



Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap

Johannes Wagner, Andreas Triantafyllopoulos , *Student Member, IEEE*, Hagen Wierstorf, Maximilian Schmitt ,
Felix Burkhardt , Florian Eyben, and Björn W. Schuller , *Fellow, IEEE*

Abstract—Recent advances in transformer-based architectures have shown promise in several machine learning tasks. In the audio domain, such architectures have been successfully utilised in the field of speech emotion recognition (SER). However, existing works have not evaluated the influence of *model size* and *pre-training data* on downstream performance, and have shown limited attention to *generalisation*, *robustness*, *fairness*, and *efficiency*. The present contribution conducts a thorough analysis of these aspects on several pre-trained variants of wav2vec 2.0 and HuBERT that we fine-tuned on the dimensions arousal, dominance, and valence of MSP-Podcast, while additionally using IEMOCAP and MOSI to test cross-corpus generalisation. To the best of our knowledge, we obtain the top performance for valence prediction without use of explicit linguistic information, with a concordance correlation coefficient (CCC) of .638 on MSP-Podcast. Our investigations reveal that transformer-based architectures are more robust compared to a CNN-based baseline and fair with respect to gender groups, but not towards individual speakers. Finally, we show that their success on valence is based on implicit linguistic information, which explains why they perform on-par with recent multimodal approaches that explicitly utilise textual information. To make our findings reproducible, we release the best performing model to the community.

Index Terms—Affective computing, speech emotion recognition, transformers.

I. INTRODUCTION

AUTOMATIC speech emotion recognition (SER) is a key enabling technology for facilitating better human-to-machine interactions [1]. SER research is dominated by two conceptual paradigms: *discrete emotions* [2] and *emotional dimensions* [3]. The first investigates emotional categories like *happy* or *sad*, while the latter focuses on the dimensions of *arousal*, *valence*, and *dominance* [3].

Manuscript received 20 May 2022; revised 23 January 2023; accepted 26 March 2023. Date of publication 31 March 2023; date of current version 4 August 2023. This work was supported by DFG's Reinhart Koselleck Project under Grant 442218748 (AUDIONOMOU). Recommended for acceptance by K. Livescu. (*Corresponding author: Andreas Triantafyllopoulos.*)

Johannes Wagner, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, and Florian Eyben are with the audeERING GmbH, 82205 Gilching, Germany (e-mail: jwagner@audeerling.com; hwierstorf@audeerling.com; mschmitt@audeerling.com; fburkhardt@audeerling.com; fe@audeerling.com).

Andreas Triantafyllopoulos is with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany (e-mail: andreas.triantafyllopoulos@uni-a.de).

Björn W. Schuller is with the audeERING GmbH, 82205 Gilching, Germany, and with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany, and also with the GLAM – the Group on Language, Audio, and Music, Imperial College, SW7 2BX London, U.K. (e-mail: schuller@informatik.uni-augsburg.de).

Digital Object Identifier 10.1109/TPAMI.2023.3263585

A SER system achieves this through the linguistic (*what* has been said) or the paralinguistic (*how* it has been said) stream [1], [4], [5]. The linguistic stream is better suited for valence recognition [6], [7] and can draw from recent advances in automatic speech recognition (ASR) and natural language processing (NLP) [8], but might be limited to a single language. Paralinguistics works better for arousal and dominance [6], [7] and has the potential to generalise across different languages. Both paradigms can be combined in *bimodal* architectures [4], which require to execute several different models. Instead, we aim towards a model that only implicitly utilises the linguistic information stream during deployment, and does not require access to ASR and NLP frontends.

Although the field has seen tremendous progress in the last decades [1], three major challenges remain for real-world paralinguistics-based SER applications: a) improving on its inferior valence performance [7], [9], b) overcoming issues of generalisation and robustness [10], [11], and c) alleviating individual- and group-level fairness concerns, which is a pre-requisite for ethical emotion recognition technology [12], [13]. Previous works have attempted to tackle these issues in isolation, but combining them is not straightforward.

In recent years, the artificial intelligence (AI) field is undergoing a major paradigm shift, moving from specialised architectures trained for a given task to general-purpose *foundation models* that can be adapted to several use-cases [14]. Such models have seen tremendous success in computer vision [15], [16], NLP [17], and computer audition [18], [19], including SER [20], [21]. Among others, wav2vec 2.0 [18] and HuBERT [19] have emerged as foundation model candidates for speech-related applications. We evaluate several publicly-available pre-trained variants of those models for dimensional SER, and show that they can achieve state-of-the-art results for valence. We further analyze the influence of the model architecture, the pre-training data, how well the models generalise, their robustness, fairness, and efficiency. Moreover, we make our best performing model publicly available [22]. To our best knowledge this is the first transformer-based dimensional SER model released to the community. For an introduction on how to use it, please visit: <https://github.com/audeerling/w2v2-how-to>.

The remainder of this paper is organised as follows. Section II discusses related work, Section III presents the models, databases, and evaluation methods. Section III-D shows the results and investigates why transformer models are able to close the valence gap and improve performance with respect

TABLE I

STATE-OF-THE-ART 4-CLASS EMOTION RECOGNITION PERFORMANCE ON IEMOCAP USING TRANSFORMER-BASED ARCHITECTURES RANKED BY UNWEIGHTED AVERAGE RECALL (UAR) / WEIGHTED AVERAGE RECALL (WAR). THE TABLE ENCODES WHETHER THE BASE (B) OR LARGE (L) ARCHITECTURE WAS USED AS WELL AS WHETHER THE PRE-TRAINED MODEL WAS FINE-TUNED FOR SPEECH RECOGNITION (FT-SR). THE COLUMN FT-D MARKS IF THE TRANSFORMER LAYERS WERE FURTHER FINE-TUNED DURING THE DOWN-STREAM CLASSIFICATION TASK

| | Work | Model | L | FT-SR | FT-D | UAR | WAR |
|----|-------|----------|---|-------|------|------|------|
| 1 | [23] | w2v2-L | ✓ | | | 60.0 | |
| 2 | [24]* | w2v2-L | ✓ | | | 62.5 | 62.6 |
| 3 | [20] | w2v2-b | | | | | 63.4 |
| 4 | [25] | w2v2-b | | | | 63.4 | |
| 5 | [26] | w2v2-b | | ✓ | | 63.8 | |
| 6 | [20] | hubert-b | | | | | 64.9 |
| 7 | [25] | hubert-b | | | | 64.9 | |
| 8 | [20] | w2v2-L | ✓ | | | | 65.6 |
| 9 | [25] | w2v2-L | ✓ | | | 65.6 | |
| 10 | [26] | w2v2-b | | | | 67.2 | |
| 11 | [20] | hubert-L | ✓ | | | | 67.6 |
| 12 | [25] | hubert-L | ✓ | | | 67.6 | |
| 13 | [27] | w2v2-b | | | ✓ | 69.9 | |
| 14 | [28] | w2v2-L | ✓ | | | 70.7 | |
| 15 | [20] | w2v2-b | | ✓ | ✓ | | 73.8 |
| 16 | [27] | w2v2-b | | | ✓ | 74.3 | |
| 17 | [20] | hubert-b | | | ✓ | | 76.6 |
| 18 | [20] | w2v2-L | ✓ | ✓ | ✓ | | 76.8 |
| 19 | [20] | w2v2-b | | | ✓ | | 77.0 |
| 20 | [20] | w2v2-L | ✓ | | ✓ | | 77.5 |
| 21 | [20] | hubert-L | ✓ | ✓ | ✓ | | 79.0 |
| 22 | [20] | hubert-L | ✓ | | ✓ | | 79.6 |

* For a fair comparison we report the result on the utterance level. Authors report better performance on the phonetic level.

to robustness and fairness. Section V investigates efficiency improvements, before Section VI summarises the results, and Section VII concludes the paper.

II. RELATED WORK

The focus of our work is the recognition of emotional dimensions. However, most related studies target emotional categories. Since the approaches are closely related, we consider both in this section.

In Table I, we provide a summary of recent works based on wav2vec 2.0 and HuBERT on the IEMOCAP dataset [29], on which most prior works have focused. Results are ranked by unweighted average recall (UAR) / weighted average recall (WAR) on the four emotional categories of anger (1103 utterances), happiness (+ excitement) (1636), sadness (1084), and neutral (1708), which is the typical categorical SER formulation for IEMOCAP. Most of the works apply leave-one-session-out cross validation (5 folds), except [24], using leave-one-speaker-out cross validation (10 folds), and [20], who do not explicitly mention which folds they used. Even though authors have used different head architectures and training procedures in their studies, we can draw some general observations:

- 1) Fine-tuning pre-trained weights yields a 10% boost.
- 2) Additional ASR fine-tuning does not help with SER (e. g. row 15 vs row 19 -3.2%).

- 3) The large architecture is typically better than the base one (e. g. row 17 vs row 22 $+3.0\%$), but differences can be quite small (e. g. row 19 vs row 20 $+5\%$).
- 4) HuBERT outperforms wav2vec 2.0 (e. g. row 22 vs row 20: $+2.1\%$).
- 5) When performing a fine-tuning of the transformer layers, a simple average pooling in combination with a linear classifier built over wav2vec 2.0 or HuBERT as proposed by [20] seems sufficient and shows best performance in the ranking. However, some of the more complex models like the cross-representation encoder-decoder model proposed by [28] only report results without fine-tuning the pre-trained model during the down-stream task.

While the aforementioned studies have focused on emotional categories, there also exist several ones which concentrate on dimensions. The most comparable to ours is that of [30], who fine-tuned wav2vec 2.0 / HuBERT on arousal, dominance, and valence. Their results show that pre-trained models are particularly good in predicting valence. When additionally joining audio embeddings from the fine-tuned models and text representations obtained with a pre-trained BERT model, they got a concordance correlation coefficient (CCC) for valence of. 683 on the MSP-Podcast corpus [31]. Furthermore, they were able to distill the multi-model system to an audio-only model using student-teacher transfer learning, while still reaching a concordance correlation coefficient (CCC) of. 627 (a massive improvement compared to the previous state-of-the-art performance of only. 377 [32]). However, this improvement was the result of cross-modal transfer learning, and it remains unclear whether speech-based architectures are by themselves able to reach such performance level – a fact we further explore in our work.

The presented results demonstrate the great potential of wav2vec 2.0 and HuBERT for emotion recognition. However, the influence of pre-training data quantity and domain remains unclear. For instance, even though the large model shows consistently better performance, it is unclear if that can be attributed to the additional layers or to an 60 fold increase of training data compared to the base model. Likewise there is little understanding on the impact of language, as previous work focused in pre-training on English speech data. In this contribution, we present a systematic comparison of different models pre-trained under various conditions (e. g. including noisy speech) and evaluate them on several datasets (in-domain and cross-corpus).

Moreover, it is important to show that SER models work well under noisy conditions. [10], [11], [33], [34] have shown that previous SER models suffer from robustness issues. We systematically investigate robustness of transformer-based models against a variety of augmentations that do not change the human perception of the underlying emotion [33].

Finally, we consider fairness an important, but challenging topic for machine learning models. Discussions in the speech processing community focus mainly on group fairness, e. g. gender [35]. For SER models, only a few evaluations are available. [36] found a decrease in CCC for females compared to males for arousal in MSP-Podcast (v1.3) of. 234. Besides group fairness, this contribution investigates individual fairness by estimating

TABLE II

TRANSFORMER-BASED MODELS INCLUDED IN THIS STUDY AND DETAILS ON THE DATA USED DURING PRE-TRAINING. MODELS COMPRISED OF TWO ARCHITECTURE DESIGNS (WAV2VEC 2.0 AND HUBERT), EACH WITH TWO DIFFERENT VARIANTS (BASE AND LARGE). FOR EACH MODEL, WE LIST INCLUDED DATASET(S), TOTAL NUMBER OF HOURS (H), NUMBER OF LANGUAGES (ENG IF ONLY ENGLISH), AND COVERED DOMAINS (READ SPEECH, TELEPHONE CONVERSATIONS, PARLIAMENTARY SPEECH, YOUTUBE)

| Model | Datasets | h | Lang | Domain |
|--------------------|---|------|------|------------|
| w2v2-b [18] | LibriSpeech | 960 | eng | R |
| hubert-b [19] | LibriSpeech | 960 | eng | R |
| w2v2-L [18] | Libri-Light | 60k | eng | R |
| hubert-L [19] | Libri-Light | 60k | eng | R |
| w2v2-L-robust [38] | Libri-Light (60k) Fisher (2k) CommonVoice (700) Switchboard (300) | 63k | eng | R, T |
| w2v2-L-vox [39] | VoxPopuli | 100k | 23 | P |
| w2v2-L-xls-r [40] | VoxPopuli (372k) ML LibriSpeech (50k) CommonVoice (7k) VoxLingua107 (6.6k) BABEL (1k) | 436k | 128 | R, T, P, Y |

the influence of the speaker on the model performance, which is a known problem for speaker verification models [37].

III. EXPERIMENTAL SETUP

A. Pre-Trained Models

Throughout the paper, we discuss results obtained with transformer-based models pre-trained on large amounts of unlabelled data. We investigate two main variants: wav2vec 2.0 [18] and HuBERT [19]. The network architecture of both models is the same. As input, it expects a raw waveform normalised to have zero mean and unit variance, which is fed into a *feature encoder* consisting of 7 convolutional layers that extracts *feature vectors* over time, with a dimensionality of 512 and a step size of 20 ms. These features are projected to a higher dimension (768 or 1024 hidden units, see below) and then fed into the *encoder*. The *encoder* is a series of *transformer layers*, each of them consisting of a *multi-head self-attention module* and several *fully-connected layers*. In order to inject temporal information, the output of a *convolutional layer* is added at the input of the *encoder*.

The only difference between the main variants is the way they are pre-trained on unlabelled data. In wav2vec 2.0, the features of a certain ratio of time steps are masked, by replacing them with a learnt fixed feature vector at the input of the *encoder*. A *contrastive loss* between the *encoder outputs* and a *quantised version* of the input features is then minimised [18]. In order to avoid learning too simple representations, the quantisation is done using a codebook, whose *diversity loss* is minimised as well. In contrast, HuBERT minimises a *cross-entropy loss* for the masked time steps, where the targets are not trained simultaneously with the model. The pre-training is performed in several steps, where in the first step, clusters obtained by *k-means clustering* of MFCCs are employed as targets and in later steps, clusters of the outputs of certain transformer layers are taken into account [19]. In following these strategies, the models try

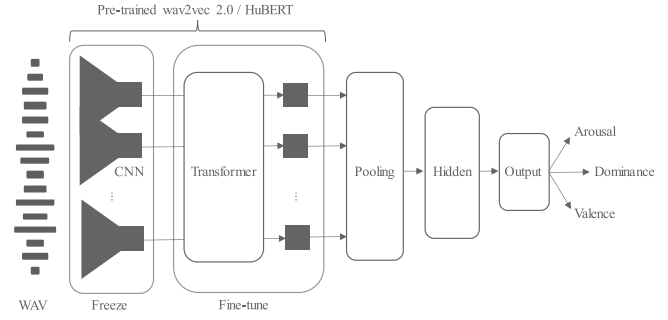


Fig. 1. Proposed architecture built on wav2vec 2.0 / HuBERT.

to learn the *structure* of speech, resulting in a reduced need for labelled task-specific training data.

Both wav2vec 2.0 and HuBERT exist in two forms: a base architecture with 12 transformer layers of 768 hidden units each (95 M parameters), and a large architecture with 24 transformer layers of 1024 hidden units each (317 M parameters). Apart from that, we further distinguish them by the data used for pre-training. We included the four models found in previous work (cf. Section II), which are pre-trained on English audiobooks, namely wav2vec2-base (w2v2-b), hubert-base-ls960 (*hubert-b*), wav2vec2-large (w2v2-L), hubert-large-ls60 k (*hubert-L*); the wav2vec2-large-robust model (w2v2-L-robust), additionally trained on telephone speech; the wav2vec2-large-100k-voxpupuli model (w2v2-L-vox), trained only on parliamentary speech in multiple languages; and the wav2vec2-xls-r-300 m model (w2v2-L-xls-r), trained on more than 400 k hours across all domains and multiple languages. Compare Table II for citations and an overview of the included data. We did not include models fine-tuned on speech recognition as previous work showed that this does not lead to better performance. Also note that we refer to their fine-tuned versions when we report results (cf. Section III-B).

B. Architecture

Inspired by [20] we apply average pooling over the hidden states of the last transformer layer and feed the result through a hidden layer and a final output layer (see Fig. 1). For fine-tuning on the downstream task, we use the ADAM optimiser with CCC loss, which is the standard loss function used for dimensional SER [9], [32], [41], and a fixed learning rate of $1e-4$. We run for 5 epochs with a batch size of 32 and keep the checkpoint with best performance on the development set.

During training, we freeze the CNN layers but fine-tune the transformer ones. According to [20], such a partial fine-tuning yields better results. When using the term fine-tuning, we will henceforth refer to this partial fine-tuning. These models are trained using a single random seed, for which the performance is reported.

We compare results to a 14-layer Convolutional Neural Network (CNN14) as a standard baseline we have been using for SER in previous work [9], [42]. It follows the architecture proposed by [43] for audio pattern recognition. Different to the transformer-based models, which operate on the raw audio

signal, this takes log-Mel spectrograms as input. CNN14 has 6 convolutional blocks with two layers each, each followed by max pooling. Convolution layers have a 3×3 kernel and a stride of 1×1 , whereas max pooling layers use a stride of 2×2 . After the last convolution layer, features are pooled using both mean and max pooling, and subsequently fed into two linear layers. Dropout with a probability of 0.2 is applied after every each convolution block. Log-Mel spectrograms are computed with 64 Mel bins, a window size of 32 ms, and a hop size of 10 ms. Note that the *CNN14* model is not pre-trained, i. e. it is always trained from scratch in our experiments. We train for 60 epochs, with a learning rate of .01, and a batch size of 64 using stochastic gradient descent (SGD) with a Nesterov momentum of .9. We select the model that performs best on the validation set.

C. Datasets

We used the *MSP-Podcast* corpus [31] (v1.7) to run multitask training on the three dimensions of arousal, dominance, and valence for speech from podcast recordings. The original labels cover a range from 1 to 7, which we map into the interval of 0 to 1. Its *train* split contains 62 hours of recordings. In-domain results are reported on the *test-1* split, which contains 21 hours of audio provided by 12,902 samples (54% female / 46% male) from 60 speakers (30 female / 30 male). The samples per speaker vary between 42 and 912.

We report cross-domain results *IEMOCAP* (Interactive Emotional Dyadic Motion Capture) dataset [29], which contains 12 hours of scripted and improvised dialogues by ten speakers (5 female / 5 male). It provides the same dimensional labels as *MSP-Podcast*, but in a range of 1 to 5, which we map to the interval 0 to 1. Since we use the dataset only during evaluation, we do not apply the usual speaker cross-validation, but treat the corpus as a whole. It includes 10,039 samples (49% female / 51% male).

Finally, we report cross-corpus results for valence on the test set of the Multimodal Opinion Sentiment Intensity (*MOSI*) [44] corpus. The dataset is a collection of YouTube movie review videos spoken by 41 female and 48 male speakers. They are annotated for sentiment on a 7-point Likert scale ranging from -3 to 3 , which we map to the interval 0 to 1. The test set contains 1 h audio recordings given as 685 samples (51% female / 49% male), annotated for sentiment. As the gender labels are not part of the distributed database, we re-annotated them ourselves [45].

While sentiment is a different concept than valence, as the former corresponds to an attitude held towards a specific object and the latter more generally characterises a person's feeling [46], there is evidence that sentiment annotations can be decomposed to two constituents: intensity and polarity [47], which roughly correspond to arousal and valence. We therefore expect some correlation between (predicted) valence and (annotated) sentiment scores.

D. Evaluation

Machine learning models for speech emotion recognition are expected to work under different acoustic conditions and for different speakers. To cover this, we evaluate them for correctness, robustness, and fairness [48].

Correctness measures how well predictions match the ground truth. The concordance correlation coefficient (CCC) provides an estimate of how well the predicted distribution matches the ground truth one [49], and is the typical measure for evaluating dimensional SER models [50].

Robustness (cf. Section IV-H) measures how model performance is affected by changes to the input signals such as adding background noise. Applying changes to the input signals must be carefully done for SER, as they might affect the ground truth label [33], [51]. We focus on testing the robustness of the models against data augmentations that do not change the human perception of the underlying emotion. We select the following five augmentations from [33] to enable direct comparison with previous results: *Natural soundscapes* adds a randomly selected sample from the natural class of the ESC-50 dataset [52] with a signal-to-noise ratio (SNR) of 0 dB, 10 dB or 20 dB; *Human, non-speech* adds a randomly selected sample from the human class of the ESC-50 dataset with a SNR of 0 dB, 10 dB or 20 dB; *Interior/domestic* adds a randomly selected sample from the interior class of the ESC-50 dataset with a SNR of 0 dB, 10 dB or 20 dB; *Speed up segment* selects a random segment of 10% to 20% length within the utterance and increases its speed by 1.25; *Fade-in/fade-out* decreases or increases the amplitude of the signal by 2% every second.

Fairness (cf. Section IV-I) evaluates if the model predictions show biases for certain protected attributes like race, gender, or age [53]. We focus on gender due to the lack of sufficient available information and/or datasets for other attributes. For regression problems, there is no clear definition how to measure fairness, but most approaches try to achieve an equal average expected outcome for population A and B [54]. We measure fairness by estimating the gender fairness score as the difference in the correctness metric (CCC) between female and male groups. A positive gender fairness score indicates a better performance of the model for female speakers.

IV. EVALUATION

We begin our investigation with a thorough evaluation of transformer-based models. We show that valence is the primary beneficiary of pre-training as it enables the models to implicitly learn linguistic information during the fine-tuning of the transformer layers. Utilising a comprehensive testing scheme, we attempt to identify how different aspects of foundation models impact performance and generalisation. We place particular emphasis on *robustness* and *fairness*, which are critical considerations for SER systems targeted to real-world applications.

A. Can Foundation Models Close the Performance Gap for Valence?

Answer: The best models achieve a similar performance for arousal and dominance as non-transformer architectures [32], but improve the CCC score for valence by .26 and close the performance gap for valence.

Details: In Fig. 2, we show in-domain and cross-domain CCC performance for different wav2vec 2.0 and HuBERT models as well as for the *CNN14* baseline.

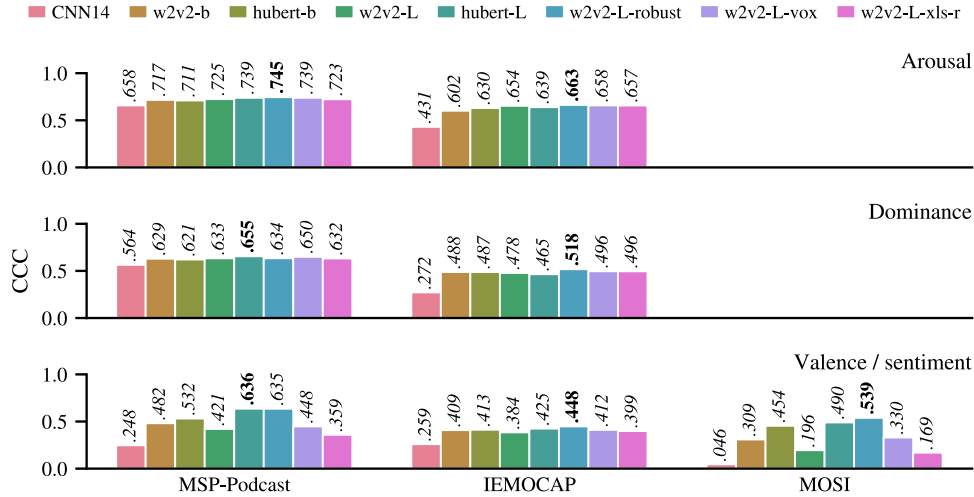


Fig. 2. CCC scores for arousal, dominance, valence (MSP-Podcast / IEMOCAP), and sentiment (MOSI). All models have been trained for emotional dimension prediction using multitasking only on MSP-Podcast, and subsequently evaluated on its test set (in-domain), as well as to the test set of MOSI and the entire IEMOCAP dataset (cross-corpus).

We first focus on arousal and dominance. For MSP-Podcast (in-domain) and IEMOCAP (cross-domain), all transformer-based models score very similar, with *w2v2-L-robust* showing the overall best performance by reaching a CCC score of .745/.634 (arousal/dominance) on MSP-Podcast, and .663/.518 on IEMOCAP. For MSP-Podcast, results are similar compared to the CCC scores of .745/.655 achieved by [32] and .757/.671 achieved by [30].

For valence, we see a larger fluctuation of CCC scores for different transformer models ranging from .359 for *w2v2-L-xls-r* to .636 for *hubert-b*, both on MSP-Podcast. Overall, *w2v2-L-robust* shows again the best overall performance by reaching a CCC score of .635 on MSP-Podcast, .448 on IEMOCAP, and .539 for predicting sentiment on MOSI. For MSP-Podcast, results are better compared to the CCC score of .377 achieved by [32] and similar to .627 by [30] achieved with a model distilled from an audio + text based teacher.

B. Does Explicit Linguistic Information Further Improve Performance?

Answer: Adding linguistic information does not improve predictions for arousal and dominance, and only in some cases for valence. However, especially models pre-trained on multiple languages seem to benefit when tested on English speech.

Details: To evaluate whether adding linguistic information improves the predictions, the following experiment is conducted: a regression head is pre-trained, using as input pooled BERT embeddings in addition to the pooled states of the fine-tuned transformer models.

BERT (Bidirectional Encoder Representations from Transformers) is a transformer model for natural language, pre-trained on English language corpora consisting of more than 3 billion words [55]. The BERT embeddings have a dimensionality of 768 and are extracted from the transcriptions generated by the

wav2vec2-base-960h speech recognition model¹. The fusion is done by concatenating the representations of both modalities. As regression head, exactly the same architecture as for the fine-tuning of *wav2vec 2.0* and *HuBERT* models is employed. For training, the weights of both models are frozen. The training is done with multi-target CCC-loss for a maximum of 100 epochs, with early stopping based on CCC development set performance.

In Fig. 3, we report deviations from the results achieved with the fine-tuned acoustic models alone (cf. Fig. 2). We can see that a fusion with embeddings from the text domain helps with valence, but not with arousal and dominance, where performance actually deteriorates. This is in line with our previous findings, where we also found that introducing linguistic information sometimes hampered performance for those two dimensions on MSP-Podcast [9]. What is interesting, though, are the relatively large differences between the models, and that, especially, our best models *hubert-L* and *w2v2-L-robust* do not improve. The models that benefit most are the two multi-lingual models *w2v2-L-vox* and *w2v2-L-xls-r*, showing that models pre-trained on multiple languages gain from a fusion with text features from the test set domain language.

C. Do the Models Implicitly Learn Linguistic Information?

Answer: The models implicitly capture linguistic information from the audio signal. The extent in which they learn sentiment during fine-tuning depends on the data used for pre-training (e. g. multi-lingual data makes it more difficult). Generally, we see that valence performance correlates with a model's ability to predict sentiment.

Details: Previous findings suggest that during fine-tuning, the models implicitly learn linguistic information. To assess

¹<https://huggingface.co/facebook/wav2vec2-base-960h>

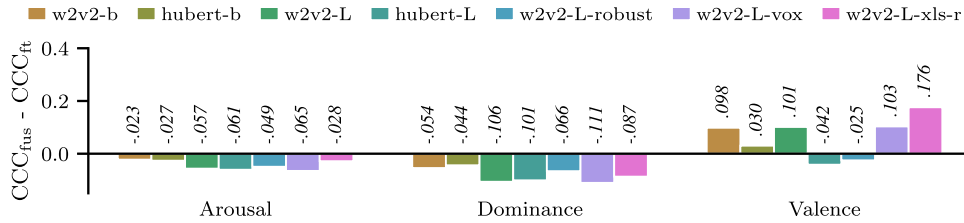


Fig. 3. Text and audio fusion results for arousal, dominance, and valence prediction on MSP-Podcast. Embeddings from the already fine-tuned models are concatenated with BERT embeddings extracted from automatic transcriptions, whereupon a two-layer feed-forward neural network is trained. We show the difference to results with the fine-tuned (ft) models from Fig. 2.

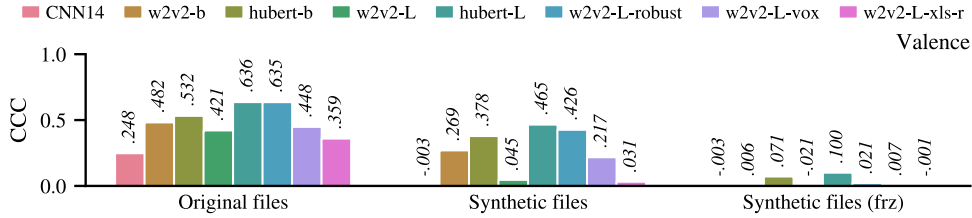


Fig. 4. CCC performance for valence on the original and synthetic files on MSP-Podcast. We see that models with a high performance on the original files are more sensitive to sentiment (cf. left and center section). To prove that a fine-tuning of the transformer layers is required to learn linguistic content, we additionally show the correlation for models where the transformer layers were frozen (frz) during training (cf. Section IV-D).

how sensitive the models are to linguistic content, we generated a synthesised version of a subset of the test set from the transcriptions of MSP-Podcast.² In Fig. 4, we finally show CCC performance for valence on the original and synthesised files for all models. We see that performance gaps between the models in Fig. 2 are directly linked with their ability to predict sentiment. Models reaching a high performance on the original files also do so on their synthetic versions and vice versa. However, to learn linguistic content, a fine-tuning of the transformer layers is essential. If we predict the synthetic test set with models where the transformer layers were frozen during training (cf. Section IV-D), correlation drops to almost zero.

This finding is also important for works doing in-domain training on IEMOCAP, as parts of the conversations are scripted which results in a leakage of text information that may result in overoptimistic results [56] when that text information is exploited by transformer models. Furthermore, our models may inherit similar biases as those found in NLP models [57].

D. How Important is a Fine-Tuning of the Transformer Layers?

Answer: Fine-tuning the transformer layers is necessary to obtain state-of-the-art performance, in particular for the valence dimension. The highest gain is observed for *hubert-L* and *w2v2-L-robust*, which are the models that benefit the least from a fusion with text.

Details: So far, we have fine-tuned all transformer layers along with the added output layer. However, practitioners often choose to use a pre-trained model as a frozen feature extractor,

²Partial audio transcripts are available with MSP-Podcast v1.9 and cover 55% of the *test-1* split from v1.7 we used for our experiments.

and subsequently train just an output layer on the generated embeddings. Nevertheless, prior studies have shown that it is necessary to fine-tune several or all layers on the target task to get good downstream performance [20], [42], [43], [58]. In this sub-section, we experiment with training only the last output layer and keeping all others frozen. This is compared to our previous experiments where we jointly fine-tune the last layer and the transformer layers.

Fig. 5 shows the difference between CCC values for the fine-tuned and frozen models. We observe performance gains when fine-tuning in all cases, demonstrating that fine-tuning of the transformer layers is necessary. Moreover, the models that see the biggest performance gain are *hubert-L* and *w2v2-L-robust*. In Section IV-B, these models were found to benefit less from additional text information. These findings indicate that a fine-tuning of the transformer layers enables the models to capture the linguistic information needed to perform well on valence.

E. Do the Models Generalise Better Across Different Domains?

Answer: Transformer-based models generalise better than a non-transformer baseline.

Details: As we see a similar trend for different transformer-based models between in-domain and cross-corpus results in Fig. 2, we focus on the best-performing one (*w2v2-L-robust*). The drop in CCC between in-domain and cross-corpus results for *w2v2-L-robust* on IEMOCAP is 11% for arousal, 21% for dominance, and 30% for valence on IEMOCAP, and 15% for sentiment on MOSI. For *CNN14*, the drop in CCC is 34% for arousal, and 52% for dominance, while for valence, we do not estimate the drop in cross-domain performance as the in-domain

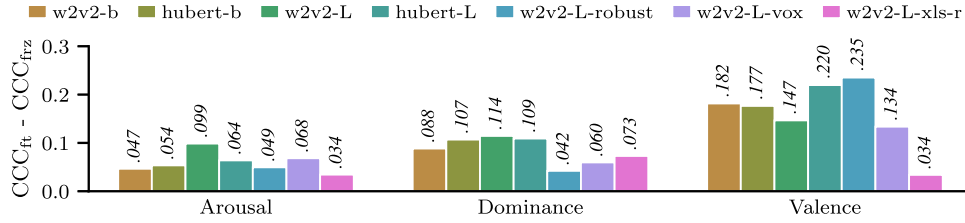


Fig. 5. Difference of fine-tuned (ft) to frozen (frz) CCC performance for arousal, dominance, and valence prediction on MSP-Podcast. The fine-tuned results are from Fig. 2, where transformer and output layers are jointly trained. For the frozen results, we keep all transformer layers frozen and simply train the output head. Results show that fine-tuning the transformer layer is worth the computational cost it incurs.

CCC is already very low. The drop in CCC is smaller for *w2v2-L-robust* for arousal and dominance, indicating that transformer-based models generalise better. For valence, we cannot derive a final conclusion, but the trend we see for sentiment in MOSI seems very promising.

F. Does More Data During Pre-Training Lead to Better Performance?

Answer: For arousal and dominance, all tested models perform equally well, whereas with respect to valence / sentiment the data used for pre-training has a strong effect. Mixing data from several domains leads to a considerable improvement for *w2v2-L-robust* compared to *w2v2-L*, which is only trained on clean speech. However, *hubert-L*, which uses the same pre-training data as *w2v2-L*, still performs as good as *w2v2-L-robust*. For models pre-trained on multi-lingual data, we see a performance drop when tested on English speech.

Details: To understand what influence the size and domain of the pre-training data have on downstream performance, we included several wav2vec 2.0 models with same large architecture but different pre-training (see Table II).

The results in Fig. 2 show only differences in terms of CCC between the transformer models for valence and sentiment, not for arousal or dominance. Previous studies uniformly report that HuBERT outperforms wav2vec 2.0 which is replicated by our results with *w2v2-L* showing a smaller CCC than *hubert-L* for the valence task on MSP-Podcast and IEMOCAP, and for the sentiment task on MOSI. The increase in performance for *w2v2-L-robust* is therefore most likely explained by the additional 3 k hours of telephone conversations used for pre-training. However, by comparing *w2v2-L-vox* and *w2v2-L-xls-r*, it also becomes clear that more data does not necessarily lead to better results. Though both models are trained on significantly more data than *hubert-L* and *w2v2-L-robust* (100 k and 463 k vs 63 k hours), they perform clearly worse. Notably, both were pre-trained on multiple languages. Since the databases we use for evaluation contain only English speakers, this could be a disadvantage to models that are exclusively pre-trained on English.

G. Does a Larger Architecture Lead to Better Performance?

Answer: A larger architecture does not lead to better performance per se. Larger architectures using different data during pre-training might perform worse than smaller architectures.

Details: We cannot directly answer what influence the size of the architecture has on performance, as we do not have transformer models with different architectures pre-trained on the same data in our evaluation (Fig. 2). We can draw some indirect conclusions, though. The size of the architecture, i. e. base vs large, seems not to be the decisive point: the small models *w2v2-b* and *hubert-b* have comparable performance to the large models *w2v2-L*, *w2v2-L-vox*, and *w2v2-L-xls-r* for arousal and dominance, both in- and cross-domain. For valence, the small models outperform *w2v2-L*, *w2v2-L-vox*, and *w2v2-L-xls-r* in most cases for MSP-Podcast and MOSI, and achieve a similar performance on IEMOCAP.

H. Are the Models Robust Against Changes to the Input Signals?

Answer: The tested models are reasonably robust against changes to the input signals, with *w2v2-L-robust* showing the highest and *hubert-b* the lowest robustness.

Details: Fig. 6 summarises the average CCC scores of the models averaged over all augmentations described in Section IV-H. All models show a drop in CCC compared to the CCC scores for the clean data from Fig. 2. *w2v2-L-robust* has now the highest CCC score for all datasets and all dimensions. The average change in CCC for *w2v2-L-robust* is -0.068 . The model with the highest average change in CCC is *hubert-b* (-0.108). The model with the lowest average change in CCC is *CNN14* (-0.047), which is mostly due to its results for IEMOCAP for which it shows no impairment of its relatively low performance by augmentations.

Table III shows changes in CCC for single augmentations for each dataset and dimension for the best performing model *w2v2-L-robust*. The performance of the model is only slightly affected (absolute change in CCC score below .05) for added background sounds with a SNR of 20 dB or a fade-in/fade-out of the signal. When speeding up parts of the signal or adding background sounds with more severe SNRs the change in CCC can be up to $-.278$. The model investigated on the same augmentations by [33] shows an equal drop in unweighted average recall (UAR) when adding background sounds with 0 dB, 10 dB, 20 dB SNR of at least $-.30$. *w2v2-L-robust* is more robust when adding background sounds with a moderate SNR. It shows a drop in CCC of up to $-.28$ for 0 dB SNR, but only a drop in CCC of up to $-.036$ for 20 dB SNR. Whereas the model investigated by [33] is similarly affected by adding *human*,

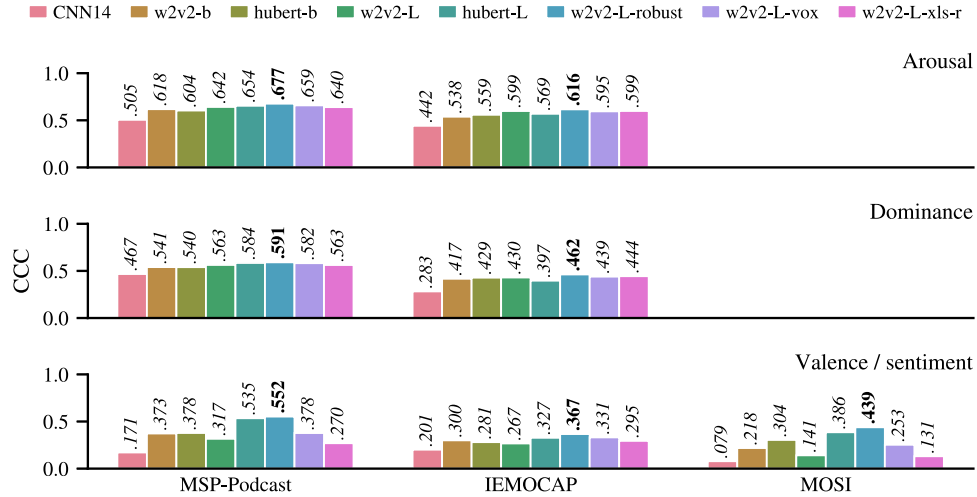


Fig. 6. CCC scores for arousal, dominance, valence (MSP-Podcast/IEMOCAP), and sentiment (MOSI) when augmenting the test data. The scores are averaged over all eleven different augmented versions of the test data.

TABLE III
CHANGE IN CCC FOR w2v2-L-ROBUST PREDICTIONS ON AUGMENTED DATA COMPARED TO ITS PREDICTIONS ON CLEAN DATA

| Augmentation | SNR | Arousal | | Dominance | | Valence / sentiment | | |
|---------------------|-------|-------------|---------|-------------|---------|---------------------|---------|--------|
| | | MSP-Podcast | IEMOCAP | MSP-Podcast | IEMOCAP | MSP-Podcast | IEMOCAP | MOSI |
| Natural soundscapes | 0 dB | -0.145 | -0.109 | -0.081 | -0.119 | -0.221 | -0.183 | -0.210 |
| | 10 dB | -0.038 | -0.018 | +0.001 | -0.035 | -0.051 | -0.064 | -0.071 |
| | 20 dB | -0.019 | -0.009 | +0.012 | -0.018 | -0.003 | -0.015 | -0.025 |
| Human, non-speech | 0 dB | -0.254 | -0.205 | -0.195 | -0.167 | -0.278 | -0.214 | -0.264 |
| | 10 dB | -0.053 | -0.036 | -0.036 | -0.058 | -0.080 | -0.092 | -0.143 |
| | 20 dB | -0.011 | -0.001 | +0.005 | -0.017 | -0.007 | -0.030 | -0.036 |
| Interior/domestic | 0 dB | -0.124 | -0.089 | -0.081 | -0.097 | -0.170 | -0.160 | -0.149 |
| | 10 dB | -0.026 | -0.011 | -0.003 | -0.029 | -0.031 | -0.049 | -0.059 |
| | 20 dB | -0.011 | -0.004 | +0.007 | -0.014 | -0.005 | -0.015 | -0.030 |
| Speed up segment | | -0.080 | -0.030 | -0.099 | -0.056 | -0.062 | -0.069 | -0.113 |
| Fade-in/fade-out | | -0.001 | -0.001 | .000 | -0.001 | .000 | .000 | -0.001 |

non-speech, *interior/domestic*, or *natural sounds* as background sounds, w2v2-L-robust is the most affected when adding *human*, *non-speech* sounds (average drop in CCC of -0.103), and the least when adding *interior/domestic* sounds (average drop in CCC of -0.055).

I. Are the Models Fair Regarding the Gender of the Speaker?

Answer: Models are more fair for arousal and dominance than for valence. For valence, most models show a higher CCC for females than for males.

Details: Fig. 7 shows gender fairness scores for the speakers in MSP-Podcast, IEMOCAP, and MOSI. As introduced in Section III-D, the gender fairness score is expressed by the difference in CCC between female and male speakers with positive values indicating higher values of the underlying metric for females. For MSP-Podcast, nearly all models show a slightly worse female CCC for arousal and dominance. For IEMOCAP, nearly all models show a slightly better female CCC for arousal and dominance.

For valence in MSP-Podcast and IEMOCAP, most models show a better CCC for female speakers than male ones – with the exception of CNN14. For sentiment in MOSI, the CNN14

model shows a bias towards better performance for male speaker, whereas all other models show very small biases in the different direction.

Averaging over all databases and dimensions the model with the best gender fairness score is w2v2-L with .007, followed by w2v2-L-vox with .015, w2v2-L-xls-r with .018, w2v2-L-robust, with .019, hubert-b with .025, hubert-L with .027, and w2v2-b with .029 up to CNN14 with -0.043 .

J. Is Performance Equal Across All Speakers?

Answer: Performance for the best foundation models is similar between most speakers in MSP-Podcast, but can deteriorate to low CCC values for some speakers.

Details: The performance of speech processing is dependent on individual speaker characteristics [37]. This has led several prior SER works to target personalisation to different speakers [59], [60], [61]. To investigate this phenomenon for transformer-based models, we examine the *per-speaker performance*, where instead of computing a global CCC value over all test set values, we compute one for each speaker. As discussed (cf. Section III-C), the MSP-Podcast test set consists of 12902 samples from 60 speakers; however, the samples are not equally

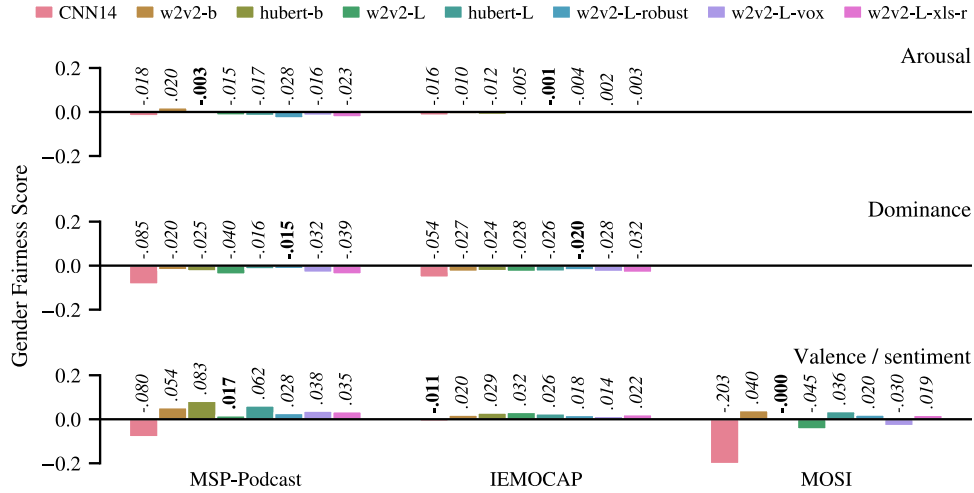


Fig. 7. Gender fairness scores for arousal, dominance, valence (MSP-Podcast / IEMOCAP), and sentiment (MOSI). The gender fairness score is given by $CCC_{female} - CCC_{male}$. A positive value indicates that the model under test performs better for female speaker and a negative value that it performs better for male speaker. A model with desired equal performance would have a gender fairness score of 0.

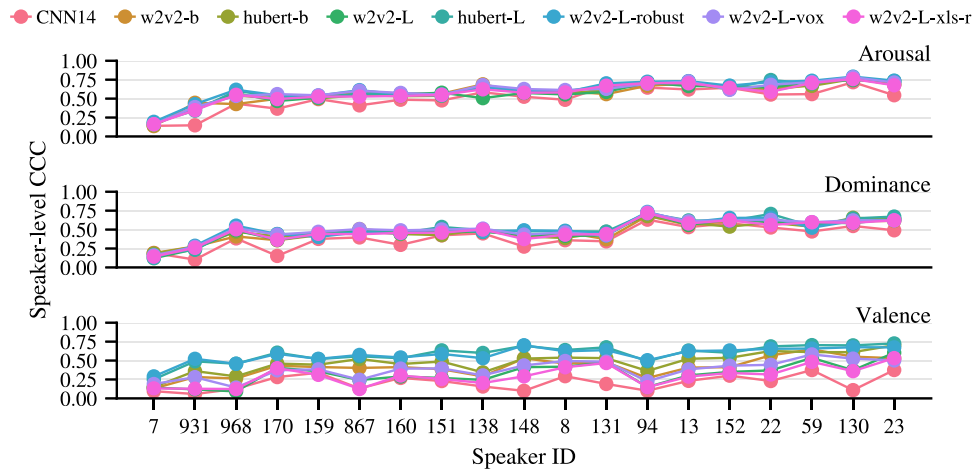


Fig. 8. Speaker-level performance (CCC) on MSP-Podcast for the different models. We only use speakers with at least 200 test set samples for robust CCC estimates. All models show low CCC for at least one speaker on all 3 tasks. Speakers have been ordered according to the mean CCC over all dimensions and models.

distributed across them (minimum samples: 41, maximum samples: 912). In order to make our subsequent analysis more robust, we only keep speakers with more than 200 samples, resulting in 19 speakers. We use bootstrapping, where we randomly sample (with replacement) 200 samples from each speaker to compute the CCC. This process is repeated 1000 times, and we report the mean value.

Our results are presented in Fig. 8. For visualisation purposes, we ordered speakers based on the average CCC value over all models and across arousal, dominance, and valence. For arousal and dominance models perform well for most speakers, and show similar performance. For speakers 7 and 931 all models show a low CCC, whereas for speaker 931 the *CNN14* model performs worse than the others. For valence, CCC values per speaker differ between models replicating the findings of Fig. 2. The best model (*w2v2-L-robust*) performs relatively similar for

most of the speaker groups and shows only a drop for speaker 7, a similar result as for valence and dominance.

Different models broadly, but not perfectly, agree on ‘good’ and ‘bad’ speakers, with pairwise Spearman correlations ranging from .960 to .725 for arousal, .972 to .825 for dominance, and .947 to .333 for valence. This could be a manifestation of the underspecification phenomenon plaguing machine learning architectures [62], as models which have similar performance on the entire test set, nevertheless behave differently across different subsets of it.

K. Why Do Foundation Models Generalise So Well?

Answer: Even without pre-training, the latent space provided by the transformer architecture generalises better than *CNN14*, as it abstracts away domain and speaker. Pre-training marginally

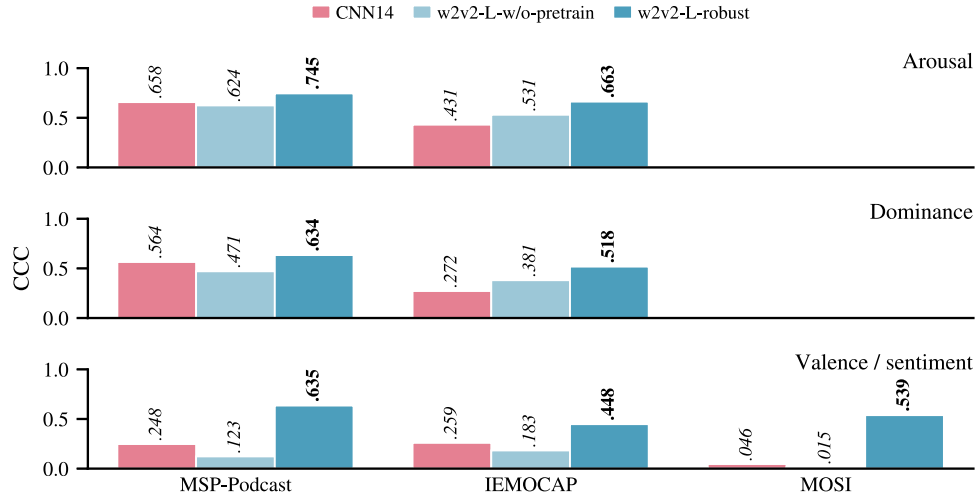


Fig. 9. CCC performance of randomly-initialized wav2vec 2.0 model (*w2v2-L-w/o-pretrain*) on in-domain and cross-corpus arousal, dominance, valence / sentiment prediction. We compare the performance with that of *CNN14* and *w2v2-L-robust*. We observe that valence and sentiment benefit massively from pre-training, without which wav2vec 2.0 performs worse than a classic CNN approach.

improves arousal and dominance performance but is critical for valence.

Details: So far, we were able to confirm the superiority of transformer-based models. However, even though pre-training seems important, it remains unclear to what extent the transformer architecture itself contributes to that success. To shed more light into this, we trained wav2vec 2.0 from a random initialisation. As our architecture, we chose the large wav2vec 2.0 architecture, which is also used by the best performing model *w2v2-L-robust*. In the following, we will refer to this model as *w2v2-L-w/o-pretrain*.

We trained the model for 50 epochs and selected the best checkpoint according to the performance on the development set (epoch 17).³ In Fig. 9, we compare in- and cross-domain performance with *CNN14* and *w2v2-L-robust*. We see that especially valence / sentiment detection benefits massively from pre-training (both in-domain and cross-domain), and that without pre-training wav2vec 2.0 performs in most cases worse than *CNN14*.

In the introduction of wav2vec 2.0, [18] postulate that pre-training helps learn more general representations that abstract away from speaker or background information. However, it is not entirely clear if these benefits are a result of pre-training or are a consequence of the specific inductive biases introduced by the architecture. To investigate this, we compare embeddings extracted with *CNN14*, *w2v2-L-w/o-pretrain*, and *w2v2-L-robust*,⁴ which are shown in Fig. 10. The embeddings are projected to two dimensions using t-SNE [63] and different information is chromatically superimposed.

³Even though we used the same data (MSP-Podcast) for fine-tuning, we expected it would take longer for the model to convert if we start from scratch. Also, this time we trained all encoder layers (including the CNN ones). Apart from that we followed the methodology described in Section III-B.

⁴We use average pooling on the output of the last CNN layer for *CNN14* and the last transformer layer for wav2vec 2.0.

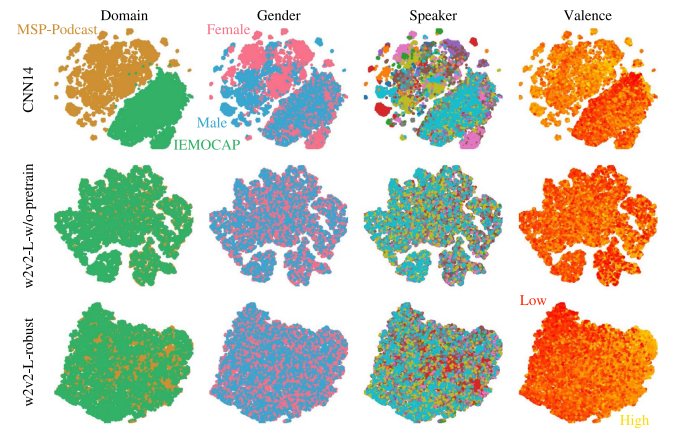


Fig. 10. Visualization of embeddings extracted with different models overlaid with meta information for a combined dataset of MSP-Podcast and IEMOCAP. We observe that the latent space of wav2vec 2.0 offers a better abstraction from domain, gender, and speaker compared to the *CNN14* baseline – even without pre-training. However, only a pre-trained model is able to separate low from high valence. To reduce the dimensionality of the latent space, we applied T-SNE [63].

For *CNN14*, two main clusters almost perfectly separate the two data sources MSP-Podcast and IEMOCAP, whereas several smaller blobs represent gender groups and individual speakers. In fact, speaker and domain are more pronounced than valence information. Hence, similar emotional content can translate into entirely different latent representations. In contrast, the latent space of both wav2vec 2.0 models shows no clusters for domain, gender, or speaker. The architecture itself seems to introduce specific inductive biases which are well-suited to learning robust representations. Nevertheless, only the pre-trained model (*w2v2-L-robust*) shows a smooth transition from low to high valence scores, showing that pre-training is still necessary for good downstream performance. Moreover, the strong speaker dependency presented in Section IV-J shows that the two dimensional t-SNE visualisations help comparing generalisation

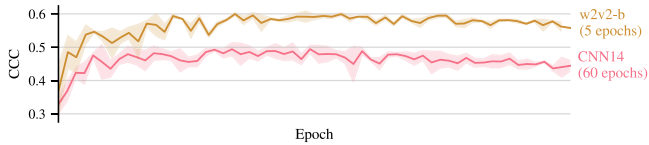


Fig. 11. Mean and standard deviation of development set performance on MSP-Podcast across three training runs. Compared to *CNN14*, *w2v2-b* converges earlier and shows less fluctuation.

abilities between models, but are not necessarily sufficient for deriving conclusions w. r. t. generalisation over different factors.

V. EFFICIENCY

For our last experimental evaluation, we focus on efficiency. We concentrate on three facets: *optimisation stability*, *computational complexity*, and *data efficiency*.

A. Does Pre-Training Help With Training Stability and Convergence?

Answer: A pre-trained model reduces the number of epochs needed to converge and improves performance stability across training runs with different seeds.

Details: To balance the effects of randomness (either in the initialisation of network weights or the data sampling), it is a common strategy to perform several runs with different random seeds. Starting from pre-trained weights, however, we expect less volatility [64], [65]. Fig. 11 shows the mean and standard deviation over the performance on the development set across three trials for *CNN14* and *w2v2-b*. *CNN14* shows a constant jittering across all 60 epochs, whereas *w2v2-b* converges faster and we can reduce the number of epochs to 5.

B. How Many Transformer Layers Do We Really Need?

Answer: We can reduce the number of transformer layers to 12 without a degradation in performance. With less than 12 layers we begin to see a negative effect on valence.

Details: In Section IV-G, we mentioned that *w2v2-b* and *hubert-b* outperform some of the large models. From that, we concluded that the size of the architecture seems less important, but it is rather the data used for pre-training that determines success. If this is really the case, we should be able to partially reduce the size of a model without losing performance.

[66] investigated different layer pruning strategies and identified top-layer dropping as the best strategy offering a good trade-off between accuracy and model size. Inspired by their findings, we set up an experiment where we successively removed transformer layers from the top of the original pre-trained model before fine-tuning. In Fig. 12, we report the effect on CCC for *w2v2-L-robust* (our overall best performing model). Results show that half of the layers can be removed without a loss in performance. We denote the resulting 12-layer model as *w2v2-L-robust-12*. Only with 10 or less layers we actually begin to see a drop for valence / sentiment on IEMOCAP and MOSI. For arousal and dominance, we still achieve good performance with only 8 layers.

C. Can We Reduce the Training Data Without a Loss in Performance?

Answer: A reduction of training samples without loss in performance is only possible for arousal and dominance.

Details: Reducing the amount of training data offers another way to speed up model building. To find out what effect the removal of training samples has, we conducted an experiment where we fine-tuned several versions of the same pre-trained model with different fractions of the training set (MSP-Podcast). We leave development and test set untouched.

Fig. 13 shows CCC for arousal, dominance, valence / sentiment on MSP-Podcast, IEMOCAP and MOSI. For efficiency, we start from the reduced 12-layer architecture and therefore compare results to *w2v2-L-robust-12* (cf. Section V-B). There is no noteworthy degradation for arousal and dominance when keeping close to the entire training set. The only exception is dominance on IEMOCAP, where we achieve best results with just 75% of the data. For these dimensions, however, performance already saturates at 25% yielding a loss of less than .02 on MSP-Podcast, whereas for IEMOCAP, even 12.5% of the training samples seem sufficient to stay within a margin of .05.

Once again, it is a different story for valence. For MSP-Podcast, we see a constant improvement that only begins to decrease when reaching 75% of the data. For MOSI, we even see a boost in CCC of almost .1 for the remaining 25%. However, in light of our findings from Section IV-C, this does not come as a surprise. Providing more linguistic diversity makes it more likely a model can detect associations between key words and emotional context. What is a surprise, though, is that on IEMOCAP, using just 7.5% of the data, results in a drop of less than .05. A possible explanation is that the vocabulary of IEMOCAP does not resemble that of MSP-Podcast and that, therefore, the impact of linguistic information is limited. This would also explain why the differences in valence performance are less pronounced for IEMOCAP (cf. Fig. 2).

VI. SUMMARY

We explored the use of (pre-trained) transformer-based architectures for speech emotion recognition. In the previous sections, we dealt with several questions in isolation. We now attempt a unified summary by collectively considering all findings.

Effect of pre-training: pre-training is essential to get good performance (Section IV-F), especially for the valence dimension. This is particularly evident when training wav2vec 2.0 from a random initialisation (Section IV-K): the model performs substantially worse on all three dimensions, and its embeddings are unable to capture valence information. In addition, pre-training serves as a form of regularisation which helps stabilise the training (Section V-A), thus resulting in models which require less iterations, and less data to train on (Section V-C). However, we were unable to determine a clear relationship of the form ‘more pre-training data leads to better performance’. In fact, downstream performance can be negatively impacted by the introduction of more data, as seen by the comparison between *w2v2-L-vox* and *w2v2-L-xls-r*, which differ only in the fact that *w2v2-L-xls-r* has been trained on more (and more diverse) data, yet performs worse on all three dimensions.

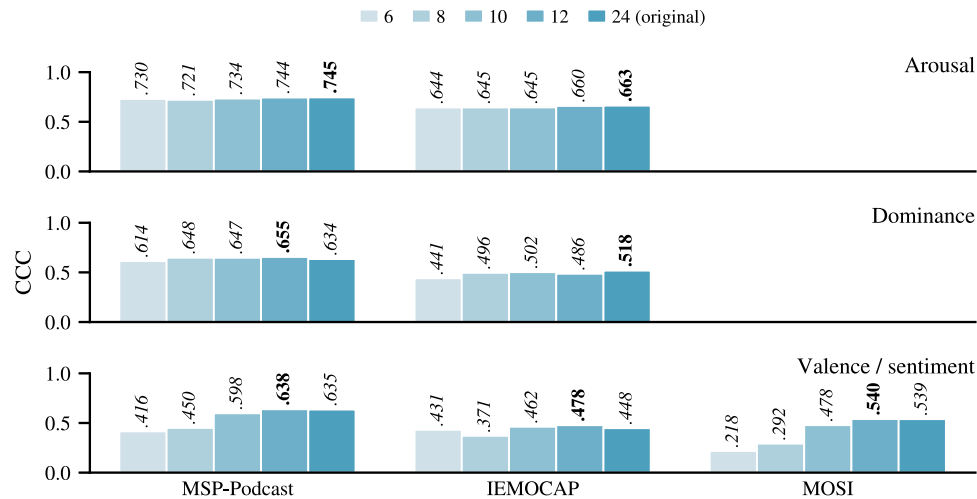


Fig. 12. CCC scores for arousal, dominance, and valence / sentiment for w2v2-L-robust and pruned versions. The legend shows the number of bottom layers kept during fine-tuning. We see that half of the layers can be removed without any loss in performance.

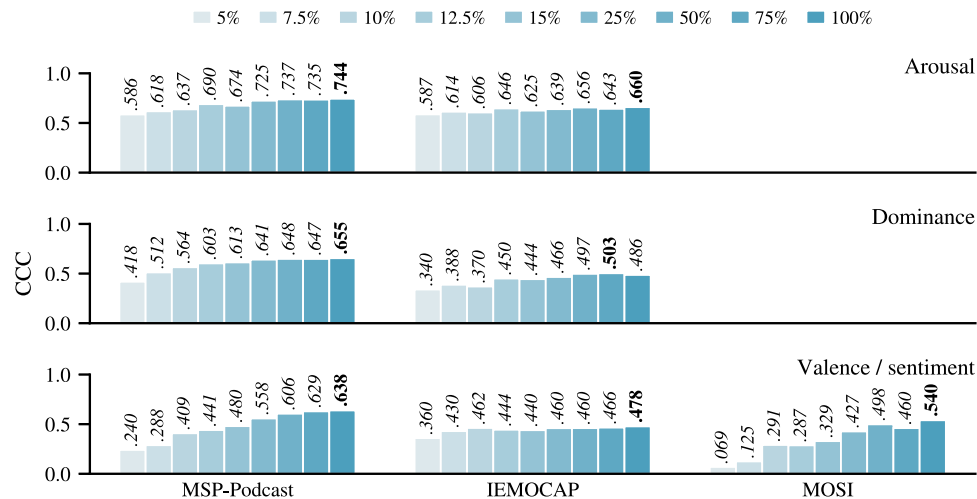


Fig. 13. CCC scores for arousal, dominance, and valence / sentiment for w2v2-L-robust on sparse training data. The legend shows the fraction of data used for fine-tuning. Please note that steps are not linear.

Generalisation: transformer-based models show very good cross-corpus generalisation (Section IV-F), robustness (Section IV-H), and appear invariant to domain, speaker, and gender characteristics (Section IV-K). These are all very important traits for any model that is intended for production use in realistic environments. However, they seem to stem primarily from the architecture rather than the pre-training as they are also evident in models initialised from random weights (Section IV-K). We also showed that several self-attention layers can be removed without hampering downstream performance (Section V-B), though they might still be necessary for successful pre-training.

Fairness: fairness remains a challenging topic for contemporary machine learning architectures. Community discussions primarily concern the issue of *group fairness*. In the present, we investigate this for the only group variable available in our datasets: gender (Section IV-I), where we observe that transformer-based architectures are more fair than the *CNN14*

baseline. However, we argue that *individual fairness* is important for SER. This refers to how models perform across different speakers; a feat which proves challenging even for the top-performing models investigated here (Section IV-J). We consider this an important topic which has not been sufficiently investigated for SER, though it is long known to impact other speech analysis models [35], [37].

Integration of linguistic and paralinguistic streams: finally, one of our most intriguing findings is that transformers seem capable of integrating both information streams of the voice signal. This is evident in how well-performing valence prediction models retain their effectiveness for synthesised speech lacking emotional intonation (Section IV-C) and fail to benefit from fusion with explicit textual information (cf. Section IV-B). Interestingly, this is only possible when fine-tuning the self-attention layers (Section IV-D), as keeping them frozen results to complete failure for synthesised speech (Section IV-C). This draws

attention to an under investigated aspect of fine-tuning, namely, how it qualitatively affects the nature of internal representations. Common understanding sees it as a mechanism through which to obtain better performance, but our analysis shows that it leads to a fundamental change in how the underlying signal is represented (moving from almost no sensitivity to linguistic content to increased reactivity to it). This mechanism may be crucial in the pursuit of paralinguistic and linguistic integration which is key to a holistic understanding of human communication. However, this integration might prove problematic in cases where the two modalities disagree, e. g. in cases of irony [67]. Our results also highlight that good valence performance might be language dependent as models pre-trained on a variety of languages perform worse for valence compared with comparable models pre-trained only for English (Section IV-A).

VII. CONCLUSION

Transformers have already revolutionised a very diverse set of artificial intelligence tasks, including speech emotion recognition. The present contribution goes beyond previous works that already established their effectiveness for SER by conducting a thorough evaluation and analysis of prominent transformer-based speech models for dimensional emotion recognition. We obtain state-of-the-art valence recognition performance on MSP-Podcast of. 638 without using explicit linguistic information, and manage to attribute this exceptional result to implicit linguistic information learnt through a fine-tuning of the self-attention layers. We release our best performing model (*w2v2-L-robust-12*) to the community [22].⁵ Transformer architectures are more robust to small perturbations, fair on the (gender) group- if not on the individual-level, and generalise across different domains. Our findings demonstrate that a new era is dawning in speech emotion recognition: that of pre-trained, transformer-based foundation models, which can finally lead to the coveted integration of the two dominant information streams of spoken language, linguistics, and paralinguistics.

REFERENCES

- [1] B. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, 2018.
- [2] P. Ekman, "An argument for basic emotions," *Cogn. Emotion*, vol. 6, no. 3/4, pp. 169–200, 1992.
- [3] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Pers.*, vol. 11, no. 3, pp. 273–294, 1977.
- [4] B. T. Atmaja, A. Sasou, and M. Akagi, "Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion," *Speech Commun.*, vol. 140, pp. 11–28, 2022.
- [5] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
- [6] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.
- [7] J. Kossaiifi et al., "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1022–1040, Mar. 2021.
- [8] S. Sahu, V. Mitra, N. Seneviratne, and C. Y. Espy-Wilson, "Multi-modal learning for speech emotion recognition: An analysis and comparison of ASR outputs with ground truth transcription," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria: ISCA, 2019, pp. 3302–3306.
- [9] A. Triantafyllopoulos, U. Reichel, S. Liu, S. Huber, F. Eyben, and B. W. Schuller, "Multistage linguistic conditioning of convolutional layers for speech emotion recognition," 2021, *arXiv:2110.06650*.
- [10] C. Oates, A. Triantafyllopoulos, I. Steiner, and B. W. Schuller, "Robust speech emotion recognition under different encoding conditions," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria: ISCA, 2019, pp. 3935–3939.
- [11] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. W. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria: ISCA, 2019, pp. 1691–1695.
- [12] A. Batliner, S. Hantke, and B. W. Schuller, "Ethics and good practice in computational paralinguistics," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1236–1253, Third Quarter 2022.
- [13] J. Cheong, S. Kalkan, and H. Gunes, "The hitchhiker's guide to bias and fairness in facial affective signal processing: Overview and techniques," *IEEE Signal Process. Mag.*, vol. 38, no. 6, pp. 39–49, Nov. 2021.
- [14] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.
- [15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, Vienna, Austria (virtual), 2020, pp. 1597–1607.
- [16] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2022.
- [17] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5998–6008.
- [18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2020, pp. 12449–12460.
- [19] W.-N. Hsu et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.
- [20] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/Hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," 2021, *arXiv:2111.02735*.
- [21] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Trans. Affect. Comput.*, Sep. 21, 2021, doi: [10.1109/TAFFC.2021.3114365](https://doi.org/10.1109/TAFFC.2021.3114365).
- [22] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. W. Schuller, "Model for dimensional speech emotion recognition based on wav2vec 2.0," Zenodo, 2022, doi: [10.5281/zenodo.6221127](https://doi.org/10.5281/zenodo.6221127).
- [23] D. N. Krishna, "Using large pre-trained models with cross-modal attention for multi-modal emotion recognition," 2021, *arXiv:2108.09669*.
- [24] J. Yuan, X. Cai, R. Zheng, L. Huang, and K. Church, "The role of phonetic units in speech emotion recognition," 2021, *arXiv:2108.01132*.
- [25] S. wen Yang et al., "Superb: Speech processing universal performance benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198, doi: [10.21437/Interspeech.2021-1775](https://doi.org/10.21437/Interspeech.2021-1775).
- [26] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 3400–3404.
- [27] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition," 2021, *arXiv:2110.06309*.
- [28] M. R. Makiuchi, K. Uto, and K. Shinoda, "Multimodal emotion recognition with high-level speech and text features," 2021, *arXiv:2111.10202*.
- [29] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
- [30] S. Srinivasan, Z. Huang, and K. Kirchhoff, "Representation learning through cross-modal conditional teacher-student training for speech emotion recognition," 2021, *arXiv:2112.00158*.
- [31] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Fourth Quarter 2019.
- [32] M. Li et al., "Contrastive unsupervised learning for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Toronto, ON, Canada, 2021, pp. 6329–6333.
- [33] M. Jaiswal and E. M. Provost, "Best practices for noise-based augmentation to improve the performance of emotion recognition 'in the wild'," 2021, *arXiv:2104.08806*.

⁵<https://github.com/audeering/w2v2-how-to>

- [34] R. Pappagari, J. Villalba, P. Želasko, L. Moro-Velazquez, and N. Dehak, "CopyPaste: An augmentation method for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6324–6328.
- [35] S. S. Rajan, S. Udeshi, and S. Chattopadhyay, "AequoVox: Automated fairness testing of speech recognition systems," 2021, *arXiv:2110.09843*.
- [36] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane, "Gender de-biasing in speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria: ISCA, 2019, pp. 2823–2827.
- [37] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. A. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation," in *Proc. 5th Int. Conf. Spoken Lang. Process.*, Sydney, Australia: ISCA, 1998, pp. 1–4.
- [38] W.-N. Hsu et al., "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," 2021, *arXiv:2104.01027*.
- [39] C. Wang et al., "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, Bangkok, Thailand, 2021, pp. 993–1003.
- [40] A. Babu et al., "XLS-R: Self-supervised cross-lingual speech representation learning at scale," 2021, *arXiv:2111.09296*.
- [41] G. Trigeorgis et al., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Shanghai, China, 2016, pp. 5200–5204.
- [42] 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. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Toronto, ON, Canada, 2021, pp. 7268–7272.
- [43] 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 Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.
- [44] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov./Dec. 2016.
- [45] H. Wierstorf, "Gender annotations for multimodal opinion-level sentiment intensity dataset (MOSI)," Zenodo, 2023, doi: [10.5281/zenodo.7554349](https://doi.org/10.5281/zenodo.7554349).
- [46] M. Munzero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Trans. Affect. Comput.*, vol. 5, no. 2, pp. 101–111, Second Quarter 2014.
- [47] L. Tian, C. Lai, and J. Moore, "Polarity and intensity: The two aspects of sentiment analysis," in *Proc. Grand Challenge Workshop Hum. Multimodal Lang.*, Melbourne, Australia: ACL, 2018, pp. 40–47.
- [48] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, "Machine learning testing: Survey, landscapes and horizons," *IEEE Trans. Softw. Eng.*, vol. 48, no. 1, pp. 1–36, Jan. 2020.
- [49] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.
- [50] F. Ringeval et al., "AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proc. Audio/Visual Emotion Challenge Workshop*, 2018, pp. 3–13.
- [51] E. Parada-Cabaleiro, A. Baird, A. Batliner, N. Cummins, S. Hantke, and B. Schuller, "The perception of emotions in noisified nonsense speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Stockholm, Sweden: ISCA, 2017, pp. 3246–3250.
- [52] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd Annu. ACM Conf. Multimedia*, ACM Press, 2015, pp. 1015–1018. [Online]. Available: <http://dl.acm.org/citation.cfm?doid=2733373.2806390>
- [53] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," 2018, *arXiv: 1808.00023*.
- [54] J. K. Fitzsimons, A. A. Ali, M. A. Osborne, and S. J. Roberts, "A general framework for fair regression," *Entropy*, vol. 21, 2019, Art. no. 741.
- [55] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Minneapolis, MN, USA, 2019, pp. 4171–4186.
- [56] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Barcelona, Spain, 2020, pp. 6484–6488.
- [57] A. Triantafyllopoulos et al., "Probing speech emotion recognition transformers for linguistic knowledge," *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2022, pp. 146–150.
- [58] A. T. Liu, S.-W. Yang, P.-H. Chi, P.-C. Hsu, and H.-Y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Barcelona, Spain, 2020, pp. 6419–6423.
- [59] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Sci. Robot.*, vol. 3, no. 19, pp. 1–11, 2018.
- [60] A. Triantafyllopoulos, S. Liu, and B. W. Schuller, "Deep speaker conditioning for speech emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, Shenzhen, China, 2021, pp. 1–6.
- [61] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," 2022, *arXiv:2201.07876*.
- [62] A. D'Amour et al., "Underspecification presents challenges for credibility in modern machine learning," 2020, *arXiv: 2011.03395*.
- [63] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.
- [64] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, Sardinia, Italy: PMLR, 2010, pp. 201–208.
- [65] B. Neyshabur, H. Sedghi, and C. Zhang, "What is being transferred in transfer learning?," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2020, pp. 512–523.
- [66] H. Sajjad, F. Dalvi, N. Durrani, and P. Nakov, "Poor man's BERT: Smaller and faster transformer models," 2020, *arXiv: 2004.03844*.
- [67] F. Burkhardt, B. Weiss, F. Eyben, J. Deng, and B. Schuller, "Detecting vocal irony," in *Proc. Int. Conf. Lang. Technol. Challenges Digit. Age*, Cham: Springer International Publishing, 2018, pp. 11–22.



Johannes Wagner received a doctoral degree from the University of Augsburg, in 2015 for his research on multimodal behaviour analysis and has been working in research projects funded by the European Union (Humaine 2004–2007, Callas 2006–2010, CEEDS 2010–2015, Ilhaire 2011–2014, Kristina 2015–2018, AriaValuspa 2015–2018). Since 2019 he is a senior researcher with audEERING GmbH.



Andreas Triantafyllopoulos (Student Member, IEEE) received the diploma in ECE from the University of Patras, Greece, in 2017. He is working toward the doctoral degree with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg. His current focus is on deep learning methods for auditory intelligence and affective computing.



Hagen Wierstorf received the doctoral degree from Technische Universität Berlin, in 2014 on the topic of sound field synthesis and continued as a post-doc. In 2016 he moved as a post-doc to Technische Universität Ilmenau and was a visiting professor with Filmuniversität Babelsberg KONRAD WOLF. In 2017 he was a post-doc with the University of Surrey. Since 2018 he is a senior researcher with audEERING GmbH.



Maximilian Schmitt received the diploma degree in electrical engineering from RWTH Aachen University, in 2012, and the doctoral degree in computer science from the University of Augsburg/Germany, in 2022. Afterwards, he worked as a research assistant for the University of Music Detmold and the University of Passau. His research focus is on intelligent audio analysis and multimodal affect recognition.



Florian Eyben received the doctorate degree in electrical engineering from Technische Universität München (TUM). He is co-founder and chief technical officer (CTO) with audEERING GmbH. He is an expert in the field of digital signal processing, speech and music analysis, and machine learning. His 100+ publications have 13000+ citations (h-index 49). He is responsible for the audio analysis tools openSMILE and openEAR.



Felix Burkhardt received the doctoral degree from the TU Berlin, in 2000 on the simulation of emotional speech by machines and has been working with the Deutsche Telekom AG for 18 years as a speech and text technology expert. He is the director of research with audEERING GmbH and part-time lecturer with the TU Berlin on auditory human machine interaction.



Björn W. Schuller (Fellow, IEEE) received the diploma, doctoral, and habilitation degrees in electrical engineering/information technology from the Technical University of Munich, Munich/Germany. He is a full professor and chair of embedded intelligence for health care and wellbeing with University of Augsburg/Germany, heads Imperial College London's Group on Language Audio and Music (GLAM), and is co-founding CEO and current CSO of audEERING GmbH. He is a Fellow of the AAAC, BCS, and ISCA.