A Comparative Study of Pre-trained Speech and Audio Embeddings for Speech Emotion Recognition

Orchid Chetia Phukan
Dept. of CSE
IIIT Delhi, India
orchidp@iiitd.ac.in

Arun Balaji Buduru
Dept. of CSE
IIIT Delhi, India
arunb@iiitd.ac.in

Rajesh Sharma
Institute of Computer Science
University of Tartu, Estonia
rajesh.sharma@ut.ee

Abstract—Pre-trained models (PTMs) have shown great promise in the speech and audio domain. Embeddings leveraged from these models serve as inputs for learning algorithms with applications in various downstream tasks. One such crucial task is Speech Emotion Recognition (SER) which has a wide range of applications, including dynamic analysis of customer calls, mental health assessment, and personalized language learning. PTM embeddings have helped advance SER, however, a comprehensive comparison of these PTM embeddings that consider multiple facets such as embedding model architecture, data used for pre-training, and the pre-training procedure being followed is missing. A thorough comparison of PTM embeddings will aid in the faster and more efficient development of models and enable their deployment in real-world scenarios. In this work, we exploit this research gap and perform a comparative analysis of embeddings extracted from eight speech and audio PTMs (wav2vec 2.0, data2vec, wavLM, UniSpeech-SAT, wav2clip, YAMNet, x-vector, ECAPA). We perform an extensive empirical analysis with four speech emotion datasets (CREMA-D, TESS, SAVEE, Emo-DB) by training three algorithms (XGBoost, Random Forest, FCN) on the derived embeddings. The results of our study indicate that the best performance is achieved by algorithms trained on embeddings derived from PTMs trained for speaker recognition followed by wav2clip and UniSpeech-SAT. This can relay that the top performance by embeddings from speaker recognition PTMs is most likely due to the model taking up information about numerous speech features such as tone, accent, pitch, and so on during its speaker recognition training. Insights from this work will assist future studies in their selection of embeddings for applications related to SER.

Keywords: Pre-trained models, Speech Emotion Recognition, Transformers, Convolutional Neural Networks.

I. INTRODUCTION

Pre-trained models (PTMs) are widely available in the speech and audio signal processing domain. Pre-training is carried out on large-scale speech (Librispeech (LS) [1]) or non-speech (AudioSet (AS) [2], VGGSound (VS) [3]) databases. They find application in various narrow-domain tasks in different ways: from feature extractors for downstream models to the whole model being fine-tuned on task-specific data. Their model architectures can be of varied nature, it can be Convolution Neural Network (CNN) based such as AlexNet, VGG, Inception, ResNet [4], etc., and also, attention-based such as AALBERT [5], CAV-MAE [6], etc. Pre-training is executed using different approaches: supervised [7] or self-supervised fashion [8]. Embeddings exploited from PTMs are used for different tasks, for example, covid-19 detection [9], music emotion recognition [10], speech emotion recognition (SER) [11].

In this work, we focus on SER, an important task for human-machine interaction. It has gained traction in recent times due to its prospective applications in a wide span of different domains, for instance, psychology, healthcare, and fields that often include customer interactions, such as customer service providers, call centers, and so on. A variety of methods have been applied for SER, ranging from fuzzy methods [12], Hidden Markov Model (HMM) based methods [13], classical machine learning-based approaches [14], deep learning-based methods [15] to embeddings from PTMs such as wav2vec 2.0 [16], HuBERT [17]. The availability of a large number of PTMs has resulted in significant progress in the field of SER. As they were trained on vast amounts of data and learned detailed and nuanced representations of speech, the embeddings extracted from them have proven beneficial for emotion recognition.

However, it is not clear which PTM embeddings are best for SER. Keesing et al. [18] provided a comparison between acoustic and neural (speech and audio) embeddings by training downstream classifiers such as SVM, RF, MLP, etc. on various speech emotion databases. Atmaja et al. [19] assessed representations of PTMs for SER that were pre-trained in a self-supervised manner on speech data by training an FCN classifier on the representations as input features. But a comprehensive comparison of embeddings extracted from a broad variety of PTMs with consideration of their model architectures, pre-training methodologies, and pre-training datasets has not been carried out for SER. We address this research gap by conducting a comparative study of embeddings extracted from eight PTMs by training low-level models with the embeddings as input features.

To summarize, the following are our main contributions:

- Compiling PTM embeddings that could be useful for performing downstream SER tasks. We consider many diverse PTMs (wav2vec 2.0, data2vec, wavLM, UniSpeech-SAT, wav2clip, YAMNet, x-vector, ECAPA) with varied model architectures, pre-training data, and pre-training procedures.
- Comprehensive comparative analysis of different PTM embeddings through downstream classifiers (XGBoost, Random Forest, Fully Connected Network) which are

trained and evaluated on four public datasets (CREMA-D, TESS, SAVEE, Emo-DB).

- Our study has found that embeddings from PTMs trained for speaker recognition tasks perform better than embeddings from other categories of Speech/Audio PTMs. Our hypothesis is that this could be speaker recognition training procedures enabling models to learn various aspects of speech such as tone, accent, pitch.

This paper is divided into six sections. Section II discusses past works on PTMs followed by Section III which elaborates on the different speech emotion databases taken into consideration for carrying out our experiments. In Section IV, we provide brief information on PTM embeddings considered for our analysis and the reason behind the consideration. Section V focuses on the lower-level classifiers, their implementation, training, and results obtained for the comparative analysis. Finally, Section VI concludes the work presented and gives prospective directions for future work.

II. Related Works

Initially, PTM architectures were mostly CNN-based, for instance, SoundNet [20], a 1D CNN trained on a massive amount of unlabeled videos collected from Flickr. It was trained in collaboration with a visual recognition network via discriminative knowledge transfer. Later, the trained model's representations were used as features combined with posterior classifiers to classify acoustic scenes. With the availability of a large-scale labeled audio dataset, AS, various models such as VGGish [4], L3-Net [21], PANNs [22], and etc. were proposed. VGGish is based on VGG architecture and was trained in a supervised manner to classify 527 sound events. L3-Net is also based on the VGG network and was pre-trained in a self-supervised manner for audio-visual correspondence. Gong et al. [23] trained EfficientNet for audio tagging on AS that was first trained on ImageNet (IM). They also discussed how pre-training in a different modality boosts performance. Niizumi et al. [24] extended Bootstrap your own latent (BYOL) approach initially given for vision to BYOL for audio (BYOL-A). BYOL-A presents a novel generalized self-supervised approach for generating audio representation and employs a CNN as an encoder. It was pre-trained on AS by removing the labels and achieved competitive results on various low-level tasks such as speaker identification, language identification, etc. with baseline models. Schneider et al. [25] proposed a novel pre-trained multilayer CNN model wav2vec, trained on unlabeled speech data for speech recognition. wav2vec reported the lowest WER for character-based speech recognition compared to past works.

Mockingjay [26], a multi-layer bidirectional transformer model was pre-trained on LS using masked modeling, where 15% of the input frames were masked to zero and it outputs the masked frames. They observed that pre-training Mockingjay in this manner resulted in improved performance in downstream supervised activities. Baevski et al. proposed wav2vec 2.0 [27], where the initial layer is a convolutional layer that acts as a feature encoder followed by transformer layer. It is trained

in a self-supervised way where masking of a few parts of the feature encoder outputs is done. Unlabeled LS is used as pre-training data and it improves upon wav2vec for phoneme recognition. HuBERT [28], a BERT-like architecture with self-supervised training was also devised that achieves comparable performance with wav2vec 2.0 for speech recognition in LS. The first fully attention-based convolution-devoid architecture named Audio-Spectrogram transformer (AST) was presented in [7] for audio classification tasks. It accepts mel-spectrogram as input. AST uses the advantages of pre-trained ViT for image classification tasks, and it is afterward trained on AS. Over previous investigations, AST reported state-of-the-art (SOTA) performance on the AS, ESC-50, and Speech Commands V2 databases. Gong et al. [29] enhanced AST further by training it in a self-supervised pattern through joint discriminative training and masked spectrogram modeling. This kind of training improved performance in lower-level tasks over the supervised version. Various encoder-decoder architectures, such as Audio-MAE [30] and MaskSpec [31], was also proposed.

Embeddings from PTMs such as YAMNet and wav2vec trained on audio and speech data, respectively, were used as input features to classifiers for SER [32]. Using models pre-trained on large databases and exploiting embeddings of them as features and applications of transfer learning by finetuning holds a promising future for SER. However, no comparison of embeddings recovered from a wide range of PTMs has been conducted for SER taking into account their model architectures, pre-training procedures, and pre-training datasets. To fill this knowledge gap, we conduct a comparative investigation of embeddings retrieved from eight diverse PTMs pre-trained on speech and audio data.
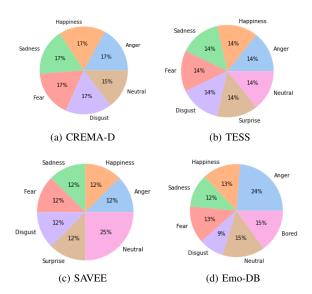

(a) CREMA-D (b) TESS

(c) SAVEE (d) Emo-DB

Fig. 1: Distribution of Emotions across different corpora

III. Speech Emotion Corpora

We experiment with four openly accessible benchmark speech emotion databases: Crowd-Sourced Emotional Mul-

TABLE I: Basic information related to various speech emotion corpora

Corpus	Lanaguage	# of utterances	# of speakers	Labeled Emotions
CREMA-D	English	7442	91	Anger, Happiness, Sadness, Fear, Disgust, Neutral
TESS	English	2800	2	Anger, Happiness, Sadness, Fear, Disgust, Neutral, Surprise
SAVEE	English	480	4	Anger, Happiness, Sadness, Fear, Disgust, Neutral, Surprise
Emo-DB	German	535	10	Anger, Happiness, Sadness, Fear, Disgust, Neutral, Bored

timodal Actors Dataset (CREMA-D) [33], Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [34], Toronto Emotional Speech Set (TESS) [35], Surrey Audio-Visual Expressed Emotion (SAVEE) [36], and German Emotional Speech Database (Emo-DB) [37]. Essential information and distribution of emotions for each corpus are given in Table I and Figure 1 respectviely.

Additional information related to the databases can be found below:

- **CREMA-D:** The audio snippets feature 48 male and 43 female performers from various ethnic origins. They talked from a list of 12 phrases. With a diverse range of ages, genders, and ethnicities, CREMA-D is a high-quality data source for SER.
- **TESS:** It is recorded by two female actors. Both actresses were fluent in English and cherry-picked 200 words were spoken by the actresses for the seven emotions.
- **SAVEE:** It comprises recordings of four male actors in British English accents. For each emotion, the actors delivered phrases that were phonetically balanced.
- **Emo-DB:** Recordings are from 5 male and 5 female actors. The actors were given a selection of ten distinct scripts from which to speak.

IV. PRE-TRAINED MODEL EMBEDDINGS

Embeddings derived from PTMs capture the semantic and aural information of the input clip. We intend to evaluate how effective these embeddings are at capturing emotional content by comparing embeddings retrieved from various PTMs. For selecting PTMs whose embeddings are to be used in our study, we follow two benchmarks: Speech processing Universal PERformance Benchmark (SUPERB) [38] and Holistic Evaluation of Audio Representations (HEAR) [39].

SUPERB consists of various speech-related tasks ranging from speaker identification, speech emotion recognition, speech recognition, voice separation, speaker diarization, etc. We select models with top performance in SUPERB and are openly available such as wav2vec 2.0, data2vec, wavLM, and UniSpeech-SAT. For wav2vec 2.0, we choose the base[1] version for our experiments that contains 12 transformer blocks in its architecture. On SUPERB, data2vec delivers slightly lower results than the model with the best performance i.e wavLM. data2vec [40] aims for bridging the gap in learning methods by proposing a generalized learning framework for different input modalities. wavLM [41] outperforms every other counterpart except UniSpeech-SAT on SUPERB. UniSpeech-SAT is a

contrastive loss model with multitask learning. UniSpeech-SAT pre-training is done in a speaker-aware format whereas wavLM learns masked speech prediction and denoising concurrently during pre-training. This assists wavLM in dealing with multidimensional information contained in speech, such as speaker identity, spoken content, and so on. wavLM base+[2] version is used for carrying out our experiments and it is made of a total of 12 transformer encoder layers and was pre-trained on 94k hours data from various diverse speech databases including LibriLight, VoxPopuli, and GigaSpeech. We choose the base+ version for wavLM as it has achieved slight improvement over the base version on SUPERB with a similar number of parameters. For wav2vec 2.0, data2vec, wavLM, and UniSpeech-SAT the last hidden states are extracted and with the application of pooling average, they are converted to a vector of 768-dimension for each audio file to be used as input features for low-level classifiers. The input audio is sampled to 16KHz for all the self-supervised PTMs. We work with the base versions of wav2vec 2.0, data2vec[3], and UniSpeech-SAT[4] due to computational constraints and they were pre-trained on 960 hours of speech data from LS. wav2vec 2.0 is the lowest-performing model on SUPERB among all the self-supervised models under consideration, however, it has been applied for SER and proven to be effective in both English and multilingual formats [42].

As SUPERB is primarily concerned with speech processing tasks, PTMs pre-trained on speech data and in self-supervised manner, we chose various other PTMs with presence in HEAR such as wav2clip and YAMNet. Presence of wav2vec 2.0 can also be seen in HEAR leaderboard. wav2clip and YAMNet doesn't achieve SOTA performances on HEAR leaderboard and are mostly dominated by transformer-based architectures pre-trained in a self-supervised fashion. However, we added them in our evaluation as we wanted to access the effectiveness of their embeddings for SER as they were pre-trained using different methodologies and differed in terms of the data used for pre-training. wav2clip[5] [43] is pre-trained using knowledge distillation from CLIP and employs ResNet-18 as an audio encoder and uses VGGSound, an audio-visual Youtube database as pre-training data. Each audio file is transformed to a 2D sprectrogram for input to ResNet and converted to a vector of 512-dimension by average pooling. Similar to its parent architecture CLIP, wav2clip also transfers the audio embeddings to a joint embedding space. wav2clip embeddings as input features with supervised models have

---

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

[2]https://huggingface.co/docs/transformers/model_doc/wavlm
[3]https://huggingface.co/docs/transformers/model_doc/data2vec
[4]https://huggingface.co/docs/transformers/model_doc/unispeech-sat
[5]https://pypi.org/project/wav2clip/

TABLE II: Comparison of XGBoost trained on different PTM embeddings

Audio PTM	CREMA-D		TESS		SAVEE		Emo-DB	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
wav2vec 2.0	40.29	40.47	69.76	69.61	40.28	29.44	49.38	46.90
data2vec	49.33	49.52	76.90	76.37	37.50	29.29	49.38	48.22
wavLM	45.48	45.85	83.10	82.41	50.00	42.73	54.32	52.06
UniSpeech-SAT	56.13	56.35	83.57	83.40	45.83	32.86	69.70	61.42
wav2clip	47.45	46.77	95.00	94.95	55.56	52.79	72.84	66.31
YAMNet	46.82	46.49	92.38	92.35	50.00	41.17	58.02	51.41
x-vector	**60.16**	**60.09**	**97.86**	**97.77**	**68.06**	**62.17**	**83.95**	**80.07**
ECAPA	54.34	54.02	97.14	97.05	55.56	50.09	75.31	69.70

TABLE III: Comparison of Random Forest trained on different PTM embeddings

Audio PTM	CREMA-D		TESS		SAVEE		Emo-DB	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
wav2vec 2.0	37.69	37.47	57.38	56.68	38.89	25.50	56.79	51.11
data2vec	44.58	44.37	68.33	67.46	36.11	23.45	58.02	54.28
wavLM	40.64	41.01	76.67	75.79	45.83	35.78	50.62	47.99
UniSpeech-SAT	49.06	48.93	78.33	77.99	45.83	32.35	60.49	49.07
wav2clip	44.94	44.16	94.52	94.50	59.72	55.24	67.90	63.55
YAMNet	43.87	42.49	88.57	88.54	51.39	39.16	53.09	50.12
x-vector	**52.01**	**51.64**	98.33	98.28	**61.11**	**49.89**	81.48	78.40
ECAPA	44.05	43.05	**98.57**	**98.45**	48.61	36.09	**83.95**	**80.98**

TABLE IV: Comparison of Fully Connected Network trained on different PTM embeddings

Audio PTM	CREMA-D		TESS		SAVEE		Emo-DB	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
wav2vec 2.0	46.02	45.81	84.76	84.40	41.67	31.98	60.49	57.70
data2vec	53.89	53.76	86.67	86.08	43.06	33.41	64.20	63.35
wavLM	55.77	55.57	95.00	94.80	50.00	32.27	62.96	59.63
UniSpeech-SAT	64.28	64.43	96.67	96.65	61.11	49.71	82.72	79.04
wav2clip	47.18	46.92	96.90	96.79	61.11	51.81	74.07	75.42
YAMNet	48.25	48.22	96.19	96.09	55.56	41.52	61.73	59.46
x-vector	**65.80**	**65.64**	98.81	98.79	**70.83**	**64.90**	87.65	87.01
ECAPA	61.15	60.95	**99.52**	**99.50**	61.11	54.11	**88.89**	**87.09**

(a) CREMA-D (b) TESS (c) SAVEE (d) Emo-DB

Fig. 2: t-SNE plots of wav2vec 2.0 embeddings across different speech emotion corpora

(a) CREMA-D (b) TESS (c) SAVEE (d) Emo-DB

Fig. 3: t-SNE plots of data2vec embeddings across different speech emotion corpora

(a) CREMA-D   (b) TESS   (c) SAVEE   (d) Emo-DB

Fig. 4: t-SNE plots of wavLM embeddings across different speech emotion corpora

(a) CREMA-D   (b) TESS   (c) SAVEE   (d) Emo-DB

Fig. 5: t-SNE plots of UniSpeech-SAT embeddings across different speech emotion corpora

(a) CREMA-D   (b) TESS   (c) SAVEE   (d) Emo-DB

Fig. 6: t-SNE plots of wav2clip embeddings across different speech emotion corpora

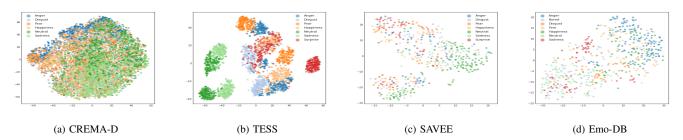(a) CREMA-D   (b) TESS   (c) SAVEE   (d) Emo-DB

Fig. 7: t-SNE plots of YAMNet embeddings across different speech emotion corpora

shown to be better than representations from other PTMs pre-trained on audio data in most datasets except FSD50K, where YAMNet representations performed better. YAMNet[6] pre-training is done in a supervised fashion on AS mainly for audio

classification and is based on MobileNet V1 CNN architecture. YAMNet generates frame-level embeddings that are average pooled into 1024-dimension clip-level embeddings.

To broaden our assessment, we also considered PTMs for speaker recognition, as knowledge gained for speaker recognition can be beneficial for SER. Evidence suggests that

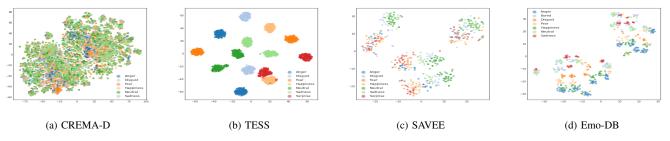[6]https://github.com/tensorflow/models/tree/master/research/audioset/yamnet

(a) CREMA-D   (b) TESS   (c) SAVEE   (d) Emo-DB

Fig. 8: t-SNE plots of x-vector embeddings across different speech emotion corpora

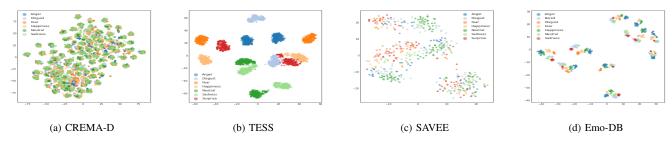(a) CREMA-D   (b) TESS   (c) SAVEE   (d) Emo-DB

Fig. 9: t-SNE plots of ECAPA embeddings across different speech emotion corpora

information gained for speaker recognition can help in SER [44]. Researchers have also advocated inserting knowledge about the speaker identity to network devoted to the primary job of SER [45] to boost performance for SER. So, we select x-vector [46] and ECAPA [47] to validate the efficacy of speaker recognition system for SER. x-vector, a time delay neural network (TDNN) improves over previous speaker recognition system, i-vector and Emphasized Channel Attention, Propagation and Aggregation (ECAPA) approach inserts several modifications to the x-vector model architecture. We pick off-the-shelf x-vector[7] and ECAPA[8] models. Both were pre-trained on a combination of voxceleb1 and voxceleb2 in a supervised manner. For pre-training of x-vector and ECAPA, all of the input audio files were sampled at 16Khz single-channel. We extract 512 and 192-dimension embeddings using Speechbrain [48] for x-vector and ECAPA respectively.

V. EXPERIMENTS

A. Downstream Classifier

We experiment with two classical machine learning approaches XGBoost (XGB), and Random Forest (RF), and a fully connected network (FCN). FCN is a simple neural network with three dense layers, batch normalization and dropout in between. Activation function being used is *relu* in all the dense layers and followed by *softmax* in the output layer which outputs the probabilities for different emotions. The same models are trained and evaluated with all the embeddings taken under consideration.

[7]https://huggingface.co/speechbrain/spkrec-xvect-voxceleb
[8]https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

All four speech emotion corpora are splitted to 85:15 ratio with 15% being used for testing. Out of the remaining 85%, 10% is kept for validation and the rest is used for training the classifiers. Hyperparameters are selected based on the performance of the classifiers on the validation set using GridSearchCV from *sklearn* library. We train the FCN for 50 epochs with a learning rate of 1e-3 and the optimizer being used is *Adam*. In addition, learning rate decay and early stopping are also applied while training the FCN.

B. Experimental Results

We compared the performance of eight PTMs embeddings across four speech emotion databases with two popular metrics accuracy and F1-score (macro). Table II, Table III and Table IV shows the results of XGB, RF, and FCN for different PTMs embeddings across different datasets respectively.

Among self-supervised embeddings (wav2vec 2.0, data2vec, wavLM, UniSpeech-SAT), UniSpeech-SAT performed the best. It achieved the highest performance on CREMA-D, TESS, Emo-DB in Table II followed by CREMA-D, TESS, Emo-DB in Table III, and CREMA-D, TESS, SAVEE, Emo-DB in Table IV. Speaker-aware pre-training may have contributed to these findings. The second is wavLM embeddings that outperformed UniSpeech-SAT embeddings on SAVEE in Table II and III. A diverse dataset and the approach for pre-training where denoising is concurrently involved might adhere to this outcome. Among data2vec and wav2vec 2.0, data2vec embeddings perform better than wav2vec 2.0, however, the data used for pre-training belongs to the same dataset (LS), this can be the result of the architectural difference between data2vec and wav2vec 2.0.

wav2clip embeddings perform better than the self-supervised embeddings excluding UniSpeech-SAT across almost all the databases. This could be resultant of the learned knowledge achieved from CLIP and also during its pre-training in a multi-modal format which aims to push all the modalities to a single embedding space. YAMNet embeddings achieved moreover comparable results w.r.t its self-supervised counterparts, sometimes higher and sometimes lower, for example, in Table II, YAMNet embeddings proved to be more effective in capturing emotion on TESS and Emo-DB. YAMNet reported lower performance than wav2clip across all the datasets except only in one instance in Table IV.

Embeddings from speaker recognition PTMs outperformed all other embeddings from different speech/audio PTMs across all spoken emotion datasets. This might be a manifestation of the information learned to identify speakers, where it is trained to recognize and distinguish between unique speakers. As a result, they learned to recognize distinctive elements of a person's speech patterns, such as rhythm, tone, and pitch, as well as linguistic and behavioral variables. x-vector achieves the top performance in comparison to ECAPA in most instances except on TESS and Emo-DB in Table III and IV.

We also present t-SNE plots of raw embeddings extracted from different PTMs to understand the emotion-wise cluster. Figures 2, 3, 4, 5, 6 7, 8, and 9 illustrates the t-SNE plots for wav2vecv 2.0, data2vec, wavLM, UniSpeech-SAT, wav2clip, YAMNet, x-vector, and ECAPA embeddings respectively. These figures support the results obtained from the tables above, it can be seen the embeddings extracted from PTMs for speaker recognition have far better emotion clusters with the highest distance between them than embeddings from other PTMs, especially for TESS corpus followed by wav2clip, YAMNet, and UniSpeech-SAT embeddings. For CREMA-D and TESS, the clusters formed by all eight PTM embeddings are almost inseparable. The results from the tables as well as the t-SNE plots show that models pre-trained with knowledge of the speaker performs best in SER, as evidenced by the performance of UniSpeech-SAT among self-supervised PTMs and the overall performance of x-vector and ECAPA.

VI. CONCLUSION

PTMs have been useful in various speech and audio-related tasks. Pre-train it on vast amount of labeled or unlabeled data and these models or the derived features from it can be highly beneficial for a wide range of tasks. Out of the variety of speech processing tasks, SER is a hard task to recon with, as due to various factors comes into play including difference in voice, tone, and accent. Past literature have shown the usage of different speech/audio PTMs embeddings for SER. However, previous studies haven't presented an extensive comparison of PTMs for SER with inclusion of various perspectives such as architectures of the PTMs, data utilized during pre-training phase, and pre-training technique followed. Our studies tries to narrow down this research gap by comparing embeddings derived from eight PTMs (wav2vec 2.0, data2vec, wavLM, UniSpeech-SAT, wav2clip, YAMNet,

x-vector, ECAPA) by training three classifiers (XGB, RF, FCN) on top of these features for four speech emotion datasets (CREMA-D, TESS, SAVEE, Emo-DB). Classifiers trained on embeddings extracted from models pre-trained for speaker recognition attained top performance in all corpora. Our findings suggest that the knowledge acquired for speaker recognition, such as recognition of tone and accent, provides benefits for SER. Embeddings generated from self-supervised PTMs have achieved SOTA performance across a wide range of downstream applications, with architectures such as wavLM and UniSpeech-SAT coming out on top. However, the results of our investigation show that embeddings from simpler CNN PTM like YAMNet still hold solid ground in terms of performance for SER. The outcomes of this study can be used to guide future studies in selecting appropriate embeddings for speech-emotion detection applications.

Future Work: We considered eight PTMs, and in the future, we plan to extend our work by incorporating more diverse speech/audio PTM architectures. We investigated four speech emotion corpora in this study, three in English and one in German; in the future, we aim to include more databases not just in English but also in other languages.

REFERENCES

[1] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[2] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[3] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[4] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[5] P.-H. Chi, P.-H. Chung, T.-H. Wu, C.-C. Hsieh, Y.-H. Chen, S.-W. Li, and H.-y. Lee, "Audio albert: A lite bert for self-supervised learning of audio representation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 344–350.

[6] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. Glass, "Contrastive audio-visual masked autoencoder," *arXiv preprint arXiv:2210.07839*, 2022.

[7] Y. Gong, Y. Chung, and J. R. Glass, "AST: audio spectrogram transformer," *CoRR*, vol. abs/2104.01778, 2021. [Online]. Available: https://arxiv.org/abs/2104.01778

[8] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[9] M. G. Campana, A. Rovati, F. Delmastro, and E. Pagani, "L 3-net deep audio embeddings to improve covid-19 detection from smartphone data," in *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 2022, pp. 100–107.

[10] E. Koh and S. Dubnov, "Comparison and analysis of deep audio embeddings for music emotion recognition," *arXiv preprint arXiv:2104.06517*, 2021.

[11] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 373–380.

[12] A. A. Razak, R. Komiya, M. Izani, and Z. Abidin, "Comparison between fuzzy and nn method for speech emotion recognition," in *Third International Conference on Information Technology and Applications (ICITA'05)*, vol. 1. IEEE, 2005, pp. 297–302.

[13] B. Vlasenko and A. Wendemuth, "Tuning hidden markov model for speech emotion recognition," *Fortschritte der akustik*, vol. 33, no. 1, p. 317, 2007.

[14] T. Iliou and C.-N. Anagnostopoulos, "Comparison of different classifiers for emotion recognition," in *2009 13th Panhellenic Conference on Informatics*. IEEE, 2009, pp. 102–106.

[15] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[16] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[17] M. Pastor, D. Ribas, A. Ortega, A. Miguel, and E. Solano, "Cross-corpus speech emotion recognition with hubert self-supervised representation," *Proceedings of the IberSPEECH*, pp. 76–80, 2022.

[18] A. Keesing, Y. S. Koh, and M. Witbrock, "Acoustic features and neural representations for categorical emotion recognition from speech." in *Interspeech*, 2021, pp. 3415–3419.

[19] B. T. Atmaja and A. Sasou, "Evaluating self-supervised speech representations for speech emotion recognition," *IEEE Access*, vol. 10, pp. 124 396–124 407, 2022.

[20] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," *Advances in neural information processing systems*, vol. 29, 2016.

[21] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

[22] 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.

[23] Y. Gong, Y.-A. Chung, and J. Glass, "Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3292–3306, 2021.

[24] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Byol for audio: Self-supervised learning for general-purpose audio representation," 2021. [Online]. Available: https://arxiv.org/abs/2103.06695

[25] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *CoRR*, vol. abs/1904.05862, 2019. [Online]. Available: http://arxiv.org/abs/1904.05862

[26] A. T. Liu, S. wen Yang, P.-H. Chi, P. chun Hsu, and H. yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2020. [Online]. Available: https://doi.org/10.1109%2Ficassp40776.2020.9054458

[27] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[28] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *CoRR*, vol. abs/2106.07447, 2021. [Online]. Available: https://arxiv.org/abs/2106.07447

[29] Y. Gong, C. J. Lai, Y. Chung, and J. R. Glass, "SSAST: self-supervised audio spectrogram transformer," *CoRR*, vol. abs/2110.09784, 2021. [Online]. Available: https://arxiv.org/abs/2110.09784

[30] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," 2022. [Online]. Available: https://arxiv.org/abs/2207.06405

[31] D. Chong, H. Wang, P. Zhou, and Q. Zeng, "Masked spectrogram prediction for self-supervised audio pre-training," 2022. [Online]. Available: https://arxiv.org/abs/2204.12768

[32] A. Keesing, Y. S. Koh, and M. Witbrock, "Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech," in *Proc. Interspeech 2021*, 2021, pp. 3415–3419.

[33] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[34] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[35] M. K. P.-F. Kate Dupuis, "Toronto emotional speech set (TESS) | TSpace Repository — tspace.library.utoronto.ca," https://tspace.library.utoronto.ca/handle/1807/24487, 2010, [Accessed 06-Nov-2022].

[36] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[37] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[38] 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 *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[39] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally *et al.*, "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 2022, pp. 125–145.

[40] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[41] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pretraining for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[42] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6907–6911.

[43] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2clip: Learning robust audio representations from clip," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4563–4567.

[44] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7169–7173.

[45] C. L. Moine, N. Obin, and A. Roebel, "Speaker attentive speech emotion recognition," *arXiv preprint arXiv:2104.07288*, 2021.

[46] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[47] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[48] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.