Convolutional and Recurrent Neural Networks for Spoken Emotion Recognition

Aaron Keesing
School of Computer Science
University of Auckland
New Zealand
akee511@aucklanduni.ac.nz

Ian Watson
School of Computer Science
University of Auckland
New Zealand
ian@cs.auckland.ac.nz

Michael Witbrock
School of Computer Science
University of Auckland
New Zealand
m.witbrock@auckland.ac.nz

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.

1 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 (Peter and Beale, 2008; Koolagudi and Rao, 2012). 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 there has been a large amount of research in this field, a lot of results come from testing only on one or two datasets, which may or may not be publicly available. Additionally, different methodologies are often used, reducing direct comparability of results. Given the wide variety of neural architectures and testing methodologies, there is need for a common testing framework to help comparisons.

This study aims to test some SER models proposed in the literature on a discrete emotion classification task, and promote reproducibility of results by using public and academic licensed datasets. In addition, the code is publicly hosted on GitHub[1] under an open source license, so that our results may be verified and built upon. Our work has two main benefits. First, it serves as a baseline reference for future research that uses datasets present in this study. Second, it allows for comparisons between datasets to see which of their properties may influence classification performance of different models.

The paper is structured as follows. In Section 2 related work is given, and in Section 3 we list the datasets used in this study. The tested methods are outlined in Section 4, and the results given in Section 5. We briefly discuss these results in Section 6 and a conclude in Section 7.

2 Related Work

There has been some previous work in comparing SER techniques on a number of datasets. In Schuller et al. (2009a), Schuller et al. compare a hidden Markov model/Gaussian mixture model (HMM/GMM) and a SVM classifier for emotion class, arousal and valence prediction on nine datasets. For HMM/GMM, 12 MFCC, log-frame-energy, speed and acceleration features, are extracted per frame. For SVM, 6552 features are extracted based on 39 statistical functionals of 56 low-level descriptors (LLDs). Testing was done in a leave-one-speaker-out (LOSO) or leave-one-speaker-group-out (LOSGO) cross-validation setup. The only three datasets in common with the present study are EMO-DB, eNTERFACE and SmartKom, for which unweighted average recall (UAR) of 84.6%, 72.5%, and 23.5% were achieved, respectively. We use a similar methodology in the present paper.

The Schuller et al. work is expanded in Stuhlsatz et al. (2011), where multi-layer stacks of restricted Boltzmann machines (RBMs) are pre-trained in an unsupervised manner, then fine-tuned using backpropagation as a feed-forward neural network. The same datasets and configurations are used

[1] https://github.com/Broad-AI-Lab/emotion

	Emotion									
Dataset	ang.	hap.	sad.	fear	sur.	dis.	neu.	oth.	unk.	Total
CaFE	144	144	144	144	144	144	72			936
CREMA-D	1271	1271	1270	1271		1271	1087			7441
DEMoS	246	167	422	177	203	140		209		1564
EMO-DB	127	71	62	69		46	79	81		535
EmoFilm	232	240	254	221		168				1115
eNTERFACE	215	212	215	215	215	215				1287
IEMOCAP	1103	1636	1084				1708			5531
JL-corpus	240	240	240				240	240		1200
MSP-IMPROV	792	2644	885				3477			7798
Portuguese	63	46	59	41	64	35	60			368
RAVDESS	192	192	192	192	192	192	96	192		1440
SAVEE	60	60	60	60	60	60	120			480
ShEMO	1059	201	449		225		1028			2962
SmartKom	99	118	54		9		1786	183	46	2295
TESS	400	400	400	400	400	400	400			2800

Table 1: Dataset emotion distribution. The number of clips in each of the 'big six' emotions along with neutral and other, is given, as well as the total number of clips in each dataset. *oth.* = other (dataset specific); *unk.* = unknown

as in Schuller et al. (2009a), but the all-class emotion classification results are better on only some of the datasets. In particular, GerDA performs slightly better on average for SmartKom, but slightly worse for EMO-DB and eNTERFACE. In the current work, we compare many more methods on many more datasets; we also include more recent datasets.

3 Datasets

Fifteen datasets are used in this study, some of which are open datasets, while others require a signed EULA to access. All of the datasets have a set of categorical emotional labels. A question arises when using acted datasets with additional annotations, such as CREMA-D, as to whether to use the actor's intended emotion as 'ground truth' for training a classifier or instead use a consensus of annotators with majority vote. For MSP-IMPROV and IEMOCAP, the label assigned by annotators is used, consistent with previous work. For CREMA-D we have opted to use the actors intended emotion, rather than any annotator assigned labels. A table describing the emotion distribution in each dataset is given in Table 1.

3.1 Open Datasets

Open datasets are those under a free and permissive license, and are able to be downloaded with requesting permission or signing an academic license. The open datasets used in this study are: *Canadian-French emotional* dataset (Gournay et al., 2018), *Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D)* (Cao et al., 2014), *EMO-DB* (Burkhardt et al., 2005), *eNTERFACE* dataset (Martin et al., 2006), *JL corpus* (James et al., 2018), *Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)* (Livingstone and Russo, 2018), *Sharif Emotional Speech Database (ShEMO)* (Mohamad Nezami et al., 2019), and the *Toronto Emotional Speech Set (TESS)* (Dupuis and Pichora-Fuller, 2011).

3.2 Licensed Datasets

Licensed datasets are those that require signing an academic or other license in order to gain access to the data. The licensed datasets used in this study are: *Database of Elicited Mood in Speech (DEMoS)* (Parada-Cabaleiro et al., 2019), *EmoFilm* (Parada-Cabaleiro et al., 2018), *Interactive Emotional Dyadic Motion Capture (IEMOCAP)* (Busso et al., 2008), *MSP-IMPROV* (Busso et al., 2017), *Surrey Audio-Visual Expressed Emotion (SAVEE)* database (Haq et al., 2008), and the *SmartKom* corpus, public set (Schiel et al., 2002).

4 Methodology

4.1 Models

We implement four neural network models that have been proposed in previous literature. These

Model	Input	CNN	RNN	Att.	# Params
Aldeneh	2D	✓			4.4M–13.7M
Latif	1D	✓	✓		1.2M
Zhang	1D	✓	✓	✓	0.7M
Zhao	2D	✓	✓	✓	0.6M

Table 2: Summary table of model parameters. CNN: convolutional neural network. RNN: recurrent neural network. Att.: attention pooling. The number of parameters for the Aldeneh model depends on the number of frequency bands in the input.

models were selected with the goal of having a variety of model types (convolutional and recurrent), variety of input formats (spectrogram and raw audio), and recency (within the past few years). After each model is introduced with citation, it will subsequently be referred to by the primary author's surname. A summary table of model types and number of parameters is given in Table 2. Each model outputs are probability distribution over N classes.

We implement the final model from Aldeneh and Mower Provost (2017). This consists of four independent 1D convolutions, followed by maxpooling. The resulting vectors are concatenated into a feature vector which is passed to two fully-connected layers. The Aldeneh model takes a 2D sequence of log Mel-frequency spectrograms as input.

The model from Latif et al. (2019) consists of 3 independent 1D convolutions of with batch normalisation and maxpooling. The filters are concatenated feature-wise and a 2D convolution is performed, again with batch normalisation and maxpooling. The final 1920-dimensional feature sequence is passed through a LSTM block, followed by 30% dropout and a fully-connected layer. The Latif model takes 1D raw audio as input.

The model from Zhao et al. (2019) consists of a convolutional branch and a recurrent branch that act on 2D spectrograms. The recurrent branch consists of a bidirectional LSTM with a single layer, whereas in the paper they used two layers. The convolutional branch consists of three sequential 2D convolutions, with batch normalisation, max-pooling and dropout. The filters and kernel sizes are different across convolutions and the resulting time-frequency axes are flattened and passed through a dense layer. The convolutional and recurrent branches are individually pooled using weighted attention pooling, concatenated and finally passed through a dense layer.

The model proposed in Zhang et al. (2019) acts on a raw audio waveform. The audio is framed with a frame size of 640 samples and shift of 160 samples. Two 1D convolutions with maxpooling are calculated along the time dimension. The features are then pooled in the feature dimension and flattened to a 1280-dimensional vector per frame. The sequences are fed into a 2-layer GRU, before weighted attention pooling, as in the Zhao model. Although this model was originally designed to perform multi-task discrete valence and arousal classification, we apply it to the single-task emotion label classification.

4.2 Cross-validation

We perform leave-one-speaker-out (LOSO) or leave-one-speaker-group-out (LOSGO) cross-validation for all tests. Before testing, we perform per-speaker standardisation of feature columns, as in (Schuller et al., 2009a). If a dataset has more than 12 speakers, then 6 random speaker groups are chosen for cross-validation. For IEMOCAP and MSP-IMPROV, each session defines a speaker group. All models are trained for 50 epochs with the Adam optimiser (Kingma and Ba, 2017) and a learning rate of 0.0001. The batch size used for the Aldeneh and Latif models was 32, for the Zhao model was 64, and for the Zhang model was 16. Each was trained using sample weights inversely proportional to the respective class sizes, so the each class had equal total weight. The sample weights were used to scale the cross-entropy loss. The metric reported is 'unweighted average recall' (UAR), which is simply the mean of the per-class recall scores. This incorporates all classes equally even if there is a large class bias, and minimises the effect of class distribution on the reported accuracy, so that models can't simply optimise for the majority class. Each test is repeated 3 times and averaged, except for the Zhang model, which was only tested once, because it took too long to train.

All models were implemented in Python using the TensorFlow[2] Keras API. Testing was run on a machine with 64GB of RAM, an AMD Ryzen 3900X CPU, and two NVIDIA GeForce RTX 2080 Super GPUs, each with 8GB of VRAM. Each training run used only one GPU, however.

For the Zhang et al. (2019) and Latif et al. (2019) models we use the raw time domain signals. These are clipped to a maximum length of 80,000 samples (5 seconds at 16,000 kHz), but not padded, unlike the fixed spectrograms. For the Zhao et al. (2019) model we input a 5 second log-mel spectrogram with 40 mel bands calculated using a frame size of 25ms and frame shift of 10ms. Audio is clipped/padded to exactly 5 seconds. For the Aldeneh and Mower Provost (2017) model we test three different inputs: a 5 second 240 mel band spectrogram, 240 log-mel bands without clipping/padding, and 40 log-mel bands without clipping/padding. The log-mel bands are variable length sequences and are length-padded to the nearest larger multiple of 64, before batching. This way the models train with different sequence lengths.

5 Results

A table of results is given in Table 3 below. All combinations of dataset and model+features were tested. For comparison, we also report on the performance of the 'IS09' standard feature set introduced in the first INTERSPEECH emotion competition (Schuller et al., 2009b). For this we use a support vector machine (SVM) with radial basis function (RBF) kernel, with SVM parameter C and kernel parameter γ optimised using LOS(G)O cross-validation. We also report human accuracy where it has either been mentioned in the corresponding citation, or can be calculated from multiple label annotations provided with the dataset.

6 Discussion

From the results we see that the models using raw audio as input tend to perform worse than those using spectrogram input. There are also cases, such as on the Portuguese dataset, where the Zhang model performs the best of the four, and such as on the JL corpus, where the raw audio models are better than the fixed-size spectrogram models but worse than the variable length log-mel models.

There are many possible reasons for this, and due to time constraints, more thorough investigation was not able to be done. One reason is likely the lack of hyperparameter tuning. Hyperparameters like number of training epochs, learning rate, batch size, and model specific hyperparameters such as the number of convolution kernels or number of LSTM units, can have a moderate effect on the performance of each model. These would need to be optimised per-dataset using cross-validation, before testing. Another possible reason is the tendency for models to overfit. We found that the raw audio models were overfitting quite badly and achieving worse performance on the test set as a result, even though they have a moderate number of parameters. Regularisation techniques can help with this, such as dropout and regularisation loss, along with batch normalisation. Finally, while we tried to make our models as similar as possible to the original papers, there are likely implementation differences that negatively influence the performance of our models. The design of the Zhang model was for discrete arousal/valence prediction, and it is likely that a slightly modified architecture would better suit categorical emotion prediction. The other models were also tested with slightly different methodologies from ours, which would influence difference in reported results.

We also see a dependence on both dataset and features used. The Aldeneh model with 240 log-mels tended to be better than with only 40 log-mels, but also better than a fixed size 240 mel-band spectrogram, but this was dependent on dataset. It's possible that the zero-padding and -60dB clipping of the spectrograms negatively impacted the performance. The Zhao model performs best out of the four on the SmartKom dataset, achieving a UAR better than chance level, but still worse than the SVM with IS09 features. It's possible that in this instance the separate LSTM and convolutional branches have a greater effect. Unfortunately we were not able to test all combinations of spectrogram features with the Zhao model. In future we aim to complete this, as well as compare using spectrograms with different frame size and clipping parameters.

Finally, the time taken to train these models is quite long due to using full cross-validation. An argument can be made for predefined training/validation/test sets of larger datasets, but these are often created ad hoc and can vary between studies, so collective agreement would be needed for using these as a common standard.

[2] https://www.tensorflow.org/

Corpus	A1	A2	A3	L	N	O	SVM-IS09	Human
CaFE	53.8	54.0	52.1	22.3	32.3	48.0	**57.2**	
CREMA-D	66.6	**67.0**	63.4	42.4	48.4	57.9	65.0	40.0
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	45.0	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	54.0	61.2	46.6	66.2	69.1
MSP-IMPROV	47.2	47.5	46.2	35.2	38.0	48.6	**52.4**	77.8
Portuguese	38.3	39.0	41.5	37.4	43.3	39.9	**50.0**	73.2
RAVDESS	32.5	39.5	60.0	29.6	32.9	43.0	**60.6**	62.5
SAVEE	58.4	**59.6**	48.5	34.8	33.0	30.1	57.0	66.5
ShEMO	54.6	**55.7**	50.7	43.6	48.4	51.8	51.3	
SmartKom	15.8	16.8	17.5	16.0	16.7	22.6	**28.5**	
TESS	48.7	49.5	**55.1**	38.5	30.6	48.4	45.9	82

Table 3: Table of results. All values are given in UAR. 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. Human accuracy is the average accuracy of a human rater, either tested in the relevant citation, or calculated directly from annotations (e.g. CREMA-D).

7 Conclusion

In this paper we have presented an evaluation of different neural network models proposed for emotion recognition, and compared their performance for discrete emotion classification on 15 publicly available and academic datasets. We used a consistent methodology across all datasets, and have kept hyperparameters very similar across the proposed models. The results show differences in the performance of the models which sometimes depends on the evaluated dataset. We also showed that the models requiring raw audio input tended to perform worse than the ones requiring spectrogram input, however more testing is required, with hyperparameter tuning and regularisation techniques, to determine the cause of this performance difference. In general, our work serves as a baseline for comparison for future research.

In future, we aim to additionally test models using utterance level features as input, and compare with non-neural network models such as SVM and random forests. We also aim to test feature generation methods such as bag-of-audio-words and unsupervised representation learning.

Acknowledgements

The authors would like to thank the University of Auckland for funding this research through a PhD scholarship. We would like to thank in particular the School of Computer Science for providing the computer hardware to train and test these models. We would also like to thank the anonymous reviewers who submitted helpful feedback on this paper.

References

Zakaria Aldeneh and Emily Mower Provost. 2017. Using regional saliency for speech emotion recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2741–2745. ISSN: 2379-190X.

Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. 2005. A database of German emotional speech. In *Interspeech 2005 - Eurospeech*, pages 1517–1520, Lisbon, Portugal.

C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost. 2017. MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335.

H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. 2014. CREMA-

D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Kate Dupuis and M. Kathleen Pichora-Fuller. 2011. Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set. *Canadian Acoustics*, 39(3):182–183.

Philippe Gournay, Olivier Lahaie, and Roch Lefebvre. 2018. A canadian french emotional speech dataset. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 399–402. ACM.

Sanaul Haq, Philip JB Jackson, and James Edge. 2008. Audio-visual feature selection and reduction for emotion classification. In *International Conference on Auditory-Visual Speech Processing 2008*, pages 185–190, Tangalooma Wild Dolphin Resort, Moreton Island, Queensland, Australia.

Jesin James, Li Tian, and Catherine Inez Watson. 2018. An Open Source Emotional Speech Corpus for Human Robot Interaction Applications. In *Interspeech*, pages 2768–2772.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. ArXiv: 1412.6980.

Shashidhar G. Koolagudi and K. Sreenivasa Rao. 2012. Emotion recognition from speech: a review. *International Journal of Speech Technology*, 15(2):99–117.

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Julien Epps. 2019. Direct Modelling of Speech Emotion from Raw Speech. *arXiv:1904.03833 [cs, eess]*. ArXiv: 1904.03833.

Steven R. Livingstone and Frank A. Russo. 2018. 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.

O. Martin, I. Kotsia, B. Macq, and I. Pitas. 2006. The eNTERFACE' 05 Audio-Visual Emotion Database. In *22nd International Conference on Data Engineering Workshops*, pages 8–8.

Omid Mohamad Nezami, Paria Jamshid Lou, and Mansoureh Karami. 2019. ShEMO: a large-scale validated database for Persian speech emotion detection. *Language Resources and Evaluation*, 53(1):1–16.

Emilia Parada-Cabaleiro, Giovanni Costantini, Anton Batliner, Alice Baird, and Björn Schuller. 2018. Categorical vs Dimensional Perception of Italian Emotional Speech. In *Interspeech 2018*, pages 3638–3642. ISCA.

Emilia Parada-Cabaleiro, Giovanni Costantini, Anton Batliner, Maximilian Schmitt, and Björn W. Schuller. 2019. DEMoS: an Italian emotional speech corpus. *Language Resources and Evaluation*.

Christian Peter and Russell Beale, editors. 2008. *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*. Number 4868 in Lecture Notes in Computer Science. Springer Science & Business Media. Google-Books-ID: Bwd-cWtO666EC.

Florian Schiel, Silke Steininger, and Ulrich Türk. 2002. The SmartKom Multimodal Corpus at BAS. In *Third International Conference on Language Resources and Evaluation*, pages 200–206, Las Palmas, Canary Islands, Spain. Citeseer.

B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth. 2009a. Acoustic emotion recognition: A benchmark comparison of performances. In *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, pages 552–557.

Björn Schuller, Stefan Steidl, and Anton Batliner. 2009b. The INTERSPEECH 2009 emotion challenge. In *10th Annual Conference of the International Speech Communication Association*, pages 312–315, Brighton, United Kingdom.

A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller. 2011. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5688–5691.

Z. Zhang, B. Wu, and B. Schuller. 2019. 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)*, pages 6705–6709.

Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller. 2019. Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition. *IEEE Access*, 7:97515–97525.