This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3111110, IEEE Transactions on Affective Computing



# Enforcing Semantic Consistency for Cross Corpus Emotion Prediction using Adversarial Discrepancy Learning in Emotion

Chun-Min Chang, *Student Member, IEEE,* Gao-Yi Chao, *Student Member, IEEE,* and Chi-Chun Lee, *Senior Member, IEEE*

**Abstract**—Mismatch between databases entails a challenge in performing emotion recognition on a practical-condition unlabeled database with labeled source data. The alignment between the source and target is crucial for conventional neural network; therefore, many studies have mapped two domains in a common feature space. However, the effect of distortion in emotion semantics across different conditions has been neglected in such work, and a sample from the target may be considered a high emotional annotation in the target but as low in the source. In this work, we propose the maximum regression discrepancy (MRD) network, which enforces semantic consistency in a source and target by adjusting the acoustic feature encoder to minimize discrepancy in maximally distorted samples through adversarial training. We show our framework in several experiments using three databases (the USC IEMOCAP, MSP-Improv, and MSP-Podcast) for cross corpus emotion prediction. Compared to the Source-only neural network and DANN, MRD network demonstrates a significant improvement between 5% and 10% in the concordance correlation coefficient (CCC) in cross-corpus prediction and between 3% and 10% for evaluation on MSP-PODCAST. We also visualize the effect of MRD on feature representation to shows the efficacy of the MRD structure we designed.

**Index Terms**—speech emotion recognition, generative adversarial network, cross corpus learning, semantic consistency, domain adaptation

## 1 Introduction

Affective computing is a cognitive psychological process linking the innate neuro-physiological process to the distribution of human traits intertwined with motivation, thought, personality, and temperament through computing. The perception of the behavior of an individual is further affected by idiosyncratic factors, such as personal temperament, mood, and motivation are modeled by commercial applications (e.g., natural human-computer interface [1], health care [2], and marketing [3]). The easily derived signals from human extrinsic behavior include facial landmarks, action units, and physiological signals [4].

Speech, which continues to be the most information-rich and accessible message exchange medium for humans, is employed for this kind of research. Compared with other modalities, speech is the most accessible and has the least cost for large collections. Speech emotion recognition (SER) is the conventional application of emotion recognition. It does not require bodily contact or expensive equipment (microphone only); hence, it is more convenient to detect emotion. Due to powerful techniques, deep learning algorithms have emerged. People calculate more complicated problems regarding emotions with more data-driven learning methodologies. This was not possible in the past because of the poor accessibility of emotion data and poor computational power.

- *Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan*
- *MOST Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan*
  *E-mail: cclee@ee.nthu.edu.tw*

Practical conditions on SER are difficult. The conventional approach for SER is context-dependent due to the same distribution between the source and target, leading to a model that performs well, yet is too expensive for collecting and annotating a sufficient amount of emotion data. In contrast, it is common that targets are collected in different context. This may be completely different from the specific environment of data behind the model, which could cause a decrease in the prediction accuracy due to emotion data as a source collected in a universal context. This may be completely different from the specific environment of data behind the model, which could cause a decrease in the prediction accuracy due to emotion data as a source collected in a universal context, which is set up with fixed gender and interaction.

The constraint of data-driven techniques is the phenomenon known as dataset bias or domain shift [5]. This issue of non-robustness is especially evident when learning to perform cross-corpus emotion recognition. Speech contains major variability in emotional acoustic manifestations, which are affected by context, such as the gender information, language [6], recorded environment [7], application domain [8], interaction type [9], and so on. Most real-life emotion corpora for the emotional applications are often highly contextualized. They result in a large mismatch in practical conditions between testing data (target domain) and training data (source domain).

Therefore, domain-adaptation methods have been proposed for solving this situation. Several studies have been conducted on this topic. Chang et al. trained an adversarially-enriched acoustic code vector with a universal

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3111110, IEEE Transactions on Affective Computing



context database considering the emotion information in the specific dataset and helping the prediction of the in-context database[ [10], [11]] using the same distribution with training and testing data, using other information from another database in it afterward. It helps to improve the robustness of emotion prediction but costs time. In contrast, domain adaptation involves training on more universal emotion data, testing on distinct data with a specific context, and compensating for the degradation in SER performance by transferring the related information from the labeled data source domain to the unlabeled target domain. There has been a long struggle to find a way to remedy the accuracy decline in SER performance. Schuller et al. trained cross-corpus emotion speech recognition based on speaker-dependent feature normalization methods [12]. He attempted to align marginal distributions between target and source data via a few kinds of normalization. Maximum mean discrepancy (MMD) optimization, proposed by Borgwardt et al. [13], is another method to align the cluster between the source and target. Song et al. used maximum mean discrepancy in the optimization procedure of non-negative matrix factorization to address the SER domain adaptation problem [14].

Further, Sun et al. proposed deep CORAL to learn a nonlinear transformation that aligns the correlations of layer activation in deep neural networks (DNNs) [15]. This is an effective way to align target data with source data. However, deliberately learning an indifferentiable common feature space between the source and target data could mitigate domain-specific idiosyncratic factors when performing source to target emotion recognition. Deng et al. introduced an adaptive denoising auto-encoder based on an unsupervised domain adaptation method, where prior knowledge learned from a target set is used to regularize the training on a source set [16]. Zong et al. used a domain-adaptive least-squares regression (DaLSR) model to take an additional unlabeled dataset from a target speech corpus serving as an auxiliary dataset and combined it with the labeled training dataset from the source speech corpus [17].

An adversarial learning mechanism has also been used in general domain adaptation. It assumes that, by aligning the target and source emotion data distribution repeatedly, the learned target feature representation can directly be used to transfer the source emotion label to the correct target label. This technique is based on the learning semantic representations for unlabeled target samples by aligning labeled source centroid and pseudo-labeled target centroid[ [18], [19], [20]]. For example, 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 [21]. Laradji et al. extended this idea by adding triplet loss and metric learning to improve the state-of-the-art unsupervised adaptation results for a vision task [22]. However, mapping the target and source data into an indifferentiable common space does not enforce any emotion semantic consistency (i.e., source features of high valence data may be mapped to target features of low valence data). In the task of cross-corpus emotion recognition, this semantic distortion is especially apparent in the valence attribute, as demonstrated in the previous experimental results.

Therefore, our goal is to mitigate this particular issue of emotional semantic distortion in speech. Semantic distortion stems from misjudgment samples as a singularity, which are similar to the specific source samples yet have a distinct annotation from the specific source samples. Both the emotion information and semantic distortion from the source to target should be learned in the network to increase the accuracy and recognize the singularity from the target.

Self-supervised learning is a good way to solve these problems. Self-supervised learning has been researched for several years in domain adaptation, aiming to allow the model to be aware of the data differences and to overcome them [ [23], [24]]. Sun et al. indicated that self-supervised auxiliary tasks are effective in reducing domain shifts [25]. Saito et al. proposed an entropy minimization loss to encourage neighborhood clustering in the target domain with self-supervision [26]. Moreover, it is also important to derive the right distribution of the source within the duration of self-awareness of the discrepancy of the singularity. Saito et al. showed that the severity of the distortion can be estimated using the quantified target discrepancy and incorporating this discrepancy in the procedure of learning the domain-indifferentiable feature space [27].

In this paper, we propose a self-learning mechanism called the maximum regression discrepancy (MRD) network. Two regressors as an alternative view are employed to discriminate a distorted sample and verify the distortion from the original prediction in this network. These two regressors act like two reviewers, giving a score for each sample from the source and target. In emotion information prediction, two reviewers check their answers with the ground truth for the source sample as supervised learning. If self-supervision mechanism finds the gap between these two reviewers to be a discrepancy, we must minimize it for the target samples. Learning the gap between these two reviewers as a discrepancy helps the model react when facing a distorted target sample.

We have proposed this framework in our previous work [28]. Our MRD network enforces semantic consistency when learning the common acoustic feature space with an adversarial discrepancy mechanism (i.e., minimizing the maximum cross-corpus discrepancy). This work extends beyond that work, proposing an MRD network to perform regression from speech by contributing in the following ways:

1) This paper explores the usability of the MRD network structure within three databases: the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [29], Multimodal Signal Processing (MSP-Improv) [30], and MSP-Podcast [31].
2) We compare the MRD network using the cross-corpus discrepancy for the results of activation, valence, and dominance.
3) We compare the MRD network with the DANN (unsupervised domain adaptation by back propagation) in similar database conditions.
4) We expand the structure of the MRD system with several combinations of layers and regressors. We then show the result in the largest data target.
5) We graph histogram on the prediction of the MRD network to examine the effectiveness on bipolar annotations and to observe the projection of the feature rep-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3111110, IEEE Transactions on Affective Computing



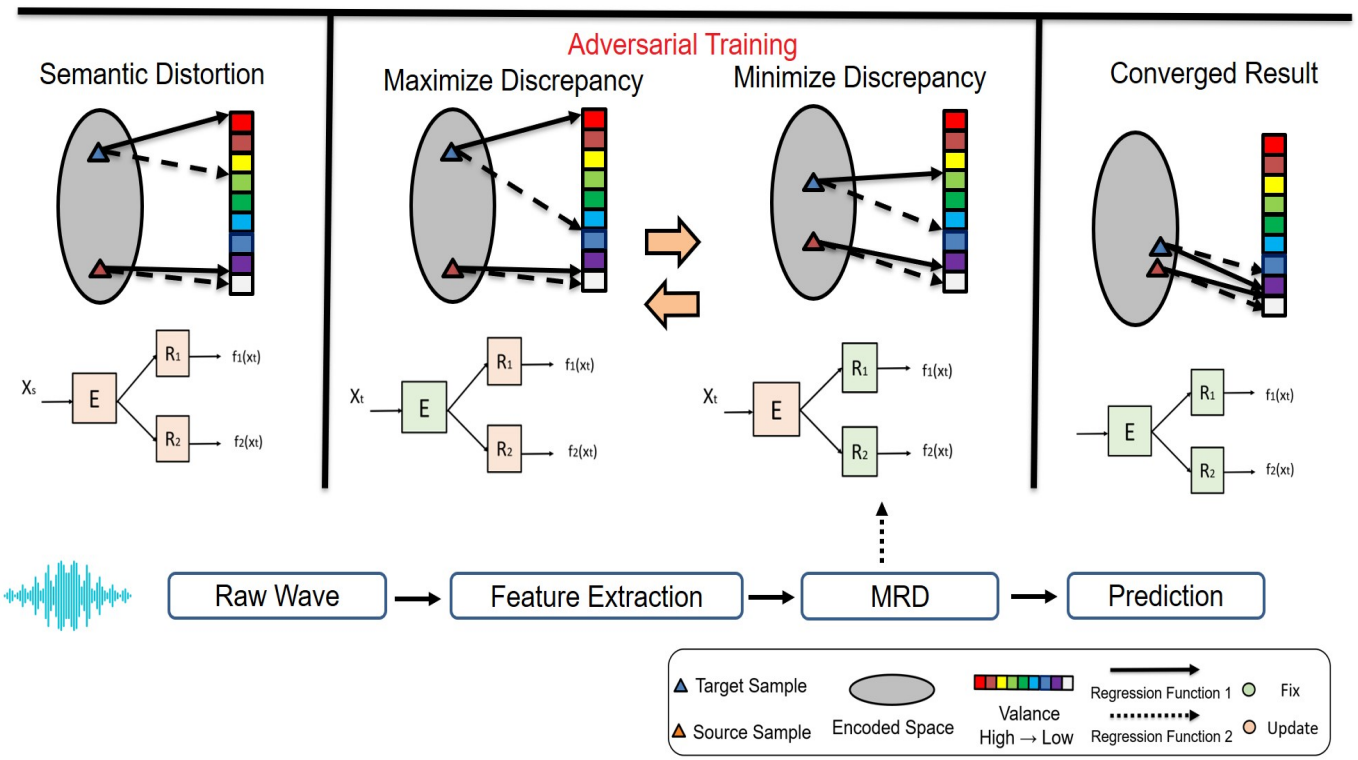


Fig. 1: Adversarial discrepancy learning procedure of MRD network. Step 1 to train the entire network (Encoder and both regressors) to consider semantic distortion on the net. Step 2 to maximize the discrepancy by regressors to shape the highly-distorted representation. Step 3 to minimize the discrepancy by adjusting common space to reduce semantic distortion within encoder. Step 4 to feed target data to evaluate MRD network

.

resentation and examine the semantic distortion with the plots.

6) We plot the feature representation with t-SNE algorithm to compare the domain adaptation from MRD network, DANN, and the vanilla adaptation methods.

In our results , the MRD network demonstrates its superior emotion regression results for these two methods. With cross-corpus validation in different specific conditions, the MRD can achieve 53%, 24%, and 36%, respectively, while the SoNN achieves 39%, 17%, 34%, respectively in terms of activation, valence, and dominance in the CCC for IEMOCAP. With distinct context validation in larger data size conditions, the MRD network can achieve 54%, 27%, and 55%, respectively in terms of activation, valence, and dominance. The rest of the paper is organized as follows. Section 2 describes the database and our MRD network. Section 3 presents the experimental setup. Section 4 describes the experiment and results. Finally, Section 5 concludes with future work.

## 2 Research Methodology

### 2.1 Emotion Databases

#### 2.1.1 Interactive Emotional Motion Capture (IEMOCAP)

The University of Southern California IEMOCAP database is an audiovisual English database. The database consists of five dyadic sessions with a total of 10 actors (five males and five females) grouped in pairs to engage in dyadic face-to-face interactions. In each session, these actors were requested to perform both scripted and spontaneous dialog interactions, collected by motion capture and audio/video data synchronously.

The design of the dyadic interactions was developed by experts to elicit natural multimodal emotion displayed by the actors. Approximately 12 hours of data are segmented into utterances (a total 10039 utterances), and annotators annotated the utterances of the actors using both categorical emotion labels (e.g., angry, happy, sad, neural, etc.) and dimensional representations (e.g., activation, valence, and dominance) on a scale of 1 to 5. The categorical labels per utterance were annotated by at least three raters, and the dimensional attributes were annotated by at least two raters.

Given the spontaneous nature of this database and the inter-evaluator agreement of around 0.4, this database remains a challenging emotion database for algorithmic advancement. For our work, we used activation, valence, and dominance labels for prediction, and each one contains an activation and valence label for scale mapping in ranges between -3 and 3.

#### 2.1.2 Multimodal Signal Processing-Improv (MSP-Improv)

There are over nine hours of audiovisual emotion recordings in the MSP-Improv database. It has six predefined dyadic scenarios. These scenarios were designed to improvise the actors emotions and perform their natural emotional behaviors and to control the emotional content of the lexicon. In addition, every scenario has a target sentence, where one of the dyadic actors performs his or her behavior to allow the main actor to speak the sentences contextualized in happy,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3111110, IEEE Transactions on Affective Computing



TABLE 1: *Performance of the SoNN, DANN and proposed MRD network on two different train and test mismatch condition.[IEMOCAP: USC-IEMOCAP corpus, MSP-improv: MSP-improv corpus] Noted: std : Standard deviation, PR: Pearson's correlation, CCC: concordance correlation coefficient, $*$ : significantly better than SoNN ($p < 0.05$), $\bullet$ : significantly better than DANN ($p < 0.05$).*

| $src$ | MSP-Improv | **Activation** | | | | **Valence** | | | | **Dominance** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $tgt$ | IEMOCAP | PR | $std$ | CCC | $std$ | PR | $std$ | CCC | $std$ | PR | $std$ | CCC | $std$ |
| | ToT | **0.61** | .014 | 0.62 | .015 | 0.44 | .007 | 0.40 | .007 | 0.46 | .014 | 0.43 | .010 |
| | MRD | $\mathbf{0.61}^{*\bullet}$ | **.006** | $0.53^{*\bullet}$ | .005 | $0.25^{*\bullet}$ | .004 | $\mathbf{0.24}^{*\bullet}$ | .005 | $\mathbf{0.44}^{*\bullet}$ | .019 | $\mathbf{0.36}^{*}$ | .016 |
| | DANN | 0.59 | .005 | **0.56** | .007 | **0.27** | .008 | 0.21 | .003 | 0.38 | .013 | 0.35 | .011 |
| | SoNN | 0.55 | .011 | 0.39 | .013 | 0.23 | .003 | 0.17 | .002 | 0.37 | .013 | 0.34 | .014 |
| $src$ | IEMOCAP | **Activation** | | | | **Valence** | | | | **Dominance** | | | |
| $tgt$ | MSP-Improv | PR | $std$ | CCC | $std$ | PR | $std$ | CCC | $std$ | PR | $std$ | CCC | $std$ |
| | ToT | 0.65 | .006 | 0.63 | .009 | 0.43 | .008 | 0.39 | .007 | 0.50 | .007 | 0.46 | .009 |
| | MRD | $\mathbf{0.62}^{*\bullet}$ | .013 | $\mathbf{0.54}^{*\bullet}$ | .019 | $\mathbf{0.30}^{*\bullet}$ | .015 | $\mathbf{0.29}^{*\bullet}$ | .016 | $\mathbf{0.41}^{*\bullet}$ | .005 | $\mathbf{0.36}^{*\bullet}$ | .006 |
| | DANN | 0.59 | .006 | 0.38 | .016 | 0.22 | .008 | 0.21 | .010 | 0.38 | .015 | 0.34 | .015 |
| | SoNN | 0.60 | .012 | 0.43 | .025 | 0.19 | .009 | 0.18 | .010 | 0.37 | .021 | 0.34 | .025 |

angry, sad, and neutral contexts. The approach allows the actor to express emotions as dictated by the scenarios, avoiding prototypical reactions that are characteristic of other acted emotional corpora. In addition, the database is segmented by speaking turns and utterances.

The database contains not only target sentences but also sentences during the improvisations and the natural interactions between actors during breaks. The MSP-Improv was segmented into the utterance level (8386 utterances in total). Each sentence is annotated with activation, valence, dominance, and emotional categories by at least five annotators and was conducted with a crowd-sourced evaluation scheme [32]. The consensus label assigned to each speech turn is the average value of the scores provided by the collector of MSP-Improv.

We chose utterances with dimensional annotation attributes (activation, valence, and dominance), total 8386 utterances. Each dimensional annotation is annotated with an integer from one to five. In this paper, we map the dimensional attribute annotation from -3 to 3.

#### *2.1.3 MSP-PODCAST*

The MSP-Podcast is an extensive speech collection database. It contains 33262 recordings with multiple speakers from audio-sharing websites that are licensed as Creative Commons. It contains different conditions for large numbers of speakers performing spontaneous conversations expressing emotional behaviors. Noise, music, and overlapped speech are not involved.

The same situation occurs with MSP-Improv. The recordings are segmented into speaking turns and last between 2.75s and 11s. The candidate segments were annotated with emotional labels using an improved version of the crowdsourcing framework proposed by Burmania et al [33]. Each sentence is annotated with activation, valence, dominance, and emotional categories by at least five annotators on a scale of 1 to 7.

We used activation, valence, and dominance labels for prediction, and each one contains the activation and valence label in scale mapping in the range of between -3 and 3. All recordings were divided into three parts training, validating, and testing sets with 19707, 4300, and 9255 data, respectively, which are denoted as $POD_{initial}$. At the same time, we use different database settings from the paper in which Abdelwahab proposed DANN [21], that is, 8084, 1844, and 4201 labeled sentences for training, validating, and testing sets, respectively, which are denoted as $POD_{adjusted}$. The $POD_{initial}$ sizes are almost twice that of the $POD_{adjusted}$.

### 2.2 Acoustic Features

The OpenSMILE toolkit is employed in feature extraction [34]. We used the INTERSPEECH 2010 Computational Paralinguistics Challenge (ComParE) feature set in this work, which consists of spectral, prosody, energy, and voicing-related low-level descriptors (LLDs) that are further processed by computing various statistical functionals (a total dimension of 1582) [35]. First, the toolkit extracts 38 frame-level descriptors (i.e., Mel-frequency cepstral coefficients (MFCCs), pitch, jitters, shimmer, etc.), smoothing these descriptors with a low-passing window. Second, all the LLDs are calculated as high-level descriptors with mean, standard deviation , skewness, kurtosis, and so on. A more detailed description can be found in the previous work. We also separately z-normalized this feature set for each corpus to eliminate the feature value gap between corpora.

### 2.3 Maximum Regression Discrepancy Network

In this paper, we further discuss the MRD network. Fig. 1 illustrates our entire framework and the adversarial discrepancy learning procedure. The procedure is divided into three steps of the framework: the encoding, discrepancy maximum, and discrepancy minimum steps. The labeled data from the source domain data, denoted as $(X_s, Y_s)$, and the unlabeled data from the target domain, denoted as $(X_t, Y_t)$, are employed. The training of the MRD network requires an encoder, $E$, and two regressors, $R_1$ and $R_2$.

Step 1 encoding: Encoder $E$ is employed to train the embedding feature for $X_s$. Two regressors, $R_1$ and $R_2$, train two predictors to regress the emotional label separately. The encoder and two regressors derive semantic information from the source samples. Both are trained well with the labeled source samples, and the loss function used in this step is the mean squared error (MSE) loss defined below:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3111110, IEEE Transactions on Affective Computing



TABLE 2: *Performance of the SoNN, DANN and proposed MRD network on testing set of MSP-Podcast database in two different setting $POD_{initial}$ and $POD_{adjusted}$ using two database being source: [IEMOCAP: USC-IEMOCAP corpus, MSP-Improv: MSP-Improv corpus] Noted: std : Standard deviation, PR: Pearson's correlation, CCC: concordance correlation coefficient, $*$ : significantly better than SoNN ($p < 0.05$), $\bullet$ : significantly better than DANN ($p < 0.05$).*

| | **Activation** | | | | **Valence** | | | | **Dominance** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $POD_{adjusted}$ | | $POD_{initial}$ | | $POD_{adjusted}$ | | $POD_{initial}$ | | $POD_{adjusted}$ | | $POD_{initial}$ | |
| | PR | CCC | PR | CCC | PR | CCC | PR | CCC | PR | CCC | PR | CCC |
| ToT | 0.76 | 0.72 | 0.73 | 0.72 | 0.31 | 0.16 | 0.31 | 0.25 | 0.70 | 0.65 | 0.65 | 0.64 |
| MSP-Improv | | | | | | | | | | | | |
| MRD | $\mathbf{0.57}^{*\bullet}$ | $\mathbf{0.51}^{*\bullet}$ | $\mathbf{0.54}^{*\bullet}$ | $\mathbf{0.40}^{*\bullet}$ | $\mathbf{0.15}^{\bullet}$ | $\mathbf{0.16}^{*\bullet}$ | $\mathbf{0.27}^{*\bullet}$ | $\mathbf{0.25}^{*\bullet}$ | $\mathbf{0.61}^{*\bullet}$ | $\mathbf{0.53}^{*\bullet}$ | $\mathbf{0.55}^{*\bullet}$ | $\mathbf{0.45}^{*\bullet}$ |
| DANN | 0.55 | 0.39 | 0.53 | 0.38 | 0.13 | 0.09 | 0.21 | 0.24 | 0.58 | 0.38 | 0.53 | 0.35 |
| SoNN | 0.54 | 0.30 | 0.50 | 0.30 | 0.13 | 0.09 | 0.22 | 0.21 | 0.58 | 0.34 | 0.48 | 0.36 |
| IEMOCAP | | | | | | | | | | | | |
| MRD | $\mathbf{0.55}^{*\bullet}$ | $\mathbf{0.54}^{*\bullet}$ | $\mathbf{0.53}^{\bullet}$ | $\mathbf{0.49}^{*\bullet}$ | $\mathbf{0.23}^{\bullet}$ | $\mathbf{0.22}^{*\bullet}$ | $\mathbf{0.29}^{*\bullet}$ | $\mathbf{0.29}^{*\bullet}$ | $\mathbf{0.52}^{*\bullet}$ | $\mathbf{0.46}^{*\bullet}$ | $\mathbf{0.43}^{*\bullet}$ | $\mathbf{0.36}^{*\bullet}$ |
| DANN | 0.53 | 0.42 | 0.52 | 0.48 | 0.18 | 0.17 | 0.24 | 0.22 | 0.49 | 0.42 | 0.38 | 0.35 |
| SoNN | 0.52 | 0.40 | 0.50 | 0.44 | 0.22 | 0.17 | 0.28 | 0.21 | 0.47 | 0.37 | 0.34 | 0.30 |

$$\min_{E,R_1,R_2} L_{mse}(X_s, Y_s)$$

Step 2 discrepancy maximum: The domain shift occurs when the target sample is input into the model, degrading the performance due to semantic distortion. Hidden distorted representations must be detected, preventing the two regressors from converging to the same output. The discrepancy distortion is estimated using the inconsistency loss [36] derived from the output of the two regressors, $f_1(x)$ and $f_2(x)$, defined below:

$$L_{dis}(X_t) = \frac{1}{K}\sum_{k=1}^{K} |f_1(x_{tk}) - f_2(x_{tk})|$$

Here, k denotes the number of batches. Note that $R_1$ and $R_2$ are initialized differently (i.e., using a different number of layers to avoid converging to the same output). We updated both regressors, $R_1$ and $R_2$, while fixing the encoder $E$, which helps to detect the hidden distorted representations. Distorted samples should be detected by the inverse discrepancy loss. The regression loss of the source should be involved in the objective function in this step due to maintaining the efficiency in Step 1. The objective function is defined below:

$$\min_{R_1,R_2} L_{mse}(X_s, Y_s) - L_{dis}(X_t)$$

Step 3 discrepancy minimum: To narrow the distance between the target sample and source domain sample, it is necessary to minimize inconsistency loss while fixing the same regressors to ensure the encoded features from encoder E preserve the least distorted semantic information. We updated the encoder, $E$, $m$ times to minimize the discrepancy while fixing the regressors. We trained a good predictor in Step 1 for source data and lowered the distortion sample of the target data when mapping it to the source data distribution. Then, we let the target data be mapped to the source distributionthat is, narrowing the distance between the source domain and target domain in the $E$ encoded space. The objective function is as follows :

$$\min_{E} L_{dis}(X_t)$$

Here, we introduced the hyper parameter, $m$, which balances the procedure of the encoder and regressors in the adversarial learning network. In this work, $m = 3$ in each epoch. This parameter was determined experimentally.

Step 4 testing: After finishing the training of the MRD network, given the test sample $x_t$, the regression value $r_t$ is obtained as follows:

$$r_t = \frac{f_1(x_t) + f_2(x_t)}{2}$$

This framework aims to eliminate the negative effect of domain adaptation by training encoder $E$, which minimizes the maximal semantic distortion from the corpora and two regressors, $R_1$ and $R_2$, trained from the source data, to predict the source data reliably. These two regressor predictions are distorted when predicting the target sample due to the domain shift.

## 3 Experimental Setup

In this work, we set up seven different experiments. Each experiment provides a comparison of the different aspects.

Experiment 1 provides a performance comparison of activation, valence ,and dominance in the unsupervised domain adapted speech regression tasks between the MSP-Improv and the IEMOCAP with several models and Experiment 2 provides a comparison of the domain adaptation to the target sample, MSP-Podcast in two setting ($POD_{adjusted}$ and $POD_{initial}$) using the source sample between MSP-Improv and the IEMOCAP with several models.

Experiment 3, 4 provides the structure analysis of the MRD network with several combinations of layers and regressors, then show the result in the largest data target.

Experiment 5 provides a experiment to observe the effectiveness of the proposed method and to examine whether the annotations of high and low are predicted rightly in the MRD network, and Experiment 6 provides the projection of the feature representation, examines the semantic distortion with the plots.

Experiment 7 provides a visualization on the distribution of the encoded space to analyze the generalization ability of MRD and the comparison from other models on activation, valence, and dominance.

The results of each experiment are reported in terms of Pearsons correlation coefficient (PR) and concordance

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3111110, IEEE Transactions on Affective Computing



TABLE 3: *Performance of MRD network of adjusting the regressor numbers on MSP-Podcast database in two different setting $POD_{initial}$ and $POD_{adjusted}$. Noted: PR: Pearson's correlation, CCC: concordance correlation coefficient*

| **# of Regrssor** | | **Activation** | | | | **Valence** | | | | **Dominance** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $POD_{adjusted}$ | | PR | $std$ | CCC | $std$ | PR | $std$ | CCC | $std$ | PR | $std$ | CCC | $std$ |
| MSP-Improv | 2 | **0.54** | .014 | **0.47** | .017 | 0.23 | .008 | 0.19 | .008 | **0.53** | .014 | **0.45** | .024 |
| | 3 | 0.52 | .014 | 0.44 | .011 | 0.23 | .004 | 0.19 | .004 | 0.52 | .012 | 0.44 | .023 |
| | 4 | 0.50 | .013 | 0.40 | .018 | 0.23 | .008 | **0.20** | .008 | 0.49 | .010 | 0.40 | .016 |
| IEMOCAP | 2 | **0.49** | .012 | **0.45** | .015 | 0.19 | .005 | 0.17 | .007 | 0.35 | .011 | 0.27 | .015 |
| | 3 | 0.48 | .010 | 0.43 | .013 | **0.19** | .007 | **0.17** | .007 | **0.36** | .015 | **0.27** | .016 |
| | 4 | 0.42 | .010 | 0.39 | .010 | 0.16 | .012 | 0.16 | .012 | 0.35 | .010 | 0.25 | .012 |
| **# of Regrssor** | | **Activation** | | | | **Valence** | | | | **Dominance** | | | |
| $POD_{initial}$ | | PR | $std$ | CCC | $std$ | PR | $std$ | CCC | $std$ | PR | $std$ | CCC | $std$ |
| MSP-Improv | 2 | 0.52 | .009 | 0.48 | .018 | 0.11 | .019 | 0.11 | .019 | **0.52** | .013 | **0.45** | .021 |
| | 3 | 0.52 | .009 | 0.47 | .014 | **0.13** | .006 | 0.12 | .015 | 0.51 | .010 | 0.44 | .029 |
| | 4 | 0.52 | .011 | 0.46 | .015 | 0.13 | .011 | **0.13** | .012 | 0.49 | .010 | 0.42 | .038 |
| IEMOCAP | 2 | **0.53** | .011 | **0.53** | .014 | **0.22** | .009 | **0.21** | .009 | **0.40** | .016 | **0.38** | .014 |
| | 3 | 0.52 | .012 | 0.51 | .012 | 0.21 | .017 | 0.21 | .018 | 0.39 | .009 | 0.37 | .009 |
| | 4 | 0.51 | .015 | 0.48 | .015 | 0.19 | .017 | 0.19 | .017 | 0.39 | .015 | 0.35 | .013 |

correlation coefficient (CCC) between the ground truth and estimated values. The variable PR is defined as follows:

$$\rho_{x,y} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y}$$

CCC is defined as :

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2}$$

Moreover, all results are trained 10 times, and we obtained the average of the results. There are a couple of methods to compare our performance. Moreover, we compare the performance of our proposed models with baselines and conduct statistical testing using the Student's t-test. (two-tailed t-test, p-value $< 0.05$)

1) ***Maximum Regression Discrepancy Network***:
For our proposed Maximum Regression Discrepancy (MRD) network, the parameters of the MRD network are listed below. The numbers of layers of the encoder and both regressors are 4, 2, and 1, respectively. All hidden layer widths is [1024, 512, 256, 128], [128,64], [128] respectively. The number of nodes are set to the power of 2 depending on the number of layers following by the network (if there are three layers, it will be [1024, 512, 256]), therefore, the nodes will vary in deep and shallow experiment of Section 4.4. We also used batch normalization, dropout (p = 0.5) and SELU as activation function in all layers in which the function output is normalized. Empirical observation has been shown the use of SELU improving the convergence of adversarial training in several researches [37], [38]. We use Adam as an optimizer to minimize the objective function and the number of epochs and the learning rate are determined according to different tasks. In this study, the maximum number of epochs is 100, and the learning rate ranges from $1e-6$ to $5e-6$.
2) ***Domain adversarial neural network***:
This is an unsupervised domain adaptation method through propagation based on the method proposed by Ganin et al [39]. We compared our results with those of Abdelwahab et al. [21]. The technical structure is similar to his previous work. We conducted three parts of Domain adversarial neural network (DANN): the encoder, task classifier, and domain classifier. The numbers of each layer are 4, 2, and 1, and the hidden layer is [1024, 512, 256, 128], [128, 64], and [128], which is similar architectures with our proposed method. We also employed batch normalization, dropout (p = 0.5) and SELU as activation function in each layer, and the batch size is 128.
3) ***Source-only neural network***:
Source-only neural network (SoNN) is trained only on the source domain and is regressed on the target domain without any direct adaptation. We trained the Source-only neural network (SoNN) in PyTorch. It is divided into an encoder part and a classifier part. The numbers of each layer are 4 and 1, and the hidden layers are set at [1024, 512, 256, 128] and [128], which is similar architectures with our proposed method. In addition, we employed batch normalization and the dropout rate (p = 0.5) in each layer. The activation layer is SELU for all layers. The number of epochs is 100, and the learning rate is $5e-5$.
4) ***Train-on-target***:
Train-on-target (ToT) is the perfect condition for training. The DNN is trained and tested on the target domain. The same technical setting is used in the Source-only neural network.

## 4 Experimental results and analysis

### 4.1 Similar Context Comparison

We explored the effect of the proposed model in different dimensions of emotion. Chao et al. presented the results for valence with their proposed model [28]. In addition, in this session, the performance was expanded to cover activation and dominance.

IEMOCAP and MSP-Improve are two emotion databases with different predefined dyadic interaction context and different actors but in a similar recorded environment. It is a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3111110, IEEE Transactions on Affective Computing



TABLE 4: *Adjustment of shallow difference of two regressors. Performance of MRD adjusting the shallow difference layer on MSP-Podcast database in two different setting $POD_{initial}$ and $POD_{adjusted}$. Noted: PR: Pearson's correlation, CCC: concordance correlation coefficient*

| **difference of Layers** | | **Activation** | | | | **Valence** | | | | **Dominance** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $POD_{adjusted}$ | | $pr$ | $std$ | $ccc$ | $std$ | $pr$ | $std$ | $ccc$ | $std$ | $pr$ | $std$ | $ccc$ | $std$ |
| MSP-imp | 2 | 0.53 | .016 | 0.45 | .017 | 0.24 | .010 | 0.19 | .013 | 0.52 | .017 | 0.46 | .022 |
| | 3 | 0.52 | .020 | 0.43 | .020 | 0.23 | .008 | 0.19 | .010 | 0.52 | .010 | 0.45 | .018 |
| | 4 | 0.50 | .024 | 0.35 | .025 | 0.22 | .010 | 0.20 | .013 | 0.48 | .012 | 0.38 | .016 |
| | 5 | 0.43 | .024 | 0.35 | .025 | 0.22 | .010 | 0.20 | .013 | 0.48 | .012 | 0.38 | .016 |
| | 6 | 0.44 | .018 | 0.32 | .020 | 0.21 | .013 | 0.20 | .012 | 0.44 | .012 | 0.32 | .016 |
| IEMO | 2 | 0.49 | .010 | 0.46 | .015 | 0.19 | .012 | 0.17 | .010 | 0.35 | .014 | 0.27 | .022 |
| | 3 | 0.48 | .013 | 0.43 | .014 | 0.19 | .008 | 0.17 | .010 | 0.35 | .012 | 0.26 | .013 |
| | 4 | 0.46 | .017 | 0.40 | 0.02 | 0.19 | .009 | 0.18 | .007 | 0.35 | .028 | 0.25 | .025 |
| | 5 | 0.45 | .017 | 0.37 | .020 | 0.19 | .012 | 0.17 | .009 | 0.35 | .019 | 0.24 | .021 |
| | 6 | 0.43 | .016 | 0.33 | .018 | 0.20 | .013 | 0.15 | .013 | 0.36 | .013 | 0.24 | .014 |
| **difference of Layers** | | **Activation** | | | | **Valence** | | | | **Dominance** | | | |
| $POD_{initial}$ | | $pr$ | $std$ | $ccc$ | $std$ | $pr$ | $std$ | $ccc$ | $std$ | $pr$ | $std$ | $ccc$ | $std$ |
| MSP-imp | 2 | 0.52 | .014 | 0.47 | .021 | 0.11 | .016 | 0.11 | .016 | 0.52 | .016 | 0.44 | .029 |
| | 3 | 0.52 | .013 | 0.47 | .023 | 0.11 | .017 | 0.11 | .017 | 0.51 | .013 | 0.46 | .020 |
| | 4 | 0.52 | .013 | 0.46 | .037 | 0.11 | .024 | 0.11 | .025 | 0.50 | .010 | 0.44 | .024 |
| | 5 | 0.51 | .015 | 0.43 | .027 | 0.10 | .023 | 0.09 | .022 | 0.48 | .019 | 0.43 | .029 |
| | 6 | 0.50 | .016 | 0.41 | .021 | 0.09 | .024 | 0.08 | .025 | 0.45 | .017 | 0.38 | .037 |
| IEMO | 2 | 0.53 | .011 | 0.52 | .012 | 0.22 | .017 | 0.22 | .018 | 0.40 | .012 | 0.39 | .012 |
| | 3 | 0.53 | .007 | 0.51 | .006 | 0.21 | .020 | 0.20 | .020 | 0.39 | .015 | 0.37 | .013 |
| | 4 | 0.52 | .010 | 0.48 | .008 | 0.20 | .014 | 0.19 | .015 | 0.39 | .013 | 0.34 | .015 |
| | 5 | 0.50 | .009 | 0.44 | .010 | 0.21 | .021 | 0.20 | .020 | 0.40 | .010 | 0.33 | .008 |
| | 6 | 0.49 | .004 | 0.41 | .005 | 0.21 | .023 | 0.18 | .017 | 0.39 | .020 | 0.32 | .018 |

preliminary robustness verification of our proposed method to train and test on each of these two databases alternatively.

Table 1 lists the performance for each condition. The upper condition is the performance for MSP-Improv as the source and IEMOCAP as the target. The lower condition is the performance with IEMOCAP as the source and MSP-Improv as the target. Each result is presented with the PR and CCC with the standard deviation. The first row presents the upper boundary CCC for domain adaptation, which indicates the independent database prediction. We also compare the difference from our proposed model with DANN, SoNN and ToT, asserting their significance with t-test at p-value $< 0.05$. There are several points regarding the upper condition. First, the result from the proposed MRD network can achieve 0.61 PR, which is equal to the upper boundary CCC for activation and is much better than the 0.59 and 0.55 PR derived from the DANN and SoNN, respectively. The same trend occurs for dominance at 0.44 PR, where the PR performance of the MRD network is less than that for Train-on-target method at 0.02 PR (4.5% of relative degradation) and is more than the other networks, increasing by 0.06 PR (13.6% of relative improvements). Second, task on domain adaptation is relatively difficult in terms of valence. The performance of these networks is much lower than the Train-on-target method due to a severe domain mismatch that degrades the performance (i.e., the SoNN achieves 0.17 CCC). Although the PR performance in the MRD is somewhat less than that for DANN at 0.25 PR and the PR performance at 0.27 is, conversely, better. The MRD achieves 0.24 CCC, which increases 41.1% of relative improvements over the SoNN and 14.2% of relative improvement compared to the DANN. In the lower condition, the performance of the MRD is better than that for the train-on-target method and DANN, with 0.02 CCC (3.3% of relative improvement) and 0.11 CCC (57.8% of relative improvement) improvements in terms of activation and valence respectively. The performance of DANN in terms of activation is lower than for the SoNN, yet the performance of the MRD network is better than that of the SoNN . Performance increases by 57.8% in CCC and 37.9% in PR of relative improvement compared to the train-on-target method for both PR and CCC in terms of valence and by 10.8% in CCC and 5.8% in PR of relative improvement in terms of dominance, respectively.

### 4.2 Distinct Context Comparison

The MSP-Podcast database contains different conditions for numerous speakers performing spontaneous conversations. Section 2 mentioned that the MSP-Podcast database is set in two different situations, with $POD_{initial}$ as the original total size of the database and $POD_{adjusted}$ as the similar database setting in the training and testing, as adjusted by Abdelwahab et al. It is crucial to evaluate model capacity with different sizes and contexts of data and to compare the results with other networks; therefore, $POD_{adjusted}$ employed 8084 samples in the training set, 1844 samples in the validating set, and 4201 samples in the testing set. In addition, $POD_{initial}$ employed 19707 samples in the training set, 4300 samples in the validating set, and 9255 samples in the testing set, which is discussed in this chapter.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3111110, IEEE Transactions on Affective Computing



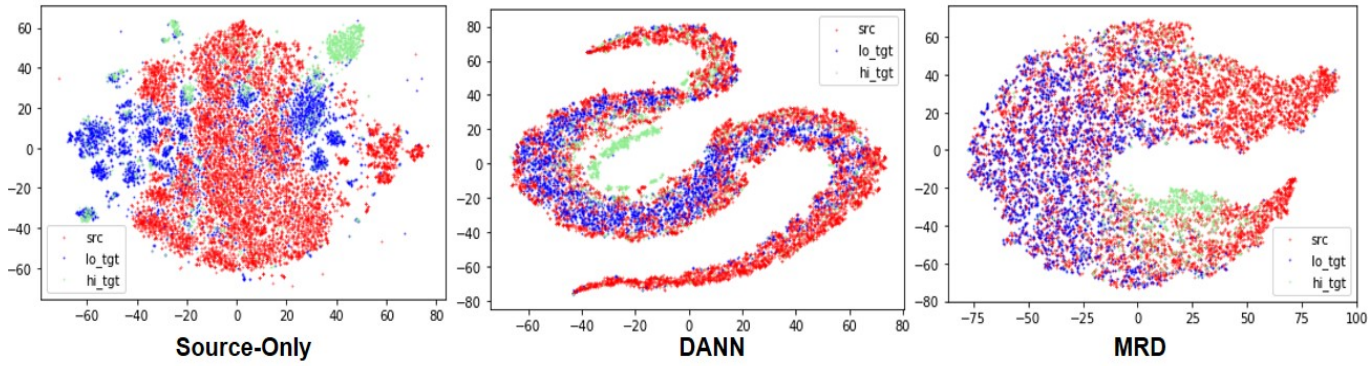


Fig. 2: The t-SNE algorithm are employed to plot feature representation transformed by the encoder from the MRD network, DANN and SoNN for activation. Training on the MSP-Improv dataset and test on the MSP-Podcast corpus. It have been marked in to three class, source, target that is marked low discrepancy distortion in the MRD network, target marked high discrepancy distortion in the MRD network.

The robustness and stability of the proposed method are proven with the performance of both situations.

Table 2 lists the results of two different settings on testing set of the MSP-Podcast, $POD_{adjusted}$ and $POD_{initial}$, with the source datasets IEMOCAP and MSP-Improv. These results were compared to the Train-on-target method where the model is trained and validated with training and validating data and testing in the testing set from the MSP-Podcast corpus (refer to the Train-on-target data in Table 2). The MSP-Improv and IEMOCAP are employed as the source data on these two datasets alternatively. We also assert the significance with t-test at p-value $< 0.05$ to compare the difference from our proposed model with DANN, SoNN and ToT in this section.

First, we assessed the $POD_{adjusted}$. The performance of the MRD is much better than that of the DANN and SoNN for these three kinds of annotations. It achieved a boost of 0.21 CCC and 0.14 CCC from the proposed framework compared to SoNN methods and is much better than the DANN over 0.12 CCC for two distinct source databases on activation. There is a similar tendency concerning dominance. The PR performance of the MRD network is 0.61 PR and 0.52 PR compared to 0.58 PR and 0.47 PR of the SoNN, which increased by 5.1% and 10.6% of relative improvement, respectively. The remaining shows relative improvements around 3.6% improvement from the DANN to the MRD network for the PR in terms of the activation, 15.3% to 27.7% improvement for the CCC from the DANN to the MRD in terms of valence, and 5.1% and 6.1% for the PR improvement from the DANN to the MRD in terms of dominance.

Second, we focus on the performance of the larger database, $POD_{initial}$. When the target database is bigger, the model capability is lower. Therefore, the three different models (MRD, DANN, and SoNN) have a decline in performance for each source of data for activation and dominance. Although the performance declined within these networks, it still significantly increases in relative improvements in terms of activation and dominance, achieving a boost of 33% (0.1 CCC) and 11.3% (0.05 CCC) for activation and 25% (0.09 CCC) and 20% (0.06 CCC) for dominance for the sources of data from MSP-Improv and IEMOCAP. This implies the stability and capability of MRD are better than those of another network. Finally, the performance for valence particularly outperforms the Train-on-target method in CCC whether on $POD_{adjusted}$ or $POD_{initial}$.

The performance from the MRD network is even better than that for the Train-on-target method (0.22 PR vs. 0.16 PR) and on $POD_{adjusted}$ (0.29 PR vs. 0.25 PR) for the source data of IEMOCAP. There may be two reasons. First, valence is difficult to predict from audio data. Second, the MSP-Podcast contains too many complex contexts for training, and it is easier to build a model from the relatively pure context database (MSP-Improv and IEMOCAP).

### 4.3 Difference Regressors

The MRD network aims to eliminate semantic distortion for each annotation from the source and target database. Two regressors were employed in the original structure. We aim to determine whether the number of regressors affects the model performance or the formation of semantic distortion. That is, can the framework be more robust in adjusting the discrepancy of the distribution if we change the regressors of the MRD network. We changed the number of regressors to verify the assumption,

Table 3 summarizes the results of these assumptions for two different settings for validating set of $POD_{adjusted}$ and $POD_{initial}$ as the target data samples. In addition, the training model uses two different source datasets (MSP-Improv and IEMOCAP for each setting condition). First, MRD network with two regressors appear to perform better for all conditions for the two sources in terms of activation. Activation is more common in different emotion databases; therefore, we only need two regressors to ensure the smallest discrepancy. Second, correlation changes because the distribution of the target domain variance becomes increasingly different from the original distribution. The model must be stable for practical situations, no matter how the distribution of the target varies. In addition, the result demonstrates that it is relatively stable for three regressors than for other numbers of regressors due to the low standard variation and its similar performance to other numbers of regressors. However, it is unclear for valence because some of the results indicate that more regressors are better than

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3111110, IEEE Transactions on Affective Computing



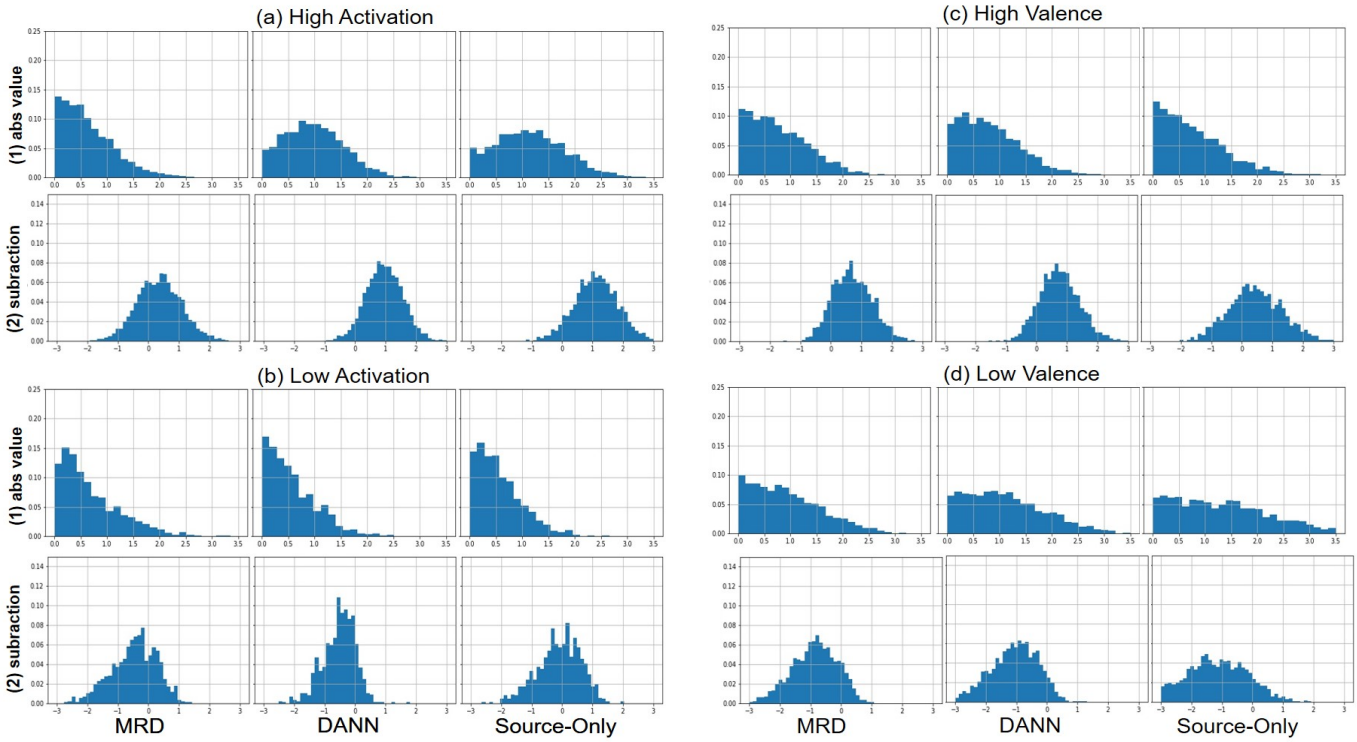


Fig. 3: Ground truth of samples from the MSP-Podcast corpus are divided into bipolar annotation. Each block (a, b, c, d) represents histogram of (1) absolute difference and (2) difference from ground truth and prediction of the MRD network, DANN, and SoNN. X-axis represents the numeric difference of ground truth and predictions. Y-axis shows the normalized data samples with respect to different emotion states.

two. It can be an adjusted number to converge the system of emotions experimentally.

### 4.4 Deep and Shallow Analysis

The MRD network trains different distributions in two different regressors to minimize the semantic distortion in the source and target. One more question is how the difference between these two regressors can eliminate the semantic distortion to improve correlation. We change the layer number of the regressor to confirm this assumption. The first regressor fixes the layer number, which is 1. The other regressor changes its layer number so that the layer difference for these two layers becomes a parameter that we can discuss, the value of which is from 2 to 6. Table 4 summarizes the results of this experiment. These results are evaluated on the validating set of $POD_{adjusted}$ and $POD_{initial}$ datasets which are used to check the adaptation of large and small data sizes, and the MRD network trains on two different source datasets (MSP-Improv and IEMOCAP) for each setting condition.

The correlation decreases when the layer difference increases. The result may lead to the assumption that different distributions are needed to minimize the discrepancy; however, results become worse if the distribution is too distinct to converge with the MRD network. The phenomenon is that the prediction increases when the source of the data is IEMOCAP when the data are larger, whereas the prediction decreases when the source is MSP-Improv when the data are greater but are more stable. It could be that the stability of variance does not vary by the layer difference and that the distribution does not vary that much when there are more data. Based on these results, Classifier with 2 layers toward the other classifier with 1 layers is the better choices which derives good performance and owns relatively stable variance.

### 4.5 Semantic Distortion Analysis

Semantic distortion is defined as the target sample similarity to the source sample with annotation for distinct emotion. The MRD framework has been proposed to detect semantic distortion using discrepancy distortion loss; thus, semantic distortion is examined in this subsection. Fig.2 displays the projection of the feature representation.

Target samples from the MSP-Podcast corpus are calculated for discrepancy distortion loss and are divided into binary categories after training the MRD. Further, the DANN and SoNN are trained well and marked with the class, source, and target with low discrepancy distortion, and with the target with high discrepancy distortion. Target samples with high discrepancy distortion are clustered in a specific area. These samples are hardly handled. As a result, the target samples with high discrepancy distortion are divided from the main distribution in the feature representation for the SoNN. In contrast, the MRD framework is aware of the sample with semantic distortion. Furthermore, the DANN and MRD cluster these samples in the main distribution. Target samples with high discrepancy distortion in DANN are slightly divided from the main contribution, while target samples with high discrepancy distortion in the MRD are packed with all target samples with high discrepancy dis-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3111110, IEEE Transactions on Affective Computing



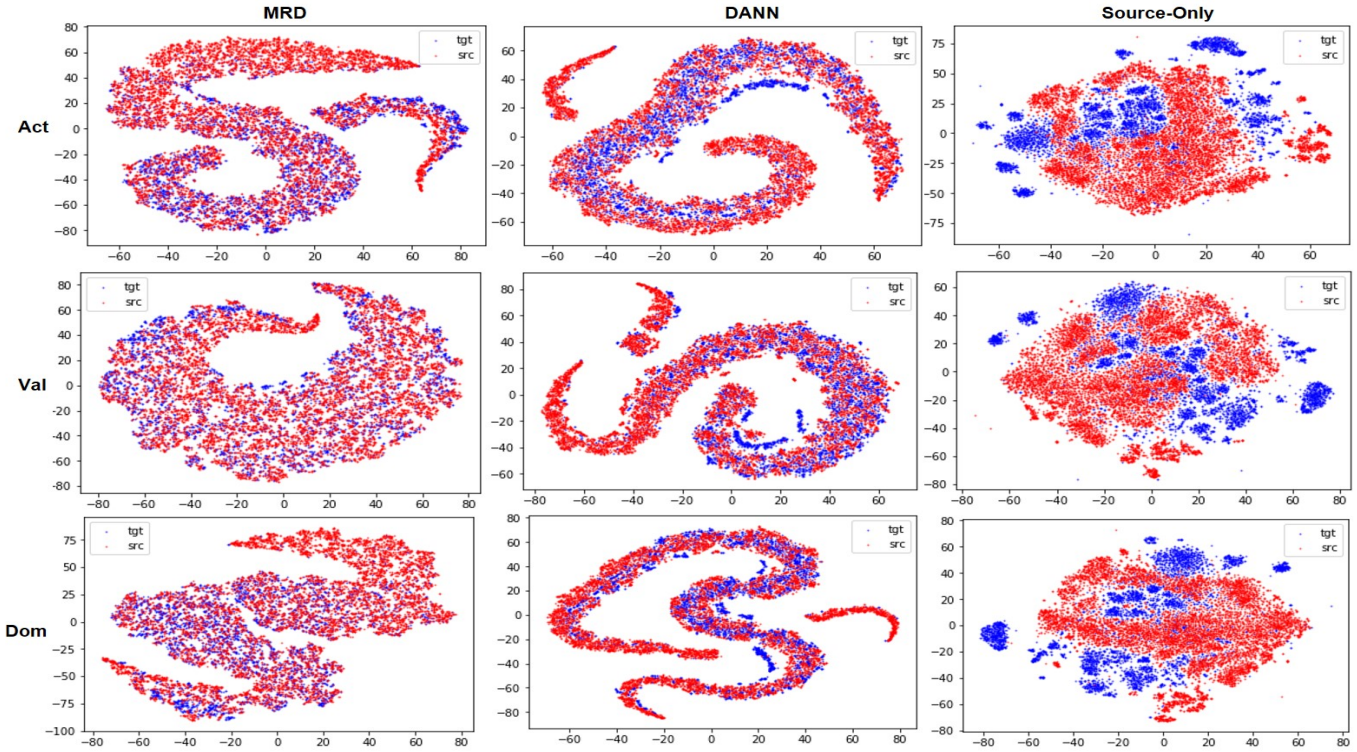


Fig. 4: Feature representation transformed by the encoder from the MRD network, DANN and SoNN for activation, valence and dominance trained on MSP-Improv corpus and tested on MSP-Podcast. These figures are produced by t-SNE algorithm.

tortion in the main distribution. This implies that semantic distortion is learned in the MRD framework.

### 4.6 Prediction Difference Analysis

The MRD network aims to decrease the emotional semantic distortion of the target data by minimizing the maximum discrepancy with the source data. We further discuss which bipolar annotations are decreased by the model. It is easier to examine this issue by dividing the annotations into high and low to visualize the difference in the ground truth from the prediction.

Fig. 3 illustrates the histogram plotted absolute difference and subtraction between prediction and label from the samples of the MSP-Podcast corpus. We subtracted each algorithm predictions from the label of the sample in terms of activation and valence respectively. For (a)-(1), the predictions from the MRD (perfectly match ground truth around 15 %) are closer to ground truth than the predictions from the SoNN and DANN(around 5 %). In (a)-(2), more detail are told. The subtraction distribution are all nearly symmetric. Although high activation are underestimated by all the model(peak of distribution is in the right side of 0), MRD seems to be more accurate compared to other methods due to a closer peak to origin. Besides, for low activation in (b)-(1), the predictions of DANN is slightly better than that of MRD. In (b)-(2), it is observed that low activation may be underestimated by all models(peak of distribution is in the left side of 0), however, center of all the distribution are nearly close. If we take (a)-(1) with (b)-(1), it is realized that MRD's distribution in high activation is closer to the origin as well as MRD's distribution in low activation is nearly the same with other compared methods, leading to better performance of MRD from DANN and SoNN.

(c) and (d) displays the histogram plotted the absolute difference between prediction and ground truth from the samples of the MSP-Podcast corpus in Valence. It seems that SoNN is slightly better than MRD and DANN in (c)-(1). The peak of distribution in SoNN is closer to the origin than others, yet its range is wider in (c)-(2). For (d)-(1), the predictions of MRD(perfectly match ground truth around 10%) are much better than others (around 6%). The peak of MRD and DANN are similar around -1, on the other hand, the peak of SoNN are close to -2. From (c)-(2) and (d)-(2), it is shown that MRD may have the same prediction situation with DANN but also better than Source-only method. However, these three models all result in suboptimal performances in regressing valence which is also shown in session 4.2 (performance in valence is much worse compared to that in activation or dominance).

These three models may all underestimate target sample which make predicted values smaller than the ground truth, e.g., prediction is -2 but ground truth is -3 or prediction is 2 but ground truth is 3. However, it can be observed that the peak of distribution in activation from MRD are more closer to 0 as well as the peak of distribution in valence from MRD are a bit closer to 0 from Figure 1. Via adversarial training on the discrepancy from the distance of target and source, MRD network are forced to align source with target and reduce the domain shift.

### 4.7 Adaptation Visualization

Given the concern regarding the wrong domain adaptation between the source and target data, the t-SNE algorithm was employed to visualize the distribution of these datasets. Fig. 4 illustrates the 2D projection of the feature representation using the t-SNE algorithm for three different algorithms, MRD, DANN, and SoNN in terms of activation, valence, and dominance. There are nine plots in this figure. Each plot is labeled with a specific caption. For activation, the differences are shown between these three algorithms. The source and target can be distinguished from the SoNN. There are still many target sample unions in the DANN figure; nevertheless, it is indistinguishable, and both the target and source are scattered over the whole plot. The domain adaptation aims to let the source and target to be located in the same distribution. The MRD network has the best adaptation for activation. The situation is similar to that for valence. The 2D projection of the DANN with valence is that the target sample focuses on the left side; however, the projection for the MRD network for valence is fully scattered over the whole graph for the target and source. The distribution of the target with the source of the DANN for dominance is somewhat distinguishable. Moreover, comparing three distinct distributions of the MRD network. The distribution from the valence is only one distribution; nevertheless, the target sample for activation and dominance seems to partially be on the lower side of the distribution.

## 5 Conclusion and Future Work

We proposed an innovative framework for emotion recognition that can solve the problem of target and source domain adaptation in the practical emotion recognition context by understanding domain adaptation and semantic consistency. The MRD network uses adversarial training methods, maximizing discrepancy in the distribution of two prediction regressors and minimizing the discrepancy of the regressors from encoders to constrain the target domain. Though the experiment, we demonstrate the cross-corpus results of the MRD first to show this structure has overcome other frameworks in two completely different emotion databases. Then, we reveal the limitations of this model in a more complex emotion database, MSP-Podcast, which contains more contextual emotion information. Whether there are fewer or more data in the podcasts, the MRD with different source domains of emotion databases can outperform the Train-on-target method by a large margin. This demonstrates the stability of the MRD network. Finally, we illustrate the best structure in the MRD network with a few different experiments. We consider the regressor numbers from the results of the same database and consider the deep shallow difference from the results.

In future work, further development is needed to explore the limitations of domain adaptation with the MRD for practical commercial applications. First, we should employ more emotion databases to make a massive emotion base to evaluate the efficiency of domain adaptation, for source domains that should have the maximum variability. Second, the MRD network can be expanded into different modalities from audio. With a visual feature, it may perform even better.

## References


[1] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2015.

[2] M. Chen, Y. Zhang, Y. Li, M. M. Hassan, and A. Alamri, "Aiwac: Affective interaction through wearable computing and cloud technology," *IEEE Wireless Communications*, vol. 22, no. 1, pp. 20–27, 2015.

[3] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita, "An affective guide robot in a shopping mall," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. ACM, 2009, pp. 173–180.

[4] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[5] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching," *Dataset shift in machine learning*, vol. 3, no. 4, p. 5, 2009.

[6] S. M. Feraru, D. Schuller *et al.*, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 125–131.

[7] A. Camurri, I. Lagerlöf, and G. Volpe, "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques," *International journal of human-computer studies*, vol. 59, no. 1-2, pp. 213–225, 2003.

[8] D. L. Roter, R. M. Frankel, J. A. Hall, and D. Sluyter, "The expression of emotion through nonverbal behavior in medical visits," *Journal of general internal medicine*, vol. 21, no. 1, pp. 28–34, 2006.

[9] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[10] 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.

[11] ——, "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] B. Schuller, B. Vlasenko, F. Eyben, M. Wllmer, 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.

[13] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[14] 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.

[15] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*. Springer, 2016, pp. 443–450.

[16] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[17] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," *IEEE signal processing letters*, vol. 23, no. 5, pp. 585–589, 2016.

[18] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *International Conference on Machine Learning*, 2018, pp. 5423–5432.

[19] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[20] Z. Luo, Y. Zou, J. Hoffman, and L. F. Fei-Fei, "Label efficient learning of transferable representations acrosss domains and tasks," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 165–177. [Online]. Available: http://papers.nips.cc/paper/6621-label-efficient-learning-of-transferable-representations-acrosss-domains-and-tasks.pdf

[21] 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.

[22] I. Laradji and R. Babanezhad, "M-adda: Unsupervised domain adaptation with deep metric learning," *arXiv preprint arXiv:1807.02552*, 2018.

[23] R. Moraffah, K. Shu, A. Raglin, and H. Liu, "Deep causal representation learning for unsupervised domain adaptation," *arXiv preprint arXiv:1910.12417*, 2019.

[24] M. Cartwright, J. Cramer, J. Salamon, and J. P. Bello, "Tricycle: Audio representation learning from sensor network data using self-supervision," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 278–282.

[25] Y. Sun, E. Tzeng, T. Darrell, and A. A. Efros, "Unsupervised domain adaptation through self-supervision," *arXiv preprint arXiv:1909.11825*, 2019.

[26] K. Saito, D. Kim, S. Sclaroff, and K. Saenko, "Universal domain adaptation through self supervision," *arXiv preprint arXiv:2002.07953*, 2020.

[27] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3723–3732.

[28] Y.-S. L. Gao-Yi Chao, Chun-Min Chang and C.-C. Lee, "Enforcing semantic consistency for cross corpus valence regression from speech using adversarial discrepancy learning," *Interspeech2019*, 2019.

[29] 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.

[30] 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.

[31] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, 2017.

[32] A. M. Turk, "Amazon mechanical turk," *Retrieved August*, vol. 17, p. 2012, 2012.

[33] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, 2015.

[34] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[35] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[36] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[37] J. D. Curtó, I. C. Zarza, F. De La Torre, I. King, and M. R. Lyu, "High-resolution deep convolutional generative adversarial networks," *arXiv preprint arXiv:1711.06491*, 2017.

[38] Z. Zhang, M. Li, and J. Yu, "On the convergence and mode collapse of gan," in *SIGGRAPH Asia 2018 Technical Briefs*, 2018, pp. 1–4.

[39] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.


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

**Gao-Yi Chao** graduated from the National Taiwan Tsing Hua University, with a Master's degree in Electronic and Electrical Engineering, specialised in speech emotion recognition(SER). His work focuses on the cross-corpus and cross-modality SER. His recent publication can be found in ISCA Interspeech and ACM ICMI conference.

**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 multimedia, health analytics, and behavior computing. 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.