Supporting Interpersonal Emotion Regulation of Call Center Workers via Customer Voice Modulation

Duri Lee*
Korea Advanced Institute of Science & Technology (KAIST)
School of Computing
Korea
durilee@kaist.ac.kr

Kyungmin Nam*
Delft University of Technology (TU Delft)
Computer Science and Engineering
Netherlands
K.Nam-2@student.tudelft.nl

Uichin Lee
Korea Advanced Institute of Science & Technology (KAIST)
School of Computing
Korea
uclee@kaist.ac.kr

ABSTRACT

Call center workers suffer from the aggressive voices of customers. In this study, we explore the possibility of proactive voice modulation or style transfer, in which a customer's voice can be modified in real time to mitigate emotional contagion. As a preliminary study, we conducted an interview with call center workers and performed a scenario-based user study to evaluate the effects of voice modulation on perceived stress and emotion. We transformed the customer's voice by modulating its pitch and found its potential value for designing a user interface for proactive voice modulation. We provide new insights into interface design for proactively supporting call center workers during emotionally stressful conversations.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

interpersonal emotion regulation, user interface design, customer voice modulation with emotion detection, emotion contagion

ACM Reference Format:
Duri Lee, Kyungmin Nam, and Uichin Lee. 2024. Supporting Interpersonal Emotion Regulation of Call Center Workers via Customer Voice Modulation. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3613905.3650968

1 INTRODUCTION

Call center workers act as the face of their organizations, often bearing the brunt of emotional interactions with customers, possibly owing to negative emotional contagion. Their primary role is to resolve customer issues and create a positive experience, a task that demands substantial emotional effort, particularly in the face of negative customer behavior [32]. Regulating emotions to follow the requirements of their roles can significantly drain their mental resources, affecting their personal well-being and organizational performance [5].

Therefore, studies have been conducted to develop tools to support customer service workers who constantly perform 'interpersonal emotion regulation' with the aim of influencing the emotions of customers through their interactions for customer satisfaction. Existing tools include systems for the continuous emotional assessment of customers during interactions, such as identifying angry customers [26] and visualizing their emotions [14]. Although these tools offer indirect relief, our study aims to explore a proactive approach in which the voice of a customer is modulated in real time.

This study examined the need for and the effectiveness of voice modulation technology in alleviating stress and emotional contagion among call center employees. Considering that most customer service interactions are computer mediated, as in call centers, we hypothesized that communication channels can act as buffers in emotional exchanges. We focused on a call center scenario, in which customers call when facing an issue, and explored the mitigation of emotional contagion by modifying communication channels. For instance, automatically adjusting voice characteristics to soften aggressive expressions or altering customer voices to reduce emotional immersion may be convincing strategies. It could be particularly relevant given that vocal communication remains a primary medium in human interaction [10].

While previous studies have focused on systems that identify angry customers through speech emotion recognition [14, 27], we propose to move one step further and develop an interface that automatically modulates the voice of the customer to protect call center workers. We conducted a formative study by interviewing call center workers to understand their needs and concerns regarding voice modulation during conversations. Based on this, as a preliminary work, we conducted experiments in which we modulated the pitch of the voice to evaluate the user experience and its effectiveness. Our proactive approach suggests the potential value of developing automatic voice modulation interfaces based on speech recognition.

2 BACKGROUND AND RELATED WORKS

2.1 Emotional contagion and Emotional labor

Emotional contagion refers to the automatic imitation of the behavior of another person, such as the tone of voice and facial expressions, to evoke similar emotions [13]. It is an important concept in interpersonal emotion regulation, particularly for individuals in the service industry [4]. They should convey positive emotions to customers intentionally [29]. Even when workers are exposed to

*Both authors contributed equally to the paper

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650968

negative customer behavior [23], they should constantly strive to separate themselves from customers to prevent emotional contagion. This process impacts their workload as it requires workers to engage in mental processes to regulate their emotions, regardless of their actual emotional state.

Emotional workload occurs when workers capture the feelings of customers, consciously recognize emotional contagion, and deliberately express positive emotion [17]. This is known as emotional labor and has been noted as a unique type of work performed in the workplace. This is a self-emotion control process in which workers regulate their emotions even after emotional events to follow the required performance in front of customers [9, 11]. Psychological studies have highlighted the study of emotional labor mechanisms and their effects on the health and job satisfaction of workers in occupational groups of service work [20]. According to studies, the workload from emotional labor is significantly related to stress when emotional dissonance occurs [18]. As excessive emotional labor is known to be highly correlated with health problems (e.g., cardiovascular diseases) [3, 19], protecting workers from excessive stress caused by emotional workload is necessary. Considering the impact of stress on health, a two-pronged approach is needed to mitigate the stress of emotional labor. This can be achieved by either reducing the demand for positive emotional expressions or preventing the contagion of negative emotions. Our study focused on preventing negative emotional contagion by controlling the medium through which workers communicate with customers.

2.2 Speech Emotion Recognition in HCI

Speech inherently carries rich emotional information, both consciously and unconsciously [16, 30]. For instance, pitch variations and mel-frequency cepstral coefficients (MFCCs) can signify emotions [12], and pitch, timing, and voice quality in speech can indicate arousal [6]. The accurate recognition and interpretation of these vocal cues significantly enhance the quality and efficacy of human–computer interactions. As voice-based interfaces gain prevalence, speech emotion recognition technologies have become integral to fostering more intuitive and empathetic user experiences [33].

This technological foundation has facilitated the design of systems that enhance human–computer communication and improve interpersonal interactions through computer-mediated channels. Studies in this area include the augmentation of speech-based emotional expressions with visual aids to assist users in recognizing and interpreting emotions. For instance, studies have explored the use of color-coded text messages to reflect the emotions detected in speech [7] and implement emotive speech bubbles [2]. In the context of accessibility, studies have been conducted to help individuals with hearing impairments perceive emotions in online meetings using specially designed subtitles [8].

Whereas existing design studies have focused on augmenting emotions that are challenging to recognize during communication, we propose a system that prevents excessive emotional expressions from being conveyed to listeners. Ultimately, we aim to develop a system that automatically recognizes emotions and converts emotionally extreme vocal expressions into normal expressions. Therefore, in this study, we selected call center workers as users who could benefit the most from such an interface and aimed to understand its feasibility within their working scenario.

3 FORMATIVE STUDY

Before starting our study, we performed preliminary interviews with 12 call center workers to understand their perceptions of the potential value of voice modulation systems in their work environments. All participants were female and varied in age (mean: 36 years, std: 4.7 years) and career (mean: 6 years, std: 4 years). It took around one hour per person and we asked user consent. The interview were semi-structured and included the following questions:

- What do you think if there is a system modulating customer's voices automatically?
- How would you feel if the system altered the customers' voices to sound dehumanized, such as a robot voice?

Through a thematic analysis of the results, we observed that most workers acknowledged the need for a system that automatically modulates voice volume, recognizing its potential to alleviate some of the stress of handling challenging customers. For instance, one participant answered "It doesn't sound bad. High-pitched voices hurt my ears. I usually listen to my speakers at full volume, but when the customer suddenly yells, I beep." Similarly, another participant said, "I have a traumatic experience in the very low tone of men. Whenever I listened to this voice, I felt stressed from the beginning of the call." The other answered that "I think it would be useful when I have to listen annoying voice in all day." However, several workers expressed concerns about dehumanizing customer voices and the potential risk of misinterpreting the emotions or intentions of customers. One answered that "Not all customers intend to express anger. Sometimes, the workers' responses matter. I think if the system makes it difficult to understand the emotions of the customer, it could result in unnecessary conflicts." Another said, "If the system changes the voice, I think system has to show additional information about what's going on in other way to understand the situation." The interview results led us to identify that, while there is an apparent necessity for automatic volume control in a call center environment, further investigation is required into the modification and its impact on understanding the emotion of additional prosodic features, such as pitch (tone height) and the application of dehumanized voice filters.

4 METHOD

Based on the needs and concerns captured in our formative study, we set the following research questions to assess the impact of voice modulation systems on stress, emotional contagion, and content comprehension.

- RQ1: Does modulating the prosodic features of a speaker's voice affect listeners' perceived stress?
- RQ2: Can voice modulation systems reduce emotion contagion in customer service interactions?
- RQ3: How does voice modulation influence listeners' comprehension of speech content?

In the user study, participants listened to both the original and modulated speech and evaluated the content in terms of stress

levels, emotional contagion, and content delivery. The following subsections describe the details of this study.

4.1 Selecting Features to Modulate

Considering the various ways to modulate voice, we reviewed the literature and examined the general characteristics of call centers to determine the most effective features for modulation. Speech emotion recognition focuses on three prosodic features: pitch (fundamental frequency), energy (pitch contour), and duration (speech rate) [1]. Applying our system to a call center environment requires real-time audio processing, as maintaining a smooth flow of conversations between call center workers and customers is important. Therefore, we prioritized pitch and energy modulation and excluded duration as it is not practical for real-time transformations. Studies on emotional speech recognition, which distinguishes emotional states by acoustic features (pitch, intensity, and timing), have shown that salient prosodic features for each emotion exist and that anger is characterized by a high mean pitch level, vocalization intensity, and energy [31]. Considering the challenges posed by angry customers in call centers, we selected pitch as the primary feature for modulation, because of its strong association with anger.

4.2 Preparing Audio Samples

Our first step was to prepare angry vocal expressions from the customers. To extract customer anger, we selected the following scenario: The customer makes unreasonable requests for unavailable services. The worker cannot offer any further solution because the customer has already agreed to the customer agreement form. The customer expresses frustration in a normal tone, but as emotions escalate, he begins to complain aggressively. Although CSRs organizations are unable to provide a direct solution, they must effectively manage the situation as representatives of the organization. We acknowledged that this is a common and stressful situation encountered by several CSRs. Therefore, we collected a corpus of customer utterances in these scenarios where a customer (actor) interacts with a call center worker in a laboratory setting. We selected audio samples from utterance segments in which the customer actor used negative expressions. The experimental scenarios were designed to distinguish between interactions in which the customer makes a rude complaint with a neutral tone and those with an aggressive tone. Using Praat, a speech analysis software, we analyzed the prosodic features of audio samples from both scenarios and found clear differences despite the conversations being about the same situation, as illustrated in Fig. 1, labeled as Versions A and B. These audio sample pairs were used for modulation.

The next step was to modulate the prosodic attributes associated with anger, using the original audio samples. Using Praat's diverse manipulation functions, three different audio samples were created using two techniques: (1) lowering the pitch frequency and (2) removing the pitch contours. The first technique reduced the pitch frequency by 50 Hz, yielding Condition 2 (C2) audio samples as listed in Table 1. For the second technique, we eliminated pitch peaks and created a uniform pitch level, leading to Condition 3 (C3) samples that sound monotone or 'robotic.' Combining both techniques generated Condition 4 (C4) audio samples. Consequently, we produced eight audio samples from the two sources, as listed in Table 1.

4.3 Participants

The study involved 13 participants, including 11 university students and 2 others. The majority were in the range of 20–24 years old (11), and the remainder were in the range of 55–57. The gender distribution was predominantly women (10 women and 3 men). Although none had prior experience in a customer service call center, their diverse backgrounds ensured comprehensiveness and enhanced the generalizability of the findings.

4.4 Questionnaires

To answer the research questions, we used six questionnaires with a 5-point Likert scale (Table 2). We assessed the stress level in Question 1 to answer RQ1. To examine the impact of emotional contagion on RQ2, we split the inquiry into two parts: first, assessing the perceived emotion of the speaker (Questions 2 and 3), and second, evaluating the emotional reaction of the listener (Questions 4 and 5). For a clear assessment of the self-report questionnaire, terms such as arousal and valence were explained in advance. We employed a two-dimensional arousal-valence space [24] to specify emotions, where arousal and valence represent the intensity and positivity/negativity of emotions, respectively. To address RQ3, the participants evaluated their comprehension of audio content to determine the influence of modulation on content delivery (Question 6).

4.5 Procedure

The participants were assigned a random sequence of the four conditions listed in Table 1. This approach prevented them from anticipating the conditions and altering their responses. The order of listening to Versions A and B under each condition was not randomized. For instance, if the random order of conditions was selected as C2, C3, C4, and C1, the audio would be presented as follows: A-2, B-2, A-3, B-3, A-4, B-4, A-1, and B-1. After listening to each audio sample, the participants answered the six questions outlined in Section 4.4. The study procedure lasted approximately 30 min per participant and we asked user consent.

5 RESULTS

Our user study revealed that modulating pitch not only reduced stress and emotional contagion but also did not significantly hinder content comprehension.

5.1 Influence of voice modulation on stress

To analyze the data collected from the first questionnaire (perceived stress level), we calculated the mean value of each condition for both versions. Thereafter, within each version, we used a Wilcoxon signed rank test (significance level $\alpha = 0.05$) to compare the base condition C1 with other modulation conditions and determine whether a significant difference in perceived stress between them exists. Additionally, we used Friedman test along with a Post Hoc Test using Conover's Method to identify the pair with the most significant difference.

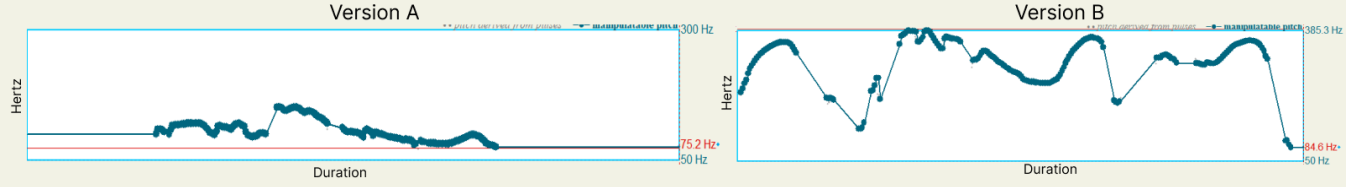

Figure 1: For the user study, we use both audio versions A and B, each containing expressions of customer anger with different prosodic features.

Table 1: Eight audio samples used in the user study

Condition\Version	A (neural voice tone)	B (aggressive voice tone)
C1 (without modulation)	A-1	B-1
C2 (lowered pitch)	A-2	B-2
C3 (removed pitch contours)	A-3	B-3
C4 (applied both filters used in C2, C3)	A-4	B-4

Table 2: Questionnaire

Topic	Question	Scale
Stress level	1) The stress level I perceived is	1 = no stress, 2 = mild stress, 3 = moderate stres,s 4 = much stress, 5 = extreme stress
Perceived arousal	2) The emotional arousal level of customer is	1 = very low, 2= low, 3 = neutral, 4 = high, 5 = very high
Perceived valence	3) The emotional valence level of customer is	1 = very negative, 2 = negative, 3 = moderate, 4 = positive, 5 = very positive
Felt arousal	4) My emotional arousal level is	1 = very low, 2= low, 3 = neutral, 4 = high, 5 = very high
Felt valence	5) My emotional valence level is	1 = very negative, 2 = negative, 3 = moderate, 4 = positive, 5 = very positive
Content delivery	6) I understand the content of the audio	1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree

As illustrated in Fig. 2, the original audio (C1) scored the highest mean stress level in both versions. Notably, stress was higher in Version B, implying that the original audio, characterized by a higher pitch and aggressive tone, tended to evoke a greater sense of stress among participants.

The results of the Wilcoxon signed-rank test showed that, in Version A, C1 was significantly higher than C3 ($z = -2.31$, $p = .02$). In Version B, C1 compared to C3 ($z = -2.71$, $p = .006$) and C4 ($z = -3.24$, $p = .001$) exhibited significantly higher values. The Friedman test confirmed that the most significant pair associated with stress reduction was C1 and C3 ($p = .07$) in Version A, whereas it was C1 and C4 ($p = .0002$) in Version B.

This analysis suggests that removing pitch contours and creating monotone played a role in stress reduction in Version A (neutral tone). For Version B (aggressive tone), both the removal of pitch contours and the lowering of pitch levels were effective, as they addressed both the prosodic features associated with aggression. The key takeaway is that removing the pitch contours has the most significant impact on stress regulation for both speech types. The absence of intonation in speech, resulting in a 'robotic' and monotone sound, could potentially dehumanize the voice, contributing to reduced stress by diminishing the sense of person-to-person interaction in the conversation.

5.2 Influence of voice modulation on emotional changes

The response data regarding arousal and valence levels from Questions 2, 3, 4, 5 in Section 4.4 were analyzed by taking the mean for each coordinate. These means were then plotted on the emotion coordinate space, with the horizontal axis denoting mean valence and the vertical axis denoting mean arousal. The corresponding standard deviations of valence and arousal are listed in Table 3.

Plot A in Fig. 3 shows the response data for Questions 2 and 3, which represent the emotion of the speaker as perceived by the listener. Data points from Version B were positioned toward the upper left side, indicating higher arousal and more negative perception compared to Version A. C4 was the most distant from C1 in both Versions A and B. This can be seen as a challenge in accurately interpreting emotions since the perceived emotion differs from the original audio.

In terms of the emotions felt by the participants (Questions 4 and 5), data points in Plot B of Fig. 3 were clustered in the lower

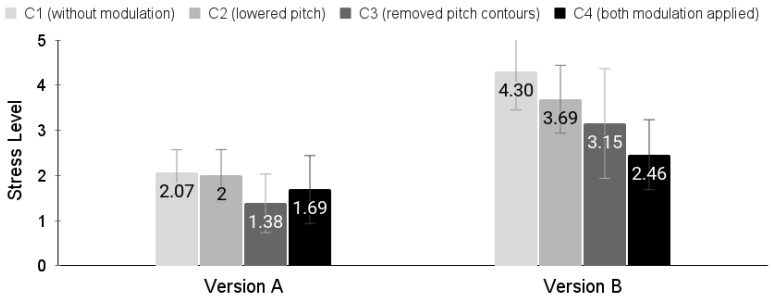

Figure 2: The mean stress levels of each case derived from participants' responses to the questionnaire - 1) What is your perceived stress level after listening to the audio sample? It shows a decrease in stress levels in the modulated audio.

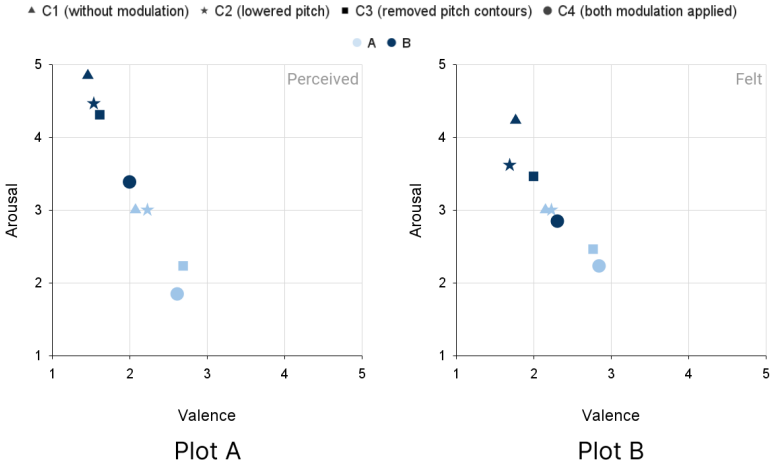

Figure 3: Valence-Arousal space of speaker's emotion perceived by the listener (Plot A) and emotion felt by the listener (Plot B). This visualizes that Version B (aggressive tone) has higher arousal and is more negative in all cases. C1 (without modulation) was the most negative and aroused.

part compared to points in Plot A, with C4 again standing out as the furthest data point from C1 in both versions. This suggests that applying a lowered pitch level and removing the pitch contour led to less negativity and lower arousal in the emotional experiences of the participants.

To examine emotional contagion, we plotted the difference between perceived and felt emotions within the same version (Plots C and D in Fig. 4). Under Version B in plot D, the notable difference vectors point distinctly right downward for all conditions, indicating less negativity and reduced arousal in felt emotions compared to perceived emotions.

Fig. 4 illustrates the difference vector and Euclidean distances between the perceived and felt emotion coordinates for each condition. In plot D of Fig. 4, which shows the emotions associated with Version B, C3 scored the highest distance (0.93) and emerged as the most optimal for emotion regulation. This condition maintained closeness to the original audio in terms of perceived emotion while showing the largest distance between perceived and felt emotions. Even with an accurate perception of a speaker's negative emotion, the distinct personal feeling suggests that it may not be as intense as perceived, leading us to interpret this as less emotionally contagious. This is a key factor in maintaining emotional authenticity, while minimizing the risk of emotional contagion.

To gain a deeper understanding of emotion contagion, we conducted a Wilcoxon signed-rank test to compare perceived and felt emotions. As summarized in Table 4, the results revealed no significant differences in valence or arousal for Version A. In the case of Version B, valence differences remained insignificant, whereas arousal differences were significant under conditions C1, C2, and C3. These findings correspond to the lengths of the difference vectors in plot D in Fig. 4, which exhibits a larger difference in arousal than in valence. These results have been confirmed in previous studies [6]. This suggests that, while the remaining linguistic features may influence emotional valence, the contagion of arousal is mitigated more effectively.

5.3 Influence of voice modulation on content comprehension

We analyzed the responses to Question 6 in Section 4.4 regarding content comprehension using the Wilcoxon signed-rank test. The

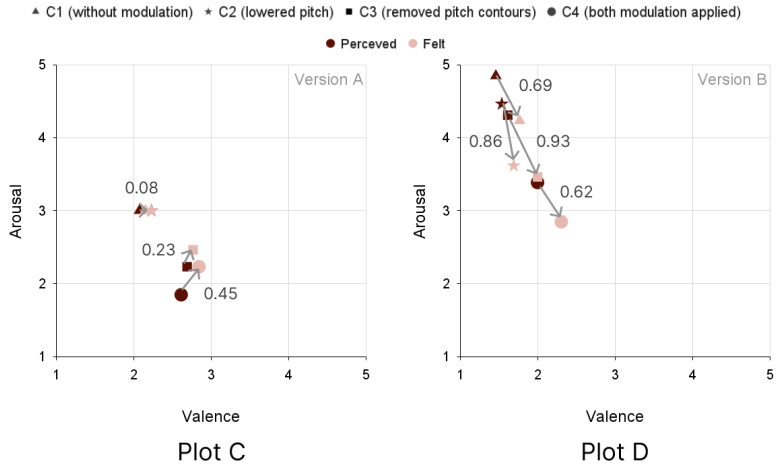

Figure 4: Difference between speaker's emotion perceived by the listener and emotion felt by the listener within version A (Plot C) and version B (Plot D). C3 in plot D scores the highest (0.93).

Table 3: The values in each cell represent the standard deviation (SD) for valence and arousal, formatted as (Valence SD, Arousal SD)

	Perceived	Felt		Perceived	Felt
A-1	(0.27, 0.81)	(0.55, 0.4)	B-1	(1.12, 0.37)	(1.09, 0.72)
A-2	(0.43, 0.81)	(0.43, 0.57)	B-2	(0.51, 0.51)	(0.48, 0.5)
A-3	(0.48, 0.72)	(0.59, 0.66)	B-3	(0.5, 0.48)	(0.57, 0.66)
A-4	(0.5, 0.68)	(0.37, 0.59)	B-4	(0.4, 1.04)	(0.63, 0.68)

Table 4: Outcomses of Wilcoxon Signed-Rank Test for emotional Valence and Arousal between Perceived and Felt emotions

	Valence	Arousal		Valence	Arousal
A-1	z = -0.45, p = .65	z = -0.1, p = .9	B-1	z = -1.63, p = .1	**z = -2.53,p = .01 < .05**
A-2	z = 0, p = 1.0	z = -0.06, p = .95	B-2	z = -1, p = .31	**z = -2.81, p = .005 < .05**
A-3	z = -0.38, p = .7	z = -1.13, p = .25	B-3	z = -1.51, p = .13	**z = -2.81, p = .005 < .05**
A-4	z = -1.73, p = .08	z = -1.89, p = .05	B-4	z = -1.63, p = .1	z = -1.9, p = .05

null hypothesis was rejected for all comparison pairs of Version A: C1 vs. C2 (z = -0.58, p = .56), C1 vs. C3 (z = -1.93, p =.053), and C1 vs. C4 (z = -1.41, p = .15). Similarly, for Version B: C1 vs C2 (z = -0.58, p = .56), C1 vs C3 (z = -1.41, p = .15), C1 vs C4 (z = -0.52, p = .60). These results suggest that voice modulation does not significantly affect the content delivery.

6 DISCUSSION

This study explored the potential need for a negative voice modulation system of a customer among call center workers and investigated the impact of voice modulation of prosodic features, which induces negative emotions, on emotional contagion. While negative linguistic features persisted, impeding a significant shift in emotional valence, we observed that modulating the selected prosodic features (mainly pitch) could prevent emotional contagion related to arousal. This highlights the potential to improve the mental health of call center workers. The following sections discuss several key points from our findings.

6.1 Effective communication by balancing out emotion contagion and accurate perception of emotion

A modulation condition under which people perceived most differently (when both the pitch was lowered, and the contours were removed) existed. Excessive use of this modulation may cause incorrect emotional perception, potentially hindering communication. However, as we verified in the comprehension test, no impact on content delivery was present. Striking a balance between reducing emotional contagion and maintaining an accurate perception of emotions will eventually lead to effective communication and ensure the well-being of call center workers. This balance can be

achieved through customizable options, as mentioned in the interviews in formative studies. While we focused on pitch-related prosodic features, further interfaces could allow workers to adjust the modulation settings according to their preferences. For instance, we can offer options to turn the modes on and off to apply automatic voice conversion or select changeable features.

6.2 Application in real-world setting

Despite advancements in voice-style transfer technology [21, 28, 34], achieving real-time applications remains a key challenge for systems designed to mitigate emotional contagion in customer service by recognizing and dampening emotions in the voices of customers. Even though deep learning-based emotion style transfer is also being explored in real time [22], it can be delayed when combined with emotion recognition algorithms. However, the prosodic features of customers are inconsistent and variable, and require real-time adaptation within a single interaction. Therefore, real-time adaptation is essential for the system and could be an alternative for developing lightweight practical algorithms using limited prosodic features that stimulate predominant emotions.

6.3 Ethical considerations

Despite the technical feasibility and potential benefits of voice modulation, its application to customer voices in a customer service environment requires careful consideration. Voice is considered a representation of personal information; therefore, generating analytics on voice-based emotions of a customer without consent has the potential to invade personal privacy [15]. Hence, informed consent should be considered.

7 LIMITATION AND FUTURE WORKS

This study demonstrated that modulated voices could reduce the spread of negative emotions, indicating the potential benefits of creating a positive work environment. However, it has some limitations: it only used a limited number of speech features; the user study was not based on actual agents; it tested modulated audio, not real-time voice modulation technology; and it only considered the perspective of the listener, not that of the customer. In addition, to measure emotional contagion, we simply selected two dimensions of emotion, but it could be more useful to use systematic methods to measure more complex emotions, such as the Genova emotion wheel [25]. Future studies should test a real-time system with more speech features and assess the effectiveness of the tool from both the perspectives of the agents and customers.

8 CONCLUSIONS

We investigated the potential of customer voice modulation in the emotional well-being of call center workers. As a preliminary study, we analyzed the effects of pitch-modulated versus unmodulated customer voices and observed that pitch modulation significantly reduced stress and improved emotional regulation, highlighting its potential as a tool for improving the workplace environment for call center workers and enhancing customer interaction quality. This study underscores the importance of auditory cues in emotional well-being and opens avenues for further exploration of technology-assisted emotion regulation strategies in high-stress occupations. Our work calls for future studies to evaluate real-time audio modulation systems using a more varied set of features with call center workers, along with consideration of ethical issues.

ACKNOWLEDGMENTS

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2022-0-00064, Development of Human Digital Twin Technologies for Prediction and Management of Emotion Workers' Mental Health Risks). The corresponding author is U. Lee.

REFERENCES

[1] Mehmet Berkehan Akçay and Kaya Oğuz. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication* 116 (2020), 56–76.

[2] Toshiki Aoki, Rintaro Chujo, Katsufumi Matsui, Saemi Choi, and Ari Hautasaari. 2022. EmoBalloon - Conveying Emotional Arousal in Text Chats with Speech Balloons. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (, New Orleans, LA, USA,) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 527, 16 pages. https://doi.org/10.1145/3491102.3501920

[3] Norah Aung and Promise Tewogbola. 2019. The impact of emotional labor on the health in the workplace: a narrative review of literature from 2013–2018. *AIMS Public Health* 6, 3 (2019), 268.

[4] Patricia B Barger and Alicia A Grandey. 2006. Service with a smile and encounter satisfaction: Emotional contagion and appraisal mechanisms. *Academy of management journal* 49, 6 (2006), 1229–1238.

[5] Roy F Baumeister, Ellen Bratslavsky, Mark Muraven, and Dianne M Tice. 1998. Ego depletion: Is the active self a limited resource? *Journal of personality and social psychology* 74, 5 (1998), 1252.

[6] Janet E Cahn. 1990. The generation of affect in synthesized speech. *Journal of the American Voice I/O Society* 8, 1 (1990), 1–1.

[7] Qinyue Chen, Yuchun Yan, and Hyeon-Jeong Suk. 2021. Bubble Coloring to Visualize the Speech Emotion. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (, Yokohama, Japan,) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 361, 6 pages. https://doi.org/10.1145/3411763.3451698

[8] Caluã de Lacerda Pataca, Matthew Watkins, Roshan Peiris, Sooyeon Lee, and Matt Huenerfauth. 2023. Visualization of Speech Prosody and Emotion in Captions: Accessibility for Deaf and Hard-of-Hearing Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (, Hamburg, Germany,) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 831, 15 pages. https://doi.org/10.1145/3544548.3581511

[9] Maureen F Dollard, Christian Dormann, Carolyn M Boyd, Helen R Winefield, and Anthony H Winefield. 2003. Unique aspects of stress in human service work. *Australian psychologist* 38, 2 (2003), 84–91.

[10] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition* 44, 3 (2011), 572–587.

[11] Alicia A Grandey. 2000. Emotional regulation in the workplace: A new way to conceptualize emotional labor. *Journal of occupational health psychology* 5, 1 (2000), 95.

[12] Kun Han, Dong Yu, and Ivan Tashev. 2014. Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech 2014*.

[13] Elaine Hatfield, John T Cacioppo, and Richard L Rapson. 1993. Emotional contagion. *Current directions in psychological science* 2, 3 (1993), 96–100.

[14] Alexander P Henkel, Stefano Bromuri, Deniz Iren, and Visara Urovi. 2020. Half human, half machine–augmenting service employees with AI for interpersonal emotion regulation. *Journal of Service Management* 31, 2 (2020), 247–265.

[15] Jacob Leon Kröger, Otto Hans-Martin Lutz, and Philip Raschke. 2020. Privacy implications of voice and speech analysis–information disclosure by inference. *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers 14* (2020), 242–258.

[16] Chul Min Lee and S.S. Narayanan. 2005. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing* 13, 2 (2005), 293–303. https://doi.org/10.1109/TSA.2004.838534

[17] Xiao-Yu Liu, Nai-Wen Chi, and Dwayne D Gremler. 2019. Emotion cycles in services: Emotional contagion and emotional labor effects. *Journal of Service Research* 22, 3 (2019), 285–300.

[18] Heyeon Park, Hyunjin Oh, and Sunjoo Boo. 2019. The role of occupational stress in the association between emotional labor and mental health: a moderated

mediation model. *Sustainability* 11, 7 (2019), 1886.

[19] Jungsun Park. 2016. Strategies to prevent work-related stress and cardiovascular diseases in South Korea. *Psychosocial Factors at Work in the Asia Pacific: From Theory to Practice* (2016), 77–86.

[20] Karen Pugliesi. 1999. The consequences of emotional labor: Effects on work stress, job satisfaction, and well-being. *Motivation and emotion* 23 (1999), 125–154.

[21] Chunyu Qiang, Peng Yang, Hao Che, Ying Zhang, Xiaorui Wang, and Zhongyuan Wang. 2023. Improving Prosody for Cross-Speaker Style Transfer by Semi-Supervised Style Extractor and Hierarchical Modeling in Speech Synthesis. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095840

[22] Yurii Rebryk and Stanislav Beliaev. 2020. Convoice: Real-time zero-shot voice style transfer with convolutional network. *arXiv preprint arXiv:2005.07815* (2020).

[23] Rita Rueff-Lopes, José Navarro, António Caetano, and Ana Junça Silva. 2015. A Markov chain analysis of emotional exchange in voice-to-voice communication: Testing for the mimicry hypothesis of emotional contagion. *Human Communication Research* 41, 3 (2015), 412–434.

[24] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[25] Vera Sacharin, Katja Schlegel, and Klaus R Scherer. 2012. Geneva emotion wheel rating study. *Center for Person, Kommunikation, Aalborg University, NCCR Affective Sciences. Aalborg University, Aalborg* (2012).

[26] Widakorn Saewong and Janjao Mongkolnavin. 2019. Classification of Anger Voice in Call Center Dialog. In *2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. 298–302. https://doi.org/10.1109/JCSSE.2019.8864217

[27] Widakorn Saewong and Janjao Mongkolnavin. 2019. Classification of anger voice in call center dialog. In *2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE, 298–302.

[28] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. 2021. An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 132–157. https://doi.org/10.1109/TASLP.2020.3038524

[29] Peter Totterdell and David Holman. 2003. Emotion regulation in customer service roles: testing a model of emotional labor. *Journal of occupational health psychology* 8, 1 (2003), 55.

[30] Panagiotis Tzirakis, George Trigeorgis, Mihalis A. Nicolaou, Björn W. Schuller, and Stefanos Zafeiriou. 2017. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE Journal of Selected Topics in Signal Processing* 11, 8 (2017), 1301–1309. https://doi.org/10.1109/JSTSP.2017.2764438

[31] Dimitrios Ververidis and Constantine Kotropoulos. 2006. Emotional speech recognition: Resources, features, and methods. *Speech Communication* 48, 9 (2006), 1162–1181.

[32] Dieter Zapf, Amela Isic, Myriam Bechtoldt, and Patricia Blau. 2003. What is typical for call centre jobs? Job characteristics, and service interactions in different call centres. *European journal of work and organizational psychology* 12, 4 (2003), 311–340.

[33] Xinlei Zhang, Zixiong Su, and Jun Rekimoto. 2022. Aware: Intuitive Device Activation Using Prosody for Natural Voice Interactions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (, New Orleans, LA, USA,) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 432, 16 pages. https://doi.org/10.1145/3491102.3517687

[34] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2021. Seen and Unseen Emotional Style Transfer for Voice Conversion with A New Emotional Speech Dataset. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 920–924. https://doi.org/10.1109/ICASSP39728.2021.9413391