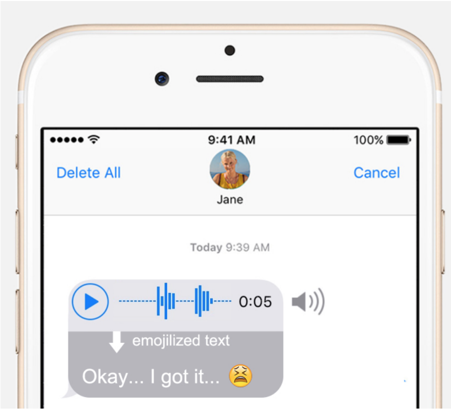

Okay... I got it... 😫

Figure 1: The example application of *emojilization*: the attached emoji compensates for the emotional expression of the voice which automatic speech recognition missed.

Emojilization: An Automated Method For Speech to Emoji-Labeled Text

Jiaxiong Hu
Qianyao Xu
Academy of Arts & Design,
Tsinghua University
Beijing, China
hujx16@mails.tsinghua.edu.cn
xuqy17@mails.tsinghua.edu.cn

Yingqing Xu
The Future Lab, Tsinghua University
Academy of Arts & Design,
Tsinghua University
Beijing, China
yqxu@mail.tsinghua.edu.cn

Limin Paul Fu
Natural HCI Lab, Alibaba
Sunnyvale, California, United States
paulfu@alibaba-inc.com

ABSTRACT

Speech To Text (STT) plays a significant role in Voice User Interface (VUI). While preserving necessary semantic information in converted text, STT generally captures no or limited emotional information. In this paper, we present an *emojilization* tool to automatically attach related emojis to the STT-generated

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3313071

texts by analyzing both textual and acoustic features in speech signals. For a given voice message, the tool selects the most representative emoji from 64 most commonly used emojis. We conducted a pilot study with 34 participants. In our study, 159 utterances were labeled with emojis by our tool. The emotion restoration effect was evaluated. The results indicate that the proposed tool effectively compensates for the *emotion loss*.

KEYWORDS Voice user interface; speech recognition; affective computing; emotions

INTRODUCTION

Voice user interface (VUI) is increasingly utilized in various scenarios. In general, many mobile applications such as WeChat (Tencent, 2011), or virtual assistants such as Alexa (Amazon, 2015) and TmallGenie (Alibaba, 2017) have embedded with VUI. In all of those scenarios, speech recognition is a key technology as it mediates between computational framework and voice user interface.

Speech recognition (Speech To Text, STT) maintains users' basic semantics[2]. However, users' emotional expression is not preserved adequately in many cases. The perceived emotion of the STT text often is not accurate, which sometimes leads to a misunderstanding between interlocutors.

Voice is an essential aspect of human social interaction [13]. From paralinguistic information in speech, listeners can judge emotions more accurately [3]. However, STT-generated texts contain no indication of emotional states for users or annotation for later Natural Language Process (NLP) in VUI. For a better discussion on the loss of emotional expression caused by STT, we define it in this paper as *emotion loss*. In this work, the *emojilization* is proposed to compensate for the *emotion loss*. By fusing semantic and acoustic features from speech signals, related emojis are attached to appropriate positions among texts. Emoji is chosen to represent emotions because of its popularity all over the world and its compatibility with Unicode Transformation Formats (UTF). We conducted a comparative study in which 34 participants rated perceived emotions from STT-generated texts, *emojilized texts*, and the original spoken audio (assumed as reference). The*emotion loss* was compared. The result shows that the *emojilization* tool compensates for the *emotion loss*.

RELATED WORKS

Speech recognition is a key technology of voice interface. Current automatic speech recognition techniques even outperform human in specific situations. E.g., for recognition of read speech, the word error rate of Amodei, Dario, et al.'s method is 3.10%, while 5.03% for human[2]. The excellent performance of STT ensures the usability of voice interface.

Numerous commercial products have adopted VUI, such as Siri(Apple, 2011). Martin Porcheron et al. studied the use of VUI in daily life. They identified the way in which VUI coordinates with the order of conversations, and proposed the design concept of VUI interactions, requests, and responses [11]. However, they have not introduced emotional information to the VUI system[7]. Essentially, the framework of these VUIs is based on natural language process, which requires the annotation of

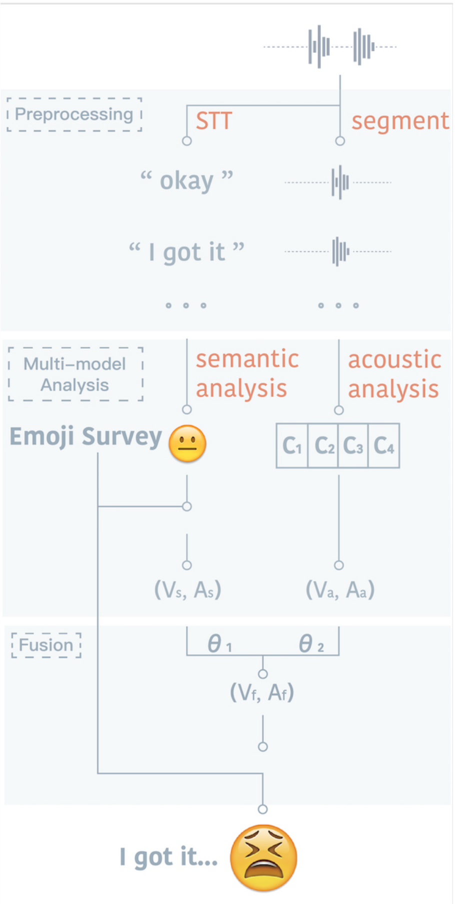

Figure 2: The system framework of *emojilization*. The *emojilization* tool considers both semantic and acoustic features in speech signals so that the generated emojis can reflect both the meaning of words and the emotional expression of voice.

emotional information [1]. The markup language in emotion-oriented computing systems is essential for data interpretation, reasoning, and behavior generation[14], which is our primary motivation of annotating text with emojis.

Latest affective computing techniques have proved the ability of computers to perceive the emotional states of users. At present, unimodal speech emotion recognition (SER) has reached the accuracy of 86.65% with time distributed convolutional neural network (with Emo-DB corpus) [8]. Multi-modal emotion recognition is a common method to perceive the user's emotional states comprehensively. Martin Wöllmer et al. have used the Bidirectional Long Short-Term Memory (LSTM) network in a multi-emotion recognition technology that is context-sensitive and based on two-modal fusion [16]. These works provide *emojilization* references about how to choose the right emoji by analyzing the speech signals.

EMOJILIZATION

To make the full use of information in speech signals, we proposed a tool to *emojilize* the STT process by labeling emotional expressions with emojis. As shown in Figure 2, speech signals are divided into texts and audio clips to take semantic and acoustic analysis separately, and then, get fused before outputting emojis. To integrate the system, emojis and emotions are both quantified through the Valence-Arousal model, a two-dimensional model, in which all emotions are distributed in[12].

Firstly, input speech signals need preprocessing where speech is converted into text by STT. The original audio is segmented into clips according to the pauses of the speech. Each audio clip and its STT text form a unit for later fusion of the two modalities.

Secondly, each unit requires acoustic and semantic analysis separately. For the audio clip, the LSTM-Recurrent Neural Network speech emotion recognition algorithm [9] is implemented and tuned to analyze the emotional states of the speaker. As for the STT text in the unit, we use DeepMoji[5] to analyze the emotional expression of the clause. Both approaches were proved effective in solving perceptual problems.

The most important part in *emojilization* is the fusion of acoustic and semantic modals. The outputs of the two approaches mentioned above are in two different forms. The SER model outputs confidence values (ranging from 0% to 100%) of four discrete emotions: joyful, sad, angry and calm. The result of DeepMoji model is a set of confidence values of five emojis that are most related to the STT text. The fusion strategy is to map the results of these two modals to the same Valence-Arousal coordinate(the VA coordinate as Figure 4). The VA coordinate is a two-dimensional valence (V; emotional pleasantness) and arousal (A; emotional activation/excitation) model [12]. The values of both V and A range from -3 to 3.

For the semantic modal, the most related emoji is selected to put back to the VA coordinate according to the result of an emoji interpretation survey that we conducted earlier referring to Tigwell

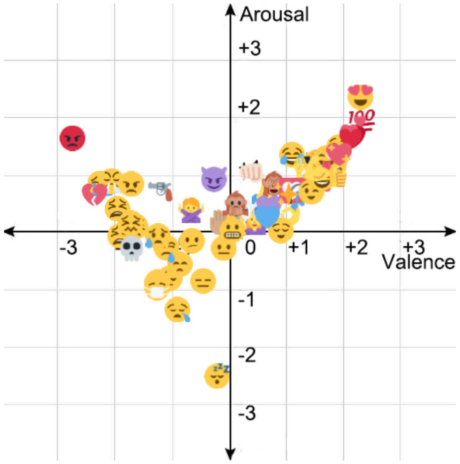

Figure 3: The result of the emoji survey.

Emoji Survey One hundred fifty-nine participants took an online survey to evaluate 64 emojis. Participants evaluate the emotional expression of each emoji on a web-based measuring tool, which is similar to the tool using in the lab study of this paper. The method of the survey is referred to Tigwell et al.'s work[15]. The result of the survey indicates that the meaning of some emojis vary a lot among participants, while some emojis are consistent. The major discovery is that emojis expressing apparent negative emotion show more substantial deviation of understanding. The emoji with most varying interpretation is "😡. Those express positive emotions tend to be comprehended consistently, e.g., "😊" is interpreted consistently. Please note that a few emojis such as "😈" are not interpreted consistently, so a point in the coordinate may not be the best representation for an emoji.

et al.'s work[15]. The result of the survey is presented as 4. For instance, if the most related emoji of the clause is "😊", then "the semantic VA coordinate" is (+1.5, +1.2). To better explain the computation process, we define the VA coordinate as "the semantic VA coordinate"(V_s, A_s). For the emojis that are interpreted more inconsistently, (V_s, A_s) could be an area in the VA coordinate. Emojis are semantically grouped by similarity according to Henning Pohl et al.'s work[10]. If the outputting emoji of the semantic model shows some specific meaning such as "👿", then the "😈" is alternative.

For the acoustic modal, the Long-Short Term Memory - Recurrent Neural Network provides four confidence values and four emotions which four base vectors are pre-defined: happy(1, 1), angry(-1, 1), sad(-1, -1), calm(1, -1) for this instance. "The acoustic VA coordinate" is the result the sum of each product between emotion confidence and its base vector, which means that the confidence values are served as the weights: $(V_a, A_a) = c_1 \times (1, 1) + c_2 \times (-1, 1) + c_3 \times (-1, -1) + c_4 \times (1, -1)$. A cubic function is optional to reduce the weight of the acoustic modal if the confidence is weak.

Finally "the fusion VA coordinate"(V_f, A_f) is the sum of "the semantic VA coordinate" and "the acoustic VA coordinate": $\theta_1 \times (V_s, A_s) + \theta_2 \times (V_a, A_a)$. In this instance, when $\theta_1 = 5$, $\theta_2 = 30$, the system performed well in the test. Consequently, the final output emoji is the one which has the least Euclidean distance to "the fusion VA coordinate"(V_f, A_f). Eventually, the chosen emoji is attached to the end of the clause. Note that if the confidence values of both modals are too weak, the tool will not attach any emoji.

Note the significant difference between DeepMoji model and proposed *emojilization* tool. If two sentences, which consist of the same words but are spoken in different tones, are inputted to the proposed tool, the output emojis of these two sentences are different. E.g., utterances such as "Okay, I got it." are neutral semantically, but if they are spoken angrily, the tool labels with a "😡" as well.

STUDY: TEXT, EMOJILIZED TEXT, AND SPEECH

In this study, we evaluated how effectively *emojilized* text can compensate for the *emotion loss* in the STT process.

Users may have different understandings of the same content delivered via text, *emojilized* text and speech. Note that the comprehension of speech is used as the ground truth. For instance, the emotional state may not be preserved as a result of converting highly aroused speech into text, which potentially letting the audience perceive contrasting emotions from the original speech and the converted text.

Three steps constituted our pilot study, each involving a type of conversation. Specifically, participants were asked to appraise their perceived emotions in response to conversations in three forms sequentially: text, *emojilized* text and speech. Speech was arranged at last because we assumed that participants perceive the most sufficient emotional information from speech. Conversations contained the same content but were presented differently: text and *emojilized* text were printed out on papers,

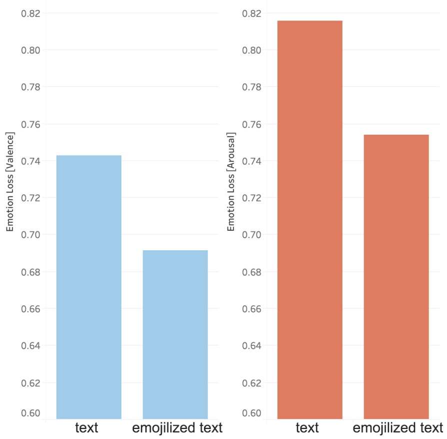

Figure 4: The average *emotion loss* reduced in both valence and arousal by *emojilization*.

Results: The intensity of emotional expression was represented as the moduli in the two-dimensional space of the VA model. Between plain text and speech, the intensity of emotional expression was significantly different (F(1,66)=27.778, p<0.05 using ANOVA), corroborating the existence of *emotion loss*.

We tested the role of *emojilization* by means of sentences as the subjects of our research. In the experiment, a total of 159 *emojilized* sentences were employed. The *emotion loss* at the valence level reduced from a mean of 0.743 to 0.692, an overall improvement of 6.9%, in which there was a significant effect(F(1,158)=4.576, p<0.05). Moreover, the *emotion loss* at Arousal level decreased from a mean of 0.816 to 0.754, an overall improvement of 7.5% in which there was a significant effect, F(1,158) = 6.756, p<0.05.

while the recording speech was played. In order to reduce the learning effect, all conversations in one form were presented before the next form. So the interval that participants assessed the same content was extended.

To measure the emotional information that participants perceived from a sentence, we used the two-dimensional valence (V; emotional pleasantness) and arousal (A; emotional activation/excitation) model [12]. In our study, we developed a web-based VA model whose two-dimensional space consisted of horizontal and vertical coordinate values ranging from -3 to 3.

To quantify the extent of *emotion loss*, we built the following formula:

$$\textit{emotion loss} = \text{diff}\,(E(x), E(audio))$$

E(x) indicates the participants' evaluations of the perceived emotional expression in the form "x", divided into two dimensions, valence and arousal.

We recruited 34 students (mean age=23.21, standard deviation=2.579, native Chinese speakers) for the study. 11 Mandarin conversations[6] with different topics, different emotional atmospheres, and by different speakers were prepared. Each conversation lasted approximately three minutes with an overall predefined target affective atmosphere (i.e., angry, happy, sad, neutral, surprise, and frustration).

CONCLUDING REMARKS

Different cultures and languages vary in the comprehension of both conversations and emojis [4]. Hence, the influence of cultural and linguistic background still needs to be studied. For the future work, *emojilization* tool needs further improvements to output a bigger range of emojis. Also, an interesting avenue of future work lies in the *emojilization*-centric combination of affective computing and voice user interface design.

Emotion loss has been defined to represent the loss of the emotional expression caused by Speech To Text. To reduce the *emotion loss*, we devised the *emojilization* tool that automatically attaches related emojis to the STT-generated text by fusing both textual and acoustic features in speech signals. The evaluation results demonstrate the effectiveness of our tool in *emotion loss* reduction. We believe that *emojilization* has the realistic potential to facilitate future work in the research area of emotionalized voice user interface.

ACKNOWLEDGEMENTS

This work was supported by the National Key Research and Development Plan under Grant No. 2016YFB1001402, and National Natural Science Foundation of China under Grant No. 61402250, 61232013. Besides, the authors wish to thank Tsinghua University-Alibaba Joint Research Laboratory for Natural Interaction Experience for its support. In particular, advice given by Haipeng Mi, Chen Zhao, Li Huang has been a great help.

REFERENCES

[1] Sameera A Abdul-Kader and JC Woods. 2015. Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications* 6, 7 (2015).

[2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*. 173–182.

[3] Jo-Anne Bachorowski. 1999. Vocal expression and perception of emotion. *Current directions in psychological science* 8, 2 (1999), 53–57.

[4] Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 531–535.

[5] Anders Sogaard Iyad Rahwan Sune Lehmann Bjarke Felbo, Alan Mislove. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524* (2017).

[6] Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee. 2017. NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus. In *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 292–298.

[7] Limin Paul Fu, James Landay, Michael Nebeling, Yingqing Xu, and Chen Zhao. 2018. Redefining Natural User Interface. In *Extended Abstracts of the 2018 CHI Conference*. 1–3.

[8] Wootaek Lim, Daeyoung Jang, and Taejin Lee. 2016. Speech emotion recognition using convolutional and recurrent neural networks. In *Signal and information processing association annual summit and conference (APSIPA), 2016 Asia-Pacific*. IEEE, 1–4.

[9] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2227–2231.

[10] 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), 6.

[11] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *ACM CHI Conference on Human Factors in Computing Systems*.

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

[13] James A Russell, Jo-Anne Bachorowski, and José-Miguel Fernández-Dols. 2003. Facial and vocal expressions of emotion. *Annual review of psychology* 54, 1 (2003), 329–349.

[14] Marc Schröder, Hannes Pirker, Myriam Lamolle, Felix Burkhardt, Christian Peter, and Enrico Zovato. 2011. Representing emotions and related states in technological systems. In *Emotion-Oriented Systems*. Springer, 369–387.

[15] Garreth W Tigwell and David R Flatla. 2016. Oh that's what you meant!: reducing emoji misunderstanding. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. ACM, 859–866.

[16] Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proc. INTERSPEECH 2010, Makuhari, Japan*. 2362–2365.