Dimensional Speech Emotion Recognition from Acoustic and Text Features Using Multitask Learning

◎ Bagus Tris Atmaja* , Masato Akagi (JAIST)

1 Introduction

Understanding human emotion is important to make a proper response in a particular situation. It is not only useful for human-human communication, but also for future human-machine communication. Emotion can be recognized from many modalities: facial expression, speech, and motion of body parts. In the absence of visual features, speech is the only way to recognize emotion such as in telephone or call center application. By identifying caller emotion automatically from a system, appropriate feedbacks can be taken quickly and precisely.

Speech is a modality where both acoustic and verbal information can be extracted to recognize human emotion. However, most speech emotion recognition (SER) systems use only acoustic features and show poor performance compared to multimodal emotion recognition systems. In this research, the use of acoustic and text features is proposed to improve the SER performance. Text can be extracted from speech and it may contributes to emotion recognition. For example, an interlocutor can perceive emotion not only from prosodic information but also from semantics.

This study on speech emotion recognition from acoustic and text features is aimed to evaluate the combination of acoustic and text features to improve the performance of dimensional automatic speech emotion recognition. Current research on pattern recognition also shows that the use of multimodal features from audio, visual, and motion capture increases the performance compared to single modal [2]. The big data research also shows that the use of more data improves the performance compared to smaller data on the same algorithm. Given the acoustic and text features, an improvement of speech emotion recognition should be obtained. In this case, many technologies such as human-computer interaction can potentially benefit from such improvement. To determine that the use of acoustic and text features improve significantly on dimensional SER over acoustic feature only, we assert significance test at $p-$value=0.05 using the two-tailed paired t-test.

2 Dataset and Features

IEMOCAP: The IEMOCAP dataset [1] consists of 10039 turns with valence, arousal, and dominance label. Each label is annotated by at least two raters. The average score among raters is used to train the features. The original scores of those emotion dimensions in range [1,5] are converted to [-1,1].
GeMAPS: The GeMAPS feature set [2] is a collection of acoustic features extracted on frame-based processing that is designed for affective application. This feature set consists of 23 acoustic features including loudness, alpha ration, Hammerberg index, MFCCs, F1, F2, F3 and others. GeMAPS feature set is extracted from IEMOCAP speech data with 25 ms of window length and 10 ms hop length.
Word Vector: Word vector is the text features extracted from IEMOCAP text data. Global Vector (GloVe) embedding based on [3] is used to weight the obtained word vector from IEMOCAP text with the size of 300 dimension.

3 Dimensional SER System

Our SER system is concatenation of two networks, acoustic network and text networks. Acoustic features are fed into acoustic network, and text features are fed into text network. Both networks used three LSTM layers, with 64 nodes on each layer. The output of both networks are concatenated and coupled with two dense layers with 64 and 32 nodes. The last dense layer is split into three dense networks with 1 node each to predict the score of valence, arousal, and dominance (VAD). This architecture of the proposed dimensional SER system is shown in Fig. 1.

4 Multitask Learning

As shown in Fig. 1, the output of the system is to predict degrees of valence, arousal, and domi-

*Corresponding author, email: bagus@jaist.ac.jp

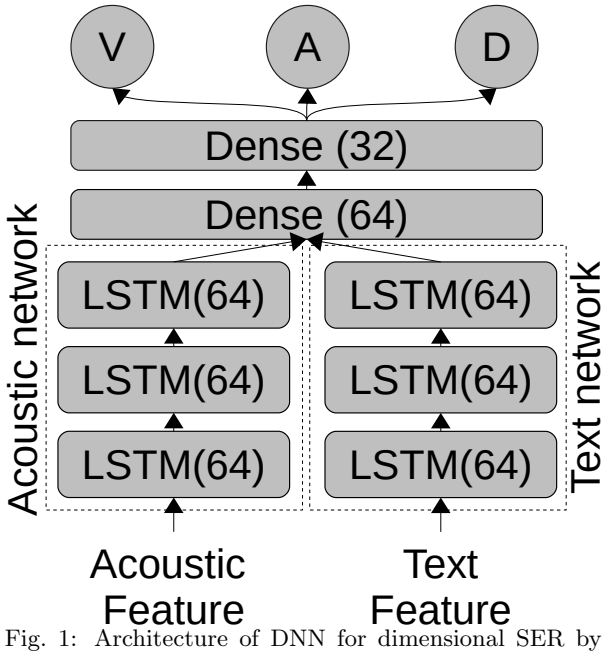

Fig. 1: Architecture of DNN for dimensional SER by combining acoustic and text features

nance (3D) simultaneously. To tackle this issue, we proposed to use concordance correlation coefficient (CCC)-based multitask learning which jointly learning to predict those three emotions. That CCC loss (CCCL) is formulated as follows,

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (1)$$

$$CCCL = 1 - CCC \quad (2)$$

where ρ is Pearson correlation between predicted emotion degree x and true emotion degree y, σ^2 is a variance and μ is a mean. As the learning process minimizes three variables, we used the following multi-task learning approach to optimize CCC score.

$$CCCL_{tot} = CCCL_V + CCCL_A + CCCL_D \quad (3)$$

Different from previous research [4], we do not use any weighting factors to adjust CCCL from three emotion dimensions.

5 Results and Discussion

Table 1 shows the result of CCC score on dimensional SER experiment from previous and current methods. The differences between Ref. [4] and *speech* method are (1) the use of smaller batch size, and (2) no weighting factor used on speech method. The use of smaller batch-size on speech-based emotion recognition improves the performance on the used dataset. In *(*speech+text) method, acoustic and text features are used. The use of text features significantly improved the performance of dimensional SER, particularly on valence dimension (p-value=4.05×10^{-20}).

Table 1: Result of CCC scores on dimensional SER

Method	V	A	D	Total
Ref. [4]	0.11	0.43	0.36	0.9
speech	0.110	0.494	0.352	1.056
speech+text	0.416	0.506	0.476	1.399

Fig. 2 shows the score of valence, arousal, and dominance in 20 experiments. It is shown that our proposed method can obtain a consistent result (mean total CCC score=1.399). The next research is directed to improve this score to close human annotation performance.

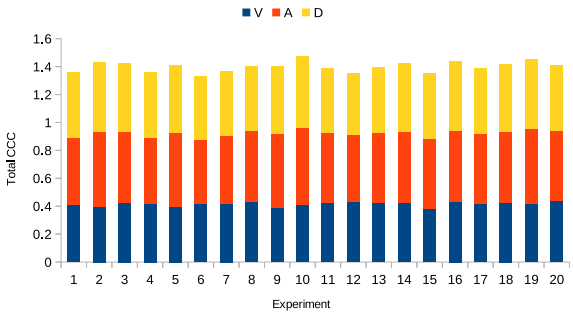

Fig. 2: Score of VAD in 20 experiments

References

[1] Busso, Carlos *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database." Language resources and evaluation 42, no. 4 (2008): 335.

[2] Eyben, Florian *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing." IEEE Transactions on Affective Computing 7, no. 2 (2015): 190-202.

[3] Pennington, Jeffrey *et al.*, "Glove: Global vectors for word representation." In Proc. of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

[4] Atmaja, B.T. and Akagi, Masato. "RNN-based Dimensional Speech Emotion Recognition.", in Proc. ASJ Autum Meeting. 2019.