

FUSION APPROACHES FOR EMOTION RECOGNITION FROM SPEECH USING ACOUSTIC AND TEXT-BASED FEATURES

Leonardo Pepino*

Pablo Riera*

Luciana Ferrer*

Agustín Gravano*[†]

*Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina

[†]Departamento de Computación, FCEyN, Universidad de Buenos Aires (UBA), Argentina

ABSTRACT

In this paper, we study different approaches for classifying emotions from speech using acoustic and text-based features. We propose to obtain contextualized word embeddings with BERT to represent the information contained in speech transcriptions and show that this results in better performance than using Glove embeddings. We also propose and compare different strategies to combine the audio and text modalities, evaluating them on IEMOCAP and MSP-PODCAST datasets. We find that fusing acoustic and text-based systems is beneficial on both datasets, though only subtle differences are observed across the evaluated fusion approaches. Finally, for IEMOCAP, we show the large effect that the criteria used to define the cross-validation folds have on results. In particular, the standard way of creating folds for this dataset results in a highly optimistic estimation of performance for the text-based system, suggesting that some previous works may overestimate the advantage of incorporating transcriptions.

Index Terms— speech emotion recognition, fusion, deep learning, BERT

1. INTRODUCTION

Speech emotion recognition (SER) is an active research area with important applications in the field of human-computer interaction. SER is a complex task even for humans [1]. In fact, in spite of recent advances enabled by deep learning models and the release of larger emotion datasets, the performance of SER systems is still relatively poor, with average recall rates usually well below 70% on the most realistic datasets, indicating that it remains an open problem.

Most SER systems use low-level descriptors (LLD) extracted from the audio signal such as MFCCs, pitch and voice quality features [2], or features learned automatically from spectrograms using deep neural networks [3, 4]. The excellent performance of current automatic speech recognition systems (ASR) also allows us to extract reliable text transcriptions from the speech without the need for human annotators. A few works have incorporated this information into SER systems. In some of these studies, emotional word-based vectors were computed from word occurrences in each emotion class [5], or using external lexicons [6]. Similarly, emotional vectors can be extracted from phonemes [7]. In some works [5], the word-based vectors are used as input to SVM classifiers together with high-level statistics of the acoustic LLDs. Another approach is to train text- and audio-based classifiers separately and combine their outputs to

make a final prediction [8]. Recently, deep neural networks have been used to learn audio-linguistic embeddings [9] and to train emotion classifiers in an end-to-end framework combining text and audio modalities [10, 11, 12, 13].

In this paper, we study different ways of fusing audio and linguistic information, using early and late fusion techniques and comparing different training approaches, including (1) initializing two individual branches with models trained separately for audio and text and further fine-tuning the last few layers, (2) fixing the text and audio branches and training only the fusion parameters, and (3) training the whole combined neural network from scratch. For the audio branch, we use a standard approach based on MFCC, pitch, loudness, jitter, shimmer and logHNR features. For the text branch, we use contextualized word embeddings [14] instead of the standard word embeddings like Glove [15] used in most of the previous works [12, 10]. Standard word embeddings like those obtained with Glove are extracted independently of the context in which the words are found. For example, the word “sad” would be assigned the same embedding whether the phrase was “I am very sad” or “I am not sad at all”. On the other hand, contextualized word embeddings like those extracted by BERT take into account the whole phrase in which the word is found. As a consequence, the embedding corresponding to the word “sad” in those two phrases would most likely be different. We hypothesized that this characteristic should positively impact SER performance. To our knowledge, [16] is the only work in which word embeddings extracted with BERT have been used for SER. In that paper, authors propose a shared representation of audio, text and video modalities through deep canonical correlation analysis. A comparison with other types of embeddings is not shown in that work.

The proposed models are tested on the well-studied IEMOCAP dataset [17], as well as on the more challenging MSP-PODCAST dataset [18]. Our first contribution is to show that linguistic information gives significant improvements in performance when combined with acoustic information on the MSP-PODCAST dataset. As far as we know, this is the first time that linguistic information has been used on this dataset. Second, we show that the use of contextualized word embeddings obtained with BERT results in significant improvements with respect to using standard word embeddings obtained with Glove. Third, we propose a novel way to fuse audio and text information by pretraining the neural network in audio and text modalities and then fine-tuning the fused model. Finally, we show that creating folds by speaker is not sufficient to obtain fair performance predictions on IEMOCAP, since the data contains scripted dialogues which greatly affect the performance of text-based systems when the same script is observed in training and testing. This being such a widely used dataset, we believe this observation is of great importance to the research community.

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-18-1-0026. We gratefully acknowledge Carlos Busso for giving us access to the TSP-PODCAST dataset, and NVIDIA Corporation for the donation of the Titan Xp GPU 2 used for this research. Correspondence: lpepino@dc.uba.ar

2. MODELS

This section describes the models used in the experiments. First, the individual models for each modality are introduced. Then, models that combine the text and audio information are described. All models are trained to optimize cross-entropy loss for four emotion classes: happy, sad, angry and neutral.

2.1. Text-based model

Recently, a language model called BERT [14], trained with large amounts of data has been released to the community. This model can be fine-tuned or used as a feature extractor for downstream tasks, achieving state-of-the-art results on many of them. BERT is based on the Transformer [19] – a network capable of modeling long contextual information, generating word embeddings that are conditioned on the phrase in which the word is found. In this study, a sequence of word embeddings is extracted from speech transcriptions using the pretrained BERT base uncased model, which consists of 12 layers, 12 attention heads and 110M parameters. The word embeddings are formed by adding the activations of the last 4 layers of the pretrained BERT model without fine-tuning. The resulting features are used as input to the text model shown on the left of Figure 1.

The first layer (L_{T1}) in our text-based model operates on each embedding in the sequence reducing its dimensionality from 768 to 128. Then, 2 convolutional layers (L_{T2} and L_{T3}) model relationships across neighboring elements of the sequence. Finally, an average over time is taken, resulting in an embedding that summarizes all the information in the sample. A final dense layer with softmax activation predicts emotion probabilities $P(C_k)$. We applied batch normalization in all layers.

To make a comparison with non-contextualized word embeddings, we trained the same model using 300-dimensional Glove embeddings [15].¹

2.2. Audio-based model

Each speaker utterance was divided into 32ms segments, using a hop length of 10ms. The following acoustic features were extracted from each window using openSMILE [20]: pitch, jitter, shimmer, logHNR, loudness, and the first 13 MFCCs. These features were normalized to have a mean of 0 and standard deviation of 1, using the global statistics. Finally, first-order differences were added for all features to form a sequence of 36-dimensional feature vectors that are the input to the neural network shown on the right of Figure 1.

The audio model consists of two convolutional layers L_{A1} and L_{A2} that model the temporal evolution of the input sequence followed by mean-pooling over time. A final dense layer with softmax activation returns the emotion probabilities $P(C_k)$.

2.3. Fusion models

In this section, we describe the strategies we implemented to combine audio and text information. In all cases, the fusion model consists of two parallel branches processing audio and text separately up to a layer where the information from the two branches is merged. The models differ on the location of the merging layer, on the network appended after merging, and on the training approach.

2.3.1. Early Fusion

In the early fusion (EF) approach, the fixed-size embeddings obtained after mean pooling in the audio and text models (Figure 1,

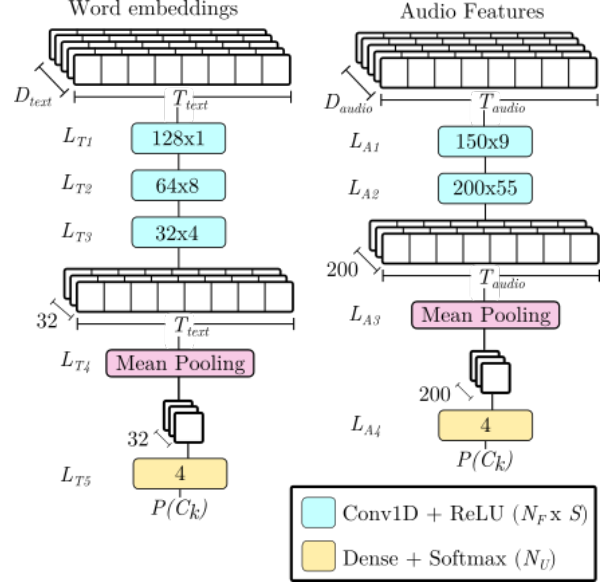


Fig. 1: Text-based and audio-based architectures. T_{text} and T_{audio} are the sequence lengths of the model inputs and D_{text} and D_{audio} are the number of features for each input. N_F is the number of convolutional filters, S is the kernel size and N_U is the number of neurons in dense layers. 1D-Convolutional layers operate on the time axis.

layers L_{T4} and L_{A3}) are concatenated resulting in a multi-modal embedding of 232 dimensions. This embedding is input to a feed-forward neural network with a hidden dense layer of 128 units with ReLU activation and an output layer with 4 units and softmax activation.

2.3.2. Late Fusion

In the late fusion model (LF), the logits (pre-softmax) of the audio and text models are concatenated (Figure 1, layers L_{T5} and L_{A4}) resulting in an 8-dimensional vector that is used as input to a dense layer of 4 units with softmax activation. This dense layer learns to combine the logits of audio and text modalities to generate the final output probabilities $P(C_k)$. We have also explored learning a scalar weight for each system instead of a full dense layer but the resulting performance was slightly worse.

2.3.3. Training strategies

We trained our fusion models in 3 different ways:

- Cold-start (CS): Use Xavier uniform initialization [21] for all layers of the fusion model and train the model jointly from scratch.
- Pre-trained (PT): Train the audio and text models separately and use the trained weights to initialize the corresponding layers of the audio and text branches in the fusion model. The layers after merging are initialized with Xavier uniform initialization. Only these layers are trained, keeping the layers up to the merging point frozen.
- Warm-start (WS): Initialize all layers as in the PT approach but instead of training only the layers after merging, train also the layers right before pooling for each branch (L_{T3} and L_{A2}), keeping the first layers (L_{T1} , L_{T2} and L_{A1}) frozen, as in the PT approach. This procedure, in contrast to PT, allows the layers immediately before the pooling to change their weights.

¹The pretrained Glove model we used can be downloaded from <http://nlp.stanford.edu/data/glove.42B.300d.zip>.

3. EXPERIMENTAL SETUP AND DATASETS

Our experiments were performed on the IEMOCAP and MSP-PODCAST datasets. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [17] has a length of approximately 12 hours and consists of scripted and improvised dialogues by 10 speakers. It is composed of 5 sessions, each including speech from an actor and an actress. Annotators were asked to label each sample choosing one or more labels from a pool of emotions. In this work, we used 4 emotional classes: anger, happiness, sadness and neutral, and following [22], we relabeled excitement samples as happiness. Instances from other classes and with no annotator agreement were discarded.² For this dataset, human transcriptions are used for the text-based system.

To test our models we used 5-fold cross-validation, organizing the folds so that training and test sets do not share actors or scripts. This last point is very important for the text-based model, as has been noted in [7], because dialogues from the same script are very similar. We show the effect of the criteria used for making the folds on both text and audio models in Section 4.1.

The MSP-PODCAST dataset v1.4 [18] contains speech segments from podcast recordings, annotated using crowdsourcing. After discarding the instances not belonging to any of the 4 emotional classes under study, the training set contains 12078 speech segments from 601 speakers, while the test set contains 5557 utterances from 50 speakers not present in the training set. The training and test set definitions used in this paper are the ones provided with the dataset. The test set is gender balanced. Speech transcriptions were extracted using the Google Cloud Speech-to-Text API.³

To counteract the effect of class imbalance present in both datasets, a cost-sensitive training strategy was applied by multiplying the loss of each instance with the inverse of the frequency of the class it belongs to. The models were optimized using Adam [24] with a learning rate of 0.0007, except for the fine-tuning case in which the learning rate was decreased to 0.0001, and the case of late fusion using pre-trained branches where learning rate was increased to 0.01. During the first 40 steps, the learning rate was linearly increased from 0 to the final value, except for the late fusion system with pretraining (LF-PT). We applied dropout with 0.5 probability at the input of layer L_{A2} only for the audio branch. As the input sequences have variable length, we padded them with zeros up to a maximum sequence length and then masked the padded values.

We report two different metrics: average recall (AvRec), and the average area under the ROC (AvAUC). Average recall is used instead of accuracy since both datasets have significant imbalance across classes. Both averages are computed over the four target emotions, considering a one-vs-all problem in order to compute individual recall and AUC values.

We observed that using early stopping in IEMOCAP led to inconsistent results as the data are scarce to generate a validation fold large enough. For this reason, the number of epochs for training each model was selected by optimizing the median AvAUC value on IEMOCAP over 5 seeds. The architectures and hyperparameters were also selected based on IEMOCAP results (sometimes using a single seed). The final results on IEMOCAP were obtained over 10 seeds, including the 5 used for the optimization of the epoch and the hyperparameters tuning. This leads to possibly optimistic results on this dataset. On the other hand, the results on MSP-PODCAST

²Note that discarding no-agreement samples and samples from non-target emotions is not an ideal practice [23]. Here, we decided to do this since it is standard practice in SER literature, facilitating comparisons across papers.

³<https://cloud.google.com/speech-to-text/>

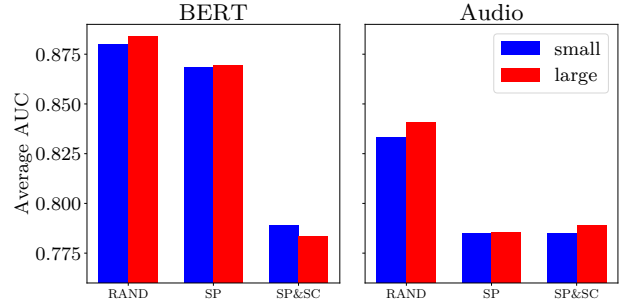


Fig. 2: Effect of different criteria for defining the folds in IEMOCAP on audio- and text-based systems for two different model sizes (small and large). RAND: random folds, SP: by-speaker folds, SP&SC: by-speaker and by-script folds.

were obtained using the same number of epochs and hyperparameters chosen for IEMOCAP, also averaging over 10 seeds. All of our models were trained using Keras [25].

4. RESULTS AND DISCUSSION

In this section we report results for individual and fused systems. We start by showing the effect that the criteria used to define the folds for cross-validation on IEMOCAP has on the two individual systems.

4.1. Effect of partition criteria for IEMOCAP folds

Figure 2 shows the AvAUC for three different criteria used to define the folds on IEMOCAP. Results are computed on the merged test scores for all folds. In all cases, 5 folds are used. We compare: (1) random folds (RAND), where no information about speakers or scripts is used to define the folds; (2) folds by speaker (SP) where each fold contains the two speakers from one of the sessions; and (3) fold by speaker and script (SP&SC) where the folds are defined as in the previous case, but only script 3 is used for testing while all other scripts are used in training. Note that this last option includes less data for each fold. Finally, we compare two different sizes of models: one using half of the nodes in the models from Figure 1 (small), and one using twice the number of nodes (large). We note that most papers use by-speaker folds [11, 12, 26], while some use random folds [10]. We are not aware of any work that splits by script, though some works discard the scripts altogether using only improvisation instances for testing [7, 13].

Figure 2 clearly shows that both random and by-speaker folds result in optimistic performance for the text-based systems. For the audio-based system, both by-speaker and by-speaker-and-script options lead to similar performance (indicating that the effect of by-speaker-and-script folds having less data is limited), while the random splits result in an optimistic estimation of performance. Furthermore, the conclusion of which model size is optimal for BERT features changes depending on the fold criteria, as a large model is more likely to overfit, but this effect can only be observed when using folds that do not repeat speakers or scripts between training and test sets. Given these results, we believe it is essential to define the folds for IEMOCAP carefully, not allowing speakers and scripts seen in training to be repeated in testing. In the remaining experiments, we use folds by speaker and script.

4.2. Audio- and text-based models

Figure 3 shows the performance obtained with the different systems on both datasets. Average AUC is reported using box and whiskers

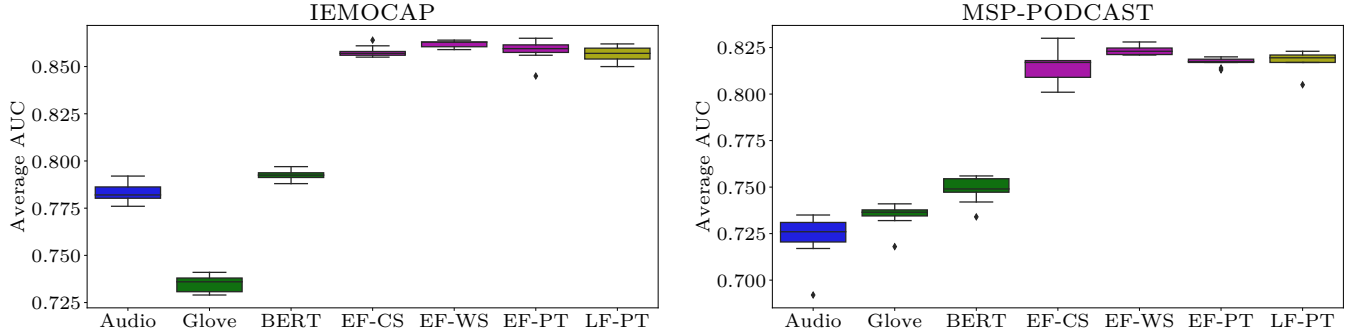


Fig. 3: Results for IEMOCAP and MSP-PODCAST dataset. Average AUC distributions for 10 different initialization seeds for different systems: audio model, Glove and BERT based text models, early fusion with cold-start (EF-CS), pretraining (EF-PT) and warm-start (EF-WS) and late fusion with pretraining (LF-PT) models.

Table 1: Median of evaluation metrics \pm interquartile range obtained using 10 seeds for all the tested models in both datasets.

| | IEMOCAP | | MSP-PODCAST | |
|-------|--------------------------------|---------------------------------|--------------------------------|---------------------------------|
| | AvRec (%) | AvAUC | AvRec (%) | AvAUC |
| Audio | 56,0 \pm 1,9 | .782 \pm .006 | 45,7 \pm 1,4 | .726 \pm .010 |
| Glove | 47,8 \pm 0,1 | .736 \pm .007 | 49,8 \pm 0,4 | .736 \pm .003 |
| BERT | 55,2 \pm 1,0 | .792 \pm .003 | 51,0 \pm 0,9 | .749 \pm .007 |
| EF-CS | 65,1\pm0,5 | .857 \pm .002 | 58,2 \pm 2,4 | .817 \pm .009 |
| EF-WS | 64,7 \pm 1,6 | .863\pm.002 | 59,1\pm1,8 | .823\pm.003 |
| EF-PT | 64,9 \pm 1,0 | .859 \pm .004 | 56,5 \pm 0,3 | .817 \pm .002 |
| LF-PT | 63,9 \pm 0,5 | .857 \pm .006 | 58,0 \pm 0,7 | .819 \pm .004 |

plots to show the variation in performance for 10 different seeds used to initialize the DNN weights. Table 1 shows both AvAUC and AvRec values. We can see that our proposed text model based on BERT embeddings shows slightly better performance than the audio model on IEMOCAP, while Glove embeddings give significantly worse performance. This contradicts previous results on IEMOCAP, where the text-based models significantly outperform audio models [10, 11]. As we showed in Section 4.1 this is explained by the way we have defined the folds, preventing the text model from being trained in dialogues very similar or identical to the ones present in the test set and avoiding unrealistically good performance estimates for these systems.

The effect of using BERT versus Glove to represent word information can be seen in Figure 3 and Table 1. BERT embeddings outperform Glove ones in both datasets with relative UAR improvements of 15.5% and 2.4% in IEMOCAP and MSP-PODCAST datasets, respectively. We attribute this performance gain to the contextual information imbued in the pretrained BERT model. While our text model could potentially learn contextual information from standard word embeddings like Glove, learning to represent negations or modification values would require a significant amount of data. We hypothesize that this is the reason why Glove performance is closer to BERT in MSP-PODCAST than in IEMOCAP, since the size and variability of dialogues in MSP-PODCAST may be allowing the text model to learn contextual information even from standard word embeddings.

Finally, we note the large effect that the seed has on our systems. In many cases, the ranking of systems changes significantly depending on the seed (results not shown due to lack of space), which thus highlights the critical importance of using several seeds in order to reach more solid conclusions.

4.3. Fusion models

As it has been noted in previous works, adding text information to audio-based SER systems gives significant performance improvements [10, 12]. This is also observed in our fusion experiments where for both MSP-PODCAST and IEMOCAP datasets, the AvRec improves 16% relative to the best performing single model. All fusion approaches perform similarly, in agreement with previous results in the literature [27, 11]. Only the late fusion approach with pre-training is shown here, due to space considerations. The other two training approaches gave similar results.

A small advantage of the warm-start approach can be observed for both datasets with the early fusion architecture, indicating that this direction may be worth further exploration. In the future, we plan to explore approaches where the fusion is made before or at the pooling layer. We believe this has the potential to give additional benefits since the interaction between both modalities is most likely happening at short time intervals rather than at phrase level.

5. CONCLUSIONS

We presented different approaches for emotion recognition from speech using audio features and transcriptions. We showed results on two publicly available datasets: IEMOCAP and MSP-PODCAST. We demonstrated the positive effect of representing linguistic information using contextualized word embeddings extracted with BERT compared to using standard word embeddings like those extracted with Glove. We also showed, in agreement with previous works, that the fusion of audio- and text-based information leads to significant improvements of approximately 16% on both datasets relative to using the best single modality. To our knowledge, these are the first published results using linguistic information on MSP-PODCAST, a very large, naturalistic and challenging emotion dataset.

Several fusion strategies were tested, including early and late fusion using different training procedures. Results were not significantly different for the different methods, which again agrees with previous observations in the literature.

As an additional contribution, we highlighted the importance and impact of how folds are defined for the IEMOCAP dataset, showing how the standard procedure of splitting by session leads to highly optimistic results on our text-based system. We hope that our proposed criteria, which avoids repeating scripted dialogues between training and test sets, or the alternative of discarding scripted dialogues, will be adopted in future works on the IEMOCAP dataset, specially for text-based systems.

6. REFERENCES

- [1] Laurence Devillers, Laurence Vidrascu, and Lori Lamel, “Challenges in real-life emotion annotation and machine learning based detection,” *Neural Networks*, vol. 18, no. 4, pp. 407–422, May 2005.
- [2] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos, “Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011,” *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, Feb. 2015.
- [3] W. Q. Zheng, J. S. Yu, and Y. X. Zou, “An experimental study of speech emotion recognition based on deep convolutional neural networks,” in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, Sept. 2015, pp. 827–831.
- [4] Aharon Satt, Shai Rozenberg, and Ron Hoory, “Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms,” in *Interspeech*, Aug. 2017, pp. 1089–1093.
- [5] Q. Jin, C. Li, S. Chen, and H. Wu, “Speech emotion recognition with acoustic and lexical features,” in *ICASSP*, April 2015, pp. 4749–4753.
- [6] Ze-Jing Chuang and Chung-Hsien Wu, “Multi-modal emotion recognition from speech and text,” in *International Journal of Computational Linguistics & Chinese Language Processing: Special Issue on New Trends of Speech and Language Processing*, August 2004, vol. 9, pp. 45–62.
- [7] Kalani Wataraka Gamage, Vidhyasaharan Sethu, and Eliathamby Ambikairajah, “Salience based lexical features for emotion recognition,” in *ICASSP*, 2017, pp. 5830–5834.
- [8] B. Schuller, G. Rigoll, and M. Lang, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture,” in *ICASSP*, May 2004, vol. 1, pp. I–577.
- [9] Albert Haque, Michelle Guo, Prateek Verma, and Li Fei-Fei, “Audio-linguistic embeddings for spoken sentences,” in *ICASSP*, 2019, pp. 7355–7359.
- [10] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, “Multi-modal speech emotion recognition using audio and text,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 112–118.
- [11] Jilt Sebastian and Piero Pierucci, “Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts,” in *Interspeech 2019*, Sept. 2019, pp. 51–55.
- [12] Saurabh Sahu, Vikramjit Mitra, Nadee Seneviratne, and Carol Espy-Wilson, “Multi-Modal Learning for Speech Emotion Recognition: An Analysis and Comparison of ASR Outputs with Ground Truth Transcription,” in *Interspeech*, Sept. 2019, pp. 3302–3306.
- [13] Biqiao Zhang, Soheil Khorram, and Emily Mower Provost, “Exploiting Acoustic and Lexical Properties of Phonemes to Recognize Valence from Speech,” in *ICASSP*, May 2019, pp. 5871–5875.
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT, Minneapolis, USA*, 2019, pp. 4171–4186.
- [15] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, “Glove: Global vectors for word representation,” in *EMNLP*, 2014.
- [16] Zhongkai Sun, Prathusha K. Sarma, William Sethares, and Erik P. Bucy, “Multi-Modal Sentiment Analysis Using Deep Canonical Correlation Analysis,” in *Interspeech*, Sept. 2019, pp. 1323–1327.
- [17] Carlos Busso, 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*, vol. 42, no. 4, pp. 335, Nov 2008.
- [18] Reza Lotfian and Carlos Busso, “Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings,” *IEEE Transactions on Affective Computing*, vol. PP, pp. 1–1, 08 2017.
- [19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [20] Florian Eyben, Martin Wöllmer, and Björn Schuller, “Opensmile: The munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM International Conference on Multimedia*, New York, NY, USA, 2010, pp. 1459–1462.
- [21] Xavier Glorot and Yoshua Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS’10)*, 2010.
- [22] Haytham M. Fayek, Margaret Lech, and Lawrence Cavedon, “Evaluating deep learning architectures for Speech Emotion Recognition,” *Neural Networks*, vol. 92, pp. 60–68, 2017.
- [23] Pablo Riera, Luciana Ferrer, Agustín Gravano, and Lara Gauder, “No sample left behind: Towards a comprehensive evaluation of speech emotion recognition system,” in *Proc. Workshop on Speech, Music and Mind 2019*, 2019.
- [24] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference for Learning Representations*, 2015.
- [25] François Chollet et al., “Keras,” <https://keras.io>, 2015.
- [26] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li, “Learning alignment for multimodal emotion recognition from speech,” in *Interspeech*, 2019, pp. 3569–3573.
- [27] Efthymios Georgiou, Charilaos Papaioannou, and Alexandros Potamianos, “Deep Hierarchical Fusion with Application in Sentiment Analysis,” in *Interspeech 2019*, Sept. 2019, pp. 1646–1650.