Convolutional and Recurrent Neural Networks for Spoken Emotion Recognition

Aaron Keesing, Ian Watson, Michael Witbrock
School of Computer Science, University of Auckland

Abstract

We test four models proposed in the speech emotion recognition (SER) literature on 15 public and academic licensed datasets in speaker-independent cross-validation. Results indicate differences in the performance of the models which is partly dependent on the dataset and features used. We also show that a standard utterance-level feature set still performs competitively with neural models on some datasets. This work serves as a starting point for future model comparisons, in addition to open-sourcing the testing code.

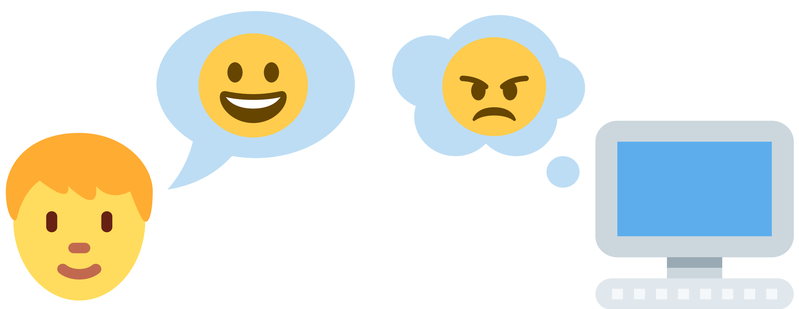
Introduction

Speech emotion recognition (SER) is the analysis of speech to predict the emotional state of the speaker, for which there are many current and potential applications. As speech-enabled devices become more prevalent, the need for reliable and robust SER increases, and also the need for comparability of results on common datasets. While much research has been done in this field, there are varying methodologies and datasets employed, reducing comparability.

In this paper, we aim to:

- Test SER models proposed in the literature on a discrete emotion classification task.
- Promote reproducibility of results by using public and academic licensed datasets
- Provide open-source code at `github.com/Broad-AI-Lab/emotion`

The two main benefits of our research are: baseline results for future research and comparison of results between datasets to see which of their properties may influence performance.



Methodology

We test on 15 public datasets under either open or academic licenses:

- **Open datasets**: CaFE, CREMA-D, EMO-DB, eNTERFACE, JL corpus, a Portuguese dataset [1], RAVDESS, ShEMO, TESS
- **Academic datasets**: DEMoS, EmoFilm, IEMOCAP, MSP-IMPROV, SAVEE, SmartKom.

We test 4 neural network architecture previously mentioned in the SER literature, from: Aldeneh and Mower Provost [2], Latif et al. [3], Zhang et al. [4], Zhao et al. [5].

Each model is tested on each dataset in leave-one-speaker-out (LOSO) or leave-one-speaker-group-out (LOSGO) cross-validation. LOSGO is used in the IEMOCAP and MSP-IMPROV datasets based on sessions. If the dataset has more than 12 speakers, 6 random speaker groups are used. Each model was trained for 50 epochs with a learning rate of 0.0001 and model-dependent batch size between 16 and 64.

Results

A table of results is show below, all numbers are % unweighted average recall. Human accuracy is mentioned where relevant.

Corpus	A1	A2	A3	L	N	O	SVM-IS09	Human
CaFE	53.8	54	52.1	22.3	32.3	48	**57.2**	
CREMA-D	66.6	**67**	63.4	42.4	48.4	57.9	65	40
DEMoS	61.4	**61.9**	61.5	25.5	26.9	45.7	51.2	61.1
EMO-DB	73.2	74.6	72.7	45.2	49.7	53.7	**82.1**	84.3
EmoFilm	49.6	49.7	49.4	40.2	45.6	44.7	**53.2**	73
eNTERFACE	77.9	**79.4**	77.4	38.6	45	66.4	76.3	
IEMOCAP	**61.1**	60.5	58.2	46.2	49.2	58.3	59.8	73.8
JL	65.8	**67.8**	47.9	54	61.2	46.6	66.2	69.1
MSP-IMPROV	47.2	47.5	46.2	35.2	38	48.6	**52.4**	77.8
Portuguese	38.3	39	41.5	37.4	43.3	39.9	**50**	73.2
RAVDESS	32.5	39.5	60	29.6	32.9	43	**60.6**	62.5
SAVEE	58.4	**59.6**	48.5	34.8	33	30.1	57	66.5
ShEMO	54.6	**55.7**	50.7	43.6	48.4	51.8	51.3	
SmartKom	15.8	16.8	17.5	16	16.7	22.6	**28.5**	
TESS	48.7	49.5	**55.1**	38.5	30.6	48.4	45.9	82

A1: Aldeneh model with variable 40 log-mels. A2: Aldeneh model with variable 240 log-mels. A3: Aldeneh model with fixed 5s 240-mel spectrogram. L: Latif model with 5s raw audio. N: Zhang model with 5s raw audio. O: Zhao model with fixed 5s 40-mel spectrogram

Discussion

- Models using raw audio as input tend to perform worse than those using spectrogram input, with exceptions such as the Portuguese and JL datasets. One reason might be suboptimal hyperparameters used.
- Models using raw audio also tended to overfit much more than the other models even with a moderate number of parameters. Regularisation techniques may help mitigate this.
- Three of the models perform quite poorly. This is likely an implementation issue, as results in the respective papers are higher even with slightly different methodologies. We are continuing testing and are discussing with respective authors.

Conclusion

We have compared the performance of 4 different neural network models for discrete emotion classification on 15 datasets, using a consistent methodology. Results show performance dependent on both classifier and dataset, although more thorough testing is required to explain the lower-than-expected performance of some of the models.

Future work:

- Test models using utterance level features
- Compare with classifiers such as SVM and random forests.
- Test feature generation methods such as bag-of-audio-words and representation learning.

References

[1] S. L. Castro and C. F. Lima, 'Recognizing emotions in spoken language: A validated set of Portuguese sentences and pseudosentences for research on emotional prosody', Behavior Research Methods, vol. 42, no. 1, pp. 74–81, Feb. 2010.
[2] Z. Aldeneh and E. Mower Provost, 'Using regional saliency for speech emotion recognition', in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Mar. 2017, pp. 2741–2745.
[3] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, 'Direct Modelling of Speech Emotion from Raw Speech', in Interspeech 2019, Graz, Sep. 2019, pp. 3920–3924.
[4] Z. Zhang, B. Wu, and B. Schuller, 'Attention-augmented End-to-end Multi-task Learning for Emotion Prediction from Speech', in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019, pp. 6705–6709.
[5] Z. Zhao et al., 'Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition', IEEE Access, vol. 7, pp. 97515–97525, 2019.