Improving Speech Emotion Recognition with Mutual Information Regularized Generative Model

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

Abstract—Although speech emotion recognition (SER) research has been advanced, thanks to deep learning methods, it still suffers from obtaining inputs from large quality-labelled training data. Data augmentation methods have been attempted to mitigate this issue, generative models have shown success among them recently. We propose a data augmentation framework that is aided by cross-modal information transfer and mutual information regularization. Mutual information based metric can serve as an indicator for the quality. Furthermore, we expand this data augmentation scope to multimodal inputs, thanks to mutual information ensureing dependency between modalities. Our framework was tested on three benchmark datasets: IEMOCAP, MSP-IMPROV and MSP-Podcast. The implementation was designed to generate input features that are fed into last layer for emotion classification. Our framework improved the performance of emotion prediction against existing works. Also, we discovered that our framework is able to generate new inputs without any cross-modal information.

Index Terms—data augmentation, generative adversarial network, mutual information, speech emotion recognition, multimodal deep learning.

1 Introduction

Speech emotion recognition (SER) had significant improvement in recent years thanks to application of deep learning architectures [1], [2]. Training of deep learning models rely heavily on high-quality labeled datasets, while typical dataset has smaller volume than counterparts in computer vision (ex: MNIST dataset). This scarcity triggers the issue of generalizability of SER models. In this context, data augmentation is promising as it can increase the volume of the dataset [3], [4].

The simplest scheme of data augmentation is to copy a perturbed version of an original data sample. This perturbation can come in various forms, such a simple way as signal transformation, adding gaussian noise [5] or sophisticated way as mix-up [6]. In contrast, generative models have attracted researchers as they can synthesize new data samples after being trained with the original data. Generative Adversarial Networks (GANs) [7] have emerged among other generative models [3], [4], thanks to their ability to generate high-quality synthetic data. The key difference between two methodology is that perturbation merely replicates data with same information with additive noise, while generative models extract information from data and synthesize new samples from extracted information.

- *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.*
- *S. Sivadas is with National Computer Systems, Singapore.*
- *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*

Data augmentation methods, via generative models, involve conditioned generative models. As SER problems are mostly classification problems, class conditioned generative models are most common methods [4], [8]. Thus, crucial challenge is, how to reconstruct samples based on extracted information so that synthesized samples are similar as possible to real samples. In this paper, we hypothesize that conditional generative model naively assumes deterministic mapping of emotional labels to generated sample. However, such mapping is unrealistic, for this reason, the quality of generated samples cannot be ensured.

Mutual information regularization can be a sound alternative against conditional generative models [9], [10]. Mutual information can provide a quantitative measurement to observe the dependency between generated samples and class labels, which is a crucial requirement for data augmentation method. We train generative model to approximate data distribution, while regularizing with mutual information [10].

Many generative data augmentation methods focused on generating samples while conditioning on emotional class labels. Recently, researchers started to incorporate information that are not sourced from audio, which is text information [11], [12]. With cross-modal information, it is natural to expand the scope to multimodal SER. Not limited to data augmentation, many recent research adopted the framework of cross-modal information transfer [12]–[15]. However, most of the work focused on supplementing or recovering noisy or missing modality input respectively. Multimodal augmentation framework to explicitly improve multimodal emotion recognition has not yet been explored thoroughly.

In this paper, we propose generative data augmentation

TABLE 1
Summary of literature survey

Title	Year	Generative Augmentation	Cross-modal transfer	Cross-modal alignment	Mutual Information	Multimodal augmentation
Latif et al. [16]	2020	✓				
Yi et al. [8]	2020	✓				
Latif et al. [4]	2020	✓				
Sahu et al. [9]	2022	✓			✓	
Wang et al. [17]	2022	✓				
Latif et al. [18]	2023	✓	✓			
Malik et al. Model [11]	2023	✓	✓			
Kim et al. [19]	2024	✓	✓			
Goncalves et al. [13]	2022		✓			
Chen et al. [20]	2023		✓	✓		
Wang et al. [21]	2023		✓			
Meng et al. [14]	2023		✓			
Wang et al. [12]	2023	✓	✓			
Liu et al. [15]	2024		✓	✓		
Ours	2025	✓	✓	✓	✓	✓

framework for SER that exploits text features and mutual information. Our data augmentation method adopt InfoGAN [10] to generate audio features while regularizing GAN via mutual information (or penalizing independence). Mutual information regularizing module of InfoGAN is de facto encoder which outputs test feature from audio feature input, thus, our method can synthesize text feature as well. Therefore, we expanded our augmentation framework to multimodal inputs. With real and generated features for both audio and text, we trained audio-text fusion emotion classification layer with every four combinations for real and generated features. Through experiments on benchmark datasets, we observed the improvement at both SER and multimodal SER, thanks to our framework.

Our contribution can be summarised as follows:

1) We propose generative model based data augmentation framework, that is applicable for both SER and multimodal SER.
2) We introduce combining of cross-modal transfer and mutual information for SER data augmentation methodology.
3) We discovered that mutual information regularizer can offer observables to validate the dependency of generated data with emotion and text information.

2 Related Works

The positioning of our contribution can be summarized into a table as in Table 6. Our novelty lies on generative model for augmentation with deliberate combination of cross-modal transfer and mutual information. Furthermore, we expand it toward multimodal augmentation to improve multimodal emotion prediction.

Generative adversarial learning has been widely employed in speech emotion recognition (SER) to leverage generated samples that are expected to convey emotional characteristics. In an early study by Latif et al. [16], an autoencoder-based framework was proposed, where the bottleneck representation (or encoder output) served as the generator, and the discriminator was trained to distinguish between prior and generated samples. This approach, commonly referred to as an adversarial autoencoder, provided a foundation for combining generative models with SER tasks. Similarly, Yi et al. [8] utilized an autoencoder-based approach but introduced a separate generator in their framework.

In another study, Latif et al. [4] integrated a generative adversarial network (GAN) with an autoencoder, utilizing datasets augmented via the mix-up method. Sahu et al. [9] advanced this approach by employing an autoencoder scheme along with the mutual information maximization principle, demonstrating that InfoGAN [10] could enable emotion-controllable synthesis in SER. Wang et al. [17] also used a GAN for data augmentation but addressed the issue of imbalanced class distributions. They incorporated triplet loss guidance to learn more robust emotional features.

More recently, Latif et al. [18] explored text-to-speech models as a data augmentation method, capitalizing on advancements that have simplified the training of end-to-end text-to-speech systems using emotional speech datasets. Malik et al. [11] adopted diffusion models to generate mel-spectrograms from text embeddings, while Kim et al. [19] combined diffusion models with variational autoencoders (VAEs) to enhance data generation.

Meanwhile, multimodal deep learning research has gained much attention recently [22], [23], many works intended to strengthen the representation from one modality by using the input from other modalities, referred as co-learning. Fusion of representation from different modalities can be considered as implict co-learning in broad sense, but in this work we focus where one modality is explicitly augmented from other modalities. In other words, we focus on cross-modal transfer or cross-modal alignment.

Several recent works addressed to improve the robustness in the case where some modality inputs might be missing or noisy. Goncalves et al. [13] used multitask learning scheme to expose the model to the situation where only one modality input is available. In addition, typical cross-modal attention layer is employed to align the representation of both modalities. Wang et al. [21] also handled missing modality case by doing random modality masking of input. However, these approaches are implicit, neither gaining additional information from other modalities nor reconstructing missing modality is feasible.

On the other hand, aligning representation between different modalities can be a better approach to enforce

information sharing between them. As briefly mention in previous paragraph, cross-modal attention [13] is the most common method, but rather focused on fusion than co-learning. Alternatively, optimizing metrics between representations of different modality has been explored as in the work of Chen et al. [20]. Recently, contrastive learning has gained much attention in this context as well. Liu et al. [15] used contrastive learning loss between different modalities to enforce modality-invariance. Yet, alignment approach does not enable reconstruction of missing modality.

Reconstruction approach is intuitive co-learning scheme, Meng et al. [14] trains multimodal autoencoder to generate reconstructed inputs. Also, Wang et al. [12] specifically assumes the modality missing scheme, that missing modality input is reconstructed via conditional diffusion model. Generative models, such as GAN or diffusion models, are special case of reconstruction methodology. However, generative models are expressive and versatile, as generative models extract rich information from the data to enable generation of noisy samples or learn mutual information between modalities for alignment.

3 Proposed Framework

Our framework is intended for the situation where the one, who has a decent SER model (and also a text emotion recognition model for multimodal SER case), wishes improve the performance by increasing the training data samples. In this chapter, we demonstrate the usage of our framework with an example: we train the generative model to generate features before the classification layer, rather than generating raw audio, to reduce the complexity of the experiment (visual summary is depicted in Fig 1.

First, we will prepare the decent SER model, pre-trained audio transformer is fine-tuned for benchmark SER dataset, simultaneously training with text feature (or embedding), which is an output from frozen text transformer, to be aligned with audio feature (or embedding) before linear classification layer. Second, we train InfoGAN to generate audio features extracted from the training dataset with mutual information regularizer. Finally, the audio features (and text features for multimodal SER case) and generated features are prepared and fed into final linear classification and trained with emotion classification loss, with larger number of data sample to enhance performance.

3.1 Baseline preparation

We adopted two types of pre-trained audio transformers as the example: Audio Spectrogram Transformer (AST) [24] and Wav2Vec2 [25]. The audio transformer will get raw audio input after necessary feature extraction and output final hidden states, which is the audio feature that we intend to generate. A linear layer with softmax activation function is added. The output of this classification layer will be referred as prediction of our baseline model.

The audio feature before the classification layer is referred to as h:

$$h = f_a(x_a). \tag{1}$$

where f_a denotes transformer that consists feature encoder and x_a denoting audio input after the preprocessing function. The linear classifier from the encoded feature is a fully-connected layer predicting emotional class c:

$$\hat{y} = \text{softmax}(W_y h + b_y). \tag{2}$$

The weights W_y and biases b_y of the fully connected layers which can be further fine-tuned during InfoGAN training and augmented data fine tuning. Training objective is minimizing the cross-entropy loss $\mathcal{L}_{SER}$:

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

y denotes the emotional label corresponding audio input x_a, and E denotes the expectation.

In addition, we add cross-modal alignment module to baseline model. Let us define the text feature encoding module with text transformer to obtain text feature t as:

$$t = f_t(x_t). \tag{4}$$

where f_t denote pre-trained transformer model and x_t denoting textual input, which is taken from the ground truth transcript, provided from the training dataset.

Inspired by CLAP [26], we exploit contrastive learning to align audio feature h and text feature t into same representation space:

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

where i and j denotes the index within the minibatch.

Furthermore, we maximize the mutual information between h and t, we employ contrastive learning to maximize mutual information as in [27], via InfoNCE loss. InfoNCE loss shares identical structure with $\mathcal{L}_{CL}$, however, the positive pair is formulated in a way that log ratio of conditional probability of $p(t|h)$ to marginal probability $p(t)$. Thus, we use another linear layer to project h to $\hat{t}$ and get InfoNCE loss ($\mathcal{L}_{MI}$) as:

$$\hat{t} = W_t h + b_t \tag{6}$$

$$\mathcal{L}_{MI} = -log \frac{\exp\left(sim(t^i, \hat{h}^i)/\tau\right)}{\sum_{j=1}^{B} \exp\left(sim(t^i, \hat{h}^j)/\tau\right)}. \tag{7}$$

In result, we fine-tune the baseline model with training dataset simultaneously with all the loss terms above.

3.2 InfoGAN

In the second stage, we will employ InfoGAN that will generate $\hat{h}$ from the implicitly estimated probability distribution function of h, while regularizing generator module with mutual information between $\hat{h}$ with t and emotion labels.

3.2.1 Generative Adversarial Network (GAN)

GAN is commonly employed as a data augmentation to improve SER performances. Generation of synthetic data from GAN has sufficient quality to mimic emotion in original data, which is also explainable in theoretic analysis. Our GAN framework consists of the generator and the discriminator. The generator emits synthetic audio features (or

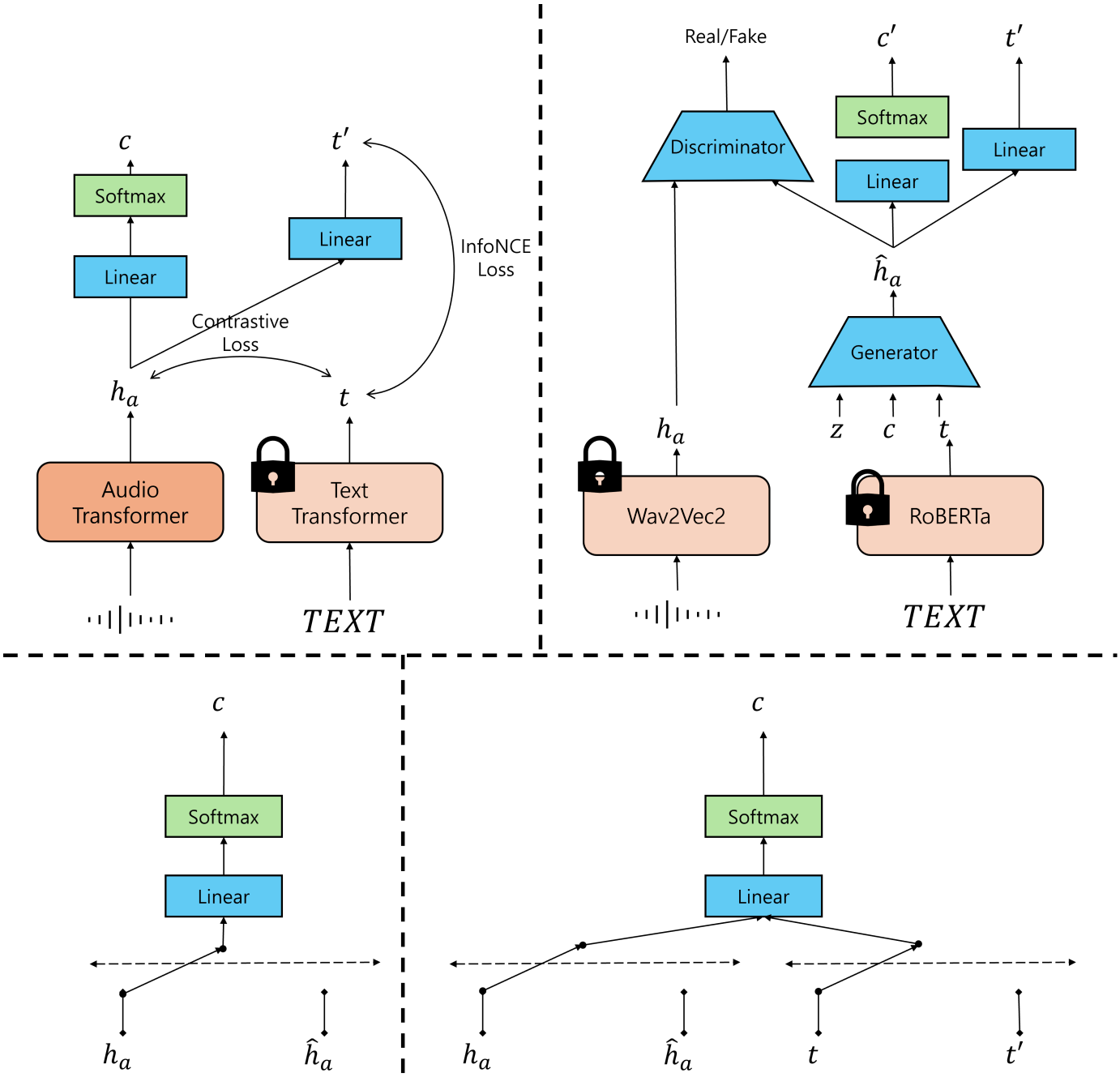

Fig. 1. The visual summary of proposed augmentation framework to improve SER, consisted of three stages. First stage (top left): baseline model is trained with contrastive loss and InfoNCE loss. Second stage (top right): InfoGAN, GAN with mutual information module that predicts latent c' and t' from generated $\hat{h}$, is trained and mutual information module re-uses the same prediction layers in the first stage. Third stage (bottom): we have two parallel stream in the final stage, SER case (bottom left) and multimodal SER case (bottom right), which is training the linear classification module with all possible input combinations by switching between h or $\hat{h}$ and t or t'. linear classifier layer is fine tuned with generated samples.

embedding) that mimic the original features extracted from raw speech. Generator is a network that takes random noise vectors (and other conditioning vectors if available) as input and outputs feature representations. On the other hand. the discriminator network is a binary classifier that discriminate between real and generated data. The generator and discriminator are trained with an objective that is described as minimax game, where they are adversarially optimized. We denote h as real data sampled from the data distribution, following probability distribution function (pdf) of $p_{\text{data}}(h)$. Also, z is a noise vector, sampled from pdf of $p_z(z)$. And discriminator and generator is depict as functions, $D(h)$ and $G(z)$. Finally, training objective function of GAN ($V(D,G)$) is as follows:

$$\min_G \max_D V(D,G) = \mathbb{E}_{h\sim p_{\text{data}}(h)}[\log D(h)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]. \tag{8}$$

Training consists two alternate phase, each phase is responsible in updating either D or G.

1) **Discriminator updating phase:** D is optimized to maximize $V(D, G)$.
2) **Generator updating phase:** G is optimized in opposite direction, minimizing $V(D, G)$, identical to minimizing $\log(1-D(G(z)))$ as the other term is not affected by updating G. However, it is implemented to maximize $\log D(G(z))$, for the sake of efficient gradient descent (refer to original paper for more details).

The training yields optimal discriminator as follows:

$$D^*(h) = \frac{p_{\text{data}}(h)}{p_{\text{data}}(h) + p_g(h)}. \tag{9}$$

We denote p_g as the distribution of output of $G(\boldsymbol{z})$ sampled from $\boldsymbol{z} \sim p_{\boldsymbol{z}}$. Take note that, $G(z)$ can be replaced with h, which denotes generated sample. Thus, the random variable h can be considered as a union of original data and synthesized data. Objective function can be re-arranged with optimal discriminator as follows.

$$\begin{aligned} &V(G, D) \\ &= \int_{\boldsymbol{h}} p_{\text{data}}(\boldsymbol{h}) \log(D(\boldsymbol{h}))dh + \int_z p_{\boldsymbol{z}}(\boldsymbol{z}) \log(1 - D(g(\boldsymbol{z})))dz \\ &= \int_{\boldsymbol{h}} p_{\text{data}}(\boldsymbol{h}) \log(D(\boldsymbol{h})) + p_g(\boldsymbol{h}) \log(1 - D(\boldsymbol{h}))dh. \end{aligned} \tag{10}$$

Above derivation holds true under assumption that G is a deterministic mapping from z to h. Thus, following identity

$$p_g(h|z) = \delta(h - G(z)), \tag{11}$$

is used convert integration over z to integration over h.

After rewriting objective function, it turns out that discriminator updating phase resembles noise constrastive estimation, thus, discriminator updating phase results in D to implicitly parametrizing original data distribution. Yet, hidden assumption has to be taken noted: pdf of generated sample is distinctive to pdf of original data samples.

Finally, it is intuitive to follow that generator updating phase is, in effect, approximating p_g to p_{data}. If generator takes conditioning vectors other that z, the GAN training will be performed under assumption that there is a determistic mapping from conditioning vector (emotional labels or text embedding) as well. For this reason, there is no observable quantity to measure dependency between conditioning vector and generate sample, which is crucial measure to ensure quality of augmented data.

3.2.2 Information Maximizing Generative Adversarial Network (InfoGAN)

InfoGAN is an extension of GAN proposed for controllable generation via exploitation of disentangled latent vector alongside with random noise vector.

1) **GAN:** The original training objective of GAN is defined as:

$$\begin{aligned} &\min_G \max_D V(D, G) = \\ &\quad \mathbb{E}_{h \sim p_{\text{data}}(h)}[\log D(h)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \end{aligned} \tag{12}$$

2) **InfoGAN:** Generator of InfoGAN requires two parts: z is the noise vector as in convention and c represents the latent code (or vector) that is expected to be disentangled and interpretable (we will use two latent variables c and t as defined in baseline model during experiment, but here we simply denote as c for sake of simplicity withing this subsection only). The objective function of InfoGAN does not modifiy existing objective but simply adds a mutual information term $I(c; G(z, c))$ with hyperparamter λ:

$$\min_{G,I} \max_D V(D, G) - \lambda I(c; G(z, c)). \tag{13}$$

Mutual information is defined using entropy H as follows:

$$I(c; G(z, c)) = H(c) - H(c|G(z, c)). \tag{14}$$

Here, the first entropy term, $H(c)$, can be ignored in optimization as we can sample latent vector c from simple distribution (ex: random noise) or we use text embedding which is fixed during training. In result, mutual information maximization is achieved by minimizing the second term, which is conditional entropy. As conditional entropy is non-negative, minimum is 0, which implies that there exists mapping from generated sample ($G(z, c)$) to latent vector (c). From this intuition, we can train addition network, denoted as Q, to predict c when given $G(z, c)$ as input. While, this is informal derivation, more formal description of this technic is variational mutual information maximization, which is maximizing the lower bound of mutual information with auxiliary distribution function Q (please refer to original paper for more information).

InfoGAN training can be considered as vanilla GAN training while regularizing via mutual information. During training this regularization term can be computed to infer mutual information. The loss of vanilla GAN computed from either generator or discriminator does not indicate the quality of generated sample. On the other hand, the loss value from mutual information can be a direct indicator of generated sample.

In our method, we follow conventional training procedure to train GAN as in [7]. But for mutual information loss, we will describe our definition of loss here. In conventional InfoGAN, mutual information loss is simply cross entropy loss of predicting c from generate sample of mean squared error loss to predict c from generated sample. This is a variational approximation to maximize mutual information as c are typically chosen to be a random variable with simple pdf that is known, so that we can easily sample. Our method has two latent variable c and t, which is sampled from the dataset, to generate sample $\hat{h} = G(z, c, t)$. To maximize mutual information for c, as y follows simple categorical distribution, we use the same scheme of variational

approximation. Thus, minimizing following terms achieves maximum mutual information:

$$\hat{y_g} = \text{softmax}(W_y\hat{h} + b_y), \tag{15}$$

$$\mathcal{L}_{I_y} = -E\{\sum_c 1(y=c)log(\hat{y_g})\}. \tag{16}$$

Note that the same weights from the baseline model is used to enforce latent variable to be coded with emotional information. If not, InfoGAN training objective will code arbitrary information of variations within the dataset (similar to clustering results). Similar approach is taken for latent variable t, but use InfoNCE loss instead of mean squared error loss, as we are optimizing embedding model rather than regression model.

$$\hat{t_g} = W_t\hat{h} + b_t \tag{17}$$

$$\mathcal{L}_{I_t} = -log\frac{\exp\left(sim(t^i, \hat{t_g}^i)/\tau\right)}{\sum_{j=1}^{B}\exp\left(sim(t^i, \hat{t_g}^j)/\tau\right)}. \tag{18}$$

Again, that the same weights from the baseline model is used to enforce latent variable to be coded with textual information. Furthermore, using the weights used in baseline model allows us to maximize mutual information not only with generated samples but also with real samples.

3.3 Data augmentation for SER

After InfoGAN is successfully trained, we can train emotion prediction model by training data samples, together with generated data samples that resembles their distribution. As final stage, all modules, except emotion classification module, are frozen.

The case for SER is rather straightforward. After extracting h from training dataset, we generate same number of $\hat{h}$, and then fed them through same emotion classification layer and compute loss to further fine-tune them. In effect, the emotion predictor has been trained with dataset that has size of double from original training dataset.

For the case of multimodal SER, h, $\hat{h}$, t and t' has to be prepared. And new classification layer that takes concatenated feature of audio feature and text feature has to be defined. Then we can compose inputs with all possible concatentation with original and generated features. Thus, we can achieve the data sample size increase by factor of four.

4 Experiment

4.1 Dataset

4.1.1 IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset

The IEMOCAP dataset is a commonly used dataset for SER research, which consists of approximately 12 hours of recordings of video with audio, featuring interactions between 5 pairs of male and female actors. These pairs of actors were set in dyadic sessions to engage in dialogues to elicit emotional expressions. Multimodal raw data was collected and we used speech recording and transcript for our experiments. Each recording was segmented per utterance and manually annotated by multiple human evaluators. Originally, there are nine categorical emotional labels but only four major emotions were chosen, which are: neutral, happiness (merged with excited), sadness and anger.

4.1.2 MSP-IMPROV (Multimodal Signal Processing - Interactive Emotion Dyadic Motion Capture) dataset

The MSP-IMPROV dataset is a multimodal dataset for research in SER, which share similar structures with IEMOCAP dataset. The MSP-IMPROV dataset consists of approximately 9 hours of recordings 12 actors (6 male and 6 female) interacting with each other. Each recording is segmented into the unit of utterances, and annotated by multiple human evaluators. Annotation for categorical emotion labels yielded four basic emotion category: hapiness, sadness and neutral.

4.1.3 MSP-Podcast [28]

This corpus is created by collecting English speech data from existing podcast recording. Comparable to IEMOCAP and MSP-IMPROV, it is not collected from lab environment, thus, it is more close to 'in-the-wild' speech dataset. The dataset comes predefined partition of training, development and test set. MSP-PODCAST data comes with more emotion category than MSP-IMPROV. But to match with our experiment setting, we only used four categories: neutral, happiness, sadness and anger.

4.2 Data augmentation for SER

4.2.1 Baseline model

We adopt the pre-trained AST [24] from 'https://huggingface.co/MIT/ast-finetuned-audioset-10-10-0.4593' as backbone. And add linear classifier after pooled output from AST. For text feature encoder, pre-trained BERT [29], that is 'bert-uncased-base' version. For SER experiment, we do not consider multimodal SER accuracy, thus, BERT is frozen and we only use it to extract text features from ground-truth transcript.

4.3 InfoGAN architecture

We train InfoGAN to generate feature encoding, which is in the dimension of 768. So we take the simplest form of architecture for Discriminator and Generator which is a linear layer without any activation function. Also the network for mutual information maximization is also linear layer only. Especially, the linear layer that maximizes mutual information with y uses the same layer from emotion classifier. In this way, we can prevent latent code vector to encode something other than emotion, as there are many other complex variations inherent in the training dataset. Emotional code vector is trained with cross entropy to maximize the mutual information. For the latent embedding given from BERT embedding we use InfoNCE loss, which is a contrastive learning loss trying attract the BERT embedding and predicted text embedding from generated sample. Lastly we adopted mix-up strategy for training generator and discriminator as proposed in [6], for stable training of generator.

TABLE 2
Performance comparison with existing data augmentation methods on IEMOCAP dataset

Methods	Without augmentation	With augmentation
Sahu et al. [9]	59.42	60.29
Bao et al. [3]	59.48±0.71	60.37±0.70
Latif et al. [31]	60.51±0.57	61.05±0.68
Malik et al. [11]	58.62±2.11	61.22±1.85
Ours	60.81±4.83	63.40±2.52

4.4 Data augmentation for multimodal SER

We adopt the pre-trained Wav2Vec2 [30] from 'https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim' for multimodal SER experiment. For text feature encoder, pre-trained RoBERTA [29], that is 'roberta-uncased-large' version. Fof multimodal SER experiment, RoBERTa is separately trained to predict emotion labels from transcript before the baseline stage. After that RoBERTa model weight is frozen during training baseline and onwards. For MSP-Podcast experiment, we use Llama 3.0 instead of RoBERTa for text features, but same Wav2Vec2 for audio features. Other than the change of feature size, due to change of back-bone model, configuration for InfoGAN stay the same.

5 Result

The experiment was conducted in two rounds. First, our data augmentation framework was tested on SER case. In first round, we observe the performance difference before and after data augmentation, and ablation study to check the contributing effects of each modules in the model. Finally, we run the multimodal SER experiments. The second round, we compare the performance of the model using proposed framework with existing works.

5.1 SER experiment

We experimented our method on commonly used dataset IEMOCAP. The experiment followed leave-one-speaker-out cross validation scheme. Unweighted average recall is reported per each speaker and we statistically average and present mean and standard deviation in the Table 6 as follows.

The results depict that our augmentation method achieves the state-of-the-art performance among existing methods. Furthermore, our augmentation method shows highest improvement against its baseline after augmentation, which is 2.6%, which is same amount of improvement with Malik et al. [11]. Sahu et al. employed typical emotional label conditional GAN and Bao et al. [3] CycleGAN approach to generate additional data to train SER model. Latif et al. [31] achieved generalization improvement by using unlabelled dataset during training. Still, the improvement of accuracy against their baseline was marginal. Recently, the margin between baseline and data augmented SER was improved when data generation model was conditioned on text input as in Malik et al. [11]. And our model also achieved same margin and state-of-the-art performance by using mutual information between generated data and text data.

TABLE 3
Experiment results from ablation study

Methods	UAR
Full model	63.40±2.52
Without cross-modal alignment	62.31±3.65
Without cross-modal alignment & text embedding	61.07±2.45
Without cross-modal alignment & mutual information maximization	61.70±2.58
Baseline without cross-modal alignment	60.73±3.92

TABLE 4
Multimodal SER experiments on IEMOCAP dataset

Model	Unweighted Accuracy		
	Audio	Text	Audio + Text
sc_LSTM [32]	55.96	63.51	68.45
bc_LSTM [32]	57.43	67.69	73.89
AE [33]	56.79	25.44	26.25
CRA [34]	56.79	27.96	30.13
MMIN [35]	58.23	67.52	74.72
CIF-MMIN [15]	58.44	69.26	75.65
Ours	72.83	65.51	76.54

5.1.1 *Ablation study*

We performed ablation study to examine the effect of each modules we proposed. The results are depicted in Table 3.

First, we present the results when cross-modal alignment component(text embedding component and corresponding losses) ofbaseline model is removed. This result show that the cross-modal alignment module during baseline model training improves the quality of data augmentation. We hypothesize that the loss terms corresponding to cross-modal alignment enforces strong dependence between audio embedding and text embedding, providing better initial state to train InfoGAN. Next, we observed the contribution of text embedding and mutual information maximization loss in InfoGAN training. The result depict that both component has significant contribution in quality of data augmentation. Lastly, we experiment the case where we do not use data augmentation and baseline model is without cross-modal alignment, which is the SER performance on fine-tuning on vanilla AST. The result shows similar baseline performance as in Table 3. This implies that cross-modal alignment helps improve the quality of generated samples from InfoGAN, not directly improving SER performances.

5.2 Multimodal SER Experiment

Multimodal SER experiment was performed on three datasets. This experiment is an use-case-scenario, where you are provided with two emotion prediction models, that one takes speech input and text input for the other. Through this experiment, we intend to show that even with simple data augmentation method via generative model, we improve the performances to the decent amount.

5.2.1 *IEMOCAP*

The experiment result revealed the advantage of our data augmentation method, in terms of accuracy. CRA [34], MMIN [35] and CIF-MMIN [15] exploit the dependency between features from different modality, thus achieving

TABLE 5
Multimodal SER experiments on MSP-IMPROV dataset

Model	Unweighted Accuracy		
	Audio	Text	Audio + Text
sc_LSTM [32]	40.91	32.01	41.36
bc_LSTM [32]	42.53	57.39	58.32
AE [33]	42.69	27.62	38.73
CRA [34]	38.96	28.37	37.97
MMIN [35]	42.71	56.49	60.98
CIF-MMIN [15]	41.56	58.57	61.72
Ours	57.85	40.29	62.84

TABLE 6
Multimodal SER experiments on MSP-Podcast dataset

Model	Unweighted Accuracy		
	Audio	Text	Audio + Text
emoDARTS [36]	61.15	-	-
Ours	56.76	49.01	60.61

decent performance among other existing works. Inspired from these results, we adopted similar methodology, not to recover missing feature (modality) but to generate new multimodal feature sets.

5.2.2 MSP-IMPROV

We ran same experiment with MSP-IMPROV dataset. As IEMOCAP and MSP-IMPROV are datasets with similar characteristics, similar trend with IEMOCAP experiment result is observed.

5.2.3 MSP-Podcast

Unlike, IEMOCAP and MSP-IMPROV, MSP-Podcast is more naturalistic dataset with much larger number of samples. As there weren't any existing works that ran experiment with similar protocol, we compared our method with emoDARTS [36], which uses neural architecture search and only audio input. Our method performed up to similar level of success to emoDARTS. Compared to emoDARTS, our model can be simpler choice to offer during implementation.

Notable point is, that MSP-Podcast has severe class imbalance problem, thus, we generated data samples for the lesser count of emotional classes to make same numbers of samples per emotion. This scheme of data generation is unlike the generation in existing works, as [12], which requires other modality features to transfer emotional information. In this case, we generated audio feature and text feature from scratch, with only emotional label provided. Still, out data augmentation method synthesized decent sample and we were able to mitigate the class-imbalance problem.

6 Conclusion

We proposed a new framework of data augmentation for speech emotion recognition (SER) using GAN. Unlike previous approaches, we employ the combination of mutual information regularization and cross-modality transfer (from text to be specific). The advantage of mutual information regularization is: mutual information regularizing loss can serve as an observable metric to depict the quality of generated data sample, in the form of dependency measure. Furthermore, we propose multimodal data augmentation using our framework. Previous works were limited in using cross-modal information transfer or cross-modal generative model for supplementing noisy of missing modality inputs. In our work, we propose to use our cross-modal generative model to generate more multimodal inputs by preparing combination of real inputs and generated inputs. We conducted experiments to test our framework in both, SER and multimodal SER cases. The framework was implemented to generate audio features or text features that is fed into final classification layer. Our method was evaluated on IEMOCAP, MSP-IMPROV and MSP-Podcast. Thanks to our method, we observed decent amount of improvement. We conclude, that when one has two separate decent models to predict emotion from audio and text, our framework can be implemented to generate more inputs further improve the performance in multimodal emotion recognition. Additionally, we discovered that with mutual information reglarization, one can generate more decent inputs without any information from other modality.

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] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1634–1654, 2021.

[3] 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.

[4] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Augmenting generative adversarial networks for speech emotion recognition," *arXiv preprint arXiv:2005.08447*, 2020.

[5] 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.

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

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[8] L. Yi and M.-W. Mak, "Improving speech emotion recognition with adversarial data augmentation network," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 1, pp. 172–184, 2020.

[9] S. Sahu, R. Gupta, and C. Espy-Wilson, "Modeling feature representations for affective speech using generative adversarial networks," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 1098–1110, 2020.

[10] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, 2016.

[11] I. Malik, S. Latif, R. Jurdak, and B. W. Schuller, "A preliminary study on augmenting speech emotion recognition using a diffusion model," *Proceedings of Interspeech, Dublin, Ireland, August, 2023*, 2023.

[12] Y. Wang, Y. Li, and Z. Cui, "Incomplete multimodality-diffused emotion recognition," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] L. Goncalves and C. Busso, "Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2156–2170, 2022.

[14] T. Meng, Y. Shou, W. Ai, N. Yin, and K. Li, "Deep imbalanced learning for multimodal emotion recognition in conversations," *IEEE Transactions on Artificial Intelligence*, 2024.

[15] R. Liu, H. Zuo, Z. Lian, B. W. Schuller, and H. Li, "Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities," *IEEE Transactions on Affective Computing*, 2024.

[16] 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.

[17] S. Wang, H. Hemati, J. Guðnason, and D. Borth, "Generative data augmentation guided by triplet loss for speech emotion recognition," in *Interspeech*, 2022.

[18] S. Latif, A. Shahid, and J. Qadir, "Generative emotional ai for speech emotion recognition: The case for synthetic emotional speech augmentation," *Applied Acoustics*, vol. 210, p. 109425, 2023.

[19] Y.-J. Kim and S.-P. Lee, "A generation of enhanced data by variational autoencoders and diffusion modeling," *Electronics*, vol. 13, no. 7, p. 1314, 2024.

[20] C. Chen, H. Hong, J. Guo, and B. Song, "Inter-intra modal representation augmentation with trimodal collaborative disentanglement network for multimodal sentiment analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1476–1488, 2023.

[21] 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.

[22] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[23] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations & trends in multimodal machine learning: Principles, challenges, and open questions," *ACM Computing Surveys*, vol. 56, no. 10, pp. 1–42, 2024.

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

[25] 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.

[26] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 873–883.

[33] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, 2006.

[34] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1405–1414.

[35] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2608–2618.

[36] T. Rajapakshe, R. Rana, S. Khalifa, B. Sisman, B. W. Schuller, and C. Busso, "emodarts: Joint optimisation of cnn & sequential neural network architectures for superior speech emotion recognition," *IEEE Access*, 2024.