Interspeech 2022
18-22 September 2022, Incheon, Korea

Probing Speech Emotion Recognition Transformers for Linguistic Knowledge

Andreas Triantafyllopoulos [1], Johannes Wagner [2], Hagen Wierstorf [2], Maximilian Schmitt [2], Uwe Reichel [2], Florian Eyben [2], Felix Burkhardt [2], Björn W. Schuller [1,2,3]

[1] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2] audEERING GmbH, Gilching, Germany
[3] GLAM – Group on Language, Audio, & Music, Imperial College, UK

andreas.triantafyllopoulos@uni-a.de

Abstract

Large, pre-trained neural networks consisting of self-attention layers (transformers) have recently achieved state-of-the-art results on several speech emotion recognition (SER) datasets. These models are typically pre-trained in self-supervised manner with the goal to improve automatic speech recognition performance – and thus, to understand linguistic information. In this work, we investigate the extent in which this information is exploited during SER fine-tuning. Using a reproducible methodology based on open-source tools, we synthesise prosodically neutral speech utterances while varying the sentiment of the text. Valence predictions of the transformer model are very reactive to positive and negative sentiment content, as well as negations, but not to intensifiers or reducers, while none of those linguistic features impact arousal or dominance. These findings show that transformers can successfully leverage linguistic information to improve their valence predictions, and that linguistic analysis should be included in their testing.

Index Terms: speech emotion recognition, transformers

1. Introduction

Recently, deep neural networks (DNNs) consisting of self-attention layers (i. e., *transformers*) have provided state-of-the-art results for speech emotion recognition (SER) and have substantially improved valence prediction [1, 2, 3, 4, 5]. These models are typically pre-trained on large corpora in a self-supervised fashion, with the main goal of improving automatic speech recognition performance; thus, they capture a large amount of linguistic information that is beneficial for that task. Accordingly, valence is often easier to predict from text rather than audio information [6, 7]. This raises the question whether transformer models fine-tuned for SER partially rely on that information for improving their valence performance, as opposed to utilising exclusively paralinguistic cues. Furthermore, if transformers turn out to leverage linguistic information, they might also be fallible to the same risks and biases faced by natural language processing (NLP) models [8, 9, 10].

Preliminary findings indicate that transformers make use of linguistic information for valence prediction; we found that WAV2VEC2.0 variants fine-tuned to predict arousal, valence, and dominance retain a large part of their valence, but not its arousal or dominance performance when tested on neutral speech synthesised from transcriptions [1]. The models additionally exhibited high reactivity to the sentiment of the text in their valence predictions. Interestingly, both these trends were only evident after fine-tuning the self-attention layers for SER, and not when simply training an output head on the pre-trained embeddings. Moreover, the fine-tuning of those layers proved crucial in obtaining state-of-the-art valence recognition. Overall, this suggests that linguistic information is needed for obtaining good valence performance (while not so for arousal or dominance) and that this information is uncovered by fine-tuning the transformer layers.

Previous works that analysed the representations of WAV2VEC2.0 lend further evidence to the hypothesis that its intermediate layers contain traces of linguistic information. These works rely on the process of *feature probing* [11, 12, 13, 14], whereby a simple model (i. e., probe) is trained to predict interpretable features using the intermediate representations of the model to be tested. For example, Shah *et al.* [13] found evidence of linguistic knowledge in the middle and deeper layers of the *base* WAV2VEC2.0 model (*w2v2-b*), with acoustic knowledge being more concentrated in the shallower layers. For the *large* variant (*w2v2-L*), Pasad *et al.* [12] found that it follows a similar pattern, with shallower layers focusing more on acoustic properties and middle ones on linguistics; however, the trend is reversed towards the last layers with the transformer layers exhibiting an autoencoder-style behaviour and reconstructing their input, thus placing again an emphasis on acoustics. This is consistent with the pre-training task of WAV2VEC2.0 [15], masked token prediction, which tries to reconstruct the (discretised) inputs of the transformer. They further found that automatic speech recognition (ASR) fine-tuning breaks this autoencoder-style behaviour by letting output representations deviate from the input to learn task-specific information.

The main contribution of this work relies on providing comprehensive, reproducible probing processes based on publicly-available tools with an emphasis on the linguistic information learnt by SER models. It is based on three probing methodologies: (a) re-synthesising speech signals from automatic transcriptions of the test partition using ESPNET [16, 17], (b) using CHECKLIST [18] to generate a test suite of utterances that contain text-based emotional information, which is also synthesised using ESPNET, and (c) feature probing, where we follow the work of Shah *et al.* [13] to detect traces of acoustic and linguistic knowledge in the intermediate representations of our model. We use this process to characterise the behaviour of our recent, state-of-the-art SER model [1], and contrast it to the behaviour of the original embeddings (i. e., freezing the transformer layers) in order to better understand the impact of fine-tuning. In particular, this lets us investigate whether the fine-tuning is necessary for adapting to acoustic mismatches between the pre-training and downstream domains, as previously shown for convolutional neural networks (CNNs) [19], or to better leverage linguistic information. This type of behavioural testing goes beyond past work that typically investigates SER models' robustness with respect to noise and small perturbations [20, 21, 22] or fairness [23, 24], thus, providing better insights into the inner workings of SER models.

10.21437/Interspeech.2022-10371

2. Methodology

2.1. Model training

We begin by briefly describing the process used to train the models probed here. More details can be found in Wagner *et al.* [1]. The model follows the *w2v2-L* architecture [15] and has been pre-trained on 63k hours of data sourced from 4 different corpora, resulting in a model which we refer to as *w2v2-L-robust* [25]. *w2v2-L-robust* is adapted for prediction by adding a simple head, consisting of an average pooling layer which aggregates the embeddings of the last hidden layer, and an output linear layer. It is then fine-tuned on multitask emotional dimension prediction (arousal, valence, and dominance) on MSP-Podcast (v1.7) [26]. The dataset consists of roughly 84 hours of naturalistic speech from podcast recordings. The original labels are annotated on a 7-point Likert scale, 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 male). The samples have a combined length of roughly 21 hours, and vary between 1.92 s and 11.94 s per sample.

For fine-tuning on the downstream task, we use the Adam optimiser with concordance correlation coefficient (CCC) loss, which is commonly used as loss function for dimensional SER [27], 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. Training instances are cropped/padded to a fixed length of 8 seconds.

In order to understand the impact of fine-tuning several layers, we experiment with two variants: *w2v2-L-emo-frz* and *w2v2-L-emo-ft*. Both are using the same output head, but for the former, we only train this added head, whereas for the latter, we additionally fine-tune the transformer layers (while always keeping the original CNN weights). According to Wang *et al.* [3], such a partial fine-tuning yields better results than a full fine-tuning including the CNN-based feature encoder. These models are trained using a single random seed, for which the performance is reported, as we found that fine-tuning from a pre-trained state leads to stable training behaviour [1].

2.2. Probing #1: Re-synthesised transcriptions

The first probing experiment is motivated by Wagner *et al.* [1], where we synthesised neutral-sounding speech using the transcriptions of MSP-Podcast. In the present work, instead of using Google Text-to-Speech and the manual transcriptions (which cover only a subset of the dataset), we use open-source tools to automatically transcribe and re-synthesise each utterance. For transcriptions, we use the *wav2vec2-base-960h* speech recognition model.[1] While these are less accurate than manual transcriptions (word error rate on the 50 334 transcribed samples is 34.7%) , they have the added benefit of a) covering the entire dataset, and b) allowing us to extract linguistic features for probing (Section 2.4). The resulting transcriptions are synthesised using a transformer model trained with a guided attention loss and using phoneme inputs [17], which gave the highest MOS scores when trained on LJ Speech [28]. This model is freely available through ESPNET[2] [16, 17].

ESPNET is able to synthesise realistic-sounding neutral speech which contains some prosodic fluctuations resulting from sentence structure, but this variation is not (intentionally) carrying any emotional information, as it has only been trained to synthesise the target utterance agnostic to emotion. Thus, on average, any emotion would manifest only in the text, rather than in the paralinguistics; and, therefore, any SER model that performs well on the resulting samples would have to utilise linguistic information. This is tested by computing the CCC performance on the synthesised samples.

2.3. Probing #2: CHECKLIST and TTS

We further use the CHECKLIST toolkit[3] [18] as another means of gauging model dependence on linguistics. CHECKLIST is a toolkit which allows the user to generate automatic tests for NLP models. It contains a set of expanding templates which allows the user to fill in keywords and automatically generates test instances for the intended behaviour. Ribeiro *et al.* [18] use this functionality to benchmark several NLP models, including sentiment models, and measure their success and failure rate. They have used data from the airlines domain (e. g., "That was a wonderful aircraft."). To be comparable with previous work, we use the test suites generated by the original authors. Even though this does not fit the domain our models were trained on (podcast data), we expect our model to generalise to a reasonable extent.

CHECKLIST works by generating a set of test sentences. However, our models rely on the spoken word. We thus use ESPNET to synthesise them. The same considerations as in Section 2.2 apply here; namely, that any emotional information will be attributable to text, rather than acoustics.

Contrary to Ribeiro *et al.* [18], we do not use CHECKLIST for *testing*, but for *probing*. That is, we do not a-priori expect the model to produce a specific outcome (e. g., high valence for positive words). Rather, we investigate what the model predicts in an attempt to better gauge its reliance on linguistic content for making predictions. Therefore, any emotional information is (on average) explicitly attributed to linguistics.

Our probing begins with negative, neutral, and positive words in isolation (e. g., "dreadful", "commercial", "excellent"); this tests the behaviour of the models when the relationship of linguistics to sentiment is straightforward. Then, we introduce context (e. g., "That was a(n) dreadful/commercial/excellent flight"); this does not influence the sentiment, but adds more prosodic fluctuation as the utterances become longer. As a more fine-grained test, we add intensifiers/reducers (e. g., "That was a really/somewhat excellent flight") to positive/negative phrases, which are expected to impact (increase/decrease) valence and, potentially, arousal. Finally, we add negations to negative, neutral, and positive words in context; this inverts the sentiment for negative/positive and leaves neutral unchanged. Note that the sentiment test suite proposed by Ribeiro *et al.* [18] includes additional tests (e. g., for robustness to small variations in the text or fairness with respect to linguistic content). These we exclude, as we do not consider them relevant for our main question, which is whether (and to what extent) our models rely on linguistics to make their predictions.

2.4. Probing #3: Feature probing

Feature probing has emerged as an interesting paradigm for understanding what auditory DNNs are learning [11, 12, 13]. In the present study, we follow the recipe of Shah *et al.* [13]. We train a 3-layer feed-forward neural network (with hidden sizes [768, 128], Adam optimiser with learning rate 0.0001, batch size of 64, 100 epochs, and exponential learning rate on validation loss plateau with a factor of 0.9 and a patience of 5

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

[2] https://github.com/espnet/espnet

[3] https://github.com/marcotcr/checklist

Table 1: *CCC for (A)rousal, (V)alence, and (D)ominance prediction when evaluating* w2v2-L-emo-frz *and* w2v2-L-emo-ft *on the original test recordings of MSP-Podcast vs TTS samples generated with ESPNET from automatic transcriptions created with wav2vec2.0.*

Data	Model	A	V	D
Original*	***w2v2-L-emo-ft***	.745	.634	.635
	w2v2-L-emo-frz	.696	.592	.400
Synthesised	***w2v2-L-emo-ft***	.041	.386	.048
	w2v2-L-emo-frz	.014	.015	.024

* Results taken from Wagner *et al.* [1].

epochs) on the output representation of each transformer layer of *w2v2-L-emo-ft* and *w2v2-L-emo-frz* to predict the following set of acoustic and linguistic features, which are proposed by Shah *et al.* [13]. As linguistic features, we use the number of: 1. **unique words**, 2. **adjectives**, 3. **adverbs**, 4. **nouns**, 5. **verbs**, 6. **pronouns**, 7. **conjunctions**, 8. **subjects**, 9. **direct objects**, as well as 10. **type/token ratio** (hence referred to as "word complexity" as in Shah *et al.* [13]), and 11. **the depth of the syntax tree**. We additionally add the number of **negations**, as this turned out important during our CHECKLIST probing. These features are all extracted using the Stanford CoreNLP toolkit[4] [29]. As acoustic features we use: 1. **total signal duration**, 2. **zero crossing rate**, 3. **mean pitch**, 4. **local jitter**, 5. **local shimmer**, 6. **energy entropy**, 7. **spectral centroid**, and 8. **voiced to unvoiced ratio**. The acoustic features are extracted using the ComParE2016 [30] feature set of our openSMILE toolkit[5] [31] – with the exception of duration, which is obtained with *audiofile*.[6] We evaluate predictive performance using root mean squared error (RMSE).

3. Results and discussion

Our discussion begins with the results of our first probing: testing the performance on re-synthesised transcriptions. Table 1 shows the performance of *w2v2-L-emo-ft* and *w2v2-L-emo-frz* on the original test set of MSP-Podcast and its re-synthesised version. Consistent with our previous results [1], the fine-tuned model obtains a competitive valence performance on the re-synthesised transcriptions (CCC: .386). This is comparable to previous state-of-the-art works like that of Li *et al.* [32], which reported a valence CCC of .377 on original data, showing that linguistic information alone is sufficient for obtaining good performance on that dimension. This is not true for arousal and dominance – which is consistent with previous findings showing that linguistics are not as competitive for those two dimensions [6]. Interestingly, the valence results only hold after in-domain fine-tuning of transformer layers on the original data; when keeping the original pre-trained weights, the model performance drops to chance-level. This is surprising given the fact that *w2v2-L-robust* must contain at least surface-level linguistic information (i. e., phonemes, words) as it yields state-of-the-art ASR results with minimal fine-tuning [25]. Nevertheless, the results of our first probing experiment show that *w2v2-L-emo-ft* has some dependence on linguistic knowledge.

Figure 1 then shows an overview of our second probing process. It shows the distributions of predicted emotional dimensions for (negative/neutral/positive) words in isolation, in context, and in the presence of negations for *w2v2-L-emo-ft*. We are primarily interested in two things: (a) a comparison of how model predictions change for each word category in isolation and in context for each of the three emotional dimensions (i. e., the same colour should be compared across all columns for the first and second rows), and (b) a comparison of how model predictions change when adding negations (i. e., the same colour should be compared within each column between the second and third row). These comparisons are quantified by statistical tests (pairwise t-tests between negative-neutral and neutral-positive for the first two rows, or negative-negative etc. for negations; 5% 95% CIs obtained with bootstrapping), while also being qualitatively described through visual inspections. Consistent with our previous results, we did not observe any large or significant differences for arousal and dominance, so we only discuss changes to valence for brevity. Furthermore, *w2v2-L-emo-frz* showed little reactivity to most probing experiments; the only substantial (but not statistically significant) difference is seen between valence predictions of negative (mean: .526; CI: [.376-.683]) and neutral words in isolation (mean: .602; CI: [.499-.732]). All other differences were marginal, showing that *w2v2-L-emo-frz* depends little on linguistic information. In contrast, *w2v2-L-emo-ft* shows several interesting trends, which we proceed to discuss in the following paragraphs.

Negative words in isolation obtain lower valence scores (mean: .412; CI: [.172-.655] than neutral (mean: .509; CI: [.430-.584]) or positive ones (mean: .588; CI: [.400-.711]); the difference between negative and positive was significant. Valence is lower for negative (mean: .395; CI: [.180-.626]) than for neutral words in context (mean: .565; CI: [.437-.681]), with the difference being significant. Accordingly, positive words in context are predicted more positively (mean: .692; CI: [.394-.820]) than neutral words in context (mean: .565; CI: [.437-.681]) – but this difference is not significant.

Surprisingly, negations seem to have a consistently negative impact on valence scores – even for negative utterances which should lead to more positive scores. Both, positive (mean: .509; CI: [.355-.751]) and negative phrases (mean: .372; CI: [.223-.542]), are scored lower than their counterparts without negation, but these differences are also not significant. Interestingly, adding negations to neutral words does result in a statistically significant reduction of valence predictions (mean: .450; CI: [.357-.606]). We return to the impact of negations later.

The last part of this probing experiment concerns intensifiers and reducers. These largely leave all dimensions unaffected (CIs overlap, p-values $> .05$). The only exception are the valence predictions of negative words, which are somewhat impacted by intensifiers (mean: .439; CI: [.180-.745]), but this difference is not significant, either. Thus, these higher-level semantics seem to leave the model overall unaffected.

Our last probing methodology sheds more light onto the inner workings of the self-attention layers, and how they are impacted by fine-tuning. Figure 2 shows the RMSE ratio between *w2v2-L-emo-ft* and *w2v2-L-emo-frz* when probing their intermediate representations with various linguistic features. This shows *relative* changes caused by fine-tuning. Values below $100\,\%$ mean that the *w2v2-L-emo-ft* model is better at predicting features than *w2v2-L-emo-frz*. We hypothesise that the network will increase its dependence (thus decreasing the ratio) on the features that are most useful for making predictions, leave unaffected the amount of information it contains for features that are already present in its representations to a sufficient extent, and decrease it for any that are potentially harmful.

Most features are unaffected by fine-tuning, with their

[4]https://stanfordnlp.github.io/CoreNLP
[5]https://audeering.github.io/opensmile
[6]https://github.com/audeering/audiofile

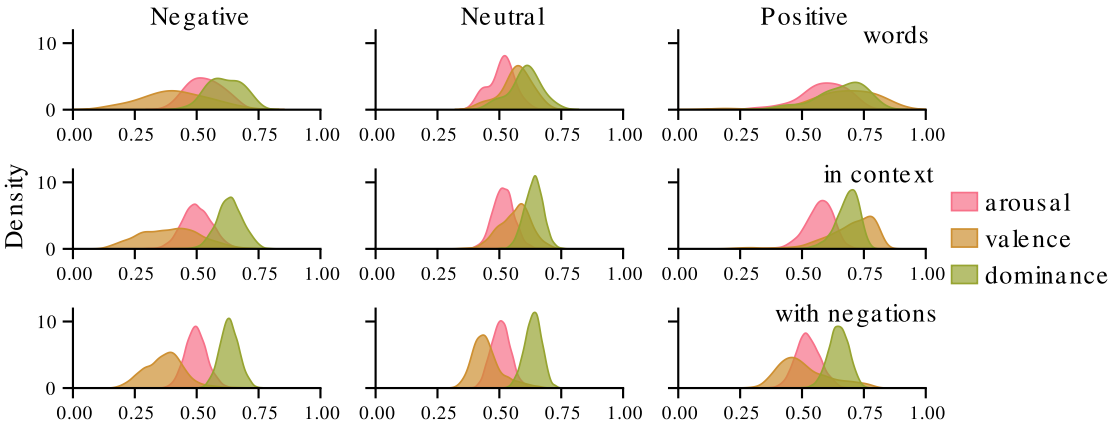

Figure 1: w2v2-L-emo-ft *behaviour on negative/neutral/positive text samples generated with* CHECKLIST *and synthesised with* ESPNET. *The resulting utterances are all synthesised without emotional intonation - thus, variability is attributed to the linguistic content.*

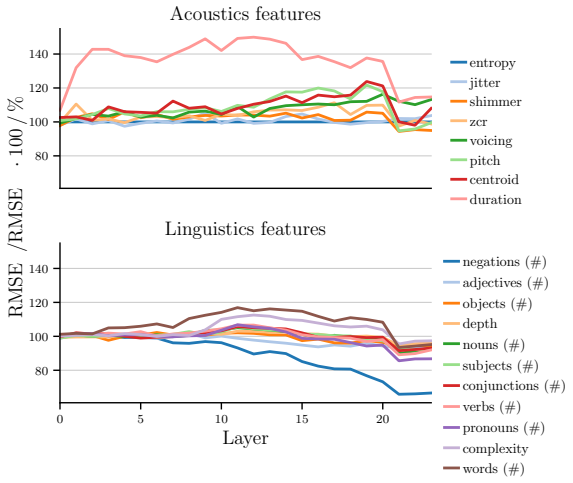

Figure 2: *RMSE ratio (in percentage) for acoustic (top) and linguistic (bottom) feature prediction using fine-tuned vs frozen (ft/frz) embeddings for all self-attention layers. Values below 100% mean that the* w2v2-L-emo-ft *model is better at predicting features than* w2v2-L-emo-frz.

RMSE ratio fluctuating around 100 %. The only ones showing substantial change are negations, word count and complexity, and duration. The network seems to decrease its dependence on the 'surface-level' features of word count and complexity [13], indicating that those are not needed for emotion recognition. This reduction is only evident in the middle layers (8-20).

The most outstanding changes in information for a given feature are seen for negations (RMSE ratio decreases by 70 %) and duration (RMSE ratio increases by 150 %). Evidently, the network considers the latter an uninformative feature (potentially because MSP-Podcast contains utterances of different lengths but with similar labels thus making duration a confounding feature). In contrast, the former is considered an important feature for its downstream tasks – which is consistent with the high reactivity to negations seen for CHECKLIST. We further investigate this by computing the Pearson correlation coefficient (PCC) between negations and the valence error on the (original) MSP-Podcast test set ($y_{\text{true}} - y_{\text{pred}}$). The PCC shows a small positive trend (.132) for valence, but not for arousal (.005) or dominance (−.003). This means that *w2v2-L-emo-ft* tends to under-predict ($y_{\text{true}} > y_{\text{pred}}$) as the number of negations increases. We further computed the PCC between the number of negations and ground truth valence annotations on the training set: these show a small, but non-negligible negative trend (−.142) - whereas no such trend exists between negations and arousal (.033) or dominance (.018). We hypothesise that *w2v2-L-emo-ft* picks up this spurious correlation between negations and valence; which explains why negations lead to lower valence scores in CHECKLIST tests.

In summary, valence predictions of *w2v2-L-emo-ft* are impacted by linguistic information, while arousal and dominance are unaffected by it. Furthermore, fine-tuning its self-attention layers is necessary to exploit this linguistic information. This explains previous findings that linguistics are not as suitable as acoustics for arousal/dominance prediction on MSP-Podcast [6], and that distilling linguistic knowledge to an acoustic network helps with valence prediction [2]. It also shows that using tests from the NLP domain will become necessary as speech 'foundational' models [10] become the dominant paradigm for SER.

4. Conclusion

We presented a three-stage probing methodology for quantifying the dependence of SER models on linguistic information, and used it to analyse the behaviour of a recent state-of-the-art model. Our approach demonstrates that the success of transformer-based architectures for the valence dimension can be partially attributed to linguistic knowledge encoded in their self-attention layers. It further helped us uncover a potentially spurious correlation between valence and negations which could hamper performance in real-world applications. As our probing pipeline is based on open-source libraries and is thus fully reproducible, we expect it to prove a useful tool for analysing future SER models. Future work could extend our methodology by expanding the set of probing features or utilising emotional voice conversion [33] to control the emotional expressivity of synthesised samples as another parameter [34].

5. Acknowledgements

This work has received funding from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS).

6. References

[1] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *arXiv preprint arXiv:2203.07378*, 2022.

[2] 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.

[3] 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.

[4] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," Brno, Czech Republic, 2021, pp. 3400–3404.

[5] S. Liu, A. Mallol-Ragolta, E. Parada-Cabeleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller, "Audio self-supervised learning: A survey," *arXiv preprint arXiv:2203.01205*, 2022.

[6] 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.

[7] 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 ACM International Conference on Multimedia (ACM MM)*, Chengdu, China: ACM, 2021, pp. 5706–5707.

[8] C. Aspillaga, A. Carvallo, and V. Araujo, "Stress test evaluation of transformer-based models in natural language understanding tasks," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, 2020, pp. 1882–1894.

[9] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" In *Proceedings FAccT*, Virtual Event, Canada, 2021, pp. 610–623.

[10] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[11] D. Ma, N. Ryant, and M. Liberman, "Probing acoustic representations for phonetic properties," in *Proceedings ICASSP*, IEEE, Toronto, Canada, 2021, pp. 311–315.

[12] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proceedings ASRU*, Cartagena, Colombia, 2021, pp. 914–921.

[13] J. Shah, Y. K. Singla, C. Chen, and R. R. Shah, "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021.

[14] Y.-A. Chung, Y. Belinkov, and J. Glass, "Similarity analysis of self-supervised speech representations," in *Proceedings ICASSP*, IEEE, Toronto, Canada, 2021, pp. 3040–3044.

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings NeurIPS*, Vancouver, BC, Canada, 2020, pp. 12 449–12 460.

[16] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings INTERSPEECH*, Hyderabad, India, 2018, pp. 2207–2211.

[17] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proceedings ICASSP*, IEEE, Barcelona, Spain, 2020, pp. 7654–7658.

[18] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of nlp models with checklist," in *Proceedings ACL*, Seattle, USA, 2020, pp. 4902–4912.

[19] 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 ICASSP*, IEEE, Toronto, Canada, 2021, pp. 7268–7272.

[20] 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 INTERSPEECH*, Graz, Austria, 2019, pp. 1691–1695.

[21] C. Oates, A. Triantafyllopoulos, I. Steiner, and B. W. Schuller, "Robust speech emotion recognition under different encoding conditions," in *Proceedings INTERSPEECH*, Graz, Austria, 2019, pp. 3935–3939.

[22] 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.

[23] M. Mohamed and B. Schuller, "Normalise for fairness: A simple normalisation technique for fairness in regression machine learning problems," *arXiv preprint arXiv:2202.00993*, 2022.

[24] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane, "Gender de-biasing in speech emotion recognition," in *Proceedings INTERSPEECH*, Graz, Austria, 2019, pp. 2823–2827.

[25] 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.

[26] 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.

[27] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proceedings INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1103–1107.

[28] K. Ito and L. Johnson, *The LJ speech dataset*, https://keithito.com/LJ-Speech-Dataset/, 2017.

[29] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings ACL: system demonstrations*, Baltimore, USA, 2014, pp. 55–60.

[30] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini, *et al.*, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proceedings INTERSPEECH*, San Francisco, USA, 2016, pp. 2001–2005.

[31] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[32] 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 ICASSP*, IEEE, Toronto, Canada, 2021, pp. 6329–6333.

[33] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[34] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *arXiv preprint arXiv:2201.03967*, 2022.