DAWN OF THE TRANSFORMER ERA IN SPEECH EMOTION RECOGNITION: CLOSING THE VALENCE GAP

Johannes Wagner[1], Andreas Triantafyllopoulos[2], Hagen Wierstorf[1], Maximilian Schmitt[1], Felix Burkhardt[1], Florian Eyben[1], Björn W. Schuller[1,2,3]
[1] audEERING GmbH, Gilching, Germany
[2] EIHW, University of Augsburg, Augsburg, Germany
[3] GLAM, Imperial College, London, UK

March 17, 2022

ABSTRACT

Recent advances in transformer-based architectures which are pre-trained in self-supervised manner have shown great promise in several machine learning tasks. In the audio domain, such architectures have also 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. Furthermore, our investigations reveal that transformer-based architectures are more robust to small perturbations compared to a CNN-based baseline and fair with respect to biological sex groups, but not towards individual speakers. Finally, we are the first to show that their extraordinary success on valence is based on implicit linguistic information learnt during fine-tuning of the transformer layers, which explains why they perform on-par with recent multimodal approaches that explicitly utilise textual information. Our findings collectively paint the following picture: transformer-based architectures constitute the new state-of-the-art in SER, but further advances are needed to mitigate remaining robustness and individual speaker issues. To make our findings reproducible, we release the best performing model to the community.

1 Introduction

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

The goal of an automatic SER system is to analyse the voice signal and derive a prediction of an emotional category or dimensional value. This can be done either through the linguistic (*what* has been said) or the paralinguistic (*how* it has been said) stream [4] – or both. The former is primarily contained in textual information (e. g. the transcription of an input utterance) while the latter in the acoustic and prosodic information contained in the raw audio signal. Each stream comes with its pros and cons: linguistics is better suited for valence recognition [5, 6, 7, 8] and can heavily draw from recent advances in automatic speech recognition (ASR) and natural language processing (NLP) [9, 10, 11], but might be limited to a single language. Paralinguistics works better for arousal and dominance [5, 6, 7, 8] and has the potential to generalise across different languages, while typically suffering from low valence performance [4, 8]. Of course,

the strengths of both paradigms can be utilised in complementary fashion in *multimodal*[1] architectures. However, this entails a combination of several different models, which puts a potentially prohibitive strain on computational resources, while still suffering from a major limitation of linguistic-based approaches, which is that they are limited to a single language. For this reason, we aim towards a model that only implicitly (if at all) utilises the linguistic information stream during deployment, and does not require access to ASR and NLP frontends.

Although this 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 [4, 8], b) overcoming issues of generalisation and robustness [12, 13], and c) alleviating individual- and group-level fairness concerns, which is a prerequisite for ethical emotion recognition technology [14, 15]. Previous works have attempted to tackle these issues in isolation, e. g. by using cross-modal knowledge distillation to increase valence performance [16], speech enhancement or data augmentation to improve robustness [12, 13], and de-biasing techniques to mitigate unfair outcomes [17]. However, each of those approaches comes with its own knobs to twist and hyperparameters to tune, making their combination far from 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 easily adapted to several use-cases [18]. Typically, these foundation models are trained on large datasets, often using *proxy tasks* to avoid dependencies on hard-to-acquire labels, and then fine-tuned on (small) sets of labelled data for their intended tasks. Such models have seen tremendous success in computer vision [19], NLP [20], and computer audition [21, 22] – including SER [23, 24, 25].

Among others, wav2vec 2.0 [21] and HuBERT [22] have emerged as foundation model candidates for speech-related applications. Prior works have successfully utilised one or both of them for (primarily categorical) SER (cf. Section 2).

In this work, we evaluate several publicly-available pre-trained models for dimensional SER (cf. Section 4). We analyze the success of the models (cf. Section 5), and investigate their efficiency (cf. Section 6). Hereby we answer several questions, which are organised as sub-sections:

- Can transformer-based models close the performance gap for valence? (Section 4.1)
- Do the models generalise better across different domains? (Section 4.2)
- Does more (and more diverse) data during pre-training lead to better performance? (Section 4.3)
- Does a larger architecture lead to better performance? (Section 4.4)
- Are the models robust against small perturbations to the input signals? (Section 4.5)
- Are the models fair regarding the biological sex of the speaker? (Section 4.6)
- Is performance equal across all speakers? (Section 4.7)
- Does explicit linguistic information further improve performance? (Section 4.8)
- Why do transformer-based models generalise so well? (Section 5.1)
- How important is a fine-tuning of the transformer layers? (Section 5.2)
- Do the models implicitly learn linguistic information? (Section 5.3)
- Does pre-training help with training stability and convergence? (Section 6.1)
- How many transformer layers do we really need? (Section 6.2)
- Can we reduce the training data without a loss in performance? (Section 6.3)

Moreover, we make our best performing model publicly available [26]. To our best knowledge this is the first transformer-based dimensional SER model released to the community. For an introduction how to use it, please visit: https://github.com/audeering/w2v2-how-to.

The remainder of this paper is organised as follows. Section 2 discusses related work, Section 3 presents the models, databases, and evaluation methods. Section 4 shows the results that are then further analysed in Section 5. Section 6 investigates efficiency improvements, before Section 7 summarises the results, and Section 8 concludes the paper.

2 Related Work

In Table 1, we provide a summary of recent works based on wav2vec 2.0 and HuBERT on the IEMOCAP dataset [33], where most prior works have focused. Results are ranked by unweighted average recall (UAR) / weighted average re-

[1]Technically speaking, the term *multistream* would be more correct, as linguistics and paralinguistics do not constitute different modalities per se, but different information streams derived from the same modality, namely speech. Nevertheless, we adopt the term multimodal, as it is more often used in literature.

Table 1: *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 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	Krishna [27]	w2v2-L	✓			60.0	
2	Yuan *et al.* [28]*	w2v2-L	✓			62.5	62.6
3	Wang *et al.* [23]	w2v2-b					63.4
4	Yang *et al.* [29]	w2v2-b				63.4	
5	Pepino *et al.* [30]	w2v2-b		✓		63.8	
6	Wang *et al.* [23]	hubert-b					64.9
7	Yang *et al.* [29]	hubert-b				64.9	
8	Wang *et al.* [23]	w2v2-L	✓				65.6
9	Yang *et al.* [29]	w2v2-L	✓			65.6	
10	Pepino *et al.* [30]	w2v2-b				67.2	
11	Wang *et al.* [23]	hubert-L	✓				67.6
12	Yang *et al.* [29]	hubert-L	✓			67.6	
13	Chen and Rudnicky [31]	w2v2-b			✓	69.9	
14	Makiuchi *et al.* [32]	w2v2-L	✓			70.7	
15	Wang *et al.* [23]	w2v2-b		✓	✓		73.8
16	Chen and Rudnicky [31]	w2v2-b			✓	74.3	
17	Wang *et al.* [23]	hubert-b			✓		76.6
18	Wang *et al.* [23]	w2v2-L	✓	✓	✓		76.8
19	Wang *et al.* [23]	w2v2-b			✓		77.0
20	Wang *et al.* [23]	w2v2-L	✓		✓		77.5
21	Wang *et al.* [23]	hubert-L	✓	✓	✓		79.0
22	Wang *et al.* [23]	hubert-L	✓		✓		79.6

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

call (WAR) on the four emotional categories of anger (1103 utterances), happiness (1636), sadness (1084), and neutral (1708), which is the typical categorical SER formulation for IEMOCAP. Since we are dealing with an unbalanced class problem, UAR and WAR can diverge. However, Yuan *et al.* [28] report both yielding almost identical values. We therefore assume that a ranking over both metrics is still meaningful. Most of the works apply leave-one-session-out cross validation (5 folds), except Yuan *et al.* [28] using leave-one-speaker-out cross validation (10 folds) and Wang *et al.* [23] who do not explicitly mention which folds they used. The results are obtained with the base architecture (*w2v2-b* / *hubert-b*) or the large architecture (*w2v2-L* / *hubert-L*) in a down-stream classification task (for more details on the models, see Section 3.2). Even though, authors have used different head architectures and training procedures in their studies, we can draw some general observations from Table 1:

1. We see a roughly 10% better performance with models where the weights of the pre-trained model were not frozen during the down-stream task.
2. Using a pre-trained model fine-tuned for speech recognition does not help with the down-stream task (e. g. row 15 vs row 19 -3.2%).
3. When the base and the large architecture of the same model type are tested within the same study, the large one yields better results (e. g. row 17 vs row 22 $+3.0\%$), though the difference can be quite small (e. g. row 19 vs row 20 $+.5\%$).
4. Likewise, in that case 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 Wang *et al.* [23] 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 Makiuchi *et al.* [32] only report results without fine-tuning the pre-trained model during the down-stream task.

Table 2: *State-of-the-art concordance correlation coefficient (CCC) performance using transformer-based architectures on MSP-Podcast for arousal, dominance, and valence (sorted by the latter). The table encodes whether the base or large (L) architecture was used. In all cases, the pre-trained models were further fine-tuned during the downstream task.*

	Work	Model	L	A	D	V
1	Srinivasan *et al.* [16]	w2v2-b		.728	.636	.363
2	Srinivasan *et al.* [16]	w2v2-L	✓	.735	.654	.472
3	Srinivasan *et al.* [16]	hubert-b		.733	.640	.485
4	Srinivasan *et al.* [16]	hubert-L	✓	.752	.674	.547

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 Srinivasan *et al.* [16], 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 – a feat which has long escaped audio-based models. 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 [34]. Furthermore, they were able to distill the multi-model system to an audio-only model using student-teacher transfer learning, while still reaching a CCC of .627 (a massive improvement compared to the previous state-of-the-art performance of only .377 [35]). In Table 2, we summarise their results for *w2v2-b*, *hubert-b*, *w2v2-L*, and *hubert-L* without cross-modal distillation. The numbers back up two of our earlier findings: the large architecture is superior to the base model and HuBERT outperforms wav2vec 2.0. Their CCC performance surpasses both that of Triantafyllopoulos *et al.* [4] (.515), who proposed a multimodal fusion of pre-trained BERT embeddings with an untrained CNN model, and of Li *et al.* [35] (.377) who pre-train a CRNN model on LibriSpeech using Contrastive Predictive Coding and subsequently fine-tuned it on MSP-Podcast.

The presented results clearly demonstrate the great potential of wav2vec 2.0 and HuBERT for emotion recognition. However, it remains unclear what influence the amount and domain of the data used for pre-training really has. For instance, even though the large model consistently shows better performance, it is unclear if that can be attributed to the additional layers or the fact that it was trained on 60 times more data compared to the base model. Likewise – since the models used in previous work were all pre-trained on read English speech – there is little understanding on the impact that the use of speech from other domains may have. In this contribution, we therefore 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).

Besides investigating performance of SER models on clean test data, it is important to show that they also work well under more challenging conditions. Even though augmentation methods have been used to improve performance on clean test data [36, 37], only a few studies have evaluated performance on augmented test data as well. Jaiswal and Provost [38] and Pappagari *et al.* [39] have shown that previous SER models show robustness issues, particularly for background noise and reverb. In this contribution, we systematically investigate robustness of transformer-based models against a variety of augmentations, focusing on small perturbations of the input signal as larger changes can modify the perceived emotion [38, 40].

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. biological sex in automatic speech recognition [41]. For SER models, only a few evaluations are available. Gorrostieta *et al.* [17] found a decrease in CCC for females compared to males for arousal in MSP-Podcast (v1.3) of around .234 for their convolutional model. Besides group fairness, this contribution investigates individual fairness by estimating the influence of the speaker on the model performance, which is a known problem for other speaker verification models [42].

3 Experimental setup

3.1 Architecture

Inspired by Wang *et al.* [23], we use a simple head architecture, which we build on top of wav2vec 2.0 [21] or HuBERT [22] (see Figure 1): 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 (the pooled embeddings and the hidden layer outputs are dropped out). For fine-tuning on the downstream task, we use the ADAM optimiser with CCC loss, which is the

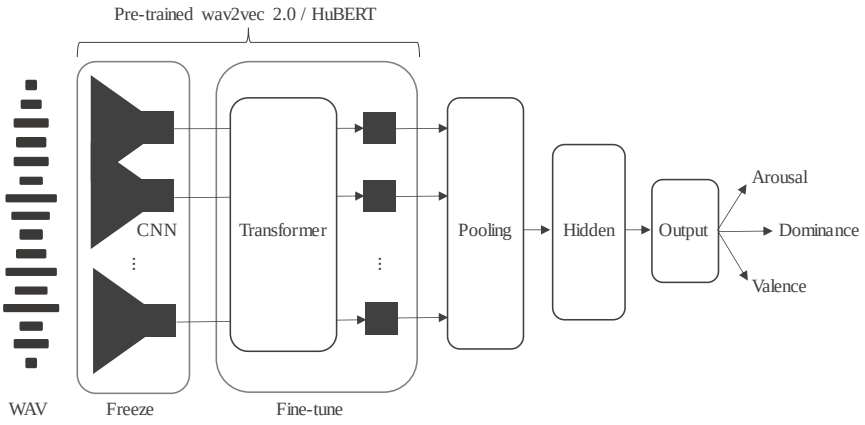


Figure 1: *Proposed architecture built on wav2vec 2.0 / HuBERT.*

standard loss function used for dimensional SER [4, 35, 43], 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 layers. According to Wang *et al.* [23], such a partial fine-tuning yields better results than a full fine-tuning including the CNN-based feature encoder. Note that in the following, when we use the term fine-tuning, we actually refer to a partial fine-tuning, except where otherwise specified. 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*) we have been successfully using for SER in previous work [4, 44]. It follows the architecture proposed by Kong *et al.* [45] for audio pattern recognition. Different to the transformer-based models, which operate on the raw audio signal, this takes log-Mel spectrograms as input. Note that this model is *not* pre-trained, i. e. it is always trained from scratch in our experiments. We used 60 epochs, a learning rate of $.01$, and a batch size of 64 using stochastic gradient descent (SGD) with a Nesterov momentum of $.9$. We selected the model that performed best on the validation set.

3.2 Pre-trained models

Throughout the paper, we will discuss results obtained with transformer-based models pre-trained on massive amount of unlabelled data. Basically, we investigate two variants: wav2vec 2.0 and HuBERT, which share a similar design, but follow a different training procedure. Both exist in two forms: a base architecture with 12 transformer layers and 768 hidden units (95M parameters), and a large architecture with 24 transformer layers and 1024 hidden units (317M parameters). Apart from that, we further distinguish them by the data used for pre-training. For the sake of readability, we will refer to the models by aliases introduced in Table 3a. Also note that – unless otherwise stated – we refer to their fine-tuned versions when we report results (cf. Section 3.1).

The core idea of a transformer-based model is to transform speech data into a sequence of discrete units, similar to the words in a text sentence. To learn such a representation, the models are pre-trained in a self-supervised way, i. e. without any labels. Hence, basically any speech dataset can be used. One goal of this paper is to gain a better understanding what influence the pre-training data has on the performance of the fine-tuned model. Table 3b provides an overview of the data used for pre-training. Beside the four models we found in previous work (cf. Section 2) and which are pre-trained on English audiobooks (1-4), we picked a model additionally trained on telephone speech (5), a model trained only on parliamentary speech in multiple languages (6), and a model trained on more than 400k hours across all domains (also in multiple languages) (7). We did not include models fine-tuned on speech recognition but trust our earlier assumption that they will not lead to better performance.

3.3 Datasets

We used the MSP-Podcast corpus [34] (v1.7) to run multitask training on the three dimensions of arousal, dominance, and valence. The dataset consists of roughly 84 hours of naturalistic speech from podcast recordings. The original labels cover a range from 1 to 7, which we normalise into the interval of 0 to 1. In-domain results are reported on the *test-1* split. The *test-1* split contains $12,902$ samples (54% female / 46% male) from 60 speakers (30 female / 30

Table 3: *Transformer-based models included in this study.*

(a) *Names of the original pre-trained models*[a] *and aliases used throughout the paper. Models comprised of two architecture designs (wav2vec 2.0 and HuBERT), each with two different variants (base and large).*

	Work	Original name	Alias
1	Baevski *et al.* [21]	wav2vec2-base	w2v2-b
2	Hsu *et al.* [22]	hubert-base-ls960	hubert-b
3	Baevski *et al.* [21]	wav2vec2-large	w2v2-L
4	Hsu *et al.* [22]	hubert-large-ll60k	hubert-L
5	Hsu *et al.* [46]	wav2vec2-large-robust	w2v2-L-robust
6	Wang *et al.* [47]	wav2vec2-large-100k-voxpopuli	w2v2-L-vox
7	Babu *et al.* [48]	wav2vec2-xls-r-300m	w2v2-L-xls-r

[a]Models are publicly available at https://huggingface.co/facebook

(b) *Details on the data used during pre-training. For each model, we list included dataset(s), total number of hours (h), number of languages (*eng *if only English), and covered domains (**R**ead speech, **T**elephone conversions, **P**arliamentary speech, **Y**outube).*

	Model	Datasets	h	Lang	R	T	P	Y
1	w2v2-b	LibriSpeech	960	eng	✓			
2	hubert-b	LibriSpeech	960	eng	✓			
3	w2v2-L	Libri-Light	60k	eng	✓			
4	hubert-L	Libri-Light	60k	eng	✓			
5	w2v2-L-robust	Libri-Light (60k) Fisher (2k) CommonVoice (700) Switchboard (300)	63k	eng	✓	✓		
6	w2v2-L-vox	VoxPopuli	100k	23			✓	
7	w2v2-L-xls-r	VoxPopuli (372k) Multilingual LibriSpeech (50k) CommonVoice (7k) VoxLingua107 (6.6k) BABEL (1k)	436k	128	✓	✓	✓	✓

male). The samples per speaker are not balanced and vary between 42 and 912. The samples have a combined length of roughly 21 hours, and vary between 1.92 s and 11.94 s per sample.

We report cross-domain results for the IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset [33], which consists of roughly 12 hours of scripted and improvised dialogues by ten speakers (5 female / 5 male). It provides the same dimensional labels as MSP-Podcast corpus, but in a range of 1 to 5, which we normalise to the interval 0 to 1. Since we use the dataset only during evaluation, we do not apply the usual speaker cross-folding, but treat the corpus as a whole. It includes $10,039$ samples (49% female / 51% male) with a varying length between .58 s and 34.14 s.

Finally, we additionally report cross-corpus results for valence on the Multimodal Opinion Sentiment Intensity (MOSI) [49] corpus. The dataset is a collection of 4 h 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 normalise to the interval 0 to 1. As the gender/sex labels are not part of the distributed database, we re-annotated them ourselves. We report results on the test set that contains 685 samples (51% female / 49% male) with a total duration of 1 hour and varying sample length between .57 s and 33.13 s.

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 [50], there is nevertheless evidence to suggest that sentiment annotations can be decomposed to two constituents: intensity and polarity [51], which we consider to roughly correspond to arousal and valence. We therefore expect some correlation between (predicted) valence and (annotated) sentiment scores. As our primary interest is a between-model comparison for out-of-domain generalisation, and not the absolute sentiment prediction performance itself, we consider the use of MOSI for cross-corpus experiments well-motivated from a practical, if not necessarily a theoretical, point of view.

3.4 Evaluation

Machine learning models for speech emotion recognition are expected to work under different acoustical conditions and for different speakers. To cover this, we evaluate them for correctness, robustness, and fairness [52].

Correctness measures how well the model predictions match the ground truth labels. The concordance correlation coefficient (CCC) [53] provides an estimate how well the distribution of the model predictions corresponds to the distribution of the ground truth data. This is a typical measure to rank different models on dimensional SER benchmarks [54], and is used as the main ranking criterion in this work as well. For MSP-Podcast, the correctness analysis is additionally extended to single speakers.

Robustness (cf. Section 4.5) measures how stable the model predictions are against perturbations to the input signals, which do not affect the ground truth labels. Applying stronger changes to the input signals must be carefully done for SER, as they might affect the ground truth label [38, 40]. We focus instead of testing the invariance of the model against subtle perturbations. Robustness in this case is not defined by the change in the correctness metric, but given by the percentage of samples that show an absolute difference between model output for a given clean and augmented input signal below a defined threshold [55, 56]. As a threshold we select $.05$, which reflects a change of less than 5% on the regression scale. A robustness of $.95$ would indicate that 95% of all samples show a difference in the model output that is below $.05$.

As perturbations, we independently apply the following augmentations. They were all developed by applying them to a single high quality speech recording at 16 kHz, and ensuring that they are only slightly audible and do not change the perceived emotion: *Additive Tone* adds a sinusoid with a frequency randomly selected between 5000 Hz and 7000 Hz, with a peak signal-to-noise ratio randomly selected from 40 dB, 45 dB, 50 dB; *Append Zeros* adds samples containing zeros at the end of the input signal with the number of samples randomly selected from 100, 500, 1000; *Clip* clips a given percentage of the input signal with the percentage randomly selected from $.1\%$, $.2\%$, $.3\%$; *Crop Beginning* removes samples from the beginning of an input signal with the number of samples randomly selected from 100, 500, 1000; *Gain* changes the gain of an input signal by a value randomly selected from -2 dB, -1 dB, 1 dB, 2 dB; *Highpass Filter* applies a high pass Butterworth filter of order 1 to the input signal with a cutoff frequency randomly selected from 50 Hz, 100 Hz, 150 Hz; *Lowpass Filter* applies a low pass Butterworth filter of order 1 to the input signal with a cutoff frequency randomly selected from 7500 Hz, 7000 Hz, 6500 Hz; *White Noise* adds Gaussian distributed noise to the input signal with a root mean square based signal-to-noise ratio randomly selected from 35 dB, 40 dB, 45 dB.

Fairness (cf. Section 4.6) evaluates if the model predictions show biases for certain protected characteristics or attributes like race, biological sex, or age [57]. We focus on biological sex 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 [58]. We measure fairness by estimating the difference in the correctness metric (CCC) and expect it to be equal for male and female groups. We name it sex fairness score, which can be formulated as

$$\text{Sex fairness score} = \text{CCC}_{\text{female}} - \text{CCC}_{\text{male}}\,, \tag{1}$$

where $\text{CCC}_{\text{female}}$ is the CCC for all female samples, and CCC_{male} the CCC for all male samples in the test datasets. A positive sex fairness score indicates a better performance of the model for female speakers.

In addition, we assume that the average arousal, dominance, and valence values for the male and female groups are very similar. As we see differences in the ground truth labels between the male and female groups, we measure the difference relative to the ground truth labels. We name it sex fairness bias and define it as

$$\text{Sex fairness bias} = \overline{\hat{y}_{\text{female}} - y_{\text{female}}} - \overline{\hat{y}_{\text{male}} - y_{\text{male}}}\,, \tag{2}$$

where $\hat{y}_{\text{female}}$ are the predictions for all female samples, y_{male} the truth values for all male samples, and $\overline{(\cdot)}$ the mean. A positive sex fairness bias would indicate that the difference in mean arousal, dominance, or valence between females and males has changed into the direction of females for the predictions compared with the ground truth.

4 Evaluation

We begin our investigation with a thorough evaluation of transformer-based models. Utilising a comprehensive in-domain and cross-corpus testing scheme, we attempt to identify how different aspects of foundation models (e. g. model size and pre-training data) impact performance and generalisation. In addition, we place particular emphasis on

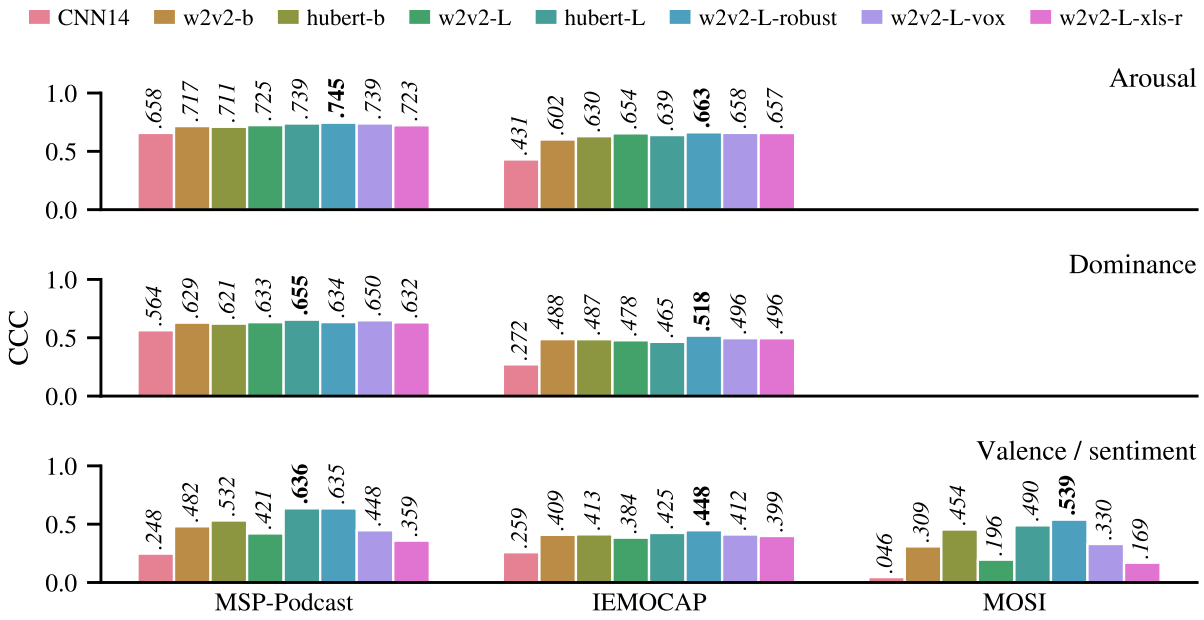


Figure 2: *CCC scores for arousal, dominance, valence (MSP-Podcast / IEMOCAP), and sentiment (MOSI). All models have been trained for emotional dimension prediction using multitasking on MSP-Podcast, and subsequently evaluated on its test set (in-domain), as well as to the test set of MOSI and the entire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

robustness and *fairness*, which are critical considerations for SER systems targeted to real-world applications. Finally, we investigate if explicit linguistic information can help improve model performance.

4.1 Can transformer-based models close the performance gap for valence?

In Figure 2, we show in-domain and cross-domain performance in terms of CCC scores for wav2vec 2.0 models with base and large architecture and different pre-training data, as well as for HuBERT models with base and large architecture and the *CNN14* baseline.

We first focus on the results for arousal and dominance. In-domain (MSP-Podcast), all transformer-based models score within a very narrow range ($\sim$.03). Best performance is achieved with *w2v2-L-robust* for arousal (.745) and with *hubert-L* for dominance (.655). *CNN14* performs slightly worse with a $\sim$.07 drop in average. On cross-domain data (IEMOCAP), the performance range for transformer models is approximately doubled and decreases by $\sim$.09 for arousal and $\sim$.15 for dominance compared to in-domain. The gap to *CNN14* is further increased to an average of $\sim$.21. Again, *w2v2-L-robust* achieves the best performance: arousal (.663) and dominance (.518).

For valence and MSP-Podcast, *hubert-L* (.636) and *w2v2-L-robust* (.635) are the best performing models. They are both clearly above the currently best reported performance of .547 using *hubert-L* [16], an increase of about .089. The gap to the *CNN14* model is even larger with .388. On cross-domain data, *w2v2-L-robust* is the best performing model reaching .448 on IEMOCAP. The gap to the *CNN14* model (.259) is lowered to 0.189. For MOSI, *hubert-L* and *w2v2-L-robust* are the clear winners, with the latter again achieving the highest correlation of .539. The gap to *CNN14* increases to .493 as the *CNN14* model achieves only a CCC of .046.

> The best performing models achieve a similar performance for valence and dominance for in-domain and cross-corpus data. This indicates that a transformer-based model can close the performance gap for valence without explicit linguistic information.

4.2 Do the models generalise better across different domains?

As we see a similar trend for different transformer-based models between in-domain and cross-corpus results in Figure 2, we focus on *w2v2-L-robust* to represent the transformer-based models for this analysis. The drop in CCC between in-domain and cross-corpus results for *w2v2-L-robust* is 11% for arousal, 21% for dominance, 30% for valence, all on IEMOCAP, and 15% for sentiment on MOSI. For *CNN14*, the drop in CCC is 34% for arousal, and 52% for dominance. For valence, we do not evaluate cross-domain performance as the in-domain 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.

Transformer-based models generalise better than the non-transformer baseline (*CNN14*).

4.3 Does more (and more diverse) data during pre-training lead to better performance?

As discussed in Section 2, previous studies report results only for *w2v2-b*, *hubert-b*, *w2v2-L*, and *hubert-L*, which were pre-trained only on clean speech, whereby 60 times more data was used to pre-train *w2v2-L* and *hubert-L*. This makes it difficult to draw conclusions about what influence size and domain of the pre-training data really have on downstream performance. In our study, we therefore included several wav2vec 2.0 models with large architecture and different pre-training (see Table 3).

For the in-domain valence results in Figure 2, we see an almost 10 times increase of CCC range ($\sim$.3) transformer models fall into than for arousal / dominance. The choice of the architecture and the data used during pre-training seems to be more crucial for valence detection. Previous studies uniformly report that HuBERT outperforms wav2vec 2.0 when comparing *hubert-L* and *w2v2-L*, which is replicated by our results with *w2v2-L* showing a $\sim$.2 smaller CCC than *hubert-L*. The increase in performance for *w2v2-L-robust* is therefore most likely explained with the additional 3k hours of telephone conversations used for pre-training. However, if we look at *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* (100k and 463k vs 63k hours), they perform clearly worse. Interestingly, they do not even match the performance of *w2v2-b*. For *w2v2-L-vox*, we may explain its low performance with the fact that it was trained on a single type of speech (parliamentary debates) and perhaps will only perform well within that specific context. However, *w2v2-L-xls-r* has been trained on the most diverse mix of data among the tested models (though again parliamentary speech forms the vast majority of $\sim$85%). Notably, both models 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 – a fact that can be further explored by multi-language evaluations.

We next turn to cross-domain results for valence and sentiment. We begin with IEMOCAP. Since it contains rather prototypical emotions expressed by a small number of actors, it can be regarded as a quite homogeneous corpus. This may explain that the performance of the transformer models again falls within a narrow range of $\sim$.06. For MOSI, however, we see a pattern that looks very similar to the one of MSP-Podcast, except that the extremes lie further apart, now covering a range of $\sim$.4. Once more, *hubert-L* and *w2v2-L-robust* are the clear winners, with the latter again achieving the highest correlation of .539. They are followed by the two base models and *w2v2-L-vox*. The models *w2v2-L* and *w2v2-L-xls-r* lag clearly behind, trailed only by *CNN14* which achieves almost zero correlation.

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 again a performance drop (at least when tested on English speech).

4.4 Does a larger architecture lead to better performance?

We cannot directly answer what influence the size of the architecture has on the performance, as we do not have transformer models with different architectures pre-trained on the same data in our evaluation (Figure 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 similar performance as the large models *w2v2-L*, *w2v2-L-vox*, and *w2v2-L-xls-r* for arousal and dominance, in-domain and cross-corpus. For valence, the small models outperform

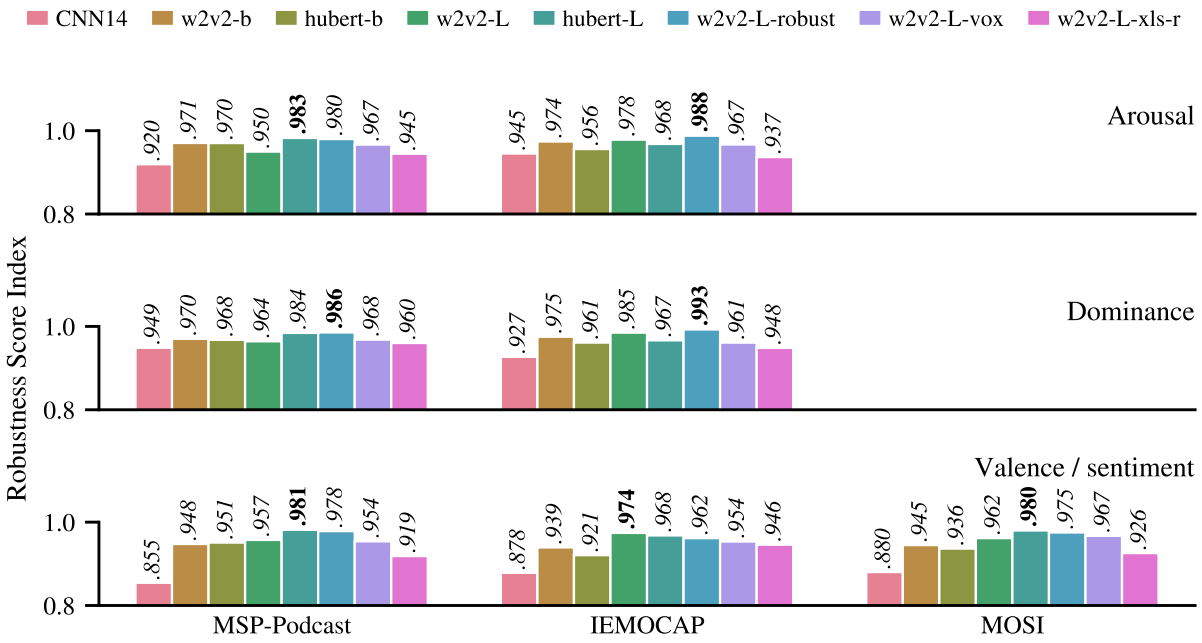


Figure 3: *Robustness scores averaged over all augmentations for arousal, dominance, valence (MSP-Podcast / IEMOCAP), and sentiment (MOSI). The robustness score is given by the ratio of samples that did not change more than* .05 *on a scale from* 0 *to* 1 *between the clean and augmented signal.*

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.

> 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.

4.5 Are the models robust against small perturbations to the input signals?

Figure 3 summarises the average robustness scores of the models over all augmentations described in Section 4.5. The robustness score is given by the ratio of samples that did not change more than .05 on a scale from 0 to 1. *hubert-L* and *w2v2-L-robust* show the highest robustness scores for MSP-Podcast and MOSI, *w2v2-L* and *w2v2-L-robust* for IEMOCAP. Averaged over all data sets and dimensions *w2v2-L-robust* and *hubert-L* show the highest robustness scores with .980 and .976. *w2v2-L-xls-r* and *CNN14* show the lowest average robustness scores with .908 and .940. The robustness score averaged over all models and data sets is .961 for arousal, .967 for dominance, and .944 for valence.

For most augmentations, the robustness score averaged over all models, datasets, and dimensions is larger than .97, with .998 for *Append Zeros*, .996 for *Clip*, .978 for *Crop Beginning*, .995 for *Gain*, .995 for *Highpass Filter*, and .994 for *Lowpass Filter*. The two augmentations leading to the strongest changes in model output are *Additive Tone* with an robustness score of .827, and *White Noise* with .863. This is in line with Jaiswal and Provost [38], who found SER models sensitive to environmental noise and reverb.

> The tested models are generally robust against most small perturbations to the input signals, with w2v2-L-robust and hubert-L showing the highest robustness. Only when adding white noise or high frequency tones to the input signals, the output of the models becomes less stable.

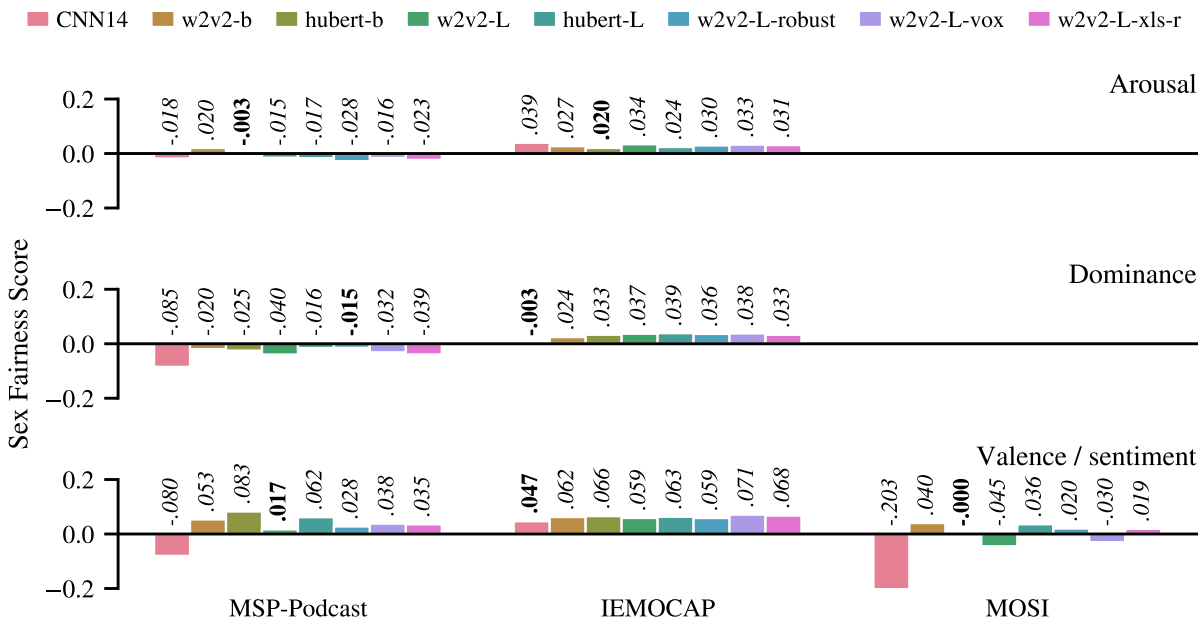


Figure 4: *Sex fairness scores for arousal, dominance, valence (MSP-Podcast / IEMOCAP), and sentiment (MOSI). The sex fairness score is given by $CCC_{female} - CCC_{male}$. A positive value indicates that the model under test performs better for female speaker and a negative value that it performs better for male speaker. A model with desired equal performance would have a sex fairness score of 0.*

4.6 Are the models fair regarding the biological sex of the speaker?

Figure 4 shows sex fairness scores for the speakers in MSP-Podcast, IEMOCAP, and MOSI. As introduced in Section 3.4, the sex 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 speaker than for male, again with the exception of *CNN14*. For sentiment in MOSI, the *CNN14* shows a bias towards better performance for male speaker, whereas all other models show very small biases in different directions.

Averaging over all databases and dimensions the model with the best sex 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 $-.043$.

We also investigated if the models show a bias by predicting higher average values compared to the ground truth for one of the sexes as given by the sex fairness bias value (Section 3.4). The sex fairness bias is in general low, reaching its largest scores with .086 for *w2v2-L* and .066 for *w2v2-L-xls-r* both for valence on MSP-Podcast. On the same database *w2v2-b* shows the largest sex mean shift for arousal (.028) and dominance (.019). For IEMOCAP no model shows a sex fairness bias larger than .007 or smaller than $-.005$. For MOSI, *CNN14* shows the largest sex fairness bias for valence ($-.061$), followed by *w2v2-b* ($-.029$).

Averaging over all databases and dimensions, the models with the best sex fairness bias values are *w2v2-L-robust* and *hubert-L* with $-.003$, followed by *w2v2-L-vox* and *hubert-b* with .006, over *w2v2-b* with .007, *CNN14* with $-.010$, *w2v2-L-xls-r* with .011 up to *w2v2-L* with .014.

Most models show good sex fairness score and sex mean shift values for arousal and dominance. For valence, most models show a higher CCC for females than for males. Overall, *w2v2-L-vox* and *w2v2-L-robust* show the fairest performance.

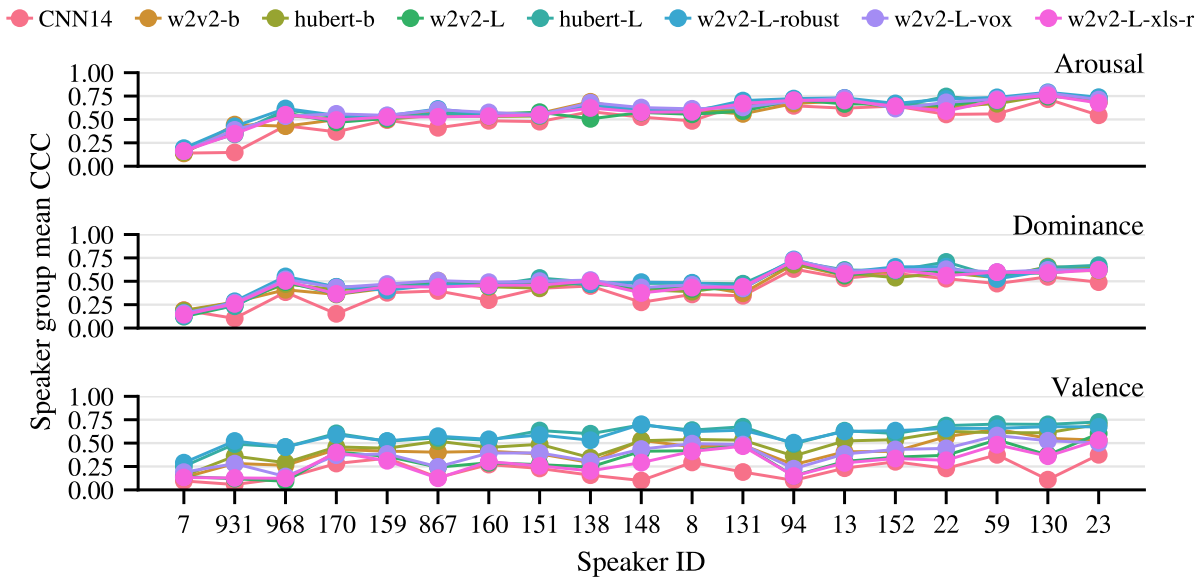


Figure 5: *Speaker-level performance (CCC) on MSP-Podcast for the different models. We only use speakers with at least 200 test set samples for robust CCC estimates. All models show low CCC for at least one speaker on all 3 tasks. Speakers have been ordered according to the mean CCC over all dimensions and models.*

4.7 Is performance equal across all speakers?

The performance of speech processing is dependent on individual speaker characteristics [42]. 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 3.3), the MSP-Podcast test set consists of 12902 samples from 60 speakers; however, the samples are not equally 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. The highest standard deviations of the mean CCC across the 1000 runs are .057, .061, and .064 for arousal, dominance, and valence, respectively.

Our results are presented in Figure 5. For visualisation purposes, we ordered speakers based on the average CCC value over all models and across arousal, dominance, and valence. CCC performance varies across speakers; even for arousal, where the models reach their highest performance, there are speakers for which the models perform well (CCC $>$ .7) and one, for which performance is substantially lower (CCC $<$ .2). The situation is similar for dominance and valence. The CCC shows similar values for the different models, and most the speakers. 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 Figure 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.

> We investigated the performance variation across different speakers in the MSP-Podcast test set and conclude that performance for the best models is similar between most speakers, but can deteriorate to low CCC values for some speakers.

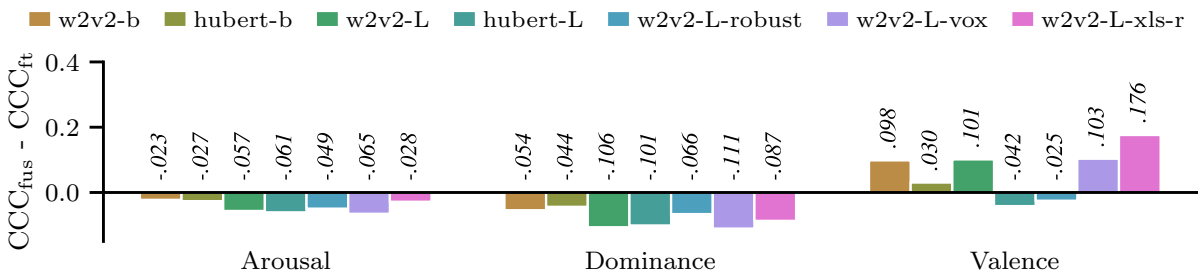


Figure 6: *Text & audio fusion results for arousal, dominance, and valence prediction on MSP-Podcast. Embeddings from the already fine-tuned models are concatenated with BERT embeddings extracted from automatic transcriptions, whereupon a two-layer feed-forward neural network is trained. We show the difference to results with the fine-tuned (ft) models from Figure 2.*

4.8 Does explicit linguistic information further improve performance?

To evaluate whether adding linguistic information improves the predictions, the following experiment is conducted: a regression head is re-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 [63]. The BERT embeddings have a dimensionality of 768 and are extracted from the transcriptions generated by the *wav2vec2-base-960h* speech recognition model[2]. 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, consisting of two layers, where the size of the first layer has exactly the same size as the (fused) embedding space. For training, the weights of both the acoustic and the linguistic transformer models are frozen. The training is done with multi-target CCC-loss for a maximum of 100 epochs. The results on the test partition of MSP-Podcast are evaluated for the model epoch with the highest CCC on the development set.

In Figure 6, we report deviations from the results achieved with the fine-tuned acoustic models alone (cf. Figure 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 actually hampered performance for those two dimensions on MSP-Podcast [4]. 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*, which gives some evidence that it is in particular models pre-trained on multiple languages that gain from a fusion with text features. Given that the employed test set contains only English speech, it can be concluded that adding more in-domain (w. r. t. language) knowledge might be beneficial.

> 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.

5 Analysis

The previous section has provided a holistic evaluation of transformer-based architectures and established their efficacy for dimensional speech emotion recognition. However, there still remain several open questions as to why they are so effective, and in particular why they perform so well for the valence dimension. In this section, we shed more light into this question by examining the embedding space of those models, discussing the importance of fine-tuning, as well as identifying the type of information they use to make their predictions.

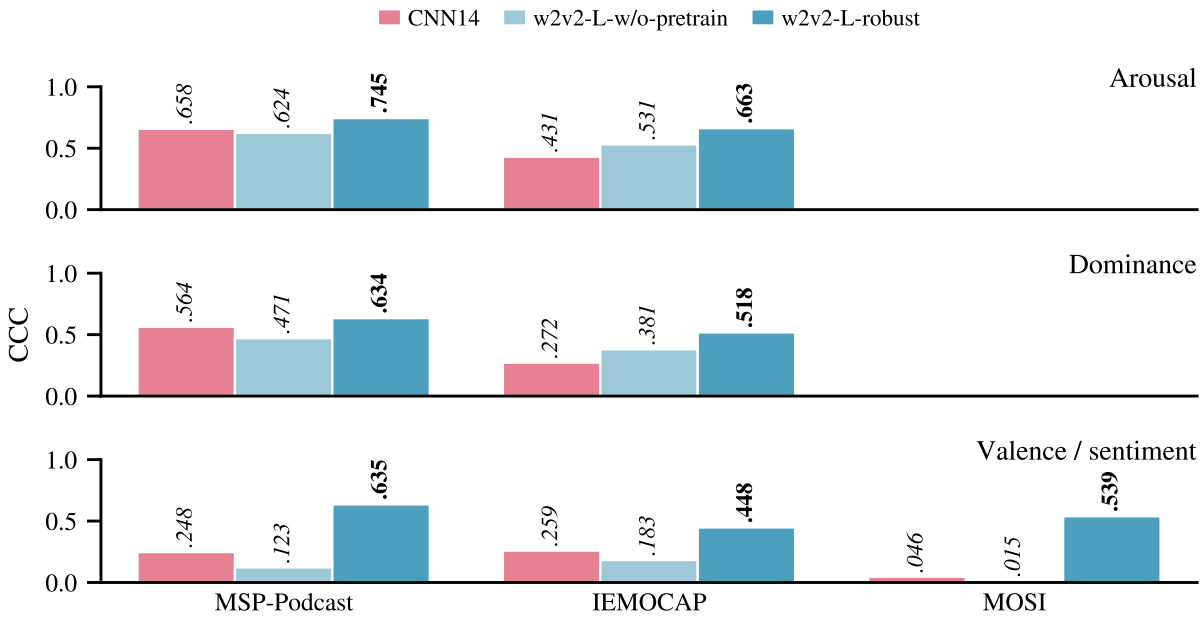


Figure 7: *CCC performance of randomly-initialised wav2vec 2.0 model (w2v2-L-w/o-pretrain) on in-domain and cross-corpus arousal, dominance, valence / sentiment prediction. We compare the performance with that of CNN14 and w2v2-L-robust. We observe that valence and sentiment benefit massively from pre-training, without which wav2vec 2.0 performs worse than a classic CNN approach.*

5.1 Why do transformer-based models generalise so well?

In the previous section, we were able to confirm the superiority of transformer-based models over classic approaches like *CNN14*, especially on valence and in cross-corpus settings. However, even though we saw that the data used for 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*. This enables us to evaluate the impact of pre-training separately from that of the architecture. 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 Figure 7, 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, Baevski *et al.* [21] postulate that the quantisation of latent representations when used as targets for self-supervised 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 Figure 8. The embeddings are projected to two dimensions using t-SNE [64] and different information is chromatically superimposed on them.

For *CNN14*, we can see two main clusters almost perfectly separating the two data sources (MSP-Podcast and IEMOCAP), and several smaller blobs representing gender groups and individual speakers. In fact, in the embeddings of *CNN14*, speaker and domain are more pronounced than the information for valence we actually want to

[2] https://huggingface.co/facebook/wav2vec2-base-960h

[3] Even 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 transformer layers (including the CNN ones). Apart from that we followed the methodology described in Section 3.1.

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

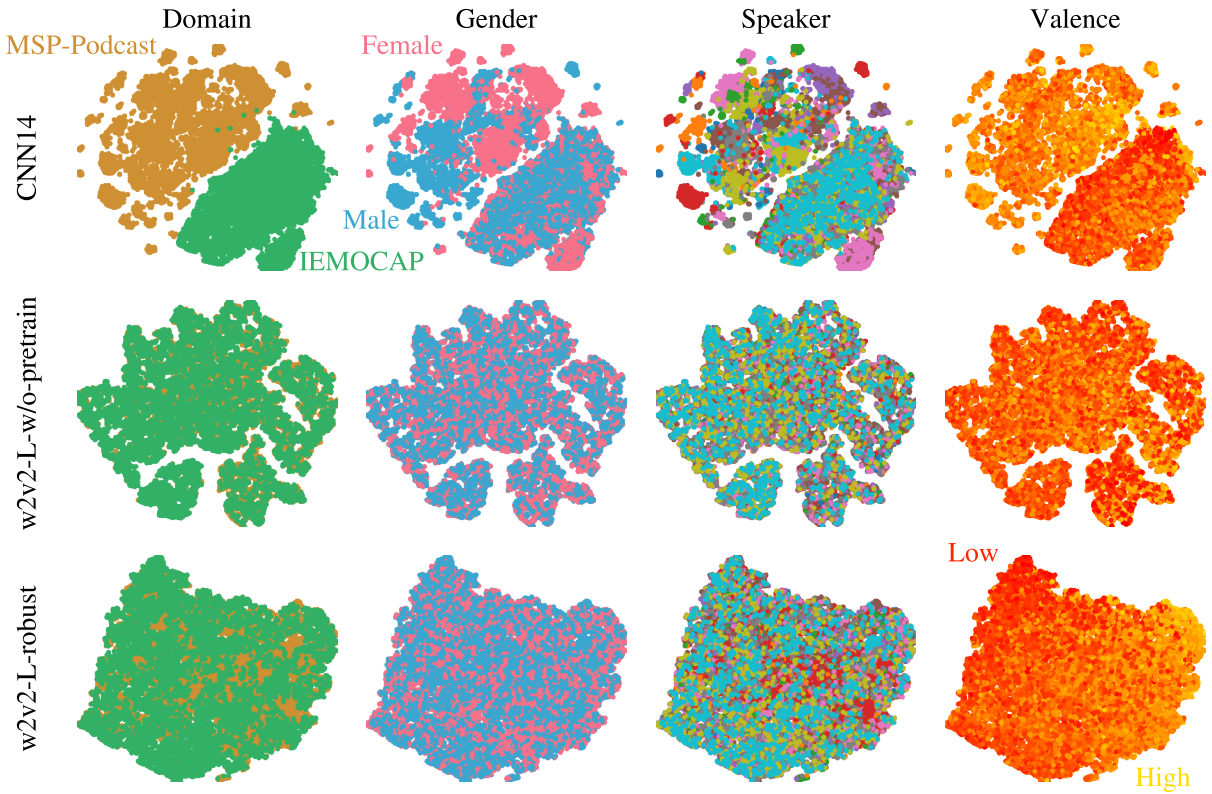


Figure 8: *Visualisation of embeddings extracted with different models overlayed with meta information for a combined dataset of MSP-Podcast and IEMOCAP. We observe that the latent space of wav2vec 2.0 offers a better abstraction from domain, gender, and speaker compared to the* CNN14*baseline – even 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 [64].*

model. Hence, depending on the context, 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. This is independent of starting from a pre-trained state or a random initialisation, indicating that the architecture itself introduces specific inductive biases which are well-suited to learning representations that are not influenced by factors not relevant to the task. This is in line with recent work showing that the inductive biases of transformer-based architectures, and specifically their self-attention layers, are primarily responsible for generalisation, and not large-scale pre-training [65]. Nevertheless, only the pre-trained model (*w2v2-L-robust*) shows a smooth transition from low to high valence scores, showing that pre-training is necessary for good downstream performance. Moreover, the strong speaker dependency presented in Section 4.7 of the models shows that the two dimensional t-SNE visualisations help comparing generalisation abilities between models, but are not necessarily sufficient for deriving conclusions w. r. t. generalisation over different factors.

> Even without pre-training, the latent space provided by the transformer architecture generalises better than CNN14, as it abstracts well away from domain and speaker. Pre-training primarily helps improving the performance for arousal and dominance. In case of valence, however, a pre-training is necessary, as otherwise, prediction fails.

5.2 How important is a fine-tuning of the transformer layers?

So far, in our work, we have fine-tuned all transformer layers along with the added output layer. However, in an attempt to reduce the computational overhead of an experiment, practitioners often choose to use a pre-trained model as a frozen feature extractor, and subsequently train simply the output layer on the generated embeddings. Nevertheless,

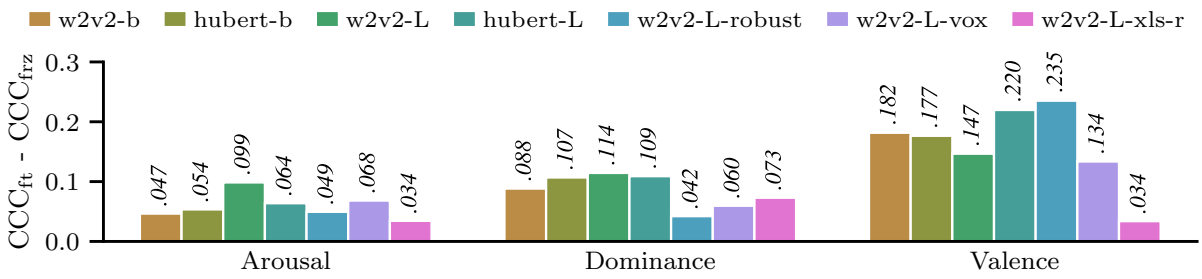


Figure 9: *Difference of fine-tuned (ft) to frozen (frz) CCC performance for arousal, dominance, and valence prediction on MSP-Podcast. The fine-tuned results are from Figure 2, where transformer and output layers are jointly trained. For the frozen results, we keep all transformer layers frozen and simply train the output head. Results show that fine-tuning the transformer layer is worth the computational cost it incurs.*

prior studies have shown that it is necessary to fine-tune several, or sometimes all, layers on the target task to get good downstream performance [23, 45, 66]. Moreover, previous work on convolutional neural networks (CNNs) has shown that earlier layers see more adaptation than later ones, which potentially explains the need to fine-tune all of them to get good performance [44]. In this sub-section, we investigate whether this is also needed for the models investigated here. 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.

Figure 9 shows a comparison between CCC values obtained with these two settings, where we show the difference between fine-tuned and frozen CCC. We observe a large performance gain for valence, and a lesser, but still considerable one for dominance, while arousal is less susceptible to adaptation of the transformer layers. On valence, we see a maximum improvement for *w2v2-L-robust* by .235 demonstrating that fine-tuning of the transformer layers is necessary and worth the computational cost it incurs. Moreover, the models that see the biggest performance gain due to an adaptation of the self-attention layers are *hubert-L* and *w2v2-L-robust*. In Section 4.8, these models were found to benefit least from additional text information in the form of BERT embeddings. 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.

> 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 least from a fusion with text.

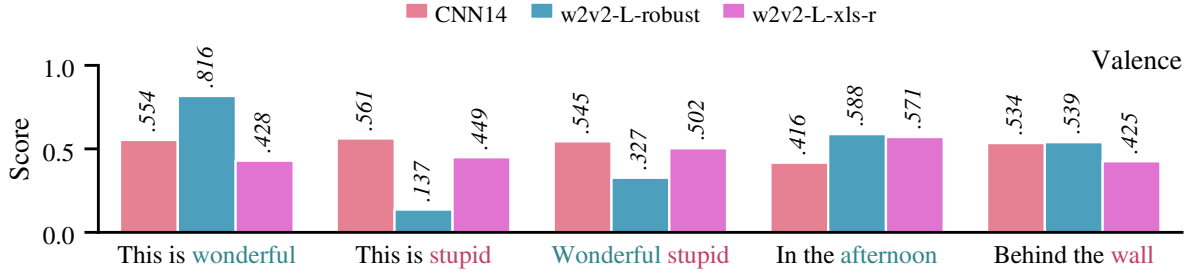
5.3 Do the models implicitly learn linguistic information?

In Section 4.8, we investigated to what extent the models benefit from a fusion with textual information; and, we made the surprising observation it only improved some models like the multi-lingual *w2v2-L-xls-r*. Actually, even with linguistic information, *w2v2-L-xls-r* still performs worse than the mono-modal *w2v2-L-robust*. In the previous section, we then showed that a fine-tuning of the transformer layers is required to achieve a high performance on the valence dimension. These findings suggest that during the fine-tuning, the models implicitly learn sentiment.

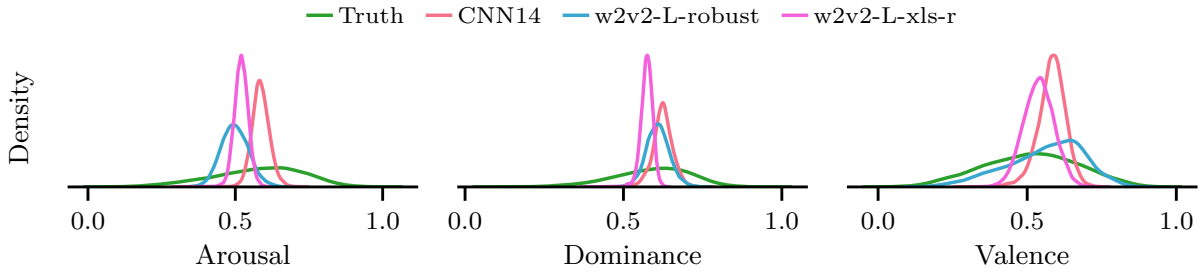
To prove this assumption, we conducted the following experiment. From the manual transcriptions of the training set of MSP-Podcast, we selected from all words with at least ten occurrences those 50 words that have, on average, the lowest (negative) / highest (positive) rating with respect to valence (see wordclouds in Figure 10a). We then picked four words and made up short sentences, which we transformed into audio files with a text-to-speech engine[5] and predicted their valence score. Since the synthesised files sound neutral, we would expect values around .5. As we see in Figure 10b, this is in fact the case for *CNN14* and *w2v2-L-xls-r*. But, *w2v2-L-robust* predicts the sentence *"This is wonderful"* as clearly positive (.816) and *"This is stupid"* as clearly negative (.137). This we can only explain with the fact that the model is not just listening to the paralinguistic part of the message, but also takes linguistic information into account. Basically, it means that transformer models are capable of detecting key words (e. g. *"wonderful"* and *"stupid"*) and assigning them a positive or negative meaning. And, once again it seems that the pre-training data determines to what extent a model is able to learn such relations. In Section 4.8, we speculated that models pre-

[5]We use the publicly available gTTS (Google Text-to-Speech): https://pypi.org/project/gTTS/

(a) *Word clouds with the 50 most positively / negatively rated words from the MSP-Podcast training set. The size of the words expresses their frequency.*



(b) *Valence predictions for exemplary synthesised sentences including positive and negative words from the word clouds. As the TTS system is not expressing emotion in its paralinguistic stream, the variation in prediction is mostly attributed to the linguistic information. The examples are meant to demonstrate the relative sensitivity of w2v2-L-robust to linguistic content relative to CNN14 and w2v2-L-xls-r.*



(c) *Distribution of predictions on the synthesised MSP-Podcast test set. We compare the ground truth distribution with the predictions obtained for the synthesised transcriptions created with a text-to-speech engine. For the synthesised sentences, we expect a spiky Gaussian near .5, as seen for arousal and dominance. For valence, however, we get a much flatter distribution for w2v2-L-robust, which can be explained by its sensitivity to the linguistic content of an utterance.*

Figure 10: *To find out if transformer-based models are able to learn sentiment from audio alone, we investigate their performance on neutral sounding sentences generated with a TTS engine.*

trained on multiple languages have a disadvantage when evaluated on English only. When we compare the response of *w2v2-L-xls-r* and *w2v2-L-robust* in Figure 10b, we can now provide a convincing explanation: since *w2v2-L-xls-r* is less specialised on English speech it is not able to learn linguistic content in the same way as *w2v2-L-robust*, which leads to a lower performance on valence.

The wordclouds in Figure 10a also include words we would not necessarily perceive as positive or neutral, probably a side-effect of the context in which they occur. This, of course, bears the risk that the model learns unwanted associations. However, as we see in Figure 10b, *w2v2-L-robust* does not perceive all words from the wordclouds as positive or negative. For instance, the sentences including *"afternoon"* and *"wall"* are still predicted as rather neutral.

To asses 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.[6] In Figure 10c, we plot the predictions on the synthesised files and compare

[6]Partial 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.

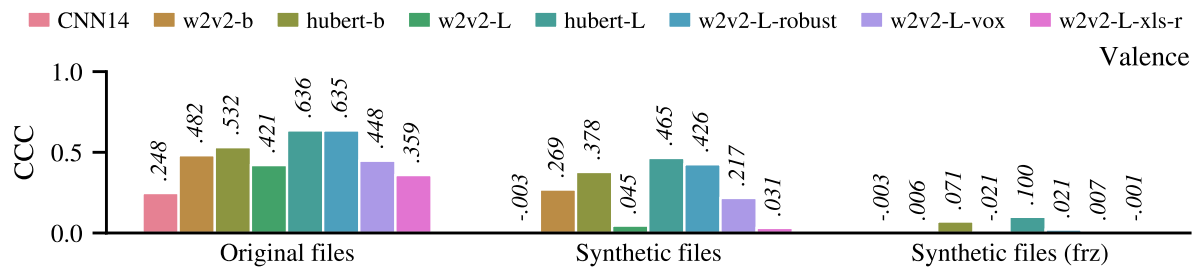


Figure 11: *CCC performance for valence on the original and synthetic files on MSP-Podcast. We see that models with a high performance on the original files are more sensitive to sentiment (cf. left and center section). To prove that a fine-tuning of the transformer layers is required to learn linguistic content, we additionally show the correlation for models where the transformer layers were frozen (frz) during training (cf. Section 5.2).*

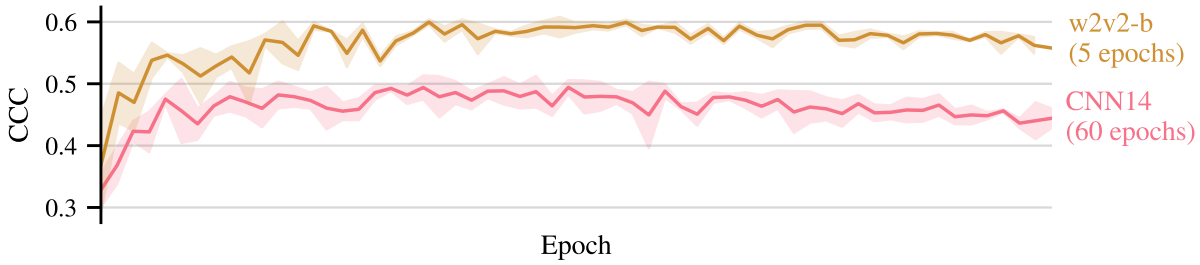


Figure 12: *Mean and standard deviation of development set performance on MSP-Podcast across three training runs. Compared to CNN14, w2v2-b requires less steps to converge and already shows lower fluctuation after the first epoch. To compensate for the fewer number of epochs in case of w2v2-b we run the evaluation every 100 steps, which corresponds to 12 measurements per epoch.*

them to the ground truth of the original files. While *CNN14* and *w2v2-L-xls-r* show the expected Gaussian peak near .5 across all dimensions, for valence, *w2v2-L-robust* generates a flatter distribution resembling that of the ground truth. We can take this as evidence that *w2v2-L-robust* is indeed sensitive to the linguistic content of an utterance.

In Figure 11, 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 Figure 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 5.2), correlation drops to almost zero.

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

6 Efficiency

For our last experimental evaluation, we focus on the efficiency of transformers as foundation models. We concentrate on three facets of efficiency: *optimisation stability*, *computational complexity*, and *data efficiency*, and show how we can improve on all three.

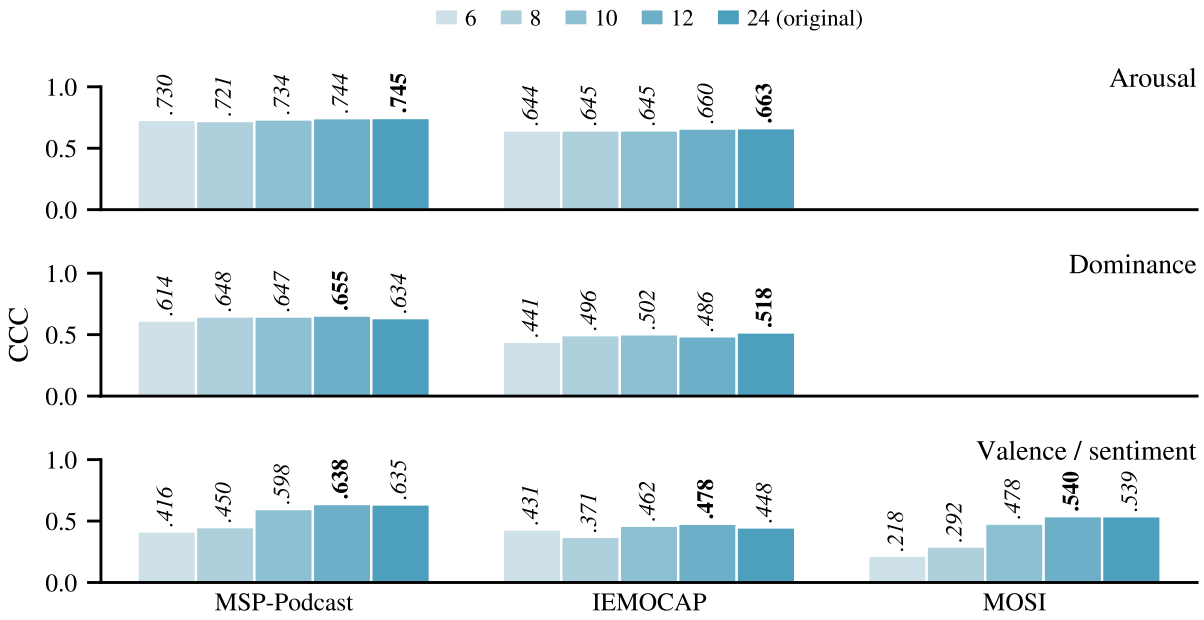


Figure 13: *CCC scores for arousal, dominance, and valence / sentiment for w2v2-L-robust and pruned versions. The legend shows the number of bottom layers kept during fine-tuning. We see that half of the layers can be removed without any loss in performance.*

6.1 Does pre-training help with training stability and convergence?

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 training runs starting from different random seeds. Starting from pre-trained weights, however, we expect less volatility [67, 68]. Figure 12 shows the mean and standard deviation over the performance on the development set across three trials for *CNN14* and *w2v2-b*. For *CNN14*, we observe almost constant jittering across all 60 epochs. For *w2v2-b*, we have a quite different picture: the model converges faster and we can reduce the number of epochs to 5. Except for the first epoch, when the effect of randomised batches leads to some variability, we observe a steady performance between different runs. When starting from a pre-trained transformer model, it therefore seems reasonable to run on a limited number of epochs and report results from a single run.

> Starting from a pre-trained model reduces the number of epochs needed to converge and improves performance stability across training runs with different seeds.

6.2 How many transformer layers do we really need?

In Section 4.4, we mentioned that *w2v2-b* and *hubert-b*, both having 12 transformer layers, outperform some of the large models with 24 transformer layers. 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 reduce the size of a model to some extent without losing performance.

Sajjad *et al.* [69] 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 Figure 13, we report the effect on CCC scores for *w2v2-L-robust* (our overall best performing model). The results show that half of the layers can be removed without a loss in performance. Except for a slightly higher number of hidden units (1024 instead of 768), the resulting 12-layer version of *w2v2-L-robust* – in the following denoted as *w2v2-L-robust-12* – resembles exactly the architecture of *w2v2-b*. Only with 10 or less layers we actually begin to

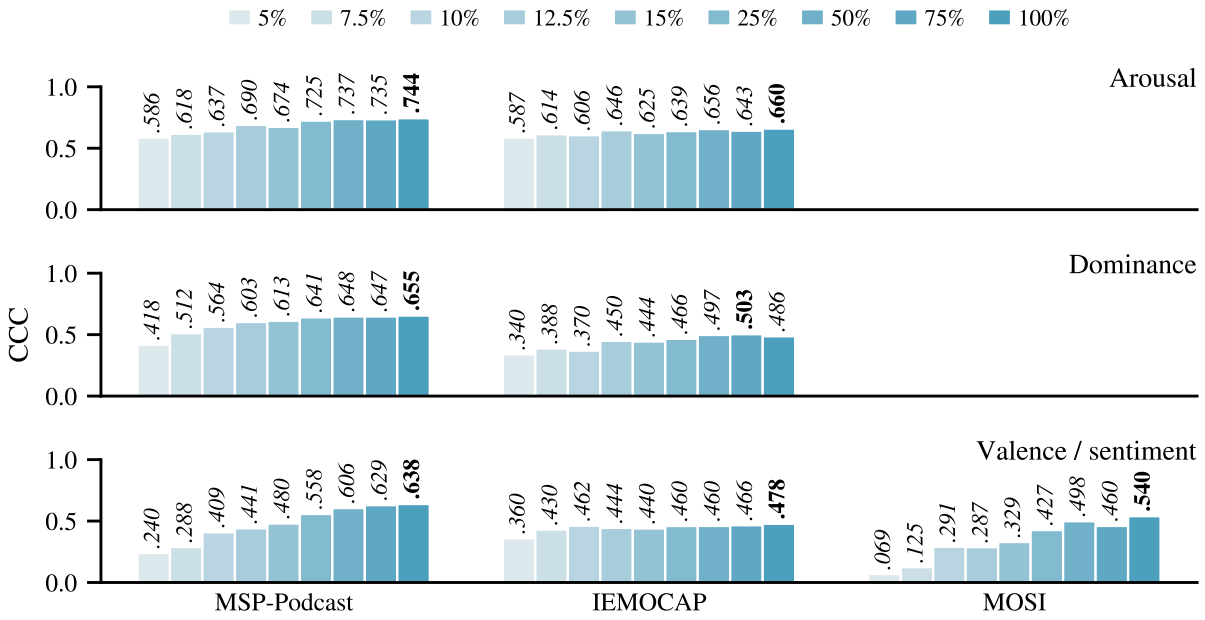


Figure 14: *CCC scores for arousal, dominance, and valence / sentiment for w2v2-L-robust on sparse training data. The legend shows the fraction of data used for fine-tuning. Please note that steps are not linear.*

see a drop for valence / sentiment on IEMOCAP and MOSI. For arousal and dominance, we still achieve a steady performance with only 8 layers.

> 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.

6.3 Can we reduce the training data without a loss in performance?

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). For instance, we randomly drop 95% of the training set and build the model with the remaining 5%. We leave development and test set untouched.

In Figure 14 we report results for such a sparse training. For efficiency, we start from the reduced 12-layer architecture and therefore compare results to *w2v2-L-robust-12* (cf. Section 6.2). Again, we show CCC scores for the three dimensions on MSP-Podcast (in-domain), as well as, IEMOCAP and MOSI (cross-domain). At a first glance, we see that there is no noteworthy degradation when sticking to the full training set. The only exception is dominance on IEMOCAP, where we achieve best results with just 75% of the data. For these dimensions, however, we actually seem to reach saturation already at 25% yielding a performance loss of less than .02 on MSP-Podcast, whereas, in case of 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 narrow when reaching 75% of the data. For MOSI, we even see a boost in CCC of almost .1 for the remaining 25%. However, in the light of our findings from Section 5.3, namely that performance on valence depends on the ability to predict sentiment, this does not come as a surprise. Providing more linguistic content makes it more likely a model can detect associations between key words and emotional context. What is a surprise, though, is that on IEMOCAP, with just 7.5% of the data, we miss performance on the full training set by less than .05. A possible explanation is that the vocabulary of IEMOCAP does not resemble that of MSP-Podcast very much and that therefore, the impact of linguistic information is limited. This would also explain why the differences in performance on valence are less pronounced for IEMOCAP compared to the other two databases (cf. Figure 2).

A reduction of training samples without loss in performance is only possible for arousal and dominance. With respect to valence, there is no sweet point in our data.

7 Summary and outlook

We have explored the use of (pre-trained) transformer-based architectures for speech emotion recognition from a set of different perspectives. In the previous sections, we have dealt with several probing questions in isolation, in an attempt to identify the most pertinent factors to each different aspect of our investigation. We now attempt a unified summary by collectively considering all findings presented in the preceding sections.

- *Effect of pre-training:* we have determined that pre-training is essential to get good performance (Section 4.3), and in particular for the valence dimension. This is particularly evident when training the wav2vec 2.0 from a random initialisation (Section 5.1): 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 6.1), thus resulting in models which require less iterations, and less data to train on (Section 6.3). However, we were unable to determine a clear relationship of the form 'more pre-trained 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.

- *Generalisation:* transformer-based models show very good cross-corpus generalisation (Section 4.3), robustness to small perturbations (Section 4.5), and appear invariant to domain, speaker, and gender characteristics (Section 5.1). 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 type of architecture used, rather than the form of pre-training data and regiment, as they are also evident in models initialised from random weights (Section 5.1). This finding has been observed in other domains and remains under active investigation by the community [65]. In the context of this work, we showed that several self-attention layers can be removed without hampering downstream performance (Section 6.2) – an indication that they may not be needed for good downstream performance (though they might still be necessary for successful pre-training).

- *Fairness:* fairness remains a challenging topic for contemporary machine learning architectures, SER ones included. Community discussions primarily concern the issue of *group fairness*. In the present, we investigate this for the only group variable available in our datasets: biological sex (Section 4.6), where we observe that transformer-based architectures are fairer than the CNN14 baseline. However, we argue that *individual fairness* is important for SER (and machine learning tasks pertaining to human behaviour analysis in general). This refers to how the models perform across different speakers; a feat which proves challenging even for the top-performing models investigated here (Section 4.7). 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 [41, 42].

- *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 5.3) and fail to benefit from fusion with explicit textual information (cf. Section 4.8). Interestingly, this is only possible when fine-tuning the self-attention layers (Section 5.2), as keeping them frozen results to complete failure for synthesised speech (Section 5.3). 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). 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 [70, 71]. 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 4.1).

8 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 [26][7]. Transformer architectures are more robust to small perturbations, fair on the (biological sex) 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.

9 References

[1] B. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] P. Ekman, "An argument for basic emotions," *Cognition & 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," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[4] A. Triantafyllopoulos, U. Reichel, S. Liu, S. Huber, F. Eyben, and B. W. Schuller, "Multistage linguistic conditioning of convolutional layers for speech emotion recognition," *arXiv preprint arXiv:2110.06650*, 2021.

[5] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 7–19, 2010.

[6] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010.

[7] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decision Support Systems*, vol. 115, pp. 24–35, 2018.

[8] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Messner, E. Cambria, G. Zhao, and B. W. Schuller, "The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress," in *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, Chengdu, China: ACM, 2021, pp. 5706–5707.

[9] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece: IEEE, 2018, pp. 112–118.

[10] 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 *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria: ISCA, 2019, pp. 3302–3306.

[11] S. Amiriparian, A. Sokolov, I. Aslan, L. Christ, M. Gerczuk, T. Hübner, D. Lamanov, M. Milling, S. Ottl, I. Poduremennykh, E. Shuranov, and B. W. Schuller, "On the impact of word error rate on acoustic-linguistic speech emotion recognition: An update for the deep learning era," *arXiv preprint arXiv:2104.10121*, 2021.

[12] C. Oates, A. Triantafyllopoulos, I. Steiner, and B. W. Schuller, "Robust speech emotion recognition under different encoding conditions," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria: ISCA, 2019, pp. 3935–3939.

[13] 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 *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria: ISCA, 2019, pp. 1691–1695.

[14] A. Batliner, S. Hantke, and B. W. Schuller, "Ethics and good practice in computational paralinguistics," *IEEE Transactions on Affective Computing*, 2020.

[15] 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 Processing Magazine*, vol. 38, no. 6, pp. 39–49, 2021.

[16] S. Srinivasan, Z. Huang, and K. Kirchhoff, "Representation learning through cross-modal conditional teacher-student training for speech emotion recognition," *arXiv preprint arXiv:2112.00158*, 2021.

[17] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane, "Gender de-biasing in speech emotion recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria: ISCA, 2019, pp. 2823–2827.

[18] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[7] https://github.com/audeering/w2v2-how-to

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the International Conference on Machine Learning (ICML)*, Vienna, Austria (virtual), 2020, pp. 1597–1607.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, 2020, pp. 12 449–12 460.

[22] 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 Processing*, vol. 29, pp. 3451–3460, 2021.

[23] 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," *arXiv preprint arXiv:2111.02735*, 2021.

[24] S. Ottl, S. Amiriparian, M. Gerczuk, V. Karas, and B. Schuller, "Group-level speech emotion recognition utilising deep spectrum features," in *Proceedings of the 8th EmotiW – Emotion Recognition In The Wild Challenge (EmotiW 2020), 22nd ACM International Conference on Multimodal Interaction (ICMI)*, Utrecht, The Netherlands: ACM, 2020, pp. 821–826.

[25] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 12, 2021.

[26] 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*, 2022. DOI: 10.5281/zenodo.6221127.

[27] D. N. Krishna, "Using large pre-trained models with cross-modal attention for multi-modal emotion recognition," *arXiv preprint arXiv:2108.09669*, 2021.

[28] J. Yuan, X. Cai, R. Zheng, L. Huang, and K. Church, "The role of phonetic units in speech emotion recognition," *arXiv preprint arXiv:2108.01132*, 2021.

[29] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, *Superb: Speech processing universal performance benchmark*, 2021.

[30] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3400–3404, 2021.

[31] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition," *arXiv preprint arXiv:2110.06309*, 2021.

[32] M. R. Makiuchi, K. Uto, and K. Shinoda, "Multimodal emotion recognition with high-level speech and text features," *arXiv preprint arXiv:2111.10202*, 2021.

[33] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[34] 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, 2019.

[35] M. Li, B. Yang, J. Levy, A. Stolcke, V. Rozgic, S. Matsoukas, C. Papayiannis, D. Bone, and C. Wang, "Contrastive unsupervised learning for speech emotion recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, ON, Canada: IEEE, 2021, pp. 6329–6333.

[36] S. Padi, S. O. Sadjadi, R. D. Sriram, and D. Manocha, "Improved speech emotion recognition using transfer learning and spectrogram augmentation," in *Proceedings of the International Conference on Multimodal Interaction*, 2021, pp. 645–652.

[37] M. Xu, F. Zhang, X. Cui, and W. Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Ontario, Canada, 2021, pp. 6319–6323.

[38] M. Jaiswal and E. M. Provost, "Best practices for noise-based augmentation to improve the performance of emotion recognition "in the wild"," *arXiv preprint arXiv:2104.08806*, 2021.

[39] R. Pappagari, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, "Copypaste: An augmentation method for speech emotion recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6324–6328.

[40] E. Parada-Cabaleiro, A. Baird, A. Batliner, N. Cummins, S. Hantke, and B. Schuller, "The perception of emotions in noisified nonsense speech," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden: ISCA, Aug. 2017, pp. 3246–3250.

[41] S. S. Rajan, S. Udeshi, and S. Chattopadhyay, "Aequevox: Automated fairness testing of speech recognition systems," *arXiv preprint arXiv:2110.09843*, 2021.

[42] 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 *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia: ISCA, 1998, pp. 1–4.

[43] G. Trigeorgis, F. Ringeval, R. Brückner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China: IEEE, 2016, pp. 5200–5204.

[44] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, ON, Canada: IEEE, 2021, pp. 7268–7272.

[45] 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.

[46] 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," *arXiv preprint arXiv:2104.01027*, 2021.

[47] C. Wang, M. Riviere, 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 *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Bangkok, Thailand, 2021, pp. 993–1003.

[48] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[49] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[50] M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 101–111, 2014.

[51] L. Tian, C. Lai, and J. Moore, "Polarity and intensity: The two aspects of sentiment analysis," in *Proceedings of the Grand Challenge and Workshop on Human Multimodal Language*, Melbourne, Australia: ACL, 2018, pp. 40–47.

[52] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, "Machine learning testing: Survey, landscapes and horizons," *IEEE Transactions on Software Engineering*, vol. 48, no. 1, pp. 1–36, 2020.

[53] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[54] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, *et al.*, "Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proceedings of the 2018 on audio/visual emotion challenge and workshop*, 2018, pp. 3–13.

[55] S. Zheng, Y. Song, T. Leung, and I. Goodfellow, "Improving the robustness of deep neural networks via stability training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, 2016, pp. 4480–4488.

[56] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of NLP models with checklist," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, virtual: ACL, 2020, pp. 4902–4912.

[57] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, 2018.

[58] J. K. Fitzsimons, A. A. Ali, M. A. Osborne, and S. J. Roberts, "Equality constrained decision trees: For the algorithmic enforcement of group fairness," *arXiv preprint arXiv:1810.05041*, 2018.

[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," *Science Robotics*, 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 *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Shenzhen, China: IEEE, 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," *arXiv preprint arXiv:2201.07876*, 2022.

[62] A. D'Amour *et al.*, "Underspecification presents challenges for credibility in modern machine learning," *arXiv preprint arXiv:2011.03395*, 2020.

[63] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, Minneapolis, MN, USA: Association for Computational Linguistics (ACL), 2019, pp. 4171–4186.

[64] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research (JMLR)*, vol. 9, pp. 2579–2605, 2008.

[65] Y. Bai, J. Mei, A. L. Yuille, and C. Xie, "Are transformers more robust than cnns?" In *Advances in Neural Information Processing Systems (NeurIPS)*, Sydney, Australia (virtual), 2021, pp. 1–13.

[66] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain: IEEE, 2020, pp. 6419–6423.

[67] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?" In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Sardinia, Italy: PMLR, 2010, pp. 201–208.

[68] B. Neyshabur, H. Sedghi, and C. Zhang, "What is being transferred in transfer learning?" In *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, 2020, pp. 512–523.

[69] H. Sajjad, F. Dalvi, N. Durrani, and P. Nakov, "Poor man's BERT: smaller and faster transformer models," *arXiv preprint arXiv:2004.03844*, 2020.

[70] D. Wilson and D. Sperber, "Explaining irony," *Meaning and relevance*, pp. 123–145, 2012.

[71] F. Burkhardt, B. Weiss, F. Eyben, J. Deng, and B. Schuller, "Detecting vocal irony," in *Language Technologies for the Challenges of the Digital Age*, G. Rehm and T. Declerck, Eds., Cham: Springer International Publishing, 2018, pp. 11–22.