Article

Synthetic versus human voices in audiobooks: The human emotional intimacy effect

new media & society
2023, Vol. 25(7) 1746–1764
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14614448211024142
journals.sagepub.com/home/nms
SAGE

Emma Rodero
Media Psychology Lab, Department of Communication, Pompeu Fabra University (UPF), Spain

Ignacio Lucas
Psychiatry and Mental Health Group, Neuroscience Program, Institut d'Investigació Biomèdica de Bellvitge - IDIBELL, Spain

Abstract

Human voices narrate most audiobooks, but the fast development of speech synthesis technology has enabled the possibility of using artificial voices. This raises the question of whether the listeners' cognitive processing is the same when listening to a synthetic or a human voice telling a story. This research aims to compare the listeners' perception, creation of mental images, narrative engagement, physiological response, and recognition of information when listening to stories conveyed by human and synthetic voices. The results showed that listeners enjoyed stories narrated by a human voice more than a synthetic one. Also, they created more mental images, were more engaged, paid more attention, had a more positive emotional response, and remembered more information. Speech synthesis has experienced considerable progress. However, there are still significant differences versus human voices, so that using them to narrate long stories, such as audiobooks do, is difficult.

Keywords

Audiobooks, cognitive processing, human voice, perception, physiological response, synthetic voice, talking books

Corresponding author:
Emma Rodero, Department of Communication, Pompeu Fabra University (UPF), Roc Boronat, 138, 08108 Barcelona, Spain.
Email: emma.rodero@upf.edu

Introduction

We are living in the golden age of audio. As many people are overwhelmingly busy, they have changed their consumption habits and started listening to podcasts or audiobooks while doing other activities. If you do not have time to read, you can listen to an audiobook while cooking or driving. More and more people listen to audiobooks or audio stories, especially by using earphones (Have and Pedersen, 2020; Tattersall Wallin and Nolin, 2020). According to an Edison Research survey (2019), the average number of audiobooks listened to per year has increased from 6.8 in 2019 to 8.1 in 2020. Most audiobooks are listened to through mobile phones using earbuds. However, new media forms, like smart speakers, become social actors that encourage these new habits (Humphry and Chesher, 2020). Smart speakers allow people to do everyday tasks as listening by using a voice command. The most popular voice-activated devices are Amazon Echo, Google Home, and Apple HomePod (Edison Research, 2020). According to an Edison Research study (2020), 24% of Americans own a smart speaker. In 2020, these owners used these devices more often (43%) than in 2019. In this context, many people have started to consume audio in smart speakers (23%), almost in the same proportion as radio (25%). Most of the owners use these devices for playing music (85%), but regarding audio also for listening to the radio (45%) and for reading short stories (17%). In a recent survey, the Audio Publishers Association's (APA) annual sales survey (2020) showed that 46% of smart speaker owners had used it to listen to an audiobook, 31% more than in 2018.

An audiobook is a recording of a published book. The APA in 1994 established this term as the industry standard (APA, 2021). However, initially, they were named talking books. This name came from the program that the American Foundation for the Blind launched for veterans injured during World War I and blind adults. Today, books read for blind or print disabled people are known as "talking books" (Thoet, 2017). These books are mainly listened to by using a device that facilitates navigation throughout the book so that blind people can listen to a synthetic voice reading, for example, the chapter, the page, or the footnotes. The text of the book is generally narrated by a human voice, although there also are synthetic-narrated titles. Another possibility is to have the text of the book that is read by a text-to-speech system. Therefore, artificial voices have been used in talking books for some years, and blind people and individuals with reading impairments could be more used to listening to them. However, they are not common in commercial audiobooks for the general population. There is a growing debate about this subject, as speech synthesis technology has improved substantially, and artificial voices do not sound robotic anymore. For audiobook companies, using artificial voices can be more feasible, and listeners also can choose a text-to-speech technology to listen to them; Amazon Alexa offers this option. For audiobook producers, this is an easier and cheaper solution than paying a human narrator. However, the question that arises is whether the experience is the same when listening to a synthetic voice telling a story, compared to a human voice. Listening to an artificial voice in brief messages to navigate throughout the book (which we can understand as a functional use) is one thing, but to listen to the entire narration with this kind of voice (which we can conceive as a creative or artistic use) is quite another. As far as we know, no studies have analyzed the cognitive processing of human versus synthetic voices when narrating stories measuring the participants' physiological response. Therefore, this research aims to compare the listeners' perception, physiological response, and recognition of information when listening to stories conveyed by human and synthetic voices (through Alexa). Although Alexa can play audiobooks narrated by human or artificial voices, in this study, we tested the synthetic voice with this smart speaker, as this device can easily read any book with a voice command, and people are used to listening to this voice. We analyzed the stories' likability and enjoyment, creation of mental images (imagery), narrative engagement, attention levels (via heart rate), physiological arousal (via electrodermal activity), emotional valence (via facial detection), and recognition of information (via a memory test).

This study makes two main contributions: (1) to determine how human versus artificial voices are processed when narrating a story, showing that a "human emotional intimacy effect" is produced and (2) to demonstrate how human versus artificial voices affect the individuals' physiological response.

Theoretical background

Perception of synthetic voices

Some studies have demonstrated differences in perception between synthetic and human voices, mainly due to the sound's low quality (Chen, 2006; Syrdal et al., 1994; Winters and Pisoni, 2004). The human voice's superiority has been called the "voice effect" or "voice principle" (Mayer, 2014). According to this effect, learning is more effective in terms of retention and transfer outcomes when a human voice delivers the content compared to a machine voice. Therefore, human voices are considered more effective than synthesized for teaching.

There is no doubt that technology plays a critical role in the perception of these voices. The main elements that characterize synthetic voices' quality are intelligibility and naturalness (Paris et al., 2000). As some studies have demonstrated (Craig and Schroeder, 2017, 2019) perception improves when synthetic voices have better sound quality. Current technology has considerably improved in the past years, and synthetic voices now sound more intelligible and natural (Wolters et al., 2014). Some companies are now using Artificial Intelligence (A.I.) to improve intelligibility. This advance has significantly enhanced synthetic voices to the point that it is hard to differentiate if they are human or artificial. However, speech naturalness, conveyed through prosody, is still a challenge, especially when the speech's content is a creative format like a story. Listening in smart speakers to a synthetic voice delivering instructions or brief information, like the weather, can be tolerable, as some studies have shown (Rodero, 2017), but listening to a synthetic voice in an audiobook for hours may not be an entirely satisfactory experience. Synthetic voices are not yet as expressive as human voices, and this poor performance becomes a relevant problem when the stimulus is artistic or aesthetic as in a long story, usually listened to while listeners are moving (Tattersall Wallin and Nolin, 2020). Therefore, it is an activity that demands prosody variations to avoid a drop in listeners' attention.

Moreover, human voices possess a specific quality. They are embodied voices. The speaker's body is completely involved in generating it (Barker, 2015). This means that when we hear it, we listen to the sound and the body that produces it, the grain of the voice in Barthes's (1985) words. Some studies have shown that listeners have a remarkable ability to recognize characteristics, including the speaker's age, gender, height, and weight (Assmann et al., 2006; Mulac and Giles, 1996; van Dommelen and Moxness, 1995). "When we listen to a voice, we draw a specific physical portrait of its owner in our minds, based on its vocal features." (Rodero, 2020: 18). This presence of the body is perceived in the vocal features, especially in the voice tone, and in the non-articulated sounds such as respirations, vocalizations, or fillers (e.g. umm). These voice sounds (e.g. saliva, tongue clicking, lips popping) inherent to a human voice are amplified by microphones, making them expressive (Connor, 2000). Therefore, when we listen to a human voice, this voice calls for a body (Chion, 1994), and this characteristic makes each voice unique, "an expression of one distinct embodied individual to another" (Cavarero, 2005: 197–210). "The voice is thus always written by the body" (Di Matteo, 2015). However, we do not have these human references in synthetic voices. From this point of view, technology denatures artificial voices by decoupling them from the body. This lack of naturalness can be why synthetic voices are still rejected (Parker, 2013).

This weaker position of synthetic voices can also be the reason why they usually are perceived as poor. Stern et al. (1999) found that human voices, both male and female, were perceived as more favorable and less negative than computer-synthesized voices. Mullennix et al. (2003) showed that female synthetic voices were perceived as less pleasing, less convincing, and less truthful, squeakier, more accented, and nasal than female human voices. Also, human voices are more trusted than artificial voices and, therefore, robots with more human features influence user's social responses to a greater extent (Xu, 2019). Recently, Rodero (2017) concluded that human voices were better perceived than synthetic voices in advertising messages. Specifically, human voices were assessed as more effective, more appropriate, pleasant, credible, persuasive, understandable, and clearer than synthetic voices. They were also considered more suitable for emotional communication, which is especially important to tell a story. If the perception of these human voices is more positive, as these studies have shown, we may also deduce that the experience of listening to a story will be better when delivered by a human voice and, therefore, that listeners will enjoy stories narrated by human voices more. These considerations lead to the first hypothesis of this study:

> *H1*. Participants will enjoy audio stories told by a human voice more than audio stories delivered by a synthetic voice.

Voice perception can also affect other factors that, in turn, may condition listeners' cognitive processing. Regarding storytelling, two of these relevant factors to process a story are imagery and engagement.

Mental imagery is the experience in which listeners create mental images in their minds in the absence of an object or stimulus (Bone and Ellen, 1992; Kosslyn et al., 2001; Kosslyn, 1994). Thus, this experience implies the encoding, processing, and evocation in memory of the stimulus by creating a semi-perceptual representation (Babin and Burns, 1998). Some studies have suggested that creating mental images in people's minds improves cognitive processing, as the brain is actively working to draw these images (Paivio, 1991). Regarding audio and advertising, some studies have concluded that high-imagery commercials are more effective, as they improve recall of information (Bolls and Lang, 2003; Goosens, 1994; Miller and Marks, 1997). In audio stories, Rodero (2012) has shown that some audio features can stimulate the creation of mental images, although, in this study, the verbal stories attained the lowest imagery and attention levels compared with the use of sound effects or sound shots. From this idea, we can derive that if there is deeper cognitive processing elaborating these mental images, which enhances memory, individuals will be more engaged with the story and the characters (Mar and Oatley, 2008). Along with the type of narration, mental imagery's potential to increase engagement is supported by some studies (Green and Brock, 2000; Green et al., 2004). There is a strong relationship between engagement and imagery. When individuals are enjoying the narrative, feel connected with the characters, are immersed in the story, and engaged, the number of mental images created in their minds is higher (Bolls and Muehling, 2007; Busselle and Bilandzic, 2008). There are no studies measuring imagery and narrative engagement of synthetic voices to the best of our knowledge. However, based on previous research and bearing in mind that some studies have shown that artificial voices are not as well perceived and comprehended, we posit the next hypothesis:

> *H2.* The audio stories narrated by a human voice will achieve a higher level of imagery and narrative engagement than the stories with a synthetic voice.

If human voices are better perceived, listeners create more mental images and feel more connected to the story with a strong narrative engagement is easy to infer that the attention levels to the story, the emotional response, and the recall of information also will be higher than listening to a story narrated with synthetic voices. We analyze these aspects in the next section.

Psychological response to synthetic voices

The lack of naturalness of synthetic voices compared to human ones can also negatively affect the listeners' physiological response. Some studies have shown that the encoding process of the speech delivered by synthetic voices is more demanding and requires more time than human voices (Roring et al., 2007). Consequently, this difficulty means that listeners may need to allocate more cognitive resources to process the message correctly, and this effort could imply a higher cognitive load (Luce et al., 1983; Taake, 2009). Winters and Pisoni (2004) found that participants took longer to respond to a synthetic voice than a human voice, requiring more effort from the listener. However, the recent improvement in synthetic speech might have reduced the energy charge to process artificial voices. Craig and Schroeder, (2017, 2019) showed that the cognitive load of human and synthetic voices was equivalent when modern technology was used. However, in these experiments, the authors used a virtual human, adding human appearance and gestures, not only a voice sound. As artificial voices do not sound as expressive as human ones, we can deduce that listening to an audiobook for a long time narrated by a synthetic voice may produce a higher cognitive effort and provoke an overload and a drop of attention (Delogu et al., 1998). Considering these previous studies, we expected that listeners pay less attention to the stories conveyed by a synthetic voice than a human voice. This lack of attention and interest toward synthetic voices could also be reflected in a less intense emotional response, both in arousal and valence, the two main dimensions of emotion (Bradley and Lang, 1994). Arousal represents the intensity of the emotion, from calmness to excitement, while valence is the pleasure of this feeling regarding negative or positive emotion. If the perception of synthetic voices is more negative than human voices and the levels of enjoyment, narrative engagement and imagery are lower, then we can infer that the levels of arousal will be also lower and the emotion more negative. As far as we know, there are no studies that analyze synthetic voices by measuring attention using heart rate, arousal indexed by electrodermal activity, and emotional valence with facial recognition. There also are no studies of this type related to audiobooks. However, based on the literature review, we can postulate the third hypothesis of this study:

> *H3.* The audio stories narrated by a human voice will attain higher attention, physiological arousal, and more positive emotion than audio stories delivered by a synthetic voice.

Recognition of information

A critical part of cognitive processing is memory. If listeners cannot remember or comprehend the information they hear, then the communication process will have failed. Some studies have found that synthetic voices are more difficult to understand than human voices (Lai et al., 2000). Due to artificiality, listeners need more cognitive resources to process synthetic voices, and therefore, the cognitive effort is greater than when processing human voices. This artificiality affects the recall process (Wolters et al., 2014). However, this extensive use of cognitive resources could be beneficial. If more resources are allocated to process the stimulus, listeners will pay more attention, and then recall or recognition of information could obtain optimal levels, according to Luce (1981). This hypothesis seems to be proven in some studies in which there were no differences in recall between synthetic and human voices, although these studies are not recent (Pisoni and Hunnicut, 1980; Paris et al., 2000). This lack of differences was more evident when the content to recall was straightforward (Jenkins and Franklin, 1982). However, this is not the case with audiobooks. The narrative is simple, but the length and variety of actions and characters can be complicated to process when narrated by a monotonous voice. Prosody variations are decisive for memory, as some studies have shown (Rodero et al., 2017). These melodic changes are essential for the recall and comprehension of synthetic voices (Paris et al., 2000; Sanderman and Collier, 1997). If processing synthetic voices is challenging due to the lack of naturalness and becomes a prolonged activity, individuals can grow tired, become inattentive, and then abandon the task. In this case, they should need more resources to process the message. Lang et al. (2015) showed that if the level of structural features is low and the information introduced is high, then recognition accuracy is low. This could be the case with synthetic voices and audiobooks. The structural features level is low (one artificial voice), but the information introduced is high (a mechanical voice). Therefore, the result could be low recognition, according to these authors. In these circumstances, listeners can suffer a cognitive overload, as Rodero (2016) has demonstrated when listeners process fast speech. Consequently, we can hypothesize that if listeners enjoy less the stories narrated by synthetic voices, and the engagement, the creation of mental images, the attention, arousal, and valence are lower, then recognition of information should be affected. This leads us to the last hypothesis:

> *H4.* The audio stories narrated by a human voice will achieve higher recognition of information than audio stories delivered by a synthetic voice.

Method

Experimental design

The experiment was a mixed factorial design with two kinds of voices (human and synthetic), two stories, and two presentation orders. We recruited a professional female voice actor as the human voice and used Amazon's Alexa as the synthetic voice. The two stories were delivered by both the human female voice and the female synthetic voice (Alexa). The order was the only between-subjects factor. Half of the participants first listened to story one with a human voice and then story two with a synthetic voice. The other half first listened to story one with a synthetic voice and then story two with a human voice. As participants performed a recognition of information test, each subject had to listen to different stories.

All statistical analyses of scales (enjoyment, imagery, narrative engagement) and recognition of information were submitted to a two (voices) by two (stories) analysis of variance (ANOVA). The order was not significant. All statistical analyses of physiological data (arousal and attention) were submitted to a two (voices) by 3 minutes (length of the stories) repeated measures ANOVA. The methodology of this research employs a multidimensional approach analyzing self-reported data, psychophysiological measures, and memory tests (Potter and Bolls, 2012). The self-report scales measure the participants' perception, the psychophysiological data (heart rate, skin conductance, and facial recognition), attention, arousal, valence, and the recognition of information quantify the encoding process. This triple operationalization helps to be more accurate in understanding the participants' processing of human compared to synthesized voices.

Participants

Sixty participants ($n=60$) took part in this study. Participants included 35 females and 25 males, aged between 19 and 35 ($M=25$). The criteria to participate in this study were to be a habitual audiobook listener and native in the stories' language. The exclusion criteria were age below 18 years old and hearing impairments. We recruited the participants from the population through announcements in different outlets and social networks. Once we collected the volunteers (70), we removed those that did not meet the criteria. Then, using email lists, we assigned them an order of stimulus presentation. The sample size was according to the psychophysiological studies in media (Potter and Bolls, 2012). All the participants were exposed to all the conditions (human and synthetic voices).

Stimuli

The materials were composed of two audio stories in Spanish, as the experiment was conducted in Spain. A professional author wrote the stories specifically for this study to control all aspects of the content. The participants needed not to know the stories, as this factor could affect attention and especially recognition of information. Both stories contained exactly the same words and number of exclamatory and interrogative sentences. The main character in both stories was a woman. Therefore, we recruited a professional female voice actress as the human voice and used Amazon's Alexa, which uses a female voice, as the synthetic voice. For selecting the female voice, we conducted a previous casting. The duration of both stories was about 3 minutes.

Dependent variables

Enjoyment. Participants were asked two questions in two 7-point scales: "Did you like the story?" and "Did you enjoy the story?" Both scales formed the enjoyment level. The Cronbach's alpha coefficient for this scale was .76; thus, a high coefficient.

Imagery scale. We used the imagery scale by Ellen and Bone (1991) to measure the quality and quantity of mental images created in the participants' minds. The stimulation of mental images was operationalized by measuring the vividness and the quantity of the images. First, vividness is the clarity of the images that an individual creates in mind. Vividness is a scale of 7 points formed by the following items—how clear, detailed, vivid, well-defined, and lifelike the images are—with a scale for each value. The average of all the items on this scale was the vividness level. Second, quantity/ease is the number of images stimulated in the participants' minds after listening to each story. The next questions compose the 7-point scale: to what extent you generated mental images; how many images were stimulated; what degree of difficulty they experienced in generating the images in their mind; how fast they created these images; and whether they would agree that they had generated images without difficulty. The average of the scale formed the quantity/ease level. The average of the two scales (vividness and quantity) formed the imagery variable. Cronbach's alpha coefficient was .90; thus, with a high coefficient in the reliability tests.

Narrative engagement. We used the narrative engagement scale by Busselle and Bilandzic (2008). This scale has been applied to different narrative studies. The narrative engagement scale measures four dimensions—attentional focus, narrative understanding, emotional engagement, and narrative presence—asking for different aspects. Values were reversed to have higher scores suggesting a more robust feeling. Cronbach's alpha coefficient was .72.

Physiological data. The measures were attention and emotional response (arousal and valence). Attention was measured using heart rate in beats per minute (BPM). A polygraph Biopac MP-160 (Biopac Systems) registered the electrocardiogram (ECG) from which we extracted the participants' heart rate using the R-wave peaks. ECG was recorded with a sampling rate of 1000 Hz a band-pass filter of 0.5–35 Hz. Two Beckman standard Ag-AgCl electrodes (8 mm sensor diameter) were placed on the participants' chest and abdominal part with a ground electrode on the non-dominant side. Interbeat intervals were visually inspected for artifacts and premature beats. Segments containing artifacts were excluded. The heart rate's deceleration represents an increase in attention and cognitive effort (Potter and Bolls, 2012).

Arousal was measured with electrodermal activity, which was based on recording the eccrine glands' activity found mainly in hand. Two 8-mm Ag-AgCl electrodes were attached to the participants' non-dominant hand fingers with a constant voltage (.5 V). The values were registered by a polygraph Biopac MP-160 (Biopac Systems). For each participant, the signal was calibrated before the experimental session and was recorded with a sampling rate of 1000 Hz and Low pass filters (LP: 66.5 Hz, Q=0.5 and LP: 38.5 Hz, Q=1).

The facial analysis measured the emotional dimension of valence. The valence dimension refers to the positive or negative emotions of an individual exposed to a stimulus. The Facial Analysis recognition software, *Facereader* (Noldus), identified the concrete emotions that the participants felt while listening to the audio stories (neutral, happiness, sadness, anger, contempt, disgust, surprise, and fear). From these data, we extracted the general valence level.

Responses in heart rate, skin conductance, and facial expression were determined by averaging across segments of 5 seconds. The 10 seconds previous to each audio presentation were used as the baseline. The response was measured for a total of 190 seconds (38 segments) after the audio onset. Response scores were calculated by subtracting the baseline values from the response signal.

Recognition of information. Once the participants listened to each story, the recognition of information was measured using a multiple-choice test. Participants had to answer five questions by each story for which they had three possible answers. Recognition accuracy was scored as 1 for hits and 0 for errors.

Experimental procedure

Once a participant showed interest in taking part in the experiment, he or she was assigned a day and time and an experimental condition (version 1 or 2). Then, each subject went to the lab to participate in the experiment. When they arrived, the first step was to sign the Consent Statement. Once the participants signed the documents, we placed sensors attached to their bodies to get the psychophysiological response—sensors in the chest for heart rate (ECG) and electrodes in their hands for electrodermal activity (EDA). Physiological data acquisition was performed continuously during the experiment using Biopac MP-160 (Biopac Systems). A camera in front of the participants registered the facial expressions to measure valence. The participants listened to the two different

Table 1. Descriptive statistics for all the variables.

Variable	Type of voice	*M*	*SD*
Like	Human	3.47	.57
	Synthetic	1.33	.47
Enjoy	Human	3.13	.77
	Synthetic	1.60	.56
Enjoyment	Human	3.30	.48
	Synthetic	1.46	.39
Vividness	Human	3.15	.52
	Synthetic	1.50	.38
Quantity	Human	2.97	.68
	Synthetic	1.57	.53
Imagery	Human	3.06	.54
	Synthetic	1.54	.39
Focus	Human	2.87	1.40
	Synthetic	1.20	.40
Understanding	Human	2.80	1.32
	Synthetic	1.53	.97
Emotional	Human	3.10	1.51
	Synthetic	1.30	.53
Presence	Human	4.20	.84
	Synthetic	1.23	.56
Attention	Human	−1.59	.80
	Synthetic	1.85	.52
Arousal	Human	−.46	.21
	Synthetic	−.96	16
Valence	Human	.14	.15
	Synthetic	−.40	.17
Recognition	Human	3.48	1.09
	Synthetic	2.31	1.19

stories in two different orders to avoid that the participants' tiredness could affect the same story. After listening to each story, they had to rate the story using the enjoyment, imagery, narrative engagement scales, and five questions for the recognition test. The total duration of the experiment was 30 minutes. Each participant received €10 in cash for his or her participation in this study.

Results

Table 1 shows all the descriptive statistics.

H1. Enjoyment

H1 established that participants would enjoy audio stories told by a human voice more than audio stories delivered by a synthetic voice. Main effects were found for the type of voice in the variables like, enjoyment, and the general level of enjoyment. The story narrated by a human voice was more liked, and participants enjoyed more, therefore, the enjoyment level was higher than the same story with a synthetic voice. The story and the interaction between the type of voice and the story were not significant in any variable. The data support H1.

H2. Imagery and engagement

H2 posited that the audio stories narrated by a human voice will achieve a higher level of imagery and engagement than the stories with a synthetic voice. The variable imagery was composed of the vividness and quantity of the participants' mental images. Main effects were found for vividness, quantity, and imagery in the type of voice. The story narrated by a human voice created more vivid images, more mental images, and a higher level of imagery than the same story with a synthetic voice. The story and the interaction between the type of voice and the story were not significant in any variable.

Second, the variable narrative engagement was composed by the focus on the story, the understanding, the emotional response, and the narrative presence. There were significant results for the type of voice in all the variables. Participants focused more on the story narrated by a human voice, understood it better, had a stronger emotional response, and felt more present in the narrative than in the same story with a synthetic voice. The story and the interaction between the type of voice and the story were not significant in any variable. The data support H2.

H3. Physiological response

H3 posited that the audio stories narrated by a human voice will attain a higher level of attention, physiological arousal, and positive emotion compared to audio stories delivered by a synthetic voice.

For attention, measured with heart rate, there were significant results in the type of voice and the interaction between the type of voice and the time of the story, $F(1, 59) = 2.17$, $p < .001$. The human voice registered a lower heart rate than the synthetic. This means that participants paid more attention and had less of a cognitive load to human voices than to synthetic ones. Figure 1 shows the results.

There were significant results in the interaction between the type of voice and the time of the story regarding arousal, measured with electrodermal activity, $F(1, 59) = 1.75$, $p = .004$. Participants had higher levels of emotional activation throughout the story listening to human voices than to synthetic ones. Figure 2 shows the results.

In the valence variable, measured with facial recognition, there were significant results for the type of voice. The story and the interaction between the type of voice and the story were not significant. The happiness, sadness, anger, surprise and disgust emotions were significant (all p values $< .001$). Participants had a positive valence in the story narrated by a human voice and a negative one in the same story narrated by a synthetic voice. Therefore, they felt more happiness in the stories conveyed by a human voice than with the synthetic voice. Conversely, they felt more sadness, anger, surprise

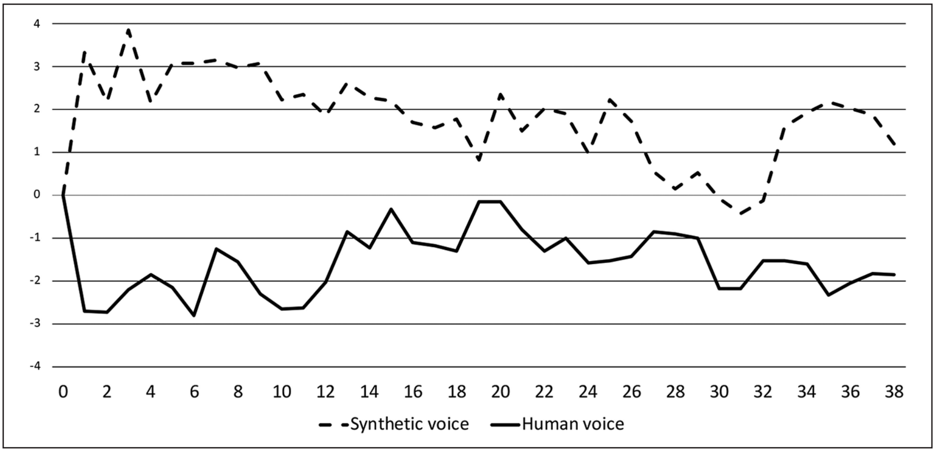

Figure 1. Heart rate.

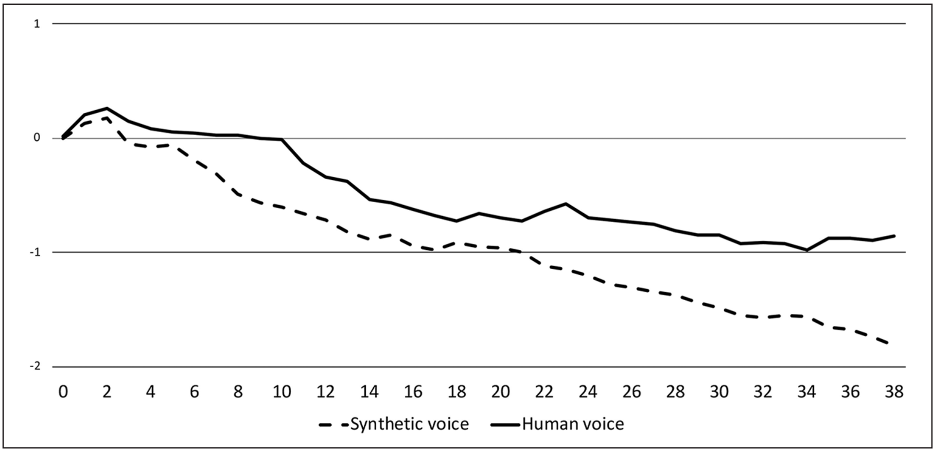

Figure 2. Skin conductance.

and disgust with the artificial voice. Stories were not significant in any case. The data support H3.

H4. Recognition of information

H4 suggested that the audio stories narrated by a human voice would achieve higher recognition of information compared to audio stories delivered by a synthetic voice. In this variable, there were significant results for the type of voice. Participants recalled more information in the story narrated by a human voice than in the same story narrated by a synthetic voice. The story and the interaction between the type of voice and the story

Table 2. Results of the analysis of variance (ANOVA).

Variable	Voice		
	F	*p*	η^2
Enjoyment			
Like	253.58	.001	.819
Enjoy	75.86	.001	.575
General enjoyment	265.83	.001	.826
Imagery			
Vividness	195.81	.001	.778
Quantity	76.70	.001	.578
Imagery	153.21	.001	.732
Narrative engagement			
Attentional focus	37.20	.001	.399
Narrative understanding	17.10	.001	.234
Emotional engagement	36.38	.001	.394
Narrative presence	246.86	.001	.815
Heart Rate	12.22	.002	.368
Electrodermal activity	1.75	.004	.121
Facial Recognition	177.61	.001	.587
Recognition of information	15.01	.001	.211

were not significant. Therefore, we can conclude that H4 is also supported. Table 2 shows the ANOVA results for all the variables.

Discussion

This study analyzed the listeners' perception, physiological response, and recognition of information when listening to audio stories, like audiobooks, conveyed by human and synthetic voices (Amazon Alexa). In particular, we analyzed the participants' enjoyment, imagery, engagement, attention, physiological arousal, emotional valence, and recognition of information.

The results of the study were conclusive. Human voices—in this case, a female voice—obtained a better result in all the variables than the artificial voice (Alexa), indicating that the deeper level of processing had a stronger emotional response. The self-perception of enjoyment, imagery, and engagement; the physiological levels of arousal, attention, and valence, and the recognition of information were significantly superior in the stories narrated by the human compared to the synthetic voice when they were exactly the same. Therefore, these results can be explained according to the "human emotional intimacy effect" that we have proposed in this study. This principle is also in line with the "voice effect" (Mayer, 2014).

First, the perception of the two stories narrated with a human voice was more positive. Participants reported that these stories were more liked, and they enjoyed more than when Alexa conveyed them. Some studies have shown that human voices are perceived as more persuasive, convincing and truthful, and squeakier than synthetic voices (Mullennix et al., 2003; Rodero, 2017; Stern et al., 1999). According to these studies, if the perception was more positive, then logically, the stories narrated by a human voice also were the best enjoyed. Therefore, these results are aligned with previous research.

Second, participants felt more engaged with the stories narrated by a human voice. If participants enjoyed the human stories more, we could reasonably conclude that the self-reported engagement was greater. The narrative understanding, presence, attentional focus, and emotional engagement were significantly higher when participants listened to the actress compared to Alexa. As the listeners felt more connected with the characters and immersed in the story, they also created more mental images as well as more vivid ones than when listening to the stories with a synthetic voice, according to the results of different studies about storytelling and audio (Bolls and Lang, 2003; Goosens, 1994; Green et al., 2004; Miller and Marks, 1997; Rodero, 2012). The data suggest that the encoding process was more effective with the human voice than Alexa (Mar and Oatley, 2008), as the psychological results showed.

Third, listeners paid more attention to the stories with the human voice than the artificial one. The data suggest that participants allocated more cognitive resources to process stories told by a human voice. Also, the emotional activation that we measured with electrodermal activity was higher with the human voice. This high arousal reinforces the results of the self-perception in which participants reported more emotional engagement. For the stories conveyed by a human voice, the arousal was greater with a stronger emotional response than listening to Alexa. Along with this, the emotional valence was more positive for the human than the synthetic voice. Participants felt more happiness with the human voice and more sadness, anger, surprise, and disgust with the synthetic voice. Therefore, they felt more negative emotions and were more surprised with Alexa. Altogether, the physiological results showed greater activation of the autonomic nervous system, stronger and more positive emotional response, and better cognitive processing with a higher attention level in the human voice case.

Fourth, this better cognitive processing was finally shown with the results of the recognition test. Participants recalled more information about the stories in the case of human voices compared to Alexa. According to previous research, the memory results of this study showed better recall and comprehension levels for human voices (Lai et al., 2000; Wolters et al., 2008). Due to the artificiality, for listeners, processing an artificial voice was difficult, and, therefore, the recognition was affected. According to the studies by Lang et al. (2015), as the level of structural features was low (one artificial voice), but when the information introduced was high (a mechanical voice), recognition accuracy was low. Prosody variations are decisive for memory, and consequently, melody changes are essential for recall and comprehension of synthetic voices, as some studies have shown (Paris et al., 2000; Rodero et al., 2017; Sanderman and Collier, 1997).

All in all, this study results reveal the power of listening to a human voice when narrating stories, causing what we call a "human emotional intimacy effect." When people listen to a human voice, they experience a closeness and connection feeling that activates a solid and positive emotional response, as shown in the physiological response. Human voices are strongly connected with our brain and emotions, as they represent our primary means of communication. Moreover, they are embodied voices, and their sound transmits a human body's presence (Barker, 2015; Di Matteo, 2015; Rodero, 2020). This body reference is perceived both in the vocal features and in non-articulated sounds, such as respirations, vocalizations, or fillers (umm) that can make the narration in an audiobook more human and expressive (Connor, 2000). Therefore, this closeness of human voices increased the participants' emotional engagement and then they created more mental images. The human and body presence and the emotional connection enhanced the attention levels, aided by an increment in expressivity. Human beings use their voices to express thoughts and feelings and to form social relationships. Furthermore, human voices are a very rich and versatile instrument. The different prosody variations that modify voice can produce a great variety of different expressions. These changes conveyed through voice, sometimes very subtle, help individuals provide the meaning of the speech and their intention and emotion. In the narrative, these prosody variations are crucial to give the tone and sense to the story's content and especially to deliver the necessary doses of expressivity to the narration and interpretation. Therefore, we can conclude that prosody and narrative are inherently connected. Narrators must use different prosody features to engage listeners, maintain attention throughout the story, and create emotional connections (Paris et al., 2000; Rodero et al., 2017; Sanderman and Collier, 1997). Finally, all these factors improved the listeners' cognitive processing, as they recalled more information about the stories with a human voice.

However, while for human voices there are no limits to combining the different features and forming of very different sounds to give all the nuances to the narration of a story, synthetic voices have limited options. Furthermore, they are without a human body, and, in consequence, they have no human references (e.g. respirations, fillers, saliva, tongue clicking, lips popping). These can be some of the reasons why, still today, artificial voices are not as expressive as human ones. The lack of expressive performance conveyed through the appropriate use of prosody variations could leave our participants with a less positive perception, engagement, and creation of mental images. This problem could also hinder their cognitive processing, with less attention, autonomic arousal, positive emotion, and recognition of information. Ultimately, the "human emotional intimacy effect" connected to a human body was not produced, and there was no affective connection or closeness. These could be some of the reasons for explaining these results.

This study's results can be of great interest to the audiobook industry and for all the sound industry in general, particularly this devoted to fiction formats, such as fiction podcasts or radio/audio dramas, for example. The findings clearly can help the audiobooks and talking books companies to understand the relevance of the human voice in creative works as a book narration is. Artificial voices can help develop functional tasks, like delivering brief messages or instructions (Rodero, 2017). This functional practice can make life easier for the general population by giving brief information, such as the traffic or the forecast, and making accessible formats for the elderly, people who are blind or have reading impairments, for instance, navigating in talking books. However, narrating a story or a book is a creative process that requires high doses of expressivity to engage listeners. Professional actors are specifically trained to interpret the narration with enough prosody variations to elicit a stronger emotional response in the listeners compared to synthetic voices, as this study has demonstrated. Therefore, if the audiobooks or talking books companies' goal is to create an emotional impact, make this experience enjoyable, and improve people's attention and memory, the recommendation is to narrate the stories with human professional actors. The same suggestion can be applied to all the audio formats. Although synthetic voices do not usually narrate audio fiction podcasts or radio dramas, the technology progress can open this possibility in the future.

This research only can be interpreted in the context of a laboratory experiment. The first limitation of this study is the use of Alexa to narrate the stories. Future research should include other synthetic voices and, therefore, a variety of different technologies to test if there are differences. Related to this aspect, another consideration is that the listeners' experience with synthetic voices cannot be considered universal and objective, but this relationship should be interpreted culturally and historically. The study is limited to the Western world as we have employed the standard of the audiobook industry in these countries and the most sold smart speaker, Alexa. Moreover, some individuals can be more used to listening to artificial voices, such as blind people using talking books. Therefore, further research should explore these aspects. Finally, we only analyzed a female voice. The effect of other voices should be examined in future studies to reinforce these results.

Without a doubt, technology will continue improving artificial voices, and during the next few years, they could tell stories captivating listeners. From this point of view, speech processing and synthesis can benefit from this study to enhance the artificial voices' expressivity, which is the main component to improve. Until then, for a synthetic voice, it will not be easy to compete with the richness of a professional human actor narrating a story.

Authors note

Emma Rodero is also affiliated with UPF-Barcelona School of Management, Spain and Ignacio Lucas is also affiliated with Department of Psychiatry, Hospital Universitari de Bellvitge, Spain.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by Leonardo Grants for Researchers and Cultural Creators, BBVA, Spain.

ORCID iD

Emma Rodero https://orcid.org/0000-0003-0948-3400

References

Audio Publishers Association (APA) (2021). The Voice of the Industry. Available at: https://www.audiopub.org/ (accessed 23 May 2021).

Assmann P, Neary T and Dembling S (2006) Effects of frequency shifts on perceived naturalness and gender information in speech. In: *Proceedings of the 9th international conference on spoken language processing*, Pittsburgh, PA, 17–21 September.

Babin LA and Burns AC (1998) A modified scale for the measurement of communication-evoked mental imagery. *Psychology & Marketing* 15(3): 261–278.

Barker P (2015) With one voice: disambiguating sung and spoken voices through a composer's experience. In: Thomaidis K and Macpherson B (eds) *Voice Studies: Critical Approaches to Process, Performance and Experience*. London: Routledge, pp. 16–26.

Barthes R (1985) *The Responsibility of Forms: Critical Essays on Music, Art, and Representation*. New York: Hill and Wang.

Bolls PD and Lang A (2003) I saw it on the radio: the allocation of attention to high-imagery radio advertisements. *Media Psychology* 5(1): 33–55.

Bolls PD and Muehling DD (2007) The effects of dual-dask processing on consumers' responses to high- and low-imagery radio advertisements. *Journal of Advertising* 36(4): 35–34.

Bone PF and Ellen PS (1992) The generation and consequences of communication-evoked Imagery. *Journal of Consumer Research* 19: 93–103.

Bradley MM and Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25(1): 49–59.

Busselle RV and Bilandzic H (2008) Transportation and transportability in the cultivation of genre-consistent attitudes and estimates. *Journal of Communication* 58(3): 508–529.

Cavarero A (2005) *FoR More Than One Voice: Toward a Philosophy of Vocal Expression*. Palo Alto, CA: Stanford University Press.

Chen F (2006) *Designing Human Interface in Speech Technology*. Boston, MA: Springer.

Chion M (1994) *Audio-vision: Sound on Screen*. New York: Columbia University Press.

Connor S (2000) *Dumbstruck: A Cultural History of Ventriloquism*. Oxford: Oxford University Press.

Craig SD and Schroeder NL (2017) Reconsidering the voice effect when learning from a virtual human. *Computers and Education* 114: 193–205.

Craig SD and Schroeder NL (2019) Text-to-Speech software and learning: investigating the relevancy of the voice effect. *Journal of Educational Computing Research* 57(6): 1534–1548.

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

Di Matteo P (2015) Performing the Entre-Deux: the capture of speech in (dis)embodied voices. In: Thomaidis K and Macpherson B (eds) *Voice Studies: Critical Approaches to Process, Performance and Experience*. London: Routledge, pp. 104–119.

Edison Research (2019) The Infinite Dial 2019. Available at: https://www.edisonresearch.com/infinite-dial-2019/

Edison Research (2020) The smart audio report 2020. Available at: https://www.edisonresearch.com/the-smart-audio-report-2020-from-npr-and-edison-research/

Ellen PS and Bone PF (1991) Measuring communication-evoked imagery processing. *Advances in Consumer Research* 18: 806–812.

Goosens G (1994) Enactive imagery: information processing, emotional responses, and behavioral intentions. *Journal of Mental Imagery* 18(3/4): 119–150.

Green MC and Brock TC (2000) The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology* 79(5): 701–721.

Green MC, Brock TC and Kaufman GE (2004) Understanding media enjoyment: the role of transportation into narrative worlds. *Communication Theory* 14: 311–327.

Have I and Pedersen BS (2020) The audiobook circuit in digital publishing: voicing the silent revolution. *new media & society* 22(3): 409–428.

Humphry J and Chesher C (2020) Preparing for smart voice assistants: cultural histories and media innovations. *new media & society*. Epub ahead of print 22 May. DOI: 10.1177/1461444820923679.

Jenkins JJ and Franklin LD (1982) Recall of passages of synthetic speech. *Bulletin of the Psychonomic Society* 20(4): 203–206.

Kosslyn SM (1994) *Image and Brain: The Resolution of the Imagery Debate*. Cambridge, MA: The MIT Press

Kosslyn SM, Ganis G and Thompson WL (2001) Neural foundations of imagery. *Nature Reviews Neuroscience* 2: 635–642.

Lai J, Wood D and 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*, The Hague, 1–6 April, pp. 321–328.

Lang A, Gao Y, Potter RF, et al. (2015) Conceptualizing audio message complexity as available processing resources. *Communication Research* 42: 759–778.

Luce PA (1981) *Comprehension of fluent synthetic speech produced by rule*. Research on Speech Perception Progress, Report No. 7, pp. 229-242. Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce PA, Feustel TC and Pisoni DB (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.

Mar RA and Oatley K (2008) The function of fiction is the abstraction and simulation of social experience. *Perspectives in Psychological Science* 3: 173–192.

Mayer RE (2014) Principles based on social cues in multimedia learning: personalization, voice, image, and embodiment principles. *The Cambridge Handbook of Multimedia Learning* 16: 345–370.

Miller DW and Marks LJ (1997) The effects of imagery evoking radio advertising strategies on affective responses. *Psychology and Marketing* 14: 337–361.

Mulac A and Giles H (1996) ‘You’re only as old as you sound’: perceived vocal age and social meanings. *Health Communication* 8: 199–215.

Mullennix JW, Stern SE, Wilson SJ, et al. (2003) Social perception of male and female computer synthesized speech. *Computers in Human Behavior* 19(4): 407–424.

Paivio A (1991) Dual coding theory: retrospect and current status. *Canadian Journal of Psychology/ revue Canadienne De Psychologie* 45(3): 255.

Paris CR, Thomas MH, Gilson RD, et al. (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.

Parker B (2013) Should you hire a computer to narrate your audiobook? *The Book Designer*. Available at: https://www.thebookdesigner.com/2013/11/ispeech/

Pisoni DB and Hunnicutt S (1980) Perceptual evaluation of MITalk: the MIT unrestricted text-to-speech system. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, Denver, CO, 9–11 April, pp. 572–575. New York: IEEE.

Potter RF and Bolls P (2012) *Psychophysiological Measurement and Meaning: Cognitive and Emotional Processing of Media*. London: Routledge.

Rodero E (2012) See it on a radio story: sound effects and shots to evoked imagery and attention on audio fiction. *Communication Research* 39(4): 458–479.

Rodero E (2016) Influence of speech rate and information density on recognition: the moderate dynamic mechanism. *Media Psychology* 19(2): 224–242.

Rodero E (2017) Effectiveness, attention, and recall of human and artificial voices in an advertising story. Prosody influence and functions of voices. *Computers in Human Behavior* 77: 336–346.

Rodero E (2020) A voice you can’t see. *The Unesco, Courier* 1: 18–19.

Rodero E, Potter RF and Prieto P (2017) Pitch range variations improve cognitive processing of audio messages. *Human Communication Research* 43(3): 397–413.

Roring RW, Hines FG and Charness N (2007) Age difference in identifying words in synthetic speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49: 25–31.

Sanderman AA and Collier R (1997) Prosodic phrasing and comprehension. *Language and Speech* 40: 391–409.

Stern SE, Mullennix JW, Dyson CL, et al. (1999) The persuasiveness of synthetic speech versus human speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 41(4): 588–595.

Syrdal AK, Bennett RW and Greenspan SL (Eds) (1994) *Applied Speech Technology*. Boca Raton, FL: CRC Press.

Taake KP (2009) *A comparison of natural and synthetic speech: with and without simultaneous reading*. Thesis, Washington University, St. Louis, MI.

Tattersall Wallin E and Nolin J (2020) Time to read: exploring the timespaces of subscription-based audiobooks. *new media & society* 22(3): 470–488.

Thoet A (2017) A short history of the audiobook, 20 years after the first portable digital audio device. Available at: https://www.pbs.org/newshour/arts/a-short-history-of-the-audiobook-20-years-after-the-first-portable-digital-audio-device (accessed 23 May 2021).

van Dommelen WA and Moxness BH (1995) Acoustic parameters in speaker height and weight identification: sex-specific behaviour. *Language and Speech* 38: 267–287.

Winters SJ and Pisoni DB (2004) Research on spoken language processing: perception and comprehension of synthetic speech. *Progress Report* 26. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.298.1410&rep=rep1&type=pdf

Wolters MK, Johnson C, Campbell PE, et al. (2014) Can older people remember medication reminders presented using synthetic speech? *Journal of the American Medical Informatics Association* 22: 35–42.

Xu K (2019) First encounter with robot Alpha: how individual differences interact with vocal and kinetic cues in users' social responses. *new media & society* 21(11–12): 2522–2547.

Author biographies

Emma Rodero is a researcher and professor at Pompeu Fabra University (Spain), PhD. in Communication and PhD. in Psychology. She is the director of the Media Psychology Lab, devoted to analyzing cognitive processes, emotions, and behavior underlying media and technology interactions.

Ignacio Lucas is a postdoctoral researcher in the Behavioural Addictions and Eating Disorders Unit of the Department of Psychiatry of the Bellvitge University Hospital/ Bellvitge Biomedical Research Institute (IDIBELL). His research interests are focused on human health and behaviour from Neuropsychological and Psychophysiological perspectives.