International Journal for Educational Media and Technology
2025, Vol.19, No. 1, pp.101-113

Examining Synthetic Speech in Corporate Training: Focusing on Impression Formation and Likability

Shingo Marubayashi
Media Opus Plus Inc., Japan
marubayashi@mediaopusplus.com

Mizuki Ohtori
Media Opus Plus Inc., Japan
ohtori@mediaopusplus.com

Hiroki Murakawa
Nihon Fukushi University, Japan
murakawa@n-fukushi.ac.jp

Fumito Kitamura
Nagasaki University, Japan
kitamuraf@nagasaki-u.ac.jp

Norio Setozaki
Nagasaki University, Japan
setozaki@nagasaki-u.ac.jp

This study aims to identify the characteristics of synthetic speech that tend to elicit higher levels of likability by examining evaluations and impressions of synthetic speech. To achieve this objective, two practices were conducted. Practice 1 involved a confirmation test, subjective evaluation, and an impression formation survey on video teaching materials featuring an instructor's human voice and a synthetic speech. Practice 2 focused only on evaluating synthetic speech (four types in total: two male and two female), to survey on evaluating and impression formation surveys. In Practice 1, the results indicated no significant differences in test scores or subjective evaluations between the two materials. However, in the impression formation survey, human voices were significantly more highly evaluated for characteristics such as "Friendly," "Pleasant," "Sensible," and "Approachable." In Practice 2, synthetic speech perceived as "Pleasant," "Approachable," and "Kind" tended to receive higher evaluation scores.

Keywords: Corporate training, Impression formation, Synthetic speech, Trait adjective scales

Introduction

Corporate training is said to play an important role in adult education. For example, Ben-Hur (2014) states that "learning is one of the key elements for companies to survive, succeed, and maintain competitive advantage." In other words, people are an important asset in companies, and improving the quality of human resources is an important management issue for companies. In addition, Harashchenko (2019) points out the importance of corporate education amid social change, and Nakahara (2012) emphasizes the importance of rethinking the workplace as a "learning environment" rather than simply a place to perform work.

As such, corporate training plays an important role in adult education. In companies where diverse human resources work in a variety of ways, face-to-face education poses challenges in terms of cost, time, and human resources. Synthetic speech is expected to be an effective educational medium in addressing these challenges. The use of synthetic speech in corporate training offers several advantages: (1) it facilitates the creation of teaching materials such as videos based on text scripts; (2) it reduces costs, including human resources, associated with instructors or narrator changes and revisions of materials due to content updates; (3) it allows for the provision of voice options tailored to learners' preferences; and (4) contributes to multilingual support and universal learning. Points (1) to (4) enable more flexible and inclusive training approaches that go beyond the limitations of traditional face-to-face education, better accommodating a diverse workforce and work styles.

However, for companies to consider utilizing synthetic speech, it must be possible to conduct corporate training using synthetic speech, with no difference in confirmation tests and evaluations equal to or better than those given to the instructor's human voice.

As a precedent study on synthetic speech, Mitani (2014) conducted an experiment using in-house broadcasts at a lifelong learning facility and found that human voices were easier to understand than synthetic speech at the time. However, it is necessary to consider the background of this comparison, namely that the human voices were those of professional announcers, that synthetic speech technology was less developed than it is today, and that the characteristics of the speech design had an impact. On the other hand, Dinçer (2022) concluded that, with the development of text-to-speech conversion technology, modern text-to-speech conversion engines can achieve the same effect as human voices regarding learning outcomes and cognitive load. As such, synthetic speech is being verified in stages, and the difference between human voices and synthetic speech is narrowing due to technological developments, suggesting that synthetic speech has the potential to increase its effectiveness as an educational medium.

Furthermore, research on the use of such synthetic speech in the context of higher education is limited, although a few studies have been conducted. Compared to research focusing on natural speech—for instance, Yamasumi et al. (2005), who developed a new scale for objectively evaluating impressions of lecture speech—studies on synthetic speech are relatively scarce. Ikenoue and Kitazawa (2023) compared teachers' natural voices with synthetic speech of the same teachers and synthetic speech of other people, and found that the support rate was highest for the natural voices of teachers, followed by synthetic speech of other people, and then synthetic speech of teachers themselves.

These results suggest that familiarity with the instructor's human voice may have caused discomfort toward the synthetic speech. However, it is also possible that the differences in the impressions formed by the teachers' human voices, the teachers' synthetic speech, and the synthetic speech of other people affected the approval ratings.

Furthermore, there was a lack of research on the differences in speech and preferences arising from differences in the technical foundations used to generate synthetic speech. Schwab et al. (2012) conducted a study on word recognition training using synthesized speech and found that the group trained with synthetic speech performed better on word recognition tests than those trained with natural voices or not trained at all. As such, research on the use of synthetic speech in education is progressing in higher education. However, many studies in higher education are conducted within the same university, which tends to result in a high degree of homogeneity among participants.

On the other hand, companies employ a diverse workforce with diverse work styles, so the diversity of participants differs from that in higher education institutions. Furthermore, corporate training is conducted as tasks that must be performed in the course of business, so the nature of the training also tends to differ. For this reason, there has been a demand from the business community for research based on practice in companies.

However, although there are studies that discuss the importance of sound and voice in e-learning (Rautela 2024), there is a lack of research based on the investigation of confirmation tests and subjective evaluations through the practice of corporate training, differences in voice caused by differences in the technical basis for generating synthetic speech, impression formation, and likability.

In this way, when synthetic speech is used in corporate training, it is necessary to ensure that the training is effective and that there is no difference in the results of confirmation tests, and that the synthetic speech is evaluated as being equal to or better than the instructor's human voice. However, the question of what kind of synthetic speech should be used in corporate training has not yet been resolved.

The research question of this study is: To what extent does the use of synthetic speech preferred by participants enhance the effectiveness of corporate training?

Therefore, this study aims to identify the characteristics of synthetic speech that tend to be highly evaluated to investigate whether differences in impression formation toward speech affect the evaluation of synthetic speech through practice in companies. Two practices were established to achieve this objective.

Practice 1 aimed to clarify and evaluate the differences between the instructor's human voice and synthetic speech in video teaching materials used in corporate training. To this end, we conducted a confirmation test on the content of the video materials, subjective evaluations, and an impression formation survey.

Practice 2, to eliminate the influence of the instructor's human voice, we compared only synthetic speech, rather than known human voices, to investigate the possibility of problems with the speech itself, such as the tone and pronunciation of the instructor's human voice. In the comparison, from the perspective of impression formation, which showed significant differences in Practice 1, we compared four types of synthetic speech generated using the same technical foundations in order to eliminate differences in speech due to technical differences in synthetic speech generation and reduce the influence of speech design characteristics. In Practice 2, we aimed to identify impression formation questions highly correlated with synthetic speech's evaluation scores. To this end, we investigated the impression formation and likability of four types of synthetic speech.

Comparison of human voice and synthetic speech (Practice 1)

Purpose of Practice 1

One area where synthetic speech is increasingly expected to be utilized is in corporate training through video teaching materials. In recent years, there has been growing demand for video teaching materials in corporate training due to considerations for viewing environments and the need to accommodate universal learning. The advantages of synthetic speech include reduced costs, including labor costs associated with changing instructors and narrators and making corrections to update teaching materials. In video teaching materials, if synthetic speech is evaluated as equivalent to or more effective than the instructor's human voice, it has the potential to fulfill the objectives of corporate training.

Practice 1 aimed to clarify and evaluate the differences between the instructor's human voice and synthetic speech in video teaching materials used in corporate training.

Methods and subjects

In the first practice, we obtained the cooperation of 35 individuals aged 19 to 60 (average age 32.91, SD=13.23) working at Japanese ICT venture companies in the education sector. Only participants who provided informed consent were included in the survey. Among the 35 participants, 33 were native Japanese speakers, 1 was a native Chinese speaker, and 1 was a native Vietnamese speaker. Additionally, the two participants whose native language was not Japanese had sufficient language proficiency to perform their work in Japanese without difficulty. One of these individuals had obtained a N2 certification in the Japanese Language Proficiency Test (JLPT).

In Practice 1, training was conducted using materials based on ISO 27001-compliant ISMS (Information Security Management System). The ISMS training utilized lecture-style video materials with slides. After the training, participants completed a confirmation test on the video materials to assess their understanding of the content. The confirmation test consists of 10 questions. Each worth 10 points, for a total of 100 points.

Subsequently, as part of informed consent, participants were informed that participation in the study was voluntary and that non-participation would not result in any disadvantage. Written consent was then obtained. The document clearly stated the results might be published, for example in academic papers, in a form that would not disclose individual identification. Furthermore, the first item in the survey included a question confirming whether the participant had read the research information sheet and consent form, and only participants whose gave consent was confirmed through this item were included in the study. Participants who provided informed consent were asked about their age, years of service, number of training sessions attended, employment status within the company, and the type of video teaching materials they viewed. Following this, a survey was conducted to gather subjective evaluations and impressions regarding the synthesized speech used in the training materials.

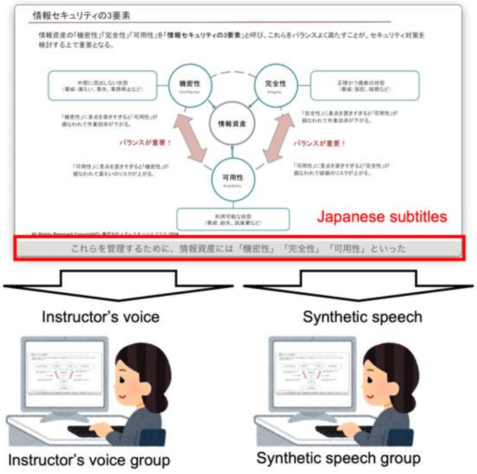
The video materials were produced in two formats: one featuring the instructor's human voice and the other featuring synthetic speech. Both video materials had the same slides, narration script, and subtitles, with only difference in narration. The video materials were produced in Japanese, including the subtitles.

Synthetic speech was used as the text-to-speech feature of Microsoft Word for Windows (version 2406) by Microsoft Corporation. The reason for this is that the software is widely used in companies, is easy to use, and is likely to become one of the options when considering synthetic speech for corporate training in the future. Since there are studies showing that men and women may perceive speech differently (Re et al., 2012; McAleer et al., 2014), in Practice 1, we compared the human voice of a male instructor with synthetic speech to eliminate gender effects. The video teaching materials using the two types of speech were randomly assigned to participants. 18 participants who watched the video with the instructor's human voice were classified as the "instructor voice group", while the 17 participants who watched the video with the synthetic speech were classified as the "synthetic speech group".

In both groups, subjective evaluations of the video materials themselves were created based on Nagahama et al. (2018).

However, since the instructor's face was not displayed in the video materials for Practice 1, the question "The instructor's face was visible" was changed to "It was better if the instructor's face was visible" in accordance with Nagahama & Morita (2017), and the question "Did you pay attention to the instructor's video during the lecture?" was deleted. Additionally, the term "lecture format consisting of slides, instructor video, subtitles, and audio" was revised to remove "instructor video." Furthermore, considering the possibility that the impression formed by audio may influence video materials, we used the adjective scale developed by Hayashi (1978). The overview of the survey is shown in Figure 1.

Figure 1
Practice 1 Overview of the survey


情報セキュリティの3要素

情報資産の「機密性」「完全性」「可用性」を「情報セキュリティの3要素」と呼び、これらをバランスよく満たすことが、セキュリティ対策を検討する上で重要となる。

外部に流出しない状態
（脅威：漏えい、紛失、業務停止など）

正確かつ最新の状態
（脅威：誤記、破損など）

「機密性」に重点を置きすぎると「可用性」が損なわれて作業効率が下がる。

「完全性」に重点を置きすぎると「可用性」が損なわれて作業効率が下がる。

バランスが重要！

バランスが重要！

「可用性」に重点を置きすぎると「機密性」が損なわれて漏えいのリスクが上がる。

「可用性」に重点を置きすぎると「完全性」が損なわれて破損のリスクが上がる。

利用可能な状態
（脅威：紛失、誤廃棄など）

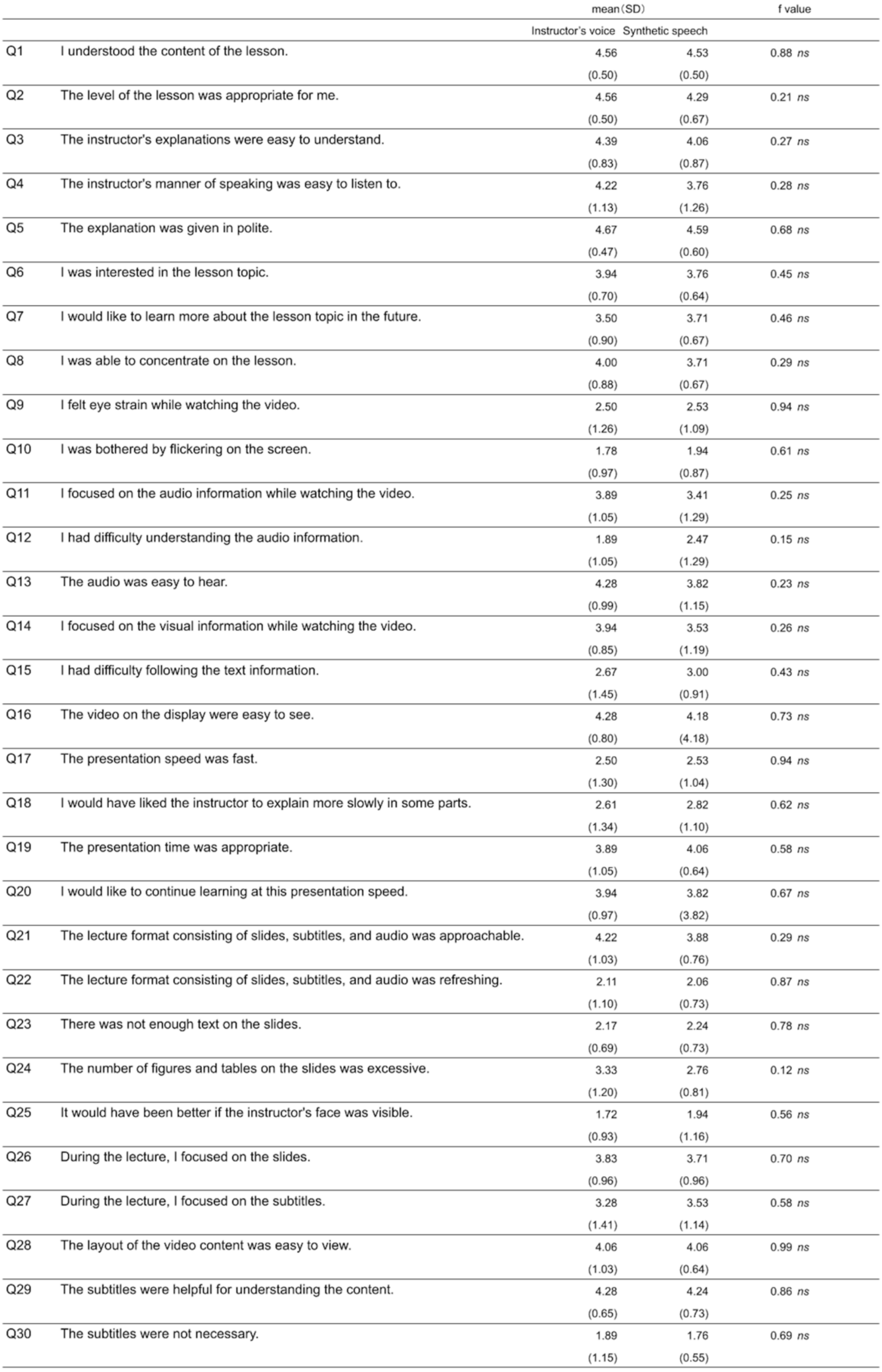
これらを管理するために、情報資産には「機密性」「完全性」「可用性」といった

Results and Discussion

The confirmation test is a requirement for completing the company's training program, with a score of 80 or higher out of 100 points, and may be administered multiple times. Therefore, the scores from the first test were analyzed for target. The average scores were 89 points for the instructor-voiced group and 92 points for the synthetic speech group. A t-test was conducted between the two groups, and no significant difference was found (t(33) = 0.94, p = .36). While the purpose of the confirmation test was to assess the level of understanding of the video materials, which is a relatively simple task, and consideration of the ceiling effect was necessary, both groups achieved the objectives of the corporate training program. In the subjective evaluation questions regarding the video materials themselves, a 5-point scale ranging from "strongly agree (5)" to "strongly disagree (1)" was used, and the average score for each question was calculated. A t-test was conducted between the two groups, and as shown in Table 1, no significant differences were found in any of the questions.

The absence of significant differences between the two groups in the confirmation test indicates that the understanding of the video materials was equivalent. Additionally, the lack of significant differences in subjective evaluations between the two groups suggests that there is no clear difference between the instructor's human voice and the synthetic speech. These results suggested that corporate training utilizing synthetic speech can potentially achieve its objectives.

Table 1
Subjective evaluation questions



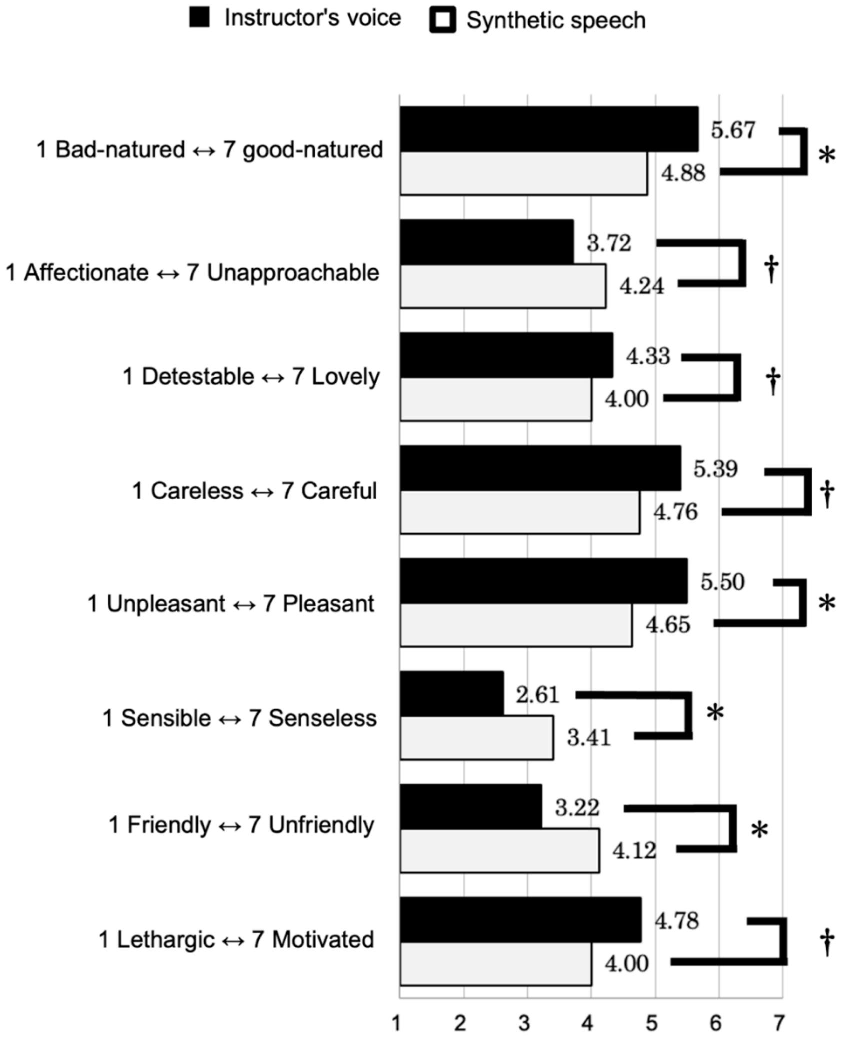
		mean（SD）		f value
		Instructor's voice	Synthetic speech	
Q1	I understood the content of the lesson.	4.56 (0.50)	4.53 (0.50)	0.88 *ns*
Q2	The level of the lesson was appropriate for me.	4.56 (0.50)	4.29 (0.67)	0.21 *ns*
Q3	The instructor's explanations were easy to understand.	4.39 (0.83)	4.06 (0.87)	0.27 *ns*
Q4	The instructor's manner of speaking was easy to listen to.	4.22 (1.13)	3.76 (1.26)	0.28 *ns*
Q5	The explanation was given in polite.	4.67 (0.47)	4.59 (0.60)	0.68 *ns*
Q6	I was interested in the lesson topic.	3.94 (0.70)	3.76 (0.64)	0.45 *ns*
Q7	I would like to learn more about the lesson topic in the future.	3.50 (0.90)	3.71 (0.67)	0.46 *ns*
Q8	I was able to concentrate on the lesson.	4.00 (0.88)	3.71 (0.67)	0.29 *ns*
Q9	I felt eye strain while watching the video.	2.50 (1.26)	2.53 (1.09)	0.94 *ns*
Q10	I was bothered by flickering on the screen.	1.78 (0.97)	1.94 (0.87)	0.61 *ns*
Q11	I focused on the audio information while watching the video.	3.89 (1.05)	3.41 (1.29)	0.25 *ns*
Q12	I had difficulty understanding the audio information.	1.89 (1.05)	2.47 (1.29)	0.15 *ns*
Q13	The audio was easy to hear.	4.28 (0.99)	3.82 (1.15)	0.23 *ns*
Q14	I focused on the visual information while watching the video.	3.94 (0.85)	3.53 (1.19)	0.26 *ns*
Q15	I had difficulty following the text information.	2.67 (1.45)	3.00 (0.91)	0.43 *ns*
Q16	The video on the display were easy to see.	4.28 (0.80)	4.18 (4.18)	0.73 *ns*
Q17	The presentation speed was fast.	2.50 (1.30)	2.53 (1.04)	0.94 *ns*
Q18	I would have liked the instructor to explain more slowly in some parts.	2.61 (1.34)	2.82 (1.10)	0.62 *ns*
Q19	The presentation time was appropriate.	3.89 (1.05)	4.06 (0.64)	0.58 *ns*
Q20	I would like to continue learning at this presentation speed.	3.94 (0.97)	3.82 (3.82)	0.67 *ns*
Q21	The lecture format consisting of slides, subtitles, and audio was approachable.	4.22 (1.03)	3.88 (0.76)	0.29 *ns*
Q22	The lecture format consisting of slides, subtitles, and audio was refreshing.	2.11 (1.10)	2.06 (0.73)	0.87 *ns*
Q23	There was not enough text on the slides.	2.17 (0.69)	2.24 (0.73)	0.78 *ns*
Q24	The number of figures and tables on the slides was excessive.	3.33 (1.20)	2.76 (0.81)	0.12 *ns*
Q25	It would have been better if the instructor's face was visible.	1.72 (0.93)	1.94 (1.16)	0.56 *ns*
Q26	During the lecture, I focused on the slides.	3.83 (0.96)	3.71 (0.96)	0.70 *ns*
Q27	During the lecture, I focused on the subtitles.	3.28 (1.41)	3.53 (1.14)	0.58 *ns*
Q28	The layout of the video content was easy to view.	4.06 (1.03)	4.06 (0.64)	0.99 *ns*
Q29	The subtitles were helpful for understanding the content.	4.28 (0.65)	4.24 (0.73)	0.86 *ns*
Q30	The subtitles were not necessary.	1.89 (1.15)	1.76 (0.55)	0.69 *ns*

On the other hand, in the questions of the trait adjective scale that asked about impression formation, 20 pairs of adjectives were presented using a 7-point SD scale, and respondents were asked to select the one that best matched their impression of the instructor's human voice. The 7-point scale was converted directly into scores, and the average score for each question item was calculated. A t-test was conducted between the two groups, and as shown in Figure

2, there were significant differences in the questions "1 Bad-natured ⇔ 7 Good-natured," "1 Unpleasant ⇔ 7 Pleasant," and "1 Sensible ⇔ 7 Senseless." "Friendly ⇔ Unfriendly" (t(33) = 2.16, p < .05; t(33) = 2.63, p < .05; t(33) = 2.14, p < .05; t(33) = 2.09, p < .05). In addition, there were significant differences in the questions "1 Affectionate ⇔ 7 Unapproachable," "1 Detestable ⇔ 7 Lovely," "1 Careless ⇔ 7 Careful," and "1 Lethargic ⇔ 7 Motivated" (t(33) = 1.83, p < .10; t(33) = 2.00, p < .10; t(33) = 1.99, p < .1 there were significant trends (t(33) = 1.83, p < .10; t(33) = 2.00, p < .10; t(33) = 1.96, p < .10; t(33) = 1.69, p < .10).

Differences in the characteristic adjective scales were found in items such as "1 Sensible ⇔ 7 Senseless" and "1 Friendly ⇔ 7 Unfriendly," suggesting that the instructor's human voice and synthetic speech may give different impressions to learners. These characteristics may have been influenced by differences between the instructor's human voice and synthetic speech, such as problems with the voice itself, such as the tone and pronunciation of the instructor's human voice and whether the voice is that of a colleague known to the learner or an unfamiliar synthetic speech, as well as the relative discomfort of synthetic speech compared to human voices and the characteristics of the voice design.

Figure 2
Results of characteristic adjective scale



（＊＊：p<.01,　＊：p<.05,　†：.05<p<.10）

As shown above, in Practice 1, no significant differences were observed between the instructor's natural voice and the synthetic speech regarding confirmation test scores and subjective evaluations of the video materials. This suggests that for slide-based lecture video materials, the difference attributable to using either a natural or synthetic speech was minimal. Subtitles, in particular, represent an initiative supporting universal learning by assisting learners such as non-native Japanese speakers and those with hearing impairments. It is possible that subtitles contributed to mitigating the perceived differences between the natural and synthetic speech.

On the other hand, an examination of impression formation between the instructor's human voice and synthetic speech revealed significant differences or trends in several items, suggesting that differences in impression formation may exist.

Verification of impression formation of synthetic speech (Practice 2)

Practice 2 Objectives

Practice 1 showed that there was no significant difference between the actual voice and synthetic speech in terms of the training confirmation test scores and subjective evaluation of the video teaching materials themselves. However, there were significant differences or significant trends in some of the items on the adjective scale used to measure impression formation, indicating the possibility of differences in impression formation.

However, it was not possible to clearly determine whether the differences in impression formation were due to differences between human voices and synthetic speech, or differences between the human voices of colleagues known to the participants and the synthetic speech of unknown voices, and therefore it was not possible to clearly identify whether the differences in impression formation led to differences in the evaluation of synthetic speech.

Therefore, in Practice 2, to eliminate the influence of the instructor's human voice, such as the possibility of problems with the voice itself, such as the tone and pronunciation of the instructor's human voice, only synthetic speech was compared, rather than known human voices. In addition, for synthetic speech, we selected synthetic speech of unknown characters to eliminate the influence of impressions formed by known characters associated with the voices.

Furthermore, using synthetic speech generated with the same technical foundations, we eliminated the influence of technical differences and reduced the influence of voice design characteristics. By comparing four types of synthetic speech generated with the same technical standards rather than human voices, we aimed to identify impression formation questions that were highly correlated with the evaluation scores for synthetic speech from the perspective of impression formation, where significant differences were found in Practice 1.

Methods and subjects

In the second practical survey, a total of 78 individuals employed at educational ICT venture companies in Japan participated. Of these, 52 participants (21 men and 31 women), aged 19 to 65 (M = 35.27, SD = 14.47), who provided informed consent were included in the study. Of the 52 participants, 51 were native Japanese speakers and 1 was a native Vietnamese speaker. The Vietnamese speaker had obtained a N2 level on the Japanese Language Proficiency Test (JLPT) and was capable of understanding Japanese sufficiently. Approximately 20 of the survey participants had experience using synthetic speech in their work.

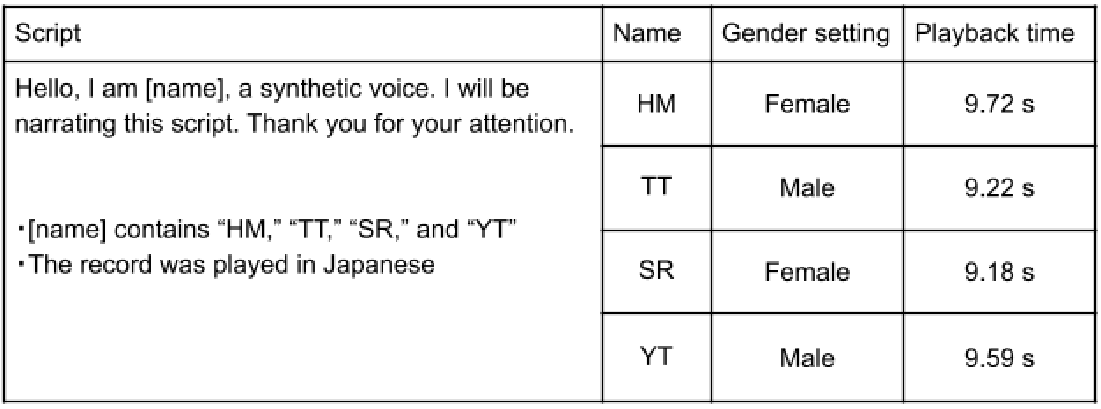
Similar to Practice 1, informed consent was obtained, and the participants were asked to provide their employee ID numbers, gender, and age. They were then asked about their acceptability of synthetic speech in corporate training. The question asked was, "Do you think it is desirable to use synthetic speech in corporate training?" and respondents were asked to choose one of five answers: "Strongly agree," "Agree," "Neither agree nor disagree," "Disagree," or "Strongly disagree." After that, we conducted a survey on the listening of synthetic speech and impression formation. The employee ID numbers and responses obtained in Practice 2 were analyzed as statistical data that could not be used to identify individuals, as in Practice 1.

In the synthetic speech listening test, four types of synthetic speech were recorded in Japanese. For the speech synthesis, "Future Voice Crayon," a Japanese synthetic speech service developed by NTT TechnoCross Co., Ltd. utilizing deep learning, was adopted. "Future Voice Crayon" is known for its ability to produce natural and expressive synthetic speech. The reason for its adoption is that it is a synthetic speech technology based on the Japanese language, and it is equipped with more than 50 patterns of synthetic speech characters generated using the same technical foundations. This makes it possible to eliminate the influence of impressions of known characters and differences due to the superiority or inferiority of the technical foundations, and to reduce the influence of voice design features. In addition, it is possible to systematically change the settings of the synthetic speech quality, speech rate, intonation, and voice pitch, which will enable further development in the future. The text scripts used for the synthesized speech were identical, with only the voice names differing. Furthermore, the scripts were intentionally kept simple to minimize the influence of content on participants' impressions.

In Practice 2, the four types of synthetic speech were symbolically named "HM," "TT," "SR," and "YT" to eliminate the influence of the names on the impressions. The genders of the four types of synthetic speech were set as male for "TT" and "YT" and female for 'HM' and "SR." In Practice 2, we compared not only male voices but also female voices to ensure gender balance proposed by Kobayashi and Kurakata (2023).The playback times for the four types of speech were 9.72 seconds for "HM," 9.22 seconds for "TT," 9.18 seconds for "SR," and 9.59 seconds for "YT." The contents of the reading script, the names of the synthetic speech, the gender settings, and the playback times of the synthetic speech are shown in Table 2.

Table 2

Summary of four types of synthetic speech

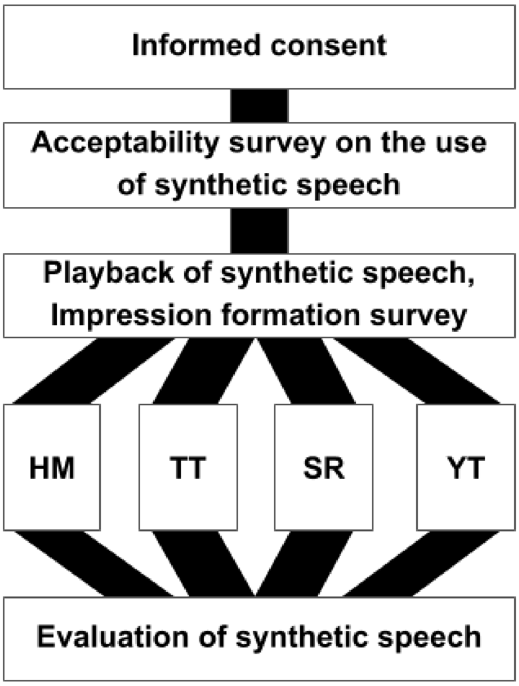


Script	Name	Gender setting	Playback time
Hello, I am [name], a synthetic voice. I will be narrating this script. Thank you for your attention. ・[name] contains "HM," "TT," "SR," and "YT" ・The record was played in Japanese	HM	Female	9.72 s
	TT	Male	9.22 s
	SR	Female	9.18 s
	YT	Male	9.59 s

In addition, to eliminate order bias when presenting the four types of synthetic speech to survey participants, each participant was assigned a randomized listening order. Participants listened to the four types of synthetic speech in the specified order and responded to impression-related questions for each voice immediately after listening. The impression survey employed the same items used in Practice 1, based on the trait adjective scale developed by Hayashi (1978), consisting of Questions 1 through 20.

After all participants had listened to all four synthetic speech and answered the impression formation questions, each participant was asked to evaluate the synthetic speech to explore the possibility that differences in impression formation were linked to evaluations of likability. The question was: "How much do you like the synthetic speech [name] (where [name] is contained with the name of the synthetic speech, 'HM,' 'TT,' 'SR,' or 'YT')?" Participants were asked to rate their liking on a scale of 1 to 10. Since there was a possibility that participants might forget the voices or impressions after listening to the four types of synthetic speech and answering the impression formation questions, they were allowed to take notes if necessary and could review the synthetic speech they wanted to confirm as many times as needed. The overview of the survey is shown in Figure 3.

Figure 3
Practice 2: Overview of the survey


Informed consent
Acceptability survey on the use of synthetic speech
Playback of synthetic speech, Impression formation survey
HM
TT
SR
YT
Evaluation of synthetic speech

Results and Discussion

In an acceptability survey on the use of synthetic speech, when asked, “Do you think it is a good idea to use synthetic speech in corporate training?”, Of the 52 valid responses, three respondents answered “Strongly agree,” 25 answered “Agree,” 19 answered “Neither agree nor disagree,” three answered “Disagree,” and two answered “Strongly disagree.” A chi-square test was conducted to assess the fit of the response results, and a significant association was observed ($p < .01$). Multiple comparisons of the number of responses showed that the “Agree” response was significantly different from the “Strongly agree,” “Disagree,” and “Strongly disagree” responses ($\chi^2(52) = 3.97$, $p < .01$; $\chi^2(52) = 3.97$, $p < .01$; $\chi^2(52) = 4.23$, $p < .01$). In addition, there was a significant difference between the “Neither agree nor disagree” response and the “Strongly agree,” “Disagree,” and “Strongly disagree” responses ($\chi^2(52) = 3.20$, $p < .01$; $\chi^2(52) = 3.20$, $p < .01$; $\chi^2(52) = 3.49$, $p < .01$). Based on these results, it can be said that the majority of respondents answered “Agree” or “Neither agree nor disagree.” This suggest that the use of synthetic speech in corporate training is tolerated to a certain extent by participants.

When the responses to the questions were analyzed in terms of age, with “Strongly agree” as 5, “Agree” as 4, “Neither agree nor disagree” as 3, “Disagree” as 2, and “Strongly disagree” as 1, there was no correlation between age and the acceptability of synthetic speech ($r = .00$). Therefore, we can say that the participants in the survey did not show a tendency to evaluate synthetic speech less favorably with increasing age, and that corporate training using synthetic speech is generally accepted to a certain extent. However, the participants in the survey were employed by Japanese ICT venture companies, and their acceptability of AI narration, including ICT literacy, may differ from that of other industries and companies. In addition, some of the participants in the survey use synthetic speech in their work, and their familiarity with synthetic speech may have affected the results.

Following the acceptability survey on the use of synthetic speech, the participants were asked to listen to four types of synthetic speech. After that, as an impression formation survey, 20 questions on characteristic adjectives were presented using a 7-point SD scale. In the impression formation survey, the participants were asked to select the adjective that best described their impression of each synthetic speech. The numerical responses were calculated as scores, and a correlation analysis was performed between the average of each question item and the evaluation scores obtained on a 10-point scale for the likability of the speech. Table 3 shows the correlation coefficients between the characteristic adjective scales and the evaluation scores for the four types of synthetic speech.

Table 3

Correlation coefficients between adjectival scales describing the characteristics of four synthetic speech and evaluation scores

	correlation coefficient			
Question	HM	TT	SR	YT
Q1(1 Active ⇔ 7 Passive)	-0.01	-0.24	-0.14	-0.02
Q2(1 Bad-natured ⇔ 7 Good-natured)	0.25	0.51	0.61	0.17
Q3(1 Not impudent ⇔ 7 Impudent)	-0.19	-0.21	-0.41	0.17
Q4(1 Affectionate ⇔ 7 Unapproachable)	-0.21	-0.27	-0.30	-0.11
Q5(1 Detestable ⇔ 7 Lovely)	0.40	0.19	0.30	0.25
Q6(1 Broad-minded ⇔ 7 Narrow-minded)	-0.21	0.03	-0.21	-0.17
Q7(1 Unsociable ⇔ 7 Sociable)	0.23	0.34	0.10	-0.03
Q8(1 Responsible ⇔ 7 Irresponsible)	-0.40	-0.37	-0.20	-0.08
Q9(1 Careless ⇔ 7 Careful)	0.19	0.07	0.15	-0.01
Q10(1 Shameless ⇔ 7 Shy)	-0.10	-0.11	-0.24	0.07
Q11(1 Dignified ⇔ 7 Frivolous)	-0.09	-0.12	-0.08	-0.14
Q12(1 Cheerless ⇔ 7 Cheerful)	0.08	0.32	-0.04	0.14
Q13(1 Assertive ⇔ 7 Sneaky)	-0.19	-0.24	-0.31	-0.10
Q14(1 Unpleasant ⇔ 7 Pleasant)	0.48	0.46	0.58	0.26
Q15(1 Sensible ⇔ 7 Senseless)	-0.08	-0.11	0.05	-0.07
Q16(1 Friendly ⇔ 7 Unfriendly)	-0.31	-0.28	-0.38	-0.25
Q17(1 Lethargic ⇔ 7 Motivated)	0.16	0.25	0.29	0.15
Q18(1 Unconfident ⇔ 7 Confident)	0.07	0.23	0.23	0.15
Q19(1 Patient ⇔ 7 Impatient)	-0.14	-0.16	-0.22	-0.06
Q20(1 Unkind ⇔ 7 Kind)	0.43	0.34	0.49	0.31

The results of the correlation analysis between these characteristic adjective scales and the evaluation scores showed that there was a moderate or weak correlation in all four types of statements: question 14 (1 Unpleasant ⇔ 7 Pleasant), question 16 (1 Friendly ⇔ 7 Unfriendly), and question 20 (1 Unkind ⇔ 7 Kind). Therefore, statements with characteristics such as “pleasant,” “friendly,” and “kind” are associated with likability and tend to receive higher evaluation scores.

These characteristics are desirable in corporate settings, as they foster a sense of security and trust while reducing discomfort and stress. Therefore, such attributes should be considered when developing synthetic speech for use in corporate training.

In addition, in order to identify adjective pairs that are less correlated with evaluation scores in the correlation between evaluation scores for synthetic speech and impression formation questions, we extracted those with low correlation in each synthetic speech.

As a result, we found that the adjective pairs in question 9 (1 Careless ⇔ 7 Careful) (HM: r = .19, p = .16; TT: r = .07, p = .62; SR: r = .11, p = .42; YT: r = .15, p = .27), question 11 (1 Dignified ⇔ 7 Frivolous) (HM: r = .09, p = .51; TT: r = .12, p = .40; SR: r = .08, p = .56; YT: r = .14, p = .33), question 15 (1 Sensible ⇔ 7 Senseless) (HM: r = .08, p = .59; TT: r = .11, p = .42; SR: r = .05, p = .71; YT: r = .07, p = .65) showed no correlation in any of the four types of audio. Therefore, in Practice 2, regardless of which evaluation increased, the likability score did not increase or decrease for questions such as "Careless ⇔ Careful," "Dignified ⇔ Frivolous," and "Sensible ⇔ Senseless."

Characteristics such as "Careless ⇔ Careful," "Dignified ⇔ Frivolous," and "Sensible ⇔ Senseless" may exist in a psychologically independent dimension from the comprehensive evaluation of "likability" that listeners have toward speech. If so, it suggests that these items are evaluated from a different perspective than likability, and therefore do not correlate with likability. Examples include social trust in speech and the listener's personality traits. To verify this point, it will be necessary to use factor analysis or other methods to clarify the underlying structure of the evaluation scale and separate the core factors that constitute "likability" from other factors. On the other hand, it is also possible that the listeners could not make a judgment due to the perceptual limitations of the speech stimuli. The synthetic speech used in this study did not reach a level where its acoustic characteristics strongly evoked specific impressions such as "Careless" or "Dignified," and it is possible that the evaluators were unable to perceive clear differences in these dimensions. In other words, these characteristics did not affect likability, but were not sufficiently perceived from the speech stimuli in the first place. In future research, it will be necessary to systematically manipulate voice quality, speaking speed, intonation, pitch, and other factors, or modify the text being read aloud, to identify the conditions under which these impressions are perceived. In Practice 1, the significant difference between the instructor's human voice and synthetic speech in terms of "sensible ⇔ senseless" became less pronounced, suggesting that there may have been problems with the instructor's human voice itself, such as tone and pronunciation, the difference between the human voices of colleagues known to the participants and the synthetic speech of unknown speakers, the relative unfamiliarity of synthetic speech compared to human voices, and the characteristics of the speech design may have influenced the difference between human and synthetic speech.

Summary

The research question of this study is: To what extent does the use of synthetic speech preferred by participants enhance the effectiveness of corporate training?

This study aims to identify the highly evaluated synthetic speech characteristics and investigate whether differences in impression formation toward speech affect the evaluation of synthetic speech through practice in companies. To achieve this objective, two practices were established.

Practice 1 aimed to clarify and evaluate the differences between the instructor's human voice and synthetic speech in video teaching materials used in corporate training. To this end, we conducted a confirmation test on the content of the video materials, subjective evaluations, and an impression formation survey.

Practice 2, from the perspective of impression formation, where there were significant differences in Practice 1, we compared four types of synthetic speech generated using the same technical foundations in order to eliminate differences in speech due to technical differences in synthetic speech generation and reduce the influence of speech design characteristics. To this end, we investigated the impression formation and likability of four types of synthetic speech.

In Practice 1, there was no significant difference between the scores on the training confirmation test and the subjective evaluation of the video materials themselves between participants employed by Japanese educational ICT venture companies.
This suggests that even in corporate training utilizing synthetic speech, it is possible to achieve the training objective of understanding the content of the materials. There were differences in the impressions formed by the instructor's human voice and synthetic speech in items such as "Sensible ⇔ Senseless" (1) and "Friendly ⇔ Unfriendly" (1) on the adjective scale used to measure impression formation. These differences were due to issues with the voice itself, such as the tone and pronunciation of the instructor's human voice, whether it is the human voice of a colleague the participants know, or a synthetic speech they do not know, and other factors.

In Practice 2, we first examined the acceptability of synthetic speech in corporate training by conducting a survey on the listening and impression formation of four types of synthetic speech. The results showed no significant trend suggesting that evaluations of synthetic speech decrease with age, suggesting that the use of synthetic speech in corporate training is generally acceptable to participants. However, it is important to note that the survey was conducted in Japanese with participants whose native language is Japanese, and the average age of participants was 35.27 years (SD=14.47), with the sample consisting of individuals employed at Japanese education ICT venture companies. Different results may be observed in companies with employees whose native languages are more diverse, companies where the language used is not the employees' native language, or in companies with younger or older average ages, those in different industries, or organizations with different corporate cultures.

In addition, in a study of synthetic speech perception and impression formation, we compared synthetic speech alone, rather than known human voices, to explore the possibility that differences in impression formation lead to evaluations of synthetic speech, investigated how impression formation affects their evaluation and aimed to identify voice characteristics associated with higher favorability.

As a result, in the survey of four types of synthetic speech listening and impression formation, the correlation coefficients between the evaluation scores and the characteristic adjective scales were calculated, revealing moderate to weak correlations in questions 14 (1 Unpleasant ⇔ 7 Pleasant), question 16 (1 Friendly ⇔ 7 Unfriendly), and question 20 (1 Unkind ⇔ 7 Kind) for all four types of speech. Characteristics such as "pleasant," "friendly," and "kind" are positive elements that are sought after in companies, as they are associated with a lack of discomfort or stress and a sense of security and trust. These characteristics also reduce the auditory burden on the listener. This suggests that they should be considered when evaluating synthetic speech for corporate training. Conversely, questions such as "Careless ⇔ Careful," "Dignified ⇔ Frivolous," and "Sensible ⇔ Senseless" showed little correlation with likability scores regardless of which evaluation increased and decreased. This can be considered a characteristic that is evaluated independently of the listener's overall evaluation of the pleasantness of the voice.

Characteristics such as "Careless ⇔ Careful," "Dignified ⇔ Frivolous," and "Sensible ⇔ Senseless" may have been relatively less important than other characteristics directly related to communication in organizations based on functional human relationships. Based on these results, we will consider the use of synthetic speech in corporate training, taking into account characteristics such as "pleasant," "friendly," and "kind," which have the potential to reduce the burden on learners and improve training effectiveness. We will continue to conduct further research in the future.

On the other hand, it is possible that the synthetic speech used in this study did not convey impressions such as "Careless ⇔ Careful," "Dignified ⇔ Frivolous," and "Sensible ⇔ Senseless". It is worth investigating whether other synthetic speech would produce similar results. In Practice 1, the association between "sensible ⇔ senseless," a significant difference between the instructor's human voice and the synthetic speech became weak. Such a result may be due to factors such as there may have been a problem with the tone or pronunciation of the instructor's human voice, whether it was the human voice of a colleague the participants knew or a synthetic speech they did not know, or the difference between a human voice and a synthetic speech.

A correlation analysis between the characteristic adjective scales of four types of synthetic speech and evaluation scores suggested that differences in the characteristic adjective scales were related to evaluation scores in some questions regarding impression formation. This indicates that synthetic speech with higher evaluation scores may share common elements. However, the specific characteristic adjectives and the magnitude of the correlation remain unclear. Going forward, we will continue to investigate specific descriptive adjectives and the magnitude of related differences, while also considering the possibility that listeners may not have been able to make judgments due to the perceptual limits of auditory stimuli. We aim to systematically manipulate voice quality, speech rate, intonation amplitude, and pitch using audio stimuli. Additionally, we consider that further research is needed to investigate the relationship between differences in speech content and training content, as well as the relationship with playback speed (Nagahama et al.2017;Nagahama et al.2018) and speech speed, similar to the findings for videos. By advancing these studies, we will be able to reflect more accurate emotions and tones in prompts and settings in response to the rapid evolution of synthetic speech, which will lead to the generation of synthetic speech that meets the objectives of corporate training. At the same time, we will consider conducting research that focuses on the effectiveness of corporate training.

In addition, although this study focused on using synthetic speech, we would like to explore the potential of digital twins, virtual avatars (Mizuho et al. 2024) changes in the learning environment in the future, with the goal of finding the optimal synthetic speech for learners.

Note

This paper is based on the content presented by Marubayashi et al. (2024) at the 31st Annual Conference of the Japan Association for Educational Media Study.

References

Ben-Hur, S. (2014). The business of corporate learning: Insights from practice. Eiji Publishing.

Dinçer, N. (2022). The voice effect in multimedia instruction revisited: Does it still exist? Journal of Pedagogical Research,6,3.

Harashchenko, L., Komarovska, O., Matviienko, O., Ovsiienko, L., Pet'ko, L., Shcholokova, O., & Sokolova, O. (2019). Models of corporate education in the United States of America. Journal of Entrepreneurship Education, 22(3), 1–6.

Hayashi, F. (1978). Adjective scales for traits. In Hori, H. (Ed.), Psychological Measurement Scale Collection II (pp. 5–9). Science-sha

Ikenoue, Y., & Kitazawa, T. (2023). Comparison of Teacher's Synthetic speech and Different Person's Synthetic speech in Video Materials: Results of Investigation on Possibility of Substitution from Natural Voice to Synthetic speech in Video Materials. Journal of the Japan Association for Education of Information Studies, 16(1), 75–83.

Kobayashi, M., & Kurakata, K. (2023). Which is easier to hear, a female or male voice? Journal of the Acoustical Society of Japan, 79(2), 85–93.

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'hello'? Personality impressions from brief novel voices. PLoS ONE*, 9*(3), e90779. https://doi.org/10.1371/journal.pone.0090779

Mitani, M. (2014). Which site-broadcasting is better in lifelong learning facilities? Comparison between two systems, one with natural voice and the other with artificially synthetic speech by hearing experience. Humans and Nature, 25,63−74.

Mizuho, T., Amemiya, T., Narumi, T., & Kuzuoka, H. (2024). Virtual Omnibus Lecture: Effects of Remote Lecturing Using Various Lecturer Avatars on Students Memory. The Virtual Reality Society of Japan，29(3),109−118.

Nagahama, K., Kanno, H., & Morita, Y. (2018). The effects of high-speed presentation of video content on learning outcomes: Focusing on learning style and dual-channel model. Japan Society for Educational Technology, 41(4), 345–362.

Nagahama, K., & Morita, Y. (2017). Analysis of learning effects from high-speed presentation of video content. Japan Society for Educational Technology, 40(4), 291–300.

Nakahara, J. (2012). The workplace as a learning environment. The Japanese Journal of Labour Studies, 618, 35–45.

Rautela, V. (2024). Enhanced learning outcomes with audio in e-learning: An analysis. International Journal of Advanced Corporate Learning, 17(4), 69–79, https://doi.org/10.3991/ijac.v17i4.48547

Re, D. E., O'Connor, J. J., Bennett, P. J., & Feinberg, D. R. (2012). Preferences for very low and very high voice pitch in humans. PLoS ONE, 7(3), e32719. https://doi.org/10.1371/journal.pone.0032719

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. Human Factors, 27(4), 395–408.

Yamasumi, K., Kagomiya, T., Maki, Y., & Maekawa, K. (2005). Impression evaluation scales for lecture speech. The Journal of the Acoustical Society of Japan, 61(6), 303–311.