Multi-corpus Affect Recognition with Emotion Embeddings and Self-Supervised Representations of Speech

Sina Alisamir
LIG
Univ. Grenoble Alpes, Atos
Grenoble, France
sina.alisamir@univ-grenoble-alpes.fr

Fabien Ringeval
Grenoble INP, LIG
Univ. Grenoble Alpes, Inria, CNRS
Grenoble, France
fabien.ringeval@univ-grenoble-alpes.fr

François Portet
Grenoble INP, LIG
Univ. Grenoble Alpes, Inria, CNRS
Grenoble, France
francois.portet@univ-grenoble-alpes.fr

***Abstract*—Speech emotion recognition systems use data-driven machine learning techniques that rely on annotated corpora. To achieve a usable performance in real-life, we need to exploit multiple different datasets since each one can shed the light on some specific expression of affect. However, different corpora use subjectively defined annotation schemes, which poses a challenge to train a model that can sense similar emotions across different corpora. Here, we propose a method that can relate similar emotions across corpora without being explicitly trained for it. Our method relies on self-supervised representations, which can provide us with highly contextualised speech representations, and multi-task learning paradigms. This allows to train on different corpora without changing their labelling schemes. The results show that by fine-tuning self-supervised representations on each corpus separately, we can significantly improve the state of the art within-corpus performance. We further demonstrate that by using multiple corpora during the training of the same model, we can improve the cross-corpus performance, and show that our emotion embeddings can effectively recognise the same emotions across different corpora.**

***Index Terms*—affective computing, emotion recognition, emotion embedding, multi-task learning, self-supervised representation**

I. Introduction

Speech Emotion Recognition (SER) has lately become a popular field of research, with applications in many domains such as human-centered services, education, and health. Current state of the art SER exploits supervised Deep Learning (DL) techniques to compute different emotion targets for specific corpora separately [1]–[3]. However, existing corpora available for emotion recognition rarely exceed ten hours, while DL methods require large amounts of labeled data to generalise well on unseen data. Thus, training on only one corpus would result in a model that has not seen enough expressions of affect as represented in real life. This issue is often referred to as data scarcity and several corpora are usually combined to address this problem.

However, emotion corpora are build with targets defined in a subjective manner, using different annotation paradigms, mainly following Ekman's or Russell's theories of affect [4], [5]. For example, RECOLA [6] uses arousal and valence dimensions for annotating emotion, whereas EmoDB [7] and GEMEP [8] use two different sets of emotion categories (c.f. Table I). This inconsistent labelling schemes across corpora poses a challenge for SER, as current state of the art DL methods use a fixed set of categories to train a model.

In order to use multiple corpora with different annotation schemes, emotion labels are often unified across corpora. This is achieved either by mapping them into a common subset of emotion categories, or ignoring a subset of categories. This solution has been shown to cause catastrophic information loss due to the inherent subjective nature of the emotion labels across corpora [9], [10]. Even when the unified emotions have the same or similar psychological meaning, a significant performance loss is observed compared to simply considering different classifiers for different corpora [11].

On the other hand, Multi-Task Learning (MTL) paradigms have been successfully used without the need to unify labels. MTL usually involves sharing hidden layers across different tasks, while using separate classifiers for each one. This means that we can have one shared model across different tasks (or different corpora), where the target of each task (or corpus) is defined differently. MTL has been successfully used in SER, either on one corpus, by considering different classifiers for different emotion dimensions [12], [13], or by using different corpora, where each corpus set of labels is considered a different target [9], [14].

The issue of data scarcity in SER is not only limited to inconsistent definition of emotion categories across different corpora. It is also related to having different speakers, microphones, and environments, which is referred to as domain mismatch. Unlike variable emotion targets, domain mismatch issues on the side of the data do not require labels or supervised DL techniques to be addressed. Thus, one can use unsupervised learning techniques such as self-supervised representations to gain some robustness against domain mismatch in the data.

Self-Supervised Learning (SSL) methods such as the contrastive loss objective used to build a W2V2 model [15], do not need any labels to learn contextualised abstractions of speech,

and can thus benefit from the abundance of unlabelled data. Therefore, by training SSL models on large amounts of data, we can achieve highly contextualised representations of speech that are robust against domain mismatch issues [16]–[18]. The robustness against unseen data can also be further improved by training on data from several different domains [19]. In SER, W2V2 representations have already shown superior performance compared to more traditional features such as Mel-scale Filter Bank (MFBs) or eGeMAPS, over different corpora and by using different classifiers [20]–[23].

Contributions

W2V2 representations can address domain mismatch issues in the data to a great degree. On the other hand, MTL can address the inconsistent way of representing the emotion across corpora. Here, by using both W2V2 and MTL, we propose a method that can compute a generalised emotion embedding that is shared across different corpora, and encompass different representations of affect in a single vector. To summarise our contributions, we achieve the following objectives:

- Evaluation of MTL with W2V2 representations for SER
- Fine-tuning W2V2 representations and analysing the performance in both within-corpus and cross-corpus settings
- Training and evaluation of an emotion embedding that represents similar emotions across multiple corpora

II. Related work

We describe in what follows the work that has been accomplished in the use of MTL for SER, with a focus on cross-corpus settings. We also discuss the concept of emotion embedding, and the interest of self-supervised representations.

Multi-Task Learning for Speech Emotion Recognition

MTL for SER first started as a mean to exploit different ways of annotating emotion for one specific corpus to improve the overall recognition performance. For example, using Arousal and Valence dimensions as an auxiliary task have shown improvements in multiple works [12], [13]. Benefiting from other modalities like video and text, in addition to audio, MTL has also shown to be effective for predicting several emotion categories alongside a sentiment dimension, achieving state of the art performance [24]. As many other works in SER, their MTL system relies on Gated Recurrent Unit (GRU) models, which can learn a joint representation between related tasks.

Multi-Task Learning in Cross-corpus Settings

As there usually exists several domains mismatches between different datasets, several studies have evaluated the performance of MTL for cross-corpus emotion prediction, meaning that the model is trained on one corpus and tested on another. Authors specifically focused on comparing MTL to Single-Task Learning (STL), where the target is only one task, and demonstrated that MTL can provide models with better generalisation capabilities over different emotions [25], providing that there exists a correlation between them [26].

Training multiple SER corpora in MTL was investigated in [9]. Authors used labels of nine different corpora as they were originally defined and remarkably increased the performance compared to STL. They did not map different emotion categories from different corpora into the same subspace on purpose, as this would come with information loss. In a later work [14], other para-linguistic tasks in addition to emotion were also considered, including 18 different classification and regression tasks. A task relatedness matrix was introduced to more efficiently benefit from related tasks. They also showed that, compared to STL, their MTL approach significantly improved the performance over several different tasks. However, the focus of their study was not specific to SER, or to find a representation of emotion that can generalise well over different annotation schemes, but rather to leverage a holistic view of different speech related tasks. Furthermore, a recent study on six different corpora, showed that multi-corpus training can ameliorate the performance of cross-corpus SER, as this approach is more suited to deal with mismatched conditions [27].

Multi-Task Learning for Multi-lingual SER

Using multiple corpora containing different languages and cultures for training a model may reduce SER performance, as emotions might be expressed differently depending on the language and culture [1]. For instance, multi-lingual MTL was investigated on gender, emotion and language tasks for two different Japanese and English datasets and authors reported that multi-lingual models did not work better than mono-lingual models [28]. Moreover, some emotion dimensions like Valence are more sensitive to language [29], especially when using only the audio modality [1]. Despite the reported drop in performance in some studies, it has been shown in many others that using multiple corpora with different languages can still not only achieve reasonable performance but also be beneficial to deal with rare events that occur frequently in real-life [1], [9], [14], [29]. Thus, we can still benefit from using multiple corpora annotated for emotion even if they contain expressions from different languages.

Emotion embedding

Emotion can be described through a multi-dimensional space represented by a numerical vector, which is usually referred to as an emotion embedding. Thus, the emotion embedding's space contains information related to an expressed affect, without being limited to a specific corpus's labelling scheme. Several studies have explored the idea of benefiting from an emotion embedding to improve SER. For example, in [30], authors improved cross-corpus SER by proposing a method that places both the source and the target features into the same subspace containing emotion label information. Exploiting deep ResNet models from the field of vision, in [31], different task-specific classifiers mapped the emotion embedding, computed by a shared frozen model's output, to different targets based on different corpora. However, for multi-corpus experiments, authors mapped the emotion labels

TABLE I
DETAILS OF THE CORPORA USED IN THIS WORK; DURATION IS GIVEN IN HH:MM FORMAT. WE ALSO PROVIDE MAPPINGS OF THE ORIGINAL EMOTION TARGETS OF EACH CORPUS TO NEGATIVE, NEUTRAL, AND POSITIVE CLASSES, AS WERE USED IN OUR CROSS-CORPUS EVALUATIONS.

Corpus	Duration	# Utterances	Negative emotions	Neutral emotions	Positive emotions
CaFE	01:09	936	anger, disgust, fear, sadness	neutral, surprise	joy
EmoDB	00:25	535	anger, anxiety, boredom, disgust, sadness	neutral	happiness
GEMEP	00:51	1260	anger, despair, fear, irritation, sadness, worry	-	fun, interest, joy, pleasure, pride, relief
RAVDESS	01:29	1440	anger, disgust, fear, sadness	neutral, surprise	calmness, happiness

of each corpus to the same target subspace and did not use them as they were originally defined.

Emotion embeddings have also been specifically studied in [10], with the help of an emotional encoder based on convolutional recurrent layers. Then, multiple corpus-dependent classifiers mapped the emotion embeddings to corpus specific emotion classes. Authors showed that by using an adversarial process to remove corpus-specific non-emotional information, they can obtain an emotion embedding that contains cross-corpus emotional information. However, their results were obtained only on two corpora who shared almost the same exact labels of emotion.

Self-supervised representations

Despite their novelty, self-supervised representations have already become the popular choice to describe speech data instead of traditional features such as MFBs, due to their superior performance on many speech related tasks. For example, in [32], W2V2 representations have shown performance improvement for dimensional emotion prediction compared to MFB features. Authors showed that W2V2 representations allow the use of less complex models, compared to MFB features, concluding that W2V2 representations provide contextualised information of speech that are robust in different contexts. Thus, by using self-supervised representations, less labelled data are needed for the downstream task, which is beneficial for SER, as it is highly susceptible to data scarcity issues [18].

Moreover, non-quantised Wav2Vec representations were benchmarked over 17 different SER corpora using nine different machine learning methods, and results showed their superior performance over different acoustic features [21], although traditional representations showed a statistically less variant performance across different corpora, which was also observed in [20]. Another study showed that W2V2 representations achieved better performance than traditional acoustic feature sets such as eGeMAPS [22], and also showed that using the pretrained W2V2 representations performed better for the SER task than representations fine-tuned for ASR.

III. METHOD

In this section, we first explain the corpora and the W2V2 representations that were used in this work. Then, we describe the architecture of our model and its training strategy in detail.

A. Corpora

We used four different acted emotion corpora, which contain speech data expressed in three different languages (French, German, English), with four different accents (Swiss, Canadian, German, American). We focused on acted emotion to more clearly study the interplay of our system, before moving on to using data recorded in-the-wild, which would be harder to analyse, as it contains more subtle emotions as well as environmental noise. A summary of the corpora used here is provided in Table I. For all the corpora that we used, there were no official partitioning of data for training a machine learning model. Thus, excepted for Leave One Speaker Out (LOSO) evaluations, where we use one speaker for evaluation and the rest for training the model, we used our own partitioning, where we keep a balance between male and female speakers, and use a distribution of 70%-15%-15% for choosing the training-development-test partitions.

1) CaFE: the Canadian French Emotional (CaFE) [33] dataset includes about one hour of emotional speech from twelve actors who read six different French sentences with basic emotions of anger, disgust, happiness, neutral, fear, surprise and sadness. Here, audio files from actors 9 and 10 are used for development and files from actors 11 and 12 are used for testing. The rest were used for the training partition.

2) EmoDB: the Berlin database of emotional speech (EmoDB) [7] contains about half an hour of 500 acted German phrases in a happy, angry, anxious, fearful, bored, disgusted, and neutral way from ten different German actors. Here, we used speech from actors 11 and 13 for development, 15 and 16 for testing, and the rest were used for training.

3) GEMEP: The GEneva Multimodal Emotion Portrayals (GEMEP) [8] contains about one hour of emotional speech that contain no verbal information (the words do not mean anything) uttered from ten actors with a Swiss French accent. The original data contains 18 different emotional states. However, to follow other works [8], [34] we use the 12 core emotions of anger, despair, worry, irritation, fear, sadness, amusement, joy, pride, interest, pleasure, and relief. We used speech from actors 5 and 9 for development, 8 and 10 for testing, and the rest were used for training.

4) RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [35] contains 7356 files, from which we only use the emotional speech utterances. The database contains 24 professional actors in a neutral North American accent. It also contains eight emotion categories of anger, calm, disgust, fear, happiness, neutral, sadness, surprised. We used speech from actors 19, 20, and 21 for development, 22, 23, and 24 for testing, and the rest were used for training.

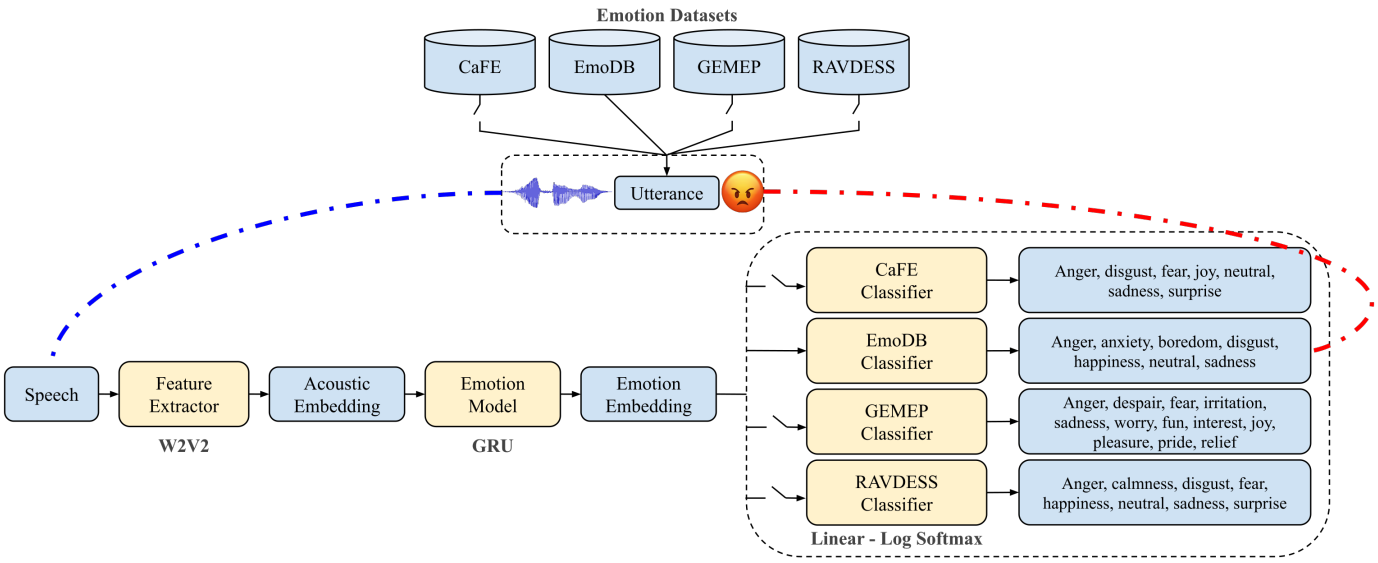

Fig. 1. The proposed multi-corpus emotion recognition model. The feature extractor and the emotion embedding models are shared for all datasets. However, for each set of emotion categories for each corpus, a specific classifier is used to map the emotion embedding to the probabilities of the emotion classes.

B. W2V2 Representations

One of the most popular SSL models is W2V2, which exploits a contrastive predicting loss to predict masked frames of speech [15]. SSL models are usually trained on huge amounts of data to provide speech representations that are more contextualised and less impacted by noises, compared to traditional acoustic features such as MFBs. Here, since our data contain different languages, We used a multilingually trained W2V2 model[1]. We did not use any normalisation scheme for our representations as we used a large architecture, which already contains a normalisation layer.

C. Model architecture

The proposed model as depicted in Figure 1 consists of three main parts:

1) Feature extractor: Different audio representations such as MFB or W2V2 can be used here to provide an acoustic embedding, given a speech wave of an utterance.
2) Emotion model: Given an acoustic embedding, the emotion model, which is the same for every corpus used, calculates the emotion embedding. Here, we can use a GRU for the emotion model and pick the last frame's output as the emotion embedding.
3) Classifier: We use different classifiers to map the emotion embedding to different emotion categories based on the specifications of the used corpora. Each classifier has a linear layer followed by a log-softmax function to estimate class probabilities of each corpus' set of emotions.

For an utterance from a given corpus, the model first calculates the acoustic embedding, from which, the emotion model predicts the emotion embedding. Then, by using different classifiers, the emotion embedding can be mapped into the set of emotion classes of a target corpus. Thus, during the training, the emotion model is shared across all the corpora, while each corpus uses a different classifier. The hypothesis is that by sharing the emotion model during the training, it can learn to represent an "*understanding*" of the underlying perceivable emotion across the different corpora. We tested this hypothesis by both visualising the emotion embedding and quantitatively evaluating our method in both within-corpus and cross-corpus settings.

D. Training strategy

The shared emotion model and the classifiers are trained jointly. We first randomly pick and copy the files for the under-represented corpora so that all datasets have the same number of utterances for training and development partitions. Then, we randomly pick an utterance from the pool of all the utterances of all the corpora. The utterance is then given to the model and depending on the dataset of the chosen utterance, the appropriate classifier is used, and then the loss is backwarded through the classifiers as well the emotion model. The loss in our case is a cross-entropy loss, which is commonly used for classification tasks. In this way, the emotion model would continue to be trained, regardless of which corpus the utterance input belongs to. Only the classifiers are optimised according to the corpus to which the chosen utterance belongs.

Fine-tuning: In our baseline experiments the weights of the W2V2 model are frozen during the training. However, in our Fine-Tuning (FT) experiments, the loss is also backwarded through the W2V2 model. Thus, based on the gradients calculated for each training iteration, we allow the weights of the W2V2 model to be updated alongside the weights of the emotion model and the classifiers. In this way, the acoustic representations are influenced by the utterances used during the training.

[1]https://huggingface.co/voidful/wav2vec2-xlsr-multilingual-56

TABLE II

RESULTS OF OUR WITHIN-CORPUS EXPERIMENTS FOR CAFE, EMODB, GEMEP AND RAVDESS DATASETS. WE USED OUR OWN PARTITIONING TO REPORT ON THE BASELINE AND FINE-TUNING RESULTS. WE ALSO USED LEAVE ONE SPEAKER OUT (LOSO) CROSS VALIDATION TO BE ABLE TO FAIRLY COMPARE OUR RESULTS TO STATE OF THE ART.

Method	Evaluation Metric	CaFE	EmoDB	GEMEP	RAVDESS	Average
Baseline						
Single-corpus	UAR	**63.7 %**	57.5 %	**39.8 %**	60.4 %	55.3 %
Multi-corpus	UAR	60.1 %	**67.4 %**	39.4 %	60.4 %	**56.8 %**
Baseline + W2V2 Fine-tuning						
Single-corpus + FT	UAR	**76.2 %**	**72.5 %**	**52.8 %**	71.4 %	**68.2 %**
Multi-corpus + FT	UAR	75.0 %	69.3 %	44.9 %	**71.9 %**	65.3 %
State of the Art						
Ours (Single-corpus + FT)	UAR (LOSO)	**77.2 %**	**90.5 %**	**55.8 %**	**82.2 %**	**76.4 %**
Subspace learning and extreme learning [34]	UAR (Random Partitioning)	-	-	43.3 %	-	
Prosodic and spectral features + SVM [36]	Accuracy (10 fold Cross-Validation)	70.6 %	86.0 %	-	70.6 %	-
MFCC/GTCC features with echo state network [37]	UAR (LOSO)	-	86.8 %	-	73.1 %	-

TABLE III

RESULTS (UAR) OF CROSS CORPUS EXPERIMENTS FOR EITHER THREE MAPPED CLASSES OF NEGATIVE, NEUTRAL AND POSITIVE, OR THE COMMON EXISTING EMOTION CLASSES BETWEEN THE UTTERANCE'S CORPUS AND THE CLASSIFIER'S CORPUS. THE EFFECT OF FINE-TUNING (FT) THE WEIGHTS OF THE W2V2 MODEL, ON BOTH WITHIN-CORPUS AND CROSS-CORPUS CASES, IS ALSO REPORTED. STATISTICS ARE GIVEN IN THE FOLLOWING FORMAT: MEAN (STANDARD DEVIATION).

Method	Within-corpus	Cross-corpus
Negative, neutral, and positive classes		
Single-corpus	67.2% (4.7%)	42.0% (8.3%)
Single-corpus + FT	72.3% (5.5%)	44.6% (12.0%)
Multi-corpus	69.1% (4.5%)	**50.8% (11.6%)**
Multi-corpus + FT	**73.8% (4.0%)**	42.3% (12.9%)
Common classes		
Single-corpus	55.3% (10.7%)	41.2% (16.2%)
Single-corpus + FT	**68.2% (10.5%)**	45.3% (10.8%)
Multi-corpus	56.8% (12.1%)	**47.5% (10.7%)**
Multi-corpus + FT	65.3% (13.8%)	32.1% (25.1%)

IV. EXPERIMENTS AND RESULTS

A. Setup

All the experiments were done using Pytorch [38] and the SpeechBrain toolkit [39] with seeds set to zero manually. The computer's OS was Debian GNU/Linux 10, and we used an NVIDIA Quadro RTX 6000 with 23 Giga-bytes of memory, CUDA version 11.3.

To define the setup of our experiments, we used different possible models and parameters that are commonly used for SER, considering different ranges of complexity. Then, to choose the best setup, we ran a grid search over different possible setups with the following parameters:

- Emotion model: GRU, Transformer (8 heads)
- Feature: MFB, W2V2
- Hyper-parameter: 1 layer with 64 nodes, 2 layers with 128 nodes, 4 layers with 256 nodes
- Learning rate: 0.01, 0.001, 0.0001

After training each possible setup for 50 epochs, and evaluation on both GEMEP and RAVDESS datasets independently, we found the setup of the W2V2 representation with the GRU model using 1 layer with 64 nodes, and learning rate of 0.0001 to be the best in terms of performance. We chose the classifier to be a linear layer followed by a log-softmax layer to obtain the probabilities of different classes. We also tried replacing the linear layers of the classifiers with GRUs, however we did not observe any improvement.

B. Within-corpus evaluation

To test our method, we first compared training and testing on one corpus only (single-corpus) *vs.* training on multiple corpora (multi-corpus). Then, we tested each model for each corpus separately. Results are quantified with the Unweighted Average Recall (UAR) and are given in Table II.

1) Baseline: For our baseline experiments, we froze the W2V2 weights and used it solely to provide representations of speech. Results show that, by using multi-corpus instead of single-corpus training, we can have an overall significant improvement in the UAR. We also found that multi-corpus MTL can improve the performance on a specific corpus (EmoDB) by using other corpora during the training, which is inline with previous studies [9], [10], [14].

2) Fine-tuning: We also did not freeze the W2V2 weights in order to have an end-to-end MTL paradigm through the fine-tuning of SSL representations. Results show that on average, with FT, training and testing only for one corpus achieves better results than when we utilise multiple corpora. We think that by training our system end-to-end, W2V2 models learn to predict more corpus specific acoustic representations instead of generic cross-corpus features. These corpus specific representations can perform better on the corpus they were trained for, but would not be able to generalise well across corpora. This is further discussed in Section IV-C, where we evaluate the cross-corpus performance of our method in more detail.

3) Comparison to state of the art: Moreover, we added a Leave One Speaker Out (LOSO) cross-validation in order to be able to fairly compare our work to the state of the art. The

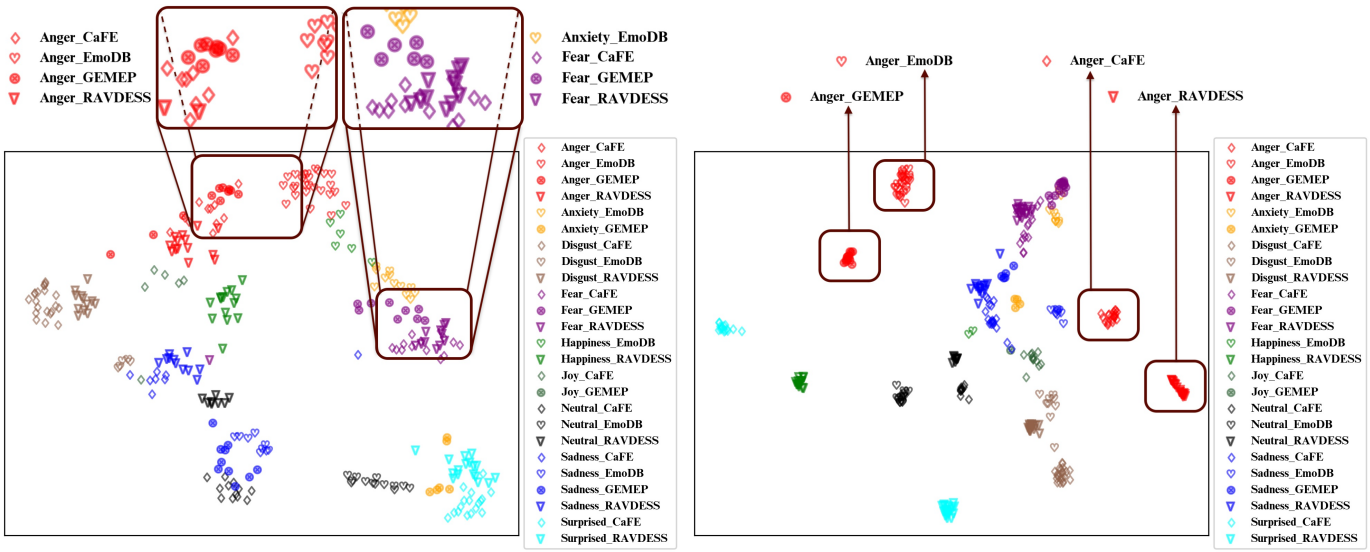

Fig. 2. Emotion embeddings of different correctly classified utterances of the test partitions of the studied corpora. Only emotions that were used at least in two different corpora are shown. On the left: emotion embeddings of the baseline model with W2V2 weights frozen during the training; clusters of similar emotions across different datasets can be identified in this space. On the right: emotion embeddings of the baseline model with fine-tuning of the W2V2 weights during the training; clusters of emotion are now specific to each dataset and similar expressions are located in different parts of the emotion space.

results show that our best model (fine-tuned single-corpus) can achieve better performance than the state of the art, which we think is mainly thanks to using highly contextualised speech representations in our method.

C. Cross-corpus evaluation

The generalisation of emotion targets across corpora can be studied by putting an utterance of a certain corpus as input and observe the output through the classifier of another corpus. Since different corpora do not use the same set of classes, we can not quantify our cross-corpus results the same way as for the within-corpus experiments. Thus, we considered either mapping the prediction of each utterance to be either a negative, neutral or a positive class (c.f. Table I) or only use the utterances with common emotion classes between the corpus of the utterance and the corpus of the classifier. For example, CaFE and GEMEP datasets both have four common emotion classes of anger, fear, joy, and sadness. Results for both modes of evaluation are presented in Table III.

1) Baseline: Cross-corpus results show that the multi-corpus method significantly outperform the single-corpus method. We used UMAP [40] to reduce the dimensions of our emotion embeddings and present them in a two dimensional space, cf. Figure 2. One can easily observe that for our baseline multi-corpus training paradigm (no FT), the emotion embeddings of the utterances of different corpora are mostly put closer to each other in the embedding space, when they represent the same or similar emotions. This means that the proposed multi-corpus method can in most cases obtain a sense of the underlying emotion across different corpora. For example, different utterances representing anger are put closer to each other on top part of the figure. Interestingly, We can also see that utterances labelled as anxiety in EmoDB are put close to utterances labelled as fear in the other corpora. We can further observe that utterances labelled as happy and joy are close to the ones labelled as anger. We think this can be because both happiness and anger are associated with a high level of arousal according to Russell's theory of core affect.

We further provide a confusion matrix, where we predicted the utterances of CaFE using the GEMEP classifier, in Figure 3. Here, the baseline model without FT is mostly correct for the four common emotions between the two datasets (anger, fear, joy, and sadness). Interesting results are further observed when analysing emotion labels that are different between the two datasets. For example, disgust utterances of the CaFE dataset are mainly labelled as irritation, which indicates that the suggested method, is able to some degree, generalise across different emotion labels in different corpora.

2) Fine-tuning: Results in Table III suggest that while FT can be effective when training and testing on the same corpus, it significantly drops the performance in domain mismatched conditions. Thus, the best results in cross-corpus settings is achieved with the baseline method, where the weights of the W2V2 model are frozen.

Visualisations of the emotion embeddings given in Figure 2 shows that, after FT, different utterances from the same corpus tend to be closer to each other in the embedding space, rather than similar emotion categories from different corpora. This means that after FT the self-supervised representations, the emotion embeddings tend to become corpus-dependent, which can be the result of the acoustic embeddings becoming corpus-dependent. One can also see that FT makes the MTL method less focused on sensing the underlying emotion across the corpora. For example, sadness utterances of the CaFE dataset are mostly labelled as joy with the GEMEP classifier, and anger as pleasure, cf. Figure 3.

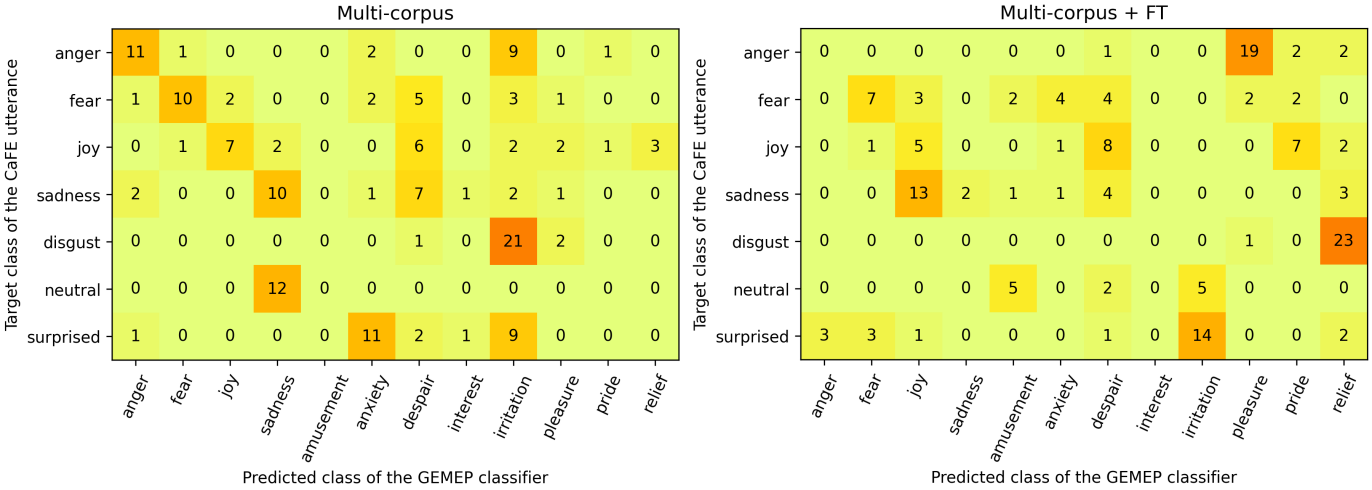

Fig. 3. Confusion matrix of the GEMEP classifier's predictions from the CaFE utterances. Here, the original target label for each utterance is based on the set of emotion labels of the CaFE corpus, whereas the prediction of the model is based on the set of emotion labels of the GEMEP corpus. On the left: The confusion matrix of the suggested multi-corpus method without fine-tuning. On the right: The confusion matrix of the method, where W2V2 weights are fine-tuned during the training of the model.

Visualisations of the emotion embeddings in Figure 2 alongside the quantitative results reported in Table III and Figure 3, suggest that FT drops the performance in cross-corpus settings and put the emotion embeddings of the same corpora closer to each other while being distant to other corpora. This means that fine-tuning the W2V2 acoustic representations, as used in our MTL approach, would make the model too specific to the speech content of the used datasets.

On the other side, if we use non-contextualised generic representations of speech such as MFBs, the multi-corpus paradigm works poorly compared to using W2V2 representations. We think that this is because we did not have enough training data to learn both contextualised representations of speech and different emotion labels. This shows that when having limited training data, the high contextualisation of W2V2 representations without fine-tuning can be especially beneficial for the multi-corpus MTL paradigm, to learn generalised representations of emotion.

V. Conclusions

We expanded upon and evaluated the idea of using an emotion embedding that can get a sense of emotion, by observing that in the embedding space, the same emotion categories across different corpora are placed closer to each other. We showed that through this method, we can not only improve cross-corpus emotion recognition accuracy but also provide a framework that can address the problem of data scarcity of emotion labels in SER, by allowing the training of a shared neural network for multiple corpora with different annotation schemes.

We also showed that using highly contextualised W2V2 representations are especially effective for SER, and by fine-tuning them in single-corpus settings in an end-to-end manner, we can improve state of the art performance for SER. However, by fine-tuning W2V2s, we lose cross-corpus information in the acoustic embedding space, which makes the method less relevant in cross-corpus settings.

Furthermore, our preliminary results of evaluation on unseen corpora show that this method cannot yet generalise very well beyond the used corpora, which is not surprising since we only used four rather small datasets. Thus, since the results clearly show that this method can reconcile different emotion labeling paradigms across the trained corpora, we see one future path to be utilising more labelled corpora in our multi-corpus MTL method. We would also look into integrating different corpora that use dimensional annotation schemes as their emotion targets, as well as, emotional data gathered in the wild to test robustness over environmental noises.

Acknowledgment

This work is part of a Cifre thesis with the grant number 2019/0729 from National Association for Research and Technology (NART) and was partially supported by MIAI@Grenoble-Alpes (ANR-19-P3IA-0003).

References

[1] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner *et al.*, "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, 2019, pp. 3–12.

[2] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Deep representation learning in speech processing: Challenges, recent advances, and future trends," *arXiv preprint arXiv:2001.00378*, 2020.

[3] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, 2021.

[4] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[5] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[6] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–8.

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

[8] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the geneva multimodal expression corpus for experimental research on emotion perception." *Emotion*, vol. 12, no. 5, p. 1161, 2012.

[9] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4990–4994.

[10] Z. Zhu and Y. Sato, "Reconciliation of multiple corpora for speech emotion recognition by multiple classifiers with an adversarial corpus discriminator." in *Proc. Interspeech*, 2020, pp. 2342–2346.

[11] R. da Silva, M. Valter Filho, and M. Souza, "Interaffection of multiple datasets with neural networks in speech emotion recognition," in *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*. SBC, 2020, pp. 342–353.

[12] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on affective computing*, vol. 8, no. 1, pp. 3–14, 2015.

[13] N. K. Kim, J. Lee, H. K. Ha, G. W. Lee, J. H. Lee, and H. K. Kim, "Speech emotion recognition based on multi-task learning using a convolutional neural network," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 704–707.

[14] Y. Zhang, F. Weninger, S. Björn, and R. Picard, "Holistic affect recognition using panda: paralinguistic non-metric dimensional analysis," *IEEE Transactions on Affective Computing*, vol. 13, pp. 769–780, 2019.

[15] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[16] C.-C. Lee, K. Sridhar, J.-L. Li, W.-C. Lin, B.-H. Su, and C. Busso, "Deep representation learning for affective speech signal analysis and processing: Preventing unwanted signal disparities," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 22–38, 2021.

[17] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, pp. 992–1004, 2020.

[18] S. Alisamir and F. Ringeval, "On the evolution of speech representations for affective computing: A brief history and critical overview," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 12–21, 2021.

[19] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve *et al.*, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, 2021.

[20] S. Evain, H. Nguyen, H. Le, M. Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet *et al.*, "Task agnostic and task specific self-supervised learning from speech with lebenchmark," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[21] A. Keesing, Y. S. Koh, and M. Witbrock, "Acoustic features and neural representations for categorical emotion recognition from speech," in *Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Brno, Czech Republic*, 2021, pp. 3415–3419.

[22] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[23] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition," *arXiv preprint arXiv:2008.06682*, 2020.

[24] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," *arXiv preprint arXiv:1905.05812*, 2019.

[25] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proc. Interspeech*, 2017, pp. 1103–1107.

[26] A. M. Oliveira, M. P. Teixeira, I. B. Fonseca, and M. Oliveira, "Joint model-parameter validation of self-estimates of valence and arousal: Probing a differential-weighting model of affective intensity." *Proceedings of Fechner Day*, vol. 22, pp. 245–250, 2006.

[27] N. Braunschweiler, R. Doddipatla, S. Keizer, and S. Stoyanchev, "A study on cross-corpus speech emotion recognition and data augmentation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 24–30.

[28] S.-w. Lee, "The generalization effect for multilingual speech emotion recognition across heterogeneous languages," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5881–5885.

[29] M. Neumann *et al.*, "Cross-lingual and multilingual speech emotion recognition on english and french," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5769–5773.

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

[31] M. Gerczuk, S. Amiriparian, S. Ottl, and B. W. Schuller, "Emonet: A transfer learning framework for multi-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, 2021.

[32] S. Evain, H. Nguyen, H. Le, M. Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet, A. Allauzen, Y. Estève, B. Lecouteux, F. Portet, S. Rossato, F. Ringeval, D. Schwab, and L. Besacier, " LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech," in *Proc. Interspeech 2021*, 2021, pp. 1439–1443.

[33] P. Gournay, O. Lahaie, and R. Lefebvre, "A canadian french emotional speech dataset," in *Proceedings of the 9th ACM multimedia systems conference*, 2018, pp. 399–402.

[34] X. Xu, J. Deng, E. Coutinho, C. Wu, L. Zhao, and B. W. Schuller, "Connecting subspace learning and extreme learning machine in speech emotion recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 795–808, 2018.

[35] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[36] M. El Seknedy and S. Fawzi, "Speech emotion recognition system for human interaction applications," in *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE, 2021, pp. 361–368.

[37] H. Ibrahim, C. K. Loo, and F. Alnajjar, "Speech emotion recognition by late fusion for bidirectional reservoir computing with random projection," *IEEE Access*, vol. 9, pp. 122 855–122 871, 2021.

[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, Z. Lin, N. Gimelshein, and L. Antiga, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of the thirty-third Conference on Neural Information Processing Systems (NIPS)*. Vancouver, Canada: Neural Information Processing Systems Foundation, 2019, pp. 8026–8037.

[39] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[40] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.