INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts

Jilt Sebastian, Piero Pierucci

Telepathy Labs GmbH, Zurich, Switzerland
{jilt.sebastian, piero.pierucci}@telepathy.ai

Abstract

In human perception and understanding, a number of different and complementary cues are adopted according to different modalities. Various emotional states in communication between humans reflect this variety of cues across modalities. Recent developments in multi-modal emotion recognition utilize deep-learning techniques to achieve remarkable performances, with models based on different features suitable for text, audio and vision. This work focuses on cross-modal fusion techniques over deep learning models for emotion detection from spoken audio and corresponding transcripts.

We investigate the use of long short-term memory (LSTM) recurrent neural network (RNN) with pre-trained word embedding for text-based emotion recognition and convolutional neural network (CNN) with utterance-level descriptors for emotion recognition from speech. Various fusion strategies are adopted on these models to yield an overall score for each of the emotional categories. Intra-modality dynamics for each emotion is captured in the neural network designed for the specific modality. Fusion techniques are employed to obtain the inter-modality dynamics. Speaker and session-independent experiments on IEMOCAP multi-modal emotion detection dataset show the effectiveness of the proposed approaches. This method yields state-of-the-art results for utterance-level emotion recognition based on speech and text.

Index Terms: emotion recognition, multi-modal, fusion techniques, deep learning

1. Introduction

Recognizing and responding to various emotional states of human is essential for effective interactions. Emotions are expressed through multiple channels such as linguistic content, voice characteristics, facial expressions, and gestures. These correspond to emotional content in text, speech, and video modalities. The emotional state could be represented as either categorical or dimensional. Basic emotions such as happy, angry, and sad are categorical labels whereas the degree of emotion is represented using dimensions such as valence, arousal, and dominance. All categorical emotions can be embodied in the dimensional plane. Emotion recognition can be performed either at utterance-level or at dialog-level. An utterance-level recognizer only uses a single utterance to infer the emotional state, while a dialog-level classifier makes use of self and/or inter-speaker influences for decision making.

Text is an important modality for emotion recognition [1, 2]. Specific words are often used for expressing the intended emotion. Sentiment-analysis from the text closely resembles the emotion recognition task from speech. However, the emotional cues present in speech are different from that present in the linguistic content. Emotional content in speech is characterized by the variations in voice characteristics such as pitch, energy, vocal effort, loudness, and other frequency-related measures. Studies on emotion recognition from speech rely on such low-level descriptors (LLDs) and their high-level statistical functions (HSF) [3]. Frame-level and utterance-level features are used for modeling various emotional states [4], sometimes with attention mechanisms [5, 6] for speech-based emotion recognition. Audio features are used in conjunction with visual features for audio-visual emotion recognition [7, 8, 9].

Recent developments in deep learning techniques are reflected in emotion recognition research. Various approaches based on deep [7], convolutional [10] and recurrent neural networks [3] have been proposed for speech and video based emotion recognition and text-based sentiment analysis. Variants of such networks with memory blocks [11] and attention mechanisms [12] were developed recently for multi-modal emotion recognition.

Multi-modal approaches make use of complementary information from multiple modalities to improve the emotion detection system. Multi-modal analysis mostly considers feature-level fusion (a.k.a., early fusion). This is done by concatenating the available features at the input level before passing it to the model. As a single system is developed for all the modalities together, early fusion method does not necessarily capture the variability within a domain [10, 13, 14, 15]. Rather, such models are useful in modeling complex inter-modality dynamics. ICON model [11] uses early fusion and dialog-level decision making by considering the histories of the dyadic-interactions. However, such systems require long contextual information which is not readily available in typical real-time interactions between humans and computer.

Decision-level fusion, on the contrary, is characterized by building separate models for each of the domains. This results in effectively capturing the intra-modality dynamics via domain-specific models [16, 17]. The final classification is performed via decision voting over a weighted averaging fusion. When the feature spaces are different, such techniques have limited performance as they fail to capture the inter-modality dynamics. Tensor-Fusion Network [18] uses embedding-based sub-networks for each modality and a tensor fusion layer along with an inference layer at the output, providing a fair balance between the inter and intra-modality dynamics.

This paper presents neural network models for emotion recognition focusing on text and speech domains. We employ various fusion techniques [19] to provide relevance to inter-modality dynamics, while keeping the separate models to capture the intra-modality dynamics. Inter-modality dynamics are further improved by using a joint-model for text and speech signals. Uni-modal architectures are thus enhanced by both early and late fusion techniques. The originality of the work lies in efficiently combining early and late fusion methods over deep-learning based uni-modal emotion detection systems. The Proposed approaches achieves state-of-the-art results for utterance-

http://dx.doi.org/10.21437/Interspeech.2019-3201

based emotion recognition from speech and corresponding transcripts.

This paper is organized as follows. Section 2 describes the proposed method for emotion recognition from text. Section 3 presents the method for classifying the emotions based on speech. Section 4 presents the various fusion strategies adopted. Section 5 discusses the experimental evaluation. Conclusions are presented in Section 6.

2. Emotion Recognition from Text

This section presents the feature extraction and the proposed neural network framework for text-based emotion recognition. This is shown as "uni-modal LSTM for text" in the block diagram of the proposed approach (Figure 1). The feature extraction module provides a representation for each utterance whose contextual dependencies are modeled in the LSTM based neural network.

2.1. Feature Extraction

Features are extracted from the utterance transcriptions by employing a convolutional neural network (CNN). Neural network based feature extractors such as CNN learn abstract representations of the input sentence which contain the semantic meaning based on words and word-probabilities. A simple CNN with a convolutional and max-pooling layer is used as the feature extractor [20].

The input representation to CNN are 300-dimensional pretrained word embeddings. These are extracted from the FastText embeddings [21]. The convolution layer consists of three filters with sizes f_1,f_2, f_3 with f_{out} feature maps each. We perform 1D convolutions using these filters followed by max-pooling on its output. The pooled features are finally projected onto a dense layer with dimension D_T and its activations are used as the textual representation $T \in R^{D_T}$.

2.2. LSTM RNN framework

The neural network model for classifying the emotional state of the input sentence consists of a long short-term memory (LSTM) based recurrent neural network (RNN). The architecture consists of an LSTM recurrent layer and 3 fully connected layers. The recurrent connection captures the relevant contextual information for classifying given text of the utterance. This helps in identifying the emotion where the consecutive words provide additional cue for the emotional class. The D_T dimensional features per utterance are fed to the LSTM layer with N_1 recurrent connections. The fully connected layers have descending number of hidden units N_2, N_3 and N_4, the latest being the number of emotional categories.

3. Emotion Recognition from Speech

Acoustic features are extracted for each utterance in the feature extraction stage and are used for building CNN for emotion recognition from acoustic features. This CNN is shown as "joint-CNN model" in Figure 1. This is because, we will utilize this model for early fusion in Section 4, which has an improved performance.

3.1. Feature Extraction

Feature extraction from speech signal is performed using openSMILE toolkit [22]. The IS13 ComParE1 challenge [23] based feature extraction results in 6373 features per utterance. The feature set consists of LLDs such as pitch, Mel-spectra, loudness, mel frequency cepstral coefficients, etc., and their statistical measures (HSDs) such as mean, standard deviation, minimum, maximum, etc. Min-Max normalization and L2-based feature selection are performed to reduce the feature dimension to D_s. This lower-dimensional speech-based features $S \in R^{D_s}$ are used as the input to CNN.

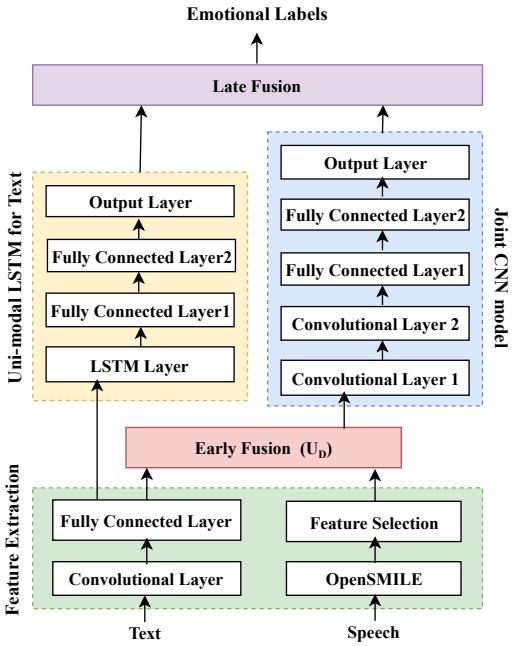

Figure 1: *Block diagram of the proposed approach.*

3.2. CNN Framework

The neural network for obtaining the emotional classes for speech consists of two convolutional layers with ReLU activations, each of which is followed by a max-pooling layer. It is then followed by three fully-connected layers. Each of the convolutional layers has N_f number of filters with a width of N_w each. The convolutional layers work on unit stride, and the filters are learned for the emotional categories. The output of the second convolution layer is flattened before feeding into two fully connected layers having sizes N_{c2} and N_{c3} respectively. The output layer has the length of the number of emotional labels.

LSTM-based model explained in Section 2 can be exploited for speech as well. We observed that it also provides similar performance. Nevertheless, this work sticks to using CNN for speech because it yields better performance on fusion than using LSTM-based systems, probably because of a completely different modelling approach.

4. Fusion Techniques

The proposed method combines the model posteriors for the late fusion and concatenates text and acoustic features for the early fusion.

4.1. Early Fusion

Early fusion is one of the most common fusion techniques. In the feature-level fusion, we combine the information obtained via feature extraction stages of text and speech [24]. The final input representation of the utterance is,

$$U_D = tanh((W^f[T;S] + b^f)) \quad (1)$$

The CNN model for speech described in Section 3 is also considered as a joint-model for emotion detection. This is shown in Figure 1. The convolution layer is fed with the feature vector of size $U_D = (T;S)$ which is obtained by early fusion. This helps in capturing the inter-modality dynamics between the text and speech within the same model. We term this model as joint-CNN.

4.2. Late Fusion Techniques

Three types of decision level fusions are considered at the output of uni-modal and joint-CNN emotion recognizers. We consider late fusion of the joint model with a domain specific model. They help in capturing the inter-modality dynamics, in addition to that captured in the joint-CNN model.

4.2.1. Late Fusion-I

This decision-level fusion is performed by combining the scores of multiple systems. The sum combination rule is beneficial if different models are built on similar feature space, e.g, ensemble methods. The output score of the given utterance in average fusion is given by,

$$Score = \frac{Score(T) + Score(S,T)}{2} \quad (2)$$

Late fusion-I gives equal preferences to the posteriors coming from LSTM-based text emotion recognizer and the joint-CNN model.

4.2.2. Late Fusion-II

In this approach, we use late fusion of posterior probabilities from LSTM-based text emotion recognizer and the joint-CNN (early fusion) system with different weight values. The output score of the given utterance in weighted average fusion is given by,

$$Score = w_1 * Score(T) + w2 * Score(S,T) \quad (3)$$

where, $w1, w2 \leq 1$, and $w_1 + w_2 = 1$. The weights are determined based on the performance on the validation data using trial and error method.

4.2.3. Late Fusion-III

The posterior probabilities can also be combined using a product rule [19]:

$$Score_j = \frac{Score_j(T) * Score_j(S,T)}{\sum_{j'} Score_{j'}(T) * Score_{j'}(S,T)} \quad (4)$$

where $j' \neq j$, and j represents the class index. The underlying assumption is that the feature spaces are different and class-conditionally independent [19]. This is useful as we are building one domain specific model and combining it with joint-CNN with a different feature representation. The inter-modality dynamics of the joint-CNN system is further improved by using the late fusion techniques.

5. Experimental Evaluation

5.1. Dataset

Muti-modal emotion recognition is performed using Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [25]. It is a database consisting of audio, video, transcription, and motion capture data obtained from dyadic interactions between pairs of 10 speakers. This is one of the most commonly used dataset for multi-modal emotion recognition task. The subjects are involved in emotional interactions which are grouped into five sessions, each of which has a pair of subjects. The videos are segmented into utterances with annotations of emotions spanning over 9 categorical and 3 dimensional labels. The IEMOCAP dataset has approximately 12 hours of spoken content, and is one of the largest publicly available dataset for this task. We consider six major categories of emotions for the classification (anger, happiness, sadness, neutral, excitement, and frustration). Each recording is labelled by at least 3 annotators. The recordings in which majority of them have an agreement upon the emotion label are used for the experiments.

5.2. Experimental Procedure

we consider six emotional categories to remain consistent with the recent literature on multi-modal emotion recognition. Most of the existing literature on speech-based emotion recognition [3, 5] only consider four of these categories (excluding excitement and frustration). However, the performance degrades with six classes for these models owing to the larger degree of confusion between classes, and the variation in the distribution of classes. Some of the multi-modal systems only consider four or five classes [26, 27, 28]. The proposed system and the baselines outperform such systems under similar evaluations (four class classification task), hence they are not included for comparison in this paper.

The training and validation data are created using the first four sessions consisting of 8 speakers in 120 videos (5810 utterances). Session 5, consisting of 31 videos (1623 utterances) is used for the testing. By this approach, we make sure that the test speakers and session are unseen by the trained models. Validation set is chosen as 20% of the training data.

Table 1: *Hyperparameter values for the models (ED-Emotion Detection).*

Model	Architecture(s)	Hyperparameters	Values
Feature extractor (text)	CNN	(f^1, f^2, f^3, f_{out})	(3, 4, 5, 64)
Feature extractor Output	-	(D_T, D_S)	(100, 100)
Text ED	LSTM	Number of LSTM units (N_1)	128
Text ED	LSTM	(N_2, N_3, N_4)	(512, 128, 6)
Speech ED	CNN	(N_f, N_w)	(128, 5)
Speech ED	CNN	(N_{c2}, N_{c3}, N_{c4})	(512, 128, 6)
Late Fusion-II	CNN, LSTM	$(w_1, w2)$	(0.3, 0.7)

The hyperparameter details of text and speech-based emotion recognition and other parameters of the proposed approaches are shown in Table 1. For the uni and bi-modal architectures, Adam optimizer with an initial learning rate of 0.001 is used with cross entropy loss. Training is performed for 20 epochs for CNN, and 40 epochs for LSTM with an early-stopping criterion. The validation loss is monitored with a patience factor of 6. 20% dropout is applied to the LSTM layers for regularization. The feature extraction from text uses a single layer CNN with the hyperparameters as shown in Table 1. LSTMs and CNNs were implemented using the Keras toolkit [29].

Experiments are performed to allow comparisons with

state-of-the-art multi-modal emotion recognition systems, with similar input feature representations, same number and type of categorical labels, and a similar evaluation scheme. Multi-modal emotion recognition is usually evaluated in the literature using weighted accuracy (WA) or, unweighted accuracy (UA) or F1 score (F1). We consider all of these measures to have a fair comparison with the literature.

5.3. Baseline Systems

We compare the performance of the proposed approaches with state-of-the-art utterance-level systems. Additionally, dialog-level systems are considered to analyze the difference in the performance owing to the contextual information supplied in-addition to the utterance-level features to such systems.

1. **Tensor fusion network (TFN)** [18]: It is a fusion-based approach which explicitly models the intra- and inter-modality dynamics. Uni, bi- and tri-modal interactions are aggregated in a specially designed fusion layer and an inference layer.

2. **Memory fusion network (MFN)** [12]: This approach makes use of a fusion mechanism known as delta-memory attention network to achieve the multi-view sequential learning. This model provides the state-of-the-art results for utterance-level multi-modal emotion recognition system. We consider TFN and MFN as our baseline systems, though they also use visual features in addition to speech and text.

3. **Bi-directional contextual LSTM (cLSTM)** [24]: It is a dialog-level emotion recognizer which classifies utterances by hierarchically modeling the contextual uni-modal and multi-modal features using separate LSTMs. The decision is influenced by the neighbouring utterances as well.

4. **Interactive conversational memory network (ICON)** [11]: This approach hierarchically models the self and inter-speaker influences using global memories. It consists of self-influence module, dynamic global influence module, and multi-hop memory for attaining this. This dialog-level model provides the state-of-the-art dialog-level emotion recognition. However, dialog-level systems such as ICON require history of utterances which is not readily available in real-time human-computer interactions.

5.4. Results

Table 2 shows the performance of the proposed approaches with various fusion techniques, the parameters of which are described in Table 1. Mean classification performance is evaluated through weighted accuracy (WA) and F1 score. IEMOCAP has 384 utterances for "neutral" emotion whereas only 144 utterances for the "happy" emotion in the test set. Unweighted accuracy (UA) computes the average of class-specific recalls, and is beneficial for unbalanced datasets such as IEMOCAP.

Early fusion resulted in improved measures over the uni-modal systems. Late fusion of uni-modal systems captures more inter-modality dynamics than the early fusion. Combinations of early and late fusions further improves the performance. Weighted sum combination of joint-CNN and LSTM systems provides best unweighted accuracy, whereas the product combination of these systems gives the best overall accuracy. The class-wise performance is minimum for "happy" category which has the minimum support, and maximum for "sad" category, across all the systems. Maximum confusion was 33% between "happy" and "excited" categories.

The proposed approaches are compared to the baseline systems in Table 3. Both of the proposed approaches beats TFN for all of the performance measures. Late Fusion-III performs superior to state-of-the-art baseline (MFN) for all of the evaluation measures. It should be noted that the proposed methods only use text and speech modality whereas the baselines also use visual features for classification. Nevertheless, the proposed fusion methods achieve improvement over the baselines.

Table 2: *Performance of the proposed approaches on IEMOCAP dataset.*

SI No.	Model	Type	UA	WA	F1
(1).	CNN	Audio	47.1	46.0	45.0
(2).	LSTM	Text	53.0	53.5	53.6
(3).	Early Fusion	joint-CNN (Eqn. 1)	55.6	57.2	57.2
(4).	Late Fusion-I	Average (Eqn. 2)	58.8	59.5	59.7
(5).	Late Fusion-II	Weighted (Eqn. 3)	**60.2**	59.6	59.6
(6).	Late Fusion-III	Product (Eqn. 4)	59.3	**61.2**	**61.2**

Table 3: *Comparison with the baselines and dialog-level systems (marked by (*)). A-Audio, T-Text and V-Video.*

Model	Features	UA	WA	F1
TFN	(A+T+V)	55.9	58.8	58.5
MFN	(A+T+V)	58.3	60.1	59.9
Late Fusion-II	Weighted (A+T)	**60.2**	59.6	59.6
Late Fusion-III	Product (A+T)	59.3	**61.2**	**61.2**
cLSTM*	(A+T+V)	56.1	59.8	59.0
ICON*	(A+T)	60.0	**63.0**	**63.0**

We consider dialog-level emotion recognizers, ICON and cLSTM, to sought insights about the differences in performance with respect to our utterance-level systems. This is also shown in Table 3. The performance of the proposed fusion methods are superior to cLSTM (which also uses the visual features). The results of product-rule based fusion are gracefully below the ICON model. This shows that the state-of-the-art utterance-level performance is competitive with that of the dialog-level measures. The experimental evaluation validates that state-of-the-art utterance-level results can be achieved by selecting the models which when combine together at the decision-level, captures inter-modality dynamics.

6. Conclusions

We presented novel fusion techniques on deep learning models for improved emotion recognition in multi-modal scenarios. Proposed combination of early and late fusion techniques on text and speech based models utilise complementary information available across lingustic and spoken content. They achieve state-of-the-art utterance-level recognition performance when evaluated on a standard bench-marking dataset. The performance is close to the best available model for multi-modal emotion recognition, which utilizes self and other speaker influences in the decision making. This suggest that adequate modeling and fusion methods serve as a good direction for emotion recognition from multiple modality. Future research would focus on incorporating visual features and models specific to video segments to aid in the decision making via fusion.

7. Acknowledgements

The authors would like to acknowledge Dr. Thi Truc Vien Nguyen for having started this research while at Telepathy Labs GmbH, Zurich in 2018. We also thank the members of Speech Technology Group at Telepathy Labs GmbH for their constructive feedback and reviews.

8. References

[1] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 2005, pp. 579–586.

[2] C. Strapparava and R. Mihalcea, "Annotating and identifying emotions in text," in *Intelligent Information Access*. Springer, 2010, pp. 21–38.

[3] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[4] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[5] G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018.

[6] B. S. Zixing Zhang, Bingwen Wu, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *Accepted for publication in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[7] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[8] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[9] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency, "Emoreact: a multimodal approach and dataset for recognizing emotional responses in children," in *Proceedings of the 18th acm international conference on multimodal interaction*. ACM, 2016, pp. 137–144.

[10] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 2016, pp. 439–448.

[11] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2594–2604.

[12] M. N. P. S. C. E. M. L. Zadeh A, Liang PP, "Memory fusion network for multi-view sequential learning," in *Proceedings of the thirty-Second AAAI Conference on Artificial Intelligence 2018 April*, 2018.

[13] D. P. Morency LP, Mihalcea R, "Towards multimodal sentiment analysis: Harvesting opinions from the web." in *In Proceedings of the 13th international conference on multimodal interfaces ACM.*, 2011, pp. 169–176.

[14] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2013, pp. 973–982.

[15] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.

[16] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing, "Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis," *arXiv preprint arXiv:1609.05244*, 2016.

[17] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[18] 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.

[19] D. M. Tax, M. Van Breukelen, R. P. Duin, and J. Kittler, "Combining multiple classifiers by averaging or by multiplying?" *Pattern recognition*, vol. 33, no. 9, pp. 1475–1485, 2000.

[20] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[21] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[22] 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*. ACM, 2010, pp. 1459–1462.

[23] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[24] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "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)*, vol. 1, 2017, pp. 873–883.

[25] 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.

[26] S. Tripathi and H. Beigi, "Multi-modal emotion recognition on iemocap dataset using deep learning," *arXiv preprint arXiv:1804.05788*, 2018.

[27] Y. Gu, S. Chen, and I. Marsic, "Deep multimodal learning for emotion recognition in spoken language," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5079–5083.

[28] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," *Proc. Interspeech 2018*, pp. 247–251, 2018.

[29] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.