INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Dual Memory Fusion for Multimodal Speech Emotion Recognition

Darshana Priyasad, Tharindu Fernando, Sridha Sridharan, Simon Denman, Clinton Fookes

Signal Processing, AI and Vision Technologies (SAIVT), Queensland Univeristy of Technology,
Brisbane, Australia

{dp.don, t.warnakulasuriya, s.sridharan, s.denman, c.fookes}@qut.edu.au

Abstract

Deep learning has been widely used in multi-modal Speech Emotion Recognition (SER) to learn sentiment-related features by aggregating representations from multiple modes. However, most SOTA methods use attentive fusion or late fusion of data which ignores the possibility of long-term dependencies among data. In this study, we propose a transformer-based SER architecture that fuses modality representations through explicit memory modules, where the information from current inputs is integrated with historical information allowing the model to understand the relative importance of modes over time. We have used Wav2Vec2 and BERT models to extract audio and text features which are then fused together by aggregating features from individual modes with information stored in memory, followed by downstream classification. Following state-of-the-art methods, we evaluate our proposed method on the IEMOCAP dataset and results indicate that memory-based fusion can achieve substantial improvements.

Index Terms: emotion recognition, memory networks, multi-modal fusion, multi-task learning

1. Introduction

Over the past several years, Speech Emotion Recognition (SER) has evolved into an integral component of Human-Computer Interaction (HCI) systems, facilitating natural interactions with machines [1, 2, 3]. Human intentions are often expressed through verbal and non-verbal cues [4], which are captured by SER and HCI systems to precisely identify emotions, and trigger accurate feedback. However, the characteristics of human emotional responses vary from one person to the next due to personality and the level of stimuli [5], which makes subject-independent emotion analysis challenging. Furthermore, human emotional responses are often expressed through multiple traits such as speech, facial attributes and spoken language, thus multimodal approaches are also commonly used along with uni-modal approaches in intelligent systems.

Existing SER and HCI systems have employed deep learning over conventional machine learning approaches due to deep learning's robustness, its ability to self-engineer features, and its superior performance [6, 7, 8]. However, deep learning requires larger datasets to achieve generalizability, which makes emotion recognition a challenge due to the smaller datasets available [9], resulting in the widespread use of transfer learning. The majority of recent applications use CNNs (i.e SincNet [10, 11]) and transformers (i.e. Wav2Vec2 [12, 13]) to learn robust-features for acoustic emotion recognition from raw waveforms. Several applications have aggregated multi-level acoustic information captured through spectrograms and MFCCs together with the unprocessed audio to exploit complementary information to improve performance [14]. However, audio transformers have revolutionized SER as it eliminates the need for recurrent connections and convolutions, simultaneously yielding better representations and higher performance [7, 13]. Similarly, NLP-based transformer models such as BERT [15] are widely used to capture language representations from transcripts for emotion classification [16, 17, 18, 19].

Compared to uni-modal approaches discussed above, multimodal deep learning aggregates emotion traits from multiple sources together to improve overall performance by exploiting complementary information in heterogeneous data [11, 20, 21]. However, fusion doesn't guarantee increased performance, and thus needs to be carefully designed. Different fusion strategies including early and late fusion are often incorporated in literature to enhance the performance of SER systems [16, 22, 23]. Attention is often used along with multi-modal fusion to direct the network towards salient features that maximise the overall objective in a flexible manner [11, 14], and different variations of attention such as self-attention [24], co-attention [14] and cross-attention [11] have been widely used in SER. The notion of historical context in multi-modal fusion has also been applied to learn relationships and dependencies among data modalities [25, 26]. However, most of these methods (i.e. LSTMs) fail to capture long-term dependencies among modalities, and as a solution explicit memory networks can be incorporated in fusion [27, 28, 29].

Even though significant advances have been made in multimodal SER, the applicability of memory networks for modality fusion is not well explored [30]. We argue that the performance of SER systems can be improved by capturing relationships and dependencies among training data which can be incorporated during inference. In this paper, we present a novel memory-based fusion approach for SER (with audio and text modalities) which effectively stores historic information in explicit memory, and aggregates historic knowledge with current input data to improve recognition performance. Using both audio speech and the correspondingly transcribed text for emotion recognition has been investigated by several researchers [6, 16] since these two modalities carry complementary information. Experiments were conducted on the IEMOCAP dataset [9] to enable fair comparisons with state-of-the-art methods and substantial performance improvement is achieved in terms of recognition accuracy.

2. Methodology

In this paper, we present a novel fusion framework for multimodal SER using memory networks to boost performance over conventional naive fusion. The proposed fusion mechanism is capable of learning long-term dependencies and relationships among data during training, and this knowledge is used to im-

10.21437/Interspeech.2023-1090

prove performance during inference. Figure 1 illustrates the high-level architecture of the proposed model. The network takes two inputs, an audio sample and the corresponding text, which are passed through separate transfer-learned encoders to obtain low-dimensional embeddings. The resultant embeddings are passed through the proposed memory fusion module, and the output of the memory is fed to the classification head. A detailed description of the architecture and its configuration is given in Sections 2.1, 2.2 and 2.3.

2.1. Transformer-based Encoder Network

In the proposed approach, we have used audio and text modalities for SER. Due to the limited availability of large public datasets for emotion recognition, we have used transformer networks and transfer learning to extract robust latent features from input data, where a Wav2Vec2 [12] model with the SUPERB configuration [31] is used as the audio encoder (pre-trained for speaker id), and a BERT [15] model is used as the text encoder. In both transformer architectures, we have used the sequence classification variant of both architectures. First, we preprocess each audio sample and its corresponding text sample using the feature extractor and tokenizer associated with each encoder. We have selected 5s long audio segments sampled at 16000 Hz from the beginning of the audio sample, and 30 tokens from each text sample as inputs to the encoders. Due to the higher computational cost associated with transformer networks, we have used the "base" architecture that contains only 12 stacked transformers for both Wav2Vec2 and BERT. The encoder networks were trained (fine-tuned) end-to-end with the memory module and the classification heads, keeping only the "feature_extractor" (Wav2Vec2) and "embedding" (BERT) layers frozen.

2.2. Multi-Modal Neural Memory Fusion

The proposed neural memory-based fusion layer consists of two explicit memory blocks ($M_a \in \mathbb{R}^{n_a \times d_a}$ and $M_t \in \mathbb{R}^{n_t \times d_t}$ where n and d represents the number of memory slots and memory dimension that capture audio and text memory respectively) and three sub-modules termed the reader, composer and writer, which carry out memory operations as illustrated in Figure 2. The number of memory slots in both memory modules was fixed at 120 (n_a=n_t=120), and the memory dimension was set at 768 (d_a=d_t=768) which is the output dimension of the "base" configuration of the encoder transformers. During initialization, each memory is filled with values from a Xavier normal distribution with a gain of 1. The proposed memory layer takes two inputs, the audio embedding (x_a) and the text embedding (x_t), and it outputs the fused output (x_f) and the transformed memory states.

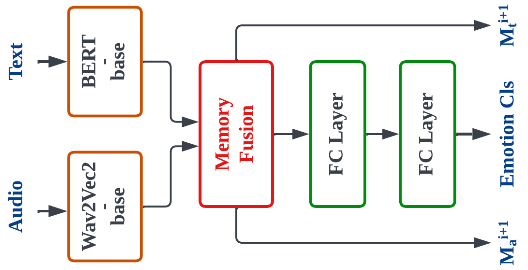

Figure 1: *High-level system architecture. The input audio sample and the corresponding text are passed through two separate encoder networks to generate low-dimensional embeddings, which are then passed through the proposed fusion module. The fusion module outputs updated memory states and the composed output (after fusion), and the latter is then passed through an MLP followed by the classification.*

The embedding vectors from encoders (x_a and x_t) are first passed through the reader sub-module along with the memory blocks (M_a^s and M_t^s where s represents the state at a training iteration) to generate associated memory vectors, which are later used by the composition sub-module to compose the fusion output. However, contrary to conventional memory networks [27, 28, 32], we use an embedding from one modality to obtain a complementary and semantically related memory representation of the other modality through a cross-attention mechanism.

First, we pass x_a and M_t^s through the reader sub-module, and these are multiplied together to generate the vector x_{tmp} (x_{tmp}= $x_a \times M_t^s$). Unlike conventional self-attention where a softmax (or sigmoid) activation is used to calculate the attention weights, we use a Gumbel softmax activation [33] due to its ability to manipulate the distribution of attention weights ranging from a one-hot to uniform weighting, by adjusting the temperature (τ). We calculate τ_t by passing x_{tmp} through a fully connected layer (with $W_a \in \mathbb{R}^{1 \times d_a}$ and b_a as weight and bias respectively) which is trained along with the network. We limit τ_t to be between $1e^{-4}$ and τ_{max} (fixed at 10) by taking the absolute value and clamping. The resultant temperature value is used to calculate the Gumbel-softmax weights, which are then multiplied with x_{tmp} to get the text memory representation from the audio embeddings (x_{a_t}) as per Equations 1 and 2. Similar steps are followed to get the audio memory representation from text embeddings (x_{t_a}) as shown in Figure 2.

$$\tau_t = min(max(|x_{tmp} \times W_a^T + b_a|, 1e^{-4}), \tau_{max}) \qquad (1)$$

$$x_{a\_t} = \sum_{i=1}^{n_a} M_a^s(i).\frac{exp((log(x_{tmp}(i)) + g_i)/\tau_t}{\sum\limits_{l=1}^{n_a} exp((log(x_{tmp}(l)) + g_l)/\tau_t} \qquad (2)$$

The resultant feature vectors, x_{t_a} and x_{a_t}, along with the encoded embeddings x_a and x_t, are then passed through the composer sub-module where fusion occurs. We use a shared Compact Bilinear Pooling layer (CBP) [34, 35], which offers the discriminating abilities of bilinear pooling with fewer parameters, to fuse the embeddings of one modality with the memory representation from the other modality, as shown by Equation 3. First, we pass x_a and x_{a_t} through the CBP layer followed by x_{t_a} and x_t (the order of audio and text is preserved when passing through the CBP). The pooled outputs, x_{p_1} and x_{p_2}, represent fused feature vectors with different information. Each pooled vector is passed through a separate self-attention layer (to identify salient information) with batch normalization. Finally, the fusion output (x_f) is obtained by averaging the feature outputs of the two normalization layers as shown in 2.

$$x_{p_1} = CBP(x_a, x_{a\_t}) \ \& \ x_{p_2} = CBP(x_{t\_a}, x_t) \qquad (3)$$

In the proposed approach, the writer sub-module is used to update the memory state based on the corresponding encoder embedding. Unlike the previous sub-modules, the operations in the writer sub-module are carried out by considering each mini-batch of embeddings as one feature as explained below. First, we concatenate the current memory state M_a^s with x_a along the zeroth axis (if the mini-batch dimension is (m, d_a) and the memory block dimension is (n_a, d_a), the resultant feature dimension is $(m$+$n_a, d_a)$), to expand the memory with current modality information. The resultant feature vector is then transformed into a new latent space using a transformer layer which

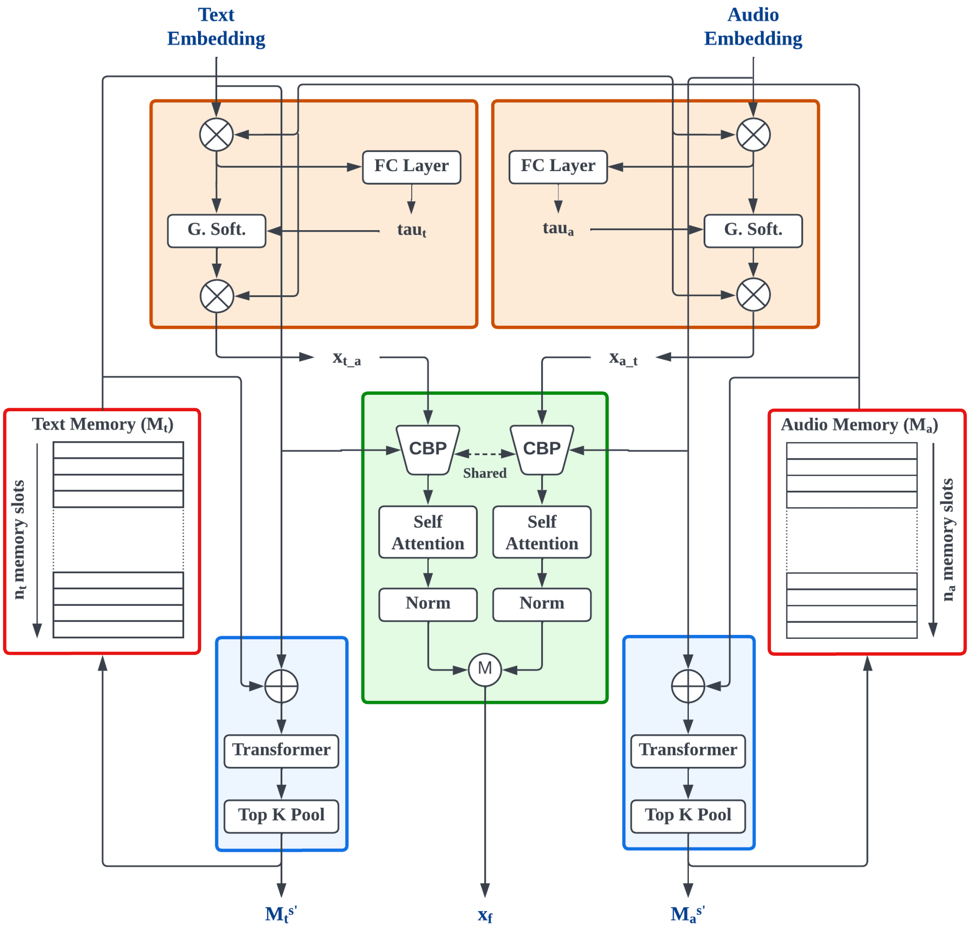

Figure 2: *Proposed memory fusion architecture. The memory architecture contains three sub-modules: reader (brown), composer (green) and writer (blue); and two explicit memory blocks (red), each to learn and store dependencies related to a specific modality. The functionality of the respective sub-modules is discussed in Section 2.2. The memory takes embeddings from audio and text encoders as inputs and outputs the fused feature vector along with new memory states, which is later used to calculate the overall loss.*

learns relationships and dependencies among each vector in the expanded memory. Since, the memory stores information from past iterations (during training), this may learn long-term relationships among data. However, to get the number of memory slots back to n_a, we use the differentiable Top-K pooling layer (k is set to n_a) which weights and selects the k most salient feature vectors from the expanded memory. The selected k memory slots are then used as the new memory state ($M_a^{s'}$). As shown in Figure 2, the same operations are carried out for the text modality, resulting in the new memory state ($M_t^{s'}$). Even though the new memory state is calculated during the forward pass of the training, we assign it back to the memory variable only at the beginning of the next iteration to avoid any instabilities during back-propagation. Finally, these calculated new memory states are used to calculate the overall loss during the training process (see Section 2.3). However, the memory update process is not carried out during inference to avoid the memory storing any information from the evaluation set, and to ensure a fair evaluation.

2.3. Classification Network and Objective Functions

The fused feature x_f is passed through the classification network which consists of two fully connected layers with 1024 and 512 units followed by the classification head resulting in log-probability vector $x_o \in \mathbb{R}^{N\times 4}$, where N refers to mini-batch size. To train the writer sub-module end-to-end along with the complete network, we have introduced a pairwise similarity loss which seeks to achieve orthogonality among vectors in the memory block while increasing the informativeness of the stored features. First, for $M_a^{s'}$, a similarity matrix $S_a \in \mathbb{R}^{n_a \times n_a}$ is obtained where $S_a(i,j)$ represents the cosine similarity between the i^{th} and j^{th} memory vectors of $M_a^{s'}$. The mean cosine similarity is calculated by averaging the upper triangle of the similarity matrix, excluding the diagonal which is always 1. The loss for $M_t^{s'}$ is calculated similarly. and then added to the categorical cross-entropy loss calculated on x_o, obtaining the total loss (L) as per Equation 4.

$$L = \sum_{l\in\{a,t\}} \frac{\sum\limits_{i=1}^{n_l}\sum\limits_{\substack{j=1\\i<j}}^{n_l} S_l(i,j)}{n_l(n_l-1)} + \frac{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{4} -T_j log(x_o(i,j))}{N} \quad (4)$$

3. Experiments

3.1. Dataset and Experiment Setup

We have evaluated our proposed memory architecture for emotion recognition on the IEMOCAP dataset, containing utterances from 10 unique speakers divided across five recordings sessions. We have selected samples from four basic emo-

Table 1: *Performance comparison with SOTA*

Model	WA	UA
Isolated Gaussian Reg. [36]	69.3%	68.1%
Co-attention based Fusion [16]	69.8%	71.0%
Attentive Time-Frequency NN [1]	73.4%	64.5%
Key-Sparse Transformers [6]	74.3%	75.3%
Modality Calibration [20]	75.6%	77.6%
ResNet + BERT [37]	75.8%	76.1%
Block and Token Attention [3]	73.2%	75.2%
Speechformer++ [7]	70.5%	71.5%
Ours	76.8%	77.3%

tions: anger, sadness, happiness (samples labelled excitement are merged with happiness as per [6]) and neutral; to be consistent with the SOTA. The resultant dataset contains 5,531 utterances with 1,103, 1,636, 1,084 and 1,708 samples representing anger, happiness, sadness and, neutral respectively. We used a Leave-One-Session-Out (LOSO) cross-validation protocol for evaluations following [6]. The model was trained with an SGD optimizer with learning rates of 0.01 (complete network) and 0.001 (memory module). Experiments were conducted on a high-performance computing cluster with an NVidia M40 and T4 GPUs, $30GB$ of memory, and 6 CPU cores.

3.2. Results and Analysis

A comparison of our proposed fusion method with the SOTA are given in Table 1. The performance is evaluated using Weighted Accuracy (WA) and Unweighted Accuracy (UA) for a LOSO cross-validation protocol on the IEMOCAP dataset. The proposed memory-based multi-modal fusion approach has been able to achieve 76.8%, 77.3%, and 0.77 in terms of WA, UA and F1-score respectively, showing a substantial improvement over the SOTA.

Transfer learning of pre-trained transformers has been widely used for feature encoding in SOTA methods [14, 37] due to its ability to learn better features as opposed to training from scratch. However, the task that the transformer is pre-trained on can have a significant impact on the overall performance and generalizability. Contrary to [14], which has used a Wav2Vec2 network pre-trained on speech-to-text translation, we have utilized a Wav2Vec2 variant fine-tuned on speaker identification (SUPERB setting [31]). Thus, the audio encoder can generate embeddings that are subject-independent and carry emotional traits which result in better generalization and performance.

Furthermore, the input data has a significant impact on what is learnt by the network. Spectrogram-based networks can fail to capture specific frequency bands associated with SER, or fail to model emotion-related correlations in the frequency domain during model training [1]. Therefore, compared to pre-computed features such as spectrograms and MFCCs [1, 3, 14], deep networks can learn more robust and informative features which helps lead to the improved performance of the proposed method. In multi-modal SER, fusion can be carried out at different levels (early, intermediate or late), each of which may learn different representations. Contrary to [37] that has used score fusion that promotes independent learning from modalities, we have used intermediate feature fusion that learns a joint representation across both modalities. With this approach, both the encoders learn a rich representation space by mutually exploiting complementary information, resulting in improved performance.

Deep networks learn the features from input samples to maximise the SER performance during training, but some learnt information may not be related to the expressed emotion and may in fact be redundant, or noisy. Therefore, key-sparse transformers [6] have been developed to remove noise and focus on emotion-related information. However, in our proposed method, we use attention within fusion layers to focus on salient information while utilizing the explicit memory blocks to store task-specific information which is not redundant. Self-guided modality calibration methods [20] have been proposed to jointly maintain word-to-sentence dependence and uni-modal independence. Similarly, in our proposed method, we have explicit memory blocks to learn modality-specific long-term dependencies while jointly learning multi-modal interactions through intermediate fusion and the downstream classification network.

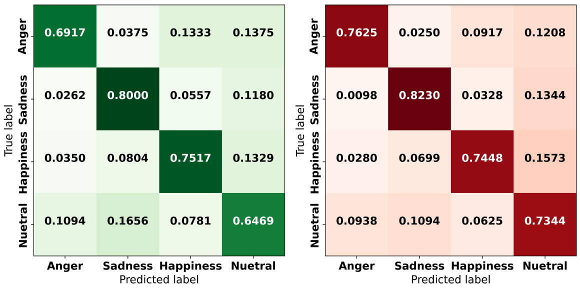

Figure 3: *Confusion matrix for an average performing session (Session 3) with LOSO-CV for Left : Naive Fusion, Right : Memory Fusion*

For a fair analysis of the proposed memory-based fusion architecture, we compared it with a naive fusion model with the same configuration. In naive fusion, we simply concatenate the feature embeddings, x_a and x_t, while keeping the same downstream model and classification head. We achieved an average WA, UA and F1 score of 74.3%, 75.2% and 0.74 respectively which is a 2.5% drop compared to the memory fusion (see Table 1) highlighting the superiority of the proposed memory fusion approach. Figure 3 shows confusion matrices from the same testing split for memory fusion and naive fusion. It is observed that higher performance has been achieved with "anger" and "sadness" which are the least represented emotions in the dataset, highlighting the importance of the proposed memory. The average inference time for one sample is 63 ms and 91 ms for naive and memory fusion respectively. Furthermore, since the fusion layer takes fixed-size inputs irrespective of the input length (a transformer with the base setting always outputs a vector of size $\mathbb{R}^{1\times768}$), the time-complexity remains constant making this layer scalable for longer utterances without needing of additional computational resources.

4. Conclusion

This paper proposes a multi-modal fusion architecture for SER using explicit memory to improve recognition performance. This method hypotheses that substantial improvements can be achieved by learning and incorporating long-term dependencies among multi-modal data along with the inputs. The proposed memory module is capable of capturing and storing relationships which are aggregated together with inputs during training and inference. Furthermore, memory fusion generates a robust and rich latent feature representation which results in substantial performance improvements compared to naive feature fusion. Furthermore, the use of transformers that are pre-trained on a relevant domain, i.e. speaker identification (specific speech traits can affect persons' way of expressing emotions), can also have contributed to the performance improvement over the state-of-the-art in terms of Weighted Accuracy (WA).

5. References

[1] C. Lu, W. Zheng, H. Lian, Y. Zong, C. Tang, S. Li, and Y. Zhao, "Speech Emotion Recognition via an Attentive Time–Frequency Neural Network," *IEEE Transactions on Computational Social Systems*, 2022.

[2] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Affect recognition from scalp-EEG using channel-wise encoder networks coupled with geometric deep learning and multi-channel feature fusion," *Knowledge-Based Systems*, p. 109038, 2022.

[3] J. Lei, X. Zhu, and Y. Wang, "BAT: Block and token self-attention for speech emotion recognition," *Neural Networks*, vol. 156, pp. 67–80, 2022.

[4] R. O. Stanley and G. D. Burrows, "Varieties and functions of human emotion," *Emotions at work: Theory, research and applications in management*, pp. 3–19, 2001.

[5] G. H. Ice, "Measuring emotional and behavioral response," 2007.

[6] W. Chen, X. Xing, X. Xu, J. Yang, and J. Pang, "Key-Sparse Transformer for Multimodal Speech Emotion Recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 6897–6901.

[7] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "SpeechFormer++: A Hierarchical Efficient Framework for Paralinguistic Speech Processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[8] D. Krishna and A. Patil, "Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks," in *Interspeech*, 2020, pp. 4243–4247.

[9] 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, pp. 335–359, 2008.

[10] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT).* IEEE, 2018, pp. 1021–1028.

[11] D. Priyasad, T. Fernando, S. Denman, C. Fookes, and S. Sridharan, "Attention driven fusion for multi-modal emotion recognition," *arXiv preprint arXiv:2009.10991*, 2020.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[13] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 7967–7971.

[14] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech Emotion Recognition with Co-Attention Based Multi-Level Acoustic Information," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 7367–7371.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] Z. Zhao, Y. Wang, and Y. Wang, "Multi-level Fusion of Wav2vec 2.0 and BERT for Multimodal Emotion Recognition," *arXiv preprint arXiv:2207.04697*, 2022.

[17] J. Santoso, T. Yamada, K. Ishizuka, T. Hashimoto, and S. Makino, "Performance improvement of speech emotion recognition by neutral speech detection using autoencoder and intermediate representation," in *Proc. Interspeech*, 2022.

[18] M. Perez, M. Jaiswal, M. Niu, C. Gorrostieta, M. Roddy, K. Taylor, R. Lotfian, J. Kane, and E. M. Provost, "Mind the gap: On the value of silence representations to lexical-based speech emotion recognition," *Proc. Interspeech 2022*, pp. 156–160, 2022.

[19] P. Kumar, V. Kaushik, and B. Raman, "Towards the Explainability of Multimodal Speech Emotion Recognition," in *Interspeech*, 2021, pp. 1748–1752.

[20] M. Hou, Z. Zhang, and G. Lu, "Multi-Modal Emotion Recognition with Self-Guided Modality Calibration," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 4688–4692.

[21] P. Koromilas and T. Giannakopoulos, "Deep multimodal emotion recognition on human speech: A review," *Applied Sciences*, vol. 11, no. 17, p. 7962, 2021.

[22] A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowledge-Based Systems*, vol. 244, p. 108580, 2022.

[23] J. Li, S. Wang, Y. Chao, X. Liu, and H. Meng, "Context-aware Multimodal Fusion for Emotion Recognition," *Proc. Interspeech 2022*, pp. 2013–2017, 2022.

[24] Y. Li, P. Bell, and C. Lai, "Fusing ASR Outputs in Joint Training for Speech Emotion Recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 7362–7366.

[25] R. Beard, R. Das, R. W. Ng, P. K. Gopalakrishnan, L. Eerens, P. Swietojanski, and O. Miksik, "Multi-modal sequence fusion via recursive attention for emotion recognition," in *Proceedings of the 22nd conference on computational natural language learning*, 2018, pp. 251–259.

[26] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream lstm: A deep fusion framework for human action recognition," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV).* IEEE, 2017, pp. 177–186.

[27] J. Weston, S. Chopra, and A. Bordes, "Memory networks," *arXiv preprint arXiv:1410.3916*, 2014.

[28] T. Fernando, S. Denman, A. McFadyen, S. Sridharan, and C. Fookes, "Tree memory networks for modelling long-term temporal dependencies," *Neurocomputing*, vol. 304, pp. 64–81, 2018.

[29] D. Priyasad, A. Partovi, S. Sridharan, M. Kashefpoor, T. Fernando, S. Denman, C. Fookes, J. Tang, and D. Kaye, "Detecting heart failure through voice analysis using self-supervised mode-based memory fusion," in *Proceedings of the 23rd INTERSPEECH Conference*. International Speech Communication Association, 2022, pp. 2848–2852.

[30] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Memory based fusion for multi-modal deep learning," *Information Fusion*, vol. 67, pp. 136–146, 2021.

[31] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "SUPERB: Speech processing Universal PERformance Benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[32] S. Sukhbaatar, J. Weston, R. Fergus *et al.*, "End-to-end memory networks," *Advances in neural information processing systems*, vol. 28, 2015.

[33] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[34] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 317–326.

[35] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[36] C. Fu, C. Liu, C. Ishi, and H. Ishiguro, "An Adversarial Training Based Speech Emotion Classifier with Isolated Gaussian Regularization," *IEEE Transactions on Affective Computing*, 2022.

[37] S. Padi, S. O. Sadjadi, D. Manocha, and R. D. Sriram, "Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models," *arXiv preprint arXiv:2202.08974*, 2022.