INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Emotion Label Encoding using Word Embeddings for Speech Emotion Recognition

Eimear Stanley, Eric DeMattos, Anita Klementiev, Piotr Ozimek, Georgia Clarke, Michael Berger and Dimitri Palaz

Speech Graphics Ltd, Edinburgh, United Kingdom

{eimear.stanley, eric.demattos, anita.klementiev, piotr.ozimek, g.clarke, berger, dpalaz}@speech-graphics.com

Abstract

Speech Emotion Recognition (SER) is an important and challenging task for human-computer interaction. Human emotions are complex and nuanced, hence difficult to represent. The standard representations of emotions, categorical or continuous, tend to oversimplify the problem. Recently, the label encoding approach has been proposed, where vectors are used to represent the emotion space. In this paper, we hypothesise that using a pre-existing vector space that encodes semantic information about emotion is beneficial for the task. To this aim, we propose using word embeddings obtained from a Language Model (LM) as labels for SER. We evaluate the performance of the proposed approach on the IEMOCAP corpus and show that it yields better performance than a standard baseline. We also present a method to combine free text labels, which are unusable in conventional approaches, and by doing so we show that the model can learn more nuanced representations of emotions.

Index Terms: speech emotion recognition, word embeddings, label encoding, paralinguistics

1. Introduction

Speech Emotion Recognition (SER) uses machine learning models to classify emotion from audio, which rely on accurate labels. However, human emotions are highly complex and emotion annotation is a difficult and subjective task [1]. The two common representations of emotion are categorical and continuous. The categorical approach uses a small set of discrete labels (*angry, happy, sad,* etc.) and relies on majority voting across a number of annotators to establish one ground truth label. However, this approach may oversimplify the task as it overlooks the complexity and natural subjectivity of emotion perception. Additionally relying on majority voting may exclude relevant information in the form of individual annotations, as well as data without agreement. The continuous approach looks to represent emotions in a continuous space, most commonly using arousal, valence, and dominance as dimensions. This has the same problems as the categorical approach, as well as potentially being more challenging to annotate.

Recently, using categorical labels in the framework of label encoding has been investigated. In this framework categorical labels are used to generate a vector, which is then used as a target for training the model. Two approaches have been proposed for SER: soft labels and metric learning. In the soft labels approach [2], the vectors are probability distributions computed from the individual annotations from each annotator. The metric learning approach [3] proposes to *learn* the vector space using a triplet loss function. Both approaches have shown some improvements on the SER task and have made first steps towards a better representation of the emotion space.

In the field of Natural Language Processing (NLP), a major area of research is designing representations of words that encode semantic information as well as the relationships between them. The most successful representations to this day are word embeddings, computed by a neural networks-based Language Model (LM). We hypothesise that word embeddings could better capture the complexity and multiplicity of emotion perception as opposed to the categorical labels that result from the closed-set majority voting approach. These embeddings are known to capture linear substructures [4], which could be used to take into account different labels among annotators. Our hypothesis is that this approach could be the key to encoding emotions efficiently and bringing more nuanced emotion information to the model.

This paper is an exploratory study aiming to showcase the potential benefit of using word embeddings as labels for SER. We present a novel approach where the model is trained to predict the word embeddings of the emotion labels, as opposed to the label itself. Using the IEMOCAP corpus [5], different loss functions are compared along with a baseline using the same model architecture trained with the standard approach. We show that the proposed approach yields better performance than the baseline, and yields performance on par with recent literature. A major novelty of the proposed approach is its ability to use any free text labels without mapping, which is impossible with conventional approaches. Based on this property, we propose an approach to combine individual annotations from multiple annotators and free text comments to investigate bringing more nuanced emotion information to the model. We also present a study on leveraging natural language to analyse the trained models and show that the models learn more nuanced emotion representations.

The key contributions of the paper are: (1) we present a novel approach to use word embeddings as labels for SER and show its benefits, and (2) we propose a novel approach to combine emotion labels, including free text labels and show its potential to learn better emotion representations.

2. Related Work

In this section we present the related work for word embeddings and for label encoding approaches for SER. For more details on other approaches in SER see [6, 7].

Word embeddings are vectors representing words in a space which preserve syntactic and semantic properties to some extent, which can then be used in different NLP tasks. Word embeddings are often computed using neural networks-based models [8] and are typically learned within a LM. The most popular language models are word2vec [4], GloVe [9], BERT [10] and GPT-3, [11] as used in the popular application ChatGPT.

10.21437/Interspeech.2023-1591

Word embeddings have recently been used to represent emotions in the context of speech synthesis. They were originally introduced in [12] as style tags for the synthesised speech, and the approach was then extended to use word embeddings computed from emotion labels [13].

In the context of SER, one recent approach to representing emotion is metric learning, where a measure of distance is learned using similarity between classes. Most commonly the triplet loss function is used, where the distance between an anchor, a positive and a negative sample is optimised, which has been shown to improve SER model performance [3, 14, 15]. Recent work [16] shows further improvements when integrating an additional sentiment constraint within the triplet loss function. A denoising autoencoder using a continuous metric loss based on either activation or valence labels was proposed in [17] to improve generalisation across languages.

To the same aim, recent approaches have sought to account for annotations from multiple annotators and model the subjectiveness of the SER task. The most common approach uses soft labels, which typically converts the traditional one-hot emotion label attained through majority voting to a probability distribution taking into account the individual annotations. Numerous previous studies have successfully used soft labels on the task of SER [2, 18, 19, 20]. More recently, [21] leverages the co-occurrence of emotion labels to create a weighted matrix, while [22] investigates modeling the uncertainty of emotion labels with Dirichlet priors.

3. Methodology

3.1. Label encoding using word embeddings

In the label encoding approach, the model is trained to predict vectors corresponding to the label, rather than the label index y. In this paper, vectors are word embeddings from a dictionary obtained from a pre-trained LM. This dictionary W has an entry for all words in the English language. The vectors w used for training are obtained by getting the entry of the dictionary corresponding to the name $\mathtt{y}$: $w = W(\mathtt{y})$.

The model $f(\cdot)$ is trained by minimising the distance between its output $f(x)$ for input utterance x and the target word embedding w. The loss functions $\mathcal{L}(x, w)$ investigated in this paper are the Mean Square Error (MSE), Mean Absolute Error (MAE) and Log Cosh (LC).

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{n} (f(x_n) - w_n)^2 \tag{1}$$

$$\mathcal{L}_{MAE} = \frac{1}{N} \sum_{n} |f(x_n) - w_n| \tag{2}$$

$$\mathcal{L}_{LC} = \frac{1}{N} \sum_{n} \log(\cosh(f(x_n) - w_n)) \tag{3}$$

3.2. Combining encoded labels

Due to the complexity and subjectivity of emotion, it is common for a given utterance to have more than one label. In an attempt to represent emotion in a more fine-grained and subtle manner, we use one of the most useful properties of word embeddings: their ability to capture semantic linear substructures, which enable simple arithmetical operations. For instance, word2vec [4] reports that $W(\mathtt{Rome}) = W(\mathtt{Paris}) - W(\mathtt{France}) + W(\mathtt{Italy})$. Taking inspiration from this principle, we investigate an approach to combine encoded labels using vector geometry. The combined word embedding w_{new} is computed as the average of the different word embeddings w_i obtained from the different labels $\mathtt{y}_i$, $i = 1, ..., L$:

$$w_{new} = \frac{1}{L} \sum_{i=1}^{L} w_i = \frac{1}{L} \sum_{i=1}^{L} W(\mathtt{y}_i). \tag{4}$$

For instance, if a given utterance has three labels $\{anger, sadness, sadness\}$, the new word embedding will be on the line between the word embedding for *sadness* and the word embedding for *anger*, at $1/3$ of the distance from *sadness* and $2/3$ from *anger*.

3.3. Evaluation

In order to evaluate the output of the models trained on word embeddings, a metric measuring the distance between embeddings is first needed. Following GloVe [9], we use the cosine similarity defined as the cosine of the angle between two embeddings.

In the SER literature, two metrics are commonly used: the Weighted Accuracy (WA), and the Unweighted Accuracy (UA), which is the average of the per-class accuracies. To compute these metrics with the proposed models using word embeddings, their output needs to be mapped to the index labels. The prediction $\hat{y}$ is therefore computed by finding the closest embedding to the output w from the embedding dictionary W_Y using the cosine similarity s_{cos}.

$$\hat{y} = \underset{i}{\operatorname{argmin}}\ s_{cos}(w, W_Y(i)), \tag{5}$$

where W_Y is a subset of the word embedding dictionary W containing only entries corresponding to the label set (for example: {*anger, happiness, neutral, sadness*}).

4. Experimental Setup

4.1. Database

IEMOCAP is an audio-visual corpus totalling 12 hours of data from 10 speakers [5]. Each utterance was labeled by three annotators choosing from a discrete set of categorical emotion labels. Annotators were permitted to use one or more labels per utterance and could leave free text comments if they felt none of the labels were adequate. For the purpose of this study and in line with previous works [23], majority voting is used and only utterances with *anger, excitement, happiness, sadness,* and *neutral* ground truth labels are used, with *excitement* merged with *happiness*.

4.2. Encoded label sets

The following three label sets are used to train the models using word embeddings. All sets use the same audio data, which is selected as described in section 4.1. We use the noun form of all emotion labels.

- **Baseline set** S_1 is created by extracting the word embeddings from the ground truth labels. We refer to this as the 4-label set.
- **Annotator set** S_2 is created by extracting the individual annotations from multiple annotators and combining them as presented in Section 3.2, leading to 129 different embeddings targets. This means that labels outside of the four label set are included, for example *frustration*. The *other* label is discarded as its meaning is unclear and could confuse the model.

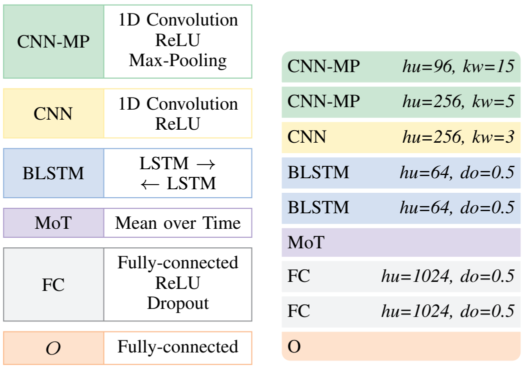
- **Comment set** S_3 is created by extracting the comments provided by the annotators as well as the individual annotations as in S_2. We split the comments on commas, discarding any comments longer than one word as well as those containing misspellings. The comments and the labels are combined in the same way as in S_2, leading to 629 different embedding targets.

4.3. Model and training details

The architecture of the model is based on [24] and its hyper-parameters are shown in Figure 1. Following the latest findings in SER [25], we use the learned representation from a pre-trained wav2vec2.0 model [26] as input features, more precisely the representation from the 14th layer. The wav2vec model, pre-trained on the Librispeech corpus, is obtained from huggingface[1].

We train three models using word embeddings on the three label sets presented above: M_1 on S_1, M_2 on S_2 and M_3 on S_3. The models are trained using 5-fold cross-validation, where for each fold one session from the corpus is considered as the test set and the rest is designated randomly to either train (80%) or validation (20%), as is usually done in the literature [23]. The models are trained for 100 epochs with a learning rate of 2×10^{-3} for the model using word embeddings and 2×10^{-4} for the baseline, using early stopping and the one_cycle scheduler [27] provided by pytorch [28]. As a baseline we train the model using categorical labels and the cross entropy loss function.

We evaluate three commonly used and readily available word embeddings: word2vec [4], GloVe [9] and BERT [10]. As we found no significant differences between them for our approach, we select the best performing: GloVe[2] which contains 2.2M entries. For the ease of reproducibility, no processing was done on the embedding space.



(a) *Network modules* (b) *Model*

Figure 1: *Model architecture. 'hu' refers to the number of hidden units, 'kw' is the kernel width and 'do' is the dropout rate.*

5. Results

5.1. Recognition studies

5.1.1. *Evaluation on the 4-label set*

In this study, the outputs of the three models are mapped to the 4-label set as described in Section 3.3. We compare them with the baseline and related work. Table 1 presents the results of the models using word embeddings, trained with the three losses presented in Section 3.1, along with the baseline. All three models outperform the baseline, showing the benefits of using word embeddings as labels. The M_1 model yields the best performance, which is expected as it is the only model trained exclusively on the 4-label set. The performance gap of M_2 and M_3 compared to M_1 is very small, which shows another benefit of the proposed approach as training models on hundreds of different labels tends to be challenging when using categorical labels. The three loss functions yield similar performances: the LC function being slightly better for M_1 and the MSE loss for M_2 and M_3. For fair comparison across all three models, going forward, we use the same loss function. We select the LC as it yields the highest singular performance.

Table 1: *Model accuracies on the 4-label set*

Baseline: 61.45% UA, 57.70% WA						
	MSE		**MAE**		**LC**	
Model	UA	WA	UA	WA	UA	WA
M_1	67.91	66.61	68.78	67.68	**69.68**	**68.47**
M_2	68.26	66.37	66.63	65.06	65.48	62.24
M_3	67.33	66.14	65.37	63.93	65.89	62.35

As seen in Table 2, M_1 yields performance on par with recent work but not reaching state-of-the-art. This is expected as this work is an exploratory study and did not include an extensive hyper-parameter search or fine-tuning of the wav2vec features.

Table 2: *Comparison with recent work*

Approach	**UA [%]**	**WA [%]**
EPATA-TDNN w/ fbank [23]	57.60	56.52
Self-attention w/ IS09 [29]	63.80	68.10
Word embeddings (this paper)	**69.68**	**68.47**
Co-attention w/ MFCC+wav2vec [30]	71.05	69.80
EPATA-TDNN w/ wav2vec (FT) [23]	77.07	76.58

5.1.2. *Evaluation on combined labels*

In order to further evaluate all three models, we compute the average cosine similarity between the predicted and target embeddings for the three label sets. The results are presented in Table 3. M_2 and M_3, trained on combined labels, have a higher similarity on S_2 and S_3, showing that the models have learned a more fine-grained representation of the word embedding space. Contrary to expectation, the best model on S_2 is not M_2 but M_3, which shows that adding the comments is beneficial. This highlights one of the key benefits of our approach, as free text comments are very difficult to use in standard approaches and so are typically discarded.

Table 3: *Average cosine similarity across the label sets*

Model	S_1	S_2	S_3
M_1	**0.789**	0.764	0.759
M_2	0.756	0.834	0.831
M_3	0.756	**0.841**	**0.839**

[1] https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

[2] https://nlp.stanford.edu/projects/glove/

Table 4: *Top 5 nearest neighbours of the target labels. Those in bold match an emotion from the target label.*

Set	Label	1st	2nd	3rd	4th	5th
S_1	*happiness*	**happiness**	joy	contentment	bliss	prosperity
S_2	*excitement, happiness, happiness, happiness,*	**happiness**	joy	contentment	**excitement**	bliss
S_3	*amazed, enthusiastic, happiness, happiness, happiness, impressed*	**happiness**	joy	enthusiasm	gratitude	happy
S_3	*bantering, excitement, happiness, happiness, joking*	**happiness**	laughter	joy	**excitement**	contentment

5.2. Exploratory studies using natural language

One of the major appeals of the proposed approach is that the word embedding space not only contains the target labels but the entire vocabulary from the word embedding dictionary. In this section, we use natural language to further assess our hypothesis that the combined label sets can enable the models to learn more nuanced and subtle representations of emotion. More specifically, we use nearest neighbours in both qualitative and quantitative analysis.

5.2.1. *Qualitative analysis*

To illustrate the benefit of using a semantic space for label encoding, we look at the closest embeddings to the target embedding. Table 4 presents the 5 nearest neighbours to a small number of example labels from S_1, S_2 and S_3. The combined sets, S_2 and S_3, encapsulate the meanings from multiple labels and better model entangled emotions. To evaluate if the models are able to capture this, we perform the same analysis of finding the closest embedding, but on the predicted embeddings. Table 5 presents the top 5 nearest neighbours to the predicted embeddings for each of the four emotion labels in S_1. First, this shows that the model is able to learn the word embedding space. On closer inspection, we can see that the nearest neighbours vary across the three models, for example, the neighbours of *neutral* in M_1 are more syntactically related while in M_2 and M_3 they are more semantically related. Also of note, is that *sadness* is a neighbour to *happiness* in M_1, and not in M_2 and M_3. This suggests that the models trained on the combined sets capture a more informed representation.

Table 5: *Top 5 nearest neighbours of the predictions*

Model	Emotion label			
	anger	*neutral*	*happiness*	*sadness*
M_1	anger frustration resentment grief rage	neutral neutrals nuetral Neutral zone	happiness joy sadness contentment feelings	sadness despair sorrow grief loneliness
M_2	frustration anger resentment fear rage	neutral sense feelings frustration feeling	joy excitement happiness enthusiasm feelings	sadness despair sorrow grief feelings
M_3	frustration anger resentment fear rage	neutral sense feelings feeling muted	excitement joy happiness enthusiasm feelings	sadness despair sorrow grief feelings

5.2.2. *Quantitative analysis*

This experiment uses the top-10 accuracy, which measures the accuracy of a given word being present in the 10 nearest neighbours. We define l_1 and l_2 as the most and second most frequent emotion label within the combined label respectively. We compute the top-10 accuracy of l_1 and l_2 for each prediction on S_2 and S_3 across the three models. The results are presented in Table 6. The top-10 accuracy on l_1 shows that training on the combined labels does not hurt the ability of the model to predict the majority label, conversely the accuracy improves for M_2 and M_3. Additionally, the top-10 accuracy on l_2 is higher for M_2 and M_3, which shows that this approach helps the models predict a secondary label. This indicates that M_2 and M_3 are able to better capture the nuance of emotions. Overall, these two findings indicate that the predictions from M_2 and M_3 are better situated in the word embedding space, showing the benefit of the combined label approach.

Table 6: *Top-10 accuracy for l_1 and l_2 on S_2 and S_3*

	S_2		S_3	
Model	l_1	l_2	l_1	l_2
M_1	59.58	31.93	59.40	28.17
M_2	71.34	**51.35**	71.23	**45.24**
M_3	**72.32**	50.22	**72.23**	42.45

6. Conclusion

In this paper we presented an exploratory study using word embeddings as labels for speech emotion recognition. This included a novel approach to combine multiple emotion labels, including free text comments. We showed that the proposed approach yields better performance than the standard approach using the same model architecture, highlighting the benefit of using a label encoding space which already encodes semantic relations between emotions. Additionally, we presented a study using natural language, which indicates that the model trained on the combined label sets learned a more informed and nuanced representation of emotion, and that the proposed approach can leverage semantic relationships between emotions. These are novel insights which could pave the way for new developments towards improving SER models and their interpretability. For future work, we will investigate applying this approach to cross-corpus training as it could address one of its main issues: the mismatch between label sets. We will also investigate filtering the word embedding space.

In terms of limitations, the approach is restricted by both the data and word embeddings, which are readily available in English, but are less available for a wider range of languages. Additionally, using word embeddings as label encodings introduces new biases to the SER task, which otherwise would not exist, as language models are known to manifest potentially harmful social biases [31].

7. References

[1] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[2] H. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *Proc. International Joint Conference on Neural Networks*. IEEE, 2016, pp. 566–570.

[3] J. Huang, Y. Li, J. Tao, and Z. Lian, "Speech emotion recognition from variable-length inputs with triplet loss function." in *Proc. Interspeech*. ISCA, 2018, pp. 3673–3677.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. International Conference on Learning Representations*, 2013, pp. 1–12.

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

[6] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[7] M. B. Akçay and K. Oğuz, "Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[8] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[9] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2014, pp. 1532–1543.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proc. Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[12] M. Kim, S. J. Cheon, B. J. Choi, J. J. Kim, and N. S. Kim, "Expressive Text-to-Speech Using Style Tag," in *Proc. Interspeech*. ISCA, 2021, pp. 4663–4667.

[13] Y. Shin, Y. Lee, S. Jo, Y. Hwang, and T. Kim, "Text-driven Emotional Style Control and Cross-speaker Style Transfer in Neural TTS," in *Proc. Interspeech*. ISCA, 2022, pp. 2314–2317.

[14] P. Kumar, S. Jain, B. Raman, P. P. Roy, and M. Iwamura, "End-to-end triplet loss based emotion embedding system for speech emotion recognition," in *Proc. International Conference on Pattern Recognition*. IEEE, 2021, pp. 8766–8773.

[15] B. Mocanu, R. Tapu, and T. Zaharia, "Utterance level feature aggregation with deep metric learning for speech emotion recognition," *Sensors*, vol. 21, no. 12, p. 4233, 2021.

[16] B. Mocanu and R. Tapu, "Emotion Recognition from Raw Speech Signals Using 2D CNN with Deep Metric Learning," in *Proc. International Conference on Consumer Electronics*. IEEE, 2022, pp. 1–5.

[17] S. Das, N. Lund, N. Lønfedlt, A. Pagsberg, and L. Clemmensen, "Continuous metric learning for transferable speech emotion recognition and embedding across low-resource languages," in *Proc. Northern Lights Deep Learning Workshop*, vol. 3, 2022.

[18] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *Proc. International Conference on Affective Computing and Intelligent Interaction*. IEEE, 10 2017, pp. 415–420.

[19] K. Sridhar, W.-C. Lin, and C. Busso, "Generative approach using soft-labels to learn uncertainty in predicting emotional attributes," in *Proc. International Conference on Affective Computing and Intelligent Interaction*. IEEE, 2021, pp. 1–8.

[20] H.-C. Chou, W.-C. Lin, C.-C. Lee, and C. Busso, "Exploiting annotators typed description of emotion perception to maximize utilization of ratings for speech emotion recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 7717–7721.

[21] H.-C. Chou, C.-C. Lee, and C. Busso, "Exploiting co-occurrence frequency of emotions in perceptual evaluations to train a speech emotion classifier," in *Proc. Interspeech*. ISCA, 2022, pp. 161–165.

[22] W. Wu, C. Zhang, X. Wu, and P. C. Woodland, "Estimating the uncertainty in emotion class labels with utterance-specific Dirichlet priors," in *Transactions on Affective Computing*. IEEE, 2022.

[23] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *Proc. International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 6922–6926.

[24] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, and G. Hofer, "Analysis of deep learning architectures for cross-corpus speech emotion recognition," in *Proc. Interspeech*. ISCA, 2019, pp. 1656–1660.

[25] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech*. ISCA, 2021, pp. 3400–3404.

[26] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[27] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*, vol. 11006. SPIE, 2019, pp. 369–386.

[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic Differentiation in PyTorch," in *Proc. of Neural Information Processing Systems Workshop*, 2017.

[29] L. Tarantino, P. N. Garner, A. Lazaridis *et al.*, "Self-attention for speech emotion recognition." in *Proc. Interspeech*. ISCA, 2019, pp. 2578–2582.

[30] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *Proc. International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 7367–7371.

[31] P. P. Liang, C. Wu, L.-P. Morency, and R. Salakhutdinov, "Towards understanding and mitigating social biases in language models," in *Proc. International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 6565–6576.