IMPROVING SPEECH EMOTION RECOGNITION WITH UNSUPERVISED REPRESENTATION LEARNING ON UNLABELED SPEECH

Michael Neumann, Ngoc Thang Vu

University of Stuttgart, Germany
{michael.neumann|thang.vu}@ims.uni-stuttgart.de

ABSTRACT

In this paper we present our findings on how representation learning on large unlabeled speech corpora can be beneficially utilized for speech emotion recognition (SER). Prior work on representation learning for SER mostly focused on the relatively small emotional speech datasets without making use of additional unlabeled speech data. We show that integrating representations learnt by an unsupervised autoencoder into a CNN-based emotion classifier improves the recognition accuracy. To gain insights about what those models learn, we analyze visualizations of the different representations using t-distributed neighbor embeddings (t-SNE). We evaluate our approach on IEMOCAP and MSP-IMPROV by means of within- and cross-corpus testing.

Index Terms— representation learning, speech emotion recognition, unsupervised learning, visualization, CNN

1. INTRODUCTION

Two major challenges in the field of speech emotion recognition (SER) are data scarcity and finding an optimally discriminative speech representation for the task, as Schuller described it in [1]: "As the quest for the optimal features has dominated the field similarly as the ever-lacking large and naturalistic databases [...]". Regarding representation learning (RL), there exist different approaches, mostly using variants of autoencoders (AE) to learn suitable features from the data in an unsupervised manner, as in [2, 3, 4, 5, 6]. In [5], variational AEs are trained and the learnt representation is then used as input to a long short-term memory (LSTM) network for emotion recognition. A similar approach is presented in [2] which closely relates to our work. The authors compared different kinds of AEs and input features. In contrast to the present work, these studies have not used any additional unlabeled speech resources. Potential ways to incorporate additional data have been presented recently in [7, 8]. In [7], four different types of AEs were trained on Librispeech data and the encoders were used to generate representations for labeled emotional speech to feed into a convolutional neural network (CNN) for SER. While this study used only the AE representations as input, Lakomkin et al. [8] also experimented with a combination of emotion-specific and ASR-specific representations in a progressive neural network.

In this work we explore one direction how unsupervised representation learning on large unlabeled speech corpora can be utilized to enhance the performance of an SER system. We train a recurrent sequence-to-sequence AE on unlabeled data and use it to generate representations for the labeled target data. These representations are incorporated in the training procedure of an attentive CNN as additional source of information for emotion classification. We show that adding this additional feature representation improves the accuracy over the baseline system for within- and cross-corpus evaluation. For 4-class emotion recognition on the IEMOCAP corpus, we report state-of-the-art comparable results. In addition, we present and analyze visualizations of the different representations (ACNN and AE) in order to gain a better understanding of what is learned by the models.

2. METHODS

For the task of speech emotion recognition, we use an attentive convolutional neural network (ACNN) with multi-view learning, proposed in [9]. As input features, 26 logMel filterbanks in the range 0 to 6.5kHz are extracted for 25ms long frames with a 10ms shift. We use the openSMILE toolkit [10] for feature extraction. The network consists of one convolutional and one max-pooling layer, followed by an attention layer, which computes weights α_i over all feature maps for each time step i. This is shown in Equation 1, where x_i is a vector of an input matrix x and $f(x) = W^T x$, with W being a trainable parameter. The output of the attention layer is a weighted sum of the input sequence.

$$\alpha_i = \frac{exp(f(x_i))}{\sum_j exp(f(x_j))} \tag{1}$$

For learning a compact latent representation from unlabeled speech as additional information source we train a time-recurrent sequence-to-sequence autoencoder on spectrograms. We utilize the auDeep toolkit [11, 12] for spectrogram extraction, autoencoder training and for generating representations with the learnt model. For spectrogram extraction

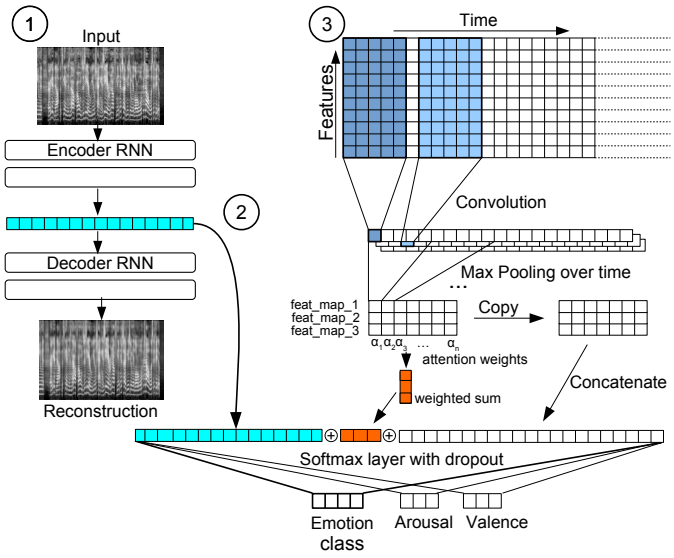


Fig. 1: Overview of the model architecture. The training procedure follows these consecutive steps: (1) autoencoder training on a large speech corpus, (2) generation of latent representations for the emotional speech samples, (3) ACNN training with those representations as additional feature vector.

we take 80ms long frames with a window overlap of 40ms and extract 128 Mel frequency bands. Figure 1 presents an overview of the architecture and shows how the representation generated by the encoder network is integrated into the CNN training. Note that the two networks are trained consecutively, as depicted in Figure 1. The encoder representations generated in step (2) are not changed by the CNN. We apply dropout for regularization on the whole concatenated feature vector, before the final softmax classification. The multi-view objective function presented in Equation (2) combines three cross-entropy losses, one for emotion classes ($\mathcal{L}_{emo}$), and one for arousal ($\mathcal{L}_{aro}$) and valence ($\mathcal{L}_{val}$) scores each. The parameters α and β control the weight of those losses. For more details about the attention mechanism and the multi-view learning procedure, the reader is referred to [9].

$$\mathcal{L} = (1 - \alpha - \beta)\mathcal{L}_{emo} + \alpha\mathcal{L}_{aro} + \beta\mathcal{L}_{val} \qquad (2)$$

3. SPEECH CORPORA

For our experiments we use two datasets annotated for emotion recognition: **IEMOCAP** [13] and **MSP-IMPROV** [14]. Both corpora have been created and annotated in a similar way. They consist of English dyadic interactions between actors and are labeled with categorical emotion classes as well as arousal and valence scores on a 5-point scale. Both corpora contain audio and video data, but we use only audio for this study.

We use samples from the four classes angry, happy, sad, and neutral (as it was done frequently in other studies [4, 15, 16, 17, 18]). Note, that for IEMOCAP we merged samples from the classes excitement and happy to form one class happy. The dataset contains 5,531 utterances (1,103 angry, 1,636 happy, 1,708 neutral, 1,084 sad) grouped into 5 sessions (one female and one male speaker per session). MSP-IMPROV consists of 6 sessions in the same manner (12 speakers) and contains 7,798 utterances (792 angry, 2,644 happy, 3,477 neutral, 885 sad). Since the input length for a CNN has to be fixed for all samples, we set the maximal length to 7.5s. Longer turns are cut at 7.5s and shorter ones are padded with zeros. Arousal and valence labels are grouped into three classes each for multi-view learning. The same range mapping as in [9, 19] is used: low: [1,2]; medium: (2,4); high: [4,5].

As additional unlabeled data for AE training we use two well-known corpora from the field of automatic speech recognition (ASR): Tedlium (release 2) [20] and Librispeech [21]. Tedlium 2 is a collection of 1,495 Ted talks comprising 207 hours of transcribed English speech. We segmented the talks according to the timing information in the transcripts, resulting in 92,973 segments. We have trained two models, one with the full dataset and one with a smaller subset consisting of 400 talks, respectively 25,303 segments. Librispeech contains 1000 hours of read English speech from audiobooks. Due to computational limitations, we use a subset of 100 hours, respectively 28,539 utterances.

4. EXPERIMENTAL RESULTS

4.1. Setup

The baseline for this study is the ACNN model without any additional representation data (right-hand side of Figure 1). We conduct 5-fold cross validation on IEMOCAP, taking samples from 8 speakers as train and development sets and the ones from the remaining 2 speakers as resprective test set.

For generating additional feature representations, we train autoencoders on four datasets with the following motivations.

- The main research question is whether additional unlabeled data can be utilized to improve the accuracy of SER. For that purpose, we train an AE on the full Tedlium 2 corpus as the main experiment.
- As control condition, we train an AE only on IEMOCAP itself (respectively MSP-IMPROV for cross-corpus evaluation). In doing so, we can verify the effect of additional data compared to just using an AE representation of the test corpus itself.
- To investigate a potential effect of the amount of additional data, we train a model on a small subset of Tedlium.

- To confirm our findings, we use another kind of additional data in form of a subset of Librispeech.

Another research question we investigate is the effect of our approach on cross-corpus evaluation. For that, we use IEMOCAP as training set and MSP-IMPROV as test set (in the same four conditions as described above).

We report mean, maximum and minimum results from ten runs of each experiment with different random seeds in order to observe variations. All results represent unweighted average recall (UAR) which is the considered a suitable measure for unbalanced datasets.

4.2. Hyper-parameters

The encoder and decoder of the AE consist both of 2 layers with 256 gated recurrent units (GRU) each. After testing several combinations of uni- and bidirectional encoders and decoders with regard to the reconstruction loss, we found that using a unidirectional encoder and a bidirectional decoder is a good choice. We apply a learning rate of 0.001 and a dropout rate of 0.2 and train the models for 64 epochs (respectively 32 epochs for the full Tedlium model).

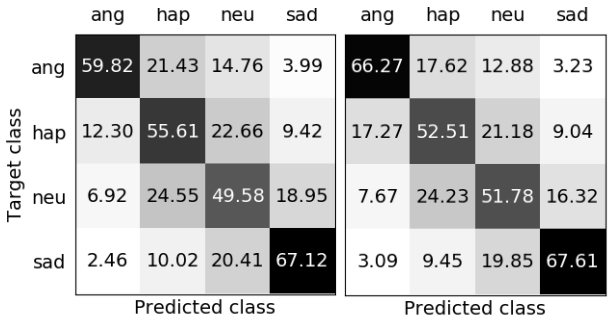
Our ACNN model is implemented with tensorflow [22]. We use stochastic gradient descent with an adaptive learning rate (Adam [23]). The model's hyper-parameters are: 200 convolutional filters of size 26x10 (spanning all 26 log-Mel filter-banks), convolutional stride of 3, pooling size of 30 (1-dimensional because convolution outputs a vector), and a Glorot uniform initialization [24] of kernel weights. The model is trained for 100 epochs and we apply a dropout rate of 0.8 for IEMOCAP and 0.7 for MSP-IMPROV to the last layer. We found that this is necessary to prevent overfitting because the datasets are relatively small. For multi-view learning, we control the influence of arousal/valence predictions on the total loss function with a weight of 0.2 for each.

4.3. Results

Table 1 presents the results of all conditions described in section 4.1. The left-hand side of the table shows the performance on IEMOCAP (5-fold cross validation) and the right-hand side the results of cross-corpus evaluation (trained on IEMOCAP and tested on MSP-IMPROV).

In both cases we observe consistent improvements over the baseline when adding the represenations generated by the different AE models. The results for the control condition are similar to (IEMOCAP) or even below (MSP-IMPROV) the baseline. This indicates that it is in fact the additional speech data which helps improving the performance. It can also be seen that adding more data increases the performance further, as the best results are achieved with the full Tedlium corpus.

For IEMOCAP, we achieve a mean UAR of 59.54% which is comparable with state-of-the-art results on this 4-class subset. To the best of our knowledge, the best results in the literature are 60.9% [17] and 62.5% [25]. However, strict comparison remains difficult because there are no standardized train and test splits and many factors affect the result. Those two studies match our conditions almost completely (emotion classes, UAR as measure, merging happy and excited) with the exception that they used leave-one-speaker-out cross validation as opposed to leave-one-session-out.



(a) ACNN baseline. (b) ACNN+AE (full Tedlium).

Fig. 2: Confusion matrix for mean results on IEMOCAP.

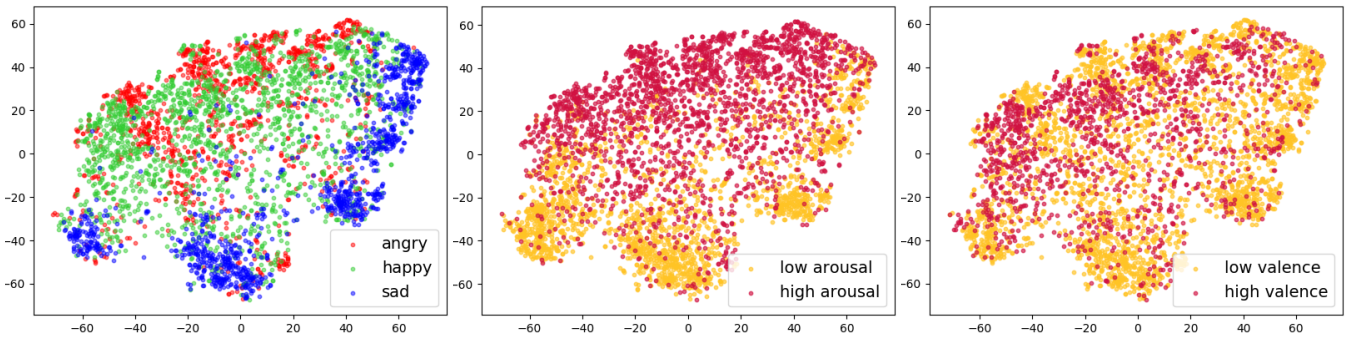
	IEMOCAP			MSP-IMPROV (cross-corpus evaluation)		
	μ	min	max	μ	min	max
BL	58.03	56.78	59.12	42.99	42.19	44.14
C	58.07	56.56	59.68	42.37	41.20	43.37
sT	58.85	57.01	60.10	45.21	43.74	46.78
Li	59.05	57.89	60.18	44.82	42.98	46.33
fT	**59.54**	58.16	60.15	**45.76**	45.02	46.69

Table 1: Results measured in UAR. BL - Baseline without additional represenation, C - control condition (AE trained on IEMOCAP/MSP-IMPROV), sT - small Tedlium subset, Li - Librispeech, fT - full Tedlium.

To gain more insights about the results, we analyze error distributions in the confusion matrix in Figure 2, showing the mean results across all ten runs on IEMOCAP. We see that the ACNN+AE model has a higher accuracy for sad (slight difference), neutral, and angry. However, for happy the accuracy drops below the baseline. The percentages of happy-angry confusions are more balanced when adding the AE representations, which indicates that the baseline model has a stronger bias for happy which is counterbalanced to a certain extent in the ACNN+AE model.

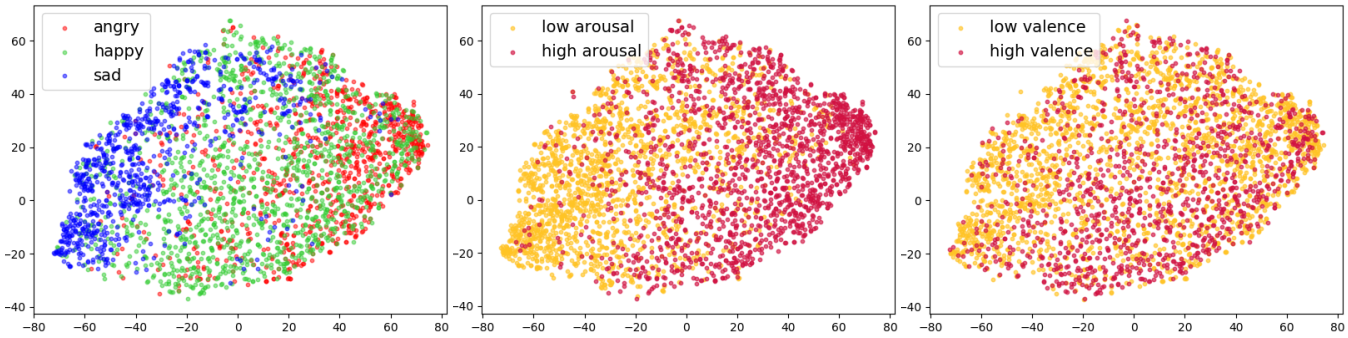
5. VISUALIZATION OF SPEECH REPRESENTATIONS

In this section, we present visualizations of both the learnt ACNN representation (last layer before softmax) and the represenation from the AE trained on Tedlium. We want to inves-



(a) Class labels. (b) Arousal scores. (c) Valence scores.

Fig. 3: t-SNE visualizations of the last hidden layer of the ACNN for IEMOCAP.



(a) Class labels. (b) Arousal scores. (c) Valence scores.

Fig. 4: t-SNE visualizations of the AE representations for IEMOCAP (AE trained on full Tedlium).

tigate what the two models learn with regard to different aspects, such as emotion class, speaker identity and gender, and arousal/valence scores. Figures 3 and 4 show 2D projections generated with t-distributed stochastic neighbor embeddings (t-SNE) [26]. In Figures 3a and 4a we excluded the class neutral for visual clarity. Neutral samples are in both cases distributed across the whole plot and do not form a well-defined cluster. This finding has also been reported in [2].

It can be seen that the ACNN is capable of separating sad from angry to a certain extent. The class happy, however, forms a high-variance cluster which largely intersects with angry, which explains the high confusion rates seen in Figure 2. The plots for arousal and valence show that the model is much more discriminative for arousal than for valence. This obervation is in line with related work on SER [2, 27, 28, 29, 30].

Interestingly, the visualizations of the AE representation in Figure 4 show similar patterns (despite no emotion labels are involved). This indicates that the autoencoder implicitly learns to separate low and high arousal, and therefore angry and sad samples can be distinguished surprisingly well (which explains the boost for angry in the results in Figure 2).

Regarding speaker gender and identity we found that both representations are invariant to these factors, i.e. no separable clusters can be found in the 2D projections. These plots are not included because of space limitations.

6. CONCLUSIONS

In this paper, we have shown that incorporating representations generated by an autoencoder that was trained on a large dataset, leads to consistent improvements in recognition accuracy of the presented SER model. Further, we presented t-SNE visualizations that reveal the discriminative strength of those representations with regard to low and high arousal. Future work includes experimentation with different variants of autoencoders and investigation in generative adversarial networks for representation learning.

7. REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, 2018.

[2] S. Ghosh, E. Laksana, et al., "Learning representations of affect from speech," *International Conference on Learning Representations (ICLR)*, 2016.

[3] S. Sahu, R. Gupta, et al., "Adversarial auto-encoders for speech based emotion recognition," *Proc. of Interspeech*, 2017.

[4] S. Ghosh, E. Laksana, et al., "Representation learning for speech emotion recognition," *Proc. of Interspeech*, 2016.

[5] S. Latif, R. Rana, et al., "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," in *Proc. of Interspeech*, 2018.

[6] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Proc. of Interspeech*, 2018.

[7] S. E. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[8] E. Lakomkin, C. Weber, et al., "Reusing neural speech representations for auditory emotion recognition," in *Proc. of the Eighth International Joint Conference on Natural Language Processing*, 2017.

[9] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *Proc. of Interspeech*, 2017.

[10] F. Eyben, F. Weninger, et al., "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of the 21st ACM international conference on Multimedia*. ACM, 2013.

[11] M. Freitag, S. Amiriparian, et al., "audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, 2017.

[12] S. Amiriparian, M. Freitag, et al., "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proc. of the DCASE 2017 Workshop*, 2017.

[13] C. Busso, M. Bulut, et al., "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, 2008.

[14] C. Busso, S. Parthasarathy, et al., "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, 2017.

[15] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on Affective Computing*, 2015.

[16] Q. Jin, C. Li, et al., "Speech emotion recognition with acoustic and lexical features," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[17] V. Rozgic, S. Ananthakrishnan, et al., "Ensemble of svm trees for multimodal emotion recognition," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2012.

[18] J. Gideon, S. Khorram, et al., "Progressive neural networks for transfer learning in emotion recognition," in *Proc. of Interspeech*, 2017.

[19] A. Metallinou, M. Wollmer, et al., "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, 2012.

[20] A. Rousseau, P. Deléglise, and Y. Esteve, "Enhancing the ted-lium corpus with selected data for language modeling and more ted talks.," in *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014.

[21] V. Panayotov, G. Chen, et al., "Librispeech: an asr corpus based on public domain audio books," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[22] M. Abadi, P. Barham, et al., "Tensorflow: a system for large-scale machine learning.," in *OSDI*, 2016, vol. 16.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference for Learning Representations (ICLR)*, 2015.

[24] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. of the thirteenth international conference on artificial intelligence and statistics*, 2010.

[25] R. Xia and Y. Liu, "Leveraging valence and activation information via multi-task learning for categorical emotion recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[26] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, 2008.

[27] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[28] B. Schuller, B. Vlasenko, et al., "Acoustic emotion recognition: A benchmark comparison of performances," in *Automatic Speech Recognition and Understanding (ASRU), Workshop on*. IEEE, 2009.

[29] G. Trigeorgis, F. Ringeval, et al., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*, 2016.

[30] F. Eyben, K. R. Scherer, et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, 2016.