ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890089

Domain-Specific Adaptation in Speech Emotion Recognition Using Emotional Distribution Alignment

Abinay Reddy Naini[1], Donita Robinson[2], Elizabeth Richerson[2], Carlos Busso[1,3]
[1]*The University of Texas at Dallas, Richardson, TX, 75080, USA*
[2]*North Carolina State University, Raleigh, NC, 27695, USA*
[3]*Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, USA*
abinayreddy.naini@utdallas.edu, drobins7@ncsu.edu, ericher@ncsu.edu, busso@cmu.edu

***Abstract*—This work addresses the challenge of building speech emotion recognition models that generalize effectively across different domains, particularly when only limited target domain data is available with or without emotional label information. Traditional models often struggle with cross-domain performance due to the variability in emotional expressions and the lack of alignment between the training and target domains. We propose a novel approach that prioritizes aligning the emotional label distribution of the training data with that of the target domain by undersampling the source domain. Even though we intentionally reduce the size of the training set from the source domain, the emotional content alignment leads to clear performance improvements, outperforming models trained with the complete training set. This strategy highlights the importance of aligning emotional attributes during training, helping to create robust emotion recognition models across diverse applications. Our findings also reveal that performance significantly improves when even a small amount of labeled target domain data is available, allowing for a more accurate assessment of the emotional distribution in the target domain.**

***Index Terms*—Speech emotion recognition, Target-domain adaptation,**

I. Introduction

Emotion recognition, particularly in speech, plays a critical role in various human-computer interaction systems, enhancing their ability to understand and respond to users' emotional states [1]–[5]. Effective emotion recognition can significantly improve the user experience in applications such as virtual assistants, mental health monitoring, and customer service by enabling systems to offer more personalized and empathetic interactions [6]–[8]. However, building models that generalize well across different domains remains a challenge due to the variability in emotional expressions across cultures, languages, and contexts [9], [10]. To increase the robustness of *speech emotion recognition* (SER), it is essential to develop models that perform robustly not only on the datasets they are trained on but also on unseen domains, making cross-domain emotion recognition a key focus area in the field [11]–[15].

Despite the growing interest in cross-domain emotion recognition, existing approaches often fail to perform satisfactorily on new datasets. Many models tend to overfit the training data, capturing specific patterns that do not generalize well to other domains [16]. This issue is particularly problematic when the target domain has limited labeled data available for training. Previous attempts to improve cross-domain performance have included techniques such as domain adaptation and transfer learning. Even though recent *unsupervised domain adaptation* (UDA) strategies showed significant improvements in cross-domain SER tasks, the overall performance is still very low compared to within-corpus results. Naini et al. [17] demonstrated that UDA strategies show more relative improvement when the target dataset differs significantly from the training data. Particularly, when the training and target domain datasets contain a significant shift in the emotional distribution, trained models tend to be biased towards the content in the training domain. While acoustic domain mismatches between the source and target domains are definitely essential to reduce, we argue that matching the distribution of the emotional content is equally important. SER studies often neglect this research direction.

This study proposes a novel approach to aligning the distribution of the emotional content in the source and target domains. The approach selectively uses only the training data that share a similar emotional distribution with the target domain. We implement this approach using pseudo labels obtained by predicting the emotional content in the target domain. Our approach undersamples the training set, creating training sets with emotional distributions that resemble the distribution of the target domain.

We evaluate our proposed approach for the prediction of the emotional attributes valence, and arousal. Our experiments demonstrate that this distribution-aligning approach significantly improves cross-domain emotion recognition performance. When the training data's emotional attributes are closely aligned with those of the target domain, the model is better able to capture the nuances of emotion in the new context, leading to more accurate predictions. Interestingly, we observe that even when we reduce the size of the training set to achieve this alignment, the model's performance still surpasses that of models trained on the entire training set. This finding highlights the importance of mitigating the distribution mismatch in the emotional content conveyed in the source and target domains. The results underscore the potential of strategically curating training data based on emotional attribute distribution to enhance cross-domain performance. By prioritizing the alignment of these attributes between training and target domains, we can develop more adaptable and effective emotion recognition models, paving the way for more accurate

and reliable applications in diverse real-world settings.

II. Related Work

Improving the performance of SER models across different test conditions and domains has been a focal point of research. Early efforts, such as those by Han et al. [18], focused on enhancing the consistency of emotion recognition across sub-classification tasks, demonstrating that more consistent emotion rankings can lead to better generalization. Similarly, Martinez et al. [19] showed that transforming emotional attributes into ranks rather than discrete classes can improve model performance in varied conditions. The advances in *self-supervised learning* (SSL) models such as wav2vec2 [20] have further pushed the boundaries, leveraging large amounts of unlabeled data to learn robust speech representations. However, these models often require careful adaptation to maintain performance across different domains, as highlighted by research from Hsu et al. [21]. In SER, SSL-based models have led to higher performance than models trained with traditional speech representations [22]–[24]. Naini et al. [17] showed that UDA strategies are also effective with SSL-based models.

Addressing the variability in emotional attributes across domains, recent research has begun to focus on aligning arousal, valence, and dominance distributions between training and target datasets. Traditional SER models have often neglected the importance of these distributions, which are critical for cross-domain generalization. Lotfian and Busso [25] emphasized the benefits of managing emotional attributes carefully, suggesting that aligning these distributions can enhance model robustness. Building on this insight, our work proposes a method that selectively curates training data to match the AVD distribution of the target domain, resulting in improved performance in cross-domain SER scenarios. This approach highlights the potential of emotional attribute alignment in enhancing the generalization of SER models across diverse applications.

III. Methodology

The proposed strategy improves SER model performance across domains by aligning the emotional attribute distributions of the training and target data, addressing distributional differences that hinder generalization. In the following subsections, we present the proposed distribution alignment strategy for two scenarios: one where a small portion of labeled target domain data is available, and another where no target domain labels are accessible.

A. Distribution Alignment with Labeled Target Domain Data

In this scenario, we aim to align the emotional attribute distributions between the source and target domains to improve the generalization of the SER model. The emotional attributes (arousal, valence, and dominance) are often distributed differently across domains, and these discrepancies can lead to suboptimal model performance when training on the entire source dataset without considering the target domain distribution. To address this problem, we align the source data to match the emotional attribute distribution of the target domain. The emotional score distribution of the target domain is obtained, and a subset of the source domain data is selected to mirror this distribution. The process involves dividing the target dataset into n bins based on the emotional attribute scores. The emotional attributes are often annotated using a Likert scale. First, we normalize the scores in the range -1 and 1. Then, we use the n bins to divide the space. Suppose $t_1, t_2, \ldots, t_n$ represents the set of samples in each of the target set bins, where t_1 represents all the samples in the first bin. Similarly, $s_1, s_2, \ldots, s_n$ represent the samples from the source database that are included in the bins. Then, we estimate

$$K = \min\left\{\frac{|s_1|}{|t_1|}, \frac{|s_2|}{|t_2|}, \ldots, \frac{|s_n|}{|t_n|}\right\} \tag{1}$$
$$\tilde{s_i} \subseteq s_i \quad \text{such that} \quad |\tilde{s_i}| = K \cdot |t_i|, \quad 1 \leq i \leq n$$

where $|A|$ represents the cardinality of the set A, and $\tilde{s_i}$ represents the samples in the resampled source dataset that is used for training. By aligning the emotional attributes between the training and target datasets, the model is better equipped to generalize to the target domain, resulting in improved performance. This method leverages the availability of labeled data in the target domain to create a distributional aligned training set, enabling better generalization across domains.

B. Distribution Alignment without Target Domain Labels

In the previous scenario, we assumed access to a small labeled subset of the target domain. However, in many real-world scenarios, obtaining labeled target domain data is not feasible due to time, cost, or privacy constraints. To address this challenge, we propose an approach that leverages pseudo-labels predicted by a state-of-the-art SER model. Instead of using actual target domain labels, this method uses a pre-trained SER model to predict the emotional attribute scores (arousal, valence, and dominance) for the target dataset. These predicted labels, denoted as $\hat{t}_i$ for each bin i, serve as a proxy for the true emotional attribute distribution. Although these pseudo-labels are estimates, they allow us to perform distribution alignment similarly to the case with labeled data. After obtaining the pseudo-labels, we replace $\hat{t}_i$ in place of t_i in Eq. 1 to obtain a resampled source dataset that is used for training.

We further enhance this approach by introducing an iterative procedure. After obtaining the initial pseudo-labels and resampling the source dataset to match the predicted target distribution, we train the model using this resampled source data. Once trained, the updated model is used to generate a second iteration of pseudo-labels for the target dataset, potentially improving the accuracy of the predicted emotional attribute scores. We then resample the source dataset again, using the newly predicted labels to align the source distribution with the updated target pseudo-labels. This iterative process continues, with the source data being resampled and the model retrained until the performance improvement plateau on a development set of the source domain. The iterative refinement

of pseudo-labels allows the model to gradually better align the source distribution with the target domain, resulting in enhanced generalization performance.

IV. Experimental Setting

A. Resources

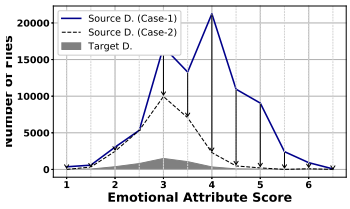
In our experiments, we leverage four public corpora. We use the MSP-Podcast dataset [26] as the source domain. We use the VAM [27], RECOLA [28], and WHiSER [29] datasets as target domains. These corpora are selected since they have different mismatches with the source domain (e.g., language, emotional content, environmental recordings).

The MSP-Podcast corpus [26] is the primary dataset used in our study for English-speaking data. We work with version 1.11 of this corpus, which includes 151,654 speaking turns gathered from various audio recordings, all available under Creative Commons licenses. For training, we use a subset consisting of 84,030 speaking turns. The dataset also includes a development set with 19,815 segments and a test set, referred to as test set 1, which contains 30,647 segments. Each speaking turn in the dataset has been annotated by at least five different raters, who provided ratings on the emotional attributes valence (negative versus positive), arousal (calm versus active), and dominance (weak versus strong) using a 7-Likert scale. They also annotated the primary and secondary emotions. We focus on the prediction of emotional attributes.

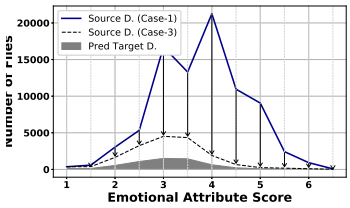
This study uses three different datasets as target domains to evaluate our methodology. The first dataset, VAM [27], consists of 12 hours of audiovisual recordings from a German TV talk show, where participants openly discussed their personal and relational issues. The dataset includes 947 utterances capturing spontaneous emotions expressed by 47 participants in unscripted conversations. Due to the nature of the program, the emotional distribution is predominantly negative, making it a challenging dataset for emotion recognition. The second dataset, RECOLA [28], is part of the REmote COLlaborative and Affective database, which was originally used in the *Audio/Visual Emotion Challenge* (AVEC). This study uses the full RECOLA dataset, which contains 46 recordings divided into training, development, and testing sets, balanced by gender, age, and native language. The VAM and RECOLA datasets are used to evaluate our approach to predicting the emotional attributes arousal, and valence. We also include the WHiSER corpus [29], which comprises 5,427 speech segments derived from President Nixon's Oval Office recordings between 1971 and 1973 [30]. These segments, ranging from 3 to 11 seconds, provide a valuable test set for evaluating emotional attributes in challenging conditions, such as distant speech and noisy environments.

B. Experimental Setup

This section describes the four experimental cases considered in our study. Each case explores different approaches to target domain adaptation, from using the full source domain dataset to more advanced techniques involving distribution alignment and pseudo-labeling. Fig. 1 illustrates the experimental setup, showcasing the differences between the various cases we considered. **Case 1** is a model trained on the entire MSP-Podcast source dataset without any modifications to the data distribution. The model predicts the emotional attributes (arousal, and valence) based on the source domain. The model is then evaluated on the target datasets (VAM, RECOLA, and WHiSER) without considering any alignment between the source and target domain distributions. This approach serves as a baseline to observe how well the model generalizes across domains without any distribution adaptation. This case's performance illustrates the challenges of directly applying a model trained on one domain to another without domain adaptation. In **Case 2**, the performance is improved by aligning the emotional attribute distributions of the source dataset with that of the target domain. We assume that a representative portion of the target domain is labeled. The target domain's emotional attributes are divided into bins, and a subset of the source dataset is resampled to match the target distribution. As shown in Fig. 1(a), the solid line represents the source distribution, while the dotted line represents the resampled subset of source data. Training the model on this aligned source data better reflects the target domain's emotional characteristics, leading to improved generalization. In **Case 3**, we perform distribution alignment using pseudo-labels, which are generated by a pre-trained SER model (as shown in Fig. 1(b)). Here, no labeled data from the target domain is assumed to be available. Instead, the model predicts the emotional attributes of the target dataset, and these predicted labels, or pseudo-labels, are used to estimate the emotional attribute distribution of the target domain. The source data is then resampled to match this pseudo-distribution. Fig. 1(b) illustrates this, where the solid line represents the full source data, and the dotted line represents the resampled source data based on the pseudo-



(a) Case 1 vs Case 2



(b) Case 1 vs Case 3

Fig. 1: (a) Illustration of source data used for training in Case 1 vs Case 2 (b) Illustration of source data used for training in Case 1 vs Case 3

TABLE I: *concordance correlation coefficient* (CCC) of the baselines (case-1) and proposed methods for arousal, valence and categorical. The table reports results for Case-1, Case-2, case-3 and Case-4 (Sec. IV-B). The symbol * indicates that using the proposed framework significantly improves the corresponding baseline method.

	Case 1	Case 2	Case 3	Case 4
	Arousal (CCC)			
VAM	0.301	0.314*	0.310*	0.312*
RECOLA	0.538	0.557*	0.551*	0.551*
WHiSER	0.299	0.308*	0.303	0.303
	Valence (CCC)			
VAM	0.261	0.316*	0.298*	0.307*
RECOLA	0.493	0.509*	0.503*	0.507*
WHiSER	0.392	0.417*	0.408*	0.411*

labels. This process enables the model to generalize better to the target domain, even in the absence of actual target labels. **Case 4** builds on this strategy by introducing an iterative pseudo-labeling approach. After training the model using the resampled source data from Case 3, the newly trained model generates a second iteration of pseudo-labels for the target dataset. The source data is then refined and resampled again to match these updated pseudo-labels. This iterative process continues until the performance improvement plateaus, further refining the source-target distribution alignment and enhancing generalization across domains.

C. Implementation Details

In our experiments, we utilized the wavLM-large model [31], a state-of-the-art self-supervised learning (SSL) model from the Hugging Face library [32], trained on English speech data. We select this model due to its robustness in speech-related tasks and its ability to generalize across domains. For emotional attribute prediction (arousal, valence, dominance), we fine-tuned the wavLM-large model by adding a downstream *deep neural network* (DNN) head. The fine-tuning process was guided by a cost function that optimized the *concordance correlation coefficient* (CCC) for all three attributes. After fine-tuning, we freeze the model's parameters, and the SSL model is used as a feature extractor to train the SER model on the source domain data.

We utilized two high-performance resources to training and testing the models. We performed fine-tuning and model training on an EC2 g5.4xlarge instance equipped with an NVIDIA A10G GPU while we ran other experiments on an NVIDIA GeForce RTX 3090 GPU. We employed the Adam optimizer [33] with a learning rate of 10e-5, ensuring that the model converged efficiently.

V. Experimental Results

To determine if the results are statistically significant, we employed a one-tailed t-test, considering significance at a p-value less than 0.05. This statistical analysis verified that the observed improvements across different cases were not due to random chance. Table I shows the results. The proposed method of aligning the emotional attribute distributions of the source and target domains led to better performance.

We observed notable improvements in CC across the VAM, RECOLA, and WHiSER datasets. For arousal, Case 2, which involved matching the training data distribution with the target domain using actual labels, achieves significant improvements over the results in Case 1. The approximate gains were $\sim 4.3\%$ for VAM, $\sim 3.5\%$ for RECOLA, and $\sim 3.0\%$ for WHiSER. These results indicate that aligning the emotional distribution with the target domain yields consistent benefits. Case 3, where we use pseudo-labels to estimate the target domain's distribution, also showed improvements over Case 1 across all datasets. However, the gains were slightly smaller than the results for Case 2. For WHiSER, the improvement in arousal was more modest ($\sim 1.3\%$) since the arousal distribution of the WHiSER corpus is relatively similar to that of its distribution on the source domain (i.e., MSP-Podcast). Case 4, which applied the iterative refinement process, resulted in marginal gains compared to Case 3, particularly in the VAM and RECOLA databases, where the iteration process helped improve the performance further. For valence, the results are more significant. Case 2 produced the highest gains, with an improvement of $\sim 21\%$ for the VAM corpus, which is the dataset with the most distinct valence distribution compared to the source domain. We observe improvements of $\sim 3.2\%$ for the RECOLA database and $\sim 6.3\%$ for the WHiSER database. The WHiSER corpus benefited more in the valence prediction task than in arousal, as the valence distribution of WHiSER is very different from its distribution in the source domain. Case 3 also resulted in moderate improvements across all datasets. Using pseudo-labels leads to better performances than Case 1, approaching the CCC values observed in Case 2, where we use the actual labels from the target domain. Case 4 showed a slight improvement over Case 3, but the overall improvements were smaller compared to Case 2. These results reinforce the importance of aligning the training data's emotional distribution with that of the target domain.

VI. Conclusions

In this study, we proposed a novel approach to improving cross-domain speech emotion recognition by aligning the emotional attribute distributions of training data with that of the target domain. We evaluated our methodology through a series of experiments exploring scenarios where we have some labels from the target domain or when we estimate the target domain distribution using pseudo labels predicted by SER models. The results demonstrated that our approach significantly enhances the prediction of the emotional attributes arousal, and valence, especially when the training data is carefully curated to match the emotional landscape of the target domain. The proposed iterative refinement process further underlines the importance of continual adaptation and fine-tuning in achieving robust SER models. Future work could explore integrating more sophisticated techniques for estimating emotional distributions and apply this approach to other modalities and languages, broadening the applicability and effectiveness of the method in real-world scenarios.

REFERENCES

[1] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford University Press, Nov. 2013, pp. 110–127.

[2] J. Acosta, "Using emotion to gain rapport in a spoken dialog system," Ph.D. dissertation, University of Texas at El Paso, El Paso, TX, USA, December 2009.

[3] C.-C. Lee, K. Sridhar, J.-L. Li, W.-C.Lin, B.-H. Su, and C. Busso, "Deep representation learning for affective speech signal analysis and processing: Preventing unwanted signal disparities," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 22–38, Nov. 2021.

[4] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*. Calgary, AB, Canada: IEEE, April 2018, pp. 5084–5088.

[5] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1215–1227, April-June 2023.

[6] Y. Lei and H. Cao, "Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels," *IEEE Transactions on Affective Computing*, vol. Early Access, 2023.

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

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

[9] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 585–589, May 2016.

[10] 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.

[11] 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.

[12] 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.

[13] A. Reddy Naini, M. 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.

[14] 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.

[15] P. Mote, B. Sisman, and C. Busso, "Unsupervised domain adaptation for speech emotion recognition using K-Nearest neighbors voice conversion," in *Interspeech 2024*, Kos Island, Greece, September 2024, pp. 1045–1049.

[16] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *ArXiv e-prints (arXiv:1704.04861)*, pp. 1–9, April 2017.

[17] A. Reddy Naini, M. Kohler, E. Richerson, D. Robinson, and C. Busso, "Generalization of self-supervised learning-based representations for cross-domain speech emotion recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024)*, Seoul, Republic of Korea, April 2024.

[18] W. Han, T. Jiang, Y. Li, B. Schuller, and H. Ruan, "Ordinal learning for emotion recognition in customer service calls," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, Barcelona, Spain, May 2020, pp. 6494–6498.

[19] H. Martinez, G. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 314–326, July-September 2014.

[20] 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)*, vol. 33, Virtual, Dec. 2020, pp. 12 449–12 460.

[21] W.-N. Hsu *et al.*, "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.

[22] L. Goncalves, A. Salman, A. Reddy Naini, L. Moro-Velazquez, T. Thebaud, P. Garcia, N. Dehak, B. Sisman, and C. Busso, "Odyssey 2024 - speech emotion recognition challenge: Dataset, baseline framework, and results," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, Quebec, Canada, June 2024, pp. 247–254.

[23] J. Wagner *et al.*, "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. 10 745–10 759, September 2023.

[24] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using Wav2vec 2.0 embeddings," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 3400–3404.

[25] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5205–5209.

[26] ——, "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] 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.

[28] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Shanghai, China, April 2013, pp. 1–8.

[29] A. Reddy Naini, L. Goncalves, M. Kohler, D. Robinson, E. Richerson, and C. Busso, "WHiSER: White House Tapes speech emotion recognition corpus," in *Interspeech 2024*, Kos Island, Greece, September 2024.

[30] "Oval 741-2; june 23, 1972; white house tapes; richard nixon presidential library and museum, yorba linda, california." 1972. [Online]. Available: https://www.nixonlibrary.gov/index.php/white-house-tapes

[31] 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.

[32] T. Wolf *et al.*, "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv e-prints (arXiv:1910.03771v5)*, pp. 1–8, October 2019.

[33] D. 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.