Attentively-Coupled Long Short-Term Memory for Audio-Visual Emotion Recognition

Jia-Hao Hsu and Chung-Hsien Wu
Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, TAIWAN
E-mail: {sky84003, chunghsienwu}@gmail.com

***Abstract*— There have been more and more studies on emotion recognition through multiple modalities. In the existing audio-visual emotion recognition methods, few studies focused on modeling emotional fluctuations in the signals. Besides, how to fuse multimodal signals, such as audio-visual signals, is still a challenging issue. In this paper, segments of audio-visual signals are extracted and considered as the recognition unit to characterize the emotional fluctuation. An Attentively-Coupled long-short term memory (ACLSTM) is proposed to combine the audio-based and visual-based LSTMs to improve the emotion recognition performance. In the Attentively-Coupled LSTM, the Coupled LSTM is used as the fusion model, and the neural tensor network (NTN) is employed for attention estimation to obtain the segment-based emotion consistency between audio and visual segments. Compared with previous approaches, the experimental results showed that the proposed method achieved the best results of 70.1% in multi-modal emotion recognition on the dataset BAUM-1.**

I. INTRODUCTION

Previous research has been conducted for developing and evaluating the methods for automated emotion recognition. In the interaction between people, facial images and voice signals contribute most of the emotional expression in daily communication [1], so this paper selected facial images and voice signals as the modalities of emotion recognition.

In the study of emotion recognition, in addition to feature extraction and classification, there is one more work that has been paid more attention and discussion. It is the unit of emotion recognition. In the past, most studies dealt with the whole signal or file, which can be called object-based or file-based recognition unit. However, more and more studies believed that the entire signal may contain different emotions, showing emotional fluctuations in the signal. It means that a signal may contain more than one emotion. So the system using segment-based recognition units [2] was proposed to more accurately classify the emotions of different signal segments, or to improve the accuracy of the entire signal identification with segmented emotions. Considering the above factor, this paper takes segments as the recognition unit using the segmentation method proposed in [3]. The speech processing tool [4] and prosodic feature algorithms [5] are adopted for signal segmentation assuming that there is only one emotion in each segment.

In previous studies, there were many studies on speech emotion recognition using convolutional neural network (CNN) for feature extraction. [6, 7] used the long short-term memory (LSTM) [8] to classify the emotional features and effectively process the context of signals to enhance the speech emotion recognition performance. In a conversation, there may be more than one emotion, so they detect the sentence breakpoints to distinguish emotion change [9]. More studies on speech emotion recognition focused on non-verbal segments or whispered sound [10]. For audio part, this paper uses the audio clip as the unit for audio emotion recognition and uses deep neural network (DNN) for feature extraction and classification considering the non-verbal features of audio signals. For facial emotion recognition, as the VGG (visual geometry group model) [11] performed excellently well and it has been applied to many facial expression recognition systems [12], this paper uses VGG to extract the facial expression features for emotion recognition.

It can be seen from past studies that the multi-modal fusion methods are mainly divided into three types [13, 14]: feature-level fusion, model-level fusion, and decision-level fusion. Feature-level fusion is the most common method. [15] and [16] took the entire segment of the signal directly to extract different modal features by deep belief network (DBN) and Autoencoder (AE), and then concatenated the features for classification. In recent years, there have also been many studies using feature-level fusion. [17] used outer products to fuse multiple features individually. [18] used gated recurrent units (GRUs) to fuse multiple features and applied attention mechanism to focus on key sentences. Most of them performed well, but the influence of temporal context is usually not considered in the model. However, it is unlikely to highlight the importance of each modality in feature-level fusion. For decision-level fusion, [19] processed and classified the signals from each modality independently. The weighted sum of the outputs from all modalities was then used as the final output. Yet, the relationship and interaction between different modalities [13] were not considered. This is also the biggest problem in decision-level fusion method, especially when there are modal dependencies or temporal dependencies in the signal. In contrast, the flexibility of model-level fusion can alleviate these problems. [20-22] used multiple HMMs to characterize the features from different modalities and established the relationship between HMMs to obtain the final result. Nevertheless, HMM is not suitable for long-term dependency emotion recognition. [23] considered the temporal order of the modalities of the human brain when sensing emotions, and used a hierarchical fusion model to fuse the features. But they also did not consider the interaction between the multimodal signals in the specific segment. Most of all, the influence of

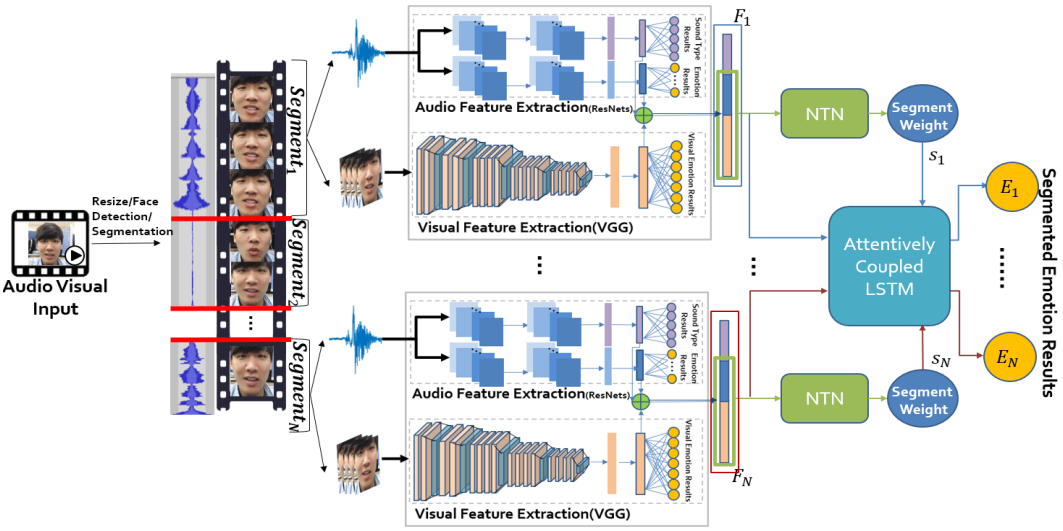

Fig. 1 System Architecture of the Proposed Approach

interaction between different modal features and the influence of context are the key to the fusion of multi-modal features in this paper. We try to improve the existing model-level fusion method, and considers the attention mechanisms to emulate the function of human brain for emotion recognition.

We have sorted out the following problems in the existing multi-modal emotion recognition systems, and the corresponding contributions of this paper are summarized as follows. First, in order to simulate the emotion fluctuation in a signal, segment-based emotion recognition is employed. Second, most of the multi-modal systems did not consider the influence of the emotion context and the dependence between the modalities for fusion. We propose an Attentively-Coupled LSTM that could process and fuse segmented signals efficiently. Furthermore, it can consider the dependencies between audio and visual modalities at the same time and use the attention mechanism to achieve attentive fusion for segment-based multimodal emotion recognition.

II. DATABASE

The existing audio-visual databases are summarized in Table 1. The database collection methods can be roughly divided into two types. The first is to give a preset emotional situation and ask participants to perform the emotion, and the second is to give participants to watch some videos or pictures to induce participants to explain their experience. Because this paper considers both verbal and non-verbal expressions of the speakers in the audio, we select the database with spontaneous expression. In the visual part, we focus more on the changes in facial expressions, so the database with full facial images is chosen. We also consider the duration of the databases, image resolution and lighting, and finally choose BAUM-1 [24] as the database for this study.

The selected Turkish audio-visual database, BAUM-1, was recorded by 31 Turks (13 females, 18 males). The database has a total of 1457 audio-visual files, with an average length of 4.66 seconds. According to the emotional expression, the database is divided into two categories: acted expression class and spontaneous expression class. They were labeled with two different groups of emotional label types.

TABLE. 1
AUDIO-VISUAL DATABASES

Database	Lang.	Elicitation	Time
GEMEP [25]	French	Acted	00:43
eNTERFACE'05 [26]	English	Induced	01:00
IEMOCAP [27]	English	Acted	02:11
RECOLA [28]	French	Acted	01:55
BAUM-1	Turkish	Both	01:56

In order to meet the requirement for the data format in the segment-based multimodal emotion recognition in this paper, each segment in the BAUM-1 database needs to be re-labeled for each modality (Audio or Visual). The seven emotional labels are anger, fear, sadness, surprise, neutral, disgust and happiness [29]. The sound type labels are categorized as laugh, breath, shout, silence and verbal [3]. We use the segmentation model proposed in [3] to segment the data, and re-label the data using audio-only, visual-only and audio-visual modalities. Segmentation steps include silence detection [30], verbal/non-verbal segment detection, and prosodic phrase segment detection [5]. Three annotators marked the emotions of each segment, and used voting to get the label of each segment. Fig. 2 shows the statistics of the segment-based training data.

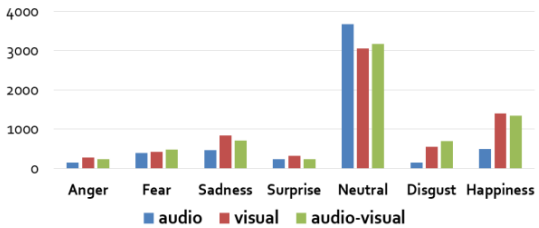

Fig. 2 Statistics of the Segment-Based Training Data

III. METHODS

A. Pre-Processing

Because of the need to consider the influence of the sound types of non-verbal segments in the audio, and the possible emotion change in the verbal segments, we need to segment the raw audio signal into segments, each of which contains only one emotion, in the pre-processing step. The segmentation points in the audio and visual data are used to divide the signal into segments.

B. Feature Extraction

The convolutional neural networks (CNNs)[31] are used to extract the features from audio-only signal. In addition to considering the features of non-verbal segments for recognition, we train two CNN models to perform audio feature extraction. The verbal segments are used to train a CNN model with emotions as the target outputs. Then, the non-verbal segments and the appropriate verbal segments are used to train another CNN of the sound type. Two models are shown in Fig. 3.

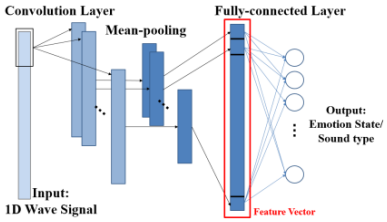

Fig. 3 The Audio Feature Extraction Model

We use VGG to extract facial emotion features. In the training phase, the image segment emotion label is used as the emotion classification target of all image frames in the segment, and the VGG is trained to recognize the single facial image emotion. In the test phase, each facial image in the segment is fed to the VGG, and finally the hidden layer vector is obtained from each image frame. The average of each dimension between vectors is taken as the facial emotion feature vector of this segment.

C. Calculation of Segment Attention weights

The Neural tensor network (NTN)[32] shown in Fig. 4 is used to train the model that can evaluate the degree of correlation between two input vectors, and this model is trained independently of other models. The architectural concept is described in (1). Given the audio emotion feature vector F_A^t and the parallel facial emotion feature F_V^t, the attention weight score s^t of the t^{th} segment in the emotion expression statement is calculated using the NTN.

$$s^t = u_R^T f(F_A^t M_R F_V^t + V_R \begin{bmatrix} F_A^t \\ F_V^t \end{bmatrix} + b_R) \quad (1)$$

Where M_R and V_R denote the transformation matrix and the vector, respectively. u_R is an activation function. R is the types of relationship between the two features, or known as tensor dimension. We can define the tensor dimension by ourselves and use this formula to calculate the degree of correlation between two features.

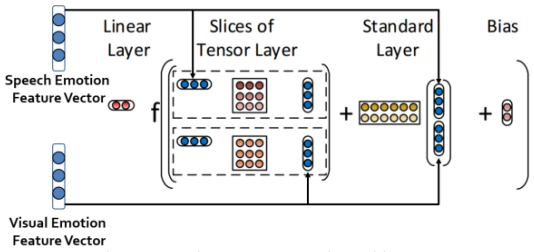

Fig. 4 Neural Tensor Network Architecture

If the two emotion labels of the audio and visual feature vectors are the same, the training target is 1. Otherwise, it is set to 0. We use this framework to assess whether this segment can fully express the emotions, and whether it can have a significant impact on the preceding and succeeding segments, and use this emotion consistency score as the attention weight of the succeeding segment for fusion.

D. Segment-Based Emotion Recognition Model

The coupled long short-term memory model proposed in [33] is selected for emotion recognition. The coupled LSTM shown in Fig. 5 consists of two LSTM models. There are two memory cells at each time step. In the coupled LSTM, Cell 1 is the audio cell which processes the audio features, while Cell 2 is the image cell to process the facial emotional features.

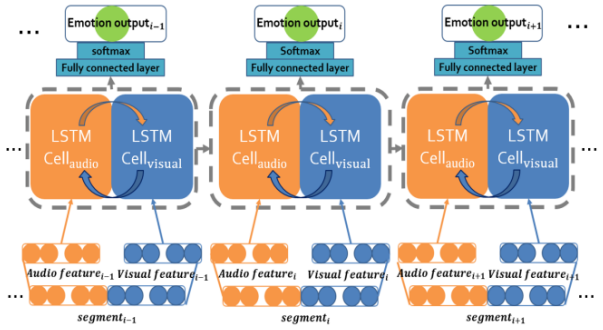

Fig. 5 Coupled LSTM Architecture

The difference from the general LSTM model lies in the way in which the Coupled LSTM model update the cell state. Generally, the LSTM model update cell state value as shown in (2), and at each time when the cell state is updated, the previous cell needs to be considered. Where i^t and f^t denote input gate and forget gate at the t-th time-step. c^t and z^t denote cell state and input sum at the t-th time-step. The Coupled LSTM model has two cells operating in parallel, and its update needs to consider the state values of the previous audio cell and the image cell simultaneously so as to achieve the interaction between two different modal cells. The key steps of the phase mixing method are shown in (3). The updated cell state value is the summation of information of input and previous cell state values. And the hidden state h_{cell}^t is obtained by the product of output gate o_{cell}^t and cell state c_{cell}^t, shown as (4).

$$c^t = i^t \otimes z^t + f^t \otimes c^{t-1} \quad (2)$$

$$c_{cell}^{t} = i_{cell}^{t} \otimes z_{cell}^{t} + f_{cell}^{t} \otimes \frac{1}{2} c_{cell}^{t-1} + f_{cell}^{t} \otimes \frac{1}{2} c_{\neg cell}^{t-1} \quad (3)$$

$$h_{cell}^{t} = o_{cell}^{t} \otimes tanh(c_{cell}^{t}) \quad (4)$$

Since the Coupled LSTM model is passed through each segment, the cell state value and the cell state are completely transmitted, but not all messages can fully express the same amount of information. So, we apply attention mechanism and design the Attentively-Coupled LSTM shown in Fig. 6 to assist the model in making subsequent identification based on the emotion consistency between the previous audio and visual segments. The emotion consistency is calculated by the NTN. With the output score as the segment attention weight, each segment is weighted during recognition, so that the model can focus on the important segment information with higher attention weight when identifying emotions. When the model is updated, the weight of the segment is given, and (3) is replaced by (5). The state value of the cell is completely passed in, but the state value from the counterpart cell needs to be weighted by the segment weight. This model sums the values of h_{cell}^{t} and uses two linear layers and softmax to obtain the emotions of each segment.

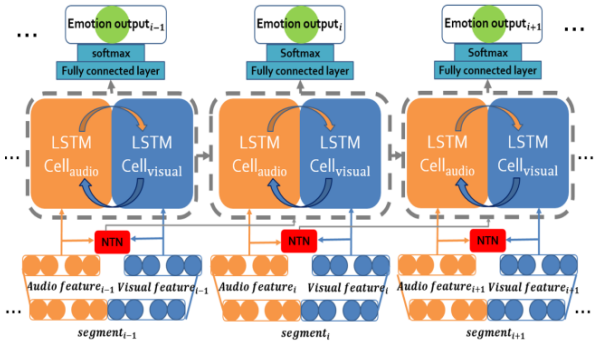

Fig. 6 Attentively-Coupled LSTM architecture

$$c_{cell}^{t} = i_{cell}^{t} \otimes z_{cell}^{t} + (f_{cell}^{t} \otimes \frac{1}{1+s^{t-1}} c_{cell}^{t-1} + f_{cell}^{t} \otimes \frac{s^{t-1}}{1+s^{t-1}} c_{\neg cell}^{t-1}) \quad (5)$$

IV. Experimental Results

First, we adjusted the model parameters from the experimental results to find the best parameters of the proposed system. The parameters we adjusted were the kernel size and filter number of CNN models, input size and number of layers of VGG, tensor dimension of NTN model and hidden size of LSTM. Table 2 shows the best model parameters of our proposed system. Fig. 7 shows the confusion matrix of the output of the proposed system.

We compared our system with other bimodal emotion recognition systems using the segmented and re-labeled BAUM-1database. The methods for comparison were Feature-level fusion method, Decision-level fusion method, Coupled

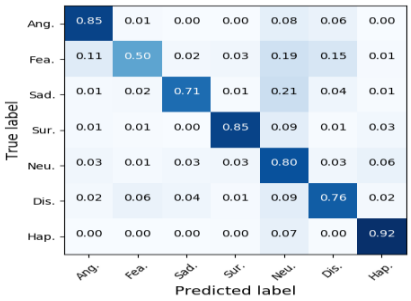

Fig. 7 Confusion Matrix of Attentively-Coupled LSTM

TABLE. 2
THE BEST PARAMETER SETTINGS IN OUR SYSTEM

Models	Settings
Audio emotion CNN model	Kernel size = 64, Number of filters = 300
Sound type CNN model	Kernel size = 32, Number of filters = 700
VGG-19 model	Input size = 72*72
NTN model	Tensor dimension = 5
Attentively-Coupled LSTM	Hidden dimension = 128

LSTM, AVEF [34] and the proposed Attentively-Coupled LSTM. Attention-based LSTM was used as the classification model of the Feature-level fusion method and Decision-level fusion method. The audio and visual feature extraction models of the Feature-level fusion, Decision-level fusion and Coupled LSTM were the same as our proposed system. AVEF used AlexNet[35] as audio feature extraction model, C3D[36] network as visual feature extraction model and Deep Belief Network (DBN)[37] as classification model, and it performed well on the BAUM-1 corpus. Table 3 shows the experimental results of the five emotion recognition systems testing in BAUM-1.

TABLE. 3
RESULTS OF BIMODAL EMOTION RECOGNITION SYSTEMS

Methods	Acc.
Feature-level fusion	66.4% ± 1.5%
Decision-level fusion	61.3% ± 2.3%
Coupled LSTM	60.4% ± 2.5%
AVEF	59.2% ± 1.2%
ACLSTM	70.1% ± 1.3%

V. Conclusions

From the experiments, it could be seen that consideration of using segment-based emotion consistency as attention weights improved the performance of Coupled LSTM model for segment-based multimodal emotion recognition. Compared with some existing audio-visual bimodal emotion recognition architectures, Attentively-Coupled LSTM proposed in this paper, achieved the best results. This shows the importance of applying the segment weighting attention mechanism to the model-level fusion model for segment-based multimodal emotion recognition. The method we propose can be easily applied to existing segment-based systems and effectively improve system performance.

REFERENCES

[1] M. Pantic and L. J. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* no. 12, pp. 1424-1445, 2000.

[2] M. Neumann and N. T. 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.

[3] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019: IEEE, pp. 5866-5870.

[4] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010: ACM, pp. 1459-1462.

[5] M. Dominguez, M. Farrús, and L. Wanner, "An automatic prosody tagger for spontaneous speech," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 377-386.

[6] C.-W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 2017: IEEE, pp. 583-588.

[7] K.-Y. Huang, C.-H. Wu, and M.-H. Su, "Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses," *Pattern Recognition,* vol. 88, pp. 668-678, 2019.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, no. 8, pp. 1735-1780, 1997.

[9] A. Schirmer and T. C. Gunter, "Temporal signatures of processing voiceness and emotion in sound," *Social cognitive and affective neuroscience,* vol. 12, no. 6, pp. 902-909, 2017.

[10] J. Deng, S. Frühholz, Z. Zhang, and B. Schuller, "Recognizing emotions from whispered speech based on acoustic feature transfer learning," *IEEE Access,* vol. 5, pp. 5235-5246, 2017.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[12] B. Yang, J. Cao, R. Ni, and Y. Zhang, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access,* vol. 6, pp. 4630-4640, 2017.

[13] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence,* vol. 31, no. 1, pp. 39-58, 2008.

[14] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies," *APSIPA transactions on signal and information processing,* vol. 3, 2014.

[15] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proceedings of the 19th ACM international conference on multimodal interaction*, 2017: ACM, pp. 553-560.

[16] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE transactions on cybernetics,* vol. 49, no. 3, pp. 1110-1122, 2018.

[17] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250,* 2017.

[18] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 2122-2132.

[19] S. Sahoo and A. Routray, "Emotion recognition from audio-visual data using rule based decision level fusion," in *2016 IEEE Students' Technology Symposium (TechSym)*, 2016: IEEE, pp. 7-12.

[20] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T. S. Huang, and S. Levinson, "Audio-visual affect recognition through multi-stream fused HMM for HCI," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 2: IEEE, pp. 967-972.

[21] T.-H. Yang, C.-H. Wu, K.-Y. Huang, and M.-H. Su, "Coupled HMM-based multimodal fusion for mood disorder detection through elicited audio–visual signals," *Journal of Ambient Intelligence and Humanized Computing,* vol. 8, no. 6, pp. 895-906, 2017.

[22] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Two-level hierarchical alignment for semi-coupled HMM-based audiovisual emotion recognition with temporal course," *IEEE Transactions on Multimedia,* vol. 15, no. 8, pp. 1880-1895, 2013.

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

[24] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Transactions on Affective Computing,* vol. 8, no. 3, pp. 300-313, 2016.

[25] T. Bänziger and K. R. Scherer, "Introducing the geneva multimodal emotion portrayal (gemep) corpus," *Blueprint for affective computing: A sourcebook,* pp. 271-294, 2010.

[26] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, 2006: IEEE, pp. 8-8.

[27] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation,* vol. 42, no. 4, p. 335, 2008.

[28] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, 2013: IEEE, pp. 1-8.

[29] P. Ekman, "An argument for basic emotions," *Cognition & emotion,* vol. 6, no. 3-4, pp. 169-200, 1992.

[30] J.-P. Goldman, "EasyAlign: an automatic phonetic alignment tool under Praat," 2011.

[31] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998.

[32] X. Qiu and X. Huang, "Convolutional neural tensor network architecture for community-based question answering," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
[33] M.-H. Su, C.-H. Wu, K.-Y. Huang, and T.-H. Yang, "Cell-Coupled Long Short-Term Memory With L-Skip Fusion Mechanism for Mood Disorder Detection Through Elicited Audiovisual Features," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 31, no. 1, pp. 124-135, 2019.
[34] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (AVEF): A deep efficient weighted approach," *Information Fusion,* vol. 46, pp. 184-192, 2019.
[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.
[36] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489-4497.
[37] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation,* vol. 18, no. 7, pp. 1527-1554, 2006.