Deep Convolutional Neural Networks for Feature Extraction in Speech Emotion Recognition

Chapter *in* Lecture Notes in Computer Science · June 2019

DOI: 10.1007/978-3-030-22643-5_9

Panikos Heracleous
National Institute of Advanced Industrial Science and Technology
89 PUBLICATIONS **657** CITATIONS

Yasser Mohammad
NEC Corporation
134 PUBLICATIONS **1,107** CITATIONS

Akio Yoneyama
KDDI Research
84 PUBLICATIONS **782** CITATIONS

Deep Convolutional Neural Networks for Feature Extraction in Speech Emotion Recognition

Panikos Heracleous[1], Yasser Mohammad[2], and Akio Yoneyama[1]

[1] KDDI Research, Inc.
2-1-15 Ohara, Fujimino-shi, Saitama 356-8502, Japan
{pa-heracleous,yoneyama}@kddi-research.jp
[2] Artificial Intelligence Research Center, AIST
2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan
yasserm@aun.edu.eg

Abstract. Speech emotion recognition is a task designed to automatically identify human emotions in spoken utterances. The current study focuses on speech emotion recognition based on deep convolutional neural networks (DCNNs) and extremely randomized trees. Specifically, we propose a method based on DCNN, which extracts informative features from the speech signal, and those features are then used by an extremely randomized trees classifier for emotion recognition. The CNNs are a special variant of conventional feed-forward deep neural networks (DNNs), and have been used in many speech applications. Another method is also proposed which integrates DCNN with i-vectors for emotion recognition. The proposed methods were evaluated using the state-of-the-art English IEMOCAP and FAU Aibo German emotional corpora for the recognition of four and five emotions, respectively. When using the IEMOCAP English corpus and DCNN with extremely randomized trees, a 63.9% unweighted average recall (UAR) was obtained. In the case of using the German children's Aibo corpus, a 61.8% UAR was achieved. These results are very promising showing the effectiveness of the proposed methods in speech emotion recognition. The proposed methods were compared with a baseline approach based on support vector machines (SVM), and they showed superior performance.

Keywords: speech emotion recognition, deep convolutional neural networks, informative features, i-vectors, extremely randomized trees

1 Introduction

Emotion recognition plays an important role in human-computer interaction and is attracting a high level of attention because of its real world applications [1]. Emotion recognition can be applied in human-robot interaction to detect the user's emotions, or in call-centers to identify the caller's emotional state. In particular, in cases of emergency, emotion recognition can provide feedback to the operator so that he or she can respond in an appropriate way. Furthermore, the emotional state of the caller may be very informative concerning the level of customer satisfaction.

The current study focuses on emotion recognition based on the speech modality. A method is proposed which uses DCNN to extract informative features from each layer of the network, and the extracted features are then flattened and used by extremely randomized trees [2] for emotion recognition. Extremely randomized trees are similar to random forest [3], but with random tree splitting. The motivation for using extremely randomized trees is due to the lower computational cost, and additionally the method shows a high level of performance in the case of a small number of features.

A CNN [4, 5] is a special variant of conventional neural networks consisting of convolution and pooling layers. Many studies have reported results for speech emotion recognition [6], image classification [7], and sentence classification [8] based on CNNs. In particular, CNNs are very popular in image classification and most of the recent related studies are based on CNNs. In the current study, CNNs are used because of their simplicity compared to a conventional feed-forward DNN. Due to parameter sharing, computational and memory costs are lower.

In addition to DCNN with extremely randomized trees, another method based on conventional CNNs is also experimentally investigated. In this case, instead of using extremely randomized trees for classification, a fully connected layer is added on the top of convolutional layers of the DCNN, and emotion recognition is performed using the features of the last layer. When using the two methods, the neural networks are fed with frame-level spectral features. Furthermore, for more comprehensive investigations, DCNNs fed with i-vectors [9] are also applied in speech emotion recognition. In the i-vector paradigm, the spoken utterance is represented by a small number of factors, which comprise the variability of speaker, channel, emotion, or language. Although i-vectors have been successfully used in speech emotion recognition [10–12], the integration of deep learning (DL) and i-vectors has not been investigated comprehensively so far, and only very few studies addressed this issue [13]. As a result, DL and i-vectors for speech emotion recognition are still an open research area and further investigations are necessary.

In a previous study [14], the authors demonstrated experimental results on far-field speech emotion recognition using a DCNN for feature extraction and extremely randomized trees for classification. In the current study, the DCNN architecture is simplified by excluding network pre-training, and, also, by using the features of all convolutional layers to select the learned features used in classification. The motivation for using the features from all layers lies in the fact that lower-level features may be also very informative resulting in higher classification rates when included. Furthermore, in the proposed methods are evaluated using also the English IEMOCAP corpus [15] for classification of four emotions.

Regarding the emotional data used, the proposed methods are evaluated using the state-of-the-art English IEMOCAP and German FAU Aibo [16] corpora for the classification of four and five emotions, respectively. For comparison purposes, a baseline speech emotion recognition experiment using the popular SVM classifier [17] with i-vectors was also conducted.

2 Related Work

Previously, several studies addressed the problem of emotion recognition using different modalities, classifiers, and feature extraction methods. Emotion recognition can be performed using speech signal [18], visual/facial information [19], electroencephalography (EEG) signals [20], and also using physiological signals such as, blood volume pulse (BVP), electromyography (EMG), skin conductance (SC), skin temperature (SKT) and respiration (RESP) [21].

Speech emotion recognition using Gaussian mixture models (GMMs) was reported in [22, 23]. In [24], hidden Markov model- (HMM) based speech emotion recognition was presented. SVM is among the most popular classifier used in speech emotion recognition [25, 26]. More recent studies are based on neural networks (NN) [27, 28]. Currently, speech emotion recognition using deep neural networks is being investigated [29, 30].

Mel-frequency cepstral coefficients (MFCC) [31] are very commonly and widely used features in speech emotion recognition. In addition to MFCC features, shifted delta-cepstral (SDC) coefficients [32, 33] can also be applied. Originally, SDC coefficients were used in spoken language identification showing superior performance compared with the sole use of MFCC features. In the current study, SDC coefficients are concatenated with MFCC features to form the basic feature vectors. In many recent studies, low-level descriptors (LLD) and functionals [34] are used as features. Considering the success of i-vectors in speaker recognition and spoken language identification [35], however, studies on speech emotion recognition using i-vectors and neural networks have also been presented in a small number of studies. In the current study, the i-vectors used by conventional CNN architecture are extracted from concatenated MFCC features and SDC coefficients.

3 Methods

3.1 Emotional Corpora

In the current study, the FAU Aibo and the IEMOCAP data are used. The FAU Aibo German corpus consists of 9 hours of speech uttered by 51 children while interacting with Sony's Aibo robot. The spontaneous Aibo speech was recorded using a close-talking microphone, and was annotated into 11 categories by five human annotators. However, in the current study, the 5-class task is considered, and data for the emotions angry, emphatic, joyful, neutral, and rest were used for the classification. For training, 590 utterances for each emotion were used, and for testing, 299 utterances for each emotion were used. The data were randomly selected from the entire data set.

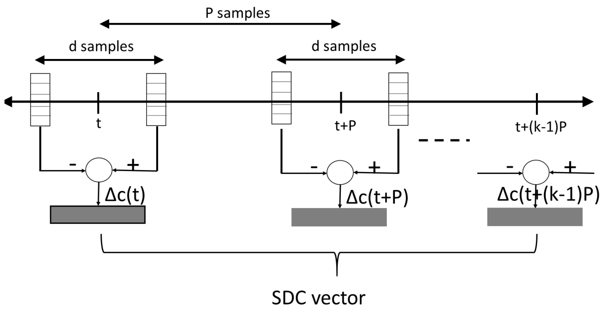

Fig. 1. Computation of shifted delta cepstral (SDC) coefficients.

The IEMOCAP database was collected at the SAIL lab of the University of Southern California. It contains 12 hours of audiovisual data produced by 10 actors. The data were annotated into categorical labels as well as dimensional labels. In the current study, categorical labels were used to classify the emotional states of neutral, happy, angry, and sad. To avoid unbalanced data, 250 utterances for training and 70 utterances for testing randomly selected from each emotion were used.

3.2 Feature Extraction

Cepstral Features MFCCs are the basic features used in the current study. The MFCC features are extracted every 10 ms using a window-length of 20 ms.

In addition to MFCC features, SDC coefficients are also used. The SDC feature vectors are obtained by concatenating delta cepstra across multiple, and they are described by the *N* number of cepstral coefficients, *d* time advance and delay, *k* number of blocks concatenated for the feature vector, and *P* time shift between consecutive blocks. For each SDC final feature vector, *kN* parameters are used. In contrast, in the case of conventional cepstra and delta cepstra feature vectors, *2N* parameters are used. The SDC is calculated as follows:

$$\Delta c(t + iP) = c(t + iP + d) - c(t + iP - d) \tag{1}$$

The final vector at time t is given by the concatenation of all $\Delta c(t + iP)$ for all $0 \leq i < k$, where $c(t)$ is the original feature value at time t. Fig. 1 shows the computation procedure for the SDC coefficients. Therefore, in modeling the emotions being classified, this study also used MFCC features, concatenated with SDC coefficients to form feature vectors of length 112. In the case of using CNN and i-vectors, the concatenated MFCC/SDC features were used to extract the i-vectors used by the classifier. In the other two cases, neural networks fed by blocks of MFCC/SDC (center frame ± 10 frames) features were applied.

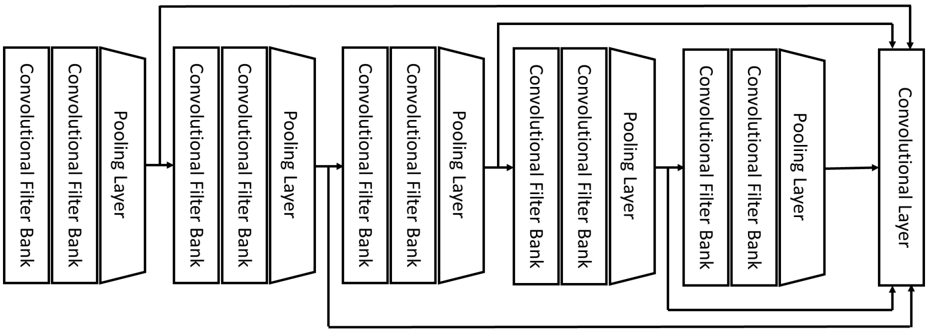
Convolutional Filter Bank
Convolutional Filter Bank
Pooling Layer
Convolutional Filter Bank
Convolutional Filter Bank
Pooling Layer
Convolutional Filter Bank
Convolutional Filter Bank
Pooling Layer
Convolutional Filter Bank
Convolutional Filter Bank
Pooling Layer
Convolutional Filter Bank
Convolutional Filter Bank
Pooling Layer
Convolutional Layer

Fig. 2. The architecture of the deep feature extractor along with the classifier used during feature learning.

I-vector features A widely used classification approach in speaker recognition is based on GMMs with universal background models (UBM). In this approach, each speaker model is created by adapting the UBM using maximum a posteriori (MAP) adaptation. A GMM supervector is constructed by concatenating the means of the adapted models. As in speaker recognition, GMM supervectors can also be used for emotion classification.

To overcome the limitations of the high dimensionality of GMM supervectors, the i-vectors model the variability contained in the supervectors with a small set of factors. In this case, an input utterance can be modeled as:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{2}$$

where $\mathbf{M}$ is the emotion-dependent supervector, $\mathbf{m}$ is the emotion-independent supervector, $\mathbf{T}$ is the total variability matrix, and $\mathbf{w}$ is the i-vector. Both the total variability matrix and emotion-independent supervector are estimated from the complete set of training data.

Proposed feature extraction and selection approach In this paper, DCNN for learning informative features from the speech signal that is then used for emotion classification is investigated. The MFCC and SDC features are calculated using overlapping windows with a length of 20 ms. This generates a multidimensional time-series that represent the data for each session. The proposed method is a simplified version of the method recently proposed in [36] for activity recognition using mobile sensors.

The proposed classifier consists of a DCNN followed by extremely randomized trees instead of the standard fully connected classifier. The motivation for using extremely randomized trees lies in previous observations showing their effectiveness in the case of a small number of features. The network architecture is shown in Fig. 2, and consists of a series of five blocks, each of which consists of two convolutional layers (64 5×5) followed by a max-pooling layer (2×2). Outputs from each block are then combined and flattened to represent the learned features.

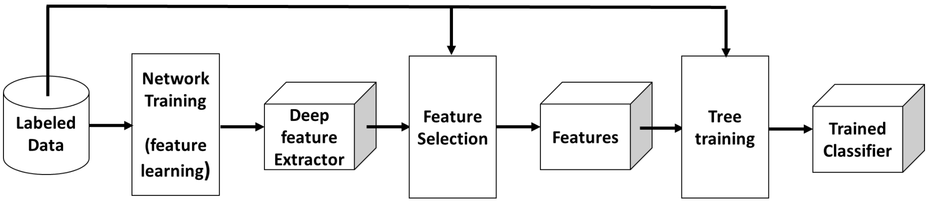

Fig. 3. The proposed training process showing the three stages of training and the output of each stage.

The main idea behind the proposed approach is to use deep networks as feature learners only, but not classifiers, and then utilize feature selection to determine a small set of neurons that provide maximal information for an efficient activity recognizer. This approach combines elements from standard deep learning methodology. However, the approach treats the problem of generating an efficient feature extractor given an accurate, yet inefficient deep neural network, as a feature selection problem instead of using standard neural network compression techniques like optimal brain damage [37].

The training process is shown in Fig. 3, and consists of two main stages: feature learning and feature selection. In the feature learning stage, labeled data are used to train a deep neural architecture as a feature extractor. The goal of this step is to produce a large set of features that are as informative about the emotion recognition problem as possible. To achieve that, no attempt is made at this stage to optimize the computational cost, resulting thus in a slow feature extractor. Feature selection is then used to keep a small fraction of the learned features in the fast feature extractor upon which a classifier can be trained to solve emotion recognition problem. During recognition, only the fast feature extractor and final classifier are kept.

During training, a classifier consisting of a fully connected network (two layers with 16 and 16 $ReLU$ neurons) followed by a sigmoid layer with the number of target emotions takes the output of all layers in the architecture through bypass connections. These bypass connections allow the fully connected network to utilize low level features extracted in early convolution filters instead of having to rely on the higher level features learned by the capping CNN. The obvious problem with this design is that the number of inputs to the classifier increases dramatically. For this reason, we employ L_1 regularization during the training process to generate a sparse representation in the classifier by setting most of the weights in its earlier layers to zero. This reduces the effective input size. Other forms of pruning can be used at this stage to reduce the computational cost of the classifier [38].

Furthermore, during training, neurons in early stages are subjected to multiple updates at every gradient-based weight update due to the use of bypassing connections. For this reason, the learning rate used for a neuron in layer i (η_i)

is calculated from the base learning rate (η) as:

$$\eta_i = \frac{\eta}{n - i}, \tag{3}$$

where n is the total number of layers.

Although, it is possible to just use the described classifier for emotion recognition (i.e., conventional DCNN), in the proposed approach, the final shallow classifier, after this training is completed, is removed and replaced with an extremely randomized trees classifier trained on a subset of the neurons that is selected.

The feature extractor learned through the proposed method will be impractical for a real applications (e.g., applications on a mobile devices) due to its large size resulting from using bypassing connections from all neurons to the output. This problem can be alleviated while improving the generalization capacity of the system using feature selection.

Selecting appropriate bypass connections from the slow feature selector can be thought of as a standard feature selection problem which is solved in this paper using a multi-criteria wrapper method. Each feature (neuronal output i) is assigned a total *quality* ($Q\left(i\right)$) according to Equation 4 where $\bar{I}_j\left(i\right)$ is z-score normalized feature *importance* ($I_j\left(i\right)$) according to a base feature selection method.

$$Q\left(i\right) = \sum_{j=0}^{n_f} w_j \bar{I}_j\left(i\right), \tag{4}$$

The raw *importance* measure is calculated as a weighted summation of multiple base feature selector importance measures after z-score normalization. In this work, we utilize two base selectors: randomized logistic regression [39], and extremely randomized trees. Random linear regression (RLR) estimates feature importance by randomly selecting subsets of training samples and fitting them using a L_1 sparsity inducing penalty that is scaled for a random set of coefficients. The features that appear repeatedly in such selections (i.e., features with high coefficients) are assumed to be more important and are given higher scores. The second feature selector employs extremely randomized trees. During fitting decision trees, features that appear at lower depths are generally more important. By fitting several such trees, feature importances can be estimated as the average depth of each feature in the trees.

Feature selection uses n-fold cross validation to select an appropriate number of neurons to keep in the final (i.e., fast) feature extractor. For each fold, the quality of each neuron is calculated using Equation 4 employing its training set and then an extremely randomized tree classifier is fitted to the training set and evaluated on the validation set. The process is repeated recursively on the top-half of the neurons until a single neuron is kept in the feature set. The number of features/neurons that maximizes the F_1-measure on the validations sets is finally kept.

Table 1. Recalls for individual emotions when using MFCC features with/without SDC coefficients [%] (FAU Aibo).

Feature for	Emotion					
i-vector extraction	Angry	Emphatic	Joyful	Neutral	Rest	UAR
MFCC	46.5	35.1	53.5	33.8	28.8	39.5
MFCC+SDC	55.5	62.9	71.2	68.2	41.1	59.8

Table 2. EERs for individual emotions when using MFCC features with/without SDC coefficients [%] (FAU Aibo).

Feature for	Emotion					
i-vector extraction	Angry	Emphatic	Joyful	Neutral	Rest	Average
MFCC	32.8	35.0	26.8	35.4	41.8	34.4
MFCC+SDC	20.1	19.7	16.1	19.7	29.8	21.8

4 Results

This section presents the results obtained using the FAU Aibo and IEMOCAP corpora. The proposed method based on DCNN and extremely randomized trees is compared with three other classifiers namely, DCNN with a fully-connected layer on top fed with MFCC/SDC features, DCNN fed with i-vectors, and SVM fed also with i-vectors. In this section, the improvements when using SDC coefficients along with MFCC features compared to the sole use of MFCC are also described.

For evaluation, the equal error rate (EER) and the UAR are used. The UAR is defined as the mean of the recalls of the individual classes.

4.1 Emotion recognition using the German FAU Aibo corpus

Table 1 shows the recalls and the UAR when using DCNN with/without SDC coefficients in the i-vector extraction. The results show that when using MFCC only in the i-vector extraction, the UAR was as low as 39.5%. When MFCC features were concatenated with SDC coefficients, the UAR improved to 59.8% showing a 20.3% absolute improvement. As show, the emotion *joyful* shows superior performance, and the emotion *rest* shows the lowest recall. A possible reason might be the fact that the class *rest* consists of several emotions not belonging to other classes. The results obtained when using also SDC are very promising and superior to the results obtained in similar studies [40]. The results also show the effectiveness of integrating i-vectors and CNN for speech emotion recognition using only 590 training i-vectors for each emotion.

Table 2 shows the EERs of the five emotions when using the FAU Aibo corpus. When using MFCC features only in the i-vectors extraction, the average

Table 3. Recalls for individual emotions when using three different classifiers [%] (FAU Aibo).

Classifier	Emotion					
	Angry	Emphatic	Joyful	Neutral	Rest	UAR
Deep convolutional neural networks + Extremely randomized trees	62.9	63.5	61.9	60.2	60.5	**61.8**
Deep convolutional neural networks + Fully connected layer	51.5	53.8	51.8	52.5	52.5	52.4
Support vector machine	55.2	44.5	62.2	35.5	46.5	48.8

EER was 34.4%. When SDC coefficients were also concatenated, the EER improved to 21.8% showing an absolute reduction of 13.4%. The lowest EER was obtained in the case of the emotion *joyful*, and the highest EER was achieved in the case of the emotion *rest*. Tables 1 and 2 show the effectiveness of using SDC coefficients in speech emotion recognition. Therefore, in the following experiments, MFCC features concatenated with SDC coefficients will be used.

Table 3 shows the recalls of the individual classes, and also the UAR obtained. As shown, using DCNN for feature extraction and extremely randomized trees for classification, a 61.8% UAR was obtained. This is the highest UAR among the four classifiers. In the case of using conventional DCNN with a fully connected layer on top, the UAR was 51.4%. Finally, when using SVM, a 48.8% UAR was achieved. The results show, that in the two cases of using DCNN with extremely randomized trees and with a fully-connected layer, similar recalls were obtained across the five emotions. In the case of using SVM with i-vector features, the emotion *joyful* was classified with the highest recall, and the emotions *neutral* and *rest* showed the lowest recalls. This is similar to the case when DCNN with i-vectors was used. Previous studies reported that when short utterances were used for speaker recognition, the extracted i-vectors become unreliable [41]. Also, in the case of using i-vectors, the optimal case is when long training and long test utterances are used. It may happen, therefore, that in the current study training and test utterances of different lengths were randomly selected resulting in a higher recall variability. Note, however, that when using DCNN with i-vectors, the second highest UAR was obtained, and i-vectors can still be considered to be a very effective feature extraction method in speech emotion recognition. Tables 4, 5, 6, and 7 show the confusion matrices when using the four classifiers. As shown, a higher variability in misclassification is obtained when i-vectors were used.

4.2 Emotion recognition using the English IEMOCAP corpus

Table 8 shows the recalls of the four emotions in the case of using the IEMOCAP corpus. In this case, DCNN fed with i-vectors were used. For i-vector extraction

Table 4. Confusion matrix [%] of five emotions recognition when using DCNN with i-vectors (FAU Aibo).

	Angry	Emphatic	Joyful	Neutral	Rest
Angry	55.5	18.1	7.0	6.4	13.0
Emphatic	11.4	62.9	0.3	18.4	7.0
Joyful	2.7	2.7	71.2	5.7	17.7
Neutral	1.0	13.0	4.1	68.2	13.7
Rest	11.4	10.0	20.1	17.4	41.1

Table 5. Confusion matrix [%] of five emotions recognition when using DCNN and extremely randomized trees (FAU Aibo).

	Angry	Emphatic	Joyful	Neutral	Rest
Angry	62.9	13.0	9.0	6.7	8.4
Emphatic	8.7	63.5	11.7	10.0	6.1
Joyful	11.4	9.0	61.9	9.7	8.0
Neutral	9.0	11.1	11.0	60.2	8.7
Rest	12.4	6.4	12.7	8.0	60.5

Table 6. Confusion matrix [%] of five emotions recognition when using DCNN with a fully connected layer (FAU Aibo).

	Angry	Emphatic	Joyful	Neutral	Rest
Angry	51.5	15.1	11.4	10.0	12.0
Emphatic	10.8	53.8	14.7	12.7	8.0
Joyful	13.4	12.0	51.8	12.1	10.7
Neutral	11.7	13.4	12.4	52.5	10.0
Rest	13.7	8.0	16.4	9.4	52.5

Table 7. Confusion matrix [%] of five emotions recognition when using SVM with i-vectors (FAU Aibo).

	Angry	Emphatic	Joyful	Neutral	Rest
Angry	55.2	15.7	6.0	7.0	16.1
Emphatic	16.7	44.5	3.3	17.1	18.4
Joyful	3.3	2.7	62.2	4.7	27.1
Neutral	7.7	12.4	13.6	35.5	30.8
Rest	11.4	9.4	18.7	14.0	46.5

MFCC features alone and also MFCC features concatenated with SDC coefficients were used. As shown, when using MFCC features only, an UAR of 55.5% was obtained. When SDC coefficients were also concatenated, the UAR improved to 62.0%. The results also show that in most of cases (three out of four) the SDC

Table 8. Recalls for individual emotions when using MFCC features with/without SDC coefficients [%] (IEMOCAP).

Feature for	Emotion				
i-vector extraction	Neutral	Happy	Angry	Sad	UAR
MFCC	46.0	40.0	66.0	70.0	55.5
MFCC+SDC	48.0	36.0	88.0	76.0	62.0

Table 9. EER for individual emotions when using MFCC features with/without SDC coefficients [%] (IEMOCAP).

Feature for	Emotion				
i-vector extraction	Neutral	Happy	Angry	Sad	Average
MFCC	30.0	36.0	18.0	22.0	26.5
MFCC+SDC	32.0	26.7	12.0	18.0	22.2

Table 10. Recalls for individual emotions when using three different classifiers [%] (IEMOCAP).

Classifier	Emotion				
	Neutral	Happy	Angry	Sad	UAR
Deep convolutional neural networks + Extremely randomized trees	55.7	68.6	67.1	64.3	**63.9**
Deep convolutional neural networks + Fully connected layer	48.6	64.3	61.4	62.9	59.3
Support vector machine	26.0	39.0	45.0	37.0	36.8

coefficients resulted in higher recalls. The highest rates of recognition were for the *angry* and *sad* emotions. In contrast, the lowest recall was achieved in the case of the emotion *happy*.

Table 9 shows the EERs when using DCNN fed with i-vectors. In the case of using MFCC features only, the average EER was 26.5%. When SDC coefficients were also used, a 22.2% EER was obtained. The results show that when also using SDC coefficients, significant improvements were obtained. Therefore, in the following experiments, MFCC features concatenated with SDC coefficients will be considered.

Table 10 shows the recalls and the UARs in the case of the IEMOCAP corpus and when using three different classifiers. As shown, when using DCNN for feature extraction and extremely randomized trees for classification, a 63.9% UAR was obtained, which is the highest among the four classifiers. This result is

Table 11. Confusion matrix [%] of four emotions recognition when using DCNN with i-vectors (IEMOCAP).

	Neutral	Happy	Angry	Sad
Neutral	48.0	18.0	14.0	20.0
Happy	34.0	36.0	14.0	16.0
Angry	4.0	4.0	88.0	4.0
Sad	12.0	8.0	4.0	76.0

Table 12. Confusion matrix [%] of four emotions recognition when using DCNN with extremely randomized trees (IEMOCAP).

	Neutral	Happy	Angry	Sad
Neutral	55.7	15.7	20.0	8.6
Happy	14.3	68.6	10.0	7.1
Angry	8.6	14.3	67.1	10.0
Sad	12.9	11.4	11.4	64.3

Table 13. Confusion matrix [%] of four emotions recognition when using DCNN with a fully-connected layer (IEMOCAP).

	Neutral	Happy	Angry	Sad
Neutral	48.6	12.8	24.3	14.3
Happy	15.7	64.3	11.4	8.6
Angry	14.3	12.9	61.4	11.4
Sad	15.7	12.9	8.6	62.8

Table 14. Confusion matrix [%] of four emotions recognition when using SVM with i-vectors (IEMOCAP).

	Neutral	Happy	Angry	Sad
Neutral	37.1	32.9	14.3	15.7
Happy	8.6	55.7	17.1	18.6
Angry	10.0	15.7	64.3	10.0
Sad	8.5	24.3	14.3	52.9

very promising and superior to the results obtained in similar studies [42, 43]. The results also show the effectiveness of the proposed method when DCNN is used for informative feature extraction. When using DCNN with a fully connected layer on top, a 59.3% UAR was achieved. Finally, the UAR in the case of SVM with i-vectors was as low as 36.8%. Tables 11, 12, 13, and 14 show the confusion matrices in the case of the IEMOCAP corpus and when using the four classifiers described previously. As shown, the emotion *neutral* is classified with the lowest recall in all cases.

5 Discussion

A limitation of the current study is the small volume of training data used in the classification experiments. Specifically, in the case of using the FAU Aibo corpus, 590 training utterances for each emotion were used, and in the case of using the IEMOCAP corpus, 250 training utterances were used, respectively. Considering that DL based methods require a large amount of training data for accurate parameter estimation, further improvements may be possible by increasing the amount of data. The features used in the current study were based on MFCC and SDC coefficients, and also on the i-vectors. Although, several alternatives were considered (e.g., bottleneck features, LLD, etc.), well-known and very effective features were selected. Also, in particular the authors were interested in investigating the use of i-vectors with CNN due to the very small number of studies that have addressed this issue.

6 Conclusions

The current study focused on speech emotion recognition based on deep learning. We proposed a method based on DCNN, which extracts informative features used by extremely randomized trees for emotion recognition. When using the German FAU Aibo corpus for the recognition of five emotions, the proposed method achieved a 61.8% UAR. In the case of the IEMOCAP corpus, a 63.9% UAR was obtained. These results are very promising and show the effectiveness of the proposed method in speech emotion recognition. Additionally, several other classification and features extraction methods were experimentally investigated. The proposed method, however, showed superior performance. Currently, speech emotion recognition in adverse environments is being investigated.

References

1. Busso, C., Bulut, M., Narayanan, S.: Toward Effective Automatic Recognition Systems of Emotion in Speech. In Gratch, J., Marsella, S., eds.: Social emotions in nature and artifact: emotions in human and human-computer interaction. Oxford University Press, New York, NY, USA (November 2013) 110–127
2. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees . Machine Learning **63, Issue 1** (2006) 3–42
3. Ho, T.K.: Random Decision Forests. In Proc. of the 3rd International Conference on Document Analysis and Recognition (1995) 278–282
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., eds.: Advances in Neural Information Processing Systems 25, Curran Associates, Inc. (2012) 1097–1105
5. Abdel-Hamid, O., Mohamed, A.r., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing **22** (2014) 1533–1545

6. Lim, W., Jang, D., Lee, T.: Speech Emotion Recognition Using Convolutional and Recurrent Neural Networks. in Proc. of Signal and Information Processing Association Annual Summit and Conference (APSIPA) (2016)
7. Rawat, W., Wang, Z.: Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review . Neural Communication **29** (2017) 23522449
8. Kim, Y.: Convolutional Neural Networks for Sentence Classification. in Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014) 1746–1751
9. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-End Factor Analysis for Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing **19(4)** (2011) 788–798
10. Gomes, J., El-Sharkawy, M.: i-Vector Algorithm with Gaussian Mixture Model for Efficient Speech Emotion Recognition. in Proc. of International Conference on Computational Science and Computational Intelligence (CSCI) (2015) 476–480
11. Liu, R.X.Y.: Using i-vector space model for emotion recognition. in Proc. of Interspeech (2012) 2227–2230
12. Gamage, K.W., Sethu, V., Le, P.N., Ambikairajah, E.: An i-vector GPLDA System for Speech based Emotion Recognition. in Proc. of APSIPA Annual Summit and Conference (2015) 289–292
13. Zhang, T., Wu, J.: Speech Emotion Recognition with i-vector Feature and RNN model. in Proc. of ChinaSIP (2015) 524–528
14. Heracleous, P., Mohammad, Y., Takai, K., Yasuda, K., Yoneyama, A., Sugaya, F.: A Study on Far-field Emotion Recognition Based on Deep Convolutional Neural Networks. in Proc. of International Conference on Computational Linguistics and Intelligent Text Processing (2018)
15. Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S., Narayanan, S.: IEMOCAP: Interactive emotional dyadic motion capture database. Journal of Language Resources and Evaluation (2008) 335–359
16. Steidl, S.: Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech. Logos Verlag, Berlin (2009)
17. Cristianini, N., S.-Taylor, J.: Support vector machines. Cambridge University Press, Cambridge (2000)
18. Basu, S., Chakraborty, J., Bag, A., Aftabuddin, M.: A Review on Emotion Recognition Using Speech. International Conference on Inventive Communication and Computational Technologies (ICICCT) (2017) 109–114
19. Metallinou, A., Busso, C., Lee, S., Narayanan, S.: Visual Emotion Recognition Using Compact Facial Representations and Viseme Information. In Proc. of ICASSP (2010) 2474–2477
20. Alarcao, S.M., Fonseca, M.J.: Emotions Recognition Using EEG Signals: A Survey. IEEE Transactions on Affective Computing (2017)
21. Maaoui, C., Pruski, A.: A comparative study of SVM kernel applied to emotion recognition from physiological signals. In Proc. of 5th International Multi-Conference on Systems, Signals and Devices (2008)
22. Tang, H., Chu, S., Johnson, M.H.: Emotion Recognition From Speech Via Boosted Gaussian Mixture Models. in Proc. of ICME (2009) 294–297
23. Xu, S., Liu, Y., Liu, X.: Speaker Recognition and Speech Emotion Recognition Based on GMM. 3rd International Conference on Electric and Electronics (EEIC 2013) (2013) 434–436
24. Schuller, B., Rigoll, G., Lang, M.: Hidden Markov Model-based Speech Emotion Recognition. in Proc. of the IEEE ICASSP **I** (2003) 401–404

25. Pan, Y., Shen, P., Shen, L.: Speech Emotion Recognition Using Support Vector Machine. International Journal on Smart Home **6 (2)** (2012) 101–108
26. Chavhan, Y., Dhore, M.L., Yesaware, P.: Speech Emotion Recognition Using Support Vector Machine. International Journal of Computer Applications (0975 - 8887) **1, No. 20** (2010) 6–9
27. Nicholson, J., Takahashi, K., Nakatsu, R.: Emotion Recognition in Speech Using Neural Networks. Neural Computing & Applications **9, Issue 4** (2000) 290–296
28. Shaw, A., Vardhan, R.K., Saxena, S.: Emotion Recognition and Classification in Speech using Artificial Neural Networks . International Journal of Computer Applications (0975 8887) **145, No.8** (2016) 5–9
29. Stuhlsatz, A., Meyer, C., Eyben, F., Zielke1, T., Meier, G., Schuller, B.: Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks. in Proc. of ICASSP (2011) 5688–5691
30. Han, K., Yu, D., Tashev, I.: Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. in Proc. of Interspeech (2014) 2023–2027
31. Sahidullah, M., Saha, G.: Design, Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition. Speech Communication **54 (4)** (2012) 543–565
32. Bielefeld, B.: Language Identification Using Shifted Delta Cepstrum. In Fourteenth Annual Speech Research Symposium (1994)
33. T.-Carrasquillo, P., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A., , Jr., J.D.: Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features. in Proc. of ICSLP2002-INTERSPEECH2002 (2002) 16–20
34. Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. in Proc. of Interspeech (2007) 2253–2256
35. Ranjan, S., Yu, C., Zhang, C., Kelly, F., Hansen, J.H.L.: Language Recognition Using Deep Neural Networks With Very Limited Training Data. in Proc. of ICASSP (2016) 5830–5834
36. Mohammad, Y., Matsumoto, K., Hoashi, K.: Deep feature learning and selection for activity recognition. In: in Proc. of the 33rd ACM/SIGAPP Symposium On Applied Computing. ACM SAC (2018) 926–935
37. LeCun, Y., Denker, J.S., Solla, S.A.: Optimal brain damage. In: Advances in Neural Information Processing Systems 2. (1990) 598–605
38. Yu, J., Lukefahr, A., Palframan, D., Dasika, G., Das, R., Mahlke, S.: Scalpel: Customizing dnn pruning to the underlying hardware parallelism. In: Proceedings of the 44th Annual International Symposium on Computer Architecture, ACM (2017) 548–560
39. Friedman, J., Hastie, T., et al., R.T.: Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics **28**(2) (2000) 337–407
40. Attabi, Y., Alam, J., Dumouchel, P., Kenny, P., Shaughnessy, D.O.: Multiple windowed spectral features for emotion recognition . in Proc. of ICASSP (2013) 7527–7531
41. Zhang, J., Inoue, N., Shinoda, K.: I-vector Transformation Using Conditional Generative Adversarial Networks for Short Utterance Speaker Verification . in Proc. of Interspeech (2018) 3613–3617

42. Tzinis, E., Potamianos, A.: Segment-Based Emotion Recognition Using Recurrent Neural Networks. in Proc. of ACII (2017) 190–195

43. Huang, C.W., Narayanan, S.: Attention Assisted Discover of Sub-Utterance in Speech Emotion Recognition. in Proc. of Interspeech (2016) 1387–1391