Capstone Proposal:
Speech Emotion Recognition using Deep Learning

Kris Tost
Machine Learning Nanodegree Program
Udacity

March 20, 2018

1 Domain Background

Emotion Recognition in computing involves the detection of human emotion from one or more modalities such as audio speech, facial features in images or videos, or the semantic and linguistic analysis of transcribed speech [1]. Speech Emotion Recognition (SER) utilises acoustic features present in speech to perform emotion classification through purely audio-based modalities [2].

In a future where robotics [3], assistive living devices [4, 5] and virtual companions are the norm [6, 7], it is imperative that such applications can detect and respond intelligently to human emotion [8, 9], particularly in situations where other modalities such as facial features and language semantics aren't available due to device constraints or even privacy restrictions.

1.1 Personal Motivation

The author recently completed a comparative study of various Support Vector Machines (SVM) coupled with the Bag-of-Audio-Words (BoAW) approach [10] for SER tasks. However, the author was unable to explore state-of-the-art Deep Learning techniques for SER within the scope of that study. Therefore, this capstone project seeks to explore the use of Deep Learning techniques in classifying human emotion from audio-only speech samples using Emo-DB, a well-known and highly cited emotional speech corpus [12].

2 Problem Statement

Emotion classification models used in SER typically use discrete categorical labels [13], often based on the big 6 emotions identified by Paul Ekman [14]. For example, the labels 'anger', 'happiness', and 'sadness' are frequently used categorical labels employed to describe emotional speech. Predicting and classifying emotional speech in such an instance can therefore be considered a multi-classification supervised learning problem.

The challenge for this capstone project is to apply supervised learning algorithms to develop a model capable of classifying audio recordings of emotionally-charged human speech by outputting a discrete emotion class label, and/or a set of probabilities for the emotion classes under consideration. The goal is to treat the SER problem as an image classification task by training a Convolutional Neural Net (CNN) on spectrogram images of the spoken utterances [15, 16, 17].

3 Data set and Inputs

This project will utilise the Emo-DB[1] emotional speech corpus [12]. It contains 535 utterances of acted emotional speech from 10 German-speaking actors (5 female, 5 male). A total of 7 emotion classes are represented in the corpus consisting of the emotions anger, boredom, disgust, fear, happiness, sadness,

[1] `http://www.emodb.bilderbar.info/download/download.zip`

and neutral. Emo-DB is an oft-cited benchmark speech corpus [2, 18, 19] and is freely available for download with no license requirements. The emotions expressed are highly acted and recorded under ideal, noise-free acoustic conditions, making it a suitable benchmark dataset for establishing the viability of a classifier and relating its performance against other approaches documented in the literature [18].

A common data extraction technique for audio processing tasks such as speech recognition and speech emotion recognition is the creation of spectrogram images of audio [15, 20]. A spectrogram is a 2D time-frequency representation of the changes in signal strength of various frequencies and how they vary in time [21]. Frequency is typically represented on the y-axis, with time on the x-axis. Amplitude at a given time and frequency is indicated by the intensity of colour, with low amplitudes represented by darker colours and higher amplitudes by lighter colours [15] [21]. Spectrograms can be generated from the raw audio by segmenting the signal into sequences of frames and then applying Short-time Fast Fourier Transforms (STFFT) to the audio frames [20]. Mel-scaled filter banks are applied to the transformed frames and finally rendered out as 2D RGB images of frequency vs time to produce a spectrogram [20].

4 Solution Statement

This project will attempt to develop a CNN-based classifier model trained on the Emo-DB corpus. It will either use the LLDs extracted from the audio samples using openSMILE, or it will use 'end-to-end' learning by using CNNs to perform feature extraction on the raw audio data as represented by spectrogram images produced from the audio files. In the case of spectrograms and image classification tasks, CNNs have been shown to be very suitable at SER classification tasks, surpassing results obtained by SVM using acoustic functionals on the same corpus [22, 23, 24]. A CNN using a categorical cross-entropy loss function to compute accuracy in accordance with multi-classification coding conventions for CNNs and other deep learning models. The resulting classifier will be trained, tuned, and evaluated using KFold cross-validation, and a comparative analysis with a benchmark model will be conducted to compare the developed solution against those in the literature.

5 Benchmark Model

Poria et al. (2017) report in their review [4] that to the best of their knowledge, the best *speaker-independent*[2] accuracy achieved on Emo-DB is 81% by Atassi and Esposito (2008)[3] using a mix of acoustic and prosodic vocal features engineered and extracted from the audio signal. Their review however precedes some of the most recent deep learning and CNN-based classifiers reported in the literature for Emo-DB, as well as alternative feature engineering techniques such as Bag-of-Audio-Words (BoAW) [25].

For example, using the BoAW technique for engineering features and using an SVM classifier, the OpenXBOW [26] toolkit achieves roughly 81.4% accuracy on Emo-DB using Sequential Minimal Optimisation (SMO)[4] and Weka[5].

For CNN-based benchmarks using Emo-DB and spectrograms, Badshah et al. (2017) [15], Stoler et al. (2017) [16], and Weibkirchen et al. (2017) [17] establish benchmark performance metrics on EmoDB using CNNs. Stolar et al. report an average accuracy of 79.68% and 76.79% for female and male speakers respectively, comparing their results against other papers utilising spectrograms and Emo-DB [27, 28]. Badshah et al. don't report accuracy or recall values but they do provide confusion matrices showing the precision on a per-class basis. Weibkirchen et al. report Unweighted Average Recall values (explained in the next section) of 71% for the top 1 class probability in a speaker-independent assessment.

[2]Speaker-independence refers to models trained on one set of human speakers and tested on a different set of speakers.

[3]`http://ieeexplore.ieee.org/document/4669768/`

[4]`http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SMO.html`

[5]`https://www.cs.waikato.ac.nz/ml/weka/`

6 Evaluation Metrics

In the case of SER, Weighted Average Recall (WA) and Unweighted Average Recall (UAR) are the typical metrics used in the community for establishing performance [3, 22]. WA is the mean of all the recall scores for each class, where each class' recall score is weighted by the number of instances for that class.

Let rc_i be the recall score where TP is true positives and FN is false negatives for class i:

$$rc_i = \frac{TP}{TP + FN}$$

and let WA be:

$$WA = \frac{(rc_i * |c_i|)...(rc_n * |c_n|))}{|c_i| + ...|c_n|}$$

where n is the number of classes, and $|c_i|$ is the total number of instances for a given class i. WA is effectively the same as the Accuracy metric, expressed as the number of correctly identified samples divided by the total number of samples:

$$\frac{\#\ correct\ predictions}{\#\ total\ samples}$$

The Unweighted Average Recall is used for unbalanced data sets where there isn't an even distribution of classes represented in the data set. It is defined as the mean value over all of the recall scores for each class:

$$\frac{1}{n}\sum_{i=1}^{n} rc_i$$

In keeping with the above metrics used in the literature and to allow for reasonable comparison between the proposed solution and the benchmark model(s), the WA and UAR metrics will be used to measure and evaluate the solution. In addition, a confusion matrix and classification report are both useful to show per-class performance on a multi-class problem.

7 Project Design

At a high level, the general workflow for this project is expected to consist of the following stages:

- Data collection, analysis and pre-processing
- Generating spectrograms and data augmentation
- Scaling and Normalisation
- Data Splitting (Train/Validation/Test)
- Algorithm Training (KFold cross-validation)
- Hyperparameter Tuning
- Evaluation and assessment against test data

The proposed solution involves approaching SER as an image classification problem, solved by training a CNN model on spectrograms generated from raw audio inputs. The convolutional layers of the the CNN are expected to perform the function of feature extraction, while the later fully-connected layers learn the patterns and adjust the network's weights to achieve minimal loss on the error function. The data pre-processing steps will utilise mel filter banks, MFCCs, windowing and framing to produce Mel-spectrogram images for input into the CNN. Similar to [17], a 'data augmentation' process will be undertaken to increase the number of samples in the training data from 535 to approximately 1800

or more by splitting the utterances into smaller spectrograms. For example, short utterances less then say, 2 seconds in length will be left as-is, whereas longer utterances in excess of 2 seconds may be split into multiples of 2 second fragments. The precedent for a 2 second fragment is based on a sampling of the literature [17, 23].

For evaluation purposes, a Leave-One-Speaker-Out (LOSO) speaker-independent KFold cross-validation strategy will be applied. Each training fold will consist of n-1 speakers with the validation fold containing samples from a single speaker, such that no training fold will contain samples of the speaker from the test fold. This will ensure that the model generalises to a speaker-independent test set, as is typical for SER tasks.

7.1 Software and Libraries

The following software libraries and toolkits are expected to be used to complete the project:

- Python 3.5
- Keras
- TensorFlow
- Scikit-Learn
- Jupyter notebooks

References

[1] S. Mariooryad and C. Busso, "Exploring Cross-Modality Affective Reactions for Audiovisual Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 183–196, Apr. 2013.

[2] T. Vogt, E. André, and J. Wagner, "Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation," C. Peter and R. Beale, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 75–91. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-85099-1_7

[3] M. Tahon and L. Devillers, "Towards a Small Set of Robust Acoustic Features for Emotion Recognition: Challenges," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 16–28, Jan. 2016.

[4] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, Sep. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253517300738

[5] A. Hong, Y. Tsuboi, G. Nejat, and B. Benhabib, "Affective Voice Recognition of Older Adults1," *Journal of Medical Devices*, vol. 10, no. 2, pp. 020 931–020 931–2, May 2016. [Online]. Available: http://dx.doi.org/10.1115/1.4033226

[6] G. Acampora, D. J. Cook, P. Rashidi, and A. V. Vasilakos, "A Survey on Ambient Intelligence in Healthcare," *Proceedings of the IEEE*, vol. 101, no. 12, pp. 2470–2494, Dec. 2013.

[7] P. Rashidi and A. Mihailidis, "A Survey on Ambient-Assisted Living Tools for Older Adults," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 3, pp. 579–590, May 2013.

[8] K. Johnson, "Amazon's Alexa wants to learn more about your feelings," Dec. 2017. [Online]. Available: https://venturebeat.com/2017/12/22/amazons-alexa-wants-to-learn-more-about-your-feelings/

[9] H. Pérez-Espinosa, J. Martínez-Miranda, I. Espinosa-Curiel, J. Rodríguez-Jacobo, and H. Avila-George, "Using acoustic paralinguistic information to assess the interaction quality in speech-based systems for elderly users," *International Journal of Human-Computer Studies*, vol. 98, pp. 1–13, Feb. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S107158191630129X

[10] M. Schmitt and B. W. Schuller, "openXBOW - Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *arXiv:1605.06778 [cs]*, May 2016, arXiv: 1605.06778. [Online]. Available: http://arxiv.org/abs/1605.06778

[11] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, Dec. 2008. [Online]. Available: https://link.springer.com/article/10.1007/s10579-008-9076-6

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *in Proceedings of Interspeech, Lissabon*, 2005, pp. 1517–1520.

[13] c. Wikipedia, *Emotion classification — Wikipedia, The Free Encyclopedia*, 2017. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Emotion_classification&oldid=802307984

[14] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[15] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," in *2017 International Conference on Platform Technology and Service (PlatCon)*, Feb. 2017, pp. 1–5.

[16] M. N. Stolar, M. Lech, R. S. Bolia, and M. Skinner, "Real time speech emotion recognition using RGB image classification and transfer learning," in *2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS)*, Dec. 2017, pp. 1–8.

[17] N. Weißkirchen, R. Böck, and A. Wendemuth, "Recognition of Emotional Speech with Convolutional Neural Networks by Means of Spectral Estimates," Oct. 2017.

[18] C.-n. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *The Artificial Intelligence Review; Dordrecht*, vol. 43, no. 2, pp. 155–177, Nov. 2012. [Online]. Available: http://search.proquest.com.proxy.library.dmu.ac.uk/docview/1647316503/abstract/19FE966497E74B1BPQ/1

[19] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language—State-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, Jan. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230812000162

[20] H. Fayek, "Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between," Apr. 2016. [Online]. Available: http://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

[21] Wikipedia, "Spectrogram," Mar. 2018, page Version ID: 830276272. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Spectrogram&oldid=830276272

[22] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech," in *Proc. of the 25th ACM International Conference on Multimedia, MM*, 2017.

[23] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks*, Mar. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S089360801730059X

[24] A. Satt, S. Rozenberg, and R. Hoory, "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms," 2017.

[25] M. Schmitt, F. Ringeval, and B. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech," Sep. 2016, pp. 495–499. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2016/abstracts/1124.html

[26] openXBOW, "openXBOW: the Passau Open-Source Crossmodal Bag-of-Words Toolkit," Jan. 2018, original-date: 2016-05-21T23:32:39Z. [Online]. Available: https://github.com/openXBOW/openXBOW

[27] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech Emotion Recognition Using CNN," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 801–804. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654984

[28] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Dec. 2016, pp. 1–4.