Special Issue Paper | Published: 28 March 2022

Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and features fusion

Rashid Jahangir, Ying Wah Teh, Ghulam Mujtaba, Roobaea Alroobaea, Zahid Hussain Shaikh & Ihsan Ali

Machine Vision and Applications **33**, Article number: 41 (2022)

Abstract

Speech emotion recognition (SER) is one of the most challenging and active research topics in data science due to its wide range of applications in human–computer interaction, computer games, mobile services and psychological assessment. In the past, several studies have employed handcrafted features to classify emotions and achieved good classification accuracy. However, such features degrade the classification accuracy in complex scenarios. Thus, recent studies employed deep learning models to automatically extract the local representation from given audio signals. Though, automated feature engineering overcomes the issues of handcrafted feature extraction approach. However, still there is a need to further improve the performance of reported techniques. This is because, in reported techniques, single-layer and two-layer convolutional neural networks (CNNs) were used and these architectures are not capable of learning optimal features from complex speech signals. Thus, to overcome this limitation, this study proposed a novel SER framework, which applies data augmentation methods before extracting seven informative feature sets from each utterance. The extracted feature vector is used as input to the 1D CNN for emotions recognition using the EMO-DB, RAVDESS and SAVEE databases. Moreover, this study also proposed a cross-corpus SER model using the all audio files of common emotions of aforementioned databases. The experimental results showed that our proposed SER framework outperformed existing SER frameworks. Specifically, the proposed SER framework obtained 96.7% accuracy for EMO-DB with all utterances in seven emotions, 90.6% RAVDESS with all utterances in eight emotions, 93.2% for SAVEE with all utterances in seven emotions and 93.3% for cross-corpus with 1930 utterances in six emotions. We believe that our proposed framework will bring significant contribute to SER domain.

Abbreviations

SER: Speech emotion recognition

HCI: Human–computer interaction

MFCC: Mel frequency cepstral coefficient

RAVESS: Ryerson audio-visual database of emotional speech and song

SAVEE: Surrey audio-visual expressed emotion database

CNN: Convolutional neural network

CL: Convolutional layer

ReLU: Rectifier linear unit

ZCR: Zero cross rate

HNR: Harmonics-to-noise ratio

MEDC: Mel energy spectrum dynamic coefficients

k-NN: K-nearest neighbor

SVM: Support vector machine

References

1. Chen, L., Su, W., Feng, Y., Wu, M., She, J., et al.: Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. Inf. Sci. **509**, 150–163 (2020)

2. Zheng, W., Zheng, W., Zong, Y.: Multi-scale discrepancy adversarial network for crosscorpus speech emotion recognition. Virtual Real. Intell. Hardw. **3**(1), 65–75 (2021)

3. Hansen, J.H., Cairns, D.A.: Icarus: Source generator based real-time recognition of speech in noisy stressful and lombard effect environments☆. Speech Commun. **16**(4), 391–422 (1995)

4. Koduru, A., Valiveti, H.B., Budati, A.K.: Feature extraction algorithms to improve the speech emotion recognition rate. Int. J. Speech Technol. **23**(1), 45–55 (2020)

5. Schuller, B., Rigoll, G., Lang, M.: Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I-577 (2004)

6. Spencer, C., Koç, İ.A., Suga, C., Lee, A., Dhareshwar, A.M., et al.: A comparison of unimodal and multimodal measurements of driver stress in real-world driving conditions. (2020)

7. France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., Wilkes, M.: Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Eng. **47**(7), 829–837 (2000)

8. Uddin, M.Z., Nilsson, E.G.: Emotion recognition using speech and neural structured learning to facilitate edge intelligence. Eng. Appl. Artif. Intell. **94**, 103775 (2020)

9. Jahangir, R., Teh, Y.W., Hanif, F., Mujtaba, G.: Deep learning approaches for speech emotion recognition: state of the art and research challenges. Multimed. Tools Appl. **80**, 1–66 (2021)

10. Abdel-Hamid, O., Mohamed, A.-R., Jiang, H., Deng, L., Penn, G., et al.: Convolutional neural networks for speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(10), 1533–1545 (2014)

11. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., et al.: 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), pp. 5200–5204 (2016).

12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural. Inf. Process. Syst. **25**, 1097–1105 (2012)

13. Fu, L., Mao, X., Chen, L.: Speaker independent emotion recognition based on SVM/HMMs fusion system. In: 2008 international conference

on audio, language and image processing, pp. 61–65 (2008).

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

15. Pawar, M.D., Kokate, R.D.: Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients. Multim. Tools Appl. **80**, 1–25 (2021)

16. Zhang, S., Tao, X., Chuang, Y., Zhao, X.: Learning deep multimodal affective features for spontaneous speech emotion recognition. Speech Commun. **127**, 73–81 (2021)

17. Issa, D., Demirci, M.F., Yazici, A.: Speech emotion recognition with deep convolutional neural networks. Biomed. Signal Process. Control **59**, 101894 (2020)

18. Sajjad, M., Kwon, S.: Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access **8**, 79861–79875 (2020)

19. Badshah, A.M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., et al.: Deep features-based speech emotion recognition for smart affective services. Multimed. Tools Appl. **78**(5), 5571–5589 (2019)

20. Er, M.B.: A novel approach for classification of speech emotions based on deep and acoustic features. IEEE Access **8**, 221640–221653 (2020)

21. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden Markov models. Speech Commun. **41**(4), 603–623 (2003)

22. Nicholson, J., Takahashi, K., Nakatsu, R.: Emotion recognition in speech using neural networks. Neural Comput. Appl. **9**(4), 290–296 (2000)

23. Noroozi, F., Sapiński, T., Kamińska, D., Anbarjafari, G.: Vocal-based emotion recognition using random forests and decision tree. Int. J. Speech Technol. **20**(2), 239–246 (2017)

24. Jahangir, R., Teh, Y.W., Memon, N.A., Mujtaba, G., Zareei, M., et al.: Text-independent speaker identification through feature fusion and deep neural network. IEEE Access **8**, 32187–32202 (2020)

25. Aljuhani, R.H., Alshutayri, A., Alahdal, S.: Arabic speech emotion recognition from saudi dialect corpus. IEEE Access **9**, 127081–127085 (2021)

26. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of German emotional speech. In: Ninth European Conference on Speech Communication and Technology (2005).

27. 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 **13**(5), e0196391 (2018)

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

29. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks

with cutout. arXiv preprint arXiv:1708.04552 (2017).

30. Chen, S., Dobriban, E., Lee, J.H.: A group-theoretic framework for data augmentation. J. Mach. Learn. Res. **21**(245), 1–71 (2020)

31. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., et al.: Deep speech: scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014).

32. Wei, S., Zou, S., Liao, F.: A comparison on data augmentation methods based on deep learning for audio classification. In: Journal of Physics: Conference Series, p. 012085, (2020).

33. Domingos, P.: A few useful things to know about machine learning. Commun. ACM **55**(10), 78–87 (2012)

34. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., et al.: librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference, pp. 18–25 (2015).

35. Palo, H.K., Chandra, M., Mohanty, M.N.: Recognition of human speech emotion using

variants of mel-frequency cepstral coefficients. In: Advances in Systems, Control and Automation. Springer, pp. 491-498 (2018)

36. Shahamiri, S.R., Thabtah, F.: An investigation towards speaker identification using a single-sound-frame. Multimed. Tools Appl. **79**(41), 31265–31281 (2020)

37. Wang, H.-C., Syu, S.-W., Wongchaisuwat, P.: A method of music autotagging based on audioand lyrics. Multimed. Tools Appl. **80**(10), 15511–15539 (2021)

38. Beigi, H.: Speaker recognition. In: Fundamentals of Speaker Recognition, pp. 543–559. Springer, Boston, MA (2011). https://doi.org/10.1007/978-0-387-77592-0_17

39. Harte, C., Sandler, M., Gasser, M.: Detecting harmonic change in musical audio. Presented at the Proceedings of the 1st ACM workshop on Audio and music computing multimedia, Santa Barbara, California, USA, 2006. [Online]. https://doi.org/10.1145/1178723.1178727.

40. Nweke, H.F., Teh, Y.W., Al-Garadi, M.A., Alo, U.R.: Deep learning algorithms for human activity recognition using mobile and wearable

sensor networks: state of the art and research challenges. Expert Syst. Appl. **105**, 233–261 (2018)

41. Garcia-Ceja, E., Riegler, M., Kvernberg, A.K., Torresen, J.: User-adaptive models for activity andemotion recognition using deep transfer learning and data augmentation. User Model User-Adap Inter. **30**, 365–393 (2020)

42. Nie, W., Ren, M., Nie, J., Zhao, S.: C-GCN: correlation based graph convolutional network for audio-video emotion recognition. IEEE Trans. Multimed. **23**(3793), 3804 (2020)

43. Gholamy, A., Kreinovich, V., Kosheleva, O.: Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation. Departmental Technical Reports (CS) 1209 (2018). https://scholarworks.utep.edu/cgi/viewcontent.cgi?article=2202&context=cs_techrep

44. Hajarolasvadi, N., Demirel, H.: 3D CNN-based speech emotion recognition using K-means clustering and spectrograms. Entropy **21**(5), 479 (2019)

45. Farooq, M., Hussain, F., Baloch, N.K., Raja,

F.R., Yu, H., et al.: Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. Sensors **20**(21), 6008 (2020)

46. Heracleous, P., Yoneyama, A.: A comprehensive study on bilingual and multilingual speech emotion recognition using a two-pass classification scheme. PLoS ONE **14**(8), e0220386 (2019)

47. Zhao, Z., Li, Q., Zhang, Z., Cummins, N., Wang, H., et al.: Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-Based discrete speech emotion recognition. Neural Netw. **141**, 52–60 (2021)

48. Kwon, S.: Att-Net: Enhanced emotion recognition system using lightweight self-attention module. Appl. Soft Comput. **102**, 107101 (2021)

Acknowledgements

The authors are grateful to the Taif University Researchers Supporting Project (Number TURSP-2020/36), Taif University, Taif, Saudi Arabia.

Author information

Authors and Affiliations

Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, 50603, Kuala Lumpur, Malaysia

Rashid Jahangir, Ying Wah Teh & Ihsan Ali

Department of Computer Science, COMSATS University Islamabad, Vehari Campus, Pakistan

Rashid Jahangir

Center of Excellence for Robotics, Artificial Intelligence and Blockchain, Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan

Ghulam Mujtaba

Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif, 21944, Saudi Arabia

Roobaea Alroobaea

Department of Mathematics, Sukkur IBA University, Sukkur, Pakistan

Zahid Hussain Shaikh

Corresponding authors

Correspondence to Rashid Jahangir or Ying Wah Teh.

Additional information

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Rights and permissions

Reprints and Permissions

About this article

Cite this article

Jahangir, R., Teh, Y.W., Mujtaba, G. *et al.* Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and features fusion. *Machine Vision and Applications* **33,** 41 (2022). https://doi.org/10.1007/s00138-022-01294-x

Received	Revised	Accepted
05 April 2021	23 December 2021	22 February 2022

Published

28 March 2022

DOI

https://doi.org/10.1007/s00138-022-01294-x

Keywords

Speech emotion recognition **Cross-corpus**

Signal processing **Deep learning**

Convolutional neural networks