INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Impact of Emotional State on Estimation of Willingness to Buy from Advertising Speech

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

[1]NTT Media Intelligence Laboratories, Japan
[2] NTT Communication Science Laboratories, Japan
mizuki.nagano.az@hco.ntt.co.jp

Abstract

The characteristics of a speaker's voice can affect the perceived impression or behavior of the listener. Previous studies of consumer behavior have shown that this can be well explained by the emotion-mediated behavior model. However, few studies of the emotion-mediated behavior model have used advertising speech. In this paper, we examine whether the stimulus-organism-response theory using emotional state can explain willingness to buy from advertising speech stimulus. The subjects listened to speech with modified speech features (mean F0, speech rate, spectral tilt, or standard deviation of F0) and rated their willingness to buy the products advertised in the speech and their own perceived emotions (pleasure, arousal, dominance). We found that the emotions partially mediate the influence of speech features on the willingness to buy. These results will be useful for developing a method of speech synthesis to increase people's willingness to buy.

Index Terms: the willingness to buy, advertising speech, emotion

1. Introduction

The characteristics of a speaker's voice are known to affect the perceived impression or behavior of the listener. For example, men with lower-pitched voices tend to be trusted more [1], and candidates for political office with lower-pitched voices are more likely to be elected [2]. However, voice characteristics do not simply determine perceived impressions, e.g., we have different impressions even of men with higher-pitched voices depending on the situation [3,4]. This is because high-level processing in the brain is involved in perceiving impressions from speech [5].

In studies of consumer behavior, an external stimulus, such as a congestion situation and background music, affects consumer behavior and willingness to buy [6]. To better explain a response from a stimulus, a hierarchical model based on the stimulus-organism-response (S-O-R) theory has been proposed [6] (Figure 1). The PAD (pleasure, arousal, dominance) emotional state model is often used as the organism [7]. As the stimulus, congestion situation and background music have been used [8, 9]. However, few studies of the S-O-R theory with or without the PAD model have been conducted using speech as a stimulus. Poon *et al.* investigated the influence of differences in pause length and voice quality on willingness to buy through a perceived personality state, not the PAD model. However, they did not examine speech features that may affect emotions or show statistical evidence for the importance of the organism state as is often done in consumer behavior study [10].

In this study, we examine whether the S-O-R theory using the PAD model can explain willingness to buy from speech stimulus of electric appliance advertisements spoken by a professional narrator. To verify the effect of the difference in the speech features, speech was generated with the converted static and dynamic features of F0, speech rate, and spectral tilt. These parameters have been used to study speech features and emotions [11]. Two large-scale subjective evaluations were conducted via crowdsourcing. Subjects were asked to evaluate the emotion on PAD dimensions and the willingness to buy when listening to advertising speech. We conducted analysis of variance (ANOVA), path analysis, and mediation analyses using the obtained data of speech features of speech stimuli, emotions, and the willingness to buy.

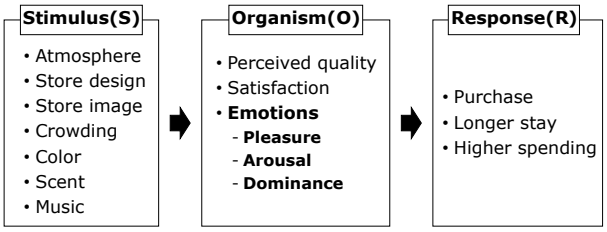

Figure 1: *S-O-R theory.*

2. Stimulus-Organism-Response (S-O-R) theory

The S-O-R theory [6] comprises three dimensions: stimulus (S), organism (O), and response (R) (Figure 1). Stimulus incorporate all external environmental factors in the store such as atmosphere [12], design [13, 14], brand image [15], crowding [8], color, scent, and music [16]. These stimuli stimulate various consumer responses. In previous studies, the responses were investigated by using the perceived quality [14], satisfaction [16], and emotions [17]. These consumer responses result in approach or avoidance behavior. Approach behavior refers to a positive attitude toward the environment such as staying in a place. In contrast, avoidance behavior refers to a negative attitude toward the environment such as escaping from a place. In consumer behavior research, approach or avoidance behavior is confirmed by indicators such as purchase, longer stay time [13] and higher spending [18].

The PAD model specifically focuses on the emotion as the organism (O) in S-O-R theory. There are three emotional states: pleasure, arousal, and dominance (PAD) [7]. Pleasure refers to the degree of feeling joy, satisfaction, and happiness with a situation. Arousal refers to the degree of feeling excited, passionate, and active about the situation. Dominance refers to the degree of feeling that an individual has influence over a situation and can control it. The PAD model has been verified with various external stimuli, and the evidence from previous studies has shown that this model reflects the relations between these stimuli, consumers' emotions, and consumers' responses. It is

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-827

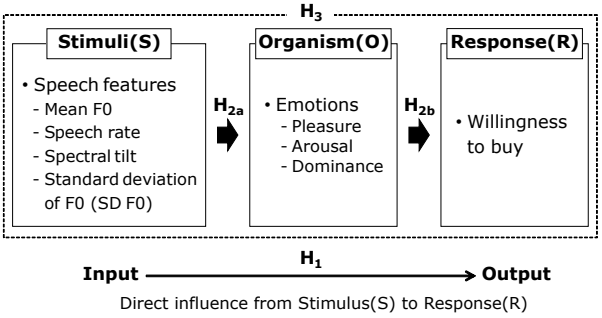

Figure 2: *Conceptual model in this study.*

expected that the PAD model can also explain the willingness to buy caused by speech stimuli as in previous studies.

Thus, the paper attempts to verify the conceptual model (Figure 2) and the following four hypotheses:

H_1: Speech features will directly influence the willingness to buy (S-R).

H_{2a}: Speech features will influence emotions (S-O).

H_{2b}: Emotions will influence the willingness to buy (O-R).

H_3: The S-O-R theory that is mediated by emotion will be able to explain the willingness to buy caused by speech stimuli (S-O-R).

3. Speech material

Four sentences of neutral speech stimuli spoken by a Japanese female professional narrator were used. Each sentence was intended to promote electrical appliances. The sampling frequency of the recorded speech was 22.05 [kHz]. The average sentence duration, F0, and speech rate were 14.87 [s], 255.53 [Hz], and 7.02 [mora/s]. Each phoneme boundary was manually segmented.

To verify the influence of speech features on emotion and the willingness to buy, we manipulated mean F0 [Hz], speech rate [mora/s], spectral tilt, and standard deviation of F0 (SD F0) of the original speech. Mean F0 and speech rate were manipulated in both Studies 1 and 2. Mean F0 was converted by a factor of 0.9439 (low) or 1.0595 (high) to the average F0 of the sentence using WORLD [19], and speech rate was converted by a factor of 1.1225 (slow) or 0.8909 (fast) by a pointer interval controlled overlap and add (PICOLA) [20]. The conversion rate for mean F0 was determined on the logarithmic axis to take into account human auditory characteristics. In Study 1, the spectral tilt was transformed by performing de-emphasis (steep) and pre-emphasis (flat) with $\alpha = 0.2$ [21]. In Study 2, SD F0 was converted by a factor of 1.5 (more variation) or 0.6667 (less variation) [22]. These parameters were determined by preliminary experiments. To reduce the influence of sound quality deterioration due to speech modification mentioned above, all stimuli used in the subjective evaluation were analyzed and synthesized using WORLD: a process of analysis-by-synthesis by WORLD was also performed even for original speech that was used as a baseline. As a result, we used 27 types of speech stimuli per sentence in the combination of mean F0, speech rate, and spectral tilt in Study 1 and mean F0, speech rate, and SD F0 in Study 2. The average intensity level of all stimuli was 62 dB.

4. Subjective evaluation and analysis

4.1. Study 1

The subjective evaluation about purchase intension and emotions was examined with a 3 (mean F0: low vs. neutral vs. high) × 3 (speech rate: slow vs. neutral vs. fast) × 3 (spectral tilt: steep vs. neutral vs. flat) within-subjects experimental design using speech stimuli as described in the previous section. The subjects were 201 Japanese persons (119 males and 82 females; mean age = 37.91 years, *SD* = 14.37), who were gathered via crowdsourcing. Two of the four sentences (54 stimuli) were selected for each subject, taking counter balance into account. The subjects listened to the speech stimuli only once, rated the degree of their willingness to buy the advertised product, and then rated their own perceived emotions in the browser. The order of evaluated items was designed on the basis of previous study [18]. The willingness-to-buy response was removed from the display when the subjects answered about their emotions.

The willingness to buy was rated on a 7-point Likert scale (1: not at all willing to buy-7: very willing to buy). Subjects were given the following instructions: "You don't really have to think about whether or not to buy it because we want you to evaluate "motivation". Please answer not how you feel about the manufacturer or brand, but how you feel about the narrator's way of speaking." The subject's emotion was rated in each of three dimensions on a 7-point Likert scale: [pleasure (pleasant-unpleasant), arousal (calm-excited), and dominance (dominant-submissive)]. Subjects were given the instructions used in a previous study [23] to make it easier to understand the emotional dimensions (e.g. "Pleasure refers to how good or bad you feel."). The experiment lasted about 30 minutes.

We verified the hypotheses by conducting three analyses: ANOVA for the validation of H_1, path analysis for H_{2a} and H_{2b}, and mediation analysis for H_3. The mean F0 [Hz] and speech rate [mora/s] in the speech interval and the first-order LPC coefficient as spectral tilt in the vowel interval of each stimulus were used as speech features. The evaluation value for the emotion and that for willingness to buy collected in the subjective evaluation experiment were used. Nine subjects who had incomplete data were excluded from the subsequent analysis.

4.1.1. A three-way ANOVA

To analyze the direct influence of speech features on the willingness to buy, a three-way ANOVA was conducted. There were main effects of mean F0 ($F(2, 10366) = 45.408, p<.01$) and speech rate ($F(2, 10366) = 1340.981, p<.01$), but no main effect of spectral tilt ($F(2, 10366) = 2.108, p = .122$). There were no first-order interaction effects between mean F0 and speech rate ($F(4, 10364) = .649, p = .628$), between mean F0 and spectral tilt ($F(4, 10364) = .212, p = .932$), or between the speech rate and spectral tilt ($F(4, 10364) = 1.011, p = .400$). There was also no second-order interaction effect ($F(8, 10360) = .208, p = .990$). These results indicate that speech features influence the willingness to buy: H_1 is supported. High mean F0 or fast speech rate tended to increase the willingness to buy.

4.1.2. Path analysis

Path analysis [24] was conducted to examine direct dependencies between speech features and emotions and between emotions and willingness to buy. Figure 3 shows a path diagram with a path coefficient. Some fit indices in Table 1 indicate that the model is valid. As shown in Figure 3, all speech features had a significant positive effect on all dimensions of the emotion, except for a minor difference between spectral tilt and pleasure. Thus, H_{2a} was supported. The speech rate had a greater path coefficient than other speech features. Pleasure and arousal had

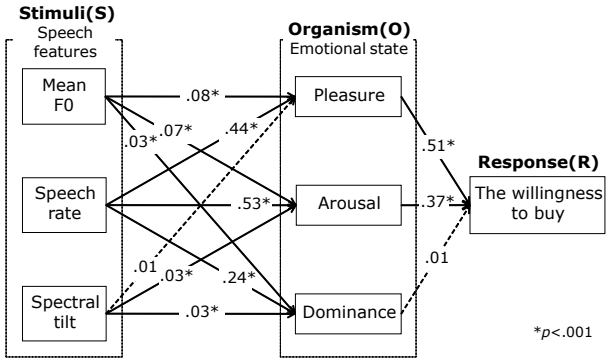

Figure 3: *Path analysis of conceptual model for Study 1.*

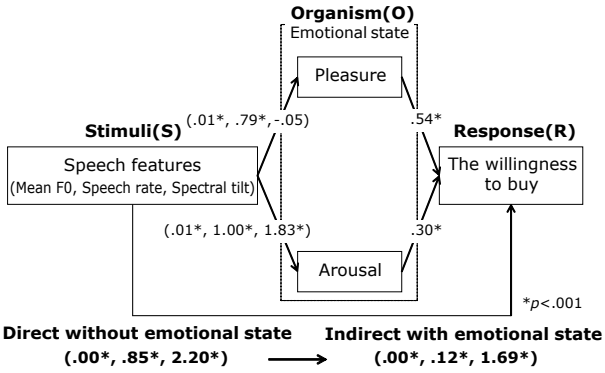

Figure 4: *Mediation analysis for Study 1.*

Table 1: *Fit indices for path analysis. GFI, AGFI, CFI, and RMSEA respectively indicate Goodness of Fit Index, Adjusted GFI, Comparative Fit Index, and Root Mean Square Error of Approximation.*

Fit indices	Accepted value	Study 1	Study 2
Absolute fit measures			
χ^2 (Chi-square)		15.019	43.026
df		3	3
GFI	>0.9	1.000	0.999
RMSEA	<0.10	0.045	0.037
Incremental fit measures			
AGFI	>0.9	0.998	0.995

a significant positive effect on the willingness to buy, but dominance did not. Therefore, H_{2b} was supported for pleasure and arousal.

4.1.3. Mediation analysis

ANOVA and path analysis are insufficient to verify the importance of considering emotional states in order to estimate willingness to buy from speech features. Thus, the mediation effect of emotions on the willingness to buy was verified in the mediation analysis. The mediation analysis assumes the mediating variable M that has a potential influence between the independent variable X and the dependent variable Y when X affects Y. This analysis is to determine the existence and the influence of M. According to previous study [14], the mediation effects of emotions can be examined by assessing path coefficients with and without mediators in the model. In this study, the speech features were entered as the independent variables, the emotion as the mediating variables, and the willingness to buy as the dependent variable.

Figure 4 shows the results of mediation analysis. The statistical significance of mediation effects was concluded with 1,000 bootstrap samples. A significant correlation was observed between speech features and emotions and between emotions and willingness to buy, similar to path analysis. Let the partial regression coefficient of S-R in the S-R path be direct without a mediator and that in the S-O-R + S-R path be indirect with a mediator. The coefficient of direct without a mediator was significant. The coefficient of indirect with a mediator was also significant. The smaller value of S-R in indirect with a mediator than direct without a mediator indicates that the S-O-R path was valid, which suggests that pleasure and arousal mediate the influence of speech features on the willingness to buy. Thus, H_3 was supported.

4.2. Study 2

In Study 1, we found that the mean F0 affected the willingness to buy, but in Study 2, we investigated whether the variation of F0 in a sentence affects the willingness to buy. Spectral tilt, which did not affect it in Study 1, was excluded. The standard deviation of F0 [Hz] in a sentence was used as SD F0 [25]. Study 2 was examined with a 3 (mean F0: low vs. neutral vs. high) × 3 (speech rate: slow vs. neutral vs. fast) × 3 (SD F0: large vs. neutral vs. small) within-subjects experimental design. The subjects were 202 Japanese persons (127 males and 75 females; mean age =38.76 years, *SD* = 13.86), who were gathered via crowdsourcing. The experimental procedure and analysis method were the same as in Study 1. Twenty-one subjects who had incomplete data were excluded from the subsequent analysis.

4.2.1. A three-way ANOVA

There were main effects of mean F0 ($F(2, 9772) = 52.811, p<.01$), speech rate ($F(2, 9772) = 970.240, p<.01$), and SD F0 ($F(2, 9772) = 399.209, p<.01$). The first-order interaction effects were observed between speech rate and SD F0 ($F(4, 9770) = 6.907, p<.01$), but not between mean F0 and speech rate($F(4, 9770) = 1.522, p = .193$) or between mean F0 and SD F0 ($F(4, 9770) = 2.339, p = .0529$). The significant first-order interaction effect was analyzed for simple main effects. The result showed that each factor was significant. The faster speech rate enhanced the willingness to buy in all SD F0 conditions (in less condition: $F(2, 3255) = 361.1, p<.01$; in neutral condition: $F(2, 3255) = 361.8, p<.01$; in more condition: $F(2, 3255) = 253.5, p<.01$). The more SD F0 enhanced the willingness to buy in all speech rate conditions (in slow condition: $F(2, 3255) = 130.6, p<.01$; in neutral condition: $F(2, 3255) = 193.6, p<.01$; in fast condition: $F(2, 3255) = 85.58, p<.01$). The second-order interaction was not observed ($F(8, 9766) = 0.859, p = .551$). The result revealed mean F0, speech rate, and SD F0 influence the willingness to buy. Thus, H_1 was supported. High mean F0, fast speech rate, or large SD F0 tended to increase the willingness to buy.

4.2.2. Path analysis

Path analysis was not designed to take interaction into account because of being based on previous studies [13, 14]. As shown in Figure 5, all speech features had a significant positive effect on all dimensions of the emotions. Some fit indices in the Table 1 indicate that the model is valid. Thus, H_{2a} was supported. Like Study 1, the path coefficient in speech rate was the largest, and that in SD F0 was larger than mean F0. As for the

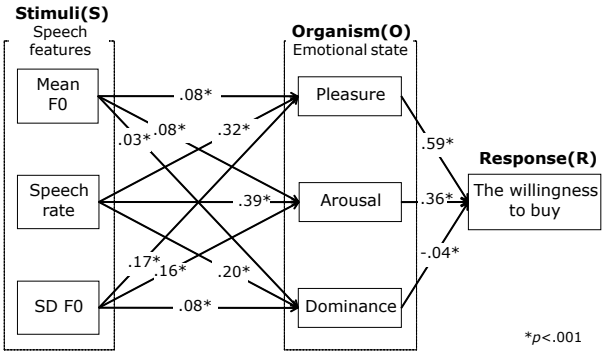

Figure 5: *Path analysis of conceptual model for Study 2.*

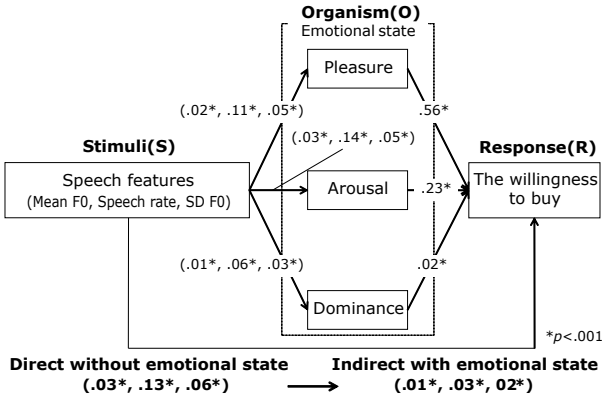

Figure 6: *Mediation analysis for Study 2.*

effect of the emotions on the willingness to buy, all dimensions of emotions including dominance were significant. Therefore, H_{2b} was supported.

4.2.3. Mediation analysis

The results of the mediation analysis are shown in Figure 6. The mediation analysis revealed a mediation effect of all emotions including dominance between the speech features and the willingness to buy. The results indicate that pleasure, arousal, and dominance mediate the influence of speech features on the willingness to buy and that the conceptual emotion-mediated model is valid. Therefore, H_3 was supported.

5. Discussion

From the results of ANOVA, we found that speech features of mean F0, speech rate, and SD F0 affected the willingness to buy (H_1). Path analysis confirmed that there was a dependency between speech features and emotions and between emotions and willingness to buy like in previous studies [17, 26] (H_{2a},H_{2b}). Through mediation analysis, we found that there was a mediation effect of emotions between speech features and willingness to buy (H_3). These results suggested that the S-O-R theory using the PAD model can explain willingness to buy from advertising speech.

Although few studies have examined the relationship between speech rate and willingness to buy, Peterson *et al.* have examined the relationship between mean F0 and SD F0 of salespeople and sales performance [25]. They showed that there was no correlation between these features and performance, which is different from our result, but we speculate that this is due to the difference in experimental methods between speech conversion in this study and speech analysis of multiple salespeople in [25].

Studies 1 and 2 differed in the effectiveness of the path from dominance to willingness to buy. The significance of this path is still controversial, with many studies indicating there is no significant difference [7, 12], and others indicating there is [27]. However, our results suggest that all the speech features in this experiment influenced pleasure and arousal more than dominance, consistent with previous studies that pleasure and arousal have the most significant effect. We speculate that the significance of this path may be due to the interaction between speech features, but similar experiments need to be conducted for other speech features to confirm the effects on emotions in more detail.

The results of ANOVA were consistent with previous results of research into speech that show mean F0, speech rate, and SD F0 affect behavior [28, 29] but not the spectral tilt. This is because the results of the path analysis in Study 1 showed that spectral tilt was correlated with dominance, but dominance did not affect the willingness to buy as in previous studies [7, 12]. In addition, since the experiment was conducted via crowdsourcing, the listening environment of the subjects was not controlled, so the responses of the spectral tilt, which are easily affected by the playback device and headphones, may have differed among the subjects. To reveal the effect of speech on behavior, it is necessary to examine the effects of other speech features not verified in this study and control the listening environment.

The results of the mediation analysis showed that emotions had a partial mediation effect because the significance of the direct effect of the speech features was observed. If the mediation effect of the emotions was sufficient as a mediator, the significance of the direct effect should be lost. Thus, there are other mediators besides emotions in the effect of the speech features on the willingness to buy. To identify other factors, it may be effective to subdivide the relationship between the speech features and emotions using impressions of speech [5]. In addition, auditory characteristics change with age [30]. The influence of individual factors such as gender or age should also be verified.

6. Conclusions

In this paper, we analyzed the relationship between speech features, emotions, and the willingness to buy on the basis of a consumer behavior model. Two large-scale subjective evaluation data were collected via crowdsourcing. The subjects listened to speech with different speech features (mean F0, speech rate, spectral tilt, or SD F0) and rated their willingness to buy the products advertised in the speech and their own perceived emotions (pleasure, arousal, dominance). The result of a three-way ANOVA showed that the speech features affect the willingness to buy. As the result of path analysis and mediation analysis, the emotions were revealed to function as a partial mediator in the influence of speech features on the willingness to buy, and the emotion-mediated model can be used. In particular, increasing pleasure and arousal can be expected to enhance the willingness to buy. It was suggested that the emotion-mediated model may be applied to other behaviors besides the willingness to buy. In future work, we will investigate the effect of other speech features, other mediators, and individual factors such as gender or age.

7. References

[1] M. S. Tsantani, P. Belin, H. M. Paterson, and P. McAleer, "Low vocal pitch preference drives first impressions irrespective of context in male voices but not in female voices," *Perception*, vol. 45, no. 8, pp. 946–963, 2016.

[2] C. A. Klofstad, "Candidate voice pitch influences election outcomes," *Political Psychology*, vol. 37, no. 5, pp. 725–738, 2016.

[3] K. J. Montano, C. C. Tigue, S. G. Isenstein, P. Barclay, and D. R. Feinberg, "Men's voice pitch influences women's trusting behavior," *Evolution and Human Behavior*, vol. 38, no. 3, pp. 293–297, 2017.

[4] J. J. M. O'Connor and P. Barclay, "High voice pitch mitigates the aversiveness of antisocial cues in men' s speech," *British Journal of Psychology*, vol. 109, no. 4, pp. 812–829, 2018.

[5] X. Li and M. Akagi, "A three-layer emotion perception model for valence and arousal-based detection from multilingual speech," in *Proceedings of Interspeech 2018*, vol. 2, no. 6, 2018, pp. 3643–3647.

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

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

[8] 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.

[9] T. Sunaga, "How the sound frequency of background music influences consumers' perceptions and decision making," *Psychology and Marketing*, vol. 35, no. 4, pp. 253–267, 2018.

[10] 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.

[11] M. Schröder, "Emotional speech synthesis: a review." in *Proceedings of the 7th European Conference on Speech*, 2001, pp. 561–564.

[12] 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.

[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] 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.

[16] 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.

[17] 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.

[18] 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.

[19] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[20] N. Morita and F. Itakura, "Time-scale modification algorithm for speech by use of pointer interval overlap and add (picola) and its evaluation," in *Proceedings of Annual Meeting of Acoustical Society of Japan*, vol. 86, 1986, pp. 9–16.

[21] F. Fauziya and G. Nijhawan, "A comparative study of phoneme recognition using GMM-HMM and ANN based acoustic modeling," *International Journal of Computer Applications*, vol. 98, no. 6, pp. 12–16, 2014.

[22] I. Fukuoka, H. Takatsu, S. Fujie, Y. Hayashi, and T. Kobayashi, "Prosodic analysis for utterance sequence in spoken dialogue based information," in *Proceedings of the 30th Annual Conference of the Japanese Society for Artificial Intelligence*, 2016, pp. 1–4.

[23] H. Mori, H. Aizawa, and H. Kasuya, "Consistency and agreement of paralinguisticinformation annotation for conversational speech," *The Journal of the Acoustical Society of Japan*, vol. 61, no. 12, pp. 690–697, 2005.

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

[25] R. Peterson, M. Cannito, and S. Brown, "An exploratory investigation of voice characteristics and selling effectiveness," *Journal of Personal Selling and Sales Management*, vol. 15, no. 1, pp. 1–15, 1995.

[26] A. Tursunov, S. Kwon, and H. S. Pang, "Discriminating emotions in the valence dimension from speech using timbre features," *Applied Sciences*, vol. 9, no. 12, pp. 1–18, 2019.

[27] R. F. Yalcha and E. R. Spangenbergb, "The effects of music in a retail setting on real and perceived shopping times," *Journal of Business Research*, vol. 49, pp. 139–147, 2000.

[28] A. De Waele, A. S. Claeys, and V. Cauberghe, "The organizational voice: The importance of voice pitch and speech rate in organizational crisis communication," *Communication Research*, vol. 46, no. 7, pp. 1026–1049, 2019.

[29] A. Rosenberg and J. Hirschberg, "Charisma perception from text and speech," *Speech Communication*, vol. 51, no. 7, pp. 640–655, 2009.

[30] C. Füllgrabe, B. C. J. Moore, and M. A. Stone, "Age-group differences in speech identification despite matched audiometrically normal hearing: contributions from auditory temporal processing and cognition," *Frontiers in Aging Neuroscience*, vol. 6, p. 347, 2015.