IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 September 2023, accepted 6 October 2023, date of publication 19 October 2023, date of current version 25 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3326071

RESEARCH ARTICLE

Speech Emotion Recognition and Deep Learning: An Extensive Validation Using Convolutional Neural Networks

FRANCESCO ARDAN DAL RÍ[1,2], FABIO CIFARIELLO CIARDI[2], AND NICOLA CONCI[1], (Senior Member, IEEE)

[1]Department of Information Engineering and Computer Science (DISI), University of Trento, Povo, 38122 Trento, Italy
[2]Department of Electronic Music, Conservatory of Music F. A. Bonporti, 38122 Trento, Italy

Corresponding author: Francesco Ardan Dal Rí (francesco.dalri-2@unitn.it)

ABSTRACT The domain of Speech Emotion Recognition (SER) has experienced a tremendous revolution due to the outbreak of deep learning, which has contributed, as in many other research areas, to a significant boost in terms of model accuracy. SER refers to a branch of Human-Computer Interaction (HCI), which deals with recognizing emotional states from human speech. Although being a thriving field of research, SER still poses several non-trivial challenges, mainly due to the lack of shared best practices and high-quality datasets that can make the developed models suitable for their application in real environments. In this paper, we implement a CNN-based model combined with a Convolutional Attention Block, and conduct a series of experiments involving a selection of four English datasets popularly used for SER applications: RAVDESS, TESS, CREMA-D, and IEMOCAP. After testing the proposed pipeline on individual datasets, achieving a mean accuracy of 83%, 100%, 68% and 63% respectively, we perform an extensive cross-validation between common emotional classes belonging to single datasets or combinations of them, with the aim to investigate the generalization abilities of the extracted features.

INDEX TERMS Speech emotion recognition, affective computing, deep learning.

I. INTRODUCTION

Emotions play a central role in human communication: emotional connotations such as facial expressions, body movements, and voice tones largely contribute to the way in which people interact. The possibility of automatically recognizing the emotional state of a subject is of interest for a number of application fields, especially related to Human-Computer Interaction (HCI), and in particular, in areas like robotics [57], mobile services [26], healthcare [6], as well as autonomous/intelligent systems [75]. Being vocal expressions one of our primary forms of interaction, Speech Emotion Recognition (SER) is an area that has constantly received a considerable amount of attention [3], [18], methodologically following the evolution of learning approaches.

Although deep learning has significantly revolutionized the way in which we tackle the research problems at hand, by largely improving the results in most research domains [45], SER still poses a number of non-trivial issues to be addressed.

According to the general theory on the subject, emotions occupy a continuous multidimensional space [51], within which it is possible to identify a series of pertinent areas to define a discrete subdivision [66]. The majority of commonly used datasets adopt this subdivision in terms of discrete emotional classes, labelled according to basic emotions. Literature has found evidence for both universal and culture-specific principles in vocal emotion recognition. Listeners decode emotion-relevant prosodic cues relying on similar inference rules across cultures [53]. However, it is worth noting that a discrete subdivision is prone to different interpretations, due to cultural [31], subjective [17], or linguistic [33] reasons. For instance, on the audio portion

The associate editor coordinating the review of this manuscript and approving it for publication was Alessandro Floris.

of the RAVDESS dataset (see Section IV-A1), the degree of accuracy of human validators, with respect to the emotion that the actors intended to express, is reported to be only 67% [37].

Furthermore, many works (e.g. [9], [20], [35]) indicate a general lack of suitable datasets to tackle emotion recognition; moreover, the content of these datasets is often very specific and difficult to adapt to real-life scenarios, in which a correct recognition of the emotional content from raw audio only has proven to be very difficult [29].

As such, we're interested in investigating the generalization capabilities of a selection of publicly available datasets frequently used in SER: RAVDESS, TESS [48], CREMA-D [16], and IEMOCAP [14]. To this aim, we implement a classification pipeline based on a Convolutional Neural Network (CNN) architecture combined with an Attention module. The proposed solution, which also outperforms the existing state-of-the-art, has been validated on RAVDESS to establish our initial benchmark. Next, maintaining the network architecture unaltered, we define a large set of cross-validation experiments, considering both single datasets and combinations of them, to assess the applicability of the learned models in real-world scenarios.

The paper is structured as follows: in Section II, we provide an overview of the main features and architectures used in SER, with specific regard to CNN-based approaches similar to the one we present in this work. In Section III, we briefly describe our pipeline and the implementation of the model. Then, in Section IV, we introduce the four datasets used for comparison and the conducted experiments; the obtained results are reported in Section V and discussed in Section VI. Concluding remarks are drawn in Section (VII).

II. RELATED WORK

A. SPEECH EMOTION RECOGNITION

The goal of Speech Emotion Recognition (SER) is to investigate how to automatically detect and identify emotional content in human speech audio signals [18].

Although some approaches in the literature operate directly on the raw signal (e.g. [64]), it is more common to exploit different representations, in order to shrink the data dimensionality while preserving the relevant features.

2D time-frequency representations, especially Mel-Spectrogram and MFCCs, are commonly adopted [45], [70], as well as the use of Chromagram, Zero-Crossing Rate, Spectral Contrast, and Tonnetz [19], [28].

Originally, SER tasks were accomplished using hand-crafted features and traditional machine learning, such as SVMs [41], GMMs [27], or HMMs [44]. In the last decade, however, deep learning-based approaches have contributed to this research area, introducing new reference benchmarks, as in the work by Han et al. [24], who firstly applied a single-hidden-layered Deep Neural Network (DNN) with many hidden units to segment-level features (MFCCs, pitch period, harmonics-to-noise ratio, and their respective delta) extracted from the IEMOCAP [14] database.

Several methods and architectures have been tried: Niu et al. [43] extracted features from spectrograms via an AlexNet-based model and passed them to a DPARIP algorithm (a data augmentation technique based on the principle of retinal imaging and convex lens imaging), on six classes from IEMOCAP; Li et al. [34] used a CNN-SSAE on Mel and IMel spectrograms from three separated datasets; Tripathi et al. [65] implemented a ResNet supervised by Focal Loss to address the class imbalance in IEMOCAP; Wang et al. [69] combined a CNN-BiLSTM model with multiple stacked Transformers creating well-defined features clusters in the latent space; Sultana et al. [59] validated a series of CNNs/LSTMs-based architectures troughout multilungual experiments conducted on IEMOCAP and SUBESCO [60] (respectively English and Bangla); Su et al. [58] applied a Graph Attentive GRU to 78-dimensional acoustic descriptors representing four classes from IEMOCAP and MSP-IMPROV [15]; Latif et al. [32] proposed a hybrid architecture composed of Dense blocks and LSTM on spectrograms combining two speech datasets with real environmental noises from DEMAND [62] in order to improve noise robustness; Wu et al. [72] utilized Capsule Network along with recurrent connections also on IEMOCAP; Sahu et al. [52] passed a complex space of 1582 features extracted with the OpenSmile toolkit [21] into an Adversarial AutoEncoder; Mohan et al. [40] recently achieved remarkable results with a decision-tree-based ensemble model with a gradient boosting framework (XG Boosting) using only MFCCs as input features.

B. CONVOLUTIONAL NEURAL NETWORKS AND SER

Convolutional Neural Networks (CNNs) have been widely applied in SER, both as the main model or as feature extractors on top of/combined with other architectures (e.g. [1], [11], [55]). Indeed, as mentioned in Section II-A, the general trend in SER is to use 2D representations of audio signals, which contain frequency information over time, for which CNNs proved to be particularly effective.

Badshah et al. [7] compared the results from a freshly and pre-trained/fine-tuned AlexNet on spectrograms extracted from the Berlin dataset [13], which contains stimuli from four speakers over seven emotional classes. The authors achieved an accuracy of 84.3% with the freshly trained model, suggesting that fine-tuning a pre-trained model did not yield satisfactory results.

Issa et al. [28] tried five different CNN-based models over three different datasets (RAVDESS [37], EMODB [13], and IEMOCAP [14]). The models have been fed with a feature vector containing five different representations of the same audio: MFCCs, Mel-scaled Spectrogram, Chromagram, Spectral Contrast, and Tonnetz. The authors achieved top accuracy of 71.61%, 86.1%, and 64.3% on the three datasets, respectively, outperforming most existing models with relatively simple architectures.

Asiya and Kiran [5] also experimented with 1D CNNs trained with multiple features (Zero-Crossing Rate, Mel Spectrogram, Chroma, MFCCs, and Root Mean Square). In order to achieve better results, they applied data augmentation (noise injection, time shifting, pitching, and stretching) on the audio signals before extracting the features. This results in a top accuracy of 68% on RAVDESS, 75% on RAVDESS with gender recognition, and 89% on joint RAVDESS and TESS [48] over eight different emotions.

García-Ordás et al. [22] adopted a slightly different approach, using a Fully-Convolutional Neural Network (FCN) with no dense layers, allowing them to process variable-length audio samples. They trained the model with both Mel-Spectrograms and 100 MFCCs on three datasets separately (RAVDESS, TESS, EMODB). On all three datasets, MFCCs outperformed Mel-Spectrograms, and the authors achieved a mean accuracy of 75.28%, 92.71% and 99.03%, respectively, over five cross-validation folds.

Xu et al. [73] developed a CNN-based model, which concatenates horizontal and vertical features in the first layer using two different kernel sizes (10×2 and 2×8), plus an attention mechanism before the fully connected layer. By using noise injection as data augmentation before extracting MFCCs, they achieved a weighted accuracy of 71.18% on IEMOCAP and 77.8% on RAVDESS.

III. METHODOLOGY

A. PREPROCESSING

The human auditory system can perceive frequencies ranging from ∼20 Hz to ∼20 kHz, and our ears are traditionally considered particularly sensitive in the range between ∼100 Hz and ∼4.5 kHz [56]. Therefore, higher-frequency components are often considered redundant in speech signals, and it is a common practice to filter out such components. Thus, we apply an 8th-order Chebyshev filter [49] at $sr/2$ in order to avoid aliasing, and downsample the audio signals to 16 kHz, so as to discard the spectral components over 8 kHz. Furthermore, in order to feed the audio files to the neural network, trimming/padding is necessary, making sure that the files are all of the same length. We fix their length to 3.5 s.

B. DATA AUGMENTATION

Considering the limited size of the available datasets used for the experiments, data augmentation is applied to the signals in the training splits of the datasets (see IV-B). Although more sophisticated techniques could have been applied, for the purpose of this work we prefer to only consider basic transformations, so as to not compromise the feature space significantly. As such, we implement the following augmentations:

- time shift: the signal is shifted along the x-axis in the range $\pm$ 350 ms;
- noise injection: a white noise with absolute amplitude in the range 0-0.2 is added to the signal.

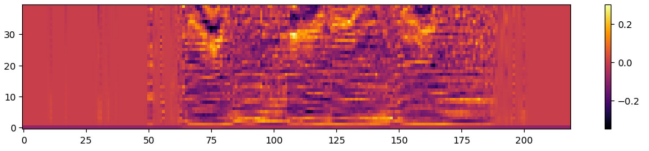

FIGURE 1. An example of the extraction of MFCCs (referred to RAVDESS/07-01-02-02-01.wav, padded at 3.5 s).

At each iteration and for each file, a single augmentation along with a random value in the corresponding range is randomly chosen and applied to the input signals, before extracting the features.

C. FEATURE EXTRACTION

The literature has shown how Mel-Frequency Cepstral Coefficients (MFCCs) are considered the most robust features for speech recognition and proved to be efficient for a variety of tasks, such as speech recognition, speech enhancement, and blind source separation [54]. As such, they have also been successfully exploited in SER [12], [42], [45]. In order to extract the MFCCs, the input signal is framed using a window size $ws = 512$ and hop size $hs = 256$, resulting in chunks of ∼32 ms with a ∼16 ms overlap. This is in line with the recent SER literature, which tends to prefer a more detailed representation over higher framing values adopted in most other speech applications (e.g. [50]). Although the most representative MFCCs for audio are generally the first 10 [63], and considering that most libraries for cepstral coefficients extraction use by default 13-20 coefficients, the literature has demonstrated that using a larger set can be beneficial [22], [23]. Although Patni et al. [47] reported that using more than 16 coefficients not only was redundant but even decreased the accuracy performance of their system, we have not observed this behaviour throughout our experiments. Therefore, in line with the studies proposed in [10] and [46], in our implementation, we extract 40 MFCCs (Figure 1).

As such, the resulting input tensor representing each sample has a shape of $[1 \times 40 \times 218]$.

D. BASELINE MODEL

The model architecture is summarized in Figure 2 and consists of three main components: a set of convolutional blocks, an attention module, and a linear classifier. Each Convolutional Block (CB) is composed of a 2D convolutional layer and batch normalization, followed by an activation function. Given the small number of samples in the datasets (see Section IV-A), we chose a GELU activation function [25] in order to smoothly regularize the output.

The input is passed to the first CB, which aims to shrink the input dimensionality along the x-axis with learnable weights. The output is then passed to two parallel CBs with larger $[5 \times 5]$ kernels, one of which has a dilation factor $d = 2$, whose purpose is to learn contextual features. The resulting outputs are concatenated along the channel dimension and

passed to another CB, followed by an Average Pooling with $[2 \times 3]$ kernel, which returns a $[9 \times 9]$ spatial representation of the learned features. At this point, a Convolutional Block Attention Module (CBAM) [71] infers attention maps along channel/spatial dimensions, and multiplies them with the previous feature map. Then, another CB along with a Max Pooling further reorganize the features, while the last CB shrinks their dimensionality. The latent representation is then sent to the classifier, composed of a Flatten layer and two Fully Connected layers, each of them composed of a Linear Layer with ReLU activation and a Dropout layer with $p = 0.4$. The resulting model is therefore quite portable, with a total of 1,409,966 trainable parameters.

IV. EXPERIMENTAL SETUP

A. DATASETS

For the evaluation of our methodology and datasets cross-validation, we rely on four different datasets, popularly used for SER applications: RAVDESS, TESS, CREMA-D, and IEMOCAP. These datasets share similar properties: all utterances are in English, with a mean duration of ∼3-4 s. Here we report a detailed description of the datasets.

1) RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [37] dataset, which constitutes our main baseline, has become a consolidated standard for SER applications. The dataset is composed of 7356 audio-video files of 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements ("*Kids are talking by the door*" and "*Dogs are sitting by the door*") in a *neutral* North American accent. The speech part contains 1440 utterances and eight emotion classes, namely *neutral*, *calm*, *happiness*, *sadness*, *anger*, *fear*, *surprise*, and *disgust*. The files are in.wav format, mono, and sampled at 48 kHz, 16-bit.

2) TESS

The Toronto Emotional Speech Set (TESS) [48] is another popular dataset used for SER tasks. It is composed of 2800 files spoken by two actresses aged 26 and 64 years, distributed over seven emotion classes: *anger*, *disgust*, *fear*, *happiness*, *surprise*, *sadness*, and *neutral*. Each class contains 200 stimuli by each actress (400 in total), with 200 different target words spoken in the carrier phrase "*Say the word …*". The files are in.wav format, mono, and sampled at 24.414 kHz, 16-bit.

3) CREMA-D

The Crowd-source Emotional Multimodal Actors Dataset (CREMA-D) [16] is an audio-visual dataset containing 7442 stimuli from 91 actors, 48 males and 43 females from different ethnicities, aged between 20 and 74 years. The dataset encompasses six emotional classes, namely *anger*, *disgust*, *fear*, *happiness*, *neutral* and *sadness*, each of which contains spoke from a selection of 12 sentences with four different emotion levels. The files are in.wav format, mono, and sampled at 48 kHz, 16-bit.

TABLE 1. Samples per class for each of the four datasets considered.

	RAVDESS	TESS	CREMA-D	IEMOCAP
Anger	192	400	1271	1477
Happiness	192	400	1271	2301
Sadness	192	400	1271	1759
Neutral	96	400	1087	2127
Disgust	192	400	1271	//
Fear	192	400	1271	//
Surprise	192	400	//	//
Calm	192	//	//	//

4) IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [14] is a multimodal dataset containing ∼12h of audiovisual data (video, speech, motion capture of face, text transcriptions), subdivided into 5 sessions, in which two different actors (male and female) perform both improvised and scripted dialogues. These are split into sentences, and their emotional content is categorically annotated by multiple evaluators. The resulting dataset is then composed of samples over nine emotional labels: *anger*, *happiness*, *excitement*, *sadness*, *frustration*, *fear*, *surprise*, *neutral*, and *other*. The distribution of classes is quite unbalanced, and the evaluations do not always agree: therefore, according to existing literature such as [36] and [73], it is a common trend to (1) drop the poorly represented labels, (2) merge *happiness* with *excitement* (due to their prosodic similarities), and (3) only consider the samples equally labelled by half or more evaluators. It is worth noting that the lengths of the utterances in IEMOCAP are also unbalanced, with respect to the other datasets considered. As such, similarly to [73], we preprocessed the dataset by slicing long samples in smaller segments of 3.5 s with 0.5 s overlapping, keeping their original label. After filtering, we obtained a database containing 7664 samples over four emotional classes: *anger*, *happiness*, *sadness*, and *neutral*. The files are in.wav format, mono, and sampled at 22.05 kHz, 16-bit.

The resulting number of samples across different emotional classes with respect to each dataset (possibly after filtering/merging/slicing) are reported in Table 1.

B. VALIDATION PIPELINE

In our work, we consider four different sets of experiments:

- We first evaluate our model on RAVDESS (8 classes) to establish an initial benchmark;
- We evaluate our model on the other three datasets in their entirety for reference purposes;
- We perform a number of cross-experiments involving RAVDESS, TESS and CREMA-D (common 6 classes) for further analysis;

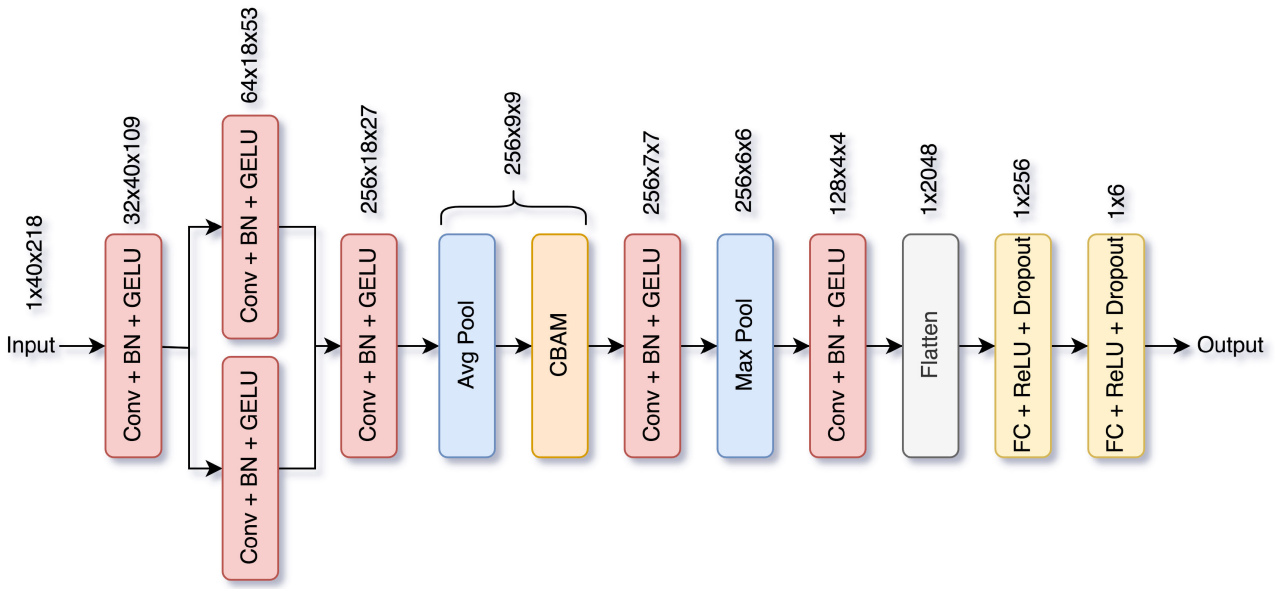

FIGURE 2. The model architecture. On top of each block, the respective output dimensionality is shown. The colour coding denotes the Convolutional layers (including Batch Normalization and GELU activation) in red, the pooling layers in blue, the Convolutional Block Attention Module (CBAM) in orange, the Flatten layer in grey, and the Fully Connected layers (including ReLu activation and Dropout) in yellow. The output of the whole network consists of the probability distribution for each class.

- We evaluate the generalization abilities of the proposed solution with respect to IEMOCAP and vice-versa, considering only the 4 common classes of the entire corpora, and fine-tuning each model on the respective dataset.

The entire pipeline is implemented in PyTorch. We choose for training the Categorical Cross-Entropy Loss, optimized with Adam [30] with a learning rate $lr = 1e^{-3}$ and a batch size $bs = 24$, similarly to previous works such as [59] and [73]. Due to the relatively small number of samples in the datasets (see Section IV-A), and in order to avoid overfitting, we introduce two regularization mechanisms: (1) an early-stopping callback function with the maximum number of epochs to 500 and a patience value $P = 10$, and (2) a learning rate scheduler, with a minimum value of $1e^{-5}$, a multiplying factor of $m = 0.5$ and a patience value $P = 4$. The relatively high value of P is chosen because of the observed fluctuations in the validation loss curves (see Section VI and Figure 5).

The datasets are subdivided considering a random 80%/10%/10% train/validation/test splits, using a 5-fold validation for training and 3-fold for fine-tuning. In the case of experiments with multiple datasets, we split each of them before concatenation, making sure to keep the same percentage of each dataset. We are aware that some previous works (e.g. [4]) suggest testing the models in a speaker/sentence-independent fashion, but this approach does not apply to our scenarios (for instance, RAVDESS having only two sentences, TESS only two speakers). Moreover, being our work primarily focused towards cross-validating these datasets, we argue that a random split is a fair option for our purposes.

Although minor improvements can be achieved on the single datasets by tuning the network parameters ad-hoc, for the sake of consistency, all hyperparameters have been maintained unaltered throughout all the experiments. The entire pipeline, including sample loading, augmentations and features extraction, has been implemented in CUDA using two NVIDIA RTX3090 GPUs. A complete training cycle lasts approximately between ∼5 and ∼9 minutes, depending on the dataset considered and on the number of epochs, while a single batch of 24 samples takes on average ∼5 ms. The inference forward pass for a single sample takes ∼1.5 ms.

Further details of the experiments and the achieved numerical results are reported in the next section.

V. RESULTS

In this section, we present the results of the experiments conducted in our research, following the procedure discussed in Section IV-B, along with a short commentary. For conciseness, the confusion matrices and loss curves are always relative to the best fold, while in the tables, mean values along with the respective standard deviations are reported.

A. BASELINE MODEL ON RAVDESS

To verify the effectiveness of our pipeline, we first experiment with RAVDESS, which among the selected datasets is the one with the largest number of emotional classes.

The model is trained for an average of 84 epochs, achieving a mean weighted accuracy (WA) of 82.97% ($std = 2.49\%$, $top = 86.04\%$) and a mean unweighted accuracy (UA) of 82.38% ($std = 2.35\%$, $top = 84.92\%$). Overall, the model proved robust in correctly identifying especially the

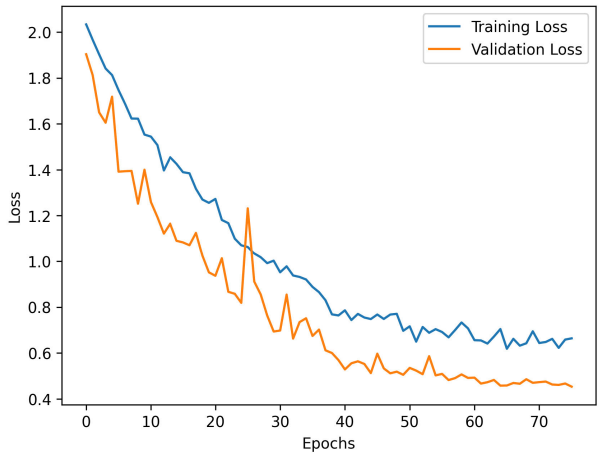

FIGURE 3. Training and validation losses on RAVDESS. In our experiments, the model tends to overfit the training set. Therefore, early-stopping helps with generalization. Despite a few jumps in the early stages of the training, probably due to the small number of samples in the dataset and to the high initial learning rate, we observe a certain stabilization in the curves after ∼40-50 epochs.

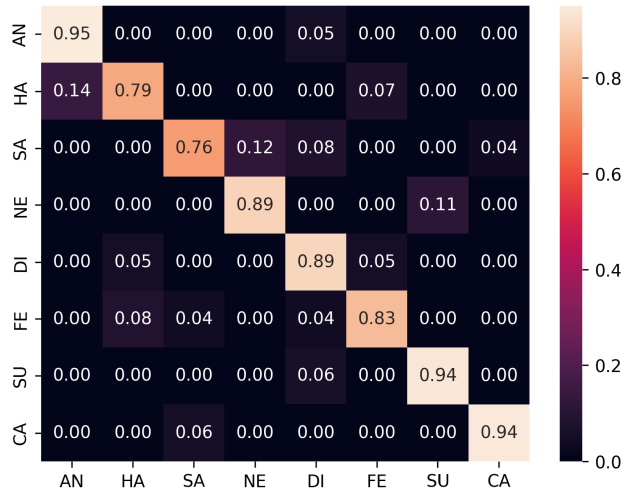

FIGURE 4. Confusion matrix related to the baseline model trained on the entire RAVDESS dataset. A similar behaviour has been observed in all folds. KEYS: AN = *anger*, FE = *fear*, HA = *happiness*, DI = *disgust*, SA = *sadness*, SU = *surprise*, NE = *neutral*, CA = *calm*.

emotional classes of *anger* and *neutral*; the architecture tends instead to misclassify samples belonging to *calm* and *surprise*, incorrectly labelling them as *neutral* and *happiness*, respectively. We argue that these results are justified by the prosodic similarities between those classes.

Our approach outperformed the state-of-the-art regarding similar architectures (see Section II-B), as summarized in Table 2. Figure 3 and Figure 4 show representative training/validation loss curves and confusion matrix related to the best fold.

B. REFERENCE ON TESS, CREMA-D, IEMOCAP

We now provide the results obtained by appliying the proposes to the other datasets considered.

TABLE 2. Accuracy comparison with existing CNN-based models on the RAVDESS dataset.

Method	Year	WA%	UA%
Baseline Vanilla AlexNet	//	67.68	64.9
Issa et al. [28]	2020	71.61	//
Asiya & Kiran [5]	2021	75	74
García-Ordás et al. [22]	2021	75.28	//
Xu et al. [73]	2021	77.8	77.4
Ours	**2023**	**82.97**	**82.38**

TABLE 3. Accuracy results of the model trained on the four datasets individually and in their entirety.

		WA%		UA%	
Dataset	**Classes**	**Mean**	**Top**	**Mean**	**Top**
RAVDESS	8	82.97±2.49	86.04	82.38±2.35	84.92
TESS	7	100.0±0.0	100	100.0±0.0	100
CREMA-D	6	68.3±1.63	70.3	68.22±1.62	70.03
IEMOCAP	4	63.18±1.32	64.74	64.5±1.21	65.02

With TESS (7 classes), the model is trained for an average of 52 epochs, achieving a mean WA of 100.0% ($std = 0.0\%$, $top = 100.0\%$) and a mean UA of 100.0% ($std = 0.0\%$, $top = 100.0\%$).

With CREMA-D (6 classes), the model is trained for an average of 82 epochs, achieving a mean WA of 68.3% ($std = 1.63\%$, $top = 70.3\%$) and a mean UA of 68.22% ($std = 1.62\%$, $top = 70.03\%$).

With IEMOCAP (4 classes), the model is trained for an average of 47 epochs, achieving a mean WA of 63.18% ($std = 1.32\%$, $top = 64.74\%$) and a mean UA of 64.5% ($std = 1.21\%$, $top = 65.02\%$).

The obtained results are reported in detail in Table 3, while Figure 5 and Figure 6 show the respective loss curves and the confusion matrices of the three datasets.

C. CROSS EXPERIMENTS ON RAVDESS, TESS, CREMA-D

In this set of experiments, we only consider the six common classes of the three datasets involved: *anger*, *disgust*, *fear*, *happiness*, *neutral* and *sadness*.

First, we train the model on single datasets, using the remaining ones as test sets. With these experiments, we want to verify if the features extracted from the datasets are somehow comparable. The results obtained in this phase are in line with our expectations: indeed, the highest WA is 37%, reached by the model trained on RAVDESS and tested on TESS, and the worst is 18%, reached by the model trained on TESS and tested on RAVDESS. Overall, the average accuracy values barely surpasses 25%.

Then, we repeat the experiment using combinations of two datasets in the training phase. The achieved results are similar to the previous tests. The model trained jointly on RAVDESS+CREMA-D even decreases the accuracy on TESS with respect to the models trained separately on RAVDESS and CREMA-D.

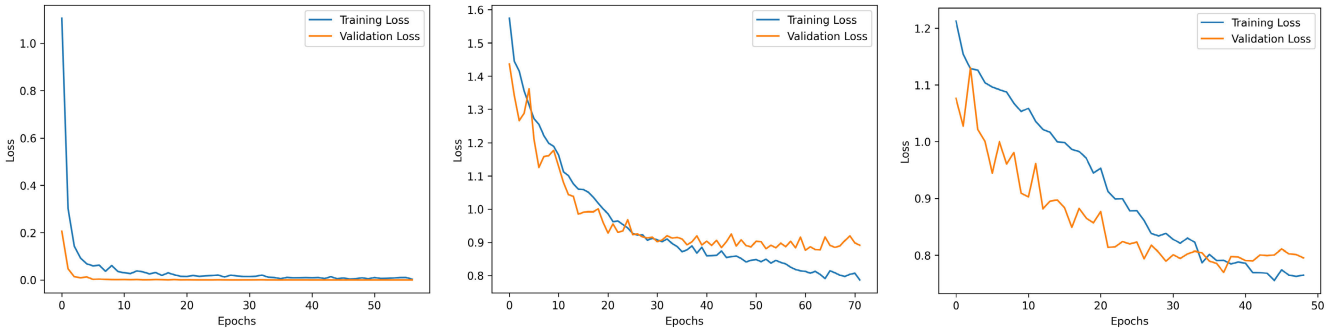

FIGURE 5. Loss curves of (left-to-right) TESS; CREMA-D; and IEMOCAP. The curves show that TESS tends to converge almost immediately, and after just ∼10 epochs the improvement is minimal. The training loss with CREMA-D is still quite smooth, however, the validation loss seems to reach a plateau after ∼40-50 epochs. Finally, the validation loss with IEMOCAP shows clear jumps, while the decrease in the training loss is overall linear. This suggests that the model could possibly still improve on IEMOCAP, however at the expense of its generalization capabilities.

FIGURE 6. Confusion matrices of (left-to-right) TESS; CREMA-D; and IEMOCAP. Every sample in TESS has been classified correctly. In CREMA-D, *disgust* is occasionally predicted as *sadness*. Overall, *sadness* is the better-classified emotion, while *anger* and *fear* always yielding the lowest and highest number of false positives respectively. In IEMOCAP, the classification is slightly different from fold to fold, but overall, *anger* still is the less ambiguous class. KEYS: AN = *anger*, HA = *happiness*, SA = *sadness*, NE = *neutral*, DI = *disgust*, FE = *fear*, SU = *surprise*.

TABLE 4. Accuracy values of the cross-validation experiments conducted on RAVDESS, TESS and CREMA-D (6 classes). When an entire dataset is used for testing in a single pass, the standard deviation is not provided.

Train	Test	WA%	UA%	Fine-Tuning WA%	Fine-Tuning UA%
RAVDESS	TESS	36.57	36.55	100±0.0	100±0.0
	CREMA-D	24.97	24.81	44.62±0.3	45.2±0.51
TESS	RAVDESS	18.21	16.59	42.71±2.18	37.8±3.76
	CREMA-D	25.86	25.49	46.37±0.92	47.02±1.04
CREMA-D	RAVDESS	25.88	24.22	32.30±3.08	31.42±2.87
	TESS	34.21	34.31	99.74±0.04	98.85±0.04
RAVDESS+TESS	RAVDESS+TESS	93.81±2.34	93.66±1.45	//	//
	CREMA-D	25.12	24.58	46.3±0.83	46.08±1.33
RAVDESS+CREMA-D	RAVDESS+CREMA-D	66.78±0.96	66.84±1.45	//	//
	TESS	29.28	29.29	99.78±0.04	99.6±0.03
TESS+CREMA-D	TESS+CREMA-D	75.66±0.5	77.0±1.28	//	//
	RAVDESS	21.55%	19.78%	39.46±2.99	36±3.13
RAVDESS+TESS+CREMA-D	RAVDESS	76.62±6.68	75.62±7.14	//	//
	TESS	99.83±0.22	99.83±0.22	//	//
	CREMA-D	64.75±1.05	64.76±0.9	//	//

These has motivated us to add some additional experiments, applying a fine-tuning procedure on the trained models. This resulted in an improvement in the accuracy of CREMA-D (from 26% up to 46% with the model trained on TESS). With RAVDESS, instead, the accuracy after fine-tuning tends to improve much less and settle on accuracy values between 32-43%. The behaviour of TESS is still different, as its accuracy after fine-tuning improves in all

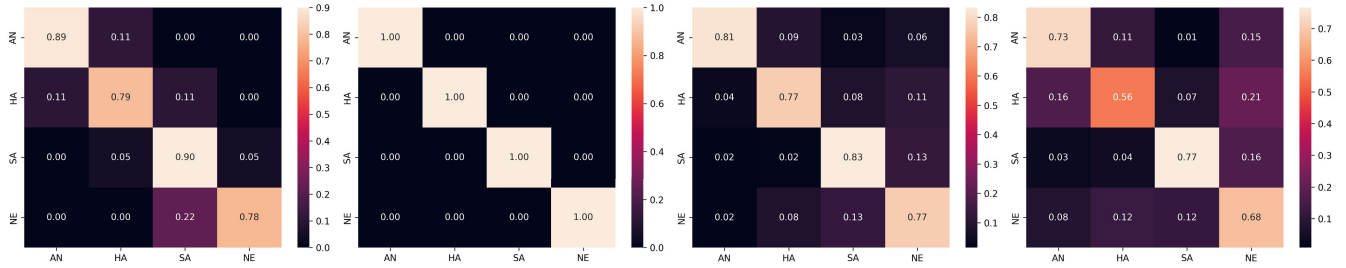

FIGURE 7. Confusion matrices of (left-to-right) RAVDESS; TESS; CREMA-D; and IEMOCAP (4 classes) with respect to the best fold of each individual dataset. KEYS: AN = *anger*, HA = *happiness*, SA = *sadness*, NE = *neutral*.

cases until it reaches levels close to 100%. This suggests that the internal features of each dataset are quite peculiar, and models trained on some of them struggle to adapt to the others, the only exception being TESS, which always reaches top accuracies. Further elaborating on these results, we finally train and test our model with all three datasets. We observe similar accuracy values with respect to the model trained on single datasets (see Table 3), with the sole exception of RAVDESS, penalized by the smaller number of samples and by the dropping of the *surprise* and *calm* classes. The corresponding results are summarized in Table 4.

D. FINE-TUNING USING IEMOCAP

We finally propose two additional sets of experiments, both involving IEMOCAP. We argue that this dataset is the most challenging in our selection: despite having only four classes, samples are extremely varied in terms of sentence length, background noises, and number of different words. Therefore, we consider this dataset rather disjoint from the other three considered, as it is closely resembles a real-life scenario.

According to the outcomes shown in the previous subsections, we conduct cross-experiments by training the model on the four common classes (*anger*, *happiness*, *neutral* and *sadness*). We train the model with RAVDESS, TESS and CREMA-D and perform fine-tuning on IEMOCAP, and vice-versa. Interestingly, the process of fine-tuning with IEMOCAP yields more stable results, with an average accuracy of ∼50% regardless of the dataset used for training. The reverse process (training on IEMOCAP and fine-tuning on the others) also improves the performances with respect to the single datasets. The results of these experiments are shown in Table 5.

Lastly, we train the model on the entire datasets corpora. Besides TESS, which again reaches 100%, with CREMA-D we obtain a mean accuracy of 76% (higher than the model trained on just itself), observe a slight improvement with IEMOCAP, and RAVDESS still achieved a mean accuracy of 77%, despite consisting just of the 4.5% of the entire corpora (only 672 samples from a total of 14836).

The results are summarized in Table 6, while individual confusion matrices are shown in Figure 7.

TABLE 5. Results of the cross-experiments with the model trained on RAVDESS, TESS and CREMA-D and fine-tuned on IEMOCAP, and vice-versa.

Train	Fine Tuning	WA%	UA%
RAVDESS	IEMOCAP	44.02±1.45	43.32±3.52
TESS		51.01±2.02	50.96±1.77
CREMA-D		53.3±2.97	52.26±5.12
IEMOCAP	RAVDESS	50.0±1.65	48.01±4.75
	TESS	100.0±0.0	100.0±0.00
	CREMA-D	61.93±2.62	61.64±3.09

TABLE 6. Results of the model trained on the entire corpora of the selected datasets an tested on the individual test splits.

Train	Test	WA%	UA%
ALL DATASETS	RAVDESS	77.45±3.06	77.03±1.87
	TESS	100.0±0.0	100.0±0.0
	CREMA-D	76.4±2.24	76.16±2.33
	IEMOCAP	65.15±2.18	65.94±2.12

VI. DISCUSSION

In this section, we provide a short commentary on the results obtained in our experimental validation.

The purpose of the first two sets of experiments (see Section IV-B) was to verify the viability of our pipeline, testing the developed model on all datasets separately. As for RAVDESS, our approach outperformed the previous literature adopting similar architectures, with a mean WA of 83% (see Section V-A). Also with TESS, CREMA-D and IEMOCAP, our pipeline proved to be robust in correctly classifying the emotions, and the achieved results are in line with the state-of-the-art, such as [39] and [74], showing a mean WA of respectively 100%, 68% and 64% (see Section V-B).

The third and fourth sets of experiments constitutes, in our view, the main contribution of our work, as we are interested in investigating the compatibility of the features extracted from the different datasets. Indeed, many previous works have used two or more concatenated datasets to train neural networks (e.g. [38], [61], [67]), or extensively validated different approaches on the same dataset (e.g. [4]). However, to the best of our knowledge, none of them has investigated such cross-compatibility, which, in our opinion, is crucial in order to deploy these learning models in a real-life scenario.

FIGURE 8. PCA performed on the features extracted by the model on RAVDESS (top-left); TESS (top-right); CREMA-D (bottom-left); and IEMOCAP (bottom-right). The colour code has been maintained unaltered for the common classes. TESS exhibits well-defined feature islands, and also in RAVDESS they maintain an overall clear segmentation. In CREMA-D and IEMOCAP, although it is still possible to observe a certain clustering, features tend to overlap. KEYS: AN = *anger*, HA = *happiness*, SA = *sadness*, NE = *neutral*, DI = *disgust*, FE = *fear*, SU = *surprise*, CA = *calm*.

As such, a series of cross-validation experiments have been performed over both single datasets and combinations of them. The results summarized in Tables 4 and 5 show that none of the models trained on specific datasets or combinations of them is immediately able to correctly classify samples belonging to other datasets. Fine-tuning, however, improves the models' performance overall. In particular, we always obtained a significant increase in accuracy after fine-tuning on TESS. This is also in line with the literature, which reports how models trained on TESS tend to easily score almost 100% in accuracy with performances generally superior to other datasets (e.g. [2], [8], [76]).

To better understand the motivations of such behaviour, we perform individual Principal Component Analysis (PCA) on the features extracted by our models trained on the four datasets used, as shown in Figure 8. The dimensionality reduction is applied to the feature vector after the flattening layer, of shape [1 × 2048]. It is possible to note how the RAVDESS (a) and IEMOCAP (d) datapoints are quite homogeneously distributed within the feature space. However, the eight classes in RAVDESS are overall more clustered with respect to the just four in IEMOCAP, and therefore, it is reasonable that the model struggles in achieving good performance on the latter. The increase in accuracy after fine-tuning with CREMA-D (b) appears to be justified by a more clustered, albeit overlapping, feature map. PCA also revealed that TESS (c) presents well-defined feature islands, which explains the ease of classification of this dataset. We support the fact that the presence of stimuli produced by only two actresses makes the tonal representation obtained through MFCCs much clearer and more focused on the different emotional classes, making the classification process easier. To further confirm the limited complexity of TESS, we also performed t-Distributed Stochastic Neighbor Embedding (t-SNE) [68] on the features extracted by our model. The outcome of this analysis, shown in Figure 9, highlights not only a clear subdivision between emotional classes, but also two separate and well-defined groups for each class, referring to the two speakers.

As a final consideration, although it is evident that the general lack of suitable datasets still hinders a robust and effective application of SER systems in a real-life scenario (as also expressed in [9], [20], and [35]), the model trained on three/four datasets returns an overall acceptable accuracy and does not seem to penalize certain datasets over others (see Table 4 and Table 6). As such, we suggest that a concatenation of datasets is a good starting point to promote the adoption of such systems in real-world contexts.

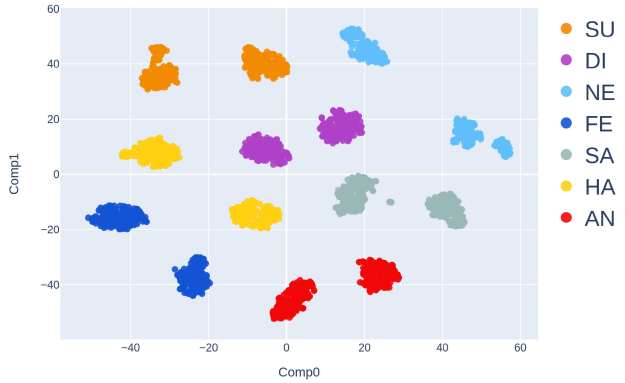

FIGURE 9. 2D t-SNE performed on the features extracted from TESS. The two separate feature islands for each emotional class reveal that the model, besides being able of effectively classify the different emotions, has also learned the different vocal characteristics of the two speakers. KEYS: AN = *anger*, HA = *happiness*, SA = *sadness*, NE = *neutral*, DI = *disgust*, FE = *fear*, SU = *surprise*.

VII. CONCLUSION AND FUTURE WORK

In this paper, we proposed extensive validation of four popular and publicly available English datasets used in SER (RAVDESS, TESS, CREMA-D and IEMOCAP), with the aim of verifying their cross-generalization capabilities. Following a well-established pipeline, we rely solely on a CNN-based model, extracting MFCCs as input features.

The preliminary validation on RAVDESS demonstrated the viability of the proposed solution, outperforming the existing state-of-the-art. The results achieved on the other three datasets are also in line with the existing literature. Despite the model achieving good results on the test split of the datasets (or combinations of datasets) with which it was trained, the lower accuracies obtained by testing it on the remaining ones demonstrates the peculiarity of the respective stimuli and, consequently, of the extracted features. The different degree of accuracy with respect to each dataset is also confirmed by the outcome of the PCA. However, fine-tuning on target datasets always returned noticeable improvements in accuracy, showing that the model's classifier is still able to adapt, to a certain extent, to the intrinsic characteristics of new data. Finally, despite the aforementioned specificity of the employed selection of datasets, models trained on multiple datasets proved robust in achieving individual accuracy results similar, or even slightly superior, to the models trained on single datasets. This suggests that SER benefits from a high number of datapoints with diverse characteristics.

As future work, following this comparative approach, we plan to experiment with other datasets, possibly testing the performance of our models with natural dialogues.

REFERENCES

[1] A. A. Abdelhamid, E. M. El-Kenawy, B. Alotaibi, G. M. Amer, M. Y. Abdelkader, A. Ibrahim, and M. M. Eid, "Robust speech emotion recognition using CNN+LSTM based on stochastic fractal search optimization algorithm," *IEEE Access*, vol. 10, pp. 49265–49284, 2022.

[2] M. R. Ahmed, S. Islam, A. K. M. M. Islam, and S. Shatabda, "An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition," *Expert Syst. Appl.*, vol. 218, May 2023, Art. no. 119633.

[3] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, Jan. 2020.

[4] N. Antoniou, A. Katsamanis, T. Giannakopoulos, and S. Narayanan, "Designing and evaluating speech emotion recognition systems: A reality check case study with IEMOCAP," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[5] A. U A and K. V K, "Speech emotion recognition—A deep learning approach," in *Proc. 5th Int. Conf. I-SMAC (IoT Social, Mobile, Analytics Cloud) (I-SMAC)*, Nov. 2021, pp. 867–871.

[6] N. Azam, T. Ahmad, and N. Ul Haq, "Automatic emotion recognition in healthcare data using supervised machine learning," *PeerJ Comput. Sci.*, vol. 7, p. e751, Dec. 2021.

[7] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. Int. Conf. Platform Technol. Service (PlatCon)*, Feb. 2017, pp. 1–5.

[8] M. Bansal, S. Yadav, and D. K. Vishwakarma, "A language-independent speech sentiment analysis using prosodic features," in *Proc. 5th Int. Conf. Comput. Methodolog. Commun. (ICCMC)*, Apr. 2021, pp. 1210–1216.

[9] F. Bao, M. Neumann, and N. T. Vu, "CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition," in *Proc. Interspeech*, Sep. 2019, pp. 2828–2832.

[10] M. Baruah and B. Banerjee, "Speech emotion recognition via generation using an attention-based variational recurrent neural network," in *Proc. Interspeech*, Sep. 2022, pp. 4710–4714.

[11] J. L. Bautista, Y. K. Lee, and H. S. Shin, "Speech emotion recognition based on parallel CNN-attention networks with multi-fold data augmentation," *Electronics*, vol. 11, no. 23, p. 3935, Nov. 2022.

[12] A. Bhavan, P. Chauhan, Hitkul, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowl.-Based Syst.*, vol. 184, Nov. 2019, Art. no. 104886.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Sep. 2005, pp. 1517–1520.

[14] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[15] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan. 2017.

[16] 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. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct. 2014.

[17] N. Ding, V. Sethu, J. Epps, and E. Ambikairajah, "Speaker variability in emotion recognition—An adaptation based approach," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 5101–5104.

[18] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[19] M. B. Er, "A novel approach for classification of speech emotions based on deep and acoustic features," *IEEE Access*, vol. 8, pp. 221640–221653, 2020.

[20] S. E. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5099–5103.

[21] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia*, Oct. 2013, pp. 835–838.

[22] M. T. García-Ordás, H. Alaiz-Moretón, J. A. Benítez-Andrades, I. García-Rodríguez, O. García-Olalla, and C. Benavides, "Sentiment analysis in non-fixed length audios using a fully convolutional neural network," *Proc. Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102946.

[23] U. Garg, S. Agarwal, S. Gupta, R. Dutt, and D. Singh, "Prediction of emotions from the audio speech signals using MFCC, MEL and chroma," in *Proc. 12th Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Sep. 2020, pp. 87–91.

[24] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. Interspeech*, Sep. 2014, pp. 1–5.

[25] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[26] M. S. Hossain and G. Muhammad, "An emotion recognition system for mobile applications," *IEEE Access*, vol. 5, pp. 2281–2287, 2017.

[27] H. Hu, M.-X. Xu, and W. Wu, "GMM supervector based SVM with spectral features for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2007, p. 413.

[28] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101894.

[29] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[30] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015, doi: arXiv:1412.6980ss.

[31] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross lingual speech emotion recognition: Urdu vs. western languages," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, Dec. 2018, pp. 88–93.

[32] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition," 2020, *arXiv:2005.08453*.

[33] P. Laukka, D. Neiberg, and H. A. Elfenbein, "Evidence for cultural dialects in vocal emotion expression: Acoustic classification within and across five nations," *Emotion*, vol. 14, no. 3, p. 445, 2014.

[34] J. Li, X. Zhang, L. Huang, F. Li, S. Duan, and Y. Sun, "Speech emotion recognition using a dual-channel complementary spectrogram and the CNN-SSAE neutral network," *Appl. Sci.*, vol. 12, no. 19, p. 9518, Sep. 2022.

[35] J. Liu, Z. Liu, L. Wang, Y. Gao, L. Guo, and J. Dang, "Temporal attention convolutional network for speech emotion recognition with latent representation," in *Proc. Interspeech*, Oct. 2020, pp. 2337–2341.

[36] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, "Speech emotion recognition with local–global aware deep representation learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7174–7178.

[37] 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, May 2018, Art. no. e0196391.

[38] E. Mishra, A. K. Sharma, M. Bhalotia, and S. Katiyar, "A novel approach to analyse speech emotion using CNN and multilayer perceptron," in *Proc. 2nd Int. Conf. Advance Comput. Innov. Technol. Eng. (ICACITE)*, Apr. 2022, pp. 1157–1161.

[39] B. Mocanu and R. Tapu, "Emotion recognition from raw speech signals using 2D CNN with deep metric learning," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2022, pp. 1–5.

[40] M. Mohan, P. Dhanalakshmi, and R. S. Kumar, "Speech emotion classification using ensemble models with MFCC," *Proc. Comput. Sci.*, vol. 218, pp. 1857–1868, Jan. 2023.

[41] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1057–1070, Jul. 2011.

[42] P. Nantasri, E. Phaisangittisagul, J. Karnjana, S. Boonkla, S. Keerativittayanun, A. Rugchatjaroen, S. Usanavasin, and T. Shinozaki, "A light-weight artificial neural network for speech emotion recognition using average values of MFCCs and their derivatives," in *Proc. 17th Int. Conf. Electr. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON)*, Jun. 2020, pp. 41–44.

[43] Y. Niu, D. Zou, Y. Niu, Z. He, and H. Tan, "Improvement on speech emotion recognition based on deep convolutional neural networks," in *Proc. Int. Conf. Comput. Artif. Intell.*, Mar. 2018, pp. 13–18.

[44] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, Nov. 2003.

[45] S. K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, "Deep learning techniques for speech emotion recognition: A review," in *Proc. 29th Int. Conf. Radioelektronika (RADIOELEKTRONIKA)*, Apr. 2019, pp. 1–6.

[46] S. Patnaik, "Speech emotion recognition by using complex MFCC and deep sequential model," *Multimedia Tools Appl.*, vol. 82, no. 8, pp. 11897–11922, Mar. 2023.

[47] H. Patni, A. Jagtap, V. Bhoyar, and A. Gupta, "Speech emotion recognition using MFCC, GFCC, chromagram and RMSE features," in *Proc. 8th Int. Conf. Signal Process. Integr. Netw. (SPIN)*, Aug. 2021, pp. 892–897.

[48] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," Borealis, Toronto, ON, Canada, V1, 2020.

[49] P. Podder, M. Hasan, M. Islam, and M. Sayeed, "Design and implementation of butterworth, Chebyshev-I and elliptic filter for speech signal analysis," 2020, *arXiv:2002.03130*.

[50] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Proc. Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 103107.

[51] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Personality*, vol. 11, no. 3, pp. 273–294, Sep. 1977.

[52] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," 2018, *arXiv:1806.02146*.

[53] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *J. Cross-Cultural Psychol.*, vol. 32, no. 1, pp. 76–92, Jan. 2001.

[54] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Appl. Acoust.*, vol. 158, Jan. 2020, Art. no. 107020.

[55] A. Slimi, H. Nicolas, and M. Zrigui, "Hybrid time distributed CNN-transformer for speech emotion recognition," in *Proc. 17th Int. Conf. Softw. Technol.*, 2022, pp. 11–13.

[56] W. B. Snow, "Audible frequency ranges of music, speech and noise," *Bell Syst. Tech. J.*, vol. 10, no. 4, pp. 616–627, Oct. 1931.

[57] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human-robot interaction: Recent advances and future perspectives," *Frontiers Robot. AI*, vol. 7, p. 145, Dec. 2020.

[58] B.-H. Su, C.-M. Chang, Y.-S. Lin, and C.-C. Lee, "Improving speech emotion recognition using graph attentive bi-directional gated recurrent unit network," in *Proc. Interspeech*, Oct. 2020, pp. 506–510.

[59] S. Sultana, M. Z. Iqbal, M. R. Selim, M. M. Rashid, and M. S. Rahman, "Bangla speech emotion recognition and cross-lingual study using deep CNN and BLSTM networks," *IEEE Access*, vol. 10, pp. 564–578, 2022.

[60] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "SUST Bangla emotional speech corpus (SUBESCO): An audio-only emotional speech corpus for Bangla," *PLoS ONE*, vol. 16, no. 4, Apr. 2021, Art. no. e0250173.

[61] G. Tamulevičus, G. Korvel, A. B. Yayak, P. Treigys, J. Bernatavičienė, and B. Kostek, "A study of cross-linguistic speech emotion recognition based on 2D feature spaces," *Electronics*, vol. 9, no. 10, p. 1725, Oct. 2020.

[62] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proc. Meetings Acoust. (ICA)*, vol. 19, 2013, Art. no. 035081.

[63] D. Torres-Boza, M. C. Oveneke, F. Wang, D. Jiang, W. Verhelst, and H. Sahli, "Hierarchical sparse coding framework for speech emotion recognition," *Speech Commun.*, vol. 99, pp. 80–89, May 2018.

[64] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5200–5204.

[65] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, "Focal loss based residual convolutional neural network for speech emotion recognition," 2019, *arXiv:1906.05682*.

[66] M. Trnka, S. Darjaa, M. Ritomský, R. Sabo, M. Rusko, M. Schaper, and T. H. Stelkens-Kobsch, "Mapping discrete emotions in the dimensional space: An acoustic approach," *Electronics*, vol. 10, no. 23, p. 2950, Nov. 2021.

[67] S. Ullah, Q. A. Sahib, Faizullah, S. Ullahh, I. U. Haq, and I. Ullah, "Speech emotion recognition using deep neural networks," in *Proc. Int. Conf. IT Ind. Technol. (ICIT)*, Oct. 2022, pp. 1–6.

[68] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[69] X. Wang, M. Wang, W. Qi, W. Su, X. Wang, and H. Zhou, "A novel end-to-end speech emotion recognition network with stacked transformer layers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6289–6293.
[70] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021.
[71] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.
[72] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu, and H. Meng, "Speech emotion recognition using capsule networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6695–6699.
[73] M. Xu, F. Zhang, and W. Zhang, "Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset," *IEEE Access*, vol. 9, pp. 74539–74549, 2021.
[74] A. Yadav and D. K. Vishwakarma, "A multilingual framework of CNN and bi-LSTM for emotion classification," in *Proc. 11th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2020, pp. 1–6.
[75] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, and R. W. Picard, "Driver emotion recognition for intelligent vehicles: A survey," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–30, May 2021.
[76] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor, "Recognition of emotions in speech using convolutional neural networks on different datasets," *Electronics*, vol. 11, no. 22, p. 3831, Nov. 2022.

FRANCESCO ARDAN DAL RÍ received the master's degree in electronic music from the Conservatory of Music F. A. Bonporti. He is currently pursuing the joint Ph.D. degree with the Department of Information Engineering and Computer Science (DISI), University of Trento, and the Conservatory of Music F. A. Bonporti. He is currently a musician. He has a strong background in contemporary music. His current research interests include music-related human–computer interaction to audio analysis/processing using deep learning techniques, with a particular focus on generative audio and neural synthesis.

FABIO CIFARIELLO CIARDI is a tenured Professor of composition with the Conservatory of Music F. A. Bonporti. After the academic studies of composition, electronic music, and musicology, he studied composition with Franco Donatoni at the Accademia S. Cecilia of Rome and Tristan Murail and Philippe Manoury, IRCAM, Paris. He has been a Composer-in-Residence with EMS (Stockholm) and IMEB (Bourges). His compositions have been awarded in several international competitions (e.g., "L. Russolo 1992" Varese, "MusicaNova 1993" Praha, "ICMC Cd selection 1993" Tokyo, "Olympia 1993" Athens, "Spectri Sonori93" Tulane, USA, XXV Concours Int. de Musique électroacoustique 1998 Bourges, France, and Valentino Bucchi 1999 Rome). He has developed algorithms for dissonance calculation, sound spatialization, financial data sonification, and speaking voice transcription for acoustic instruments. He is the author of articles on the sociology and psychology of music, sonification, music analysis, and theory. In 2021, his article *Strategies and Tools for the Sonification of Prosodic Data: A Composer's Perspective* won the Best Paper Award at the 26th International Conference on Auditory Display. He is the Co-Chair of the Dictionary for Multidisciplinary Music Integration.

NICOLA CONCI (Senior Member, IEEE) received the Ph.D. degree from the University of Trento, in 2007. In 2007, he was a Visiting Student with the Image Processing Laboratory, University of California Santa Barbara. From 2008 to 2009, he was a Postdoctoral Researcher with the Multimedia and Vision Research Group, Queen Mary University, London. He is currently an Associate Professor with the Department of Information Engineering and Computer Science, University of Trento. He is the author and coauthor of more than 130 articles in peer-reviewed journals and conferences. His current research interests include related to video analysis and computer vision applications for human behavior understanding, with particular dyadic and group interactions. He has been the Co-Chair of the First and Second International Workshop on Computer Vision for Winter Sports, hosted at IEEE WACV, in 2022 and 2023; the General Co-Chair of the International Conference on Distributed Smart Cameras, in 2019; the Symposium Signal Processing for Understanding Crowd Dynamics, held concurrently with IEEE AVSS in 2017; and the Technical Program Co-Chair of the Symposium Signal Processing for Understanding Crowd Dynamics at IEEE GlobalSip, in 2016.

Open Access funding provided by 'Università degli Studi di Trento' within the CRUI CARE Agreement