EmoBalloon - Conveying Emotional Arousal in Text Chats with Speech Balloons

Toshiki Aoki*
aoki-toshiki1127@g.ecc.u-tokyo.ac.jp
The University of Tokyo
Tokyo, Japan

Rintaro Chujo*
rintaro-chujo@g.ecc.u-tokyo.ac.jp
The University of Tokyo
Tokyo, Japan

Katsufumi Matsui
matsui.katsufumi@mail.u-tokyo.ac.jp
The University of Tokyo
Tokyo, Japan

Saemi Choi
saemi1.choi@samsung.com
Samsung Research
Seoul, Republic of Korea

Ari Hautasaari
a.hautasaari@iii.u-tokyo.ac.jp
The University of Tokyo
Tokyo, Japan

ABSTRACT

Text chat applications are an integral part of daily social and professional communication. However, messages sent over text chat applications do not convey vocal or nonverbal information from the sender, and detecting the emotional tone in text-only messages is challenging. In this paper, we explore the effects of speech balloon shapes on the sender-receiver agreement regarding the emotionality of a text message. We first investigated the relationship between the shape of a speech balloon and the emotionality of speech text in Japanese manga. Based on these results, we created a system that automatically generates speech balloons matching linear emotional arousal intensity by Auxiliary Classifier Generative Adversarial Networks (ACGAN). Our evaluation results from a controlled experiment suggested that the use of emotional speech balloons outperforms the use of emoticons in decreasing the differences between message senders' and receivers' perceptions about the level of emotional arousal in text messages.

CCS CONCEPTS

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

KEYWORDS

speech balloon, text chat, voice input, emotion

ACM Reference Format:
Toshiki Aoki, Rintaro Chujo, Katsufumi Matsui, Saemi Choi, and Ari Hautasaari. 2022. EmoBalloon - Conveying Emotional Arousal in Text Chats with Speech Balloons. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3491102.3501920

*Both authors contributed equally to this research.

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501920

1 INTRODUCTION

Nowadays, it is common for people to communicate over instant messaging (IM) and text chat applications in their everyday work and social life. Text chat apps are widespread and specialized for a range of communicative contexts from formal to informal. Moreover, some recent mobile applications also make use of the text-based medium as a way to aid hearing-impaired users by transcribing voice messages or telephone calls to text via voice-to-text functionality [17, 51]. However, while the range of options for text chat apps has drastically increased with the popularity of smartphones, there are still many limitations to these applications, particularly when used for socio-emotional communication [9, 13, 24, 43, 59, 62, 63, 65, 67].

The main limitation when communicating via text chat applications is that messages are restricted to the use of symbols and paralinguistic cues. That is, basic text-based chats omit message sender's facial expressions, gestures and the tone of their voice, which are available in richer mediums and face-to-face interactions [13, 43, 62, 63, 65]. The lack of these emotional cues can in turn hinder socio-emotional communication between senders and receivers, and lead to misunderstandings about the emotional tone of messages [9, 24, 59, 62, 67]. For example, if a friend sends you an ambiguous message such as "it was good" after eating in a restaurant, without the more nuanced emotional cues available in richer mediums it may be difficult to determine the intensity or intended level of emotional arousal in the message.

There are many existing solutions that aim to amend the low emotionality of text-based chats. The most common way to enhance text messages is by using emoticons or stamps as paralinguistic cues to convey more nuanced emotional information [2, 25, 65]. Emoticons often represent emotional facial expressions, but also have limitations in the sense that they do not represent the sender's experienced state but rather act as additional symbols chosen by the sender to enhance their message [61, 62, 65]. Other approaches to improve socio-emotional communication via text-based chats include the use of emotional typefaces (i.e., fonts) [12, 73], and the use of animated text [68] to enhance the emotionality of messages. Outside the field of computer-mediated communication (CMC), television programs may use emotional subtitles to enhance the experience of hearing impaired audience [49].

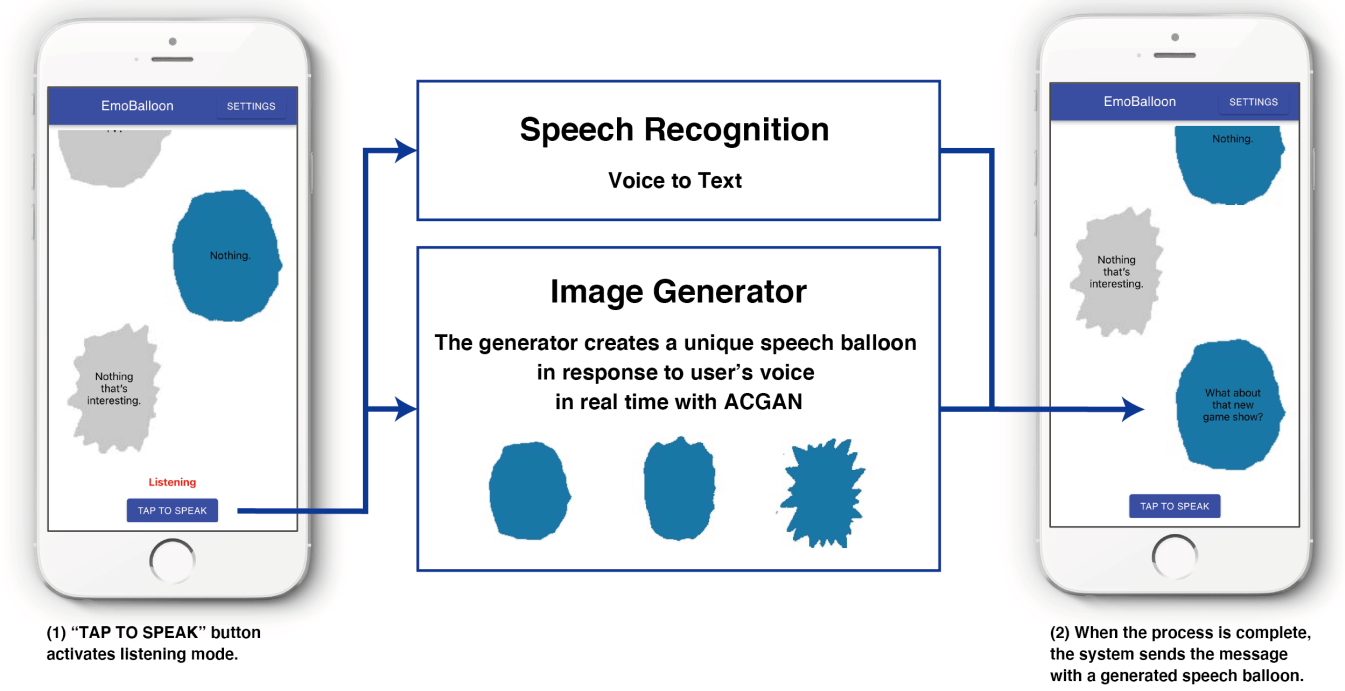


Figure 1: Overview of EmoBalloon system design

In this paper, we propose a novel method to increase the socio-emotional cues available in text-based chats with speech balloons, which are traditionally used in many chat applications to frame the textual content as well as to indicate the message sender in color-coding. EmoBalloon (Figure 1) is developed as a voice input text chat application, which detects the sender's emotional arousal from their voice and automatically generates a matching speech balloon for the text message being sent. As a first step in our system development, we investigated the relationship between emotional arousal and valence, and the shape of speech balloons in Japanese manga using the Manga109 dataset [1, 42, 46]. To implement our system for automatic speech balloon generation, we used a machine learning approach (the ACGAN) with 100,000 generated images as training data based on the speech balloon data extracted from the Manda109 dataset [1, 46]. The system was first evaluated via crowdsourcing to confirm that the automatically generated speech balloons represented emotional information.

We further evaluated the efficacy of EmoBalloon for supporting text-based socio-emotional communication in a controlled experiment, where Japanese participants used either a voice input version of EmoBalloon, a manual selection version of EmoBalloon with keyboard text input, or a text chat with emoticons to convey emotional information to their partner. The experiment results indicated that emotional speech balloons significantly increased the perceived level of emotional arousal in text messages for the receivers, as well as the agreement between message senders and receivers regarding the level of emotional arousal conveyed in a message. Furthermore, emotional speech balloons outperformed emoticons in decreasing the differences between message senders' and receivers' perceptions about the level of emotional arousal, and contrary to emoticons, speech balloons did not affect receivers' perceptions about the emotional valence in text-based messages. We discuss our results and propose future directions to further improve socio-emotional communication in text-based chat applications.

2 RELATED WORKS

Daily interactions between people are shaped by shared affective experiences. How well emotions are conveyed and how accurately they are understood is important for both interpersonal relationships and individual well-being [22]. In face-to-face settings, conversational partners can make use of the rich emotional cues in facial expressions, gestures or the tone of their voice to convey and assess the emotional nuances during interactions [7, 18].

However, computer-mediated communication (CMC) environments limit the use of the communicative cues available in face-to-face settings. One of the most prominent theories of CMC, the media richness theory, describes the abilities of different CMC mediums to allow the use of natural language, rapid feedback, the establishment of personal focus and the use of multiple types of communicative cues simultaneously [13, 43]. For example, video conferencing is considered a "rich" medium, with which interlocutors can make use of multiple cues (e.g., facial expressions, gestures, tone of voice) to convey their message and exchange feedback in real-time.

Text-based instant messaging (IM) applications lack the rich vocal and nonverbal emotional cues available in CMC environments where conversational partners can hear and see each other [15, 59, 62]. On one hand, the lack of these communicative cues affords interlocutors to craft their self-presentation when communicating with others [64]. Furthermore, as the receivers have fewer cues to make assessments about their conversational partner, they tend to form an idealized version of the person and their personality [66]. On the other hand, the fewer communicative cues available in text-based CMC can render individual messages less emotionally arousing than when communicating over richer mediums or in face-to-face settings, which may, in turn, lead to a greater chance of misunderstandings occurring between conversational partners regarding the emotional tone of a message [9, 24, 59, 62, 67].

2.1 Emotional expression in text-based communication

Text-based chats are considered a "lean" communication medium, which requires conversational partners to put more effort into facilitating mutual understanding and resolving the meaning of ambiguous messages compared to richer media such as video or audio conferencing [9, 13]. According to the social information processing (SIP) theory, social relationships are formed even through lean text-based mediums if the interlocutors expect more opportunities to interact over time [62, 63]. However, due to the limited emotional cues more messages have to be exchanged between conversational partners for similar relational effects compared to interactions in face-to-face settings or over richer CMC mediums [62].

While establishing personal relationships is possible even with limited socio-emotional cues, in the case of individual messages and message exchanges, receivers may encounter difficulties in accurately recognizing the intended emotional meaning. Byron [9] called the discrepancies between the sender and receiver's perceptions about the socio-emotional content negativity and neutrality effects. That is, receivers may perceive text-based messages as more negative than intended by the sender if the message includes any negative socio-emotional cues (e.g., words, phrases, emoticons) [9, 65]. Similarly, receivers may perceive messages intended to convey a positive tone as more neutral than intended by the sender [9].

Message senders (or authors) "hear" their own voice while writing emotional messages [39]. However, this information is not transferred to the message receiver in traditional text-based mediums, which renders the message less emotionally arousing to the receiver than to the sender [37, 39]. Moreover, message senders may overestimate how much of the socio-emotional content is accurately conveyed to the receiver [32]. Due to these differences between senders and receivers' perceptions about the emotionality of text-based messages, there is a higher chance of misunderstandings occurring between the conversational partners regarding the emotional tone (i.e., receivers perceiving messages as more negative than intended) which in turn may lead to conflict escalation [4, 21].

In order to overcome the challenges in socio-emotional communication and the limitations of text-based chats, conversational partners may make use of symbols (e.g., exclamation marks) or paralinguistic cues such as emoticons to enrich the emotional content in their messages [2, 25, 65, 72]. The use of emoticons is associated with higher levels of emotional transmission, and tend to indicate the place for nonverbal emotional expressions occurring in face-to-face interactions, such as laughter or frowning [31, 33, 45, 53, 54, 69, 74]. These cues in turn can strengthen the intensity of the written emotional message, especially in the case of negative emotional tone expressed in the written cues [16, 65]. But, how users interpret emoticons and different emoticon designs may also be tied to their cultural background [70].

Previous research has aimed to go beyond the use of written and paralinguistic cues to enhance socio-emotional text-based communication. For example, one way to increase the socio-emotional cues in text chats is to add the nonverbal cues (i.e., facial expressions, gestures) via a video feed between the conversational partners [58]. However, this type of approach still omits the emotional tone carried in the speaker's voice [39], and may be less viable than text-only chats in many communicative situations in part due to evaluation apprehension, vanity or privacy concerns stemming from the use of the video feed [14].

Recent research has also proposed novel ways to detect users' emotional states in text-only chats and convey them through familiar paralinguistic cues (i.e., emoticons). Liu and colleagues [44] proposed ReactionBot, which detects the message sender's facial expression without opening a video feed to the receiver, and automatically displays an emoticon matching the emotionality of the user's facial expression. However, this automated approach was found to increase the user's anxiety about "negative emotion leak" in text-only chats, where their facial expression is usually not visible to the receiver [44]. On the other hand, Kim and colleagues [35] proposed a system that recommends the message sender appropriate emoticons based on the context of the conversation, which may help users convey their emotional tone in paralinguistic cues more effectively as well as diversify their emoticon use in text-only conversations.

Other approaches have developed methods to detect the context of messages from the sender's digital and physical environment. Buschek and colleagues [8] proposed three systems to add personalized fonts, physiological data from the user as well automatic annotation of the user's background (e.g., music, weather) to enhance the expressiveness and emotionality of text chats. Chen and colleagues [10] used the digital environment (e.g., calendars and other external apps) on users devices to enhance text prediction, whereas the system developed by Kim and colleagues [34] analyzed the context of the textual message content to improve image suggestions during text chats with an aim to aid emotional expression through images.

Besides detecting contextual information from the message sender, previous works have also explored ways to enhance the emotionality of text itself. Choi and Aizawa [12] proposed a system that matched the emoticon chosen by a message sender with an emotional typeface (i.e., font), whereas Yonekura and colleagues [73] proposed a method to automatically select an emotional typeface matching to the emotional tone of a text-based message. Both studies also reported on the effectiveness of emotional typefaces to enhance the socio-emotional content in text-based chat messages [12, 73]. Besides emotional typefaces, the use of animated text has long roots outside the CMC domain, where the technique has been

used to enhance the television viewing experience of the hearing impaired [49]. Previous research has also investigated how translating data from biosensors attached to a message sender's body to animated text enhances text-based online communication [68].

While many previous studies have proposed ways to improve the textual and paralinguistic cues in text messages, few studies have investigated the effects of another aspect of displaying socio-emotional content in text-based chats - the speech balloon.

2.2 Speech balloons in text-based communication

Most IM and text-based chat applications make use of some type of speech balloon to frame the textual content, as well as to often indicate the message sender with color coding. However, speech balloons can also serve another function, and are used as one technique to convey emotional nuance in comics and Japanese manga [47]. Recent research has investigated the relationship between linguistic features of speech and speech balloon shape [71]. Yamanishi and colleagues [71] proposed a method to match the type of speech balloon to speech in Japanese manga cartoons (i.e., cartoon text), and concluded that especially the "explosion" type speech balloon predicted high arousal speech [71]. Another study investigated the relationship between manga characters expressions and speech balloon shape using the tf-idf method and focusing on keywords and symbols within the speech balloons [60]. Speech balloons have also been proposed as additives to subtitles in TV shows [38], as well as for real-time communication via video and automated transcripts to aid the hearing impaired to detect the characteristics of the speakers' speech beyond the verbal content (e.g., tone, speed, prosody, volume) [52].

In the field of text-based CMC, previous research by Kurlander and colleagues [41] proposed a system called Comic Chat, which converted online discussions to comic strips. However, this system focused mostly on expressing emotions via characters' facial expressions and gestures in a generated comic, and the original version included only three types of speech balloons (speech, thought and whisper) [41]. More recently, Chen and colleagues explored how the color of a speech bubble (or voice message indicator) enhances the emotional expression in voice messages sent via various IM applications [11]. However, beyond Comic Chat, research further investigating the relationship between socio-emotional communication and the use of speech bubbles in text-based chats is still scarce.

To sum, speech balloons have recently gained interest in CMC research to enhance the communicative experience. However, to the best of our knowledge, there currently exists few practical solutions for the use of speech balloons to enhance emotional communication in text-based chats. In this paper, we propose such a solution - EmoBalloon.

3 EMOBALLOON SYSTEM DESIGN

In this paper, we propose a system, EmoBalloon, to support socio-emotional communication in text chats with speech balloons. As a first step in our system development, we analyzed a Japanese manga data set (Manga109 [42]) to investigate the relationship between speech balloon shape and emotional information conveyed in speech text. Based on the results, we then trained a machine learning model with auxiliary classifier generative adversarial networks (ACGAN) by a generated data set of 100,000 speech balloon images to generate new speech balloons on a linear emotional arousal scale. Lastly, we investigated whether the speech balloons generated by EmoBalloon corresponded to emotional arousal levels through a crowdsourcing evaluation. The results from this evaluation confirmed the applicability of speech balloons generated by EmoBalloon to convey socio-emotional information in text-based chats.

3.1 Speech balloon data set

In this study, we use the Manga109 data set, which consists of 109 Japanese manga titles and a total of 10,130 pages [42]. Each page in the data set is annotated with the position of characters' faces and bodies, dialogue (speech), frame with character ID, as well as the contents of speech in text (Figure 2). However, the original data set did not provide any information about speech balloons used within the manga.

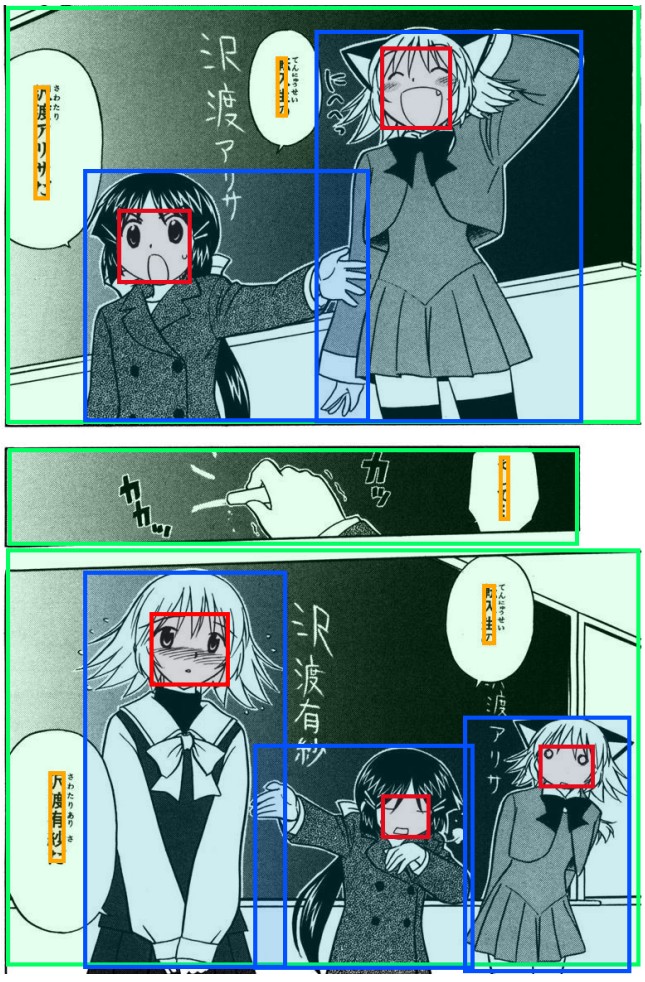


Figure 2: Sample of Manga109 data ©八神健(Ken Yagami)

As a first step, we therefore automatically extracted the speech balloons based on the location information of the speech text on each page. Every text is surrounded by a white region (i.e., balloon), and this region is separated from other regions of the page by a dark line. We firstly converted all color images to black and white in order to contrast the white balloon region from other regions. We then extracted the white regions that are adjacent to text. After masking the text region, we were able to extract clear speech balloon images from the Manga109 data set.

However, these data and speech text pairs also included image data which were not considered speech balloons, but a type of explanatory line in Manga (i.e., not including character dialogue). As we are only interested in speech balloons related to dialogue, we used the transfer learning method [50] to classify the speech balloon data unrelated to dialogue. Transfer learning is a method to learn the classification data based on a pre-trained model. We classified approximately 2000 images as either speech balloon data or non-speech balloon data as training data by hand and then used a pre-trained Resnet18 model to learn these 2000 images. With this method, all data were classified, and speech balloon image and speech text pairs were generated as a data set.

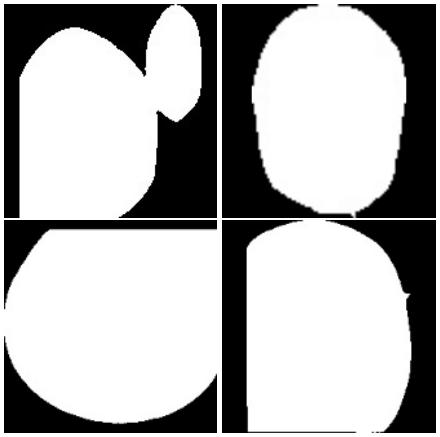


Figure 3: Examples of round shape speech balloons (top left image: ©大井昌和(Masakazu Ooi), top right image: ©八神健(Ken Yagami), bottom left and bottom right images: ©よしまさこ(Masako Yoshi))

There are many different shapes of the extracted speech balloons. Based on previous works [60, 71], we focus on two speech balloon categories, the round shape (Figure 3), and the explosion shape (Figure 4). The reason why we focus on these two types of speech balloons in this work is that, firstly, based on previous work the round shape speech balloon is the most common in Japanese manga, and can be considered emotionally neutral (i.e., does not convey any emotion) [60]. The explosion type is the second most commonly used speech balloon, and is associated with high arousal and negative emotions (e.g., anger), as well as the use of exclamation marks in speech text indicating higher arousal [60] [57] [71, 72]. Also, Japanese users are likely most familiar with these two types of speech balloons and the associated dialogue type in manga. The other less common speech balloon types include

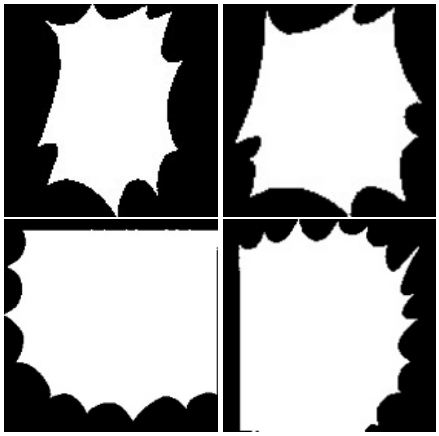


Figure 4: Examples of explosion shape speech balloons (top left and top right images: ©新沢基栄(Motoei Shinzawa), bottom left and bottom right images: ©よしまさこ(Masako Yoshi))

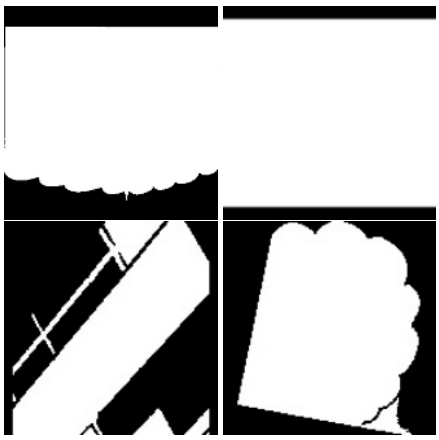


Figure 5: Examples of other shape speech balloons (©よしまさこ(Masako Yoshi))

the cloud, wave and flushing types, among others, which may also convey emotional information [60], and are considered part of our future work (Figure 5). In this work, the other speech balloon types were handled as one category.

First, we manually labeled approximately 1,000 images in each of the three categories (i.e., round, explosion, other). Leveraged by a pre-trained model [26], all the 147,944 speech balloon images in the generated data set were then classified into the three categories with the transfer learning method [50]. A total of 40,576 speech balloons were classified as the round shape and 7,560 balloons were classified as the explosion shape.

Table 1: The means and standard deviations of the arousal and valence levels

	the arousal levels	the valence levels
Explosion shape	0.388(0.321)	0.058(0.378)
Round shape	0.312(0.289)	-0.005(0.358)
p-value	<.001	<.001

Table 2: Ratio of texts in explosion and round shape speech balloons including at least one exclamation mark

	ratio
Explosion shape	0.534
Round shape	0.116

3.2 The relationship between emotional arousal, emotional valence, and speech balloon shape

Previous research regarding the shape of speech balloons and the message receiver's perception of the emotionality of the message suggested that the "explosion" shape speech balloons may be associated with high arousal [71]. Furthermore, previous research using the tf-idf method on the relationship between emotions and speech balloons used in Japanese manga suggested that the explosion shape is in general associated with negative emotions (anger, cry, scream), whereas an ordinal (i.e., round) shape was not associated with any emotions or emotional arousal [60]. However, based on previous research, it is difficult to predict whether a linear relationship between the speech balloon shape and emotional valence exists. That is, whether more explosion-like speech balloon shapes are associated with more intensely negative emotional valence.

To investigate the relationship between the shape of the speech balloon and the emotions underlying the text in detail, we employed a text emotion analysis API [1]. Since most emotion analysis methods show competitive performance in English, all the texts in the Manga109 dataset were automatically translated to English [2] before conducting the emotion analysis. Given an input text, the emotion analysis API outputs a value for arousal (i.e., magnitude indicating the overall strength of emotion) between 0 and infinity, and a value for emotional valence (i.e., score indicating how positive or negative the input sentence is) between -1 and 1.

We verified whether a relationship between the shape of a speech balloon and related speech text exists using a two-sample Kolmogorov - Smirnov test. We tested a null hypothesis: "sample X and sample Y are derived from the same distribution". Sample X is the population score for the level of arousal in text associated with a round shape speech balloon, and sample Y is the population score for the level of arousal in text associated with an explosion shape speech balloon. Based on the results from the K-S test, the null hypothesis was rejected ($p < .001$) indicating that the population arousal score for the text in the round shape speech balloon was different from the population arousal score of the text in the explosion shape speech balloon, and the explosion shape speech balloons were associated with higher emotional arousal.

Similarly, we analyzed the relationship between the valence score of the speech text associated with the round shape speech balloons and the explosion shape speech balloons using K-S test. We tested the null hypothesis: sample X and sample Y are derived from the same distribution. Sample X is the population emotional valence score for text associated with a round shape speech balloon, and sample Y is the population emotional valence score for text associated with an explosion shape speech balloon. Based on the results from the K-S test, the null hypothesis was rejected (p<.001) indicating that the population valence score for the text in the round shape speech balloon was different from the population valence score of the text in the explosion shape speech balloon.

Table 1 shows the means and standard deviations of the arousal and valence scores of the text in each shape speech balloon. We also discovered that contrary to the proposition in previous work [60], the valence score of speech text with the explosion shape speech balloon was higher (i.e., more positive) than with the round shape speech balloon. While the mean values for both shapes are close to neutral, this result suggested that the explosion shape speech balloon may be related to higher valence score of speech text and the round shape speech balloon may be related to lower valence score of speech text.

Lastly, we counted the number of exclamation marks in the text data, as we discovered that the Natural language API may not evaluate the exclamation marks in sentiment analysis. Table 2 shows the number of the texts which include at least one exclamation mark in each shape of balloon and the ratio of such text in each speech balloon shape. Exclamation marks are associated with higher level of arousal in text [72], and this analysis further indicated that the explosion shape speech balloon is related to higher emotional arousal and the round shape speech balloon is related to lower emotional arousal in speech text.

3.3 Dataset for ACGAN

In-the-wild speech balloon images in the Manga109 data set were not a good fit for both text chat apps and training generative models. Most of the speech balloons were partially occluded by the outer frame of a panel, so the generated balloon would follow the distribution of the occluded balloons, which in turn would affect

[1] https://cloud.google.com/natural-language/
[2] https://www.deepl.com/

user experience if used on actual text chat apps (Figure 3, Figure 4). Therefore, instead of using the extracted original images as is, we use graphically rendered speech balloons to generate emotional speech balloons.

As described in the previous section as well as in previous work [71], the explosion shape speech balloons represent higher emotional arousal whereas the round speech balloons are regarded as more neutral in terms of emotional arousal. We randomly selected 100 round shape (0.0 - low arousal) and 100 explosion shape (0.9 - high arousal) speech balloons not occluded by the outer frame from the Manga109 dataset, and interpolated between the two balloon type pairs using Adobe Illustrator Blend tool[3] combining the two objects in any ratio (1:0 - 0:1). Below is an example of a round speech balloon with a low arousal score (0.0) and an explosion speech balloon with a high arousal score (0.9). Each pair of speech balloons and their interpolation was repeated 10,000 times, and a total of 100,000 images were rendered with this process. The size of the rendered speech balloon images was set to 128×128 pixels.

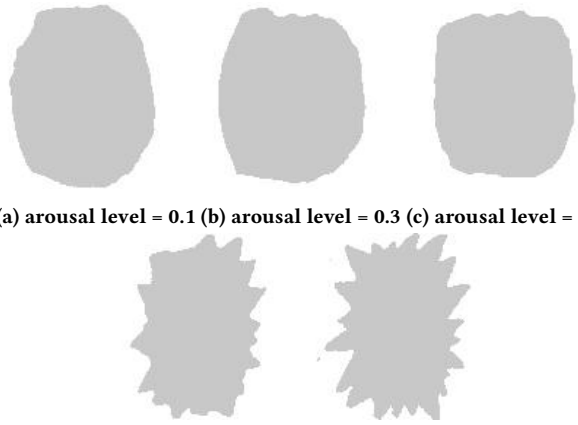
3.4 GAN and ACGAN

Generative Adversarial Networks (GANs) [23] are generative models that learn data distribution using the process of adversarial training. The generator produces fake data from trained data distribution while the discriminator distinguishes whether the input data came from trained data distribution or real. Thanks to its outstanding performance in producing photorealistic images it has been developed and applied to many applications such as art [19], fashion [29] and even human image synthesis [30], and can realize human-like randomness for generating various forms of speech balloons. However, no control is allowed in vanilla GAN to change the modes of the generated images. To solve this problem, Auxiliary Classifier GAN (ACGAN) [48] was proposed. With the ACGAN, we can control the data mode (i.e., label) to generate images. Even with the same arousal level as input ACGAN can generate many forms of speech balloons that appear as human-like drawings thanks to the innate randomness of GANs, and which are simultaneously controlled by arousal level. Furthermore, the ACGAN uses continuous values similarly to emotion detection software making it a practical choice for generating images related to emotional arousal.

3.5 Training process of ACGAN

The generator was trained by the rendered speech balloon images (Chapter 3.3) and the true images and their arousal levels were used as the labels. Here, the arousal levels are between 0.0 and 1.0. The discriminator got the speech balloon images and the fake speech balloon images, which were generated by the generator, and returned whether the input image was true or false and the prediction of arousal level. The model is trained by Adam optimizer ($\beta_1 = 0.5, \beta_2 = 0.999$) with the batch size of 64 and the learning rate of 0.0002. These hyper parameters (e.g., beta) were experimentally determined in previous studies [36, 48, 55]. We trained the model for 15 epochs.

The strength of the trained ACGAN is that it can generate a unique speech balloon image for any arousal value (0.00-1.00), and each time a value is input. As an example, five speech balloon images generated with five different arousal values are illustrated in Figure 6.



(a) arousal level = 0.1 (b) arousal level = 0.3 (c) arousal level = 0.5

(d) arousal level = 0.7 (e) arousal level = 0.9

Figure 6: Example speech balloon images generated by the generator

3.6 Evaluation of generated EmoBalloon images

The trained ACGAN was evaluated to confirm that the generated speech balloon image corresponds to emotional arousal level. That is, the ACGAN should generate a speech balloon image that is perceived as a higher arousal level by humans when the input arousal level increases. As a test data set, we generated five speech balloon images with random noise for arousal labels 0.00, 0.25, 0.50, 0.75 and 1.00. The Japanese text "おはよう", which means "good morning" in English, was placed in the middle of those images.

Figure 7: Example question in the pairwise comparison task

These five images were evaluated by 210 crowdsourcing workers hired via a Japanese crowdsourcing platform Yahoo! Crowdsourcing[4]. The crowdworkers were asked to complete a pairwise comparison of the speech balloons at different arousal levels. For this task, they answered a question "which [of the images] seems higher in arousal" in Japanese (Figure 7). Each worker completed

[3]https://helpx.adobe.com/illustrator/using/blending-objects.html/

[4]https://crowdsourcing.yahoo.co.jp/

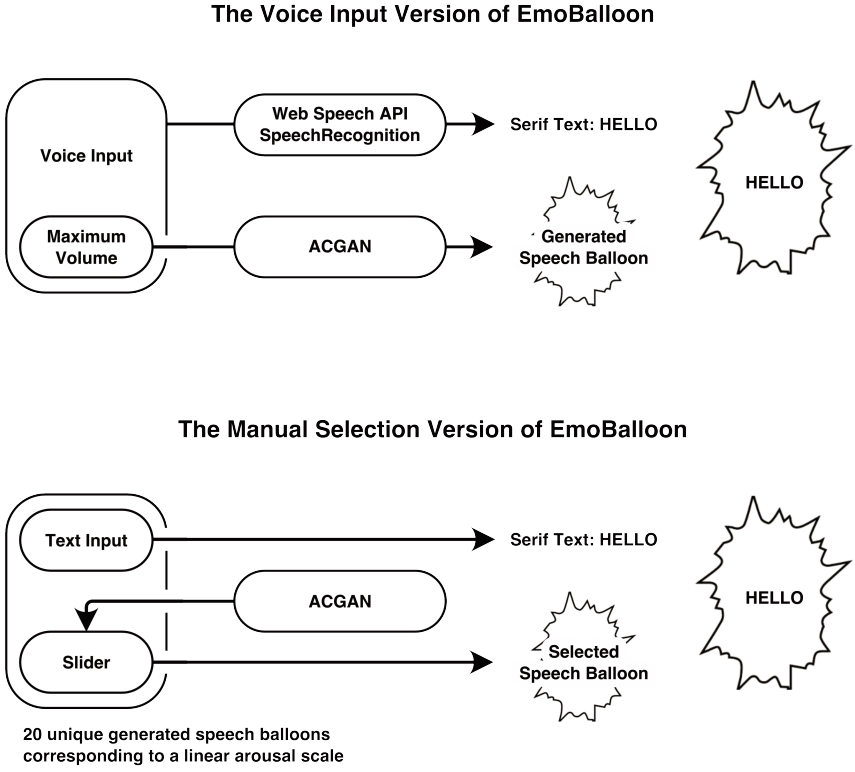


Figure 8: Procedure for sending messages with EmoBalloon

the pairwise comparison for all combinations of the five generated images (10 pairs) only once. The order of the speech balloon image pairs were randomized between seven crowdworker tasks, and 30 crowdworkers finished each task.

The results were analyzed using the Bradley-Terry-Luce (BTL) model [6] to calculate the ranking from the pairwise comparison data. By this analysis, the ranking of the speech balloons based on the level of arousal labels was in the order from highest to lowest: 1.00, 0.75, 0.50, 0.00, 0.25. Over 91% of the crowdworkers' evaluations ranked the 0.25 label's images and the 0.00 label's images as lower arousal in pairwise comparison with any of the other three labels. Furthermore, out of the pairwise comparisons between the 0.25 label's image and the 0.00 label's image, only 37% of evaluations indicated that the 0.00 label's image was higher in arousal.

These results indicated that the ACGAN's generator produced speech balloon images that correspond to high emotional arousal levels. However, it is important to also highlight that the results do not suggest that the generated speech balloon images were incorrect for low arousal levels.

4 IMPLEMENTATION

Figure 8 illustrates the overview of the EmoBalloon voice input text chat application system design. The reason we chose a voice input text chat as our target platform is two-fold. Firstly, speech to text input is already supported in many smartphone and desktop applications including text-based chats such as WeChat[5] or WhatsApp[6], where users who are unable or unwilling to listen to voice messages can read the transcribed text version instead [3]. Furthermore, speech to text technology is used in specialized applications to aid hearing impaired users to interact with voice messages or telephone calls [17, 51]. More importantly, with current smartphone and PC technology, it is possible to simultaneously convert speech to text messages and detect the emotional information based on the characteristics of a user's voice (e.g., volume, emotional tone[7]).

In the voice input EmoBalloon application, the user (message sender) speaks their message out loud, and the application detects both the speech content with speech recognition software, as well as the speaking volume. Speaking volume corresponds to the level of arousal [20, 27, 40], and furthermore, is easy to control by the

[5]https://web.wechat.com/
[6]https://play.google.com/store/apps/details?id=com.bongappstore9.voice_typing/
[7]https://webempath.net/lp-eng/

user. Hence, the speaking volume in which the message sender inputs their message is considered as the level of arousal in the current voice input implementation of EmoBalloon. Based on the detected speech content (text-based message) and speaking volume (level of arousal), the system generates a unique speech balloon image, attaches the text message inside the speech balloon and sends it to the receiver.

Figure 8 also illustrates the system design for manual selection version of EmoBalloon. We created the manual selection implementation for the purposes of this study, and to investigate differences between the two input methods (i.e., voice input vs. manual selection) for emotional speech balloons. For the manual selection, we generated 20 unique speech balloons corresponding to a linear emotional arousal scale. The user can select the speech balloon for their message using a slider before sending it to the receiver. The main difference between the two versions is that the voice input EmoBalloon generates unique speech balloons each time a user inputs a message, whereas in the manual selection version the user is limited to the pre-selected range of speech balloons that need to be manually selected.

5 USER STUDY

We conducted a controlled experiment with a mixed design, where message richness (paralinguistic cues added vs. text only) was set as the within-subjects factor and paralinguistic cue type (voice input EmoBalloon vs. manual selection EmoBalloon vs. emoticon) was set as the between-subjects factor. The objective of the user study was to test the efficacy of EmoBalloon in interactive settings, and to explore three hypotheses and two research questions regarding the relationship between speech balloon shapes and socio-emotional communication in text-based CMC.

Regarding the communicative aspects of using emotional speech balloons in text-based chats, we expected that the speech balloons generated by EmoBalloon would increase the emotional arousal conveyed in text-based messages. Previous research suggests that authors hear their own voice when authoring emotionally arousing messages [37, 39], whereas added paralinguistic cues strengthen the intensity of emotional text messages [16, 65]. In voice input text chats, the authors need to vocalize their messages instead of "hearing" them in their head, and we hypothesized that visualizing this information in the form of a speech balloon would cause senders to produce messages with higher level of emotional arousal. Furthermore, based on our own evaluation (Chapter 3.2) as well as previous work [71], we expected that the message receivers would perceive these messages as more emotionally arousing when displayed with speech balloons generated by EmoBalloon regardless of the input type.

> H1: EmoBalloon causes senders to produce voice input text messages with higher level of emotional arousal.
>
> H2: EmoBalloon increases the level of emotional arousal perceived in text-based messages by the message receiver.

In text-based communication, message senders and receivers tend to have disparate perceptions regarding the emotional tone of a message (i.e., negativity and neutrality effects) [9]. With the addition of the generated speech balloons, we hypothesized that the sender and the receiver would reach a higher level of agreement regarding the emotional arousal in a message compared to a chat without EmoBalloon functionality [71].

> H3: EmoBalloon increases the mutual understanding between senders and receivers regarding the level of emotional arousal in text-based messages.

We further wanted to explore the relationship of speech balloons to other paralinguistic cues in text chats, specifically emoticons. Emoticons are associated with higher levels of emotional transmission in text messages [31, 33, 45, 53, 54, 74], and can strengthen the intensity of the written emotional message [16, 65]. Hence, we wanted to investigate whether the use of EmoBalloon would produce similar or dissimilar effects to emoticons in text-based chats.

> RQ1: How does EmoBalloon compare to emoticon use as a paralinguistic cue to convey emotional information in text-based messages?

Lastly, we wanted to explore whether the use of EmoBalloon would influence the message senders and receivers' perceptions about the emotional valence of messages in an interactive setting.

> RQ2: How does EmoBalloon affect senders and receivers' perceptions about the emotional valence in voice-input text messages?

We implemented the three text chat interfaces (voice input EmoBalloons, manual selection EmoBalloon and emoticon) as web apps

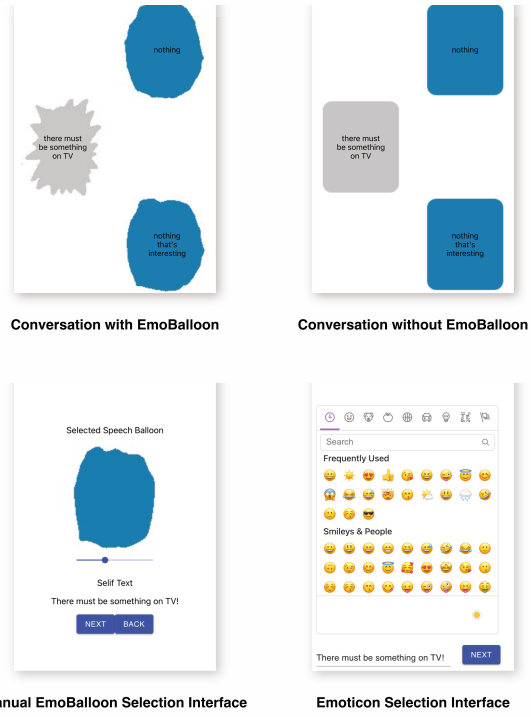


Figure 9: Example conversation from the user study with EmoBalloon and without EmoBalloon (top), and interface for manual EmoBalloon and emoticon selection (bottom)

Table 3: Example dialogue with Japanese translation

Title: What's on TV?
A: 退屈だよ(I'm bored.)
B: テレビで何やってるの？(What's on TV?)
A: 何も(Nothing.)
B: 絶対何かはやっているよ!(There must be something on TV!)
A: 面白いもの何もないんだよね(Nothing that's interesting.)
B: 新しいテレビ番組は？(What about that new game show?)
A: なんのこと？(Which one?)
B: 「ディール・オア・ノーディール」のこと("Deal or No Deal")
A: 冗談でしょう？(Tell me you're joking.)
B: この番組好きなんだよね。(I love that show.)
A: 一回見たことあるけど、もういいかな(I watched it once. That was enough.)
B: 今やってるよ。一緒にみようよ。(It's on right now. Let's watch it together.)

using Google Cloud Functions[8] with the trained ACGAN's generator implemented as API for the voice input EmoBalloon version. Figure 9 (top left) illustrates an example conversation with speech balloons generated by EmoBalloon. The manual selection EmoBalloon version used a manual slider selector for the 20 pre-generated speech balloons, whereas the emoticon selection interface was implemented using Emoji-Mart [9] with Apple-style emoticons (Figure 9, bottom).

5.1 Method

We hired 62 Japanese participants for this experiment (28 male, 34 female). The participants were randomly assigned to pairs, and 12 pairs participated in the voice input EmoBalloon condition, 10 pairs participated in the manual selection EmoBalloon condition and 9 pairs participated in the emoticon condition. Their mean age was 23.52 (SD = 4.04) years. None of the participants had any background information about the experiment before joining the study. They were compensated for their participation in accordance with the rules of our university.

We evaluated the efficacy of EmoBalloon by using a scripted role play scenario. The reason for choosing a scripted scenario instead of testing the system in a more naturalistic setting was to control the dialogue content between conditions and pairs, as well as to ensure that the evaluation data included emotionally arousing messages. The role play scenario was selected from the ESL/EFL Easy Conversation corpus, which included various dialogue examples [56]. The dialogues were translated to Japanese by the authors. Each pair completed a total of three scenarios. An example of the dialogue script is illustrated in Table 3.

All participants communicated with each other remotely, and only by sending messages with the communication tools developed for this experiment. In the voice input EmoBalloon condition, the experiment participants were asked to act out (i.e., not just read aloud) the dialogue via a voice input text chat. They were also told that a speech balloon was generated based on the volume of their voice, and that they could revise their input if desired. In the manual selection EmoBalloon and emoticon conditions, the participants were asked to write the scripted messages on a keyboard and then use a speech bubble or emoticon that best conveyed their intended emotional nuance for the message. At the time of sending each message, the sender was asked to indicate the intended level of emotional arousal and the intended level of emotional valence by selecting one of the five manikins (5-point scale), respectively, in the Self-Assessment Manikin (SAM) test developed by Bradley and Lang [5]. When the message receiver read the message, they were also asked to evaluate how they perceived the level of arousal and valence in the message using the same two scales. Once the message receiver evaluated the message, they moved on to sending the next line in the dialogue. This process was repeated until the pair finished the given script.

Each pair finished the task in two conditions: with and without paralinguistic cues (i.e., voice input EmoBalloon, manual selection EmoBalloon or emoticons). The pairs completed the second condition after 14 days or more had passed. The 14 day interval was set to allow the participants to forget their evaluations about the level of arousal and valence for each message during the first condition. The order of conditions was counterbalanced between pairs, and the roles within pairs for the dialogue (i.e., A and B in Table 3) were reversed during the second condition.

5.2 Results

To test our hypotheses and explore our research questions, we first created the measures for sender and receiver arousal and valence evaluations by calculating the average arousal and valence scores for each message (Cronbach's α >=.77). Since each pair acted out three scenarios with and without paralinguistic cues, and each scenario included a total of 12 messages, our data consists of evaluations for a total of 432 messages in the voice input EmoBalloon condition, 359[10] messages in the manual selection EmoBalloon condition, and 324 messages in the emoticon condition.

We conducted 2 (message richness: Text-only vs. Paralinguistic cues added) × 3 (Paralinguistic cue type: Voice input EmoBalloon vs. Manual selection EmoBalloon vs. Emoticon) Mixed ANOVAs (α = .05) to analyze the mean arousal evaluations for each message. The

[8] https://cloud.google.com/functions/
[9] https://github.com/missive/emoji-mart/
[10] Evaluations for one message were not recorded due to software malfunction.

Table 4: The means and standard deviations of arousal evaluations

	Voice input EmoBalloon		Manual selection EmoBalloon		Emoticon	
	With paralinguistic cues	Without paralinguistic cues*	With paralinguistic cues	Without paralinguistic cues*	With paralinguistic cues	Without paralinguistic cues*
(a) arousal evaluated by sender	3.20 (0.65)	3.14 (0.64)	3.22 (0.73)	2.99 (0.68)	3.06 (0.81)	2.76 (0.78)
(b) arousal evaluated by receiver	3.16 (0.57)	3.01 (0.69)	3.24 (0.73)	2.91 (0.64)	3.13 (0.71)	2.77 (0.72)
(c) the arousal difference between sender and receiver	0.91 (0.21)	1.06 (0.26)	0.77 (0.25)	0.99 (0.26)	0.99 (0.29)	0.97 (0.26)

data for valence evaluations was not normally distributed (Shapiro-Wilk, $p < .05$), and hence, we conducted our statistical analyses on emotional valence evaluations with non-parametric tests and report on the median values for this data.

Table 4 shows the means and standard deviations for emotional arousal evaluations in each condition.

Our first hypothesis concerned whether EmoBalloon causes senders to produce messages with higher level of arousal (Figure 10). Results from a Mixed ANOVA showed that there was a significant main effect for message richness ($F[1, 105] = 38.62$, $p < .001$) on the mean sender arousal score. The main effect of paralinguistic cue type ($F[2, 105] = 1.38$, $p = .26$) was not significant, but there was a significant interaction effect between message richness and paralinguistic cue type ($F[2, 105] = 5.16$, $p = .007$). Results from planned pairwise comparisons indicated that adding paralinguistic cues to text messages did not increase the sender emotional arousal ratings in the voice input EmoBalloon ($t[35] = 1.12$, $p = .27$) condition, but did so for the manual selection EmoBalloon ($t[35] = 3.86$, $p < .001$) and emoticon conditions ($t[35] = 5.78$, $p < .001$). H1 was not supported.

Next, we tested our second hypothesis on whether the receivers perceived the messages as more emotionally arousing with EmoBalloon than without (Figure 11). Results from a Mixed ANOVA showed a significant main effect for message richness ($F[1, 105] = 55.51$, $p < .001$) on the mean receiver arousal score. The main effect of paralinguistic cue type ($F[2, 105] = 0.48$, $p = .620$) was not significant, but there was a significant interaction effect between message richness and paralinguistic cue type ($F[2, 105] = 3.29$, $p = .041$). Results from planned pairwise comparisons indicated that adding paralinguistic cues to text messages increased the receivers' emotional arousal ratings in the voice input EmoBalloon ($t[35] = 2.17$, $p = .037$), the manual selection EmoBalloon ($t[35] = 4.95$, $p < .001$) as well as the emoticon ($t[35] = 5.93$, $p < .001$) conditions. H2 was supported.

Our third hypothesis predicted that EmoBalloon would decrease the differences in senders' and receivers' perceptions about the emotional arousal in messages (Figure 12). Results from a Mixed ANOVA showed that there was a significant main effect for message richness ($F[1, 105] = 13.05$, $p < .001$) on the sender and receiver agreement about emotional arousal in a message. Moreover, the main effect of paralinguistic cue type ($F[2, 105] = 3.47$, $p = .0.35$)

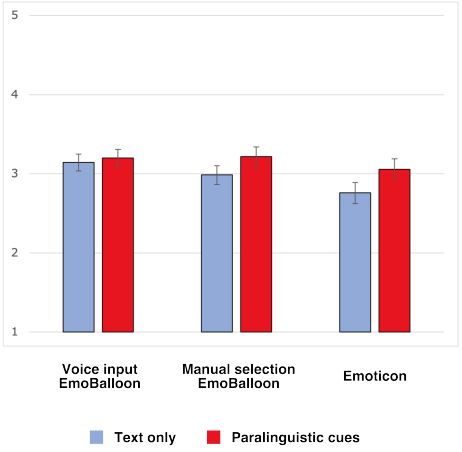


Figure 10: Mean arousal evaluations by the sender. Error bars represent the standard error of the mean.

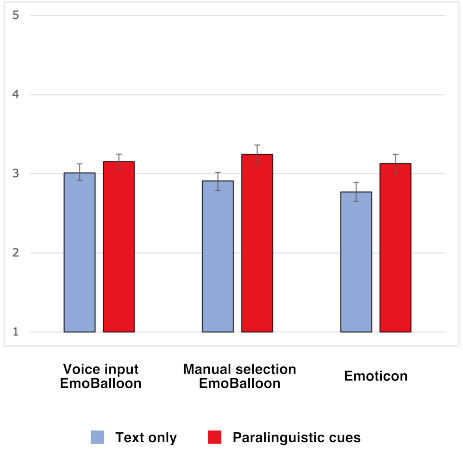


Figure 11: Mean arousal evaluations by the receiver. Error bars represent the standard error of the mean.

Table 5: The median valence evaluations

	Voice input EmoBalloon			Manual selection EmoBalloon			Emoticon		
	With paralinguistic cues	Without paralinguistic cues	p-value[11]	With paralinguistic cues	Without paralinguistic cues	p-value[11]	With paralinguistic cues	Without paralinguistic cues	p-value[11]
(a) valence evaluated by sender	2.25	2.33	.705	2.35	2.30	.480	2.44	2.00	.510
(b) valence evaluated by receiver	2.16	2.08	.237	2.40	2.20	.621	2.56	2.00	.048
(c) the valence difference between sender and receiver	0.25	0.33	.715	0.26	0.31	.857	0.24	0.24	.157

as well as the interaction effect between message richness and paralinguistic cue type ($F[2, 105] = 4.80$, $p = .010$) were statistically significant. Results from planned pairwise comparisons indicated that including paralinguistic cues in messages decreased the differences between sender and receiver arousal evaluations in the voice input EmoBalloon condition ($t[35] = 3.03$, $p = .005$) and the manual selection EmoBalloon condition ($t[35] = 3.67$, $p < .001$), but not in the emoticon condition ($t[35] = 0.31$, $p = .76$). H3 was supported.

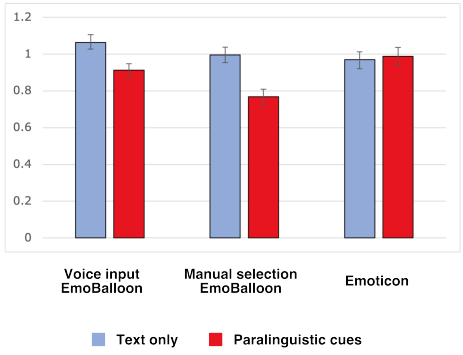


Figure 12: Mean difference in arousal evaluations between senders and receivers. Error bars represent the standard error of the mean.

The above results also partially answered our first research question regarding the differences between speech balloon shapes and emoticons used as paralinguistic socio-emotional cues. EmoBalloon was more effective in reducing the differences between message senders' and receivers' perceptions about the level of emotional arousal in text messages compared to emoticon use.

Our second research question concerned the effects of speech balloons on the conveyed and perceived emotional valence in messages. To explore this research question, we tested the differences in sender and receiver emotional valence evaluations with and without paralinguistic cues. Table 5 shows the median valence evaluations with and without paralinguistic cues in each condition. Results from Friedman tests ($\alpha = .05$) indicated no significant differences in any of the comparisons for either voice input EmoBalloon or manual selection EmoBalloon. Conversely, the receivers' valence evaluations for messages sent with emoticons were significantly higher than for messages sent without emoticons ($\chi^2[1]$=18.94, p<.001). These results further answered our RQ1 as well as RQ2. Unlike emoticons, speech balloons generated by EmoBalloon did not have a significant effect on the senders or receivers' perceptions regarding the emotional valence in text-based messages.

6 DISCUSSION

We introduced EmoBalloon, a voice input text chat application which translates a message sender's emotional arousal to a corresponding speech balloon. Besides describing our system design, our goal was to shed light on how speech balloon shapes may affect socio-emotional communication in text-based CMC, where the lack of communicative cues makes it challenging for conversational partners to convey and accurately detect the emotional nuances in messages [59, 62]. Our results demonstrated how the explosion shape speech balloon is associated with high arousal speech, and how the use of speech balloon shapes as paralinguistic cues in text chats can influence senders and receiver's perceptions about the emotional tone of text-based messages. Moreover, our analyses confirmed the efficacy of conveying emotional arousal in speech balloons generated based on message senders' speech volume in voice input text chats.

The results from our analysis on the relationship between emotional tone of speech in Japanese manga and the shape of speech balloons associated with the written speech text contributed to the findings in previous work [71]. That is, using an independent data set from [71] we found that compared to a round shape speech balloon, the explosion shape speech balloons were associated with high emotional arousal dialogue (Table 1), and these dialogues included more exclamation marks (Table 2) which further indicated a higher level of emotional arousal in text [2, 25, 65]. Moreover, our

[11] Friedman test

analysis suggested a linear relationship between more "explosion-like" speech balloons and level of emotional arousal conveyed in the design (Figure 6).

In our controlled experiment, we compared the effects of speech balloons and emoticons used as paralinguistic cues on senders and receivers' perceptions of emotional arousal and valence in text messages. Firstly, we expected that emotional speech balloons generated with voice input would cause senders to produce messages with higher level of arousal as they are able to hear and monitor their own voice and the generated speech balloon shape when authoring emotional messages [37, 39]. Contrary to our first hypothesis, we found no evidence that voice input would affect the senders' evaluations of the level of emotional arousal in text messages, but rather our results suggested that manually inputting paralingustic cues (either speech balloons or emoticons) increased the senders' perceived emotional arousal (Figure 10). One possible explanation for this result may be that the speech balloons generated by voice input EmoBalloon acted as feedback to the message sender. That is, when users manually select either a speech balloon or emoticon, these paralinguistic cues are used to selectively convey emotional information to the receiver who craft a mental image of the person based on the limited and curated cues [64, 66]. But, speech balloons generated by voice input may be perceived first as a reflection of their own emotional state by the sender, and only secondly as a paralinguistic cue to enhance a text-based message to be sent to the receiver. In other words, senders using the voice input EmoBalloon may have been worried about a type of "negative emotional leak", as described in [44], since they were aware that the receiver knew the speech balloons were generated based on the senders voice rather than selected by them manually. However, this finding should be investigated in more detail in future works as well as in other CMC contexts, such as giving feedback to users writing high arousal messages to online message boards (e.g., flaming, harassment) with a system similar to EmoBalloon.

Our second hypothesis concerned the effects of speech balloon shapes on the message receiver's perception of the emotional arousal in a text message. Our results were in line with previous works on both speech balloons [71] and emoticon use[16, 31, 33, 45, 53, 54, 65, 74], where we found that both speech balloons generated by EmoBalloon and emoticons increased the level of emotional arousal in a message for the receivers (Figure 11). This result confirmed the efficacy of our approach. That is, emotional speech balloons can be used as paralinguistic cues similarly to emoticons to influence the message receiver's perception regarding the emotional arousal in a message regardless whether the speech balloons are automatically generated or selected by the sender.

However, when investigating the disparities between speech balloons and emoticons as paralinguistic cues in more detail (RQ1 and RQ2), we found that these cues may function differently in influencing the sender and receiver perceptions regarding the level of emotional arousal and valence in text messages. Firstly, our analysis results showed that using speech balloons to convey emotional arousal significantly increased the agreement between senders and receivers regarding the level of arousal in messages (H3), but this was not the case for emoticons (Figure 12). Conversely, we found that speech balloons did not have an effect on senders or receivers' perceptions regarding the emotional valence in text messages (i.e., whether message is more positive or more negative). However, similarly to findings in previous research [16, 31, 33, 45, 53, 54, 65, 74], emoticons increased the intensity of receivers' emotional valence evaluations (Table 5).

These results highlight the potential of speech balloon shapes to convey specific emotional information in text based chats - emotional arousal. Presently, emoticons are ubiquitous on text chat platforms, and have a long history in enhancing online communications. However, the richness of emoticons may also be seen as one drawback. That is, humans tend to map emoticons on both emotional valence and arousal scales [28], and as seen from our experiment results, emoticons can affect the receivers' perceptions on both emotional valence and arousal simultaneously. Moreover, the interpretation of specific emoticons is also dependent on the user's background [70]. Contrary to emoticons, our results indicated that speech balloons can be used to enhance emotional arousal exclusively. This property could, for example, allow users to convey their level of excitement via text without influencing the receiver's positive-negative evaluations with emoticon use (i.e., accidental sarcasm).

Overall, our results suggest that EmoBalloon can enhance the emotional arousal in both voice input and traditional text chats (e.g., WeChat[1], WhatsApp [2]). Particularly, hearing impaired users could benefit from textual transcription of voice messages and phone calls, where the level of emotional arousal of the speaker/sender is preserved in speech balloons [17, 51]. However, it is also important to consider the level of control that users have over their emotional expression in text-based CMC. In lean CMC mediums, individuals are able to manipulate their self-presentation (i.e., how others view them) due to lack of communicative cues, and automated approaches that convey more detailed emotional information may lead to fears of "negative emotion leak" [44] [64]. That is, while EmoBalloon may be useful for message receivers to determine the level of emotional arousal in text, it is also important to consider how different input methods influence the message senders perceptions and behavior in future works.

7 LIMITATIONS AND FUTURE WORK

In the present study, we focused on two types of speech balloon shapes based on previous works; the explosion shape and the round shape [60, 71]. In future work, we are interested in exploring the effects of different speech balloon shapes on socio-emotional communication in text-based CMC, particularly the effects of the cloud and polygon shapes which are associated with positive emotions [60]. Furthermore, since current affective computing technologies, such as emotion analysis in text, perform the best with English language, we automatically translated our dataset for analysis. In the future, we aim to implement our approach without translating the text data and by using emotion detection software tuned for the original language.

We are also interested in combining our system with supporting technologies aimed to improve emotional valence detection in text-based CMC. Namely, our results suggested that the speech balloon shape may not affect how emotional valence is perceived in text messages. Hence, aside for emoticons, combining EmoBalloon with novel systems that influence the receivers' perceptions about the

emotional valence of messages by using emotional typefaces could be a viable approach to decrease the negativity and neutrality effects in text-based CMC [9, 12, 73]. Furthermore, we are interested in investigating how other types of visualizations, such as color of the speech bubble [11], may affect the users' perceptions on the emotional valence and level of arousal in text-based messages sent with EmoBalloon.

Our experimental design included some limitations in terms of the generalizability of our results to other languages and cultures. Since our system was trained with data from Japanese manga (Manga109 [42]), the experiment participants, all being Japanese, were aware of the speech balloon styles, but future studies should include users from other language and cultural backgrounds as well. Furthermore, Western comics may have different speech balloon styles conveying different emotions or level of emotional arousal, which if used as training data could lead to different results. Furthermore, the participants' cultural background may have an effect on how they perceive the emoticons used in this study. As reported in previous studies, Japanese users may interpret emoticons differently from Western users [70], and hence, future studies should investigate whether the reported effects would persists in cross-cultural settings.

We are also planning on conducting a more naturalistic studies using EmoBalloon. That is, we designed our experiment to produce emotionally arousing messages by the use of emotional scripts in a controlled environment (i.e., no background noise, static location), and asking the participants to act out a dialogue. The participants were also aware that their messages were being evaluated for arousal and valence, which may further influence the results. Hence, our plan is to conduct a longitudinal study with voice input EmoBalloon, where the experiment participants would use the system for daily communication with people whom they already have a relationship with in a naturalistic environment while also covering more complex emotion models beyond valence and arousal.

8 CONCLUSION

In this paper, we introduced EmoBalloon; a system which automatically translates a message sender's level of emotional arousal to speech balloons in voice input text chats. Our system was designed based on analysis results regarding the relationship between speech balloon shape and the emotional arousal and valence of dialogue text in Japanese manga. The system was realized with the ACGAN, which was trained with a dataset of speech balloons derived from Manga109 dataset [42] and evaluated through crowdsourcing. We evaluated the efficacy of our system to support socio-emotional text-based communication in a controlled experiment. Our results suggested that speech balloon shape can affect the message receiver's perception regarding the level of emotional arousal conveyed in a message, as well as increase agreement about the arousal level between a message sender and a receiver. Moreover, our results highlighted the differences between speech balloons and emoticons used as paralinguistic cues in text based chats. Our findings contributed to our understanding of the social information processing in text-based CMC mediums, and gave insights on future directions to enhance socio-emotional communication in text chat applications.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP18K18085. This research is part of the results of Value Exchange Engineering, a joint research project between Mercari, Inc. and the RIISE.

REFERENCES

[1] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. 2020. Building a Manga Dataset "Manga109" with Annotations for Multimedia Applications. *IEEE MultiMedia* 27, 2 (2020), 8–18. https://doi.org/10.1109/mmul.2020.2987895

[2] Saima Aman and Stan Szpakowicz. 2007. Identifying Expressions of Emotion in Text. In *Text, Speech and Dialogue*, Václav Matoušek and Pavel Mautner (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 196–205.

[3] Tom Arnstein. 2019. Never Listen to Another Voice Message Again: WeChat Rolls Out Awesome Transcription Function. https://www.thebeijinger.com/blog/2019/09/12/never-listen-another-voice-message-again-wechat-rolls-out-awesome-transcription

[4] Sigal G Barsade. 2002. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative science quarterly* 47, 4 (2002), 644–675. https://doi.org/10.2307/3094912

[5] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[6] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.

[7] Tony W Buchanan, Kai Lutz, Shahram Mirzazade, Karsten Specht, N Jon Shah, Karl Zilles, and Lutz Jäncke. 2000. Recognition of emotional prosody and verbal components of spoken language: an fMRI study. *Cognitive Brain Research* 9, 3 (2000), 227–238. https://doi.org/10.1016/S0926-6410(99)00060-9

[8] Daniel Buschek, Mariam Hassib, and Florian Alt. 2018. Personal Mobile Messaging in Context: Chat Augmentations for Expressiveness and Awareness. *ACM Trans. Comput.-Hum. Interact.* 25, 4, Article 23 (aug 2018), 33 pages. https://doi.org/10.1145/3201404

[9] Kristin Byron. 2008. Carrying too heavy a load? The communication and miscommunication of emotion by email. *Academy of Management Review* 33, 2 (apr 2008), 309–327. https://doi.org/10.5465/AMR.2008.31193163

[10] Fanglin Chen, Kewei Xia, Karan Dhabalia, and Jason I. Hong. 2019. *MessageOnTap: A Suggestive Interface to Facilitate Messaging-Related Tasks.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300805

[11] 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.* Association for Computing Machinery, New York, NY, USA, Article 361, 6 pages. https://doi.org/10.1145/3411763.3451698

[12] Saemi Choi and Kiyoharu Aizawa. 2019. Emotype: Expressing emotions by changing typeface in mobile messenger texting. *Multimedia Tools and Applications* 78, 11 (jun 2019), 14155–14172. https://doi.org/10.1007/s11042-018-6753-3

[13] Richard L. Daft and Robert H. Lengel. 1986. Organizational Information Requirements, Media Richness and Structural Design. *Management Science* 32, 5 (may 1986), 554–571. https://doi.org/10.1287/mnsc.32.5.554

[14] Jose Eurico De Vasconcelos Filho, Kori M. Inkpen, and Mary Czerwinski. 2009. Image, appearance and vanity in the use of media spaces and videoconference systems. In *GROUP'09 - Proceedings of the 2009 ACM SIGCHI International Conference on Supporting Group Work.* ACM Press, New York, New York, USA, 253–261. https://doi.org/10.1145/1531674.1531712

[15] Alan R Dennis and Susan T Kinney. 1998. Testing media richness theory in the new media: The effects of cues, feedback, and task equivocality. *Information systems research* 9, 3 (1998), 256–274. https://doi.org/10.1287/isre.9.3.256

[16] Daantje Derks, Arjan E R Bos, and Jasper Von Grumbkow. 2008. Emoticons and Online Message Interpretation. *Social Science Computer Review* 26 (2008), 379–388. https://doi.org/10.1177/0894439307311611

[17] NTT DOCOMO. 2021. みえる電話[in Japanese]. https://www.nttdocomo.co.jp/service/mieru_denwa/index.html

[18] Paul Ekman. 1982. Methods for measuring facial action. *Handbook of Methods in Nonverbal Behavior Research*, 44–90.

[19] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. 2017. CAN: Creative Adversarial Networks Generating "Art" by Learning About Styles and Deviating from Style Norms *. arXiv:1706.07068v1

[20] Robert W. Frick. 1985. Communicating Emotion. The Role of Prosodic Features. 97, 3 (may 1985), 412–429. https://doi.org/10.1037/0033-2909.97.3.412

[21] Raymond A Friedman and Steven C Currall. 2003. Conflict escalation: Dispute exacerbating elements of e-mail communication. *Human relations* 56, 11 (2003), 1325–1347.

[22] Susan R Fussell. 2002. Introduction and overview. In *The verbal communication of emotions*, Susan R Fussell (Ed.). Psychology Press, New York, NY, USA, 9–24.

[23] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) *(NIPS'14)*. MIT Press, Cambridge, MA, USA, 2672–2680.

[24] Jeffrey T. Hancock, Kailyn Gee, Kevin Ciaccio, and Jennifer Mae-Hwah Lin. 2008. I'm Sad You're Sad: Emotional Contagion in CMC. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (San Diego, CA, USA) *(CSCW '08)*. Association for Computing Machinery, New York, NY, USA, 295–298. https://doi.org/10.1145/1460563.1460611

[25] Jeffrey T. Hancock, Christopher Landrigan, and Courtney Silver. 2007. Expressing Emotion in Text-Based Communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 929–932. https://doi.org/10.1145/1240624.1240764

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 770–778. https://doi.org/10.1109/CVPR.2016.90

[27] George L Huttar. 1968. Relations between prosodic variables and emotions in normal American English utterances. *Journal of Speech and Hearing Research* 11, 3 (1968), 481–487. https://doi.org/10.1044/jshr.1103.481

[28] Sara R. Jaeger, Christina M. Roigard, David Jin, Leticia Vidal, and Ares Gaston. 2019. Valence, arousal and sentiment meanings of 33 facial emoji: Insights for the use of emoji in consumer research. *Food Research International* 119 (may 2019), 895–907. https://doi.org/10.1016/j.foodres.2018.10.074

[29] Nikolay Jetchev and Urs Bergmann. 2017. The Conditional Analogy GAN: Swapping Fashion Articles on People Images. arXiv:1709.04695 [stat.ML]

[30] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. arXiv:1812.04948 [cs.NE]

[31] Shogo Kato, Yuuki Kato, and Kanji Akahori. 2006. Study on Emotional Transmissions in Communication Using Bulletin Board System. In *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2006*, Thomas Reeves and Shirley Yamashita (Eds.). Association for the Advancement of Computing in Education (AACE), Honolulu, Hawaii, USA, 2576–2584. https://www.learntechlib.org/p/24095

[32] Boaz Keysar and Anne S Henly. 2002. Speakers' overestimation of their effectiveness. *Psychological Science* 13, 3 (2002), 207–212. https://doi.org/10.1111/1467-9280.00439

[33] Sujay Khandekar, Joseph Higg, Yuanzhe Bian, Chae Won Ryu, Jerry O. Talton Iii, and Ranjitha Kumar. 2019. Opico: A Study of Emoji-first Communication in a Mobile Social App. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 450–458. https://doi.org/10.1145/3308560.3316547

[34] Joongyum Kim, Taesik Gong, Kyungsik Han, Juho Kim, JeongGil Ko, and Sung-Ju Lee. 2020. Messaging Beyond Texts with Real-Time Image Suggestions. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services* (Oldenburg, Germany) *(MobileHCI '20)*. Association for Computing Machinery, New York, NY, USA, Article 28, 12 pages. https://doi.org/10.1145/3379503.3403553

[35] Joongyum Kim, Taesik Gong, Bogoan Kim, Jaeyeon Park, Woojeong Kim, Evey Huang, Kyungsik Han, Juho Kim, Jeonggil Ko, and Sung-Ju Lee. 2020. No More One Liners: Bringing Context into Emoji Recommendations. *Trans. Soc. Comput.* 3, 2, Article 9 (apr 2020), 25 pages. https://doi.org/10.1145/3373146

[36] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[37] Ned Kock. 2005. Media richness or media naturalness? The evolution of our biological communication apparatus and its influence on our behavior toward e-communication tools. *IEEE transactions on professional communication* 48, 2 (2005), 117–130. https://doi.org/10.1109/TPC.2005.849649

[38] Yuko Konya, Akihiro Nakatani, Itaru Sato, and Itiro Siio. 2013. Paralinguistic 表現を用いた聴覚障害者向け吹き出し型字幕提示方法[in Japanese] Paralinguistic Caption Presentation method with Speech bubble for hearing impaired person. 研究報告エンタテインメントコンピューティング*(EC)* 2013-EC-29, 4 (2013), 1–6.

[39] Justin Kruger, Nicholas Epley, Jason Parker, and Zhi-Wen Ng. 2005. Egocentrism over e-mail: Can we communicate as well as we think? *Journal of personality and social psychology* 89, 6 (2005), 925. https://doi.org/10.1037/0022-3514.89.6.925

[40] Ito Kumiko. 1985. 感情を含む音声に関する研究II 合成単母音[え] による音響パラメータ評価[in Japanese]. 人間工学 21, 2 (1985), 81–87. https://doi.org/10.5100/jje.21.81

[41] David Kurlander, Tim Skelly, and David Salesin. 1996. Comic chat. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996*. Association for Computing Machinery, Inc, New York, New York, USA, 225–236. https://doi.org/10.1145/237170.237260

[42] Aizawa Yamasaki Matsui Lab. 2015. Manga109. http://www.manga109.org/ja/index.html

[43] Robert H Lengel and Richard L Daft. 1988. The selection of communication media as an executive skill. *Academy of Management Perspectives* 2, 3 (1988), 225–232. https://doi.org/10.5465/ame.1988.4277259

[44] Miki Liu, Austin Wong, Ruhi Pudipeddi, Betty Hou, David Wang, and Gary Hsieh. 2018. ReactionBot: Exploring the Effects of Expression-Triggered Emoji in Text Messages. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW, Article 110 (nov 2018), 16 pages. https://doi.org/10.1145/3274379

[45] Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the Ubiquitous Language: An Empirical Analysis of Emoji Usage of Smartphone Users. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 770–780. https://doi.org/10.1145/2971648.2971724

[46] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2017. Sketch-based Manga Retrieval using Manga109 Dataset. *Multimedia Tools and Applications* 76, 20 (2017), 21811–21838. https://doi.org/10.1007/s11042-016-4020-z

[47] Fusanosuke Natsume. 1997. マンガはなぜ面白いのか: その表現と文法*[in Japanese]*. NHK Publishing, Tokyo, Japan.

[48] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional Image Synthesis with Auxiliary Classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 2642–2651.

[49] James Ohene-Djan, Jenny Wright, and Kirsty Combie-Smith. 2007. Emotional Subtitles: A System and Potential Applications for Deaf and Hearing Impaired People. In *Proceedings of the Conference and Workshop on Assistive Technologies for People with Vision and Hearing Impairments: Assistive Technology for All Ages, CVHI-2007*. 415.

[50] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. In *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22. 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[51] Pedius. 2021. Pedius. https://www.pedius.org/

[52] Yi Hao Peng, Ming Wei Hsu, Paul Taele, Ting Yu Lin, Po En Lai, Leon Hsu, Tzu Chuan Chen, Te Yen Wu, Yu An Chen, Hsien Hui Tang, and Mike Y. Chen. 2018. Speechbubbles: Enhancing captioning experiences for Deaf and hard-of-hearing people in group conversations. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2018-April. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3173867

[53] Henning Pohl, Christian Domin, and Michael Rohs. 2017. Beyond just text: semantic emoji similarity modeling to support expressive communication. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 1 (2017), 1–42. https://doi.org/10.1145/3039685

[54] Robert R Provine, Robert J Spencer, and Darcy L Mandell. 2007. Emotional expression online: Emoticons punctuate website text messages. *Journal of Language and Social Psychology* 26, 3 (2007), 299–307. https://doi.org/10.1177/0261927X06303481

[55] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:1511.06434 [cs.LG]

[56] rong-chang ESL Inc. 2021. Easy Conversations For ESL/EFL Beginners. http://www.eslfast.com/easydialogs/

[57] James Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39 (dec 1980), 1161–1178. https://doi.org/10.1037/h0077714

[58] Jeremiah Scholl, John McCarthy, and Rikard Harr. 2006. A Comparison of Chat and Audio in Media Rich Environments. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work* (Banff, Alberta, Canada) *(CSCW '06)*. Association for Computing Machinery, New York, NY, USA, 323–332. https://doi.org/10.1145/1180875.1180925

[59] John Short, Ederyn Williams, and Bruce Christie. 1976. *The social psychology of telecommunications*. John Wiley & Sons, Hoboken, New Jersey, USA.

[60] Hideki Tanaka, Ryosuke Yamanishi, and Junichi Fukumoto. 2015. Relation Analysis between Speech Balloon Shapes and their Serif Descriptions in Comic. In *2015 IIAI 4th International Congress on Advanced Applied Informatics*. IEEE, New York, NY, USA, 229–233. https://doi.org/10.1109/IIAI-AAI.2015.235

[61] Philip A. Thompsen and Davis A. Foulger. 1996. Effects of pictographs and quoting on flaming in electronic mail. *Computers in Human Behavior* 12, 2 (1996), 225 – 243. https://doi.org/10.1016/0747-5632(96)00004-0

[62] Joseph B. Walther. 1992. Interpersonal Effects in Computer-Mediated Interaction. *Communication Research* 19, 1 (feb 1992), 52–90. https://doi.org/10.1177/009365092019001003

[63] Joseph B. Walther. 1994. Anticipated Ongoing Interaction Versus Channel Effects on Relational Communication in Computer-Mediated Interaction. *Human Communication Research* 20, 4 (jun 1994), 473–501. https://doi.org/10.1111/j.1468-

2958.1994.tb00332.x

[64] Joseph B. Walther. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication research* 23, 1 (1996), 3–43. https://doi.org/10.1177/009365096023001001

[65] Joseph B. Walther and Kyle P. D'Addario. 2001. The Impacts of Emoticons on Message Interpretation in Computer-Mediated Communication. *Social Science Computer Review* 19, 3 (aug 2001), 324–347. https://doi.org/10.1177/089443930101900307

[66] Joseph B. Walther, Yuhua Liang, David DeAndrea, Stephanie Tong, Caleb Carr, Erin Spottswood, and Yair Amichai-Hamburger. 2011. The Effect of Feedback on Identity Shift in Computer-Mediated Communication. *Media Psychology* 14 (mar 2011), 1–26. https://doi.org/10.1080/15213269.2010.547832

[67] Joseph B. Walther, Tracy Loh, and Laura Granka. 2005. Let me count the ways: The interchange of verbal and nonverbal cues in computer-mediated and face-to-face affinity. *Journal of language and social psychology* 24, 1 (2005), 36–65. https://doi.org/10.1177/0261927X04273036

[68] Hua Wang, Helmut Prendinger, and Takeo Igarashi. 2004. Communicating Emotions in Online Chat Using Physiological Sensors and Animated Text. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (Vienna, Austria) *(CHI EA '04)*. Association for Computing Machinery, New York, NY, USA, 1171–1174. https://doi.org/10.1145/985921.986016

[69] Hisako W. Yamamoto, Misako Kawahara, Mariska E. Kret, and Akihiro Tanaka. 2020. Cultural Differences in Emoticon Perception: Japanese See the Eyes and Dutch the Mouth of Emoticons. *Letters on Evolutionary Behavioral Science* 11, 2 (dec 2020), 40–45. https://doi.org/10.5178/lebs.2020.80

[70] Hisako W. Yamamoto, Misako Kawahara, Mariska E. Kret, and Akihiro Tanaka. 2020. Cultural Differences in Emoticon Perception: Japanese See the Eyes and Dutch the Mouth of Emoticons. *Letters on Evolutionary Behavioral Science* 11, 2 (dec 2020), 40–45. https://doi.org/10.5178/lebs.2020.80

[71] Ryosuke Yamanishi, Hideki Tanaka, Yoko Nishihara, and Junichi Fukumoto. 2017. Speech-balloon Shapes Estimation for Emotional Text Communication. *Information Engineering Express* 3, 2 (2017), 1–10. https://doi.org/10.52731/iee.v3.i2.168

[72] Dezhi Yin, Samuel D. Bond, and Han Zhang. 2017. Keep Your Cool or Let it Out: Nonlinear Effects of Expressed Arousal on Perceptions of Consumer Reviews. *Journal of Marketing Research* 54, 3 (2017), 447–463. https://doi.org/10.1509/jmr.13.0379

[73] Ryota Yonekura, Saemi Choi, Ryota Yoshihashi, Katsufumi Matsui, and Ari Hautasaari. 2019. Automated Font Selection System based on Message Sentiment in English Text-Based Chat [in Japanese]. *IEICE Technical Report* 118, 502 (2019), 131–136.

[74] Rui Zhou, Jasmine Hentschel, and Neha Kumar. 2017. Goodbye Text, Hello Emoji: Mobile Communication on WeChat in China. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 748–759. https://doi.org/10.1145/3025453.3025800