4461

SERAB: A multi-lingual benchmark for speech emotion recognition

N. Scheidwasser-Clow[1, 2], M. Kegler[3], P. Beckmann[1], M. Cernak[2]

[1]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland [2]Logitech Europe S.A., Lausanne, Switzerland

[3]Imperial College London, London, United Kingdom

Code

Paper

Motivation

- How to tackle speech emotion recognition (SER) problems?
  - **Small datasets** → task-specific models generalize poorly
  - Approaches:
    - Handcrafted feature sets
    - **Self-supervised** DNNs as audio/speech representations
- SER performance can vary with the evaluation protocol
- Multi-dataset benchmarks already exist for:
  - Computer Vision [1]
  - Natural Language Processing [2]
  - Non-semantic speech processing [3]
- **SERAB** = a benchmark for SER:
  - **9** datasets
  - **6** languages
  - Different **sizes** (500-7,500 samples)
  - Different **emotion classes** (anger, fear, sadness...)

SERAB: tasks

Dataset	Language	Classes	Samples
AESDD [4]	Greek	5	604
CaFE [5]	French	7	864
CREMA-D [6]	English	6	7442
EmoDB [7]	German	7	535
EMOVO [8]	Italian	7	588
IEM4 [9]	English	4	5531
RAVDESS [10]	English	8	1440
SAVEE [11]	English	7	480
ShEMO [12]	Persian	6	3000

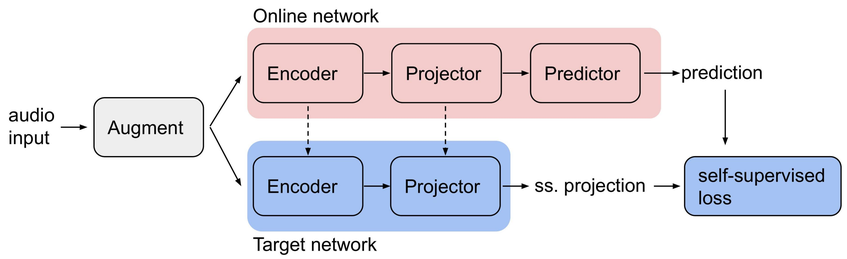
Baseline approaches

Compare well-established/state-of-the-art frameworks:

- Acoustic feature set: openSMILE [13]
- Speech representation: TRILL [3]
- Audio representations: VGGish [14], YAMNet [15], BYOL-A [16]

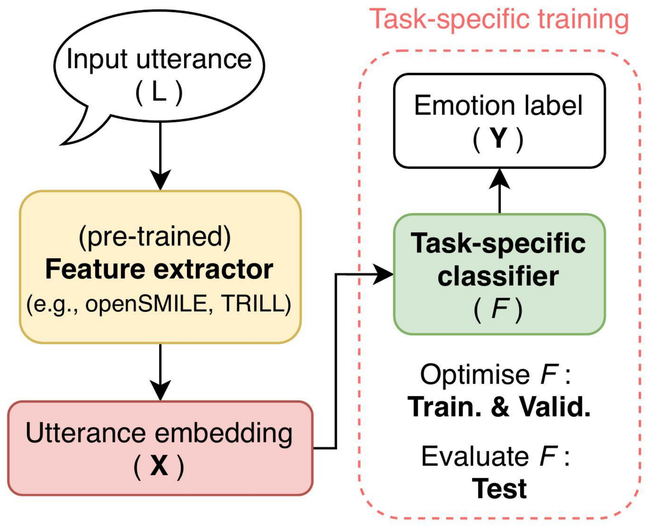
Proposed:

- BYOL-A retrained on speech samples only → **BYOL-S**
- BYOL-A with CvT [17] encoding → **BYOL-S/CvT**

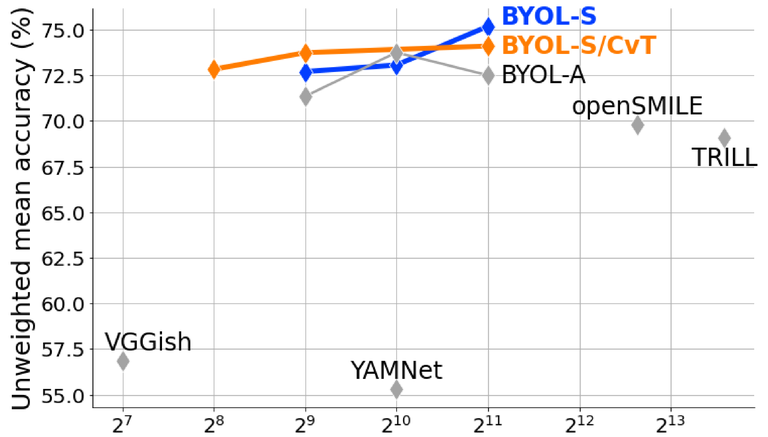


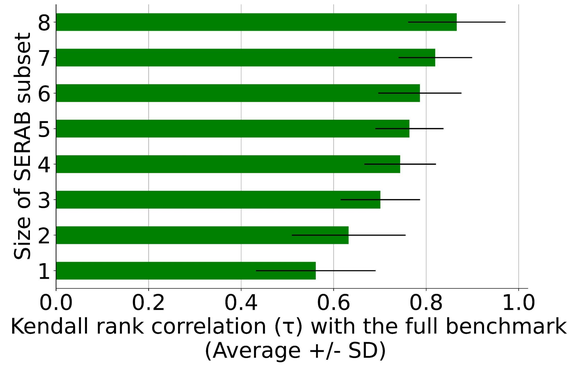
BYOL-A framework. Adapted from [16]

SERAB: evaluation



Results





References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei- Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[2] A.Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *ICLR*, 2018.

[3] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. C. Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards Learning a Universal Non-Semantic Representation of Speech," in *Proc. Interspeech*, 2020, pp. 140–144.

[4] N. Vryzas, R. Kotsakis, A. Liatsou, C. A. Dimoulas, and G. Kalliris, "Speech emotion recognition for performance interaction," *J. Audio Eng. Soc.*, vol. 66, no. 6, pp. 457–467, 2018.

[5] P. Gournay, O. Lahaie, and R. Lefebvre, "A Canadian French emotional speech dataset," in *Proc. ACM Multimedia Systems*, 2018, pp. 399–402.

[6] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, 2014

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, 2005.

[8] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO corpus: an Italian emotional speech database," in *LREC*, 2014, pp. 3501–3504

[9] 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," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[10] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, 2018.

[11] S. Haq, P. J. Jackson, and J. Edge, "Speaker-dependent audio-visual emotion recognition," in *AVSP*, 2009, pp. 53–58.

[12] O. M. Nezami, P. J. Lou, and M. Karami, "ShEMO: a large-scale validated database for Persian speech emotion detection, *Lang. Resour. Eval.*, vol. 53, no. 1, pp. 1–16, 2019.

[13] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia*, 2010, pp. 1459–1462.

[14] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al., "CNN architectures for large-scale audio classification," in *ICASSP*, 2017, pp. 131–135.

[15] M. Plakal and D. Ellis, "YAMNet," https://github.com/tensorflow/models/tree/master/research/audioset/yamnet, 2020

[16] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for Audio: Self-supervised learning for general-purpose audio representation," in *IJCNN*, 2021.

[17] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," *arXiv preprint arXiv:2103.15808*, 2021.