Speaker and Emotion Recognition of TV-Series Data Using Multimodal and Multitask Deep Learning

Sashi Novitasari[1] Quoc Truong Do[1] Sakriani Sakti[1,3] Dessi Lestari[2] Satoshi Nakamura[1,3]
[1]Graduate School of Information Science, Nara Institute of Science and Technology
[2]Department of Informatics, Bandung Institute of Technology
[3]RIKEN AIP

[1]{sashi.novitasari.si3, do.truong.dj3, ssakti, s-nakamura}@is.naist.jp
[2]{dessipuji}@informatika.org

1 Introduction

In our daily lives, the interaction between humans involves both linguistic and paralinguistic aspect to make a better understanding of the interaction. Several studies previously were conducted to enable recognition of the nonverbal aspect by machine, for the example are the recognition of speech emotion [2] and speaker [6], which are the important nonverbal aspects in the conversation.

In recent time, study on multimodal and multitask recognition in human-machine communication field have gained attention. Emotion [8] or speaker [9] recognition using multi-modality had also been done by incorporating features from various aspects of video and resulted in the improvement. Studies on the multitask recognition also have been conducted, for the examples are recognition on speech and emotion [7]. In spite of these achievements, the study on multitask recognition that utilize multimodal features is still limited.

In this work, we conducted a study on both multimodal and multitask recognition by using deep learning to recognize the speaker and emotion of a speech from TV-series. The utilized multimodal features consist of acoustic, lexical, and facial features, which can be extracted from TV-series data. We explored two methods to combine the multimodal features to see the better choice for utilizing multimodal features. The applied methods include features concatenation and hierarchical fusion.

2 Model Architectures

2.1 Multimodal Classification

We evaluated two combination methods to combine the multimodal features. In the first method, feature concatenation, we concatenated the multimodal features into one feature set for the classification. In the second method, feature hierarchical fusion, the first layer of the neural network will take the first type of features, and its output will be concatenated with another type of features, and so on so the features will be fused hierarchically. Based on our experiment, the best fusion order was lexical, acoustic, then facial.

2.2 Multitask Classification

The multitask classification model classify three tasks that consist of speaker name, valence, and arousal of the input speech. Valence and arousal are emotion dimensions which are commonly used in the emotion recognition systems. Since valence and arousal are originally represented as a continuous value, we segmented each dimension into three categories based on the value: positive, negative, and neutral, to simplify the recognition.

3 Features

The utilized features include acoustic, lexical, and facial features, with the acoustic features as our main

features. Each feature extracted as utterance-level features to enable the utterance-level classification in our experiment. The acoustic features were extracted from the speech audio by using the openSMILE toolkit[3]. We used INTERSPEECH 2010 feature configuration[4] for the classification. The lexical features were extracted from the subtitle of video as an average vector of word-vectors for each utterance. The word-vectors were generated by using Google Word2Vec[5] model. The facial features were extracted from the speaker's facial structure in the video by using the openFace toolkit[1].

4 Experiment

4.1 Dataset

We utilized dataset that constructed from an English TV-series data, which has been broadcast and also has been released in DVD commercially. The dataset consists of 2760 English utterances that collected from several random episodes. Each utterance was annotated with its speaker and emotion information after the collection process. The annotation resulted in 57 classes of speaker identity, while the emotion dimensions consist of 3 classes each. Labels of each task were not evenly distributed because of the uncontrolled environment of the source data. In the experiment, we utilized 2460 utterances as the train data and 300 utterances as the test data. Due to the limitation on the data size, the test data only included 10 speakers with 30 samples each.

4.2 Results

We conducted the experiment by using multilayer perceptron models. The multimodal and multitask models were also evaluated by comparing its performance to unimodal and single-task models. The results can be seen in Table 1. In this table, the symbols 'A', 'F', and 'L' denotes acoustic features, facial features, and lexical features respectively. The symbol 'U', 'C', and 'H' denotes unimodal classification and the multimodal features combination methods which consist of concatenation and hierarchical fusion respectively.

Table 1: Experiment results (F1-score %).

Feature(s)	**Speaker**	**Valence**	**Arousal**
Single-task			
U: A	<u>62.77</u>	54.31	44.59
C: A+F	56.73	55.45	45.78
C: A+F+L	56.23	54.10	47.66
H :A+F+L	61.21	52.65	47.87
Multitask			
U: A	60.13	<u>56.21</u>	<u>49.71</u>
C: A+F	58.51	53.84	48.66
C: A+F+L	61.01	54.10	47.77
H :A+F+L	52.71	53.51	46.44

These results show that the multitask model, which only used acoustic features, resulted in the improvement than the single-task model for emotion recognition. The speaker, however, was best identified by the unimodal and single-task model. Unfortunately, the multitask model did not result in improvement when it utilized the multimodal features. The utilization of the multimodal features for the multitask recognition resulted in a complex model. Since the train data was limited, the simpler model or unimodal-multitask model performed better than the multimodal-multitask model.

5 Conclusion

We constructed the multimodal and multitask speaker and speech emotion recognition model by using deep learning and TV-series data. The multitask model outperformed the single-task model, especially in emotion recognition, by using acoustic features only. However, the multimodal-multitask model did not result in a significant improvement due to the limitation on the data size.

6 Acknowledgment

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.

References

[1] T. Baltrusaitis, P. Robinson, and L.P. Morency. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, March 2016.

[2] E. Bozkurt, E. Erzin, .E. Erdem, and A.T. Erdem. Interspeech 2009 emotion recognition challenge evaluation. In *2010 IEEE 18th Signal Processing and Communications Applications Conference*, pp. 216–219, April 2010.

[3] F. Eyben, M. Wollmer, and B. Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pp. 1459–1462, New York, NY, USA, 2010. ACM.

[4] W.A. Jassim, R. Paramesran, and N. Harte. Speech emotion classification using combined neurogram and interspeech 2010 paralinguistic challenge features. *IET Signal Processing*, Vol. 11, No. 5, pp. 587–595, 2017.

[5] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pp. 3111–3119, USA, 2013. Curran Associates Inc.

[6] D.A. Reynolds. Automatic speaker recognition using gaussian mixture speaker models. *The Lincoln Laboratory Journal*, pp. 173–192, 1995.

[7] Z. Tang, L. Li, and D. Wang. Multi-task Recurrent Model for Speech and Speaker Recognition. *ArXiv e-prints*, March 2016.

[8] L. Tian, J. Moore, and C. Lai. Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pp. 565–572, Dec 2016.

[9] F. Vallet, S. Essid, and J. Carrive. A multimodal approach to speaker diarization on tv talk-shows. *IEEE Transactions on Multimedia*, Vol. 15, No. 3, pp. 509–520, April 2013.