Biomedical Signal Processing and Control 59 (2020) 101894

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Speech emotion recognition with deep convolutional neural networks

Dias Issa*, M. Fatih Demirci, Adnan Yazici

Department of Computer Science, Nazarbayev University, Nur-Sultan 010000, Kazakhstan

ARTICLE INFO

Article history:
Received 31 July 2019
Received in revised form 17 January 2020
Accepted 15 February 2020

Keywords:
Speech emotion recognition
Deep learning
Signal processing

ABSTRACT

The speech emotion recognition (or, classification) is one of the most challenging topics in data science. In this work, we introduce a new architecture, which extracts mel-frequency cepstral coefficients, chromagram, mel-scale spectrogram, Tonnetz representation, and spectral contrast features from sound files and uses them as inputs for the one-dimensional Convolutional Neural Network for the identification of emotions using samples from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Berlin (EMO-DB), and Interactive Emotional Dyadic Motion Capture (IEMOCAP) datasets. We utilize an incremental method for modifying our initial model in order to improve classification accuracy. All of the proposed models work directly with raw sound data without the need for conversion to visual representations, unlike some previous approaches. Based on experimental results, our best-performing model outperforms existing frameworks for RAVDESS and IEMOCAP, thus setting the new state-of-the-art. For the EMO-DB dataset, it outperforms all previous works except one but compares favorably with that one in terms of generality, simplicity, and applicability. Specifically, the proposed framework obtains 71.61% for RAVDESS with 8 classes, 86.1% for EMO-DB with 535 samples in 7 classes, 95.71% for EMO-DB with 520 samples in 7 classes, and 64.3% for IEMOCAP with 4 classes in speaker-independent audio classification tasks.

1. Introduction

Speech emotion recognition is an important problem receiving increasing interest from researchers due to its numerous applications, such as audio surveillance, E-learning, clinical studies, detection of lies, entertainment, computer games, and call centers. Nevertheless, this problem still remains a significantly challenging task for advanced machine learning techniques. One of the reasons for such a moderate performance is the uncertainty of choosing the right features. In addition, the existence of background noise in audio recordings, such as real-world voices, could dramatically affect the effectiveness of a machine learning model [1]. Nevertheless, the advent of decent emotional speech recognition models could significantly improve the user experience in systems involving human-machine interactions, for example in the areas of Artificial Intelligence (AI) or Mobile Health (mHealth) [2]. Indeed, the ability to recognize emotions from audio samples and, therefore, the ability to imitate these emotions could have a considerable impact on the field of AI. Various virtual assistants in the field of mHealth after using such models can significantly improve their performance. In addition, emotional speech recognition systems are unpretentious in terms of the hardware requirements.

For now, deep learning models are utilized to solve recognition problems such as face recognition, voice recognition, image recognition, and speech emotion recognition [3–6]. One of the main advantages of deep learning techniques lies in the automatic selection of features, which could, for example, be applied to important attributes inherent to sound files having a particular emotion in the task of recognition of speech emotions [7].

In recent years, various models based on deep neural networks for speech emotion recognition have been introduced. While one group of these models designs the neural network with the objective of detecting significant features directly from raw sound samples [8], the other group uses only one particular representation of a sound file and input to their models, e.g., [7,1].

In this work, we extract five different features from a sound file and stack the resulting matrices in a one-dimensional array by taking mean values along the time axis. This array is then fed into the 1-D Convolutional Neural Network (CNN) model as input. We assert that the mixing of these features in the input data provides a more diverse representation of a sound file, which leads to a better generalization and a better classification during the process of recognizing emotions from speech. In addition, we utilize an incre-

* Corresponding author.
E-mail addresses: dias.issa@nu.edu.kz (D. Issa), muhammed.demirci@nu.edu.kz (M. Fatih Demirci), adnan.yazici@nu.edu.kz (A. Yazici).

https://doi.org/10.1016/j.bspc.2020.101894
1746-8094/

mental methodology for modifying our baseline model to improve its classification accuracy.

Although several speech-emotion recognition frameworks in the literature combine different feature types, the proposed feature combination using five different spectral representations of the same sound file has not been attempted by researchers so far. Specifically, the mix of features resulting in powerful identification and tracking of timbre fluctuations but poor distinguishable representations of pitch classes and harmony is further enriched with additional features to improve its representational power. As a result, our best performing model outperforms all existing frameworks, which use audio features and report their classification accuracies on the same emotion classes for both RAVDESS [9] and the IEMOCAP datasets [10], yielding the new state-of-the-art. For the EMO-DB dataset [11], our best performing model outperforms all previous work, with the exception of the study by Zhao et al. [12]. However, our model compares favorably with that one in terms of generality, simplicity, and applicability. In addition, we have noticed some inconsistencies related to [12] as we discuss them in Section 4.4.

In the next section, we present a brief review of the literature on previous work in the field of speech-related emotion recognition. After that, we present our methodology and the proposed baseline model in Section 3. The datasets, our improvements to the baseline model, and the experiments are described in the next section. After comparing our results with those of previous approaches, we draw conclusions and indicate possible future directions.

2. Literature review

The majority of speech-emotion recognition architectures that utilize neural networks are Convolutional Neural Networks (CNN), recurrent neural networks (RNN) with long-short term memory (LSTM), or their combination [8,7,2,13]. The combination of CNN and RNN could detect an essential pattern in audio files when extracting features and classifying entities [8,7]. One of the main goals in speech emotion recognition is the identification of significant features that could then be used for model training. Researchers use different approaches in order to solve this problem. Trigeorgis et al. [8] utilize the raw audio data as input for their model. The authors then used CNN for preprocessing of the audio samples with the purpose of reducing noise and emphasizing specific regions of the audio file. The architecture achieved better results than the state-of-the-art models of that time.

The different way was utilized by Tarantino et al. [14] who adopted the predefined feature set, eGeMaps [15]. Combining the extracted features with self-attention and global windowing techniques, the authors were able to achieve significant enhancement of the state-of-the-art results on IEMOCAP [10] database.

Moreover, Triantafyllopoulos et al. [16] employed previously mentioned eGeMaps [15] together with the another predefined feature set named ComParE [17]. The authors applied speech enhancement algorithms for speech emotion recognition task and showed significant improvement in performance.

Lim et al. [7] converted the raw audio data from EMO-DB [11] dataset to its two-dimensional representation by applying the short-time Fourier transform. After that, the generated signal was sent to the models, where the first of the layers was CNN. The best results were shown by the time distributed CNN, one of their proposed models.

The similar method was used by Badshah et al. [2]. The authors generated visual representations of the sound samples called spectrograms out of the raw audio data from EMO-DB emotion [11] dataset. After that, these spectrograms were passed to the CNN model. The results show that the new CNN model produces satisfactory results for most categories, in addition to the fear. Nevertheless, they performed with 52% accuracy on the test set for all emotions [2,18].

Like Badshah et al. [2], Zhao et al. [12] utilized log-mel spectrograms as input data to their 2-D CNN LSTM network. The results of their work demonstrated the best possible accuracy of 95.33% for speaker-dependent classification and 95.89% for speaker-independent classification of samples from EMO-DB dataset [11]. To the best of our knowledge, it is the most advanced state-of-the-art solution for solving the problem of speech emotion recognition on this dataset.

Chatziagapi et al. [19] in turn proposed the Generative Adversarial Network (GAN) based data augmentation approach in order to enhance the performance of speech emotion recognition models. The authors applied GAN to spectrograms in order to produce synthetic ones. This technique was applied to the samples of IEMOCAP [10] database and allowed Chatziagapi et al. [19] to achieve 10% relative performance gain.

Another promising method was presented by Demircan and Kahramanli [20] that used Mel Frequency Cepstral coefficients (MFCCs) produced from 520 samples taken from EMO-DB dataset [11] for their models. The framework then implemented the feature reduction utilizing the fuzzy C-means clustering. The authors developed several classifiers: artificial neural network (ANN), support vector machines (SVM) and k-nearest neighbors (kNN). The classification accuracies on the test set were 90%, 92.86%, and 92.86%, respectively.

Yoon et al. [21] applied a multimodal approach for speech emotion recognition. For this task they utilized audio and text data from IEMOCAP [10] database. The authors utilized MFCCs and text tokens as the input features for their model. The multimodal approach led to the 71.8% accuracy on the testing set.

Unlike the previous approaches, Huang et al. [22] built the model that uses deep learning combined with the classic machine learning technique to classify the 800 entries of EMO-DB dataset [11]. The hybrid model, called semi-CNN, consists of CNN input layers dedicated for the affect-salient feature learning with the SVM classifier in the last layer for categorization. In the same way as the authors mentioned above, Huang et al. [22] utilized spectrograms as input to their semi-CNN. The results of their experiments indicate 88.3% accuracy for the speaker-dependent classification and 85.2% for speaker-independent classification on the test set.

Furthermore, Wu et al. [23] implemented only traditional machine learning techniques for classification of samples from EMO-DB [11] dataset with 800 entries. The authors proposed the new type of sound features, called modulation spectral features (MSFs). Utilizing MSFs in combination with prosodic features as input to a multi-class linear discriminant analysis (LDA) classifier, Wu et al. [23] obtained 85.8% accuracy for speaker-independent classification on the test set.

Lampropoulos and Tsihrintzis [24] combined MPEG-7 descriptors, MFCCs and Timbral features extracted from the utterances of EMO-DB dataset [11] and utilized this mixture as an input to their model. The authors used SVM with RBF kernel in order to classify the audio samples, and as a result, achieved the accuracy of 83.93% using Leave One Out evaluation [24].

Wang et al. [25] proposed a new type of sound features called Fourier Parameter (FP) features that are estimated by Fourier Analysis. The authors used only 6 out of 7 emotion classes of EMO-DB dataset [11] by eliminating the “disgust” class. Wang et al. [25] extracted FP and MFCC features from the dataset and utilized them as input to SVM classifier with the average accuracy of 73.3%.

Shegokar and Sircar [26] also utilized SVM for classifying male speech samples from RAVDESS dataset [9]. The authors applied the Continuous Wavelet Transform (CWT) when selecting features and fed the selected features to different types of SVM classifiers. The

best result with 60.1% accuracy was obtained with Quadratic SVM with a 5-fold cross-validation technique [26].

Another approach, called multi-task learning, has been demonstrated by Zhang et al. [27] in their paper. They combined features from the songs and the speech samples from RAVDESS [9] dataset by indicating that classifiers using the song-to-speech relationship can achieve higher accuracy [27]. The authors utilized only 4 classes of emotions out of 8: angry, happy, neutral and sad. As a result, they achieved the accuracy of 57.14% using the group multi-task feature selection (GMTFS) model.

The same idea was implemented by Zeng et al. [28] using a deep neural network (DNN). Using the spectrograms generated from the songs and the speech utterances of RAVDESS [9] dataset as input to their multi-task gated Residual Networks (GResNets), the authors achieved the accuracy of 65.97% and reported that the model outperforms the task-specific ones, trained separately for song and speech [28]. On the other hand, instead of constructing their own model, Popova et al. [29] used a fine-tuned DNN to classify the melspectrograms obtained from the speech samples of RAVDESS dataset [9]. The authors obtained the accuracy of 71% using Convolutional Neural Network VGG-16 as a classifier [29].

3. Datasets and methodology

We use three different audio datasets, RAVDESS [9], EMO-DB [11], IEMOCAP [10], which are widely employed by researchers in emotion recognition. After presenting the datasets, we describe the proposed framework, which starts with feature extraction followed by the baseline deep learning model. While the baseline model is ideal for RAVDESS, we present some additional deep learning models generated by different hyper parameter settings of the baseline and slight modifications to its architecture in order to adapt it for EMO-DB and IEMOCAP. The datasets, feature extraction, and the baseline model are presented next.

3.1. Datasets

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [9] is chosen as one of the datasets for our model because of its great availability. This dataset contains audio and visual recordings of 12 male and 12 female actors pronouncing English sentences with eight different emotional expressions. For our task, we utilize only speech samples from the database with the following eight different emotion classes: sad, happy, angry, calm, fearful, surprised, neutral and disgust. The waveform of each emotion is depicted in Fig. 1. The overall number of utterances is 1440.

The second dataset we use in our framework is EMO-DB [11], which is widely used by researchers in the field of speech-based emotion recognition, allowing us to draw more comprehensive comparisons with previous works. The dataset contains 535 audio utterances in German divided into 7 emotion classes: anger, sadness, fear/anxiety, neutral, happiness, disgust, and boredom.

While the two previous datasets are acted, we also employ the IEMOCAP dataset generated from improvisational data [10]. This dataset is comprised of audio, video and face motion capture samples collected from five pairs of male and female actors. The samples are distributed among five sessions, each containing data from a particular pair. The actors performed by using theatrical scripts, or by improvising affective scenarios. The audio files of the dataset are categorized into ten emotion classes: angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, other. To measure the performance of the proposed framework on improvised data, we utilize only the improvised prototypical (complete agreement on the affective state from evaluators) audio samples of the IEMOCAP dataset. In addition, since the majority of previous frameworks using IEMOCAP utilize only 4 emotion classes, i.e., angry, happy, neutral, and sad, we also use these classes in order to fairly compare our results. The number of audio files used in this setting is 889.

3.2. Feature extraction

Feature extraction plays a crucial role in the success of any machine learning model. Appropriate feature selection could lead to a better trained model, while inappropriate features would significantly hinder the training process [8]. For the feature extraction process, we utilized Librosa audio library [30]. Specifically, we use five different spectral representations of the same record as the input for our deep learning model:

- Mel-frequency Cepstral Coefficients (MFCCs)
- Mel-scaled spectrogram
- Chromagram
- Spectral contrast feature
- Tonnetz representation

Mel-scaled spectrogram and MFCCs are widely utilized in the field of sound classification and speech emotion recognition [31]. These features mimic to a certain extent the reception pattern of sound frequency intrinsic to a human. In particular, MFCCs collectively make up the mel-frequency cepstrum, which is defined as the representation of the short-term power spectrum of a sound. The Fourier transform and the energy spectrum are obtained and mapped into the Mel-frequency scale. Although both Mel-scaled spectrogram and MFCCs are decent in identification and tracking of timbre fluctuations in a sound file, they tend to be poor in a distinguishable representation of pitch classes and harmony [32]. Chromagrams are applied in order to deal with this problem. In this work, we used the obtained chromagram using short-time Fourier transform (STFT) [30]. The spectral contrast feature provides a more detailed spectral proof of a sound with respect to MFCCs and Mel-scaled spectrograms. According to the literature [33], methods based on detailed spectral information outperform techniques using Mel-scale in the field of music genre classification. The Tonnetz representation of a sound is similar to the chromagrams with respect to the representation of harmony and pitch classes. The method measures the tonal centroids of a sound in a six-dimensional pitch space called Tonal Centroid Space introduced by Harte et al. [34]. The tonal centroid is based on the Harmonic Network showing a planar representation of pitch relations such that pitch classes with close harmonic relations such as major/minor thirds have smaller Euclidean distances on the plane.

As mentioned before, the use of several different audio features, instead of just one, combines different sound characteristics such as pitch, timbre, harmony, etc., into one training utterance. This leads to a richer description of a sound sample, which improves the performance of speech-related emotion recognition models.

3.3. Proposed baseline model

In the proposed framework, we use the convolutional neural network (CNN) for the classification of emotions based on features extracted from a sound file. Our baseline model includes one-dimensional convolutional layers combined with dropout, batch normalization, and activation layers. The first layer of our CNN receives 193×1 number arrays as input data. The initial layer is composed of 256 filters with the kernel size of 5×5 and stride 1. After that, batch normalization is applied, and its output is activated by Rectifier Linear Units layer (ReLU). The next convolutional

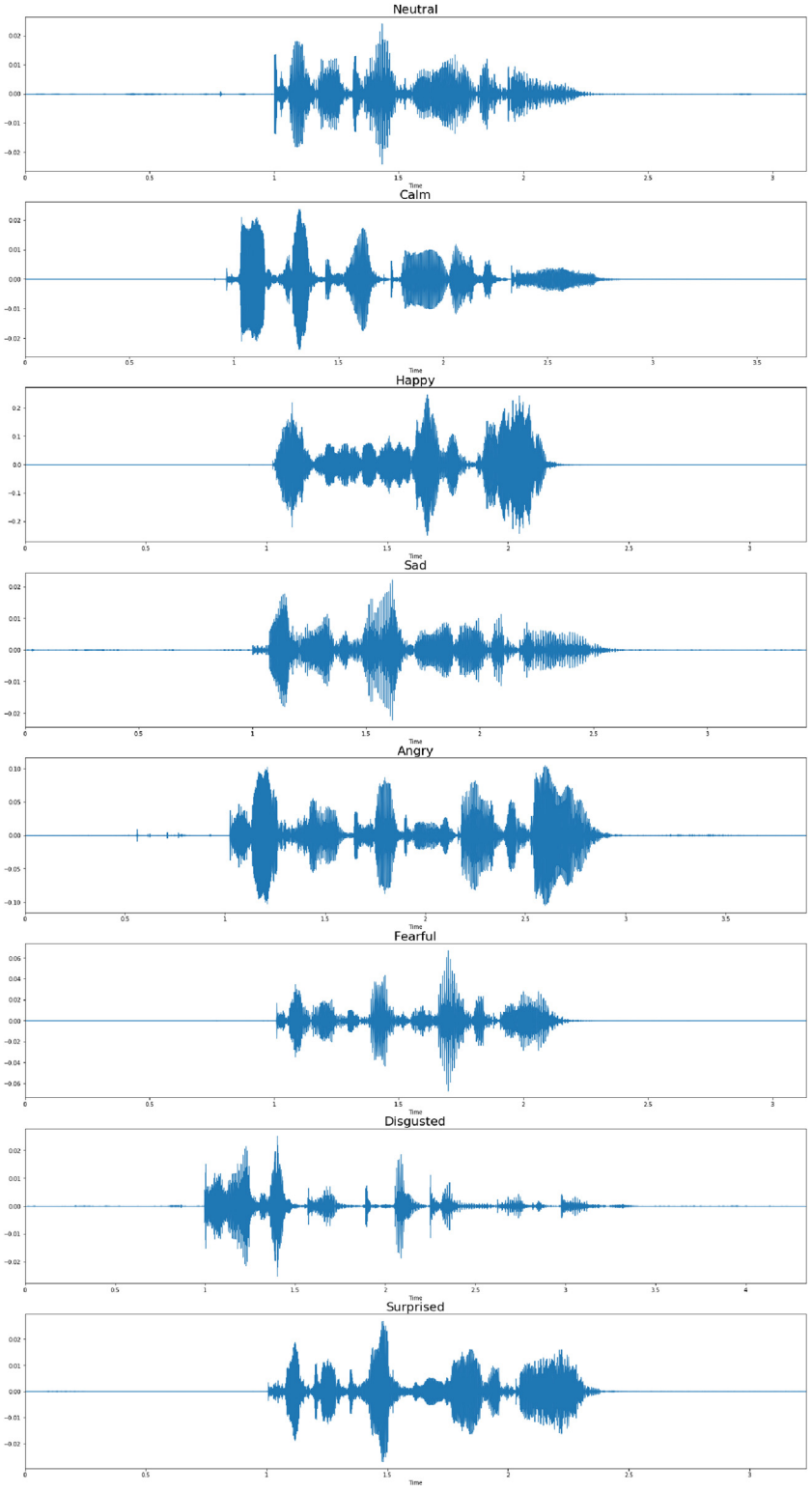

Fig. 1. The waveforms of eight emotions from RAVDESS dataset.

layer consisting of 128 filters with the same kernel size and stride receives the output of a previous input layer. The output of this layer is also activated by ReLU, and then dropout with the rate of 0.1 is applied. Next, batch normalization is implemented, feeding its output to the max-pooling layer with a window size of 8. Next, 3 convolution layers with 128 filters of size 5×5 are located, two of which are followed by ReLU activation layers and finally followed by batch normalization, ReLU activation and dropout layer with the rate of 0.2. The final convolutional layer with the same parameters is followed by the flattening layer and dropout with the rate of 0.2. The output of the flattening layer is received by a fully connected layer with 8, 7, or 2 units, depending on the number of predicted classes. After that, batch normalization and softmax activation are applied. Our model uses RMSProp optimizer with the learning rate

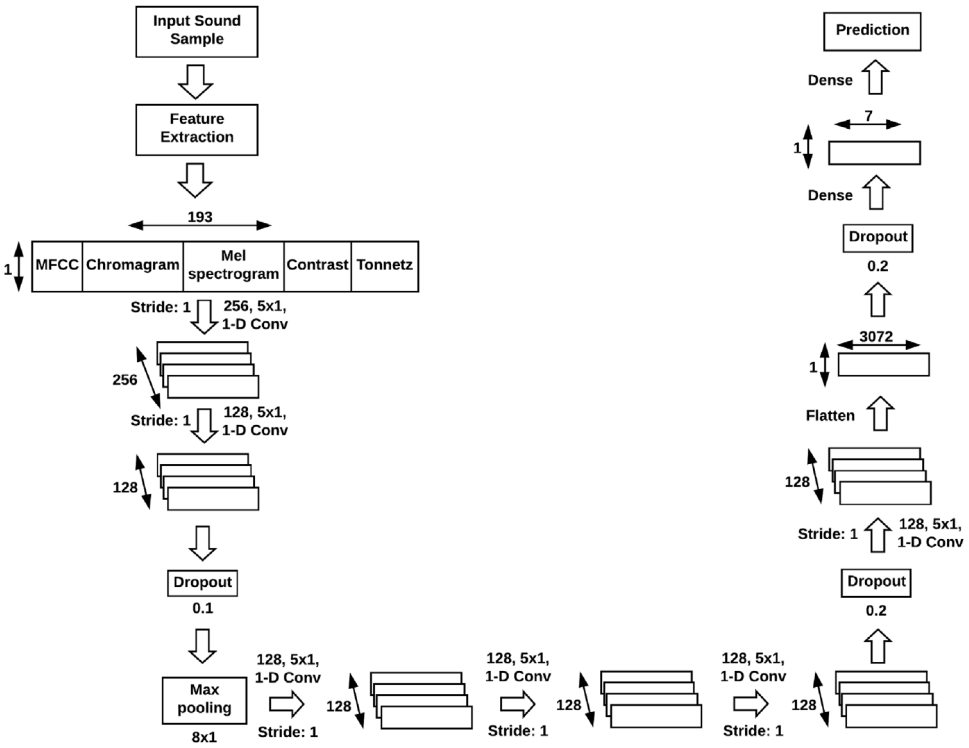

Fig. 2. The baseline topology of CNN.

of 0.00001 and the decay rate of 1e-6. The proposed architecture is illustrated in Fig. 2.

The proposed topology described above serves as the basis for all the models proposed in this paper. However, each of the models is fine-tuned later for better performance of their own classification tasks. We tune models simply by adding, removing or modifying dropout rates, and by removing some of the layers. More detailed information about the changes for a particular model is provided in the next section.

4. Model variations, and experiments

For the classification of emotions, we have implemented several incremental models using three datasets mentioned before. We will discuss these models in detail below.

4.1. RAVDESS model

The first implemented model has a topology described in the previous section with eight units in the fully connected final layer, allowing us to predict eight different emotions that exist in the RAVDESS dataset (Fig. 2). We applied five-fold cross validation, in which the dataset was randomly divided into five groups of equal size and 80% of the dataset was used as training while the rest 20% for testing. We performed this schema five times and computed the average classification accuracy on test sets. Since data partitioning is performed randomly, the classification is speaker-independent. After 700 epochs, the emotion classification performance of our model is recorded as 71.61%. The confusion matrix is given in Fig. 3 clearly shows that the model confidently identifies the strong emotions like “angry”. However, it confuses some close emotions like “calm” and “sad” or “happy” and “surprised”.

4.2. EMO-DB models

For the classification of the EMO-DB dataset, we present several models constructed incrementally. Before describing the models, it would be useful to add that we have implemented some data augmentation techniques to increase the number of samples used in training. Our data augmentation techniques include alterations like moving the beginning of the sound file by some small amount, speeding up and slowing down the file by 1.23% and 0.81% of its normal speed, respectively, and adding random noise to the 25% of its length. As in the RAVDESS Model, we use five-fold cross validation. Applying the data augmentation method increases the amount of training data from 425 samples to 2125 instances. The size of test set (110) is not modified. Similar to the previous model, our classification is speaker-independent due to the random partitioning of the dataset.

4.2.1. The model with 7 classes (Model A)

The topology of our first model (Model A) is different from the baseline model shown in Fig. 3. More precisely, we make the following modifications to the baseline: elimination of batch normalization layer after the first dropout and after the fully connected layer, removal of two subsequent convolution layers after the max pooling, removal of convolution layer preceding the flattening layer, and elimination of dropout before the fully connected layer. The topology of this model is presented in Fig. 4. After 300 epochs of training, our first model yields an accuracy of 82.86%. The confusion matrix illustrated in Fig. 5 clearly indicates that most classes are classified correctly, while “boredom” and “neutral”, and “disgust” and “fear” are confused with each other.

4.2.2. The model with 5 classes (Model B)

Based on the preliminary classification results of Model A, we decide to check the performance of our model without the two most misclassified “disgust” and “boredom” classes. We also change the

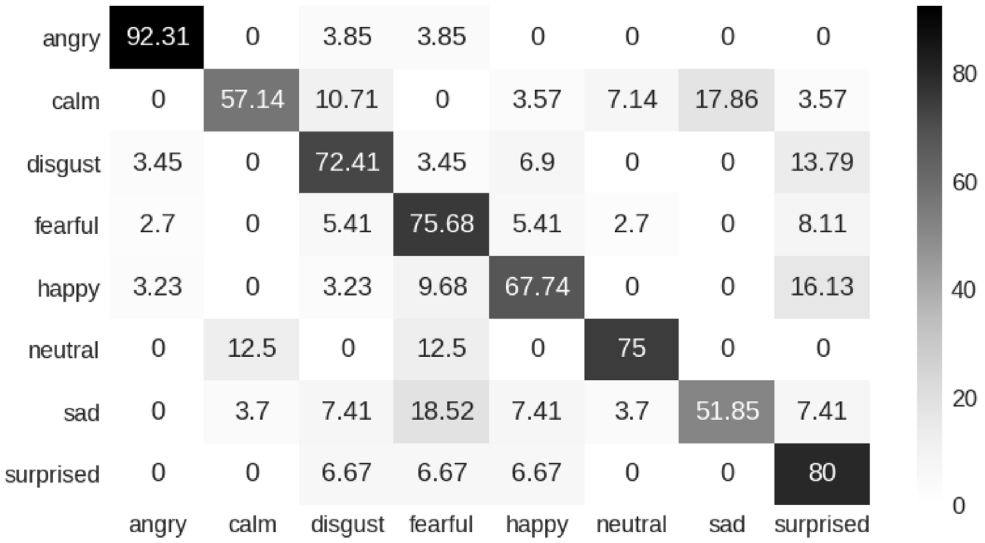

Fig. 3. Confusion matrix for the RAVDESS model.

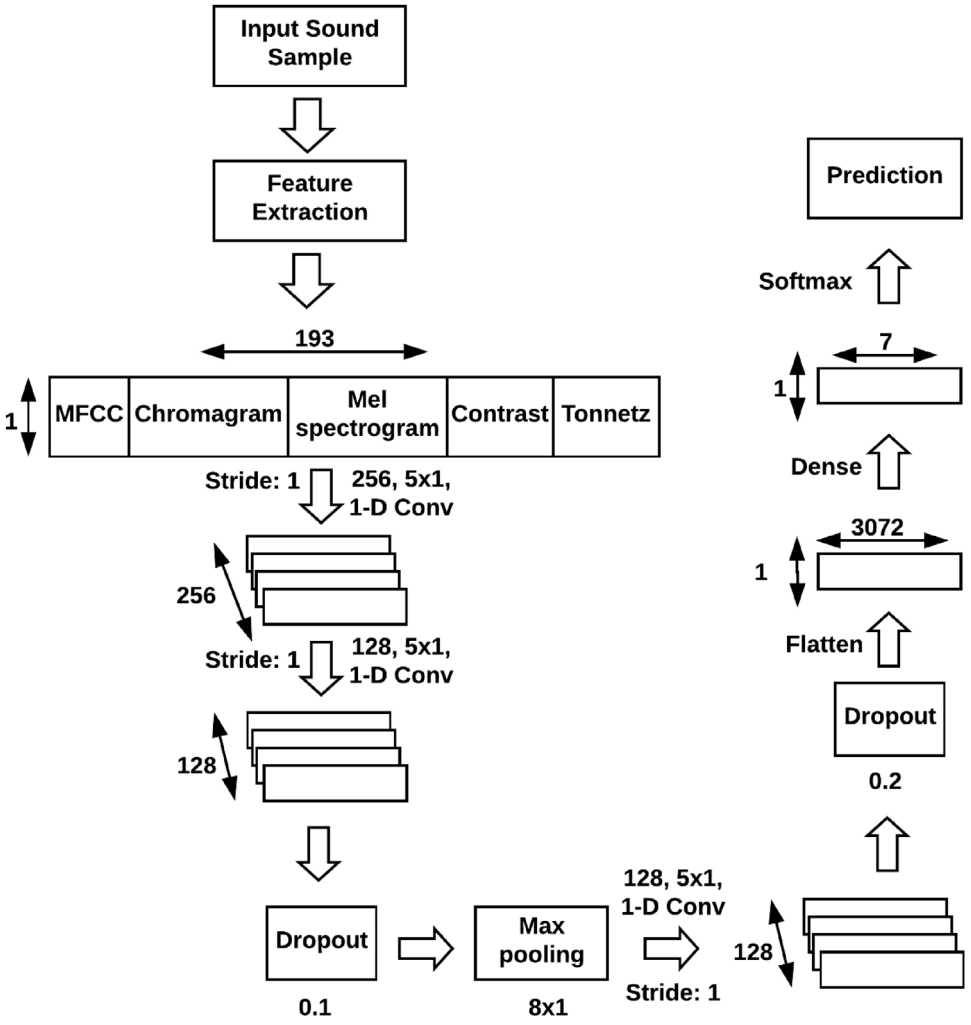

Fig. 4. The topology of the model with 7 classes (Model A).

topology of Model A: additional convolution layer before the flattening layer and additional dropout with the rate of 0.25 after the fully connected layer. With these new modifications, our new model (Model B) reaches a 96.34% accuracy on the subset of EMO-DB with five classes. It is clear that the two eliminated "disgust" and "boredom" classes have added a considerable amount of confusion to Model A, motivating us to increase the performance with the models specifically trained for particular emotions.

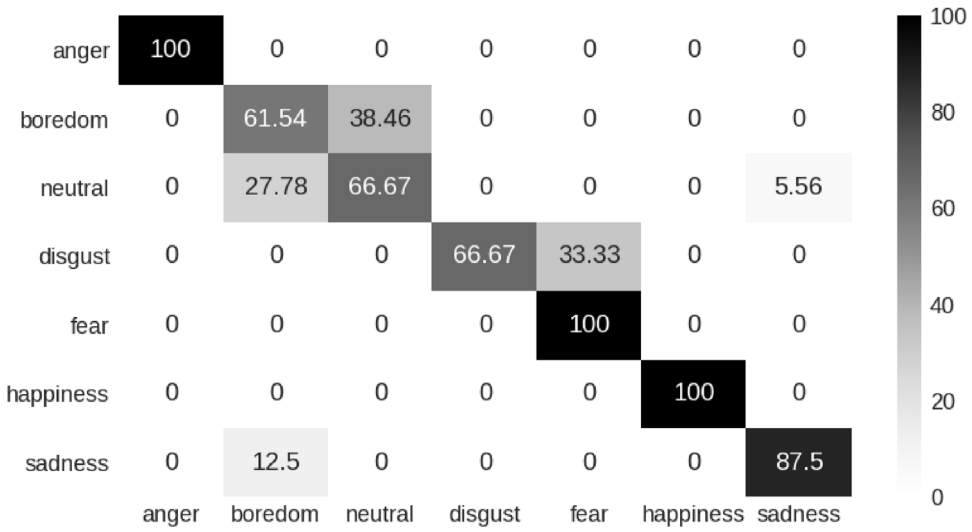

Fig. 5. Confusion matrix for Model A.

Table 1
The accuracy rates of the 7 binary classifiers used in Model C.

Emotion class	Accuracy (%)
Happiness	98.75
Sadness	98.75
Anger	97.50
Fear	97.50
Disgust	97.50
Boredom	90.00
Neutral	86.25

4.2.3. The ensemble of 7 models (Model C)

The next approach employed in this paper is to ensemble 7 binary classifiers, each of which is designed for one emotion category. The goal of each of these classifiers is to categorize a sound file into one of two classes: the class of a particular emotion for which the classifier has been trained for, and "the other" class. We organize the classifiers in decreasing order of accuracy. In this way, the classifier is to predict the class of an utterance only if the previous classifier has classified it as an instance of "the other" class with high accuracy. Otherwise, we assume that the utterance belongs to the class of the previous classifier. All binary classifiers are trained, validated and tested on the same sets of samples as those described at the beginning of this section. In addition, all classifiers have almost the same structures as Model A with only a few slight modifications: additional convolution layer after the maximum pooling layer with the same parameters as the previous one and additional dropout with the rate of 0.1 after the fully connected layer with two neurons.

The two emotions with the highest accuracy rate are "happy" and "sad". The next place is divided among "angry", "fearful" and "disgusted" categories, while the last two places are taken by "bored" and "neutral (calm)". The accuracy rates for each binary classifiers are depicted in Table 1. Although individual accuracies are fairly high, the final accuracy of the ensembled model is recorded as 82.4%, almost the same accuracy as Model A. It is unexpected that classifiers performing significantly promising results do not reach a similar performance in an ensemble. After investigating this issue more closely, we observe that the main reason for the lack of increased accuracy is the accumulation of errors caused by each binary model. Namely, the majority of the incorrect labelings comes from false negatives, i.e., the cases where an utterance is incorrectly classified as "the other" class.

4.2.4. The ensemble of 3 models (Model D)

The new idea that we have employed to improve the classification accuracy is to combine Model B with the binary models for "boredom" and "disgust" in an ensemble. Since Model B does not contain these two classes, we decide to organize them in a similar manner as in Model C. Specifically, our current model (Model D) starts with the binary "disgust" model followed by "boredom", and finally ends with Model B. The structure of the ensemble is illustrated in Fig. 6. Based on the overall results, this model increases the classification accuracy to 84.76%.

4.2.5. The ensemble of Model C and Model D (Model E)

Since Model D generates a better score then the previous models, we use the probabilities given by Model C and Model D to improve the accuracy. More precisely, given an audio sample file, we compare the classification probabilities of Model C and Model D, and use the classification with the highest probability as its final classification. In other words, we use the classification suggested by the following equations:

$$P_{Model\ E} = \max(P_{Model\ C},\ P_{Model\ D}), \tag{1}$$

$$P_{Model\ C} = \max(P_i), \text{ where } i \text{ is an emotion class.} \tag{2}$$

This strategy has improved our accuracy to 86.1%. The confusion matrix for this model (Model E) is illustrated in Fig. 7. Overall, the classification accuracies of our models are shown in Fig. 8.

4.3. IEMOCAP model

For categorizing the samples from IEMOCAP database, we utilized almost the same topology of the neural network as presented in Model A. The only differences are in optimizer function and in dropout value before the max-pooling layer. We utilized Adam optimizer instead of RMSProp due to its better convergence. We also increased the dropout from 0.1 to 0.2 with the purpose of avoiding overfitting. In order to train and test the model, the five-fold cross-validation technique was implemented, making our classification method speaker-independent. Specifically, since the dataset was divided into five sessions, each of which consisted of recordings of different pairs of speakers, we used sessions 1–4 for training and session 5 for testing. We then used the sessions 2–5 for training and session 1 for testing, etc., until all different sessions were used for training and testing. With this setup, our model achieved 64.30% of classification accuracy.

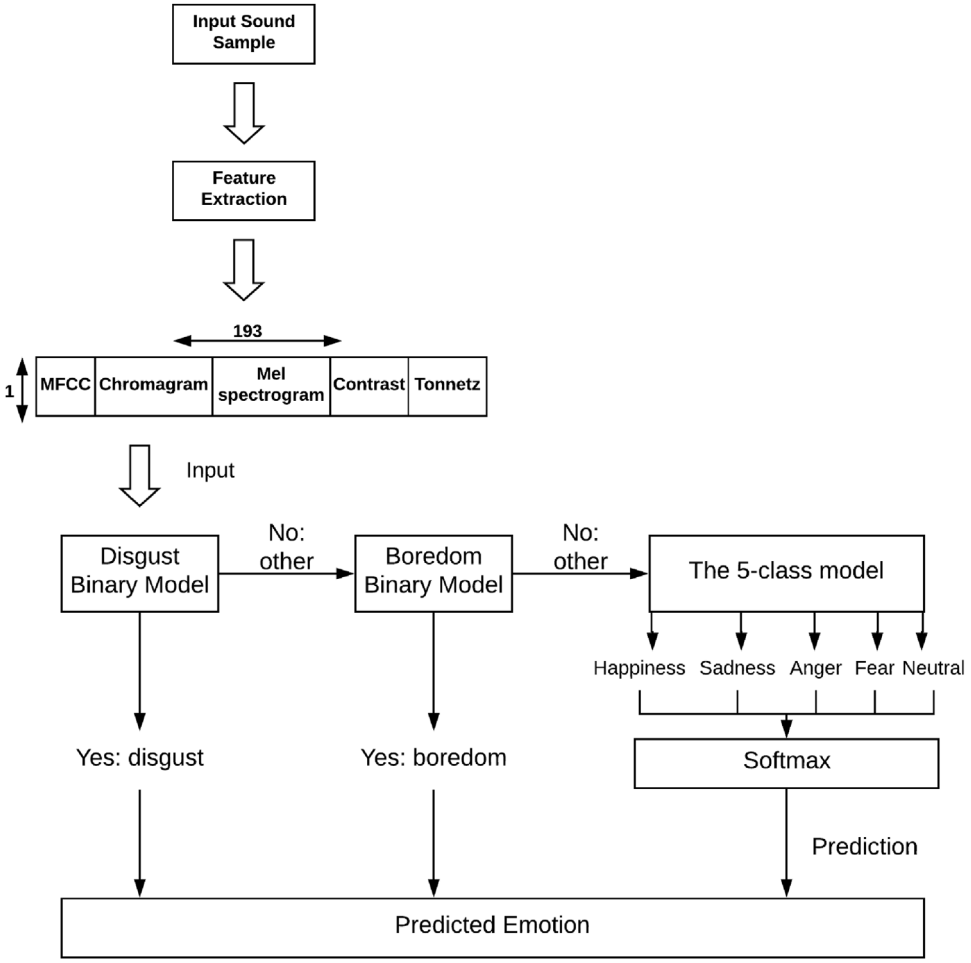

Fig. 6. The architecture of Model D.

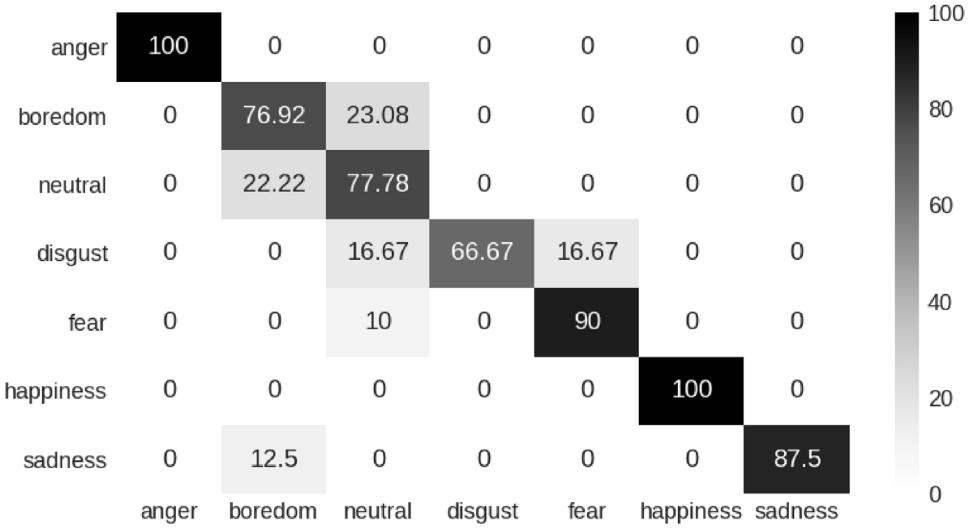

Fig. 7. Confusion matrix for Model E.

4.4. Comparison with previous work

4.4.1. RAVDESS dataset

The current human accuracy rate for this dataset has been reported as 67% [9], which indicates that the classification of emotions for this dataset is not a simple and straightforward task even for human beings. Fig. 9 illustrates the comparison of our model with the previous works on speech emotion recognition using RAVDESS dataset. According to the results, the proposed framework considerably outperforms the models of Shegokar and Sircar [26], Zeng et al. [28], and human accuracy [9].

Fig. 8. Different models implemented in this work for classification of utterances of EMO-DB dataset. Models A, C, D, and E consists of seven classes.

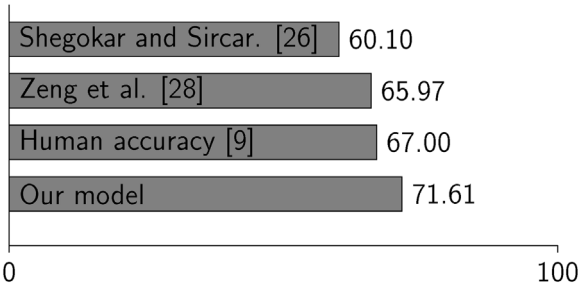

Fig. 9. RAVDESS: comparison with the previous works.

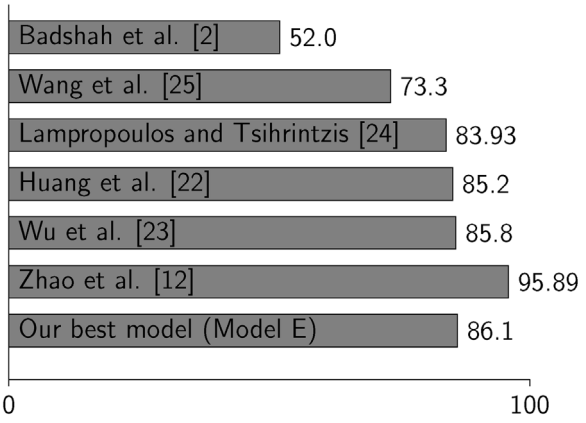

Fig. 10. Comparison with the previous works using EMO-DB with 535 samples.

In the literature, Popova et al. [29] report 71% accuracy on RAVDESS. However, they only utilize 1368 samples instead of 1440. That's why we do not include that method in the comparison chart. In fact, when we compare our framework, we only use previous frameworks, which utilize all samples from the dataset with the same number of classes. In addition, we should note that a video file is associated with each sound in the dataset. Some existing approaches also use these video files in combination with sound files to perform both feature extraction and classification processes. Here, we focus on methods that classify emotions with sound files only. Based on the overall results of the emotion classification, our framework sets the new state-of-the-art in audio sample classification for RAVDESS.

4.4.2. EMO-DB dataset

Comparing our work with previous speaker-independent approaches for the EMO-DB dataset [11] is given in Fig. 10. According to the results of the classification, our best model (Model E) greatly surpasses the works of Badshah et al. [2] and Wang et al. [25], and scores higher than Lampropoulos and Tsihrintzis [24], Wu et al. [23], and Huang et al. [22]. On the other hand,

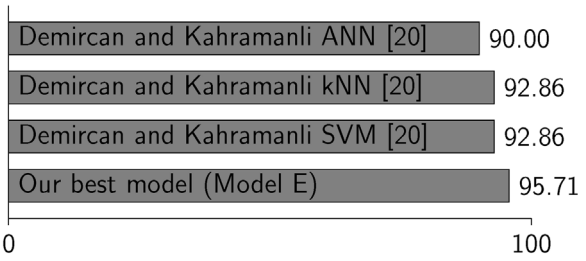

Fig. 11. EMO-DB with 520 samples: comparison with the previous works.

as shown in the comparison chart, the most recent approach of Zhao et al. [12] gets the best result in the speaker-independent speech emotion recognition domain for this dataset. However, their architecture is more sophisticated in comparison with our models. More specifically, the authors construct two convolutional neural network and long short-term memory (CNN LSTM) networks, one 1D CNN LSTM network and a 2D CNN LSTM network to learn emotion-related features from speech and log-mel spectrogram. Nevertheless, the advantages of our architecture mainly come from its simplicity, applicability, and generality, which is clearly seen from the fact that the use of models with almost the same architecture, even for different languages, makes it possible to achieve/outperform the state-of-the-art results. In addition, we believe that the study by Zhao et al. pose some inconsistency issues. They present the performance of their framework for both speaker-dependent and speaker-independent experiments. According to the previous studies, for example, Wu et al. [23], and Huang et al. [22], speaker-dependent models are expected to achieve higher emotion classification results than speaker-independent models. Even this is the case for Zhao et al., when using another sound dataset (see Table 9 in [12]). On the other hand, we note that their speaker-dependent result for EMO-DB is worse than their speaker-independent performance, giving an inconsistent result with respect to the literature.

By examining the literature more closely, we may notice more existing frameworks presenting their performances on this dataset. However, as in the RAVDESS experiments, we exclude them from our comparison chart mainly because they use only a part of this dataset. For instance, Demircan and Kahramanli [20] utilize only 520 samples as oppose to 535 without giving the reason for such a reduction and obtain an accuracy of up to 92.86%. We should also note that the second problem of inconsistency that emerges from the work of Zhao et al. [12] is that they directly compare their precision with Demircan and Kahramanli [20] and declare that they increase speaker-independent accuracy of 92.86% to 95.89%. However, we think that these frames are not comparable because only a subset of EMO-DB is utilized in [20].

To equitably compare our framework to [20], we eliminate some samples from the dataset based on the following hypothesis: if Model A, Model C, and Model D predict the same class for a particular sample, and the true class of that sample is not the predicted one, and then we remove the sample from the dataset. As a result of this procedure, the number of the remaining audio files in the dataset also becomes 520. In this case, the classification accuracy of Model E increases to 95.71%. Therefore, with 520 samples, Model E achieves a higher performance than the one proposed in [20]. This comparison is given in Fig. 11.

4.4.3. IEMOCAP dataset

Fig. 12 shows the comparison of our work to the previous speaker-independent approaches using the improvised audio samples of IEMOCAP database with the same number of emotion classes. Our model outperforms both the work of Lee and Tashev [35] and the work Tripathi and Beigi [36]. At the same time, the

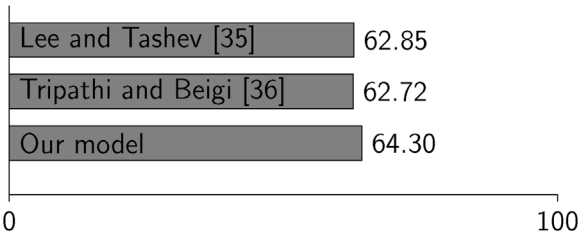

Fig. 12. Comparison with the previous works using IEMOCAP database.

configuration of our model is significantly simpler in comparison with the LSTM approach of [36] and the RNN approach of [35].

While various previous frameworks report their performances on these three datasets, the reason for not including more previous frameworks for comparison mainly comes from incomparable evaluation methods or different numbers of emotion classes used by the previous approaches. For instance, the work of Chen et al. [37] utilizes unweighted average recall (UAR) as opposed to the accuracy measurement used in this paper. Zheng et al. [38] use a different number of emotion classes, while Kim et al. [39] employ audio features in combination with visual features. Others, like Lakomkin et al. [40] train and test their models on both improvised and acted data.

5. Discussion and conclusion

Speech emotion recognition is a complex task, which involves two essential problems: feature extraction and classification. In this paper, we propose a new framework for speech emotion recognition using one-dimensional deep CNN with the combination of five different audio features as input data. Our model outperforms state-of-the-art approaches for the RAVDESS and IEMOCAP datasets. For EMO-DB, we incrementally present a set of models based on our initial framework to improve the performance. Our best-performing model (Model E) achieves higher accuracy than all previous work, except one approach [12]. However, our approach compares favorably with that one in terms of generality, simplicity, and applicability. All of the proposed models work directly with raw sound data without the need for a conversion to visual representations, unlike some other approaches, [2,7,12,22], for example.

Nevertheless, we think that more research on this topic can be done. The inclusion of other types of features or the use of an auxiliary neural network to achieve high-level features could significantly improve the accuracy of our models. In addition, applying more comprehensive sets of data augmentation techniques to increase the size of training data could also improve performance. Finally, the use of additional layers of LSTM can also improve the accuracy of the classification.

It should be noted that the order of stacking sound features play an important role in the final performance. Therefore, changing the order could result in different classification accuracy. The order we used in this paper was discovered experimentally by identifying the best performance using a small subset of the datasets. Determining the optimal feature order is also one of our future plans.

CRediT authorship contribution statement

Dias Issa: Conceptualization, Methodology, Software, Visualization, Investigation, Validation, Writing - original draft, Data curation, Writing - review & editing. **M. Fatih Demirci:** Methodology, Supervision, Project administration, Conceptualization, Writing - original draft, Writing - review & editing. **Adnan Yazici:** Methodology, Supervision, Project administration, Conceptualization, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare no conflicts of interest.

References

[1] I.T. Kun Han, D. Yu, Speech emotion recognition using deep neural network and extreme learning machine, Interspeech (2014) 223–227.
[2] A.M. Badshah, J. Ahmad, N. Rahim, S.W. Baik, Speech emotion recognition from spectrograms with deep convolutional neural network, in: 2017 International Conference on Platform Technology and Service (PlatCon), IEEE, 2017, pp. 1–5.
[3] S. Mittal, S. Agarwal, M.J. Nigam, Real time multiple face recognition: a deep learning approach, in: Proceedings of the 2018 International Conference on Digital Medicine and Image Processing, ACM, 2018, pp. 70–76.
[4] H.-S. Bae, H.-J. Lee, S.-G. Lee, Voice recognition based on adaptive mfcc and deep learning, in: 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), IEEE, 2016, pp. 1542–1546.
[5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 770–778.
[6] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, Y.-H. Chen, Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds, in: ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 5866–5870.
[7] W. Lim, D. Jang, T. Lee, Speech emotion recognition using convolutional and recurrent neural networks, in: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific, IEEE, 2016, pp. 1–4.
[8] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, S. Zafeiriou, Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 5200–5204.
[9] S.R. Livingstone, 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 13 (2018) e0196391.
[10] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, Iemocap: interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (2008) 335.
[11] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of german emotional speech, Ninth European Conference on Speech Communication and Technology (2005).
[12] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1d & 2d cnn lstm networks, Biomed. Signal Process. Control 47 (2019) 312–323.
[13] Y. Niu, D. Zou, Y. Niu, Z. He, H. Tan, Improvement on speech emotion recognition based on deep convolutional neural networks, Proceedings of the 2018 International Conference on Computing and Artificial Intelligence (2018) 13–18.
[14] L. Tarantino, P.N. Garner, A. Lazaridis, Self-attention for speech emotion recognition, Proc. Interspeech 2019 (2019) 2578–2582.
[15] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. André, C. Busso, L.Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan, et al., The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, IEEE Trans. Affect. Comput. 7 (2015) 190–202.
[16] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, B. Schuller, Towards robust speech emotion recognition using deep residual networks for speech enhancement, Proc. Interspeech 2019 (2019) 1691–1695.
[17] B.W. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J.K. Burgoon, A. Baird, A.C. Elkins, Y. Zhang, E. Coutinho, K. Evanini, The interspeech 2016 computational paralinguistics challenge: deception, sincerity & native language, Interspeech 2016 (2016) 2001–2005.
[18] N. Weißkirchen, R. Bock, A. Wendemuth, Recognition of emotional speech with convolutional neural networks by means of spectral estimates, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), IEEE, 2017, pp. 50–55.
[19] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, S. Narayanan, Data augmentation using gans for speech emotion recognition, Proc. Interspeech 2019 (2019) 171–175.
[20] S. Demircan, H. Kahramanli, Application of fuzzy c-means clustering algorithm to spectral features for emotion classification from speech, Neural Comput. Appl. 29 (2018) 59–66.
[21] S. Yoon, S. Byun, K. Jung, Multimodal speech emotion recognition using audio and text, in: 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2018, pp. 112–118.
[22] Z. Huang, M. Dong, Q. Mao, Y. Zhan, Speech emotion recognition using cnn, Proceedings of the 22nd ACM International Conference on Multimedia (2014) 801–804.
[23] S. Wu, T.H. Falk, W.-Y. Chan, Automatic speech emotion recognition using modulation spectral features, Speech Commun. 53 (2011) 768–785.
[24] A.S. Lampropoulos, G.A. Tsihrintzis, Evaluation of mpeg-7 descriptors for speech emotional recognition, in: 2012 Eighth International Conference on

Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), IEEE, 2012, pp. 98–101.
[25] K. Wang, N. An, B.N. Li, Y. Zhang, L. Li, Speech emotion recognition using fourier parameters, IEEE Trans. Affect. Comput. 6 (2015) 69–75.
[26] P. Shegokar, P. Sircar, Continuous wavelet transform based speech emotion recognition, in: 2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS), IEEE, 2016, pp. 1–8.
[27] B. Zhang, E.M. Provost, G. Essi, Cross-corpus acoustic emotion recognition from singing and speaking: a multi-task learning approach, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 5805–5809.
[28] Y. Zeng, H. Mao, D. Peng, Z. Yi, Spectrogram based multi-task audio classification, Multimed. Tools Appl. (2017) 1–18.
[29] A.S. Popova, A.G. Rassadin, A.A. Ponomarenko, Emotion recognition in sound, in: International Conference on Neuroinformatics, Springer, 2017, pp. 117–124.
[30] B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: audio and music signal analysis in python, Proceedings of the 14th Python in Science Conference (2015) 18–25.
[31] S.S. Stevens, J. Volkmann, E.B. Newman, A scale for the measurement of the psychological magnitude pitch, J. Acoust. Soc. Am. 8 (1937) 185–190.
[32] H. Beigi, Fundamentals of Speaker Recognition, Springer Science & Business Media, 2011.
[33] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, L.-H. Cai, Music type classification by spectral contrast feature, in: 2002 IEEE International Conference on Multimedia and Expo, 2002. ICME'02. Proceedings, vol. 1, IEEE, 2002, pp. 113–116.
[34] C. Harte, M. Sandler, M. Gasser, Detecting harmonic change in musical audio, Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia (2006) 21–26.
[35] J. Lee, I. Tashev, High-level feature representation using recurrent neural network for speech emotion recognition, Sixteenth Annual Conference of the International Speech Communication Association (2015).
[36] S. Tripathi, H.S.M. Beigi, Multi-Modal Emotion Recognition on IEMOCAP Dataset Using Deep Learning, 2018 CoRR abs/1804.05788.
[37] M. Chen, X. He, J. Yang, H. Zhang, 3-d convolutional recurrent neural networks with attention model for speech emotion recognition, IEEE Signal Process. Lett. 25 (2018) 1440–1444.
[38] W. Zheng, J. Yu, Y. Zou, An experimental study of speech emotion recognition based on deep convolutional neural networks, in: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2015, pp. 827–831.
[39] Y. Kim, H. Lee, E.M. Provost, Deep learning for robust feature generation in audiovisual emotion recognition, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 3687–3691.
[40] E. Lakomkin, C. Weber, S. Magg, S. Wermter, Reusing Neural Speech Representations for Auditory Emotion Recognition, 2018 arXiv preprint arXiv:1803.11508.