INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Interaction and Transition Model for Speech Emotion Recognition in Dialogue

Ruo Zhang[1,2], *Ando Atsushi*[1], *Satoshi Kobashikawa*[1], *Yushi Aono*[1]

[1]NTT Media Intelligence Laboratories, NTT Corporation, Japan
[2]School of Electrical and Computer Engineering,
Georgia Tech, North Ave NW, Atlanta, 30332, GA, USA

zhangruo1117@gatech.edu, {ando.atsushi, kobashikawa.satoshi, aono.yushi}@lab.ntt.co.jp

Abstract

In this paper we propose a novel emotion recognition method modeling interaction and transition in dialogue. Conventional emotion recognition utilizes intra-features such as MFCCs or F0s within individual utterance. However, human perceive emotions not only through individual utterances but also by contextual information. The proposed method takes in account the contextual effect of utterance in dialogue, which the conventional method fails to. Proposed method introduces Emotion Interaction and Transition (EIT) models which is constructed by end-to-end LSTMs. The inputs of EIT model are the previous emotions of both target and opponent speaker, estimated by state-of-the-art utterance emotion recognition model. The experimental results show that the proposed method improves overall accuracy and average precision by a relative error reduction of 18.8% and 22.6% respectively.

Index Terms: speech emotion recognition, contextual information, long short-term memory, end-to-end model

1. Introduction

Emotion plays an important role in human communication, as it helps to convey and understand actual messages. This information can be contained in acoustic audio, visual expression, and the underlying lexical meaning of speech as generated by humans [1]. In recent years, emotion recognition has become more and more important in the research topic area and focus is being placed on classifying emotion in audio signals [2, 3]. Emotion recognition is seen as critical in Human Computer Interface (HCI) when it comes to applications in the domains of software engineering, website customization, education and gaming, for it helps the machine better understand humans [4]. The aim of this research is speech emotion recognition in human-to-human dialogues.

Numerous studies have examined speech emotion recognition. The most popular approach is based on heuristic features extracted from individual utterances [5]. The statistics of low-level descriptors (LLDs) such as fundamental frequency (F_0), energy, Mel-Frequency Cepstral Coefficients (MFCCs), zero-crossing ratio in an utterance are used as heuristic features [6]. In addition to these, visual [7] and linguistic characteristics [8] are also employed. However, one of the problems of this approach is that it is difficult to find effective features for estimation [9]. Therefore, in recent years, some researchers have tried to obtain features automatically. Convolutional Neural Network (CNN) [10] and Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) [11] were used to extract utterance features. The recently proposed combination of LSTM-RNN with full-connected neural networks as sequence-to-sequence with attention model shows the best performance in terms of individual utterance emotion recognition [12, 13].

In addition to emotion estimation in individual utterances, the contextual information of dialogues is another focus of attention. Contextual features such as change rates of prosodic features between a target and the previous utterance of a speaker was shown to improve performance [14]. In contrast to this feature-based approach, several studies explicitly modeled the emotion transition of target speakers by employing Hidden Markov Models (HMMs) or LSTM-RNN [15]. One key advantage of the modeling approach is that it can deal with larger sets of contextual information than feature-based methods.

In this paper, we expand a conventional single-speaker emotion transition model to a multi-speaker emotion model to utilize the information generated by interactive dialogues. According to the emotion generation theory [16], human emotion is affected by not only self-contextual information but also the surroundings. Therefore, interactive information such as previous utterance of the opposite speaker will improve the recognition of the emotion of the target speaker.

Our contributions of this paper are as follows:

- We reveal the effectiveness of interactive information for emotion recognition in dialogues by analyses of a dialogue speech dataset.
- LSTM-RNN is employed to model both the interaction and transition of the emotions of two speakers; the conventional model utilizes only self-transition.
- State-of-the-art automatic feature extraction and emotion recognition from individual utterances is applied; the conventional method utilizes heuristic features.

The outline of this paper is as follows. The analysis and modeling of interactive information for emotion recognition are elucidated in Section 2. Section 3 introduces the proposed method. Section 4 and 5 cover the experiments and discussions.

2. Analysis of Emotional Interaction

This section analyses the emotional interaction information present in the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [17].

IEMOCAP contains natural conversational speech between a male and a female. There are 5 pairs and 149 dialogues in total. Total utterance number is 10039 and the average duration is 4.5 seconds. Each utterance is tagged with both 10-class categorical emotion labels and 3-dimensional continuous emotion values; valence, arousal and dominance.

This paper tackles 5-class emotion classification; *angry, happy, sad, neutral* and *others*. The labels of *angry, happy, sad, neutral* in the database are used unchanged, while the other labels are taken as *others*. The classes have 1103, 595, 1084, 1708, and 5549 utterances, respectively. Note that 45% of *others* are utterances that lacked a majority class decision, e.g.one

http://dx.doi.org/10.21437/Interspeech.2017-713

Table 1: *Bi-grams of emotions between those of current and previous of target speakers (left) and those of current target speaker and previous opponent speaker (right).*

		Previous Emo. (Target Spk.)					Previous Emo. (Opposite Spk.)				
		angry	*happy*	*sad*	*neutral*	*others*	*angry*	*happy*	*sad*	*neutral*	*others*
Current Emo. (Target Spk.)	*angry*	**0.68**	0.00	0.01	0.02	0.29	**0.38**	0.00	0.02	0.24	0.36
	happy	0.00	**0.47**	0.00	0.10	0.43	0.00	**0.45**	0.02	0.13	0.40
	sad	0.02	0.04	**0.69**	0.08	0.16	0.03	0.01	**0.42**	0.33	0.21
	neutral	0.02	0.03	0.06	**0.61**	0.28	0.17	0.04	0.13	**0.18**	0.47
	others	0.06	0.06	0.04	0.14	**0.71**	0.12	0.06	0.04	0.20	**0.58**

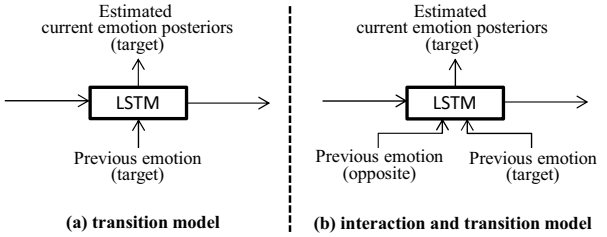

Figure 1: *Models for analysis of emotion estimation from contextual information.*

annotator assigned the label *happy* while another gave it *neutral*.

We first analyzed the characteristics of emotion transition and interaction from the bi-grams of the emotions of target speaker and opposite speaker. The result is shown in Table 1. The left part of the table shows that one trait of emotion transition is high continuity, which matches the conventional method. Furthermore, the right part of the table shows that there is also some synchronicity between previous emotion of opposite speaker and current emotion of target speaker. This indicates that humans often come to adopt the same emotion as the opposite speaker; i.e. sympathy.

A second analysis confirmed that this contextual information can be utilized for emotion recognition. Two simple models shown in Figure 1, which enable us to treat just contextual information, were used for this analysis. They estimated the current emotion of target speaker from the previous emotion state of target speaker (transition model), or those of both the target and opposite speaker (interaction and transition model). The transition model is regarded as the same as the conventional single-speaker emotion transition approach. LSTM-RNN was employed for both models. The numbers of layers and units of LSTM-RNN were 1 and 64, respectively. Ground-truth emotion labels and continuous values of individual utterances were used as emotion state inputs. Dialogues of 4 pairs in IEMOCAP were used for model training while the remaining pair data was tested.

The accuracies of estimating emotion from contextual information are shown in Table 2. *Random* and *Copy previous* means selecting the emotion class randomly and selecting the previous emotion of target speaker in a dialogue, respectively. From the table, the interaction and transition model has better estimation accuracy than the transition model. Therefore, interactive information matches conventional self-transition, while considering both the transition and interaction of emotions will improve emotion recognition accuracy.

Table 2: *Accuracies of estimating emotions from contextual information.*

	Acc. [%]
Random	20.0
Copy previous	49.1
(a) Transition model	50.6
(b) Interaction and transition model	**63.3**

3. Emotion Recognition using Interaction and Transition Model

In this section, we propose a new framework, the Emotion Interaction and Transition (EIT) model, to utilize contextual information for emotion recognition. The proposed model is overviewed in Figure 2.

The structure of EIT is similar to that in Figure 1, except for the time pooling layer and emotion posterior probabilities of the current utterance. The time pooling layer stores the emotion posterior probabilities of the previous utterance of each speaker in a dialogue. Emotion posterior probabilities of the current utterance represent the individual results for the target speaker. These posteriors are estimated individually by the state-of-the-art individual emotion recognition method, Attention-Based-Sequence-to-Sequence (ABS2S) model [12]. These posteriors are concatenated as a single vector and fed to the LSTM layer. This structure allows EIT to re-estimate emotion posteriors of the current utterance from not only individual results of the current utterance but also the influences of both target speaker transition and opposite speaker interaction.

Since the EIT part and individual emotion estimation part can be regarded as a single network, we proposed two different training scenarios: either jointly or separately training the two parts. The separate scenario uses the best model of the already trained individual emotion recognition method to fit EIT, calculates loss, and then back-propagates the loss in EIT alone. The flow of this scenario is shown in Figure 3. In the joint scenario, only the individual emotion recognition part is trained in the first few epochs, and then the entire network is trained. The aim is to let the individual emotion recognition part learn roughly correct emotion results before optimizing whole network, which will yield stable training. The number of these initial epochs is determined empirically by the best model of baseline.

4. Experiments

4.1. Experiment Setup

We evaluated the proposed method using the IEMOCAP database. In the experiments, dialogues of 4 pairs were used as training data and those of the remaining pair was tested.

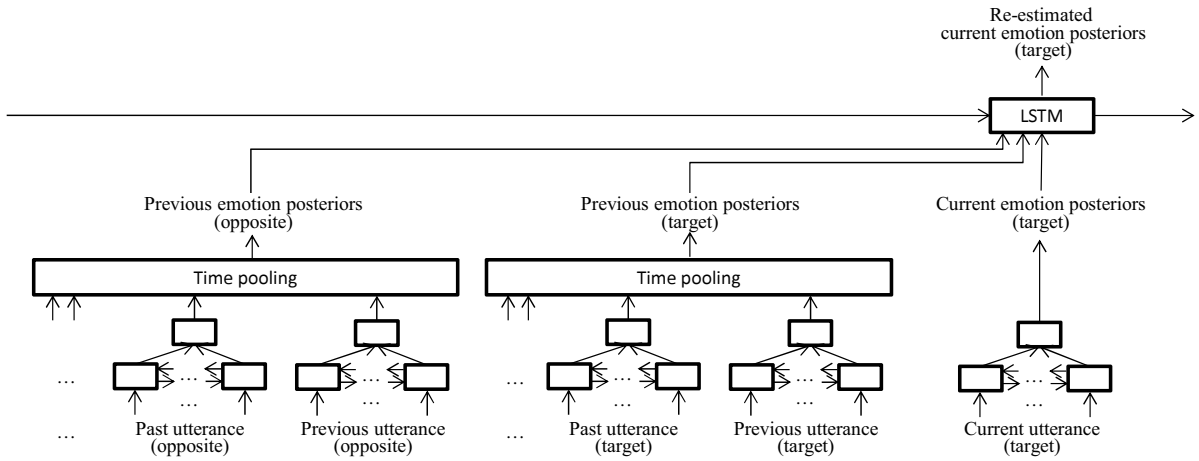

Figure 2: *Structure of proposed EIT model.The lower part is the same as conventional individual emotion recognition model.*

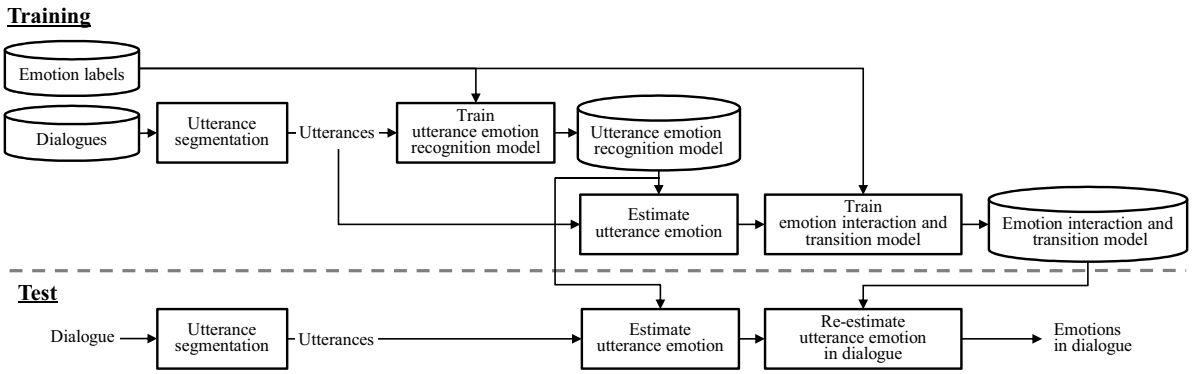

Figure 3: *Flow of separated EIT training.*

The training and test data had 8220 and 1819 utterances, respectively. 32-dimensional LLDs were extracted as frame-level acoustic features, including fundamental frequency (F_0), energy, 12-dimensional MFCCs, voice probability, zero-crossing ratio and their first order derivatives. For performance comparisons, two other methods were also evaluated; emotion recognition from individual utterances and emotion recognition based on the emotion transition (ET) model. ET model is the proposal, EIT, with the removal of the posteriors of the previous utterance of the opposite speaker. This is the same approach as the conventional self-transition based method [15] thus we regarded it as the baseline. ABS2S model was employed for emotion recognition from individual utterances and the lower part of EIT and ET model. It consisted of a 1-layer bidirectional LSTM with 64 hidden units and attention module with 64 hidden units. Both EIT and ET model used 1-layer LSTM with 64 hidden units. Overall accuracy (*Acc.*) and average precision of each emotion (*AvP.*) were employed to measure performance.

4.2. Results and discussions

The results of the emotion estimation experiments are shown in Table 3. The EIT and ET model shown in this table used the separated training scenario. The table reveals that the two contextual-information based methods were superior to the individual estimation of emotions. Furthermore, EIT improves both accuracy and average precision compared with the conventional ET model. This indicates that not only transition but also interaction information are important in estimating speaker's emotion in a dialogue.

Table 3: *Accuracies and average precisions of the methods.Indiv.means emotion recognition from individual utterances.*

	Acc. [%]	*AvP.* [%]
Indiv.	51.7	36.8
ET model (Baseline)	57.6	44.1
EIT model (Proposed)	**60.8**	**51.1**

In our experiments, joint training was less effective than separated training. We suspect this is due to the overly-complicated procedure for loss back-propagation. Because of three connections of individual emotion recognition networks and time pooling layer, the routes used to propagate loss to the individual emotion recognition network are too complex. This made training of the entire model more difficult. One possible solution to this problem is to limit the route of backpropagation; for example, losses are not propagated to the past utterances of the target and the opposite speaker.

Finally, we plot an example of emotion estimation results in a dialogue in Figure 4. As we can see in the first several utterances in a dialogue, sad posteriors of EIT and ET model are

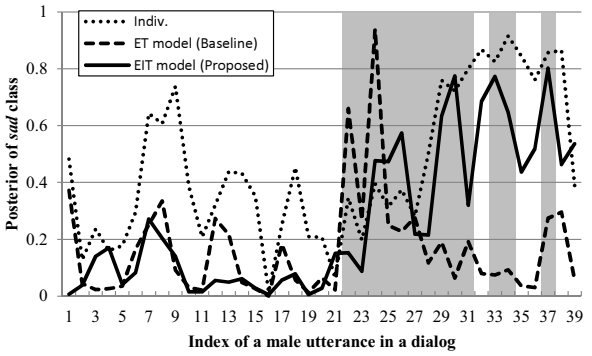

Figure 4: *Posteriors of sad emotion in male utterance in a dialogue.Gray background means that correct emotion is sad in the utterance.*

lower than those determined in individual emotion estimation. It is considered that contextual information gives certain constraints of emotional changes, which prevents emotion recognition error. On the other hand, the results of the last ten utterances show that EIT yield better sad emotion estimates than the ET model. We suspect that EIT is more robust than ET model in past emotion recognition errors of target speaker because EIT is affected by emotions of interlocutor. These indicate that EIT is able to effectively ameliorate the weaknesses of individual emotion recognition.

5. Conclusions

In this paper we proposed a novel emotion recognition method for dialogues. Our proposal utilizes both interactive and transitional features in contextual information, an aspect that conventional methods tend to utilize only partially. Through the analyses of a major emotion dialogue corpus, we showed that modeling both emotion interaction and transition is a more effective way of estimating the current true emotion. Our proposed method, EIT, utilizes end-to-end LSTMs to model both interaction and transition and employs a state-of-the-art individual emotion recognition method. Experiments showed that the proposed model achieved an improvement of 18.8% in overall accuracy and 22.6% in average precision relative to the current state-of-the-art individual emotion recognition method.

Future works include analyses of the effects of the emotion categories. EIT currently uses *other* class as a tag to cover several different types of emotions. It is required to investigate the optimal set of emotion categories for utilizing contextual information in emotion recognition. Other works include evaluating EIT with other datasets and to integrate multimodal information such as visual and linguistic features into our framework.

6. References

[1] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[2] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[3] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10–11, pp. 787–800, 2007.

[4] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel, *Emotion Recognition and Its Applications*, 2014, pp. 51–62.

[5] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[6] R. Banse and K. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[7] P. Ekman, *Emotion in the Human Face*. Cambridge University Press, 1982.

[8] L. Devillers, L. Lamel, and I. Vasilescu, "Emotion detection in task-oriented spoken dialogues," in *Proceedings of ICME*, vol. 3, 2003, pp. 549–552.

[9] C. Brester, E. Semenkin, and M. Sidorov, "Multi-objective heuristic feature selection for speech-based multilingual emotion recognition," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 6, no. 4, pp. 243–253, 2016.

[10] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 801–804.

[11] R. B. George Trigeorgis, Fabien Ringeval, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proceedings of ICASSP*, 2016, pp. 5200–5204.

[12] C. W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition," in *Proceedings of INTERSPEECH*, 2016.

[13] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proceedings of ICASSP*, 2017, pp. 2227–2231.

[14] J. Liscombe, G. Riccardi, and D. Hakkani-Tür, "Using context to improve emotion detection in spoken dialog systems," in *Proceedings of INTERSPEECH*, 2005, pp. 1845–1848.

[15] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *Proceedings of INTERSPEECH*, 2010, pp. 2362–2365.

[16] J. J. Gross and L. F. Barrett, "Emotion generation and emotion regulation: One or two depends on your point of view." *Emotion review*, vol. 3, no. 1, pp. 8–16, 2011.

[17] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.