Review

Depression detection model using multimodal deep learning

Hyewon Yoo [1], Hayoung Oh [2,*]

[1] Statistic and Computer Science; Sungkyunkwan University; Seoul; 03603; Korea; mariahwy@g.skku.edu
[2] College of Computing and Informatics; Sungkyunkwan University; Seoul; 03603; hyoh79@skku.edu
* Correspondence: hyoh79@skku.edu; Tel.:

Abstract: This study compares the performance of existing studies on multimodal emotion recognition, and proposes a model that fuses two modalities with the speaker's text and voice signals as input values and detects depression. Based on the DAIC-WOZ dataset, voice features were extracted using CNN, text features were extracted using Transformers, and two modalities were fused through a tensor fusion network. We also build a model to detect whether the speaker is depressed or not using LSTM in the final layer. This study suggests the possibility of increasing access to mental illness diagnosis by enabling patients to detect depression on their own in daily conversations. If the model proposed in this study is developed and the voice conversation system is connected, it will be easier for patients who cannot visit the hospital periodically or who are reluctant to visit the hospital to check their condition and seek recovery. Furthermore, it can be expanded to multi-label classification for various mental diseases and used as a simple self-mental disease diagnosis tool.

Keywords: depression detection; fusion; feature extraction; deep learning

1. Introduction

According to the 2021 Covid-19 National Mental Health Survey released by the Ministry of Health and Welfare, the proportion of people who thought of suicide in 2021 increased by 40% compared to March 2020, the beginning of the Covid-19 outbreak, and one in five showed a risk of depression [1]. The disconnection from the world negatively affected public sentiment, resulting in an increase in the overall number of depression patients. As such, depression is no longer a rare disease for modern people, and it is already a social problem in itself because severe depression disorders can lead to suicide. Only a few people have access to high-quality mental health services around the world, and the majority do not recognize that they are suffering from mental illness [2]. Furthermore, there are people who are reluctant to visit hospitals and be diagnosed with negative perceptions of mental illness [3]. For this reason, research related to artificial intelligence technology that can help with the initial diagnosis of depression and the treatment of depression patients has been actively conducted. Among them, studies are mainly conducted to detect depression using deep learning or to predict the serious level of depression. Biological signals used to diagnose depression include the speaker's facial expression, voice, and conversation context revealed in the image. In this study, a deep learning model was built by paying attention to the speaker's text and voice signals, and a study was conducted to compare the method of fusion of text and voice signals.

2. Previous works

Text data has traditionally been used in the field of emotional recognition. The language model is learned by predicting the probability of the next word from the given words. Natural language processing reflects order and interrelationship information between words. Therefore, an end-to-end learning method (seq2seq) that can take context into account is used [4]. The RNN-based language model has the advantage of being able to obtain output values considering the previous context. However, RNNs are difficult to imply information about long sequences due to their fixed context vector size. That is,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2023 doi:10.20944/preprints202305.0663.v1

when the input sequence is too long, there is a disadvantage that the information of the initial token is diluted. An attention model has been proposed to solve the chronic problem of these RNN [5]. The attention model initially maintained the structure of the RNN and fine-tuned parameters through a backpropagation algorithm, but the RNN structure was time-consuming to eliminate the RNN structure and parallelize the self-attention layer to best express input values. Transformer models that frequently appear in recent natural language processing studies are end-to-end learning models created by stacking multiple encoders and decoders using such multi-head attention. In our study we use BERT, a pre-trained Transformer encoder-based language model, to perform well in emotion classification [6]. In the study, we identified the problem that the BERT model was less learned and built a new EmoBERTa specialized for emotion classification using the RoBERTa model that applied an improved learning method [7-8].

In recent studies in the field of emotion recognition, not only text data but also speech data are frequently used. Since voice data is an analog signal, it is necessary to convert it into a digital signal and input it to a computer. In this process, a method of extracting samples in units of one second is used to express sound waves in numbers. The feature must then be extracted to determine the characteristics of the speech. The voice has the form of a wave, and the wave consists of a time domain and a frequency domain. Therefore, it is necessary to reduce the dimension through fast Fourier transforms that convert the time domain of the wave into the frequency domain. Features to be extracted are various, such as spectrum, mel spectograms (MS), and MFCC. In order to extract features, audio signals are first divided by frame to obtain spectra by applying high-speed Fourier transform [9]. Mel spectrum (MS) is the application of mel filter bank to this spectrum [10]. Furthermore, MFCC can be obtained by applying cepstral analysis to the mel spectrum [11]. In the study, we compared which feature is best to use when analyzing voice data [12]. Mel Spectogram (MS) and MFCC were found to have the best performance. Recent deep learning models using voice data convert feature-specific values or heat maps into images and use them. It is to learn a CNN model using a three-dimensional RGB image as an input value. The CNN model is a deep learning model frequently used in the field of computer vision, which is a useful neural network for finding patterns in images or images. Typical CNN models include ResNet, which skips layers and transfers residuals as an invariant function to reduce bottlenecks [13]. The study compared the audio classification performance by CNN model and revealed that voice feature learning using ResNet-50 is the most effective [14].

Emotional classification fields such as depression detection are evaluated in a natural way of thinking like humans to use various modalities together rather than simply using one modality. Therefore, multimodal learning methods that utilize different types of data at the same time are frequently used in the field of emotional recognition. The multimodal fusion scheme has a late fusion scheme that creates an independent model for each modality and simply ensembles each output value in the last layer [15]. This method has the advantage of being able to adopt and use different feature learning structures to best suit each modality. However, there is a disadvantage that each modality cannot consider the external correlation (inter-modality dynamics) that occurs in the process of interacting with other modalities. Next, there is an early fusion method that combines and sorts features for each modality in advance and uses them as input values for the classification layer [16]. This method is not efficient because it does not reflect the intra-modality dynamics by modality. However, research emphasizes that a lot of information is lost because existing prior studies do not reflect both these external and internal correlations, and proposes a tensor fusion network (TFN) that can reflect both external and internal correlations [17]. Tensor fusion networks are fusion networks that can reflect all correlations between up to three modalities by performing a three-layer Cartesian product operation in embedding each modality.

3. Datasets

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2023 doi:10.20944/preprints202305.0663.v1

In this study, we used DAIC-WOZ dataset [18]. The DAIC-WOZ dataset is part of the DAIC (Distress Analysis Interview Corp.) and is a clinical interview designed to help diagnose psychological conditions such as anxiety, depression, and post-traumatic stress disorder. It contains interviews between virtual interviewer 'Ellie' and participants. At this time, the participants are labeled as people with and without depression. This database is widely used in multimodal psychiatric research because it provides all facial features, voice, and text information for participants. A Patient Health Questionnaire (PHQ-8) presurvey was used to label participants for depression on the DAIC-WOZ dataset. The PHQ-8 survey is a measure of depression consisting of eight questions, an established survey item based on valid diagnosis and severity of depression in large clinical studies. There are 189 participants in total, and each participant provides one voice file with an average length of 7 to 33 minutes and an average length of 16 minutes. All voice files were recorded at 16 kHz. Participants filled out a PHQ-8 preliminary questionnaire before the interview, and labeled it after judging that if the score of the questionnaire was 10 or higher, there was no depression if it was less than 10 points. After removing missing and outliers from the DAIC-WOZ dataset, counting the numbers showed that it consisted of 107 learning data, 47 testing data, and 35 verification data. In addition, 133 participants showed depressive symptoms and 75 participants judged that they did not have depressive symptoms, indicating that more participants had depressive symptoms. Table 1 presents the number of training, validation and test data in DAIC-WOZ dataset.

Table 1. DAIC-WOZ dataset..

Label	Train	Dev	Test
Depression	77	23	33
Non-depression	30	12	33

By the authors.

4. Model

The model structure proposed in this study is largely composed of four elements. First, in the voice feature extraction layer, a feature is extracted by converting a voice into a digital signal. Next, in the text feature extraction layer, features are extracted from natural language. The modality fusion layer fuses voice and text features. Finally, depression is classified in the depression diagnosis layer. Figure 1 schematically shows the structure of the entire model.

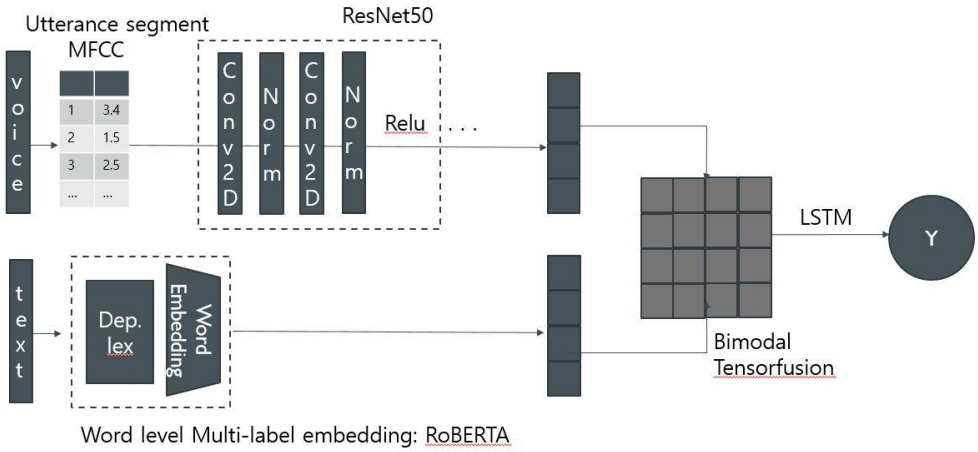

Figure 1. Overall model structure.

4.1. Audio Feature Extraction

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2023 doi:10.20944/preprints202305.0663.v1

In the speech feature extraction layer, the librosa library was used to convert speaker-specific speech into digital signals and extract features. Librosa is a representative library that handles voice data, allowing users to load files in the form of wav and handle them directly, and provides voice conversion functions. First, 16,000 samples were extracted per second from the original voice file provided by the DAIC-WOZ dataset. In addition, Ellie's voice was deleted from the original voice file after additionally recalling the script file, considering that the voice required to detect depression was voice related to the participant's answer. Next, in order to make the participant's answer the input value, the voice interval was separated based on the start time and completion time of the answer. In this study, MFCC features that show good performance for speech classification were extracted and used. The librosa library basically extracts 20 MFCCs per second, but 100 MFCCs per second were extracted to better capture the complex features contained in human speech. Since the lengths of all voice intervals are different, padding was applied to match the size of the input value. Since the average length of the entire voice section was about 2 seconds, the padding was applied as 200. Figure 2 is visualized after applying padding to the MFCC extracted from the first conversation of participant 300. The voices of all participants through the voice feature extraction layer proposed in this study are separated into sections corresponding to the script, and a preprocessing process is performed with input values in the form of (100, 200) as shown in Figure 3.

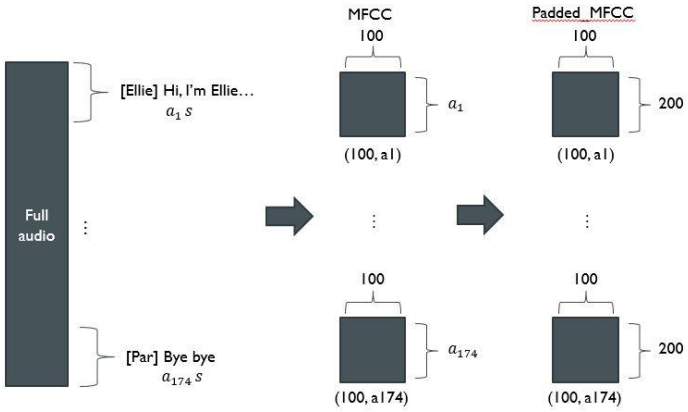

Figure 2. Example of participant 300 audio preprocessing.

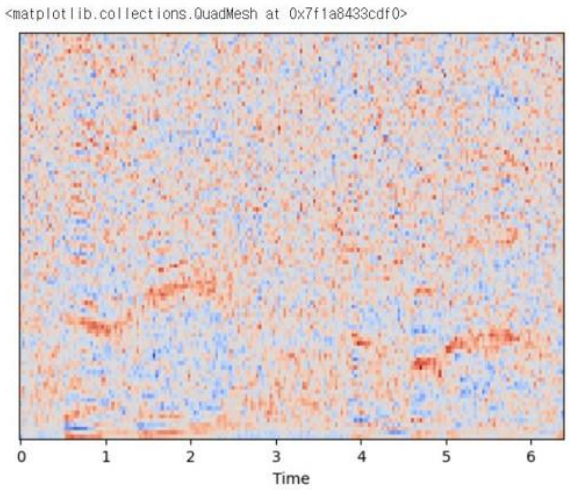

Figure 3. Example of participant 300 MFCC.

4.2. Text Feature Extraction

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2023 doi:10.20944/preprints202305.0663.v1

In the text feature extraction layer, a script for each segment is tokenized and then a feature is extracted through a language model. Since the DAIC-WOZ dataset is available in English, tokenization was carried out with Word Tokenizer of the NLTK package provided by Python. As a result of examining the scripts of the DAIC-WOZ dataset, there were many missing values, and there were also many values in parentheses that indicated actions or responses other than participants' answers, such as "<laughter>, so missing values, exclamations, special characters, and response descriptions were eliminated in advance. Next, by referring to NLTK's list of non-terms, non-terms such as I and you were removed, and words that were difficult to contain less than three characters were deleted. The root identification process was performed using the Porter Stemmer of the NLTK package. Root identification is the process of unifying words with the same meaning but slightly different forms. This is done by cutting the end of the word through a set rule. Therefore, it is highly likely that the word derived as a result of root identification is a word that does not exist in the dictionary. Next, the text feature was entered into the language model to extract it. First, token embedding, segment embedding, and positional embedding were applied to add location information of each word to the BERT model that does not receive word input sequentially. BERT is a pre-trained language model based on Transformers released by Google. After completing the feature embedding in this way, the feature was extracted using the BERT model. After that, it was flatten and expressed as a feature. Figure 4 is a diagram illustrating how text features are represented as one vector.

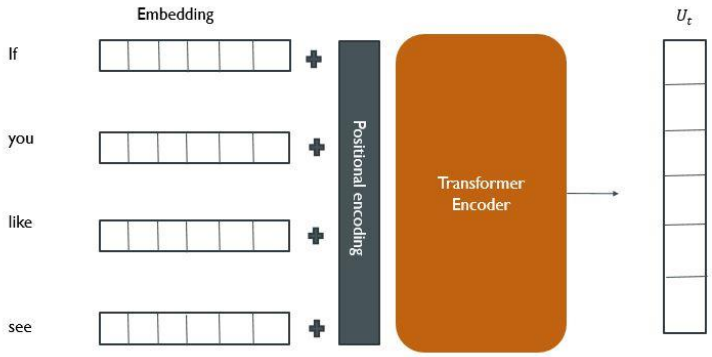

Figure 4. Extraction text features.

4.3. Modality Fusion

In studies fusing two or more modalities, simple merging of features in three ways is often used [19]. The add method is a method of simply adding two vectors. The concat method connects two vectors, and unlike the other two methods, the length of the vector increases. Next, the multiply method is a method of multiplying two vectors. In addition, in recent studies, fusion through cross attention is also underway. In this study, a fusion method using a tensor fusion network was used. Tensor fusion networks connect features extracted from each modality to a fully connected layer and represent them as one feature vector. In this process, voice and text features were displayed in a two-dimensional space by performing a two-layer Cartesian product set operation. Since the two-layer Cartesian product set calculates the product between all possible sets, it can capture correlations without loss of information, and because it is a simple extrinsic operation, learning parameters do not increase, so increasing the dimension of the feature does not increase the likelihood of overfitting. Tensor fusion networks are used to reduce the amount of computation, and both internal and external correlations between modalities can be considered.

4.4. Depression Detection

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2023 doi:10.20944/preprints202305.0663.v1

In the depression diagnostic layer, classification is performed by linking the final multimodal feature vector merged into one feature in the modality fusion layer to the LSTM model [20]. The LSTM model is a type of RNN model, which has been introduced to compensate for the shortcomings of RNN that do not reflect long-term memory when performing tasks that require a lot of context. The reason why LSTM was used in this study is to add time series characteristics because it is reasonable to classify depression considering the entire context of the interview. As a function of the output layer, sigmoid was used because the proposed model performs binary classification into two groups: 0 which is a group that does not show depressive symptoms and 1 which is a group that shows depressive symptoms.

5. Experiments

In this study, two experiments were conducted. First, the performance of the modality fusion method was compared. Next, the performance of the classification model used in the classification layer was compared.

First, the method of fusing modalities compared the performance of the tensor fusion network (TFN) used in the model proposed in this study, which simply adds two modalities, concat increases the length by connecting two modalities, multiply multiplying the two modalities. Table 2 is a table comparing classification performance according to the fusion method.

Table 2. Classification accuracy by fusion method.

Fusion method	Accuracy
Add	0.6275
Concat	0.6837
Multiply	0.6878
TFN	0.8012

By the authors.

Among the early fusion methods, add, concat, and multiply, the multiply method showed the best performance. However, it can be seen that the fusion method using the tensor fusion network showed superior performance to the late fusion method.

The logistic regression model used for binary classification at the classification layer, the support vector machine (SV), a classification model that determines which of the two classes a new data point belongs to, and the XGBoost (XGB) and several learning machines using boosting techniques to minimize errors were compared. Next, the performance of the classification model using LSTM, the method proposed in this study, was compared. Table 3 is a table comparing classification performance according to the classification model.

Table 3. Classification accuracy by classification model.

Model	Accuracy
SVM	0.5873
RF	0.5692
LR	0.6931
XGB	0.6024
LSTM	0.8012

By the authors.

 doi:10.20944/preprints202305.0663.v1

Comparing other machine learning models, it was found that the logistic regression model showed the best performance. However, as proposed in this study, it was found that the classification method using the LSTM model showed superior performance than the classification method using the logistic regression model.

6. Discussions

In this study, research trends in the field of emotional classification using multimodal deep learning were examined, and a depression detection model using a multimodal deep learning model was proposed. When fusing two features in the field of emotion classification, it is more effective to use tensor fusion networks than to use early fusion or late fusion methods. In addition, if the dataset used in the classification layer is an ordered multi-turn conversation, it is found that using a model that can reflect time series characteristics is more effective in terms of performance. Furthermore, in this study, a depression detection model was proposed using emotion classification technology. The model proposed in this study can be used in the process of preventing, diagnosing, treating, and post-management of mental disorders. It has the advantage of being able to easily collect users' voices and texts using the Internet of Things (IoT) device, which is rapidly developing recently. In addition, by applying the depression detection model proposed in this study, it can be simply self-diagnosed for depression and can be used as a tool to diagnose the condition of patients who are difficult to visit the hospital frequently. In future studies, after extracting patients' voice and text modalities using the Internet of Things, the model proposed in this study will be used to diagnose depression and design a framework that connects them with the voice conversation system. Not only can this framework self-diagnose depression, but it can also be expected to have therapeutic effects in the process of voice conversation between patients and voice conversation systems. In addition, further research will be conducted using models of good performance that have achieved SOTA among multimodal deep learning models in the future. However, in the process of using multimodal deep learning to detect depression, there is a side effect that false negatives and false positives can provide wrong health information to patients. Therefore, this model should be used as an aid for prevention or diagnosis rather than independently using it to detect depression, and accurate diagnosis of depression should follow the opinion of a professional doctor. It is also necessary to review the adverse effects of multimodal deep learning depression diagnostic dialogue systems from the early stages of framework development, and to clearly establish and comply with ethical guides or implementation guidelines.

Author Contributions: Writing—original draft, conceptualization, methodology, software, investigation, formal analysis, H.Y.; and validation, supervision, project administration, H.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jungshin, L. Ministry of Health and Welfare in Korea. 2021, pp. 1-3. Available online: http://www.mohw.go.kr/upload/viewer/skin/doc.html?fn=1641812737137_20220110200537.hwp&rs=/upload/viewer/result/202304/. (accessed on 26 April 2023).
2. Colleen, L. National Institute of Mental Health. 2021. Available online: https://www.nimh.nih.gov/health/statistics/mental-illness/. (accessed on 26 April 2023).
3. Puspitasari, I. M.; Garnisa, I. T.; Sinuraya, R. K.; Witriani, W. Perceptions, knowledge, and attitude toward mental health disorders and their treatment among students in an Indonesian University. *Psychology Research and Behavior Management*, 2020, pp. 845-854.

4. Sutskever, I.; Vinyals, O.; Le, Q. V. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 2014, pp. 27.
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Polosukhin, I. Attention is all you need. Advances in neural information processing systems, 2017, 30.
6. Devlin, J.; Chang, M. W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
7. Taewoon, K.; Piek, V. Emoberta: Speaker-aware emotion recognition in conversation with Roberta. *arXiv preprint arXiv:2108.12009*, 2017.
8. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
9. Weisstein, E. W. Fast fourier transform. 2015. Available online: https://mathworld. wolfram. com/. (accessed on 26 April 2023).
10. Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Wu, Y. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779-4783.
11. Mansour; Abdelmajid. H.; Gafar. Z. A. S.; Khalid A. M. Voice recognition using dynamic time warping and mel-frequency cepstral coefficients algorithms. *International Journal of Computer Applications*, 116.2, 2015.
12. Faustino, P.; Oliveira, J.; Coimbra, M. Crackle and wheeze detection in lung sound signals using convolutional neural networks. *In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 345-348.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2016, pp. 770-778.
14. Hershey, S.; Chaudhuri, S; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Wilson, K. CNN architectures for large-scale audio classification. *In 2017 ieee international conference on acoustics, speech and signal processing (icassp)*, IEEE, 2017, pp. 131-135.
15. Trong, V. H.; Gwang-hyun, Y.; Vu, D. T.; Jin-young, K. Late fusion of multimodal deep neural networks for weeds classification. *Computers and Electronics in Agriculture*, 2020, 175, 105506.
16. Gadzicki, K.; Khamsehashari, R.; Zetzsche, C. Early vs late fusion in multimodal convolutional neural networks. *In 2020 IEEE 23rd international conference on information fusion (FUSION)*. IEEE, 2020, pp. 1-6.
17. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L. P. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.
18. Jonathan, G.; Ron, A.; Gale, L.; Giota, S.; Stefan, S.; Angela, N.; Rachel, W.; Jill, B.; David, D.; Stacy, M.; David, T.; Skip, R.; Louis, P. M. The Distress Analysis Interview Corpus of Human and Computer Interviews. *In Proceedings of Language Resources and Evaluation Conference (LREC)*, 2014.
19. Yoon, J.; Kang, C.; Kim, S.; Han, J. D-vlog: Multimodal Vlog Dataset for Depression Detection. *In Proceedings of the AAAI Conference on Artificial Intelligence*. 2022, 36, 11, pp. 12226-12234.
20. Ma, J.; Tang, H.; Zheng, W. L.; Lu, B. L. Emotion recognition using multimodal residual LSTM network. *In Proceedings of the 27th ACM international conference on multimedia*. 2019, pp. 176-183.