Computers in Human Behavior 77 (2017) 336–346

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Effectiveness, attention, and recall of human and artificial voices in an advertising story. Prosody influence and functions of voices

Emma Rodero

Departament of Communication, Pompeu Fabra University (UPF), Roc Boronat, 138. 08108, Barcelona, Spain

ARTICLE INFO

Article history:
Received 25 August 2016
Received in revised form
22 August 2017
Accepted 24 August 2017
Available online 31 August 2017

Keywords:
Human and artificial voices
Prosody
Effectiveness
Attention
Recall

ABSTRACT

Many users are exposed every day to artificial voices in their different devices. Because of this, there is a growing interest both in improving the quality of these voices and in analyzing how they are perceived and processed. However, very little research has been conducted to examine nonverbal elements such as prosody. Accordingly, the first purpose of this study is to determine how artificial voices compared to human voices are processed in a narrative advertising story modifying prosody regarding effectiveness, attention, concentration, and recall. The second objective is to evaluate their functions for different applications, advertising among them. The results show that human voices are assessed as more effective and achieved a better level of effectiveness, attention, and recall with less concentration. Concerning the functions, the more important and complex a function is, the more a human voice is preferred over an artificial one.

1. Introduction

We are immersed in a technological world of devices that use synthetic or artificial voices with which people interact. Every day, many people are exposed to synthetic voices on their smartphones, websites, and at home. The applications of artificial speech in these devices are diverse – from the easiest such as those used by telecommunications services, to more sophisticated ones that aid physically challenged people, deliver language education, aid in corporate communication, handle transactions, and appear on audio books, documentaries, or the news. One of the most visible areas where these voices are used is advertising. Some companies, especially local businesses, use synthetic voices to announce their products, as the production is cheaper and faster than paying a voice-over artist. Therefore, there is a growing interest both in improving the quality of these voices and in analyzing how they are processed, especially as some authors have shown that prosody parameters can affect credibility, attitudes toward the ad, and intent to buy a product (Gélinas-Chebat, Chebat & Vaninsky, 1996).

This research has two main goals: a) to analyze the effectiveness, attention and recall of human and artificial voices narrating an advertising story, where prosody is modified; and b) to study the evaluation of these voices depending on their functions in these devices. The research questions are as follows: a) what type of voice (human or artificial) is better perceived and processed in an advertising story in which prosody plays an important role; and b) what functions both human and artificial voices better carry out?

2. Prosody quality

Intelligibility and naturalness are the two main parameters by which an artificial voice is valued (Delogu, Conte, & Sementina, 1998; Nass & Min Lee, 2001; Paris, Thomas, Gilson, & Kincaid, 2000). Intelligibility is the parameter that makes a voice understandable while naturalness is the quality by which a voice sounds more similar to human speech (Nusbaum, Francis, & Henly, 1995). Synthetic voices are artificial voices that can be automatically built using different methods. The parametric speech systems are based on a model of human speech production. These systems were later improved and, in the 1990s, unit-selection synthesis systems were developed (Hinterleitner, Norrenbrock, Moller, & Heute, 2012, pp. 240–245). As a result, current synthesizers are now able to produce a more natural and intelligible sound. Wolters, Johnson, Campbell, DePlacido, and McKinstry (2014) showed differences among speech synthesis systems when measuring recall of messages. For this reason, this study measures two kinds of artificial voices: synthetic, using a computer-generated voice; and human manipulated voice, transforming a human voice.

In spite of overcoming the problem with intelligibility, all

E-mail address: emma.rodero@upf.edu.

http://dx.doi.org/10.1016/j.chb.2017.08.044
0747-5632/

synthetic systems still have some degradation, meaning a lower quality of the signal compared to human voices (Hinterleitner, Norrenbrock, Möller, & Heute, 2014). Consequently, synthetic voices are still perceived as less natural than human voices (Gong & Lai, 2003; Wolters et al., 2014; Hinterleitner et al., 2014). Mirenda, Eicher, and Beukelman (1989) concluded that an artificial voice should have two qualities – it has to be not only highly intelligible but also natural. Therefore, naturalness is now one of the main concerns (Hinterleitner et al., 2014), and this naturalness is in part provided by prosody quality.

Voice is the instrument used by speakers, providing speaker identification (how the speaker sounds); in contrast, prosody is how this speaker uses the voice parameters (how he speaks). Prosody plays an essential role when we listen to a voice, as it provides expressiveness (sense and emotion) to speech and consequently naturalness, one of the most important traits in human communication (Giles, 1973; Sreenivasa Rao, 2012; Pauletto, Balentine & Pidcock, 2013). Prosody, as a relevant component in human interaction (Pittam, 1994), may influence the listener's information processing (Hirschberg, 2005; Levi & Pisoni, 2007; Rodero, 2015a). Prosodic cues, composed by intonation, stress and rhythm, provide the melody and the parsing of speech (Nooteboom, 1997). For instance, some studies have considered that speech rate improves intelligibility in synthetic voices (Delogu et al., 1998; Marics & Williges, 1988; Slowiaczek & Nusbaum, 1985). These prosody features help segment the discourse and provide an acoustic signal for listeners to understand speech, which ultimately may enhance cognitive processing (Rodero, 2015a; Sanderman & Collier, 1997). All in all, this host of prosody traits – if employed efficiently – can help to lend expressiveness to a speaker's message, and, consequently, draw attention and facilitate understanding (Rodero, 2006, 2007).

However, prosody representation in synthetic voices is still poor (Grice, Vagges, & Hirst, 1991; Salza, Fabbrizio, Oreglia, et al., 1993; Delogu et al., 1998). Synthetic voices cannot emulate the prosody patterns of human voices, at least with current technology. Consequently, expressiveness is affected. Hennig and Chellali (2012) considered that expressiveness, conveyed by prosody, is an important characteristic in artificial voices. Mayo, Clark, and King (2011), as well as Vainio, JaÅNrvikivi, and Werner (2002), showed the weight of intonation for synthetic speech naturalness. This is highly relevant since naturalness and expressiveness are important factors in some current applications utilizing artificial voices, especially in narrative texts such as advertising, documentaries and audio books. Cabral, Oliveira, Raimundo, et al. (2006) found that people rated human voices as better than synthetic voices when a story is narrated. Their results indicate that the type of voice was significantly relevant for the proper understanding of the story, the expression of appropriate emotional content, the credibility of the voice, and the satisfaction with the voice. However, these authors did not study the listener's information processing. The question thus arises about how artificial voices, as compared to human ones, are processed when prosody is a decisive factor, as, for example, in a narrative text where expressiveness is very relevant. Therefore, the first goal of this study is to analyze the perception and processing of these voices (effectiveness, attention, concentration, and recall) when modifying prosody in an advertising story.

3. Information processing of artificial voices

During the past 40 years, numerous studies have analyzed the perception of artificial voices, especially by comparing them to human voices (Syrdal, Bennett, & Greenspan, 1994; Winters & Pisoni, 2004; Chen, 2006). These works have shown perceptual differences between both voices, which can be explained due to the poorest signal and prosodic quality in artificial voices comparing to human ones. A study by Terken and Lemeer (1988) concluded that the presence of prosody cues in synthetic speech improved speech's attractiveness. These results are interesting because the authors assumed that the perception of prosody was strongly dependent on segmental intelligibility. When intelligibility is poor, listeners do not have enough time to perceive prosodic differences. However, when intelligibility improves, prosody becomes more influential. Therefore, prosody seems to act at a second, higher level. Nusbaum, Schwab, and Pisoni (1984, pp. 391–408) showed that listeners rated the human voice with more positive attributes (friendly, smooth, easy, clear, pleasant, etc.) than they did the synthetic voice. Listeners also were more confident with a human voice when asked if they comprehended the passages correctly.

In the first part of this study, we extend the current research by analyzing how effective artificial voices are compared to human ones. The first research question is which voice – artificial or human – is better perceived regarding effectiveness? Effectiveness is defined in this study as the ability to produce a better effect or result. To answer this question, we used a self-reported scale of effectiveness. We asked participants how clear, correct, pleasant, credible, persuasive and comprehensible each voice was. Considering the previous studies in which human voices were evaluated, we can hypothesize that human voices will be assessed as more effective than the artificial ones. This leads to formulate the first hypothesis:

H1. Human voices will be rated as more effective than artificial ones, both human manipulated and synthetic, when prosody is the differential factor among them.

Effectiveness is the first dependent variable analyzed in this study. The second different variable, not interrelated, is attention. A poor signal and the lack of naturalness of artificial voices comparing to human ones could affect the encoding phase of the listener's information processing (Pisoni, 1997; Winters & Pisoni, 2004). This factor is important to listeners. As the identification of every phoneme in synthesis speech grows more difficult, the listener must apply a greater cognitive effort (Gorenflo & Gorenflo, 1997; Roring, Hines, & Charness, 2007; Taake, 2009; Winters & Pisoni, 2004), and this effort can affect attentiveness (Delogu et al., 1998). This effort has been shown in different tasks where the listener's response was longer than the listener's time response to human voices (Winters & Pisoni, 2004). Information processing of artificial voices may require more time and more cognitive resources than the processing of human voices (Delogu et al., 1998; Duffy & Pisoni, 1992; Luce, Feustel, & Pisoni, 1983). Therefore, listeners feel that understanding of synthetic voices is more difficult (Lai, Cheng, Green, & Tsimhoni, 2001). The key question to ask is what would happen in certain challenging situations when listeners do not have much time to process information or when they are not willing to devote much effort to process the message. Examples might be a phone call answered by a synthetic voice, or an ad, audiobook, or YouTube documentary narrated by an artificial voice (Stern, Chobany, Patel, & Tressler, 2014). This is the reason why this study analyzes a narrative text instead of isolated sentences or words, as in previous research.

The second research question is what type of voice – artificial or human – will result in more attention and concentration when narrating a story? In this study, we measure these two variables – attention and concentration – with a retrospective scale (Potter & Choi, 2006). Since it is a narrative story, we expected that listeners pay less attention, due to a less interest, when artificial voices are used, but they feel more effort in the form of concentration. This leads us to the study's second hypothesis:

H2. Human voices will achieve greater attention than artificial voices, both human manipulated and synthetic, when prosody is a differential factor among them. They also will require less concentration than the latter.

Finally, the last part of the processing process is recall, which affects the retrieval process. Some studies have not found differences when analyzing recall and comprehension of synthetic and human voices (Nye, Ingemann & Donald, 1975; Pisoni & Hunnicutt, 1980; Pisoni, Manous, & Dedina, 1987; Paris, Gilson, Thomas, & Silver, 1995; Delogue et al., 1998; Lai, Wood, & Considine, 2000; Taake, 2009). Luce (1981, pp. 229–242) suggested that this equal level of comprehension might be a consequence of a higher degree of cognitive effort. As the listener devotes more time and more cognitive resources to encoding synthetic speech, the subsequent recall and comprehension may be enhanced (Moody & Joost, 1986; Winters & Pisoni, 2004), especially if the text to recall is simple (Jenkins & Franklin, 1982). Paris et al. (2000) found beneficial recall effects of prosody when applied to synthetic voices. These authors concluded that prosodic representation is a powerful guide that eases the listener's information processing, and that prosody is decisive when rating speech as natural, influential and intelligible. Sanderman and Collier (1997) showed the influence of prosody in the comprehension of synthetic speech. This study found that inappropriate prosodic contours hindered comprehension. Wolters et al. (2014) measured recall of different messages containing medication names. These authors concluded that, for short messages (reminders of medications), a high-quality synthetic voice was as well recalled as a human voice. When the participants heard repetitions of the medication names, the authors found that the unit selection system obtained a higher level of recall than statistical parametric synthesis. The key question here is what would happen when the message is not a simple instruction. Therefore, the third research question of this study is how recall is affected when these artificial voices narrate a story where prosody is important. Based on these studies, our hypothesis is as follows:

H3. Human voices will achieve a better level of recall than artificial voices, both human manipulated and synthetic, when prosody is the differential factor among them.

Finally, as companies, nonprofits or government organizations increasingly use artificial voices in a growing number of areas, analyzing how they are rated depending on their use or functions is crucial. This is the second goal of this research. In addition, as advertising more frequently uses artificial voices, we added a question to rate the suitability of these voices for advertising and to gauge the usefulness of the various voices. The fourth research question is what functions can fulfill the two types of voices – artificial and human – related to their use in devices? To answer this question, the participants rated the extent to which they would be persuaded by these voices to buy a product, go to a destination, carry out bank transactions, request information, and use as an audio-guide, all of which are common functions in which artificial voices are used. The participants previously evaluated the level of complexity of these functions (how difficult was each one). Our hypothesis here is as follows:

H4. Human voices will be preferred over artificial ones (human manipulated or synthetic) when they fulfill a more complex function.

We completed the evaluation part with a questionnaire about the knowledge and preferences for artificial voices.

4. Method

4.1. Participants

A gender-balanced sample of university students ($N = 200$) in communication (21–23 years old) was selected to listen to the corpus of this study. For a MANOVA F test, ($f^2 = 0.06$) the power analysis indicated that the estimated sample size would be 160 participants; 94 males and 106 females students were randomly selected from the gender group, using alphabetical class lists. The questionnaire included an initial question about their proficiency in Spanish. All the questionnaires were valid, so this was the final sample.

4.2. Message stimuli

The message was selected from a pool of awards radio advertisements. Sixteen narrative ads were selected due to their simple narration structure, one narrator, and simple syntactic structures. These texts were modified to remove the name of the brands and then were rated by a sample of ten respondents. The final text selected was the best rated in terms of the story's interest. The message was a brief and simple story in an ad format in which the expressive prosodic effect could be applied (see Appendix). The length was short (20 s).

This text was recorded in Peninsular Spanish with all the types of voices analyzed in this study. Two synthetic voices were used. The voice for Siri (Apple) served as the first female artificial voice, and the male voice of Loquendo in Peninsular Spanish was used as the second artificial voice. Loquendo is a text-to-speech (TTS) application (Nuance). Then we used two human voices, male and female, which were modified using KaleiVoiceCope software (Mayor, Bonada, & Janer, 2009). KaleiVoiceCope (KVC) is a voice transformation technology, which modifies voice by controlling spectral and physical characteristics or modifying timbre. In this study, we transformed human voices to sound as artificial ones. Synthetic voices sounded slightly more intelligible and clear than human manipulated voices. In total, we obtained four artificial voices – two human manipulated voices (KVC1 and KVC2) and two synthetic voices (Loquendo and Siri). These four voices were assessed by a sample of 20 informants, who rated the voices in a 7-point-scale (natural-artificial voice). All the voices were rated as artificial (between 6.55 and 6.85), and there were no significant differences among them ($F = 1.39$; $p = 0.251$).

Once the two first artificial voices were selected, two more human speakers (professional voice-over artists) were chosen (male and female); they were not exceedingly different in pitch compared to the artificial ones (Siri and Loquendo) to avoid deviations owing to the type of voice. In total, we collected four human voices (two of them modified with KaleiVoiceCope).

The prosody features analyzed in this study are intonation (pitch level, pitch range, and contour), speech rate, and pauses. Pitch is the perceptual correlate of the fundamental frequency – F0 – and is measured in Hertz (Nooteboom, 1997). Pitch is a voice feature, which identifies a speaker's characteristics – gender, age, size… (Schötz, 2006). Two types of pitch parameters are used in intonation (Gussenhoven, 2004) – pitch level and pitch range. Pitch level represents the average of pitch used in the curve and has been found to correlate with persuasive speech, charisma (Rosenberg & Hirschberg, 2009; Signorello, D'Errico, Poggi, Demolin, & Mairano, 2012), and credibility (Rodero, Mas, & Blanco, 2017). Pitch range is the difference between the highest and the lowest pitch level that a speaker uses when speaking. An expanded pitch range was related to more expressiveness and charisma (Niebuhr, Voβe, & Brem, 2016). Along with this, the form that the curve of

intonation adopted is relevant for expressiveness. If a speaker wants to sound more credible and natural, he has to use a smooth intonation contour (Rodero et al., 2017). The other important prosody feature is related to duration structure. Speech rate, along with pauses, is an important feature in speech perception (Nooteboom, 1997) and influences information processing of the message (Rodero, 2015b), the speaker's credibility (Rodero et al., 2017), and charisma (Rosenberg & Hirschberg, 2009).

As the aim of this study was to minimize the risk of choices being made based on the pitch of voice, the voices should sound alike, although one was human and the other was artificial. Pitch level – that is, the mean of pitch in the entire fragment – was measured in semitones. There were no statistical differences in the average of pitch level among voices ($F = 6.42$; $p = 0.064$; partial $\eta^2 = 0.77$), between artificial and human female voices ($F = 0.088$; $p = 0.794$; partial $\eta^2 = 0.04$), and between artificial and human male voices ($F = 0.119$; $p = 0.763$; partial $\eta^2 = 0.04$). The results gleaned from this analysis are set out in Table 1.

The next step was to analyze the other prosody parameters – pitch range, contour, speech rate and pauses of the artificial voices – using Praat speech analysis software (Boersma & Weenink, 2017). This analysis served to modify these features in the narration conveyed by human voices. The professional speakers recorded the same text but in a more expressive manner varying intonation, speech rate, and pauses. The texts were recorded in a professional sound studio with optimal acoustical conditions (a Sennheiser MD 421 II microphone and a sample rate of 44.100 samples per second, 16 bits, stereo) modifying the prosody parameters according to the instructions. The results were analyzed following the same method. Therefore, the voices were different, regardless of their nature (human non-manipulated voices, human manipulated, and synthetic voices) in the prosody used to read the text (intonation, speech rate, and pauses).

Regarding intonation, the voices were distinguished in pitch range (the difference between the maximum pitch and the minimum pitch, measured in semitones), as Table 1 shows. The analysis showed statistical differences in pitch range among voices ($F = 27.35$; $p = 0.004$; partial $\eta^2 = 0.77$), between artificial and human female voices ($F = 19.57$; $p = 0.047$; partial $\eta^2 = 0.05$), and between artificial and human male voices ($F = 86.61$; $p = 0.011$; partial $\eta^2 = 0.05$). The voices also were different in intonation contour. To examine this variable, the eight voices were analyzed with the speech analysis software Praat (Boersma & Weenink, 2017) with the MOMEL and INTSINT modules (Hirst, 2007). Intonation was analyzed by obtaining the curve and labeling the pitch contour using MOMEL and INTSINT. These two plugins for the acoustic analysis software Praat (Boersma & Weenink, 2017) allowed an analysis and labeling of intonation, according to the model of Hirst (2007). This model makes a distinction between pitch segments interpreted globally with regard to the median tone of the speaker (T – top, M – mid and B – bottom) and the segmented tones that were interpreted locally in terms of the preceding tones: Higher (H), a pitch peak; Lower (L), a pitch valley; Same (S), same tone as the preceding tone; Upstep (U), an increase with a peak lower than H; and Downstep (D), a decrease with a valley lower than L.

The intonation curves were modeled using the MOMEL module and then labeled using INSINT. Fig. 1 shows the results of this analysis. The intonation of the human voices was more dynamic, and by extension more expressive than the intonation of the artificial voices, because human voices used a greater pitch range. Hence, the curve was more irregular. The intonation curve for the synthetic voices always moved within the same pitch range (visible in Fig. 1 in the repeated triangle shapes pointing to a regular intonation). The labeling shows a similar manner of intonating sentences independently from the content. On the other hand, with the human voice expressivity is heightened as the intonation varies according to the content and, therefore, the ends of blocks of meaning do not bear the same form.

Concerning speech rate and pauses, the artificial voices had a more regular speech rate and more regular pauses than the human voices. Natural speakers do not read at a regular or constant pace, and they make more irregular, longer significant pauses, a factor that does not occur with an artificial voice that is instead characterized by short pauses at regular intervals. Speech rate was measured in syllables per second including articulation rate plus pauses duration. There were statistical differences in the mean of speech rate among all the voices, ($F = 61.11$; $p < 0.001$; partial $\eta^2 = 0.93$). Statistical differences in the number of pauses and the duration of the pauses were also found. Pauses were longer than 150 ms, as this duration was the minimum length of strategic pauses (after coma or full stop) in artificial voices. Regarding the number of pauses, there were differences among all voices ($F = 16.33$; $p = 0.010$; partial $\eta^2 = 0.97$. Also, there were differences among voices in the duration of the pauses ($F = 26.14$; $p = 0.004$; partial $\eta^2 = 0.97$).

In conclusion, the voices were quite similar in their pitch level, and they were different in the type (human non-manipulated voices, human manipulated, and synthetic voices) and in the intonation – pitch range and contour – speech rate, and pauses.

4.3. Measures

Effectiveness. We use a scale of a previous study describing voices (Rodero, Larrea, & Vázquez, 2013). The most employed adjectives in this scale, scoring 0.5 or higher, were – pleasant, correct,

Table 1
Prosodic data of the voices.

Voice	Mean of pitch level (semitones)	Mean of pitch Range (semitones)	Speech Rate (syllables per second)	Pauses duration (milliseconds)	Number of pauses
Human non-manipulated Female Voice 1	8.69	10.18	5.19	487	9
Human non-manipulated Female Voice 2	12.03	11.7	5.27	404	8
Human non-manipulated Male Voice 1	2.47	8.37	5.13	450	7
Human non-manipulated Male Voice a 2	5.38	6.49	5.18	481	7
Synthetic Female Voice	9.22	0.15	7.13	161	4
Synthetic Male Voice	1.68	−11.35	6.17	224	6
Human manipulated Female Voice	13	−5.91	6.71	228	5
Human manipulated Male Voice 2	4.72	−8.17	6.19	196	5

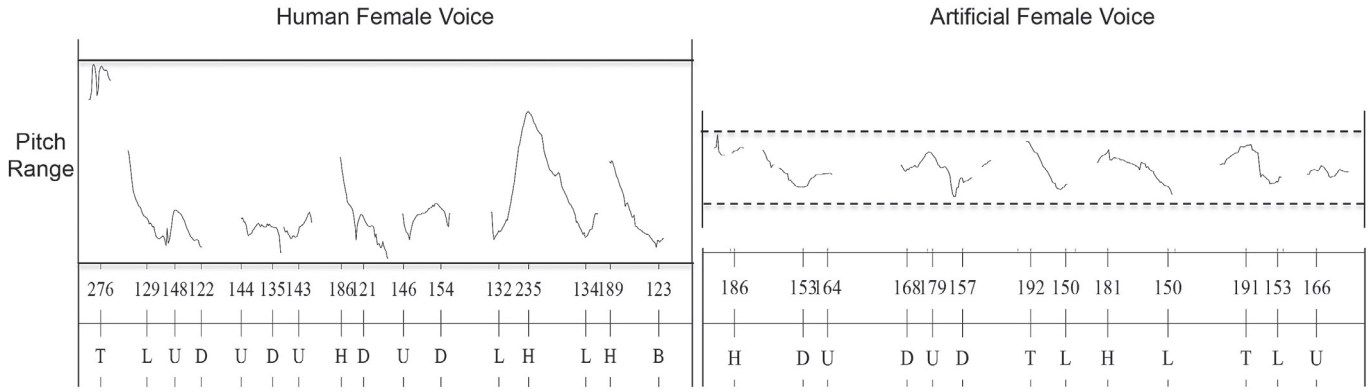

Fig. 1. Example of pitch contour, pitch range and lnsint analysis.

clear, pertinent, calm, credible, persuasive, authoritarian, natural, appropriate, comprehensible, and dynamic. We then performed a factor analysis. Five items (clear, correct, pleasant, credible, persuasive and comprehensible) had loadings that surpassed 0.74. Thus, we use a 7-bipolar scale of opposite pairs (1 was the lowest level). Participants rate how clear, correct, pleasant, credible, persuasive and comprehensible the voices were. The average of all these variables was the effectiveness index. The Cronbach's Alpha coefficient for effectiveness, $\alpha = 0.728$, showed internal consistency.

Attention and concentration. These variables were measured with a retrospective scale (Potter & Choi, 2006). This scale was employed in previous studies (Rodero, 2015a). The Cronbach's Alpha coefficient, $\alpha = 0.974$, showed a high internal consistency. The subjects had to answer two questions: how much attention they had paid throughout the story, and how much they had concentrated on it. We explained to the participants the difference between these two questions. Attention was the level of focusing on the story and concentration was the effort to follow the story. The answers were quantified on a 7-points scale (1 was the lowest level of attention and concentration). Therefore, this study only included self-perception of attention. That decision was intentional and intended to help researchers know what participants consciously thought about their attention and their effort.

Recall. The recall variable was formed by three open questions. These were content-questions with information about the story (e.g., What time did the protagonists go fishing?). For all the questions, there was only a correct answer. The answers of the participants were coded in three groups: hits, errors or non-answers. The formula applied to quantify the final recall level was the number of hits minus the errors divided by 3: H-(E/3). This is a calculation that allows quantifying correct answers, errors, and no-answers.

Evaluation. First, to complete these scales, we added a question to determine the extent to which the respondents believed the voices could be deemed as suitable for advertising. They were required to rate their suitability on a scale of 1–5. Secondly, to gauge the usefulness of the voices, the participants rated the extent to which they would be persuaded by these voices to buy a product, go to a destination, carry out bank transactions, request information, and use as an audio-guide. These functions also were assessed by the participants in order of complexity (from 1 to 5). The final order was to request information ($M = 1.12$, $SD = 0.23$), to use as an audio-guide ($M = 1.82$, $SD = 0.41$), to go to a destination ($M = 2.88$, $SD = 0.37$), to buy a product ($M = 4.15$, $SD = 0.19$), to carry out bank transactions ($M = 4.89$, $SD = 0.31$).

Finally, the second part of the survey measured their level of knowledge and preferences for artificial voices. Table 2 sets out the questionnaires and scales.

4.4. Design

The experiment was a 3 type of voices (human non-manipulated/human manipulated/synthetic voices) mixed factorial between-subjects design applied to effectiveness, attention, concentration, recall, suitability for advertising, and five different functions. Participants were randomly divided into eight groups and assigned to each group.

4.5. Procedure

The sample was divided into eight groups (for each voice in the corpus) of 25 subjects, also gender-balanced. Each group listened to one voice in the same room but in different hours. Once the participants listened to the voice, they completed the questionnaire. The participants answered the effectiveness, attention, concentration, recall and evaluation scales. Finally, they completed the knowledge and preferences questionnaires. The duration of each session was 10 min approximately. No incentives were provided.

5. Results

The results were measured by applying a variance analysis in the three types of voices (human non-manipulated/human manipulated/synthetic voices) factorial MANOVA for five dependent variables – effectiveness of the voice; attention; concentration; recall of the listener; suitability for advertising; – and five functions – to buy a product; to go to a destiny; to do bank transactions; to demand information; and to make as audio guide. Box's M test of homogeneity of covariance was significant ($p < 0.001$) as well as Levene's homogeneity of variance for all the variables except concentration ($p = 0.140$). As the principle of homoscedasticity was violated, two tests were applied to show whether or not differences between the groups were statistically significant. Both the Welch test and the Kruskal-Wallis H test were significant for all the variables. Table 3 shows the descriptive statistics.

The dependent variables had significant main effects for type of voice, $F(7, 193) = 91.3$, $p < 0.001$, partial $\eta^2 = 0.98$. Table 4 shows the zero-order correlation matrix of all final study variables.

5.1. Hypothesis 1. Effectiveness

The first hypothesis suggested that human voices would be rated as more effective than artificial ones, both human

Table 2
Questionnaires of the experiment.

Questionnaire 1. Knowledge and level of perception of artificial voices	
Question	Evaluation
Are you familiar with artificial voices?	Yes/No
What is your opinion about them?	Very negative Negative Regular Positive Very positive
Why?	Open question
What is the best function for an artificial voice?	Open question
What voice do you prefer to sell you a product?	Human voice Artificial voice The same
What kind of voice do you prefer to sell a product?	Male voices Female voices Children voice The same

Questionnaire 2. Scales and indexes								
Scale	Evaluation							
Effectiveness scale	Unclear							Clear
	Incorrect							Correct
	Unpleasant							Pleasant
	Not credible	1	2	3	4	5	6	7 Credible
	Not persuasive							Persuasive
	Incomprehensible							comprehensible
Attention scale		1	2	3	4	5	6	7
Concentration scale		1	2	3	4	5	6	7
Recall	Five questions:Hits-(Errors/3)							
Adequacy for advertising			1	2	3	4	5	
In what extent are you persuaded … human or artificial voices:								
To buy a product			1	2	3	4	5	
To go to a destination			1	2	3	4	5	
To carry out bank transactions			1	2	3	4	5	
To request information			1	2	3	4	5	
To use as an audio guide			1	2	3	4	5	

Table 3
Descriptive statistics.

	Type of voice	Mean	Deviation
Level of effectiveness	Human non-manipulated	6.09	0.63
	Synthetic	3.39	1.01
	Human manipulated	2.88	0.83
Level of attention	Human non-manipulated	4.04	0.75
	Synthetic	3.35	0.93
	Human manipulated	2.92	1.03
Level of concentration	Human non-manipulated	1.88	0.74
	Synthetic	3.92	0.89
	Human manipulated	4.10	0.94
Level of recall	Human non-manipulated	5.81	0.74
	Synthetic	3.42	0.86
	Human manipulated	2.93	1.03
Adequacy for advertising	Human non-manipulated	4.38	0.72
	Synthetic	2.14	0.93
	Human manipulated	1.76	0.80
Buy a product	Human non-manipulated	3.93	0.92
	Synthetic	2.10	0.96
	Human manipulated	1.53	0.76
Go to a destination	Human non-manipulated	3.92	0.98
	Synthetic	2.78	1.33
	Human manipulated	2.47	1.2
Carry out bank transactions	Human non-manipulated	3.64	0.87
	Synthetic	2.53	1.2
	Human manipulated	1.88	0.97
Request information	Human non-manipulated	3.92	0.87
	Synthetic	2.82	1
	Human manipulated	2.59	1
Do as audio-guide	Human non-manipulated	4.06	0.87
	Synthetic	3	1.2
	Human manipulated	2.43	1.3

Note: All the results are significant.

manipulated and synthetic, when prosody is a differential factor among them. The differences among the three types of voices were significant, $F(2, 197) = 350.49$, $p < 0.001$, partial $\eta^2 = 0.77$, with a high consistency. Human non-manipulated voices were judged as more effective ($M = 6.09$; $SD = 0.63$) than synthetic ($M = 3.45$; $SD = 0.81$) and human manipulated voices ($M = 2.89$; $SD = 0.92$). The effect sizes for the comparisons between human non-manipulated and synthetic voices (d = 3.44) and between human non-manipulated and human manipulated voices (d = 3.87) were found to exceed Cohen's (1988) convention for a large effect (d = 0.80). The comparison between human manipulated and synthetic voices (d = 0.64) had a medium effect. The post-hoc test showed that there were significant differences among the three types of voices: between human non-manipulated with human manipulated and with synthetic voices, $p < 0.001$, and between human manipulated and synthetic voices, $p = 0.005$. Therefore, the data reveal that the first hypothesis can be supported.

5.2. Hypothesis 2. Attention and concentration

The second hypothesis established that human voices would achieve greater attention than artificial voices, both human manipulated and synthetic, when prosody is the differential factor among them. They also would require less concentration than the latter. The data showed significant differences in these two variables: attention, $F(2, 197) = 29.46$, $p < 0.001$, partial $\eta^2 = 0.23$, and concentration, $F(2, 197) = 163.99$, $p < 0.001$, partial $\eta^2 = 0.62$.

Regarding attention, human non-manipulated voices obtained a higher level of attention ($M = 4.04$; $SD = 0.75$) than synthetic ($M = 3.35$; $SD = 0.93$) and human manipulated voices ($M = 2.92$; $SD = 1.03$). The effect sizes for the comparisons between human non-manipulated and synthetic voices (d = 0.81) and between human non-manipulated and human manipulated voices (d = 1.82) had a large effect. The comparison between human manipulated and synthetic voices (d = 0.43) had a medium effect. The post-hoc test showed that there were significant differences between human non-manipulated with human manipulated and with synthetic voices, $p < 0.001$, but there were no differences between human manipulated and synthetic voices, $p = 0.088$.

Regarding concentration, human non-manipulated voices obtained a lower level of concentration ($M = 1.88$; $SD = 0.74$) than synthetic ($M = 3.92$; $SD = 0.89$) and human manipulated voices ($M = 4.10$; $SD = 0.94$). The effect sizes for the comparisons between

Table 4
Means, standard deviation and Pearson correlation matrix (n = 200).

	M	SD	1	2	3	4	5	6	7	8	9	10	11
1. Type of voice	–	–	–										
2. Effectiveness	4.63	1.67	-0.85**	(0.77)									
3. Attention	3.59	0.99	-0.37**	0.48*	(0.23)								
4. Concentration	2.95	1.35	0.78**	-0.72**	-0.22**	(0.62)							
5. Recall	4.49	1.58	-0.83**	0.90**	0.43**	-0.76**	(0.70)						
6. Adequacy for advertising	3.09	1.54	-0.83**	0.83**	0.48**	-0.68**	0.82**	(0.69)					
7. Buy a product	2.87	1.40	-0.76**	0.75**	0.45**	-0.66**	0.73**	0.84**	(0.59)				
8. Go to a destination	3.27	1.32	-0.46**	0.54**	0.45**	-0.33**	0.51**	0.58**	0.65**	(0.24)			
9. Carry out bank transactions	2.92	1.25	-0.55**	0.52**	0.24**	-0.43**	0.49**	0.63**	0.66**	0.55**	(0.36)		
10. Request information	3.31	1.14	-0.44**	0.51**	0.43**	-0.33**	0.48**	0.66**	0.60**	0.62**	0.62**	(0.28)	
11. Do as audio-guide	3.39	1.31	-0.47**	0.53**	0.42**	-0.42**	0.49**	0.64**	0.60**	0.48**	0.59**	0.75**	(0.28)

Correlation is significant at the 0.01 level (2-tailed)**.
Cronbach's alphas are shown in the diagonal.

human non-manipulated and synthetic (d = 2.49) and between human non-manipulated and human manipulated voices (d = 0.19) had a large effect. The comparison between human manipulated and synthetic voices (d = 0.43) had a small effect. The post-hoc test showed that there were significant differences between human non-manipulated with human manipulated and with synthetic voices, $p < 0.001$, but there were not differences between human manipulated and synthetic voices, $p = 0.693$. Therefore, the second hypothesis can be accepted.

5.3. Hypothesis 3. Recall

The third hypothesis suggested that human voices would achieve a better level of recall than artificial voices, both human manipulated and synthetic, when prosody is the differential factor among them. The results were also significant in this variable, $F(2, 197) = 240.23$, $p < 0.001$, partial $\eta^2 = 0.70$. The content of the story narrated by human non-manipulated voices was better recalled ($M = 5.81$; $SD = 0.74$) than synthetic ($M = 3.42$; $SD = 0.86$) and human manipulated voices ($M = 2.93$; $SD = 1.03$). The effect sizes for the comparisons between human non-manipulated and synthetic (d = 2.97) and between human non-manipulated and human manipulated voices (d = 3.21) had a large effect. The comparison between human manipulated and synthetic voices (d = 0.51) had a medium effect. The post-hoc test showed that there were significant differences among the three types of voices: between human non-manipulated with human manipulated and with synthetic voices, $p < 0.001$, and between human manipulated and synthetic voices, $p = 0.036$. Therefore, hypothesis 3 can also be supported. Fig. 2 illustrates the results for all of these variables.

5.4. Hypothesis 4. Evaluation

The fourth hypothesis established that human voices would be preferred over artificial ones (human manipulated or synthetic) when they fulfil a more complex function. Regarding this aspect, the data showed that there were significant differences between artificial and human voices in all the addressed activities: to buy a product, $F(2, 197) = 142.768$, $p < 0.001$, partial $\eta^2 = 0.59$; to go to a destination, $F(2, 197) = 31.94$, $p < 0.001$, partial $\eta^2 = 0.24$; to conduct bank transactions, $F(2, 197) = 55.86$, $p < 0.001$, partial $\eta^2 = 0.36$; to request information, $F(2, 197) = 39.77$, $p < 0.001$, partial $\eta^2 = 0.28$; and to serve as an audio guide, $F(2, 197) = 38.95$, $p < 0.001$, partial $\eta^2 = 0.28$. The post-hoc test showed that there were significant differences among the three types of voices: between human non-manipulated with human manipulated and with synthetic voices, $p < 0.001$, for all the variables. Regarding the differences between human manipulated and synthetic voices, there were only significant differences in the next variables: to buy a product, $p = 0.005$, and to conduct bank transactions, $p = 0.013$. There were not significant differences in the rest of the variables: to go to a destination, $p = 0.546$, to request information, $p = 0.616$, and to serve as an audio guide, $p = 0.057$.

Therefore, as shown in Table 3, artificial voices received a lower score in all the functions, especially human manipulated voices. The functions were rated in order of complexity: to use as an audio-guide; to request information; to go to a destination; to buy a product, and to carry out bank transactions. The outcome order was the same rated previously by the participants in complexity, regardless of the two first functions: to use as an audio-guide and to request information. The lower scores were obtained by functions in which money comes into play: purchasing a product and carrying out a bank transaction. Fig. 3 shows this data.

Regarding suitability for advertising, the results of the ANOVA analysis pointed that the artificial voices were considered unsuitable, especially human manipulated voices ($M = 1.76$; $SD = 0.86$) and synthetic ($M = 2.14$; $SD = 0.93$) compared to the human non-manipulated voices ($M = 4.38$; $SD = 0.72$), with statistically significant difference, $F(2, 197) = 232.44$, $p < 0.001$, partial $\eta^2 = 0.69$. The effect sizes for the comparisons between human non-manipulated and synthetic (d = 2.69) and between human non-manipulated and human manipulated voices (d = 3.30) had a large effect. The comparison between human manipulated and synthetic voices (d = 0.42) had a medium effect. The post-hoc test showed that there were only significant differences between human non-manipulated with human manipulated and with synthetic voices, $p < 0.001$. There were no significant differences between human manipulated and synthetic voices, $p = 0.090$.

Concerning the questionnaires, the first question dealt with the degree of familiarity the respondents had with the artificial voices after being told what was understood by an artificial voice. Although these university students were young and they utilize technology in their everyday lives, most of them – 61.5% – stated they were unfamiliar with artificial voices while 38.5% stated that they were accustomed to them.

The second section of the questionnaire asked how they perceived the artificial voices. Most respondents have an intermediate or regular value of 56%; the second highest group gave a negative value (22.5%), followed by a positive assessment (16.5%) and finally a highly negative assessment (5%). No one rated these voices as being particularly positive. Again, no differences owing to gender arose from the sample ($\chi^2 = 0.395$). When asked why these voices were valued in such a manner, the most commonly cited adjectives were as follows – because they are more *insecure, non-human, unnatural, fake, annoying, distant, untrustworthy, unreliable, less understandable, impersonal, more distant, and more*

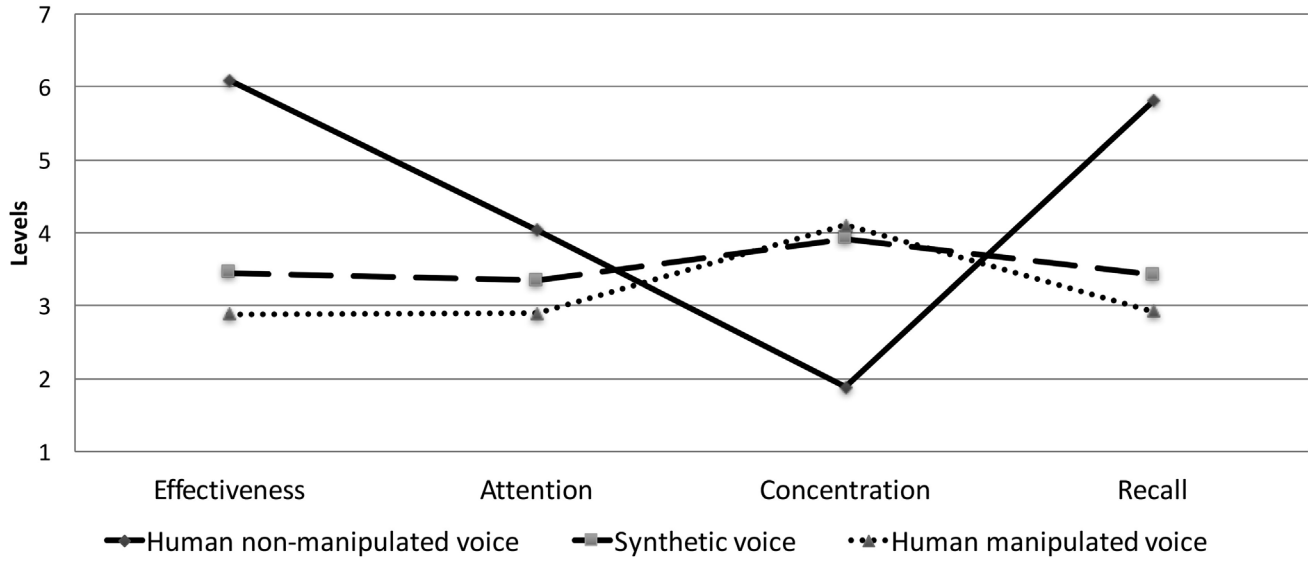

Fig. 2. Processing of human and artificial voices.

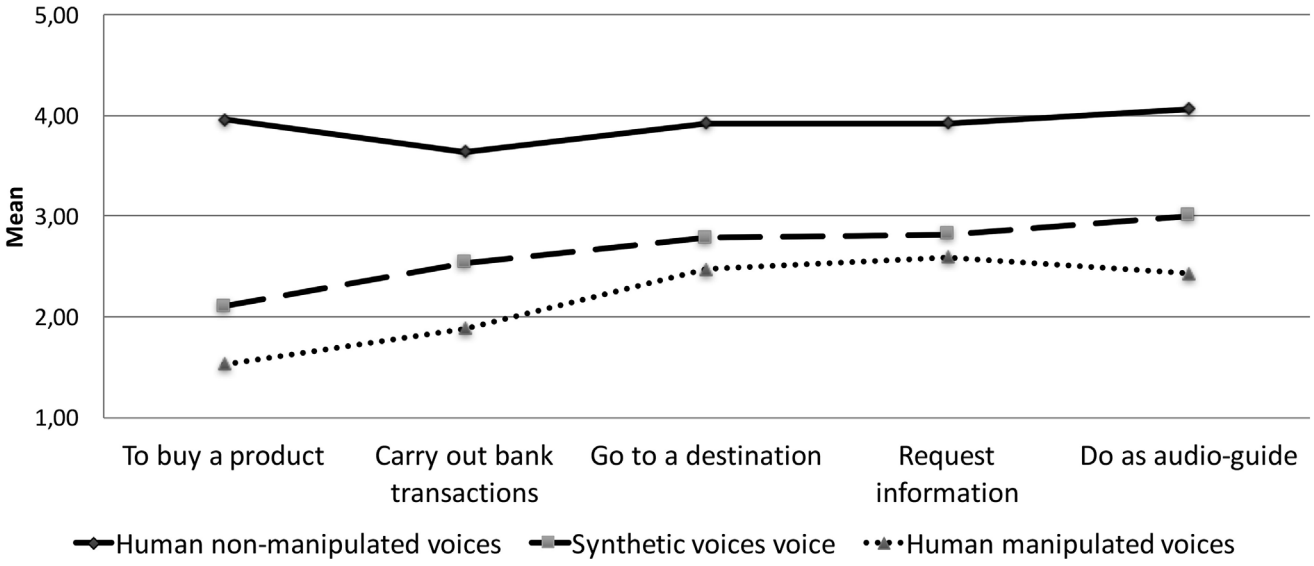

Fig. 3. Functions of human and artificial voices.

artificial.

Another question was to examine the functions that artificial voices would carry out. Here, most responses pointed to technology, devices and machines, adding that the specific functions should be linked to short, rational and simple informational, repetitive or indicative messages, according to the results by Wolters et al. (2014).

About preferences for voices when concerning the selling of a product or the offering of services, most respondents stated that human voices were preferable – 79.5% – followed by those who stated they were indifferent – 15% – and those who opted for artificial voices – 5.5%.

Finally, about the preference for voices according to gender, 45.5% of respondents stated that they preferred a male voice, 29% a female voice, and 25.5% stated they were indifferent. In this instance, there was a significant gender difference in the sample ($\chi^2 = 0.000$). A crossover effect arose as in the study by Rodero et al. (2013) and Mullennix, Stern, Wilson, and Dyson (2003) in which women gave a better rating to the male voice while men gave it the lowest rating, confirming the gender-based differences (Crowell, Scheutz, Schermerhorn, & Villano, 2009; Nass & Brave, 2005).

6. Discussion

This research contributes to studies of the processing of artificial voices, which are more and more common in different devices (phones, tablets, etc.), transportation (airports, train stations, etc.), or communication (advertising, YouTube, audiobooks, etc.). The study's objectives have been twofold: a) to analyze human and artificial voices narrating an advertising story, where prosody is modified, by measuring effectiveness, listener's attention, concentration, and recall; and b) to study the evaluation of these voices depending on their functions in these devices.

6.1. Discussion of voice analysis

Regarding the first goal, the results showed that human voices were better processed than artificial voices in terms of effectiveness, attention, concentration and recall. First, human voices were deemed as being more effective. They were rated as clearer, more natural, more pleasant, more dynamic and more persuasive, in line with the prior assessment the respondents made in the questionnaire. Regarding the artificial voices, the synthetic ones were considered as more effective than human manipulated voices. Nevertheless, both received a lower score compared to human voices. This negative perception of artificial voices may be why less attention was given to them. This is in spite of the fact that their artificial characteristics could have drawn attention to them in the first place. Even so, this was not the case, and the explanation can lie in the expressive prosodic effect and the value of expressiveness (Hennig & Chellali, 2012; Pittam, 1994). As human voices are perceived in a more positive light and have shown to be more attractive (Terken & Lemeer, 1988), they make the message easier to encode by refocusing attention on the content; in other words, the voice, which is effectively produced through prosody expressiveness, causes the listener to focus on the story (Rodero, 2007). Moreover, as it is more comprehensible (in line with the results found by Sanderman & Collier, 1997), the level of concentration the listener needs to give is lower. Consequently, sufficient cognitive resources are allocated to correctly encode the message. In contrast, when content is less understood, concentration, effort, increases, as is the case with the story delivered by artificial voices, according to previous studies (Wolters et al., 2014). In these variables, there were no significant differences between the artificial voices. Here both artificial voices caught less attention and greater concentration than human ones. Therefore, the participants required greater cognitive effort to process them, in line with the results obtained by some authors (Roring et al., 2007; Taake, 2009; Winters & Pisoni, 2004). Lastly, this improved encoding of the story narrated with human voices is also backed up by the greater level of recall obtained. When the story is told effectively, thanks to the expressive conveyed by prosody, greater attention is paid, and less effort is required to achieve greater recall. This is the opposite result achieved in other studies about recall (Nye et al., 1975; Pisoni & Hunnicutt, 1980; Pisoni et al., 1987; Delogue, Conte & Sementina, 1998; Taake, 2009). The explanation for these results, as suggested by authors, such as Winters and Pisoni (2004), is that due to decreased intelligibility of artificial voices, listeners devote more time to understanding artificial voices and more resources to processing them. As a consequence, recall is improved. However, this does not seem the explanation for our results. Intelligibility in our tested artificial voices was optimal. Thus, we can suggest that the prosody could cause this result, according to the results found by Paris et al. (2000). The self-perception of attention was greater for human voices, and the self-perception of concentration was higher for artificial voices. We can conclude that the participants were willing to pay more attention to human voices because this task required less effort. On the contrary, attention was hindered for artificial voices because this task required much effort for the participants. Therefore, the processing suffered as demonstrated by a lower level of recall. However, in these variables, there were significant differences between artificial voices. The message was better recalled with synthetic voices, according to the recall results obtained by Wolters et al. (2014). Thus, regarding recall, as effectiveness, the results suggest that the type of artificial voice matters.

We also can suggest that the type of message – an expressive story – could have influenced the results. Expressiveness was important in this story, and the lack of this characteristic in artificial voices' narration negatively affected the perception of these non-human voices. This interpretation may agree with Terken and Lemeer's study (1988). With an acceptable level of intelligibility, prosody becomes more influential, especially in an expressive text.

Regarding the second goal, the results showed that the participants preferred human voices to carry out all the analyzed functions. The analyzed functions were rated in order of complexity: to use as an audio-guide; to request information; to go to a destination; to buy a product; to carry out bank transactions. From this, we can conclude that the more complex a function is, the more a human voice will be preferred over an artificial one, especially when it is a human manipulated voice. The lowest rating was obtained by those functions in which money was involved (purchasing a product or carrying out a bank transaction) compared to merely informative functions (requesting information or using an audio guide). This data suggests that the greater the risk the user assumes in the interaction, the stronger the preference for human over artificial voices. This could be an explanation for the results obtained by Shank (2013). This research revealed that customers perceived a company as being more responsible when interaction involved humans rather than computers. Indeed, the data compiled by Fleming and Asplund (2007) suggests that humans provide far better service in business management compared to technology. In these two functions, synthetic voices obtain significant differences compared to human manipulated voices. Therefore, participants prefer a synthetic voice more than a human manipulated one when money was involved. Once again naturalness could be an important factor.

6.2. Discussion of voice familiarity

Concerning the questionnaires, contrary to what may seem *a priori*, the young university students in the sample did not seem to be very familiar with artificial voices. This in itself is an initially surprising conclusion because they are within the age range of those who have the broadest access to technology, and, moreover, their capacity as university students gives them greater exposure through various devices that would use artificial voices. Even so, they stated they were unfamiliar with them, and this could account for why they had an unclear opinion regarding the possible benefits afforded by them. The majority of those surveyed gave them an average rating, while a negative rating was the second largest group in the results. A lower degree of familiarity may have led the sample to gain a negative perception of the voices. However, the sample did reach a consensus in suggesting that artificial voices should be assigned to the technology sector (devices and machines) with the main function of issuing simple, short and rational, informative messages. It is important to highlight the fact that there was a constant reference to rationality and the informational concept, contrary to emotional communication. Under no circumstances did these young university students consider that the artificial voices could be effective in conveying emotional communication. This ties in with the classifications lent to artificial voices – insecurity, artificiality, distance, coldness, mistrust and lack of credibility. This in turn affects one of the parameters for which artificial voices were gauged – naturalness – although intelligibility also may apply since some students stated that the degree of understanding for artificial voices was lower. Therefore, the first problem is identified. The lack of intelligibility and naturalness in artificial voices leads to a lack of trust, giving rise to a negative rating. As a result, the expressive prosodic effect was decisive and, accordingly, further improvements must be made in voice synthesis at a prosodic level to increase the level of intelligibility and naturalness (Kamm, Walker, & Rabiner, 1997; Nass & Min Lee, 2001). The findings also showed that the same characteristics as those expected of human voices were demanded of the artificial voice. People evaluate a synthetic voice

interaction as they do a human interaction, upholding the theory set out in the "Media Equation" (Brave, Nass, & Hutchinson, 2005; Reeves & Nass, 1996; Rosenthal-von der Pütten et al., 2013; Van Wissen, Gal, Kamphorst, & Dignum, 2012).

The findings also showed gender differences. The majority of the participants preferred a male voice, and a crossover effect arose in which women gave a better rating to the male voice and men gave it the lowest rating. These differences can be explained due to the predominance of male voices and the persistence of a gender vocal stereotype in advertising (see Rodero et al., 2013).

In short, the primary consequence of this analysis is that the effect of expressive prosody, as a distinguishing factor linked to the types of voice, is a decisive factor in the evaluation of human-computer interaction. Consequently, research on voice synthesis should be carried out in greater depth and studies relating to prosody, and emotions should be improved to lend a greater human feel to today's synthetic voices to afford greater consistency (Isbister & Nass, 2000). Such an improvement would provide us with improved areas of application in the spheres of corporate communication, advertising, education and health.

Also, this study could be extended using more objective measures to index attention, such as heart rate (ECG) or electrodermal activity (EDA). This study only included the self-perception of attention. The decision was deliberated and intended to know what participants consciously thought about their attention. Despite it, the measure of attention in this study can be considered incomplete, and future studies should reinforce this measure. Another limitation could be the use of only one text. Future studies using this methodology should include more texts and more varied to validate the results of the research. Finally, expressiveness is a concept that should be thoroughly explored. The prosody parameters used by human voices were very expressive according to a narrative story. Future studies can analyze different prosody styles, with different levels of expressiveness, to compare the results to synthetic voices.

7. Conclusions

This study analyzed how artificial voices (synthetic and human manipulated), compared to human voices, are processed in a narrative story where prosody is modified. The dependent variables were effectiveness, attention, concentration and recall. The results showed that human voices were rated as more effective and achieved a better level of attention and recall with less concentration. Regarding the type of artificial voices, synthetic voices obtained a better result than human manipulated ones in effectiveness and recall. Human voices, more expressive due to prosody modifications, were better processed than artificial voices when telling an advertising story. We can conclude that the human voice is better perceived and processed in an advertising story because of prosody's important role.

The second goal of this research was to evaluate the functions of these voices for different applications. In this aspect, the more important and complex a function was then the more a human voice was preferred over an artificial one. The lowest rating was obtained by those functions in which money was involved (purchasing a product or carrying out a bank transaction) compared to merely informative functions (requesting information or using an audio guide). In functions where money was involved, synthetic voices obtained a better result than human manipulated ones. Therefore, the type of artificial voice matters for effectiveness, recall, and for money transactions.

Appendix 1

English version: If there's one thing I remember about my father, it's the time we spent ice fishing in Ontario. Every weekend we'd get up at 5 in the morning and head out with a heavy ice auger. So, when I had a son, I wasn't real fired up on the idea of ice fishing. I'll take him to a Snowboard Store, and I'll get him a bindings package for only 200 euros.

Spanish version: Si hay algo especial que recuerdo sobre mi padre, son aquellos momentos que pasábamos pescando en el hielo en Ontario. Cada fin de semana nos levantábamos a las 5 de la mañana y cargábamos con una hélice muy pesada para romper el hielo. Así que, cuando tenga un hijo, no lo animaré con la idea de pescar en el hielo. Lo llevaré an una tienda de Snowboard y le compraré todo el equipo por sólo 200 euros.

Appendix 2

Descriptive statistics of the effectiveness scale

	Type of voice	Mean	Deviation
Clear	Human non-manipulated	6.44	0.59
	Synthetic	3.80	1.42
	Human manipulated	2.00	0.87
Correct	Human non-manipulated	6.06	0.95
	Synthetic	3.82	1.40
	Human manipulated	3.69	1.44
Pleasant	Human non-manipulated	6.14	0.96
	Synthetic	3.20	1.24
	Human manipulated	3.10	1.22
Credible	Human non-manipulated	5.80	0.82
	Synthetic	2.96	1.27
	Human manipulated	2.65	1.24
Persuasive	Human non-manipulated	5.72	0.80
	Synthetic	2.82	1.30
	Human manipulated	2.37	1.11
Comprehensible	Human non-manipulated	6.38	0.89
	Synthetic	3.76	1.25
	Human manipulated	3.49	1.43

References

Boersma, P., & Weenink, D. (2017). *Praat: Doing phonetics by computer [software]*. Retrieved from: http://www.praat.org.

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-computer Studies, 62*(2), 161–178.

Cabral, J., Oliveira, L., Raimundo, G., & Paiva, A. (2006). What voice do we expect from a synthetic character? *Proceedings of SPECOM*, 536–539.

Chen, F. (2006). *Designing human interface in speech technology*. Sweden: Springer Science & Business Media.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Crowell, C. R., Scheutz, M., Schermerhorn, P., & Villano, M. (2009). Gendered voice and robot Entities: Perceptions and reactions of male and female subjects. In *IROS'09 proceedings of IEEE/RSJ international conference on intelligent robots and systems, NJ, USA* (pp. 3735–3741).

Delogu, C., Conte, S., & Sementina, C. (1998). Cognitive factors in the evaluation of synthetic speech. *Speech Communication, 24*(2), 153–168.

Duffy, S. A., & Pisoni, D. B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech, 35*(4), 351–389.

Fleming, J. H., & Asplund, J. (2007). *Human sigma: Managing the employee-customer encounter*. New York, NY: Gallup Press.

Gélinas-Chebat, C., Chebat, J. C., & Vaninsky, A. (1996). Voice and advertising: Effects of intonation and intensity of voice on source credibility, attitudes toward the advertised service and the intent to buy. *Perceptual and Motor Skills, 83*(1), 243–262.

Giles, H. (1973). Communicative effectiveness as a function of accented speech. *Speech Monograph, 40*, 330–331.

Gong, L., & Lai, J. (2003). To mix or not to mix synthetic speech and human speech? Contrasting impact on judge-rated task performance versus self-rated

performance and attitudinal responses. *International Journal of Speech Technology, 6*, 123–131.
Gorenflo, D. W., & Gorenflo, C. W. (1997). Effects of synthetic speech, gender, and perceived similarity on attitudes toward the augmented communicator. *Augmentative and Alternative Communication, 13*, 87–91.
Grice, M., Vagges, K., & Hirst, D. (1991). Assessment of intonation in text-to-speech synthesis systems - a pilot test in English and Italian. Genova. 24-26 September 1991 *Proceeding of Eurospeech-91, 2*, 879–882.
Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge, UK: Cambridge University Press.
Hennig, S., & Chellali, R. (2012). Expressive synthetic Voices: Considerations for human robot interaction. In *IEEE RO-MAN: The 21st IEEE international symposium on robot and human interactive communication. September 9-13, 2012. Paris, France*.
Hinterleitner, F., Norrenbrock, C., Moller, S., & Heute, U. (2012). *What makes this voice sound so bad? A multidimensional analysis of state-of-the-art text-to-speech systems*. In Spoken Language Technology Workshop (SLT), 2012 IEEE (IEEE).
Hinterleitner, F., Norrenbrock, C., Möller, S., & Heute, U. (2014). Text-to-speech synthesis. In S. Möller, & A. Raake (Eds.), *Quality of experience. Advanced concepts, applications and methods*. Berlin: Springer.
Hirschberg, J. (2005). Pragmatics and intonation. In L. Horn, & G. Ward (Eds.), *The handbook of pragmatics* (pp. 515–537). Wiley-Blackwell.
Hirst, D. (2007). A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. In *Proceedings of the XVIth international conference of phonetic sciences, saarbrücken* (pp. 1233–1236).
Isbister, K., & Nass, C. (2000). Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-computer Studies, 53*, 251–257.
Jenkins, J. J., & Franklin, L. D. (1982). Recall of passages of synthetic speech. *Bulletin of the Psychonomic Society, 20*(4), 203–206.
Kamm, C., Walker, M., & Rabiner, L. (1997). The role of speech processing in human-computer intelligent communication. In *NSF Workshop on human-centered systems: Information, interactivity, and intelligence, Arlington, VA*. Available at: www.ifp.uiuc.edu/nsfhcs/talks/rabiner.html.
Lai, J., Cheng, K., Green, P., & Tsimhoni, O. (2001). On the road and on the Web?: comprehension of synthetic and human speech while driving. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 206–212).
Lai, J., Wood, D., & Considine, M. (2000). The effect of task conditions on the comprehensibility of synthetic speech. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 321–328).
Levi, S. V., & Pisoni, D. B. (2007). Indexical and linguistic channels in speech perception: Some effects of voiceovers on advertising outcomes. In T. M. Lowrey (Ed.), *Psycholinguistic phenomena in marketing communications*. Mahwah, NJ: Lawrence Erlbaum Associates.
Luce, P. A. (1981). *Comprehension of fluent synthetic speech produced by rule. In research on speech perception progress report No. 7*. Bloomington, IN: Speech Research Laboratory, Indiana University.
Luce, P. A., Feustel, T. C., & Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and. Natural speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 25*(1), 17–32.
Marics, M. A., & Williges, B. H. (1988). The intelligibility of synthesized speech in data inquiry systems. *Human Factors, 30*(6), 719–732.
Mayo, C., Clark, R. A., & King, S. (2011). Listeners' weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis. *Speech Communication, 53*(3), 311–326.
Mayor, O., Bonada, J., & Janer, J. (2009). Kaleivoicecope: Voice transformation from interactive installations to video-games, 2009 February. In *AES 35th international conference, london, UK* (pp. 11–13).
Mirenda, P., Eicher, D., & Beukelman, D. (1989). Synthetic and natural speech preferences of male and female listeners in four age groups. *Journal of Speech and Hearing Research, 32*, 175–183.
Moody, T., & Joost, M. (1986). Synthesized speech, digitized speech, and recorded speech: A comparison of listener comprehension rates. In *Proceedings of the voice input/output society. Alexandria, VA*.
Mullennix, J. W., Stern, S., Wilson, S. J., & Dyson, C. (2003). Social perception of male and female computer synthesized speech. *Computers in Human Behavior, 19*, 407–424.
Nass, C., & Brave, S. (2005). *Wired for speech*. Cambridge, MA: MIT Press.
Nass, C., & Min Lee, K. (2001). Does computer-synthesized speech manifest Personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied, 7*(3), 171–181.
Niebuhr, O., Voße, J., & Brem, A. (2016). What makes a charismatic speaker? A computer-based acoustic-prosodic analysis of Steve Jobs tone of voice. *Computers in Human Behavior, 64*, 366–382.
Nooteboom, S. (1997). The prosody of speech: Melody and rhythm. In *The handbook of phonetic science* (pp. 640–673). Oxford: Blackwell.
Nusbaum, H. C., Francis, A. L., & Henly, A. S. (1995). Measuring the naturalness of synthetic speech. *International Journal of Speech Technology, 1*, 7–19.
Nusbaum, H. C., Schwab, E. C., & Pisoni, D. B. (1984). *Subjective evaluation of synthetic speech: Measuring preference, naturalness and acceptability. In research on speech perception progress report No. 10*. Bloomington, IN: Speech Research Laboratory, Indiana University.
Nye, P. W., Ingemann, F., & Donald, L. (1975). *Synthetic speech comprehension: a comparison of listener performances with and preferences among different speech forms*. Status Report on Speech Perception SR-41. Haskins Laboratories.
Paris, C. R., Gilson, R. D., Thomas, M. H., & Silver, N. C. (1995). Effect of synthetic voice intelligibility on speech comprehension. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(2), 335–340.
Paris, C. R., Thomas, M. H., Gilson, R. D., & Kincaid, J. P. (2000). Linguistic cues and memory for synthetic and natural speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 42*(3), 421–431.
Pauletto, S., Balentine, B., Pidcock, C., Jones, K., Botacci, L., Aretoulaki, M., et al. (2013). Exploring expressivity and emotion with artificial voice and speech technologies. *Logopedics, Phoniatrics, Vocology, 38*(3), 115–125.
Pisoni, D. B. (1997). Perception of synthetic speech. In *Progress in speech synthesis* (pp. 541–560). New York: Springer.
Pisoni, D. B., & Hunnicutt, S. (1980). Perceptual evaluation of MITalk: The MIT unrestricted text-tospeech system. In *IEEE international conference on acoustics, speech and signal processing* (pp. 572–575). New York: IEEE.
Pisoni, D. B., Manous, L. M., & Dedina, M. J. (1987). Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility. *Computer Speech & Language, 2*(3), 303–320.
Pittam, J. (1994). *Voice in social interaction: An interdisciplinary approach*. Thousand Oaks, CA: Sage.
Potter, R. F., & Choi, J. (2006). The effects of auditory structural complexity on attitudes, attention, arousal, and memory. *Media Psychology, 8*, 395–419.
Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge: Cambridge University Press.
Rodero, E. (2006). *Analysis of intonation in news presentation on television*. In ExLing-2006, Atenas.
Rodero, E. (2007). Characterization of a presentation on audiovisual media. *Estudios del mensaje periodístico, 13*, 523–542.
Rodero, E. (2015a). The principle of distinctive and contrastive coherence of prosody in radio News: An analysis of perception and recognition. *Journal of Nonverbal Behavior, 39*, 79–92.
Rodero, E. (2015b). Influence of speech rate and information density on Recognition: The moderate dynamic mechanism. *Media Psychology, 19*(2), 224–242.
Rodero, E., Larrea, O., & Vázquez, M. (2013). Male and female voices in commercials. Analysis of effectiveness, adequacy for product, attention and recall. *Sex Roles, 68*(5), 349–362.
Rodero, E., Mas, L., & Blanco, M. (2017). The influence of prosody on politician's credibility. *Journal of Applied Linguistics and Professional Practice, 11*(1).
Roring, R. W., Hines, F. G., & Charness, N. (2007). Age difference in identifying words in synthetic speech. *Human Factors, 49*, 25–31.
Rosenberg, A., & Hirschberg, J. (2009). Charisma perception from text and speech. *Speech Communication, 51*, 640–655.
Rosenthal, A. M., Krämer, N. C., Hoffman, L., Sobieraj, S., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics, 5*, 17–34.
Salza, P. L., Di Fabbrizio, G., Oreglia, M., Falcone, M., Sementina, C., & Delogu, C. (1993). Development of a context dependent methodology for text-to-speech synthesis evaluation in interactive dialogue systems. In *Esprit project 6819 (SAM-A), speech technology assessment in multilingual applications, SAM-a periodic progress report year 1*. April 1993-30-September 1993.
Sanderman, A. A., & Collier, R. (1997). Prosodic phrasing and comprehension. *Language and Speech, 40*, 391–409.
Schötz, S. (2006). Perception, analysis and synthesis of speaker age. *47*. Linguistics and Phonetics).
Shank, D. B. (2013). Are computers good or bad for business? How mediated customer computer interaction alters emotions, impressions, and patronage toward organizations. *Computers in Human Behavior, 29*, 715–725.
Signorello, R., D'Errico, F., Poggi, I., Demolin, D., & Mairano, P. (2012). Charisma perception in political speech: A case study. In *International conference on speech and corpora* (pp. 343–348). Brazil: Belo Horizonte.
Slowiaczek, L. M., & Nusbaum, H. C. (1985). Effects of speech rate and pitch contour on the perception of synthetic speech. *Human Factors, 27*(6), 701–712.
Sreenivasa Rao, K. (2012). *Predicting prosody from text for text-to-speech synthesis*. New York: Springer.
Stern, S. E., Chobany, C. M., Patel, D. V., & Tressler, J. J. (2014). Listeners' preference for computer-synthesized speech over natural speech of people with disabilities. *Rehabilitation Psychology, 59*(3), 289.
Syrdal, A. K., Bennett, R. W., & Greenspan, S. L. (1994). *Applied speech technology*. CRC press).
Taake, K. P. (2009). *A comparison of natural and synthetic speech: With and without simultaneous reading*. Thesis. Washington University.
Terken, J., & Lemeer, G. (1988). Effects of segmental quality and intonation on quality judgments for texts and utterances. *Journal of Phonetics, 16*, 453–457.
Vainio, M., JaÅNrvikivi, J., & Werner, S. (2002). Effect of prosodic naturalness on segmental acceptability in synthetic speech. In *Proceedings IEEE 2002 workshop on speech synthesis, santa monica, California*.
Van Wissen, A., Gal, Y., Kamphorst, B., & Dignum, V. (2012). Human-Agent team formation in dynamic environments. *Computers in Human Behavior, 28*(1), 23–33.
Winters, S. J., & Pisoni, D. B. (2004). Perception and comprehension of synthetic speech. *Progress Report Research on Spoken Language Processing, 26*.
Wolters, M. K., Johnson, C., Campbell, P. E., DePlacido, C. G., & McKinstry, B. (2014). Can older people remember medication reminders presented using synthetic speech? *Journal of the American Medical Informatics Association, 22*(1), 35–42. amiajnl-2014.