A Convolutional and Recurrent Neural Network-based Approach for Speech Emotion Recognition

Piotr Duch[0000–0003–0656–1215], **Izabela Wiatrowska,**
Paweł Kapusta[0000–0002–3527–7208]

[1]*Lodz University of Technology*
Institute of Applied Computer Science
Stefanowskiego 18, 90-537 Łódź, Poland
piotr.duch@p.lodz.pl
pawel.kapusta@p.lodz.pl

DOI:10.34658/9788366741928.42

Abstract. *Speech emotion recognition (SER) is a crucial aspect of human-computer interaction. In this article, we propose a deep learning approach, using CNN and RNN architectures, for SER using both convolutional and recurrent neural networks. We evaluated the approach on four audio datasets, including CREMA-D, RAVDESS, TESS, and EMOVO. Our experiments tested various feature sets and extraction settings to determine optimal features for SER. Our results demonstrate that the proposed approach achieves high accuracy rates and outperforms state-of-the-art algorithms.*

Keywords: *artificial intelligence, speech emotion recognition*

1. Introduction

Speech emotion recognition (SER) is a critical aspect of human-computer interaction, particularly as more interactions are based on spoken communication. Emotions are conveyed not only through posture, facial expressions and gestures but also through the tone, pitch, and other acoustic features of spoken language. However, recognizing emotions from speech patterns can be challenging due to the subjective nature of emotions, the difficulty of distinguishing between multiple emotions expressed in a single conversation, and the time-consuming process of collecting and classifying data. In this article, we investigate the application of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) in SER, which have potential applications in various fields, such as healthcare, psychology, criminal investigations, and customer service. The use of trained classifiers can help computers better understand human needs and respond appropriately, particularly when the conversation context is essential. Overall, SER has the potential to improve the quality and effectiveness of human-computer interactions and contribute to our understanding of emotional expression and communication.

2. Datasets and methodology

In our research on speech emotion recognition, we have selected four databases: CREMA-D [1], RAVDESS [2], TESS [3], and EMOVO [4] to evaluate the proposed algorithm. The combination of these databases enables a comprehensive evaluation of speech emotion recognition with the ability to consider various emotions, cultures, genders, and languages.

In this study, we selected three features for sound transformation from the time domain to the frequency domain to extract features from audio files. The Mel spectrogram was chosen as the first feature due to its frequent use in deep learning and ability to transform frequencies comparable to how humans perceive sound differences expressed in Hertz. The second feature are the Mel frequency cepstral coefficients (MFCCs), which consist of 13 coefficients that capture the shape of the human vocal system and tone color. The last feature is the Chromagram, a pitch-based profile of 12 pitch classes that captures harmonic and melodic sound features, resistant to changes in tone color. For RNN, the extracted features were stacked, forming a single matrix that becomes the input to the network, while for CNN, each feature was sent separately to the network to enable the convolutional layers to learn specific weights for each feature. These three features enable a more comprehensive analysis of the emotional state of the speaker, which is particularly important in the context of the proposed deep learning models (see Fig. 1).

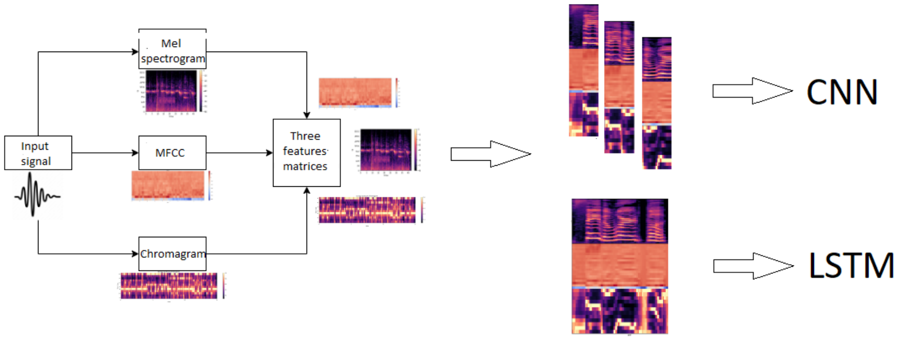

Figure 1. Architecture of the proposed algorithm. Source: own work.

During the training process, the input size had to be standardized. This was particularly important for the CNN where each input signal had to have the same dimension. In contrast, the length of the signals could differ in the RNN. To ensure compatibility with the CNN, each sample was divided into fragments that overlapped by 25%. Furthermore, the datasets were augmented using three approaches: adding noise with an amplitude of 0.035 to the sample, slowing down the speed of speech, and lowering the pitch. The CNN architecture consisted of two convolutional layers with 128 and 64 filters for the Mel spectrogram and 64 and 32 filters for the other features, respectively. The results for all three features were then combined and flattened. Two fully connected layers with 64 and 6 neurons were used to complete the network. For RAVDESS and CREMA-D, the optimal neural network consisted of one additional convolutional layer for each feature with 32 filters. The RNN architecture consisted of two Long Short-Term Memory (LSTM) layers with 128 and 64 cells for RAVDESS, TESS, and EMOVO and three LSTM layers with 128, 64, and 64 cells for CREMA-D. Similar to CNN, the RNN also used two fully connected layers with 64 and 6 neurons to complete the network.

3. Results

In our study, we evaluated the performance of our proposed deep learning models for speech emotion recognition using four datasets: TESS, RAVDESS, CREMA-D, and EMOVO. During our experiments, we tested several different feature sets and feature extraction settings to determine the optimal features for speech emotion recognition. Our final results were very promising, with the model achieving satisfactory results on all tested datasets (Table 1).

Furthermore, we have compared the accuracy of our approach with several state-of-the-art methods using three different datasets: RAVDESS, EMOVO, and CREMA-D (Table 2). Our results indicate that our proposed approach has a higher accuracy rate than other algorithms in RAVDESS and EMOVO datasets and comparable accuracy in CREMA-D.

Table 1. The classification performance on chosen datasets using CNN and RNN.

Dataset	CNN	LSTM
TESS	100%	99%
RAVDESS	84%	77%
CREMA-D	64%	62%
EMOVO	87%	89%

Table 2. Accuracy comparison with existing SER algorithms

Method	**Accuracy**	**Year**
RAVDESS dataset		
DCNN [5]	71.6%	2020
Multimodal fine-grained learning [6]	74.7%	2020
Head Fusion [7]	77.4%	2020
BiLSTM [8]	82%	2020
Our approach – LSTM	77%	2023
Our approach – CNN	84%	2023
EMOVO dataset		
Multi-Level Local Binary and Ternary [9]	73.87%	2020
Mel frequency magnitude coefficient [10]	73.81%	2021
Twine shuffle pattern [11]	79.08%	2021
Statistical Feature Extraction for Deep SER [12]	83.9%	2022
Our approach – LSTM	89%	2023
Our approach – CNN	87%	2023
CREMA-D dataset		
SE-ResNet + GhostVLAD layer + emotion constrain [13]	64.92%	2021
ANN+ReLU (MFCC) [14]	71.96%	2021
2D CNN [15]	70.1%	2022
BYOL-S, 2048 [16]	76.9%	2022
Our approach – LSTM	66.2%	2023
Our approach – CNN	64%	2023

4. Conclusions

The research presented in this article provides a comprehensive evaluation of the proposed deep-learning algorithms for speech-emotion recognition. The high accuracy of our method suggests that it is a promising technique for accurate speech emotion recognition, which can be applied in various fields, such as healthcare, psychology, and customer service. Furthermore, our approach can facilitate the development of more sophisticated human-computer interaction systems that can better understand the emotional state of the speaker and respond appropriately.

References

[1] Cao H., Cooper D.G., Keutmann M.K., Gur R.C., Nenkova A., Verma R., *Crema-d: Crowd-sourced emotional multimodal actors dataset*, *IEEE transactions on affective computing*, 2014, vol. 5, no 4, pp. 377–390.

[2] Livingstone S.R., Russo F.A., *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*, 2018, vol. 13, no 5, p. e0196391.

[3] Dupuis K., Pichora-Fuller M.K., *Toronto emotional speech set (tess) collection*, 2010, (access: 04-05-2022).
https://tspace.library.utoronto.ca/handle/1807/24487

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

[5] Issa D., Fatih Demirci M., Yazici A., *Speech emotion recognition with deep convolutional neural networks*, *Biomedical Signal Processing and Control*, 2020, vol. 59, no 101894, doi: https://doi.org/10.1016/j.bspc.2020.101894.

[6] Li H., Ding W., Wu Z., Liu Z., *Learning fine-grained multimodal alignment for speech emotion recognition*, *arXiv preprint arXiv:2010.12733*, 2020.

[7] Xu M., Zhang F., Zhang W., *Head fusion: Improving the accuracy and robustness of speech emotion recognition on the iemocap and ravdess dataset*, *IEEE Access*, 2021, vol. 9, pp. 74539–74549.

[8] Sajjad M., Kwon S., et al., *Clustering-based speech emotion recognition by incorporating learned features and deep bilstm*, *IEEE access*, 2020, vol. 8, pp. 79861–79875.

[9] Sönmez Y.Ü., Varol A., *A speech emotion recognition model based on multi-level local binary and local ternary patterns*, *IEEE Access*, 2020, vol. 8, pp. 190784–190796.

[10] Ancilin J., Milton A., *Improved speech emotion recognition with mel frequency magnitude coefficient*, *Applied Acoustics*, 2021, vol. 179, p. 108046.

[11] Tuncer T., Dogan S., Acharya U.R., *Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques*, *Knowledge-Based Systems*, 2021, vol. 211, p. 106547.

[12] Sekkate S., Khalil M., Adib A., *A statistical feature extraction for deep speech emotion recognition in a bilingual scenario*, *Multimedia Tools and Applications*, 2023, vol. 82, p. 11443–11460.

[13] Mocanu B., Tapu R., Zaharia T., *Utterance level feature aggregation with deep metric learning for speech emotion recognition*, *Sensors*, 2021, vol. 21, no 12, p. 4233.

[14] Dolka H., VM A.X., Juliet S., *Speech emotion recognition using ann on mfcc features*, [In:] *2021 3rd international conference on signal processing and communication (ICPSC)*, IEEE, pp. 431–435.

[15] Mittal R., Vart S., Shokeen P., Kumar M., *Speech emotion recognition*, [In:] *2022 2nd International Conference on Intelligent Technologies (CONIT)*, IEEE, pp. 1–6.

[16] Scheidwasser-Clow N., Kegler M., Beckmann P., Cernak M., *Serab: A multilingual benchmark for speech emotion recognition*, [In:] *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 7697–7701.