems are sorted as follows: *AT&T*, *Loquendo*, *Pico*, *Microsoft*, *Nuance*, *Ivona*, and *Google*. In contrast to *intelligibility* and *expressiveness* features, there is no clear set

differentiation among the TTS systems as they all present similarities ($p > 0.05$) with their neighboring ranked ones. In any case, *Google* is perceived as the most natural TTS system showing that there is a correlation between the features analyzed in this work.

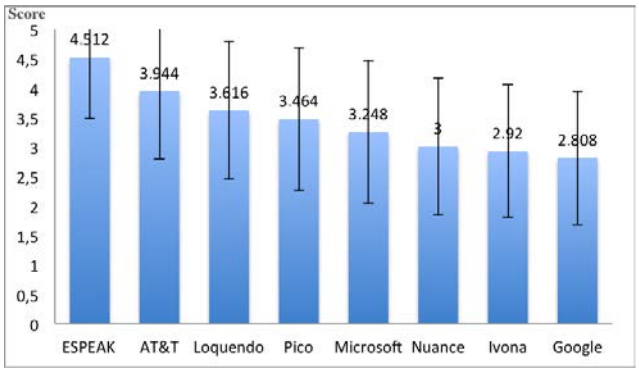

Figure 6. Ranking of TTS systems for *Artificiality*. The vertical axis represents the mean score obtained on the questionnaires, 5 being the maximum. The error bars represent the standard deviation.

4.4. Suitability: Are All TTS Systems Equally Suitable for Each Robot?

This feature tries to investigate which TTS system is perceived as the most suitable for each of the three different social robots presented in Section 3.2. This research question is considered for each robot separately:

- RQ4.1: are all TTS systems equally suitable for Maggie?
- RQ4.2: are all TTS systems equally suitable for Mbot?
- RQ4.3: are all TTS systems equally suitable for Mini?

Therefore, a one-way repeated measures ANOVA is conducted, using the scores obtained for each robot, to determine whether there are significant differences between the TTS systems in terms of their suitability for a specific robot.

4.4.1. Maggie

According to the results obtained for *Maggie*—Wilks' Lambda = 0.52, $F(7, 116) = 15.61, p < 0.001$—there are significant differences between the TTS systems. For this reason, we can say that not all the TTS systems are equally suitable for Maggie. Table 1 shows the descriptive statistics and Figure A4 presents the pairwise comparison table. In this figure, it can be observed that the most suitable one is *Google* although *Ivona*, *Loquendo*, and *Nuance* obtain similar results, $p > 0.112$. On the other hand, the worst evaluated TTS systems, being significantly different from *Google* ($p < 0.05$), are *Espeak* and *Pico* (see Figure 7).

Table 1. Descriptive statistics for *Suitability* of each TTS system for *Maggie*.

Maggie	*M*	*SD*	*N*
AT&T	3.27	1.21	44
Espeak	2.16	1.16	44
Google	4.02	0.90	44
Ivona	3.98	0.85	44
Loquendo	3.32	1.23	44
Microsoft	2.79	1.25	44
Nuance	3.52	0.99	44
Pico	2.64	0.99	44

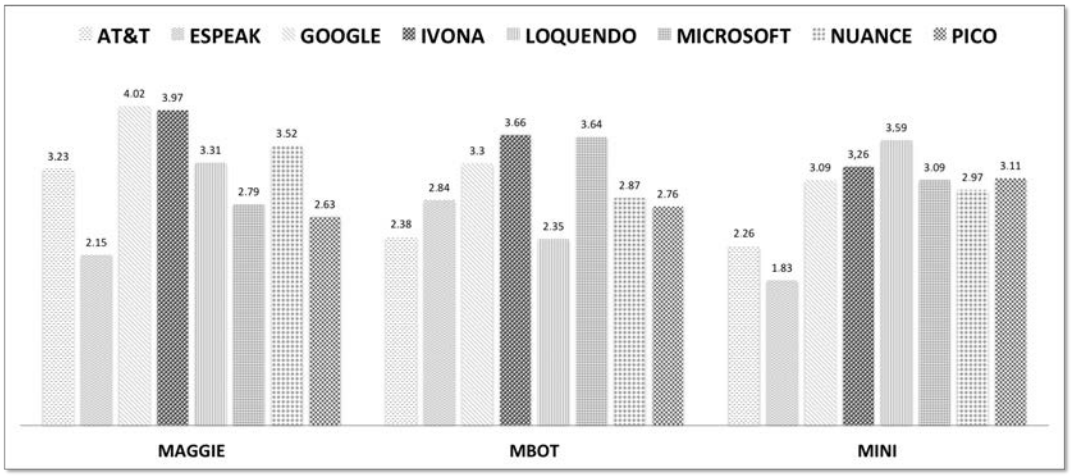

Figure 7. Ranking of suitability, grouped by robot—the TTS systems preferred for each robot. Five is the maximum score.

4.4.2. Mbot

In the case of *Mbot*, the results of the multivatiate test, $WL = 0.73, F(7, 116) = 6.22, p < 0.001$, also confirm that not all the TTS systems are perceived as equally suitable for Mbot. Table 2 presents the values of the mean and the standard deviation, and the pairwise comparison table is shown in Figure A5. For this robot, the favourite one is *Ivona*, with *Microsoft* and *Google* the second and the third best evaluated TTS systems. These three systems obtained similar results, $p = 1$. On the opposite side, *AT&T* and *Loquendo* are the TTS systems considered as significantly not well-suited for this robot ($p < 0.05$), in comparison to *Ivona*, since they were the worst evaluated ones (see Figure 7).

Table 2. Descriptive statistics for *Suitability* of each TTS system for *Mbot*.

Mbot	*M*	*SD*	*N*
AT&T	2.38	1.18	39
Espeak	2.85	1.29	39
Google	3.31	1.28	39
Ivona	3.67	1.22	39
Loquendo	2.36	1.20	39
Microsoft	3.64	0.96	39
Nuance	2.87	1.13	39
Pico	2.77	1.04	39

4.4.3. Mini

Finally, for *Mini*, $WL = 0.64, F(7, 116) = 9.21, p < 0.001$, so, again, there are significant differences between the TTS systems in terms of their suitability for this robot. The descriptive statistics are presented in Table 3. According to these results, it seems that there is no clear 'winner' on this occasion. In the pairwise comparison table, Figure A6, it is observed that just one TTS system, *Espeak*, is significantly different from the rest of the systems except for *AT&T*. These two systems are perceived as the least suitable for *Mini*, so, although we cannot affirm that all the TTS systems are equally suitable for this robot, there are no significant differences between the other ones ($p = 1$). This means that there are six TTS systems equally suitable for *Mini*.

In Figure 7, we can observe that, as has been said, although the preferred TTS system is *Loquendo*, the majority of the TTS systems obtained similar results: there are no significant differences between the TTS systems except for *Espeak* and *AT&T*, which were the worst evaluated ones.

Table 3. Descriptive statistics for *Suitability* of each TTS system for *Mini*.

Mini	*M*	*SD*	*N*
AT&T	2.26	0.96	42
Espeak	1.83	1.03	42
Google	3.09	1.12	42
Ivona	3.26	1.23	42
Loquendo	3.59	1.31	42
Microsoft	3.09	1.26	42
Nuance	2.98	1.35	42
Pico	3.12	1.23	42

4.5. Correlations between the Four Features Analyzed

To complete this study, we intended to analyze the correlations between the four features using the Pearson product–moment correlation coefficient. To do so, we performed a preliminary analysis to prove the conditions of normality, linearity and homoscedasticity. The test showed a strong positive correlation between three of the features: *intelligibility*, *expressiveness*, and *suitability* ($r > 0.476$, $p < 0.01$). Additionally, there is a negative correlation between the previous features and *artificiality* ($r < -0.225$, $p < 0.01$) as shown in Table 4.

Table 4. Pearson product-moment correlations between the four features analyzed.

Measures	*1*	*2*	*3*	*4*
(1) Intelligibility				
(2) Artificiality	−0.301			
(3) Suitability	0.476	−0.225		
(4) Expressiveness	0.548	−0.358	0.540	

It means that those questions related to *intelligibility*, *expressiveness*, and *suitability* are directly correlated. The cause could correspond to the following reasons: (i) all questions are related to the same feature or, at least, this is what participants have perceived; or (ii) there is a real relation between the analyzed features. In our opinion, this could be the actual cause. Considering the second assumption, we can infer that, if a TTS system is perceived as intelligible, it will also be perceived as expressive and, consequently, these systems will tend to be preferred for a social robot.

5. Discussion and Conclusions

In this work, we have presented a comparison of eight TTS systems considering four features: *intelligibility*, how clear the voice of the robot is; *expressiveness*, how monotonous the voice is; *artificiality*, how "robotic" the robot voice is; and *suitability*, how adequate the voice is for a robot. The first two features are usually included in these kinds of studies as the aspects to be optimized. Additionally, we have included the last two, since, in social robotics, it is important to analyze how natural and suitable for the robot the voice is perceived. The tests have been carried out after integrating these systems into three social robots.

In total, 125 participants evaluated all features for each TTS system, but each participant just considered one of the social robots. After that, we conducted a statistical analysis to see if there were significant differences in the results obtained by each TTS. The method used was a one-way repeated measure ANOVA. Regarding RQ1, RQ2, and RQ3, the statistical analysis shows that there are differences in terms of *intelligibility*, *expressiveness*, and *artificiality* for the TTS systems. This allows establishing a comparison between the systems, indicating which one is the most and least intelligible, expressive, and artificial. Moreover, the analysis indicates that a direct correlation exists between the features *intelligibility* and *expressiveness* and an inverse correlation between these ones and *artificiality*.

In general, the TTS system provided by Google is the best rated one with respect to *intelligibility* and *expressiveness*, being perceived as the least artificial. Finally, *Espeak* is at the end of the ranking, with user perception of being robotic, monotonous, and not clear.

In relation to RQ4, we observe that, although for each robot there are significant differences between the TTS systems, we cannot conclude that there is just one most suitable TTS system for each robot. In fact, there is a set of TTS systems preferred for each robot—for *Maggie*: *Google*, *Ivona*, *Nuance*, and *Loquendo*; for *Mbot*: *Ivona*, *Microsoft*, and *Google*; for *Mini*: *Loquendo*, *Ivona*, *Pico*, *Google*, *Microsoft*, and *Nuance*.

Considering the results obtained for this feature, we can make the following observations:

- For our three social robots, the most suitable TTS systems overall are *Google* and *Ivona*. In fact, *Ivona* has been, in all cases, the second best rated (with no significant differences from the first and the third ones). Therefore, this TTS system can be a good selection for these robots.
- In relation to the less suitable TTS systems, it is interesting to note that, for *Maggie* and *Mini*, the worst evaluated system is *Espeak* (it is significantly different from the most suitable one ($p < 0.05$)). On the contrary, this TTS system is not perceived as the least suitable one for *Mbot*.

  One reason could be that *Maggie* and *Mini* have more physical similarities between them (*Mini* is a small version of *Maggie*) than with *Mbot*. Another reason could be related to gender issues. One aspect about the TTS systems that has not been considered until now is the gender of the synthesized voice. This characteristic may seem to be not very relevant at first, but, considering that we give names to the robots, which people can associate with the feminine or masculine gender, this feature must be considered in order to evaluate the suitability of a particular voice with a specific robot. All TTS systems have been tested using a feminine voice except for *Espeak*, which uses a masculine voice. According to our own experience, people tend to refer to *Maggie* and *Mini* as feminine, and to *Mbot* as masculine. Therefore, it is logical that this TTS system is perceived as less suitable for *Maggie* and *Mini*, and not so unsuitable for *Mbot*.
- In general, the TTS systems that are evaluated as the most 'robotic' ones (*Espeak*, *AT&T*, and *Pico*) are also considered as less suitable for the robots. This seems to be a contradiction, but, it must be noted that these TTS systems are also the ones that were evaluated as the less clearly understood by the participants (*intelligibility*).

Limitations and Lessons Learned

The work presented in this paper has some limitations. First of all, the validity of the analysis might be influenced by the language used in the experiments: Spanish. Although this may not be a limitation per se, we limited the study to TTS system that offered that specific language. Therefore, we have missed other interesting TTS systems.

Another limitation, also related to the selection process, is that another reason to choose these eight systems was their price. As in the previous point, this may cause some good TTS systems (maybe better than the ones considered in this paper) to have been discarded.

In relation to the *suitability* feature, just three social robots were used, and, moreover, they may have some resemblance to each other: all of them have a head, eyes, similar colors, etc. This fact may explain the results and conclusions obtained in RQ4: although there are some TTS systems clearly not suitable for the robots, when selecting the most suitable one, we do not have a clear winner for each robot.

It should be noted that the participants filled the questionnaires after watching a video of the robots speaking instead of directly interacting with the robots. This limitation presented an important advantage to this study, allowing for reaching a broader number of participants. We are aware that some bias may have been introduced due to this limitation associated with the lack of interaction. In addition, the sounds registered may have been affected by some constraints such as our microphones when recording the utterances, the audio encoding in the recordings, the recording distance and

position with respect to the robot, and the sound equipment of the participants. In this regard, we acknowledge that using videos for the evaluation could have introduced some bias due to the lack of direct interactions with the robot and the system chosen for reproducing the sounds. The quality of the voice perceived by participants could be affected by some aspects as the microphone used to collect and record the audio; the audio codec used in the video; the distance and position regarding the robot; and the sound equipment used by the volunteers. For the first limitations, we made an effort to make sure that the recordings were made from the same position with respect to the robots and TTS systems and with a high-quality recording system. In addition, the sound system used by the participants in the experiments was an aspect in which we had no control.

Finally, another factor that should be taken into account is that the name of the TTS system is said in the videos. Although *Google* has a very good performance objectively, maybe participants were influenced by the name, since it is a well known name product (authority bias). In this sense, the order in which each user listens to the utterance could also be affected by the comparison bias since the users evaluating TTS systems for each robot have heard the utterances in the same order.

Author Contributions: All authors have actively contributed to the elaboration of the manuscript, more particularly F.A.M. has performed the integration of the TTS systems in the robot architecture, Á.C.-G. and M.Á.S. have focused on the statistical analysis, M.M. and J.C.C. on performing the test scenario and collect the necessary data.

Funding: The research leading to these results has received funding from the projects: "Development of social robots to help seniors with cognitive impairment (ROBSEN)", funded by the Ministerio de Economia y Competitividad; "RoboCity2030-DIH-CM", funded by Comunidad de Madrid and co-funded by Structural Funds of the EU; "Robots Sociales para estimulación física, cognitiva y afectiva de mayores (ROSES)" funded by Agencia Estatal de Investigación (AEI)

Acknowledgments: Give thanks to all the entities that have financed part of this research, as well as, to all the users who have wanted to participate and contribute to the development of this work.

Conflicts of Interest: The authors of this paper certify that they have NO affiliations with or involvement in any organization or entity with any financial interest, or non-financial interest in the subject matter or materials discussed in this manuscript.

Appendix A

Pairwise Comparisons

(I) TTS	(J) TTS	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
AT&T	Espeak	-,004	,099	1,000	-,319	,311
	Google	-2,505*	,091	,000	-2,797	-2,213
	Ivona	-2,370*	,097	,000	-2,681	-2,059
	Loquendo	-1,851*	,105	,000	-2,187	-1,515
	Microsoft	-1,488*	,114	,000	-1,851	-1,125
	Nuance	-1,644*	,112	,000	-2,003	-1,286
	Pico	-1,123*	,111	,000	-1,478	-,769
Espeak	AT&T	,004	,099	1,000	-,311	,319
	Google	-2,501*	,098	,000	-2,814	-2,189
	Ivona	-2,366*	,100	,000	-2,687	-2,045
	Loquendo	-1,847*	,106	,000	-2,185	-1,509
	Microsoft	-1,484*	,116	,000	-1,854	-1,114
	Nuance	-1,640*	,128	,000	-2,051	-1,230
	Pico	-1,120*	,114	,000	-1,484	-,755
Google	AT&T	2,505*	,091	,000	2,213	2,797
	Espeak	2,501*	,098	,000	2,189	2,814
	Ivona	,135	,050	,228	-,025	,295
	Loquendo	,654*	,089	,000	,370	,938
	Microsoft	1,017*	,090	,000	,730	1,304
	Nuance	,861*	,098	,000	,548	1,173
	Pico	1,381*	,086	,000	1,107	1,655
Ivona	AT&T	2,370*	,097	,000	2,059	2,681
	Espeak	2,366*	,100	,000	2,045	2,687
	Google	-,135	,050	,228	-,295	,025
	Loquendo	,519*	,091	,000	,227	,811
	Microsoft	,882*	,083	,000	,616	1,148
	Nuance	,726*	,095	,000	,422	1,029
	Pico	1,246*	,089	,000	,963	1,530
Loquendo	AT&T	1,851*	,105	,000	1,515	2,187
	Espeak	1,847*	,106	,000	1,509	2,185
	Google	-,654*	,089	,000	-,938	-,370
	Ivona	-,519*	,091	,000	-,811	-,227
	Microsoft	,363*	,100	,012	,042	,684
	Nuance	,206	,115	1,000	-,162	,574
	Pico	,727*	,106	,000	,388	1,066
Microsoft	AT&T	1,488*	,114	,000	1,125	1,851
	Espeak	1,484*	,116	,000	1,114	1,854
	Google	-1,017*	,090	,000	-1,304	-,730
	Ivona	-,882*	,083	,000	-1,148	-,616
	Loquendo	-,363*	,100	,012	-,684	-,042
	Nuance	-,157	,115	1,000	-,522	,209
	Pico	,364*	,111	,039	,009	,719
Nuance	AT&T	1,644*	,112	,000	1,286	2,003
	Espeak	1,640*	,128	,000	1,230	2,051
	Google	-,861*	,098	,000	-1,173	-,548
	Ivona	-,726*	,095	,000	-1,029	-,422
	Loquendo	-,206	,115	1,000	-,574	,162
	Microsoft	,157	,115	1,000	-,209	,522
	Pico	,521*	,107	,000	,179	,863
Pico	AT&T	1,123*	,111	,000	,769	1,478
	Espeak	1,120*	,114	,000	,755	1,484
	Google	-1,381*	,086	,000	-1,655	-1,107
	Ivona	-1,246*	,089	,000	-1,530	-,963
	Loquendo	-,727*	,106	,000	-1,066	-,388
	Microsoft	-,364*	,111	,039	-,719	-,009
	Nuance	-,521*	,107	,000	-,863	-,179

Based on estimated marginal means

*. The mean difference is significant at the

b. Adjustment for multiple comparisons: Bonferroni.

Figure A1. Pairwise comparisons for *Intelligibility*. Pairs of TTS systems with significative differences are highlighted in yellow. Note that due to language configuration of the system the decimal part is delimited by a comma.

Pairwise Comparisons

(I) TTS	(J) TTS	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
AT&T	Espeak	,238	,116	1,000	-,131	,608
	Google	-1,705*	,116	,000	-2,077	-1,334
	Ivona	-1,120*	,125	,000	-1,518	-,721
	Loquendo	-1,120*	,129	,000	-1,533	-,707
	Microsoft	-,928*	,123	,000	-1,322	-,533
	Nuance	-,878*	,131	,000	-1,297	-,460
	Pico	-,320	,117	,206	-,694	,055
Espeak	AT&T	-,238	,116	1,000	-,608	,131
	Google	-1,944*	,120	,000	-2,326	-1,562
	Ivona	-1,358*	,121	,000	-1,744	-,972
	Loquendo	-1,358*	,125	,000	-1,758	-,959
	Microsoft	-1,166*	,115	,000	-1,532	-,800
	Nuance	-1,117*	,141	,000	-1,567	-,667
	Pico	-,558*	,122	,000	-,947	-,168
Google	AT&T	1,705*	,116	,000	1,334	2,077
	Espeak	1,944*	,120	,000	1,562	2,326
	Ivona	,586*	,097	,000	,276	,895
	Loquendo	,586*	,110	,000	,234	,937
	Microsoft	,778*	,117	,000	,404	1,151
	Nuance	,827*	,123	,000	,434	1,219
	Pico	1,386*	,110	,000	1,035	1,737
Ivona	AT&T	1,120*	,125	,000	,721	1,518
	Espeak	1,358*	,121	,000	,972	1,744
	Google	-,586*	,097	,000	-,895	-,276
	Loquendo	,000	,122	1,000	-,391	,391
	Microsoft	,192	,103	1,000	-,138	,522
	Nuance	,241	,120	1,000	-,142	,624
	Pico	,800*	,110	,000	,450	1,151
Loquendo	AT&T	1,120*	,129	,000	,707	1,533
	Espeak	1,358*	,125	,000	,959	1,758
	Google	-,586*	,110	,000	-,937	-,234
	Ivona	,000	,122	1,000	-,391	,391
	Microsoft	,192	,116	1,000	-,180	,564
	Nuance	,241	,135	1,000	-,191	,674
	Pico	,800*	,126	,000	,399	1,201
Microsoft	AT&T	,928*	,123	,000	,533	1,322
	Espeak	1,166*	,115	,000	,800	1,532
	Google	-,778*	,117	,000	-1,151	-,404
	Ivona	-,192	,103	1,000	-,522	,138
	Loquendo	-,192	,116	1,000	-,564	,180
	Nuance	,049	,108	1,000	-,296	,395
	Pico	,608*	,107	,000	,266	,950
Nuance	AT&T	,878*	,131	,000	,460	1,297
	Espeak	1,117*	,141	,000	,667	1,567
	Google	-,827*	,123	,000	-1,219	-,434
	Ivona	-,241	,120	1,000	-,624	,142
	Loquendo	-,241	,135	1,000	-,674	,191
	Microsoft	-,049	,108	1,000	-,395	,296
	Pico	,559*	,101	,000	,235	,883
Pico	AT&T	,320	,117	,206	-,055	,694
	Espeak	,558*	,122	,000	,168	,947
	Google	-1,386*	,110	,000	-1,737	-1,035
	Ivona	-,800*	,110	,000	-1,151	-,450
	Loquendo	-,800*	,126	,000	-1,201	-,399
	Microsoft	-,608*	,107	,000	-,950	-,266
	Nuance	-,559*	,101	,000	-,883	-,235

Based on estimated marginal means

*. The mean difference is significant at the

b. Adjustment for multiple comparisons: Bonferroni.

Figure A2. Pairwise comparisons for *Expressiveness*. Pairs of TTS systems with significative differences are highlighted in yellow. Note that due to language configuration of the system the decimal part is delimited by a comma.

Pairwise Comparisons

(I) TTS	(J) TTS	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
AT&T	Espeak	-,561*	,101	,000	-,883	-,239
	Google	1,138*	,127	,000	,732	1,544
	Ivona	1,016*	,141	,000	,565	1,467
	Loquendo	,329	,132	,383	-,091	,749
	Microsoft	,703*	,140	,000	,255	1,150
	Nuance	,941*	,149	,000	,464	1,419
	Pico	,478*	,138	,020	,038	,917
Espeak	AT&T	,561*	,101	,000	,239	,883
	Google	1,699*	,128	,000	1,291	2,107
	Ivona	1,577*	,131	,000	1,159	1,994
	Loquendo	,890*	,133	,000	,464	1,316
	Microsoft	1,264*	,135	,000	,834	1,694
	Nuance	1,502*	,139	,000	1,057	1,947
	Pico	1,038*	,125	,000	,638	1,438
Google	AT&T	-1,138*	,127	,000	-1,544	-,732
	Espeak	-1,699*	,128	,000	-2,107	-1,291
	Ivona	-,122	,111	1,000	-,476	,232
	Loquendo	-,809*	,134	,000	-1,237	-,381
	Microsoft	-,435*	,134	,043	-,865	-,006
	Nuance	-,197	,132	1,000	-,618	,224
	Pico	-,661*	,140	,000	-1,107	-,214
Ivona	AT&T	-1,016*	,141	,000	-1,467	-,565
	Espeak	-1,577*	,131	,000	-1,994	-1,159
	Google	,122	,111	1,000	-,232	,476
	Loquendo	-,687*	,127	,000	-1,093	-,281
	Microsoft	-,313	,127	,412	-,717	,091
	Nuance	-,075	,122	1,000	-,465	,316
	Pico	-,539*	,133	,003	-,963	-,114
Loquendo	AT&T	-,329	,132	,383	-,749	,091
	Espeak	-,890*	,133	,000	-1,316	-,464
	Google	,809*	,134	,000	,381	1,237
	Ivona	,687*	,127	,000	,281	1,093
	Microsoft	,374	,136	,195	-,061	,809
	Nuance	,612*	,133	,000	,186	1,038
	Pico	,148	,114	1,000	-,217	,514
Microsoft	AT&T	-,703*	,140	,000	-1,150	-,255
	Espeak	-1,264*	,135	,000	-1,694	-,834
	Google	,435*	,134	,043	,006	,865
	Ivona	,313	,127	,412	-,091	,717
	Loquendo	-,374	,136	,195	-,809	,061
	Nuance	,238	,121	1,000	-,148	,625
	Pico	-,225	,129	1,000	-,637	,186
Nuance	AT&T	-,941*	,149	,000	-1,419	-,464
	Espeak	-1,502*	,139	,000	-1,947	-1,057
	Google	,197	,132	1,000	-,224	,618
	Ivona	,075	,122	1,000	-,316	,465
	Loquendo	-,612*	,133	,000	-1,038	-,186
	Microsoft	-,238	,121	1,000	-,625	,148
	Pico	-,464*	,126	,009	-,865	-,062
Pico	AT&T	-,478*	,138	,020	-,917	-,038
	Espeak	-1,038*	,125	,000	-1,438	-,638
	Google	,661*	,140	,000	,214	1,107
	Ivona	,539*	,133	,003	,114	,963
	Loquendo	-,148	,114	1,000	-,514	,217
	Microsoft	,225	,129	1,000	-,186	,637
	Nuance	,464*	,126	,009	,062	,865

Based on estimated marginal means

*. The mean difference is significant at the

b. Adjustment for multiple comparisons: Bonferroni.

Figure A3. Pairwise comparisons for *Artificiality*. Pairs of TTS systems with significative differences are highlighted in yellow. Note that due to language configuration of the system the decimal part is delimited by a comma.

Pairwise Comparisons

TTS	TTS	Mean Difference	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
AT&T	Espeak	1,114*	,190	,000	,506	1,721
	Google	-,750*	,229	,038	-1,481	-,019
	Ivona	-,705	,225	,060	-1,422	,013
	Loquendo	-,045	,247	1,000	-,835	,744
	Microsoft	,477	,231	1,000	-,260	1,215
	Nuance	-,250	,219	1,000	-,948	,448
	Pico	,636*	,185	,023	,044	1,229
Espeak	AT&T	-1,114*	,190	,000	-1,721	-,506
	Google	-1,864*	,237	,000	-2,620	-1,107
	Ivona	-1,818*	,228	,000	-2,546	-1,091
	Loquendo	-1,159*	,254	,000	-1,969	-,349
	Microsoft	-,636	,233	,201	-1,380	,107
	Nuance	-1,364*	,261	,000	-2,198	-,529
	Pico	-,477	,227	1,000	-1,202	,247
Google	AT&T	,750*	,229	,038	,019	1,481
	Espeak	1,864*	,237	,000	1,107	2,620
	Ivona	,045	,162	1,000	-,472	,563
	Loquendo	,705	,240	,112	-,062	1,472
	Microsoft	1,227*	,220	,000	,524	1,931
	Nuance	,500	,210	,529	-,171	1,171
	Pico	1,386*	,224	,000	,671	2,101
Ivona	AT&T	,705	,225	,060	-,013	1,422
	Espeak	1,818*	,228	,000	1,091	2,546
	Google	-,045	,162	1,000	-,563	,472
	Loquendo	,659	,257	,326	-,163	1,481
	Microsoft	1,182*	,192	,000	,569	1,794
	Nuance	,455	,207	,838	-,206	1,115
	Pico	1,341*	,209	,000	,673	2,009
Loquendo	AT&T	,045	,247	1,000	-,744	,835
	Espeak	1,159*	,254	,000	,349	1,969
	Google	-,705	,240	,112	-1,472	,062
	Ivona	-,659	,257	,326	-1,481	,163
	Microsoft	,523	,258	1,000	-,301	1,347
	Nuance	-,205	,246	1,000	-,990	,581
	Pico	,682	,221	,070	-,023	1,387
Microsoft	AT&T	-,477	,231	1,000	-1,215	,260
	Espeak	,636	,233	,201	-,107	1,380
	Google	-1,227*	,220	,000	-1,931	-,524
	Ivona	-1,182*	,192	,000	-1,794	-,569
	Loquendo	-,523	,258	1,000	-1,347	,301
	Microsoft	-,727*	,198	,010	-1,359	-,095
	Pico	,159	,195	1,000	-,462	,781
Nuance	AT&T	,250	,219	1,000	-,448	,948
	Espeak	1,364*	,261	,000	,529	2,198
	Google	-,500	,210	,529	-1,171	,171
	Ivona	-,455	,207	,838	-1,115	,206
	Loquendo	,205	,246	1,000	-,581	,990
	Microsoft	,727*	,198	,010	,095	1,359
	Pico	,886*	,182	,000	,306	1,467
Pico	AT&T	-,636*	,185	,023	-1,229	-,044
	Espeak	,477	,227	1,000	-,247	1,202
	Google	-1,386*	,224	,000	-2,101	-,671
	Ivona	-1,341*	,209	,000	-2,009	-,673
	Loquendo	-,682	,221	,070	-1,387	,023
	Microsoft	-,159	,195	1,000	-,781	,462
	Nuance	-,886*	,182	,000	-1,467	-,306

Based on estimated marginal means
*. The mean difference is significant at the
b. Adjustment for multiple comparisons: Bonferroni.

Figure A4. Pairwise comparisons for *Suitability* and the robot *Maggie*. Pairs of TTS systems with significative differences are highlighted in yellow. Note that due to language configuration of the system the decimal part is delimited by a comma.

Pairwise Comparisons

TTS	TTS	Mean Difference	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
AT&T	Espeak	-,462	,202	,676	-1,107	,184
	Google	-,923*	,243	,006	-1,699	-,147
	Ivona	-1,282*	,239	,000	-2,044	-,520
	Loquendo	,026	,263	1,000	-,813	,865
	Microsoft	-1,256*	,245	,000	-2,040	-,473
	Nuance	-,487	,232	1,000	-1,229	,255
	Pico	-,385	,197	1,000	-1,014	,244
Espeak	AT&T	,462	,202	,676	-,184	1,107
	Google	-,462	,252	1,000	-1,265	,342
	Ivona	-,821*	,242	,026	-1,593	-,048
	Loquendo	,487	,269	1,000	-,373	1,348
	Microsoft	-,795*	,247	,047	-1,584	-,005
	Nuance	-,026	,277	1,000	-,912	,861
	Pico	,077	,241	1,000	-,693	,847
Google	AT&T	,923*	,243	,006	,147	1,699
	Espeak	,462	,252	1,000	-,342	1,265
	Ivona	-,359	,172	1,000	-,908	,190
	Loquendo	,949*	,255	,008	,134	1,763
	Microsoft	-,333	,234	1,000	-1,081	,414
	Nuance	,436	,223	1,000	-,277	1,149
	Pico	,538	,238	,709	-,221	1,298
Ivona	AT&T	1,282*	,239	,000	,520	2,044
	Espeak	,821*	,242	,026	,048	1,593
	Google	,359	,172	1,000	-,190	,908
	Loquendo	1,308*	,273	,000	,435	2,181
	Microsoft	,026	,204	1,000	-,625	,676
	Nuance	,795*	,220	,012	,093	1,497
	Pico	,897*	,222	,003	,188	1,607
Loquendo	AT&T	-,026	,263	1,000	-,865	,813
	Espeak	-,487	,269	1,000	-1,348	,373
	Google	-,949*	,255	,008	-1,763	-,134
	Ivona	-1,308*	,273	,000	-2,181	-,435
	Microsoft	-1,282*	,274	,000	-2,157	-,407
	Nuance	-,513	,261	1,000	-1,347	,322
	Pico	-,410	,234	1,000	-1,159	,338
Microsoft	AT&T	1,256*	,245	,000	,473	2,040
	Espeak	,795*	,247	,047	,005	1,584
	Google	,333	,234	1,000	-,414	1,081
	Ivona	-,026	,204	1,000	-,676	,625
	Loquendo	1,282*	,274	,000	,407	2,157
	Nuance	,769*	,210	,010	,098	1,440
	Pico	,872*	,207	,001	,212	1,532
Nuance	AT&T	,487	,232	1,000	-,255	1,229
	Espeak	,026	,277	1,000	-,861	,912
	Google	-,436	,223	1,000	-1,149	,277
	Ivona	-,795*	,220	,012	-1,497	-,093
	Loquendo	,513	,261	1,000	-,322	1,347
	Microsoft	-,769*	,210	,010	-1,440	-,098
	Pico	,103	,193	1,000	-,514	,719
Pico	AT&T	,385	,197	1,000	-,244	1,014
	Espeak	-,077	,241	1,000	-,847	,693
	Google	-,538	,238	,709	-1,298	,221
	Ivona	-,897*	,222	,003	-1,607	-,188
	Loquendo	,410	,234	1,000	-,338	1,159
	Microsoft	-,872*	,207	,001	-1,532	-,212
	Nuance	-,103	,193	1,000	-,719	,514

Based on estimated marginal means
*. The mean difference is significant at the
b. Adjustment for multiple comparisons: Bonferroni.

Figure A5. Pairwise comparisons for *Suitability* and the robot *Mbot*. Pairs of TTS systems with significative differences are highlighted in yellow. Note that due to language configuration of the system the decimal part is delimited by a comma.

Pairwise Comparisons

TTS	TTS	Mean Difference	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
AT&T	Espeak	,429	,195	,831	-,194	1,051
	Google	-,833*	,234	,015	-1,581	-,086
	Ivona	-1,000*	,230	,001	-1,735	-,265
	Loquendo	-1,333*	,253	,000	-2,142	-,525
	Microsoft	-,833*	,236	,017	-1,588	-,078
	Nuance	-,714	,224	,050	-1,429	,000
	Pico	-,857*	,190	,000	-1,463	-,251
Espeak	AT&T	-,429	,195	,831	-1,051	,194
	Google	-1,262*	,242	,000	-2,036	-,488
	Ivona	-1,429*	,233	,000	-2,173	-,684
	Loquendo	-1,762*	,260	,000	-2,591	-,933
	Microsoft	-1,262*	,238	,000	-2,023	-,501
	Nuance	-1,143*	,267	,001	-1,997	-,289
	Pico	-1,286*	,232	,000	-2,027	-,544
Google	AT&T	,833*	,234	,015	,086	1,581
	Espeak	1,262*	,242	,000	,488	2,036
	Ivona	-,167	,166	1,000	-,696	,363
	Loquendo	-,500	,246	1,000	-1,285	,285
	Microsoft	4,441E-16	,225	1,000	-,720	,720
	Nuance	,119	,215	1,000	-,568	,806
	Pico	-,024	,229	1,000	-,756	,708
Ivona	AT&T	1,000*	,230	,001	,265	1,735
	Espeak	1,429*	,233	,000	,684	2,173
	Google	,167	,166	1,000	-,363	,696
	Loquendo	-,333	,263	1,000	-1,175	,508
	Microoft	,167	,196	1,000	-,460	,794
	Nuance	,286	,212	1,000	-,391	,962
	Pico	,143	,214	1,000	-,541	,826
Loquend o	AT&T	1,333*	,253	,000	,525	2,142
	Espeak	1,762*	,260	,000	,933	2,591
	Google	,500	,246	1,000	-,285	1,285
	Ivona	,333	,263	1,000	-,508	1,175
	Microsoft	,500	,264	1,000	-,344	1,344
	Nuance	,619	,252	,430	-,185	1,423
	Pico	,476	,226	1,000	-,245	1,198
Microsoft	AT&T	,833*	,236	,017	,078	1,588
	Espeak	1,262*	,238	,000	,501	2,023
	Google	-4,441E-16	,225	1,000	-,720	,720
	Ivona	-,167	,196	1,000	-,794	,460
	Loquendo	-,500	,264	1,000	-1,344	,344
	Nuance	,119	,202	1,000	-,528	,766
	Pico	-,024	,199	1,000	-,660	,612
Nuance	AT&T	,714	,224	,050	,000	1,429
	Espeak	1,143*	,267	,001	,289	1,997
	Google	-,119	,215	1,000	-,806	,568
	Ivona	-,286	,212	1,000	-,962	,391
	Loquendo	-,619	,252	,430	-1,423	,185
	Microsoft	-,119	,202	1,000	-,766	,528
	Pico	-,143	,186	1,000	-,737	,451
Pico	AT&T	,857*	,190	,000	,251	1,463
	Espeak	1,286*	,232	,000	,544	2,027
	Google	,024	,229	1,000	-,708	,756
	Ivona	-,143	,214	1,000	-,826	,541
	Loquendo	-,476	,226	1,000	-1,198	,245
	Microsoft	,024	,199	1,000	-,612	,660
	Nuance	,143	,186	1,000	-,451	,737

Based on estimated marginal means
*. The mean difference is significant at the
b. Adjustment for multiple comparisons: Bonferroni.

Figure A6. Pairwise comparisons for *Suitability* and the robot *Mini*. Pairs of TTS systems with significative differences are highlighted in yellow. Note that due to language configuration of the system the decimal part is delimited by a comma.

References

1. Van Bezooijen, R.; Pols, L.C. Evaluating text-to-speech systems: Some methodological aspects. *Speech Commun.* **1990**, *9*, 263–270. [CrossRef]
2. Handley, Z. Is text-to-speech synthesis ready for use in computer-assisted language learning? *Speech Commun.* **2009**, *51*, 906–919. [CrossRef]

3. O'Malley, M. Text-to-speech conversion technology. *Computer* **1990**, *23*, 17–23. [CrossRef]
4. Klatt, D.H. Review of text-to-speech conversion for English. *J. Acoust. Soc. Am.* **1987**, *82*, 737. [CrossRef] [PubMed]
5. Pappas, C. Top 10 Text to Speech (TTS) Software for eLearning. 2019. Available online: https://elearningindustry.com/top-10-text-to-speech-tts-software-elearning (accessed on 12 December 2019).
6. Comparison of Speech Synthesizers. 2018. Available online: https://en.wikipedia.org/wiki/Comparison_of_speech_synthesizers (accessed on 12 December 2019).
7. Dutoit, T.; Pagel, V.; Pierret, N.; Bataille, F.; Van der Vrecken, O. The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In Proceedings of the Fourth International Conference on Spoken Language Processing, ICSLP'96, Philadelphia, PA, USA, 3–6 October 1996; Volume 3, pp. 1393–1396.
8. Cao, H.; de Perre, G.V.; Simut, R. Enhancing My Keepon robot: A simple and low-cost solution for robot platform in Human-Robot Interaction studies. In Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication (ROMAN), Edinburgh, UK, 25–29 August 2014; pp. 555–560.
9. Wada, K.; Ikeda, Y.; Inoue, K.; Uehara, R. Development and preliminary evaluation of a caregiver's manual for robot therapy using the therapeutic seal robot Paro. In Proceedings of the 19th International Symposium in Robot and Human Interactive Communication, Viareggio, Italy, 13–15 September 2010; pp. 533–538.
10. Fujita, M. On activating human communications with pet-type robot AIBO. *Proc. IEEE* **2004**, *92*, 1804–1813. [CrossRef]
11. Shamsuddin, S.; Ismail, L.I.; Yussof, H.; Zahari, N.I.; Bahari, S.; Hafizan, H.; Jaffar, A. Humanoid robot NAO: Review of control and motion exploration. In Proceedings of the 2011 IEEE International Conference on Control System, Computing and Engineering, Penang, Malaysia, 25–27 November 2011; pp. 511–516.
12. Lafaye, J.; Gouaillier, D.; Wieber, P.B. Linear model predictive control of the locomotion of Pepper, a humanoid robot with omnidirectional wheels. In Proceedings of the 2014 IEEE-RAS International Conference on Humanoid Robots, Madrid, Spain, 18–20 November 2014; pp. 336–341.
13. Tsagarakis, N.; Metta, G.; Sandini, G. iCub: The design and realization of an open humanoid platform for cognitive and neuroscience research. *Adv. Robot.* **2007**, *21*, 1151–1175. [CrossRef]
14. Metta, G.; Sandini, G.; Vernon, D. The iCub humanoid robot: An open platform for research in embodied cognition. In Proceedings of the PerMIS '08, Workshop on Performance Metrics for Intelligent Systems, Gaithersburg, MD, USA, 19–21 August 2008; pp. 50–56.
15. Group, A. Acapela. 2019. Available online: http://www.acapela-group.com (accessed on 12 December 2019).
16. Kenmochi, H.; Ohshita, H. VOCALOID-commercial singing synthesizer based on sample concatenation. In Proceedings of the INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, 27–31 August 2007; pp. 4009–4010.
17. Kenmochi, H. VOCALOID and Hatsune Miku phenomenon in Japan. In Proceedings of the Interdisciplinary Workshop on Singing Voice, Tokyo, Japan, 1–2 October 2010.
18. Tachibana, M.; Nakaoka, S.; Kenmochi, H. A singing robot realized by a collaboration of VOCALOID and Cybernetic Human HRP-4C. In Proceedings of the Interdisciplinary Workshop on Singing Voice (InterSinging 2010), Tokyo, Japan, 1–2 October 2010.
19. Apple. Siri. 2019. Available online: http://www.apple.com/ios/siri (accessed on 12 December 2019).
20. Google. Google Now. 2019. Available online: https://www.google.com/landing/now (accessed on 12 December 2019).
21. Amazon. Kindle. 2019. Available online: https://kindle.amazon.com (accessed on 12 December 2019).
22. Corporation, M. Cortana. 2019. Available online: http://windows.microsoft.com/es-es/windows-10/getstarted-what-is-cortana (accessed on 12 December 2019).
23. Roehling, S.; MacDonald, B.; Watson, C. Towards expressive speech synthesis in english on a robotic platform. In Proceedings of the Australasian International Conference on Speech Science and Technology, Auckland, New Zealand, 6–8 December 2006; pp. 130–135.
24. Bakhsh, N.K.; Alshomrani, S.; Khan, I. A comparative study of Arabic text-to-speech synthesis systems. *Int. J. Inf. Eng. Electron. Bus.* **2014**, *6*, 27. [CrossRef]
25. Shruthi, G.; Kumar, P. Comparative study of text to speech system for indian language. *Int. J. Adv. Comput. Inf. Technol.* **2012**, *1*, 199–209.

26. Francis, A.; Nusbaum, H. Evaluating the quality of synthetic speech. In *Human Factors and Voice Interactive Systems*; Springer: Boston, MA, USA, 1999; pp. 63–97.
27. Handley, Z.; Hamel, M. Establishing a methodology for benchmarking speech synthesis for computer-assisted language learning (CALL). *Lang. Learn. Technol.* **2005**, *9*, 99–120.
28. ITU-T. Transmission Quality Subjective Opinion Tests. A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices. Available online: https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-P.85-199406-I!!PDF-E&type=items (accessed on 12 December 2019).
29. MOS Scale. 2019. Available online: https://en.wikipedia.org/wiki/Mean_opinion_score (accessed on 12 December 2019).
30. Viswanathan, M. Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale. *Comput. Speech Lang.* **2005**, *19*, 55–83. [CrossRef]
31. King, S. Measuring a decade of progress in text-to-speech. *Loquens* **2014**, *1*, 6. [CrossRef]
32. Alonso-Martin, F. Sistema de Interacción Humano-Robot Basado en Diálogos Multimodales y Adaptables. Ph.D. Thesis, Universidad Carlos III de Madrid, Madrid, Spain, 2014.
33. Alonso-Martín, F.; Castro-González, A.; Luengo, F.; Salichs, M. Augmented Robotics Dialog System for Enhancing Human–Robot Interaction. *Sensors* **2015**, *15*, 15799–15829. [CrossRef] [PubMed]
34. Gonzalez-Pacheco, V.; Ramey, A.; Alonso-Martin, F.; Castro-Gonzalez, A.; Salichs, M.A. Maggie: A Social Robot as a Gaming Platform. *Int. J. Soc. Robot.* **2011**, *3*, 371–381. [CrossRef]
35. Salichs, M.; Barber, R.; Khamis, A.; Malfaz, M.; Gorostiza, J.; Pacheco, R.; Rivas, R.; Corrales, A.; Delgado, E.; Garcia, D. Maggie: A Robotic Platform for Human-Robot Social Interaction. In Proceedings of the 2006 IEEE Conference on Robotics, Automation and Mechatronics, Bangkok, Thailand, 1–3 June 2006; pp. 1–7.
36. Castro-González, Á.; Castillo, J.C.; Alonso-Martín, F.; Olortegui-Ortega, O.V.; González-Pacheco, V.; Malfaz, M.; Salichs, M.A. The Effects of an Impolite vs. a Polite Robot Playing Rock-Paper-Scissors. In Proceedings of the International Conference on Social Robotics, Kansas City, MO, USA, 1–3 November 2016; pp. 306–316.
37. González-Pacheco, V.; Castro-González., Á.; Malfaz, M.; Salichs, M.A. Human-Robot Interaction in the MOnarCH project. In Proceedings of the 13th Robocity2030 Workshop, Madrid, Spain, 11 December 2015; pp. 1–8.
38. Monarch European Project. 2019. Available online: http://monarch-fp7.eu (accessed on 12 December 2019).
39. Google. Google Forms. 2019. Available online: https://www.google.es/intl/es/forms/about (accessed on 12 December 2019).
40. IBM. SPSS. 2019. Available online:http://www-01.ibm.com/software/es/analytics/spss (accessed on 12 December 2019).