The Persuasive Design of AI-synthesized Voices

Hannah Chang
Singapore Management University
Anirban Mukherjee
Cornell University

Acknowledgements:
This research is supported by the Ministry of Education, Singapore, under its Tier 2 Grant (MOE-T2EP40221-0008).

Cite as:
Chang Hannah, Mukherjee Anirban (2023), The Persuasive Design of AI-synthesized Voices. *Proceedings of the European Marketing Academy*, 52nd, (113984)

Paper from the 52nd Annual EMAC Conference, Odense/Denmark, May 23-26, 2023

The Persuasive Design of AI-synthesized Voices

Abstract:

We investigate the impact of AI-based, machine-synthesized narrating voices on consumer cognitions and behavior in media-rich environment. Across four studies (plus pretests), we show that the design of AI voices systematically and predictably affects consumer cognition and behavior. Specifically, the designs of AI voices have differential effects in early versus later stages of consumer purchase journey. In situations where the consumers' attention is already directed to the message, we find that marcomm with more AI voices generates a smaller proportion of favorable thoughts, which leads to a lower purchase likelihood. These results support our conceptualization that hearing more AI voices narrate a message is more cognitively effortful for listeners to process compared to hearing a single AI voice narrate the same message. Moreover, this effect is attenuated for consumers who enjoy expending cognitive effort and detrimental in consumption contexts where the consumer is more familiar with the product category. Substantive and theoretical implications are discussed.

Keywords: voice, voice assistants, video, persuasion

Track: Consumer Behavior

1. Introduction of Paper

Artificial intelligence technology seeks to emulate humans. One aspect is AI-synthesized voices, used in voice assistants (such as Amazon Alexa, Apple Siri, and Google Assistant) and other applications. With the extensive availability and enhanced accuracy of AI-synthesized voices, consumer research is starting to examine the impact of AI-synthesized voices on consumer information processing and decision making. The extant literature, however, is relatively limited for two key reasons. First, despite its proliferation, AI voice technology is still relatively new. Previous studies on how consumers process information conveyed in spoken speech tended to focus on human voices. Second, the effect of sound and narrator's voice on consumer behavior remains fairly under-researched, perhaps due to methodological challenges in the design of appropriate experimental stimuli (cf. Krishna and Schwarz 2014; Dahl 2010).

AI-synthesized voices aim to emulate human voices and differ in several acoustic dimensions. Human voices differ in many key acoustic properties (e.g., timbre, pitch) due to differences in physical features of the vocal tract (Fant 1960). These acoustic properties map onto higher-level language processes, such as word recognition (Pisoni and Luce 1987). Moreover, beyond the conveyed verbal content, a human voice carries additional (nonverbal) information for the brain to process, which facilitates social functioning. Through paralinguistic cues (e.g., loudness, tone), listeners receive indexical information such as a speaker's gender, emotional state, and age (Belin, Fecteau, and Bédard 2004).

In this research, we posit that the use of more narrating AI voices should prompt consumers to process the overall spoken message in a more cognitively effortful manner compared to the use of a single AI voice in persuasive videos. We examine the implications of this phenomenon on consumer purchase likelihood and delineate the consumer contexts and conditions in which it is likely to aid/hinder purchase. We test our predictions in four experiments, spanning different decision domains, product categories, and outcomes. Experiment 1 examines the effect of the same versus different narrating AI voices on persuasion, depending on whether consumers are in earlier or later stages of consumer purchase journey. Prior findings suggest that changes in narrator voices can help capture consumer attention (Cherry 1953). Building on prior research, we posit and test that hearing multiple AI voices convey a message can affect how consumers process it depending on whether consumer attention is already directed

(vs. not directed) toward the message, which in turn affects persuasion. The next two experiments focused on the effect of narrating voices on processing after consumer attention is obtained. Experiment 2 assesses if processing difficulty increases with more AI voices. Experiment 3 examines if it is possible for consumers to exhibit more favorable responses after attending to product message narrated by more AI voices. Experiment 4 examines whether consumer characteristics relating to attentional allocation moderate the effect of number of AI voices. Across the studies, we find that the effect is mediated by the favorability of cognitive responses (Experiments 1, 3, and 4). Due to space constraint, we focus on reporting Experiments 1, 2, and 4 in the space below.

2. Study 1: Consumer Attention and AI-synthesized Voices

The purpose of this experiment was to examine whether the same versus different AI voices help or hinders persuasion, depending on whether consumer attention is already directed to the message narrated by the AI-based voice-over. We varied participants' attention to the product messaged narrated by one or five voices in a video's voice-over. We predicted that when attention is obtained, participants' purchase likelihood for the product would be higher when the product message is narrated by the same AI voice than by different AI voices. In contrast, when attention is not obtained, participants' purchase likelihood would be higher when the message is narrated by five voices than by one voice.

2.1 Method

Participants and Design. A total of 468 U.S. participants (53.9% women; $M_{age} = 40.3$) who qualified (scoring more than 1 on a 7-point scale of general interest) completed the study for a small monetary compensation. They were randomly assigned to one of four conditions of a 2 (attentional orientation: control vs. attention) × 2 (number of AI voices: 1 vs. 5) between-subjects design.

Procedure and Measures. In the main decision task, participants were asked to imagine that they were looking for a wireless charger and came across a new product in an online marketplace. Before they were shown the product video, we varied participants' attentional focus toward the spoken product message. Half of the participants (attention conditions) were given the additional

instructions: “In this study, we are interested in understanding the extent to which you listen to and think about the product message in the video’s voiceover. Please focus on what the voiceover is saying while watching the video.” To ensure that participants in the attention conditions read the instructions, they were asked to indicate their agreement with the statement: “I will pay attention to the video’s voiceover and listen carefully to what is said about the new product” (1=“disagree”; 7=“agree”). In contrast, the remaining participants (control conditions) were not given these instructions.

All participants were shown a video about the target product. The original video was an actual product video posted on Kickstarter, the world’s largest crowdfunding website. The video utilized voice-over narration, a popular production technique where a voice (from an off-screen narrator) discusses information about a brand or product. It did not have an on-screen narrator speaking to the camera but only visually showcased the product design and usage, which was conducive to our key manipulation of the number of AI voices discussing the product. We created different versions of the video that were identical in visual and spoken messages; the only difference was in the number of AI voices in the voice-over. Half of the participants watched the versions of the video with the same voice narrating the entire product message throughout the video; the other half watched the versions of the video with five different synthetic voices narrating the same overall message sequentially (each AI voice narrated about $1/5^{th}$ of the overall product message before a new voice carries on the next part of the message). We used five voice-synthesis (text-to-speech) models based on US English from a leading cloud service to create the five synthetic voices. The voices across the versions were counterbalanced.

After watching the video, participants indicated their purchase likelihood on a single item from 1 (definitely will not buy it) to 7 (definitely will buy it), which served as the main dependent measure. Participants were asked to write the thoughts they had as they watched the video (adapted from Cacioppo and Petty 1981) to shed light on their thought responses. We wrote a script for the thought-listing procedure such that as participants typed a thought in a text box, a new text box would appear below. On the ensuing screen, we displayed the thoughts participants had written and asked them to code each thought as negative, neutral, positive, or irrelevant to the product and/or the video clip.

We collected a demand check and confounding checks of task involvement and mood. Participants indicated their (a) task involvement on three items (e.g., “I took the task of evaluating

the product very seriously") from 1 ("strongly disagree") to 7 ("strongly agree") (α = .76); and (b) mood on four seven-point items (e.g., "unpleasant/pleasant," "bad/good"; α = .92). As a manipulation check for number of AI voices in the video, participants selected a number from 1 to 7. We also included an instructional manipulation check (IMC) to help detect participant satisficing to increase statistical power (Oppenheimer, Meyvis, and Davidenko 2009). Finally, participants reported background information (gender, age) and whether the video and audio track loaded properly for the product-evaluation task.

2.2 Results and Discussions

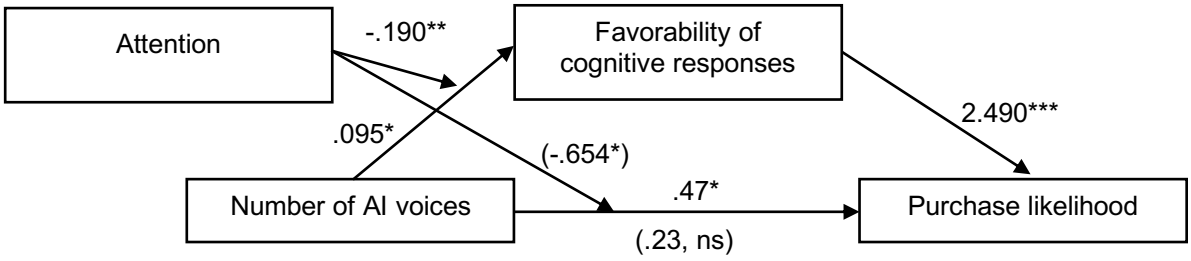
Preliminary Analyses. No participant correctly guessed the purpose of the study, but three were removed because they reported that the video did not load and 25 participants failed the IMC. The manipulation of number-of-AI-voices was successful ($F(1, 436) = 334.65$, $p < .0001$). No differences in task involvement (ps > .13) were found across conditions. There was an unexpected interaction of number of AI voices × attentional orientation ($F(1, 436) = 5.66$, $p = .0178$) on mood. Other effects in these models were nonsignificant.

Purchase likelihood. An ANOVA of participants' purchase likelihood yielded only a significant interaction of number of AI voices × attentional orientation ($F(1, 436) = 12.76$, $p = .0004$). As shown in Figure 1, the results show that participants in the control conditions were more likely to purchase the product after hearing more voices narrate the product message ($M_5 = 4.89$, $M_1 = 4.42$; $F(1, 436) = 4.29$, $p = .0389$). Central to our theorizing, participants in the attention conditions were *less* likely to do so after hearing more AI voices narrate the product message ($M_5 = 4.50$, $M_1 = 5.17$; $F(1, 436) = 8.92$, $p = .0030$). These substantive results remained after controlling for participants' mood.

Proportion of Favorable Thoughts. Analysis of participants' cognitive responses revealed a significant interaction between number of AI voices and attentional orientation ($F(1, 436) = 8.19$, $p = .0044$). Participants in the control conditions had higher proportion of favorable thoughts when they heard a message narrated by more voices ($M_5 = .56$, $M_1 = .47$; $F(1, 461) = 4.03$, $p = .0453$). Participants in the attention conditions had lower proportion of favorable thoughts when they heard a message narrated by more voices ($M_5 = .44$, $M_1 = .54$; $F(1, 461) = 4.16$, $p = .0419$).

Mediation. As shown in Figure 1 Panel (A), we conducted a moderated-mediation analysis to examine the extent to which participants' cognitive responses influenced the indirect effects of

number of AI voices on purchase likelihood across attentional orientations. We applied a standard bootstrap procedure (PROCESS Model 8; Hayes 2022) and specified a confidence level of 95% with 10,000 bootstrap resamples. Results show that the indirect effect associated with the number of AI voices × attentional orientation interaction through the proportion of favorable thoughts was significant (index of moderated mediation = -.4734; 95% CI [-.8195, -.1430]). Results show that the conditional indirect effect of the number of AI voices on purchase likelihood through proportion of favorable thoughts was (a) significant and positive for participants in the control conditions (indirect effect = .2365; 95% CI [.0053, .4785]) and (b) significant and negative for participants in the attention conditions (indirect effect = -.2369; 95% CI [-.4702, -.0061]).



Note. The path coefficient are unstandardized betas. Values in parentheses are the effects on the DV after controlling for the mediator. *$p < .05$, **$p < .01$, ***$p < .001$

Figure 1. Study 1 results

3. Study 2: Perceived Ease of Processing Information Narrated by AI Voices

Study 1 provides initial evidence that hearing different AI voices narrate a message undermines persuasion compared to hearing a single AI voice narrate the same message. Our conceptual rationale is that processing a message spoken by multiple voices is more cognitively effortful than processing the same message spoken by a single AI voice. Indirect support comes from neuroscience and memory studies, which found that hearing more voices activates more brain regions (e.g., von Kriegstein et al. 2003). We directly assesses this premise using AI voices in the context of consumer cognition and persuasion this experiment.

3.1 Method

Participants and Design. We recruited 200 U.S. participants (50% women; M_{age} = 40.5) from the Cloudresearch online panel. They were randomly assigned to one of two experimental conditions (number of AI voices: 1 vs. 5).

Procedure and Measures. Participants completed the same product-evaluation task as in Study 1, except in three aspects. First, we focused on the attention condition where participants were instructed to "watch the entire video clip and listen carefully to its product description." Second, after watching the video, participants reported how easy (1 = "very easy") or difficult (7 = "very difficult") it was to understand the product information, which was the main dependent measure. Third, given the study objective, we excluded the thought-listing task. The same demand check, manipulation check (of the number of AI voices), video/audio loading measure, and background information (gender, age, general interest in product innovations) as in Study 1 were collected.

3.2 Results and Discussions

None of the participants' guesses related number of AI voices to processing difficulty. Seven participants were removed prior to analyses as the video did not load and three for lack of general interest in crowdfunding. Subsequent analyses were based on 190 observations. The manipulation of number of AI voices was successful ($F(1, 188) = 101.62$, $p < .0001$, $\eta^2 = .35$), with the average estimate of 1.15 in the 1-voice condition and 2.59 in the 5-voice condition.

An ANOVA of the reported ease or difficulty of understanding the product information revealed a significant effect of number of AI voices ($F(1, 188) = 6.22$, $p = .0135$, $\eta^2 = .03$). Consistent with our theorizing, participants indicated greater difficulty in understanding the product information when it was narrated by five voices (M = 2.05) than when it was narrated by one voice (M = 1.62). In Experiment 3, we examine the effect using consumer characteristics. Given space constraint considerations, we discuss Study 4 next.

4. Study 4: AI Voices and Product-category Familiarity

The purpose of this experiment was two-fold: (1) to examine whether consumer characteristics relating to attentional allocation would moderate the effect of number of narrating voices and (2) to extend the effect to another product category. Past research has suggested that

subjective familiarity—how much people think they know about a product (Lichtenstein and Fischhoff 1977) or category (Gregan-Paxton, Hoeffler, and Zhao 2005)—can affect their attentional allocation and cognitive processes in evaluating new products. Compared to less-familiar consumers, more-familiar consumers are also more likely to attend to relevant and important information (Alba and Hutchinson 1987). We thus used participants' familiarity to the product category as a proxy of their attention to the product messaged narrated by the video's AI voice-over (with one or five AI voices). We predicted that the facilitative effect of voice numerosity would be more likely when participants' attention is explicitly directed to the narrated product message than when their attention is not explicitly directed to the message.

4.1 Method

Participants and Design. A total of 260 U.S. participants (52.9% women; M_{age} = 41) who qualified (scoring more than 1 on a 7-point scale of general interest) completed the study for a small monetary compensation. They were randomly assigned to one of the two experimental conditions (number of AI voices: 1 vs. 5).

Procedure and Measures. The main decision task is similar to Experiment 1 except that we used a different product—a heated smart mug that keeps hot beverages at their suitable drinking temperature—to allow us to extend the effect to a different category. All participants watched a short video clip discussing the product. Similar to Experiment 1, we created different versions of the product video in which the spoken and visual messages are identical across the versions, the only difference being the AI voice(s) narrating the spoken message. We used five voice-synthesis models (which convert text to speech) to create the voices; they were counterbalanced using a Latin-square design to ensure comparability across the versions. We created modified videos by combining audio tracks from the voice-synthesis models with visual frames of the original video. After watching the video, participants stated their purchase likelihood from 1 (definitely will not buy it) to 7 (definitely will buy it), which served as the main dependent measure. (Participants completed the same thought-listing task as in Study 1; due to space constraint, in this study we focus on the main results.)

4.2 Results and Discussions

Preliminary analyses. No participant correctly guessed the purpose of the study; one participant who failed the IMC was removed from further analyses. Subsequent analyses were based on 259 observations. Participants noticed the number of different voices in the video, with the average estimate of 1.09 in the one-voice condition and 2.26 in the five-voice condition ($F(1, 257) = 127.99$, $p < .0001$, $\eta^2 = .33$). The manipulation was successful.

Purchase likelihood. We analyzed participants' purchase likelihood using a multiple linear regression with participants' familiarity (M = 2.70, SD = 1.63, Min = 1, Max = 7), number of AI voices (dummy coded; 1-voice condition as the reference group), and their interaction term as predictors. The regression showed a significant main effect of narrating voices: participants had higher purchase likelihood when they heard more voices narrating the product message ($\beta = 1.06$, $t(255) = 2.71$, $p = .0074$, $\eta^2 = .028$). The main effect of familiarity was also significant ($\beta = .19$, $t(255) = 2.13$, $p = .034$, $\eta^2 = .017$). More germane to our hypothesis, the interaction between familiarity and number of AI voices was significant ($\beta = -.35$, $t(255) = -2.84$, $p = .0049$, $\eta^2 = .031$). A floodlight analysis (Spiller et al. 2013) identified the range of familiarity scores for which the simple effect of the number-of-voices manipulation was significant. A significant effect of number of AI voices was revealed for any familiarity score less than 1.67 ($\beta_{JN} = .47$, SE = .24, $p = .05$) and a significant effect for any familiarity score greater than 4.90 ($\beta_{JN} = -.67$, SE = .34, $p = .05$). In other words, participants who were less familiar with the category stated higher purchase likelihood for the product in the five-voice condition than in the one-voice condition, whereas participants who were more familiar with the category stated lower purchase likelihood in the five-voice condition than the one-voice condition.

5. General Discussion

AI-based voice communication—voiceover narration, voice assistant, and accessibility tool—is becoming more important to consumers in the marketplace. Yet, research on the voice effect on consumer behavior is relatively under-researched (Krishna & Schwarz, 2014; Dahl, 2010). We aim to contribute by examining the impact of AI-synthesized voices on consumer behavior. In particular, we find that the number of AI voices has a significant and important effect on consumer cognition. In situations where the consumers' attention is already directed to the message, marcomm with more AI voices generates a smaller proportion of favorable thoughts,

which leads to a lower purchase likelihood. The results support our conceptualization that hearing more AI voices narrate a message is more cognitively effortful for listeners to process than hearing a single voice. This backfire effect is attenuated when consumers have a greater need for cognition (i.e., beneficial for consumers who enjoy expending cognitive effort), and amplified in consumption contexts where the consumer is more familiar with the product.

References.

Alba, Joseph, and Wes Hutchinson (1987), "Dimensions of Consumer Expertise," *Journal of Consumer Research,* 13(4), 411–54.

Belin, Pascal, Shirley Fecteau, and Catherine Bédard (2004), "Thinking the Voice: Neural Correlates of Voice Perception," *Trends in Cognitive Sciences,* 8 (3), 129–35.

Cacioppo, John T. and Richard E. Petty (1981), "Social Psychological Procedures for Cognitive Response Assessment: The Thought Listing Technique," in *Cognitive Assessment,* ed. Thomas Merluzzi, Carol Glass, and Myles Genest, New York: Guilford, 309–42.

Cherry, Colin E. (1953), "Some Experiments on the Recognition of Speech with One and with Two Ears," *Journal of the Acoustical Society of America*, 25, 975–79.

Fant, Gunnar (1960), *Acoustic Theory of Speech Production,* The Hague: Mouton.

Fischhoff, Baruch, Paul Slovic, and Sarah Lichtenstein (1977), "Knowing with Certainty: The Appropriateness of Extreme confidence," *Journal of Experimental Psychology: Human Perception and Performance,* 3(4), 552–64.

Gregan-Paxton, J., S. Hoeffler, and Min Zhao (2005), "When Categorization is Ambiguous: Factors that Facilitate the Use of a Multiple Category Inference Strategy," *Journal of Consumer Psychology,* 15(2), 127–40.

Hayes, Andrew F. (2022), *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-based Approach,* Vol. 3. New York, NY: The Guilford Press.

Krishna, Aradhna and Norbert Schwarz (2014), "Sensory Marketing, Embodiment, and Grounded Cognition: A Review and Introduction," *Journal of Consumer Psychology*, 24 (2), 159–68.

Oppenheimer, Daniel. M., Tom Meyvis, and Nicolas Davidenko (2009), "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power," *Journal of Experimental Social Psychology,* 45(4), 867–72.

Pisoni, David B., and Paul A. Luce (1987), "Acoustic-phonetic Representations in Word Recognition," *Cognition,* 25 (1-2), 21–52.

Statista (2022), "Online video usage in the United States" (accessed August 3, 20212, https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/.

von Kriegstein, Katharina, Evelyn Eger, Andreas Kleinschmidt, and Anne Lise Giraud (2003), "Modulation of Neural Responses to Speech by Directing Attention to Voices or Verbal Content," *Cognitive Brain Research*, 17 (1), 48–55.