Data-Driven Generation of Text Balloons based on Linguistic and Acoustic Features of a Comics-Anime Corpus

Sho Matsumiya, Sakriani Sakti, Graham Neubig, Tomoki Toda, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology, Japan
{sho-m,ssakti,neubig,tomoki, s-nakamura}@is.naist.jp

Abstract

Most automatic speech recognition systems existing today are still limited to recognizing what is being said, without being concerned with how it is being said. On the other hand, research on emotion recognition from speech has recently gained considerable interest, but how those emotions could be expressed in text-based communication has not been widely investigated. Our long-term goal is to construct expressive speech-to-text systems that conveys all information from acoustic speech, including verbal message, emotional state, speaker condition, and background noise, into unified text-based communication. In this preliminary study, we start with developing a system that can convey emotional speech into text-based communication by way of text balloons. As there exist many possible ways to generate the text balloons, we propose to utilize linguistic and acoustic features based on comic books and anime films. Experimental results reveal that expressive text is more preferable than static text, and the system is able to estimate the shape of text balloons with 87.01% accuracy.

Index Terms: data-driven approaches, expressive text generation, linguistic and acoustic features

1. Introduction

Closed captioning is the process of conveying spoken utterances by text on a television, video screen, or other visual display to provide additional or interpretive information. It is mostly created for hard of hearing individuals to assist in comprehension [1]. Recently, the demand for broadcast closed captioning services has greatly increased. However, due to the quantity of the demand and the cost of the process, manual closed captioning is no longer feasible. Broadcast companies are seeking for more efficient closed captioning alternatives.

As technology of automatic speech recognition (ASR) has progressed from simple machines that respond to a small set of sounds to more sophisticated systems that respond to real spoken language, the ability of speech recognition to assist closed captioning for TV programs has been increasing. To date, for several years, the British Broadcasting Corporation (BBC) and Japan Broadcasting Corporation (NHK) have been using ASR technology for closed captioning, which significantly reduces the time taken for a manual close captioning to complete a program [2, 3, 4]. Unfortunately, most ASR systems existing today are still limited to recognizing what is being said without being concerned with how it is being said. Thus, expression of emotions, which play an important role during communication, is not generally achieved by these systems.

On the other hand, research on emotion recognition from speech has recently gained considerable interest in the fields of human-machine communication and multimedia retrieval [5]. Numerous official challenges on emotion recognition [5, 6, 7, 8] have been conducted in the last decade trying to improve features and classifiers. Furthermore, approaches of emotional speech classification in anime films that involve characteristic sounds and prosody based on speaking styles and emotional expressions have also been proposed [9]. However, these studies only aim to recognize/classify the type of emotions that lies within the speech. How those emotions could be expressed in text-based communication has not been widely investigated.

Our long-term goal is to construct expressive speech-to-text systems as illustrated in Fig.1 which convey all information from acoustic speech, including verbal message, emotional state, speaker condition, and background noise, into unified text-based communication. In this preliminary study, we start with developing a system that can convey emotional speech into text-based communication by way of text balloons. As there exist many possible ways to generate the text balloons, we propose to utilize linguistic and acoustic features based on comic books and anime films. The main reason we take this approach is because: (1) These genres are widely known (2) It is possible to construct data-driven generation systems trained from speech-text data of comic books and anime films.

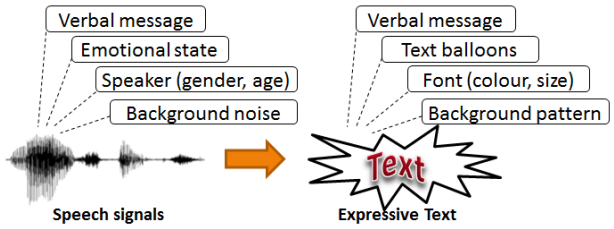

Figure 1: Overview of an expressive speech-to-text system.

2. Related Work

The inability to convey emotional speech into text-based communication greatly limits the effectiveness of communication from a social perspective [10]. Numerous attempts have been made to address the issue. Emoticons are one way to express one's emotional state in text-based messaging systems. Several studies [11, 12] have investigated the effects of using emoticons, and showed that emoticons are effective for remote emotional communication. However, emoticons only express emotion in a static manner, while various tones of voice and intensities of emotion cannot be depicted.

Kinetic typography is an alternative solution that expresses emotion in text-based communication with variations in color, size, or position over time [13, 14, 15]. However, this approach has met with limited success in practice, largely due to com-

plexity and difficulties in interaction. A study by Kalra et al. attempted to address these issues by developing "TextTone," which can semantically indicate tones within a simple intermediate plain-text format [16].

These methods of expressing emotion in text are mainly developed for text-based messaging systems with a limited number of pre-defined symbols. Comics, on the other hand, use a diverse variety of expressions created with their own visual language or iconography to convey character's emotion within the stories. Thus, the expressive symbols have more variety and are direct related to speech-based communication. The most basic element of expression in a comic is text balloons (also known as speech bubbles, dialogue balloons, or word balloons), which express various types of speeches and thoughts. Many comic stories have been adopted into anime films, and many comics have came popular worldwide.

There has been much research on the visual aspect of comics. However, most of these works focus on either balloon detection [17], balloon extraction [18, 19], and component retrieval [20]. None of them have been on text balloon generation. In this preliminary study, we attempt to utilize both comics and anime films for text balloon generation in order to improve expressiveness speech-to-text systems.

3. Overview of Text Balloons Generation

Figure 2 shows an overview of our system architecture. The system includes two components:

- automatic speech recognition that recognizes the verbal message given a speech utterance, and
- automatic generation of text balloons that deliver non-verbal communication based on linguistic and acoustic features.

The system then combines the output from both components resulting in the verbal message with text balloons. Note that, as the current focus here is to automatically generate text balloons, the speech recognition component will not be discussed further in this paper.

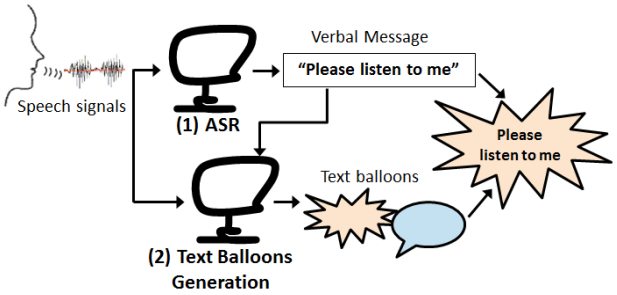

Figure 2: Overall system architecture with two components: (1) automatic speech recognition, and (2) automatic generation of text balloons

4. Comics-Anime Corpus Construction

To enable the system to generate text balloons, we compiled a corpus based on acoustic speech data from anime films and the corresponding transcription including text balloons from comic books. First, we investigate several of the world's most well-known Japanese comics, such as "Dragon ball," "Doraemon," and "One Piece," chosen considering the following factors:

- the amount of comics-anime data that can be utilized,
- the variants of text balloons within the comic books, and
- the acoustic conditions within the anime films.

Based on our analysis, we found most utterances in "Doraemon" were covered only by round text balloons, and the acoustics of speech in "Dragon ball" was too noisy. Thus, in this work, we utilize anime films and comic books of "One Piece." Furthermore, we propose to use two types of text balloon, circular and jagged, as the system's target classes, because 98% of the "One Piece" comic-anime scenes are occupied by either circular (72.98%) and jagged (27.02%) balloons.

The overall collection is done with following procedure:

- We collected anime films and the corresponding comic books of "One Piece."
- We segmented the acoustic speech of the anime and extracted only the acoustic speech that contains conversation and the target text balloons.
- For all anime speech segments, we manually annotated these speech with corresponding transcriptions from comic books.

In total, there are 2,025 utterances, where 1,478 utterances were annotated with circular balloons, and 547 utterances were annotated with jagged balloons. An example of a jagged balloon is illustrated in Figure 3.

Figure 3: Example of a jagged balloon.

5. Features and Classification of Expressive Texts

Now that we have training data consisting of input speech waveforms, and correct labels corresponding to the actual shape of the text balloon, we can proceed to build a classifier to automatically estimate the text balloons from speech. We use a feature set consisting of linguistic and acoustic features. As the linguistic features, we use unigrams, bigrams, or trigrams. Various acoustic feature sets are also investigated, including:

- an acoustic feature configuration based on the INTERSPEECH 2009 Emotion Challenge [21] as shown in Table 1 (denoted as "IS09"). It consists of 16 low-level descriptors with their first order regression coefficients and 12 functionals, resulting in a total of 384 features;
- an acoustic feature configuration based on the INTERSPEECH 2010 Paralinguistic Challenge [6] as shown in Table 2 (denoted as "IS10"). It consists of 38 low-level

descriptors with their first order regression coefficients and 21 functionals, resulting in a total of 1582 features; and

- an additional larger feature set provided by openSMILE in 2010 [22] as shown in Table 3 (denoted as "IS10+"). It consists of 56 low-level descriptors with their first order regression coefficients and 39 functionals, resulting in a total of 6552 features.

More details of description of these features can be found in [5, 6, 22].

Table 1: Feature set of Emotion Challenge 2009 (denoted as "IS09").

LLD ($16 \cdot 2$)	Functionals (12)
ZCR	Mean
RMS energy	Standard deviation
F0	Kurtosis, Skewness
HNR	Extremes: Value, Rel. position, Range
MFCC 1-12	Linear regression: Offset, Slope, MSE

Table 2: Feature set of Paralinguistic Challenge 2010 (denoted as "IS10").

LLD ($38 \cdot 2$)	Functionals (20)
PCM loudness	Position max/min
MFCC 0-14	Arith. mean, Std. deviation
Log Mel freq band 0-7	Skewness, Kurtosis
LSP Frequency	Lin. regression coeff. 1/2
F0	Lin. regression error Q/A
F0 envelope	Quartile 1/2/3
Voicing prob.	Quartile range 2-1/3-2/3-1
Jitter local	Percentile 1/99
Jitter consec. frame pairs	Percentile range 99-1
Shimmer local	Up-level time 75/90

Table 3: Large feature set provided in OpenSMILE (denoted as "IS10+").

LLD ($56 \cdot 2$)	Functionals (39)
Log energy	Var, Std. dev., Kurtosis, Skewness
MFCC 1-12	Extremes: Value, Range, Pos max/min
MELSPEC 0-25	Extremes: Mean dist. max/min
ZCR	Lin. regression centroid, coeff. 1/2
F0	Lin. regression error Q/A
F0 envelope	Lin. quad. regression coeff. 1/2/3
Voicing prob.	Lin. quad. regression error Q/A
Spectral bands 0-4	Crossing (ZCR), Quartile, Iqr
Spectral roll-off 0-3	Percentile 0.95/0.98, interp
Spectral flux	peaks: Num, Overlap, Mean dist
Spectral centroid	Peak mean, Peak mean dist
Spectral max	Mean arith/abs/quad
Spectral min	Non-zero: Num, Mean abs/quad/geo

6. Experimental Set-Up

Experiments on automatic generation of text balloons were done based on the constructed "One Piece" comic-anime corpus described in Section 4. Here, from the total of 2,025 utterances, we perform 4-fold cross validation with 1,519 utterances as a training set, using the remainder of the corpus as a test set.

Linguistic features were extracted using morphological analysis conducted by MeCab [23, 24], while acoustic features were extracted using the *openSMILE* toolkit[1] [22]. Here, we used Support Vector Machines (SVM) based on LIBSVM [25] as the machine learning technique for classifying two types of text balloons (circular or jagged) given a speech utterance.

The experiments were conducted in order to (1) evaluate the performance of automatic generation of text balloons, (2) analyze the usefulness of text balloons for expressive speech-to-text generation, and (3) analyze the relationship with emotion classification. Detailed results and discussions are described in next section.

7. Results and Discussion

7.1. Evaluation of Expressive Text Generation

First, we performed an evaluation of SVM-based automatic generation of text balloons using linguistic and acoustic features. As described in Section 4, the chance rate is 72.98%. Table 4 shows the classification results. Using only linguistic features, the best performance was achieved by the bigram model with 82.34% accuracy. Using only acoustic features, the best performance was achieved by IS10+ features with 86.35% accuracy. By the use of both feature sets, we were able to classify circular and jagged balloons with 87.01% accuracy. The results reveal that the classifier with combination features performs better than the classifier with only linguistic or acoustic features. Among these acoustic features, MFCC, energy and loudness are the top three significant features that distinguish utterances into circular and jagged text balloons.

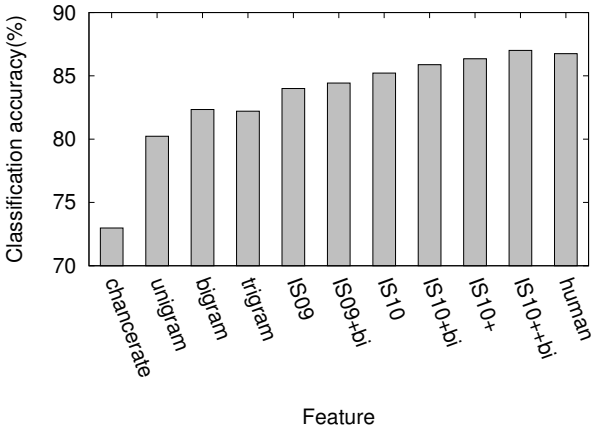

Figure 4: Accuracy of classification of text balloons

As a comparison, we also performed classification based on subjective judgement. Here, five subjects who know comics and anime well participated in the experiments. Given a speech utterance, they were requested to select the appropriate shape of the text balloons. The results of those five people are shown in Table 4. The best one achieved 88.26% accuracy, and the average score was 86.75% accuracy. The average was also included in the Fig. 4. As can be seen, our system performs better than the average of human annotators ($p < 0.05$).

[1] http://opensmile.sourceforge.net/

Table 4: Result of human annotators.

subject	rate of classification (%)
No.1	88.26
No.2	88.22
No.3	86.13
No.4	86.01
No.5	85.13
average	86.75

7.2. Usefulness of Text Balloons

Next, we perform a subject evaluation to investigate the usefulness of text balloons. Seven subjects participated in the experiments. For every video segments of anime, we subtitled with: (1) circular balloons, (2) jagged balloons, and (3) static text. Then, subjects were requested to answer ("yes" or "no") whether or not they feel an emotion appropriate for the scene while watching those film segments of anime. The experiments were done both when the annotator could hear the actual speech, and when the sound was muted.

Table 5: Comparison of static text and text balloons with speech.

		Correct Label	
		circular	jagged
Representation in the video segments	circular	83%	20%
	jagged	14%	84%
	text	35%	42%

Table 6: Comparison of static text and text balloons without speech.

		Correct Label	
		circular	jagged
Representation in the video segments	circular	85%	24%
	jagged	21%	78%
	text	37%	40%

Experimental results are shown in Table 5 (with sound) and Table 6 (without sound). "Correct Label" represents the actual shape of the text balloon that appeared in the comic, while "Representation in the video segments" represents the variety used in the video segments of anime. The percentage expresses the rate with which subjects replied "yes." The results indicate that when subtitling using the correct label, the subjects could feel emotion at a higher rate. It also shows that subtitling with text balloons are better than static text. Moreover, investigating the condition with and without sounds, the results with appropriate text balloons are still high. This reveals that text balloons are useful to express the emotion regardless of whether sound can be heard.

7.3. Relationship with Emotion Classification

Finally, we investigated the relationship between text balloon classification and emotion classification. Here, we used 5 types of emotions, including "Anger," "Sadness," "Happiness," "Fear" and no emotions. We then selected utterances that contain these emotions resulting 514 utterances in total. The distribution of emotions within those utterances is shown in Table 7.

Table 7: Distribution of emotions in the data.

Emotion	# Utterances
No emotion	351
Angry	85
Sad	44
Happy	12
Fear	24

Table 8: Distribution of text balloons within each emotion.

Emotion	Circular	Jaggy
No emotion	323	28
Angry	28	57
Sad	29	15
Happy	10	2
Fear	9	13

For all utterances within each emotion, we determined the correct label of the text balloons. The results of the text balloon distribution within each emotion is shown in Table 8. It reveals that one type of text balloon does not always represent one type of emotion. The reason for this is that in each emotion, there is always a possibility to express with either circular balloons or jagged balloons. For example, in the case of "Anger," the distribution rate of balloons is 35:65 for circular:jagged, respectively. Looking in more detail, we found that circular balloons express cold anger, while jagged balloons express hot anger. A related study on hot-anger versus cold-anger can also be found in [26].

Adopting the definition of emotion proposed in [27], there are four main dimensions that construct an emotion, including *valence, power, arousal*, and *expectancy*. Our results indicate that text balloon is a way to express the degree of *arousal* within each emotion. Thus, when the degree of arousal is low, the speech is express with circular balloons, while when the degree of arousal is high, the speech is express with jagged balloons.

8. Conclusion

In this study, we investigated the use of a comic-anime corpus to automatically generate text balloons based on linguistic and acoustic features. In addition, we also performed an analysis of the proposed system on its usefulness as well as its relation with emotion classification. The results reveal that an SVM classifier with combination features performs better than the classifier with linguistic or acoustic features only, and can automatically generate text balloons with 87.01% accuracy, slightly better than human annotators. It also shows that subtitling with text balloons is better than that with static text. Moreover, investigating conditions with and without sound, the results with appropriate text balloons is still high. This reveals that text balloons are useful even with or without sounds. Overall, text balloons are a way to express the degree of *arousal* in emotion. In the future, we will investigate other factors in order to improve expressive modality within a speech-to-text system.

9. References

[1] NHK STRL, "Speech recognition for real-time closed captioning," Broadcast Technology, 2012.

[2] M. Evans, "Speech recognition in assisted and live subtitling for television," in *Proc. of TAO Workshop on Universal Design for Information, Communication and Broadcasting Technologies*, Tokyo, Japan, 2003.

[3] A. Ando, T. Imai, A. Kobayashi, H. Isono, and K. Nakabayashi, "Real-time transcription system for simultaneous subtitling of Japanese broadcast news programs," *IEEE Transactions on Broadcasting*, vol. 46, no. 3, pp. 189–196, 2000.

[4] T. Imai, A. Matsui, S. Homma, T. Kobayakawa, K. Onoe, S. Sato, and A. Ando, "Speech recognition with a re-speak method for subtitling live broadcasts," in *Proc. INTERSPEECH*, Denver, Colorado, USA, 2002, pp. 1757–1760.

[5] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. INTERSPEECH*, Brighton, United Kingdom, 2009, pp. 312–315.

[6] B. Schuller, S. Steidl, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 2794–2797.

[7] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011 - the first international audio/visual emotion challenge," in *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*, Memphis, Tennessee, 2011, pp. 415–424.

[8] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012 - the continuous audio/visual emotion challenge," in *Proc. of ACM International Conference on Multimodal Interaction*, Santa Monica California, USA, 2012, pp. 449–456.

[9] Y. Hara and K. Itou, "Classification of emotional speech in anime films by using automatic temporal segmentation," in *Proc. of the second International Conference on Creative Content Technologies (CONTENT)*, Lisbon, Portugal, 2010, pp. 61–68.

[10] J. Walther and K. DÁddario, "The impacts of emoticons on message interpretation in computer-mediated communication," *SSCR*, vol. 19, no. 3, pp. 323–345, 2001.

[11] K. Rivera, N. Cooke, and J. Bauhs, "The effects of emotional icons on remote communication," in *Proc. CHI*, Vancouver, Canada, 1996, pp. 99–100.

[12] L. Rezabek and J. Cochenour, "Visual cues in computer-mediated communication: Supplementing text with emoticons," *Journal of Visual Literacy*, pp. 201–215, 1998.

[13] J. Lee, S. Jun, J. Forlizzi, and S. Hudson, "Using kinetic typography to convey emotion in text-based interpersonal communication," in *Proc. of the ACM Conference on Designing Interactive Systems*, Pennsylvania, USA, 2006, pp. 41–49.

[14] J. Forlizzi, J. Lee, and S. Hudson, "The Kinedit system: Affective messages using dynamic texts," in *Proc. CHI*, Florida, USA, 2003, pp. 377–384.

[15] R. Rashid, J. Aitken, and D. Fels, *Computers Helping People with Special Needs*, ser. Lecture Notes in Computer Science. Springer, 2006, vol. 4061, ch. Expressing Emotions Using Animated Text Captions, pp. 24–31.

[16] A. Kalra and K. Karahalios, *Human-Computer Interaction - INTERACT 2005*, ser. Lecture Notes in Computer Science. Springer, 2005, vol. 3585, ch. TextTone: Expressing Emotion Through Text, pp. 966–969.

[17] C. Rigaud, D. Karatzas, J. Weijer, J.-C. Burie, and J.-M. Ogier, "An active contour model for speech balloon detection in comics," in *Proc. of International Conference on Document Analysis and Recognition*, Washington, USA, 2013, pp. 1240–1244.

[18] K. Arai and H. Tolle, "Method for real time text extraction of digital manga comic," *International Journal of Image Processing (IJIP)*, vol. 4, no. 6, pp. 669–676, 2011.

[19] A. Ho, J.-C. Burie, and J.-M. Ogier, "Panel and speech balloon extraction from comic books," in *IAPR International Workshop on Document Analysis Systems*, Gold Coast, Queensland, Australia, 2012, pp. 424–428.

[20] W. Sun and K. Kise, "Similar manga retrieval using visual vocabulary based on regions of interest," in *Proc. of International Conference on Document Analysis and Recognition*, Beijing, China, 2011, pp. 1075–1079.

[21] K. Scherer, "Expression of emotion in voice and music," *Journal of Voice*, vol. 9, no. 3, pp. 235–248, 1995.

[22] M. W. F. Eyben and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia (MM)*, Florence, Italy, 2010, pp. 1459–1462.

[23] T. Kudo, "Mecab: Yet another Japanese dependency structure analyzer," http://mecab.sourceforge.net/, 2008.

[24] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," in *Proc. EMNLP*, Barcelona, Spain, 2004, pp. 230–237.

[25] R. Fan, P. Chen, and C. J. Lin, "Working set selection using second order information for training SVM," *Journal of Machine Learning Research*, vol. 6, pp. 1889–1918, 2005.

[26] J. M. Montero, J. Gutie'rrez-arriola, J. Cola's, E. Enri'quez, and J. M. Pardo, "Analysis and modelling of emotional speech in Spanish," in *Proc. ICPhS*, San Francisco, USA, 1999.

[27] J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth, "The world of emotion is not two-dimensional," *Psychological Report*, vol. 18, no. 12, pp. 1050–1057, 2007.