Hindawi
Human Behavior and Emerging Technologies
Volume 2023, Article ID 8858268, 15 pages
https://doi.org/10.1155/2023/8858268

Research Article

A Synthetic Voice for an Assistive Conversational Agent: A Survey to Discover Italian Preferences regarding Synthetic Voice's Gender and Quality Level

Marialucia Cuciniello,[1] **Terry Amorese**,[1] **Claudia Greco**,[1] **Zoraida Callejas Carrión**,[2] **Carl Vogel**,[3] **Gennaro Cordasco**,[1] **and Anna Esposito**[1]

[1]*Università degli Studi della Campania "Luigi Vanvitelli", Italy*
[2]*Universidad de Granada, Spain*
[3]*Trinity College Dublin, The University of Dublin, Ireland*

Correspondence should be addressed to Marialucia Cuciniello; marialucia.cuciniello@unicampania.it

Received 23 May 2023; Revised 28 October 2023; Accepted 11 December 2023; Published 28 December 2023

Academic Editor: Mirko Duradoni

Based on a previous investigation, a quantitative study aimed to identify user' preferences towards four synthetic voices of two different quality levels (classified through the sophistication of the synthesizer: low vs. high) is proposed. The voices administered to participants were developed considering two main aspects: the voice quality (high/low) and their gender (male/female). 182 unpaid participants were recruited for the study, divided in four groups according to their age, and therefore classified as adolescents, young adults, middle-aged, and seniors. To collect data regarding each voice, randomly audited by participants, the shortened version of the Virtual Agent Voice Acceptance Questionnaire (VAVAQ) was exploited. Outcomes of the previous study revealed that the voices of high quality, regardless of their gender, received a higher acclaim by all participants examined rather than the corresponding two voices assessed as lower quality. Conversely, findings of the current study suggest that the four new groups of participants involved agreed in showing their strong preference towards the high-quality voice gendered as female compared to all the other considered voices. Regarding the two voices gendered as male, the high-quality one was considered as more original and capable to arouse positive emotional states than the low-quality one. Moreover, the high-quality male voice was judged as more natural than the female low-quality one. Results provide some insights for future directions in the user experience and design field.

1. Introduction

In the domestic and corporate context in recent years, thanks to the rapid diffusion of voice assistants (VA), we have witnessed the advancement and consolidation of a very important issue relating to the use of artificial intelligence (AI) [1]. Voice assistants are considered as voice-controlled intelligent personal assistants (VIPAs) that have become part of common and widespread use such as Amazon Echo, Google, Apple Siri, Microsoft Cortana, and Samsung Bixby [2]. These smart devices can instantaneously elaborate almost any human request, reproducing and simulating natural human communication through natural language processing (NLP) or natural language understanding (NLU) [3].

It is undeniable that these devices have completely revolutionized the way people lead and conceive their lives by interacting and making use of the support of technology not only to make phone calls or send messages to get in touch faster, but thanks to the advances in machine learning, achieved in recent years in particular in neural networks, today it is possible to use voice-based technologies to assist health professionals [4] and provide diagnostic support to seniors in managing their daily routine [5]. Developing new applications for monitoring and managing mental disorders, dedicated to the treatment of chronic and specific conditions and promoting an overall healthier lifestyle, is among the current goals pursued by developers and researchers.

There is a substantial increase in investment in chatbots relating to the healthcare industry as witnessed by devices and applications such as Woebot, Babylon, and ADA Health [6]. Motivating such trends are several studies and randomized trials supporting that conversational agents (CAs) could effectively represent a valuable facilitation in healthcare delivery promoting and achieving positive outcomes in terms of mental health [7].

1.1. Theoretical Background. Although it is important to define and develop new paradigms that allow the automation of the vocal or textual speech of the conversational agent to make interaction with the user as natural as possible, it is equally of crucial importance to explore and identify information relating which features that the synthetic voice of the agent should have to engage end users. This represents a still almost unsolved issue and a challenge at the same time, in the development, deployment, and use of CAs. Therefore, a strong need has emerged in the scientific community to investigate the preferences of potential users, possibly of different age groups, towards some features of the synthetic voice, first of all the preference regarding the voice's gender. Some evidence supports the thesis that there is an innate human preference for female voices [8].

Unfortunately, results achieved so far, in addition to being ambiguous, are scarce, denoting a strong need to investigate the vocal gender preferences for digital artificial intelligence assistants by potential users. Some authors [9] describe gender stereotypes as overgeneralized popular beliefs about traits presumably typical of each gender. Occupationally, women tend to fill people-oriented service roles as opposed to things denoting competitive assignments, traditionally reserved exclusively for men [10]. Hentschel et al. [11] report that there are traits stereotypically associated with men and others stereotypically associated with women. The ambitious, assertive, competent, dominant, and independent temperament is traditionally referred to the man while the caring, emotional, friendly, kind, and understanding attitude is typically related to the woman. Evidence seems to show that these stereotypes are generalized and applied also to synthetic voices and disembodied chatbots [12].

A study [13] explored the preferences of individuals from different age groups and genders regarding both natural and computer-generated synthetic speech within a variety of communication contexts. The study involved listeners of different ages and genders, with each age group consisting of five males and five females. The listeners were asked to assess their preferences for twelve distinct voices, including four natural voices and eight synthetic voices, using a 5-point Likert scale. These preferences were evaluated within six specific communication contexts, which were determined by the intended user of the voice: adult female, adult male, child female, child male, computer, and self. Study's findings indicated that, among the synthetic voices, the Smoothtalker 3.0 male and RealVoice female voices received the highest ratings from the participants. However, there were consistently significant differences in the ratings when comparing natural and synthetic speech. This suggests that listeners had distinct preferences for these two types of voices in various contexts. The results of this investigation have raised several important issues concerning how age and gender appropriateness collectively influence the perception of natural and synthetic speech.

Another study addressing the issue of the impact that gender could have on synthetic voice perception [14] investigated whether the way people perceive human speech and computer-generated text-to-speech (TTS) is influenced by both the voice's gender and the listener's gender. To do this, participants were exposed to a convincing argument delivered by either a female or male human voice or a synthetic voice. The researchers then assessed the participants' attitude change and their evaluations of various speech qualities. Results of the study revealed that human female voices were generally preferred over synthetic female voices, indicating a preference for authenticity in female voices. In contrast, male synthetic voices were found to be more appealing than female synthetic voices in some cases. The level of persuasion was similar for both human and synthetic voices, highlighting their comparable effectiveness in conveying persuasive messages. The study suggested that gender-related stereotypes and expectations may apply similarly to both human and synthetic voices.

Another study [15] addresses the topic of the use of technology-based warnings, specifically those involving speech-warning statements, which can be personalized for different users and situations. The study is aimed at helping select a synthesized voice for subsequent personalized technology-based warnings delivered through virtual reality. Participants evaluated different voices that had been altered in pitch and ranked their preferences. The results showed that high-pitched female voices were the most preferred, and these voices also scored the highest in the evaluation.

As regards as the quality of synthesized voices, we tend to take for granted that the higher the quality of the voice, the greater the perceived pleasantness, but what is rarely taken into account is the context. In this regard, an interesting paper by [16] argues against the idea of a single "neutral" or "perfectly natural" speaking style as a reference for evaluating synthetic speech, challenging the common assumption that human-read speech serves as the gold standard. They suggest that the appropriateness of a speaking style within a particular context should be the primary measure of its suitability. They relate this idea to issues in human-machine interaction, like the "uncanny valley," where user expectations may not align with the machine's expression.

Some studies show that users may prefer a more "robot-like" synthetic voice in specific contexts, while others indicate that human voices are preferred in more complex tasks. There is a study [17] focusing on the importance that the influence of synthetic voices' quality could have on users' assessment. Their primary objective was to test the perceptual aspects of different voices concerning essential factors in human-computer interaction, specifically users' expectations and acceptance. The study focused on investigating the influence of synthetic voices' quality and gender on user preferences. The sample was composed of 40 participants from Northern Ireland, divided into two groups: individuals

experiencing depressive or anxiety disorders and mental health experts. Six synthetic voices, equally divided by gender, each characterized by varying quality levels, were developed for the research using free online voice synthesizers. To collect data on preferences for different synthetic voices, the Virtual Agent Voice Acceptance Questionnaire (VAVAQ) was employed [18]. The findings of the study revealed two main key points: high-quality voices were favored over lower-quality ones and the quality of a synthetic voice appeared to have a stronger impact on user evaluations compared to the voice's gender.

1.2. Aims of the Study. This new investigation is based on a previous study [19] regarding participants' gender preferences towards male and female synthetic voices of high-quality voice (*hereafter*, HQ voice) and low-quality voice level (*hereafter*, LQ voice), respectively.

The distinction between the qualities of the voices used, which were categorized as either LQ or HQ, does not rely on the physical characteristics of the voices themselves but rather on the choice of the synthesizers used to generate them. The two LQ voices assessed were developed through *NaturalReader*, an AI text-to-speech synthesizer, available at https://www.naturalreaders.com/. Instead, the HQ voices were created by the *Acapela Group* (https://www.acapela-group.com/), a company with several years of experience in top-tier synthetic voices. The former is described as a free and nonprofessional synthesizer, while the Acapela synthesizer employed for the HQ voices is recognized for its consideration of various suprasegmental and linguistic features of the utilized language.

This means that the HQ voices encompass all the paralinguistic and prosodic aspects, such as duration, clear pronunciation, empty and filled pauses, and intonation, that makes the synthetic voice less eerie and more natural, while the LQ voices do not account for these features.

To this regard, some studies [17, 20–22] have suggested that users preferred synthetic voices which can mirror human conversational skills. This implies that the proposed voice has to entail the above-mentioned suprasegmental and linguistic features defining the quality of the developed voice, which become particularly fundamental in practical applications such as customer or retail services, e-health systems, and speech synthesis, by affecting the communication's effectiveness.

The previous study involved participants assigned to 4 different age groups (adolescents, young adults, middle-aged, and seniors, respectively). The main outcomes revealed that H-QVs were considered to be notably more enjoyable, manageable, capable of eliciting positive feelings, and effective in engaging interactions with users, compared to synthetic voices of lower quality. Therefore, these data suggest that the quality of a voice may have had a greater impact on results, more than the voice gender. To collect these data, a previous and long version of the above-mentioned VAVAQ has been used. This tool is derived from the *Virtual Agent Acceptance Questionnaire* (VAAQ) [18]. Therefore, the present work shares the aims of this first previous survey to explore the preferences of potential users with respect to high- and low-quality synthetic voices of either gender and aims to test and validate the new shortened version of the questionnaire. To accomplish these aims, the current study adopted the same methodology of the previous one by recruiting a new sample composed by four groups divided among adolescents, young adults, middle-aged, and seniors. A novelty, introduced in the present study, concerns also the presence within the questionnaire of a further section (named *section 4*) developed to investigate the impact of age-related aspects of the voice on users' preferences. The previous study showed the percentage values relating to only a single item of section 4 related to the preferred age while in the current study, the other two items concerning the influence of age and the age attributed by the participants to the voices will also be investigated, respectively.

2. Methodology

2.1. Sample. This current research is aimed at exploring possible differences related to the age of the different groups of participants involved. Four groups of participants, all Italian nationality (seniors, middle-aged, young adults, and adolescents), were required to state their preferences towards the four voices (two male and two female). The reason behind the recruitment of different age groups is linked to the need to test whether preferences towards synthetic voices change with age (tested with a cross-sectional rather than longitudinal design) and, therefore, the effect that this variable (age) has on preferences towards this type of assistive technology.

As indicated earlier, two voice synthesizers were used, one of high quality and one of low quality, consequently determining the proposed voices' quality (high vs. low). Moreover, participants were required to indicate their preferences about the voices' perceived age and the tasks that they would have assigned to the proposed voices. Group 1 which is comprised of 47 Italian adolescents (25 females, mean = 15.04; SD = ±0.88), group 2 which is comprised of 45 Italian young adults (22 females, mean = 25.09; SD = ± 3.64), group 3 which consisted of 45 Italian middle-aged participants (22 females, mean = 49.11; SD = ±4.42); and group 4 which consisted of 45 Italian seniors (25 females, mean = 72.64; SD = ±5.48) were compared, respectively.

2.2. Ethical Aspects. The research was carried out in full compliance with the ethical principles of privacy and confidentiality, and researchers safeguarded participants' privacy, protecting their personal information and ensuring that data were anonymized. All participants were unpaid volunteers, and before starting the experiment, they signed the informed consent that specifies information about privacy and data protection, according to the current Italian and European laws. The ethical committee of the Università degli Studi della Campania "Luigi Vanvitelli," at the Department of Psychology, gave the approval with the protocol number 25/2017.

2.3. Synthetic Voices. Faithful to the protocol of the previous survey [19], also in this case, the 4 different synthetic voices were administered, each lasting from 4 to 7 seconds, equally balanced by gender and level of voice quality. The level of

voice quality was represented by two LQ voices named as Edoardo and Clara, respectively, and the two voices of Antonio and Giulia were assessed as higher quality. The LQ voices have been developed using NaturalReader synthesizer, and then, a free audio software (i.e., Audacity, http://www.audacityteam.org) has been exploited to record them. The development of HQ voices was instead entrusted to Acapela Group, a European company, with several years of leading experience in the development of top-quality synthetic voices. All the voices have been created within the context of the H2020-funded *Empathic* project (http://www.empathic-project.eu) devoted to design an empathic and personalized virtual coach able to support elders in everyday life. Each Italian voice enunciates the following sentence: *Ciao sono Antonio/Giulia/Edoardo/Clara. Se vuoi posso aiutarti nelle tue attività quotidiane* (Hi, my name is Antonio/Giulia/Edoardo/Clara. If you allow me, I can assist you in your daily activities). Members of Acapela company and BeCogSys lab of the Università degli Studi della Campania "Luigi Vanvitelli" focused on the voices' assessment.

2.4. VAVAQ. To carry out the current study, participants' assessment was collected through the digitalized and shortened version of the VAVAQ, administered in Italian language. In a previous study [19], further details about this version can be found. The questionnaire was digitalized through a dedicated software, which checks that the questionnaires are filled out correctly in every part and randomizes the questionnaire's sections' presentation order.

The shortened questionnaire includes six sections: section 1 of the questionnaire is dedicated to collect participants' sociodemographic data and explore their technological knowledge level and easiness of use of different devices, such as smartphones, tablets, and laptops.

Section 2 assesses participants' willingness to interact with the synthetic voice through a single question (*Su una scala da 1 (L'interazione con il Sistema mi sembra altamente probabile) a 5 (L'interazione col Sistema mi sembra altamente improbabile) scelga la risposta che meglio possa descrivere la sua interazione con la voce che ha ascoltato*/please rate from 1 (very likely) to 5 (very unlikely) your willingness/desire to interact with the voice you have listened).

Section 3 of VAVAQ contains 4 sets, each one composed by 6 items (examples are reported), evaluating on a 5-point Likert scale (*1 = fortemente d'accordo, 2 = d'accordo, 3 = non lo so, 4 = in disaccordo, 5 = fortemente in disaccordo*/1 = strongly agree, 2 = agree, 3 = I do not know, 4 = disagree, 5 = strongly disagree) the following features:

(i) Pragmatic qualities (PQ): the practicality, usefulness, controllability, and effectiveness perceived of the audited synthetic voice (e.g., *Penso che la comunicazione con la voce potrebbe essere difficile da gestire*/I think the communication with the voice could be unmanageable)

(ii) Hedonic qualities-identity (HQI): the inventiveness and enjoyable quality referring to the audited voice (e.g., *Penso che la voce sia rassicurante*/I think the voice is reassuring)

(iii) Hedonic qualities-feeling (HQF): potentially aroused by voice (e.g., *Penso che la comunicazione con la voce potrebbe essere noiosa e*/I think that communicating with the voice could be boring)

(iv) Attractiveness (ATT): the voice's capacity to engage the users (e.g., *Penso che la comunicazione con la voce potrebbe essere coinvolgente*/I think that communicating with the voice could be engaging)

Section 4 is composed by 3 items exploring participants' opinion on synthetic voices' attributed and preferred age:

(i) Item 1: *Secondo te, quanti anni ha la voce che hai ascoltato?* (how old do you think the voice you heard is?)

(ii) Item 2: *Per favore indichi se l'età della voce potrebbe influenzare la sua volontà di interagire.* (please indicate whether the voice's age would influence your willingness to interact with her/him). This item requires a dichotomic response (yes/no)

(iii) Item 3: *Per favore indichi l'età che preferirebbe la voce avesse in base alle fasce d'età elencate di seguito* (please indicate your preferred voice's age according to the age ranges listed below)

For both items 1 and 3, participants had to choose among the following age ranges: 19-28 years old, 29-38 years old, 39-48 years old, 49-58 years old, 59+ years old.

Section 5 of the instrument consists of four items, regarding the tasks' type that users would assign to the audited voices by considering the following occupations: housekeeping, healthcare, front office tasks, and protection and security (*Quanto ritiene adatta questa voce alle seguenti mansioni?*/please rate how much you judge the voice suitable in performing the following occupations). This section requires a response on a 5-point Likert scale, *1 = non adatto, 2 = abbastanza adatto, 3 = non lo so, 4 = abbastanza adatto, 5 = molto adatto* (1 = unsuitable, 2 = hardly suitable, 3 = I do not know, 4 = quite suitable, 5 = very suitable), where high suitability for the task corresponds to high scores.

Finally, section 6 (comprising 6 items) assesses the expressiveness, intelligibility, and naturalness of the proposed voice (some item examples were as follows: *La voce è molto chiara e comprensibile* (the voice is very clear and understandable) and *La voce ha un modo di parlare davvero atipico* (the voice sounds really atypical). Items in this section were rated on the same 5-point Likert scale of section 3.

Regarding sections 3 and 6 of VAVAQ, the items with negative acceptation are inversely corrected, meaning that low total scores reflect more positive evaluations than higher total scores.

2.5. Procedure. The study was developed using Lab.js, an online study builder, successively exported on JATOS, a tool allowing the generation of the links that have been given to

the participants. Each participant was provided with a link to be opened from a laptop and carried out the experiment. Raw data were extracted from JATOS (.json files) and exported into an Excel spreadsheet. As mentioned before, after being informed regarding the study's aims, participants were presented with the informed consent and asked to sign it. Next, they were invited to carry out the experiment from their own well-connected laptop by clicking on a link provided by the experimenter via email. Then, the recordings of the 4 synthetic voices appeared randomly, and after listening to each voice, participants had to fill out the VAVAQ. Questionnaire' sections appeared randomly.

3. Results

3.1. Data Analysis and Results of Comparisons among Age Groups. Repeated measures ANOVA statistical models were conducted on the scores of questionnaire's sections to evaluate participants' preferences towards the 4 voices (HQ female voice, HQ male voice, LQ female voice, and LQ male voice).

Participants' gender and their age group were inserted in the model as between-subject factors, and the scores at section 2 (willingness to interact), section 3 (PQ, HQ-I, HQ-F, and ATT), and section 6 (voice features) were included as within-subject factors. Since negative items were inversely corrected, high scores correspond to a negative voice evaluation, while low scores indicate the opposite. Additional repeated measures ANOVAs were conducted to examine differences among suitability' scores of each voice for the entrusted tasks (healthcare, housework, protection, and front office—section 5 of the VAVAQ). As well as in the previous statistical model, between-subject factors were the age group and participants' gender, whereas the scores associated with each entrusted task were considered as within-subject factor. In this section, high and low scores correspond to high and low attributed suitability, respectively.

The significance level for all the analyses was set at $\alpha = 0.05$, and Bonferroni's post hoc tests were applied to assess differences among means. Regarding section 4 of the questionnaire (attributed and preferred age range), Tables 1 and 2 show the percentage values of the synthetic voices' preferred age range for each age group. In addition, percentage values regarding the influence of the age range of the synthetic voices on participants' willingness to interact (item 2 of section 4) will be shown in Table 3. It should be clear that this item entailed a dichotomic response: positive (yes, it affects) or negative (no, I do not care). In addition to descriptive statistics, data extracted from section 4 were also analyzed through a bivariate correlational analysis to determine the relationship between the age of participants and the voices' attributed and preferred ages.

Statistical details are available in the appendix. The following paragraphs summarize the main findings.

3.1.1. Willingness to Interact. No significant differences were observed due to the participants' gender. Significant effects emerged due to the age groups. Results showed that seniors were more willing to interact with the voices compared to middle-aged participants. The quality and gender of the voice affect the willingness to interact: the HQ female voice was associated with greater scores compared to all the other voices (see Figure 1 for these results).

3.1.2. Pragmatic Qualities (PQ). Results reported that participants' gender did not affect the evaluation of pragmatic qualities (PQ), while significant differences were observed due to age groups. Results showed that seniors evaluated the four voices as more effective than adolescents judged them. Also here, voices' quality and gender affect the PQ scores. Specifically, the HQ voice gendered as female was assessed as more effective than the others. These data are depicted in Figure 2.

3.1.3. Hedonic Qualities-Identity (HQI). No differences were observed for the hedonic qualities-identity (HQI) scores due to participants' gender. A significant effect of the age group emerged. Statistical results showed that seniors judged the four proposed voices as more pleasant compared to the all the other age groups. Significant differences were found in the HQI scores among the 4 voices. As occurred for the pragmatic qualities and the willingness to interact, the HQ voice gendered as female was assessed as more pleasant than the all the other three voices. In addition, the HQ male voice was judged as more pleasant and original compared to its LQ counterpart. These data are displayed in Figure 2.

3.1.4. Hedonic Qualities-Feeling (HQF). Responses to the HQF items were affected by participants' gender. Statistical results revealed that males better rated the ability of the four voices to elicit positive emotional states than females judged the voices.

Also, the age groups affected the results. Analyses revealed that seniors assessed the four proposed voices as better capable of emotional engagement compared to middle-aged users. The voices' quality and gender affected the HQF scores. Analyses revealed that the HQ voice gendered as female was evaluated as better capable to elicit positive feelings compared to all the other voices. Furthermore, the HQ voice gendered as male was considered as more capable of eliciting positive feelings compared to its LQ counterpart. These data are illustrated in Figure 2.

For what concerns the gender differences in the HQF scores' attribution to the four voices, males assessed the LQ female voice as significantly more captivating than females. By looking at the differences among the four voices in HQF scores within each gender group, male participants evaluated both HQ voices as more emotionally engaging compared to the LQ voice gendered as male. Conversely, female participants judged the HQ female voice more able to elicit positive feelings compared to both LQ voices. Moreover, females judged the HQ voice gendered as male to be better capable of arousing positive emotional states compared to the LQ female voice.

3.1.5. Attractiveness (ATT). Participants' gender did not affect attractiveness scores, while age groups significantly differed in their attribution. Analysis revealed that middle-aged participants rated the proposed synthetic voices as less

Table 1: Attributed age range values (%) of the voices for each age group. Scores varied from 1 to 5 and reflected age ranges (1 = 19–28 years old; 2 = 29–38 years old; 3 = 39–48 years old; 4 = 49–5 8 years; 5 = 59+ years old).

Age attribution (%)	19–28 years	29–38 years	39–48 years	49–58 years	59+ years
Adolescents	14.13%	35.08%	35.08%	13.09%	2.62%
Young adults	11.67%	47.22%	32.78%	7.22%	1.11%
Middle-aged	11.11%	36.11%	32.78%	14.44%	5.56%
Seniors	3.33%	30%	35.56%	25%	6.11%

Table 2: Preferred age range values (%) of the voices for each age group. Scores varied from 1 to 5 and reflected age ranges (1 = 19–28 years old; 2 = 29–38 years old; 3 = 39–48 years old; 4 = 49–58 years; 5 = 59+ years old).

Age preference (%)	19–28 years	29–38 years	39–48 years	49–58 years	59+ years
Adolescents	45.21%	37.77%	12.77%	3.72%	0.53%
Young adults	31.67%	53.89%	12.78%	1.67%	0.00%
Middle-aged	16.11%	40.56%	30%	11.67%	1.67%
Seniors	7.22%	26.11%	40.56%	20%	6.11%

Table 3: Age range influence responses' percentages attributed to the voices according to age groups. This item of section 4 entailed a dichotomic response: yes, it affects/no, I do not care.

Age influence (%)	Yes, it affects	No, I do not care
Adolescents	32.45%	67.55%
Young adults	36.67%	63.33%
Middle-aged	18.89%	81.11%
Seniors	36.11%	63.89%

engaging than young adults and seniors. The four proposed voices were associated with significantly different ATT scores. Statistical results showed that the HQ voice gendered as female was considered as more appealing compared to the all the other voices (see Figure 2 for these results).

3.1.6. Voice Features. The assessment of voice features such as intelligibility, expressiveness, and naturalness did not differ as a function of the participants' gender. The 4 age groups evaluated significantly different these features. Analysis revealed that these differences were due to seniors considering the proposed synthetic voices as more intelligible, natural, and expressive compared to the other three groups. The four voices were differently evaluated regarding the voice features. Statistical analysis highlighted that participants assessed the HQ voice gendered as female as endowed with superior voice features compared to all the others. Moreover, participants assigned better ratings to the HQ male voice than the LQ female one. Figure 3 illustrates these results.

3.2. Attributed and Perceived Age. Participants' responses to section 6 of the VAVAQ have been extracted, and the percentages of the attributed and preferred age ranges of the proposed voices were calculated for each age group. Descriptive statistics of attributed and perceived age are reported in Tables 1 and 2, respectively.

In addition to descriptive statistics, a bivariate correlational analysis has been performed to test the association between the participants' age and the attributed and perceived age of the four voices. Participants' age was included in the analysis as a continuous variable, whereas the other correlation term consisted of the five above-mentioned age ranges.

A positive correlation emerged between participants' age and the attributed age for all the proposed voices, except for the HQ male one. Concerning the preferred age, significant positive correlations were observed for all the proposed voices. Tables 4 and 5 show correlation coefficients and significance between participants' age and attributed and preferred age, respectively. It should be noted that both analyses are associated with strong correlation coefficients, especially those concerning the preferred age, suggesting that synthetic voice's age plays a role its evaluation, depending on the users' age.

3.2.1. Influence of the Age Range. This paragraph reports the percentages of the dichotomic responses (yes/no) to the single item about whether the age would affect or not the willingness to interact with the voice. Table 5 summarizes these results. Age influence responses' percentages are reported in Table 5 for each age group.

3.3. Entrusted Occupations. This paragraph reports the results concerning the suitability scores for the different occupations. All statistical details are available in the appendix.

Participants were required to assess the suitability of the voices for the following tasks: healthcare, housekeeping, front office, and protection and security tasks. By only listening to the voices, they had to report how much they considered the voice suitable for that specific occupation.

3.3.1. Healthcare. Results showed that participants' gender did not affect the suitability scores for the healthcare tasks. Statistically significant differences were observed among age groups. Results revealed that these differences were due to seniors who considered the voices as more suited for healthcare occupations than young adults.

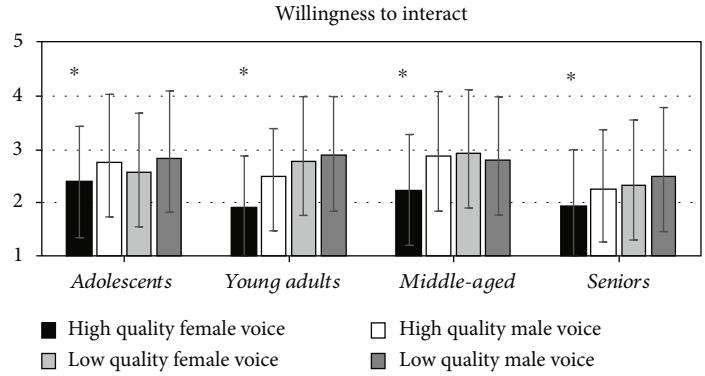

Figure 1: Adolescents, young adults, middle-aged, and seniors' willingness to interact with high-quality female and male voices and low-quality female and male voices, respectively. The symbol "*" above the bars indicates that scores are significantly different. Mean ranges go from 1 (interaction with the systems is very likely) to 5(interaction with the system seems very unlikely). Low scores correspond to high willingness to interact, while high scores correspond to low willingness to interact.

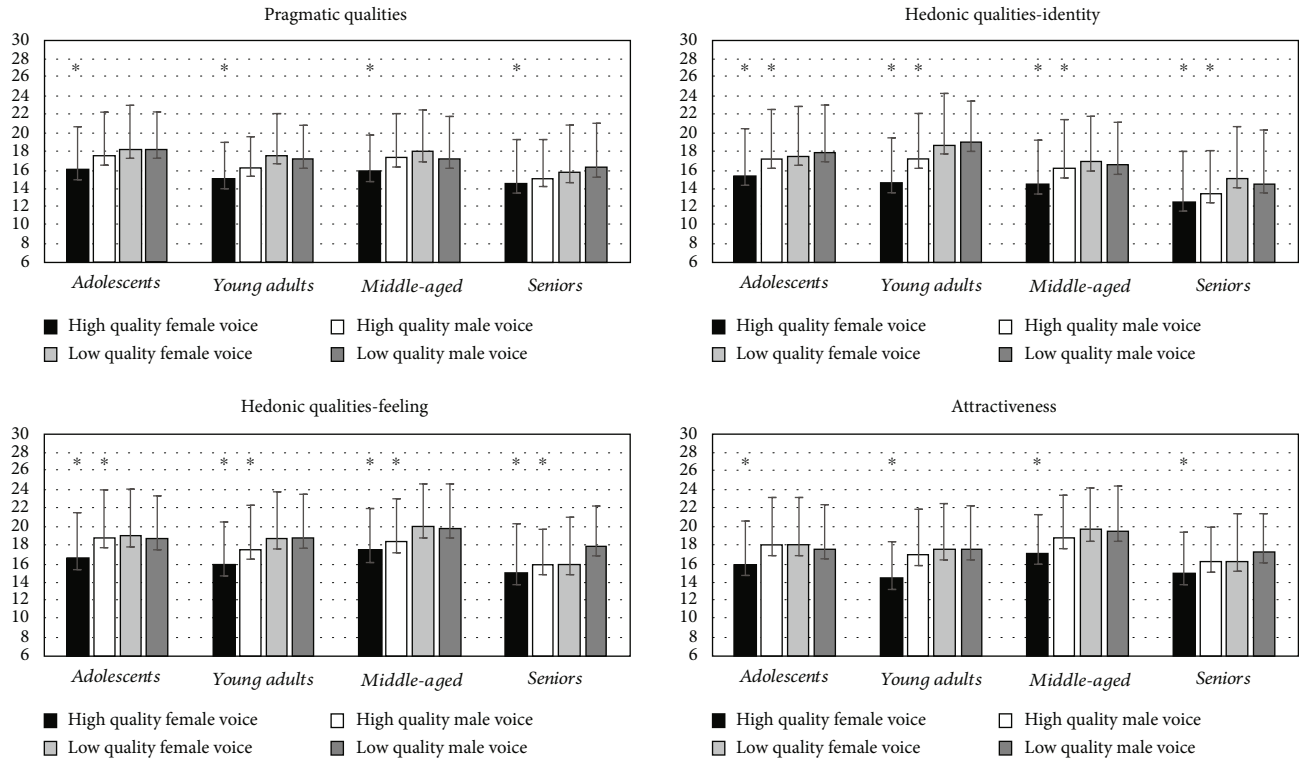

Figure 2: Section 3 scores attributed by each age group to HQ female and male voices and LQ female and male voices, respectively. The symbol "*" above the bars indicates the statistically significant comparisons. Responses vary between 6 and 30 (where higher scores correspond to negative evaluation, while lower scores reflect positive evaluation), since the total score is calculated by summing the participants' responses to 6 questions.

Suitability scores of healthcare assistance significantly differed due to the four proposed voices. Analyses revealed that the HQ voice gendered as female was rated as more qualified than the other three voices in performing healthcare occupations. Figure 4 illustrates these results.

3.3.2. Housework. Likewise, participants' age and gender did not differ due to housework tasks' suitability scores. However, the 4 voices were associated with significantly different scores. In this context, results showed that both HQ and LQ voices gendered as females were judged as more appropriate than their male counterparts in accomplishing housework. Figure 4 illustrates these results.

Moreover, depending on their age, participants assigned statistically different scores to the voices. Concerning differences between participants' groups, analyses did not reveal significant effects. Regarding differences within each age group, results highlighted that middle-age group evaluated

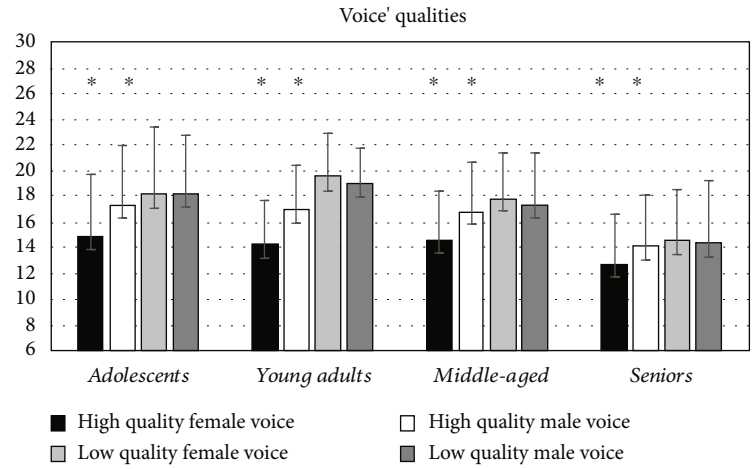

Figure 3: Differences among adolescents, young adults, middle-aged participants, and seniors' assessments in voice quality scores attributed to the high-quality female and male voices and low-quality female and male voices, respectively. The symbol "*" above the bars indicates that scores are significantly different. Mean ranges vary between 6 and 30 (where lower scores correspond to positive evaluation, while higher scores reflect negative evaluation), since the total score is obtained by adding the participants answers (from 1 to 5) to 6 questions.

Table 4: Correlation analysis results between participants' age and attributed age.

		High-quality female voice	High-quality male voice	Low-quality female voice	Low-quality male voice
Participants' age	Pearson's correlation	0.370**	0.132	0.178*	0.190*
	p value	0.000	0.075	0.016	0.010
	N	182	182	182	182

Table 5: Correlation analysis results between participants' age and preferred age.

		High-quality female voice	High-quality male voice	Low-quality female voice	Low-quality male voice
Participants' age	Pearson's correlation	0.498**	0.464**	0.487**	0.416**
	p value	0.000	0.000	0.000	0.010
	N	182	182	182	182

the LQ female voice as more appropriate for housework than both HQ and LQ voices gendered as male. Furthermore, seniors judged the female voices (both HQ and LQ) as more suited to accomplishing housework than their male counterparts.

3.3.3. Protection and Security. No differences were found in suitability evaluation of protection and security occupations due to age groups and participants' gender. Instead, they differed among the 4 voices. Specifically, the LQ female voice was rated significantly less appropriate for these tasks, compared to the other voices. Figure 4 displays these results.

3.3.4. Front Office. Front office tasks' suitability scores differed due to age groups. More in details, analyses revealed that these differences were due to seniors who better rated the proposed voices than young adults and adolescents. Differently, no differences were observed in these scores due to the participants' gender.

Among the proposed voices, there were significant differences in suitability to front office tasks. Results revealed that the best suitability for this type of tasks was associated with HQ female voice that obtained higher scores compared to all the other voices (see Figure 4).

Concerning significant differences among age groups in the front office tasks' suitability ratings, analyses highlighted that middle-aged group judged the HQ voice gendered as female as more suited in carrying out front office tasks than adolescents judged them. For what regards seniors, they better evaluated both the male voices compared to the adolescent and young adult groups. Moreover, seniors rated the LQ female voice as more qualified in accomplishing this kind of tasks than young adults and middle-aged users.

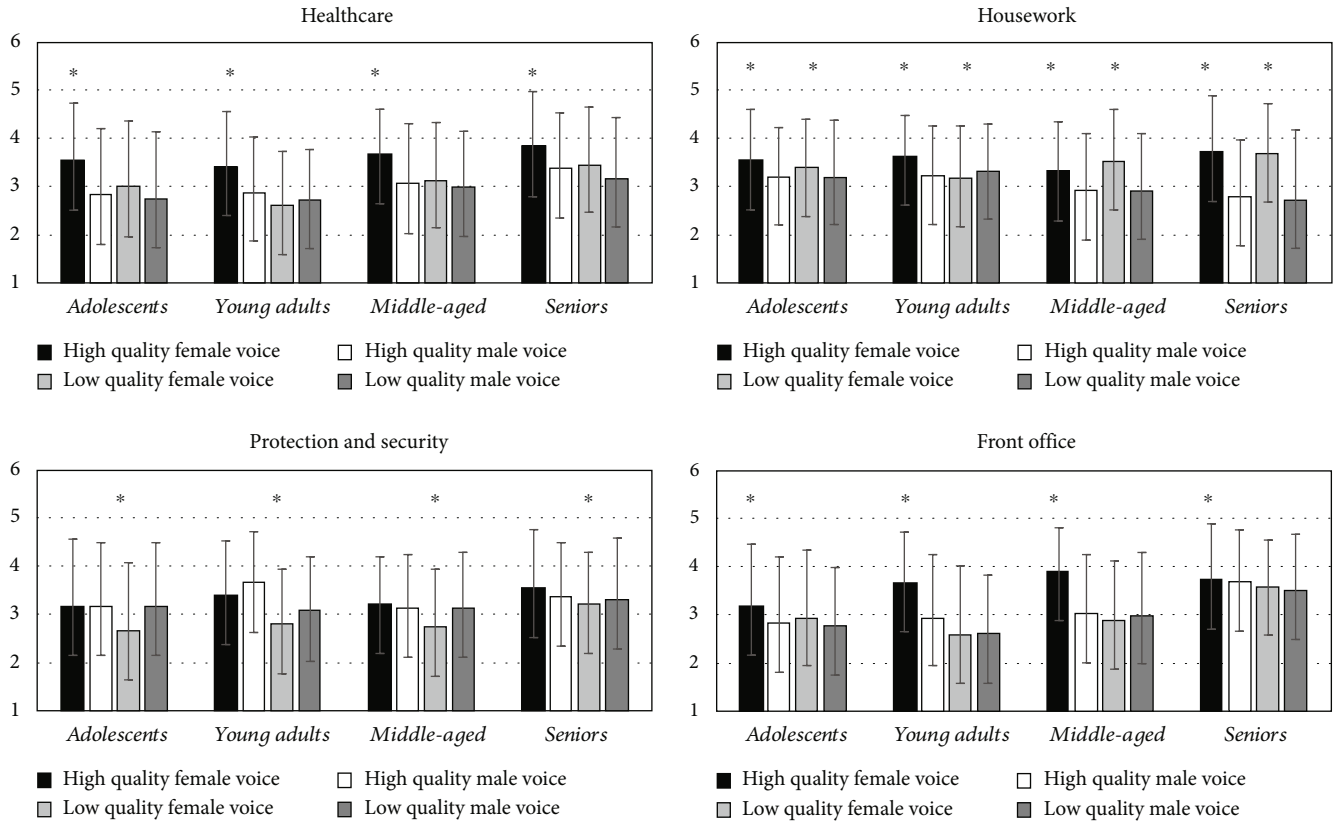

Figure 4: Suitability scores assigned to the HQ and LQ voices for healthcare, housework, protection and security, and front office tasks by each age group. The symbol "*" above the bars indicates the statistically significant comparisons. Means vary from 1 to 5, where high scores correspond to low suitability and low scores correspond to high suitability.

Concerning the differences within each age group in front office suitability evaluation, analyses revealed that these differences were due to young adults and middle-aged groups who better rated the HQ female voice than all the others in performing these tasks.

4. Discussion

With the rapid advancements in natural language processing, artificial intelligence, and automatic speech recognition, the use and availability of conversational agents' (CA) systems have gradually increased [23]. Also, the users' expectations towards such technologies have become more and more demanding, and indeed, users assume that CAs respond to their requests in an intelligent way, by learning from their preferences and providing a clear and natural interaction, reflecting the rules guiding conversations among humans [24]. These expectations suggest that future generation of CAs would require to be developed following a user-centered approach, by considering the users' assessment of the features that characterized the proposed voice assistant [25].

To this regard, the present work underlines that the four groups here considered agreed in being more willing to be involved in a potential long-lasting interaction with the HQ voice gendered as female than the HQ male voice and the LQ voices, by considering the former as more effective, pleasant, natural, expressive, and able to arouse positive emotional states and to engage the potential end users than the other proposed voices. Regarding the HQ voice gendered as male, it was judged as more enjoyable, original, and capable of eliciting positive feeling than its corresponding LQ voice, whereas participants considered the HQ voice gendered as male as more natural than the female LQ voice.

The better assessment associated with the HQ female voice is in line with those studies showing that users perceive female voices as more accommodating and pleasant, compared to male ones [26–28]. However, such preference may be ascribed to the prevalence of the automated systems gendered as female, which may lead the users to better assess a female voice, due to the societal stereotype of considering an assistant as female, rather than to their actual preferences [29, 30].

The obtained results also pointed out that the voice quality plays a pivotal role in the assessment' process. Indeed, HQ voices, regardless of their gender, are associated with more positive evaluation compared to their LQ counterparts. Such better evaluations reflect the advantages of interacting with HQ voices, rather than LQ ones. To this regard, some studies [17–20–22] have suggested that users preferred synthetic voices which can mirror human conversational skills. This implies that the proposed voice has to entail paralinguistic and prosodic aspects, such as human-like intonation, filled and empty pauses, and clear pronunciation. These

suprasegmental and linguistic features define the quality of the developed voice, which become particularly fundamental in practical applications such as customer or retail services, e-health systems, and speech synthesis, by affecting the communication's effectiveness.

In terms of gender differences, findings reported that male participants assessed the proposed voices as more capable of arousing positive emotional states, compared to female participants judged them. Such result complies with the human-computer interaction (HCI) literature sustaining that gender plays a moderating role in the technology acceptance and adoption, even though such technological gender gap is diminishing over the years (for a review, see [31, 32]).

As regards the participants' age group differences, seniors considered the four proposed voices as more emotionally engaging and attractive than middle-aged participants while young adults judged the voice as more attractive than middle-aged. Additionally, seniors considered the proposed synthetic voices as more intelligible, natural, and expressive than young adults, middle-aged participants, and adolescents. Such results are not in line with the common misconception reported in literature that elders are not willing to accept or adopt automated systems (for a review, see [33]). Conversely, our findings agree with those studies reporting more positive evaluations of CAs (or other related technologies) formulated by older users, compared to younger ones, in terms of trust, likability, and frequency of use [34, 35].

A possible interpretation is that different age groups may focus on different features of the system and be driven by different motivations during the assessment. To this regard, it has been found that older adults tend to perceive the CA as a companion with whom they can share some activities, whereas younger people consider this type of system as a tool to facilitate everyday tasks and avoid human contact [36]. Indications that older participants are more accepting of the voices in healthcare roles than other groups (see paragraph 3.3.1, above) are consistent with a greater consciousness than in other groups of what is involved in needing healthcare assistance and of the possibility that few trustworthy humans are guaranteed to be available during any period of need.

Regarding the age attribution to the voices, middle-aged participants, young adults, and adolescents disclosed the same opinions by choosing more the age ranges between 29-38 years and 39-48 years, while seniors revealed a more heterogenous choice by adding to these two mentioned age ranges and the 49-58 years age range. However, when the participants are required to reveal their preferences toward the age range, each group of participants has their own preferences respectively: adolescents and young adults mostly selected the age ranges between 19-28 years and 29-38 years (with a slight preference of adolescents for the first age range and a greater preference of young adults for the second age range); middle-aged participants disclosed their willingness to be assisted by a conversational agent aged between 29-38 years and 39-48 years (therefore, corresponding to their reference age range). Lastly, seniors, once again, disclosed their preferences toward three of the five age ranges proposed, respectively, between 39-48 years, 29-38 years, and 49-58 years. Interestingly, seniors compared the other three groups of participants did not select their reference age range by revealing their predisposition to be assisted by younger synthetic voices.

As in a previous study [19] adopting the same methodology, also in this case, the majority of the participants answered that the age of the synthetic voice would not influence the way they would interact with the voices. To this regard, it would be interesting to examine the motivation that pushes the participants to answer for no versus yes, since the concern towards a synthetic voice's age is an aspect which has not been addressed in the existing literature on the topic and should be taken into account in our future investigations. Regarding the entrusted tasks to the synthetic voices, outcomes revealed that the HQ voice gendered as female was considered as better suited than the HQ male voice and both the LQ voices (female and male, respectively), in achieving front office and healthcare tasks. With regard to housework, a typical gender stereotype seems evident, according to which female-gendered CAs are considered more suitable for nurturing and caring occupations, while CAs gendered as male are preferred for those tasks requiring authority, competence, and agency [37, 38]. Supporting this speculation is the fact that both HQ and LQ female voices were found to be better suited to perform household tasks than HQ and LQ male voices. Curiously, regarding the protection and security tasks, participants did not seem to focus on the voice's gender, rather they gave more importance to the quality of the voice, by judging the LQ voice gendered as female as significantly less skilled in performing this type of tasks than the HQ female voice and both the HQ and LQ male voices.

A possible interpretation of this difference due to the voice's quality could be that the vocal features associated to the higher quality sounded more reliable for this type of task, compared to the lower one. Lastly, considering participants' age groups differences, results revealed that seniors deemed the synthetic voices as more fit for healthcare occupations than young adults. Furthermore, seniors better rated the proposed voices than young participants and adolescents in accomplishing front office tasks.

5. Conclusions

The current research contributes knowledge in the field of user experience research by providing an analysis of users' preferences towards four synthetic voices, through the self-reported VAVAQ [18]. The assessment took into account both user- and voice-related characteristics, such as gender, age, and voice's quality.

Due to the emerging nature of this research topic, the existing literature aimed at identifying the vocal features a CA should present to be accepted by the final users is still partial and needs to be extended [23]. To this regard, the assessment of the users' perceptions towards such technologies becomes relevant for the subsequent phases of design, development, and, finally, adoption of these systems in various domains. In this context, the present study presents

original results, which may provide some insights to the topic and contribute to bridge the literature gap related to the users' preferences towards their digital assistants.

Nevertheless, some limitations should be pointed out. For instance, it should be noted that the questionnaire used for the voices' assessment in the current investigation is not a validated tool to measure users' acceptance, although this instrument has been already adopted in previous studies examining the same topic (for a review, see [18]), by revealing that it has the potential to provide a comprehensive analysis of the users' attitudes towards technology by considering several aspects related to systems. Furthermore, in the present study, the statistical model carried out for the analyses does not allow to disentangle the effects of voice's gender and quality on the questionnaires' scores. The justification for this methodological choice is that the identity of the four proposed voices was defined by simultaneously considering both variables.

This occurred since the aim of the study was to identify which vocal agent would provide the best experience for the users, rather to separately investigate the effects of the two features.

Future studies should take into account such limitations in order to provide more specific information about the factors influencing the users' assessment of the proposed CAs. Moreover, it would be interesting to extend the investigation to clinical populations, for instance, by involving participants with depressive and anxiety disorders, with the aim to provide useful information to the developers of psychological monitoring systems aimed at improving and increasing mental health in potential end users.

Appendix

A. Details of Statistical Analyses Carried Out on the VAVAQ's Scores

The collected data relative to the group of Italian young adults were compared with the data obtained from three other Italian groups formed by adolescents, middle-aged, and seniors, respectively.

The statistical model adapted to test whether the VAVAQ's scores depend on users' and voice's features was repeated measures ANOVA. In detail, participants' gender (male/female) and age group (seniors/middle-aged/young adults/adolescents) were included in the model as between-subject factors, while the within-subject factor was the voice type (HQ male/HQ female/LQ male/LQ female). Dependent variables were the scores assigned to the four voices at the items of section 2 (willingness to interact), section 3 (PQ, HQI, HQF, and ATT), and section 6 (voice features).

Additional repeated measures ANOVA were conducted to evaluate differences in suitability scores for the five entrusted occupations, attributed to each voice by participants (section 5). Also, for these analyses, age group and participants' gender were included as between-subject factors in the model, while the voice type was the within-subject factor. Dependent variables were the suitability scores for each occupation.

The significance level of all the analyses was set at $\alpha = 0.05$, and post hoc tests with Bonferroni's correction (BC) were applied to assess the comparisons among means. The interaction effects were examined through the analysis of simple effects with Bonferroni's adjustment for pairwise comparisons.

Please note that scores vary according to the different section of the questionnaire used: for willingness to interact, means range from 1 (very likely) to 5 (very unlikely) because the score was obtained from one single question; for PQ, HQI, HQF, ATT, and voice features, responses' mean varies between 6 and 30 (low scores = positive evaluation and high scores = negative evaluation), since the total score is obtained by summing the responses (from 1 to 5) to the 6 items composing these sections. For the entrusted occupations, responses' mean varies from 1 to 5 (low scores = low suitability and high scores = high suitability assigned to the voices in achieving the considered occupations).

A.1. Willingness to Interact (Section 2). No significant effects for participants' gender ($F(1,174) = 2.142$, $p = 0.145$). A significant effect of the age groups emerged ($F(3,174) = 2.811$, $p = 0.041$). Post hoc tests with BC showed that seniors (mean = 2.200) were more willing to interact with the voice, regardless of their gender or quality compared to middle-aged (mean = 2.692, $p = 0.046$). There was an effect of the voice type on the section 1 responses ($F(3,522) = 20.298$, $p < <0.01$). Regardless of their gender or age, participants were more willing to interact with the HQ female voice (mean = 2.084, $p < <0.01$) compared to the HQ male voice (mean = 2.568), the LQ female voice (mean = 2.622), and the LQ male voice (mean = 2.734). Figure 1 reports these results.

A.2. Pragmatic Qualities (PQ) (Section 3). There was no effect of gender in PQ scores ($F(1,174) = 2.887$, $p = 0.091$). Instead, the age group significantly affected PQ scores ($F(3,174) = 3.557$, $p = 0.016$). Post hoc tests with BC showed that synthetic voices' PQ scores were higher in the senior group (mean = 15.291) compared to the adolescents one (mean = 17.498, $p = 0.014$).

Significant differences in the PQ scores were observed due to the voice type ($F(3,522) = 14.931$, $p < <0.01$). Post hoc tests with BC showed that HQ female voice (mean = 15.340) was rated significantly better than the HQ male voice (mean = 16.545, $p = 0.001$), the LQ female voice (mean = 17.314, $p < <0.01$), and the LQ male voice (mean = 17.197, $p < <0.01$). Data are reported in Figure 2.

A.3. Hedonic Qualities-Identity (HQI) (Section 3). Participants' gender did not affect the hedonic qualities-identity (HQI) scores ($F(1,174) = 0.962$, $p = 0.328$), while significant differences were observed due to participants' age ($F(3,174) = 11.467$, $p < <0.01$). Post hoc tests with BC reported that HQI scores of the four proposed voices were significantly higher in the senior group (mean = 13.758) compared to the adolescent (mean = 16.967, $p < <0.01$), young adult (mean = 17.353, $p < <0.01$), and middle-aged ones (mean = 15.995, $p = 0.007$).

The four voices were associated with significantly different HQI scores ($F(3,522) = 22.705$, $p < <0.01$). Post hoc tests with BC showed that the HQ female voice (mean = 14.151, $p < <0.01$) was significantly considered as more pleasant than the HQ male voice (mean = 15.973), the LQ female voice (mean = 16.988), and the LQ male voice (mean = 16.961). Furthermore, the HQ male voice (mean = 15.973) was judged as more enjoyable than the LQ male voice (mean = 16.961, $p = 0.040$). These results are displayed in Figure 2.

A.4. Hedonic Qualities-Feeling (HQF) (Section 3). Hedonic qualities-feeling (HQF) scores significantly differ due to participants' gender ($F(1,174) = 9.337$, $p = 0.003$). In details, pairwise comparisons with Bonferroni's correction showed that male participants (mean = 16.778) assigned to the four voices significantly higher HQF scores than female participants (mean = 18.530, $p = 0.003$).

There was also a significant effect of the age groups ($F(3,174) = 4.319$, $p = 0.006$). Post hoc tests with BC reported that senior participants (mean = 16.009) considered the voices as more capable to elicit positive emotional states compared to their middle-aged counterparts (mean = 18.831, $p = 0.004$).

Significant differences in HQF scores also were observed due to the voice type ($F(3,522) = 17.620$, $p < <0.01$). Post hoc tests with BC showed that the HQ female voice (mean = 16.056) was significantly considered as more able to arouse positive feelings than the HQ male voice (mean = 17.584, $p = 0.002$), the LQ female voice (mean = 18.276, $p < <0.01$), and the LQ male voice (mean = 18.701, $p < <0.01$). Moreover, the HQ male voice (mean = 17.584) was associated with better capacities to elicit positive feelings compared to the LQ male voice (mean = 18.701, $p = 0.028$). Results are displayed in Figure 2.

A significant interaction effect between participants' gender and the voice type ($F(3,522) = 3.566$, $p = 0.014$) emerged. Pairwise comparisons with BC revealed the following:

(i) Regarding the gender differences in attributing HQF scores to the different voices, male participants (mean = 16.623) considered the LQ female voice significantly more captivating compared to female participants (mean = 19.930, $p < <0.01$)

(ii) Regarding the differences among the voice types within each gender group in HQF score assignment, male participants attributed to the LQ male voice (mean = 18.172) significantly lower HQF scores compared to those they assigned to the HQ female voice (mean = 15.395, $p < <0.01$) and the LQ female voice (mean = 16.623, $p = 0.028$). Furthermore, female participants assigned to the HQ female voice (mean = 16.717, $p < <0.01$) significantly higher HQF scores than those assigned to the LQ female voice (mean = 19.930) and the LQ male voice (mean = 19.230). Similarly, the HQ male voice (mean = 18.244) received higher HQF scores from female participants than the LQ female voice (mean = 19.930, $p = 0.016$)

A.5. Attractiveness (ATT) (Section 3). Attractiveness (ATT) did not differ according to participants' gender ($F(1,174) = 2.958$, $p = 0.087$), whereas there was an effect of the age groups ($F(3,174) = 4.339$, $p = 0.006$). Post hoc tests with BC revealed that middle-aged participants (mean = 18.671) evaluated the voices as less attractive compared to young adults (mean = 16.565, $p = 0.040$) and seniors (mean = 16.083, $p = 0.006$).

The four voices were associated with significantly different ATT scores ($F(3,522) = 18.082$, $p < <0.01$). Post hoc tests with BC showed that ATT evaluation related to the HQ female voice (mean = 15.530, $p < <0.01$) was significantly more positive than the ATT evaluation of the HQ male voice (mean = 17.429), the LQ female voice (mean = 17.791), and the LQ male voice (mean = 17.909). These data are reported in Figure 2.

A.6. Voice' Features (Section 6). There was no significant effect of the participants' gender ($F(1,174) = 0.154$, $p = 0.695$). Conversely, an effect due to the participants' age was found ($F(3,174) = 14.958$, $p < <0.01$). Post hoc test with BC showed that these differences were due to seniors (mean = 13.860, $p < <0.01$) providing a better evaluation of the voice features compared to adolescents (mean = 17.163), young adults (mean = 17.465), and middle-aged participants (mean = 16.608).

Voice' features scores were significantly different among the four voices ($F(3,522) = 36.942$, $p < <0.01$). Post hoc tests with BC showed that voice features' scores of HQ female voice (mean = 14.094, $p < <0.01$) were significantly higher compared to those assigned to HQ male voice (mean = 16.335), the LQ female voice (mean = 17.463), and the LQ male voice (mean = 17.205). In addition, participants assessed the HQ male voice (mean = 16.335) better than the LQ female one (mean = 17.463, $p = 0.005$). Results are displayed in Figure 3.

A.7. Entrusted Occupations to the Synthetic Voices (Section 5). Next subparagraphs report the results of repeated measures ANOVA performed to test differences among the suitability scores assigned to the voices to the five considered occupations: healthcare, housework, protection and security task, and front office.

Section 5 scores ranged from 1 to 5, where high scores means high suitability and low scores correspond to low suitability that seniors, middle-aged, young adults, and adolescents assigned to the proposed voices in performing the considered occupations.

A.8. Healthcare. There was no significant effect of the participants' gender ($F(1,174) = 3.332$, $p = 0.070$). Differences were observed among age groups ($F(3,174) = 3.956$, $p = 0.009$). Post hoc tests with BC revealed that young adults (mean = 2.893) assigned lower scores compared to seniors (mean = 3.483, $p = 0.008$).

Significant differences were found among the four voices ($F(3,522) = 20.112$, $p < <0.01$) in suitability scores for healthcare services. Post hoc tests with BC showed that the HQ female voice (mean = 3.616, $p < <0.01$) was evaluated

as more eligible to achieve this type of task, compared to HQ male voice (mean = 3.032), the LQ female voice (mean = 3.059), and the LQ male voice (mean = 2.899). Figure 4 illustrates these results.

A.9. Housework. There were no significant effects for participants' gender ($F(1,174) = 0.492$, $p = 0.484$) and age group ($F(3,174) = 0.639$, $p = 0.591$). There was a significant difference in the suitability scores for housework occupation due to the voice type ($F(3,522) = 14.289$, $p < <0.01$). Post hoc tests with BC showed that the HQ female voice (mean = 3.531, $p < <0.01$) was judged as more eligible for this type of task, compared to HQ male voice (mean = 3.005) and LQ male voice (mean = 3.020). Similarly, the LQ voice gendered as female (mean = 3.435) was evaluated as more suitable compared to the HQ male voice (mean = 3.005, $p < <0.01$) and the LQ male voice (mean = 3.020, $p = 0.001$) for housework. Figure 4 illustrates these results. An interaction effect between age group and the voice type ($F(9,522) = 2.738$, $p = 0.004$) on the housework suitability scores was observed. Simple effect analysis was carried out for each single variable (age group and voice type). Pairwise comparisons showed the following:

(i) Regarding housework suitability scores differences within each age group, the middle-aged one assigned greater suitability scores to LQ female voice (mean = 3.493) compared to the HQ male voice (mean = 2.893, $p = 0.029$) and the LQ male voice (mean = 2.887, $p = 0.033$). Moreover, seniors attributed greater scores to the HQ female voice (mean = 3.685, $p < <0.01$) than the HQ male voice (mean = 2.740) and the LQ male voice (mean = 2.715). Similarly, seniors rated as more suitable for this kind of task the LQ female voice (mean = 3.705, $p < <0.01$) than the HQ male voice (mean = 2.740) and the LQ male voice (mean = 2.715)

(ii) For what concerns housework suitability scores differences between the differently aged groups examined, no significant effects were found

A.10. Protection and Security Tasks. Participants' gender ($F(1,174) = 0.885$, $p = 0.348$) and age group ($F(3,174) = 1.596$, $p = 0.192$) exert no effect on the scores related to this type of task. Protection and security tasks' suitability scores were statistically different among the four voice types ($F(3,522) = 8.753$, $p < <0.01$). Post hoc tests with BC reported that LQ female voice (mean = 2.863) was assessed as less qualified than the HQ female voice (mean = 3.322, $p < <0.01$), the HQ male voice (mean = 3.324, $p < <0.01$), and the LQ male voice (mean = 3.165, $p = 0.010$) for this kind of occupations (see Figure 4 for these results).

A.11. Front Office. The gender of participants did not affect the front office suitability scores ($F(1,174) = 0.904$, $p = 0.343$). Instead, there was a significant effect of the age groups ($F(3,174) = 6.524$, $p < <0.01$). Post hoc tests with BC showed that seniors (mean = 3.631) assigned greater scores to the voices compared to young adults (mean = 2.941, $p = 0.001$) and adolescents (mean = 2.920, $p = 0.001$). The voice type affected the front office's suitability scores ($F(3,522) = 15.651$, $p < <0.01$). Post hoc tests with BC reported that the HQ female voice (mean = 3.602, $p < <0.01$) was associated with higher scores compared to HQ male voice (mean = 3.115), LQ female voice (mean = 3.006), and LQ male voice (mean = 2.964). These results are reported in Figure 4.

The interaction effect between age groups and voice type was significant ($F(9,522) = 2.335$, $p = 0.014$). Simple effect analysis was carried out for each factor (participants' age group and suitability scores front office tasks). Results were the following:

(i) For what concerns differences among age groups in front office suitability evaluation of the four voice types, adolescents (mean = 3.165) rated as less suitable the HQ female voice for front office tasks compared to middle-aged participants (mean = 3.889, $p = 0.014$). Moreover, seniors (mean = 3.700) attributed higher scores to the HQ male voice compared to adolescents (mean = 2.806, $p = 0.006$) and young adults (mean = 2.927, $p = 0.026$). Also, seniors (mean = 3.615) provided greater scores to the LQ female voice compared to young adults (mean = 2.578, $p = 0.001$) and middle-aged participants (mean = 2.888, $p = 0.047$). Similarly, seniors (mean = 3.495) attributed higher scores to the HQ male voice than young adults (mean = 2.621, $p = 0.006$) and adolescents (mean = 2.764, $p = 0.031$)

(ii) Regarding the differences in the four voice suitability evaluation for front office tasks, within each age group, pairwise comparisons showed that young adults assigned greater scores to the HQ female voice (mean = 3.637) compared to HQ male voice (mean = 2.927, $p = 0.005$), LQ female voice (mean = 2.578, $p < <0.01$), and LQ male voice (mean = 2.621, $p < <0.01$). Moreover, the middle-aged group judged the HQ female voice (mean = 3.889, $p < <0.01$) more eligible for this task, compared to the HQ male voice (mean = 3.027), the LQ female voice (mean = 2.888), and the LQ male voice (mean = 2.977)

Data Availability

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy reason.

Ethical Approval

The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Dipartimento di Psicologia at Università degli Studi della Campania "Luigi Vanvitelli" (protocol code 25/2017, date of approval 15/12/2017).

Consent

Patients are not involved in the current research. Written consent was obtained from every participant.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

The research leading to these results has received funding from the European Union Horizon 2020 research and innovation programme under grant agreement N.769872 (EMPATHIC) and N. 823907 (MENHIR); from the project SIROBOTICS that received funding from Ministero dell'Istruzione, dell 'Università, e della Ricerca (MIUR), PNR 2015-2020, D.D. 1735/2017; from the project ANDROIDS that received funding from Università della Campania "Luigi Vanvitelli" inside the programme V:ALERE 2019, funded with D.R. 906/2019; and from the project SALICE that received funding from Università della Campania "Luigi Vanvitelli" inside the programme Giovani Ricercatori (DR 509/2022), funded with DR 834/2022.

References

[1] A. L. Guzman, "Voices in and of the machine: source orientation toward mobile virtual assistants," *Computers in Human Behavior*, vol. 90, pp. 343–350, 2019.

[2] V. Petrock, *US voice assistant users 2019-who, what, where and why*, eMarketer, 2019, March 2021, https://www.emarketer.com/content/us-voice-assistant-users-2019.

[3] M. B. Hoy, "Alexa, Siri, Cortana, and more: an introduction to voice assistants," *Medical Reference Services Quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[4] L. Laranjo, A. G. Dunn, H. L. Tong et al., "Conversational agents in healthcare: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 2018.

[5] K. O'Brien, A. Liggett, V. Ramirez-Zohfeld, P. Sunkara, and L. A. Lindquist, "Voice-controlled intelligent personal assistants to support aging in place," *Journal of the American Geriatrics Society*, vol. 68, no. 1, pp. 176–179, 2020.

[6] A. B. Kocaballi, E. Sezgin, L. Clark et al., "Design and evaluation challenges of conversational agents in health care and well-being: selective review study," *Journal of Medical Internet Research*, vol. 24, no. 11, article e38525, 2022.

[7] A. A. Abd-Alrazaq, A. Rababeh, M. Alajlani, B. M. Bewick, and M. Househ, "Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis," *Journal of Medical Internet Research*, vol. 22, no. 7, article e16021, 2020.

[8] B. Krahé, A. Uhlmann, and M. Herzberg, "The voice gives it away," *Social Psychology*, vol. 52, no. 2, pp. 101–113, 2021.

[9] A. H. Eagly and W. Wood, "Social role theory," *Handbook of Theories of Social Psychology*, vol. 2, 2012.

[10] R. A. Lippa, K. Preston, and J. Penner, "Women's representation in 60 occupations from 1972 to 2010: more women in high-status jobs, few women in things-oriented jobs," *PLoS One*, vol. 9, no. 5, article e95960, 2014.

[11] T. Hentschel, M. E. Heilman, and C. V. Peus, "The multiple dimensions of gender stereotypes: a current look at men's and women's characterizations of others and themselves," *Frontiers in Psychology*, vol. 10, 2019.

[12] A. P. Chaves and M. A. Gerosa, "How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design," *International Journal of Human–Computer Interaction*, vol. 37, no. 8, pp. 729–758, 2021.

[13] M. Crabtree, P. Mirenda, and D. Beukelman, "Age and gender preferences for synthetic and natural speech," *Augmentative and Alternative Communication*, vol. 6, no. 4, pp. 256–261, 1990.

[14] J. W. Mullennix, S. E. Stern, S. J. Wilson, and C. L. Dyson, "Social perception of male and female computer synthesized speech," *Computers in Human Behavior*, vol. 19, no. 4, pp. 407–424, 2003.

[15] S. Machado, E. Duarte, J. Teles, L. Reis, and F. Rebelo, "Selection of a voice for a speech signal for personalized warnings: the effect of speaker's gender and voice pitch," *Work*, vol. 41, Supplement 1, pp. 3592–3598, 2012.

[16] P. Wagner, J. Beskow, S. Betz et al., "Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program," in *Proceedings of the 10th Speech Synthesis Workshop (SSW10)*, Vienna, Austria, 2019.

[17] T. Amorese, G. McConvey, M. Cuciniello et al., "Assessing synthetic voices for mental health chatbots," in *International Congress on Information and Communication Technology*, pp. 61–75, Springer Nature Singapore, Singapore, 2024.

[18] A. Esposito, T. Amorese, M. Cuciniello, A. M. Esposito, and G. Cordasco, "Do you like me? Behavioral and physical features for socially and emotionally engaging interactive systems," *Frontiers in Computer Science*, vol. 5, article 1138501, 2023.

[19] M. Cuciniello, T. Amorese, G. Cordasco et al., "Identifying Synthetic Voices' Qualities for Conversational Agents," in *Applied Intelligence and Informatics. AII 2022*, vol. 1724 of Communications in Computer and Information Science, pp. 333–346, Springer, Cham.

[20] C. J. Stevens, N. Lees, J. Vonwiller, and D. Burnham, "On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference," *Computer Speech & Language*, vol. 19, no. 2, pp. 129–146, 2005.

[21] H. J. D. Wiener and T. L. Chartrand, "The effect of voice quality on AD efficacy," *Psychology & Marketing*, vol. 31, no. 7, pp. 509–517, 2014.

[22] M. Patrizi, M. Vernuccio, and A. Pastore, ""Hey, voice assistant!" how do users perceive you? An exploratory study," *Sinergie Italian Journal of Management*, vol. 39, no. 1, pp. 173–192, 2021.

[23] E. C. Ling, I. Tussyadiah, A. Tuomi, J. L. Stienmetz, and A. Ioannou, "Factors influencing users' adoption and use of conversational agents: a systematic review," *Psychology & Marketing*, vol. 38, no. 7, pp. 1031–1051, 2021.

[24] E. Luger and A. Sellen, ""Like having a really bad PA" the gulf between user expectation and experience of conversational agents," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5286–5297, Silicon Valley, San Jose, California, 2016.

[25] A. Rapp, L. Curti, and A. Boldi, "The human side of human-chatbot interaction: a systematic literature review of ten years

of research on text-based chatbots," *International Journal of Human-Computer Studies*, vol. 151, article 102630, 2021.

[26] D.-C. Toader, G. Boca, R. Toader et al., "The effect of social presence and chatbot errors on trust," *Sustainability*, vol. 12, no. 1, p. 256, 2020.

[27] C. H. Ernst and N. Herm-Stapelberg, *The Impact of Gender Stereotyping on the Perceived Likability of Virtual Assistants*, Americas Conference on Information Systems, 2020, https://aisel.aisnet.org/amcis2020/cognitive_in_is/cognitive_in_is/4/.

[28] K. L. Goodman and C. B. Mayhorn, "Pitch perfect: influence of perceived voice agent gender and vocal pitch on trust and reliance," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, pp. 1529-1530, 2021.

[29] J. Cambre and C. Kulkarni, "One voice fits all?," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, pp. 1–19, 2019.

[30] J. Feine, U. Gnewuch, S. Morana, and A. Maedche, "Gender bias in chatbot design," in *Chatbot Research and Design. Conversations 2020*, A. Følstad, Ed., vol. 11970 of Lecture Notes in Computer Science, pp. 79–93, Springer, Cham, 2020.

[31] A. Goswami and S. Dutta, "Gender differences in technology usage—a literature review," *Open Journal of Business and Management*, vol. 4, no. 1, pp. 51–59, 2016.

[32] Z. Cai, X. Fan, and J. Du, "Gender and attitudes toward technology use: a meta-analysis," *Computers & Education*, vol. 105, pp. 1–13, 2017.

[33] M. Zhang, "Older people's attitudes towards emerging technologies: a systematic literature review," *Public Understanding of Science*, vol. 32, no. 8, pp. 948–968, 2023.

[34] A. Hosseinpanah, N. C. Krämer, and C. Straßmann, "Empathy for everyone? The effect of age when evaluating a virtual agent," in *HAI '18: Proceedings of the 6th International Conference on Human-Agent Interaction*, pp. 184–190, Southampton, United Kingdom, 2018.

[35] Y. H. Oh, K. Chung, and D. Y. Ju, "Differences in interactions with a conversational agent," *International Journal of Environmental Research and Public Health*, vol. 17, no. 9, p. 3189, 2020.

[36] M. J. van der Goot and T. Pilgrim, "Exploring age differences in motivations for and acceptance of chatbot communication in a customer service context," in *Chatbot Research and Design. Conversations 2020*, A. Følstad, Ed., vol. 11970 of Lecture notes in computer science, Springer, Cham, 2020.

[37] F. Fossa and I. Sucameli, "Gender bias and conversational agents: an ethical perspective on social robotics," *Science and Engineering Ethics*, vol. 28, no. 3, p. 23, 2022.

[38] M. H. A. Bastiansen, A. C. Kroon, and T. Araujo, "Female chatbots are helpful, male chatbots are competent? The effects of gender and gendered language on human-machine communication," *Publizistik*, vol. 67, no. 4, pp. 601–623, 2022.