SPEAKER VGG CCT: Cross-corpus Speech Emotion Recognition with Speaker Embedding and Vision Transformers

Alessandro Arezzo  
University of Florence  
Florence, Italy  
alessandroarezzo96@gmail.com

Stefano Berretti  
University of Florence  
Florence, Italy  
stefano.berretti@unifi.it

ABSTRACT

In recent years, Speech Emotion Recognition (SER) has been investigated mainly transforming the speech signal into spectrograms that are then classified using Convolutional Neural Networks pre-trained on generic images and fine-tuned with spectrograms. In this paper, we start from the general idea above and develop a new learning solution for SER, which is based on Compact Convolutional Transformers (CCTs) combined with a speaker embedding. With CCTs, the learning power of Vision Transformers (ViT) is combined with a diminished need for large volume of data as made possible by the convolution. This is important in SER, where large corpora of data are usually not available. The speaker embedding allows the network to extract an identity representation of the speaker, which is then integrated by means of a self-attention mechanism with the features that the CCT extracts from the spectrogram. Overall, the solution is capable of operating in real-time showing promising results in a cross-corpus scenario, where training and test datasets are kept separate. Experiments have been performed on several benchmarks in a cross-corpus setting as rarely used in the literature, with results that are comparable or superior to those obtained with state-of-the-art network architectures. Our code is available at https://github.com/JabuMIDev/Speaker-VGG-CCT

CCS CONCEPTS

• Information systems → Speech / audio search. • Computing methodologies → Speech recognition.

KEYWORDS

Speech emotion recognition, spectrograms, visual transformers, compact convolutional transformers, speaker embedding, cross-corpus test

1 INTRODUCTION

The human-machine communication reached truly remarkable achievements so far. For example, an impressive progress has been achieved in the conversion of human speech signals in text, with a growing variety of voice assistants now available on the market for disparate uses that go from home automation to the automotive one. Despite of these advances, an obstacle that does not allow the man-machine interaction to be completely indistinguishable from what occurs between two people is that machines still are not ready to fully understand the emotion of the interlocutor. This constitutes a major limitation given the importance that emotions play in dialogues between two human beings.

Motivated by these premises, recognizing the emotion of a speaker from her/his audio track is a task which is receiving an increasing attention. In particular, recent developments in the field of real-time SER exploited the advances in deep learning and transformed SER to an image classification task [17]. In these works, the audio track is usually transformed into a spectrogram image, which is then classified using Convolutional Neural Networks (CNN) or a concatenation of CNNs and Long-Short Term Memory (LSTM) networks [26]. However, there are several unsolved problems that still make this task very challenging such as the lack of usable data in the desired language, the difficulty in obtaining from an audio track features that, at the same time, can be extracted in real-time and characterize the emotion, and the dependence between the representation of the emotion and the subject that expresses it. Further, most of the studies focused on the in-corpus scenario, while the more realistic cross-corpus protocol, where training and test data are completely separated has been rarely tested with.

Developing on the above considerations, in this paper we aimed to overcome some of the aforementioned difficulties, and developed an approach capable of predicting, in real-time, the emotion of a person starting from her/his speech track. One main contribution of this work is that of using Vision Transformers (ViTs) [9] in place of traditional CNN for SER. ViTs proved their effectiveness in vision related tasks when used alone or in combination with CNNs, achieving performance that can surpass CNNs in many image classification tasks [8, 37], but they did not find yet application to SER. To this end, we used Compact Convolutional Transformers (CCTs) [12], which represent one variant of the ViT that is simpler to train when example data are scarce. Subsequently, we extended this model to the Speaker VGG CCT that exploits the ViT self-attention mechanism in an architecture that makes use of a representation of the speakers (speaker embedding) that partially compensates for the fact that each person potentially expresses emotion in a different way. As a further contribution, we target a
cross-corpus scenario. To validate the proposed approach, we performed a comprehensive set of experiments on the most commonly used speech datasets. First, we presented results for the challenging cross-corpus classification protocol and compared our method with state-of-the-art CNN image classification architectures fine-tuned for SER. Then, we also performed an in-corpus experiment to make our approach comparable with state-of-the-art solutions.

2 RELATED WORK

Voice signals with associated different emotions are characterized by changes in frequency and intensity of their waveform [2]. Therefore, several works focused on capturing those variations using different discriminating acoustic features. Until few years ago, naive (i.e., hand-crafted) features were proposed in various studies [11]. These features can be categorized on the basis of various aspects such as the domain of extraction, time or frequency, timing, local or global, type, continuous or quality of the voice, etc., using classifiers that go from Hidden Markov Models (HMMs) to Support Vector Machines (SVMs) and decision trees [16, 27, 30]. Several challenges were also organized to understand the better features to use [31, 32, 36]. More recently, deep learning (DL) solutions that are driven from the data were proposed [20]. Among different approaches, spectrogram-based solutions emerged as of particular interest for the fact spectrograms can be extracted quickly for real-time solutions, and can be also used as direct input to CNN and LSTM models pre-trained for the image classification task. This has the further advantage of exploiting the huge amount of available image data compared to the scarcity of speech tracks.

In the following, we focus on DL methods that used spectrograms as input data (a summary on spectrogram generation from audio tracks is given in the supplementary material).

Deep learning. One of the first works that experimented a DL approach for SER was proposed in [1]. In this work, spectrograms were used as features of audio data by comparing two CNN solutions: in the first one, a network was trained from scratch or with randomly initialized weights while, in the second one, a fine-tuning operation was performed on an Alexnet [15] pre-trained on the Imagenet dataset [28, 29] for the image classification task. In [33], the first work was presented that applied CNNs for real-time SER. This study compared two approaches based on the extraction of RGB spectrograms of the audio data, and on fine-tuning an AlexNet taken as a reference CNN. In the first architecture, named AlexNet-SVM, a pre-trained CNN was used to extract from the spectrograms, the features generated by its second fully connected layer, which are then given as input to an SVM classifier. In the second approach, instead, a fine-tuning was applied to the AlexNet (FTAlexNet) to take advantage of what the model learned to detect during pre-training, while adapting, at the same time, the weights of the last layers to the domain represented by spectrograms. The FTAlexNet shown better performance. This work was deepened in [17], where it is illustrated the performance decline of the models proposed, while varying the sampling frequency used to acquire the signals.

SER cross corpus. In SER, few papers focused on cross-corpus evaluation, which is referred to as the case where datasets used for training (one or more) are different from those used in the test. This corresponds to a more realistic application scenario than that used by in-corpus evaluations. In these studies, increasing the data available for training was used to improve the performance. This was obtained either using data augmentation in the time or frequency domain, or forming a training set by the union of different datasets that can also contain examples in different languages.

In [26], a CNN, an LSTM, and an architecture given by their combination were used in a cross-corpus scenario to predict the emotions from spectrograms given as input representation of the audio tracks. In order to evaluate the performance, the IEMOCAP [4], EMOVO [7], EmoDB [3], EPST [19], RAVDESS [21], SAVEE [34] and TESS [10] datasets were used in a three-class classification task, where each emotion label is mapped into the positive, negative or neutral class on the basis of its membership in the target domain. In a first experiment, the models were trained on IEMOCAP, the largest dataset among those considered, and evaluated on each of the remaining sets. In a second one, the models were in turn trained using n-1 datasets and tested on the remaining one. This allowed authors to evaluate the enhancement deriving from a cross-corpus training. The accuracy ranged from 30% to 50% for the first experiment, while in the second one the accuracy was around 50% in all cases but Emo-DB and SAVEE, where it reached 69.7% and 72.7%, respectively. The best performing model was obtained by concatenating the CNN and LSTM, that resulted slightly better than the CNN model, and much better than the LSTM one.

Attention in SER. One current trend in DL is that of empowering the proposed architectures with attention mechanisms. An overview of DL-based methods in SER, with specific analysis of variants that use attention mechanisms is given in [20].

In [35], the authors proposed a self-attention mechanism to reduce the dependence of the emotion on the speaker who expresses it. To this end, a particular architecture known as ACRNN, was presented for the first time in SER in [6]. This receives in input 3D spectrograms generated by calculating the Mel spectrograms of each audio track and their respective first and second derivatives. Such 3D-representations were computed by a 3D CNN based on the application of 3D convolutions. The outputs of this network were subsequently processed by a Bidirectional LSTM (BLSTM). The output of this module, which represents a sequence with the temporal dependencies extracted from the first part of the architecture is then processed by a level of self-attention that returns the descriptor used by a fully-connected layer to classify the emotion. The architecture was tested with an in-corpus protocol on the Att-HACK [22] and IEMOCAP datasets, and compared with the basic model given by the ACRNN. Promising results were reported with the accuracy passing, respectively, on Att-HACK and IEMOCAP, from 66.4% and 86.4% for ACRNN, to 88.9% and 96.1% for the proposed architecture.

In [35], a self-attention mechanism embedded into a Capsule Network, and a Transfer Learning based Mixed Task Net (CtLMNet) were proposed to deal with both the single corpus and cross-corpus SER tasks simultaneously. This work shares the self-attention idea with our proposal but without using ViT and speaker embedding. Moreover, in their cross-corpus tests only one dataset was used for training, while we trained our proposed network on the union of different datasets.
Following some recent works [18], we propose a real-time SER method that first transforms the speech signal to a spectrogram image, then trains a neural network architecture to classify the spectrograms. Inspired by recent results on image analysis, we propose a Vision Transformer (ViT) [14] based architecture. We did not find in the literature any previous attempt of using ViTs for SER. One reason for this is that in the image domain such architecture showed superior image classification results with respect to state-of-the-art ResNet-based models [13] when trained on very large amount of data. Since in SER there is not the same data abundance as in the image domain, we started from the Compact Convolutional Transformer (CCT) [12] that introduces convolution in ViT to make their performance superior to those of convolutional architectures even when trained on smaller amount of data.

3.1 Speaker VGG CCT

In our solution, we used a CCT-14/7x2 model trained on ImageNet [15], where 14 is the number of filters in the Transformer Encoder, and 2 is the number of convolutional layers each with filters of size 7 × 7. These layers are applied before the ViTs layers and receive as an input spectrogram images of size 224 × 224 generated from the audio track.

In order to adapt the CCT-based architecture to SER, we investigated the possibility to better represent the dependence of a speaker emotion and the speaker him/her-self. For example, not all people express a state of happiness or frustration in the same way, causing the spectrograms corresponding to the same spoken text and emotional state to be very different when passing from one subject to another. This can lead to overfitting, where models tend to learn how to classify emotions for subjects in the training dataset, but do not generalize well in testing on spectrograms extracted from speakers that were never seen before. So, our idea here is to enhance the basic CCT model with a descriptor of the speakers by combining the representation through a self-attention module. We got inspiration from the work in [23], where a 3D Attention-based Convolutional Recurrent Neural Networks (ACRNN) was trained to classify the speakers (see also Section 2 for more details on [23]).

We propose two architectures based on the VGG-Vox [24] speaker embedder. This is basically a CNN based on the Visual Geometry Group (VGG)-M [5] network trained for the speaker recognition task starting from the spectrograms of the speech signal of the VoxCeleb dataset [24]. These were extracted in the same way as for all the models presented in this paper by applying a DFT with a Hamming window of size 25ms and stride of 10ms. The resulting embedder is a vector with dimension 1024.

The first model, denoted as Speaker VGG CCT works in a similar way as in [23]. In this method, the speaker embedding extracted as reported above is combined downstream of the CCT architecture, making it possible to learn the dependence between speaker and emotion. This is illustrated on the left of Figure 1.

In the second model, instead, we defined an end-to-end solution, called Speaker VGG CCT end-to-end, where it is exploited the fact that with CCT the self-attention module is already integrated into the classifier. Therefore, the architecture ensures that the speaker embedding is given as input to the Transformers Encoder block so that it directly learns the dependence between speaker and emotion. This solution is illustrated in the middle of Figure 1. In this case, the speaker embedding is passed directly to the Transformer Encoder block of CCT. Similar to traditional CCTs, the image is processed by a sequence of convolutional blocks that output a set of feature maps. The speaker embedding is then linked with such feature maps and the entire sequence is given as an input to the Transformer Encoder. The output is a set of vectors equal to the number of feature maps plus the one associated with the representation of the speaker. A pooling layer is then used to obtain a single vector in the dimension of the transformer starting from the exit sequence. This output is then used by an MLP network for classification. Therefore, the learning of the dependence between speaker and emotion underlying the speech represented by the spectrogram occurs directly in the self-attention modules internal to the Transformer Encoder. In this way, an additional vector is returned that represents the attention to be given to each patch in interpreting the speaker embedding.

Given the importance of this speaker embedding vector, with a different meaning than those representing the feature maps, and getting inspiration from the original implementation of ViTs, we defined a further model, called Speaker VGG CCT end-to-end Speaker Token. This model differs from the previous architecture in the classification methodology. In this case, there is no pooling layer.
Table 1: Main characteristics of the datasets used in the experiments. Labels: Anger (A), Boredom (B), Disgust (D), Excitement (E), Fear (F), Happiness (H), Neutral (N), Sadness (S), Surprise (Sr)

| Dataset     | Language | # subj. | # samp. | labels | Anger | Disgust | Fear | Happy | Sad | Surprise | Neutral |
|-------------|----------|---------|---------|--------|-------|---------|------|-------|-----|----------|---------|
| EmoDB [3]   | German   | 5F / 5M | 454     | A, B, D, F, H, N, S | 127  | 46      | 69   | 71    | 62  | –        | 79      |
| IEMOCAP [4] | English  | 5F / 5M | 10,039  | A, E, H, N, S, Sr | 1,103| –       | –    | 595   | 1,084| 107      | 1,708   |
| RAVDESS [21]| English  | 12F / 12M | 1,248 | A, D, F, H, N, S, Sr | 192  | 192     | 192  | 192   | 192 | 192      | 96      |
| SAVEE [34]  | English  | 4M      | 480     | A, D, F, H, S, Sr, N | 60   | 60      | 60   | 60    | 60  | 60       | 120     |
| EMOVO [7]   | Italian  | 3F / 3M | 588     | A, D, F, H, N, S, Sr | 84   | 84      | 84   | 84    | 84  | 84       | 84      |
| DEMOS [25]  | Italian  | 2F / 2M | 8,236   | A, D, F, H, S, Sr  | 1,477| 1,678   | 1,156| 1,395 | 1,530| 1,000    | –       |

and the last level discriminates the class associated with the sample directly from the output of the Transformer Encoder associated with the speaker embedding. Given the similarity with the ViTs, in which the input corresponding to the output used to classify was with the speaker embedding. Given the similarity with the ViTs, in which the input corresponding to the output used to classify was defined as class token, we denoted the speaker embedding provided by the PCA as a speaker token. In this way, we used this particular vector that somehow represents the dependence between the speaker embedding and all the feature maps extracted from the convolutional blocks of the CCT.

3.2 Speaker Embedding

Our architecture extracts a representation of the speakers from each voice signal with a VGG-Vox network [24], and a subsequent combination of such descriptors with information detected from spectrograms. To evidence this, for each set described in Section 4 and Table 1, first, the embeddings associated with each audio track were generated, thus resulting into vectors of 1,024 elements extracted from the last fully-connected layer of the VGG-Vox network. Then, Multi-dimensional Scaling (MDS) was used to represent the embeddings in a planar domain as illustrated in Figure 2.

![Figure 2: Plots illustrate, for each dataset, the speaker embeddings with 1,024 elements using MDS in a 2D plane, emphasizing the separation for different subjects.](image)

We note the dimension of the speaker embedding vector was larger than the internal size of the used CCTs. In addition, the self-attention module used in each of the proposed architectures requires an equal size of the input features. So, PCA was applied to reduce the speaker embedding size to 384 elements. In this way, the desired size was obtained, while keeping the most discriminatory features among the 1,024. PCA effectively allowed us to reduce the size of the embedding, while keeping its discriminatory capability of the subjects.

4 EXPERIMENTAL VALIDATION

We performed a comprehensive set of experiments to validate the proposed architectures in a real-time cross-corpus scenario. We selected the ResNet50 network pre-trained on Imagenet as the reference model to compare with. ResNet is a state-of-the-art residual architecture for image classification. The computational cost associated to its 50 layers is similar to that of the used CCT-14/7x2.

Datasets. To verify the cross-corpus performance of the proposed methods, we considered some of the most used datasets in SER as summarized in Table 1. All the sets taken into consideration contain examples for the seven archetypal emotions or for a subset of them, except for IEMOCAP and EmoDB, which also include a few examples for the emotional states associated with boredom and excitement. Given this categorization, we restricted SER to four classes, namely, Anger (A), Happiness (H), Sadness (S), and Neutrality (N). In SER, this is often a way to reduce the difficulty of the cross-corpus task, while also reducing the problem to the most important emotions among all the primary ones. Cross-corpus tests for three classes are reported in the supplementary material.

We used the larger datasets for training, while testing was performed on the remaining sets. In particular, we considered EmoDB, EMOVO and SAVEE as test sets for cross-corpus validation, which also allowed us to evaluate the general performance for three languages, i.e., German, Italian and English, respectively. We also note all these datasets include speech tracks of emphasized sentences associated to an emotional state reproduced by actors.

Data Balancing. In SER, it is quite common that audio data include many examples for certain classes and a smaller number of samples for other labels. This generally leads to the problem of class collapse: a model tends to always predict examples with the labels associated with the classes that are more popular in the training dataset. This also occurs for the datasets used in this work, whose imbalance is shown in Table 1. It results that only the EMOVO dataset is perfectly balanced for archetypal emotions, while the others are somehow unbalanced. This problem tends to get even worse when aggregating multiple different datasets in a cross-corpus training procedure.

To address this issue and improve performance, two approaches were considered. In the first one, we implemented an undersampling technique that, for each class, removes a number of randomly selected samples equal to the difference between the samples of
the aforementioned label compared to those associated with the minority class. This allows training the models on a perfectly balanced dataset; however, this significantly reduces the data available for training, thus making more severe the scarcity of data and decreasing the generalization capability of the network.

As an alternative technique, we used a data augmentation procedure. The idea was to first apply this operation to each sample of the training set in order to obtain multiple versions associated with it. Then, for each class with fewer examples, additional samples were added in numbers equal to those missing to reach the cardinality of the majority class. As typical for audio computation, we first considered transformations directly applied in the time domain, specifically designed to increase a voice signal by speeding it up, slowing it down, introducing noise into it, shifting it over time, varying its tone and normalizing it. Another technique widely used for spectrograms derived from audio tracks consists in transforming them by acting directly in frequency. Doing so, it is possible to add vertical or horizontal bars to mask, respectively, certain moments in time rather than particular frequency bands. This allows the models to adapt less to the given training data, thus improving their generalization ability on never observed data. Examples of augmented spectrograms for both the approaches are reported in the supplementary material.

Measures. The following measures have been used in the evaluation, being \( t_p \) and \( t_n \) the true positive and true negative, respectively, and \( p \) and \( n \) positive and negative total:

- **Accuracy**: ratio between the number of correct predictions and the total number of predictions;
- **Unweighted Average Recall (UAR)**: it accounts for unbalancing of the test dataset. For the binary case, it is defined as 
  \[
  UAR = \frac{t_p}{t_p + \frac{1}{\text{classes}}} \cdot \frac{t_n}{t_n + 0.5},
  \]
  where 0.5 is replaced with \( n \) for the multi-class case. Each Recall is weighed in the same way regardless of the number of samples of the respective class. This is considered the metric of reference in SER as the datasets are unbalanced in most of the cases;
- **Macro-F1**: measures the quality of results for binary or multi-class classification tasks. It computes an average of the F1 measurements associated with the various classes, where each of them is defined as 
  \[
  F_1 = \frac{t_p}{t_p + \frac{1}{\text{classes}} + \frac{1}{2} t_f},
  \]
  where \( t_f \) is the total of false predictions.

**Settings**. Each training was performed on a machine with an Nvidia RTX 3060 Ti GPU for a total of 50 epochs with Adam optimizer and a batch size of 32. The learning rate was selected by estimating the best performance obtained from the models on a particular test considered as a reference. Each architecture was trained by varying its learning rate taking as training set all datasets except EmoDB and using the latter as a test. The optimal choice of such parameter was therefore found to be 5.0e^{-4} for the ResNet50 and 5.0e^{-5} for the proposed CCT-based models.

### 4.1 Cross-corpus Results

The two “basic” models (\textit{i.e.}, ResNet50 and CCT-14/7x2) were compared for the four-class task with undersampling for balancing. The IEMOCAP dataset was used for training, while tests were performed on EmoDB, SAVEE and EMOVO. We note that IEMOCAP is the largest set among those at our disposal, whereas the test datasets are those with fewer samples and characterizing three languages (\textit{i.e.}, English, German, Italian). Results shown in Table 2 denote a clear superiority of the CCT-based model compared to the ResNet50. Results also confirm the fact widely reported in the literature, that testing the models on a dataset other than the one used for training causes a noticeable performance degradation. We note that IEMOCAP was used for training, though in the literature it is reported to not score good results when tested intra-corpus; we proceeded in this way because it is one of the largest dataset at our disposal and a similar choice was also proposed in [26].

**Table 2: Cross-corpus results: train on IEMOCAP; test on EmoDB, EMOVO and SAVEE each with four classes (Anger, Happiness, Sadness, Neutrality). Undersampling was used to balance the IEMOCAP labels (best results in bold)**

| Dataset | Models     | Accuracy | UAR   | Macro-F1 |
|---------|------------|----------|-------|----------|
| EmoDB   | Resnet50   | 44.04%   | 34.80%| 28.76%   |
|         | CCT-14/7x2 | 55.84%   | 48.32%| 45.53%   |
| EMOVO   | Resnet50   | 35.12%   | 35.12%| 27.57%   |
|         | CCT-14/7x2 | 37.36%   | 37.36%| 30.94%   |
| SAVEE   | Resnet50   | 27.57%   | 28.87%| 16.86%   |
|         | CCT-14/7x2 | 29.47%   | 31.54%| 20.62%   |

Table 3 reports the comparison between all the proposed models. We still performed a cross-corpus experiment, with four classes and with undersampling for balancing (left values for each table entry). In this case, we used a cross validation protocol, where for each test dataset the models were trained several times, each time excluding the test set and, in round, selecting as validation RAVDESS and the remaining two datasets of test among EmoDB, EMOVO and SAVEE. IEMOCAP and DEMOS that include more samples among the available datasets were used for training. From the table, it can be noted that all the metrics are generally superior to those obtained by training on IEMOCAP only for ResNet50 and for the CCT model. This highlights the fact, already reported in the literature, that cross-corpus training can bring significant benefits in SER. As for the comparison between the proposed architectures, it is evident that the learning of the dependence between speakers and emotion allows us to improve the performance of the models also in a cross-corpus test. In particular, the end-to-end architectures get the best metrics. Among the proposed models, for EmoDB the one without speaker token (which uses instead the pooling layer to classify) achieves the best results, while for the other two sets it was the other architecture to achieve the best results. On average, by focusing on the two most relevant metrics, the speaker token model has obtained a higher UAR, but with a lower value for Macro-F1.

The gap with the state-of-the-art is of about 2%. This gap increases by performing the same experiments but with training carried out on balanced datasets with data augmentation, as shown by the values reported on the right of each entry of the table. The best models are those end-to-end, with a gap between the best performing one (Speaker CCT end-to-end) compared to ResNet50...
of around 6% and 5%, respectively, for accuracy and Macro-F1. Results using augmentation for data balancing are higher for almost every model, demonstrating that using more data is beneficial for improving the methods capability of generalizing at inference time.

With reference to the results illustrated so far, it should be noted the relevance of the used metrics. In fact, it has pointed out the two reference measures are the UAR and the Macro-F1 that can be considered as complementary: the UAR indicates the accuracy obtained on average taking into account the unbalancing of the test datasets, while the Macro-F1 measures the quality of the results. Even for a lower UAR value, if the respective Macro-F1 value is higher, the quality of the results, explained as the concentration of the values along the diagonal of the confusion matrix will be better. See the supplementary material for a confusion-matrix example using the two end-to-end models. Overall, we designated the Speaker CCT end-to-end as our best model because it outperforms the others in terms of Macro-F1, and almost in all the cases for accuracy and UAR. We also point out that the accuracy values on EMODB and SAVEE do not say much because these sets are unbalanced. More relevant are UAR and Macro-F1, this latter indicating the distribution of the predictions between the different classes.

### 4.2 Intra-corpus Results

An intra-corpus experiment was performed on the EmoDB dataset, thus making the approach comparable with other solutions in the literature. We used a Leave-One-Speaker-Out (LOSO) protocol, where iteratively a speaker is used for test and models are trained only on the samples of the rest of the subjects.

Table 4 shows the results for each of the proposed architectures obtained by averaging the values detected for all the speakers. The performance shows a different trend with respect to the cross-corpus experiment, with the models using the speaker embeddings showing lower accuracy and Macro-F1. This can be explained by the fact that, to work properly, such architectures need more speakers than the 10 included in EmoDB. In fact, such methods during the training learn the dependence between the representation of the speakers and their emotions. During inference, the models should be able to match the embedding of the tested speaker with the closest one among those in the training set. Therefore, if the speakers observed during the training are too few, such association can be inaccurate. Notably, CCT and all the proposed variants achieved superior performance compared to RenNet50. This is inasmuch relevant as we propose for the first time the use of ViT in SER.

The table also reports results for the method in [17], which is a reference one for the application of CNN in real-time SER. In that study, authors classified the spectrograms using AlexNet pre-trained on Imagenet. Despite the reported results were of about 80%, we note they used a 5-fold cross-validation protocol, where the trained models were evaluated on audio produced by speakers already viewed by the network during the training, with a significant loss of generality compared to an evaluation performed with the LOSO protocol.

### Table 3: Cross-corpus results: training on IEMOCAP and DEMOS; testing on EmoDB, EMOVO and SAVEE each with four classes (Anger, Happiness, Sadness, Neutrality). For each table entry, results are reported using undersampling / augmentation (us / aug) to balance the class labels (best results in bold)

| Dataset | Models | Accuracy (us / aug) | UAR (us / aug) | Macro-F1 (us / aug) |
|---------|--------|---------------------|----------------|-------------------|
| EmoDB   | Resnet50 | 52.80% / 48.97% | 50.57% / 45.20% | 41.15% / 39.32% |
|         | CCT-14/7x2  | 51.92% / 54.38% | 48.48% / 49.00% | 39.63% / 38.73% |
|         | Speaker CCT | 53.49% / **57.03%** | 47.47% / **52.52%** | 37.15% / **41.78%** |
|         | Speaker CCT end-to-end | 53.19% / 55.16% | 48.21% / 50.29% | 38.30% / **41.78%** |
|         | Speaker CCT end-to-end token | 55.65% / 54.48% | 51.10% / 49.30% | 41.59% / 39.95% |
| EMOVO   | Resnet50 | 46.73% / 46.53% | 46.73% / 46.53% | 39.11% / 43.78% |
|         | CCT-14/7x2  | 48.91% / 49.20% | 48.91% / 49.20% | 40.89% / 47.65% |
|         | Speaker CCT | 50.29% / 46.92% | 50.29% / 46.92% | 41.64% / 41.79% |
|         | Speaker CCT end-to-end | 50.00% / 50.79% | 50.00% / 50.79% | 43.32% / **48.79%** |
|         | Speaker CCT end-to-end token | **51.69%** / 51.39% | **51.69%** / 51.39% | 43.43% / 47.36% |
| SAVEE   | Resnet50 | 32.67% / 31.78% | 37.99% / 35.62% | 32.39% / 31.10% |
|         | CCT-14/7x2  | 35.33% / 33.22% | 41.67% / 36.88% | 34.40% / 32.32% |
|         | Speaker CCT | 32.78% / 33.00% | 39.02% / 37.85% | 31.62% / 31.61% |
|         | Speaker CCT end-to-end | 36.66% / **39.89%** | 42.71% / **45.35%** | 35.91% / **40.03%** |
|         | Speaker CCT end-to-end token | 32.11% / 34.33% | 38.68% / 39.86% | 30.26% / 32.87% |

### Table 4: Intra-corpus results: LOSO protocol on EmoDB using anger, happiness, sadness, neutrality, fear, and surprise.

| Models | Accuracy | UAR | Macro-F1 |
|--------|----------|-----|----------|
| Resnet50 | 67.21% | 63.64% | 60.57% |
| CCT | 74.58% | 70.75% | 68.16% |
| Speaker CCT | 75.00% | 70.56% | 67.48% |
| Speaker CCT end-to-end | 70.18% | 67.04% | 63.86% |
| Speaker CCT end-to-end token | 68.44% | 66.85% | 63.19% |
| AlexNet [17]* | 80.5% | – | – |

*Results taken from the original paper using 5-fold cross validation
CONCLUSIONS

In this paper, we addressed the SER problem in real-time and considering a concrete application scenario where most of the evaluation is performed cross-corpus. The proposed approach has some innovative aspects. First of all, it is the first to apply a ViT model in SER through the CCT variant, also integrating the learning of dependence between speaker embedding and emotion by means of self-attention on its own or in an end-to-end architecture. Furthermore, looking for a solution as general as possible, all proposed methods were tested in a cross-corpus setting as it is rarely the case in SER. Results show evident improvements with respect to ResNet50 taken as state-of-the-art architecture for real-time SER. In particular, the proposed end-to-end architecture turned out to be the best performing one.

ACKNOWLEDGMENTS

This work was done during a research internship of Alessandro Arezzo at QuestIT, Siena, Italy. The authors would like to thank QuestIT for the support.

REFERENCES

[1] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sunq Wook Baik. 2017. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In Int. Conf. on Platform Technology and Service (PlatCon). 1–5. https://doi.org/10.1109/PlatCon.2017.7883728

[2] Suzanne Berke, Ray Wilkinson, and Jane Maxim. 2011. Survey on speech emotion recognition by means of facial and vocal expressions in North American English. PLOS ONE 13, 5 (05 2018), 1–35. https://doi.org/10.1371/journal.pone.0196391

[3] Clément Le Moine and Nicolas Oubin. 2020. Att-HACK: An Expressive Speech Database with Social Attitudes. Speech Prosody 2020 (2020).

[4] Frédo Durand and Simon Kornblith. 2021. CICL: An Emotive Speech Database. In INTERSPEECH.

[5] Lewis Hill. 2021. Escaping the Big Data Paradigm with Compact Transformers. MMAsia ’22, December 13–16, 2022, Tokyo, Japan. 5 CONCLUSIONS

In this paper, we addressed the SER problem in real-time and considering a concrete application scenario where most of the evaluation is performed cross-corpus. The proposed approach has some innovative aspects. First of all, it is the first to apply a ViT model in SER through the CCT variant, also integrating the learning of dependence between speaker embedding and emotion by means of self-attention on its own or in an end-to-end architecture. Furthermore, looking for a solution as general as possible, all proposed methods were tested in a cross-corpus setting as it is rarely the case in SER. Results show evident improvements with respect to ResNet50 taken as state-of-the-art architecture for real-time SER. In particular, the proposed end-to-end architecture turned out to be the best performing one.

ACKNOWLEDGMENTS

This work was done during a research internship of Alessandro Arezzo at QuestIT, Siena, Italy. The authors would like to thank QuestIT for the support.

REFERENCES

[1] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sunq Wook Baik. 2017. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In Int. Conf. on Platform Technology and Service (PlatCon). 1–5. https://doi.org/10.1109/PlatCon.2017.7883728

[2] Suzanne Berke, Ray Wilkinson, and Jane Maxim. 2011. Survey on speech emotion recognition by means of facial and vocal expressions in North American English. PLOS ONE 13, 5 (05 2018), 1–35. https://doi.org/10.1371/journal.pone.0196391

[3] Clément Le Moine and Nicolas Oubin. 2020. Att-HACK: An Expressive Speech Database with Social Attitudes. Speech Prosody 2020 (2020).

[4] Frédo Durand and Simon Kornblith. 2021. CICL: An Emotive Speech Database. In INTERSPEECH.

[5] Lewis Hill. 2021. Escaping the Big Data Paradigm with Compact Transformers. MMAsia ’22, December 13–16, 2022, Tokyo, Japan. 5 CONCLUSIONS

In this paper, we addressed the SER problem in real-time and considering a concrete application scenario where most of the evaluation is performed cross-corpus. The proposed approach has some innovative aspects. First of all, it is the first to apply a ViT model in SER through the CCT variant, also integrating the learning of dependence between speaker embedding and emotion by means of self-attention on its own or in an end-to-end architecture. Furthermore, looking for a solution as general as possible, all proposed methods were tested in a cross-corpus setting as it is rarely the case in SER. Results show evident improvements with respect to ResNet50 taken as state-of-the-art architecture for real-time SER. In particular, the proposed end-to-end architecture turned out to be the best performing one.

ACKNOWLEDGMENTS

This work was done during a research internship of Alessandro Arezzo at QuestIT, Siena, Italy. The authors would like to thank QuestIT for the support.

REFERENCES

[1] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sunq Wook Baik. 2017. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In Int. Conf. on Platform Technology and Service (PlatCon). 1–5. https://doi.org/10.1109/PlatCon.2017.7883728

[2] Suzanne Berke, Ray Wilkinson, and Jane Maxim. 2011. Survey on speech emotion recognition by means of facial and vocal expressions in North American English. PLOS ONE 13, 5 (05 2018), 1–35. https://doi.org/10.1371/journal.pone.0196391

[3] Clément Le Moine and Nicolas Oubin. 2020. Att-HACK: An Expressive Speech Database with Social Attitudes. Speech Prosody 2020 (2020).

[4] Frédo Durand and Simon Kornblith. 2021. CICL: An Emotive Speech Database. In INTERSPEECH.

[5] Lewis Hill. 2021. Escaping the Big Data Paradigm with Compact Transformers. MMAsia ’22, December 13–16, 2022, Tokyo, Japan.