A Data Augmentation Approach for Improving the Performance of Speech Emotion Recognition

Georgia Paraskevopoulou[1], Evaggelos Spyrou[2,3] and Stavros Perantonis[2]
[1]*Department of History & Philosophy of Science, National and Kapodistrian University of Athens, Athens, Greece*
[2]*Institute of Informatics and Telecommunications, National Center for Scientific Research - "Demokritos," Athens, Greece*
[3]*Department of Computer Science and Telecommunications, University of Thessaly, Lamia, Greece*

Keywords: Emotion Recognition, Convolutional Neural Network, Spectrograms, Data Augmentation.

Abstract: The recognition of the emotions of humans is crucial for various applications related to human-computer interaction or for understanding the users' mood in several tasks. Typical machine learning approaches used towards this goal first extract a set of linguistic features from raw data, which are then used to train supervised learning models. Recently, Convolutional Neural Networks (CNNs), which unlike traditional approaches, learn to extract the appropriate features of their inputs, have also been applied as emotion recognition classifiers. In this work, we adopt a CNN architecture that uses spectrograms, extracted from audio signals as inputs and we propose data augmentation techniques to boost the classification performance. The proposed data augmentation approach includes noise addition, shifting of the audio signal, and changing its pitch or its speed. Experimental results indicate that the herein presented approach outperforms previous work which not use augmented data.

1 INTRODUCTION

The word "emotion" is composed of the prefix "e," which means "out" and the word "motion," which means "to move." Therefore, emotions are feelings that make us *move*. That is, we experience them, we express them, we recognize them, and use them to understand each other and make decisions, consciously or not. It is worth reporting that the beginning of research related to emotions seems to originate from 1872, i.e., when Darwin published his famous work "The Expression of the Emotions in Man and Animals" (Darwin, 2015). He argued that all humans, and even other animals, use similar behaviors to express their emotions. Nowadays, many psychologists agree with Darwin that certain emotions are universal to all humans, regardless of culture. These emotions are known as the "big six," i.e., anger, fear, surprise, disgust, happiness, and sadness (Ekman and Oster, 1979; Cowie and Cornelius, 2003).

According to Cowie et al., when humans interact, two communication channels are used (Cowie et al., 2001). The speaker transmits explicit messages, while the listener transmits implicit messages about the speaker. The first explicit channel has been a very active research area, while the second implicit one has been little studied in the last two decades. As a matter of fact, emotion recognition is a difficult task due to different ways of measuring and categorizing emotions or their dependency on various factors that make even humans misinterpret them. Nevertheless, emotion recognition is a very important part of our lives and as a result, it is extended to computers in order to enrich artificial intelligence (AI) applications with emotional intelligence. There are many applications in human-computer interaction such as lie detection, clinical diagnosis of schizophrenia, voice production for synthetic agents, robots, or machines that act as personal assistants. In addition, automotive environments use the information regarding the drivers' emotional state to apply safety strategies. Emotion recognition is also applicable in gaming when the "intelligent" video game recognizes the players' mood and makes the game more interesting and intractable if e.g., they exhibit positive emotions, or easier in case of negative emotions. All these applications share the following: they take as input the users' response or reaction, detect their emotional state, and subsequently make the appropriate decision.

Emotion recognition can be realized through the use of various modalities. Visual signals such as facial expressions, gestures or body movements, and

Paraskevopoulou, G., Spyrou, E. and Perantonis, S.
A Data Augmentation Approach for Improving the Performance of Speech Emotion Recognition.
DOI: 10.5220/0011148000003289
In *Proceedings of the 19th International Conference on Signal Processing and Multimedia Applications (SIGMAP 2022)*, pages 61-69
ISBN: 978-989-758-591-3; ISSN: 2184-9471

audio signals such as verbal and non verbal sounds are widely used separately or upon the application of fusion techniques; this way multimodal recognition is accomplished. Textual features that may be derived from speech could enhance the performance of a system that recognizes emotions (Atmaja et al., 2019). Moreover, biosignals like blood volume pulse (BVP) (Jang et al., 2015) or skin conductivity (Frantzidis et al., 2008) and brain signals from electroencephalography (EEG) (Song et al., 2018) or functional magnetic resonance imaging (fMRI) (Han et al., 2015) have been used in several research works. Note that it is difficult or even impossible to collect such signals out of a laboratory environment. Obviously, speech is the most available modality and in certain applications e.g., in a call center is the only one that may be collected. Furthermore, speech carries a significant amount of information which reflects emotional content that can result in achieving high recognition rates. For this reason, there is a field of research, namely "speech emotion recognition" (SER) that aims to design systems that recognize emotions using only audio signals. Ideally, an ideal SER system should be universal and robust against language or culture differences and variations (Schuller, 2018).

A definition of a SER system may be stated as "a collection of methodologies that are used in order to process and classify speech signals, detecting emotions embedded in them" (Akçay and Oğuz, 2020). To be able to efficiently and accurately recognize emotions, a SER system requires a supervised learning method, particularly a classifier that will be trained in order to recognize emotional states. For this purpose, the availability of a certain amount of labeled data is essential. Moreover, given a dataset of speech signals, pre-processing is necessary; various important features can be extracted in order to be given as input to the classifier. Prosodic, spectral, voice quality features and features based on the Teager energy operator are common types of features (Giannakopoulos, 2015). Since these features are typically extracted using algorithmic approaches, they are often referred to as "handcrafted" features. A wide range of classifiers is commonly used such as Support Vector Machines, Hidden Markov Models, Decision Trees, or Ensemble Methods. Deep Learning techniques have been recently dominating this research area, leading to impressive results, while they are able to learn from raw speech data, i.e., without the need of handcrafted features.

The rest of the paper is organized as follows: related work is presented in Section 2. Then, in Section 3 the proposed augmentation approach is presented along with the neural network architecture and the training strategy that has been adopted. Experiments are presented in Section 4. Implementation details are presented in Section 5. Finally in Section 6, we conclude the paper with a perspective analysis of possible future work.

2 RELATED WORK

Emotion recognition approaches typically extract some low-level features and a classification algorithm is then used to map them to emotion classes. The learning algorithm uses the labeled data and approximates the mapping function, which helps predict the class of new input. Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), Support Vector machines (SVMs), and Artificial Neural Networks (ANNs) are the most common classification approaches. Classifiers based on Decision Trees (DTs), k-Nearest Neighbor (k-NN), k-means, and Naive Bayes (NB) are also often used. Lastly, several SER systems use ensemble methods that combine several classifiers so as to obtain better results.

An example of such an approach is the work of (Shen et al., 2011). They achieved 82.5% accuracy using a SVM trained on a combination of prosodic and spectral features, such as energy, pitch, Linear Prediction Cepstrum Coefficients (LPCC) and Mel Frequency Cepstrum Coefficients (MFCC). Schuller et al. (Schuller et al., 2003) compared two methods. In their first method, they used global statistics features from the raw pitch and energy contour of the speech signal and classified them using GMMs. Each emotion was modeled by one GMM and the maximum likelihood model was considered as the recognized emotion at a time throughout the recognition process. In their second method, the temporal complexity increased by applying continuous HMM and using low-level instantaneous features rather than global statistics. The average recognition accuracy of seven discrete emotional states exceeded 86% using global statistics, whereas the overall recognition rate on average of five human deciders for the same corpus was 81.3%. In (Koolagudi and Rao, 2012), the authors used auto-associative neural networks (AANN) to capture the emotion-specific information from excitation source features, GMMs for developing the models using spectral features, and SVMs to discriminate the emotions using prosodic features, separately. Then, they used fusion techniques in order to combine the three kinds of features, increasing the performance up to 79.14% for the Emo-DB dataset.

Unlike most researchers who mainly rely on the classical frequency and energy-based features there

are various works that propose new features, such as (Yang and Lugger, 2010). Therein, some new harmonic features have been proposed. Moreover, (Xiao et al., 2010) also used harmonic features, combining them with Zipf-based features to enhance speech emotion recognition. They achieved 71.52% recognition rate for the Emo-DB dataset, and 81% for the DES dataset. Similarly, (Wu et al., 2011) attained 91.6% recognition rate using prosodic and modulation spectral features. In (Bitouk et al., 2010), the authors introduced a novel set of spectral features, statistics of Mel-Frequency Cepstral Coefficients computed over three phoneme type classes of interest – stressed vowels, unstressed vowels and consonants in the utterance outperforming traditional prosodic or spectral features. Moreover, their results indicated that spectral features from consonants may outperform the ones of vowels.

A typical problem that is often faced by SER systems is that typical datasets are recorded in a noise-free studio environment. However, inputs from practical applications include noise, whose effect on the accuracy of recognition may be severe. In (Tawari and Trivedi, 2010), the authors used a locally collected audio-visual database of effective expression in a car (LISA-AVDB) presenting a more realistic scenario. They also proposed an adaptive noise cancellation scheme which significantly improved results. Moreover, in (Chenchah and Lachiri, 2016) experiments with audio recordings from IEMOCAP database (Busso et al., 2008) have been presented upon adding real world noises (e.g., car, train and airport) over samples. Results upon denoising were compared to those before denoising and those without noise to measure the system performance, obtaining better results for the former.

During the last few years, the performance of the deep learning algorithms exceeded the one of the traditional machine learning algorithms, hence many research efforts have focused on deep learning and have also become the current trend in SER research. Deep learning is a class of machine learning algorithms that uses complex architectures of many interconnected layers, each consisting of nonlinear processing units. Each unit extracts and transforms features. Each layer's input is the output of the previous one. The advantage of some of these algorithms is that there is no need for handcrafted features and feature selection. All features are automatically selected and learned from raw data and typically lead to higher performance. Of course, this is achieved with the cost of higher computational time. The most widely used deep learning algorithms and architectures, which have been successfully applied to emotion recognition domain, are Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

In (Stuhlsatz et al., 2011), Generalized Discriminant Analysis (GerDA) based on Deep Neural Networks (DNNs) has been introduced, showing a highly significant improvement over traditional SVM approaches. Moreover, in (Huang et al., 2014) the authors used a semiCNN with two stages. They achieved superior results even in complex scenes (e.g., with speaker and environment distortion) outperforming traditional methods which use hand-crafted feature extraction. Similarly, in (Trigeorgis et al., 2016) an approach that learns from the raw time representation, combining CNNs with long short-term memory networks (LSTMs) has been presented. In the work of (Zhao et al., 2019) two convolutional neural networks and long short-term memory (CNN LSTM) networks, 1-D and 2-D CNNs with LSTM network for speech emotion recognition have been presented. On the Emo-DB database, the 2-D network obtained validation accuracy of 76.64% and 82.42% with speaker-dependent and speaker-independent experiments respectively. However, validation accuracy is 62,97% and 52.14% on IEMOCAP database for speaker dependent and speaker independent cases, respectively. Finally, in (Zheng et al., 2015) CNNs have been used on audio spectrograms and their superiority over classical methods has been demonstrated.

Finally, the authors in (Papakostas et al., 2017) presented an approach that uses a Convolutional Neural Network (CNN) as a classifier for emotion recognition. CNNs were responsible for identifying the important features of the input images. In this paper, spectrograms were extracted from raw data and no other extra features were required. Hand-crafted features were only extracted for validation purposes and no linguistic model was required. For their experiments four audio datasets were used, namely Emo-DB (Burkhardt et al., 2005), EMOVO (Costantini et al., 2014) , SAVEE (Jackson and Haq, 2014) which are publicly available, and a custom made one that includes audio samples gathered from movies. Considering the absence of noise in typical datasets and its presence within the real world, they added noise in three different levels before extracting the spectrogram of each training sample. Their SER is not specific to any particular language and indicated that CNNs are able to provide superior results vs. traditional techniques that use hand-crafted features.

We should herein note that the addition of noise in the work of (Papakostas et al., 2017) played the role of data augmentation. Data Augmentation is a widely used technique to tackle the problem of overfitting

that occurs due to insufficient number of training samples. In typical applications, data augmentation may enhance the extracted information from the original dataset and is defined as an artificial increase in the size of the original training samples, through data warping or oversampling. Data warping augmentations transform existing data such that their labels are preserved. In most computer vision approaches that utilize deep learning for classification, data warping encompasses augmentations like adding some noise, horizontally flipping, random crops, or color transformations. Oversampling is the technique that creates synthetic instances and adds them to the original training set.

In this paper, we present an approach that extends the previously mentioned work of (Papakostas et al., 2017). We also train a CNN with raw data, in an attempt to confirm that deep learning is indeed able to replace typical approaches which need feature extraction. However, our work is based on audio data warping augmentation. This method is inspired from computer vision data warping techniques; audio training samples are increased not only by adding noise but also by shifting, changing pitch, or changing the speed of the acoustic signal, prior to the extraction of the spectrogram. To this point, it should be mentioned that similar augmentation methods have already been applied to a different task improving its results (Slimi et al., 2022). Public audio databases are used to train and test our model.

3 PROPOSED METHODOLOGY

3.1 Emotional Speech Dataset

For the experimental evaluation of our work we used the EMOVO dataset (Costantini et al., 2014). This database is the first corpus for emotion recognition in the Italian language. Fourteen sentences (assertive, interrogative, lists) were performed from six actors, three men, and three women. Sentences are based on the big six emotional states plus the neutral state. The semantic value of the phrases is emotionally neutral in order to abet the actors not to be biased when trying to express the right emotional state. Furthermore, for spectral analysis, some basic linguistic conditions have been satisfied. It is worth noting that only recordings with anger, fear, happiness, sadness, and neutral emotional state were used in this work because we wanted our results to be comparable with the results of (Papakostas et al., 2017). To sum up, 420 samples (i.e., 14 phrases $\times$ 6 actors $\times$ 5 emotional states) were used to train our models

3.2 Data Pre-processing and Augmentation

Before training, audio samples are randomly cropped from the original audio signal in order to have the same duration, i.e., 2 sec. It is well known that deep learning approaches require large amounts of training data, in order to overcome the problem of overfitting, which in turn leads to poor performance in unseen data. As a result, we feel that the aforementioned 420 samples were not enough to achieve satisfactory classification performance.

Assuming that the satisfactory results when using EMOVO dataset with a deep learning approach (e.g., as in (Papakostas et al., 2017) are largely due to the augmentation of the dataset (e.g., with the addition of noise on audio samples) and inspired from computer vision data warping techniques, we propose a quite different approach so as to artificially augment the dataset. To this goal, shifting, changing pitch, or changing the speed of the acoustic signals are the proposed transformations that we adopt in this work. More specifically:

i. **Adding Noise.** Let $x(t)$ be the original audio signal, $n(t)$ be a "standard normal" distributed random value of noise, and N be a noise factor which depends on the Signal-to-Noise Ratio (SNR). Then, the noisy signal $\tilde{x}(t)$ is calculated as: $\tilde{x(t)} = x(t) + N \cdot n(t)$.

ii. **Shifting.** The original audio signal is shifted to left/right with a random time unit. If the audio signal is shifted to the left for M units, a zero value will be assigned to the first M units. Similarly, if the audio signal is shifted to the right for M units, a zero value will be assigned to the last M.

iii. **Changing Pitch.** This audio deformation randomly changes the pitch of audio signals.

iv. **Changing Speed.** This audio deformation stretches time series by a fixed rate. Thus, the new audio signals are augmented by changing the speed of the speaker's voice at the original audio signals.

We should herein note that augmentation factors or parameters were appropriately chosen in order to make subtle but perceptible changes to initial audio signals. An audio signal example and its corresponding augmented examples due to different kinds of augmentation are illustrated in Fig. 1. Moreover, in Fig. 1 the corresponding spectrograms are also illustrated for each case. The spectrograms we used in this work have been extracted with 40 msec. short-term window size and 20 msec. step. In our case, we extended the initial training set which consisted of

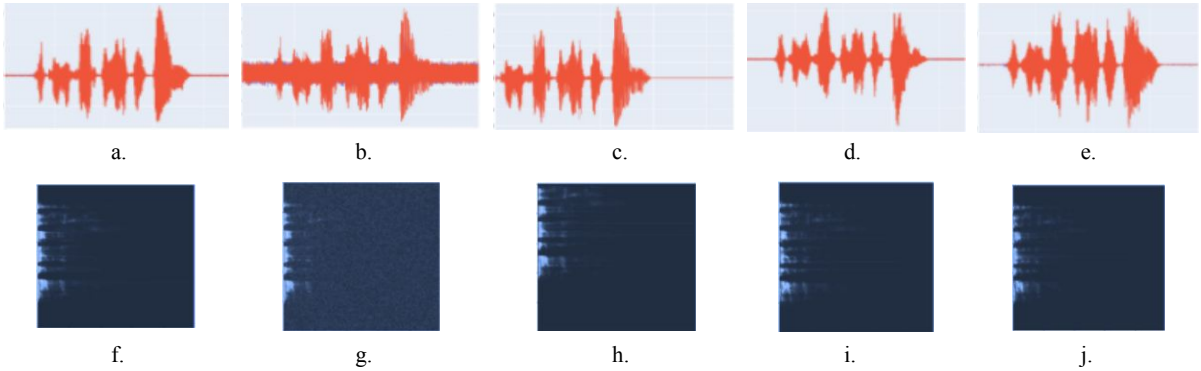

Figure 1: (a) a given (original) audio signal, (b) the original signal upon noise addition, (c) the original signal upon shifting to the left, (d) the original signal upon changing pitch, (e) the original signal upon changing speed, (f) – (j) spectrograms of (a) – (e), respectively. Note that spectrograms have been colorized for illustration purposes.

420 samples by adding 420 new samples with random SNR ratio ($N \in \{3,4,5\}$), 420 new samples with random shifting (right/left, maximum 1 sec.), 420 new samples with the random change of pitch (with a random factor $P \in [0.9, 1.1]$), and 420 new samples with the random change of speed (with a random factor $S \in [0.9, 1.1]$. Thus, the final augmented training set consists of $5 \times 420 = 2100$ audio segments of length 2 sec., each.

3.3 Neural Network Architecture and Training

We adopted the CNN architecture of (Papakostas et al., 2017), illustrated in Fig. 2. Our network consists of four stacked blocks interlaced of one convolutional layer with stride 2, a batch-normalization layer followed by the non-linearity function, and a max-pooling layer, followed by a local response normalization layer. The first convolutional layer comprises 96 feature maps and uses a kernel of size 7. The second and third convolutional layers comprise 384 filters with a kernel of size 5. The last convolutional layer again comprises 384 filters but uses a kernel of size 3. Batch-normalization transformation usage before the application of the activation function operates as a normalizer of the input batch. All pooling layers are of stride 2 with their kernel size equal to 3. The last pooling layer is stacked with two fully connected layers of 4096 units with dropout culminating in a softmax classifier.

For all the layers ReLu has been adopted as the activation function and Xavier initialization for the weights. Standard SGD algorithm has been selected for the learning process since it provides improved results compared to others, using categorical cross-entropy loss function. As a result, the output of the network is distributed on the five target classes. Step decay learning rate scheduler is also adopted, which drops the learning rate every 20 epochs by a factor equal to 0.1. The initial learning rate is equal to 0.001. In addition, l2 regularizer was added to each convolutional and each fully connected layer with parameter lambda equal to 0.008. These parameters were the result of extensive experimentation and hyperparameter tuning of our model, running 10-fold Cross-Validation for different combinations of various hyperparameters. The best performance arose from dropout equal to 0.4 and momentum equal to 0.9. The input to the network corresponds to RGB images of size 250×250 and is organized in batches of 64 samples.

For the finalized model we ran 10-fold cross-validation to test its ability to predict new data that was not used in the training process. In other words, to give an insight on how generalizable the model will be to an independent dataset. A significant decision is the stopping criterion of the training process at each round of cross-validation. That is the appropriate number of training epochs, for the purpose of preventing overfitting. A simple, effective, and widely used approach is to separate a portion of the training samples into a validation set, which is used for the evaluation of the model after every epoch of the training process. When validation performance stops improving or starts degrading, the training process should be stopped. After that point, the unseen test samples from one fold are used for the evaluation of each round of the 10-fold Cross-Validation. This method is called early stopping and was used in each round.

We split our training data before data augmentation, following well-known good practices, i.e., 80% for training, 10% for validation and the remaining 10% as the unseen test set. However, we applied the augmentation methodology on both training and validation sets, which is not a common training strategy,

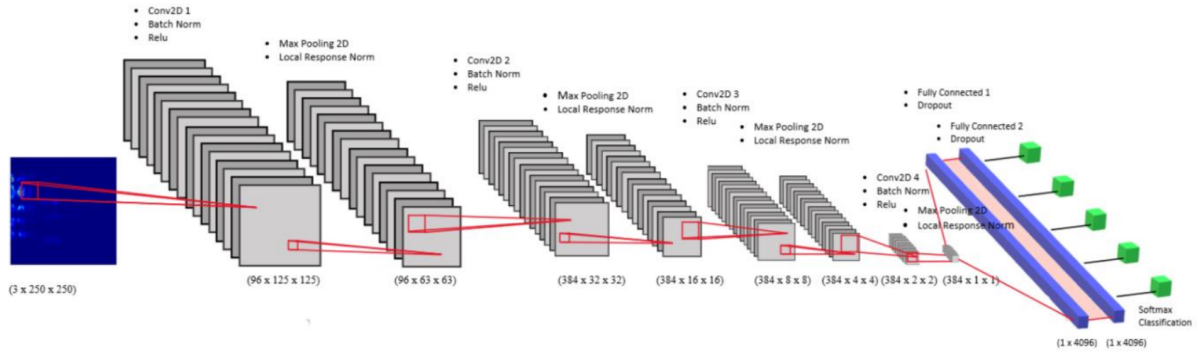

Figure 2: The Convolutional Neural Network architecture that has been used in this work.

yet it surprisingly resulted to a significant improvement of performance. Then, we proceeded with hyperparameter tuning and selection of the best model.

4 EXPERIMENTAL RESULTS

For the evaluation of our experiments, we opted for the use of the well known F_1 score, which is defined as the harmonic mean of precision (P) and recall (R), i.e., $F_1 = 2 \cdot P \cdot R/(P+R)$. Moreover, the average F_1 measure (also known as macro-averaged F_1) is defined as the arithmetic mean of the per-class F_1-scores. It is used when both true positives and true negatives are crucial. It is also considered as a very good metric for imbalanced datasets. For these reasons, the macro-averaged F_1 was chosen as metric for evaluation.

For comparison purposes two other methods have been implemented and evaluated:

- the methodology of (Papakostas et al., 2017); therein augmentation considers only 3 different SNRs (i.e., 3, 4 and 5). We have used the exact hyperparameters[1] for reproduction of the experiments.
- SVM classification based on hand-crafted features extraction from pyAudioAnalysis (Giannakopoulos, 2015). In particular, this approach uses as input fusion of 34 mid-term audio feature statistics, namely: Zero Crossing Rate, Energy, Entropy of Energy, Spectral Centroid, Spectral Spread, Spectral Entropy, Spectral Flux, Spectral Rolloff, MFCCs, Chroma Vector and Chroma Deviation.

The results in terms of the F_1 scores of the three compared approaches are presented in Table 1. Obviously, the proposed augmentation approach achieved about 10% higher performance compared to the one of (Papakostas et al., 2017) and about 15% higher performance compared to the SVM that has been trained with handcrafted features.

Table 1: Comparison of the proposed method to other methodologies, with respect to the macro-averaged f_1 scores.

	macro-averaged F_1 score
(Papakostas et al., 2017)	0.63
handcrafted features	0.60
proposed approach	**0.70**

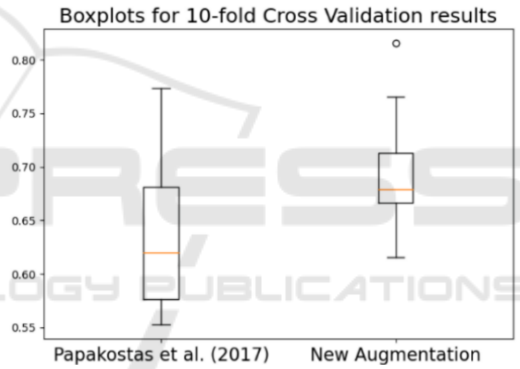

Figure 3: Box-and-whisker plot for macro-averaged F_1 scores during 10-fold cross-validation of the proposed augmentation approach and the one of (Papakostas et al., 2017).

Moreover, in Fig. 3 we illustrate box-and-whisker plots of the two deep learning-based approaches. As it may be observed for the methodology of (Papakostas et al., 2017), 50% of the rounds of 10-fold cross-validation of achieved F_1 scores between 0.58 and 0.69. Also total range is about 22% and positive skewness can be observed. In contrast, regarding the proposed approach, its box-and-whisker plot is almost symmetric and closer to the ideal. Furthermore, 50% of its cross-validation rounds achieved F_1 scores between 0.68 and 0.72 and the total range is about 8%. These observations indicate that the proposed model has small variation, and the distribution of performances has little dispersion in relation to the median.

[1] https://github.com/MikeMpapa/CNNs-Audio-Emotion-Recognition

However, it is significant to report an appreciated outlier round performance with F_1 score equal to 0.82. We believe that this was due to the randomness of splitting data during the cross-validation process. We also felt that this is a satisfactory result, indicating the possibility of the proposed augmentation approach to perform better upon more exhaustive tuning of its hyperparameters, or even a more sophisticated network architecture.

After the satisfactory performance of our proposed augmentation approach, we moved on to further evaluate its robustness with other datasets. For this purpose, we have used SAVEE and EMO-DB datasets, recorded in English and German, respectively. SAVEE (Jackson and Haq, 2014) which is a larger emotional speech dataset in English. 4 native English male speakers performed each 15 sentences per emotion, representing 6 emotional states, i.e., *disgust*, *fear*, *anger*, *joy*, *surprise*, *sadness* plus the neutral state. EMO-DB (Burkhardt et al., 2005) is an emotional speech dataset in German. 5 male and 5 female actors performed 493 utterances in total, simulating *anger*, *boredom*, *disgust*, *fear*, *happiness*, *sadness* and *neutral*. Note that at this part of the experiments we used the same CNN architecture, without any further tuning of its hyperparameters. Obviously. we used the same data augmentation methodology and the same approach for cross validation data split. The results of these experiments are depicted in Table 2. As it may be seen, the proposed approach outperformed the one of (Papakostas et al., 2017) also in the case of the EMO-DB dataset. However, in the case of the SAVEE dataset its performance was inferior. We believe that the hyperparameters which had been tuned for the EMOVO dataset caused the bad result when using the SAVEE dataset.

5 IMPLEMENTATION DETAILS

The entire project has been implemented using Python 3.6. For the data augmentation a custom python generator was implemented and the librosa[2] library was used to accomplish changing pitch and changing speed. Spectrograms were extracted using the pyAudioAnalysis library (Giannakopoulos, 2015). CNN models have been trained using the Keras deep-learning framework (Chollet et al., 2018). We also used pyAudioAnalysis to extract mid-term audio feature statistics and run SVM classification. Other python libraries that have been used were sklearn,[3] matplotlib,[4] pandas,[5] numpy[6] and scipy.[7]

6 CONCLUSIONS AND FUTURE WORK

During the last years, speech emotion recognition constitutes an area of increased interest for researchers. Many applications of voice user interfaces require emotion recognition solely by using audio information. Many machine learning approaches have already been used for this task, but most of them extract handcrafted features. Deep learning has also been used to classify emotions. In many research works (Papakostas et al., 2017) it has been demonstrated that the most important features of spectrograms that have been created by raw audio signals, may be extracted by using deep neural network architectures. This means that a linguistic model is not required.

In this work we proposed the use augmentation techniques of audio signals which further enhanced previous work and demonstrated superior performance. It is quite important to emphasize that we did not experimentally verify language independence, as it has been shown by several other studies (Schuller, 2018; Bhaykar et al., 2013). This result may be justified by the strong influence of sociocultural and linguistic features on expressing emotions in different languages. Furthermore, it is commonly accepted that humans recognize emotions not only from speech but also from gestures, posture, or even eye contact. This additional information is very important for the decision of which is the emotional state of our discussant. In addition, when a human is monolingual, she/he is trained to recognize the emotions of his native language speakers. It is difficult for her/him to understand the emotional state of other language speakers. Sometimes, there is also a difficulty in emotion recognition due to sarcasm or irony, even if the speakers speak the same language. As a result, recognition of irony or sarcasm seems to be a challenging task for future work. Another underlying reason for the difficulty of SERs in language and cross-language recognition may be the weakness of most databases for speech emotion recognition, which are simulated or elicited, to express all the essential emotional information into recorded samples.

[2]https://librosa.org/

[3]https://scikit-learn.org/stable/

[4]https://matplotlib.org/

[5]https://pandas.pydata.org/

[6]https://numpy.org/

[7]https://scipy.org/

Table 2: Experimental results of CNN with New Augmentation are in terms of the means of the achieved f1 measures during 10-fold Cross-Validation.

	EMOVO	SAVEE	EMO-DB	average
(Papakostas et al., 2017)	0.57	**0.60**	0.67	0.61
proposed	**0.70**	0.51	**0.72**	**0.64**
difference	+22.8%	-15.0%	+7.5%	+4.9%

Our future goals are focused on various directions. Firstly, we would like to increase the robustness of the proposed method in the given datasets, by further optimizing the learning process of the CNN. We also believe that speaker-dependent and speaker-independent experimental setups will lead to further improvement of results as recent research has shown. Two such examples are (Huang et al., 2014) and (Zhao et al., 2019). Within a speaker-dependent setup, samples from multiple speakers are used for training; testing takes place on different samples which belong to the same set of speakers. Moreover, within a speaker-independent setup, samples from multiple speakers are used for training and testing takes place on samples that belong to a different set of speakers. In conclusion, another future goal could be to experiment with models targeting at language or cultural information of the speech or with models that use transfer learning, which may provide another possible solution to language independence issues.

ACKNOWLEDGEMENTS

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: 1EDK-02070).

REFERENCES

Akçay, M. B. and Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76.

Atmaja, B. T., Shirai, K., and Akagi, M. (2019). Speech emotion recognition using speech feature and word embedding. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 519–523. IEEE.

Bhaykar, M., Yadav, J., and Rao, K. S. (2013). Speaker dependent, speaker independent and cross language emotion recognition from speech using gmm and hmm. In *2013 National conference on communications (NCC)*, pages 1–5. IEEE.

Bitouk, D., Verma, R., and Nenkova, A. (2010). Class-level spectral features for emotion recognition. *Speech communication*, 52(7-8):613–625.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al. (2005). A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Chenchah, F. and Lachiri, Z. (2016). Speech emotion recognition in noisy environment. In *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 788–792. IEEE.

Chollet, F. et al. (2018). Keras: The python deep learning library. *Astrophysics source code library*, pages ascl–1806.

Costantini, G., Iaderola, I., Paoloni, A., and Todisco, M. (2014). Emovo corpus: an italian emotional speech database. In *International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3501–3504. European Language Resources Association (ELRA).

Cowie, R. and Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech communication*, 40(1-2):5–32.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80.

Darwin, C. (2015). *The expression of the emotions in man and animals*. University of Chicago press.

Ekman, P. and Oster, H. (1979). Facial expressions of emotion. *Annual review of psychology*, 30(1):527–554.

Frantzidis, C. A., Lithari, C. D., Vivas, A. B., Papadelis, C. L., Pappas, C., and Bamidis, P. D. (2008). Towards emotion aware computing: A study of arousal modulation with multichannel event-related potentials, delta oscillatory activity and skin conductivity responses. In *2008 8th IEEE International Conference on BioInformatics and BioEngineering*, pages 1–6. IEEE.

Giannakopoulos, T. (2015). pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12):e0144610.

Han, J., Ji, X., Hu, X., Guo, L., and Liu, T. (2015). Arousal recognition using audio-visual features and

fmri-based brain response. *IEEE Transactions on Affective Computing*, 6(4):337–347.

Huang, Z., Dong, M., Mao, Q., and Zhan, Y. (2014). Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 801–804.

Jackson, P. and Haq, S. (2014). Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.

Jang, E.-H., Park, B.-J., Park, M.-S., Kim, S.-H., and Sohn, J.-H. (2015). Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. *Journal of physiological anthropology*, 34(1):1–12.

Koolagudi, S. G. and Rao, K. S. (2012). Emotion recognition from speech using source, system, and prosodic features. *International Journal of Speech Technology*, 15(2):265–289.

Papakostas, M., Spyrou, E., Giannakopoulos, T., Siantikos, G., Sgouropoulos, D., Mylonas, P., and Makedon, F. (2017). Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. *Computation*, 5(2):26.

Schuller, B., Rigoll, G., and Lang, M. (2003). Hidden markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 2, pages II–1. Ieee.

Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99.

Shen, P., Changjun, Z., and Chen, X. (2011). Automatic speech emotion recognition using support vector machine. In *Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology*, volume 2, pages 621–625. IEEE.

Slimi, A., Nicolas, H., and Zrigui, M. (2022). Detection of emotion categories' change in speeches. In *ICAART*.

Song, T., Zheng, W., Song, P., and Cui, Z. (2018). Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541.

Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G., and Schuller, B. (2011). Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5688–5691. IEEE.

Tawari, A. and Trivedi, M. M. (2010). Speech emotion analysis in noisy real-world environment. In *2010 20th International Conference on Pattern Recognition*, pages 4605–4608. IEEE.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). 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)*, pages 5200–5204. IEEE.

Wu, S., Falk, T. H., and Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech communication*, 53(5):768–785.

Xiao, Z., Dellandrea, E., Dou, W., and Chen, L. (2010). Multi-stage classification of emotional speech motivated by a dimensional emotion model. *Multimedia Tools and Applications*, 46(1):119–145.

Yang, B. and Lugger, M. (2010). Emotion recognition from speech signals using new harmony features. *Signal processing*, 90(5):1415–1423.

Zhao, J., Mao, X., and Chen, L. (2019). Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical signal processing and control*, 47:312–323.

Zheng, W., Yu, J., and Zou, Y. (2015). An experimental study of speech emotion recognition based on deep convolutional neural networks. In *2015 international conference on affective computing and intelligent interaction (ACII)*, pages 827–831. IEEE.