ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume **26**, Number 3–4, 2023, 375–387

Speech Emotion Recognition Using Deep Neural Networks, Transfer Learning, and Ensemble Classification Techniques

Serban MIHALACHE[1, 2, *] and Dragos BURILEANU[1]

[1]Speech and Dialogue Research Laboratory, University "Politehnica" of Bucharest, Romania
[2]Research Institute for Artificial Intelligence, Romanian Academy, Bucharest, Romania
E-mails: serban.mihalache@upb.ro*, dragos.burileanu@upb.ro
* Corresponding author

Abstract. *Speech emotion recognition* (SER) is the task of determining the affective content present in speech, a promising research area of great interest in recent years, with important applications especially in the field of forensic speech and law enforcement operations, among others. In this paper, systems based on *deep neural networks* (DNNs) spanning five levels of complexity are proposed, developed, and tested, including systems leveraging *transfer learning* (TL) for the top modern image recognition deep learning models, as well as several ensemble classification techniques that lead to significant performance increases. The systems were tested on the most relevant SER datasets: EMODB, CREMAD, and IEMOCAP, in the context of: (i) classification: using the standard full sets of emotion classes, as well as additional negative emotion subsets relevant for forensic speech applications; and (ii) regression: using the continuously valued 2D arousal-valence affect space. The proposed systems achieved state-of-the-art results for the full class subset for EMODB (up to 83% accuracy) and performance comparable to other published research for the full class subsets for CREMAD and IEMOCAP (up to 55% and 62% accuracy). For the class subsets focusing only on negative affective content, the proposed solutions offered top performance vs. previously published state of the art results.

Key-words: Convolutional neural networks; deep learning; deep neural networks; machine learning; speech emotion recognition; transfer learning.

1. Introduction

Speech emotion recognition (SER) is the research area that attempts to tackle the challenge of detecting and recognizing human emotions using only speech signals (*i.e.*, only audio data and what information can be extracted from it), which may often be the sole available type for various applications, a fact particularly true for forensic speech and law enforcement operations. Although the latter examples are the focus of this work, most SER research focuses on other

simpler, more general applied fields, *e.g.*, human-machine interfaces, virtual assistants, affective speech synthesis, etc. [1]. Specifically, this work approaches the SER task in relation to monitoring suspicious behavior for applications such as computer-aided conducting of interviews or questionings carried out by law enforcement organizations, surveillance, criminal or terrorist act prevention, etc. For the SER systems designed in this work, such applications justify having a particular focus mainly on negative emotions, often with a high intensity component.

When designing SER systems, there are the two main schools of thought in psychology that establish the conceptual modeling of emotions: (i) discrete classes, wherein each emotion (or, rather, each emotion class) is holistically distinguished from the others; and (ii) dimensional models, where a number of continuous psychological measures (*e.g.*, *arousal* and *valence*) form a multidimensional affect space (typically 2D), each emotion being a sub-zone within it.

Promising results have been reported in literature using *machine learning* (ML) and *deep learning* (DL) models and techniques, including *support vector machines* (SVMs) [2], *multilayer perceptron* (MLP) DNNs [3–4], *recurrent neural networks* (RNNs) with *long short term memory* (LSTM) cells [5–6], *convolutional neural networks* (CNNs) or *convolutional-recurrent neural networks* (CRNNs) [7]. In [8], a very large feature set was used, obtained by applying several statistical functions (*e.g.*, mean, variance, the first, second, and third quartiles, etc.) on mathematical descriptors (*e.g.*, the maximum value, the minimum value, frame-to-frame differences, etc.) computed for the log-energy, the estimated pitch, the *Mel-frequency cepstral coefficients* (MFCCs), and their delta and delta-delta coefficients, with the employed models being SVMs. The MFCCs continue to prove to be extremely robust and useful features for many tasks including SER, as demonstrated by their continued usage in state-of-the-art literature. SVMs were also adopted with similar input feature sets in [9] after applying a simple linear threshold classification to determine the gender of the speaker, and in [10] alongside GMMs. Gender data was also leveraged in [11], and data augmentation obtained through an adversarial network has been reported as a successful strategy [12]. In [13], a much smaller feature set comprising only four statistical values for the estimated pitch, the first two formants, the energy, and the *zero-crossing rate* (ZCR) were used together with feed-forward MLP neural networks, but trained with a modified backpropagation algorithm based on *genetic algorithm* (GA) principles, with a focus on negative emotions. A different approach was taken in [14], operating on the raw time-domain audio signal to extract linear prediction descriptors processed through a Gammatone filterbank before being applied to a *spiking neural network* (SNN) and *liquid state machine* (LSM) hybrid model. Feature selection techniques have also proven to be efficient in reducing the inherent noise of the input feature space [15–16]. Hybrid DNNs included using different final classifiers after a CNN feature extractor, such as metric learning-based frameworks [17]. Attention mechanisms were also successfully leveraged in other LSTM or CNN-based models [18–19]. Finally, *transfer learning* (TL) is the idea of adapting pretrained ML systems for tasks different than the ones they were initially designed for, thus leveraging the previously obtained input data space modeling and its corresponding transformations. By not having to fully train DNNs from scratch, the aim is to only fine-tune the deeper layers of the DNN (responsible for the highest-level abstractions) and/or of the top classification head. The TL methodology has been employed in slightly different ways for SER, almost always following the same fundamental framework, *i.e.*, reusing very deep image recognition neural networks [20–25] with time-frequency representations (usually spectrograms) of the audio signal.

In our previous work on SER [1], MLP-based systems were used with small input feature sets, tested on a single dataset. The main contributions of the present work include:

- Developing systems based on *deep neural networks* (DNNs) for SER, spanning five levels of complexity: single DNNs, multiple DNNs connected together following ensemble classification techniques, as well as systems leveraging transfer learning for the top modern image recognition deep learning models, either as standalone TL-DNN models, or as heterogeneous and homogeneous ensemble classifiers.
- Extending the scarce previous research on negative emotion recognition with additional approaches for SER in the context of forensic speech applications, and obtaining improved performance over most of the state-of-the-art literature previously published in the field, validated on three of the most important benchmark datasets.

The rest of the paper is organized as follows. The system architectures proposed for SER are presented in Section 2. Section 3 describes the benchmark speech datasets, as well as the experimental setup and methodology used for the approaches involved. The subsequent experimental results are presented and discussed. In Section 4, overall conclusions are drawn, and the intended future work is considered.

2. Proposed System Architectures

The proposed system architectures for the SER task were iteratively developed, and present increasing levels of complexity, falling within five categories (*approaches*):

1. single *deep neural network* (DNN) models, for classification or regression;
2. ensemble classifiers comprising multiple DNN models;
3. single DNN classification models adapted through *transfer learning* (TL);
4. heterogeneous fusion classification through TL-DNN models;
5. homogeneous ensemble classification using TL-DNN models.

The first two approaches are discussed in Subsection 2.1, with the following three TL-based systems being described in Subsection 2.2.

2.1. Direct approaches

The single DNN model approach is illustrated in Fig. 1. The DNN takes as input an extensive 2,258-dimensional set of acoustic, spectral, and cepstral features that has been used successfully in our previous work [26], in which the preprocessing and feature extraction stages have been discussed in detail. The DNN classifier is a feed-forward *fully-connected neural network* (FCNN) model, using between 1 and 4 hidden layers, with different numbers of nodes per layer, and with an output layer of size equal to either the number of classes applicable or having two neurons, corresponding to the 2D affect space dimensions (*arousal* and *valence*). The same hidden layer node structures were taken into consideration: the 'constant' architecture that consists of hidden layers each with the same number of nodes; and the 'log2dec' architecture, which incorporates progressively fewer nodes per active layer, following a decreasing log2(·) law.

The second, more advanced type of system leverages two ensemble classification techniques for multiclass problems: one-vs.-one (OvO) and one-vs.-rest (OvR). This type of system is illustrated in Fig. 2. For the former case (OvO ensemble classification), a total of $K \cdot (K-1)/2$ classifiers (where K is the number of classes) are trained independently for each pair of classes (*e.g.*, for 7-class problems, 21 classifiers), with their output values representing the probabilities of the sample belonging to each corresponding class pair. For the latter ensemble classification

technique (OvR), only K classifiers are trained independently, in each case grouping together all instances that are not part of the currently considered class, *e.g.*, Anger vs. Non-Anger, with all instances in the dataset that were labeled as any class other than Anger being relabeled as Non-Anger. In both cases, the outputs are then fed, together with their rounded values (0 or 1, the intermediate binary predictions of each DNN), to a similar DNN that performs final classification.

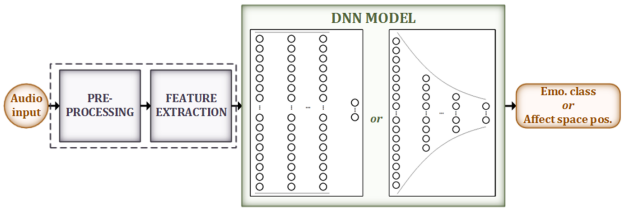

Fig. 1. *Approach 1*: single DNN model for SER.

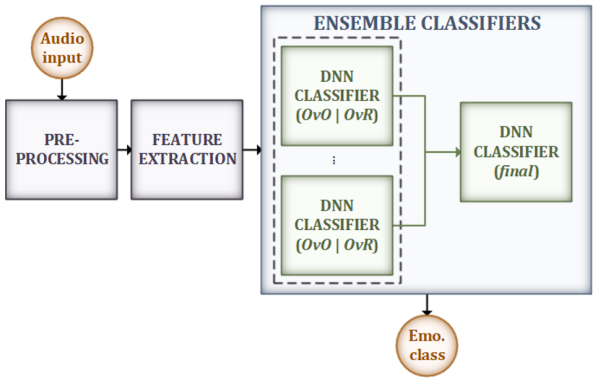

Fig. 2. *Approach 2*: ensemble classification techniques (OvO or OvR) adopted for SER using multiple DNN model classifiers.

2.2. Transfer learning approaches

The modern top-performing image recognition DNNs are: Xception [20], VGG16 and VGG19 [21], ResNet50 and ResNet50V2 [22], InceptionV3 and InceptionResNetV2 [23], NASNet [24], and EfficientNet [25]. A summarized comparison between their sizes, image recognition performance, and particular advantages that justify their choosing is presented in Table 1.

In this work, the first proposed TL-based approach consists of retraining the top layers of each of DNNs listed in Table 1 (*i.e.*, TL-DNNs) in order to develop single TL-DNN classification models. Additionally, the configuration of the classification head is changed empirically through hyperparameter tuning. An illustration of this type of proposed system is made in Fig. 3.

Going further, a form of ensemble information representation through fusion is proposed in the form of a heterogeneous TL-DNN system: the core of each of the TL-DNN models is used to extract a deep feature map representation of the input data instances. All representations are then flattened into a single *feature vector* (FV) that is subsequently fed to a DNN classifier having an FCNN architecture as presented in Subsection 2.1. This is illustrated in Fig. 4. Since the total size

of the FV is large (over 20,000), *principal component analysis* (PCA) and *linear discriminant analysis* (LDA) were also tested for dimensionality reduction, but yielded no improvement.

Table 1. Summarized comparison of the modern top-performing image recognition architectures. Top-1 and Top-5 accuracies refer to the standard model performance metrics reported on the ImageNet dataset.

Model	Accuracy [%]		No. of params.	Depth	Obs.
	Top-1	Top-5			
Xception	79.0	94.5	22.9M	81	High accuracy for medium model size.
VGG16	71.3	90.1	138.4M	16	**Smallest depth.**
VGG19	71.3	90.0	143.7M	19	Small depth, **largest model size.**
ResNet50	74.9	92.1	25.6M	107	Progressive and well-established architecture, medium model size.
ResNet50V2	76.0	93.0	25.6M	103	Improved ResNet50.
InceptionV3	77.9	93.7	23.9M	189	Progressive and well-established architecture, medium model size, very deep.
InceptionResNetV2	80.3	95.3	55.9M	449	Improved hybrid version, extremely deep.
NASNetMobile	74.4	91.9	5.3M	389	High accuracy for **smallest model size**, extremely deep.
NASNetLarge	82.5	96.0	88.9M	533	Very high accuracy, **deepest.**
EfficientNetB0	77.1	93.3	5.3M	132	Base version of highest-performing.
EfficientNetB1	79.1	94.4	7.9M	186	Second version of highest-performing.
EfficientNetB7	84.3	97.0	66.7M	438	**Best version of highest-performing,** extremely deep.

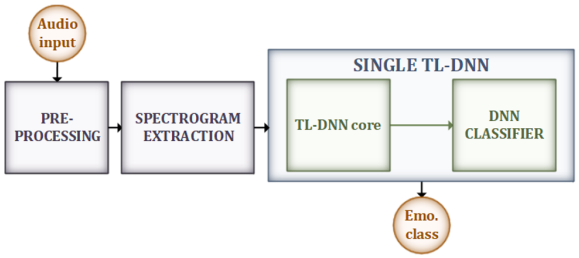

Fig. 3. *Approach 3*: single TL-DNN model for SER.

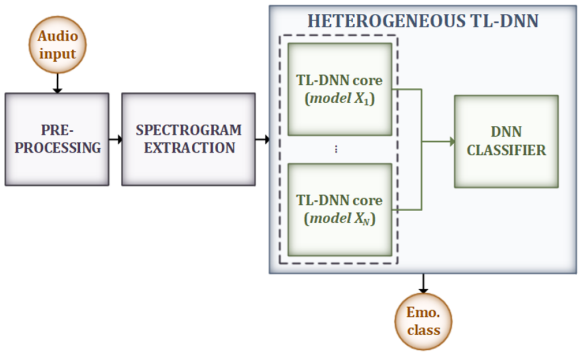

Fig. 4. *Approach 4*: heterogeneous TL-DNN ensemble classification. Each TL-DNN core provides distinct feature map representations of the underlying patterns within the input data.

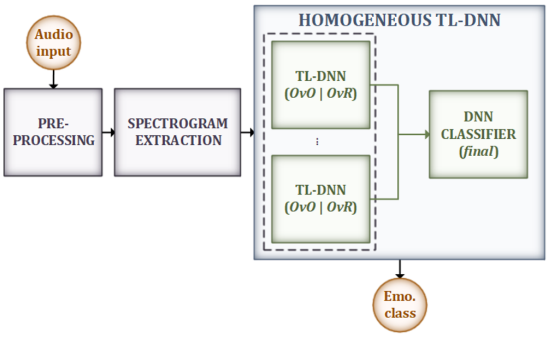

Fig. 5. *Approach 5*: homogeneous TL-DNN ensemble classification. The ensemble classification techniques (OvO or OvR) are adopted for multiple TL-DNN model classifiers.

Finally, in Fig. 5, the homogeneous TL-based approach is shown. This architecture leverages the ensemble classification techniques (OvO and OvR), but with TL-DNN models instead. The homogeneity property refers to the fact that the same model is used within a single system.

For the TL-DNN-based approaches, the input given to the networks must be in the form of spectrograms having linear or log magnitude, extracted using 25 ms Hamming windows (15 ms overlap), with linear or Mel scaling. Since the input size must be fixed, all audio input instances were first divided into segments with a standard duration (and zero-padded if required).

3. Experimental Setup and Results

3.1. Datasets and partitioning

The nature of the audio content of datasets for paralinguistic tasks should be as authentic as possible. Using simulated data, as is the case of hiring actors or guiding the subjects through specific scenarios, will usually affect the system's ability to generalize. Publicly available SER datasets reported in literature are simulated datasets, and many are in fact generally inadequate for SER, either due to their very reduced sample size, small number of speakers, and/or unreliable data annotation. Out of the benchmark SER datasets, the three best were selected for this work.

3.1.1. The EMODB dataset

The *Berlin Database of Emotional Speech* (EMODB) [27] is a German language dataset comprising 535 short utterances recorded by 10 actors (5 female, 5 male) in single-channel 16 bit PCM format, sampled at 16 kHz, with an average duration of 2.5 s and a maximum of 8 s. The 7 emotion classes considered are: *Anger* (ANG), *Disgust* (DIS), *Fear* (FEA), *Sadness* (SAD), *Boredom* (BOR), *Happiness* (HAP) and *Neutral* (NEU). Since this work focuses on forensic applications, negative emotions are more relevant. As such, additional subsets were considered:

- **EMODB-7:** all 7 classes: ANG, DIS, FEA, SAD, BOR, HAP, NEU; 535 samples;
- **EMODB-4:** 4 classes: ANG, SAD, HAP, and NEU (to match the IEMOCAP standard class subset); 339 samples;
- **EMODB-5N:** 5 classes: ANG, DIS, FEA, SAD, and NEU (*i.e.*, negative emotions only); 383 samples;
- **EMODB-2N:** 2 classes: Negative (NEG; grouping together ANG, DIS, FEA, and SAD) vs. NEU; 383 samples.

3.1.2. The CREMAD dataset

The *Crowd-sourced Emotional Multimodal Actors Dataset* (CREMAD) [28] is an English language dataset for studying emotions in a multimodal (audio-visual) context. It is described as comprising 7,442 audio-visual recordings of facial and vocal affective content manifested in sentences spoken by 91 directed actors (43 female, 48 male). The encompassed 6 emotion classes were: *Anger* (ANG), *Disgust* (DIS), *Fear* (FEA), *Sadness* (SAD), *Happiness* (HAP), and *Neutral* (NEU). The average duration of the recordings is 2.5 s (minimum 1.3 s, maximum 5 s). The recordings were subsequently labeled in terms of the perceived emotions via crowd-sourcing by 2,443 evaluators, averaging the individual ratings for each instance. Similarly to the EMODB case, 4 subsets were utilized in this work:

- **CREMAD-6:** all 6 classes: ANG, DIS, FEA, SAD, HAP, and NEU; 7,441 samples;
- **CREMAD-4:** 4 classes: ANG, SAD, HAP, and NEU (to match the IEMOCAP standard class subset); 5,145 samples;
- **CREMAD-5N:** 5 classes: ANG, DIS, FEA, SAD, and NEU (*i.e.*, negative emotions only); 6,219 samples;
- **CREMAD-2N:** 2 classes: Negative (NEG; grouping together ANG, DIS, FEA, and SAD) vs. NEU; 6,219 samples.

3.1.3. The IEMOCAP dataset

The *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) dataset [29] contains 5 sessions with 10 actors (5 female, 5 male) working in pairs to solve scripted and improvised English speaking tasks, with a total number of 10,039 audio-visual recordings. The files have an average duration of 4.5 s, with 16-bit PCM audio at a 16 kHz sampling rate. It comprises 10 discrete emotion classes (Anger, Fear, Disgust, Sadness, Happiness, Frustration, Excitement, Surprise, Neutral, and Other), many of them strongly underrepresented. This results in having to group only a smaller subset into 4 new classes, *i.e.*, *Neutral* (NEU); *Sadness* (SAD); *Anger* + *Frustration* (ANG); and *Happiness* + *Excitement* (HAP). The last two pairs were grouped together due to their closeness, similar to [30]. For the continuous dimensions, *arousal* (*activation*) and *valence* were chosen. The 2 subsets used in this work were:

- **IEMOCAP-4:** 4 classes: ANG, HAP, SAD, and NEU; 7,552 samples;
- **IEMOCAP-2N:** 2 classes: Negative (NEG; grouping together ANG and SAD) vs. NEU; 5,870 samples.

3.2. Setup and methodology

All experiments were run using the Keras framework. For the single DNN models, either the 'constant' or 'log2dec' active node architecture was used. The number of hidden layers (*i.e.*, the depth) was chosen between 1 and 4, with the number of neurons for the first hidden layer being chosen from the set $\{8, 16, 32, 64, 128, 256, 512, 1024\}$, and varying the dropout rate between 0.1 and 0.5. Other hyperparameters chosen included: the *rectified linear unit* (ReLU) activation function for the hidden layers, and either the softmax (when applied for classification) or the linear output activation function (when applied for regression). The selected optimization algorithm was Adam, employing L2-norm regularization with $\lambda = 10^{-4}$, and with the categorical cross-entropy function serving as the loss function for the classification systems, the role being given to the *mean squared error* (MSE) for the regression systems. The experiments were run for up to 200 epochs, using early stopping and learning rate decay. The same configurations were tested afterwards for the OvO and OvR DNN classifiers in the ensemble classification

experiments, with the final DNN classifier having a fixed depth of either 1, 2, or 3 layers. The configurations were also employed for the fully-connected classification heads of the TL-DNNs.

To counteract the imbalances present within each dataset between the class populations, weighting was adopted for the loss function gradient computation, assigning larger weights to the gradient components corresponding to minority class instances. For all experiments, 10-fold cross-validation was employed, with an 80% / 20% training-validation split, ensuring as best as possible that each emotion class and each gender was proportionally represented in each training and/or validation subset. Speaker separation was ensured for all experiments.

The metrics used for the classification experiments were the *unweighted accuracy* (UA) and the *weighted accuracy* (WA), for overall system performance; the *precision* (P), to measure the system's ability not to misidentify emotion classes; the *recall* (R), to measure the system's ability to retrieve the emotional content for each class; and the F1-measure, as a secondary overall metric. For the regression task, apart from the loss value, *i.e.*, the MSE value, the Pearson correlation coefficient, ρ, and the concordance correlation coefficient, ρ_c [3], were also adopted.

3.3. Results and discussions

For the single-network-based systems (*Approach 1*), for each dataset (EMODB, CREMAD, IEMOCAP), higher performance was observed as the class-complexity of the multiclass problem was reduced, *i.e.*, smaller numbers of considered classes. This is not surprising, since the learned data transformations can more easily ensure separability as the number of classes is reduced. Higher accuracies (over 80%) were observed for the EMODB subsets. One definitive reason is the reduced size of the EMODB-4 subset, and the different dataset recording conditions (acoustic conditions, particular methodologies employed, annotation quality, etc.). The best configuration comprised 2 hidden layers with the 'constant' architecture, with 64 neurons in each layer.

For the ensemble classifiers employing multiple FCNN models (*Approach 2*; applicable only for the multiclass data subsets), the best-performing architectures varied for each subset. The individual FCNNs had 2 hidden layers with 64 neurons each ('constant' architecture) for the EMODB-7 and EMODB-4; similarly for the CREMAD-6 and CREMAD-4 cases, but with the 'log2dec' architecture and 128 neurons in the first hidden layer. The other cases required a single hidden layer with 64 or 128 neurons. The OvR technique proved best for the EMODB data subsets, with higher results being obtained with the OvO technique for the CREMAD and IEMOCAP subsets. In all cases, the final classifier had a depth of 1 Performance improvements over *Approach 1* were observed, with relative increases in UA between 1.3% (CREMAD-4; from 61.8% to 62.6%) and 6.5% (IEMOCAP-4; from 55.1% to 58.7%) and in WA between 1.2% (EMODB-5N; from 90.3% to 91.4%) and 12% (IEMOCAP-4; from 55.0% to 61.6%).

The experiments for the single-network TL-DNN systems (*Approach 3*) were first conducted on the EMODB-7 subset to determine the feasibility of the approaches. All the TL-DNN model classifier heads were replaced with a single 32 neuron hidden layer, trained with a dropout rate of 0.3, and a final 7-dimensional output layer. In terms of the spectrogram hyperparameters, the most successful choice was log-magnitude with linear frequency scaling. The top-performing TL-DNN model was observed to be EfficientNetB0, the version of EfficientNet with the lowest complexity in terms of the number of parameters and core depth. It is also important to note that the EfficientNetB0 model provided the best individual class accuracy (*i.e.*, it was able to better identify each class, including the underrepresented ones), apparent from the high value of the UA, 73.2%, close to the value of the WA, 74.0%, and significantly higher than for other models.

For the heterogeneous fusion classification TL-DNN systems (*Approach 4*), the best results

were obtained with complete feature map fusion (*i.e.*, concatenating all the flattened outputs of the individual TL-DNN models), leading to a 20,960-dimensional feature vector (FV). However, the performance was considerably reduced, reaching a UA of only 52.9% and a WA of only 58.7% on the EMODB-7 subset, suggesting the infeasibility of this proposed approach.

Finally, for the homogeneous ensemble classification TL-DNN systems (*Approach 5*), the top performing TL-DNN model was either EfficientNetB0 or EfficientNetB1. The input spectrograms were scaled according to the results previously determined (*i.e.*, log-magnitude spectrograms with linear frequency scaling). The classifier heads comprise a single hidden layer with 32, 64, or 128 neurons depending on the considered data subset. The same viability and preference in terms of the ensemble classification grouping strategy was observed here as for the DNN systems investigated beforehand, with the OvR technique being successful for the EMODB dataset, and the OvO technique being more suited for the CREMAD and IEMOCAP datasets. The homogeneous ensemble classification TL-DNN systems did not outperform the simpler DNN-based ensemble classifier systems in the case of the EMODB and IEMOCAP data subsets. However, increased performance was obtained for the CREMAD data subsets, with relative increases in terms of UA of 3.2% (CREMAD-6; from 50.2% to 51.8%), 5.1% (CREMAD-5N; from 62.6% to 65.8%), and 2.4% (CREMAD-4; from 53.4% to 54.7%) and in terms of WA of 2.4% (CREMAD-6; from 53.3% to 54.6%), 5.6% (CREMAD-5N; from 66.6% to 70.3%), and 2.1% (CREMAD-4; from 57.5% to 58.7%), the most significant performance boost being obtained exactly for the most important subset for this work, *i.e.*, comprising negative emotions.

A summary of the previously discussed best results obtained in this work for the proposed SER systems, alongside performance comparisons to other results reported in literature, is given in Table 2 and Table 3. For the regression task applied to the IEMOCAP dataset, the proposed solution leads to better results in terms of *arousal* estimation. For *valence*, the obtained performance was better than [3], with the concordance value given in [11] being less reliable. The main classification comparisons are for the subsets that comprise all usable classes, *i.e.*, 7 (all) for EMODB, 6 (all) for CREMAD, and 4 (standard selection) for IEMOCAP. For the EMODB-7 subset, the most widely used benchmark for SER classification tasks, the top results obtained in this work outperformed almost all other state-of-the-art systems, validating the proposed approach, especially the developed ensemble classification techniques. Slightly less favorable, but comparable results were obtained for the CREMAD-6 subset: the best-performing proposed system did not manage to improve over all other reported methods, with the observed performance gap being under 2.6%. Mixed results were obtained for the IEMOCAP-4 subset, with the proposed solution managing to fall in the middle of the state-of-the-art hierarchy, but with a larger margin between the best performance achieved in this work vs. the top system reported in literature, *i.e.*, 5% (in terms of UA; 63.7% vs. 58.7% for this work). For all other cases, the proposed systems outperformed all other corresponding results reported in literature.

Table 2. Best SER regression performance and comparison between this work and other literature

Dataset	**System**	**Perf.**					
		MSE (loss)		ρ		ρc	
		A	**V**	**A**	**V**	**A**	**V**
IEMOCAP	[3] – MLP	–	–	–	–	0.611	0.301
	[11] – DNN	–	–	–	–	0.392	**0.715**
	This work: *Approach 1* – single DNN.	**0.073**	**0.180**	**0.677**	**0.408**	**0.621**	**0.343**

Table 3. Best SER classification performance and comparison between this work and other literature.

Data subset	System	Perf.	
		UA [%]	WA [%]
EMODB-7	[15] – SVM + recursive feature elimination	–	**86.2**
	[7] – CRNN	–	82.8
	[14] – SNN + LSM + Gammatone filterbank	–	82.4
	[9] – SVM + gender recognition	–	81.5
	[31] – GMM	79.8	–
	[32] – SVM + feature selection	78.6	79.1
	[16] – GA + clustering	77.5	–
	[31] – SVM	77.0	–
	This work: *Approach 2* – ensemble classification (OvR) with multiple DNNs (FCNNs).	**82.6**	**82.9**
EMODB-4	[10] – GMM + SVM	–	84.3
	This work: *Approach 2* – ensemble classification (OvR) with multiple DNNs (FCNNs).	**88.9**	**89.1**
EMODB-5N	[13] – MLP + GA-based modified backpropagation	–	80.4
	This work: *Approach 2* – ensemble classification (OvR) with multiple DNNs (FCNNs).	**91.2**	**91.4**
EMODB-2N	[8] – SVM	–	95.8
	[10] – GMM + SVM	–	94.9
	This work: *Approach 1* – single DNN (FCNN) classifier.	**95.1**	**98.3**
CREMAD-6	[17] – SVM	–	**57.2**
	[5] – LSTM	–	**57.0**
	[18] – LSTM	–	41.5
	This work: *Approach 5* – homogeneous ensemble classification (OvO) with multiple TL-DNN models (EfficientNetB1).	**51.8**	**54.6**
CREMAD-4	**This work**: *Approach 5* – homogeneous ensemble classification (OvO) with multiple TL-DNN models (EfficientNetB1).	**65.8**	**70.3**
CREMAD-5N	**This work**: *Approach 5* – homogeneous ensemble classification (OvO) with multiple TL-DNN models (EfficientNetB0).	**54.7**	**58.7**
CREMAD-2N	**This work**: *Approach 1* – single DNN (FCNN) classifier.	**72.8**	**72.6**
IEMOCAP-4	[12] – DNN + adversarial data augmentation	**63.7**	**63.2**
	[30] – MLP + GAN-based synthetic data	**61.0**	–
	[12] – SVM + adversarial data augmentation	**60.0**	**64.7**
	[32] – SVM + feature selection	**59.4**	59.5
	[6] – MLP + LSTM + attention	58.7	**63.5**
	[19] – LSTM	48.7	57.1
	[4] – MLP + i-vectors	–	48.8
	This work: *Approach 2* – ensemble classification (OvO) with multiple DNNs (FCNNs).	**58.7**	**61.6**
IEMOCAP-2N	[2] – SVM + feature adaptation	–	69.8
	This work: *Approach 1* – single DNN (FCNN) classifier.	**69.0**	**71.2**

4. Conclusions

SER systems based on *deep neural networks* (DNNs) spanning five levels of complexity were proposed, developed, and tested: single DNNs, multiple DNNs connected together following ensemble classification techniques (one-vs.-one and one-vs.-rest), and systems leveraging *transfer learning* (TL) for the top modern image recognition deep learning models, as standalone TL-DNN models, and as heterogeneous or homogeneous ensemble classifiers. The systems were tested on the most relevant SER datasets available: EMODB, CREMAD, and IEMOCAP, both for the standard full set of emotion classes, as well as for additional negative emotion subsets relevant for suspicious behavior monitoring and other forensic speech applications.

The proposed systems achieved state-of-the-art results (up to 83% accuracy) for the EMODB all-class subset, while the performance on the corresponding CREMAD and IEMOCAP subsets was lesser (up to 55% accuracy for CREMAD and 62% accuracy for IEMOCAP), but still comparable to other published research. Additionally, for all negative-emotion-only subsets (most relevant for this work), the proposed solutions offered top performance.

Future work includes testing other deep learning and transfer learning architectures, and further research into developing language-independent systems for cross-corpus experiments.

References

[1] S. MIHALACHE, D. BURILEANU, G. POP and C. BURILEANU, *Modulation-based SER with reconstruction error feature expansion*, Proceedings of International Conference on Speech Technology and Human-Computer Dialogue, Timisoara, Romania, 2019, pp. 1–6.

[2] T. RAHMAN and C. BUSSO, *A personalized emotion recognition system using an unsupervised feature adaptation scheme*, Proceedings International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan, 2012, pp. 5117–5120.

[3] B. T. ATMAJA and M. AKAGI, *Deep multilayer perceptrons for dimensional speech emotion recognition*, Proceedings Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Auckland, New Zealand, 2020, pp. 325–331.

[4] W. RAO, Z. H. LIM, Q. WANG, C. XU, X. TIAN, E. S. CHNG and H. LI, *Investigation of fixed-dimensional speech representations for real-time speech emotion recognition system*, Proceedings International Conference on Orange Technologies, Singapore, 2017, pp. 197–200.

[5] E. GHALEB, M. POPA and S. ASTERIADIS, *Multimodal and temporal perception of audio-visual cues for emotion recognition*, Proceedings International Conference on Affective Computing and Intelligent Interaction, Cambridge, UK, 2019, pp. 552–558.

[6] S. MIRSAMADI, E. BARSOUM and C. ZHANG, *Automatic SER using recurrent neural networks with local attention*, Proceedings ICASSP, New Orleans, USA, 2017, pp. 2227–2231.

[7] M. CHEN, X. HE, J. YANG and H. ZHANG, *3-D convolutional recurrent neural networks with attention model for SER*, IEEE Signal Processing Letters **25**(10), 2018, pp. 1440–1444.

[8] S. CASALE, A. RUSSO, G. SCEBBA and S. SERRANO, *Speech emotion classification using machine learning algorithms*, Proceedings IEEE International Conference on Semantic Computing, Santa Monica, USA, 2008, pp. 158–165.

[9] I. BISIO, A. DELFINO, F. LAVAGETTO, M. MARCHESE and A. SCIARRONE, *Gender-driven emotion recognition through speech signals for ambient intelligence applications*, IEEE Transactions on Emerging Topics in Computing **1**(2), 2013, pp. 244–257.

[10] J. C. VASQUEZ CORREA, N. GARCIA, J. R. OROZCO ARROYAVE, J. D. ARIAS-LONDONO, J. F. VARGAS BONILLA and E. NOTH, *Emotion recognition from speech under environmental noise conditions using wavelet decomposition*, Proceedings International Carnahan Conference on Security Technology, Taipei, Taiwan, 2015, pp. 247–252.

[11] H. ZHAO, N. YE and R. WANG, *Transferring age and gender attributes for dimensional emotion prediction from big speech data using hierarchical deep learning*, Proceedings IEEE International Conference on Big Data Security on Cloud, Omaha, USA, 2018, pp. 20–24.

[12] L. YI and M. W. MAK, *Improving speech emotion recognition with adversarial data augmentation network*, IEEE Transactions on Neural Networks and Learning Systems **33**(1), 2022, pp. 172–184.

[13] L. HE, Y. BO and G. ZHAO, *Speech-oriented negative emotion recognition*, Proceedings Chinese Control Conference, Hangzhou, China, 2015, pp. 3553–3558.

[14] R. LOTFIDERESHGI and P. GOURNAY, *Biologically inspired speech emotion recognition*, Proceedings ICASSP, New Orleans, USA, 2017, pp. 5135–5139.

[15] L. KERKENI, Y. SERRESTOU, K. RAOOF, M. MBARKI, M.A. MAHJOUB and C. CLEDER, *Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO*, Speech Communication **114**, 2019, pp. 22–35.

[16] S. KANWAL and S. ASGHAR, *Speech emotion recognition using clustering-based GA-optimized feature set*, IEEE Access **9**, 2021, pp. 125830–125842.

[17] E. GHALEB, M. POPA and S. ASTERIADIS, *Metric learning-based multimodal audio-visual emotion recognition*, IEEE MultiMedia **27**(1), 2020, pp. 37–48.

[18] R. BEARD, R. DAS, R. W. M. NG, P. G. KEERTHANA GOPALAKRISHNAN, L. EERENS, P. SWIETOJANSKI and O. MIKSIK, *Multi-modal sequence fusion via recursive attention for emotion recognition*, Proceedings Conference on Computational Natural Language Learning, Brussels, Belgium, 2018, pp. 251–259.

[19] Z. PAN, Z. LUO, J. YANG and H. LI, *Multi-modal attention for speech emotion recognition*, Proceedings INTERSPEECH, Shanghai, China, 2020, pp. 364–368.

[20] F. CHOLLET, *Xception: deep learning with depthwise separable convolutions*, Proceedings IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017, pp. 1800–1807.

[21] K. SIMONYAN and A. ZISSERMAN, *Very deep convolutional networks for large-scale image recognition*, Proceedings International Conference on Learning Representations, San Diego, USA, 2015, pp. 1–14.

[22] K. HE, X. ZHANG, S. REN and J. SUN, *Deep residual learning for image recognition*, Proceedings CVPR, Las Vegas, USA, 2016, pp. 770–778.

[23] C. SZEGEDY, V. VANHOUCKE, S. IOFFE, J. SHLENS and Z. WOJNA, *Rethinking the Inception architecture for computer vision*, Proceedings CVPR, Las Vegas, USA, 2016, pp. 2818–2826.

[24] B. ZOPH, V. VASUDEVAN, J. SHLENS and Q. V. LE, *Learning transferable architectures for scalable image recognition*, Proceedings CVPR, Salt Lake City, USA, 2018, pp. 8697–8710.

[25] M. TAN and Q. V. LE., *EfficientNet: rethinking model scaling for convolutional neural networks*, Proceedings International Conference on Machine Learning, Long Beach, USA, 2019, pp. 1–10.

[26] S. MIHALACHE, D. BURILEANU, E. FRANTI, M. DASCALU and C.A. BRATAN, *Lasting emotions – An investigation of short- and long-term affective content remanence in speech*, Romanian Journal of Information Science and Technology **25**(1), 2022, pp. 20–35.

[27] F. BURKHARDT, A. PAESCHKE, M. ROLFES, W. SENDLMEIER and B. WEISS, *A database of German emotional speech*, Proceedings INTERSPEECH, Lisbon, Portugal, 2005, pp. 1517–1520.

[28] H. CAO, D. COOPER, M. KEUTMANN, R. C. GUR, A. NENKOVA and R. VERMA, *CREMA-D: crowd-sourced emotional multimodal actors dataset*, IEEE Transactions on Affective Computing **5**(4), 2014, pp. 377–390.

[29] C. BUSSO, M. BULUT, C.C. LEE, A. KAZEMZADEH, E. MOWER, S. KIM, J.N. CHANG, S. LEE and S.S. NARAYANAN, *IEMOCAP: Interactive emotional dyadic motion capture database*, Language Resources & Evaluation **42**(4), art. no. 335, 2008.

[30] S. LATIF, M. ASIM, R. RANA, S. KHALIFA, R. JURDAK and B. SCHULLER, *Augmenting generative adversarial networks for speech emotion recognition*, Proceedings INTERSPEECH, Shanghai, China, 2020, pp. 521–525.

[31] T. CHASPARI, D. DIMITRIADIS and P. MARAGOS, *Emotion classification of speech using modulation features*, Proceedings European Signal Processing Conference (EUSIPCO), Lisbon, Portugal, 2014, pp. 1552–1556.

[32] S. YILDIRIM, Y. KAYA and F. KILIC, *A modified feature selection method based on metaheuristic algorithms for speech emotion recognition*, Applied Acoustics **173**, art. no. 107721, 2021.