AHD ConvNet for Speech Emotion Classification

Asfand Ali Irfan  
Fast NUCES  
i191656@nu.edu.pk

Danial Nasir  
Fast NUCES  
i191861@nu.edu.pk

Hassan Jawad  
Fast NUCES  
i191799@nu.edu.pk

Abstract
Accomplishments in the field of artificial intelligence are utilized in the advancement of computing and making of intelligent machines for facilitating mankind and improving user experience. Emotions are rudimentary for people, affecting thinking and ordinary exercises like correspondence, learning and direction. Speech emotion recognition is domain of interest in this regard and in this work, we propose a novel mel spectrogram learning approach in which our model uses the datapoints to learn emotions from the given waveform voice notes in the popular CREMA-D dataset. Our model uses log mel-spectrogram as feature with number of mels = 64. It took less training time compared to other approaches used to address the problem of emotion speech recognition.

1. Problem statement
The problem we have worked upon is speech emotion classification and this study drives an increasing attention towards recognition of emotions spontaneously from speech. This is an attempt to improve the validation accuracy of the recently released SepTr + LeRaC model increasing it from 70.95 to 71+.

2. Introduction
Speech emotion recognition is the act of recognizing the emotional state of the speaker and it is a new area of research. To make interaction between machines and humans more fluent and natural it is quite necessary that machines have an idea of the emotions in a given speech. The main objective of employing speech emotion recognition is to adapt system response upon detecting a change in the speaker’s voice.

For example, emotional speech recognition could be useful in a self-driving car where the emotions in the speech of the driver could help in determining the mental state of the driver and hence help the automated system to initiate the safety of the driver in case of an emergency.

The motivation behind doing research on emotional speech recognition was because of a wide variety of its applications including diagnostic tool for therapists, automatic translation systems in which the emotional state of the speaker plays an important role in communication between parties, call center applications and mobile communication.

3. Related work
Emotion speech recognition is still a progressive research area but some approaches have brought significant results. Among them is a system for learning audio portrayals directed by the visual methodology with regards to audiovisual speech. [16] utilized a generative audio to-video training plan in which a still picture relating to a given snippet was animated which enhanced the created video to be pretty much as close as conceivable to the genuine video of the speech segment. Through this cycle, the audio encoder network learned helpful speech portrayals that were assessed on feeling acknowledgment and speech recognition. [16] accomplished cutting-edge results for feeling acknowledgment and serious outcomes for speech recognition. This shows the capability of visual oversight for learning audio portrayals as a clever way for self-managed realizing which has not been investigated before.

The proposed unaided audio elements can use an essentially limitless measure of training information of unlabeled varying media audiovisual speech and have countless possibly encouraging applications. Another approach [16] is curriculum learning which requires to sort data samples by difficulty. Then an approach of [16] learning rate curriculum LeRaC is used which allows the use of different learning rate for every layer of neural network to make data free curriculum while the initial training epochs. In [16] it assigns higher learning rate to neural layer which is near and close to input. Then it decreases the learning rate as layers are placed further apart. Learning rate increase at different paces while the first training iterations until all the values become equal. After this neural model is trained normally. This approach was proposed in 2022 and it brought a validation accuracy of 70.95.

In [10] combining multiple machine learning models into an ensemble is known to provide superior performance levels because models can help each other in taking better decisions. Instead of just combining the models, they used a self-paced ensemble learning technique in which models learn from each other. During the self-paced learning process which is based upon pseudo labeling our ensemble also gains knowledge about the target domain. To check ensemble learning (SPEL) scheme. Experiment is done on three audio tasks. The results show that SPEL significantly outperforms the baseline ensemble models.

4. AHD ConvNet
We use log Mel spectrogram with number of mels = 64 for our features. For normalization, we subtracted the mean. We also tried standardization but it performed worse than subtracting the mean.

4.1 Approach: Speaker-invariant Embeddings
We wanted to incorporate speaker invariance in our approach so that the model can generalize its emotion recognition while only training on a few speaker samples. We tried a contrastive representation learning approach in which we trained embeddings for both speaker discrimination as well as emotion discrimination. The
motive was to get a higher resemblance between embeddings of the same emotion whereas get a lower resemblance between embeddings of the same speaker. In this way, the embeddings would only represent the information regarding emotions and would not contain any information regarding speakers. For the loss we used lifted structured loss [37] and soft nearest neighbor loss [20]. Lifted structured loss is similar to triplet loss but uses all possible triplets in the batch. It ended up being too computationally expensive for us as training was too slow. When using soft nearest neighbor loss, our embeddings didn’t train but more epochs may be required.

4.2 Approach: Activation Regularization
In this approach, we used 2 models with the same architecture that are trained in parallel. One model is trained on speaker discrimination and the other model is trained on emotion discrimination. The goal is to regularize the activations of the first layers of both the models so that they pick up different features from the input. It is desired that the regularization will result in the emotion discrimination model only picking up features relevant to the emotion of the audio and disregarding the features relevant to the speaker of the audio. Equivalently, the speaker discrimination model would only pick up features relevant to the speaker of the audio and disregard the features relevant to the emotion of the audio. The loss function we used for the regularization term were Euclidean distance between the activations, cosine distance between the activations, and l2 norm of the difference between the activations. This approach only resulted in worse accuracy compared to training a simple classifier. This may be due to there being a base shared representation that must be picked up by both the models.

4.3 Approach: Classification
Finally, we tried straight-forward classification with cross-entropy loss. We used both log mel spectrogram with 64 mels and raw audio as features. Raw audio has been used successfully in [21] but it did not perform well for us. More data may be required for training on raw audio. We tried various architectures on log mel spectrogram features. Our final architecture has 5 convolutional layers, 2 max pooling layers, and a fully connected layer. We used Relu activation function. This resulted in our highest accuracy of 60.38.

5. Evaluation and Experiments
5.1 Dataset
CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities. Actors spoke from a selection of 12 sentences and these sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified). We used the Librosa library in python for analysis of the dataset and made different graphical representations of a sound wave vibration overtime for different emotions. This is a more difficult dataset for emotion speech recognition compared to other datasets. The same sentence is spoken in different emotions which helps in text context invariance. It can be seen that there is a significant difference between wave forms for both emotions but not for the same sentence. Hence, it is not a head to head comparison but it still gives us a good overview.

Figure 1. SOUND WAVE for a happy emotion track

Figure 2. SOUND WAVE for a fear emotion track

5.2 Architectural Changes
We used different architectures including conv, lstm, conv-lstm, transformer, and conv-transformer. The Transformer failed to train when we did layernorm after self-attention and feed-forward but it did train when layernorm was done before self-attention and feed-forward. Accuracy was about 20 percent less when we didn’t provide the positional embedding to the transformer as it acted as a bag of words approach without positional embeddings. We changed the architecture of the model several times to improve the accuracy of the model. The convolutional model worked best. Inserting dropout layers in between convolutional layers even decreased the validation accuracy further. ReLu gave better results compared to Soft-Max activation function. The batch size was 125.

5.3 Evaluation
We evaluated all models in terms of the validation accuracy. We repeated the training process for 100 epochs and reported the maximum accuracy which was 60.38.

Figure 3. Loss Graph

Figure 4. Train Graph
6. Conclusion

In this paper we used a mel spectrogram based convolutional neural network to classify emotions. We carried out different experiments including LSTM and transformer-based approaches but the convolutional network gave better validation accuracy results compared to other approaches. Models training time was also less compared to other approaches. We will keep improving on the given approach and make updates to the open source code available.

References

[1] R. Ali, U. Farooq, U. Arshad, W. Shahzad, and M. O. Beg. Hate speech detection on twitter using transfer learning. Computer Speech & Language, 74:101365, 2022.

[2] H. M. Alvi, H. Sahar, A. A. Bangash, and M. O. Beg. Ensights: A tool for energy aware software development. In 2017 13th International Conference on Emerging Technologies (ICET), pages 1–6. IEEE, 2017.

[3] H. M. Alvi, H. Majeed, H. Mujtaba, and M. O. Beg. Mlee: Method level energy estimation—a machine learning approach. Sustainable Computing: Informatics and Systems, 32:100594, 2021.

[4] T. Anwar and O. Baig. Tae at semeval-2020 task 12: Ensembling approach for multilingual offensive language identification in social media. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pages 2177–2182, 2020.

[5] M. U. Arshad, M. F. Bashir, A. Majeed, W. Shahzad, and M. O. Beg. Corpus for emotion detection on roman urdu. In 2019 22nd International Multitopic Conference (INMIC), pages 1–6. IEEE, 2019.

[6] M. Asad, M. Asim, T. Javed, M. O. Beg, H. Mujtaba, and S. Abbas. Deepdetect: detection of distributed denial of service attacks using deep learning. The Computer Journal, 63(7):983–994, 2020.

[7] M. N. Awan and M. O. Beg. Top-rank: a topicalpostionrank for extraction and classification of keyphrases in text. Computer Speech & Language, 65:101–116, 2022.

[8] A. A. Bangash, H. Sahar, and M. O. Beg. A methodology for relating software structure with energy consumption. In 2017 IEEE 17th International Working Conference on Source Code Analysis and Manipulation (SCAM), pages 111–120. IEEE, 2017.

[9] M. F. Bashir, A. R. Javed, M. U. Arshad, T. R. Gadekallu, W. Shahzad, and M. O. Beg. Context aware emotion detection from low resource urdu language using deep neural network. Transactions on Asian and Low-Resource Language Information Processing, 2022.

[10] M. Beg. Critical path heuristic for automatic parallelization. University of Waterloo, David R. Cheriton School of Computer Science, Technical Report CS-2008-16, 2008.

[11] M. Beg and P. v. Beek. A constraint programming approach for integrated spatial and temporal scheduling for clustered architectures. ACM Transactions on Embedded Computing Systems (TECS), 13(1):1–23, 2013.

[12] M. Beg and M. Dahlin. A memory accounting interface for the java programming language. Technical Report CS-TR-01–40, University of Texas at Austin, 2001.

[13] M. Beg and P. Van Beek. A graph theoretic approach to cache-conscious placement of data for direct mapped caches. In Proceedings of the 2010 international symposium on Memory management, pages 113–120, 2010.

[14] M. Beg and P. Van Beek. A constraint programming approach for instruction assignment. In 2011 15th Workshop on Interaction between Compilers and Computer Architectures, pages 25–34. IEEE, 2011.

[15] M. O. Beg, M. N. Awan, and S. S. Ali. Algorithmic machine learning for prediction of stock prices. In FinTech as a Disruptive Technology for Financial Institutions, pages 142–169. IGI Global, 2019.

[16] F.-A. Croitoru, N.-C. Ristea, R. T. Ionescu, and N. Sebe. Lerac: Learning rate curriculum, 2022. URL https://arxiv.org/abs/2205.09180.

[17] N. Dilawar, H. Majeed, M. O. Beg, N. Ejaz, K. Muhammad, I. Mehmood, and Y. Nam. Understanding citizen issues through reviews: A step towards data informed planning in smart cities. Applied Sciences, 8(9):1589, 2018.

[18] M. U. Farooq, M. O. Beg, et al. Bigdata analysis of stack overflow for energy consumption of android framework. In 2019 International Conference on Innovative Computing (ICIC), pages 1–9. IEEE, 2019.

[19] M. U. Farooq, S. U. R. Khan, and M. O. Beg. Melta: A method level energy estimation technique for android development. In 2019 International Conference on Innovative Computing (ICIC), pages 1–10. IEEE, 2019.

[20] N. Frosst, N. Papernot, and G. Hinton. Analyzing and improving representations with the soft nearest neighbor loss. In K. Chaudhuri and R. Salakhutdinov, editors, Proceedings of the 35th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 2012–2020. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/frosst19a.html.

[21] A. Gazneli, G. Zimerman, T. Ridnik, G. Sharrir, and A. Noy. End-to-end audio strikes back: Boosting augmentations towards an efficient audio classification network. ArXiv, abs/2204.11479, 2022.

[22] S. Ismail, H. Mujtaba, and M. O. Beg. Spems: A sustainable parasitic energy management system for smart homes. Energy and Buildings, 252:111419, 2021.

[23] A. R. Javed, M. O. Beg, M. Asim, T. Baker, and A. H. Al-Bayatti. Alphalogger: Detecting motion-based side-channel attack using smartphone keystrokes. Journal of Ambient Intelligence and Humanized Computing, pages 1–14, 2020.

[24] A. R. Javed, M. U. Sarwar, M. O. Beg, M. Asim, T. Baker, and H. Tawfi. A collaborative healthcare framework for shared healthcare plan with ambient intelligence. Human-centric Computing and Information Sciences, 10(1):1–21, 2020.

[25] H. T. Javed, M. O. Beg, H. Mujtaba, H. Majeed, and M. Asim. Fairness in real-time energy pricing for smart grid using unsupervised learning. The Computer Journal, 62(3):414–429, 2019.

[26] M. S. Javed, H. Majeed, H. Mujtaba, and M. O. Beg. Fake reviews classification using deep learning ensemble of shallow convolutions. Journal of Computational Social Science, 4(2):883–902, 2021.

[27] M. Karsten, S. Keshav, S. Prasad, and M. Beg. An axiomatic basis for communication. ACM SIGCOMM Computer Communication Review, 37(4):217–228, 2007.

[28] H. S. Khawaja, M. O. Beg, and S. Qamar. Domain specific emotion lexicon expansion. In 2018 14th International Conference on Emerging Technologies (ICET), pages 1–5. IEEE, 2018.

[29] A. Majeed, H. Mujtaba, and M. O. Beg. Emotion detection in roman urdu text using machine learning. In Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering Workshops, pages 125–130, 2020.

[30] H. Majeed, A. Wali, and M. Beg. Optimizing genetic programming by exploiting semantic impact of sub trees. Swarm and Evolutionary Computation, 65:100923, 2021.
[31] B. Naeem, A. Khan, M. O. Beg, and H. Mujtaba. A deep learning framework for clickbait detection on social area network using natural language cues. *Journal of Computational Social Science*, pages 1–13, 2020.

[32] S. Naeem, M. Iqbal, M. Saqib, M. Saad, M. S. Raza, Z. Ali, N. Akhtar, M. O. Beg, W. Shahzad, and M. U. Arshad. Subspace gaussian mixture model for continuous urdu speech recognition using kaldi. In *2020 14th International Conference on Open Source Systems and Technologies (ICOSSST)*, pages 1–7. IEEE, 2020.

[33] S. Qamar, H. Mujtaba, H. Majeed, and M. O. Beg. Relationship identification between conversational agents using emotion analysis. *Cognitive Computation*, pages 1–15, 2021.

[34] N.-C. Ristea, R. T. Ionescu, and F. S. Khan. Septr: Separable transformer for audio spectrogram processing. 2022. URL https://arxiv.org/abs/2203.09581.

[35] H. Sahar, A. A. Bangash, and M. O. Beg. Towards energy aware object-oriented development of android applications. *Sustainable Computing: Informatics and Systems*, 21:28–46, 2019.

[36] S. Singh and D. Schicker. Seven basic expression recognition using resnet-18, 2021. URL https://arxiv.org/abs/2107.04569.

[37] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. *CoRR*, abs/1511.06452, 2015. URL http://arxiv.org/abs/1511.06452.

[38] M. Tariq, H. Majeed, M. O. Beg, F. A. Khan, and A. Derhab. Accurate detection of sitting posture activities in a secure iot based assisted living environment. *Future Generation Computer Systems*, 92:745–757, 2019.

[39] A. Uzair, M. O. Beg, H. Mujtaba, and H. Majeed. Weec: Web energy efficient computing: A machine learning approach. *Sustainable Computing: Informatics and Systems*, 22:230–243, 2019.

[40] A. Zafar, H. Mujtaba, M. O. Beg, and S. Ali. Deceptive level generator. In *AIIDE Workshops*, 2018.

[41] A. Zafar, H. Mujtaba, S. Ashiq, and M. O. Beg. A constructive approach for general video game level generation. In *2019 11th Computer Science and Electronic Engineering (CEEC)*, pages 102–107. IEEE, 2019.

[42] A. Zafar, H. Mujtaba, M. T. Baig, and M. O. Beg. Using patterns as objectives for general video game level generation. *ICGA Journal*, 41 (2):66–77, 2019.

[43] A. Zafar, H. Mujtaba, and M. O. Beg. Search-based procedural content generation for gvg-lg. *Applied Soft Computing*, 86:105909, 2020.