Exploring Attention Mechanisms for Multimodal Emotion Recognition in an Emergency Call Center Corpus

Preprint · June 2023

DOI: 10.48550/arXiv.2306.07115

3 authors:

Théo Deschamps-Berger
University of Paris-Saclay
10 PUBLICATIONS **91** CITATIONS
SEE PROFILE

Lori Lamel
French National Centre for Scientific Research
423 PUBLICATIONS **14,642** CITATIONS
SEE PROFILE

Laurence Devillers
Computer Science Laboratory for Mechanics and Engineering Sciences
191 PUBLICATIONS **8,008** CITATIONS
SEE PROFILE

EXPLORING ATTENTION MECHANISMS FOR MULTIMODAL EMOTION RECOGNITION IN AN EMERGENCY CALL CENTER CORPUS

Théo Deschamps-Berger[⋆] *Lori Lamel*[⋆] *Laurence Devillers*[†]

[⋆] LISN - CNRS, Paris-Saclay University, Orsay, France
[†] LISN - CNRS, Sorbonne University, Orsay, France

ABSTRACT

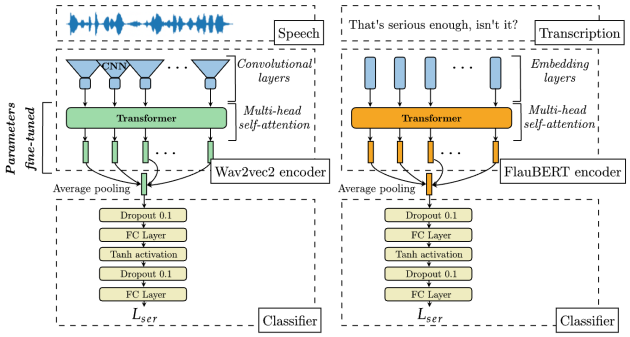
The emotion detection technology to enhance human decision-making is an important research issue for real-world applications, but real-life emotion datasets are relatively rare and small. The experiments conducted in this paper use the CEMO, which was collected in a French emergency call center. Two pre-trained models based on speech and text were fine-tuned for speech emotion recognition. Using pre-trained Transformer encoders mitigates our data's limited and sparse nature. This paper explores the different fusion strategies of these modality-specific models. In particular, fusions with and without cross-attention mechanisms were tested to gather the most relevant information from both the speech and text encoders. We show that multimodal fusion brings an absolute gain of 4-9% with respect to either single modality and that the Symmetric multi-headed cross-attention mechanism performed better than late classical fusion approaches. Our experiments also suggest that for the real-life CEMO corpus, the audio component encodes more emotive information than the textual one.

Index Terms— real-life emotional corpus, emergency call center, speech emotion recognition, Transformer-based encoders, cross-attention fusion, late fusion, model-level fusion

1. INTRODUCTION

In an emergency call center, an agent has to make quick decisions that can have great consequences. Automatic emotion recognition is a key component of human decision makings. The corpus used in this research (CEMO) was collected in a French emergency call center. In application-specific data, emotions are scarce accounting from 10% (banking context) [1] to 30% (emergency call context) of speech turns/segments [2]. In this study we used and fine-tuned two pre-trained Transformer encoders, trained on French data: LeBenchmark wav2vec2 [3] which is based on the wav2vec2 model designed to be trained on audio [4]; and FlauBERT [5], trained on text and based on the BERT model [6]. These models have demonstrated their capabilities to produce relevant representations for several downstream tasks [7, 8]. Transformers encoder-decoder [9] have triggered novel approaches, thanks to the self-attention mechanism [10] and to the cross-attention mechanism [11] that mixes two different embedding sequences which may come from different modalities. The attention mechanism has been used in speech emotion recognition with frame-level speech features [12] and with knowledge transfer [13]. The non-sequential characteristic of the attention mechanism enables Transformer to better capture long dependencies compared to Long-Short Term Memory. Specifically, Transformer encoders have been widely used to create domain-relevant representations in sequence processing. They were naturally first used on Written Language Processing tasks, particularly with language representation models such as BERT [6] designed to serve as a pre-trained core model. It is trained in a self-supervised mode on huge amounts of data with the aim of being able to fine-tune it to specific applications and low-resource domains. In our case, the data are telephone conversations (speech and transcriptions) between call center agents and patients or relatives. In this real data recording context, data exhibiting emotion are rare and often have a wide diversity in the manner it is expressed: sometimes in the voice characteristics (prosody, fluency), in the choice of words, or a combination of both. Multimodal deep learning is well-suited to handling this problem and has garnered much interest for emotion detection [14, 15, 16, 17]: The paralinguistic representation provides low-level emotion cues, and the semantic features bring content and context to the sentence. The flexibility of deep learning systems supports several multi-modal fusion levels: early [18, 19], intermediate [20, 21, 22], and late [23, 24, 25]. In this paper, we focus on late fusion, which introduces the multimodal information in the later layers of the network, allowing the earlier layers to specialize the unimodal pattern learning and extraction. We compare three late fusions strategies: the first is a score average of each specific-modality encoder (Score fusion) [23]; the second is a shared neural network fed by the concatenation of both encoders outputs (Concatenation fusion) [20, 26] and the third is a Symmetric multi-head cross-attention fusion which takes in context the semantic and the paralinguistic representations to weight the final output. A similar multi-head Symmetric cross-attention method was used in a multimodal transformer for acoustics and vision in [27] and for speech emotion recog-

Fig. 1. Parameter fine-tuning of the wav2vec2 and FlauBERT transformer layers. The Convolutional and Embedding layers are frozen. The wav2vec2 and the FlauBERT produce paralinguistic (H_p) and semantic representations (H_s), respectively.



nition with MFCC and GloVe embeddings [28]. This paper extends our previous studies on the adaptation of Transformer encoders for speech emotion recognition and on late multimodal fusions [29, 30]. To the best of our knowledge, the use of Symmetric cross-attention powered by fine-tuned pretrained encoders for speech emotion recognition has not been realized before. The rest of the paper is organized as follows. The next section explains how the encoders are fine-tuned for emotion recognition. Then, we detail the fusion strategies, and present the CEMO corpus and the experimental results.

2. FINE-TUNED PRE-TRAINED ENCODERS

In prior results [30], several pre-trained models for speech emotion recognition were adapted with the CEMO corpus [2]. We selected among the available models the leBenchmark model (Wav2Vec2-FR-3K) [3] trained on 3k hours of French and the FlauBERT model [5] trained on 24 French subcorpora (71Gb of text). This selection is justified by the performances on our CEMO corpus [30]. The wav2vec2-FR-3K model includes in its training database some spontaneous dialogues, telephone-recorded data, and emotional content, which have similar characteristics to the CEMO corpus. The FlauBERT model applies a basic tokenizer for French to the input sentence and embeds it with a vocabulary of 50K subword units built using the Byte Pair Encoding (BPE) algorithm [31]. In order to adapt the wav2vec2-FR-3K and FlauBERT models, we added a classifier on top of them, adapted to our four emotional classes. The multi-head self-attention layers of the Transformer encoders were fine-tuned for speech emotion recognition with the CEMO corpus, as shown in Figure 1, making the assumption that the first layers (Convolutional and Embedding) are reliable for this task [7, 30].

3. MULTIMODAL FUSIONS STRATEGIES

The multimodal fusion variants are applied to the outputs of both encoders, fine-tuned on speech or transcriptions, respectively. The wav2vec2 encoder produces a matrix composed of paralinguistic features per frame window ($H_p \in \mathbb{R}^{d_{\text{frames}} \times d_{\text{model}}}$) while the FlauBERT encoder yields a matrix of semantic features per word ($H_s \in \mathbb{R}^{d_{\text{subwords}} \times d_{\text{model}}}$). For the Base configuration the encoders have $d_{\text{model}} = 768$ and $d_{\text{model}} = 1024$ for the Large size. We tested three methods to extract the most relevant information from both feature spaces.

3.1. Score fusion

The output matrix of each encoder is averaged across the model dimension (mean(H_p) and mean(H_s) $\in \mathbb{R}^{d_{\text{model}}}$). The Score fusion consists of two fully connected parallel layers processing the average outputs of the two encoders.

$$\tilde{\boldsymbol{y}}_p = \sigma(tanh(W_p \times \text{mean}(H_p) + b_p)),$$
$$\tilde{\boldsymbol{y}}_s = \sigma(tanh(W_s \times \text{mean}(H_s) + b_s)),$$

where W_p and $W_s \in \mathbb{R}^{n_{\text{classes}} \times d_{\text{model}}}$. The softmax outputs generated by each classifier are then averaged by class.

3.2. Concatenation fusion

The Concatenation fusion consists of shared fully connected layers fed by the concatenation of the encoder's output matrices and averaged across the model dimension.

$$\boldsymbol{h}_{\text{concat}} = [\text{mean}(H_p), \text{mean}(H_s)]$$
$$\tilde{\mathbf{y}}_c = \sigma(tanh(W_c \times \boldsymbol{h}_{\text{concat}} + b_c))$$

where $W_c \in \mathbb{R}^{n_{\text{classes}} \times d_{\text{model}}}$.

3.3. Symmetric multi-head cross-attention fusion

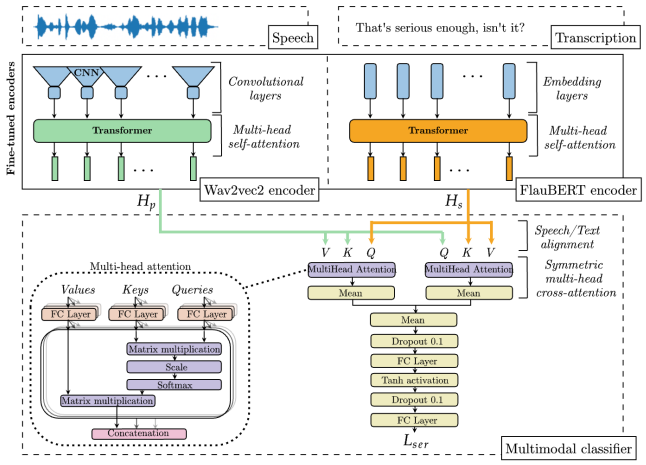
The attention algorithm relies on the idea that each element of an input matrix can be weighted by its importance in the context. To this end, the input matrix (which we call the Key matrix) is multiplied by a context matrix (denoted as Query) and then softmax and scaled to produce an attention matrix. This attention matrix is finally multiplied by the input matrix (called Value) to produce a matrix whose elements are weighted by the context; see equation 1. In self-attention, the context matrix (Q) is equal to the input matrix (K, V); in cross-attention, the context matrix (Q) comes from another modality.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

where d_k is the dimension of queries and keys.
Following Vaswani [9], we used multi-head attention (eq. 2) in order to increase the number of attention mechanisms on different representation subspaces, as shown in Figure 2.

$$\text{MHead}(Q, K, V) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_\text{h}\right) W^O \qquad (2)$$
$$\text{where head } = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

Fig. 2. Multimodal system with Symmetric multi-head cross-attention fusion.



And where the projections are parameter matrices: $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$. and d_v is the dimension of values. In this work, we used 16 attention heads ($h = 16$) for all fusions with Base model encoders and 32 heads ($h = 32$) for Large encoders.

To perform cross-attention, since the dimensions of the encoder outputs for wav2vec2 and FlauBERT are different, we experimented with two alignment methods. The first method is coarsely aligning the two embedding matrices to create equivalent matrices size. To this end, the number of frames in H_p is divided by the number of subwords in H_s, and the aligned frames are averaged over each subword (entry #subwords in Table 4). In the second method, the number of frames associated with a subword depends on the number of characters in it, and in order to give more weight to longer subwords, the frames are summed rather than averaged. This is essentially equivalent to aligning at the character level and summing the frames in each character (entry #characters in Table 4). We experimented with three fusions: Paralinguistic, Semantic, and Symmetric multi-head cross-attention fusions. The Paralinguistic and Semantic attention fusions consist of a multi-head attention with respectively: $(Q = H_s, K = H_p, V = H_p)$ and $(Q = H_p, K = H_s, V = H_s)$.

We propose a Symmetric multi-head cross-attention fusion, obtained by averaging the outputs of the two first multi-head cross-attention, as detailed in Figure 2. We refer to these fusions as: Paralinguistic, Semantic, and Symmetric cross-attention in Table 4.

4. EXPERIMENTAL RESULTS

All experiments were assessed using a subset of the CEMO corpus described in the next subsection. The evaluation is performed on five-fold with a classical cross-speaker folding strategy that is speaker-independent between training, validation, and test sets. During each fold, system training is optimized on the best Unweighted Accuracy (UA) validation set. The outputs of each fold are combined for the final results.

Table 1. Details of the CEMO subset of speech signals and manual transcripts. ANG: Anger, FEA: Fear, NEU: Neutral, POS: Positive, Total: Total number of segments.

CEMO$_s$	**ANG**	**FEA**	**NEU**	**POS**	**Total**
#Speech seg.	1056	1056	1056	1056	4224
#Speakers	149	537	450	544	812
#Dialogues	280	504	425	516	735
Duration (min, s)	39	52	49	20	160
Duration (mean, s)	2.2	2.9	2.8	1.1	2.3
Vocabulary size	1146	1500	1150	505	2600
Avg. word count	9.3	11.9	7.9	3.8	8.2

4.1. Corpus

The speech corpus was collected in a French emergency call center [2]. Data preparation is crucial to achieving good performance and robustness; here, we describe the selection of a balanced subset of the CEMO data used for model training, validation and test, as detailed in Table 1. The selected CEMO subset (2h40) is comprised of 4224 segments, with durations ranging from 0.4 to 7.5 seconds equally distributed over the four main emotion classes: ANGER, FEAR, NEUTRAL and POSITIVE. To this end, the Fear and Neutral classes were subsampled, maintaining 1056 samples for each class for which the annotators agreed while prioritizing a broad representation of speaker diversity. The minority class Anger was completed with segments from the agents. Since the patient Anger class is composed of over 85% fine-grained segments of annoyance and impatience, the selected segments in the agents were selected from the same fine-grained labels. As seen in Table 1, there are fewer speakers for Anger compared to the other classes. It can be noted in the Positive class has the most significant number of speakers and dialogues, potentially being richer and more heterogeneous than the other classes. Manual transcriptions of the CEMO corpus were performed by two coders, with guidelines similar to those proposed in the Amities project [1]. The transcriptions contain about 2499 nonspeech markers, primarily pauses, breath, and other mouth noises. The vocabulary size is 2.6k, with a mean and median of about 10 words per segment (min 1, max 47).

4.2. Results

Table 2 provides the unimodal results, which serve as a baseline for this study. During training, we use cross-entropy loss, Adam optimization [32], with a fixed learning rate of 10^{-5}, and a mini-batch of size 8. For encoder fine-tuning, due to a large number of parameters, we used gradient norm clipping

Table 2. Unimodal results with fine-tuning for speech emotion recognition with the CEMO corpus. Configuration (Config.), Fine-tuned parameters (Fine-tuned p.), UA: %

Models	Config.	Fine-tuned p.	ANG	FEA	NEU	POS	Total
Wav2Vec2-FR-3K [3]	Base	90 M	70.4	70.5	74.8	83.2	74.7
	Large	311 M	71.6	69.8	71.7	88.4	75.4
FlauBERT [5]	Base	85 M	57.9	64.9	74.1	83.5	70.1
	Large	303 M	61.0	60.7	73.2	85.6	70.1

Table 3. Multimodal fusions comparative experiments. Configuration (Config.), Trainable parameters (Train p.), UA: %

Fusion	Config.	Train. p.	ANG	FEA	NEU	POS	Total
Score	Base	2 M	69.9	73.8	75.7	84.9	76.1
	Large	4 M	79.7	70.4	77.0	89.0	79.0
Concatenation	Base	2 M	70.5	73.8	75.7	85.0	76.3
	Large	4 M	78.7	69.4	77.2	88.3	78.4

of 1, to facilitate learning stability, improving our prior results for Large encoders [30].

Both sized audio encoder configurations give an absolute gain of 4-5% with respect to the fine-tuned FlauBERT encoders, (c.f. Table 2). This result suggests that, for our real-life CEMO corpus, the paralinguistic model encodes more emotive information than the textual one. The baseline fusion methods, Multimodal Score and Concatenation, with both encoder sizes (Base and Large) outperform the single modality encoders (compare Tables 2 and 3). In Table 4, we explored the cross-attention mechanisms with three attention-based fusion strategies for the last fusion. The Paralinguistic cross-attention fusion almost always performed better than the Semantic fusion with both alignment strategies. The better results of the unimodal wav2vec2 compared to the FlauBERT encoder show that the paralinguistic encoder produces relevant features for speech emotion recognition. In addition to including acoustic information, the wav2vec2 encoder also implicitly includes lexical information. The encoder size influences the multimodal fusion performance. The Base and Large unimodal fine-tuned encoders have similar results for the Paralinguistic encoders (around 0.7% difference in Table 2), and the same results with the Semantic encoders. However, the Large configuration yielded a 2.9% improvement over the Base model for the Symmetric cross-attention fusion method with alignment based on subwords, as shown in Table 4. For a given condition (fusion method and model size), the #subwords alignment method generally performs slightly better than the #characters alignment.

Table 4. Multimodal multi-head cross-attention fusion comparison. Configuration (Config.), Trainable parameters (Train p.), UA: %

Fusion	Alignment	Config.	Train. p.	ANG	FEA	NEU	POS	Total
Paralinguistic cross-attention	#subwords	Base	2 M	69.6	74.5	74.1	84.6	75.7
		Large	4 M	80.2	71.2	74.2	87.3	78.2
	#characters	Base	2 M	69.5	74.1	74.1	84.1	75.5
		Large	4 M	79.1	73.4	74.2	87.1	78.5
Semantic cross-attention	#subwords	Base	2 M	69.8	71.0	74.1	86.8	75.5
		Large	4 M	78.3	71.1	75.9	87.9	78.3
	#characters	Base	2 M	69.2	69.7	74.3	85.8	74.8
		Large	4 M	77.7	69.7	73.8	86.8	77.0
Symmetric cross-attention	#subwords	Base	5 M	70.2	75.2	75.5	84.8	76.4
		Large	8 M	81.9	70.5	77.0	87.8	79.3
	#characters	Base	5 M	68.6	74.5	75.9	84.8	75.9
		Large	**8 M**	**82.4**	**71.5**	**75.6**	**87.9**	**79.4**

The CEMO corpus contains data samples from 812 speakers, expressive voices, and nonspeech markers comprising sentences' most frequent emotional cues. The corpus has a vocabulary of only 2600 different words, of which only a few predict emotions. In particular, Fear and Anger are poorly detected by the Semantic models. These patterns make the fusion more complex. The Symmetric cross-attention fusion outperformed single modality encoders with both alignment and encoder sizes and performed slightly better than the classical fusion Score and Concatenation fusions. The best performance of the Symmetric cross-attention multimodal fusion is 79.3% UA compared with 74.5% UA (unimodal fine-tuned wav2vec2) or 70.1% UA (unimodal fine-tuned FlauBERT).

5. CONCLUSION

This paper explored different fusion mechanisms in a multimodal context with real-life recordings from a French emergency call center. In this application, the agent aims to gather as much objective information about the situation as possible to make an informed decision. At the same time, the patient is more concerned with explaining the emergency, their pain, or their anxiety. These characteristics make the detection of emotion from speech very complex, as the emotion cues can be present in audio and/or lexical information. The multimodal fusions were powered by unimodal pre-trained encoders that were fine-tuned for speech emotion recognition using the CEMO corpus. Three primary multimodal fusions were tested; Score, Concatenation, and Symmetric cross-attention. The best emotion recognition results for CEMO were obtained using a Symmetric fusion associated with an adequate alignment. Further experiments will be carried out using transcriptions produced by an automatic speech recognizer instead of manual references.

Ethics and reproducibility

The use of the CEMO database or any subsets of it, carefully respected ethical conventions and agreements ensuring the anonymity of the callers. All the experiments were carried out using Pytorch on GPUs (Tesla V100 with 32 Gbytes of RAM). To ensure the reproducibility of the runs, we set a random seed to 0 and prevent our system from using non-deterministic algorithms. This work was performed using HPC resources from GENCI–IDRIS (Grant 2022-AD011011844R1).

6. REFERENCES

[1] H. Hardy, K. Baker, L. Devillers, et al., "Multi-layer Dialogue Annotation for Automated Multilingual Customer Service," *ISLE*, 2003.

[2] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural networks: INNS*, 2005.

[3] S. Evain, H. Nguyen, H. Le, et al., "LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech," in *INTERSPEECH*, 2021.

[4] A. Baevski, Y. Zhou, A. Mohamed, et al., "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Inform. Process. Systems*, 2020.

[5] H. Le, L. Vial, J. Frej, et al., "FlauBERT: Unsupervised Lang. Model Pre-training for French," in *Twelfth Lang. Resources and Evaluation Conf.*, May 2020.

[6] J. Devlin, M.-W. Chang, K. Lee, et al., "BERT: Pre-training of Deep Bidirectional Transformers for Lang. Understanding," in *NAACL: Human Lang. Technologies*, 2019.

[7] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, et al., *Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap*, 2022.

[8] F. Acheampong, H. Nunoo-Mensah, and W. Chen, *Transformer Models for Text-based Emotion Detection: A Review of BERT-based Approaches*, *AI Review*, 2021.

[9] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is All you Need," in *Advances in Neural Inform. Process. Systems*, 2017.

[10] J. Cheng, L. Dong, and M. Lapata, "Long Short-Term Memory-Networks for Machine Reading," in *Proc. of Empirical Methods in Natural Lang. Process.*, 2016.

[11] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine trans. by jointly learning to align and translate," in *ICLR*, 2015.

[12] Y. Xie, R. Liang, Z. Liang, et al., "Speech Emotion Classification Using Attention-Based LSTM," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.

[13] Z. Zhao, Z. Bao, Z. Zhang, et al., "Self-attention transfer networks for speech emotion recognition," *Virtual Reality & Intelligent Hardware*, 2021.

[14] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[15] P. Tzirakis, A. Nguyen, S. Zafeiriou, and B. W. Schuller, "Speech Emotion Recognition Using Semantic Inform.," in *ICASSP*, 2021.

[16] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-Attention for Speech Emotion Recognition," in *INTERSPEECH*, 2019, pp. 2578–2582.

[17] J. Han, Z. Zhang, Z. Ren, and B. Schuller, "EmoBed: Strengthening Monomodal Emotion Recognition via Training with Crossmodal Emotion Embeddings," *IEEE Trans. on Affect. Comput.*, 2021.

[18] M. Wimmer, B. Schuller, D. Arsic, et al., *Low-Level Fusion of Audio, Video Feature for Multi-Modal Emotion Recognition.*, 2008.

[19] C. Busso, S. Yildirim, M. Bulut, et al., *Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Inform.*, 2004.

[20] J. Han, Z. Zhang, N. Cummins, et al., "Strength Modelling for Real-World Automatic Continuous Affect Recognition from Audiovisual Signals," *Image and Vision Computing*, pp. 76–86, 2017.

[21] J.-B. Delbrouck, N. Tits, and S. Dupont, "Modulated Fusion using Transformer for Linguistic-Acoustic Emotion Recognition," in *Proc. of the First International Workshop on Natural Lang. Process. Beyond Text*, 2020, pp. 1–10.

[22] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, et al., "Efficient Low-rank Multimodal Fusion With Modality-Specific Factors," in *Proc. of the 56th ACL*, 2018, pp. 2247–2256.

[23] Z. Zeng, M. Pantic, and G. Roisman, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE transactions on pattern analysis and machine intelligence*, *IEEE Trans. on Pattern Analysis and Machine Int.*, 2009.

[24] N.-H. Ho, H.-J. Yang, S.-H. Kim, et al., "Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network," *IEEE Access*, 2020.

[25] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Inform. Fusion*, 2021.

[26] J. Han, Z. Zhang, F. Ringeval, et al., "Prediction-based learning for continuous emotion recognition in speech," in *IEEE ICASSP*, 2017.

[27] A. Nagrani, S. Yang, A. Arnab, et al., "Attention Bottlenecks for Multimodal Fusion," in *Advances in Neural Inform. Process. Systems*, 2021.

[28] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal Cross- and Self-Attention Network for Speech Emotion Recognition," in *IEEE ICASSP*, 2021.

[29] T. Deschamps-Berger, L. Lamel, and L. Devillers, "End-to-End Speech Emotion Recognition: Challenges of Real-Life Emergency Call Centers Data Recordings," in *ACII*, 2021.

[30] T. Deschamps-Berger, L. Lamel, and L. Devillers, "Investigating Transformer Encoders and Fusion Strategies for Speech Emotion Recognition in Emergency Call Center Conversations.," in *ICMI*, 2022.

[31] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Trans. of Rare Words with Subword Units," in *Proc. of the 54th ACL*, 2016, pp. 1715–1725.

[32] D. P. Kingma and L. J. Ba, "Adam: A Method for Stochastic Optimization," *ICLR*, 2015.