On efficacy of Meta-Learning for Domain Generalization in Speech Emotion Recognition

Raeshak King Gandhi*, Vasileios Tsouvalas† and Nirvana Meratnia‡
Department of Mathematics and Computer Science, Eindhoven University of Technology
Email: *raeshak1000@gmail.com, †v.tsouvalas@tue.nl, ‡n.meratnia@tue.nl

***Abstract*—Speech Emotion Recognition (SER) refers to the recognition of human emotions from natural speech, vital for building human-centered context-aware intelligent systems. Here, domain shift, where models' trained on one domain exhibit performance degradation when exposed to an unseen domain with different statistics, is a major limiting factor in SER applicability, as models have a strong dependence on speakers and languages characteristics used during training. Meta-Learning for Domain Generalization (MLDG) has shown great success in improving models' generalization capacity and alleviate the domain shift problem in the vision domain; yet, its' efficacy on SER remains largely explored. In this work, we propose a "*domain-shift aware*" MLDG approach to learn generalizable models across multiple domains in SER. Based on our extensive evaluation, we identify a number of pitfalls that contribute to poor models' DG ability, and demonstrate that log-mel spectrograms representations lack distinct features required for MLDG in SER. We further explore the use of appropriate features to achieve DG in SER as to provide insides to future research directions for DG in SER.**

***Index Terms*—deep learning, meta-learning, speech emotion recognition, domain generalization, domain shift**

I. Introduction

Human emotions are impulses that influence our daily decisions. Speech Emotion Recognition (SER) has recently attracted growing attention to realize human-centered context-aware intelligent systems in many fields, such as customer support call review and analysis [1], mental health surveillance [2], and smart vehicles [3]. The task of SER is to classify human emotional states by analyzing speech utterances. As emotional state are not clearly defined from a linguistics' perspective, researchers have utilized the "*palette theory*" [4] to compose a set of core emotions (i.e., anger, fear, joy, surprise, sadness and disgust), that can be combined to express any emotional state (termed compound emotion).

In the past, classification and regression algorithms were applied to a set of hand-crafted features (e.g. zero-crossing rate) to produce a probability distribution over a predefined set of emotions. With the advent of deep learning, the research focus has shifted towards end-to-end deep learning SER networks, which can now perform both feature extraction and classification. Here, log-mel spectrogram have become the "*de facto*" audio feature due to its ability to compactly capture both temporal and spatial relations [5], [6]. Authors of [7], [8] utilized CNN-based models to determine relations between sequences of input utterances to produce a compressed vector to effectively capture local features and characteristics of the input data. LSTM architectures have also been explored to capture both long-term contextual dependencies and local cues, greatly improving SER across multiple speakers [5]. In [9], an LSTM with an attention mechanism was proposed to focus on specific parts of a spoken utterance that contains vital emotional information. A common limitation of these approaches is strong dependency of the models on the utterances and languages. In particular, the performance of models trained with specific ethnic groups significantly deteriorates when models are exposed to different speakers or languages [10]. This poor generalization performance while performing cross-lingual (training across languages) or cross-corpus (training across different datasets) training is the main limiting factor in SER applicability, which is further intensified by the sparsity of available speech emotion data.

Meta-learning for Domain Generalization (DG) is a learning paradigm that aims to eliminate the well-known performance degradation problem occurred when a model trained in one source domain is applied to a target domain with different statistics ("*domain shift*"). For this purpose, MAML [11] was proposed to improve generalization and to achieve fast adaptation to new domains in a model-agnostic fashion. MLDG [12] adjusted the learning objective of MAML to concentrate on generalization over multiple domains rather than few-shot learning over multiple tasks. Nonetheless, meta-learning for DG approaches have yet to be explored in SER, where the quantity of available data and generalization of proposed methods introduce two crucial challenges significantly affecting SER applicability in most real-life applications [13].

In this regard, we investigate applicability of meta-learning to mitigate the effect of "*domain shift*" introduced by multiple corpora, speakers' characteristics, and distinct environmental settings by utilizing MLDG [12]. Unfortunately, due to a set of pitfalls discussed in Section V, the proposed approach did not yield the expected improvement in model's generalizability performance. Our main contributions are as follows:

- To the best of our knowledge, this work is the first to improve SER across multiple corpora and speaker characteristics by proposing a Domain-shift Aware Meta-Learning for Domain Generalization (`DA-MLDG`).
- Through vast evaluations with diverse public datasets, we demonstrate that `DA-MLDG` is able to improve the models' generalizability in SER; yet, improvement remains inferior to one obtained in the image domain.

- We identify a set of contributing pitfalls and present new research paths for allowing meta-learning approaches for DG in SER to flourish.
- Our open-source framework[1] enables further study on generalizability and related problems in SER.

II. Related Work

Authors of [14] first investigated the use of deep learning in SER by utilizing a Bi-directional Long Short-Term Memory (BLSTM) architecture to capture time-related relationships across log-mel spectrograms of input data. Along this path, CNN-based architectures were also explored to perform SER [7], [8]. It was shown in [15] that pre-training a deep learning model can reduce the influence of small-scale datasets and improve SER performance. Recently, multi-corpora SER has attracted growing attention [10], [16], [17], where SER models performance has shown significant fluctuations due to differences in pronunciations and enunciations across distinct languages. Here, the reported accuracies do not exceed 65% [16], [17]. Authors in [5] reported similar performance deterioration between speaker dependent and independent experiments, using a CNN-LSTM architecture. Recently, a framework for evaluating the generalization capacity of utterance-level SER deep learning models has been proposed in [10], highlighting the fact that using a limited number of datasets and samples can lead to biased evaluation. In [18], [19], an ensemble of multiple trained models over multiple SER tasks was proposed to improve models' generalization capacity. However, this approach has significant computational overhead and suffers from scalability issues.

Recently, numerous meta-learning approaches have been proposed in the vision domain to achieve generalization and fast adaptation, when sparse labeled data exist in a particular task. MAML [11] performs few-shot learning by fine-tuning the source model across multiple tasks through simulating transfer learning across the tasks. While improving the performance of few-shot learning, it also improves generalization. MetaReg [20] adds a regularizer to the loss function. The regularizer is updated regularly using a meta-learning objective to prevent overfitting and to enable DG over a wider domain distribution. MASF [21] uses meta-learning by introducing an episodic training model to simulate domain shifts. This method is then used to minimize errors from two objectives - Global Class Alignment Objective and Local Sample Clustering Objective. Doing this aligns data from similar domains to form clusters independent of the datasets. MLDG [12] is similar to MAML, except that it performs domain generalization by fine-tuning the source model across multiple datasets by simulating domain shift across the source domains.

MLDG in SER context. While meta-learning for improving generalization has been extensively explored in the vision domain, it has only recently attracted attention in SER [22], [23]. In [22] and MetaASR [23] two different variations of MAML were introduced to achieve high emotion classification accuracy over multiple languages. Authors of [22] proposed to relax some original restrictions of MAML by fixing a number of predefined classes along with varying number of classes for every task to guide the model to learn differences between the two classes. Authors of [23] used meta-learning with a Bi-LSTM encoder to extract language independent features. The aforementioned approaches have solely focused on domain adaptation (i.e. fast adaptation to small-scale data for a particular language or speakers) and have not considered improving models' generalization capability. Furthermore, they suffer when the test domain is entirely unavailable during training, which is the case in most real-life applications as domain shifts start after deployment. Utilizing meta-learning to perform domain generalization in SER can overcome such issues and widen SER applicability in real-life applications.

III. Methodology

Problem formulation. We aim to alleviate the "*domain shift*" problem in SER, where models trained on a fixed set of "*domains*" poorly perform when exposed to a new unseen domain. Our aim is to learn cross-corpus invariant features from speech emotion data that originated from distinct "*domains*" (i.e., speakers from diverse ethnic groups). If successful, the amount of available training data from a particular "*domain*" can be substantially decoupled from the predictive power of models. Such separation will mean that SER can be expanded on corpora with little to no label data (e.g., most non-english corpus), which currently is the prominent barrier in improving SER models' generalizability [24].

Formally, under the DG setting, we have a set of "*source domains*", denoted by $\mathcal{S} = \{S_i\}_{i=1}^{D}$ (D is number of "*source domains*"), from which we aim to learn common invariant features to improve the performance on a closely-related (i.e., different statistics) "*target domain*", denoted by $\mathcal{T}$. Here, in all considered "*domains*" (both $\mathcal{S}$ and $\mathcal{T}$), we aim to perform the same SER task E, i.e., to classify speech utterances to C core emotions. Our objective is to learn the model's parameters Θ, which are able to learn the task E across all domains of $\mathcal{S}$, such that it will generalize to the "*target domain*" $\mathcal{T}$. Specifically, the objective function we aim to minimize is:

$$\Theta = \arg\min_{\theta} \sum_{i=1}^{D} \mathcal{L}_{\theta}\left(S_i\right), \tag{1}$$

where $\mathcal{L}(S_i)$ denotes the loss function on domain S_i, and D is the number of domains present in $\mathcal{S}$.

DA-MLDG. We propose a domain-shift Aware MLDG (`DA-MLDG`) approach (see Algorithm 1), which adaptively computes the "*domain shift*" present between the meta-train and meta-test, and further motivates learning cross-domain features to improve models' generalization capacity. Let $p_{\theta}\left(y|x\right)$ be a neural network parameterized by weights θ that predicts softmax outputs $\widehat{y}$ for a given input x. To learn domain-invariant features across all domains in $\mathcal{S}$, at each training iteration, we aim to simultaneously minimize losses from pairs

[1] https://github.com/vtsouval/DA_MLDG

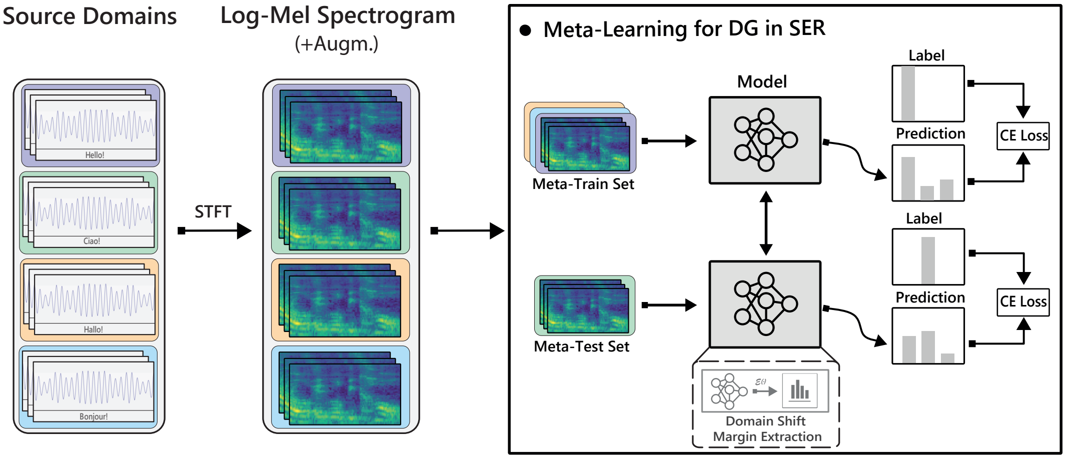

Fig. 1: Domain-shift Aware Meta-Learning for Domain Generalization (`DA-MLDG`) in SER. In each training iteration, we execute the meta-train step, followed by the computation of "*domain-shift margin*" between the meta-train/test set. Then, model parameters are updated in a "*domain-shift aware*" meta-test step.

of non-overlapping meta-train/test subsets of $\mathcal{S}$ (referred to as $\mathcal{S}'$ and $\mathcal{V}$) to resemble the "*domain shift*" between $\mathcal{S}$ and $\mathcal{T}$; thus, achieving model generalization over numerous training iterations. Accordingly, we minimize the following:

$$\begin{aligned} &\min_{\theta} \mathcal{L}_{\theta} = \mathcal{L}_{\theta}\left(\mathcal{S}'\right) + \gamma \mathcal{L}_{\theta'}\left(\mathcal{V}\right), \text{ where} \\ &\mathcal{L}(\cdot) = \frac{1}{D}\sum_{i=1}^{D} \mathcal{L}_{CE}\left(y, p\left(y|x\right)\right) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_i^j \log\left(f_i\left(x_j\right)\right) \end{aligned} \tag{2}$$

Here $\mathcal{S}'$ and $\mathcal{V}$ correspond to the meta-training and meta-test sets, $\mathcal{L}(\cdot)$ denotes the average standard cross-entropy loss ($\mathcal{L}_{CE}$) over number of domains (D) present in each set, and N is number of samples in each domain. One may note that the meta-test loss is minimized with respect to the updated model parameters θ', computed from the gradient ∇_{θ} with respect to the meta-train loss as $\theta' = \theta - \eta \nabla_{\theta}\left(\mathcal{L}\left(\mathcal{S}'\right)\right)$. In essence, we minimize the meta-test loss with respect to the update step introduced from the meta-training set (in the form of the second derivative of θ). It is important to note that subsets $\mathcal{S}'$ and $\mathcal{V}$ are randomly selected from $\mathcal{S}$ in each iteration.

Domain-shift Margin. We weigh the influence of meta-test step in each training iteration by introducing an adaptive "*domain shift*" score, γ, in Equation 3. In particular, γ aims to capture the margin of "*domain shift*" present between the meta-train and meta-test subsets by computing the energy score [25] after the meta-train step is performed. Energy score has proven to be an effective metric for detecting out-of-domain samples, and in this work we propose to use it as a proxy to identify the "*domain shift*" across multiple corpora. Hence, we compute γ in each training step, as follows:

$$\gamma = \left|\mathcal{E}(\mathcal{S}', p_{\theta'}) - \mathcal{E}(\mathcal{V}, p_{\theta'})\right|, \text{ where } \mathcal{E}(\cdot) = \left|\frac{1}{N}\sum_{i=1}^{N} log\left(\sum_{i}^{\mathcal{C}} e^{z/T}\right)\right|, \tag{3}$$

where z is logits produced by the neural network p_{θ} for a $\mathcal{C}$-way classification problem, T is a temperature scalar equal to 1 [25], and N is number of samples present on the considered set. In essence, $\mathcal{E}(\cdot)$ is the average *logsumexp* operator over the logits of the meta-training and meta-test domains.

Algorithm 1 `DA-MLDG`: Domain-shift Aware Meta-Learning for Domain Generalization in Speech Emotion Recognition. Here, parameter $gamma$ indicates the "*domain-shift margin*" between meta-train ($\mathcal{S}'$) and meta-test ($\mathcal{V}$) sets, while η_s and η_t are the learning rate for the meta-train/test steps.

```
Initialization of model with pre-trained parameters $\theta$.
for epoch $e = 1, 2, \ldots, E$ do
    for batch $B \in \mathcal{S}$ do
        Random split of domains: $\mathcal{S}'$ and $\mathcal{V} \leftarrow B$
        Meta-train Step: $\theta' = \theta - \eta_s \nabla_{\theta}\left(\mathcal{L}\left(\mathcal{S}'\right)\right)$
        Cross-Domain Shift Margin: $\gamma = |\mathcal{E}(\mathcal{S}', p_{\theta'}) - \mathcal{E}(\mathcal{V}, p_{\theta'})|$
        Meta-optimization: Update $\theta$
            $\theta = \theta - \eta_t\left(\mathcal{L}_{\theta}\left(\mathcal{S}'\right) + \gamma \cdot \mathcal{L}_{\theta'}\left(\mathcal{V}\right)\right)$
    end for
end for
```

Model pre-training strategy. Pre-training models with massive out-of-domain datasets has become popular as they equip the deep models with a better prior (i.e., initialization weights) to solve downstream tasks [23], [26]. Apart from the performance gain, pre-training on large-scale datasets boosts models' representation power; thus, amplifying model's ability to learn domain-invariant features across a wide-range audio-related tasks [15]. To this end, we pre-train our model architectures using VoxCeleb [27], which consists of over 150,000 samples obtained from 1,251 speakers. We use spectograms as inputs, computed from 2-second raw audio samples using a hamming window of 25ms width and 10ms step, as proposed in [27], while SGD with momentum (0.9) and learning rate of 0.01 was utilized with a batch size of 64 for 50 epochs. In Section IV, when referring to a pre-trained model, we implicitly refer to such a pre-trained model.

IV. Experiments

Datasets. As no public dataset is available for meta-learning-based SER, we have constructed a dataset by considering 10 publicly available SER datasets, namely CREMA-D [28], AESDD [29], CaFE [30], EmoDB [31], RAVDESS [32], TESS [33], SAVEE [34], ShEMO [35], EMOVO [36], and IEMOCAP [37]. We choose a set of five common emotions, namely happiness, sadness, anger, fear and disgust, present across all datasets for our classification tasks. With each dataset corresponding to a separate domain, these datasets were carefully selected to capture a wide range of speakers' characteristics and introduce realistic multi-linguistic (i.e., English, French, German, Greek, Persian, Italian) problems in SER. For all considered datasets, we utilized a sampling rate of 16 KHz and two seconds audio segments during training by randomly cropping the audio sample, similar to [24]. Furthermore, we randomly performed time-stretch and time-shift augmentations by factors sampled randomly from [0.8-1.25] and [-0.1,0.1] respectively; both proven to improve performance in a wide-range of audio-related tasks [38].

Evaluation Strategy. We use the same ResNet18 architecture across all baselines with VoxCeleb [27] weights (following the training strategy presented in Section III) to enable a fair comparison between their performance. For all experiments performed on PACS [39] (image classification with 7 classes), we use the same ResNet18 architecture with pre-trained weights from ImageNet [40], similar to [12]. Here, standard image augmentations (e.g., as random flipping, cropping) were utilized. We train our models with a batch size of 32 using an SGD optimizer with learning rate of 0.05 (η_t = 0.05) and momentum of 0.9, and set η_s to 0.01 based on preliminary experiments. We perform ten distinct trials (i.e., running an entire experiment) in each setting and report the average accuracy over all ten runs.

We evaluate domain generalization performance of DA−MLDG using seven scenarios, as presented in Table I. In Table I, we report model's performance on individual datasets after 50 epochs, where a speaker-independent 80/20 train-test split is considered, similar to [10]. In the first three scenarios we investigate *single language cross-corpus DG* performance by utilizing English-based datasets for both training and test sets. The goal of scenario 4 is *cross-lingual cross-corpus DG* over languages belonging to the same family (Indo-European). Scenario 5 aims at *cross-lingual cross-corpus DG* over different language family (Indo-European and Indo-Iranian). Scenario 6 performs *cross-lingual cross-corpus DG* across two dissimilar languages (English and Persian), with the test domain from one language family and the source domain from another. With scenario 7 we aim to observe if corpora with relatively simple SER task (i.e. high model accuracy) tends to perform better in domain generalization. For scenarios 6 and 7, ShEMO has no samples corresponding to the disgust emotion class; thus, introducing a highly-imbalance extreme scenario.

Baselines. As DG remains mostly unexplored in SER, we perform a thorough evaluation versus a *Deep-All* approach, where standard supervised training is performed on the aggregated data from all "*source*" domains. Standard supervised training has proven to outperforms many DG methods proposed in the vision domain [12]. Furthermore, we perform experiments with vanilla MLDG with γ=1 (denoted as *MLDG*) to investigate effect of γ in model's generalization. From the regularization perspective, we consider Representation Self-Challenging [41] (*RSC*), where DG to out-of-domain data is achieved by discarding dominant features and forcing the network to "*look*" at the remaining feature space.

Table I: Datasets utilized in the six domain generalization scenarios. Here, ○ and ● indicate the train and test domains, respectively.

Scenario	1	2	3	4	5	6	7	Average Accuracy
IEMOCAP	●							58.69 ± 1.22
CREMA-D	○	○	○			○		61.03 ± 1.07
SAVEE	○	○	○	○		○		35.33 ± 7.42
RAVDESS	○	●	○	○		○		71.15 ± 1.97
TESS		○	●				○	99.75 ± 0.04
AESDD					●			68.52 ± 2.98
CaFE				●				49.58 ± 2.76
EMOVO					○			43.21 ± 4.26
ShEMO					○	●	●	79.02 ± 2.06
EmoDB				○	○		○	88.75 ± 5.49

Table II: Performance evaluation of DA−MLDG in image classification and SER tasks. Average accuracy over ten distinct trials on test set is reported.

Dataset	Target Domain	Deep-All	MLDG	DA-MLDG
PACS	Art Painting	73.70	78.01	**78.47**
	Cartoon	75.60	**75.58**	75.44
	Photo	83.50	91.67	**94.73**
	Sketch	64.70	65.22	**66.32**
	Avg.	74.37	77.62	**78.74**
SER (English)	IEMOCAP	26.95	27.61	**31.35**
	RAVDESS	31.70	36.28	**37.75**
	CREMA-D	39.45	40.75	**42.22**
	SAVEE	22.73	**22.84**	21.43
	Avg.	30.20	32.74	**36.29**

Performance Gap vs. Image Classification Task. Since to the best of our knowledge, no meta-learning for DG approach has been explored in SER, we conducted experiments on a set of four widely-used English-based SER datasets (IEMOCAP, RAVDESS, CREMA-D, SAVEE) to determine the feasibility of performing DG in SER. Here, the indicated dataset corresponds to the "*target domain*", whereas the remaining three datasets are used for training. Furthermore, we perform similar experiments on PACS [39] to study the performance gap, while moving from image to emotion classification. The results are presented in Table II, where we notice a significant performance difference between DA−MLDG and two baselines, *Deep-All* and *MLDG* (5.87% / 19.77% and 1.44% / 10.84% in PACS / SER). Despite the performance improvement, we observe that SER models' performance is deficient across all dataset (with exception of CREMA-D) compared to the image classification counterparts. This behavior indicates the presence of unique, easily identifiable features per class in images (i.e., trunk of an elephant, mane of a horse, etc), which are present across PACS domains. This is evident in *Deep-All* row of Table II, where the DG performance gap between PACS and SER is substantial. Consequently, we notice that distinguishing domain-invariant features in images is relatively straightforward compared to the extremely varied patterns of log-mel spectrograms.

Evaluation in multi-corpora SER. We evaluated DA−MLDG performance across all seven considered scenarios, where we reported the performance in terms of unweighted, weighted (by the number of samples) and

Table III: Performance evaluation of DA−MLDG in multi-corpora SER. Average accuracy over ten distinct trials on test set is reported. Additionally, we report Unweighted Mean (**UM**), Weighted Mean (**WM**) by number of utterances across all "**target domains**", and Geometric Mean (**GM**) across all scenarios.

Scenario	1	2	3	4	5	6	7	**UM**	**WM**	**GM**
Deep-All	26.95	43.40	**47.03**	43.16	60.65	31.54	40.19	41.83	36.70	40.61
RSC	**32.36**	**44.78**	43.77	39.97	56.69	**40.70**	**40.27**	42.65	**39.17**	**42.13**
MLDG	27.61	38.86	46.29	42.57	60.45	37.87	34.51	41.16	36.47	40.11
DA-MLDG	31.35	39.41	46.80	**43.21**	**62.26**	38.79	37.11	**42.69**	38.51	41.81

Table IV: Performance evaluation of DA−MLDG using TRILLsson [42] embeddings as model inputs. Average accuracy over ten distinct trials on test set is reported.

Scenario	1	2	3	4	5	6	7
Deep-All	28.57	**42.61**	**50.16**	**45.09**	61.17	33.16	**41.15**
DA-MLDG	**32.11**	39.82	48.46	43.29	**64.11**	**40.22**	36.97

geometric mean (similar to [10]) in Table III. Here, we can see that DA−MLDG achieve an average improvement across scenarios; however this remains impractical for most applications. Further investigation of the DA−MLDG models predictions across scenarios have shown that emotions with similar arousal and valence values are often misclassified. This is also evident in the overlap between model's embedding in Figure V. In addition, we notice that *RSC*, which performs regularization to learn from hidden features present in model's inputs, can achieve better generalization than DA−MLDG; although the performance remains unsatisfactory. Subsequently, these problems reveal the lack of representation power for every emotions in the utilization of log-mel spectrogram for DG in SER.

Model Inputs Importance. To study the effect of model inputs in the classification performance, we leverage TRILLsson (v3, EfficientNetv2-B3) [42] to extract embeddings from the raw audio signals, which we utilize afterwards to perform SER with the same ResNet18 architecture. Our results across all seven scenarios are shown in Table IV, where we notice general improvements across the scenarios, even in the *Deep-All* baseline. This shows the need to concentrate on the form of input data and the requirement of having proper strategies in feature extraction. It also indicates that log-mel spectrograms may not be sufficient to improve the generalization capacity of models in SER. As these input embeddings were generated from heavily pre-trained model for paralinguistic tasks, embeddings contain domain-invariant features useful for DG in SER over different speakers and languages. We believe such direction may prove fruitful for research directions.

V. Pitfall Discussion

Limitations of log-mel spectrograms. In our experiments, we notice that models tend to over-predict emotions with high arousal values (i.e. fear, happiness and anger). This is in line with [43], where authors showed that emotions with obviously distinguishable features (i.e. high intensity) are often easier to predict. On the contrary, many emotions lack a set of distinct features (i.e. disgust), making them challenging to distinguish. Utilizing log-mel spectrograms, which represent the "*loudness*" (i.e., energy) of a signal over time at various frequencies, as model inputs often leads to misinterpretations between certain emotional classes. Adding the lack of similarity in *vocals* between languages and the unique characteristics of speakers, the distinguishable features between the same emotions across datasets widely differ, which is clearly illustrated in Figure V. Hence, selecting proper emotion classes (without overlapping arousal and valence values) is essential for improving SER, yet it remains infeasible in many practical applications. Our findings showcase the limitations (i.e the lack of proper representation for every emotions) of log-mel spectrograms when used for DG in SER. In that direction, utilizing embeddings from pre-trained deep learning models may help overcome these limitations.

Lack of in-the-wild SER datasets. The lack of proper SER datasets remains a major problem in SER. As speech tends to have many variations, a wider variety of speech patterns and emotions classes are required to encapsulate all the biases in SER. Considering IEMOCAP dataset, which contains improvised dialogues rather than simply reusing specific sentences, as "*target domain*", we notice that models' show an extreme poor generalization ability (due to the huge variation in data). On the contrary, utilizing IEMOCAP as a "*training domain*" has greatly improved models' generalization due to the presence of a variety of data to learn in spite of the bias. Thus, for meta-learning for DG in SER to advance, there is a great need for large-scale in-the-wild SER datasets to greatly boost models' performance. In that direction, human errors in SER annotations may be the largest obstacle to overcome.

VI. Conclusions and Future Work

Domain generalization in the field of audio, especially SER, is still under development. In this work, we studied the efficacy of meta-learning in SER, and proposed DA−MLDG to improve models' generalizability. While DA−MLDG demonstrates an improvement over the considered baselines, the resulting models' performance is insufficient to be practical. Through our experimentation, we identified a number of pitfalls that contribute to poor generalization ability, namely, the lack of distinct features in input representations from log-mel spectrograms, and the absence of large-scale in-the-wild SER datasets with rich speech patterns and emotions classes. From

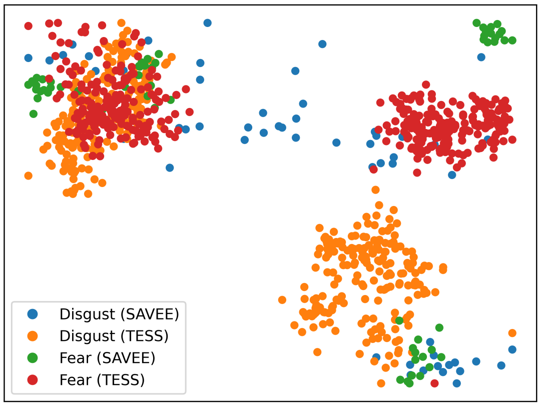

Fig. 2: t-SNE on DA−MLDG model's embeddings (logits before classifier), trained on Scenario 3, for TESS and SAVEE. Insufficient representation power of log-mel spectrograms for DG in SER can be observed.

these pitfalls, future directions for development of adequate model inputs with plentiful distinguishable representations may greatly advance DG in SER. However, for such a direction to be realized, availability of large-scale amounts of speech emotion data is a necessity. Here, providing annotations based on arousal and valence values rather than emotional classes may enable better differentiation between emotions.

Acknowledgment

This work is partially performed in the context of the Distributed Artificial Intelligent Systems project supported by the ECSEL Joint Undertaking.

References

[1] P. Gupta and N. Rajput, "Two-stream emotion recognition for call center monitoring," in *Proc. Interspeech 2007*, 2007, pp. 2241–2244.

[2] A. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Lee, S. Kwon, and S. Baik, "Deep features-based speech emotion recognition for smart affective services," *Multimedia Tools and Applications*, 2019.

[3] H.-J. e. Vögel, "Emotion-awareness for intelligent vehicle assistants: A research agenda," 05 2018, pp. 11–15.

[4] Y. Susanto, A. G. Livingstone, B. C. Ng, and E. Cambria, "The hourglass model revisited," *IEEE Intelligent Systems*, vol. 35, no. 5, 2020.

[5] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomed. Signal Process. Control.*, vol. 47, pp. 312–323, 2019.

[6] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3d log-mel spectrograms with deep learning network," *IEEE Access*, vol. PP, pp. 1–1, 08 2019.

[7] L. Zheng, Q. Li, H. Ban, and S. Liu, "Speech emotion recognition based on convolution neural network combined with random forest," 06 2018.

[8] J. Umamaheswari and A. Akila, "An enhanced human speech emotion recognition using hybrid of prnn and knn," 02 2019, pp. 177–183.

[9] J.-H. Hsu, M.-H. Su, C.-H. Wu, and Y.-H. Chen, "Speech emotion recognition considering nonverbal vocalization in affective conversations," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. PP, pp. 1–1, 04 2021.

[10] N. Scheidwasser-Clow, M. Kegler, P. Beckmann, and M. Cernak, "Serab: A multi-lingual benchmark for speech emotion recognition," *arXiv preprint arXiv:2110.03414*, 2021.

[11] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," 2017. [Online]. Available: https://arxiv.org/abs/1703.03400

[12] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *AAAI Conference on Artificial Intelligence*, 2018.

[13] K. Noh, C. Y. Jeong, J. Lim, S. Chung, G. Kim, J. Lim, and H. Jeong, "Multi-path and group-loss-based network for speech emotion recognition in multi-domain datasets," *Sensors*, vol. 21, p. 1579, 02 2021.

[14] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech 2015*. ISCA - International Speech Communication Association, September.

[15] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," *CoRR*, vol. abs/1901.09960, 2019. [Online]. Available: http://arxiv.org/abs/1901.09960

[16] S.-W. Lee, "Domain generalization with triplet network for cross-corpus speech emotion recognition," 01 2021, pp. 389–396.

[17] J. Gideon, M. Mcinnis, and E. Mower Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization," *IEEE Trans. on Affective Computing*, vol. PP, 05 2019.

[18] K. Noh, J. Lim, S. Chung, G. Kim, and H. Jeong, "Ensemble classifier based on decision-fusion of multiple models for speech emotion recognition," 10 2018, pp. 1246–1248.

[19] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 05 2020.

[20] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "Metareg: Towards domain generalization using meta-regularization," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[21] Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[22] A. Naman, C. Sinha, and L. Mancini, "Fixed-maml for few shot classification in multilingual speech emotion recognition," 2021. [Online]. Available: https://arxiv.org/abs/2101.01356

[23] J. Hsu, Y. Chen, and H. Lee, "Meta learning for end-to-end low-resource speech recognition," *CoRR*, vol. abs/1910.12094, 2019. [Online]. Available: http://arxiv.org/abs/1910.12094

[24] V. Tsouvalas, T. Ozcelebi, and N. Meratnia, "Privacy-preserving speech emotion recognition through semi-supervised federated learning," in *IEEE Int. Conference on Pervasive Computing and Communications Workshops*, 2022.

[25] W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," 2020. [Online]. Available: https://arxiv.org/abs/2010.03759

[26] S. Chopra, P. Mathur, R. Sawhney, and R. R. Shah, "Meta-learning for low-resource speech emotion recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6259–6263.

[27] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech 2018*. ISCA, sep 2018.

[28] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE trans. on affective computing*, vol. 5, no. 4, 2014.

[29] N. Vryzas, R. Kotsakis, A. Liatsou, C. Dimoulas, and G. Kalliris, "Speech emotion recognition for performance interaction," *Journal of the Audio Engineering Society. Audio Engineering Society*, vol. 66, pp. 457–467, 06 2018.

[30] V. LoBue and C. Thrasher, "The child affective facial expression (cafe) set: validity and reliability from untrained adults," *Frontiers in Psychology*, vol. 5, 2015.

[31] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[32] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," Apr. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1188976

[33] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020. [Online]. Available: https://doi.org/10.5683/SP2/E8H2MF

[34] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining frame and turn-level information for robust recognition of emotions within speech," 01 2007, pp. 2249–2252.

[35] O. Mohamad Nezami, P. Jamshid Lou, and M. Karami, "Shemo: a large-scale validated database for persian speech emotion detection," *Language Resources and Evaluation*, vol. 53, no. 1, 2019.

[36] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "Emovo corpus: an italian emotional speech database," in *Int. Conference on Language Resources and Evaluation*, 2014.

[37] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower Provost, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 12 2008.

[38] T.-S. Nguyen, S. Stueker, J. Niehues, and A. Waibel, "Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation," 2019. [Online]. Available: https://arxiv.org/abs/1910.13296

[39] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," 2017. [Online]. Available: https://arxiv.org/abs/1710.03077

[40] a. Deng, "Imagenet: a large-scale hierarchical image database," in *IEEE Conference on computer vision and pattern recognition*, 2009.

[41] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in *ECCV*, 2020.

[42] J. Shor and S. Venugopalan, "Trillsson: Distilled universal paralinguistic speech representations," 2022. [Online]. Available: https://arxiv.org/abs/2203.00236

[43] R. Banse and K. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of personality and social psychology*, vol. 70, 04 1996.