Speech Emotion Recognition Using Speech Feature and Word Embedding

Bagus Tris Atmaja *†, Kiyoaki Shirai †, and Masato Akagi †
* Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
Email: bagus@ep.its.ac.id
† Japan Advanced Institute of Science and Technology, Nomi, Japan
Email: {bagus, kshirai, akagi}@jaist.ac.jp

***Abstract*—Emotion recognition can be performed automatically from many modalities. This paper presents a categorical speech emotion recognition using speech feature and word embedding. Text features can be combined with speech features to improve emotion recognition accuracy, and both features can be obtained from speech. Here, we use speech segments, by removing silences in an utterance, where the acoustic feature is extracted for speech-based emotion recognition. Word embedding is used as an input feature for text emotion recognition and a combination of both features is proposed for performance improvement purpose. Two unidirectional LSTM layers are used for text and fully connected layers are applied for acoustic emotion recognition. Both networks then are merged by fully connected networks in early fusion way to produce one of four predicted emotion categories. The result shows the combination of speech and text achieve higher accuracy i.e. 75.49% compared to speech only with 58.29% or text only emotion recognition with 68.01%. This result also outperforms the previously proposed methods by others using the same dataset on the same modalities.**

I. Introduction

Emotion can be automatically recognized from many modalities: face, speech, and motion of the body's parts. In absence of visual feature (face and motion), speech is the only way to recognize emotion such as in telephone line, voice message and call center application [1]. By knowing caller emotion automatically from a system, feedback can be taken quickly and wisely. However, most speech emotion recognition (SER) systems show poor performance and most of them use acoustic features only. Here, an idea to use acoustic and text features are proposed to improve SER performance. This idea comes from that text can be extracted from speech and it contributes to emotion recognition. As an example, an interlocutor can perceive emotion not only from heard speech but also from the meaning of spoken words. Moreover, people tend to use specific words to express their emotion in spoken dialog because they have learned how some words are related to the corresponding emotions [6]. The current research on pattern recognition also shows that the use of multimodal features increases the performance compared to single modality [2]. The big data research also shows that the use of more data will improve the performance comparing to small data on the same algorithm [3]. Given the acoustic and text features, an improvement of SER should be obtained based on those motivations and many technologies can take benefits such as for more natural human-computer interaction by recognizing expressiveness in speech.

Combination of speech and text for speech emotion recognition is not new. In the previous years, there is an increase of paper reported that use speech and text to improve the performance of speech emotion recognition such as in [5], [4], [2]. Min Lee et al. proposed to use text and acoustic feature with logical "OR" function as the decision level from a fusion of acoustic and language information [6]. Qin Jin et al. proposes to merge acoustic and lexical features and train those features with SVM classifier to recognize its emotion category [7]. Recent papers on speech emotion recognition made use of deep learning as emotion classifier. Griol et al. use various machine learning classifiers to train three different databases to obtain user emotion category from spoken utterances [8].

This paper reports the use of a simple method to improve accuracy on SER by using a combination of speech and text features. To evaluate the proposed method using a combination of speech feature and word embedding, we perform speech emotion recognition using acoustic feature only and text emotion recognition using word embedding only from speech transcription. Speech feature, in this paper, is defined as a set of acoustic features which are extracted from the speech segments of an utterance after silence removal preprocessing. By combining speech and text features, we expect to outperform the highest performance result from either acoustic and text feature. Besides that, we want to outperform the previous research result on the same modalities and on the same Interactive Emotional Motion Capture (IEMOCAP) dataset [14]. Paper [2] used three modalities (speech, text, and mocap) and obtained 71% score as the highest performance from validation. When using two modalities (acoustic and text), they achieved 69.74% score of accuracy. Yenigalla et al. [5] used speech spectrogram and phoneme embedding and they reached 73.9% as the highest accuracy score. A deep learning approach by evaluating some systems have been proposed by Cho et al. [4] by combining audio and its transcript for IEMOCAP dataset. Their best unweighted accuracy is 65.9% for all class emotion category. Here, using speech segments of an utterance and word embedding, we intended to exceed their maximum accuracy.

Our contributions described in this paper are as follows:

- Feature extraction based on speech region of utterance

which is obtained by silence removal technique based on threshold and minimum duration of silence.
- Combination of speech feature above with word embedding from speech transcription to improve recognition rate of speech emotion recognition in categorical views.

The rest of this paper is organized as follows. In Section 2, we describe the dataset used in this research. Section 3 highlights the proposed method with each configuration for speech emotion recognition using speech feature, text emotion recognition using word embedding and a combination of those two. In Section 4, we present and discuss results obtained by previously explained approach. Finally, in Section 5 we conclude our works in this research and suggest some directions for future research.

II. Dataset

IEMOCAP dataset [14] is used to evaluate the proposed method. The IEMOCAP corpus consists of five sessions from both scripted and spontaneous act. From total 9 categories of emotion, only 3 emotion categories are used i.e. anger, excitement, neutral and sadness. The total number of used utterances is 4936 out of 10039 turns. The distribution of utterances for each class is almost identical to make the dataset balance. From many modalities, only speech and text are used on this research. Speech signals in the dataset are processed at 16 kHz of sampling rate with average length 4.5 s. For the text, average words per turn is 11.4 words while the longest utterance has 554 words. Speech and text are not aligned when it is processed (for feature extraction). The processing for both modalities is performed independently and simultaneously through its networks.

III. Proposed Method

A. Speech-based emotion recognition

Speech emotion recognition can be performed in two ways, by direct end-to-end processing, by taking raw speech to predict the emotions, or by step-by-step processing including pre-processing, feature extraction, classification, and, if needed, post-processing. We choose to follow the second approach as it gives more interpretation on which steps gives a better result and which step makes a worse result. Processing whole speech yields expensive computation and unnecessary information processing within all utterances that may result in poor performance. One solution to deal with that issue is by using the segmented speech part of the utterance by removing silence for feature extraction. Although this idea is not new, most research paper used whole speech utterances for feature extraction such as in [2]. The use of speech only segment for speech recognition is also criticized and not used by some researcher as they argued that silence is effective cues for emotion recognition [9], [10]. We extract acoustic features from the speech segment in this research based on previously explained advantages.

To start acoustic feature extraction from speech segment, speech files within the dataset first are read as a vector. For each utterance (each file), we perform silence removal to obtain speech segments. We perform silence removal based on two parameters: minimum threshold and minimum duration. The algorithm to remove silence part then can be summarized as follows:

1) Define `threshold` and `minimum duration`.
2) Scan along the samples, if amplitude of n-th sample below the threshold, count it one (n=1).
3) Perform step 2 and counted (n=n+1) until amplitude of n-th sample above threshold found.
4) Check the total number of accumulative samples below threshold, if the number of $n \geq$ `minimum duration`, remove those n samples.

The main findings on this SER using an acoustic feature from speech segment are the value of the threshold and minimum duration parameters. By some experiments, a threshold of 0.001% and minimum duration of 100 ms. Note, that there is no normalization on silence removal process, hence, the found threshold value is very small due to the wide dynamics of the speech signal. After getting the speech segment of each utterance, we perform feature extraction based on those speech segments. Each speech utterance is split into frames with a hamming window and moved it by overlap steps. The total of 34 features is extracted for each frame consist of 3 time domain features (zero crossing rate, energy, and the entropy of energy), 5 spectral domain features (spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off), 13 MFCCs, and 13 chromas. This feature extraction process is performed on each frame with length of 200 ms each and 50% overlap and concatenated for each utterance. For each utterance, we take 100 number of windows/segments resulted (100, 34) feature size as input for the acoustic classifier.

Four types of speech emotion recognition systems with/without speech feature are evaluated to see the effectiveness of proposed speech-based speech emotion recognition against whole speech input. Those systems are:

1) Whole speech using 2 stacks bidirectional LSTM networks.
2) Speech segments using 2 stack bidirectional LSTM networks.
3) Whole speech using 2 stack bidirectional LSTM networks and attention model.
4) Speech segments using 2 stack bidirectional LSTM networks and attention model.

The attention model used on the third and fourth architectures above is implemented from [11]. In that paper, the attention model is used to retrieve the only important information from the previous layer by attention weight to obtain better language translation. In another paper [15], the attention-based model also shows superiority among other approaches for speech recognition task, where the model is required to map utterances from spoken to the written domain. Borrowing the success of attention model on machine translation and speech recognition, we expect a similar improvement for this speech emotion recognition as the task is also similar.

For the first and second architectures, two bidirectional

TABLE I
EXAMPLE OF UTTERANCE AND ITS LABEL FROM IEMOCAP DATASET TRANSCRIPTION.

Utterance	Label
"Excuse me."	neutral
"That's out of control."	angry
"Did you get the mail? So you saw my letter?"	sad
"Did you get the letter?"	excitement

LSTM (BLSTM) with a number of 512 and 256 hidden neural units are stacked. Two fully connected layers then are added with 512 and 4 units. The last unit reflects the number of emotion category. Those first two architectures are the same except for the input. The number of trainable parameters on those architectures is 2,041,348 parameters for each system.

The third and fourth architectures are also the same except for the input. A bidirectional LSTM with 256 hidden neural units is used in the first layer. The second layer is attention decoder with a size of (128, 128) same as attention layer on text emotion recognition explained below. The same two fully connected (dense) layers are also added after this attention layer. Both dense layers use ReLU and softmax activation function. The total number of trainable parameters on those architectures are 14,788,612 for each system.

B. Word embedding-based emotion recognition

Even though at first it is difficult to express and recognize emotion in the text; however, by the advancement of textual analysis from natural language processing (NLP), it is possible to detect emotion inside a text. The example of manual transcription and its label is given in Table I. Based on reference, three evaluators annotated each utterance [14]. The label shown in Table I is the majority of annotation (two) among three annotators. For converting utterances into vectors, we use the basic word embedding as representations of words in each utterance. Both features and labels are fed into deep learning. The whole text emotion recognition can be explained as follows.

First, we acquire each utterance text from the manual transcription and save it in a variable. For each utterance, we did tokenization to obtain words within one utterance. The sample of utterance and its label is shown in Figure I. The tokenized words then are converted into a sequence and padded with a maximum length of 537 tokens instead of a number of fixed lengths. The implementation of this word embedding is based on [18]. This text feature is the input to the text classifier.

For the classifier, we tried three different architectures among other approaches as they show higher performance than others. The first is Convolutional Neural Networks (CNN) with four one-dimension (1D) convolution layers. After the embedding layer, the number of hidden units are 256, 128, 64 and 32. Each layer has a kernel size of 3 and dropout of 0.2 (20%) with the same activation function i.e. ReLU. For the last layer, a dense layer is added with 4 units correspond to the total number of the emotion category. The total number of trainable parameters for this network is 5,278,288.

The second network is long short-term memory (LSTM) with two layers, the first one contains 256 hidden neural units and the second one contains 256 hidden neural units. After the second layer, we add two fully connected layers with 512 units and 4 units with ReLU and softmax activation function. The total number of trainable parameters in this network are 3,410,288 parameters.

The last classifier is LSTM with attention decoder. For this network, we use an LSTM with 128 unit of hidden neural and replaced the second LSTM with attention decoder. This attention decoder receive input from the first LSTM layer as encoder and retrieve the weight from with attention weight. Two parameters for this attention function is the dimension of the hidden state (as it acts as a layer) and the length of attention matrices. We choose a value of 128 for both parameters as used in speech emotion recognition and added with two fully connected layers with 512 and 4 units as previous model. For this model, the total number of parameters is 34,057,840 parameters.

For all model, we train the input with the label at batch size of 64, number of epochs of 50 and optimized with RMSprop function. We split the data 80:20 for training and validation. The obtained result is shown in Table IV.

C. Combined speech feature and word embedding

The main proposal presented in this paper is the combination of speech-based acoustic feature and word embedding for categorical emotion recognition. The combination of feature proposed here is the concatenation of two models in early fusion, acoustic model and text model, where each model consist of different networks. Some compositions of layer combinations have been carried out to find the best model (variation of dense, LSTM and CNN). To accommodate two main different concepts in deep learning, CNN and LSTM, we use both in our models. We found the three models shown in Table II are the best obtained results among others.

TABLE II
DEEP LEARNING MODELS USED FOR COMBINATION OF SPEECH FEATURE AND WORD EMBEDDING FOR SPEECH EMOTION RECOGNITION

#	text model	speech model	concatenation model
1	CNN	Dense	Dense
2	LSTM	BLSTM	Dense
3	LSTM	Dense	Dense

For model 1, four 1D convolution layers are stacked followed by a dense layer with the number of units 256, 128, 64, 128, and 256 respectively. For speech networks, three dense (fully connected) layers are used with 1024, 512, and 256 units. Those two networks (acoustic and text) are concatenated and added with two dense layers with 256 and 4 units using ReLU and softmax activation functions. These last layers after concatenation are same for all combination models.

The second model uses two LSTM layers with 256 units each for text input. A dense layer with 256 unit is added to that network. For the speech network, a bidirectional LSTM

is encoded at first layer and attention decoder layer is added after that layer with the size of (128, 128). A dense layer with 512 unit is also added at the end of the speech network. The concatenation network adds two dense layers as the previous model.

The last model uses two unidirectional LSTM layers for text input the same as the second model. Here, for speech network,

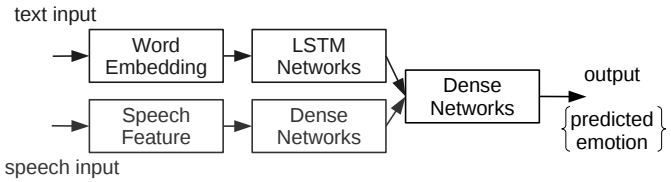

Fig. 1. Proposed speech - word embedding speech emotion recognition.

IV. Result and Discussions

A. Accuracy of proposed method

All models are evaluated in the same metric i.e. maximum accuracy. Table III shows performance result in that term for four speech models. From whole speech using LSTM to speech-based input using BLSTM with attention, there is accuracy improvement. The model using speech segment-based feature extraction and BLSTM with attention achieve the highest maximum accuracy among all speech models. However, to achieve that result, the number of trainable parameters is seven times from the first LSTM model. The computation time to achieve this result also equivalents to that comparison. This 58.29% accuracy in speech-based emotion recognition outperforms the result reported in [19], [2] which achieved 56.10% as the highest among them.

Table IV shows the accuracy of text emotion recognition from word embedding input. This result shows that the last model obtains the best result with 68% of accuracy as it uses greater trainable parameters. The result of CNN vs LSTM is slightly similar, but LSTM uses less trainable parameters than CNN. In this case, we conclude that LSTM works better compared to CNN for text emotion recognition using word embedding as an input feature. This result on text emotion recognition outperforms previous result on the same dataset by Tripathi et al. [2] which achieves 65.78% accuracy.

For the combination of acoustic and text feature, the result can be shown in Table V. Three models are evaluated as explained in the previous section. The best result is achieved by model 3 which consists of LSTM networks for text input and dense networks for speech input with dense networks for combination. This simple network reaches 75.48% of accuracy with only 5,213,060 of trainable parameters.

[1]http://github.com/bagustris/Apsipa2019_SpeechText

TABLE III
Accuracy result of emotion recognition using speech feature.

Model	Accuracy
Whole speech + BLSTM	52.83%
Speech segment + BLSTM	55.26%
Whole speech + BLSTM + Attention	53.64%
Speech segment + BLSTM + Attention	58.29%

TABLE IV
Accuracy result of text emotion recognition using word embedding.

Model	Accuracy
CNN	65.69%
LSTM	66.59%
LSTM + Attention	68.01%

B. Discussions

Combining acoustic and text feature can be seen as a way to maximize the performance of emotion recognition from speech as both features are related and can be derived from speech. How to integrate those features can be explored like in [4] by using LSTM for an acoustic feature and multi-resolution CNN for text feature. Instead of word embedding, a phoneme embedding can be used, as proposed in [5]. However, the main goal of this task is to achieve the highest accuracy to recognize emotion category. Using simple method, LSTM for text and fully connected networks for speech, we achieve higher accuracy than results obtained on the previous research.

Table VI shows a comparison of our best result with others on the same modalities and the same dataset (IEMOCAP). Our combination of acoustic and text feature close to the approach by Tripathi et al. [2]. While they used three modalities: speech, text, and mocap to obtain the best one, we only use speech and text features to exceed their result. As in [2], we also used GloVe embedding [13] for weighting word embedding which improves the accuracy about 1-2%. Using bimodal text and audio features, the authors of [2] reported 69.74% accuracy.

The authors of paper [5] obtained 73.9% as their best accuracy by using only spontaneous session data. Based on reference [19], the spontaneous data gives higher accuracy than all data on IEMOCAP dataset (Figure 2 on that reference). Therefore, we believe that our result will higher if we processed on spontaneous data only. The third comparison by Yoon [12] used Multimodal Dual Recurrent Encoder with Attention (MDREA) by utilizing Recurrent Neural Networks (RNN) for both text and audio network. While they used automatic speech recognition (ASR) to obtain text, we used manual transcription for obtained word embedding which may

TABLE V
Accuracy result of emotion recognition using combined speech feature and word embedding.

Model	Accuracy
Model #1	68.83%
Model #2	69.13%
Model #3	75.49%

TABLE VI
ACCURACY COMPARISON WITH OTHER PAPERS (SAME TEXT MODEL + SPEECH MODEL).

References	Method	Accuracy
Tripathi [2]	LSTM (GloVe Embedding) + Dense	69.74%
Yenigalla [5]	CNN (Spectrogram) + CNN (Phoneme Embedding)	73.9%
Yoon [12]	RNN + RNN	71.8%
Ours	LSTM + Dense	75.49%

leads to higher accuracy. The use of ASR is left for future work for comparison.

Another finding while experimenting on this research is that the best model for each modal (speech or text), is not the best when combined. For speech, we achieve the best model with BLSTM and attention model, and for text, the best one is also LSTM with attention model. However, when those two models are combined, the result shows a low accuracy, i.e. about 55% of accuracy (not shown on the Table as lower than other models). We still haven't understood yet why this combination of the best models resulting in poor performance. On finding the best model, we rely on some trials over some combinations of hyper-parameters values.

The intuition to use speech segment only of an utterance for feature extraction is that this part will give better emotion recognition compared to whole speech including noises and silence parts. As predicted, for speech emotion recognition using the acoustic feature only, the speech segments give higher accuracy than whole speech. This speech emotion from speech segment-based acoustic feature extraction also outperforms the previous results on using speech feature only such as results presented in [16], [2], and [17]. The human perception on speech emotion not only relies on the vocal tone that corresponds to acoustic feature but also verbal context on the meaning of what is spoken. This intuition leads us to use word embedding for an additional feature. By using deep learning, we can easily experiment on some architectures and changing some hyper-parameters values.

A simple model to improve the SER accuracy is obtained by integrating LSTM with fully connected (dense) layers. After running some experiments, an accuracy of 75.49% is achieved. However, due to random initialization of GPU and CPU used for computation, it needs to be performed several times to obtain the similar result although a fixed random number is initialized at the top of the computer program for deep learning computation. A strategy to obtain consistent high accuracy is needed for future research to enable benchmarking with other speech emotion recognition studies.

V. CONCLUSIONS

This paper reported speech emotion recognition using speech feature and word embedding. A set of acoustic features were extracted from the speech part of the utterances, and a set of sequences were generated from word embedding. For classification, LSTM was used for text input while the fully connected network was used for speech input; both are concatenated with fully connected layers. The achieved accuracy by the proposed method is 75.49% which outperforms previous research on the same dataset and similar modalities. For future research, the proposed method could be used for other datasets to check the consistency of the obtained result.

REFERENCES

[1] Petrushin, Valery. "Emotion in speech: Recognition and application to call centers." In Proceedings of artificial neural networks in engineering, vol. 710, p. 22. 1999.

[2] Tripathi, Samarth, and Homayoon Beigi. "Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning." arXiv preprint arXiv:1804.05788 (2018).

[3] Halevy, Alon, Peter Norvig, and Fernando Pereira. "The unreasonable effectiveness of data." (2009).

[4] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, Deep neural networks for emotion recognition combining audio and transcripts, Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2018-September, pp. 247–251, 2018.

[5] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, Speech Emotion Recognition Using Spectrogram & Phoneme Embedding, September, pp. 3688–3692, 2018.

[6] C. M. Lee, S. S. Narayanan, L. Angeles, and R. Pieraccini, Combining Acoustic and Language Information for Emotion Recognition, Icslp 2002, vol. 2002, pp. 6–9, 2002.

[7] Q. Jin, C. Li, S. Chen, and H. Wu, Speech emotion recognition with acoustic and lexical features, ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2015–August, pp. 4749–4753, 2015.

[8] D. Griol, J. M. Molina, and Z. Callejas, Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances, Neurocomputing, pp. 1–9, 2017.

[9] H. M. Fayek, M. Lech, and L. Cavedon, Evaluating deep learning architectures for Speech Emotion Recognition, Neural Networks, vol. 92, pp. 6068, 2017.

[10] Tian, Leimin, Catherine Lai, and Johanna Moore. "Recognizing emotions in dialogues with disfluencies and non-verbal vocalisations." In Proceedings of the 4th Interdisciplinary Workshop on Laughter and Other Non-verbal Vocalisations in Speech, vol. 14, p. 15. 2015.

[11] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

[12] Yoon, Seunghyun, Seokhyun Byun, and Kyomin Jung. "Multimodal Speech Emotion Recognition Using Audio and Text." arXiv preprint arXiv:1810.04635 (2018).

[13] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

[14] Busso, Carlos, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. "IEMOCAP: Interactive emotional dyadic motion capture database." Language resources and evaluation 42, no. 4 (2008): 335.

[15] Prabhavalkar, Rohit, Kanishka Rao, Tara N. Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly. "A Comparison of Sequence-to-Sequence Models for Speech Recognition." In Interspeech, pp. 939-943. 2017.

[16] Zhao, Yue, Xingyu Jin, and Xiaolin Hu. "Recurrent convolutional neural network for speech processing." In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5300-5304. IEEE, 2017.

[17] Chernykh, Vladimir, Grigoriy Sterling, and Pavel Prihodko. "Emotion recognition from speech with recurrent neural networks." arXiv preprint arXiv:1701.08071 (2017).

[18] Gal, Yarin, and Zoubin Ghahramani. "A theoretically grounded application of dropout in recurrent neural networks." In Advances in neural information processing systems, pp. 1019-1027. 2016.

[19] Neumann, Michael, and Ngoc Thang Vu. "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech." arXiv preprint arXiv:1706.00612 (2017).