Enhancing Resilience to Missing Data in Audio-Text Emotion Recognition with Multi-Scale Chunk Regularization

Wei-Cheng Lin
University of Texas at Dallas
Richardson, Texas, USA
wei-cheng.lin@utdallas.edu

Lucas Goncalves
University of Texas at Dallas
Richardson, Texas, USA
goncalves@utdallas.edu

Carlos Busso
University of Texas at Dallas
Richardson, Texas, USA
busso@utdallas.edu

ABSTRACT

Most existing audio-text emotion recognition studies have focused on the computational modeling aspects, including strategies for fusing the modalities. An area that has received less attention is understanding the role of proper temporal synchronization between the modalities in the model performance. This study presents a transformer-based model designed with a *word-chunk* concept, which offers an ideal framework to explore different strategies to align text and speech. The approach creates chunks with alternative alignment strategies with different levels of dependency on the underlying lexical boundaries. A key contribution of this study is the multi-scale chunk alignment strategy, which generates random alignments to create the chunks without considering lexical boundaries. For every epoch, the approach generates a different alignment for each sentence, serving as an effective regularization method for temporal dependency. Our experimental results based on the MSP-Podcast corpus indicate that providing precise temporal alignment information to create the audio-text chunks does not improve the performance of the system. The attention mechanisms in the transformer-based approach are able to compensate for imperfect synchronization between the modalities. However, using exact lexical boundaries makes the system highly vulnerable to missing modalities. In contrast, the model trained with the proposed multi-scale chunk regularization strategy using random alignment can significantly increase its robustness against missing data and remain effective, even under a single audio-only emotion recognition task. The code is available at: https://github.com/winston-lin-wei-cheng/MultiScale-Chunk-Regularization

CCS CONCEPTS

• **Computing methodologies** → **Modeling methodologies**.

KEYWORDS

multimodal emotion recognition, robust modeling

ICMI '23, October 09–13, 2023, Paris, France

ACM ISBN 978-1-4503-XXXX-X/18/06.
https://doi.org/10.1145/3577190.3614110

ACM Reference Format:
Wei-Cheng Lin, Lucas Goncalves, and Carlos Busso. 2023. Enhancing Resilience to Missing Data in Audio-Text Emotion Recognition with Multi-Scale Chunk Regularization. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3577190.3614110

1 INTRODUCTION

Automatic emotion recognition is a crucial technique that can significantly enhance the user experience in *human-computer interaction* (HCI) [19] and facilitate healthcare applications such as the detection of mental disorders (e.g., Autism [33]). Considering the multimodal nature of the externalization of human emotions, previous studies have proposed multimodal emotion recognition solutions to improve performance over other unimodal methods [9, 13, 14]. With the rapid development of *deep learning* (DL), many frameworks have been proposed in recent years focusing on the fusion of modalities, including speech, text, face, gestures, and physiological signals [12, 20, 49, 51]. A research question that remains less explored is how important is to provide precise temporal synchronization between the modalities. Is it essential to perfectly align the speech frames co-occurring with a given word to improve an audio-text recognition system?

Studies have shown a strong intercorrelation between modalities while developing multimodal emotion recognition systems [5, 6]. Currently, the most popular approach to combine multimodal signals in affective computing is to fuse their information at the model level using DL frameworks [15, 39, 52]. Feature representations are individually obtained from each modality, which are later combined using DL strategies. Given that the modalities have different sampling rates, processing frame-level features will result in loosely synchronized streams. These approaches implicitly assume that the DL models can deal with the temporal synchronization across different modalities, without any specific alignment information injected into the model as input (e.g., word-level alignments between speech frames and spoken words). Current models typically rely on attention mechanism modules such as the cross-modal attention layers [11, 12, 17, 30, 45] to achieve synchronization. Is there any benefit in explicitly providing to the model the actual timing synchronization across modalities? Is the attention mechanism enough to compensate for the mismatch in temporal synchronization across the modalities? Relying on a simple feature-level concatenation (i.e., early fusion [4, 15, 53]) between the feature maps is not appropriate to resolve these questions.

This study proposes a novel transformer-based framework that provides the flexibility to model the alignment level between the input of multimodal signals using a *word-chunk* concept. The approach is implemented for bimodal emotion recognition systems

trained with text and speech. The core idea is to introduce an additional pre-segmentation module for the audio modality, which pre-splits the speech frames into data chunks aligned to their corresponding spoken words (e.g., using the word-level boundaries obtained with forced alignment). After this step, we apply *temporal pyramid pooling* [16, 43] to obtain the chunk-level audio feature map that is temporal aligned with the corresponding text embeddings. The text and aligned chunk-level audio representation are used as the input of the main transformer model, which adopts the cross-modal attention fusion technique [45]. The pre-segmentation module offers the flexibility to determine how important the audio-text alignment is for multimodal emotion recognition systems. We compare three alignment strategies by varying the dependency on the exact lexical boundaries in the segmentation: strict, partial, or random alignment. The strict alignment strategy uses the word-level forced alignment results to split data into chunks (i.e., start and end timestamps of each word). The partial alignment strategy segments the data with a fixed window size, following the temporal order of the data. The random alignment strategy is a major contribution in this study. It splits the data chunks using randomly sampled window sizes that do not depend on the lexical boundaries. Importantly, a random alignment is created for each sentence at every epoch, providing different views of the data during the training. This multi-scale chunk strategy serves as a temporal regularization, which could be expected to increase the model robustness against missing data. The proposed framework with the three alignment strategies enables us to investigate the role of temporal synchronization in audio-text emotion recognition.

Our experimental results based on the MSP-Podcast corpus [29] indicate that strict temporal synchronization in audio-text emotion recognition plays a minor role in performance. This result suggests that attention mechanisms are effective in compensating for the lack of perfect alignment between the modalities. Interestingly, we observe that the model trained with strictly aligned information becomes extremely vulnerable to missing partial information from the multimodal features. The strong temporal dependency between modalities exacerbates the major drawback of multimodal modeling, which struggles to deal with missing modalities (Sec. 2.2). In contrast, the model trained with random alignments shows a multi-scale chunk regularization characteristic, which naturally restricts the model from learning heavy temporal dependency across modalities. The proposed multi-scale chunk regularization strategy offers a cheap and built-in solution to increase the robustness to missing data of a multimodal model, which neither adds extra model complexity nor requires other compatible systems. Furthermore, the approach remains effective even under an audio-only modeling scheme, where the proposed approach can simply serve as a data augmentation strategy to improve model performance. In summary, the main contributions of this study are:

- We propose a word-chunk transformer framework which can model the alignment level between audio and text data, facilitating a better understanding of the role of their temporal synchronization.
- We propose a multi-scale chunk regularization with random alignments between audio and text, which can significantly increase the model robustness against missing partial data from the modalities.

2 RELATED WORKS

This study focuses on multimodal emotion recognition systems (Sec. 2.1). Our analysis demonstrates that the different strategies for combining audio and text can affect the system's robustness against missing data. Therefore, this section also discusses methods designed to improve robustness of multimodal emotion recognition systems against missing data (Sec. 2.2).

2.1 Multimodal Emotion Recognition

The core problem in multimodal emotion recognition is determining the optimal approach to leverage information from different modalities. An effective modeling framework can extract meaningful inter- and intra-modality relations from multimodal signals, improving the emotion recognition performance [2]. Traditional approaches rely on simple feature-level or decision-level fusion techniques [4, 15, 47]. The feature-level fusion, referred to as early fusion, pre-concatenates handcraft sentence feature vectors from different modalities as the model input, while the decision-level fusion, referred to as late fusion, produces emotion decisions based on the majority voting result from modality-specific classifiers [42, 54]. The main drawback of these approaches is the lack of modeling flexibility to leverage temporal dynamics across modalities capturing their relationships. Recently, DL-based frameworks have provided powerful and flexible alternatives for multimodal learning, offering various strategies for better model-level fusion. Instead of concatenating features, model-level fusion combines information in the intermediate hidden outputs of the DL feature encoders [15, 27]. By leveraging jointly learnable feature representations, we can capture cross-modality dynamics along with emotional-relevant patterns [39]. Applying an additional attention mechanism [31] can further enhance the model summarization ability by modeling long-term dependency and modality-specific information [21, 52].

The current *state-of-the-art* (SOTA) fusion approaches are based on the transformer architecture [12, 18, 41, 49]. This strategy utilizes the cross-modal attention framework [45], which is a dot-product attention mechanism that systematically defines attention rules by performing dot-product multiplications of the query, key, and value matrices associated with different regions of a sequence or distinct modalities. Specifically, in a multimodal context, this approach enables the model to capture comprehensive global cross-modal correlations by using the query (Q) from one modality to compute attention scores with the keys (K) and values (V) from another modality. This process allows the model to learn associations between different modalities and integrate information from both sources into the resulting representations, leading to a more robust understanding and synthesis of information across diverse data types. Therefore, we can perform bidirectional attention between audio-visual or audio-text modalities, enabling the model to attend to either modality based on the other. In this study, we adopt the SOTA cross-modal attention model [12, 45] as the backbone architecture, proposing a transformer framework based on the *word-chunk* concept for audio-text emotion recognition. We make appropriate modifications to fit this model to our problem. We describe our multimodal architecture in Section 3.

2.2 Robustness Against Missing Data

While multimodal emotion recognition can effectively improve the model performance by leveraging complementary discriminative information conveyed by the individual modalities, these systems are less flexible when modalities are missing. The common assumption is that all the relevant modalities need to be available for the multimodal system to operate. This problem is especially prevalent in in-the-wild applications [35]. For instance, speech signals can suffer from packet losses or noise interferences [7], and significantly degrade the *automatic speech recognition* (ASR) performance [22, 23], affecting the text transcriptions generated by the ASR system. As a result, both speech and text data are compromised, leading to multimodal emotion recognition systems with worse performance [55]. Additional ad-hoc techniques are required to compensate for the model, such as missing data cancellation and generation [10, 23], cascaded enhancement and recognition models [44], or multimodal joint representation learning [38].

Various studies have been proposed to handle missing total or partial data from modalities for robust emotion recognition. Investigation from early studies showed that simple feature interpolation of missing values is helpful [40]. Another straightforward approach is to perform ensemble fusion based on model confidence from different modalities [40, 47] (e.g., weighted majority voting). The emotion classifier from the modality with missing data is expected to produce less confident predictions, and, therefore, has a lower contribution to the final decision. For recent DL frameworks, one research direction focuses on training strategies to handle missing modalities. Goncalves and Busso [12] proposed an optimized training strategy where 20% of the data from one modality was replaced by zeros in some batches (e.g., modality dropout). Parthasarathy and Sundaram [37] proposed a similar approach in which they randomly zero out visual data with a given probability during the training stage for an audiovisual emotion recognition task. This data strategy can be done at the clip or frame levels, resulting in performance gains of up to 17% for the transformer model. Zuo *et al.* [55] attempted to learn a modality-invariant feature space to compensate for the mismatch introduced by the missing data. Another approach is to explicitly introduce a system for canceling out the missing data, such as *packet-loss concealment* (PLC) for speech [23]. Mohamed and Schuller [36] presented an end-to-end PLC framework concatenated with the downstream *speech emotion recognition* (SER) task, where the model makes emotion prediction after it reconstructs the lost frames. Different from the aforementioned methods, our proposed solution to improve model robustness relaxes the cross-modality (i.e., audio-text) temporal dependency by randomizing the lexical segmentation during training.

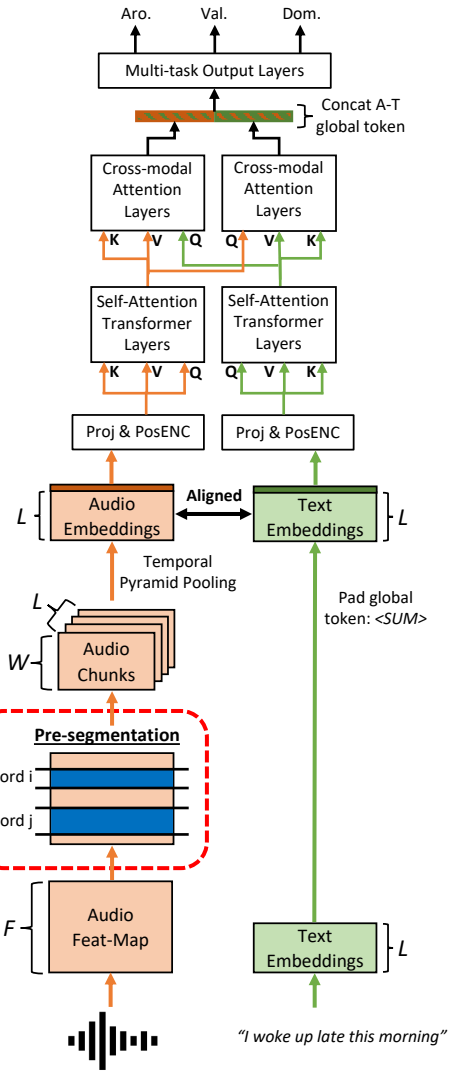


Figure 1: Overview of the proposed transformer-based architecture for audio-text emotion recognition. The key novelty of the model is the word-chunk approach to split a sentence into chunks. The approach provides the flexibility to introduce alternative alignment strategies with different dependencies on lexical boundaries to combine words with the co-occurring speech frames.

3 PROPOSED METHODOLOGY

We show the proposed framework in Figure 1, which consists of three major components: 1) pre-segmentation module, 2) input preprocessing, and 3) the main transformer-based architecture. We describe these blocks in this section.

3.1 Pre-Segmentation of Audio Chunks

Studies have shown the benefits of using chunk-based segmentation, splitting sentences into smaller segments [24, 25]. Therefore, the first step is to extract the feature representation for the input modalities of audio and text data. The frame-level audio feature map and the word-level text embeddings are denoted by the variables $\mathbf{X_A} \in \mathbb{R}^{F \times D_A}$ and $\mathbf{X_T} \in \mathbb{R}^{L \times D_T}$. The sequence lengths are F and L, where $F \gg L$ since the number of audio frames is always greater than the number of spoken words in a sentence. D_A and D_T represent the extracted feature dimension for the corresponding modalities. The goal of the pre-segmentation step is to segment the audio data into C chunks according to the lexical boundary, obtaining an aligned audio-text feature pair (i.e., make $C = L$, creating

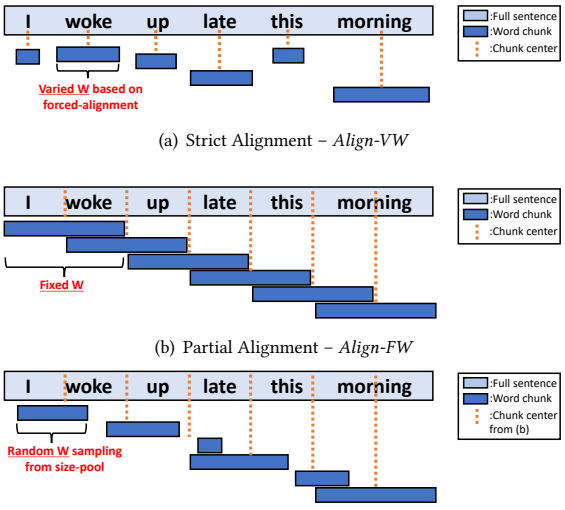
a one-to-one map between the speech and word representations). The advantage of this approach is the ability to control the strictness of word boundaries when defining the alignment between modalities. This flexibility enables us to investigate the role of temporal synchronization between lexical and acoustic information in multimodal emotion recognition systems. We consider three alignment levels: strict, partial, and random alignment. The key difference between these strategies is how we define the data chunk window size W (i.e., word boundary). Figure 2 illustrates these three strategies.

• *Strict Alignment- Fig. 2(a)*: This approach segments data chunks by strictly following the forced alignment results. It crops the audio frames co-occurring with the word using the starting and ending timestamps in the word-level alignment information. We denote this approach to as *Align-VW*, since it produces varied sizes for the length of the chunks.

• *Partial Alignment- Fig. 2(b)*: The sentence is uniformly divided into a number of chunks with a fixed duration W, according to the number of words in the sentence. This approach is inspired by the segmentation approach proposed by Lin and Busso [26]. Equation 1 shows the key formula, where the shifting between data chunks Δl depends on the sentence duration T and the number of words L. With the fixed W, each data chunk is likely to at least cover a partial region of the word corresponding to that chunk. For instance, we can see in Figure 2(b) that the first chunk covers the first word, "I", and the second chunk also covers the second word, "woke". We denote this strategy to as *Align-FW*, since it produces a fixed duration W for all the chunks.

$$\Delta l = \frac{T - W}{L - 1} \tag{1}$$

• *Random Alignment- Fig. 2(c)*: This approach randomly segments the data into chunks of different durations. The process starts with the segmentation provided by the partial alignment strategy *Align-FW* (i.e., a fixed number of chunks with the same duration). The approach crops a sub-region from each data chunk by randomly selecting the starting point of the chunk and its duration. We denote the cropped sub-region as s, which is randomly sampled from a defined set $\mathcal{S}$ of reasonable durations, where $s \leq W$. Section 4.3 provides the values used for the implementation in this study. For instance, the first data chunk in Figure 2(c) is a sub-region sampled from the first data chunk in Figure 2(b). This process is conducted for each epoch during the training process, creating different random chunk segmentations for each sentence during the training. This random cropping strategy has the least strict alignment requirements. However, this strategy proves to be an effective mean of temporal dependency regularization by randomly selecting s at each epoch. Notice that this segmentation method still preserves the minimal lexical boundary information, since the number of words determines the number of chunks. Figure 2(c) shows an example of this strategy, where the data chunks cover portions of the corresponding word. We refer to this approach as *Align-MultiScaleW*, since this approach produces different sizes of W using a multi-scale strategy. The multi-scale part of the model comes from implementing this chunk segmentation strategy at each epoch, presenting different *alignments* between speech and text during the training process.



(a) Strict Alignment – *Align-VW*

(b) Partial Alignment – *Align-FW*

(c) Random Alignment – *Align-MultiScaleW*

Figure 2: Visualization of three different chunk segmentation approaches explored to understand the role of modality synchronization in the performance of a multimodal emotion recognition system.

3.2 Input Preprocessing

After obtaining the audio chunks, we adopt the *temporal pyramid pooling* (TPP) [43] to flatten out the temporal dimension of the data chunks. Figure 3 visualizes the TPP procedure. Unlike conventional mean-pooling operation, which results in a single average vector, TPP performs mean-pooling by further subdividing the data chunk into different scales along with the temporal axis (i.e., W). Then, the approach concatenates these partial pooling results to obtain a chunk-based representation. With this approach, TPP captures richer information at different dynamic temporal resolutions. Then, we can obtain the aligned audio-embeddings $\mathbf{X_A} \in \mathbb{R}^{L \times D_A}$ by simply stacking the flattened outputs created by the TPP. In our emotion recognition task, we deal with sentence-level labels. To better represent this information, we generate a special token, denoted as <*SUM*>, for every sequence, which serves as the aggregate sentence-level representation for emotion recognition. This token is designed to encapsulate the overall sentence information and is prepended to the input of the model, as in Devlin *et al.* [8]. The final hidden state corresponding to the <*SUM*> token is then utilized to recognize emotions. To incorporate this token into the model, we use a one-vector initialization for both audio and text inputs. Specifically, we initialize <*SUM*> with a one-vector of dimensions $\mathbf{1}^{1 \times D_A}$ for audio and $\mathbf{1}^{1 \times D_T}$ for text to train the model. Notice that our input sequence length is dramatically decreased from F (frames) to L (words), which significantly reduces the complexity of the transformer model for the audio stream (i.e., $O(L^2)$).

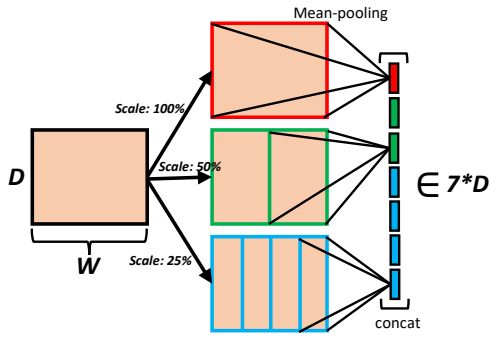


Figure 3: Visualization of temporal pyramid pooling. Our implementation exactly follows this strategy, splitting the chunks into two and four blocks.

3.3 Main Transformer Model Architecture

Our main model follows the standard setup of the transformer encoder [46]. The input embeddings are linearly projected into a fixed hidden dimension. We add positional encodings using the sine/cosine functions. Then, we implement the multi-head self-attention block to capture the intra-modality information. We refer to this block in Figure 1 as the self-attention transformer layer. Then, we implement the cross-modal attention layer to capture the cross-modality correlation. The major difference between the self- and cross-modal attention is the query (Q) term in the scaled dot-product attention formula (Eq. 2). The query of the cross-modal attention model comes from another modality [45] (see cross-modal attention layer in Fig. 1).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Lastly, we concatenate the hidden output of the global tokens from the two modalities, which are fed into the multitask output layers for predicting emotions. We build our model to recognize the three conventional emotional attributes: arousal, valence, and dominance. Since these labels are represented by continuous values, the architecture is formulated as a regression problem. We use the *concordance correlation coefficient* (CCC) to define the multitask loss function to train the model (Eq. 3). We also use CCC as the evaluation metric to assess model performance.

$$\mathcal{J} = \frac{1}{3}\big((1 - \text{CCC}_{\text{aro}}) + (1 - \text{CCC}_{\text{val}}) + (1 - \text{CCC}_{\text{dom}})\big) \quad (3)$$

4 EXPERIMENTAL SETTINGS

4.1 The MSP-Podcast Corpus

We use the release version 1.10 of the MSP-Podcast corpus [29]. The dataset collects spontaneous recordings that are available on audio-sharing websites. A series of automatic processes including speaker diarization, and noise and music detection are applied to split the recordings into clean speaking turns. The file duration of the speaking turns ranges from 2.75 to 11 secs. The approach relies on the retrieval-based strategy proposed by Mariooryad *et al.* [32], where only samples that are predicted to have emotional content are annotated with emotional labels. The emotional evaluation follows a modified version of the real-time crowdsourcing protocol proposed by Burmania *et al.* [3]. This approach resulted in a large-scale corpus with high-quality and spontaneous speaking turns.

The corpus provides the predefined train/development/test partitions with a total of 104,267 audio clips (≈166 hrs). We use the development set for hyperparameter tuning and the test set for our evaluation. These speaking turns are annotated for arousal (calm versus active), valence (negative versus positive), and dominance (weak versus strong) with continuous values from 1 to 7. Each file has at least five annotations and we consider the average score as the ground-truth labels. The corpus also provides the human transcriptions, and the corresponding word-level forced alignment results based on the *Montreal-forced-aligner* (MFA) [34]. We utilize these resources for building the proposed word-chunk experiments.

4.2 Audio and Text Features

This study relies on pre-trained self-supervised deep features for acoustic features, which have consistently shown better performance compared to traditional acoustic features in *speech emotion recognition* (SER) such as *mel frequency cepstral coefficients* (MFCCs) [48]. We extract frame-level features based on the *wav2vec2-large-robust* model [1]. For text, we follow a similar approach, extracting word-level embeddings with the *RoBERTa* model [28]. The last hidden state outputs of these models are the inputs for our model (D_A=1,024 for audio, and D_T=768 for text). We extract these features using the Hugging Face library [50]. Notice that these input features are fixed. We do not fine-tune or incorporate the pre-trained model architecture during training. We perform additional z-normalization on the features to facilitate the convergence speed of the models during training, where the normalization parameters (i.e., mean and std) are estimated from the train set.

4.3 Implementation Details

We set all the hidden dimensions of the transformer model to 256D using four attention heads. As we mentioned in Section 3.2, the model input lengths depend on the number of words spoken in the sentence (i.e., L). Therefore, we set the maximum input sequence length to 128 (words), which is long enough to cover the full information without truncating the sentence. We zero-pad sentences to reach the maximum length of 128 to facilitate batch training. We use a batch size of 128, the Adam optimizer with a learning rate set to 0.0001, and an early stopping criterion based on the CCC performance of the development set. We save the best model, which is evaluated on the test set.

For the segmentation setup, we use W=1 sec for the Align-FW approach. For the Align-MultiScaleW approach, we define the following random size-pool set $\mathcal{S}$={0.2 secs, 0.4 secs, 0.6 secs, 0.8 secs, 1.0 secs}. The TPP scale setting is exactly the same as illustrated in Figure 3, dividing the chunk into two and four parts. We conduct a two-tailed t-test based on the CCC metric to compare the results in the experimental evaluation. Each experiment is run five times using different network initializations. We equally split the original test set into five subsets, resulting in a total of 5×5=25 testing points

Table 1: CCC results of multimodal baselines and proposed transformer-based approach implemented with different alignment strategies. Symbol $*$ and † indicate that the system performance is significantly better than the MDRE-GRU and MDREA-GRU baseline, respectively. The bold values show the best result per emotional attribute.

Approach	Arousal	Valence	Dominance
MDRE-GRU [52]	0.6090	0.5592	0.4842
MDREA-GRU [52]	0.6029	0.5603	0.4792
Align-VW	0.6094	$0.5723^{*\dagger}$	$0.4932^{*\dagger}$
Align-FW	$0.6117^{\dagger}$	$\mathbf{0.5767}^{*\dagger}$	$0.4939^{*\dagger}$
Align-MultiScaleW	$\mathbf{0.6137}^{\dagger}$	$0.5691^{*\dagger}$	$\mathbf{0.5000}^{*\dagger}$

for the statistical analysis. We assert statistically significant when p-value < 0.05. All models are trained and tested using an NVIDIA GeForce RTX 2080 Ti GPU. The codes are implemented in PyTorch.

We compare two model-level fusion approaches with the transformer framework using cross-modal attention layers. We consider the *multimodal dual recurrent encoder* (MDRE)-GRU framework [52]. This approach concatenates the last time-step hidden output of a *gated recurrent units* (GRU) encoder from the audio and text streams. This model-level fusion vector is then utilized to predict emotions. The MDREA-GRU introduces an additional attention module on the top of the MDRE-GRU framework, fusing information with a weighted sum from all the time steps from the GRU encoder outputs (the "*A*" in the acronym stands for attention). Notice that we use the same deep features mentioned in Section 4.2 to train the baselines for fair comparison, which is different from the original implementation in this model that used MFCCs.

5 EXPERIMENTAL RESULTS AND ANALYSIS

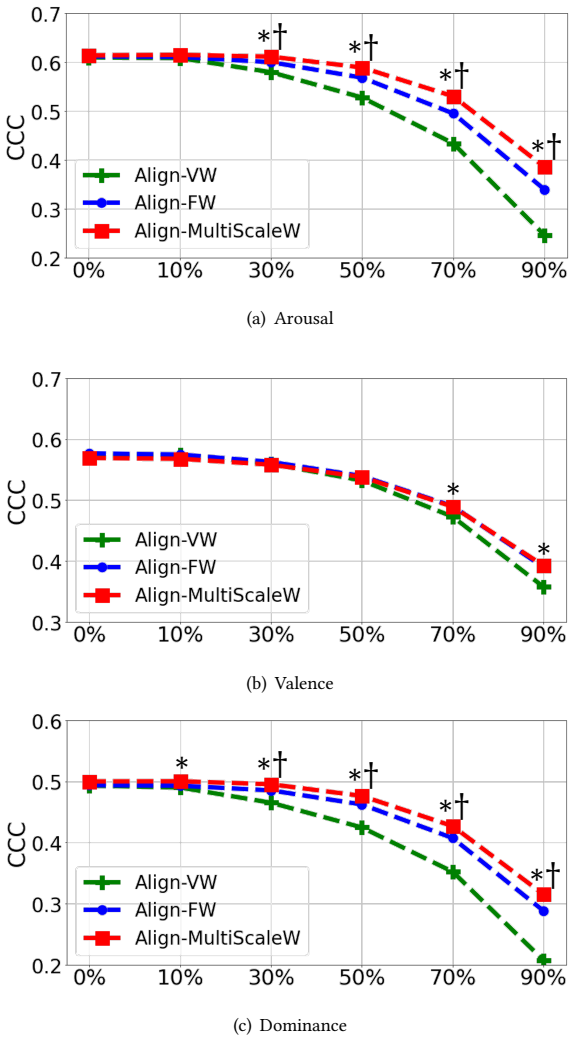
5.1 Role of Temporal Alignment

First, we compare the proposed approach with the different alignment strategies to the baselines. Table 1 lists the performances of these systems. The fusion method using cross-modal attention consistently outperforms classic model-level fusion techniques. For valence and dominance, the three implementations of the proposed approach are statistically better than the baseline models.

As mentioned before, the flexibility of the proposed model allows us to explore the change in performance as we use strict, partial, and random alignment between the audio and text feature representations. When we compare the models with the different implementations, we do not observe significant performance differences by introducing the additional alignment information. The Align-VW results do not show any performance gain over the Align-FW or Align-MultiScaleW approaches, which demonstrates the minor effect in performance of providing precise temporal synchronization between the input audio-text pair. The attention mechanism implemented in the proposed system can effectively compensate for cases when the modalities are not perfectly aligned.

5.2 Robustness to Missing Data

We evaluate the model's robustness against missing data. We simulate the missing data condition by randomly dropping x% of words



(a) Arousal

(b) Valence

(c) Dominance

Figure 4: Analysis of robustness of the multimodal systems against missing data from both modalities. The x-axis in each plot indicates the data drop ratio. The y-axis is the corresponding CCC performance. The symbols $*$ and † indicate that the results for the Align-MultiScaleW strategy are significantly better than the results achieved with the Align-VW and Align-FW strategies, respectively.

in a sentence during the inference stage. Considering the multimodal task setup, we drop both audio and text data simultaneously. Once the word is dropped, its corresponding uttered speech frames are also dropped based on the forced alignment results.

Figure 4 shows the performance when the drop rate x is set from 0% to 90% for the three emotional attributes. We add symbols to indicate when the Align-MultiScaleW strategy leads to statistically better performance than the Align-VW ($*$) and Align-FW (†) strategies. The model trained with strict alignment information (i.e., Align-VW) is vulnerable when the data is missed. This strategy is consistently the worst across emotional attributes. The

Table 2: Speech emotion recognition modeling under the packet-loss simulation evaluation. The results are shown in term of CCC values. Bold values show the best performance per attribute for each set of values for P_L and P_N. The symbols $*$ and † indicate that the results for the Align-MultiScaleW strategy are significantly better than the results achieved with the Align-VW and Align-FW strategies, respectively.

approx. drop frames (%)	(p_L, p_N)	Arousal			Valence			Dominance		
		Align-VW	*Align-FW*	*Align-MultiScaleW*	*Align-VW*	*Align-FW*	*Align-MultiScaleW*	*Align-VW*	*Align-FW*	*Align-MultiScaleW*
0%	(0.0, 1.0)	0.5906	0.5999	**0.6120***†	0.3420	0.3384	**0.3554***†	0.4776	0.4775	**0.4917***†
10%	(0.1, 0.9)	0.5931	0.6018	**0.6138***†	0.3416	0.3385	**0.3552***†	0.4790	0.4784	**0.4934***†
17%	(0.5, 0.9)	0.5934	0.6014	**0.6140***†	0.3380	0.3359	**0.3531***†	0.4774	0.4777	**0.4935***†
36%	(0.1, 0.5)	0.5950	0.6009	**0.6119***†	0.3344	0.3306	**0.3457***†	0.4795	0.4768	**0.4916***†
50%	(0.1, 0.1)	0.5887	0.5903	**0.6004***†	0.3250	0.3161	**0.3256**	0.4724	0.4668	**0.4802***†
50%	(0.5, 0.5)	0.5822	0.5873	**0.5979***†	0.3148	0.3131	**0.3235**	0.4654	0.4640	**0.4780***†
50%	(0.9, 0.9)	0.5506	0.5708	**0.5822***†	0.2899	0.2990	**0.3139***†	0.4337	0.4481	**0.4645***†
65%	(0.5, 0.1)	0.5503	0.5515	**0.5592**	**0.2828**	0.2759	0.2766	0.4362	0.4324	**0.4432**
84%	(0.9, 0.5)	0.3541	**0.3805**	0.3719*	**0.1604**	0.1572	0.1407	0.2782	**0.2934**	0.2924*
90%	(0.9, 0.1)	0.2400	**0.2684**	0.2573*	**0.0924**	0.0922	0.0763	0.1948	**0.2084**	0.2073*

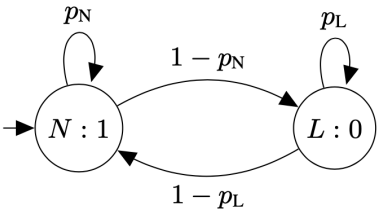


Figure 5: *Markov Chain* packet-loss model to simulate the loss of frames. State N:1 indicates that the frame is received, and state L:0 indicates that the frame is lost.

performance dramatically drops when the missing data is more than 30%, especially for arousal and dominance. The major cause for this result is that the Align-VW model learns a strong dependency between the audio and text representations relying on its temporal synchronization. This approach reduces the resiliency of the model when words are dropped. In contrast, the Align-MultiScaleW approach consistently leads to better CCC performance, showing stronger performance against missing data (see the red lines in Fig. 4). For instance, when 90% of the data is lost, the Align-MultiScaleW strategy leads to performance gains for arousal above 14% and 5% compared to the Align-VW and Align-FW strategies, respectively (Fig. 4(a)). The random alignment in the Align-MultiScaleW strategy relaxes the requirement of matching lexical boundaries, which naturally prevents the model from memorizing trivial temporal dependency. It also regularizes the model by presenting different alignments for each sentence across epochs. Interestingly, the Align-FW strategy, which has partial alignment constraints, shows better robustness than the Align-VW method, but worse performance than the Align-MultiScaleW strategy.

5.3 Audio-Only Modeling

The characteristic of the multi-scale chunk regularization is also effective for audio-only emotion recognition tasks. In this section, we simply remove the text-modality pipeline in the right-hand side of Figure 1 (i.e., green colors) to adapt the model for SER. Note that the cross-modal attention layers are replaced by self-attention layers, since we do not have another modality for the cross-attention queries. To simulate missing data testing conditions for audio, we follow the two-state *Markov Chain* packet-loss generation model commonly used to simulate missing packets on the Internet. This model was also used by Mohamed and Schuller [36] for SER tasks. Figure 5 shows this model, which is defined by the frame-level probabilities that a packet is lost (p_L) or transmitted (p_N). It produces a randomly sampled binary sequence mask for the target testing sentence. The initial frame always starts at the non-loss state.

Table 2 summarizes the full testing results under different values for p_L and p_N. The table shows that the SER model implemented with the multi-scale chunk regularization (i.e., Align-MultiScaleW) obtains better performance than the other strategies for all three emotional attributes in most of the testing cases. More importantly, we find significant CCC improvements even under non-missing data scenarios (see the 0% results), which indicates that the multi-scale chunk regularization not only improves the model robustness, but also leads to SER performance gains. The approach produces a similar effect to the random crops in computer vision tasks (i.e., we random crop the voiced portions of the speech signal), which can be considered as a data-augmentation scheme for SER to increase robustness and performance.

6 CONCLUSIONS

This paper presented a computational multimodal framework based on the transformer architecture, implemented with cross-modal attention layers for audio-text emotion recognition. The approach has the flexibility to explore different alignment constraints between the modalities. The framework is able to model the alignment level between speech and text using a word-chunk concept, which pre-segments data chunks according to the lexical boundaries. This approach facilitates the investigation of the role of temporal synchronization between the modalities in the model performance. We implement the proposed approach with strict, partial, and random alignment strategies. The random strategy is a major contribution of this study. For each sentence, the approach generates random controlled alignments between words and the co-occurring speech

frames. The alignments for each sentence change at every epoch, serving as a temporal regularization mechanism. Our experimental results based on the MSP-Podcast corpus indicate that the temporal synchronization of audio and text feature representation plays a minor role in performance, as the three strategies achieve similar results. The use of attention mechanisms, including the cross-modal attention layers, is powerful enough to compensate for potential misalignment affecting the temporal relationship of the modalities. However, we found that the model becomes extremely vulnerable to missing data when the model is trained with the strict alignment strategy. In contrast, the proposed random alignment strategy results in a multi-scale chunk regularization solution that significantly increases the robustness of the model against missing data. The study also demonstrated that this strategy is effective for SER, exhibiting similar robustness against missing acoustic frames. The random multi-scale chunk segmentation strategy also improved the model recognition performance when all the frames were available.

This study opens several research questions. Following the word-chunk concept, we can extend the investigation of temporal synchronization to three or more modalities (e.g., video, audio, and text). Also, we can explore the idea of multi-scale chunk segmentation with other modalities (e.g., fMRI or EEG). The temporal relationship between different modalities can be very complex. Therefore, having a complete understanding of its role in system performance might lead to a better solution for multimodal modeling problems. We are also intrigued by the improvements in CCC performance observed in SER when the model was implemented with the proposed multi-scale chunk segmentation. We will continue to study the use of this approach as a regularization strategy.

ACKNOWLEDGMENTS

This work was funded by National Science Foundation (NSF) under grant CNS-2016719.

REFERENCES

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, Vol. 33. Virtual, 12449–12460.

[2] T. Baltrušaitis, C. Ahuja, and L. P. Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (February 2019), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[3] A. Burmania, S. Parthasarathy, and C. Busso. 2016. Increasing the Reliability of Crowdsourcing Evaluations Using Online Quality Assessment. *IEEE Transactions on Affective Computing* 7, 4 (October-December 2016), 374–388. https://doi.org/10.1109/TAFFC.2015.2493525

[4] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. 2004. Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. In *Sixth International Conference on Multimodal Interfaces ICMI 2004.* ACM Press, State College, PA, 205–211. https://doi.org/10.1145/1027933.1027968

[5] C. Busso and S.S. Narayanan. 2006. Interplay between linguistic and affective goals in facial expression during emotional utterances. In *7th International Seminar on Speech Production (ISSP 2006)*. Ubatuba-SP, Brazil, 549–556.

[6] C. Busso and S. Narayanan. 2007. Interrelation between speech and facial gestures in emotional utterances: a single subject study. *IEEE Transactions on Audio, Speech and Language Processing* 15, 8 (November 2007), 2331–2347. https://doi.org/10.1109/TASL.2007.905145

[7] A. Cohen, I. Rimon, E. Aflalo, and H.H. Permuter. 2022. A study on data augmentation in voice anti-spoofing. *Speech Communication* 141 (June 2022), 56–67. https://doi.org/10.1016/j.specom.2022.04.005

[8] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*. Minneapolis, Minnesota, 4171–4186.

[9] S. K. D'mello and J. Kory. 2015. A Review and Meta-Analysis of Multimodal Affect Detection Systems. *Comput. Surveys* 47, 3 (April 2015), 1–36. https://doi.org/10.1145/2682899

[10] C. Du, C. Du, H. Wang, J. Li, W.-L. Zheng, B.-Liang Lu, and H. He. 2018. Semi-Supervised Deep Generative Modelling of Incomplete Multi-Modality Emotional Data. In *ACM international conference on Multimedia (MM 2018)*. Seoul, Republic of Korea, 108–116. https://doi.org/10.1145/3240508.3240528

[11] L. Goncalves and C. Busso. 2022. AuxFormer: Robust Approach to Audiovisual Emotion Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*. Singapore, 7357–7361. https://doi.org/10.1109/ICASSP43922.2022.9747157

[12] L. Goncalves and C. Busso. 2022. Robust Audiovisual Emotion Recognition: Aligning Modalities, Capturing Temporal Information, and Handling Missing Features. *IEEE Transactions on Affective Computing* 13, 4 (October-December 2022), 2156–2170. https://doi.org/10.1109/TAFFC.2022.3216993

[13] L. Goncalves and C. Busso. 2023. Learning Cross-modal Audiovisual Representations with Ladder Networks for Emotion Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*. Rhodes island, Greece, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096138

[14] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso. 2023. Versatile Audiovisual Learning for Handling Single and Multi Modalities in Emotion Regression and Classification Tasks. *ArXiv e-prints (arXiv:2305.07216)* (May 2023), 1–14. https://doi.org/10.48550/arXiv.2305.07216 arXiv:2305.07216 [cs.LG]

[15] J. Han, Z. Zhang, Z. Ren, and B. Schuller. 2019. Implicit Fusion by Joint Audiovisual Training for Emotion Recognition in Mono Modality. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*. Brighton, UK, 5861–5865. https://doi.org/10.1109/ICASSP.2019.8682773

[16] K. He, X. Zhang, S. Ren, and J. Sun. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 9 (September 2015), 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824

[17] L.A. Hendricks, J. Mellor, R. Schneider, J.-B. Alayrac, and A. Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics* 9 (July 2021), 570–585. https://doi.org/10.1162/tacl_a_00385

[18] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu. 2020. Multimodal transformer fusion for continuous emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*. Barcelona, Spain, 3507–3511. https://doi.org/10.1109/ICASSP40776.2020.9053762

[19] E. Hudlicka. 2003. To feel or not to feel: The role of affect in human-computer interaction. *International Journal of Human-Computer Studies* 59, 1-2 (July 2003), 1–32. https://doi.org/DOI:10.1016/S1071-5819(03)00047-8

[20] A. Khare, S. Parthasarathy, and S. Sundaram. 2021. Self-Supervised Learning with Cross-Modal Transformers for Emotion Recognition. In *IEEE Spoken Language Technology Workshop (SLT 2021)*. Shenzhen, China, 381–388. https://doi.org/10.1109/SLT48900.2021.9383618

[21] C. Li, Z. Bao, L. Li, and Z. Zhao. 2020. Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition. *Information Processing & Management* 57, 3 (May 2020), 102185. https://doi.org/10.1016/j.ipm.2019.102185

[22] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach. 2014. An Overview of Noise-Robust Automatic Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 4 (April 2014), 745–777. https://doi.org/10.1109/TASLP.2014.2304637

[23] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen. 2021. A Time-Domain Convolutional Recurrent Network for Packet Loss Concealment. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*. Toronto, ON, Canada, 7148–7152. https://doi.org/10.1109/ICASSP39728.2021.9413595

[24] W.-C. Lin and C. Busso. 2020. An Efficient Temporal Modeling Approach for Speech Emotion Recognition by Mapping Varied Duration Sentences into Fixed Number of Chunks. In *Interspeech 2020*. Shanghai, China, 2322–2326. https://doi.org/10.21437/Interspeech.2020-2636

[25] W.-C. Lin and C. Busso. 2022. Chunk-Level Speech Emotion Recognition: A General Framework of Sequence-to-One Dynamic Temporal Modeling. *IEEE Transactions on Affective Computing* Early Access (2022). https://doi.org/10.1109/TAFFC.2021.3083821

[26] W.-C. Lin and C. Busso. 2023. Role of Lexical Boundary Information in Chunk-Level Segmentation for Speech Emotion Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*. Rhodes island, Greece, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096861

[27] J. Liu, S. Chen, L. Wang, Z. Liu, Y. Fu, L. Guo, and J. Dang. 2021. Multimodal Emotion Recognition with Capsule Graph Convolutional Based Representation Fusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*. Toronto, ON, Canada, 6339–6343. https://doi.org/10.1109/ICASSP39728.2021.9413608

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining

Approach. *ArXiv e-prints (arXiv:1907.11692)* (July 2019), 1–12. https://doi.org/10.48550/arXiv.1907.11692 arXiv:1907.11692 [cs.CL]

[29] R. Lotfian and C. Busso. 2019. Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings. *IEEE Transactions on Affective Computing* 10, 4 (October-December 2019), 471–483. https://doi.org/10.1109/TAFFC.2017.2736999

[30] J. Lu, D. Batra, D. Parikh, and S. Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Conference on Neural Information Processing Systems (NeurIPS 2019)*. Vancouver, BC, Canada, 1–11.

[31] T. Luong, H. Pham, and C.D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. Lisbon, Portugal, 1412–1421.

[32] S. Mariooryad, R. Lotfian, and C. Busso. 2014. Building A Naturalistic Emotional Speech Corpus by Retrieving Expressive Behaviors From Existing Speech Corpora. In *Interspeech 2014*. Singapore, 238–242.

[33] C. A. Mazefsky and D. P. Oswald. 2007. Emotion perception in Asperger's syndrome and high-functioning autism: the importance of diagnostic criteria and cue intensity. *Journal of Autism and Developmental Disorders* 37, 6 (July 2007), 1086–1095. https://doi.org/10.1007/s10803-006-0251-6

[34] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017*. Stockholm, Sweden, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

[35] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. 2020. M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues. In *AAAI Conference on Artificial Intelligence (AAAI 2020)*, Vol. 34. New York, NY, USA, 1359–1367. https://doi.org/10.1609/aaai.v34i02.5492

[36] M.M. Mohamed and B.W. Schuller. 2020. ConcealNet: An End-to-end Neural Network for Packet Loss Concealment in Deep Speech Emotion Recognition. *ArXiv e-prints (arXiv:2005.07777)* (May 2020), 1–5. https://doi.org/10.48550/arXiv.2005.07777 arXiv:2005.07777 [cs.AS]

[37] S. Parthasarathy and S. Sundaram. 2020. Training Strategies to Handle Missing Modalities for Audio-Visual Expression Recognition. In *International Conference on Multimodal Interaction (ICMI 2020)*. Utrecht, The Netherlands, 400–404. https://doi.org/10.1145/3395035.3425202

[38] H. Pham, P.P. Liang, T. Manzini, L.-P. Morency, and B. Póczos. 2019. Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities. In *AAAI Conference on Artificial Intelligence (AAAI 2019)*, Vol. 33. Honolulu, HI, USA, 6892–6899. https://doi.org/10.1609/aaai.v33i01.33016892

[39] F. Qian and J. Han. 2022. Contrastive Regularization for Multimodal Emotion Recognition Using Audio and Text. *ArXiv e-prints (arXiv:2211.10885)* (November 2022), 1–5. https://doi.org/10.48550/arXiv.2211.10885 arXiv:2211.10885 [cs.SD]

[40] C. Setz, J. Schumm, C. Lorenz, B. Arnrich, and G. Tröster. 2009. Using ensemble classifier systems for handling missing data in emotion recognition from physiology: one step towards a practical system. In *International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009)*. Amsterdam, Netherlands, 1–8. https://doi.org/10.1109/ACII.2009.5349590

[41] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara. 2020. Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access* 8 (September 2020), 176274–176285. https://doi.org/10.1109/ACCESS.2020.3026823

[42] M. Soleymani, M. Pantic, and T. Pun. 2012. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing* 3, 2 (April-June 2012), 211–223. https://doi.org/10.1109/T-AFFC.2011.37

[43] S. Sudholt and G. A. Fink. 2017. Evaluating word string embeddings and loss functions for CNN-based word spotting. In *IAPR International Conference on Document Analysis and Recognition (ICDAR 2017)*. Kyoto, Japan, 493–498. https://doi.org/10.1109/ICDAR.2017.87

[44] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. Schuller. 2019. Towards Robust Speech Emotion Recognition Using Deep Residual Networks for Speech Enhancement. In *Interspeech 2019*. Graz, Austria, 1691–1695. https://doi.org/10.21437/Interspeech.2019-1811

[45] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, and R. Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Association for Computational Linguistics (ACL 2019)*, Vol. 1. Florence, Italy, 6558–6569. https://doi.org/10.18653/v1/p19-1656

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *In Advances in Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA, USA, 5998–6008.

[47] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim. 2011. Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data. *IEEE Transactions on Affective Computing* 2, 4 (October-December 2011), 206–218. https://doi.org/10.1109/T-AFFC.2011.12

[48] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B.W. Schuller. 2023. Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Early Access (2023). https://doi.org/10.1109/TPAMI.2023.3263585.

[49] Y. Wang, G. Shen, Y. Xu, J. Li, and Z. Zhao. 2021. Learning Mutual Correlation in Multimodal Transformer for Speech Emotion Recognition. In *Interspeech (2021)*. Brno, Czech Republic, 4518–4522. https://doi.org/10.21437/Interspeech.2021-2004

[50] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, and Q. Lhoest amd A.M. Rush. 2019. HuggingFace's transformers: State-of-the-art natural language processing. *ArXiv e-prints (arXiv:1910.03771v5)* (October 2019), 1–8. https://doi.org/10.48550/arXiv.1910.03771 arXiv:1910.03771 [cs.CL]

[51] X. Wu, W-L. Zheng, Z. Li, and B.-L. Lu. 2022. Investigating EEG-based functional connectivity patterns for multimodal emotion recognition. *Journal of neural engineering* 19, 11 (February 2022), 016012. https://doi.org/10.1088/1741-2552/ac49a7

[52] S. Yoon, S. Byun, and K. Jung. 2018. Multimodal Speech Emotion Recognition Using Audio and Text. In *IEEE Spoken Language Technology Workshop (SLT 2018)*,. Athens, Greece, 112–118. https://doi.org/10.1109/SLT.2018.8639583

[53] J. Zhao, R. Li, S. Chen, and Q. Jin. 2018. Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions. In *Audio/Visual Emotion Challenge and Workshop (AVEC 2018)*. Seoul, Republic of Korea, 65–72. https://doi.org/10.1145/3266302.3266313

[54] W.-L. Zheng, B.-N. Dong, and B. L. Lu. 2014. Multimodal emotion recognition using EEG and eye tracking data. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2014)*. Chicago, IL, USA, 5040–5043. https://doi.org/10.1109/EMBC.2014.6944757

[55] H. Zuo, R. Liu, J. Zhao, G. Gao, and H. Li. 2023. Exploiting Modality-Invariant Feature for Robust Multimodal Emotion Recognition with Missing Modalities. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*. Rhodes Island, Greece, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095836