A VISUALIZATION METHOD OF HUMAN SPEECH IMPRESSIONS BY USING SPEECH BALLOONS

Tomohiro Yamada and Kazunori Mano
Graduate School of Engineering and Science, Shibaura Institute of Technology
Email: mf14055@shibaura-it.ac.jp, mano@sic.shibaura-it.ac.jp

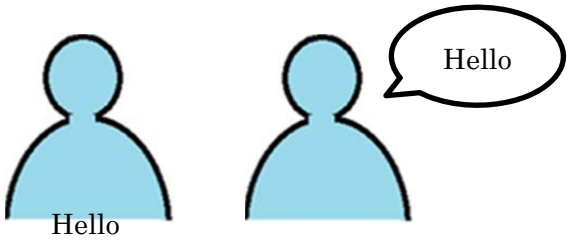
ABSTRACT

Speech balloons are very unique and useful representations in comics. They convey not only the text contents but also the talkers' emotional impressions very well. They may be used in an alternative subtitle system in the future. This paper proposes a new subtitle system that can express para-/non-linguistic information of speech through speech balloon representations with various shapes. It was experimented that how the number of edges, the angles, and the shapes are related to acoustical features of speech by subjective quality point of views. The proposed system is organized by a neural network which estimates adequate speech balloons from input speech features with various emotional impressions. In the preliminary experiment, it was confirmed that four emotion patterns can be appropriately transformed into speech balloons.

1. INTRODUCTION

Recently, welfare applications using multimedia information technology have been developed widely. For example, real time subtitles are frequently used in television programs and other video services for hearing-impaired persons. But, conventional subtitles cannot visualize para-linguistic or non-linguistic information which show intentions, emotions and individualities of talkers. Several researches have been conducted for TV viewers to recognize speech impressions without hearing actual voices. One approach is to transform para-linguistic or non-linguistic information of speech into any visual information such as colors, fonts, and so on. Animated-text representation of speech has been proposed [1, 2]. It has been developed to apply foreign language learning for visualization of pronunciation of speech. Japanese comic systems utilize another technique which uses speech or text balloons with various shapes to express para-/non-linguistic information. As well as shapes, other attributes such as colors, line widths, and line types can be utilized to transfer emotional speech impressions.

Speech balloon representations can express much emotional patterns by changing the various forms to show the feelings of dialogues in comics. A speech balloon captioning system for information support on meetings has been studied [3]. Comparing the system with the conventional TV-type subtitle system (see Fig.1(a),(b)), the latter can convey more emotional information than the former can. However, in the research, the system incorporated just a fixed shape of speech balloons.



(a) TV-type subtitle (b)Text-balloon-type subtitle

Fig. 1. Two types of captions

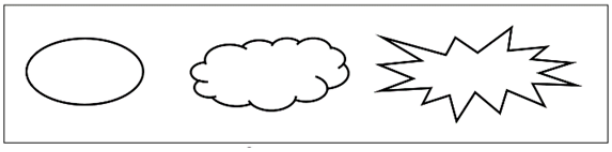
Another study of choosing a preferable balloon shape which matches the speech utterances between two types of shapes, "round" or "jagged", was reported [4]. The system was trained based on machine learning techniques to map speech sound into two types of text balloons. The result shows that 85% of the mapping trials were correctly chosen. Since the result of the same mapping experiment by human was 88%, it was revealed that the impression of speech can be extracted and transformed into speech balloon images by machine. The research only estimates one of two types. In actual human conversations much more different types of emotional, or para-/non-linguistic information are communicated.

This paper proposes a new system of speech visualization by using speech balloons. The speech balloons can have various kinds of shapes, colors, line-types, in principle. It is expected that such flexible speech balloons are useful to represent invisible speech information with emotional impression adequately.

2. OUTLINE OF SPEECH BALLOON GENERATION SYSTEM

2.1 Variations of speech balloons

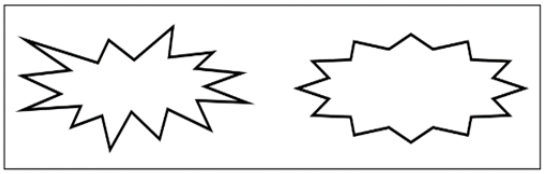
In comics, balloon-type subtitle expressions show contents and impressions of human conversations. Many Japanese comics, that is called "manga", generally use only black and white colors, but, they can represent impressions of speech sound as texts in balloons. Advantages of utilizing balloons are that (1) it can show who talks, and (2) it can show what are the emotions of talkers. There are many balloon shapes. Commonly-used 12 shapes in Japanese comics are illustrated in the Appendix. The text balloons give the different impressions of talkers. The normal balloon of an elliptic shape is always used in plain, calm or peaceful cases without any strong or weak utterances. The jagged balloon is chosen for strongly uttered, screamed, or exclaimed sound of speech. Although detailed impressions for each speech balloon are different for each person, as a beginning of this study, typical three shapes were examined, which are shown in Fig. 2.



(a) plain/elliptic (b) cloud (c) jagged

Fig. 2. Three typical speech balloons

As shown in Fig. 3, different impressions can be expressed by slightly changing the degree of jagged or bulged patterns.



(a) Strong utterance (b) Slightly strong utterance

Fig. 3. Two jagged balloons with different strongness.

2.2 Synthesis structures of speech balloons

In this research, a feature mapping between acoustic features of speech and visual features of balloons are considered. Synthesis structures of speech balloon generation are designed. To express the various shapes, based on a plain elliptic shape, jagged or cloud type balloons are generated. An elliptic balloon shapes are defined by the following three parameters of (F1) to (F3) in Fig. 4.

(F1) Horizontal length of the major axis.
(F2) Vartical length of the minor axis.
(F3) Line width.

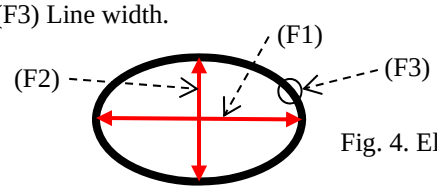


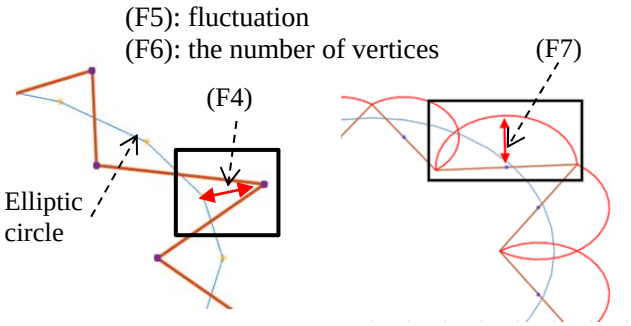
Fig. 4. Elliptic balloon.

(F4) to (F6) are the parameters to control jagged balloons. (See Fig.5 (a))

(F4) Distance from baseline elliptic shape to the vertex of a jag.
(F5) Random fluctuation of adjacent (F4) values. According to (F5) values, the variations in Fig. 3 are produced.
(F6) The number of (F4) vertices in a balloon.

The next (F7) parameter is used to make cloud shapes.

(F7) Distance from the center of each jagged edge to the top of the small circle. (See Fig.5(b)). The cloud shape is represented by bulging edges from jagged shapes.



(a) (b)

Fig. 5. Parameters of (a) a jagged balloon and (b) a cloud balloon.

These seven features of (F1) to (F7) are estimated in the proposed method. Firstly, a basic shape is determined based on (F1), (F2) and (F3) parameters. Secondly, parameters (F4), (F5), and (F6) define a jagged balloon from a basic elliptic shape, and finally (F7) generates a cloud shape balloon.

For example, Table 1 is a basic parameter set of an elliptic shape balloon and Fig. 6 is its balloon shape.

Table 1 Parameters of an elliptic balloon.

Balloon Features	F1	F2	F3	F4	F5	F6	F7
Values	1.5	1	2	0	0	0	0

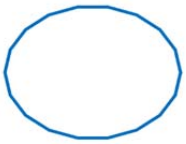


Fig. 6. Example of a generated elliptic balloon according to the parameters in Table 1.

2.3 Speech features

Speech features are defined by the following four parameters, (S1) to (S4).

(S1) Fundamental frequency (F0).
(S2)Zero crossing rate (ZCR).
(S3) Power (P).
(S4) Mel-frequency cepstral coefficients (MFCC).

For each feature, four values of maximum, average, minimum, and variance are introduced. Then, totally 16 values, that is 4 values for each of 4 features, are used in machine learning.

2.4 Generation of balloon features by using neural network mapping system

The mapping system is organized based on a machine learning method with a neural net. In the experiment, the learning data consist of 80 sentences of various emotional expressions.

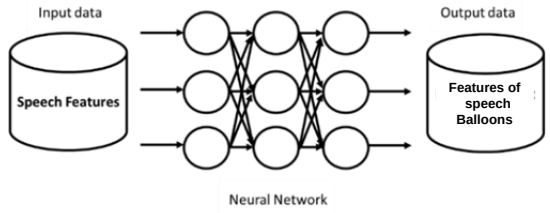


Fig. 7. A diagram of speech to balloon feature mapping.

3. EXPERIMENT

An experiment to generate various speech balloons from speech features was performed by using the neural net trained in 2.4.

3.1 Speech data and features

Speech data of four emotions, "calm", "anger", "pleasure", and "sorrow", were prepared. Fig. 8(a), (b), (c), and (d) show the average feature values obtained in the speech data.

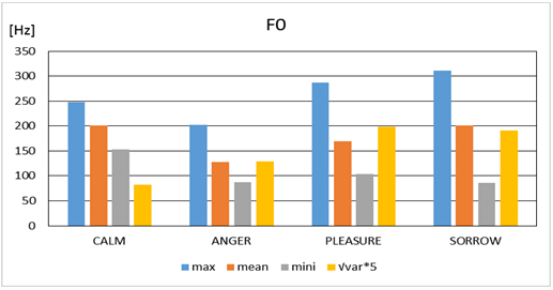


Fig.8(a). Speech fatures (S1) of fndamental frequencies (F0) of each emotion.

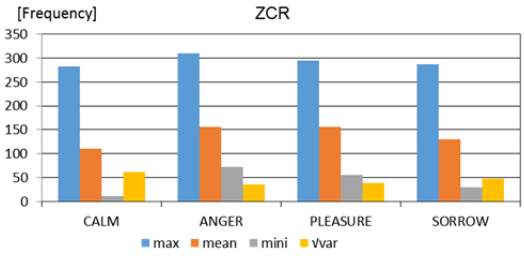


Fig.8(b). Speech fatures (S2) of zero crossing rates (ZCR) of each emotion.

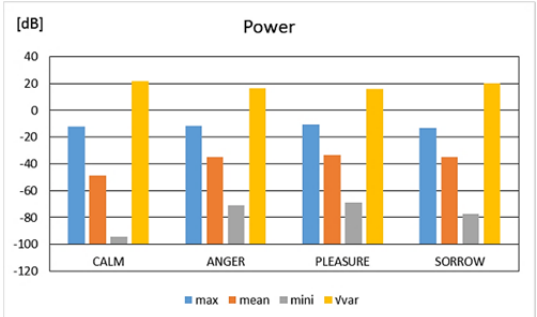


Fig.8(c). Speech fatures (S3) of power (P) of each emotion.

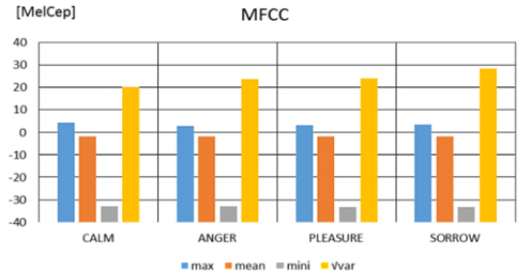


Fig.8(d). Speech features (S4) of mel-frequency cepstral coefficients (MFCC) of each emotion.

3.2 Results of generated balloons

Tables 2(a)-(d) and Fig.9(a)-(d) are the speech features and the estimated speech balloons.

(1) Generation of a speech balloon for a "calm" speech.

Table 2(a). Balloon values for "calm" speech.

Balloon Features	F1	F2	F3	F4	F5	F6	F7
Values	1.5	1	1.9	0.02	0.01	12	-0.36

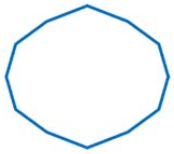


Fig. 9(a) Generated balloon for a "calm" speech.
The "calm" speech is a plain, emotionless speech.

(2) Generation of a speech balloon for an "anger" speech.

Table 2(b). Balloon values for an anger" speech

Balloon Features	F1	F2	F3	F4	F5	F6	F7
Values	1.8	1	3	0.3	0.24	18	-0.43

Fig. 9(b) Generated balloon for "anger" speech.
The "anger" speech has larger power and angry voice.

(3) Generation of a speech balloon for "pleasure" speech.

Table 2(c). Balloon values for "pleasure" speech.

Balloon Features	F1	F2	F3	F4	F5	F6	F7
Values	1.62	1	2.11	0.04	0.08	18	0.01

Fig. 9(c) Generated balloon for "pleasure" speech.
The "pleasure" speech has merrily and forceful voice.

(4) Generation of a speech balloon for "sorrow" speech.

Table 2(d). Balloon values for "sorrow" speech.

Balloon Features	F1	F2	F3	F4	F5	F6	F7
Values	1.81	1	1.22	0.22	0.25	12	0.01

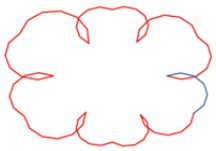


Fig. 9(d) Generated balloon for "sorrow" speech.
The "sorrow" speech has quiet and sad voice.

3.3 Discussions

For the "anger" and "pleasure" speech, the line widths are thicker than other speech sounds. The line width of the balloons are thicker in accordance with the increase of ZCR(S2) and P(S3) values. Since the balloons with cloud type shapes has mainly larger variances of the MFCC, the output shapes of "pleasure" and "sorrow" emotions are represented by cloud-type shapes. According to these results of the experiment, it is suggested that the proposed balloon generation system can convey emotional impressions from speech features to balloon features.

4. CONCLUSION

This paper proposed a new subtitle system that can express para-/non-linguistic information of speech through speech balloon representations with various shapes. The system is based on a speech balloon generation neural network which conveys emotional impressions from speech to balloons. In the preliminary experiment, it was confirmed that four emotion patterns can be transformed into corresponding balloons.

Since there are further more types of text balloons in actual Japanese comics, more detailed speech-to-balloon mapping system should be studied. In the future, not only speech impression but also linguistic information in conversations and their contexts should be considered to represent more impressive speech balloons.

REFERENCES

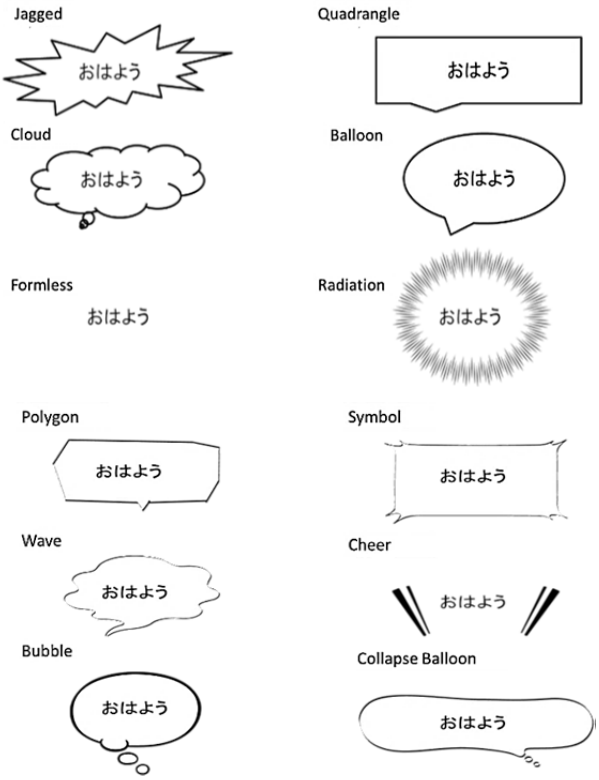
[1] S. Kimoto, K. Taira, Y. Horiuchi, S. Kuroiwa, "Transcription method to express utterance impressions, ", The Acoustical Society of Japan, Acoustical Science and Technology, pp.439-442, 2011-03.

[2]Nur Syafikah Binti Samsudin, Kazunori Mano, “Media Conversion of Paralinguistic and Nonlinguistic Speech Information into Animated Texts for Foreign Language Learning,” Proc. MJASC 2013, part3-5, 2013.

[3] A. Fujii, H. Nanjo, T. Yoshimi, "Speech balloon captioning system for information support on meetings," Information Processing Society of Japan, SIG-SLP, 2009(10), pp.75-82, 2009-01.

[4] S. Matsumiya, S. Sakti, G. Neubig, T. Toda, S. Nakamura, "Evaluation of the generation of text balloon using the acoustic features," The Acoustical Society of Japan, Acoustical Science and Technology, pp.11-12, 2014-03.

Appendix: Various text balloons in Japanese comics.



Tomohiro Yamada received his B.E. from Shibaura Institute of Technology, Japan. She is currently a Master student in Graduated School of Systems Engineering and Science at Shibaura Institute of Technology.

Kazunori Mano received his B.E., M.E. and Dr.Eng. degrees in electrical engineering from Waseda University in 1982, 1984 and 1987, respectively. From 1987 to 2008, he engaged in research on speech coding and synthesis at NTT labs. Since 2008, he has been a Professor, Department of Electronic Information Systems, Shibaura Institute of Technology. His current interests include speech processing, media coding, and communication systems.

****