INTERSPEECH 2009 Emotion Challenge Revisited: Benchmarking 15 Years of Progress in Speech Emotion Recognition

Andreas Triantafyllopoulos¹, Anton Batliner¹, Simon Rampp², Manuel Milling¹, Björn Schuller^{1,2,3}

¹CHI – Chair of Health Informatics, MRI, Technical University of Munich, Germany ²EIHW – Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany ³GLAM – Group on Language, Audio, & Music, Imperial College London, UK

andreas.triantafyllopoulos@tum.de

Abstract

We revisit the INTERSPEECH 2009 Emotion Challenge – the first ever speech emotion recognition (SER) challenge – and evaluate a series of deep learning models that are representative of the major advances in SER research in the time since then. We start by training each model using a fixed set of hyperparameters, and further fine-tune the best-performing models of that initial setup with a grid search. Results are always reported on the official test set with a separate validation set only used for early stopping. Most models score below or close to the official baseline, while they marginally outperform the original challenge winners after hyperparameter tuning. Our work illustrates that, despite recent progress, FAU-AIBO remains a very challenging benchmark. An interesting corollary is that newer methods do not consistently outperform older ones, showing that progress towards 'solving' SER is not necessarily monotonic.

Index Terms: Speech emotion recognition, Deep learning

1. Introduction

Standardised benchmarks form the backbone of reproducible science and enable the research community to showcase its progress towards a common objective. Speech emotion recognition (SER) is one subfield of speech science where several benchmarks exist, mostly in the form of challenges like: the INTERSPEECH 2009 Emotion Challenge (ComParE) [1], the first official challenge on SER, which was followed by several iterations covering a wide gamut of paralinguistic tasks; the Audio-Visual Emotion Challenge (AVEC) [2]; MuSe [3]; OMG [4]; just recently, the Odyssey 2024 SER challenge; and others.

However, while such challenges form excellent 'proving grounds' for the prevalent methods at a particular *Zeitgeist*, they are rarely revisited when newer approaches emerge. Furthermore, popular datasets oftentimes suffer from non-standardised folds (as in the case of IEMOCAP [5]) or iterative releases (like the recent MSP-Podcast [6], with a new version being released almost every year), which makes it harder to obtain a consistent comparison of all different methods.

In the present contribution – and on the occasion of its 15th anniversary – we focus exclusively on FAU-AIBO [7, 8], the dataset used for the first-ever SER challenge in 2009. The challenge contained two alternative formulations of emotion: a 2-class problem, where participants had to differentiate between negative and non-negative emotions; and a 5-class problem, where participants had to classify an utterance as angry (A), neutral (N), mothereseljoyful (P), emphatic (E), with a 5th rest (R) class. As the challenge ran before the advent of the 'deep learning (DL) era', participants never benefited from these advances. Moreover, as newer datasets emerged over time, FAU-AIBO has been relatively overlooked. Specifically, although [1] has been

referred to rather often – being a standard reference to an early paper presenting spontaneous emotions, clear partitioning, and baselines – to the best of our knowledge, in the last 15 years there has been only a limited number of studies that unequivocally used the same configuration of train and test partition with identical number of items in each class. In fact, sometimes it is explicitly mentioned that "the results presented in this paper are not directly comparable with those found using the 2009-challenge data" [9]. On a positive note, this makes it a perfect test case for a long-overdue retrospective, centred on the question of whether the community has 'solved' or at least substantially progressed on the problem of SER in the intervening years.

To answer this question, we run a large-scale study of several 'high-profile' advances that have emerged after 2009: We start from larger, more comprehensive feature sets which defined the SER landscape until ca. 2016, where we train both multi-layered perceptrons (MLPs) on their static and long short-term memory recurrent neural networks (LSTMs) on their dynamic versions. After that, we move on to the first end-to-end convolutional recurrent neural networks (CRNNs) as well as spectrogram-based convolutional neural networks (CNNs) benefiting from transfer learning. Finally, we investigate the more recent transformers pre-trained with self-supervised learning (SSL). We note that such previous large-scale experiments have only been carried out within a particular architecture family [10, 11, 12]. We additionally analyse our results with respect to inter-model agreement, examine whether hard-to-classify cases are also those where human annotators disagree the most, and try to measure whether progress is monotonic with respect to the year a model was introduced or its size (i. e., computational complexity). We note that we exclude advances on linguistics due to space limitations.

2. Previous work on FAU-AIBO

Challenge: The official challenge baseline consisted of HMM modelling of dynamic features or support vector machine (SVM) modelling of static features each combined with SMOTE oversampling to mitigate class imbalance and a separate standardisation applied on the training and test sets [1], with static features yielding moderately better performance. The challenge featured two winners: The open performance sub-challenge was won by Dumouchel et al. [13], who employed Gaussian mixture models (GMMs) trained on cepstral and expert prosodic and vocal tract features; The classifier performance sub-challenge was won by Lee et al. [14], who used the official static features but with a divide-and-conquer cascade classification approach. Kockmann et al. [15] obtained the best performance on the 5-class problem by employing a fusion of different GMMs trained on functionals (but were only 0.1% better than Lee et al. [14]). A review found an overall tendency towards smaller, carefully-designed

features over 'brute-force' approaches [16], and a fusion of the top approaches led to additional performance gains [16].

Beyond the challenge: Researchers continued to improve performance on FAU-AIBO after the end of the challenge. Closely related to our work, Cummins *et al.* [9] and Zhao *et al.* [11] both report better performance than the challenge winners using transfer-learning from CNNs pre-trained on image data; however, the former use different splits than the challenge and the latter perform a very extensive hyperparameter search (a total of 15 hyperparameters were optimised, resulting in a search space much larger than the one we employ here).

3. Methodology

3.1. Dataset

We use the official dataset of the INTERSPEECH 2009 Emotion Challenge [1], FAU-AIBO. It is a dataset of German children's speech collected in a Wizard-of-Oz scenario and annotated on the word-level for the presence of 11 emotional/communicative states by 5 raters [7, 8]. Subsequently, segmented words have been aggregated to meaningful chunks using manual semantic and prosodic criteria. Accordingly, annotated states have been mapped to 2- and 5-class categorisation using a set of heuristics, which forms a final dataset of 18216 chunks used for the challenge. The data is heavily imbalanced towards the neutral/non-negative classes. The data was collected from two schools, with one school set aside for testing (Mont) and one set aside for training (Ohm); we use the same partitioning for our experiments. Additionally, we create a small validation set comprising the last two speakers of the training set (speakers are denoted by number IDs): Ohm_31 and Ohm_32, similar to [17]. Note that the data are extensively documented in [8].

3.2. Models

The Computational Paralinguistics Challenge (ComParE) continued the first challenge for a further 14 years. Our selection of models is largely based on the ComParE series' baselines and best-performing winners of each year and on prevalent trends in the last decade of SER research. We briefly describe each model below, but also include an appendix with linked model states as well as plan to release the source code for our experiments¹.

'09-'16: openSMILE feature sets – In the follow-up iterations of the ComParE Challenge, newer versions of paralinguistic features were introduced. In general, these feature sets were larger and covered a wider gamut of acoustic and prosodic features. However, that period saw a parallel pursuit for smaller expertdriven feature sets [18]. To accommodate both, we use both the static ('functionals') and dynamic ('low-level descriptors') versions of the official IS09-IS13 and IS16 feature sets, as well as the EGEMAPS feature set [18] as provided in the latest version of the openSMILE toolkit [19]. For the dynamic features, we used a 2-layered LSTM model with 32 hidden units, followed by mean pooling over time, one hidden linear layer with 32 neurons and ReLU acivation, and one output linear layer; all hidden layers are followed by a dropout of 0.5; these models are denoted with ^d. Additionally, we train 3-layered MLPs with 64 hidden units each, a dropout of 0.5, and ReLU activation for the static features; these models are denoted with s.

'12-'23: ImageNet pre-training – Following the introduction of ImageNet and the first CNNs trained on it in 2012, such networks were subsequently introduced in the audio and speech domains

by substituting images with (pictorial representations of) spectrograms [9, 11] – a practice that is relevant to this day, with audio transformer models oftentimes initialised with states pre-trained on ImageNet [20]. We use ALEXNET, RESNET50, all versions of VGG (11,13,16,19), EfficientNet-B0 (EFFNET), the tiny, small, base, and large versions of CONVNEXT (t,b,s,l), and the tiny, base, and small versions of the SWIN Transformer (t,b,s). In all cases, we use the best-performing model state on ImageNet as available in the TORCHVISION-V0.16.0 package.As features, we always used the Mel-spectrograms generated for CNN14 (see below), i. e., 64 Mels with a window size of 32 ms and a hop size of 10 ms; the resulting matrices were then replicated over the three dimensions to generate the 3-channel input that is required by models designed for computer vision tasks.

'16-'23: End-to-end — Subsequent years saw the introduction of *end-to-end* models, i. e., models trained directly on raw audio input for the target task without any prior feature pre-processing. These models were especially successful in the case of time-continous SER, which requires predicting the emotion of very short audio frames, and essentially follow the CRNN architecture. We use two particular instantiations introduced by Tzirakis *et al.* [21] (CRNN¹⁸) and Zhao *et al.* [22] (CRNN¹⁹).

'16-'23: Supervised audio pre-training – In parallel to ImageNet pre-training, there were also efforts to collect similar largescale datasets for audio where networks could be pre-trained in a supervised fashion. Two notable examples are VoxCeleb and AudioSet, both collected from YouTube, with the former targeted to speaker identification and the latter to general audio tagging. VoxCeleb formed the basis for training speaker embedding models (i.e., 'x-vectors') using time delay neural networks (TDNNs), of which we use a more recent and improved attentionbased model (ETDNN) [23]. AudioSet in turn inspired the use of VGG-based convolutional networks, such as CNN14 introduced in pretrained audio neural networks (PANNs) [24] which was shown to also transfer well to SER tasks [25], and later, transformer-based models such as AST [20]. In addition, the introduction of the WHISPER architecture [26] led to a renaissance of supervised training for automatic speech recognition (ASR) and we thus include it in our experiments – albeit only the three smallest available variants (t,b,s) due to hardware constraints.

'20-'23: Self-supervised audio pre-training - The introduction of transformers and the advent of self-supervised pre-training for computer vision and natural language processing (NLP) also propagated to the speech and audio domain. The two dominant architectures here are wav2vec2.0 [27], which includes a convolutional backend followed by a transformer decoder trained to reconstruct its own quantised intermediate representations, and HuBert [28], a full-transformer model trained on masked token prediction. These models have yielded significant advances in SER [12], which was partially accredited to their ability to simultaneously encode linguistic and paralinguistic information [29]. In this work, we use the pretrained states from the base and large variants of wav2vec2.0 and HuBert (W2V2^{b,l}, HUB^{b,l}), a multilingual model trained on VoxPopuli [30] (W2V2m), a 'robust' version of wav2vec2.0 trained on more data [31] (W2V2^r), as well as the pruned version of that model further fine-tuned for dimensional SER on MSP-Podcast [12] (W2V2^e). Similar to Wagner et al. [12], we add an output 2-layered MLP which takes the pooled hidden embeddings of the last layer as input.

3.3. Experiments

To constrain our space of hyperparameters, we conduct two experimental phases. In the first **exploration phase**, we test all 43

¹github.com/ATriantafyllopoulos/is24-interspeech09-ser-revisited

investigated models using a fixed set of hyperparameters. Specifically, we use the Adam optimiser with a learning rate of 0.0001 and a batch size of 4 for 30 epochs. In the following tuning **phase**, we further optimise a larger set of hyperparameters for the 5 best-performing models from the exploration phase, doing a grid search over optimisers {Adam, SGD}, learning rates $\{0.01, 0.001, 0.0001\}$, and batch sizes $\{4, 8, 16\}$, while training each configuration for 50 epochs. In total, this results in 43 runs for the exploration phase and an additional 90 runs for the tuning phase (for each of the two challenge tasks). To account for variable-length sequences in training, we randomly cropped/padded all chunks to a fixed length of 3 seconds (different cropping/padding was applied on each instance across different epochs; random seed was fixed and can be reproduced) when using dynamic features (including the raw audio); during inference, we used the original utterances, only padding those shorter than 2 seconds with silence. Our loss function is the standard categorical cross-entropy, where - in order to account for the severe class imbalance – we further weigh the contribution of each instance by the inverse frequency of its true label on the training set, similar to Zhao et al. [11]. Except for W2V2 and HUB, where we freeze the feature extractors, all model parameters are fine-tuned.

In all cases, we use the defined validation set (comprising two speakers from the original training set) to select the best-performing epoch for each model, which we then proceed to evaluate on the test set. Strictly speaking, this results in different training data from some challenge participants, as they typically retrained their models on the entire training set after optimising their hyperparameters (e. g., via cross-validation). However, given the propensity of deep neural networks (DNNs) to overfit when trained too long, we found the use of such a validation set necessary for early stopping.

4. Results & Discussion

Results for the *exploration phase* are presented in Table 1. The best-performing model for the 2-class problem is CNN14, with an unweighted average recall (UAR) of .692, whereas for the 5-class problem it is RESNET50, with a UAR of .428. We further observe that some models have failed to converge and yield chance (or near-chance) performance – most likely caused by the choice of hyperparameters, which favour some models more than others. Notably, most of these results are below the challenge winners (UAR: .703/.417) and several are in the same 'ballpark' as the original baseline (UAR: .677/.382). As in Schuller *et al.* [16], a fusion of our top models (here we only take 5) improves performance.

Turning to the *tuning phase*, Table 2 shows the *best* performance for the top five models from Table 1 after optimising standard hyperparameters (optimiser, learning rate, batch size). In this case, W2V2^e yields the best performance for the 2-class problem (.717) and WHISPER^t for the 5-class problem (.454) – in both cases, we reach results better than the challenge winners, albeit with the gains for the 2-class problem being marginal. Given that WHISPER has been trained for multilingual ASR (including German), and that previous performance improvements on valence prediction for English speech heavily depended on implicit linguistic knowledge [29], we expect WHISPER's success to be also attributed to that aspect.

However, it is still the case that all models we have tested remain close to or even below the original challenge baseline and winners, and especially the fusion of the top challenge submissions. This remains so even after selecting only the best-

Table 1: UAR results for all tested models in the **exploration phase** using our standard hypermarameters (Adam, 0.01, 4) for both the 2- and the 5-class model.

Model	Year	2-class	5-class
IS09 ^s	2009	.620	.294
$ISO9^d$	2009	.670	.347
$IS10^s$	2010	.670	.348
$IS10^d$	2010	.689	.406
IS11 ^s	2011	.499	.201
$IS11^d$	2011	.670	.373
$IS12^s$	2012	.501	.201
ALEXNET	2012	.503	.200
$IS12^d$	2012	.679	.373
$IS13^d$	2013	.664	.378
$IS13^s$	2013	.498	.201
$EGEMAPS^s$	2015	.618	.245
$EGEMAPS^d$	2015	.667	.397
RESNET50	2015	.688	.428
VGG^{19}	2016	.499	.204
$IS16^s$	2016	.500	.201
VGG^{16}	2016	.646	.404
$IS16^d$	2016	.655	.382
VGG^{13}	2016	.665	.385
VGG^{11}	2016	.666	.200
EffNet	2019	.669	.353
$CRNN^{18}$	2018	.680	.392
$CRNN^{19}$	2019	.683	.372
$W2V2^l$	2020	.500	.200
$\mathrm{W2V2}^b$	2020	.500	.200
$ConvNeXt^t$	2020	.500	.400
$ConvNeXt^l$	2020	.663	.409
$ConvNeXt^b$	2020	.665	.412
$ConvNeXt^s$	2020	.674	.406
ETDNN	2020	.678	.404
CNN14	2020	.692	.394
HUB^b	2021	.500	.200
$SWIN^b$	2021	.528	.200
$SWIN^t$	2021	.530	.242
AST	2021	.535	.300
$W2V2^m$	2021	.640	.402
HUB^l	2021	.667	.418
$SWIN^s$	2021	.672	.306
$W2V2^r$	2021	.684	.411
Whisper ^s	2023	.656	.279
$W2V2^e$	2023	.684	.411
$WHISPER^t$	2023	.684	.380
$WHISPER^b$	2023	.686	.200
Late Fusion (All)		.676	.346
Late Fusion (Top-5)	_	.708	.434
		., 00	. 10 1

performing model out of all tested hyperparameters, essentially following a generally bad practice of overfitting. This was done intentionally to gauge performance under the most optimistic of settings – that of virtually unrestricted evaluation runs. We note that the original challenge participants were given 25 runs each. This shows how the gains we obtain here must be further tempered to account for more runs on our side.

Do newer/larger models perform better? Interestingly, when looking at the results of the exploration phase, there is no corre-

Table 2: UAR results for best-performing architectures in the tuning phase. Showing best results obtained after tuning hyper-parameters (batch size, optimiser, learning rate) and keeping the best-performing combination on the official test set. Also including 95% confidence intervals for our models computed with bootstrapping. SOTA results taken from original works.

Model	2-class	5-class
2009 Baseline	.677	.382
2009 Winners	.703	.417
2009 Fusion	.712	.440
Zhao et al. [11]	N/A	.454
$IS10^d$.685 [.674696]	.394 [.377411]
RESNET50	.690 [.680701]	.423 [.405441]
CNN14	.672 [.661683]	.448 [.428467]
$W2V2^e$.717 [.706728]	.448 [.431465]
$WHISPER^t$.707 [.696718]	.454 [.437472]

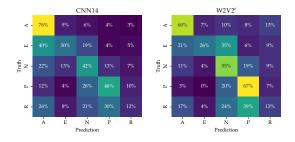


Figure 1: Test set confusion matrices (in %) for the best-performing CNN14 and W2V2^e models on the 5-class problem.

lation of UAR performance with the year of publication (Spearman's $\rho=.12/.09$ for the 2-/5-class problem), and the correlation with the amount of multiply-addition counts (MACs) and trainable parameters is also very low ($\rho=.15/.23$ and $\rho=-.08/.09$) – note that model MACs and parameters do not account for feature extraction. This further illustrates how neither more recent nor more complex models are able to surpass the prior state-of-the-art. Finally, the ranking of model performance between the 2- and 5-class problems is moderate ($\rho=.47$); this shows that models are not consistently good when given the same data but different labels (i. e., our findings are consistent with the standard "no free-lunch" theorem). Surprisingly, some models even show near chance-level performance on one task while performing well on the other.

Agreement between different models: Different models agree with one another to a moderate or good extent. The average pairwise agreement (percentage of instances where two models agree) for all models of the exploration phase is 70% and 55% for the 2- and 5-class models, which rises to 80% and 57%, respectively, when considering only the top-5 ones. Additionally, this is exemplified by considering the confusion matrices in Fig. 1 of the best-performing CNN14 and W2V2^e from the tuning phase – even though they result in an almost identical UAR, their behaviour on the test set is not very similar. For example, CNN14 shows a higher recall for the angry and emphatic classes, to the detriment of more neutral samples misclassified as such. Overall, this demonstrates that models trained on similar data do not converge to an identical solution – a finding congruent with the literature on underspecification [32].

Model vs human performance: We also investigate whether samples that are harder to classify for humans are also harder for models. The standard FAU-AIBO release comes with annotator confidences per instance, computed by taking the percentage of annotators who agree with the gold standard; we thus define 'difficulty' as 1 minus that confidence. We then make the following observations when considering all models of the tuning phase:

- a) We first adopt a model-agnostic measure of difficulty, which we define as the number of models who disagree with the max-vote computed by all models on each instance this is akin to the computation of difficulty for the human annotators. Spearman's ρ between this measure and annotator disagreement is moderate (.33 and .20 for the 2- and 5-class problems).
- b) We then adopt a model-specific measure of difficulty, defined as the cross-entropy loss for each instance, similar to Hacohen and Weinshall [33]. Different models have different rankings of instance difficulty, with average pairwise ρ being .51 and .33 for the 2- and 5-class problems.
- c) Finally, we compute the Spearman ρ between each model's UAR and its Spearman ρ with annotator disagreement; here, ρ is -.38/-.07 for the 2/5-class problem, indicating that larger agreement with human annotators does not lead to better performance (rather, the opposite). Collectively, our results indicate that models appear to learn differently than humans, with small agreement to what constitutes an easy or hard example. Limitations: Our study is obviously limited with respect to the approaches we tried; with hundreds of papers published on SER on a yearly basis, it was impossible to evaluate all of them. We thus opted for the simplest ones: fine-tuning large DNNs previously shown to be successful on other datasets using a range of standard hyperparameters. Furthermore, we have focused exclusively on one dataset; we intend to explore whether these findings generalise to other datasets in a follow-up work.

5. Conclusion

We have conducted a large-scale study of several modern DNN architectures - most of them pre-trained on large datasets - on the data of the INTERSPEECH 2009 Emotion Challenge. Given standard parameters, we were only able to marginally outperform the state-of-the-art achieved by challenge participants, with several models scoring even below the baseline, and some failing to converge altogether. Further optimising hyperparameters led us to outperform the challenge winners by small margins. Our subsequent analysis showed that performance improvements have not been consistent over time and are not caused by increased model size. Moreover, we have found that different models converge to different (sometimes complementary) solutions while differing in how challenging they find individual instances compared to human annotators. Collectively, our findings suggest that recent success achieved by DNN models must be tempered, at least when considering the FAU-AIBO setup.

6. Acknowledgements

This work was funded from the DFG's Reinhart Koselleck project No. 442218748 (AUDIONOMOUS) and the EU H2020 project No. 101135556 (INDUX-R), the Munich Data Science Institute (MDSI), and the Munich Center for Machine Learning (MCML). We would also like to remember our colleague Stefan Steidl – dedicated co-organiser of the ComParE challenges – who unexpectedly passed away at young age in October 2018.

7. References

- B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Proc. INTERSPEECH*, Brighton, UK, Sep. 2009, pp. 312–315.
- [2] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011–the first international audio/visual emotion challenge," in *Proc. ACII*, Springer, Memphis, TN, USA, 2011, pp. 415–424.
- [3] L. Stappen, A. Baird, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B. W. Schuller, I. Lefter, et al., "Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild," in Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop, 2020, pp. 35–44.
- [4] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The omg-emotion behavior dataset," in 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–7.
- [5] 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, pp. 335–359, 2008.
- [6] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.
- [7] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russell, and M. Wong, ""You stupid tin box" children interacting with the AIBO robot: A cross-linguistic emotional speech corpus," in *Proc. LREC*, Lisbon, 2004, pp. 171–174.
- [8] S. Steidl, Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech. Berlin: Logos Verlag, 2009, (PhD thesis, FAU Erlangen-Nuremberg).
- [9] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech," in *Proc.* ACM MM, Mountain View, CA, USA, Oct. 2017, pp. 478–484.
- [10] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.
- [11] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97 515–97 525, 2019.
- [12] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 09, pp. 10745–10759, 2023.
- [13] P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak, and N. Boufaden, "Cepstral and long-term features for emotion recognition," in *Proc. INTERSPEECH*, Brighton, UK, 2009, pp. 344–347.
- [14] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.
- [15] M. Kockmann, L. Burget, and J. Cernockỳ, "Brno university of technology system for interspeech 2009 emotion challenge.," in *Proc. INTERSPEECH*, Brighton, UK, 2009, pp. 348–351.
- [16] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9-10, pp. 1062–1087, 2011.

- [17] A. Triantafyllopoulos, S. Liu, and B. W. Schuller, "Deep speaker conditioning for speech emotion recognition," in *Proc. ICME*, Shenzhen, China, 2021, pp. 1–6.
- [18] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions* on affective computing, vol. 7, no. 2, pp. 190–202, 2015.
- [19] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM MM*, Firenze, Italy, 2010, pp. 1459–1462.
- [20] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. INTERSPEECH*, Brno, Czech Republic, 2021, pp. 571–575.
- [21] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. ICASSP*, Calgary, Alberta, Canada, 2018, pp. 5089–5093.
- [22] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical signal processing* and control, vol. 47, pp. 312–323, 2019.
- [23] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. INTERSPEECH*, Shanghai, China, 2020, pp. 3830–3834.
- [24] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.
- [25] A. Triantafyllopoulos and B. W. Schuller, "The role of task and acoustic similarity in audio transfer learning: Insights from the speech emotion recognition case," in *Proc. ICASSP*, Toronto, Canada, 2021, pp. 7268–7272.
- [26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, Hawaii, USA, 2023, pp. 28492– 28518
- [27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS*, vol. 33, pp. 12449–12460, 2020.
- [28] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Pro*cessing, vol. 29, pp. 3451–3460, 2021.
- [29] A. Triantafyllopoulos, J. Wagner, H. Wierstorf, M. Schmitt, U. Reichel, F. Eyben, F. Burkhardt, and B. W. Schuller, "Probing speech emotion recognition transformers for linguistic knowledge," in *Proc. INTERSPEECH*, Seoul, South Korea, 2022, pp. 146–150.
- [30] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc. ACL*, Bangkok, Thailand, 2021, pp. 993–1003.
- [31] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Proc. INTERSPEECH*, Brno, Czech Republic, 2021, pp. 721–725.
- [32] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al., "Underspecification presents challenges for credibility in modern machine learning," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 10 237–10 297, 2022.
- [33] G. Hacohen and D. Weinshall, "On the power of curriculum learning in training deep networks," in *Proc. ICML*, Long Beach, CA, USA, 2019, pp. 2535–2544.