# Supplementary Material for the Paper: ExHuBERT: Enhancing HuBERT Through Block Extension and Fine-Tuning on 37 Emotion Datasets

*Shahin Amiriparian*[1], *Filip Packań*[2], *Maurice Gerczuk*[2], *Björn W. Schuller*[1,2,3]

[1]Chair of Health Informatics, MRI, TU Munich, Germany
[2]Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[3]Group on Language, Audio, & Music, Imperial College, UK
shahin.amiriparian@tum.de

| Datasets | Performance in [% UAR] | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **W2V2 XLS-R 300M** | **W2V2 XLS-R 1B** | **Whisper Medium** | **Whisper Large v3** | **HuBERT Large** | **HuBERT XLarge** | **HuBERT Large EMOSET + 5** | **HuBERT Large EMOSET++** |
| Airplane Behaviour Corpus (ABC) [1] | 52.6 | 50.8 | 43.0 | 33.2 | 59.9 | 63.9 | **67.8** | 64.1 |
| Anger Detection (AD) [2] | 89.3 | 87.7 | 63.4 | 75.7 | **90.4** | 89.7 | 89.2 | 83.2 |
| Burmese Emotional Speech (BES) [3] | 55.0 | **60.0** | 40.0 | 42.5 | 52.5 | 45.0 | 38.7 | 41.2 |
| CASIA [4] | 48.0 | 41.8 | 39.0 | 35.2 | 46.8 | **66.4** | 63.0 | 49.8 |
| Chinese Vocal Emotions (CVE) [5] | 71.5 | 63.0 | 38.1 | 38.0 | 66.5 | 74.0 | 64.5 | **75.5** |
| Database of Elicited Mood in Speech (DEMOS) [6] | 56.1 | 69.2 | 30.0 | 32.8 | 67.7 | 57.2 | 70.8 | **90.7** |
| Danish Emotional Speech (DES) [7] | 93.0 | 88.2 | 84.9 | 93.1 | 92.0 | **97.0** | 96.0 | 96.1 |
| EA-ACT [8] | 73.0 | 85.0 | 39.0 | 39.0 | 63.0 | **85.0** | 56.0 | 70.0 |
| EA-BMW [8] | 74.8 | 71.5 | 47.4 | 54.4 | 58.1 | **82.2** | 74.8 | 54.8 |
| EA-WSJ [8] | 98.1 | 98.1 | 84.6 | 84.6 | 100 | 100 | 100 | **100** |
| Berlin Database of Emotional Speech (EMO-DB) [9] | 91.1 | 86.0 | 55.0 | 61.2 | 90.0 | 82.4 | 88.1 | **99.1** |
| eNTERFACE [10] | 66.4 | 76.1 | 39.3 | 38.6 | 93.9 | 63.2 | 92.9 | **94.6** |
| EU-Emotion Voice Database (EU-EV) [11] | 36.3 | 42.3 | 37.1 | 33.0 | 41.1 | 29.8 | 41.4 | **92.8** |
| EmoFilm [12] | 56.9 | 43.4 | 49.7 | 44.5 | 58.6 | 57.4 | 58.9 | **60.5** |
| EmotiW-2014[13] | 40.0 | 34.7 | 28.6 | 27.2 | 33.1 | 32.5 | **40.2** | 39.1 |
| FAU Aibo [14] | 36.2 | 35.3 | 31.1 | 25.8 | 38.2 | 31.5 | 39.1 | **39.8** |
| Geneva Emotion Portrayal (GEMEP) [15] | **54.1** | 48.6 | 33.7 | 32.1 | 48.8 | 51.4 | 49.0 | 53.2 |
| Geneva Vocal Emotion Express (GVESS) [16] | **43.7** | 39.3 | 20.1 | 27.9 | 49.0 | 36.8 | 38.8 | 40.8 |
| IEMOCAP [17] | 56.4 | 49.7 | 42.9 | 39.0 | 61.1 | 65.8 | 63.9 | **67.8** |
| Multimodal EmotionLines Dataset (MELD) [18] | 23.2 | 24.7 | 23.8 | 22.6 | 30.0 | 25.9 | 34.1 | **38.5** |
| Mandarin Emotional Speech (MES) [3] | 66.3 | 48.8 | 66.3 | 65.0 | 67.5 | 63.8 | **70.0** | 67.5 |
| PPMMK [2] | 45.0 | 46.1 | 33.6 | 31.1 | 51.1 | 55.9 | 51.1 | **61.4** |
| Speech in Minimal Invasive Surgery (SIMIS) [19] | 26.3 | 25.3 | 25.6 | 25.3 | 26.4 | 26.0 | 29.6 | **32.1** |
| SmartKom Multimodal Corpus (SmartKom) [20] | 20.5 | 20.5 | 20.1 | 19.6 | 21.6 | 21.4 | 21.8 | **39.4** |
| Speech under Simulated and Actual Stress (SUSAS) [21] | 37.3 | 27.3 | **51.9** | 51.0 | 45.4 | 47.0 | 33.4 | 44.7 |
| TurkishEmo [2] | 63.6 | 58.0 | 46.6 | 46.6 | 72.7 | 73.9 | 68.2 | **80.7** |
| Toronto emotional speech set (TESS) [22] | – | – | – | – | – | – | 97.9 | **99.5** |
| EU-EmoSS [23] | – | – | – | – | – | – | – | 85.1 |
| Crowd-sourced Emotional Multimodal Actors Dataset (Crema-D) [24] | – | – | – | – | – | – | 68.8 | **79.9** |
| Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [25] | – | – | – | – | – | – | 22.3 | **27.7** |
| Sharif Emotional Speech Database (ShEMO) [26] | – | – | – | – | – | – | 42.9 | **59.6** |
| Surrey Audio-Visual Expressed Emotion (SAVEE) [27] | – | – | – | – | – | – | 51.8 | **63.4** |
| URDU [28] | – | – | – | – | – | – | – | 89.5 |
| SUST Bangla Emotional Speech Corpus (SUBSECO) [29] | – | – | – | – | – | – | – | 61.7 |
| Mexican Emotional Speech Database (MESD) [30] | – | – | – | – | – | – | – | 91.5 |
| EMOVO [31] | – | – | – | – | – | – | – | 65.1 |
| Emotional Speech Dataset (ESD) [32, 33] | – | – | – | – | – | – | – | 93.1 |

Table 1: *Performance comparison of the applied Transformers on a wide range of speech emotion datasets. Our proposed fine-tuning of HuBERT Large on EMOSET++ demonstrates superior performance over all other Transformers.*

# 1. References


[1] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, "Audiovisual Behavior Modeling by Combined Feature Spaces," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 2, Apr. 2007, pp. II-733-II–736.

[2] M. Gerczuk, S. Amiriparian, S. Ottl, and B. Schuller, "EmoNet: A Transfer Learning Framework for Multi-Corpus Speech Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 13, 2022.

[3] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, Nov. 2003.

[4] - *ChineseLDC.Org* -, http://www.chineseldc.org/resource_info.php?rid=76.

[5] P. Liu and M. D. Pell, "Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli," *Behavior Research Methods*, vol. 44, no. 4, pp. 1042–1051, Dec. 2012.

[6] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. W. Schuller, "DEMoS: An Italian emotional speech corpus," *Language Resources and Evaluation*, vol. 54, no. 2, pp. 341–383, Jun. 2020.

[7] I. S. Engberg, A. V. Hansen, O. K. Andersen, and P. Dalsgaard, "Design Recording and Verification of a Danish Emotional Speech Database: Design Recording and Verification of a Danish Emotional Speech Database," *EUROSPEECH'97 : 5th European Conference on Speech Communication and Technology, Patras, Rhodes, Greece, 22-25 September 1997*, Vol. 4, pp. 1695–1698, 1997.

[8] B. Schuller, "Automatische Emotionserkennung aus sprachlicher und manueller Interaktion," Ph.D. dissertation, Technische Universität München, 2006.

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, ISCA, Sep. 2005, pp. 1517–1520.

[10] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE' 05 Audio-Visual Emotion Database," in *Proc. ICDEW*, Apr. 2006.

[11] A. Lassalle, D. Pigat, H. O'Reilly, S. Berggen, S. Fridenson-Hayo, S. Tal, S. Elfström, A. Råde, O. Golan, S. Bölte, S. Baron-Cohen, and D. Lundqvist, "The EU-Emotion Voice Database," *Behavior Research Methods*, vol. 51, no. 2, pp. 493–506, Apr. 2019.

[12] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird, and B. Schuller, "Categorical vs Dimensional Perception of Italian Emotional Speech," in *Proc. INTERSPEECH*, ISCA, Sep. 2018, pp. 3638–3642.

[13] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion Recognition In The Wild Challenge 2014: Baseline, Data and Protocol," in *Proc. ICMI*, New York, NY, USA: Association for Computing Machinery, Nov. 2014, pp. 461–466, ISBN: 978-1-4503-2885-2.

[14] A. Batliner, S. Steidl, and E. Noth, "Releasing a thoroughly annotated and processed spontaneous emotional database: The FAU Aibo Emotion Corpus," 2008.

[15] K. R. Scherer, T. Bänziger, and E. Roesch, *A Blueprint for Affective Computing: A Sourcebook and Manual*. OUP Oxford, Sep. 2010, ISBN: 978-0-19-956670-9.

[16] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[17] 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, pp. 335–359, Dec. 2008.

[18] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations*, Jun. 2019. arXiv: 1810.02508 [cs].

[19] B. Schuller, F. Eyben, S. Can, and H. Feußner, "Speech in Minimal Invasive Surgery - Towards an Affective Language Resource of Real-life Medical Operations," 2010.

[20] F. Schiel, S. Steininger, and U. Türk, "The SmartKom Multimodal Corpus at BAS," in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, M. González Rodríguez and C. P. Suarez Araujo, Eds., Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA), May 2002.

[21] J. H. L. Hansen and S. E. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. Eurospeech 1997*, 1997, pp. 1743–1746.

[22] M. K. Pichora-Fuller and K. Dupuis, *Toronto emotional speech set (TESS)*, Feb. 2020.

[23] H. O'Reilly, D. Pigat, S. Fridenson, S. Berggren, S. Tal, O. Golan, S. Bölte, S. Baron-Cohen, and D. Lundqvist, "The EU-Emotion Stimulus Set: A validation study," *Behavior Research Methods*, vol. 48, no. 2, pp. 567–576, Jun. 2016.

[24] 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 transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[25] 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, e0196391, May 2018.

[26] O. Mohamad Nezami, P. Jamshid Lou, and M. Karami, "ShEMO: A large-scale validated database for Persian speech emotion detection," *Language Resources and Evaluation*, vol. 53, no. 1, pp. 1–16, Mar. 2019.

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

[28] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross Lingual Speech Emotion Recognition: Urdu vs. Western Languages," in *2018 International Conference on Frontiers of Information Technology (FIT)*, Dec. 2018, pp. 88–93.

[29] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "SUST Bangla Emotional Speech Corpus (SUBESCO): An audio-only emotional speech corpus for Bangla," *PLOS ONE*, vol. 16, no. 4, e0250173, Apr. 2021.

[30] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, "The Mexican Emotional Speech Database (MESD): Elaboration and assessment based on machine learning," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2021, pp. 1644–1647, Nov. 2021.

[31] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO Corpus: An Italian Emotional Speech Database," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds., Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3501–3504.

[32] "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[33] K. Zhou, B. Sisman, R. Liu, and H. Li, *Seen and Unseen emotional style transfer for voice conversion with a new emotional speech dataset*, Feb. 2021. arXiv: 2010.14794 [cs, eess].