

Learning Enhanced Acoustic Latent Representation for Small Scale Affective Corpus with Adversarial Cross Corpora Integration

Chun-Min Chang, *Student Member, IEEE*, Chi-Chun Lee, *Senior Member, IEEE*

Abstract—Achieving robust cross contexts speech emotion recognition (SER) has become a critical next direction of research for wide adoption of SER technology. The core challenge is in the large variability of affective speech that is highly contextualized. Prior works have worked on this as a transfer learning problem that mostly focuses on developing domain adaptation strategy. However, many of the existing speech emotion corpora, even those considered as large scale, are still limited in size resulting in an unsatisfactory transfer result. On the other hand, directly collecting context-specific corpus often results in an even smaller data size leading to an inevitably non-robust accuracy. In order to mitigate this issue, we propose the concept of enhancing the affect-related variability when learning the *in-context* acoustic latent representation by integrating *out-of-context* emotion data. Specifically, we utilize adversarial autoencoder network as our backbone with multiple *out-of-context* emotion labels derived for each *in-context* samples that serve as an auxiliary constraint in learning the latent representation. We extensively evaluate our framework using three *in-context* databases with three *out-of-context* databases. In this work, we demonstrate not only an improved recognition accuracy but also a comprehensive analysis on the effectiveness of this representation learning strategy.

Index Terms—speech emotion recognition, adversarial network, acoustic representation, cross corpus learning

1 INTRODUCTION

ROBUST cross-context speech emotion recognition has been an active research topic in speech emotion recognition (SER) technology. The challenge has been originally addressed as a technical effort in developing a *universal* SER. For example, Bezooijen et al. used three different languages (Dutch, Taiwanese, and Japanese) to identify Dutch vocal expressions of emotion [1]. Furthermore, Schuller et al. attempted to construct a universal emotion recognizer for multiple languages simultaneously through feature normalization [2]; Deng et al. proposed to use auto-encoder to learn common representations between source and target speech samples [3]. However, these methods ignore that the large variability of affective speech in each database is highly contextualized, leading to sub-par performance of SER.

It was found with the theoretical implications that emotion perception may be universal, as evident from psychology literature [4], [5]. Hence, several advanced SER frameworks have cast the multi-context robustness issue as a transfer learning (domain adaptation problem), i.e., they have a model trained on a “relatively” large-labeled source emotion corpus, adapting the model in an unsupervised manner to the target corpus. Most prior works inherently assume that learning common (domain-invariant) representation across corpora would alleviate most of the issues and move toward a universal SER, i.e., to utilize subspace learning to find common space between datasets. For example, Liu et al. proposed domain adaptive subspace learning (DoSL) to learn a projection matrix that transforms the source and

target speech signals from the original feature space to a common subspace [6].

Several of these works have demonstrated improved recognition performances when transferring from a labeled speech emotion corpus to another unlabeled corpus. However, the accuracy is still mostly sub-par than possessing directly *in-context* data as the training data. In the real-world application of SER, one would still attempt to collect a certain amount of data for specific contexts and scenarios to train a workable SER system. It is an extremely challenging issue to have enough labeled *in-context* data. Furthermore, the labeled emotional speech data are not enough, and the variability of context-specific speech data are often not adequate for training, even for a particular context, because the limited data size does not capture well the true variability of context-specific speech emotion information. Hence, on the one hand, transfer information from a labeled speech corpus to each particular context can be sub-optimal. On the other hand, direct training on the limited labeled data (generally the case in the SER domain) often undermines the true variability of affect-related acoustic manifestation for each context.

Hence, we propose to enrich—integration of affect-related information with enhanced speech variability—the *in-context* (often small scale dataset) speech representation by integrative learning using *out-of-context* (larger in-scale than *in-context* data) dataset simultaneously. Direct learning of an enriched acoustic latent embedding by integrating *out-of-context* data would possess more representational power and require less labeled *in-context* data instead of the conventional approach in treating the development of a robust SER system as a technical effort of transfer learning. This particular scenario is also quite common in real-world ap-

• Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan
E-mail: clee@ee.nthu.edu.tw

plications, where it is possible to collect and label a limited amount of *in-context* data; however, it is difficult to extend it to large-scale collection. This concept was first proposed by [7], [8]; We developed multi-view integration algorithms for SER, specifically, for each of the samples in the target *in-context* dataset. We assigned each sample with multi-perspective emotion labels derived from the *out-of-context* dataset and fused it through a kernel fusion method.

This idea was technically advanced further using adversarial learning. The adversarial network has been proposed to be a powerful generative model learning technique [9], [10]. We previously extended the concept of integrating *out-of-context* data for *in-context* data in deriving an enriched acoustic latent code vector through an adversarial learning mechanism to improve the *in-context* emotion recognition [11]. The learning was divided into two steps. First, we introduced the multi-view emotion perspective by gathering the emotion labels from the *out-of-context* dataset for each *in-context* sample by computing cosine distance with emotion pooling. Second, the network integrated these multi-perspective by training the network with an adversarial mechanism to learn the enriched acoustic latent representation.

This work builds upon our previous research by technically extending it and validating its robustness further. Specifically, this work contributes beyond [11] with the following three major contributions:

- 1) Introducing the use of pooling *out-of-context* and *in-context* datasets for adversarial training that learns representations to be used in deriving multi-perspective emotion labels.
- 2) Comprehensive validation of our proposed *enriched* acoustic latent vector using three *in-context* databases with two different *out-of-context* databases in 3-class recognition task and one *in-context* database in binary classification task.
- 3) Extensive analysis for our framework: 1. Comparison of different manners in deriving multiple emotion perspectives. 2. Examination on visualization of the latent representations. 3. Investigation on the emotion recognition accuracy obtained as a function of different amount of training samples with different state-of-the-art method.

We evaluated our framework on three smaller in-scale in-context databases—the USC CreativeIT database [12], the VAM database [13], and the RECOLA database [14]—by leveraging three larger out-of-context databases—the USC IEMOCAP [15], the NNIME [16] and the MSP-Improv database [17]. Our experiments show that the proposed adversarially enriched latent representation show improved emotion recognition accuracy with a lower-dimensional representation in all three databases. Our experiments demonstrate the importance of integrating multiple emotion perspectives into network learning. Importantly, our analysis shows that our framework retains its robustness even when trained with much reduced in-context data, i.e., a result that is critically important in developing emotion recognition applications in a harsh context with limited data availability.

2 RELATED WORK

2.1 Cross Corpus Speech Emotion Recognition

Cross corpus SER has been studied in various research works. Conventionally, it is considered a domain adaptation problem, transferring information between datasets under similar recording conditions or from acted in-lab speech to spontaneous in-the-wild speech [18].

Most of cross corpus SER learn common representation between corpora in various ways, e.g., to center around utilizing importance weighting. For example, Hassan et al. originally proposed an importance-weighted support vector machine (IW-SVM) [19], which has been extended to three adaptive algorithms—kernel mean matching (KMM) [20], unconstrained least-squares importance fitting (uLSIF) [21], and Kullback-Leibler importance estimation procedure (KLIEP) [22]—to explicitly compensate for acoustic and speaker differences between training and testing datasets. Another common approach is to use the maximum mean discrepancy (MMD) as the metric to minimize the distribution mismatch problem between different corpora, e.g., Song et al. applied it during the optimization procedure of non-negative matrix factorization [23].

Learning common representation between corpora can also be done through deep learning following the widespread use of deep learning techniques in the SER. Gideon et al. used a progressive neural network to transfer knowledge for three tasks: emotion, speaker, and gender detection [24]. Latif et al. proposed a deep belief network to learn more effective and discriminative long-range features from source and target domain [25], [26].

2.2 Adversarial Transfer Learning for SER

Adversarial training has been proposed to be a powerful generative model learning technique related to SER transfer learning [9], [10]. Stefan et al. used it for augmenting emotion data owing to its intuitive applications as a data augmentor [27]. Furthermore, Xinyue Zhu et al. used cycle-consistent adversarial networks (CycleGAN) to augment data from the original database to train an improved emotion recognizer [28]. Eskimez et al. proposed a CNN-based generative adversarial network (GAN) generating synthetic data to solve the eyes-off information [29]; Chatziagapi et al. generated an in-class spectrogram using GAN to solve the data imbalance for SER [30]. Latif et al. utilized a data augmentation mixed up to augment GANs in representation learning as well as synthetic feature vector generation that learns better representation and generates synthetic feature vectors effectively to solve the lack of availability of larger datasets [31].

Additionally, the characteristics of the adversarial training eliminate the gap of two different databases relatively easily. It leads an alternative way as domain adaptation to the cross-corpus SER. Abdelwahab et al. used a gradient reversal layer in a multi-corpus setting with three databases to predict emotion attributes of arousal, valence, and dominance [32]. Saurabh et al. investigated the application of GANs to generate synthetic feature vectors for speech emotion recognition [33]. Moreover, Gideon et al. proposed the adversarial discriminative domain generalization (ADDOG) that makes representations learned from each dataset closer,

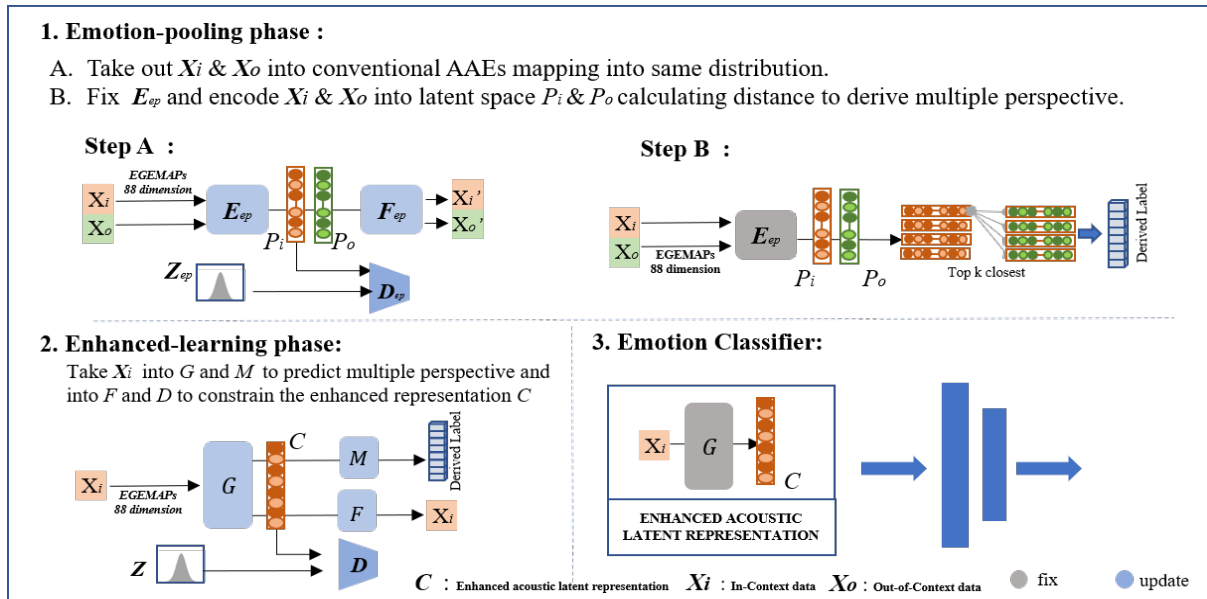


Fig. 1: Emotion-enriched adversarially acoustic latent representations for in-context emotion data learned by leveraging out-of-context emotion corpora and trained with a neural network as a classifier.

improving the emotion recognition performance in cross-corpus. Chao et al. integrated the maximum mean discrepancy (MMD) as the metric loss to enforce the representation consistency between target and source data to achieve unsupervised transfer adversarially [34], [35]. Zheng et al. used multi-scale discrepancy in adversarial network with three different time scale-kind speech features to conduct domain adaptation for cross-corpus SER [36].

Adversarial autoencoders (AAE) of the adversarial techniques have been introduced into the field of SER, unlike the studies mentioned above [9]. It conserves the advantage of autoencoder (AE), the ability of powerful learning representation of which ties the feature space between target and source. Deng et al. proposed a sparse auto-encoder-based feature for SER; they performed transfer learning by building an auto-encoder to learn from source and target and then use SVM to train a classifier [3], [37]. Moreover, Deng et al. utilized a semi-supervised autoencoder improving the emotion performance with another unlabeled emotion corpus [38]. Fu et al. proposed an adversarial autoencoder-based classifier (AAEC), augmenting the data and extended the boundary of the current data distribution [39]. Additionally, it has the advantage of adversarial learning, turning into a generative model to learn more descriptive features than other AE techniques. Sahu et al. encoded high dimensional features with AAE; these latent code vectors contained class discriminability for SER citesahu2018adversarial. Moreover, AAE is also suitable for multi-task learning. Latif et al. showed that the power of AAE is more than merely single emotion recognition; however, it improves the classification performance in a semi-supervised way. They proposed multi-task semi-supervised adversarial autoencoding to transfer knowledge within three different tasks—gender, speaker, and emotion—using AAE to constrain the feature [40]. Additionally, Fu et al. showed the ability of AAE with conversational interlocutors involved to predict gender and emotion [41]. Recently, Gao et al. proposed a feature fusion

network of multi-head attention Bi-LSTM SER, combining the representation from pre-trained AAE [42].

Inspired by these previous works, we propose an adversarially *in-context*-enriched learning enhanced latent representation structure with AAE to enrich the *in-context* (often small-scale dataset) representation by integrative information from the *out-of-context* dataset. It can transfer *out-of-context* data that solely has a different label setting into *in-context* data by directly learning an *enriched* acoustic latent embedding, requiring less labeled *in-context* data.

3 EMOTION DATABASES

We categorize the databases into two different types: *in-context* and *out-of-context* databases. The former is our main emotion recognition evaluation corpus, which contains limited data in scale. The latter is an emotion corpus that is larger in scale and can be used to aid in the learning of the enriched acoustic latent representation for the *in-context* databases. We use a total of three *in-context* databases as our main evaluation corpora and three *out-of-context* databases. These databases are introduced briefly in the following sections.

3.1 In-Context Databases

3.1.1 The USC CreativeIT database (CIT)

The USC CreativeIT database (CIT) is a publicly available emotion corpus that includes dyadic improvisations based on an established theatrical acting technique, termed the Active Analysis, to help elicit natural affective interactions [12]. There were total of 16 actors (eight males, eight females) divided into eight pairs to act out improvisations, resulting in 50 total sessions with each lasting 3– 5 min. The audio data was recorded from wireless close-up microphones synchronized with a multi-channel mixer. Each interaction was rated by three raters using a continuous-in-time annotation scheme on attributes of valence, activation,

and dominance (the scale ranged between 1 and -1). Our target label was the average of the rating. There were a total of 90 audio recordings, each of them had been previously segmented manually into utterances. There were a total of 2162 utterances in the database. This paper only focuses on the activation and valence attributes; it divided the average values of rated annotations of an utterance into three classes (class 1: [-1:-0.33], class 2: (-0.33:0.33], and class 3: (0.33:1]).

3.1.2 The VERA AM MITTAG (VAM) Corpus

The VERA AM MITTAG (VAM) corpus included recordings of a German talk show with spontaneous and emotionally rich speech [13]. There were 12 broadcasts; each included 2–5 persons involved in dialogs. The database contained audio data and facial expressions for a total of 47 speakers. Furthermore, the speakers were between 16 and 69 years. Specifically, 70% were 35 or younger at the time of recording. Each segmented sentence was rated with attributes of activation, valence, and dominance. We also divided the ratings into three classes (class 1: [-1:-0.33], class 2: (-0.33:0.33], and class 3: (0.33:1]). In summary, a total of 947 utterances were used as our target data for the VAM corpus.

3.1.3 The RECOLA Multimodal Corpus

The RECOLA multimodal corpus included 9.5 hours of continuous and synchronous multimodal recordings with audio, video, electrocardiogram (ECG), and electrodermal activity (EDA) [14]. It was a corpus of spontaneous collaborative and affective interactions. There were 46 participants (27 females, 19 males) recruited from the University of Fribourg formed in pairs to perform 23 dyadic teams work. All participants spoke French; however, they had different mother tongues: 33 were native French, 8 Italian, 4 German, and 1 Portuguese. Each interaction was rated by 6 French-speaking raters with ANNEMO web-based annotation toolkit. We used data of 34 participants data because some of the participants did not agree to share their data. The database had approximately seven hours of data. We averaged all the annotations within each utterance and divided the ratings into high versus low (class 1: [-1:0] and class 2: (0:1]) due to difficulty in dividing them into three classes. In summary, a total of 932 utterances, each with a binary label indicating high versus low for activation and valence, were used as our target data for the RECOLA corpus.

3.2 Out-of-Context Databases

Out-of-context databases are the external information we provide for *in-context* databases. The proposed method required original emotion annotation from *Out-of-context* databases.

3.2.1 The NNIME Database

The NNIME is a public multimodal Mandarin Chinese corpus that uses a similar setup as the USC CreativeIT database [16]. The NNIME database contained recordings of 44 subjects engaged in spontaneous affective interactions, which were approximately three minutes long. All the scenarios were designed by the theater professor to ensure a rich, spontaneous, and natural manifestation of affect. Like the

CIT, lapel microphones were hanged on each actor for the audio recordings in each session. There were a total of 6509 utterances segmented in the database. Four naive annotators rated the continuous-in-time activation and valence dimensions (scale ranged between -1 and 1) of each recording.

3.2.2 The USC IEMOCAP Database

The USC IEMOCAP database is a publicly available database with audio and video data. It included five dyadic face-to-face interactions from 10 actors (five males and five females) in pairs [15]. Actors interacted with their partners to engage in both scripted and spontaneous dialogues, which were collected using both motion capture and audio-video data synchronously. The interaction dynamics were designed by professional directors to aid the actors in eliciting natural and affective interactions. There were a total of 6905 utterances corresponding to approximately 12 hours of data. Each utterance was annotated with both categorical emotion labels (such as angry, happy, sad, and neural) and dimensional representations (such as valence, activation, and dominance) on a scale of 1 to 5.

3.2.3 the MSP-Improv Database(MSP)

The MSP-Improv database (MSP) is a new audio-visual emotion corpus [17]. There were over 9 hours of recordings of 6 dyadic scenarios. All of the sessions were improvisations. There was a target sentence designed for every scenario. Specifically, an actor lets his/her partner speak the targeted sentences within various contextualized emotional settings, such as happy, angry, sad, and neutral. This approach allows the actor to express emotions as guided by the scenarios and also avoids prototypical reactions. There was a total of 8438 manually segmented utterances. Each utterance was annotated by at least five raters with both categorical emotion labels (such as angry, happy, sad, and neural) and dimensional representations (such as valence, activation, and dominance) on the scale of 1 to 5. The annotations were gathered using a crowd-sourced labeling scheme. The final label assigned to each speech turn was the average value of the scores provided by the collector of MSP-IMPROV.

4 RESEARCH METHODOLOGY

We provide details of each component of the proposed framework of integrating an *out-of-context* database to learn an enhanced acoustic latent representation for improving emotion recognition for the *in-context* database.

4.1 Acoustic Features

We extracted 88-dimensional eGeMAPS acoustic features using the OpenSmile toolkit for every utterance. This set of features was used extensively in characterizing acoustic information that is relevant for emotion recognition [43]. Specifically, it contained frequency-related parameters—pitch, jitter, and shimmer—energy-related parameters—the loudness and harmonics-to-noise ratio (HNR)—and spectral parameters—alpha ratio, spectral slope, MFCC, and spectral flux. Z-normalized was also used in each corpus feature set to eliminate the value gap between corpora.

4.2 Enhanced Acoustic Latent Representation

Our complete framework derives enhanced acoustic latent representation, as shown in Fig. 1. The proposed structure involves two phases: The emotion-pooling phase and the enhanced-learning phase.

In the emotion-pooling phase, two types of data were involved: *in-context* data X_i and *out-of-context* data X_o . It pools all the available emotion data to learn a common speech representation space using AAE and derives multiple emotion perspectives for each *in-context* speech sample by leveraging *out-of-context* data.

In the enhanced-learning phase, only *in-context* data X_i is involved. An AAE is built by fine-tuning the encoder layer simultaneously to integrate multiple perspectives to obtain the final enhanced latent acoustic representation encoder. The following are the details of two phases:

- 1) Emotion-pooling phase: We first learn a joint AAE to map all available emotional speech data onto a common space to robustly derive multiple emotion perspective from *out-of-context* data X_o to *in-context* data X_i . Specifically, as depicted in Fig. 1, we pool together all the *out-of-context* data and *in-context* data to train a single AAE [44]. E_{ep} denotes the encoder, F_{ep} represents the decoder, Z_{ep} is a Gaussian distribution in which $p(z) = N(z|0, I)$, and D_{ep} denotes the discriminator. This process encodes these speech corpora into a common latent vector space, denoted as P_i and P_o , corresponding to database type. We then identify K-nearest samples (P_o^K) of the *out-of-context* dataset for every j -th sample in P_i (P_i^j) by computing cosine similarity between P_i^j to all samples in P_o . Each of the k -th identified P_o^k has an associated emotion annotation Y_{io}^k in the *out-of-context* dataset. The derived labels Y_{io}^{jk} are multi-perspective information based on P_i^j and X_o^k .
- 2) Enhanced-learning phase: We employ the proposed structure for each *in-context* corpus to learn an encoder-decoder network to derive the proposed enhanced acoustic latent representation and to integrate the *out-of-context* databases with multiple emotion perspectives, i.e., jointly learning the emotion label derived from *out-of-context* data as additional auxiliary label for the *in-context* sample [7], [8]. The trained encoder/generator is denoted as $G : X_i \rightarrow C$, which maps X_i onto C and the decoder portion is denoted as $F : C \rightarrow X'_i$; C is the learned latent acoustic representation, D denotes the discriminator, M is a recognition layer that is used in tuning G to integrate *out-of-context* emotion information (see Fig. 1). First, the *in-context* emotional acoustic code vector can be learned using the following modified reconstruction loss:

$$\mathcal{L}_{rec}(G, F) = \arg \min_{G, F} \|X'_i - X_i\|^2 \quad (1)$$

where $X' = F(G(X_i))$ represents the reconstructed features. The latent vector was further constrained by a Gaussian distribution, Z , where $p(z) = N(z|0, I)$. The adversarial loss is defined as follows :

$$\mathcal{L}_{adv}(G, D, X_i, Z) = \min_G \max_D E_{z \sim p_z} [\log(D(z))] + E_{x \sim p_{data}(X_i)} [\log(1 - D(G(X_i)))] \quad (2)$$

Then with this derived labels Y_{io}^K coming from the emotion-pooling phase, we modify the architecture to integrate this auxiliary *out-of-context* emotional information for every j -th sample for *in-context* database as an additional loss:

$$\mathcal{L}_{per}(X_i, G, M, Y_{io}) = \arg \min_{G, M} \|M(G(X_i)) - Y_{io}^K\|^2 \quad (3)$$

The complete objective function includes three different loss defined as below:

$$\mathcal{L}_{total}(G, F, D, M) = \mathcal{L}_{rec} + \mathcal{L}_{per} + \mathcal{L}_{adv} \quad (4)$$

4.3 Emotion Classifier

The discriminability of the proposed features is further validated in the SER task by training a final emotion classifier using vanilla DNN. We take $G : X \rightarrow C$ to derive the latent representation from being used as input to a DNN for three-classes emotion classifier.

5 EXPERIMENTAL SETUP

All of the experimental results are reported using unweighted average recall (UAR). We followed a strict cross-validation setting (leave one dyad out for Creative IT and RECOLA database and leave one person out for VAM database), where the learning of the enhanced acoustic latent vectors was performed only in training set for the entire pipeline. This work involves three *in-context* data—the Creative IT, the VAM, and the RECOLA—and three *out-of-context* databases—the IEMOCAP, the NNIME, and the MSP-Improv. All models were trained ten times with random initialization and the average accuracy is reported.

The number of epochs, batch size, and learning rate was 100, 64, and 0.005, respectively. Adam was used as an optimizer to minimize the objective function.

In experiment 1, we investigate the accuracy achieved using two different number of dimensions for enhanced acoustic latent representation for the labeled *in-context* databases. We treat the compared domain adaptation methods as feature-based extraction approach (denoted as 2-stage, S.2) as a feature extractor to train an emotion classifier for a fair comparison to the transfer learning methods.

In experiment 2, we compared the performance between the proposed method and the original domain adaptation method. Therefore, domain adaptation methods followed the original methods to obtain the prediction as an end-to-end approach (denoted as 1-stage, S.1). Additionally, our proposed method is turned as an end-to-end approach(S.1) to be compared with the original one(S.2).

There are four sub-network structures—encoder, decoder, discriminator, and multiple perspective classifier—within the architecture of the proposed methods (*prop.*). The encoder is a DNN architecture consisting of two fully connected layers with 128 and 256 hidden units

followed by a layer with a C hidden unit to produce the latent vector. The decoder consists of two fully connected layers with 256 and 128 hidden units followed by a layer with 88 hidden units to reconstruct the original feature. Furthermore, we used batch normalization, drop out layer with a dropout rate of 0.2, and ReLU as activation function in each fully connected layer in encoder and decoder. Discriminator consists of two fully connected layers with 64 and 32 hidden units, and ReLU as activation function followed by a layer with two hidden units and a softmax layer. Multiple perspective classifier is 3-layer DNN architecture with [size(C), 64, 32, size(K)]. Activation function between layers are leaky ReLU with parameter 0.2.

In conclude, encoded original feature (X_i) from encoder are connected to (a.)discriminator to recognize samples from latent vector or Gaussian distribution, to (b.)decoder to decode the code vector(C) to the original feature size(88) and to (c.)multiple perspective layer to predict multiple perspective emotion label (K) (in this case 10 annotations).

After that, in the S.2 condition, we consider the encoder as a feature extractor to extract the latent vector and train NN, containing 1-layers with 10 hidden units, and ReLU as activation function followed by a classification layer and a softmax layer.

Additionally, in the S.1 condition, a classifier is added to the network. It consists of two fully connected layers with 128 and 64 hidden units and ReLU as activation function followed by a classification layer with a softmax layer.

We also implement the following baseline models in deriving the latent acoustic representation:

- 1) Domain adaptation methods:
 - a) **Adversarial Discriminative Domain Generalization:** We compared our results with adversarial discriminative domain generalization (*addog*), which is a state-of-the-art domain adaptation method in adversarial training [10]. It combines encoder, classifier, and critic that measures the distance from source and target. The encoder is similar to that of (*prop.*). The classifier consisted of two fully connected layers with 128 and 64 hidden units and ReLU as activation function followed by a classification layer and a softmax layer. The critic also consisted of a fully connected layer with 64 hidden units and ReLU as activation function followed by a fully connected layer to estimate the earth mover's distance. The classifier would predict the result in S.1 condition. Moreover, the latent vector C was taken for training a NN, in S.2 condition, to compare with our proposed method in experiment 1.
 - b) **Domain Adversarial Neural Network:** *dann* proposed by Abdelwahab et al. is an unsupervised domain adaptation method with adversarial training that we take to compared with our results [32]. Domain adversarial neural network (*dann*) proposed by Abdelwahab et al. is an unsupervised domain adaptation method with adversarial training. We used to compare it with our results [32]. We conducted three parts of (*dann*): encoder, classifier, and domain classifier. The encoder and classifier were similar to that of (*addog*) for a fair comparison. The

domain classifier consisted of one fully connected layer followed by a classification with a log softmax layer. In S.1 condition, the classifier predicted the result. In S.2 condition, we considered the encoder as a feature extractor to extract the latent vector and train a NN in experiment 1.

- 2) Representation learning method:
 - a) **Pre-trained Neural Network:** A pre-trained neural network (*pre.*) indicates pre-training with *out-of-context* and taking out the encoder to extract the representation for *in-context* data to train a classifier. The pre-train neural network was divided into an encoder and classifier. The encoder and classifier were similar to those of (*addog*) for a fair comparison. In S.1 condition, the classifier predicted the result. After that, we considered the encoder as a feature extractor to extract the latent vector and train a NN, denoted as S.2 condition.
 - b) **Semi-Supervised Adversarial Autoencoding:** We also compared the proposed method with semi-supervised adversarial autoencoding (*sAA*) [25], [40], which is a state-of-the-art framework for SER. It is conducted four parts: encoder, decoder, discriminator, and classifier. The encoder, decoder, and discriminator had similar architecture to our proposed method for a fair comparison. The classifier was similar to that of (*addog*).
 - c) **Semi-Supervised Autoencoder:** Semi-supervised autoencoder (*sAE*) indicates the same framework as in [38]. In this work, *out-of-context* data is passed through a supervised classifier and decoder, and *in-context* data is passed through the only decoder during the training. After that, *in-context* data passed through the final classifier to obtain the final prediction during testing. It was divided into an encoder, decoder, and classifier. The encoder, decoder, and classifier had similar architectures to our proposed method.

We note that annotations from *out-of-context* were encoded as same as those from *in-context* in *addog*, *dann*, and *pre.* because it is needed to encode source's annotation to be as same as target's for end-to-end approaches. On the contrary, our proposed method aims to integrated *out-of-context* information to help train *in-context* databases without further adjusting annotations.

The ablation work is shown in experiment 3. All the settings of encoder, decoder, discriminator were as same as *prop.*. Therefore, we compared our framework with the following ablation models to derive the latent acoustic representation:

 - 1) **Autoencoder:** We built an vanilla autoencoder (*AE*) without a discriminator and a multiple perspective classifier as a ablation to our proposed network.
 - 2) **Adversarial Autoencoder:** We built an adversarial autoencoder(*AAE*) for each target corpus and use the latent layer as input to the NN.
 - 3) **Multiple Perspective Autoencoder:** Multiple perspective autoencoder(*MPAE*) is used, i.e.,

TABLE 1: Summary of comparison between our proposed adversarially-enriched acoustic vector and state-of-the-art for emotion classification.(UAR%) Noted: 64, 10: 64 or 10 dimensional latent vector. A. : activation and V. : valence.

In-Context: CIT															
	out-of-context: IEMOCAP					out-of-context: NNIME					out-of-context: MSP				
	sAE	sAA	dann	addog	prop.	sAE	sAA	dann	addog	prop.	sAE	sAA	dann	addog	prop.
64 A.	42.3	46.7	45.8	48.3	53.2	42.7	45.6	46.1	48	53.1	42.5	45.8	48.2	48.1	53.8
64 V.	36.9	38.2	37	37.2	40.6	36.9	38.1	36.1	36.1	40.7	37.1	38.7	36.9	36.9	41.7
10 A.	45.3	47.6	44.8	48.2	54.8	43.9	48.5	45.6	47.8	55.2	44.6	48.4	48	47.9	53.1
10 V.	37.3	37.6	35.7	35.6	39.8	36.5	37.2	34.9	35.5	39.2	36.7	37.1	37	37.8	39.5

In-Context: VAM															
	out-of-context: IEMOCAP					out-of-context: NNIME					out-of-context: MSP				
	sAE	sAA	dann	addog	prop.	sAE	sAA	dann	addog	prop.	sAE	sAA	dann	addog	prop.
64 A.	61.7	64.2	64	65.4	68.8	62.1	63.7	63.8	63.9	70	62.2	63.7	64.8	65.7	69
64 V.	43.4	43	42.2	45	45.4	41.4	42.3	37.5	39.4	46.2	42.4	42.6	44.6	43.9	47
10 A.	60.2	65.5	60.9	62.9	68.6	62.3	67	58.1	59.8	68.4	61.5	65	62.6	64	70.1
10 V.	43.3	42.5	43.8	43	46.7	42.1	43.9	39.4	39.6	43	42.7	43.1	44	43.8	46.8

In-Context: RECOLA															
	out-of-context: IEMOCAP					out-of-context: NNIME					out-of-context: MSP				
	sAE	sAA	dann	addog	prop.	sAE	sAA	dann	addog	prop.	sAE	sAA	dann	addog	prop.
64 A.	57.3	59.1	58.6	57.7	62	58	58.6	59.2	59	62.4	57.8	58.8	58	60.9	62.2
64 V.	48.1	49.1	50.2	52.1	55.8	49.1	49	50.9	50.7	53.6	48.1	48.3	51.5	51.9	55.2
10 A.	59.2	61.6	57.7	58.6	62.7	59.1	61.1	57.4	57.3	62.5	59.9	61.2	57.9	60.1	63.2
10 V.	51.1	52	50.5	50.8	56.3	49.9	50.5	51	50.9	54.7	49.5	50.8	51	52.7	57.2

without a discriminator as a ablation to our proposed network. There are two sub-networks of MPAAE: The *out-of-context* multiple perspective classifier and the *in-context* autoencoder.

4) **Adversarial Multiple Perspective Encoder:**

Adversarial multiple perspective encoder(*AMPE*) is used, i.e., without the decoder as a ablation to our proposed network.

In experiment 4, the proposed framework uses multiple emotion perspectives derived using the ‘closest’ *out-of-context* sample’s emotion rating. We additionally conducted experiments to investigate whether different manners of choosing the derived labels would have an effect on the *in-context* emotion classification accuracy. Specifically, we evaluated the following two methods in deriving multiple emotion perspectives:

1) *Inverse:*

When deriving the multiple perspectives for the *in-context* samples, we considered the emotion labels of those *out-of-context* samples that were furthest away as the derived emotion labels. We term this method as an *Inverse* perspective.

2) *Random:*

We picked random emotion labels as the derived label to be used as the multiple emotion perspectives.

Furthermore, we employed the t-SNE algorithm in experiment 5 to visualize the latent representation from the proposed method and its comparisons.

Lastly, we conducted an investigation on the effect of reduced *in-context* labeled samples in experiment 6 to observe the tendency.

6 EXPERIMENTAL RESULTS AND ANALYSIS

6.1 Comparison with State-of-the-art

We compared our result with the newly adversarial research, semi-supervised adversarial autoencoding framework [40], the domain adaption method, *dann* (S.2) [32], and *addog* (S.2) [10].

Table 1 shows the result of our proposed method and comparisons in two emotion annotation (activation and valence) when using 64-dimensional and 10-dimensional code vector paired with every *out-of-context* database. By examining results obtained using 64-dimensional vector, with the data of *out-of-context*, IEMOCAP, we obtain 53.2%, 68.8%, 62% in the activation in three *in-context* database, i.e., the CIT, the VAM, and the RECOLA respectively, as compared to 42.3%, 61.7% and 57.3 % obtained using *sAE*, which all increases around 11%, 7.1%, 4.7%. A similar trend is also observed in the valence dimension. Specifically, we achieved 40.6%, 45.4%, and 55.8% across the three *in-context* databases than achieving 36.9%, 43.4%, and 48.1% using *sAE* with an approximate increment of 3.7%, 2%, and 7.7%. It is similar when using other *out-of-context* database, NNIME and MSP. After that, we compare the proposed methods to *sAA*, which includes a classifier in its framework and utilizes *out-of-context* data to update the model unsupervised. We observe by examining results obtained using 64 dimensions that the proposed method surpasses *sAA* for the activation and valence. We obtained 53.1%, 70%, and 62.4% in the activation with *out-of-context* data, NNIME, than 45.6%, 63.7%, and 59.2% obtained using *sAA* in three *in-context* databases—the CIT, the VAM, and the RECOLA. Moreover, 40.7%, 46.2%, and 53.6% using the proposed method surpasses 38.1%, 42.3%, and 49% obtained using *sAA* in the valence. Additionally, We attempt to compare our proposed

method with domain adaptation methods (*dann* and *addog*) in similar conditions, that is, taking their encoder to extract features for the *in-context* database and train a classifier (S.2). We investigated results obtained using 64 dimensions, e.g., with *out-of-context* data, MSP, our proposed model are better than *dann* and *addog*, (53.8% versus 48.2% versus 48.1%), (69% versus 64.8% versus 65.7%), (62.2% versus 58% versus 60.9%) in activation, as well as (41.7% versus 36.9% versus 36.9%), (47% versus 44.6% versus 43.9%), (51.5% versus 51.9% versus 55.2%) in valence in three *in-context* database, i.e., the CIT, the VAM, and the RECOLA respectively.

Moreover, we note few interesting points besides the comparison between the proposed method and state-of-the-arts. First, we find that performance with *out-of-context*, NNIME, may be lower than with other *out-of-context* data from the observation in the performance of two annotations using *sAA*, *dann*, *addog* in three *in-context* databases. Similar circumstances are shown in Table 2. It may be due to the fact that these *in-context* data are closer to IEMOCAP and MSP than to NNIME. *sAA* trained straightforwardly using *out-of-context* without annotation would cause the model to include unwanted variability of *out-of-context*, leading to a similar situation with *dann* and *addog*. On the contrary, *prop.* integrated the *out-of-context* information only for those data that were similar to the *in-context* database instead of interference from the variability of *out-of-context*. Second, domain adaptation methods, such as *dann* and *addog*, force the same annotation setting between source and target, i.e., source and target annotations should have the same label space. It would then require dividing the annotation manually if the setting is different from source and target data. However, the proposed method directly uses the original annotation from the *out-of-context* to derive multiple perspective information.

Further, we report the accuracy obtained using only a 10-dimensional acoustic latent vector to evaluate the robustness of the framework. Fewer dimensions of the vector may lead to low usage of memory when applied in the real world. We obtained 54.8%, 68.6%, and 62.7% accuracy with the data of *out-of-context*, CIT, in the activation dimension than 45.3%, 60.2%, and 59.2% in the *sAE*, CIT, VAM, and the RECOLA database. For valence, we obtained 39.8%, 46.7%, and 56.3% as compared to 37.3%, 43.3%, and 51.1% in the *sAE*. Additionally, latent code constrained by Gaussian distribution may keep its discriminability in the low dimensional condition. However, the latent vectors from DA methods were constrained into a common space between source and target, which do not explicitly improve emotion discriminability in the latent. Hence, we find that the *AAE*-based architecture like *sAA* and *prop.* would decrease the accuracy only slightly or even improve the accuracy when using low-dimensional features. On the other hand, DA-based architecture such as *dann* and *addog* would decrease the accuracy mostly in activation.

6.2 Comparison with Domain Adaptation Method

Domain adaptation(DA) on SER have been proposed recent years. The purpose of DA on SER is similar to our framework, i.e., to utilize the information from another corpus (*out-of-context* database) to build an emotion classifier to be used directly in the target corpus (*in-context* database).

Specifically, Table 2 shows the results of comparing our proposed method with a series of domain adaptation methods. In this table, our proposed method *prop.* in S.2 implies that the latent code was trained on the enhanced-learning phase, and a DNN emotion classifier was then trained by these latent codes. However, *prop.* in S.1 was directly trained and predicted with the *in-context* database by adding a classifier in the enhanced-learning phase to perform prediction. Meanwhile, *pre.*, *dann* and *addog* were trained with *out-of-context* database and directly predicted the *in-context* database. There are several points to discuss.

TABLE 2: Comparison between our proposed method using 64-dimensional latent vector and domain adaptation methods.(UAR%) Noted: S.1: end-to-end approach network getting a prediction with a emotion classifier inside the network. S.2: feature based extraction approach serving as a feature extractor to train a additional emotion classifier.

64 Dim.		CIT		VAM		RECOLA	
		Act.	Val.	Act.	Val.	Act.	Val.
<i>out-of-context: IEMOCAP</i>							
S.1	<i>dann</i>	38.2	39.4	42.3	37.9	56.8	49.2
	<i>addog</i>	40.2	40.1	46.4	40.8	58.5	49.9
	<i>pre.</i>	37.5	39.5	39.9	38.5	56.2	51
	<i>prop.</i>	51.8	39	63.6	45.6	60.7	51.7
S.2	<i>pre.</i>	47.9	35.4	66.1	44.7	57.8	50.8
	<i>prop.</i>	53.2	40.6	68.8	45.4	62	55.8
<i>out-of-context: NNIME</i>							
S.1	<i>dann</i>	37	35.1	38.9	38.9	47.7	48.8
	<i>addog</i>	36.8	36.5	39.5	35.7	48.9	49.4
	<i>pre.</i>	37.9	34.9	39.9	36.5	48.1	48.2
	<i>prop.</i>	52.2	39.4	64.7	46.1	61.1	52.1
S.2	<i>pre.</i>	46.8	36.8	65.5	38.3	58	52.5
	<i>prop.</i>	53.1	40.7	70	46.2	62.4	53.6
<i>out-of-context: MSP</i>							
S.1	<i>dann</i>	43.7	36.2	50.8	38.4	49.7	47.4
	<i>addog</i>	43.9	37.9	49.9	39.7	47.8	48.6
	<i>pre.</i>	43.4	35.7	49.9	40.4	48.4	47.3
	<i>prop.</i>	52.3	39.7	64.1	44.8	61.3	51.1
S.2	<i>pre.</i>	47.5	37.4	65.8	46.2	58.7	51.3
	<i>prop.</i>	53.8	41.7	69	47	62.2	55.2

First, we compare the DA methods with our proposed methods. In the activation, it is obvious that the performance of *prop.* in S.2 outweighs that of DA a lot, i.e., the result from *out-of-context*, IEMOCAP, 53.2% of *prop.* compared to 37.5%, 38.2%, 40.2% in *pre.*, *dann* and *addog* in the CIT, 68.8% of *prop.* compared to 39.9%, 42.3%, 46.4% in the VAM, 62% of *prop.* compared to 56.2%, 56.8%, 58.5% in the RECOLA. It has a similar trend on the different *out-of-context* databases—NNIME and MSP. Moreover, in the the performance of valence , it increased to approximately 4.2% from *addog* to *prop.* in the CIT database, 5.8% from *pre.* to *prop.* from *out-of-context*, NNIME. It may be due to the difficulty in recognizing valence from speech, leading to low performance in most of the methods. Additionally, we note that the effect of domain adaptation on valence is usually better than that on activation [10], [32]. However, training deep models on the RECOLA database is much more challenging than other databases. The label distri-

TABLE 3: Ablation comparison of our proposed adversarially-enriched acoustic vector for emotion classification using 64-dimensional and 10-dimensional latent vector and domain adaptation methods in activation and valence.(UAR%)

In-Context: CIT													
		AE AAE		out-of-context: IEMOCAP			out-of-context: NNIME			out-of-context: MSP			
		MPAE	AMPE	prop.	MPAE	AMPE	prop.	MPAE	AMPE	prop.	MPAE	AMPE	prop.
Act.	64 Dim.	48.3	51.3	51.4	47.7	53.2	51.3	50.1	53.1	50.9	47.8	53.8	
	10 Dim.	47.7	51.7	52.3	49.9	54.8	53.1	49.6	55.2	51.2	49.3	53.1	
Val.	64 Dim.	37.1	37.8	37.3	37.1	40.6	37.6	37.3	40.7	37.3	35.4	41.7	
	10 Dim.	37	37.1	38.3	36.1	39.8	37.8	36.5	39.2	36.4	35.5	39.5	

In-Context: VAM													
		AE AAE		out-of-context: IEMOCAP			out-of-context: NNIME			out-of-context: MSP			
		MPAE	AMPE	prop.	MPAE	AMPE	prop.	MPAE	AMPE	prop.	MPAE	AMPE	prop.
Act.	64 Dim.	63.3	67.5	66.7	60.9	68.8	67.5	66.9	70	67.2	59.9	69	
	10 Dim.	62.4	68.1	67.2	66.1	68.6	67.7	66.5	68.4	67.9	65.9	70.1	
Val.	64 Dim.	41.1	42.4	41.3	36.1	45.4	42.5	38.8	46.2	41.5	35.2	47	
	10 Dim.	42.1	43.1	43.8	38.3	46.7	43.4	42.2	43	43.4	38.7	46.8	

In-Context: RECOLA													
		AE AAE		out-of-context: IEMOCAP			out-of-context: NNIME			out-of-context: MSP			
		MPAE	AMPE	prop.	MPAE	AMPE	prop.	MPAE	AMPE	prop.	MPAE	AMPE	prop.
Act.	64 Dim.	60.5	60.6	58.7	58.1	62	60.5	59.8	62.4	60.2	57.8	62.2	
	10 Dim.	60.4	60.8	60.6	60.1	62.7	61.1	60.6	62.5	61.1	59.7	63.2	
Val.	64 Dim.	51	52	51.9	50.5	55.8	51.6	52.6	53.6	51.7	50.8	55.2	
	10 Dim.	51.5	52.3	54	52.5	56.3	53.8	52.6	54.7	53.2	53.3	57.2	

bution is highly imbalanced, and the sample number is small. Therefore, performing SER in the RECOLA database is still difficult when the domain adaption method is used, with many results in using DA are lower than the random guessing. Overall, a transfer from a labeled speech corpus to each particular context underestimates the true variability of the distribution of limited data. In this case, the latent code integrating the annotation information from the source may be a better solution to deal with limited speech corpus in a particular context.

Second, we investigated the difference between S.1 and S.2, whether an emotion classifier is trained separately or jointly. The performance in S.2 is significantly better than in S.1 in activation for the *pre.* model in S.1 and S.2. The reason is that the network is trained for the *out-of-context* data, where the distribution is different from *in-context* data. For valence, the tendency is similar except when the IEMOCAP as the *out-of-context* data is used. We then compared our proposed method (in S.2) with *prop.* in S.1. For the activation, the performance of S.2 surpassed that of S.1 in all the situations, increasing more than 5% in the VAM database (68.8% versus 63.6%, 70% versus 64.7%, and 69% versus vs 64.1% from *out-of-context*, IEMOCAP, NNIME and MSP). For valence, the disparity between them are is not so significantly notable, i.e., it is around 1% to 3% from *prop.* of S.2 to that of S.1. Overall, we find that the *prop.* of S.2 are better than proposed research, including an emotion classifier (S.1), especially for activation. The latent code produced by the proposed approach may have further discriminability in training a simple deep neural network. However, a framework including an emotion classifier (S.1) may confuse the encoder while attempting to update the loss from the classifier, discriminator, and multiple perspective layer simultaneously. Moreover, both *prop.* perform

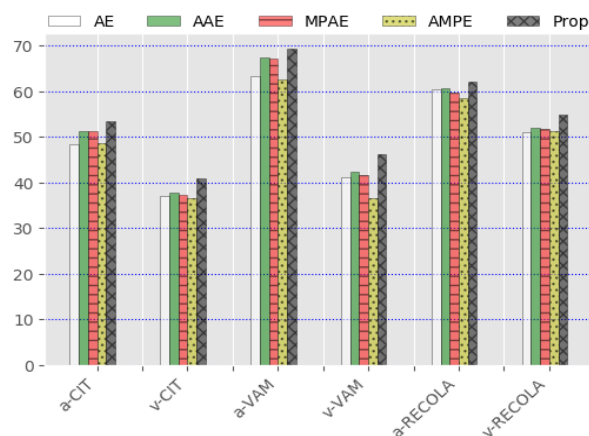


Fig. 2: The total average accuracy of each *out-of-context* databases of each label, which averages all the UARs obtained for each *in-context* data(CIT, VAM, RECOLA) of two different emotion attributes(activation(a-) and valence(v-)) for different models.

better than domain adaptation methods, which shows the effect of our framework in dealing with databases with a limited amount of labeled data.

6.3 Ablation Comparison

Table 3 shows the result of the ablations compared to our proposed method using a 64-dimensional and 10-dimensional code vector. Furthermore, each *in-context* corpus was paired with every *out-of-context* database; the results are shown in Table 3. There are two observations. First, the performance of the proposed method for the activation is better than other ablations, *AE*, *AAE*, *MPAE* and *AMPE* by either using a 64-dimensional or 10-dimensional

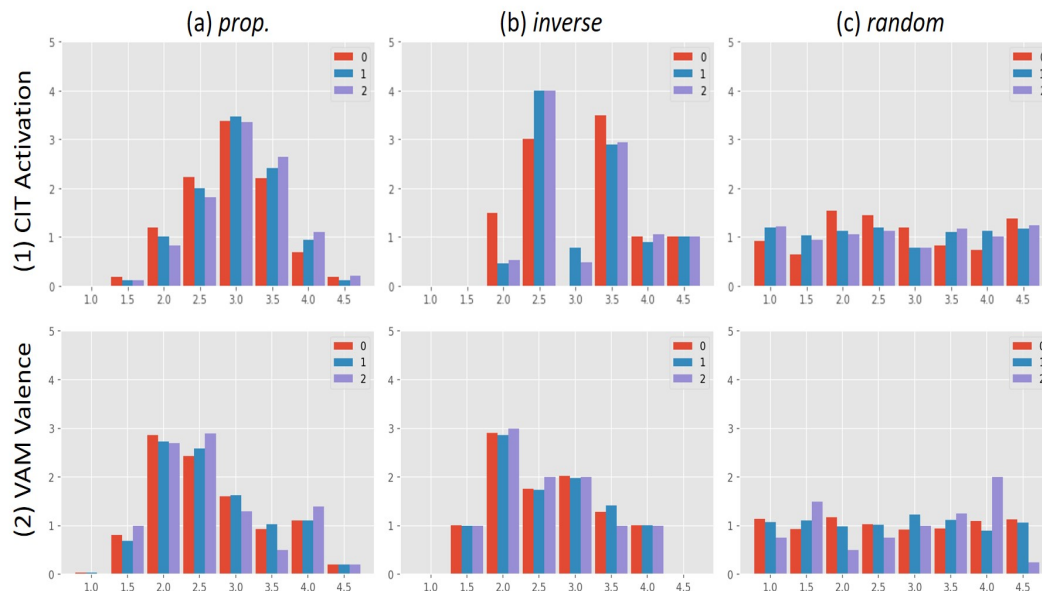


Fig. 3: The comparison of using (b)*inverse* and (c)*random* versus the proposed method ((a)*prop.*) in deriving multiple emotion perspective for (1) activation of the CIT corpus and (2) valence of the VAM corpus as the *in-context* data with the IEMOCAP as the *out-of-context* data. *random* indicates that each *in-context* sample produces 10 random derived ratings that imitate the annotation characteristic from each *out-of-context* data to be used as its derived labels. *inverse* indicates the same multiple perspective computation except using the labels of those samples that is furthest away from each *in-context* sample. The x-axis shows the derived labels plotted distributionally on the scale of 1 to 4.5. The y-axis represents the average count of selected *out-of-context* emotion annotation scale. The different bar indicates the average count of the in-context labels('0', '1' and '2' or low, mid, and high) derived or chosen for that level of *out-of-context* emotion scale ranging from 1-4.5. These histogram shows the proposition of 10 derived different perspective for specific *in-context* classes and its composition.

vector in the CIT, VAM, and RECOLA database, respectively. We find that fewer latent codes can be extracted to maintain the performance within the small database, which is helpful for the limited computing resource. For the valence, a similar trend is observed using a 64-dimensional vector besides a minor drop in the VAM database when using a 10-dimensional vector (43% versus 43.4%). Second, there is an obvious drop with *AMPE* in every condition. *AMPE* is proposed without a decoder or reconstruction layer, which may cause the latent code to lose its original representation power over the *in-context* data. Third, the performance between *AAE* and *MPAE* is approximately the same and better than *AE* in the three databases. Nevertheless, *prop.* surpasses both of them owing to combining the discriminator and multiple perspective classifier, which may imply that discriminator help latent code to be more stable to avoid the overreaction from multiple perspective classifier and vice versa. Hence, this may cause *AAE* or *MPAE* to be lower than the proposed method.

Fig. 2 shows the results, which average all the UARs obtained for each *in-context* of two different emotion attributes for different models. Results show that our adversarial learning strategy requires all components to be jointly considered in enhancing the affective information of the *in-context* latent. It is evident from the better performance *prop.* obtained than those of ablations.

TABLE 4: Comparison of different manners in deriving multiple emotion perspectives using 64-dimensional latent vector. *random* indicates producing K random derived ratings for each *in-context* sample. *inverse* indicates using labels of K furthest samples from each *in-context* sample. K is 10 here.

64 Dim.	CIT		VAM		RECOLA	
	Act	Val	Act	Val	Act	Val
<i>AAE</i>	51.3	37.8	67.5	42.4	60.6	52
<i>inverse</i>						
IEMOCAP	52.2	36.8	66.8	38.6	59.2	52.8
NNIME	51.4	37.2	68	39.9	60.4	52.6
MSP	52.6	37.2	66.8	41.1	60.3	52.6
<i>random</i>						
IEMOCAP	50.7	36.9	66.8	40.6	60.2	50.8
NNIME	50.7	37.7	65.6	41.1	60.3	50.6
MSP	50.5	37.8	65.9	40.2	60.2	49.7
<i>prop.</i>						
IEMOCAP	53.2	40.6	68.8	45.4	62	55.8
NNIME	53.1	40.7	70	46.2	62.4	53.6
MSP	53.8	41.7	69	47	62.2	55.2

6.4 Different Manners of Multiple Emotion Perspectives

Table 4 shows the results that compare the different manners in deriving multiple emotion perspectives using 64 dimensional latent vector. *AAE* indicates proposed model without any multiple emotion perspectives. *random* indicates that

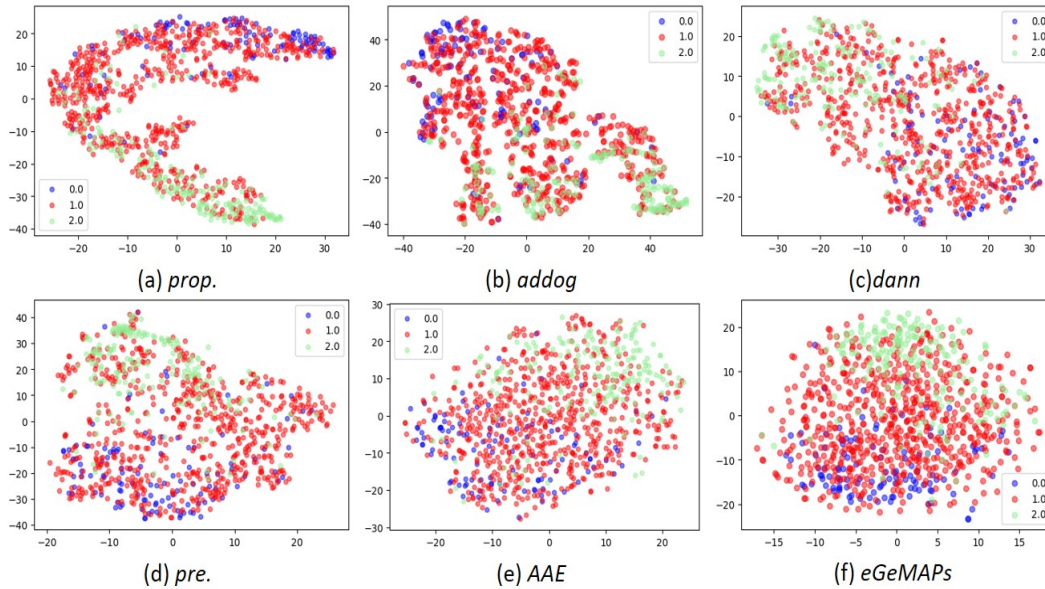


Fig. 4: Visualization of the latent representation by the t-SNE algorithm of our proposed method, *addog*, *dann*, *pre.*, *AAE* and *eGeMAPS* feature in activation with *in-context* database—the VAM corpus using 10-dimensional latent representation. It is marked in three classes—low (0), mid (1), and high (2)—with a distinct color.

we produce K random derived ratings for each *in-context* sample, imitating the annotation characteristic from each *out-of-context* data to be used as its derived labels. *inverse* indicates the same multiple perspective computation except using K labels of those furthest samples from each *in-context* sample. K is 10 here.

There are several points to note. First, we compared *random* with *AAE*. The model is clearly misled by the randomly given wrong information. The UARs of *random* were lower than *AAE* for activation. The result from *out-of-context* database, IEMOCAP, was 50.7%, 66.8%, and 60.2% than 51.3%, 67.5%, and 60.6% in the CIT, VAM, and RECOLA, respectively. There is a similar tendency for valence. The UARs from *random* are worse than *AAE*, i.e., the result from *out-of-context* database, IEMOCAP, was 36.9%, 40.6%, and 50.8% than 37.8%, 42.4%, 52%, respectively.

Second, we observe that *inverse* emotion perspectives may provide distinct information to the model. We notice while examining the *inverse* condition that most of the results perform better than the *random*, and few of the results perform better than *AAE*. For example, the result on the CIT database—IEMOCAP, NNIME, and MSP was 52.2%, 51.4%, and 52.6% compared to 51.3%. However, the *inverse* method of integrating *out-of-context* data is not as effective as the proposed method. All of these results are worse than the proposed method. For instance, we have achieved the *best* UAR on *prop.*, 53.8%, 70%, and 62.4% in activation as well as 41.7%, 47%, and 55.8% in valence than the performance of other manners.

Moreover, to see the behavior of the selection or derivation of multiple perspective from different manners; we analyze the count of multiple perspective for each *in-context* emotion class lying within *out-of-context* emotion scale for different mentioned manners. Results of this analysis are shown in Fig. 3, where the x-axis shows the *out-of-context* label scale, y-axis represents the average count of selected *in-context* emotion classes and the different bar indicates the

average count of the *in-context* labels ('0', '1' and '2') derived or chosen for that level of *out-of-context* emotion scale ranging from 1-4.5. These histogram shows the proposition of 10 derived different perspective for specific *in-context* classes and its composition. As we are choosing K or 10 perspective; the count of derived perspective for each respective class are summed to 10. For example, class '0' of *in-context* CIT database (red bar) derives multiple perspectives; more from *out-of-context* label 3.0 and less from label 1.5 and 4.5 and so on; all of them are summed to 10 in (1)-(a).

Fig. 3 (top) shows the difference between *inverse*, *random* and our proposed method for activation attribute using the CIT as the *in-context* data and the IEMOCAP as the *out-of-context* data. From (1)-(a), the *prop.* annotations were chosen from the IEMOCAP appear to be a Gaussian-like distribution. The samples in the CIT with labels '0', '1' and '2' prefer to choose 3; however, the samples in the CIT with the label of '0' (low activation) prefer low values, and the samples in the CIT with the label of '2' (high activation) prefer high values if looking into histograms of each position respectively. In (1)-(b), the *inverse* annotations were chosen from the IEMOCAP as the derived labels appear to be a little more polarized. The sample of '0' prefer to choose 3.5. Moreover, the sample of '1' and '2' all choosing 2.5 simultaneously may confuse the model to have a worse performance on prediction. It potentially leads to integrating a more diverse variability for the CIT acoustic representation (i.e., maybe missing originally), benefiting the activation recognition performances. On the contrary, most of the derived labels seem to be located evenly in (1)-(c) in the Random setting. This regularity may be detrimental (or simply useless) in affecting the recognition performances.

Fig. 3 (bottom) shows the difference between *inverse*, *random*, and our proposed method for valence attribute when using the VAM as the *in-context* data and the IEMOCAP as the *out-of-context* data. There are two observations. First, in (2)-(a), label '0' (low valence) prefer to choose the

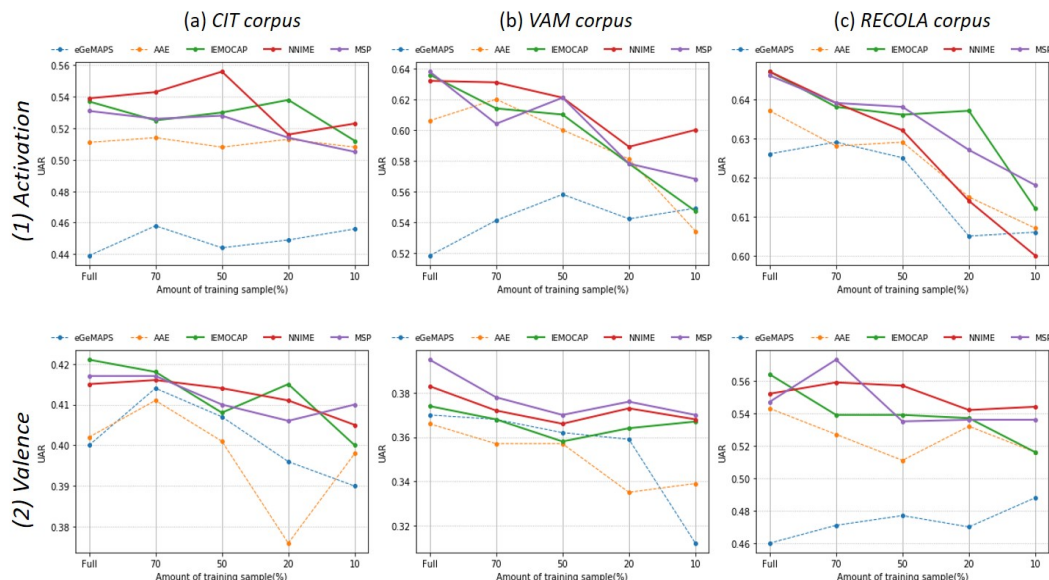


Fig. 5: Tendency of reducing sample from *in-context* data when using our proposed method, *AAE*, and *eGeMAPS*. (a), (b), (c) indicate training and testing with the CIT, the VAM, and the RECOLA corpus, respectively. (1) and (2) indicate the annotation, activation, and valence. Additionally, our proposed method is represented with a solid line, and different color indicates various *out-of-context* data. The methods used for the comparison are represented with a dotted line.

annotation 2 as its multiple emotion perspectives from the IEMOCAP. Samples in the VAM with the label of '2' (high valence) prefer to choose the annotation 2.5 from the IEMOCAP. We believe that this additional common affect-related variability that is being integrated into the acoustic latent representation would help improve the overall recognition performance. Secondly, in (2)-(b), the sample of '0' and '2' choose a nearly similar number from a scale of 1.5 to 4, leading to a worse performance than the one without multiple perspectives (38.6% (inverse) versus 42.4% (AAE), see table 4).

6.5 Visualization of Learned Representation

The t-SNE algorithm was used to visualize the distribution of *in-context* data to further assess the effect of our latent representation from *prop.*. Fig. 4 illustrates the 2D projection of the feature representation using the t-SNE algorithm to compare *prop.* with *aaddog*, *dann*, *pre.*, *AAE* and original extracted feature, *eGeMAPS*. Colors indicate the low, mid, and high of the annotation. The VAM corpus with activation using a 10-dimensional latent vector is used for the visualization.

We present some observations from the graphs. First, the polar side of the annotation (low versus high) appears to be separable in all the graphs. However, the position of mid may play an important role in determining the quality of the latent code. Specifically, the position of mid samples from (f) covers high and low samples; moreover, all samples of activation are sparse on the graph. A similar tendency can be seen in (c) and (e). The sparsity of the mid sample of (e) makes it challenging to identify samples between 0 and 2. Mid sample of (c) is also sparse; however, the low and high samples of (c) are concentrated and split with mid sample, unlike (e). Second, the clustering effect of similar annotation is important. (b) illustrates the latent vector visualization of *addog*. Compared to (b), clustered samples are more evident

in (a); however, (b) has much more clustered samples of low and high than (c), (e), and (f). Third, (a) illustrates the latent vector visualization of *prop.* The green dots (high activation) are accumulated on the bottom right, and the blue dots (low activation) are clustered on the upper right of the graph. However, the red dots (mid activation) are spread on the left side of the graph. The primary difference is that mid samples do not cover the samples of another category significantly; moreover, they have clusters samples that appear more discriminative and concentrated than the graph of other methods.

6.6 Reducing *in-context* Labeled Data

We further conduct conducted experiments to assess the limitation of our proposed framework. The core idea of our method is to enhance the acoustic latent representation of a small-scale affective *in-context* database using an *out-of-context* emotion database. An additional analysis is performed to assess the recognition rates achieved while reducing the number of *in-context* labeled samples to investigate the limitation of our framework under a severe lack of labeled data situation. Specifically, we split the *in-context* database into training and testing sets instead of cross-validation, as mentioned above. We used 20% of the data as the testing set, which belongs to different folds from the training set, e.g., there were eight folds in the CIT corpus, and two specific folds were selected as testing sets. We evaluated our complete framework using a sample size of the full set, 70%, 50%, 20%, and 10% of the training set (full set implies the entire training set or 80% of total data). Additionally, all the models were trained ten times with random initialization and then averaged the performance.

Fig 5 shows the line chart of this experiment. There are 6 slots on Fig 5. The CIT corpus(a), the VAM corpus(b) and the RECOLA corpus(c)and two annotations, activation(1) and valence(2) are compared with two baseline

(eGeMAPS, *AAE*) with dotted line as well as three *out-of-context* database with solid line.

The graph shows that there are two main situations while performing this experiment. Most of the methods suffer a loss of accuracy when decreasing the number of available samples like (1)-(b) or (1)-(c). Moreover, the methods maintain the performance when decreasing the number of available samples, such as (1)-(a) and the slots of (2). This experiment demonstrates an important characteristic of our framework, i.e., robustness against a lack of data training conditions. For example, there is an obvious drop in baseline valence accuracy in (2)-(a) of the CIT corpus when *eGeMAPS* and *AAE* methods are used; however, our proposed method maintains its performance close to 41% (From full set to 10%). A similar situation *eGeMAPS* in (1)-(a). The proposed method and *AAE* maintain the performance around 54%, while *eGeMAPS* maintain the performance of around 45%. This effect is also observed in (2)-(b) of the VAM corpus. After reducing a full set to 10%, the accuracy of the *eGeMAPS* and *AAE* model drops from approximately 36% to 34% and 32% in valence. However, the proposed method maintains an accuracy comparable at approximately 37% for valence from the full set to 10% of data samples. For the activation, it is an obvious drop from the full set to 10%; however, our proposed method surpasses *AAE* and *eGeMAPS* in most conditions.

A similar trend was observed in (2)-(c) of the RECOLA corpus. When reducing the sample size from full set to 10%, our method experiences a loss of accuracy because the number of the sample decreases. However, it still maintains a competitive recognition rate. Although there is a drop on (1)-(c), the amplitude of the drop is approximately 3 to 4%. Moreover, our proposed methods win over *AAE* and *eGeMAPS* in most conditions at the drop.

This experiment demonstrates that our proposed framework is potentially beneficial for a wide range of *rapid* adoption of SER technology across contexts where one does not have to start with a large labeled corpus; Instead, a workable recognition system could be readily achieved by learning to enrich the acoustic latent representation with existing available speech emotion corpora.

7 CONCLUSION AND FUTURE WORK

In this paper, we propose a novel framework in the adversarially enhanced acoustic latent representation of the small-scale affective corpus to improve the robustness of the SER algorithm. We propose a new learning paradigm unlike the conventional approach in treating the cross contexts robustness issue as an unsupervised transfer learning problem because there is not a single database that is large enough to cover the adequate variability in speech SER that can be suitable for transfer learning, i.e., to integrate the *out-of-context* database to aid the learning of small scale *in-context* acoustic representation, to improve the SER accuracy. This work evaluates the SER improvement when using our proposed adversarial learning strategy in three *in-context* databases with three *out-of-context* databases. We also provide an extensive analysis of the effect of various parameters of this learning strategy on the accuracy obtained, specifically focusing on the manner of deriving multiple emotion

perspectives and the amount of *in-context* data samples. Generally, we observe that the proposed method needs only a few dimensions of a latent vector with a limited amount of data, which is enough to achieve competitive recognition accuracy. This result points toward potential benefit in advancing the rapid adoption of SER across contexts.

There are several future directions. First, the results in Exp 6 show that an even lower number of samples can achieve a stable recognition accuracy. This may indicate that the manner selects which *in-context* samples are critical. It means that picking those samples that are sufficient in covering adequate affective vocal variability would be an important step. A systematic method in picking the *right* samples for the annotation would further reduce the number of samples needed in this learning paradigm. Second, this work uses derived emotion labels from the *out-of-context* database as auxiliary information to be integrated in the learning of the enhanced representation. We would like to investigate other key factors that are known to affect and modulate the speech emotion characteristics, e.g., the types of interactions, the language of the database, and the personality of the subjects, as information to be integrated into our framework. Finally, we continuously explore other modalities that can be jointly learned together. With the advancement of robust SER that can handle multiple contextual settings *rapidly*, i.e., without the need in collecting a large labeled corpus, we are hopeful in deploying SER technology widely and potentially enhancing the user experiences for these modern human-machine interfacing systems.

REFERENCES

- [1] R. Van Bezooijen, S. A. Otto, and T. A. Heenan, "Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics," *Journal of Cross-Cultural Psychology*, vol. 14, no. 4, pp. 387–406, 1983.
- [2] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, 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.
- [3] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 500–504, 2017.
- [4] J. A. Russell, "Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies." *Psychological bulletin*, vol. 115, no. 1, p. 102, 1994.
- [5] P. Ekman, "Universals and cultural differences in facial expressions of emotion." in *Nebraska symposium on motivation*. University of Nebraska Press, 1971.
- [6] N. Liu, Y. Zong, B. Zhang, L. Liu, J. Chen, G. Zhao, and J. Zhu, "Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5144–5148.
- [7] C.-M. Chang and C.-C. Lee, "Fusion of multiple emotion perspectives: Improving affect recognition through integrating cross-lingual emotion information," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5820–5824.
- [8] C.-M. Chang, B.-H. Su, S.-C. Lin, J.-L. Li, and C.-C. Lee, "A bootstrapped multi-view weighted kernel fusion framework for cross-corpus integration of multimodal emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 2017, pp. 377–382.
- [9] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," *Proc. Interspeech 2017*, pp. 1243–1247, 2017.

- [10] J. Gideon, M. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog)," *IEEE Transactions on Affective Computing*, 2019.
- [11] C.-M. Chang and C.-C. Lee, "Adversarially-enriched acoustic code vector learned from out-of-context affective corpus for robust emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7395–7399.
- [12] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The usc creativeit database: A multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, p. 55, 2010.
- [13] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 865–868.
- [14] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–8.
- [15] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.
- [16] H.-C. Chou, W.-C. Lin, L.-C. Chang, C.-C. Li, H.-P. Ma, and C.-C. Lee, "Ntime: The nthu-ntua chinese interactive multimodal emotion corpus," in *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 2017, pp. 292–298.
- [17] 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.
- [18] K. Feng and T. Chaspari, "A review of generalizable transfer learning in automatic emotion recognition," *Frontiers in Computer Science*, vol. 2, p. 9, 2020.
- [19] A. Hassan, R. Damper, and M. Niranjana, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1458–1468, 2013.
- [20] Y.-Q. Miao, R. Araujo, and M. S. Kamel, "Cross-domain facial expression recognition using supervised kernel mean matching," in *2012 11th International Conference on Machine Learning and Applications*, vol. 2. IEEE, 2012, pp. 326–332.
- [21] T. Kanamori, S. Hido, and M. Sugiyama, "A least-squares approach to direct importance estimation," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1391–1445, 2009.
- [22] Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama, "Direct density ratio estimation for large-scale covariate shift adaptation," *Journal of Information Processing*, vol. 17, pp. 138–155, 2009.
- [23] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu, and Y. Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Communication*, vol. 83, pp. 34–41, 2016.
- [24] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," *Proc. Interspeech 2017*, pp. 1098–1102, 2017.
- [25] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Cross corpus speech emotion classification-an effective transfer learning technique," *arXiv preprint arXiv:1801.06353*, 2018.
- [26] —, "Transfer learning for improving speech emotion classification accuracy," *Proc. Interspeech 2018*, pp. 257–261, 2018.
- [27] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2746–2750.
- [28] X. Zhu, Y. Liu, Z. Qin, and J. Li, "Data augmentation in emotion classification using generative adversarial networks," *arXiv e-prints*, pp. arXiv-1711, 2017.
- [29] S. E. Eskimez, D. Dimitriadis, R. Gmyr, and K. Kumanati, "Gan-based data generation for speech emotion recognition," *Proc. Interspeech 2020*, pp. 3446–3450, 2020.
- [30] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Data augmentation using gans for speech emotion recognition," in *Interspeech*, 2019, pp. 171–175.
- [31] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Augmenting generative adversarial networks for speech emotion recognition," *Proc. Interspeech 2020*, pp. 521–525, 2020.
- [32] 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.
- [33] S. Sahu, R. Gupta, and C. Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," *Proc. Interspeech 2018*, pp. 3693–3697, 2018.
- [34] G.-Y. Chao, Y.-S. Lin, C.-M. Chang, and C.-C. Lee, "Enforcing Semantic Consistency for Cross Corpus Valence Regression from Speech Using Adversarial Discrepancy Learning," in *Proc. Interspeech 2019*, 2019, pp. 1681–1685. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2019-2037>
- [35] C.-M. Chang, G.-Y. Chao, and C.-C. Lee, "Enforcing semantic consistency for cross corpus emotion prediction using adversarial discrepancy learning in emotion," *IEEE Transactions on Affective Computing*, 2021.
- [36] W. Zheng, W. Zheng, and Y. Zong, "Multi-scale discrepancy adversarial network for crosscorpus speech emotion recognition," *Virtual Reality & Intelligent Hardware*, vol. 3, no. 1, pp. 65–75, 2021.
- [37] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *2013 humaine association conference on affective computing and intelligent interaction*. IEEE, 2013, pp. 511–516.
- [38] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 31–43, 2018.
- [39] C. Fu, J. Shi, C. Liu, C. T. Ishi, and H. Ishiguro, "Aaec: An adversarial autoencoder-based classifier for audio emotion recognition," ser. *MuSe'20*. New York, NY, USA: Association for Computing Machinery, 2020, p. 45–51.
- [40] 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*, 2020.
- [41] C. Fu, C. Liu, C. T. Ishi, and H. Ishiguro, "Maec: Multi-instance learning with an adversarial auto-encoder-based classifier for speech emotion recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6299–6303.
- [42] Y. Gao, J. Liu, L. Wang, and J. Dang, "Domain-adversarial autoencoder with attention based feature level fusion for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6314–6318.
- [43] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.
- [44] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv e-prints*, pp. arXiv-1511, 2015.



Chun-Min Chang is a doctoral student at the Department of electrical engineering from National Tsing Hua University (NTHU), Taiwan. He received the B.S. degree from the Department of electrical engineering, National Tsing Hua University (NTHU), Hsinchu, Taiwan in 2015. His research interests include affective computing, machine Learning, human-centered behavioral modeling and infant behavior analysis. He was the recipient of NTHU President's Scholarship, NOVATEK Scholarship and Elite-well Scholarship. He is a student member of the IEEE Signal Processing Society.



Chi-Chun Lee (M'13, S'20) is an associate professor at the Department of Electrical Engineering with joint appointment at the Institute of Communication Engineering of the National Tsing Hua University (NTHU), Taiwan. He received his B.S. and Ph.D. degree both in Electrical Engineering from the University of Southern California, USA in 2007 and 2012. His research interests are in speech and language, affective computing, and health analytics. He is an associate editor for the IEEE Transaction on Affective

Computing (2020-), the IEEE Transaction on Multimedia (2019-2020), and a TPC member for APSIPA IVM and MLDA committee. He serves as an area chair for INTERSPEECH 2016, 2018, 2019, senior program committee for ACII 2017, 2019, publicity chair for ACM ICMI 2018, sponsorship and special session chair for ISCSLP 2018, 2020, and a guest editor in Journal of Computer Speech and Language on special issue of Speech and Language Processing for Behavioral and Mental Health.

He is the recipient of the Foundation of Outstanding Scholar's Young Innovator Award (2020), the CIEE Outstanding Young Electrical Engineer Award (2020), the IICM K. T. Li Young Researcher Award (2020), the MOST Futuretek Breakthrough Award (2018, 2019). He led a team to the 1st place in Emotion Challenge in INTERSPEECH 2009, and with his students won the 1st place in Styrian Dialect and Baby Sound subchallenge in INTERSPEECH 2019. He is a coauthor on the best paper award/finalist in INTERSPEECH 2008, INTERSPEECH 2010, IEEE EMBC 2018, INTERSPEECH 2018, IEEE EMBC 2019, APSIPA ASC 2019, IEEE EMBC 2020, and the most cited paper published in 2013 in Journal of Speech Communication. He is an IEEE senior member and a ACM and ISCA member.