GENERALIZATION OF SELF-SUPERVISED LEARNING-BASED REPRESENTATIONS FOR CROSS-DOMAIN SPEECH EMOTION RECOGNITION

Abinay Reddy Naini**, Mary A. Kohler*‡*, Elizabeth Richerson*‡*, Donita Robinson*‡*, and Carlos Busso**

*Department of Electrical and Computer Engineering, The University of Texas at Dallas
‡Laboratory for Analytic Sciences, North Carolina State University

ABSTRACT

Self-supervised learning (SSL) from unlabelled speech data has revolutionized speech representation learning. Among them, wavLM, wav2vec2, HuBERT, and Data2vec have produced benchmark performances on automatic speech recognition. However, few studies have explored the generalization of SSL-based representations to different tasks based on paralinguistic information in speech such as emotion recognition. This paper explores the generalization of all four popular SSL models for *speech emotion recognition* (SER) when trained and tested in different domains. We aim to understand how adaptable these SSL representations are when using simple domain adaptation techniques. The evaluation considers emotional speech databases that deviate in language, recording conditions, and emotional distribution, providing very different target domains. The results reveal the necessity to fine-tune the representations for the SER downstream. As the differences between the source and target domain increase, we observe that the unsupervised domain adaptation techniques are more effective. The analysis in this study provides useful insights to understand the advantages of different representations for domain adaptation in SER.

Index Terms— Speech emotion recognition, self-supervised learning, unsupervised domain adaptation

1. INTRODUCTION

A significant breakthrough in the field of speech processing is the use of *self-supervised learning* (SSL) as an appealing strategy for harnessing insights from extensive unlabeled datasets. SSL-based solutions have demonstrated their effectiveness in enhancing downstream tasks such as *automatic speech recognition* (ASR) [1–7]. These models have been successfully used in other speech tasks, including *speech emotion recognition* (SER) [8–10]. Recognizing emotions from speech plays a major role in natural human-computer interaction [11, 12], so it is important to improve SER solutions before they are deployed in practical applications. Even though recent studies in SER showed significant gains by using SSL-based models for an SER task, very few studies have explored their performance in domain conditions that differ from the ones used for training the models. One of the major barriers to obtaining a deployable SER system is its generalization across different domain conditions. Hence, it is necessary to assess the SSL-based model's performance across different domain conditions.

Various SSL-based representations have been proposed for speech-related tasks [2,5,6,13]. Limited emphasis has been given to understanding the best representation for the SER tasks, particularly, the one which generalizes across different domain conditions. While most of the recent works in SER use pre-trained SSL models as a feature extractor, there are few attempts to fine-tune the SSL-based model to SER task using a downstream head, consisting of a simple *deep neural network* (DNN) [14]. Such a fine-tuning process has shown to help achieve better performance, when tested in similar domain conditions [8]. However, it is important to understand the generalization of the fine-tuned SSL representation across different domains, particularly when tested in significantly different conditions compared to the training data used for fine-tuning. A popular way to generalize any SER model to perform best on a target domain is to perform a domain adaptation [15–19]. Studies show that a significant performance gain in SER can be achieved by using unsupervised domain adaptation strategies, which do not require any labeled data from the target domain [16, 18]. Most of these studies were conducted with hand-crafted features. It is important to understand how adaptable are SSL-based representations when adapted to a particular target domain using simple adaptation schemes.

This paper addresses these important questions by considering the generalization of four popular SSL models that have resulted in state-of-the-art performance for ASR and various other speech-related tasks: wavLM [2], wav2vec2 [13], HuBERT [5], Data2vec [6]. To understand the effect of change in language, emotional distribution, and recording conditions, we consider cross-corpus evaluations with five popular datasets used for emotion recognition tasks that present different mismatches. The SSL models are either used as they are provided, or fine-tuned to SER tasks. We formulate the SER problem as a regression task to predict the emotional attributes of arousal (calm to active), valence (negative to positive), and dominance (weak to strong). Our experiments reveal that overall wavLM-based feature representation performs best in most of the cross-domain testing conditions. We also observed that fine-tuning the SSL-based models for SER tasks can significantly improve the performance of any attribute-specific SER task. We also observed unsupervised domain adaptation strategies are more effective if the target domain is significantly different from that of the source domain, leading to average relative improvements as high as 37.34%.

2. RELATED WORK

Several SSL models have been proposed for speech inspired by the performance gains obtained by the text-based model *bidirectional encoder representations from transformers* (BERT) [20] in *natural language processing* (NLP) tasks. Most of the SSL models proposed for speech follow the wav2vec2 [4] architecture, where the raw audio inputs are directly inputted into a *convolutional neural network* (CNN). Then, several transformer layers are utilized to encode the CNN outputs into frame-level contextualized representations. The encoder model produces context representations, which are used to predict future audio frames. Then, the model is optimized using contrastive loss to maximize the agreement between predicted and actual future frames. When the SSL representation is applied to a particular speech problem, the common approach is to fine-tune the model. For example, studies on ASR use the *connectionist temporal classi-*

This work was supported by Laboratory for Analytic Sciences (LAS)

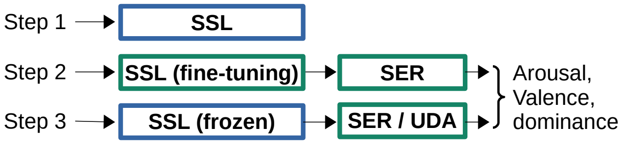

Fig. 1: Block diagram indicating different steps considered to obtain the SER models. UDA: Unsupervised domain adaptation.

fication (CTC) loss function [21] to obtain the best performance.

A few popular SSL representations were proposed following similar pre-training as in wa2vec2 [4, 13], with few differences such as using offline clustering and masked predictions to obtain HuBERT [5], and masked predictions of contextualized labels to obtain Data2vec [6]. Unlike these representations, wavLM [2] adds denoising objectives resulting in higher amounts of unlabelled data.

While the ASR task requires a frame-level prediction, SER task formulation involves an average time pooling to generate utterance-level predictions. Wang et al. [22] showed that fine-tuning the SSL representations for an SER task improves performance instead of using the frozen SSL model as a feature extractor. Hsu et al. [13] explore fine-tuning wav2vec2 representation to make it robust to various domains. However, to the best of our knowledge, this paper is the first attempt to understand various SSL representations in the context of unsupervised domain adaptation in SER.

3. SPEECH EMOTION RECOGNITION FRAMEWORK

We aim to study the performance and generalization of SSL in SER tasks. We are particularly interested in cross-corpus evaluations under different mismatches. Our analysis considers the four most popular existing SSL representations: wav2vec2 [13], wavLM [2], HuBERT [5], Data2vec [6].

Figure 1 shows the three steps we considered for obtaining various SSL-bases SER frameworks. In the first step, we consider all four pre-trained SSL representations with large models from the huggingface library [23]. In the second step, we consider a simple DNN block as a downstream head for fine-tuning the SSL representations for the SER task. We considered the *concordance correlation coefficient* (CCC) for arousal, valence, and dominance prediction as the loss function to train the SER downstream head along with fine-tuning the SSL representation. For the last step, we consider two cases. For the first case (case-1), we froze the parameters of the fine-tuned SSL block and then used it as a feature extractor for training the SER block to obtain the best performance on an individual emotional attribute (i.e., arousal, valence, or dominance). For the second case (case-2), we consider the fine-tuned SSL representations as feature extractors, like case-1. Instead of using a simple DNN for the SER task using only the CCC loss, we consider two popular *unsupervised domain adaptation* (UDA) strategies used in SER (Sec. 3.2). Case-2 explores the effectiveness of common UDA strategies on SSL-based models in different conditions for the source (train) and target (test) domains.

3.1. SSL representations

wav2vec2: This study considers the wav2vec2-large-robust model proposed by Hsu et al. [13], which considered a similar architecture as the wav2vec2-large model [4]. During the pre-training of the wav2vec2-large model, a portion of the frames within the latent speech representations generated by the CNN-based encoder undergo masking and are then inputted into the transformer encoder. The model's training objective is to minimize the contrastive loss, which maximizes the agreement between predicted and actual future frames within the latent speech representations. The resulting wav2vec2-large model is further trained using diverse domain datasets to produce the wav2vec2-large-robust model.

Data2vec: We also considered Data2vec [6], which generates contextualized targets. The training process for Data2vec involves forwarding the unmasked speech signal to a teacher model, which maintains an exponential moving average of the student model's weights. The student model is then tasked with predicting the activations of the teacher model from the masked speech signal, to minimize a regression loss.

HuBERT: The HuBERT model [5] follows a similar architecture to that of the wav2vec2. Unlike wav2vec2, HuBERT optimizes the model using a cross-entropy loss for masked time steps during the pre-training. The targets are not updated simultaneously with the model. Like the BERT model [20], a section of the mel-spectrogram input features is masked, and the model is trained to predict the masked frames using the last layer of the transformer and the L1 loss. However, HuBERT utilizes an offline clustering step to provide aligned target labels for a BERT-like prediction loss.

wavLM: wavLM [2] adopts a similar approach as HuBERT, with the addition of a speech-denoising task involving the generation of mixtures comprising speech and noise. In this pretext task, the goal is to predict the targets produced from the unmasked clean speech within the context of noisy masked speech.

3.2. Unsupervised Domain Adaptation Strategies

We consider 2 alternative methods to adapt the SSL representations:

Domain Adaptation using Ladder Network (DAL) The first UDA strategy considered in this study is the ladder network approach [24], which was used by Parthasarathy and Busso [16] for SER tasks. An encoder-decoder architecture is employed along with horizontal lateral connections at each layer. Gaussian noise is introduced in the encoder, and the objective is to reconstruct a clean version of the corresponding encoder's input. Each decoder layer aims to recover an uncorrupted version of the input that was fed into the corresponding encoder layer. The cost function consists of two components with different weights: one for the task classifier, positioned at the end of the encoder, and another for the reconstruction losses. While both losses are utilized for the labeled source domain, only the reconstruction losses are employed for the unlabeled target domain.

Adversarial Domain Adaptation (ADA) The second UDA strategy is *adversarial domain adaptation* (ADA) [25], following the method detailed by Abdelwahab and Busso [18]. This method involves employing task classifier, and domain classifier branches on top of a feature representation. In ADA, a *gradient reversal layer* (GRL) is introduced between the domain classifier and the feature representation network. The GRL compels the domain classifier to maximize classification error when distinguishing between the source and target domains. As a result, the feature representation network is trained to produce similar responses for data from both domains, reducing the domain mismatch. During training with the labeled source domain data, both task and domain classifier losses are considered. However, when training with unlabeled target domain data, only the domain classifier loss is considered. By using alternate batches of source and target domain data, this approach ensures optimal task classification accuracy while making samples from both domains indistinguishable.

3.3. Strategies Explored in the Analysis

We consider four different implementations for the experiments. In the first implementation (L), we obtain the SSL representations to

Table 1: CCC coefficient for all four SSL representations for the Large (L), Large-Robust (LR), Adversarial Domain Adaptation (ADA), and Domain Adaptation using Ladder Network (DAL) for arousal, valence, and dominance. The table also lists the within-corpus performance in the MSP-Podcast corpus. We compared the results across 4 SSL models (e.g., columns in the table). A result with *† is statistically significantly better than a result with *. A result with any symbol is statistically significantly better than a result with no symbol.

	MSP-Podcast		MSP-IMPROV				IEMOCAP				VAM				Nixon's Tapes			
	L	LR	L	LR	ADA	DAL	L	LR	ADA	DAL	L	LR	ADA	DAL	L	LR	ADA	DAL
	Arousal																	
wavLM	.625*†	**.679***	.571*	**.612***†	.588	.598*	.622*	**.683***	.681*†	.674*	.294*	.301*	.377*†	.368*	.301*	.312	.421*	.428*†
wav2vec2	.617*	.663*	.553	.601*	.599*	.590*	.628*	.671*	.662*	.669*	.281*	.291*	.352*	.337	.291*	.318	**.432***	.430*†
HuBERT	.624*†	.657*	.551	.602*	.602*	.597*	.619*	.667*	.681*†	.679*	.297*	.302*	**.379***†	.370*	.282*	.299	.401	.392*
Data2vec	.606	.639	.544	.589	.583	.579	.594	.631	.628	.622	.244	.261	.316	.322	.257	.291	.387	.368
	Valence																	
wavLM	.531*	.655*	.413*	.468	.493*	.488*	.413*	.463*	.482*	.483*	.237*	.267*	.333*	.331*	.427*	.440	.492*	.473
wav2vec2	.557*†	**.657***	.424*	.488*	**.499***	.491*	.409*	.459*	.472*	.481*	.242*	.283*†	**.353***†	.349*	.434*	.450	**.502***	.497*
HuBERT	.568*†	.646*	.407*	.461	.480*	.478*	.411*	.451*	.488*	**.490***	.233*	.259*	.347*†	.339*	.408	.436	.497*	.484
Data2vec	.512	.628	.392	.451	.461	.461	.392	.432	.441	.447	.211	.231	.297	.302	.401	.427	.459	.461
	Dominance																	
wavLM	.552*	**.628***	.403*	.438*	.477*†	**.478***	.414*†	.457*†	.469*	.469*	.267*	.281*	.332*	.332*	.332*†	.379*	.391*	.398*
wav2vec2	.537	.607	.394*	.423*	.445*	.441	.402*	.451*†	.469*	**.471***	.254*	.274*	.328*	.318	.317*	.341	.383*	**.405***
HuBERT	.528	.604	.386*	.408	.441*	.438	.394*	.439*	.459*	.461*	.250*	.276*	**.339***	.338*	.313*	.331	.377*	.390*
Data2vec	.520	.598	.371	.403	.428	.430	.377	.418	.429	.432	.237	.253	.294	.301	.297	.322	.351	.354

be used as feature extractors. Then, we train the SER block for the downstream task (Step-1 + Step-3 (case-1)). This setting does not adapt the parameters of the SSL for the SER task (i.e., Step 2). In the second implementation (LR), we obtain the Large-Robust SSL representations by fine-tuning the large models using a downstream head with the SER task (Step-1 + Step-2 + Step-3 (case-1)). In the third (ADA) and fourth (DAL) implementations, we use the LR model as a starting point, using the two UDA strategies mentioned in Section 3.2 (Step-1 + Step-2 + Step-3 (case-2).

4. EXPERIMENTAL SETTING

4.1. Emotional Databases

An important aspect of the analysis is the cross-corpus evaluation to explore diverse mismatches between the source and target domains. Our evaluation consider the MSP-podcast corpus as the source domains, and four other emotional databases as the target domains.

MSP-Podcast corpus: We consider the MSP-Podcast corpus [26] for the source domain. We use release 1.11 consisting of 151,654 English speaking turns from different audio recordings with Creative Commons licenses. The training set includes 84,030 speaking turns. All the speaking turns are obtained after removing background music, noise, and speech overlaps with a minimum duration of 2.7 seconds. Each turn is annotated by at least five annotators for emotional attributes and primary and secondary emotional categories. This study uses emotional attributes for arousal, valence, and dominance.

MSP-IMPROV: We consider the MSP-IMPROV dataset [27] as one of the target domains. The MSP-IMPROV corpus consists of dyadic interactions between 12 actors consisting of a total of 8,438 speaking turns. The MSP-IMPROV database also provides annotations for the emotional attributes of arousal, valence, and dominance. Each speech file is annotated by at least five annotators. Unlike the MSP-Podcast corpus, the audio was recorded in a closed environment, making it perfect for testing the domain mismatch. Data from the six actors from the first three sessions are used as the test set. We have reserved the remaining sessions for the training of UDA models as one of the unlabeled target datasets. The corpus is in English.

IEMOCAP corpus: The USC-IEMOCAP corpus [28] comprises interactions between actor pairs engaged in improvisational scenarios. This database encompasses 10,527 speaking turns involving ten actors across five dyadic sessions, with approximately 12 hours of audiovisual content. The annotations are provided at the turn level and encompass categorical labels (e.g., happiness, sadness, anger) along with three attributes (activation, valence, dominance) rated on a discrete scale from 1 to 5. This corpus is a common benchmark for emotion recognition tasks.

VAM Dataset: The VAM dataset [29] comprises 12 hours of audiovisual recordings captured during a German TV talk show, where families expressed their relational problems. Within this corpus, there are 947 utterances that capture spontaneous emotions expressed by 47 guests participating in unscripted, genuine conversations. Given the nature of the TV program, there is also an emotional distribution mismatch where the conversations are predominantly negatives. Therefore, the VAM dataset is ideal to understand the performance of SSL representations in the presence of language and emotional distribution mismatches. The labels include rating for arousal, valence and dominance. We have used 947 utterances in a two-fold cross-validation strategy. One set is used for testing and the other set is used as the unlabeled data to train the UDA models.

Nixon's Tapes Dataset: The Nixon's Tapes dataset [30] is part of the declassified President Nixon's Oval Office recorded conversations from 1972. The recordings are divided into smaller segments using a similar strategy than the one used for the MSP-Podcast corpus [26] to obtain 293,741 segments of speech between 3 seconds to 11 seconds. We have used a 5dB SNR filter along with a random sampling to obtain 80,000 speech segments, which are used as the unlabeled set to train the UDA models. A subset consisting of 1,427 speech segments has been annotated by at least five annotators recruited by our laboratory for emotional attributes and primary and secondary emotional categories following the protocol used for the MSP-Podcast corpus. We use the emotional attribute labels, using this set of 1,427 speech segments as the test set. This corpus is ideal to evaluate recoding mismatches, given the conditions used to record these tapes (e.g., distant speech, noisy recording). This is the first time that this corpus is used for SER tasks.

4.2. Implementation

We obtain four models for each SSL representation, as explained in Section 3: large (L), large-robust (LR), adversarial domain adaptation (ADA), and domain adaptation using ladder network (DAL). An exception is the wav2vec2 feature representation, where a large-robust representation is obtained by pruning the top 12 transformer layers from the model during the fine-tuning stage (Step-2), which is shown to preserve the recognition performance with fewer param-

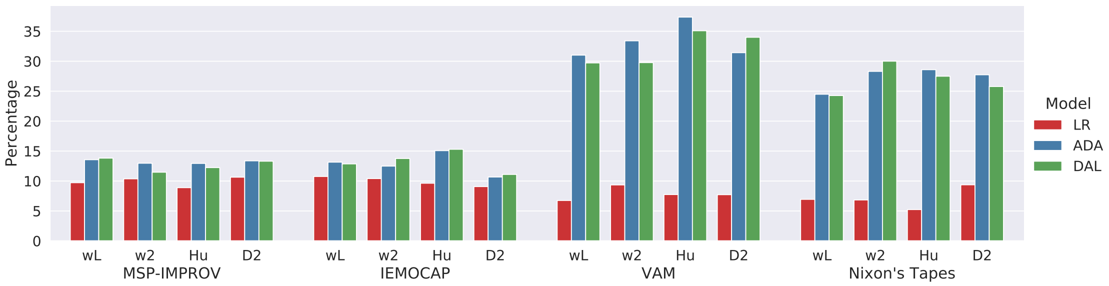

Fig. 2: Relative performance gain for each of the LR, ADA, and DAL models over the Large (L) model. The results are averaged across all three emotional attributes. wL: wavLM, w2: wav2vec2, Hu: HuBERT, D2: Data2vec.

eters [8]. The SER block considered for Step-2 and Step-3 contains two fully connected layers of 1,024 nodes with layer normalization, and the *rectified linear unit* (ReLU) as the activation function. For the fine-tuning in Step-2, we use a combination of CCC obtained from all three attributes (arousal, valence, and dominance) as a loss function. In Step-3, we consider a weighted combination of CCC loss obtained from predicting three attributes as a loss function to train each attribute-specific downstream SER task. For the ADA strategy, we consider the same SER block as before, using it as our feature representation. We add a hidden layer of 128 nodes for task and domain classifier layers with ReLU activation. For the DAL strategy, we use two hidden layers with 256 nodes, with a ReLU activation and a linear activation for the task classifier layer. For the fine-tuning process in Step-2, we have used an EC2 g5.4xlarge instance tailored with an NVIDIA A10G GPU, with Adam optimizer [31] using a learning rate of 10e-5. For all other experiments, we used an NVIDIA GeForce RTX 3090 GPU.

5. EXPERIMENTAL RESULTS

We report the performance in CCC for all the experiments. We evaluate if the results are statistically significant by randomly splitting the test set into 20 subsets of similar size for the MSP-Podcast, MSP-IMPROV, and IEMOCAP databases, and into 5 subsets for the Nixon's Tapes and VAM databases due to their limited size. We assert statistical significance at a p-value less than 0.05, using a two-tailed t-test. Table 1 shows the CCC coefficient for the within and cross-domain performances using the four SSL representations for arousal, valence, and dominance. For the first experiment, we want to understand the importance of fine-tuning SSL-based models for the SER task. The results obtained on the MSP-Podcast test set (within-corpus) are reported in the first two columns of the table. We can observe that in all cases the LR model performs significantly better than the L model. Particularly, for the valence and dominance case we observed an average relative gain of $\sim$16% in CCC. This result indicates the importance of fine-tuning SSL-based models for the SER task. For the within-corpus assessment, the wavLM-based models achieve the best result among the 4 SSL representations.

Similar to the within-corpus performance, we observe significant performance improvements for the LR models over the L models across-corpus evaluations, demonstrating the importance of fine-tuning the SER models. Overall, the wavLM, wav2vec2, and HuBERT-based models perform better than the Data2vec feature representation across all datasets. The VAM dataset produces the lowest performance among all other datasets. In addition to language and label mismatches, this result can also be explained by the average duration of its sentences, which is smaller than the average duration of the sentences in the other databases. Across the datasets without domain adaptation, the wavLM-based representations perform the best for the arousal task, and wavLM, wav2vec2, and HuBERT resulted in a similar performance for the valence and dominance tasks. The lower performance of Data2vec could be due to the differences in pre-training, which predicts contextualized latent representations instead of modality-specific data. Overall LR models perform significantly better than L models across datasets. However, the performance gain for the LR model is much higher for the MSP-IMPROV, and IEMOCAP corpora, which are closer to the training domain (MSP-Podcast), than for the VAM, and Nixon's Tapes datasets. Both unsupervised domain adaptation strategies are very effective in achieving the best performance for the target domains, particularly in the case of VAM, and the Nixon's Tapes dataset, whose domain conditions are significantly different from the conditions in the MSP-Podcast corpus.

To understand the relative performance gain using fine-tuning and UDA strategies, we computed the performance improvement for the LR, ADA, and DAL models over their corresponding L model. This analysis is conducted across all the databases, averaging the results across emotional attributes (arousal, valence, and dominance). Figure 2 shows the performance gain in percentage, where the three color bars represent the LR, ADA, and DAL models for all four representations for each dataset. The figure shows significant gains using both domain adaptation strategies for the VAM, and Nixon's Tapes databases. The relative improvements of these UDA strategies in the other two databases are positive but more limited. This result suggests that UDA strategies are more effective if the target domain is significantly different from the source domain. Among the four SSL-based models, Hubert resulted in higher gain from the LR model to the UDA-based models across all databases (i.e., the difference between the blue bars and red bars in the figure).

6. CONCLUSIONS

This study provided important insights about the use of SSL representations on SER tasks. We evaluated four common SSL feature representations in cross-corpus evaluations, using different target domains to assess their generalization across different source-target mismatches. We also explored how adaptable these models are using common unsupervised domain adaptation strategies. The analysis showed that fine-tuning the SSL models is crucial to achieve good performance in cross-corpus evaluations. We observed that wavLM-based models produced better results than the other SSL models in most experiments. The results on the VAM and Nixon's Tape recordings showed that when the mismatch between the source and target domain increases, the UDA are more successful leading to relative improvements as high as 37.34% over using the original SSL representations. These findings offer important guidance for utilizing SSL-based feature extractors in SER tasks, emphasizing the importance of adaptability and generalization across various domains.

7. REFERENCES

[1] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, October 2022.

[2] 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, July 2021.

[3] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2021)*, Cartagena, Colombia, December 2021, pp. 244–250.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS 2020)*, Virtual, December 2020, vol. 33, pp. 12449–12460.

[5] W.-N. Hsu, Y.-H. H. Tsai B. Bolte, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[6] 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 (ICML 2022)*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162, pp. 1298–1312. Proceedings of Machine Learning Research (PMLR), Honolulu, HI, USA, July 2022.

[7] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, "SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing," in *Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, Dublin, Ireland, May 2021, pp. 5723–5738.

[8] 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," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10745–10759, September 2023.

[9] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Computation and memory efficient noise adaptation of Wav2Vec2.0 for noisy speech emotion recognition with skip connection adapters," in *Interspeech 2023*, Dublin, Ireland, August 2023, pp. 1888–1892.

[10] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Adapting a self-supervised speech representation for noisy speech emotion recognition by using contrastive teacher-student learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes island, Greece, June 2023, pp. 1–5.

[11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, January 2001.

[12] R. Picard, *Affective Computing*, MIT Press, Cambridge, MA, USA, 1997.

[13] 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 e-prints (arXiv:2104.01027)*, pp. 1–9, April 2021.

[14] R. Fan, Y. Zhu, J. Wang, and A. Alwan, "Towards better domain adaptation for self-supervised models: A case study of child ASR," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1242–1252, October 2022.

[15] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Separation of emotional and reconstruction embeddings on ladder network to improve speech emotion recognition robustness in noisy conditions," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 2871–2875.

[16] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2697–2709, September 2020.

[17] A. Reddy Naini, M.A. Kohler, and C. Busso, "Unsupervised domain adaptation for preference learning based speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes island, Greece, June 2023, pp. 1–5.

[18] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, December 2018.

[19] W.-C. Lin, K. Sridhar, and C. Busso, "DeepEmoCluster: A semi-supervised framework for latent cluster representation of speech emotions," in *IEEE international conference on acoustics, speech and signal processing (ICASSP 2021)*, Toronto, ON, Canada, June 2021, pp. 7263–7267.

[20] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186.

[21] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *International Conference on Machine Learning (ICML 2006)*, Pittsburgh,PA, USA, June 2006, pp. 369–376.

[22] 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 e-prints (arXiv:2111.02735)*, pp. 1–7, November 2021.

[23] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, and Q. Lhoest amd A.M. Rush, "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv e-prints (arXiv:1910.03771v5)*, pp. 1–8, October 2019.

[24] A. Rasmusi, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in neural information processing systems (NIPS 2015)*, Montreal, Canada, December 2015, pp. 3546–3554.

[25] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, April 2016.

[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, October-December 2019.

[27] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.

[28] C. Busso and S.S. Narayanan, "Recording audio-visual emotional databases from actors: a closer look," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 17–22.

[29] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, October-November 2007.

[30] *OVAL 741-2; June 23, 1972; White House Tapes; Richard Nixon Presidential Library and Museum, Yorba Linda, California.*, 1972.

[31] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, CA, USA, May 2015, pp. 1–13.