Multitask Transformer for Cross-Corpus Speech Emotion Recognition

Chung-Soo Ahn, Rajib Rana, Carlos Busso, *Fellow, IEEE*, Jagath C. Rajapakse, *Fellow, IEEE*

Abstract—Deep learning has significantly advanced the field of Speech Emotion Recognition (SER), yet its efficacy in cross-corpus scenarios remains a challenge. To overcome this limitation, recent studies demonstrate the success of multitask learning, which uses auxiliary tasks to reduce difference between source and target dataset (or transfer knowledge from source to target datasets). Despite the efforts, the overall accuracy for cross-corpus SER is still relatively low and needs attention. To improve performance, we propose a multitask framework with SER as the primary task and contrastive learning and information maximization as auxiliary tasks. We design the auxiliary tasks innovatively to use the target data without emotional labels to develop a better understanding of the target data. The core of our multitask framework is a pre-trained transformer. While transformers have gained attention in SER, their application to cross-corpus scenarios is still limited. Multimodal approaches for cross-corpus scenario is substantially limited as well. We use text as the second modality, developing separate multitask transformers for audio and text and conduct decision-level fusion during inference. We use publicly available and widely used speech corpora, including the IEMOCAP, MSP-IMPROV and EMO-DB databases. The results demonstrate the benefits of the proposed approach, achieving improved performance on the benchmark databases in cross-corpus settings.

Index Terms—Contrastive learning, cross-corpus speech emotion recognition, speech emotion recognition, transformers

1 Introduction

Speech Emotion Recognition (SER) has become a prominent focus in research, marked by continuous advancements driven by the successful integration of deep learning technologies, as evidenced in [1]. Notably, the widespread adoption of pre-trained transformer models in diverse audio applications, such as audio representation learning [2], underscores their effectiveness in SER. This effectiveness is demonstrated through a direct fine-tuning approach, as outlined in [3], [4]. Alternatively, there is an emerging strategy involving transformers pre-trained on audio spectrograms [5], offering a viable alternative to utilizing raw waveform inputs through Wav2Vec2 [2].

A significant challenge within SER is the cross-corpora evaluation, aiming to achieve generalization across disparate datasets [6], [7]. Cross-corpus SER, a subset of SER, is a task targeted toward transferring emotion recognition knowledge from a source dataset to a target dataset [6], which is nowadays achieved by employing deep learning models [8]. Recent strategies involve multitask learning [9]–[11], wherein a deep learning model is trained with the classification loss of the SER task and other tasks that do not utilize emotion labels. A typical procedure in cross-corpus SER involves training the model initially with source data containing emotion labels and subsequently training it again (which is often considered as transfer procedure) with target data without emotional labels. Multitask learning combined with data augmentation has achieved state-of-the-art performance [11], a strategy particularly beneficial for transformer models that demand substantial data quantities. Many of recent works employed fine tuning methods of pre-trained transformer for SER [12]–[15]. Yet, as the fine tuning method is a supervised learning task, its effect on cross-corpus SER was limited. Thus, implementing transformers with multitask learning with unsupervised auxiliary tasks holds promise for improving cross-corpus SER.

The multitask learning scheme incorporates multiple training objectives to facilitate cross-corpus training. Various tasks, or training objectives, has been explored in cross-corpus SER [8], typically reducing the feature difference between source dataset and target dataset [16]–[19] or other unsupervised learning objectives that does not require emotional labels [9], [20], [21]. Among the various techniques, contrastive learning has emerged as a notable unsupervised method, drawing recent attention [22], [23]. It endeavors to learn effective representations by attracting positive pairs and repelling negative pairs. In a typical contrastive learning scheme, two augmented samples from the same data instance are designated as positive pairs; otherwise, they serve as negative pairs. This design empowers the model to glean rich information, resembling a supervised classifier, albeit without the need for labels. In effect, similar data points are attracted to each other to have close distances in the representation space. Dissimilar data points repels each other and have far distances in the representation space. This knowledge can be interpreted as the internal structure of the dataset. However, it is important to note that contrastive learning does not explicitly focus on learning the gap between clusters, where gap between clusters represents knowledge for classification. Conversely, the In-

- *C-S. Ahn and J.C. Rajapakse are affiliated with the College of Computing and Data Science at Nanyang Technological University (NTU), Singapore.*
- *R. Rana is with University of Southern Queensland (USQ), Australia.*
- *C. Busso is with the Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213. Corresponding E-mail: asjagath@ntu.edu.sg*

0000–0000/00$00.00 © 2021 IEEE

formation Maximization (IM) loss operates on the cluster assumption [24], positing that a superior classifier will learn to identify class boundaries with substantial margins. The IM loss, optimized without labels, solely maximizes information content (or minimize entropy) computed from logits.

In this paper, we leverage both contrastive learning and the IM loss, hypothesizing that these two methods can complement each other, thereby enhancing performance in classifying data from a new source without the availability of class labels. By combining the strengths of contrastive learning and IM loss, we learn rich knowledge of classification boundary for both source and target datasets, without utilizing emotional labels from the target domain.

While cross-corpus SER has predominantly focused on the audio modality, extending these approaches to include different modalities, especially text [8], presents an open challenge. A natural extension is to explore multimodal emotion recognition (MER), encompassing both audio and text. However, cross-corpus MER introduces a non-trivial challenge for fusion with cross-corpus generalization. We acknowledge the complexity arising from the higher specificity of trained fusion models, with input feature sizes doubling compared to unimodal inputs. In this paper, we address the above-mentioned challenges. Our contributions can be summarised as follows.

1) We introduce a multitask transformer framework for cross-corpus Speech Emotion Recognition (SER), incorporating contrastive learning and information maximization (IM) as auxiliary tasks to enhance generalization.
2) We extend the framework to multimodal emotion recognition (MER) with separate multitask transformers for audio and text, utilizing decision-level fusion during inference.
3) We demonstrate state-of-the-art performance on benchmark datasets, including the IEMOCAP, MSP-IMPROV, and EMO-DB databases, in cross-corpus settings.

2 Related Work

This section aims to provide an in-depth analysis of the existing literature and identify the gap that we intend to address. We discuss deep learning and multitask learning studies tackling challenges of cross-corpus SER in the first two subsections. We then delve into specific transformer-based and multimodal methods for cross-corpus SER in the following two subsections. Finally, we summarize the research gaps identified in this section.

2.1 Deep Learning for Cross-Corpus SER

Deep learning-based methods are widely used for cross-corpus SER [8]. Typical deep learning architectures, such as convolutional neural network (CNN) or recurrent neural networks (RNN), have been used for cross-corpus SER tasks [21], [25], [26]. Autoencoders [21], [25] are also often used for cross-corpus SER. Autoencoder uses reconstruction loss to learn from unlabelled target data, which can be achieved using generative models by generating synthetic data [20], [28], [33]. Another approach for cross-corpus SER is to enforce a model to learn invariant representation across domains [10], [17], [20], [34], subspace learning [18], [19], or transformation from target to source domains [40].

2.2 Multitask Learning for Cross-Corpus SER

Multitask learning with various auxiliary tasks has been proposed [9]–[11], [34] for SER as the primary task. Domain adversarial loss [17], [20], [34], [41], center loss [11], [34], and reconstruction(or denoising) loss [9] are typical choice of auxiliary tasks. We propose contrastive loss coupled with information maximization loss obtained for clustering unlabeled target data as the auxiliary tasks, which, to our knowledge, it is the first time that this strategy has been proposed.

2.3 Transformer Models for SER and Cross-Corpus SER

Transformer models have been popularly used for various tasks using speech modality [2]. They have also been actively researched for SER [31], [35], [37]–[39]. But the use of transformer for cross-corpus SER has only recently started [12], [16]. Moreover, the above studies did not adopt augmentation to increase the volume and variability of data. Also, we believe that the adoption of fine-tuning of pre-trained transformer is beneficial to cross-corpus SER [42]. Exploration of fine-tuning approach deserves thorough exploration as it can harness rich information learned during pre-training.

2.4 Multimodal Models for SER and Cross-Corpus MER

Multimodal models have gained some attention for SER [43], while many recent studies employed transformers [31], [35], [37]–[39], [44]. However, the multimodal scheme for cross-corpus emotion recognition remains unexplored. To our knowledge, only one study has covered cross-corpus MER [26]. While most literature focuses on exploiting sophisticate multimodal methodology to naively achieve state-of-the-art, little research was conducted in generalization. As practical interest lies more on generalization issue, rather than emotion recognition accuracy, cross-corpus MER demands more investigations.

2.5 Summary

We briefly summarise the existing studies in Table 1. The gaps are as follows.

1) Multitask learning has been explored for cross-corpus SER. However, we are the first to propose contrastive learning coupled with information maximization loss to cluster target data as an auxiliary task.
2) Multimodal models have been used for SER but rarely for cross-corpus MER.
3) Transformers have been used for SER, but its use is very new in the cross-corpus SER.

The primary objective of this study is to enhance the accuracy of cross-corpus emotion recognition by addressing the gaps in the existing research, focusing mostly on SER but also extending our approach to MER. As per the literature, the baseline accuracy in this context is significantly low, and our aim is to contribute towards a substantial improvement.

TABLE 1
Summary of our literature survey to compare our method with representative studies on cross-corpus SER and MER.

Author	Year	Cross-corpus	Transformer	Multimodal	Multitask learning	Augmentation	Contrastive Learning as Auxiliary Task
Gideon et al. [17]	2019	✓			✓		
Neumann et al. [25]	2019	✓			✓		
Sahu et al. [26]	2019	✓		✓			
Luo et al. [27]	2019	✓			✓		
Bao et al. [28]	2019	✓				✓	
Dissanayake et al. [21]	2020	✓			✓		
Latif et al. [10]	2020	✓			✓		
Parthasarathy et al. [9]	2020	✓			✓		
Shukla et al. [29]	2021			✓			
Ahn et al. [30]	2021	✓					
Zhou et al. [31]	2021			✓			
Chen et al. [32]	2021		✓	✓			
Su et al. [33]	2022	✓				✓	
Latif et al. [20]	2022	✓			✓	✓	
Gao et al. [34]	2022	✓			✓		
Takashima et al. [35]	2022		✓	✓			
Rajapakshe et al. [36]	2022	✓					
Latif et al. [11]	2022	✓			✓	✓	
Arezzo et al. [12]	2022	✓	✓		✓		
Hsu et al. [37]	2023		✓	✓			
Wang et al. [38]	2023		✓	✓			
Luo et al. [39]	2023		✓	✓			
Gao et al. [16]	2023	✓	✓		✓		
Ours		✓	✓	✓	✓	✓	✓

3 METHODOLOGY

We propose a multitask transformer technique to tackle cross-corpus speech emotion recognition (SER) and cross-corpus multimodal emotion recognition (MER). The centre of the technique is a multitask framework embodying a pre-trained transformer. The constrastive learning and the information maximization (IM) loss, key novelties in our multitask framework, attracts or repels data samples according to similarity or cluster membership. Which is a intuitive way to form classification boundary without emotional labels. Furthermore, the pre-trained transformer is adopted as a backbone due to capability of contextual learning [5] and generalizability of pre-trained model [3]. Our model consists of an audio and text multitask transformer, as depicted in Fig.1. Each transformer is separately trained. Multimodal emotion recognition is achieved by performing decision-level fusion by adding logits propagated from each transformer during inference. In the following subsections, we will describe the components of our model.

3.1 Audio multitask transformer

3.1.1 Data augmentation

We devised five types of augmentation ($\{a_i,\ i = 1\dots5\}$) as presented in Table 2. The augmentation is performed using APIs from 'torch-audiomentations' [45] with same name of waveform transformation methods with its parameter setting. This approach boosts the volume of data by five times. Augmentation is performed twice per data sample to obtain two augmented samples, as two parallel streams are needed for contrastive learning. Augmentation labels are assigned according to augmentation type, and later used in an augmentation-type classifier. Let us denote the augmented audio input signal as x_a. Previous work [11] showed that SER performance is affected largely due to changes in data size but a little due to the type of augmentation used. Thus, we used conventional method based on signal transformation for augmentation, not sophisticated methods such as mixup [46]. Moreover, we increased the number of data samples to multiple of five, while previous work [11] only increased to the multiple of four. We intend to reconfirm the hypothesis that the increased sample size is more significant than augmentation method. Thus, we changed augmentation methods to simpler transformation and increased from factor of four to factor of five, to ensure that the observed performance improvement is legitimately caused by the sample size increment.

3.1.2 Pre-trained Audio Spectrogram Transformer (AST) [5]

Audio Spectrogram Transformer (AST) is a transformer based model architecture that takes spectrogram as input rather than raw waveform of audio. Considering that spectrogram can be visualized as an image, AST has an architecture similar to Vision Transformer (ViT), with necessary modifications to cater to different requirement from image data obtained with spectrogram [5]. The originally proposed AST was first pretrained on ImageNet and finetuned on AudioSet [47] data, an audio event classification dataset collected from YouTube videos [5]. In our work, we adopt the AST fine-tuned on AudioSet data as pre-trained AST and further fine-tuned on SER task.

The raw input waveforms are transformed into a log-mel-spectrogram with a 128-dimension feature size, 25ms window, and 10ms stride. Input features are passed through

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3526592

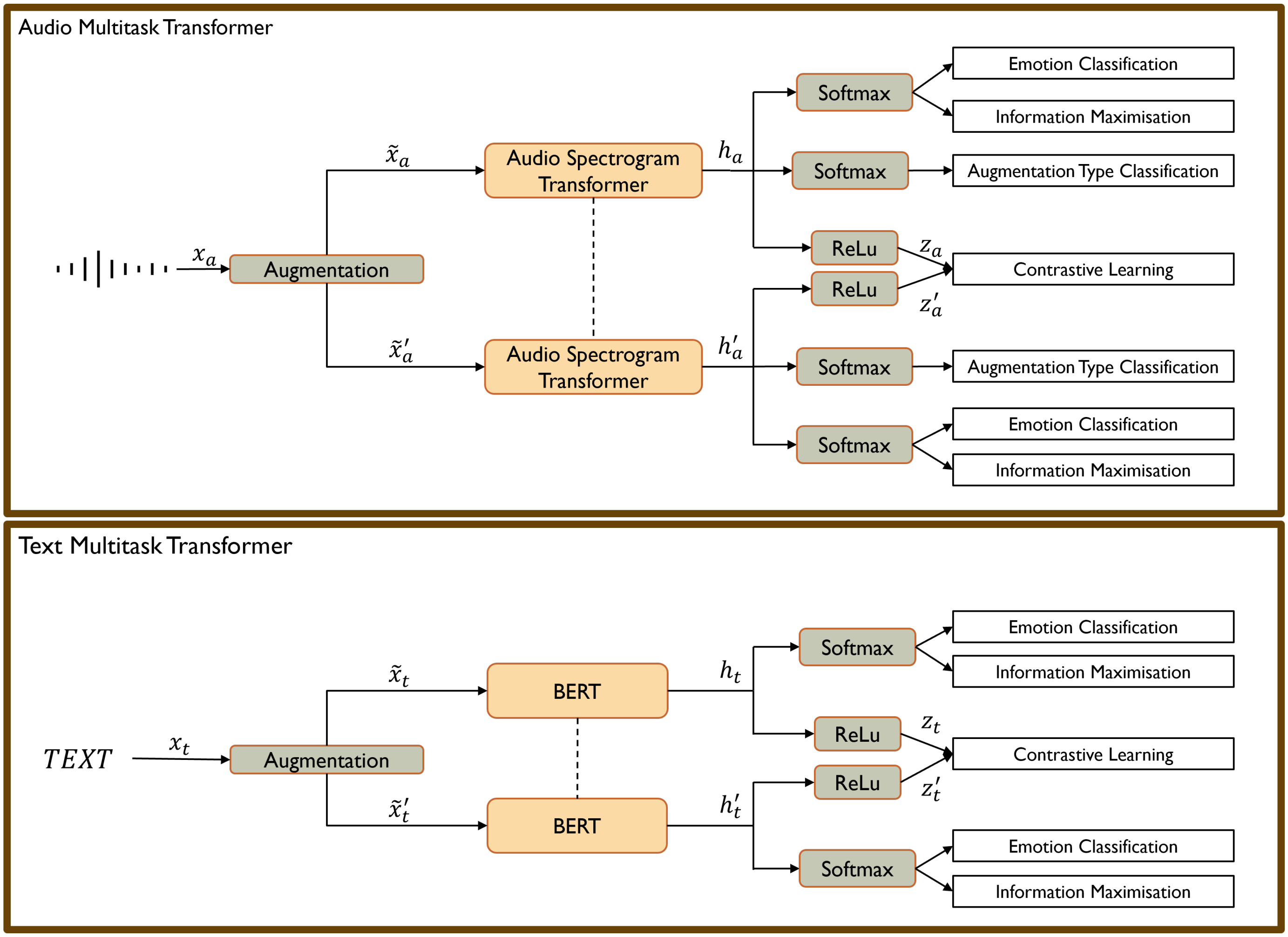

Fig. 1. The proposed multimodal approach to SER, consisting of (i) audio multitask transformer and (ii) text multitask transformer. The two multitask transformers are trained separately and their outputs are fused at the decision layer during prediction.

TABLE 2
Augmentation types used for audio multitask transformer adopted from 'torch-audiomentations' [45]

Augmentation Type	Waveform Transformation
a_1	Gain(-15db $\sim$ 5db,p=0.3), PolarityInversion(p=0.3)
a_2	Gain(-15db $\sim$ 5db,p=0.3), PolarityInversion(p=0.3),Shift(p=0.3)
a_3	Gain(-15db $\sim$ 5db,p=0.3), PolarityInversion(p=0.3), TimeInversion(p=0.3)
a_4	BandStopFilter(p=0.3), PeakNormalization(p=0.3), Shift(p=0.3)
a_5	Gain(-15db $\sim$ 5db,p=0.3), PolarityInversion(p=0.3), AddColoredNoise(p=0.3)

patch embedding layer before the transformer layers, which consists of 12 transformer layers. We sum up a total of 13 representations corresponding to the patch embedding layer and the intermediate representations of transformers layers. The sum is computed with learnable weights. The resulting representation is average-pooled to obtain the final representation before the classification layer. The final representation before the classification layer is referred to as the AST output h_a:

$$h_a = f_{ast}(x_a). \tag{1}$$

where f_{ast} denotes the AST transfer function.

3.1.3 Speech emotion recognition task

The primary task is the SER prediction of the emotional class c from the AST output h_a. The class label is predicted by a fully-connected layer, which can be formulated as

$$\hat{y}_a = \mathrm{softmax}(Wh_a + b). \tag{2}$$

The weights W and biases b of the fully connected layers are learned during training. The weights and biases are learned by minimizing the cross-entropy loss $\mathcal{L}_{SER}$:

$$\mathcal{L}_{SER} = -E\{\sum_c 1(y_a = c)log(\hat{y}_a)\}. \tag{3}$$

where y_a denotes the ground truth label for augmented input x_a, summation is taken over all class labels c, and E computes the expectation over sample space.

We introduce three other losses (tasks) to enhance performanc of the cross-corpus SER.

3.1.4 Contrastive learning loss

The network propagates two AST outputs from two augmented inputs, denoted as h_a and h'_a. Then we compute contrastive learning (CL) loss, which effectively empowers AST output to attract positive data samples and repel negative pairs [22], [23]. The positive pairs are formed by two AST outputs from two augmented samples, which are computed from single data samples. The negatives pairs are formed by pairing with other samples from the minibatch.

Inspired from Chen et al. [23], we compute the latent variable z by feeding h through a single hidden layer feed-forward network as:

$$z_a = V\sigma(Uh_a), \tag{4}$$

where σ is the rectified linear unit (ReLU) activation. Similarly, we obtain z'_a from h'_a from the second augmented input. Then, we compute the CL loss:

$$\mathcal{L}_{CL} = -log\frac{\exp\left(sim(z_a^i, {z'}_a^i)/\tau\right)}{\sum_{j=1}^{B}\exp\left(sim(z_a^i, {z'}_a^j)/\tau\right)}. \tag{5}$$

where z_a^i refers to the ith sample of the latent variable z_a in the mini batch and z_a^j refers to z_a from the forward pass of the jth samples of the minibatch. sim refers to the cosine similarity score [1], which is a inner-product between unit normalized vectors. The batch size is referred tok as B.

The hyperparameter τ is the temperature [48], which controls the smoothness of the softmax function.

3.1.5 Information Maximization (IM) loss

Our approach intends to learn clusters from the AST outputs, inspired by Krause et al. [24] and Liang et al. [49]. We compute the IM loss by computing expected entropy regularized by the empirical label's entropy.

$$\mathcal{L}_{IM} = -E\{\hat{y}_a log(\hat{y}_a)\} + E\{\hat{p}_a log(\hat{p}_a)\}, \tag{6}$$

where the empirical label distribution $\hat{p}_a$ is obtained by simply averaging over softmax output from data points $\hat{y}_a^j$ in the minibatch, containing $\hat{y}_a$: $\hat{p}_a = \frac{1}{B}\sum_j^B \hat{y}_a^j$. The first term of equation (6) is the entropy over the samples $\hat{y}_a$. In other words, the same head is used for both $\mathcal{L}_{SER}$ and $\mathcal{L}_{IM}$.

Supported by the cluster assumption [24], classifiers are expected to be more robust to differences between corpora. Note that the loss function in (6) does not need class labels, which is different from the to cross-entropy loss.

3.1.6 Augmentation-type classification loss

The third task, inspired from [11], is defined using labels assigned from the augmentation function:

$$\mathcal{L}_{aug} = -E\{\sum_{a_i} 1(y_{aug} = a_i)log(\hat{y}_{aug})\}. \tag{7}$$

where $\hat{y}_{aug}$ denotes the output from a fully connected layer predicting the augmentation label and y_{aug} represents true augmentation type label from $\{a_i,\ i = 1 \dots 5\}$ (Table 2).

3.1.7 Total loss

The total loss for the training objective is computed by summing up all the losses mentioned above for training.

$$\mathcal{L}_{total} = \mathcal{L}_{emo} + \lambda_{CL}\mathcal{L}_{CL} + \lambda_{IM}\mathcal{L}_{IM} + \lambda_{aug}\mathcal{L}_{aug} \tag{8}$$

where λ_{CL},λ_{IM}, and λ_{aug} are hyperparameters that weigh the importance of each loss and are to be empirically determined.

3.2 Text multitask transformer

Text multitask transformers share a similar architecture to the audio multitask transformer. So, in this subsection, we only describe the components that are not common between the two transformers.

3.2.1 Data augmentation

Our approach also include data augmentation for the text, but we need to ensure that the augmentation does not alter the affective information. We employ the augmentation technique 'token cutoff' [50], where 'cutoff' randomly removes the information from the input embedding matrix in a more structured manner. For example, for BERT [51], the input embedding matrix consists of tokens. To make sure that no information corresponding to the removed tokens is left, token cutoff converts embedding indices to 0. The augmentation-type classifier is not used for text, as only one type of augmentation is used. For the sake of consistency with the audio multitask transformer, the text augmentation was performed five times.

3.2.2 Pre-trained BERT [51]

BERT is well-known transformer architecture pre-trained for general NLP tasks, pre-trained on BooksCorpus [52] and text data collected from English Wikipedia pages.

A BERT-specific tokenizer tokenizes a text transcript. Similar to the audio multitask transformer, the text multitask transformer consists of total 13 representations, representation after patch embedding layer and every intermediate representations of transformer layers, are summed up with learnable weights. The input for text is denoted by y, and the corresponding BERT output is denoted by h_t.

3.2.3 Total loss

The total for the training objective for text is computed similarly to the total loss for the audio multitask transformer. The only exception is that we do not consider the augmentation-type loss, as aforementioned.

3.2.4 Decision level fusion

Our model achieves multimodal emotion recognition by fusing transformer output from each modality only during inference. Decision-level fusion is performed by adding logits forward propagated from AST and BERT outputs. To differentiate the components from each modality, we denote

1. $sim(a,b) = \frac{a^\top b}{\|a\|\|b\|}$

the modality of each term with superscript, a,t for audio and text, respectively.

$$\hat{y}_{at} = \mathrm{softmax}(W_a h_a + b_a + W_t h_t + b_t). \tag{9}$$

where W_a and b_a are the biases and weights for the audio stream, and W_t and b_t are the biases and weights for the text stream, respectively, at the fusion layer.

4 Experimental Setup

4.1 Datasets

We use three publicly available datasets: the IEMOCAP, MSP-IMPROV and EMO-DB corpora, which are widely used in SER research.

4.1.1 IEMOCAP [53]

This corpus consists of approximately 12 hours of conversation between two actors in English. Ten actors were recruited to record five sessions in a laboratory environment. Each session consists of two actors' dialogues prepared with a script or improvisation with minimal context provided. Each turn was segmented and saved as an utterance. However, not all utterances are used in our experiment, as some emotion categories have few samples. To keep emotion labels consistent with other datasets, we focused on four basic emotions: neutral, happiness (excited was incorporated into happiness), sadness, and anger. As a result, we selected 5,531 utterances consisting of neutral (1,708), happiness (1,636), sadness (1,084), and anger (1,103) for the experiment. We obtained raw 'wav' format recordings from each utterance for audio input and ground truth transcripts for text input. We randomly shuffled the 5,531 utterances and split them according to the experiment settings, which will be described later in this paper.

4.1.2 MSP-IMPROV [54]

This dataset was constructed similarly to the IEMOCAP dataset. It consists of recording dyadic conversations between 12 actors in English. Each session was recorded from two actors' dialogue and segmented per each turn. However, the emotion elicitation methodology is the difference between the IEMOCAP and MSP-IMPROV datasets. The target sentence was defined for MSP-IMPROV, and actors were required to express different emotions with the same sentence. To ensure natural elicitation, actors were asked to improvise and build up the situation where the target sentence could naturally be expressed with the intended emotion. Also, actors were asked to read the target sentence with the intended emotion to incorporate the difference between the read and acted sentences in the dataset. The dataset includes recording from the improvised sentences to elicit target emotion as well. The intended emotion categories were only four: neutral, happiness, sadness, and anger. Although a few utterances were labelled with other emotions, we only used neutral, happiness, sadness, and anger. As a result, we used 8438 utterances consisting of neutral (3477), happiness (2644), sadness (885), and anger (792). Like the IEMOCAP, we obtained raw 'wav' format recording for audio (needed resampling from 44100Hz to 16000Hz) and ground truth transcript as text as the input of the model during training.

4.1.3 EMO-DB [55]

This corpus is an emotional speech dataset collected from 10 German-speaking actors, which is considered to be a foundational dataset for the SER problem. It consists of 800 utterances with seven emotions and was recorded in a tightly controlled environment. Thus, the speech recording has the least noise and ambiguity in emotion. However, to keep a consistent emotion category, we only used 535 utterances consisting of neutral (275), happiness (71), sadness (127), and anger (62). Note that EMO-DB is in German while other datasets are in English. Its transcript were intentionally curated to be neutral. For these reasons, EMO-DB cannot be used for text-based emotion recognition experiments.

4.2 Experiment settings

The study involves two sets of experiments to evaluate the model. The first experiment is a *cross-corpus SER*, which trains an audio multitask transformer on the source data for the SER problem, adapts to the target train data without emotional labels, and then tests the model for inference on the target test data. The second experiment is a *cross-corpus multimodal emotion recognition*, where a text multitask transformer is trained on the source data for the text emotional recognition problem and adapts to the target train data without emotion labels. The model is then tested on the target test data using both audio and text transformers with decision-level fusion to predict emotion labels. The data augmentation pipeline is applied only to the training data, and it explicitly multiplies the number of samples inside each epoch.

The source data was split into 90% training set (S_{tr}) and 10% validation set (S_{va}), while 30% of the target dataset without labels was used as the training set (T_{tr}), and the remaining portion was reserved for the testing set (T_{te}). The model was trained on S_{tr} and T_{tr} alternatively per each epoch, and the training was halted when the accuracy did not improve from the previous epoch. We record the unweighted average recall (UAR) on T_{te}. Each experiment result presented in the next section is obtained by running ten rounds of randomised runs of training and testing followed by a statistical summary.

4.2.1 Audio multitask transformer model configuration

We set the hyperparameters of the total loss empirically as follows: $\lambda_{CL} = 0.5$, $\lambda_{IM} = 0.5$, and $\lambda_{aug} = 0.1$. The values were determined during preliminary experiment where we only use source data to check the validity of the model. Afterwards, value were fixed and never altered in the following experiments for cross-corpus training. To train the model, we use a pre-trained transformer downloaded from 'https://huggingface.co/MIT/ast-finetuned-audioset-10-10-0.4593'. The learning rate is set to 1×10^{-4} and decayed on a plateau by a factor of 0.1 with patience of 5. We measured the validation accuracy every 30 batch instead of calculating it per epoch. The batch size is set to 64. Data augmentation is performed before preprocessing, which involves converting the raw waveform into a spectrogram. The dimension of the spectrogram feature is set to 128, and the maximum sequence length is $1,024$. We use 12 transformer layers and 12 attention heads. The hidden dimension is 768, and the size of the latent variable z is 128.

4.2.2 Text multitask transformer model configuration

To train the model, we set the hyperparameters for the total loss weights as follows: $\lambda_{CL} = 0.5$, $\lambda_{IM} = 0.5$, and $\lambda_{aug} = 0.0$. λ_{aug} is set to zero as we do not use an augmentation type classifier for text and the others were set equivalent to audio. We download the BERT model from 'https://huggingface.co/bert-base-uncased'. The learning rate and decay scheduler follow the same structure as the audio transformer, starting with 1×10^{-4} and decaying by a factor of 0.1. However, the batch size was set to 512, so we measured validation accuracy in every $5th$ batch for the source data and every $3rd$ batch for the target data. The learning rate scheduler was set with the patience of 5. The preprocessor tokenized the sentence, and data augmentation was computed after. We performed cutoff [50] data augmentation by randomly masking the tokens with a 0 index with a probability of 0.15. The number of transformer layers, attention heads, hidden size, and latent variable size were set identically to those used for audio.

4.2.3 Multimodal transformer inference model configuration

The transformer models were trained modality-wise for multimodal inference during final testing. However, in the real world, target data will not be guaranteed to be provided with a transcript. The model was therefore evaluated using transcribed text using a pre-trained automatic speech recognition (ASR) model downloaded from 'https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english'. The downloaded model is based on the Wav2Vec2 architecture, and a language model pipeline was implemented as post-processing to obtain the final transcription.

5 Results

5.1 Cross-corpus SER experiment

5.1.1 IEMOCAP to MSP-IMPROV

The experiment for cross-corpus SER was initially carried out with the IEMOCAP dataset as the source data and the MSP-IMPROV dataset as the target data. The IEMOCAP corpus is a widely used dataset for SER and cross-corpus SER research, and it is known to have a relatively balanced distribution of emotion classes. On the other hand, the MSP-IMPROV corpus has a comparable structure to the IEMOCAP but with a larger volume of data. This experiment configuration is considered standardized, as it has been studied more extensively than other combinations of datasets [10], [11], [20], [56].

The results are presented in Table 3 and they signify the superiority of our approach against existing studies. Many of the previous works adopted adversarial learning [20], [56], either using gradient reversal layer or minimax game to learn similar representation between source and target data. Our improved results indicate that unsupervised learning based multitask learning can outperform adversarial learning based multitask learning.

TABLE 3
Cross-corpus SER experiment on IEMOCAP to MSP-IMPROV

Model	UAR
GAN [56]	43.6±1.3
Semi-supervisedAAE [10]	46.4±0.32
ADDi [20]	45.1±0.8
sADDi [20]	47.1±0.5
MTL-AUG [11]	48.1±0.30
Ours	**50.3±1.3**

TABLE 4
Cross-corpus SER experiment on MSP-IMPROV to IEMOCAP

Model	UAR
GAN [56]	45.8±1.5
ADDi [20]	48.2±0.6
sADDi [20]	49.8±0.6
Ours	**54.0±2.0**

5.1.2 MSP-IMPROV to IEMOCAP

Our method consistently outperforms state-of-the-art in the case of MSP-IMPROV to IEMOCAP experiments (refer to Table 4). Most of the previous works are observed to adopt adversarial learning, while most recent work attempted to use more tasks as in multitask learning with data augmentation [20]. Likewise, our method exploits multitask learning and data augmentation but adopts transformer and contrastive learning to achieve significant improvement of 5% over the state-of-the-art. Since the MSP-IMPROV has larger volume of data, every method in our comparison yields improvement over the case of the IEMOCAP to MSP-IMPROV experiments.

5.1.3 IEMOCAP to EMO-DB

TABLE 5
Cross-corpus SER experiment on IEMOCAP to EMO-DB

Model	UAR
GAN [56]	44.3±1.7
ADDi [20]	46.1±1.6
MTL-AUG [11]	46.8±1.4
sADDi [20]	48.3±1.5
Ours	**61.8±12.0**

TABLE 6
Impacts on duplicating data samples on IEMOCAP to EMO-DB in SER experiments

Duplication amount	UAR
Original size	52.3±7.4
Duplicated $\times 2$	54.3±9.1
Duplicated $\times 4$	58.3±7.7
Duplicated $\times 6$	59.5±8.7
Duplicated $\times 8$	61.8±12.0

Next, we run cross-corpus SER experiment from IEMOCAP to EMO-DB. EMO-DB is another common dataset based on German language. First, we observed that the number of data samples in the IEMOCAP dataset is significantly higher than that of the EmoDB corpus. So, we hypothesized that naively duplicating data samples of the

EmoDB corpus until its size is similar to that of the IEMOCAP corpus can improve performance. We progressively increased the dataset size by a factor of two, and we observed that the performance was optimal when it duplicated eight items.

As shown in Table 5 our methods achieves significant improvement in the IEMOCAP to EMO-DB experiment. The results of our method in Table 5 is obtained after additional experiments in Table 6. We progressively duplicated the EMO-DB corpus in rate of $2, 4, 6, 8$ to examine the effect of multiplicating the volume of the dataset. Although we are duplicating the dataset, we can safely assume that we are increasing dataset size as we exploit data stochastically. The overall results proves that our method outperforms existing methods even without duplication of the dataset. Duplicating the dataset size 8 times yielded the highest performance, yet it is worth to note that the variance of the performance also increased. Also, we observed that cross-corpus SER performance started to decrease when duplication factor is above 8.

5.1.4 EMO-DB to IEMOCAP

We also ran cross-corpus SER experiment from EMO-DB to IEMOCAP. Again, we ran additional experiments by progressively duplicating the data samples of EMO-DB. We observed the similar trends, when compared with IEMOCAP to EMO-DB, that

TABLE 7
Cross-corpus SER experiment on EMO-DB to IEMOCAP

Model	UAR
DANN [41]	40.5±2.0
GAN [56]	40.3±1.7
ADDi [20]	41.2±1.8
sADDi [20]	44.8±1.6
MTL-AUG [11]	41.5±1.6
Ours	**45.9±3.2**

TABLE 8
Impacts on duplicating data samples on EMO-DB to IEMOCAP in SER experiments

Duplication amount	UAR
Original size	44.1±3.3
Duplicated $\times 2$	45.5±1.3
Duplicated $\times 4$	44.2±3.2
Duplicated $\times 6$	43.0±3.0
Duplicated $\times 8$	45.9±3.2

Our results shows consistent superiority against existing studies in Table.7. EMO-DB is a small dataset and it was repeatedly augmented, but still our method outperformed state-of-the-art. The performance reached the best scores when we duplicated dataset 8 times. The performance without duplication was slightly lower then some state-of-the-art methods [20].

5.2 Multimodal cross-corpus SER experiment

Multimodal experiment results can be found in Table 9. As explained in the last paragraph, 'Audio + Text' used the same transformer for 'Audio' and 'Text' with decision-level fusion for final prediction. No training was done for multimodal fusion, and accuracy improvement was observed after decision-level fusion. Our experiment results show that our methods improve over uni-modal cross-corpus SER; we obtained 4% improvement adding the text modality to cross-corpus SER.

TABLE 9
Multimodal experiment on IEMOCAP to MSP-IMPROV

Model	UAR
Audio only	50.3±1.3
Text only	40.4±1.2
Audio + Text	54.4±0.98

TABLE 10
Comparison with existing cross-corpus multimodal emotion recogntion on IEMOCAP to MSP-IMPROV

Model	UAR
Sahu et al [26]: Audio + Text(transcribed from ASR)	40.08
Ours: Audio + Text(transcribed from ASR)	**52.9±1.0**

Although our model is trained on a dataset with the ground-truth transcript provided, this might be unrealistic. To our knowledge, only one paper has studied cross-corpus SER on audio and text [26], which also assumes that no ground-truth transcript is provided. For the sake of comparison against this work, we conducted additional experiments. We assume that the ground truth transcript is difficult to obtain, so the model is trained with ground truth transcript using Automatic Speech Recognition(ASR) output instead of ground truth during testing.

The results are presented in Table 10, showing that our method also outperforms existing studies. Furthermore, our multimodal cross-corpus SER method has only a 2% decrease in accuracy when we used transcribed text instead of ground truth transcript, depicting the robustness of our method.

5.3 Ablation study

TABLE 11
Ablation study of cross-corpus SER on IEMOCAP to MSP-IMPROV

Model	UAR
All tasks ($\lambda_{aug} = 0.1$)	**50.3±1.3**
Exclude Contrastive Learning $\mathcal{L}_{CL}$	47.0±2.4
Exclude Information Maximization $\mathcal{L}_{IM}$	47.3±1.3
Exclude Augmentation $\mathcal{L}_{aug}$	50.5±1.7
All tasks ($\lambda_{aug} = 0.5$)	47.1±2.5
Replace AST to Wav2Vec2	46.0±3.2
Training without target data	45.6±2.8

We conducted an ablation study focused on the multitask framework, in which we intend to examine the relative importance of each task in multitask learning. The results are described in Table 10. The ablation study is conducted on the IEMOCAP to MSP-IMPROV setting.

The results show that contrastive learning and information maximization (IM) loss significantly contribute to

TABLE 12
Ablation study cross-corpus text emotion recognition on IEMOCAP to MSP-IMPROV

Model	UAR
all tasks	40.4±1.2
exclude Contrastive Learning $\mathcal{L}_{CL}$	39.8±0.6
exclude Information Maximization $\mathcal{L}_{IM}$	38.8±0.6

achieving overall performance. Contrastive learning and IM loss are capable of learning the structure of the dataset, which enables efficient transfer of emotion classification knowledge to the target dataset. The exclusion of augmentation type classification yielded an increase in variance but not much change in the mean of UAR.

We vary λ_{aug} to see how performance vary and found at 0.5 and found the performance dropped significantly. We also validated our choice of using AST over other transformer model like Wav2Vec2 and confirmed that switching to AST yielded significant improvement in the performance. We suspect that Wav2Vec2 is geared toward speech recognition, focusing on the elements (most likely phonemes) within a sequence. On the contrary, AST focuses on audio classification inferring from whole sequence. Therefore, AST performs better than Wav2Vec2 in SER, where emotion is inferred from a whole utterance rather than on phonemes. Finally, we examined the significance of training with target data (without emotional labels) by removing the target data during training. Removing the target data in training resulted in lower performance.

As seen in Table 12, an ablation study was conducted on text modality as well. Similar to audio modality, both contrastive learning and information maximization loss have significant impact on the performance.

6 Conclusion and Future Work

In conclusion, the challenge of cross-corpus speech emotion recognition (SER) and cross-corpus multimodal emotion recognition (MER) has been notably addressed through the novel approach presented in this paper. Our primary contribution was introducing contrative learning and information maximization (IM) loss to the area of cross-corpus SER. With these two unsupervised tasks combined, we were able to achieve successful transfer of emotion classification knowledge from source dataset to target dataset. Furthermore, this methods were implemented on pre-trained transformer, which have been recently shown successes in SER area, as a multitask transformer for cross-corpus SER.

Secondly, we adapt the framework of multitask transformer for cross-corpus SER to text modality with minor modification. We also achieved cross-corpus MER with simple decision level fusion. Cross-corpus evaluation of MER has not been explored before, and even with simple extension from cross-corpus SER to cross-corpus MER achieved significant improvement of 4%.

Empirical results confirm the efficacy of our proposed approach, showcasing superior performance compared to the best-reported results in cross-corpus SER. Furthermore, targeted experiments reveal that the key contributors to this performance enhancement are the auxiliary tasks of contrastive learning and information maximization loss. This advances our understanding of the mechanism behind the observed improvements and underscores the potential of these auxiliary tasks in refining cross-corpus SER and MER systems.

Our experimental results have demonstrated the applicability of our approach to three widely used datasets for Speech Emotion Recognition (SER): IEMOCAP [53], MSP-IMPROV [54], and EMO-DB [55]. While these datasets are commonly used in the field, they are all collected in controlled laboratory environments, which limit their generalizability to real-world scenarios. This limitation suggests the necessity of further research using more realistic datasets, such as MSP-PODCAST [57].

MSP-PODCAST is a naturalistic dataset of substantial size, comprising 238 hours of speech recordings, compared to IEMOCAP's 12 hours. This large volume introduces additional complexity when testing cross-corpus transfer methods. Specifically, the MSP-PODCAST dataset captures more diverse emotional expressions and varied recording conditions typical of natural settings, increasing the challenge for SER models. Studies have shown that this increased variability leads to a higher variance between the training and test datasets, making it harder to achieve the same performance levels as those obtained with controlled datasets [57].

Compared to datasets like MSP-IMPROV, MSP-PODCAST exhibits a significantly larger variance between the training and test data, further complicating model generalization. This makes any performance degradation observed on MSP-PODCAST normal and expected when transitioning from controlled datasets to real-world naturalistic datasets. Given this dataset's inherent challenges, it is crucial to note that performance in these conditions often reflects the complexity of the task rather than the effectiveness of the method itself.

We conducted preliminary experiments using the IEMOCAP dataset as the source and MSP-PODCAST as the test dataset, training a multitask transformer model that incorporates both modalities (refer to Table 13). To the best of our knowledge, this represents the first attempt to apply cross-corpus Multimodal Emotion Recognition (MER) on MSP-PODCAST data. Given the challenges posed by the dataset, the results achieved are promising and provide a solid foundation for further research.

In future work, we plan to enhance our approach by exploring advanced techniques, such as aggregating multiple source datasets [33], [58], and employing data augmentation methods. These efforts aim to mitigate the inherent complexities of working with naturalistic datasets like MSP-PODCAST and further push the performance boundaries.

TABLE 13
Experiment results on IEMOCAP to MSP-PODCAST

Model	UAR
Audio only	43.3±0.7
Text only	41.7±0.9
Audio + Text	46.0±0.9

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3526592

The results in Table 13 are expected to serve as a benchmark for evaluating the effectiveness of these new methods.

REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] 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.

[3] 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.

[4] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023.

[5] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.

[6] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[7] R. Milner, M. A. Jalal, R. W. Ng, and T. Hain, "A cross-corpus study on speech emotion recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 304–311.

[8] S. Zhang, R. Liu, X. Tao, and X. Zhao, "Deep cross-corpus speech emotion recognition: Recent advances and perspectives," *Frontiers in neurorobotics*, vol. 15, p. 784514, 2021.

[9] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 2697–2709, 2020.

[10] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Transactions on Affective computing*, vol. 13, no. 2, pp. 992–1004, 2020.

[11] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Multitask learning from augmented auxiliary data for improving speech emotion recognition," *IEEE Transactions on Affective Computing*, 2022.

[12] A. Arezzo and S. Berretti, "Speaker vgg cct: Cross-corpus speech emotion recognition with speaker embedding and vision transformers," in *Proceedings of the 4th ACM International Conference on Multimedia in Asia*, 2022, pp. 1–7.

[13] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Adapting a self-supervised speech representation for noisy speech emotion recognition by using contrastive teacher-student learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] M. Tran and M. Soleymani, "A pre-trained audio-visual transformer for emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4698–4702.

[15] V. S. Alaparthi, T. R. Pasam, D. A. Inagandla, J. Prakash, and P. K. Singh, "Scser: Supervised contrastive learning for speech emotion recognition using transformers," in *2022 15th international conference on human system interaction (HSI)*. IEEE, 2022, pp. 1–7.

[16] Y. Gao, L. Wang, J. Liu, J. Dang, and S. Okada, "Adversarial domain generalized transformer for cross-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, 2023.

[17] J. Gideon, M. G. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog)," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1055–1068, 2019.

[18] J. Zhang, L. Jiang, Y. Zong, W. Zheng, and L. Zhao, "Cross-corpus speech emotion recognition using joint distribution adaptive regression," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3790–3794.

[19] H. Luo and J. Han, "Nonnegative matrix factorization based transfer subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2047–2060, 2020.

[20] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Transactions on Affective Computing*, 2022.

[21] V. Dissanayake, H. Zhang, M. Billinghurst, and S. Nanayakkara, "Speech emotion recognition 'in the wild'using an autoencoder," *Interspeech 2020*, 2020.

[22] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[24] A. Krause, P. Perona, and R. Gomes, "Discriminative clustering by regularized information maximization," *Advances in neural information processing systems*, vol. 23, 2010.

[25] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7390–7394.

[26] S. Sahu, V. Mitra, N. Seneviratne, and C. Y. Espy-Wilson, "Multimodal learning for speech emotion recognition: An analysis and comparison of asr outputs with ground truth transcription." in *Interspeech*, 2019, pp. 3302–3306.

[27] H. Luo and J. Han, "Cross-corpus speech emotion recognition using semi-supervised transfer non-negative matrix factorization with adaptation regularization." in *INTERSPEECH*, 2019, pp. 3247–3251.

[28] F. Bao, M. Neumann, and N. T. Vu, "Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition." in *Interspeech*, 2019, pp. 2828–2832.

[29] A. Shukla, S. Petridis, and M. Pantic, "Does visual self-supervision improve learning of speech representations for emotion recognition," *IEEE Transactions on Affective Computing*, 2021.

[30] Y. Ahn, S. J. Lee, and J. W. Shin, "Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation," *IEEE Signal Processing Letters*, vol. 28, pp. 1190–1194, 2021.

[31] H. Zhou, J. Du, Y. Zhang, Q. Wang, Q.-F. Liu, and C.-H. Lee, "Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2617–2629, 2021.

[32] B. Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu, and D. Zhang, "Multimodal emotion recognition with temporal and semantic consistency," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3592–3603, 2021.

[33] B.-H. Su and C.-C. Lee, "Unsupervised cross-corpus speech emotion recognition using a multi-source cycle-gan," *IEEE Transactions on Affective Computing*, 2022.

[34] Y. Gao, S. Okada, L. Wang, J. Liu, and J. Dang, "Domain-invariant feature learning for cross corpus speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6427–6431.

[35] A. Takashima, R. Masumura, A. Ando, Y. Yamazaki, M. Uchida, and S. Orihashi, "Interactive co-learning with cross-modal transformer for audio-visual emotion recognition," *Proc. Interspeech 2022*, pp. 4740–4744, 2022.

[36] T. Rajapakshe, R. Rana, and S. Khalifa, "Domain adapting speech emotion recognition modals to real-world scenario with deep reinforcement learning," *arXiv preprint arXiv:2207.12248*, 2022.

[37] J.-H. Hsu and C.-H. Wu, "Applying segment-level attention on bimodal transformer encoder for audio-visual emotion recognition," *IEEE Transactions on Affective Computing*, 2023.

[38] S. Wang, Y. Ma, and Y. Ding, "Exploring complementary features in multi-modal speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[39] M. Luo, H. Phan, and J. Reiss, "cross-modal fusion techniques for utterance-level emotion recognition from text and speech," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[40] P. Mote, B. Sisman, and C. Busso, "Unsupervised domain adaptation for speech emotion recognition using K-Nearest neighbors voice conversion," in *Interspeech 2024*, Kos Island, Greece, September 2024.

[41] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[42] A. Reddy Naini, M. Kohler, E. Richerson, D. Robinson, and C. Busso, "Generalization of self-supervised learning-based representations for cross-domain speech emotion recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024)*, Seoul, Republic of Korea, April 2024, pp. 12 031–12 035.

[43] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[44] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso, "Versatile audio-visual learning for emotion recognition," *IEEE Transactions on Affective Computing*, vol. to appear, 2024.

[45] I. Jordal, S. ES, H. BREDIN, K. Nishi, F. Lata, H. C. Blum, P. Manuel, akash raj, K. Choi, FrenchKrab, P. Żelasko, amiasato, M. L. Quatra, and E. Schmidbauer, "asteroid-team/torch-audiomentations: v0.11.0," Jun. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6778064

[46] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[47] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[48] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[49] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *International Conference on Machine Learning (ICML)*, 2020, pp. 6028–6039.

[50] D. Shen, M. Zheng, Y. Shen, Y. Qu, and W. Chen, "A simple but tough-to-beat data augmentation approach for natural language understanding and generation," *arXiv preprint arXiv:2009.13818*, 2020.

[51] 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.

[52] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[53] 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, pp. 335–359, 2008.

[54] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[55] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[56] S. Latif, J. Qadir, and M. Bilal, "Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition," in *2019 8th international conference on affective computing and intelligent interaction (ACII)*. IEEE, 2019, pp. 732–737.

[57] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[58] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards speech emotion recognition" in the wild" using aggregated corpora and deep multi-task learning," in *18th Annual Conference of the International Speech Communication Association, INTERSPEECH 2017: Situated interaction*. International Speech Communication Association (ISCA), 2017, pp. 1113–1117.

Chung-Soo Ahn received the B.S. degrees in industrial engineering from Ajou University, Suwon, Gyonggi-do, South Korea, in 2014. And he received M.Eng. degree in electrical and electrical and electronic engineering from Nanyang Technological University, Singapore in 2018. He is currently pursuing Ph.D. degree under College of Computing and Data Science with Nanyang Technological University, Singapore. His research are in the area of speech emotion recognition, multimodal deep learning, and domain adaptation.

Rajib Rana is an experimental computer scientist, Advance Queensland Research Fellow and a Senior Lecturer in the University of Southern Queensland. He is also the Director of the IoT Health research program at the University of Southern Queensland. He is recipient of the prestigious Young Tall Poppy QLD Award 2018 as one of Queensland's most outstanding scientists for achievements in the area of scientific research and communication. Rana's research work aims to capitalise on advancements in technology along with sophisticated information and data processing to better understand disease progression in chronic health conditions and develop predictive algorithms for chronic diseases, such as mental illness and cancer. His current research focus is on Unsupervised Representation Learning. He received his B. Sc. degree in Computer Science and Engineering from Khulna University, Bangladesh, with the Prime Minister and President's Gold medal for outstanding achievements and his Ph. D. in Computer Science and Engineering from the University of New South Wales, Sydney, Australia in 2011. He received his postdoctoral training at Autonomous Systems Laboratory, CSIRO before joining the University of Southern Queensland in 2015.

Carlos Busso (S'02-M'09-SM'13-F'23) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. Carlos Busso is a Professor at Language Technologies Institute, Carnegie Mellon University, where he is also the director of the Multimodal Speech Processing (MSP) Laboratory. His research interest is in human-centered multimodal machine intelligence and application, focusing on the broad areas of speech processing, affective computing, and machine learning methods for multimodal processing. He has worked on speech emotion recognition, multimodal behavior modeling for socially interactive agents, in-vehicle active safety systems, and robust multimodal speech processing. He was selected by the School of Engineering of Chile as the best electrical engineer who graduated in 2003 from Chilean universities. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. His students received the third prize IEEE ITSS Best Dissertation Award (N. Li) in 2015, and the AAAC Student Dissertation Award (W.-C. Lin) in 2024. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He received the Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian) and the Best Paper Award from IEEE Transactions on Affective Computing in 2022 (with Yannakakis and Cowie). In 2023, he received the Distinguished Alumni Award in the Mid-Career/Academia category by the Signal and Image Processing Institute (SIPI) at the University of Southern California. He received the 2023 ACM ICMI Community Service Award. He is currently an associate editor of the IEEE Transactions on Affective Computing. He is a member of AAAC and a senior member of ACM. He is an IEEE Fellow and an ISCA Fellow.

Jagath C. Rajapakse (Fellow, IEEE) is Professor of Data Science at the College of Computing and Data Science at Nanyang Technological University (NTU), Singapore. He has BSc degree in Electronics and Telecommunication Engineering from University of Moratuwa (UM), Sri Lanka, and MS and PhD degrees in Electrical and Computer Engineering from University at Buffalo (UB), USA. He was Visiting Scientist to the Max-Planck Institute of Cognitive and Brain Sciences, Germany, and the National Institute of Mental Health, USA before joining NTU. He was Visiting Professor to the Department of Biological Engineering at Massachusetts Institute of Technology (MIT).

Professor Rajapakse's research works are in the areas of explainable AI, generative AI, brain imaging, and computational and systems biology. He has published over 300 peer-reviewed research articles in high-impact journals and conferences. His current research works focus on developing computational techniques and tools for diagnosis and treatment of brain diseases by combining neuroimaging and multi-omics data; and for generating small molecule and peptide-based drugs for cancer. He is also looking into how imaging data can be integrated with multi-omics (genomics, proteomics, transcriptomics, and epigenomics) data for investigating molecular underpinnings of various diseases.

He serves as Editor for Engineering Applications in Artificial Intelligence journal (IF = 8.0) and served as Associate Editor for IEEE Transactions on medical imaging, IEEE Transaction on neural networks and learning systems, and IEEE Transactions on computational biology and bioinformatics. He was a Fulbright Scholar and elevated to IEEE Fellow in recognition of his contributions to brain image analysis.