INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A stimulus-organism-response model of willingness to buy from advertising speech using voice quality

Mizuki Nagano[1], *Yusuke Ijima*[1] *and Sadao Hiroya*[2]

[1]NTT Human Informatics Laboratories, Japan
[2] NTT Communication Science Laboratories, Japan

mizuki.nagano@ntt.com

Abstract

Speech can affect the behavior of the listener. Our previous study showed that the stimulus-organism-response theory using emotional states can explain a person's willingness to buy from advertising speech. In addition, there have been reports of the influence of voice quality in speech, which differs from other advertising stimuli, but few studies have been done on willingness to buy. In this study, we conducted an experiment to determine whether voice quality affects the willingness to buy from advertising speech. Participants listened to advertising speech and rated their willingness to buy the products advertised and their own emotions and voice quality. We found that a model constructed using voice quality as a mediator can better explain the willingness to buy from advertising speech. These results could help train salespeople in advertising speech.

Index Terms: the willingness to buy, advertising speech, voice quality

1. Introduction

Speech can influence human behavior [1, 2, 3, 4]. However, the relationship between speech and human behavior is nonlinear. Therefore, a hierarchical model has been proposed to account for this. The stimulus-organism-response (SOR) theory is a known representative model of this process [5].

Using this model, the effect of advertising speech on consumers' willingness to buy is investigated. We have shown through a mediation analysis that the emotion-mediated SOR model is effective in determining willingness to buy from advertising speech [6]. Poon et al. investigated the influence of the differences in pause length of male and female speech on the willingness to buy through a perceived personality state [7]. This indicates that consumers' internal states, which are generated by advertising speech, influence their willingness to buy.

However, it is known that both emotion and voice quality are generated from speech [8]. Voice quality refers to the speech quality of the speaker as perceived by the listener [9]. Elbarougy et al. showed that both the highest F0 and the F0 contour could influence the perception of voice calm [10] (S–O). Kobayashi et al. suggested that the perceived tenseness of speech announcements during disasters can encourage evacuation behavior [3] (O–R).

However, few studies have examined whether the voice quality-mediated SOR model can explain the effect of advertising speech on willingness to buy. Wiener et al. investigated the influence of three voice qualities (creaky, tense, and whisper) on the willingness to buy and concluded that a creaky voice was the most effective [11] (O–R), but they did not conduct a validation based on the SOR model. Since voice quality is more intuitively understandable to humans than emotion [8], clarification of this issue could be useful for training salespeople on advertising speech.

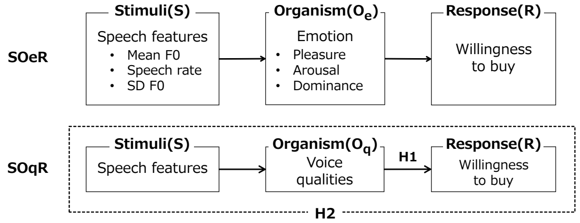

Figure 1: *Emotion-mediated SOR model (SOeR) used in [6] (top). The hypothesized SOR model mediated by voice quality (SOqR) in this study (bottom).*

In this study, we verified the influence of voice quality on the willingness to buy on the basis of the SOR model. We generated synthesized speech of multiple speakers talking about various types of products. We conducted a large-scale online survey experiment to evaluate the voice quality, emotions, and willingness to buy from advertising speech. Using the evaluation values of emotions, we confirmed whether the same results as previous studies [6] could be obtained using synthesized speech. We examined the influence of voice quality on the willingness to buy using multiple regression analysis, path analysis, and mediation analysis on the basis of the SOR model.

2. SOR model and hypothesized models

2.1. SOR model

The SOR theory consists of the three dimensions of stimulus, organism, and response [5]. Stimulus refers to all external environmental factors in a store such as brand image [12], design [13, 14], crowding [15], atmosphere [16], color, scent, and music [17]. These stimulate the organism, such as quality [14], perceived satisfaction [17], and emotions [18]. Organisms make us do approach or avoidance behavior. An approach behavior refers to a positive attitude toward the environment, such as staying in the store. In contrast, avoidance behavior refers to a negative attitude toward the environment, such as leaving the store. In consumer behavior research, approach or avoidance behaviors are confirmed by indicators such as longer staying times [13] and increased purchases [19].

We have previously shown that the emotion-mediated SOR model (SOeR) can explain the willingness to buy from advertising speech [6] (Figure 1). Emotion consists of three states: pleasure, arousal, and dominance [20]. Pleasure refers to the degree of feeling joy, satisfaction, and happiness. Arousal refers to the degree of excitement, passion, and activity. Dominance refers to the degree of feeling that an individual has influence and control over a situation.

10.21437/Interspeech.2023-1131

2.2. Voice qualities and hypothesized models

This study aimed to verify the influence of voice quality on the willingness to buy on the basis of the SOR model (SOqR). As mentioned in the previous section, the relationship between speech stimuli and voice quality has already been revealed (S-Oq) [10]. However, it is unclear whether voice quality affects the willingness to buy (Oq-R). We selected seven adjective pairs that represent voice quality commonly used in previous studies [8, 21] (Table 1). Thus, the paper attempts to verify the hypothesized model (Figure 1) and the following hypotheses:

H1:Voice quality affects the willingness to buy (Oq-R).

H2:Voice quality mediates the influence of advertising speech on the willingness to buy (S-Oq-R).

3. Experimental conditions

3.1. Speech stimuli

To analyze the relationship between voice quality and willingness to buy, it is desirable to use speech with the same silent pause location uttered by many speakers with various speaker characteristics [7]. However, there is no guarantee that the recordings of dozens of speakers will cover such variety. In addition, it would be unrealistic to record speech with the same silent pause location by several hundreds of speakers. To overcome this problem, we have generated speech stimuli using text-to-speech (TTS) synthesis. The quality of TTS has become so high that it is almost indistinguishable from human speech. For the TTS, we adopted FastSpeech2 [22] conditioned by a one-hot speaker ID. The FastSpeech2 consists of 4 and 6 feed-forward Transformer (FFT) blocks in the encoder and the mel-spectrogram decoder, respectively. The frame shift was 10 [ms]. One-hot speaker ID was converted into 256 dimensional embedding vectors via 1 full-connected layer, and fed to the variance adapter. Other parameters were followed by the original paper. The training data was 135,202 utterances by 978 Japanese speakers. At the inference time, a logarithmic mel-spectrogram was generated from input text and a one-hot speaker ID corresponding to the target speaker. Synthetic speech was obtained from the logarithmic mel-spectrogram by using a neural waveform generation method, HiFi-GAN [23].

To achieve variation in the speaker features, average log F0, standard deviation of log F0 (SD F0), and speech rate [6] were extracted for each of 978 Japanese speakers registered in the TTS system. These features were used to cluster the speakers using the k-means++ algorithm [24], and the speaker closest to the centroid of each cluster was selected. The aforementioned three speech features were then extracted from the speech of multiple sentences synthesized for the selected speakers, and the sentences with the maximum within-class variance relative to the between-class variance were selected. Considering the experimental time and subject burden, 13 male (age range 22 to 73 years) and 13 female speakers (age range 18 to 70 years) and 13 advertising sentences (e.g., electrical appliances, insurance, food) were finally selected for the experiments (338 stimuli) [1]. The locations and durations of pauses within the same sentence were consistent across speakers. The sampling frequency of the synthesized speech was 22.05 [kHz]. To calculate speech rate, the duration of each phoneme was obtained by forced alignment using a DNN-HMM acoustic model. We visually checked that there were no fatal alignment errors. The average intensity level of all stimuli was 62 [dB].

Table 1: *Voice qualities used in this study.*

(a) High pitched–Low pitched	(e)Youthful–Elderly
(b) Hoarse–Clear	(f) Bright–Dark
(c) Unstable–Calm	(g) Cold–Warm
(d) Powerful–Weak	

Table 2: *The mean and SD of the speech features of the stimuli.*

	male speech		female speech	
	mean	SD	mean	SD
Speech rate [mora/s]	8.42	0.71	7.92	0.76
mean F0 [Hz]	141.05	21.74	232.04	24.50
SD F0 [Hz]	25.65	5.46	37.75	6.67
CPP	19.05	1.16	20.91	1.49
F4 [Hz]	3621.49	152.97	4209.69	125.77
SHR	0.25	0.08	0.11	0.078

The speech features of the speech stimuli were calculated for use in the analysis. Using the method proposed by Arifianto et al. [25], we calculated the mean F0 and SD F0. The speech rate was calculated using phoneme segmentation information. As Kuang et al. [26] suggested, there is a possibility that other features beyond mean F0, SD F0, and speech rate may also affect the perception of voice quality. We investigated all of the speech features extractable using VoiceSauce (v1.37) [27]. Finally, the following parameters were selected in addition to average F0, SD F0, and speech rate: subharmonic-to-harmonic ratio (SHR), the fourth formant frequency (F4), and cepstral peak prominence (CPP) (Table 2). Since there were significant differences between male and female speech in all speech features, the speech feature and evaluation value obtained by subtracting the average values of each were used.

3.2. Experimental procedure

The experiment was conducted through an online survey using a web browser. The participants were 448 Japanese persons (237 men and 211 women). They were divided into two groups, following the approach of our previous study [6]. 222 persons listened only to male speech (mean age=46.13 years, *SD*=14.86). 226 persons listened only to female speech (mean age=46.07 years, *SD*=14.70). The participants adjusted the volume of their device so that they could listen to the speech easily using the sample speech before the experiment. The participants were not permitted to change the volume during the experiment [28]. The experimental procedure consisted of the following four steps:

(i) Listened to the speech.

(ii) Rated the voice quality of the speech.

(iii) Rated their own perceived emotion.

(iv) Rated the degree of their willingness to buy the advertised product.

Due to the large number of rated items, it is difficult to remember the speech and evaluate all items once listening. However, listening to the speech again for each item also lengthens the duration of the experiment and increases the burden on the participant. Therefore, in the experiment, items (ii), (iii), and (iv) were divided into two groups, and participants listened to the speech before each group. For example, the participants started with (i), followed by (ii) and (iv). Then, they listened to the speech again, followed by (iii). These were counterbalanced. The participants listened to 26 stimuli (13 speakers × 2 sentences) across the experiment.

[1] Samples of synthetic speech and advertising sentences are available at here: https://ntt-hilab-gensp.github.io/is2023-SOR-VQ/

Table 3: *Goodness-of-fit indices for path analysis. GFI, AGFI, CFI, RMSEA, and SRMR respectively indicate goodness-of-fit Index, adjusted GFI, comparative fit index, root mean square error of approximation, and standardized root mean square residual.*

Fit indices	Accepted value	SOeR	SOqR
CFI	≧ 0.9	0.914	0.939
GFI	≧ 0.9	0.976	0.982
AGFI	≧ 0.9	0.914	0.893
RMSEA	≦ 0.1	0.102	0.103
SRMR	≦ 0.1	0.050	0.039

Table 4: *Result of mediation analysis for SOeR. Est. and S.E. indicate estimates and standardized error, respectively.*

Mediator	Est.	S.E.	95 %CI	
			Lower	Upper
Pleasure	2.90	0.26	2.38	3.43
Arousal	0.89	0.10	0.68	1.08
Dominance	-0.07	0.02	-0.05	0.04

The voice quality was rated on a 7-point Likert scale in each of the Japanese expressions associated with voice qualities (Table 1). The participant's emotion was rated in each of the three dimensions on a 7-point Likert scale: [pleasure (pleasant-unpleasant), arousal (calm-excited), and dominance (dominant-submissive)] [6]. Participants were given instructions to make it easier to understand the emotional dimensions (e.g., "Pleasure refers to how good or bad you feel."). The willingness to buy was rated on a 7-point Likert scale (1: not at all willing to buy–7: very willing to buy). Participants were instructed not to think about whether or not to buy the item, as we want them to evaluate their "motivation," and were asked to answer on the basis of how they felt about the narrator's way of speaking instead of the manufacturer or brand. The methods and instructions for the participant's emotion and willingness to buy were designed to be similar to those in our previous study [6]. The experiment lasted about 40 minutes. This experiment was carried out with the approval of the ethic examination of Research Institute of Human Engineering for Quality Life.

3.3. Analysis methods

The influence of voice quality on the willingness to buy on the basis of the SOR model needs to be verified by mediation analysis after confirming that the preconditions are met by multiple regression and path analysis [6]. Path analysis is a method that verifies hypothesized models by analyzing direct and indirect relationships between variables [29]. The goodness of fit of the model to the experimental data is assessed using fit indices.

Mediation analysis was conducted to examine the importance of considering organism in the willingness to buy from speech features. Mediation analysis assumes the mediator (M) that has a potential influence between the independent variable (X) and the dependent variable (Y) when X affects Y [30]. This analysis is to verify how the mediator M contributes. The concept of mediation analysis is generally represented using the following equations:

$$\begin{aligned} Y &= i_1 + aX + e_1 \\ M &= i_2 + bX + e_2 \\ Y &= i_3 + a'X + cM + e_3 \end{aligned}$$

Table 5: *Correlation between voice quality and the willingness to buy.*

Voice quality	Willingness to buy
(a) High pitched–Low pitched	-0.23**
(b) Hoarse–Clear	0.41**
(c) Unstable–Calm	-0.09**
(d) Powerful–Weak	-0.25**
(e)Youthful–Elderly	-0.24**
(f) Bright–Dark	-0.39**
(g) Cold–Warm	0.43**

$^{**}p < .01$

The a coefficient represents the effect of the independent variables (X) on dependent variables (Y), is defined as the total effect. The b coefficient represents the effect of the independent variables (X) on the mediator (M). The a' coefficient represents the effect of the independent variables (X) when adjusted for the mediator (M), is defined as the direct effect. The i_1, i_2, and i_3 represent intercepts, and the e_1, e_2, and e_3 represent the residual error. The mediating effect is defined as $a - a'$ or $b * c$. The speech features were entered as the independent variables, the emotions or voice quality as the mediating variables, and the willingness to buy as the dependent variable. Only paths that were significant in the path analysis were used in the analysis. The statistical significance of the mediating effects was concluded with 1,000 bootstrap samples. The mediating effect is considered significant if zero is not included in the 95 % confidence interval calculated from the bootstrap samples.

4. Results

4.1. Validation of experimental data using an emotion-mediated SOR model

In this study, unlike the previous study [6], we used the synthesized speech of multiple speakers. Before analyzing the influence of voice quality, we confirmed whether the SOR model, which uses emotion as a mediator, is effective as in the previous study. To maintain consistency with the following analysis of voice quality, we included not only mean F0, SD F0, and speech rate but also CPP, F4, and SHR in the speech features. The evaluation value for the emotion and the willingness to buy collected in the subjective evaluation was used.

First, multiple regression analysis has shown that these speech features influence the willingness to buy (mean F0 (β=0.24, p<.05), SD F0 (β=4.73, p<.01), speech rate (β=0.19, p<.01), SHR(β=-0.54, p<.01), F4 (β=0.00048, p<.01) and CPP (β=0.09, p<.01)). Second, path analysis has shown that the goodness-of-fit indices are high enough (Table 3), and all paths were significant except for F0 and SHR on dominance. For want of space, the path coefficients were omitted. Table 4 showed the results of the mediation analysis. The mediating effects of pleasure and arousal were significant, but the mediating effect of dominance was not significant. The ratio of the mediating effect to the total effect [31] was about 79.5 %. This confirms that the SOR model mediated by emotions is effective, even when using synthesized speech stimuli of multiple speakers, as in our experiment, which is consistent with previous studies [6].

4.2. Verification of H1

Multiple regression analysis of the effect of voice quality on willingness to buy was conducted to test hypothesis H1. (e) Youthful–Elderly was found to not significantly affect the willingness to buy, so it was excluded from all analyses ($\beta =$

Table 6: *Standardized path coefficients for SOqR.*

Mediator	S → Oq						Oq → R
	Mean F0	SD F0	Speech rate	SHR	F4	CPP	Willingness to buy
(a) High pitched–Low pitched	-0.263**	-0.114**	-0.056**	0.013	-0.068**	-0.117**	-0.013
(b) Hoarse–Clear	0.011	0.153**	0.105**	-0.122**	0.069**	0.182**	0.232**
(c) Unstable–Calm	0.122**	0.076**	-0.045**	0.041**	-0.039**	-0.003	-0.028**
(d) Powerful–Weak	0.07**	-0.125**	-0.120**	-0.061**	-0.037**	-0.087**	-0.106**
(f) Bright–Dark	-0.154**	-0.159**	-0.137**	-0.001	-0.109**	-0.127**	-0.186**
(g) Cold–Warm	0.068**	0.050**	0.087**	-0.054**	0.072**	0.060**	0.263**

$^{**}p < .01$

Table 7: *Result of mediation analysis for SOqR. Est. and S.E. indicate estimates and standardized error, respectively.*

Mediator	Est.	S.E.	95 %CI	
			Lower	Upper
(b) Hoarse–Clear	0.86	0.10	0.68	1.05
(c) Unstable–Calm	0.04	0.02	0.02	0.06
(d) Powerful–Weak	0.63	0.07	0.49	0.77
(f) Bright-Dark	1.20	0.12	0.98	1.44
(g) Cold–Warm	2.35	0.25	1.86	2.85

$-0.01, p = 0.36$). The remaining six voice qualities significantly affected the willingness to buy ((a) (β=-0.02, p<.05), (b) (β=0.21,p<.01), (c) (β=-0.04,p<.01), (d) (β=-0.13, p<.01), (f) (β=-0.17, p<.01) and (g) (β=0.33, p<.01)). The correlations between these voice qualities and the willingness to buy are shown in Table 5. The most influential voice quality on the willingness to buy was (g) Cold–Warm. (b) Hoarse–Clear and (f) Bright–Dark also had an influence. Thus, H1 was supported.

4.3. Verification of H2

The influence of speech features on the willingness to buy has been confirmed in Section 4.1. We conducted verification of the influence of voice quality on the willingness to buy using path and mediation analyses. The fit indices and path coefficients are shown in Tables 3 and 6, respectively. Consistent with the previous section, the influence of voice quality on the willingness to buy was the highest for (g) Cold–Warm, followed by (b) Hoarse–Clear and (f) Bright–Dark. The speech features that most influenced these voice qualities were the speech rate, CPP, and SD F0, respectively. The fit index suggested that the SOqR model fit the experimental data as well as the SOeR model.

Table 7 showed the results of the mediation analysis. (a) High pitched–Low pitched was excluded because it had no significant effect on the willingness to buy in the path analysis. The mediating effects of all voice qualities were significant. The ratio of the mediating effect to the total effect was about 90.2 %, which is larger than that of emotion. Thus, H2 was supported.

5. Discussion

This study verified the influence of voice quality on the willingness to buy on the basis of the SOR model. We generated synthesized speech for the advertising speech of multiple speakers for various products. A large-scale experiment was conducted using an online survey to evaluate the voice quality, emotions, and willingness to buy. Before analyzing the influence of voice quality, we verified whether the SOR model was effective, as in previous studies, using the evaluation scores of emotions. We confirmed that there were no issues with using synthesized speech for this study. We verified the impact of voice quality on the willingness to buy using multiple regression, path, and mediation analyses on the basis of the SOR model.

The results of the multiple regression analysis showed that (e) Youthful–Elderly did not affect the willingness to buy. While we have previously shown that the age of the listener moderates the mediating effect of emotion [6], this result suggests that the age of the speaker has no effect on willingness to buy, but further investigation is needed in the future with a moderated mediation analysis. The results of the path analysis indicated that (g) Cold–Warm, (f) Bright–Dark, and (b) Hoarse–Clear had an impact on the willingness to buy. A number of previous studies based on the SOR model verified the satisfaction of shopping experience or service quality as organism, in addition to emotion [14, 17]. These relate to judgments about the external existence such as a store's atmosphere. It is likely that the brightness and warmth of the advertising speech in this study function similarly to these. In other words, these voice qualities may function as judgments of the quality or desirability of a clerk. The results for (b) Hoarse–Clear in this study were consistent with Zoghaib et al.'s findings that smooth voices are effective in persuasion [4]. The influence of (a) High pitched–Low pitched on the willingness to buy was not significant. This may be due to the fact that the participants in this experiment evaluated speech for the same gender. We also tested the effects of other speech features that can be extracted from VoiceSauce on voice quality, but these effects were low, resulting in a high effect of SD F0 and speech rate.

The results of the mediation analysis indicated that voice quality can explain the effect of the advertising speech on the willingness to buy to a similar or better extent as emotion. In real advertising situations, it may be easier to judge a speaker's voice quality than listeners' perceived emotions. The findings obtained in this study are considered useful for training salespeople to speak in a way that increases the willingness to buy.

6. Conclusion

This study aimed to verify the influence of voice quality on the willingness to buy on the basis of the SOR model. We generated synthesized speech for advertising speech of multiple speakers for various products. A large-scale experiment was conducted using an online survey to evaluate the voice quality, emotions, and willingness to buy. The multiple regression and path analysis results showed that the warmth, brightness, and clarity of voice particularly influenced the willingness to buy. The youthfulness of voice had no significant influence. The mediation analysis suggested that voice quality can explain the effect of advertising speech on willingness to buy to a similar or better extent than the emotion-mediated SOR model. In real advertising situations, it may be easier to judge a speaker's voice quality than listeners' perceived emotions. The findings obtained in this study could help train salespeople in advertising speech.

7. References

[1] C. Nass, I.-M. Jonsson, H. Harris, B. Reaves, J. Endo, S. Brave, and L. Takayama, "Improving automotive safety by pairing driver emotion and car voice emotion," in *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, 2005, pp. 1973–1976.

[2] A. Chattopadhyay, D. W. Dahl, R. J. Ritchie, and K. N. Shahin, "Hearing voices: The impact of announcer speech characteristics on consumer response to broadcast advertising," *Journal of Consumer Psychology*, vol. 13, no. 3, pp. 198–204, 2003.

[3] M. Kobayashi, Y. Hamada, and M. Akagi, "Acoustic features correlated to perceived urgency in evacuation announcements," *Speech Communication*, vol. 139, pp. 22–34, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639322000346

[4] A. Zoghaib, "Persuasion of voices: The effects of a speaker' s voice characteristics and gender on consumers' responses," *Recherche et Applications en Marketing (English Edition)*, vol. 34, no. 3, pp. 83–110, 2019. [Online]. Available: https://doi.org/10.1177/2051570719828687

[5] A. Mehrabian and J. A. Russell, *Approach to environmental psychology*. The MIT Press, 1974.

[6] M. Nagano, Y. Ijima, and S. Hiroya, "Perceived emotional states mediate willingness to buy from advertising speech," *Frontiers in Psychology*, vol. 13, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.1014921

[7] M. Poon, K. Chan, and E. Yiu, "The relationship between speech rate, voice quality and listeners' purchase intentions," in *Proceedings of the 9th International Conference on Speech Prosody*, 2018, pp. 468–472.

[8] C.-F. Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50, no. 10, pp. 810–828, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639308000782

[9] C. Gobl and A. Ní Chasaide, "Testing affective correlates of voice quality through analysis and resynthesis," in *Proc. ITRW on Speech and Emotion*, 2000, pp. 178–183.

[10] R. Elbarougy and M. Akagi, "Improving speech emotion dimensions estimation using a three-layer model of human perception," *Acoustical Science and Technology*, vol. 35, no. 2, pp. 86–98, 2014.

[11] H. J. D. Wiener and T. L. Chartrand, "The effect of voice quality on ad efficacy," *Psychology & Marketing*, vol. 31, no. 7, pp. 509–517, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mar.20712

[12] M. Simanjuntak, H. R. Nur, B. Sartono, and M. F. Sabri, "A general structural equation model of the emotions and repurchase intention in modern retail," *Management Science Letters*, vol. 10, no. 4, pp. 801–814, 2020.

[13] J. Y. Jang, E. Baek, S. Y. Yoon, and H. J. Choo, "Store design: Visual complexity and consumer responses," *International Journal of Design*, vol. 12, no. 2, pp. 105–118, 2018.

[14] N. Nusairat, Q. Hammouri, H. Al-Ghadir, A. M. K. Ahmad, and M. A. H. Eid, "The effect of design of restaurant on customer behavioral intentions," *Management Science Letters*, vol. 10, no. 9, pp. 1929–1938, 2020.

[15] I. Anninou, G. Stavraki, and Y. Yu, "Cultural differences on perceived crowding , shopping stress and excitement in superstores," in *Proceedings of the 51th Academy of Marketing Conference*, 2018, pp. 1–13.

[16] R. J. Donovan, J. R. Rossiter, G. Marcoolyn, and A. Nesdale, "Store atmosphere and purchasing behavior," *Journal of Retailing*, vol. 70, no. 3, pp. 283–294, 1994.

[17] H. Roschk, S. M. C. Loureiro, and J. Breitsohl, "Calibrating 30 years of experimental research: A meta-analysis of the atmospheric effects of music, scent, and color," *Journal of Retailing*, vol. 93, no. 2, pp. 228–240, 2017.

[18] A. Anwar, A. Waqas, H. M. Zain, and D. M. H. Kee, "Impact of music and colour on customers' emotional states: An experimental study of online store," *Asian Journal of Business Research*, vol. 10, no. 1, pp. 104–125, 2020.

[19] Ronald E. Milliman, "Using background affect to music behavior of the supermarket shoppers," *The Journal of Marketing*, vol. 46, no. 3, pp. 86–91, 1982.

[20] R. Donovan and J. Rossiter, "Store atmosphere: an environmental psychology approach," *Journal of retailing*, vol. 58, no. 1, pp. 34–57, 1982.

[21] H. Kido and H. Kasuya, "Extraction of everyday expression associated with voice quality of normal utterance," *Journal of the Acoustical Society of Japan*, vol. 55, no. 6, pp. 405–411, 1999 (in Japanese).

[22] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

[23] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 17 022–17 033, 2020.

[24] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," Stanford, Tech. Rep., 2006.

[25] D. Arifianto, T. Tanaka, T. Masuko, and T. Kobayashi, "Robust f_0 estimation of speech signal using harmonicity measure based on instantaneous frequency," *IEICE TRANSACTIONS on Information and Systems*, vol. E87-D, no. 12, pp. 2812–2820, 2004.

[26] J. Kuang and M. Liberman, "Integrating voice quality cues in the pitch perception of speech and non-speech utterances," *Frontiers in Psychology*, vol. 9, 2018. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02147

[27] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for voice analysis," in *17th International Congress of Phonetic Sciences*, 2011, pp. 1846–1849.

[28] M. Cooke and M. L. García Lecumberri, "How reliable are online speech intelligibility studies with known listener cohorts?" *The Journal of the Acoustical Society of America*, vol. 150, no. 2, pp. 1390–1401, 2021.

[29] O. D. Duncan, "Path analysis: sociological examples," *American Journal of Sociology*, vol. 72, pp. 1 – 16, 1966.

[30] J. J. Rijnhart, M. J. Valente, D. P. MacKinnon, J. W. Twisk, and M. W. Heymans, "The use of traditional and causal estimators for mediation models with a binary outcome and exposure-mediator interaction," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 28, no. 3, pp. 345–355, 2021.

[31] K. Preacher and K. Kelley, "Effect size measures for mediation models: Quantitative strategies for communicating indirect effects," *Psychological methods*, vol. 16, pp. 93–115, 2011.