International Journal of Informatics, Information System and Computer Engineering

Dimensional Speech Emotion Recognition from Acoustic and Text Features using Recurrent Neural Networks

Bagus Tris Atmaja[1,2] *Reda Elbarougy*[3], *Masato Akagi*[2]
[1] Sepuluh Nopember Institute of Technology, Surabaya, Indonesia
[2]Japan Advanced Institute of Science of Technology, Nomi, Japan
[3]Damietta University, New Damietta, Egypt
E-mail: bagus@ep.its.ac.id

ABSTRACTS

Emotion can be inferred from tonal and verbal information, where both features can be extracted from speech. While most researchers conducted studies on categorical emotion recognition from a single modality, this research presents a dimensional emotion recognition combining acoustic and text features. A number of 31 acoustic features are extracted from speech, while word vector is used as text features. The initial result on single modality emotion recognition can be used as a cue to combine both features with improving the recognition result. The latter result shows that a combination of acoustic and text features decreases the error of dimensional emotion score prediction by about 5% from the acoustic system and 1% from the text system. This smallest error is achieved by combining the text system with Long Short-Term Memory (LSTM) networks and acoustic systems with bidirectional LSTM networks and concatenated both systems with dense networks.

ARTICLE INFO

Article History:
Received 17 Nov 2020
Revised 20 Nov 2020
Accepted 25 Nov 2020
Available online 26 Dec 2020

Keywords:
Speech Emotion,
Neural Network,
LSTM.

1. INTRODUCTION

The demand for recognizing emotion in speech has grown increasingly as a human emotion can be expressed via speech, and many applications, such as call center, telephone communication, and voice messages, can benefit from this speech emotion recognition. The study of speech emotion recognition was established some decades ago using unsupervised learning and a small amount of data. Advancements in computation hardware and in the development of larger speech corpus have enabled us to analyze emotion in a speech

Detecting emotion is useful to investigate whether a student is confused, engaged, or certain when interacting with a tutorial system or whether a caller to help a line is frustrating or not (Jurafsky et al., 2014). By gaining knowledge of emotion from student and caller in both cases, proper action can be taken to avoid the worse condition. The degree of emotion (in the numeric score) in both cases is more relevant than the category of emotion (joy or sad, for example). These are two examples where dimensional emotion is more informative than categorical emotion.

Although research on emotion recognition has been conducted progressively, most re- search are focused on recognition of categorical emotion such as in (Griol et al., 2019; Chen et al., 2018; Atmaja et al., 2019). As shown by the previous two examples, a dimensional approach of emotion recognition is more informative in such cases. Recognizing the degree of emotion is a more challenging task as it tries to predict the numerical score rather than a category. This type of task is a class of logistic regression.

Research investigating dimensional emotion recognition in a text is reported by Calvo et al., 2020. They found by using the same classifier, i.e., non-negative matrix factorization (NMF), both categorical and dimensional emotion recognition obtain a similar result. They used emotional terms from an affective dictionary as text features for the dimensional task. In speech emotion recognition, the study of dimensional emotion recognition is reported by (Giannakopoulos et al., 2009) using a small dataset from videos, ten dimensions of acoustic features, and k-Nearest Neighbor (kNN) to estimate emotion degree. The results indicate that the resulting architecture can estimate emotion states of speech from movies with sufficient accuracy (Valence: 24%, Arousal: 35%, in terms of R^2 statistics). Both dimensional text and speech emotion recognition above used a non-deep neural network (DNN) method due to the time and size of data.

Another challenge in speech emotion recognition, besides a dimensional approach, is the strategy for extracting features. The features are the input of an emotion recognition system, and the performance of the system depends on those features. An issue to be considered when extracting features for speech emotion recognition is the necessity of combining speech (acoustic feature) with other types of features (El Ayadi et al., 2011). We choose text features as it can be extracted from speech via automatic speech recognition (ASR). The combination of these acoustic and text features is expected to improve the performance of the emotion recognition rate compared to the use of single modality i.e., acoustic feature or text feature only.

This paper presents a dimensional speech emotion recognition from a multimodal dataset. The purposes of this work are (1) to examine whether the fusion of two related features can decrease the error of dimensional emotion recognition and (2) to find the best DNN architecture for a list of DNN layer combination. A deep learning-based classifier from the category of recurrent neural network has been built for this purpose. Two types of features are used: acoustic and text features. For each feature, a set of networks is stacked. The two networks from acoustic and text features are then concatenated using late fusion architecture. The result shows that the proposed method can improve the performance compared to the method that used acoustic or text features only. The evaluation is presented in terms of mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). To extend this work, a discussion to evaluate the metric used in this research is summarized at the end of the paper.

2. DATASET

The IEMOCAP (interactive emotional dyadic motion capture) database developed by the University of Southern California was used in this research (Busso et al., 2008). A number of 10039 turns (utterances) are recorded and measured, including included speech, visual, text, and motion capture (face, head, and hand movement). From those modalities, speech signal and text transcription are used. The dimension labels are given for valence, arousal, and dominance (VAD) in a range of 1 to 5 via self-assessment manikins (SAMs). All utterances on this dataset are used in this research. From these data, 80% is used for training, and 20% is used for the test. Twenty percent of the training data is used for validation.

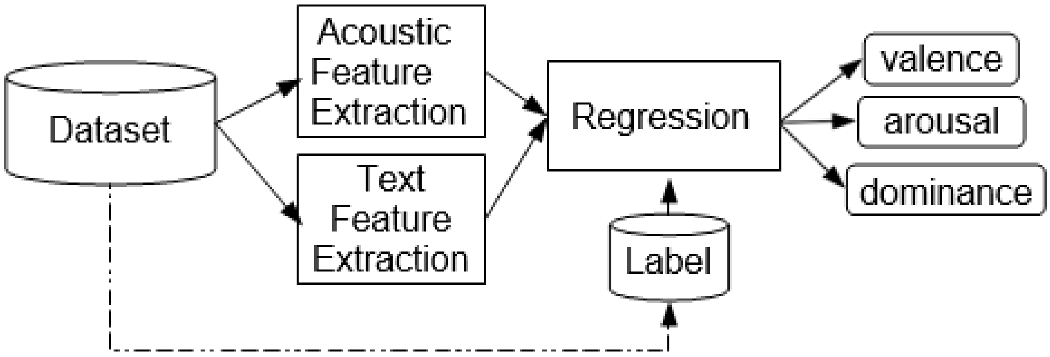

Fig. 1. Proposed dimensional speech emotion recognition from acoustic and text features. The dash line between label and dataset means that label is obtained from dataset directly.

3. PROPOSED METHOD

A proposed method of this research paper can be split into two parts: feature extraction and dimensional emotion classifier. A block diagram of the proposed system is shown in Fig. 1. From the dataset, two features are extracted: acoustic and text features. The extracted feature then is fed into a classifier where the regression process is performed by combining those two features using the late fusion method. Finally, the classifier produces the predicted emotion dimension, which will be compared to the true value label. The difference between true value label and predicted emotion dimension is the error, which is measured in three different ways.

3.1 Feature Extraction

Two sets of features from acoustic and text are used to extract emotion from speech. The following is the description of those two sets of features.

3.1.1 Acoustic Feature Extraction

A number of 31 acoustic features are used in this research. These features are,

- three time-domain features: zerocrossing rate (ZCR), energy, the entropy of energy.
- five spectral-domain features: spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off.
- 13 MFCC coefficients.
- five fundamental frequencies (for each window).
- five formants (for each window).

We limit the number of windows for each utterance to 100 with 20 ms window length and 10 ms overlap. The resulting size of the acoustic feature then is (100, 31) for a single utterance. The total size of acoustic features for all utterances within the dataset is (10039, 100, 31).

3.1.2 Text Feature Extraction

Text features can be obtained in many ways. One of the simple yet powerful methods is by word embedding (Penningtonet al., 2014). A word embedding is a vector representation of a word. A numerical value in the form of a vector is used to make the computer to be able to process text data as it only processes numerical value. This value is the points (numeric data) in the space of dimension, in which the size of the dimension is equal to the vocabulary size. The word representations embed those points in a feature space of lower dimension (Goodfellow et al., 2016). In the original space, every word is represented by a one-hot vector, a value of 1 for the corresponding word, and 0 for others. This element, with a value of 1, will be converted into a point in a range of vocabulary size.

To obtain a vector of each word in an utterance, that utterance in the dataset must be tokenized. Tokenization is a process to divide an utterance to the number of constituent words. The following is the example of a single utterance from IEMOCAP dataset with its tokenization and a resulted text vector for each word.

```
text = "Excuse me."
tokenized_text = ["Excuse",
"me"] text_vector = [832, 18]
```

To obtain the fixed length of a vector for each utterance, a set of zeros can be padded before or after the obtained vector. The size of this zeros sequence can be obtained from the longest sequence, i.e., an utterance within the dataset, which has the longest words, subtracted by the length of a vector in the current utterance. We set the longest sequence for the IEMOCAP dataset for 554 sequences.

A study to vectorize certain words has been performed by several researchers (Mikolov et al., 2013; Penningtonet al., 2014; Mikolov et al., 2017). The vector of those words can be used to weight the word vector obtained previously. The size of the dimension of each word for pre-trained word vectors is 300 (in the example above is one), shaping the size of (554, 300) text feature for each utterance, or (10039, 554, 300) for all utterances in the IEMOCAP dataset.

4. DIMENSIONAL EMOTION CLASSIFIER

Recurrent neural network (RNN) is one of the variants of the neural network that are designed to handle sequential information. These networks introduce state variables to store past information and determine the current output based on the current input. Let **H** is the output of hidden layer, **X** is the input and **W** is the weight of layer with bias b. Then,

$$\mathbf{H} = \varphi(\mathbf{X}\mathbf{W}_{xh} + \mathbf{b}_{h}). \quad (1)$$

is the output of hidden layer **H** with φ is non- linear function (activation). Having a recurrent hidden state ($\mathbf{H}_{t}$) whose activation at each time is dependent on that of the previous time ($\mathbf{H}_{t-1}$), the output of current hidden layer now is defined as,

$$\mathbf{H}_{t} = \varphi(\mathbf{X}_{t}\mathbf{W}_{xh} + \mathbf{H}_{t-1}\mathbf{W}_{hh} + \mathbf{b}_{h}). \quad (2)$$

The problem with that RNN is it always takes past time into consideration. A situation may be encountered when the early observa- tion is more/less significant to predict the fu- ture. To tackle this situation and adding some enhancements, several methods have been pro- posed by some researchers (Cho et al., 2014; Hochreiter et al., 1997). In this paper, those two RNN methods are implemented as dimensional emotion classifier.

4.1 Gated Recurrent Unit

Gated recurrent unit (GRU) enables the gating of the hidden state in RNN. This is a mechanism that is enabled for when the hidden state should be updated and when it should be reset. It is learned and addressed some limitations of RNN e.g., whether an early observation is highly significant for predicting all future observations. If the first observation is likely of great importance, it will learn not to update the hidden state after the first observation. Like- wise, it will learn to skip irrelevant temporary observations. Finally, it will learn to reset the latent state whenever needed (A. Zhang et al., 2019).

Reset unit, $\mathbf{R}_{t}$, and update unit, $\mathbf{Z}_{t}$, are the new additional units in GRU. Together with candidate unit, $\hat{\mathbf{H}}_{t}$, it updates the GRU in the following order.

$$\mathbf{R}_{t} = \sigma(\mathbf{X}_{t}\mathbf{W}_{xr} + \mathbf{H}_{t-1}\mathbf{W}_{hr} + \mathbf{b}_{r}) \quad (3)$$

$$\mathbf{Z}_{t} = \sigma(\mathbf{X}_{t}\mathbf{W}_{xz} + \mathbf{H}_{t-1}\mathbf{W}_{hz} + \mathbf{b}_{z}) \quad (4)$$

$$\tilde{\mathbf{H}}_{t} = \tanh(\mathbf{X}_{t}\mathbf{W}_{xh} + (\mathbf{R}_{t} \odot \mathbf{H}_{t-1})\,\mathbf{W}_{hh} + \mathbf{b}_{h}) \quad (5)$$

$$\mathbf{H}_{t} = \mathbf{Z}_{t} \odot \mathbf{H}_{t-1} + (1 - \mathbf{Z}_{t}) \odot \tilde{\mathbf{H}}_{t}. \quad (6)$$

where $\mathbf{H}_{t}$ now is the final update of GRUrather than reset gate of the output of RNN hidden layer unit.

4.2 Long Short-Term Memory

While GRU using two additional units, reset and update, long short-term memory (LSTM) network uses three different units to control data from current time ($\mathbf{H}_{t}$) and past time ($\mathbf{H}_{t-1}$): input, forget, and output gates. These three gates are defined as follows,

$$\mathbf{I}_{t} = \sigma(\mathbf{X}_{t}\mathbf{W}_{xi} + \mathbf{H}_{t-1}\mathbf{W}_{hi} + \mathbf{b}_{i}), \quad (7)$$
$$\mathbf{F}_{t} = \sigma(\mathbf{X}_{t}\mathbf{W}_{xf} + \mathbf{H}_{t-1}\mathbf{W}_{hf} + \mathbf{b}_{f}), \quad (8)$$
$$\mathbf{O}_{t} = \sigma(\mathbf{X}_{t}\mathbf{W}_{xo} + \mathbf{H}_{t-1}\mathbf{W}_{ho} + \mathbf{b}_{o}), \quad (9)$$

$W_{x} \in \mathbb{R}^{d\times h}$ and $W_{h} \in \mathbb{R}^{h\times h}$ is the weight parameters with bias b $\in \mathbb{R}^{1\times h}$

The complete sequence to update hidden state is defined as follow,

$$\tilde{\mathbf{C}}_{t} = \tanh(\mathbf{X}_{t}\mathbf{W}_{xc} + \mathbf{H}_{t-1}\mathbf{W}_{hc} + \mathbf{b}_{c}) \quad (10)$$

$$\mathbf{C}_{t} = \mathbf{F}_{t} \odot \mathbf{C}_{t-1} + \mathbf{I}_{t} \odot \tilde{\mathbf{C}}_{t}. \quad (11)$$

$$\mathbf{H}_{t} = \mathbf{O}_{t} \odot \tanh(\mathbf{C}_{t}). \quad (12)$$

$\tilde{\mathbf{C}}_{t}$ and $\mathbf{C}_{t}$ are candidate memory cell and memory cell, respectively. GRU and LSTM are very similar both in implementation and its result. GRU is faster due to less gates and LSTM, in many cases, slightly better than GRU due to its complexity to handle flow of data.

4.3 Model Architecture

The implementation of the regression classifier for dimensional emotion recognition can be built by stacking some RNN layers from acoustic and text input and merge both to obtain the final dimensional emotion prediction. For each modality, acoustic, and text, we varied two dense, two GRU, and two LSTM layers. For RNNs (GRU and LSTM) we also implement a bidirectional version of those networks to allow distribution of information from the past and future time (GRU and LSTM only roll information from the current and past time, see Eq. 3–6 and Eq. 10–12). Those dense, bidirectional GRU (BGRU), and bidirectional LSTM (BLSTM) layers from each modality is stacked together using two dense layers. Fig. 2 shows one of the architectures for combining acoustic and text features to obtain three emotional dimensions.

To minimize the risk of overfitting, a number of dropouts are used with value 0.4 for each acoustic and text network and 0.3 for the last dense network. Rectified linear unit (ReLU) activation is used for both dense layers in the combined network. The final dense layer with three nodes used linear activation function to obtain the score of valence, arousal, and dominance. The whole network is trained with RMSProp (Dauphin et al., 2015). optimizer with mean squared error (MSE) as a loss function. Beside MSE, we use mean absolute error (MAE) and mean absolute _percentage error (MAPE) as evaluation metrics. The implementation of this deep learning architecture is available in public repository, https://github.com/bagustris/dimensional_ser_rnn,for research reproducibility

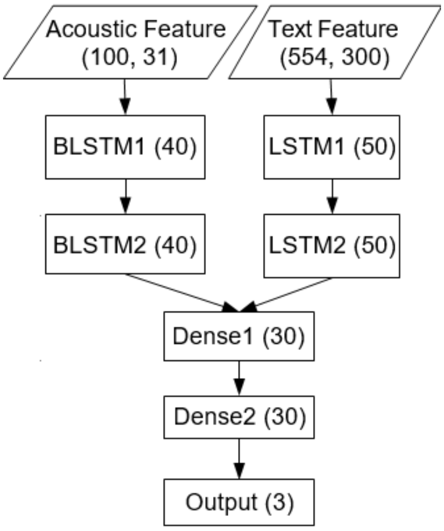

Fig. 2. Architecture of deep learning system combining acoustic and text features. The number in bracket shows the size of units/nodes on the layer.

5. RESULTS

5.1 Comparison of Acoustic, Text and Combined System

To begin our discussion, we presented the result for each different modality. Table 1 shows the performance of dimensional speech emotion recognition from the acoustic feature. Two layers of the same models are stacked and added with final a dense layer. For each model, the value of each metric is an average of five experiments (to minimize the effect of uncertainty computation due to randomness). The dense network is chosen as the baseline model. For this speech emotion recognition, the LSTM model shows modest improvement from the dense baseline layer in terms of MSE and MAE. However, other metrics i.e., MAPE, shows the different result which leads GRU/BGRU to obtain a better result. As MSE is used as a loss function, the result in MSE is relevant in this context. MAE metrics also show consistency with MSE. The MAPE metric can be used for comparison to other datasets as it has the same scale from 0 to 100.

Table 1. Performance comparison of dimensional speech emotion recognition from acoustic feature (in term of MSE, MAE, and MAPE) among different models.

Model[a]	MSE	MAE	MAPE (%)
Dense	0.652	0.66	24.188
GRU	0.648	0.655	23.69
LSTM	0.636	0.651	24.014
BGRU	0.647	0.653	23.675
BLSTM	0.656	0.66	24.109

[a]Each model is a stack of two same models.

For text-based emotion recognition, the result is shown in Table 2. As reported by other researchers (Poria et al., 2017; Tripathi et al., 2018). we obtained better performance on emotion recognition for IEMOCAP dataset by utilizing text features. In this text emotion recognition, the LSTM model shows modest improvement from the baseline and other models. This result from some experiments can be considered when combining acoustic and text features for the fusion of two networks. For this text emotion recognition, all metrics show almost consistent with each other (except GRU and BGRU, which is 0.02 different). MAE and MAPE show consistency in the order of the score among models.

Table 2. Performance comparison of dimensional speech emotion recognition from text feature (in term of MSE, MAE, and MAPE) among different models

Model[a]	MSE	MAE	MAPE (%)
Dense	0.493	0.559	20.436
GRU	0.48	0.549	19.888
LSTM	0.465	0.538	19.554
BGRU	0.482	0.548	19.881
BLSTM	0.487	0.55	19.588

[a]Each model is a stack of two same models.

Finally, we presented the result of the fusion of acoustic and text features in Table 3. Clearly, a decrement of error is shown for all MSE, MAE, and MAPE metrics. For example, using the same dense layers, the error (MAPE) decreases from acoustic (24.188%) and text (20.436%) to combined acoustic and text system (19.97%). To obtain the more decrement of error, not only the architecture of each network modalities is important but also the strategy for combining the modalities is also important (Poria et al., 2017). Although we tried several different layers after concatenation of two networks (acoustic and text), we focus on selecting the combination for each modality while keeping the use of dense layer after a combination of two networks. This focus is based on some experimentation we obtained; the simple dense layers after concatenation perform better than the more sophisticated layers (GRU, LSTM, and attention models).

Table 3. Performance comparison of dimensional speech emotion recognition from combination of acoustic and features (in term of MSE, MAE, and MAPE) among different models

Model	MSE	MAE	MAPE (%)
Dense + Dense	0.457	0.546	19.97
Dense + BGRU	0.44	0.533	19.585
BGRU + Dense	0.454	0.543	19.81
LSTM + LSTM	0.428	0.525	18.929
BLSTM + BLSTM	0.438	0.531	19.423
BLSTM + LSTM	0.419	0.517	18.713
BGRU + GRU	0.429	0.527	19.139

5.2 Design of System Architecture and Its Result

On designing the system architecture for dimensional speech emotion recognition, we rely on initial experiments using the unimodal feature, and other researcher results (Atmaja et al., 2019; Tripathi et al., 2018). From Tables 1 and 2, it is shown that the text feature gives better result on dimensional emotion recognition than acoustic feature. For both features, LSTM performs better than any other model. Using this result, we build LSTM-based networks for those two modalities and combine them with dense layers.

On choosing hyperparameters, we manually add more units to text networks as it gives a better result. The choice of using 50 units and 40 units of nodes on each LSTM layer on each modality is also obtained from experimentation; we use larger units first and decrease this number until the smaller one without decreasing the performance (error metrics). For the dense layers, the number of 30 units for each layer is also based on the experiment. The ReLU and tanh activation function in those layers perform a similar result. To avoid overfitting, we use the callback strategy besides putting the dropout layer on each network branches (acoustic, text, and combination layers). Two methods are used for callback (to stop the iteration of training): early stopping and model checkpointing. For early stopping, we use a number of 10 patiences to monitor validation loss. This means, if no decrement of validation loss (MSE) after ten epochs, the training process will stop and uses the best weight for the evaluation/prediction. The model check-pointing is a similar method to save the model (which can also be ignored if we do not want to save the model). Finally, although we obtain the best prediction of emotion dimension with BLTSM and LSTM networks, there is a room for improvement for experimenting and designing a better model architecture. In some runs, the combination of GRU performs better; however, the average result shows that a combination of LSTM is the best one. The hyperparameters optimization on future research will be done on training and development set instead of manually hand-crafted.

For the obtained improvement, a decrement of MAPE from acoustic feature-based emotion recognition is achieved up to 5.5 % when using a combined feature. For MSE and MAE, the decrement is in a range of 0.14-0.17 and 0.09- 0.11, respectively. From the text feature, the decrement of error is in range of 0.08-0.046, 0- 0.02, and 0-0.84% for MSE, MAE, and MAPE, respectively. The excerpt of the result of VAD score from the model obtained using BLSTM, and LSTM networks are presented in Table 4.

Table 4. Sample of true and predicted VAD score from model using BLSTM and LSTM Networks

Utterances	**VAD**	
	True	**Predicted**
Oh, totally. Yeah.	[4, 3, 2.5]	[3.21, 2.67, 2.60]
The craziest thing just happened to me.	[4, 3, 2.5]	[3.35, 3.02, 2.97]
This girl; she just offered me fifty thousand dollars to marry her.	[3.5, 3.5, 3]	[3.32, 3.3, 3.34]

5.3 Evaluation of Loss Function and Metrics

One of the challenging problems in dimensional emotion recognition is to choose the proper metrics for evaluation. In this paper, we used standard regression metrics i.e., MSE, MAE, and MAPE. However, when running some experiments on the same condition (system architecture), when a metric decrease, another increase the score. For example, in the second experiment, after the first, MSE gets lower, but MAPE gets higher, and so on.

Table 5 shows the raw result obtained in Table 1 for dense acoustic network. As shown in that table, the consistency of each metric is changing when re-running the experiment. In the second experiment, when MSE score decreases, MAPE score increases. In the last experiment, the MSE score increases, while MAE and MAPE decrease. To evaluate metrics, we perform a simple analysis by changing the lost function from MSE (default) to MAE and MAPE. Table 6 shows that by changing the loss function from MSE to MAE, the error result decreases slightly.

Table 5. Results of five experiments on the same models (dense layers) from speech features

Experiment#	MSE	MAE	MAPE
1	0.659	0.663	24.12
2	0.644	0.660	24.41
3	0.645	0.651	23.44
4	0.652	0.663	24.94
5	0.656	0.662	24.01

If we compare this result (our best MAE) with other research which also used acoustic and linguistic information, but with different approach (Karadoğan et al., 2012). our MAE is better than them (their best MAE is 1.28 for arousal). However, comparing the same metric across the dataset is not sufficiently comparable as the upper level bound of MAE is different for each dataset. In this case, MAPE might be more useful than MSE and MAE. Moreover, using another metric such as concordance coefficients correlation (ρ_c) as used in (Tzirakis et al., 2017) is more relevant as it has the same scale 0-1 for any datasets to measure the agreement.

Table 6. Results of BLSTM and LSTM networks from acoustic and text features with different loss function

Loss Function	MSE	MAE	MAPE
MSE	0.43	0.523	18.87
MAE	0.42	0.519	18.631
MAPE	0.469	0.543	18.835

6. CONCLUSION

We presented our work on dimensional speech emotion recognition by combining acoustic and text features using recurrent neural networks. Thirty-one acoustic features are used as input to acoustic networks, and 554- word vectors are fed to text networks. The result from unimodal shows that text-based emotion recognition performs better on IEMOCAP dataset compared to acoustic emotion recognition. The combination of acoustic and text features decreases the error of MAPE up to 5% from acoustic features only and near 1% from text feature only. For the combination among DNN layers, the use of BLSTM for acoustic network and LSTM for text network with con- catenated dense layers to combine those two features performs better compared to a list of given DNN layer combination. The choice of more advanced metric for loss function and evaluation in dimensional emotion recognition should be considered on the future research for consistency and benchmarking with other dimensional emotion recognition studies.

REFERENCES

Atmaja, B. T., Shirai, K., & Akagi, M. (2019, November). Speech emotion recognition using speech feature and word embedding. In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 519-523.

A. Zhang, Z. C. Lipton, M. Li, and A. J. Smol. (2019). Dive into Deep Learning, http://www.d2l.ai.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335.

Calvo, R. A., & Mac Kim, S. (2013). Emotions in text: dimensional and categorical models. Computational Intelligence, 29(3), 527-543.

Chen, M., He, X., Yang, J., & Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Processing Letters, 25(10), 1440-1444.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

Dauphin, Y., De Vries, H., & Bengio, Y. (2015). Equilibrated adaptive learning rates for non-convex optimization. In Advances in neural information processing systems 1504-1512.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3), 572-587.

Giannakopoulos, T., Pikrakis, A., & Theodoridis, S. (2009, April). A dimensional approach to emotion recognition of speech from movies. In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 65-68.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning, 1. Cambridge: MIT press.

Griol, D., Molina, J. M., & Callejas, Z. (2019). Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances. Neurocomputing, 326, 132-140.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Jurafsky, D., & Martin, J. H. (2014). Speech and language processing, 3.

Karadoğan, S. G., & Larsen, J. (2012, May). Combining semantic and acoustic features for valence and arousal recognition in speech. In 2012 3rd International Workshop on Cognitive Information Processing (CIP), 1-6.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, 3111-3119.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. arXiv preprint arXiv:1712.09405.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (1532-1543).

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L. P. (2017, July). Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers) (873-883).

Poria, S., Cambria, E., Hazarika, D., Mazumder, N., Zadeh, A., & Morency, L. P. (2017, November). Multi-level multiple attentions for contextual multimodal sentiment analysis. In 2017 IEEE International Conference on Data Mining (ICDM), 1033-1038.

Tripathi, S., Tripathi, S., & Beigi, H. (2018). Multi-modal emotion recognition on iemocap dataset using deep learning. arXiv preprint arXiv:1804.05788.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of Selected Topics in Signal Processing, 11(8), 1301-1309.