University of Stuttgart
Germany

Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech

Michael Neumann, Thang Vu

Summary

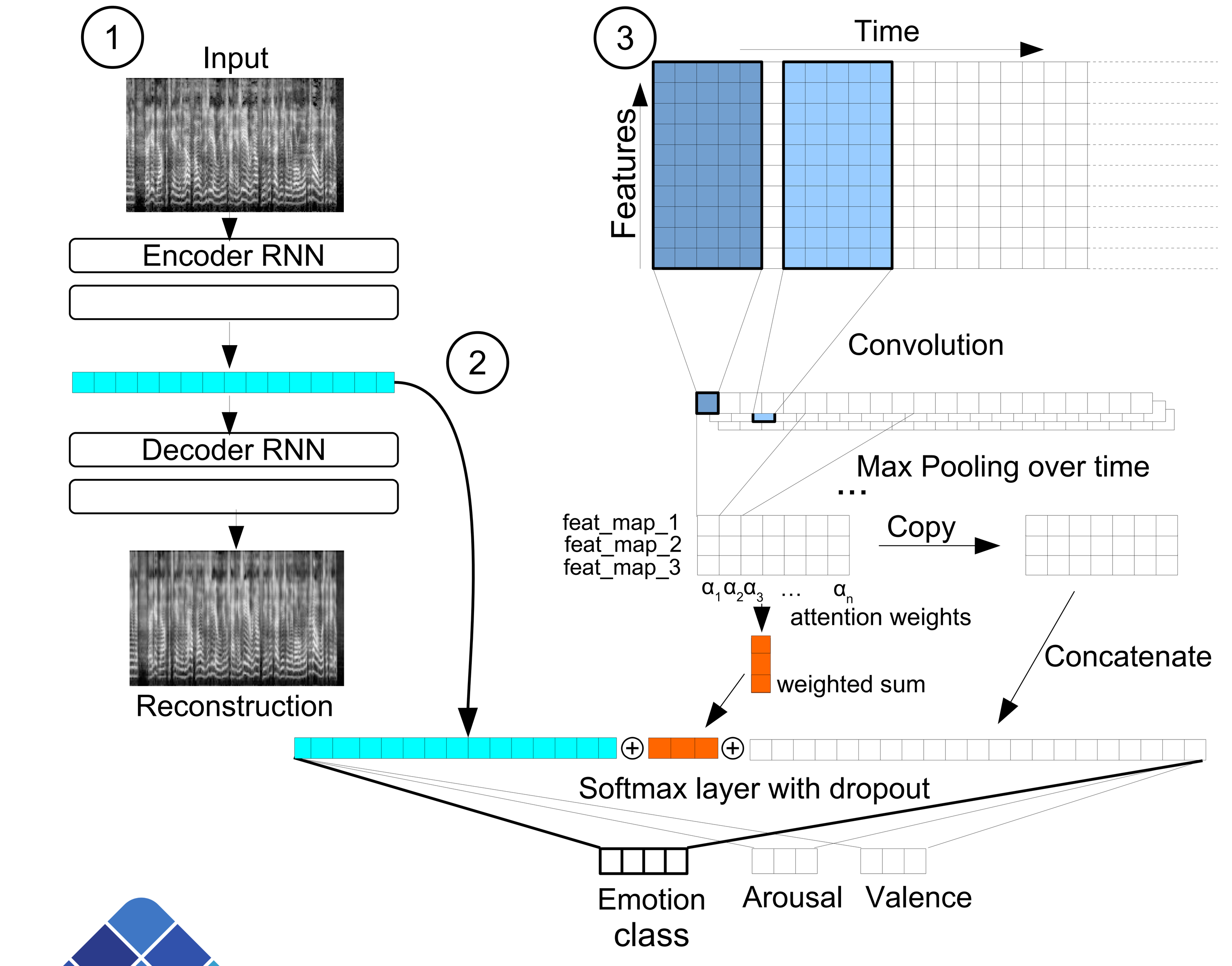
We present findings on how representation learning on large unlabeled speech corpora can be beneficially utilized for speech emotion recognition (SER). Evaluation is done by means of within- and cross-corpus testing.

Main findings:

- Integrating **representations** learned by **unsupervised autoencoder improves** emotion **classification**
- Autoencoder **representations bear emotional information** (especially arousal dimension)
- Consistent **improvements for within- and cross-corpus** evaluation

Methods

1. Train time-recurrent sequence-to-sequence autoencoder on spectrograms from large speech corpus (auDeep toolkit [1])
2. Generate latent representations for emotional speech
3. Train attentive convolutional neural network (ACNN) [2] with those representations as additional feature vector



Speech Corpora

- IEMOCAP [3]
  5,531 utterances from 10 speakers, classes {angry, happy, neutral, sad}
- MSP-IMPROV [4] (only for evaluation)
  7,798 utterances from 12 speakers, same 4 classes
- Tedlium r2 [5]
  207 hours (92,973 utterances)
- Librispeech [6]
  100 hours subset (28,539 utterances)

Experimental Results

Baseline

- ACNN without additional representations
- 5-fold cross validation (speaker-independent) for IEMOCAP

Autoencoder (AE) training on 4 datasets

a) **'Control condition'**: AE trained on IEMOCAP itself (respectively MSP-IMPROV) – no additional data source

b) **'small Tedlium'**: AE trained on subset of Tedlium (400 Ted talks, 25,303 segments)

c) **'Librispeech'**: AE trained on 100 hours Librispeech data

d) **'full Tedlium'**: AE trained on 207 hours of speech

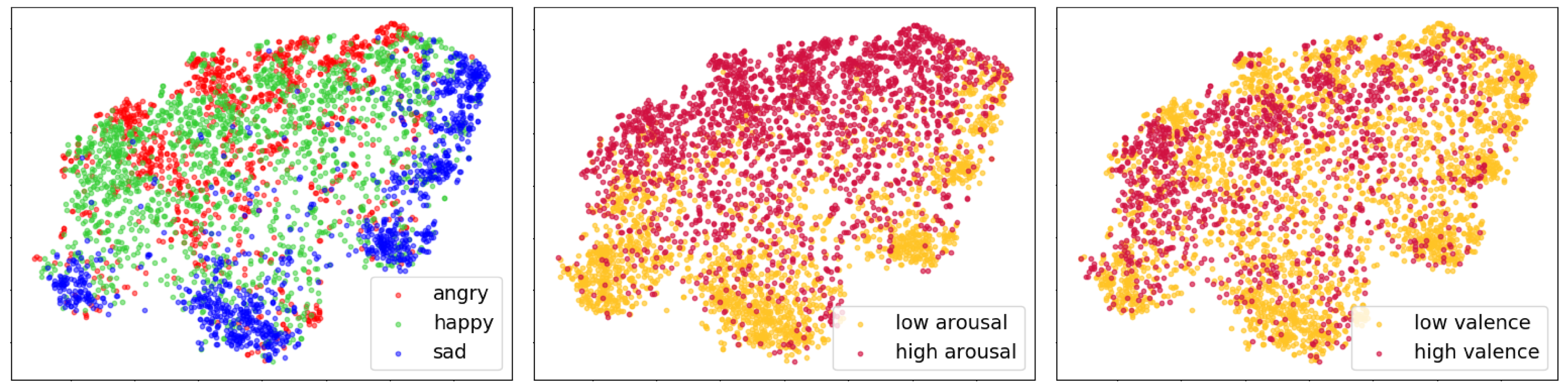
Increasing data

Unweighted average recall (UAR), averaged over 10 runs of the experiments for each setting

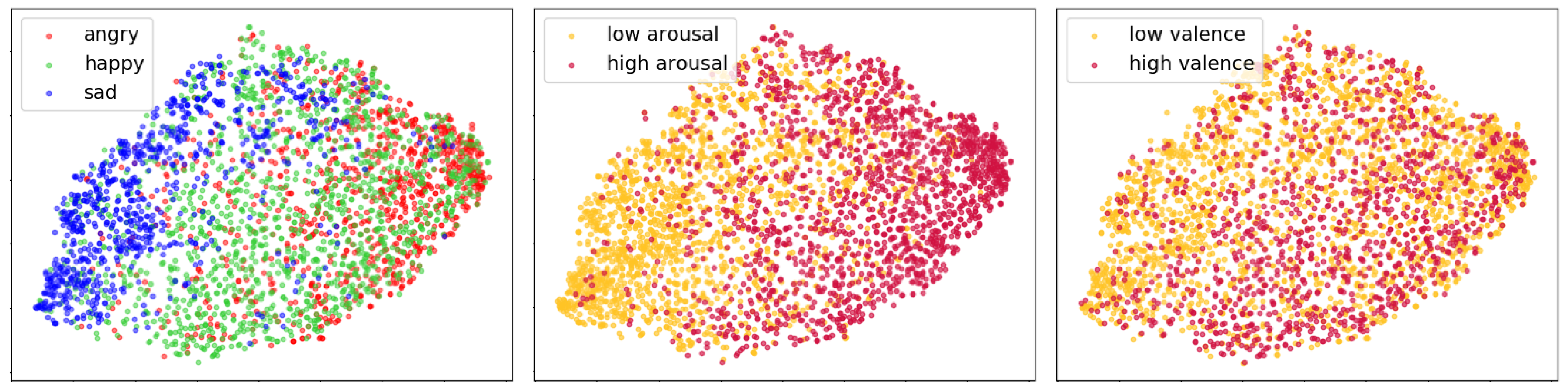
	IEMOCAP	MSP-IMPROV (cross-corpus)
Baseline	58.03	42.99
a) Control	58.07	42.37
b) small Ted	58.85	45.21
c) Librispeech	59.05	44.82
d) full Ted	**59.54**	**45.76**

→ Consistent improvements when adding representations generated by different AE models b), c), and d)

Visualization of Speech Representations



t-SNE visualizations of last hidden layer of the ACNN for IEMOCAP



t-SNE visualizations of the AE representations for IEMOCAP (AE trained on full Tedlium, no emotion information involved in training)

- ACNN: *angry* and *sad* separated to certain extend; high-variance cluster for *happy*
- ACNN: much more discriminative for arousal than for valence
- **AE: similar patterns** despite no emotion labels are involved
- → AE implicitly learns to separate low and high arousal
- Both representations are invariant to speaker sex and speaker identity (no separable clusters found in visualizations)

Selected References

[1] Michael Freitag, Shahin 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.

[2] Michael Neumann and Ngoc Thang 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.

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

[4] Carlos Busso, Srinivas 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.

[5] Anthony Rousseau, Paul Deléglise, and Yannick 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.

[6] Vassil Panayotov, Guoguo 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.

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

[8] Sefik Emre Eskimez, Zhiyao Duan, and Wendi 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.

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

www.ims.uni-stuttgart.de

Contact: michael.neumann@ims.uni-stuttgart.de