Shape of Sound: A method that enhances emotion in voice messages

Zhang Yufei
zhangyf11@shanghaitech.edu.cn
2021533169

Huang Guanjie
huanggj@shanghaitech.edu.cn
2021533181

Yao Zhengming
yaozhm@shanghaitech.edu.cn
2021531050

Peng Haoxiang
penghx@shanghaitech.edu.cn
2021531055

He Xinyi
hexy@shanghaitech.edu.cn
2021533168

ABSTRACT

Using the existing model, we propose a method that can reduce the emotional loss in the process of speech-to-text by encoding the voice emotion into emoji, which can meet the needs of users to a certain extent, and has acceptable text accuracy and emotion enhancement scores. We hope to eventually make it into a plugin called Shape of Sound.

KEYWORDS

Speech to text, Semantic analysis, Human-computer interaction

1 INTRODUCTION

In the rapidly evolving landscape of voice recognition and artificial intelligence, Shape of Sound emerges as a groundbreaking innovation that transcends conventional speech-to-text technologies. Shape of Sound is not just a tool for transcribing spoken words into text; it is a sophisticated system capable of capturing and conveying the emotional nuances embedded within the human voice. By integrating advanced voice recognition with emotion detection, Shape of Sound offers a unique and enriched textual representation that reflects the speaker's feelings through the use of emojis, varied text sizes, and colors.

The traditional tools for communication such as WeChat support transformation from voices to words merely. Although there is a function to add emojis in the words, it is difficult for one to trigger this function because the triggering condition is very hard. And there are just a few emoji for users. To solve these problems, this work tried to use some advanced algorithms to detect more emotions and visualize them. There are 6 emotions in voice information, after grading them, there is a 6-dimensional vector, and some emoji are used to correspond to the vector. In this way, the emotional expression can be more obvious and diversified.

The main innovation of this work is its achievement of multi-emotional visualization enhancement. Compared to traditional algorithms, this function is more popular with users.

2 RELATED WORK

Traditionally, frame-based models have been used to solve VC. Given the source and target speeches, their alignment is found by using DTW. Subsequently, the conversions between the acoustic features of the aligned frames are modeled. Recently, seq2seq models have been proposed in VC , where the model jointly learns alignment and frame conversion by attention mechanism without using the explicit temporal alignment.[1]

2.1 Speech Emotion Recognition

Emotion plays an important role in real-life applications. The emotion of a person can be identified by various sources of information like speech, transcript, facial expression, brain signal (EEG), and a combination of two or more of these (called multi-modal emotion recognition). Among these sources, speech is arguably the easiest to acquire. Whether it is the physical movement of the speakers or visual occlusions due to glass, mustache, beard, etc., the speech attributes are little or not affected by these, as compared to facial expressions, which can be significantly affected. The speech (acoustic) features for emotion recognition are almost similar in all languages and the same classification model can be used across languages. Even though SER has been carried out for most of the major languages, the same model can be used for other languages with acceptable accuracy. However, this is not true for linguistic features. Each language requires specific database. Compare this with the linguistic features where a language-specific model is required, which is often a hindrance for less popular or regional languages for whom labeled emotion data is not yet present. Even if labeled data is available, a language-specific model needs to be constructed, which is an obvious overhead.

Automatic SER is used in several applications. It enhances human-computer interaction (HCI) systems, such as interactive movies, storytelling, and E-tutoring applications, and retrieval and indexing of video/audio files. Speech-based emotion recognition system assists in improving the quality of service of the call attendants at call centers. Automatic emotion detection could be helpful in psychological treatments as used in. It can also be useful in the case of surveillance systems. Modern speech-based systems are designed largely using neutral speech. Here, the components of emotion can be used as an add-on to improve the accuracy in the practical applications. Nowadays, the voice-assisted search engines have become very popular. It would be beneficial for the on-line market to update a web page dynamically according to the user's emotion (detected through speech).[3]

2.2 Multimodal

The rise of social media and virtual communication has recently enabled new ways to encode meaning: The Unicode Standard emoji code, a set of now more than 3000 pictograms, which became available in most of our online messaging services. In their emoji Report 2015Footnote1 the company Emogi stated that about 92% of internet and instant messaging users utilizes emoji. The word emoji is derived from the Japanese "e" for “picture” and "moji" for “word”

and the resemblance to the English word "emotional" is coincidental (Taggart 2015). Nonetheless, many emoji are being used to express emotions and therefore all major social media platforms use sentiment analysis to understand the content of the users. This creates a great interest in research on semantic evaluation and analysis of emoji (Barbieri et al. 2016a, b, 2018).

Compared to words, emoji are characterized by a stronger iconic relationship with the referents that they designate. While the relationship between a word and its referent is mostly arbitrary, symbolic, and based on social conventions, the relationship between an emoji and the designated referent is less arbitrary, because it is based on iconicity. For example, the iconic nature of the relationship between an actual dog and is not preserved in the sign dog.

In this regard, the reader may think that emoji are comparable to Egyptian hieroglyphs. However, the pictograms that constitute the Egyptian ancient language, unlike emoji, are a sound-based phonetic system (Jespersen and Reintges 2008). For example, a sequence of hieroglyphs can be composed of the pictograms that provide the phonetics of the word BELIEF. The first hieroglyph would be a foot, the Egyptian pronunciation of the word foot would sound like "B", the next hieroglyph would depict a reed, which would be pronounced with an "E". This is followed by the lion hieroglyph, which in Egyptian is Leo and provides the "L". This can be called a rebus principle, which in principle can also be applied to emoji use, but we believe (and test in section "Automatic generation and analysis of averaged representations") that this might be a quite uncommon communicative strategy among emoji users.[4]

3 FORMATIVE STUDY

3.1 Demanding research

Our first questionnaire surveyed 204 people (N=204) about their habits and preferences in using the speech-to-text feature and included questions as shown in the table1.

From the first questionnaire survey, we can see that 40.2% of the participants had misunderstood the message because of speech-to-text, and the users who participated in the survey were willing to make up for the loss of emotion in speech-to-text through visual means. Based on the results of the questionnaire and the current investigation, we decided to design a tool that can enhance the emotion in speech-to-text through visual means.

In addition, we interviewed three representative users respectively and sorted out the existing three user needs **(N1-N3)** based on the previous questionnaire results:

N1. It is not convenient to listen to the voice in some occasions, but it is difficult to grasp the subtle emotions in text message without listening to the voice.

N2. Some people feel that the text converted from a long speech is not easy to read, and they need a tool to quickly summarize the mood of the departure.

N3. Some people do not like to listen to speech, feeling the accuracy and fun of speech-to-text need to be improved.

The results of the survey can be viewed at this link: **Need**

Table 1: Survey Results

Questions	Result distribution
1. When you receive a voice message, are you more likely to listen directly to voice/speech-to-text?	Listening to voice directly 53.43% (109) Speech to text 46.57% (95)
2. Have you ever misunderstood the content of someone's message because of voice to text?	Yes 40.2% (82) No 59.8% (122)
3. Have you ever triggered WeChat voice to text to automatically add emojis?	Yes 36.27% (74) No 63.73% (130)
4. Do you think the expression added by WeChat voice to text is consistent with the content?(0 10)	Average score: 6.11
5. If WeChat voice to text, text size will be proportional to the voice volume, do you think this is more interesting?	Yes 60.29% (123) No 39.71% (81)
6. To what extent do you think you can express emotion with words and emojis alone?(0 10)	Average score: 6.18
7. Do you like the full-screen action effect that pops up after QQ publishes love/specific text?	Like 63.73% (130) Dislike 36.27% (74)

3.2 Mode selection

To determine which visual methods were used to enhance the expression of emotion in speech-to-text, we issued a second questionnaire (Fig 1), which could be viewed at **mode**. In the second questionnaire, we hypothesized three different scenarios, first having participants read plain text and choose directly from two possible ideas. The three scenarios were then repeated, listing five existing ways to enhance emotion (emoji, bubble frame colors, text shapes, comic bubble frames, emotion-capturing and then generating animal figures), and having participants rate how much they felt the emotion of the message sender using a five-point Likert scale. In the end, emoji and animal characters received the highest average scores, 4.61, 4.66, 4.37 and 3.73, 3.56, 3.49 respectively in the three scenarios. This result also played a decisive role in our prototype design.

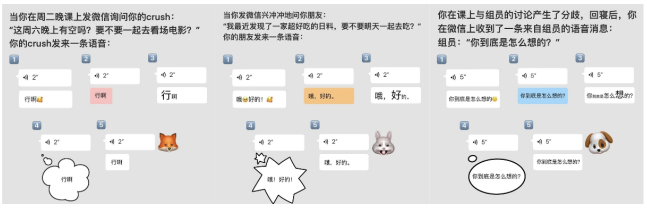


Figure 1: The second questionnaire

3.3 Prototype design

Our prototype went through two iterations, all done in Figma. The first design only included one scene. We assumed that the color and size of the text would change according to the emotion of the voice, and emojis and decorations related to the voice would be presented in the text box. However, this design was updated after the second questionnaire was released, and the updated prototype only retained the two enhanced ways of emoji and expression recognition. Users experiencing the prototype can choose either of these.

The second iteration is due to heuristic evaluation(Detail in **section 5**) during the project. We used Nielsen's heuristics to explore the problems and severity of our prototypes, and found that our design was not customizable and had problems with flexibility, and then we updated the prototypes and enhanced the flexibility. The iteration of the prototype provides guidance for our program implementation.

The final figma prototype can be view at this link: **Figma Shape of sound**.

4 SYSTEM

We hope to integrate Shape of Sound as a plug-in into WeChat, which adds more choices as transforming voice messages into texts to capture the nuance of tones and subtleties loss of emotions. However, due to capacity and time constraints, we only implemented a graphical interface that input m4a files and output txt text, which is more like a web page than an instant messaging plug-in.

Figure 2: Our system now

4.1 Interaction and Interface Design

When a user clicks into a chat interface, he can choose to enable or disable Shape of Sound after awakening Shape of Sound interface by clicking the hidden icon. There are also other buttons, which include custom settings, privacy rights and beginner's guide. Among these functions, custom settings module is used for setting customized emojis or images. In privacy rights module, users could view the required and obtained permissions for this plugin. As for beginner's guide, there are introduction and operation instructions in it.

After enable Shape of Sound, user could choose converting to text 1 or converting to text 2 to obtain different transcription emojis or expressions.

4.2 Backend Model

Based on the design requirements, we develop this tool that provides function of recognizing emotion and adding emojis, which synthesizes three main sections.

Emotion 1	Emotion 2	$\cdots$	Emotion N
n_1	n_2	$\cdots$	n_N

Table 2: Raw data

The first section is speech to text. We invoke Aliyun's API of intelligent voice interaction to convert voice message to text, using Aliyun's RAM access control and Object Storage Service as well. Aliyun's recording file identification function could also automatically divide the audio into sentences with their start time and end time within the audio, which facilitates us to split the audio by sentence.

The second section is extracting emotion from speech. We invoke emotion2vec, which is a series of fine-tuned foundational models for speech emotion recognition. We use it to recognize the split audios' emotion in six sorts — angry, fearful, happy, neutral, sad and surprised. The modal we use is "`iic/emotion2vec_base_finetuned`" of version "v2.0.4". At the same time, we fetch and output according scores of the six emotions of the split audios.

Figure 3 is the datasets employed in emotion2vec[2].

Table 1: The datasets at a glance for emotion2vec pre-training and downstream tasks.

Dataset	Pretrain	Downstream	Source	Emo	Spk	Lang	#Utts	#Hours
IEMOCAP (Busso et al., 2008)	✓	✓	Act	5	10	English	5531	7.0
MELD (Poria et al., 2019)	✓	✓	Friends TV	7	407	English	13847	12.2
CMU-MOSEI (Zadeh et al., 2018)	✓	✓	YouTube	7	1000	English	44977	91.9
MEAD (Wang et al., 2020)	✓	✗	Act	8	60	English	31792	37.3
MSP-Podcast (V1.8) (Martinez-Lucas et al., 2020)	✓	✗	Podcast	8	10000+	English	72969	113.5
Total	✓	–	–	–	–	English	169053	262.0
CMU-MOSI (Zadeh et al., 2016)	✗	✓	YouTube	7	89	English	2199	2.6
RAVDESS-Speech (Livingstone and Russo, 2018)	✗	✓	Act	8	24	English	1440	1.5
RAVDESS-Song (Livingstone and Russo, 2018)	✗	✓	Act	8	23	English	1012	1.3
SAVEE (Jackson and Haq, 2014)	✗	✓	Act	7	4	English	480	0.5
M3ED (Zhao et al., 2022)	✗	✓	TVs	7	626	Mandarin	24449	9.8
EmoDB (Burkhardt et al., 2005)	✗	✓	Act	7	10	German	535	0.4
EMOVO (Costantini et al., 2014)	✗	✓	Act	7	10	Italian	588	0.5
CaFE (Gournay et al., 2018)	✗	✓	Act	7	12	French	936	1.2
SUBESCO (Sultana et al., 2021)	✗	✓	Act	7	20	Bangla	7000	7.8
ShEMO (Mohamad Nezami et al., 2019)	✗	✓	Act	6	87	Persian	3000	3.4
URDU (Latif et al., 2018)	✗	✓	Talk shows	4	38	Urdu	400	0.3
AESDD (Vryzas et al., 2018)	✗	✓	Act	5	5	Greek	604	0.7
RESD (Lubenets et al.)	✗	✓	Act	7	200	Russian	1396	2.3

Figure 3: Datasets used by emotion2vec

The third section is our homemade emoji and emotion scores correspondence table, which bestows six scores on every emoji. Only when an emoji's all six scores accord with the detected emotion scores by emotion2vec's output score, the emoji will be printed after a sentence.

The procedure of backend modal is as Figure 4. We have implemented running the entire process with a click, only requiring to upload the audio file to the right place.

4.3 Algorithm of Emoji Quantification

Here are the steps of quantifying.

Step1 : Normalization

Processing the raw data such that the maximum value is 1.

$$\tilde{n_i} = \frac{n_i}{\sum_{i=1}^{N} n_i}$$

Step2 : Determine Thresholds

Certitude the interval of emoji as

$$(\tilde{n_i} - 0.1, \tilde{n_i} + 0.1)$$

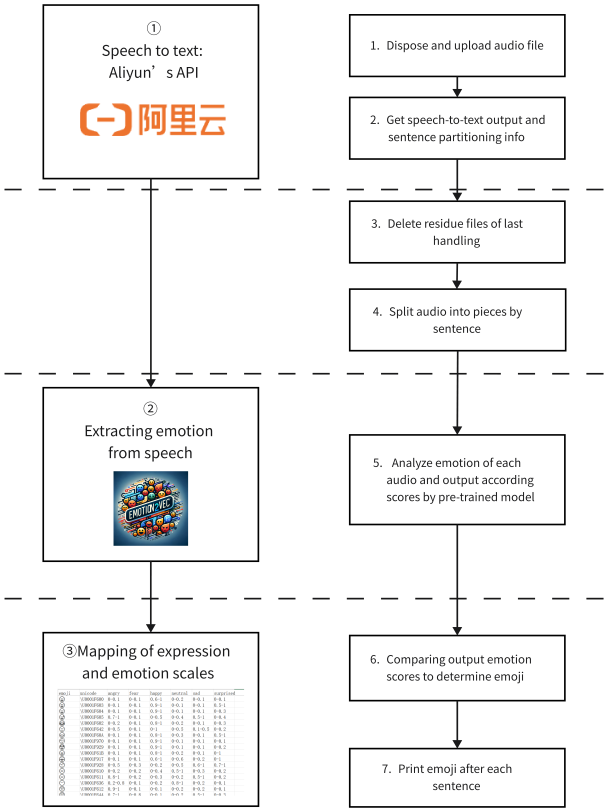


Figure 4: Pipeline of Shape of Sound system

5 EVALUATION

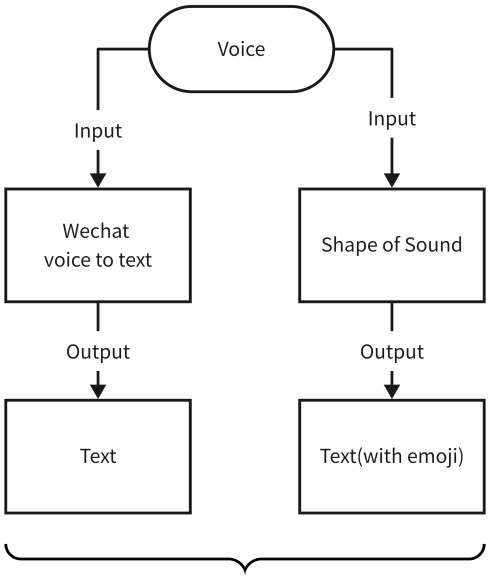
5.1 heuristic evaluation

We invited ten students to experience our prototypes separately and use Nielsen’s ten heuristics aspects to explore the problems and severity of our prototypes. The table of questions they summarized can be viewed in this document: **heuristic evaluation**

5.2 User study

We invited 15 participants (M=6, F=9) each to send three voice messages larger than 5 seconds, input the voice messages into wechat speech-to-text system and our system respectively, and then ask the participants to rate the accuracy of text translation, the difficulty of semantic understanding of sentences (which will higher if irony is in sentences) and the degree of emotional enhancement respectively. Processing is in fig 5, details can be found at this link: **user study**. Scoring was performed using a five-point Likert scale. Using wechat as a baseline, we get the following results in 3.

Analysis of the results included the following: (1) Although only three emojis (bare teeth, angry, crying) appear in wechat’s speech-to-text, the accuracy rate of the result of text conversion including emojis is almost 100%, and the average score of 1.72 is because the probability of emojis appearing in wechat text conversion is about 20%. The emotion enhancement score of the voice without emoji will be 1. (2) In the course of our research, we found that not all voice messages require emotional enhancement, so it is important to give users the "right to opt out" when designing plug-ins. Besides, different users, the expression of the meaning may be different, perhaps support for custom emoji when to appear is also a way to improve.(3) Participants did not have a uniform scoring standard for the difficulty of semantic understanding, nor did they send sentences with difficult semantic understanding, which led to the failure of the scoring.



Evaluation: (Likert scale)
1.Accuracy of speech to text
2.Semantic comprehension difficulty of the text itself (irony and similar dimensions affect this)
3.Affective enhancement

Figure 5: Method of evaluation

	SoS	WeChat
Speech-to-Text Accuracy Average Score	4.61	4.53
Emotion Enhancement Average Score	3.76	1.72
the difficulty of semantic understanding	2.3	2.3

Table 3: Results of average Scores

6 DISCUSSION, LIMITATION AND FUTURE WORK

In this section, we will discuss the practicability and effectiveness of shape of sound based on the above. And in subsection 6.2 and 6.3, we extract future design considerations from our analysis results and questionnaire feedback and explore reflect on the limitations of our work.

6.1 Discussion

The development and application of Shape of Sound have provided fascinating insights into the integration of emotional intelligence with voice recognition technology. The following is an analysis of the practicability and effectiveness of Shape of Sound.

6.1.1 The practicability of Shape of Sound. In our preliminary research, we found that in daily life, some people will directly convert speech into text because of habits, and some people can not directly listen to speech because of the workplace. In these cases, it is particularly important to enhance the emotional expression of speech-to-text. Our design is based on this kind of people. In the questionnaire survey, we finally selected two most popular methods to enhance emotional expression. Therefore, Shape of Sound can selectively generate some emojis according to the voice or generate emojis combined with facial recognition to enhance the expression of emotions. Depending on people's actual needs and preferences, the practicality of Shape of Sound has also been greatly enhanced.

6.1.2 The effectiveness of Shape of Sound. On the basis of scoring emojis on the EMOJIALL official website, we further carried out research to clarify the possible emotions contained in each emoji and the proportion of various emotions, which greatly improved the accuracy of generating emojis. In terms of page design, we adopted the suggestions of other groups, and added the button to cancel emoji or emojis and the undo button. Such a design can not only make the page more concise and controllable, but also prevent the wrong recognition of emotions after the conversion of text. Therefore, Shape of sound is effective, and it enhances the expression of emotions by improving the accuracy of emotion recognition and expression as accurately as possible.

6.2 Limitation

However, our design still has some limitations:

6.2.1 Cultural Variations. In our plug-in, the emotion recognition accuracy of Chinese-English mixed sentences is lower than that of Chinese-only sentences. In addition to this, even within the same language, different cultures express and perceive emotions in different ways. Current systems may not be able to account for these nuances, which can lead to systems not accurately conveying the speaker's emotions. Future versions of the system will therefore need to incorporate more diverse expressions of emotion and culture-specific communication styles to avoid such problems.

6.2.2 User Adaptability. In emoji interpretation and written communication, users may have different levels of familiarity with different emojis. At present, there are differences between the emojis provided by the system and those that come with wechat, and this difference may cause users to make mistakes in recognizing emojis. Therefore, in the subsequent design, we also need to conduct a questionnaire survey to identify the emojis of wechat itself.

6.2.3 Privacy Protection. In the survey results, most people show a strong interest in the function of facial recognition to generate emojis, but the discussion of its privacy is missing. Whether people will want to turn on the camera to use this feature, or how the camera will capture the user's facial features, is still up for further discussion.

6.2.4 Method Bias. In the process of emotion quantification, when we recognize an emotion, we will give it a very high score, so as to ignore other emotions. Therefore, this method will lead to the deviation of emotion recognition in some cases. In addition, in the process of collecting questionnaires in the preliminary formative study, we did not collect basic data, which may also cause some bias.

6.3 Future Work

6.3.1 Implement. We failed to achieve a complete plug-in product of integrated front-end and back-end, and hope to improve it in future research.The multi-modality based emotion enhancement model that we designed in the formative study also failed to be realized, and we hope to implement it later and use it as part of the plug-in.

6.3.2 Emotional coding. We hope to continue to investigate the emotional connotations represented by different emojis in different groups, so as to better encode the emojis after translation into text. We also hope that the Shape of Sound in the future version can extract personalized emotion recognition according to each person's past communication style, including the language expression habits of users and the version used by emojis. Provide users with the option to choose a mode of emotional expression that fits their culture and communication style.

6.3.3 Sender enhancement. In the process of research, users have reflected that the sender should know what kind of emotions he has sent and can change them. Although our current design focuses on accurately transmitting the emotions sent by the sender to reduce losses, we can also think and explore the aspects of the sender in the future.

7 CONCLUSION

In conclusion, while the Shape of Sound introduces a novel approach to enriching digital communication with emotional context, there is still room for growth and refinement. By considering these discussions and addressing these limitations, there is a clear path forward for this technology to better serve its users and finally be more widely used in communication, customer service, accessibilty, and more.

REFERENCES

[1] Tae-Ho Kim, Sungjae Cho, Shinkook Choi, Sejik Park, and Soo-Young Lee. 2020. Emotional Voice Conversion Using Multitask Learning with Text-To-Speech. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7774–7778. https://doi.org/10.1109/ICASSP40776.2020.9053255

[2] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. arXiv:2312.15185 [cs.CL]

[3] Md. Shah Fahad, Ashish Ranjan, Jainath Yadav, and Akshay Deepak. 2021. A survey of speech emotion recognition in natural environment. *Digital Signal Processing* 110 (2021), 102951. https://doi.org/10.1016/j.dsp.2020.102951

[4] Philipp Wicke and Marianna Bolognesi. 2020. Emoji-based semantic representations for abstract and concrete concepts. *Cognitive processing* 21, 4 (2020), 615–635.

Project ppt link:(including concept video of Shape of Sound):
HCI proj PPT