Impact of different voting strategies in CNN based speech emotion recognition

Nikola Simić
University of Novi Sad
Faculty of Technical Sciences
Novi Sad, Serbia
nikolasimic@uns.ac.rs

Siniša Suzić
University of Novi Sad
Faculty of Technical Sciences
Novi Sad, Serbia
sinisa.suzic@uns.ac.rs

Tijana Nosek
University of Novi Sad
Faculty of Technical Sciences
Novi Sad, Serbia
tijana.nosek@uns.ac.rs

Mia Vujović
University of Novi Sad
Faculty of Technical Sciences
Novi Sad, Serbia
miavujovic@uns.ac.rs

Milan Sečujski
University of Novi Sad
Faculty of Technical Sciences
Novi Sad, Serbia
secujski@uns.ac.rs

***Abstract*— Automatic emotion recognition systems detect the emotional state of a person by analyzing expressions that can be recorded using different modalities. Among them, speech represents one of the major modalities that can be used independently or fused with other modalities. Due to the fact that emotional state may vary over a short time and that there are similarities among features obtained for different emotions, voting strategy has a key role when observing features collected over time. The focus of this paper is to examine the impact of different voting strategies in state-of-the-art speech emotion recognition systems based on convolutional neural networks on the example of the recently recorded Serbian Emotional Amateur Corpus.**

Keywords—Speech, Emotions, Recognition, Voting, CNN.

I. Introduction

The same sentence, uttered in two different emotions, can have a completely different meaning. Emotional speech, together with facial expressions and the subject's behavior, affects the way information is transmitted and interpreted. As the fusion of these different modalities contributes to the communication between humans, communication between humans and machines can be similarly improved if machines are given the capability to recognize emotions in humans as well as to generate emotional expressions (intended for human perception).

To this date, numerous systems have been developed for the automatic recognition of emotions from speech signals. Most traditional systems are created using hand-crafted features [1,2]. While some features such as Mel-frequency cepstral coefficients (MFCC), pitch, energy and duration-based characteristics are proven to give value to these systems [1,3,4], there are still drawbacks since there are noticeable similarities between the features corresponding to different emotions. Furthermore, hand-crafted features extraction may fail to incorporate high-level features, that can lead to the loss of potentially useful information [5].

With the advancement of computing power and the increasing usage of deep neural networks in many industries, hand-crafted features have been replaced with features automatically extracted using neural networks over short periods of time. These automatically extracted features model high-level abstractions of the data, which are derived from low-level features in a hierarchical learning process [5,6]. Recurrent neural networks (RNN) such as long short-term memory (LSTM) and bi-directional LSTM, as well as convolutional neural networks (CNN) for spectrogram-based speech emotion recognition (SER), represent state-of-the-art techniques that give promising results [5,7,8].

A wide range of approaches for spectrogram-based CNN can be found in the literature. Hajarolasvadi et al. [9] proposed a method that combines the information from an 88-dimensional vector of hand-crafted features extracted at a frame level with the spectrograms for every corresponding frame. Based on hand-crafted features and using the *k*-means clustering approach, *k* most discriminant frames were selected as a summary of the speech signal. Spectrograms of these key frames were encapsulated in a 3D tensor in order to train the 3D CNN model. The results showed that the model they propose performs better than the pre-trained VGG16 2D network, where majority voting was used to create a final prediction for each audio signal. Zhao et al. [5] combined the CNN and LSTM networks in order to reap the benefits of both networks. Local correlations along with hierarchical correlations are learned by CNNs, while long-term dependencies from the learned local features are extracted using LSTM. The results showed that the networks they proposed, both 1D CNN LSTM as well as 2D CNN LSTM, achieve excellent performance in an emotion recognition task and that the 2D CNN LSTM network outperforms traditional approaches. Lim et al. [10] also proposed a method based on concatenating a deep hierarchical CNN's feature extraction architecture with LSTM layers and compared it with separated CNN and RNN. The majority voting method was used for the final decision. Badshah et al. [11] proposed a spectrogram-based CNN model and investigated the effectiveness of transfer learning for emotion recognition using a pre-trained AlexNet model. However, each audio file consisted of multiple spectrograms, and an aggregation mechanism was employed to combine the individual predictions into an overall prediction result. Besides conventional RNN and CNN layers, the focus of research has also been directed to attention layers that are shown to increase the emotion capturing capability of the models [7,12,13].

Although the improvements that deep neural networks have brought to the emotion recognition field are indubitable, emotional speech analysis is still a demanding task. The fact that emotions can vary over a short period of time is one of numerous reasons. Feature and spectrogram extraction is thus often performed over short intervals, and the final decision can be made on the basis of appropriate voting strategies. Since there could be overlap among characteristics obtained for different emotions, the choice of a voting strategy can play an important role in the system's reliability. Interestingly, while different variations of DNN models are presented in the emotion recognition studies, not much attention has been given to the comparison between voting strategies. Still, majority voting, average probability voting and maximum probability voting stood out as the commonly used strategies for various problems, from acoustic scene classification [14] and singing voice detection [15] to EEG-sleep stage scoring [16] and EMG-gesture recognition [17]. In this paper we investigate the influence of voting techniques in a speech emotion recognition task on the example of Serbian Emotional Amateur Corpus (SEAC). The database was recorded by amateur speakers using mobile phones. We analyze system performance in the case of two state-of-the art architectures. Voting strategies are applied to the frame-based inference during post-processing phase.

The rest of the paper is organized as follows. CNN-based models are described in Section II. Next, experimental setup and performances of the analyzed neural network models are presented in Section III, whereas the influence of three voting strategies, applied in the post-processing, is analyzed in Section IV. In the end, conclusions are provided in Section V.

II. SPEECH EMOTION RECOGNITION BASED ON CNN MODELS

A speech emotion recognition task can be succesfully accomplished using end-to-end convolutional neural networks [18]. For such approach, speech spectrograms are commonly used as network input. Here we obtain spectrograms by applying short-Term Fourier transform (STFT) on pre-processed audio files. Preprocessing was done as in [19], providing one-second-long segments without silence regions and with no overlap within segments.

For the analysis we use two architectures. The first one is based on VGGish architecture, presented in [12]. This architecture was originally proposed for audio classification but is successfully applied to other tasks including speech emotion recognition [13]. All the elements of this network are given in Table I. Such SER model can be considered as a very deep as there are about 205 millions of parameters in our case. The model consists of 6 convolutional layers, 4 max pooling layers, 2 fully connected layers and the output layer. The number of nodes in the output layer is set to 5, according to the number of different emotional styles in the SEAC dataset.

TABLE I. VGGISH-BASED ARCHITECTURE

Layer	Parameters
Convolution2D	filters=64, kernel=(3,3), strides=(1,1)
MaxPooling2D	pool_size=(2,2), strides=(2,2)
Convolution2D	filters=128, kernel=(3,3), strides=(1,1)
MaxPooling2D	pool_size=(2,2), strides=(2,2)
Convolution2D	filters=256, kernel=(3,3), strides=(1,1)
Convolution2D	filters=256, kernel=(3,3), strides=(1,1)
MaxPooling2D	pool_size=(2,2), strides=(2,2)
Convolution2D	filters=512, kernel=(3,3), strides=(1,1)
Convolution2D	filters=512, kernel=(3,3), strides=(1,1)
MaxPooling2D	pool_size=(2,2), strides=(2,2)
Flatten	
Dense	nodes=4096
Dense	nodes=4096
Dense	nodes=5

The second used architecture is deep stride CNN architecture (DSCNN) presented in [6]. This network is much less complex, with about 433,000 parameters in our case, so that it can be considered as suitable for constrained-precision low complexity implementations. The parameters of such a network are given in Table II.

Both networks are analyzed on the example of Serbian Emotional Amateur Corpus. This is a recently recorded corpus of emotional speech, publicly available at [20]. The corpus is recorded by amateur speakers using their cellphones. There are five emotional styles in total – neutral, anger, joy, fear and sadness. All participants had an opportunity to record 60 utterances in up to five styles according to the predefined plan. The corpus contains recordings of 55 different speakers, who recorded at least one emotional style, whereas 23 of them (11 males and 12 females) recorded utterances in all five emotional styles. Duration of recorded utterances is larger of 2 seconds, whereas most of utterances are shorter than 5 seconds. This means that a single utterance is represented using multiple spectrograms, as a single spectrogram is obtained on the one-second-long speech, as described before. Inference based on a single spectrogram can be considered valuable for real-time systems. However, making decisions based on the entire utterance (i.e. a group of spectrograms which can be considered as a frame) may improve inference, whereas delay would not be very long. Consequently, here we analyze the application of three voting techniques and explore their suitability for SER task on the example of the SEAC database. We explore majority voting, average probability voting and maximum probability voting. In all three cases, the same classifier is used for predicting the class of an individual member of a frame, whereas the final decision is made for the whole frame.

III. EXPERIMENTS

After processing recorded utterances in SEAC database, we produced 14790 spectrograms of resolution 128×70. Spectrograms are obtained by processing the equal number of utterances, produced by 23 speakers who recorded utterances in all emotional styles. For training purposes, we

TABLE II. DSCNN ARCHITECTURE

Layer	Parameters
Convolution2D	filters=16, kernel=(7,7), strides=(2,2)
Convolution2D	filters=32, kernel=(5,5), strides=(2,2)
Convolution2D	filters=32, kernel=(3,3), strides=(2,2)
Convolution2D	filters=64, kernel=(3,3), strides=(2,2)
Convolution2D	filters=64, kernel=(3,3), strides=(2,2)
Convolution2D	filters=128, kernel=(3,3), strides=(2,2)
Convolution2D	filters=128, kernel=(3,3), strides=(2,2)
Flatten	
Dense	nodes=512
Dense	nodes=5

selected 80% of utterances, whereas validation and test set consist of 10% of utterances each. We keep the train-validation-test proportion in number of utterances per every speaker. Such emotion recognition system is a speaker-dependent and those systems usually provide higher accuracy comparing to the speaker-independent models. However, the aim of the paper is to analyze the impact of several voting strategies and it is not expected the influence of emotion recognition type model. Due to different speaking dynamics of various speakers, the number of spectrograms can vary for the same utterances in various styles. In the end, we had 11867 spectrograms for training, 1528 for validation and 1395 for testing.

To evaluate the performance of models described in Table I and Table II, we observe classification accuracy, precision, recall and f1 score. VGGish-based model is trained for 30 epochs, whereas DSCNN model is trained for 120 epochs. The stopping criterion was defined by the minimum validation loss. Confusion matrix in the case of the VGGish-based architecture is presented in Table III. The overall classification accuracy of the VGGish-based architecture, obtained on the test set, is 76.99%. The performances are summed up in Table IV.

Classification accuracy of the DSCNN architecture obtained on the test set is 70.90%. Such result could be expected, as the model is much smaller comparing to the VGGish-based architecture. The corresponding confusion matrix is presented in Table V, whereas model performances are provided in details in Table VI. Similarly, as in the case of VGGish-based architecture, it can be noticed that classification accuracy of the DSCNN model is lowest in the case of anger. However, DSCNN model achieves the highest accuracy in the case of joy, whereas VGGish-based architecture provides the highest accuracy in the case of neutral speech and sadness. Such differences will be carefully considered in the future while making a novel SER system.

IV. Speech emotion recognition based on voting techniques

In this section we analyze the influence of three voting techniques on the performance of a SER system. Voting techniques are applied in the post-processing phase of the frame-based approach, so that as a classification task we consider emotion classification on the utterance level. This means that spectrograms, obtained from the same utterance, are processed separately and after that the inference is performed on the sentence level.

TABLE III. Confusion matrix in the case of VGGish-based architecture

		Predicted class				
		Anger	Neutral	Joy	Fear	Sadness
True class	Anger	190	20	27	35	5
	Neutral	5	222	10	9	14
	Joy	21	28	236	44	5
	Fear	8	14	11	203	22
	Sadness	0	22	4	17	223

TABLE IV. Performance of the VGGish-based architecture

	Precision	**Recall**	**F1 score**	**Classification accuracy [%]**
Anger	0.85	0.69	0.76	68.59
Neutral	0.73	0.85	0.78	85.38
Joy	0.82	0.71	0.76	70.66
Fear	0.66	0.79	0.72	78.68
Sadness	0.83	0.84	0.83	83.83
Weighted avg.	0.78	0.77	0.77	76.99

TABLE V. Confusion matrix in the case of DSCNN

		Predicted class				
		Anger	Neutral	Joy	Fear	Sadness
True class	Anger	171	12	59	33	2
	Neutral	13	179	36	16	16
	Joy	21	18	273	20	2
	Fear	17	13	32	185	11
	Sadness	10	29	7	39	181

TABLE VI. Performance of the DSCNN model

	Precision	**Recall**	**F1 score**	**Classification accuracy [%]**
Anger	0.74	0.62	0.67	61.73
Neutral	0.71	0.69	0.70	68.84
Joy	0.67	0.82	0.74	81.74
Fear	0.63	0.72	0.67	71.71
Sadness	0.85	0.68	0.76	68.05
Weighted avg.	0.72	0.71	0.71	70.90

The post-processing procedure for each voting technique is as follows:

- **Majority voting**: For each spectrogram within a sentence we predict the class and after that the most common vote within a frame is selected as a frame prediction. In the case that there is no clear majority of any class, we take the highest prediction value from the most common votes.
- **Average probability voting**: For each spectrogram within a sentence we predict probabilities of classes. After that, probabilities are averaged for all spectrograms within a frame and the highest average probability is chosen to predict the class of the whole sentence.
- **Maximum probability voting**: For each spectrogram within a sentence we predict probabilities of classes. After that we take the class with the highest probability of any spectrogram within the frame to predict the class of the whole sentence. In this way, we choose the class of the most reliably classified spectrogram.

The results of applying these voting techniques to the VGGish-based architecture and DSCNN model are shown in Table VII.

By observing results in Table VII, it can be clearly seen that all three voting techniques improve the overall inference of SER models. In the case of VGGish-based

TABLE VII. The influence of voting techniques on the VGGish-based architecture and DSCNN

	Classification accuracy [%]	
Voting technique	VGGish-based architecture	DSCNN
without	76.99	70.90
majority	81.65	77.06
avg. prob.	82.57	77.52
max. prob.	82.88	76.76

TABLE VIII. THE INFLUENCE OF AVERAGE PROBABILITY VOTING ON PRECISION, RECALL AND F1 SCORE

	VGGish-based architecture	**DSCNN**
Precision	0.84	0.79
Recall	0.83	0.78
F1 score	0.83	0.78

architecture, achieved gain in terms of classification accuracy is between 4.66% and 5.89%, whereas the gain in the case of DSCNN is even higher and ranges from 5.86% to 6.62%. Furthermore, we can observe that average probability voting on the frame level provides the highest gain in the case of DSCNN, whereas the gain achieved in the case of VGGish-based architecture is very near to the highest one, indicating its robustness. However, robustness should be examined for a larger number of models and datasets. Here, we provide the results of average weighted precision, recall and f1 score in the case of this voting strategy in Table VIII. By comparing results from Tables VIII, IV and VI, the one can conclude that there is a significant gain in terms of precision, recall and f1 score.

V. CONCLUSIONS

In this paper we have investigated the application of different voting strategies in some of the state-of-art speech emotion recognition models, based on convolutional neural networks. For the experiments, we have chosen a very deep VGGish-based architecture and much smaller DSCNN model. The models were tested in the case of recently recorded Serbian Emotional Amateur Corpus. In order to improve classification accuracy, we have examined the influence of three voting techniques, that are applied in a frame-based approach. In this way, predictions are made for the whole recorded utterances. The results show that voting techniques significantly improve overall classification accuracy, precision, recall and f1 score and that their application in the post-processing phase is of a great importance. Although all observed voting techniques provide significant gain, we highlight that average probability voting showed the most robust performance in the performed experiments, making it a serious candidate for further analysis with various networks and development of novel systems.

ACKNOWLEDGMENT

This research was supported by the Science Fund of the Republic of Serbia (grant #6524560, AI – S ADAPT).

REFERENCES

[1] S. G. Koolagudi, K. S. Rao, “Emotion recognition from speech: a review”, *International Journal of Speech Technology*, vol. 15, pp. 99–117, 2012.

[2] C.-N., Anagnostopoulos, T. Iliou, I. Giannoukos, “Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011”, *Artificial Intelligence Review*, vol. 43, pp. 155-177, 2015.

[3] O. W. Kwon, K. Chan, J. Hao, T.-W. Lee, “Emotion recognition by speech signals”, *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, Geneva, Switzerland, pp. 125 – 128, September 2003.

[4] V. Delić, Z. Perić, M.. Sečujski, N. Jakovljević, J. Nikolić, D. Mišković, N. Simić, S. Suzić, T. Delić, “Speech technology progress based on new machine learning paradigm”, *Computational Intelligence and Neuroscience*, 4368036, pp.1–19, 2019.

[5] J. Zhao, X. Mao, L. Chen, “Speech emotion recognition using deep 1D & 2D CNN LSTM networks*”, Biomedical Signal Processing and Control*, vol. 47, pp. 312-323, 2019.

[6] Mustaqeem, S. Kwon, “A CNN-assisted enhanced audio signal processing for speech emotion recognition”, *Sensors*, vol. 20 (1), 183, 2020.

[7] D. Li, J. Liu, Z. Yang, L. Sun, Z. Wang, “Speech emotion recognition using recurrent neural networks with directional self-attention”, *Expert Systems with Applications*, vol. 173, 114683, pp. 1-13, 2021.

[8] Z. Huang, M. Dong, Q, Mao, Y. Zhan, “Speech smotion recognition using CNN”, *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 801 – 804, November 2014.

[9] N. Hajarolasvadi, H. Demirel, “3D CNN-based speech emotion recognition using k-means clustering and spectrograms”, *Entropy*, vol. 21 (5), 479, pp. 1 – 17, 2019.

[10] W. Lim, D. Jang, T. Lee, “Speech emotion recognition using convolutional and recurrent neural networks”, *Proceedings of the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, South Korea, pp. 1-4, December 2016.

[11] M. Badshah, J. Ahmad, N. Rahim and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network”, *Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon)*, Busan, South Korea, pp. 1-5, 2017.

[12] S. Hershey, S. Chaudhur, D. P. Ellis, J. F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, K. Wilson, “CNN architectures for large-scale audio classification”, *IEEE international conference on acoustics, speech and signal processing (icassp)*, p. 131-135, 2017

[13] N. Vryzas, L. Vrysis, R. Kotsakis, C. Dimoulas, “A web crowdsourcing framework for transfer learning and personalized speech emotion recognition” . *Machine Learning with Applications*, *6*, 2021.

[14] G. Mafra, N. Duong, A. Ozerov, P. Pérez. “Acoustic scene classification: An evaluation of an extremely compact feature representation”, *Detection and Classification of Acoustic Scenes and Events 2016*, Budapest, Hungary, 2016.

[15] S.D. You, C.-H. Liu, W.-K. Chen. "Comparative study of singing voice detection based on deep neural networks and ensemble learning." *Human-centric Computing and Information Sciences*, vol. 8, 34, pp.1-18, 2018.

[16] C.-E., Kuo, G.-T. Chen, P.-Yu. Liao, "An EEG spectrogram-based automatic sleep stage scoring method via data augmentation, ensemble convolution neural network, and expert knowledge", *Biomedical Signal Processing and Control*, vol. 70, 203982, pp. 1-13, 2021.

[17] J. Chen, S. Bi, G. Zhang, G. Cao, “High-density surface EMG-based gesture recognition using a 3D convolutional neural network”, *Sensors*, vol. 20, 4, 1201, pp. 1-13, 2020.

[18] T. -W. Sun, "End-to-End Speech Emotion Recognition With Gender Information", *IEEE Access*, vol. 8, pp. 152423-152438, 2020.

[19] N. Simić, S. Suzić, T. Nosek, M. Vujović, Z. Perić, M. Savić, V. Delić, “Speaker recognition using constrained convolutional neural networks in emotional speech”, *Entropy*, 2022, in press.

[20] Serbian Emotional Amateur Corpus, available at: https://www.ktios.ftn.uns.ac.rs/sadapt/SADAPT_publications.html