PSEUDO-LABELLING AND TRANSFER LEARNING BASED SPEECH EMOTION RECOGNITION

Siddarth Venkateswaran[1], *Ronald Böck*[2], *Thomas Keßler*[1], *Ossmane Krini*[3]

[1]*Genie Enterprise Inc.*, [2]*Otto von Guericke University*, [3]*DHBW Lörrach*

{venkat,tkessler}@genie-enterprise.com, ronald.boeck@ovgu.de, krini@dhbw-loerrach.de

Abstract: This paper describes speech emotion recognition as an image classification problem using pseudo-labelling techniques, in which there is an availability of only a handful of labelled samples. Low-level acoustic features like log spectrograms and mel spectrograms, extracted from audio files, classified as different emotions were fed as RGB images to Convolutional Neural Networks (CNNs) to train the classification algorithms on. While CNNs have achieved state-of-the-art performances for image classification tasks, they however thrive on a lot of labelled data. This paper applies transfer learning by using a CNN pre-trained on a huge corpus of labelled image data, learning the distinct features on a handful of available labelled spoken data, and utilising this knowledge in iteratively generating machine-confident pseudo-labels for unlabelled acoustic data. Comparisons were made by evaluating a neural network trained using full-supervision, and that using semi-supervision with a combination of labelled and pseudo-labelled data. All the experiments were implemented on nine commonly used benchmark corpora, allowing also comparison to already published results.

1 Introduction

Speech emotion recognition (SER) is of utmost importance for human-computer interaction (HCI). While sufficient research has been conducted for tasks related to automatic speech recognition, the related SER remains still a work in progress. Speech emotions reflect unique characteristics of human speech, conveying additional information besides the content, varying in small time scales (cf. discussion in [1], p. 12 ff.). Tracking this information in real-time could lead to better performances of HCI applications. This paper tackles the issue as an image classification problem, in which acoustic features like log spectrograms and mel spectrograms are extracted from raw audio files, annotated in different emotions. These spectrograms are fed to a fine-tuned AlexNet architecture that has been pre-trained on a huge corpus of image data for object classification tasks.

Deep learning architectures have achieved state-of-the-art performances across tasks like object detection, speech-to-text conversion, semantic segmentation, etc. However, training such networks requires usually huge corpora of labelled data. Annotation, or the process of manually labelling data is a cumbersome process and requires human expertise. However, this comes with limitations in the form of fatigue, concentration lapses, etc. This paper implement semi-supervised techniques, where only a handful of ground-truth labelled data is available, while most of the data is unlabelled (the influence of unlabelled data is discussed in [2]), to overcome the aforementioned drawbacks. During the initial phase, the neural networks are trained only on the ground-truth labelled data, while pseudo-labels are generated for the unlabelled data based on the knowledge gained from training. A confidence threshold is set to classify the newly generated labels as confident pseudo-labels. During further iterations, these networks are

re-trained using a combination of ground-truth and machine-labelled confident pseudo labels, generating updated/newer pseudo-labels at the end of each iteration. This is done to alleviate the manual process of labelling data.

All the experiments in this paper were implemented on nine benchmark datasets that are popular for SER (cf. Sec. 3), being conducted gender and speaker independent based on a five-fold cross-validation strategy. Also, this research compares the performances of the neural networks when all data was labelled (full-supervision) and when only a handful of data were labelled across all classes (semi-supervision).

2 Related Work

SER is a matter of research for roughly two decades, resulting in various investigations and overviews (cf. for an brief overview e.g. [1, 3, 4, 5]). Thus, we kindly focus of those works being highly related to our approach.

A good number of studies on SER involve Hidden Markov Models (HMM) to classify speech emotions based on extracted features like the short time log frequency power coefficients (LFPC), linear prediction Cepstral coefficients (LPCC), and the mel-frequency Cepstral coefficients (MFCC) (cf. [6, 7]). [8] use hybrid Deep Neural Network - HMMs(DNN-HMM) with restricted Boltzmann Machines (RBMs) based on unsupervised pre-training, and DNN-HMMs with discriminative pre-training, comparing their performances with Gaussian mixture model based HMMs (GMM-HMMs). Further, they apply Shallow Neural Network - HMMs with two layers as well as Multi-layer Perceptrons combined with HMMs (MLP-HMMs), proving that the optimal number of hidden layers and hidden-layer units could improve the recognition accuracy of the DNN-HMMs.

Additionally, various studies have SVMs as a classifier for SER (cf. e.g. [9, 10, 11, 12]). [13] use extracted features and performs a comparison with SVM kernels like linear, polynomial, quadratic, and RBF ones. [14] apply features like statistics of pitch, energy, and MFCCs to train an SVM classifier with linear and RBF kernels using a binary tree, one vs one, and one vs rest classification strategies. Here, they implemented their experiments using a combination of different features in a gender dependent and gender independent setup.

Also SER was implemented as an image classification task (cf. e.g. [15, 16, 17]). In [16] transfer learning is used by feeding spectrograms extracted from raw audio files as RGB images to train a fine-tuned AlexNet classifier and a hybrid AlexNet-SVM classifier. [17] use acoustic characteristics based on extracted across mel, log, linear, and ERB scales as RGB images, and perform a scale-wise and channel-wise comparison of classification of emotions.

3 Corpora

In our experiments, we relied on the nine well-known benchmark corpora used in the community of affect recognition from speech. In the following, the nine corpora will be briefly introduced based on the presentation in [18].

The *Airplane Behaviour Corpus (ABC)* (cf. [19]) is intended for applications and investigations related to public transport surveillance. A collection of five moods, namely *aggressive, cheerful, intoxicated, nervous, neutral,* and *tired*, were induced using predefined scripts. These guide the subjects through a storyline. Eight speakers, being balanced in sex, took part in the data collection. The corpus contains overall 431 samples.

The *Audiovisual Interest Corpus* (AVIC) (cf. [20]) is focused to *interest* samples being uttered spontaneously. This results from the scenario setting being related to commercials. In particular, a product presenter (experimentor) leads each of the 21 subjects (ten female) through

an English commercial presentation. The level of interest is annotated for every sub-speaker turn on a three point scale (loi1 to loi3).

The *Danish Emotional Speech (DES)* (cf. [21]) data set contains speech samples of acted emotions, namely *anger, happiness, neutral, sadness,* and *surprise*. All recording were conducted in Danish, comprising full sentences, words, and chunks. These are expressed by four professional actors (two females). After data collection, each utterance was judged according to the aforementioned emotion categories.

The *Berlin Emotional Speech Database (emoDB)* (cf. [22]) is a studio recorded corpus applying an anechoic chamber to provide clear speech in high quality. The ten professional actors, being balanced in sex, utter ten German sentences with emotionally neutral content. emoDB provides speech samples in seven emotional categories, namely *anger, boredom, disgust, fear, joy, neutral,* and *sadness.* In total, 494 phrases were selected based on experts' decision in a perception test (cf. [22] for details).

The *eNTERFACE* (cf. [23]) corpus comprises samples from 42 subjects (eight female) from 14 nations providing speech recordings of pre-defined spoken content in English (this includes dialects and accents). The data collection is conducted in an office environment. The 1 277 emotional instances represent six induced emotions, namely *anger, disgust, fear, joy, sadness,* and *surprise*. Although the emotional material is acted and in categories similar to emoDB, the quality of affective content spans a much broader variety than in emoDB.

The *Belfast Sensitive Artificial Listener (SAL)* (cf. [24]) corpus provides longer samples of 25 audio-visual recordings from four speakers (two female) who interact with a virtual agent. The recording scenario is designed to let the participants work through a continuous space of emotional states in a natural interaction. In our experiments, a clustering of the continuous space was used as provided by [3]. This results in a mapping of the original arousal-valence space into four quadrants (q1 to q4).

The *SmartKom* (cf. [25]) is multi-modal corpus providing spontaneous speech. The recorded utterances are labelled in seven broader emotion categories, namely *neutral, joy, anger, helplessness, pondering, surprise,* and *unidentifiable* in German and English. The scenario is related to in interaction with a technical assistant realised in a Wizard-of-Oz setting. In our experiments, we used the German part containing 3 823 samples.

The *Speech Under Simulated and Actual Stress* (SUSAS) (cf. [26]) data set comprises both, spontaneous and acted emotional samples. The recordings are afflicted partly with field noise. In our experiments, We selected the subset providing 3 593 actual stress speech samples which present four different situations, namely *neutral, medium stress, high stress,* and *screaming*. In total seven participants (three female) are recorded in roller coaster and free fall stress situations.

The *Vera-Am-Mittag* (VAM) (cf. [27]) corpus provides a collection of audio-visual recordings being selected from a unscripted German TV talk show called "Vera am Mittag". The investigated subset includes 946 utterances from 47 participants. The affective labelling on utterance level is based on a discrete five point scale for each of the three dimensions *arousal, valence,* and *dominance*. For our experiment, we again applied a clustering related to SAL using four quadrants q1 to q4 (cf. [3]).

4 Experimental Setup

This section explains the data pre-processing steps followed by the model and training settings. The parameters had to be adjusted based on the sampling rate for each dataset.

4.1 Data Preprocessing

The raw data, available in WAV format, was passed through a voice activity detector, selecting only the voiced segments. Based on the implementations in [16], a Hamming window with a frame size of 25ms, hop size of 12.5ms, and 64 mels per frame were applied to these segments. Further, these were chunked into blocks of one second each, with an overlap of 10ms between two consecutive blocks. To further increase the availability of training material, spectrogram images were extracted for 3 different levels of maximum frequency : 7 000 Hz, 7 500 Hz, and 8 000 Hz, which was in line with the implementations in [28]. The corpora-wise break-up of the data available for training and validation are shown in Table 1.

Table 1 – No. of Mel/Log spectrogram images generated per corpus (eNT ... eNTERFACE).

Data	ABC	AVIC	DES	emoDB	eNT	SAL	SmartKom	SUSAS	VAM
No. of Images	6 612	19 440	5 004	3 519	6 099	12 177	15 402	6 285	5 961

4.2 Model Architecture and Training

All the experiments were implemented using an AlexNet architecture that was pre-trained on a huge corpus of ImageNet data (cf. [29]). The chosen architecture comprised of five convolutional blocks, followed by two fully-connected layers. All the experiments were conducted with images of dimension $224 \times 224 \times 3$. The penultimate fully-connected layer was followed by a dropout layer. The number of neurons in the final fully-connected layer varied based on the number of classes available per dataset (cf. Sec. 3), which was followed by a softmax layer to predict the probability of each class identified.

Considering the imbalance across labels in all the datasets, the models were trained to reduce the weighted categorical cross-entropy loss between each sample and the corresponding labels, in which data with lesser availability were assigned higher weights and vice versa. The networks were optimized using the Adam-based stochastic gradient descent with an initial learning rate of $1e^{-2}$. This led to a reduction by a factor of 0.1 upon it reaching a plateau in the validation accuracy for a patience of 5 epochs. The cutoff was reached with the lowest learning rate of $1e^{-5}$. The network training was stopped using an early-stopping mechanism by virtue of it, reaching a plateau in the validation accuracy with a patience of 10 epochs.

4.3 Training Strategies

Two different training strategies were implemented for all corpora:

1. **Full Supervision**: All the labelled data available was used tor training the neural network. Using the five-fold cross-validation strategy, five different combinations of training, validation, and test sets were created for each dataset, for each extracted feature.

2. **Semi Supervision**: Only 20% and 30% of all classes across all training data were considered labelled, while the rest were considered unlabelled. For each dataset, five different combinations of labelled, unlabelled, and test sets were created.

The test sets created for semi-supervision contained the same samples across each fold for semi-supervision and full-supervision, allowing a direct comparison of the performances.

4.4 Pseudo-Labelling Process

To initiate the pseudo-labelling process, the neural networks were initially trained on the handful of data which were classified as labelled. Pseudo-labels were predicted by the trained model for those samples marked as unlabelled. A confidence threshold of 99% was set to classify the newly predicted labels as confident pseudo-labels. This material was used to re-train the models from the second iteration, in combination with the available ground-truth labels.
At the end of each iteration, the the pseudo-labels were re-calculated for all the unlabelled data, while the model performances were evaluated on the hold-out test set. This process was repeated for a total of three iterations, since beyond this, it was observed during the experiments that there was a plateau in the overall performance of the models on the test set. Model performances were compared at the end of the first and the third iteration to check the impact of additional pseudo-labels for classifying speech emotion.

5 Results

The implementations were, inspired by [16] which use an AlexNet architecture, on spectrogram images extracted in the log scale (cf. Sec. 4.1) for voiced as well as non-voiced (i.e. those containing silence/pauses; these were removed in later experiments since silent segments could deteriorate the classification performances) segments. Finally, we checked the impact of retaining only voiced segments in the performance of neural networks in classifying speech emotion.

Table 2 compares the performances of the models trained under full-supervision, while Table 3 compares those trained under semi-supervision with only 30% of each class considered as labelled during the training process. Our experimental results were compared against the baseline, applying an AlexNet architecture that was pre-trained on the ImageNet dataset and tested directly on the corpora material, respectively, using voiced and non-voiced segments extracted as log scale spectrogram images.

Table 2 – Unweighted F1-Scores using full supervision, comparing the baseline on log scale to our results (eNT ... eNTERFACE).

Data	ABC	AVIC	DES	emoDB	eNT	SAL	SmartKom	SUSAS	VAM
baseline	0.91	0.89	0.93	0.96	0.81	0.78	0.90	0.96	0.86
log scale	0.92	0.94	0.94	0.94	0.89	0.92	0.90	0.97	0.91
mel scale	0.92	0.91	0.92	0.92	0.87	0.89	0.87	0.96	0.45

Table 3 – Unweighted F1-Scores at the end of the first and the third iteration using semi supervision, with only 30% of each class labelled, comparing the baseline on log scale to our results

Dataset	baseline		log		mel	
	Iteration 1	Iteration 3	Iteration 1	Iteration 3	Iteration 1	Iteration 3
ABC	0.62	0.63	0.66	**0.69**	0.69	0.67
AVIC	0.67	0.72	0.74	**0.77**	0.69	0.71
DES	0.67	**0.72**	0.63	0.66	0.64	0.67
EMODB	0.77	**0.80**	0.70	0.72	0.66	0.72
eNTERFACE	0.56	0.60	0.56	**0.62**	0.53	0.59
SAL	0.42	0.61	0.57	**0.62**	0.55	0.61
SMARTKOM	0.52	**0.58**	0.51	0.57	0.49	0.53
SUSAS	0.81	0.85	0.80	**0.86**	0.83	0.83
VAM	0.55	0.61	0.58	**0.63**	0.37	0.60

From Table 2, models trained under full-supervision under the log scale and with the retention of only voiced segments tended to outperform the baseline and the mel scale (introducing a specific auditory model, influencing the extracted images) implementation across all corpora.

For those models trained under semi-supervision (cf. Table 3), our approach could not beat the baseline for DES and emoDB after the third iteration. Again, using the log scale on voiced segments only retained higher performance. Further, in most cases adding machine generated pseudo-labels improved slightly the overall performance from the first to the third iteration.

Also, models trained under semi-supervision performed best on the SUSAS dataset, closely followed by AVIC and emoDB (cf. Table 3). Table 4 shows the best results attained on datasets when the neural networks were trained with only availability of 20% of labelled samples per class. Investigating the quality of pseudo labels generated, it was found that the number of accurate pseudo-labels generated for these three datasets at the end of the second iteration was higher as compared to the remaining datasets (2613 out of 2946 accurate pseudo-labels for SUSAS as compared to 1889 out of 2439 for VAM when trained using 30% labelled data with spectrogram images extracted in the log scale). While the presence of more labelled data increased in the overall classification performance of the neural networks (cf. Tables 2 and 3), the quality of labels also plays a significant role on classification correctness (cf. Tables 3 and 4).

Table 4 – Comparison of the unweighted F1-Scores at the end of the first and the third iteration using semi supervision for the AVIC, emoDB, and the SUSAS datasets with only 20% of each class labelled.

Dataset	baseline		log		mel	
	Iteration 1	Iteration 3	Iteration 1	Iteration 3	Iteration 1	Iteration 3
AVIC	0.60	0.58	0.68	**0.69**	0.64	0.68
emoDB	0.70	**0.73**	0.61	0.67	0.59	0.64
SUSAS	0.79	0.81	0.82	**0.83**	0.76	0.78

Finally, the removal of non-voiced segments improved in the classification performances of the neural networks, when trained under full-supervision as well as with semi-supervision (cf. Tables 2 to 4).

6 Conclusion and Outlook

This research considered the concept of SER as an image classification task with the help of pseudo-labels. We observed in all the experiments that the retention of only the voiced segments improved the overall F1-Scores as compared to the baseline. While there were improvements in the F1-Scores on models trained with pseudo-labels, as compared to those trained with only a handful of ground-truth labels, the generation of pseudo-labels with high confidence (cf. Sec. 4.4) could lead to an improved performance for SER using semi-supervised techniques. The current experiments were conducted in a gender-independent fashion, so future studies will involve pseudo-labelling of affects gender-specifically, assuming a better classification performances of the neural networks.

References

[1] BÖCK, R.: *Anticipate the User: Multimodal Analyses in Human-Machine Interaction towards Group Interactions*. TUDpress, Dresden, Germany, 2020.

[2] SCHELS, M., M. KÄCHELE, M. GLODEK, D. HRABAL, S. WALTER, and F. SCHWENKER: *Using unlabeled data to improve classification of emotional states in human computer interaction. Jour. on Multimodal User Interfaces*, 8(1), pp. 5–16, 2014.

[3] SCHULLER, B., B. VLASENKO, F. EYBEN, G. RIGOLL, and A. WENDEMUTH: *Acoustic emotion recognition: A benchmark comparison of performances*. In *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2009*, pp. 552–557. Merano, Italy, 2009.

[4] SIKKA, K., K. DYKSTRA, S. SATHYANARAYANA, G. LITTLEWORT, and M. BARTLETT: *Multiple kernel learning for emotion recognition in the wild*. In *Proc. of the 15th ICMI*, pp. 517–524. ACM, Sydney, Australia, 2013.

[5] SCHULLER, B.: *Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends*. *Communications of the ACM*, 61(5), pp. 90–99, 2018.

[6] NWE, T. L., S. W. FOO, and L. C. DE SILVA: *Speech emotion recognition using hidden markov models*. *Speech communication*, 41(4), pp. 603–623, 2003.

[7] MAO, S., D. TAO, G. ZHANG, P. C. CHING, and T. LEE: *Revisiting hidden markov models for speech emotion recognition*. In *ICASSP 2019 - 2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6715–6719. 2019.

[8] LI, L., Y. ZHAO, D. JIANG, Y. ZHANG, F. WANG, I. GONZALEZ, E. VALENTIN, and H. SAHLI: *Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition*. In *2013 Humaine Association Conf. on Affective Computing and Intelligent Interaction*, pp. 312–317. 2013.

[9] PAN, Y., P. SHEN, and L. SHEN: *Speech emotion recognition using support vector machine*. *Int. Journ. of Smart Home*, 6(2), pp. 101–108, 2012.

[10] SHEN, P., Z. CHANGJUN, and X. CHEN: *Automatic speech emotion recognition using support vector machine*. In *Proc. of 2011 Int. Conf. on Electronic & Mechanical Engineering and Information Technology*, vol. 2, pp. 621–625. IEEE, 2011.

[11] CHAVHAN, Y., M. DHORE, and P. YESAWARE: *Speech emotion recognition using support vector machine*. *Int. Journ. of Computer Applications*, 1(20), pp. 6–9, 2010.

[12] SCHULLER, B., G. RIGOLL, and M. LANG: *Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture*. In *2004 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–577. IEEE, 2004.

[13] DAHAKE, P. P., K. SHAW, and P. MALATHI: *Speaker dependent speech emotion recognition using mfcc and support vector machine*. In *2016 Int. Conf. on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, pp. 1080–1084. 2016.

[14] SINITH, M. S., E. ASWATHI, T. M. DEEPA, C. P. SHAMEEMA, and S. RAJAN: *Emotion recognition from audio signals using support vector machine*. In *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pp. 139–144. 2015.

[15] BADSHAH, A. M., J. AHMAD, N. RAHIM, and S. W. BAIK: *Speech emotion recognition from spectrograms with deep convolutional neural network*. In *2017 Int. Conf. on Platform Technology and Service (PlatCon)*, pp. 1–5. 2017.

[16] STOLAR, M. N., M. LECH, R. S. BOLIA, and M. SKINNER: *Real time speech emotion recognition using rgb image classification and transfer learning*. In *2017 11th Int. Conf. on Signal Processing and Communication Systems (ICSPCS)*, pp. 1–8. 2017.

[17] STOLA, M., M. LECH, R. S. BOLIA, and M. SKINNER: *Acoustic characteristics of emotional speech using spectrogram image classification*. In *2018 12th Int. Conf. on Signal Processing and Communication Systems*, pp. 1–5. IEEE, 2018.

[18] BÖCK, R., O. EGOROW, I. SIEGERT, and A. WENDEMUTH: *Comparative Study on Normalisation in Emotion Recognition from Speech*, pp. 189–201. No. 10688 in Lecture Notes of Computer Sciences. Springer, Cham, 2017. Best paper award.

[19] SCHULLER, B., D. ARSIC, G. RIGOLL, M. WIMMER, and B. RADIG: *Audiovisual behavior modeling by combined feature spaces*. In *Proc. of the ICASSP-2007*, pp. 733–736. IEEE, Honolulu, USA, 2007.

[20] SCHULLER, B., R. MÜLLER, B. HÖRNLER, A. HÖTHKER, H. KONOSU, and G. RIGOLL: *Audiovisual recognition of spontaneous interest within conversations*. In *Proc. of the 9th ICMI*, pp. 30–37. ACM, Nagoya, Japan, 2007.

[21] ENGBERT, I. S. and A. V. HANSEN: *Documentation of the danish emotional speech database des*. Tech. Rep., Center for PersonKommunikation, Aalborg University, Denmark, 2007.

[22] BURKHARDT, F., A. PAESCHKE, M. ROLFES, W. SENDLMEIER, and B. WEISS: *A database of german emotional speech*. In *INTERSPEECH-2005*, pp. 1517–1520. Lisbon, Portugal, 2005.

[23] MARTIN, O., I. KOTSIA, B. MACQ, and I. PITAS: *The eNTERFACE'05 audio-visual emotion database*. In *Proc. of the Workshop on Multimedia Database Management*. IEEE, Atlanta, USA, 2006. s.p.

[24] DOUGLAS-COWIE, E., R. COWIE, C. COX, N. AMIER, and D. HEYLEN: *The sensitive artificial listner: an induction technique for generating emotionally coloured conversation*. In *LREC Workshop on Corpora for Research on Emotion and Affect*, pp. 1–4. ELRA, Paris, France, 2008.

[25] STEININGER, S., F. SCHIEL, O. DIOUBINA, and S. RAUBOLD: *Development of user-state conventions for the multimodal corpus in smartkom*. In *Proc. of the Workshop on Multimodal Resources and Multimodal Systems Evaluation*, pp. 33–37. ELRA, Las Palmas, Spain, 2002.

[26] HANSEN, J. and S. BOU-GHAZALE: *Getting started with SUSAS: A speech under simulated and actual stress database*. In *Proc. of EUROSPEECH-1997*, vol. 4, pp. 1743–1746. ISCA, Rhodes, Greece, 1997.

[27] GRIMM, M., K. KROSCHEL, and S. NARAYANAN: *The Vera am Mittag German Audio-Visual Emotional Speech Database*. In *Proc. of ICME 2008*, pp. 865–868. IEEE, Hannover, Germany, 2008.

[28] WEISSKIRCHEN, N., R. BÖCK, and A. WENDEMUTH: *Recognition of emotional speech with convolutional neural networks by means of spectral estimates*. In *2017 Seventh Int. Conf. on Affective Computing and Intelligent Interaction Workshops and Demos*, pp. 50–55. IEEE, 2017.

[29] DENG, J., W. DONG, R. SOCHER, L.-J. LI, K. LI, and L. FEI-FEI: *Imagenet: A large-scale hierarchical image database*. In *2009 IEEE Conf. on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.