Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Learning More with Less: Self-Supervised Approaches for Low-Resource Speech Emotion Recognition

Ziwei Gong[1], *Pengyuan Shi*[1,*], *Kaan Donbekci*[1,*], *Lin Ai*[1], *Run Chen*[1], *David Sasu*[2], *Zehui Wu*[1], *Julia Hirschberg*[1]

[1]Columbia University, USA; [2]IT University of Copenhagen, Denmark

{zg2272,ps3391,kd2939}@columbia.edu,{lin.ai,runchen}@cs.columbia.edu, dasa@itu.dk, zw2804@columbia.edu, julia@cs.columbia.edu

Abstract

Speech Emotion Recognition (SER) has seen significant progress with deep learning, yet remains challenging for Low-Resource Languages (LRLs) due to the scarcity of annotated data. In this work, we explore unsupervised learning to improve SER in low-resource settings. Specifically, we investigate contrastive learning (CL) and Bootstrap Your Own Latent (BYOL) as self-supervised approaches to enhance cross-lingual generalization. Our methods achieve notable F1 score improvements of 10.6% in Urdu, 15.2% in German, and 13.9% in Bangla, demonstrating their effectiveness in LRLs. Additionally, we analyze model behavior to provide insights on key factors influencing performance across languages, and also highlighting challenges in low-resource SER. This work provides a foundation for developing more inclusive, explainable, and robust emotion recognition systems for underrepresented languages.

Index Terms: Speech emotion recognition, Transfer learning, Low-resources languages, Multilinguality

1. Introduction

Speech Emotion Recognition (SER) is a fundamental task in speech processing with applications in human-computer interaction, affective computing, and mental health monitoring. The goal of SER is to automatically infer emotional states from speech signals, enabling more natural and adaptive interactions between humans and machines. Early approaches to SER relied on statistical learning methods, including Support Vector Machines (SVMs), k-Nearest Neighbors (k-NNs), Gaussian Mixture Models (GMMs), and Hidden Markov Models (HMMs) [1, 2, 3, 4]. With advances in deep learning, more recent SER systems leverage architectures such as Deep Neural Networks (DNNs), Long Short-Term Memory (LSTMs), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs) [5, 6, 7, 8], achieving significant performance improvements by leveraging techniques from Automatic Speech Recognition (ASR) and Speech Understanding (SU) [9, 10].

Despite these advancements, SER models tend to perform well only in High-Resource Languages (HRLs) where large-scale annotated speech emotion datasets are available. Many Low-Resource Languages (LRLs) lack sufficient labeled data, leading to significant performance disparities across different linguistic and cultural contexts. To address this issue, prior work has explored techniques such as domain adaptation, data augmentation, and transfer learning [11, 12, 13, 14, 9]. While domain adaptation methods attempt to align feature distributions across languages, they often fail to guarantee consistent predictive performance. Data augmentation strategies, such as Generative Adversarial Networks (GANs), require large amounts of high-quality data for stable training, which is often unavailable for LRLs. This heavy dependence on labeled data limits the scalability of existing approaches and makes them less effective for low-resource scenarios.

Unsupervised learning has been proposed as an alternative to reduce reliance on labeled data, yet it remains a challenging avenue for SER. Contrastive Learning (CL) [15, 16] and self-supervised methods such as Bootstrap Your Own Latent (BYOL) [17, 18] have shown promise in representation learning for speech tasks without labeled supervision [19]. However, their effectiveness in SER is still under-explored, and challenges remain in ensuring robust feature extraction and maintaining meaningful cross-lingual generalization. Unlike tasks such as Automatic Speech Recognition (ASR), where phonetic structures provide relatively stable patterns, emotion expression is highly variable across speakers and cultures, making self-supervised learning particularly difficult in multilingual SER.

We explore the use of unsupervised learning approaches to improve speech emotion recognition in low-resource settings, providing deeper insights into cross-lingual emotion recognition. We proposal two unsupervised learning approaches, Contrastive Learning and Bootstrap Your Own Latent, and demonstrate their potential to enhance cross-lingual generalization.

Our findings show that these methods significantly improve performance in low-resource settings, achieving F1 score gains of 10.6% in Urdu, 15.2% in German, and 13.9% in Bangla. Also, we analyze model behavior through interpretability methods, including feature visualization, confusion matrix, and post-hoc explainability, identifying key factors that influence performance across languages. Beyond performance improvements, our work highlights persistent challenges in low-resource SER, such as gender bias, linguistic transfer limitations, and dataset imbalances, providing a foundation for future research in fair and effective multilingual speech emotion recognition.

2. Methodology

We choose Contrastive Learning (CL) and Bootstrap Your Own Latent (BYOL) for their effectiveness in learning robust, generalizable representations from unlabeled data, a key requirement for low-resource SER.

CL differentiates similar and dissimilar inputs using self-supervised similarity measures, relying on carefully selected negative samples to contrast with augmented positive pairs. CL is more intuitive as it explicitly separates representations but relies on high-quality negative sampling.

BYOL, in contrast, learns representations without negative samples. It maintains two networks—a target and an online network—to ensure training stability and encourage consistency

[*]These authors contributed equally.

Code released: github.com/lynneeai/Speech_Emotion

10.21437/Interspeech.2025-1006

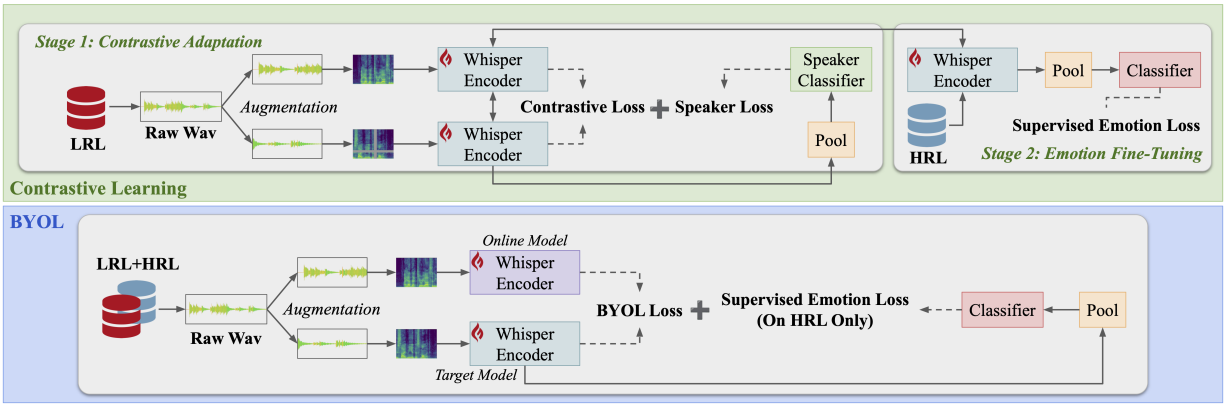

Figure 1: *Diagram of our proposed frameworks. Top: Contrastive Learning (CL). Bottom: Bootstrap Your Own Latent (BYOL). HRL and LRL signify High and Low Resource Languages.*

and invariance in learned features.

2.1. Data Processing and Augmentation

For CL, during contrastive speaker adaptation, each utterance has two views: a clean version and an augmented version, forming a positive pair since they share the same speaker label. For fine-tuning and baseline training, only the augmented views are used. We apply several data augmentations, including Gaussian noise injection (SNR sampled uniformly in 10–20 dB range), polarity inversion, gain manipulation, speed perturbation, spectral stretch, and SpecAugment [20].

For BYOL, we generate image pairs for training using time-frequency masking, mixUp, and random-resize-crop on mel-spectrograms. These augmentations, applied independently to each view, promote diverse transformations and prevent trivial memorization, fostering robust shared representations. For supervised training (2.3), only time-frequency masking and random-resize-crop are applied with reduced strength, generating three views per sample. Figure 2 shows an example of resulting spectrograms after augmentation. All feature extraction follows Whisper's default parameters.

2.2. Approach 1: Contrastive Learning (CL)

The proposed CL approach consists of two stages: speaker-contrastive adaptation and emotion fine-tuning (Figure 1).

Stage 1: Speaker-Contrastive Adaptation. We leverage speaker identity, as prior work shows a strong link between SER and speaker recognition (SR). Studies [21, 22] highlight correlations between speaker embeddings and emotional content, with pre-trained SR models like ECAPA-TDNN [23] outperforming other speech pre-training tasks. However, this connection remains unexplored in cross-lingual transfer learning. We hypothesize that enforcing embedding consistency for LRL speakers prevents the model from learning representations tied only to the HRL, whether lexical or paralinguistic. We propose a self-supervised contrastive objective on LRL datasets, forming positive/negative pairs based on speaker labels. Since speaker labels are easier to obtain than emotion labels and are commonly included in public datasets, this approach is widely applicable to other LRLs beyond our study. For the contrastive objective, we use Normalized Temperature-scaled Cross Entropy Loss [24]:

$$l_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_k \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \quad (1)$$

where τ is the temperature setting, (i, j) are indices of a positive pair (same speaker label), $\mathbf{z}$ are model embeddings, and k denotes indices with a different speaker label than i. We use cosine similarity as the similarity measure $\text{sim}(\cdot, \cdot)$, computing the loss across all positive pairs (i, j) and averaging it. To ensure valid positive/negative pairs, each batch includes 16 randomly sampled speakers from LRL datasets, with 4 utterances per speaker (without replacement). This prevents any speaker from dominating and promotes diverse negative pairs.

Stage 2: Emotion Fine-tuning. In the second stage, only English emotion-labeled HRL datasets are used. The pre-trained embedding model is fine-tuned with a cross-entropy objective for multi-class emotion classification. To address class imbalance, each batch includes an equal number of utterances per emotion class, preventing any single class from dominating.

2.3. Approach 2: Bootstrap Your Own Latent (BYOL)

BYOL [17] is a self-supervised representation learning method without negative pairs. Instead, it maintains two network branches, an *online* network and a *target* network, which are updated in a momentum fashion from *online* network. We integrate a supervised objective with a self-supervised BYOL-based objective into a single training stage, as illustrated in Figure 1. Namely, we optimize the following objectives:

- *Supervised objective (HRL data):* A cross-entropy (CE) loss is computed on HRL labeled emotion data, leveraging ground-truth emotion classes.
- *Self-supervised objective (HRL+LRL data):* Simultaneously, the BYOL objective – following the loss formulation and momentum-based weight update as described in [17] – is applied to all available utterances (both HRL and LRL), with labels ignored for HRL data in this component.

Hence, the overall objective for each mini-batch is the sum of the CE loss and BYOL loss weighted by a scalar factor λ:

$$\mathcal{L}_{\text{mixed}} = (1 - \lambda)\mathcal{L}_{\text{CE}}(\theta) + \lambda\,\mathcal{L}_{\text{BYOL}}(\theta, \xi), \quad (2)$$

where θ and ξ denote the parameters of the online and target networks, respectively. To address the issue of imbalanced data between HRL and LRL, we sample randomly from each data source at each training step. During inference, we use the trained target network with the classification head from the HRL supervised training component and evaluate it on LRL datasets.

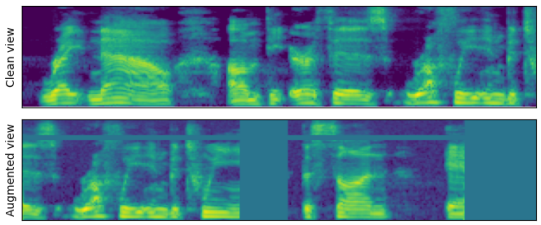

Figure 2: *Both views of the same utterance with minimal overlap, forming a positive pair in contrastive adaptation or BYOL.*

Dataset	Metric$_{(std)}$	Baseline	Contrastive	BYOL
URDU	Accuracy	$0.597_{(0.025)}$	$0.648_{(0.031)}$	$\mathbf{0.658}_{(0.052)}$
Urdu	Macro F1	$0.547_{(0.036)}$	$0.629_{(0.041)}$	$\mathbf{0.653}_{(0.060)}$
	UAR	$0.597_{(0.025)}$	$0.637_{(0.032)}$	$\mathbf{0.659}_{(0.053)}$
EMODB	Accuracy	$0.776_{(0.040)}$	$\mathbf{0.906}_{(0.017)}$	$0.794_{(0.033)}$
German	Macro F1	$0.750_{(0.059)}$	$\mathbf{0.901}_{(0.018)}$	$0.732_{(0.030)}$
	UAR	$0.758_{(0.050)}$	$\mathbf{0.896}_{(0.019)}$	$0.751_{(0.033)}$
SUBESCO	Accuracy	$0.616_{(0.015)}$	$\mathbf{0.721}_{(0.008)}$	$0.643_{(0.012)}$
Bangla	Macro F1	$0.572_{(0.016)}$	$\mathbf{0.711}_{(0.011)}$	$0.641_{(0.009)}$
	UAR	$0.616_{(0.015)}$	$\mathbf{0.721}_{(0.008)}$	$0.643_{(0.014)}$
RAVDESS	Accuracy	$0.574_{(0.020)}$	$0.580_{(0.018)}$	$\mathbf{0.582}_{(0.032)}$
English	Macro F1	$0.514_{(0.025)}$	$0.519_{(0.023)}$	$\mathbf{0.561}_{(0.029)}$
	UAR	$\mathbf{0.578}_{(0.034)}$	$0.570_{(0.026)}$	$0.572_{(0.030)}$

Table 1: *Performance$_{(std)}$ Comparison of Baseline, CL and BYOL models on different datasets along with their language*

3. Experiments

3.1. Datasets

We investigate the effectiveness of the above methods for low-resource SER on 3 LRLs, Urdu, German, and Bangla, and one HRL English as comparison.

For the HRL, We use MSP-Podcast [25] and IEMOCAP [26], both widely used for English SER. MSP-Podcast consists of speech from online podcast recordings, while IEMOCAP contains both acted and improvised dialogues from 10 actors.

For LRLs, we use EMO-DB (German) [27], URDU (Urdu) [28], and SUBESCO (Bangla) [29]. To ensure consistency across datasets, we keep only utterances from 0.5s to 12s with common labels ("happy," "sad," "neutral," "angry"). We merge the "excited" label with "happy," following [30].

For our CL approach, we train on the German, Urdu, and Bangla subsets of Common Voice [31], which provides speech samples from a larger number of unique speakers to ensure convergence to meaningful speaker embeddings without overfitting. We evaluate on all LRL datasets.

For our BYOL approach, we train on both LRLs and HRLs in the BYOL objective and HRLs in the supervision objective simultaneously. We employ a 5-fold cross-validation strategy, where the left-out target language LRL fold serves as the test set, while the left-out HRL fold is used for validation.

We follow the same train/validation/test splits as provided in Common Voice and MSP-Podcast v1.11. For the other datasets, we adopt a leave-one-session-out cross-validation strategy, resulting in 5 CV splits 5-fold CV is applied over sessions, and averaged. We use the validation split for early stopping, model checkpoint selection, and hyperparameter tuning.

3.2. Models Setup

CL. We build on Whisper's encoder [32], with three modifications: (1) Input length is reduced to 4 seconds by cropping the Positional Encoding layer, improving efficiency without performance loss. (2) Convolutional and positional encoding layers are frozen, following [33], for more stable training. (3) A classification head with two fully connected layers and Dropout is added, using Whisper embeddings (average pooled over time) to generate a fixed-length utterance representation.

BYOL. We use the Whisper's encoder described in Section 2.2, initialized from the HuggingFace checkpoint. Unlike the contrastive approach, the input length is capped at 3 seconds with cropped positional encoding. As in CL, convolutional and positional layers are frozen, and a two-layer classification head with Dropout is added to the time-averaged encoder outputs.

Baseline. We use the same Whisper-based encoder as for CL, except that we initialize the model with different weights to explicitly isolate the impact of the speaker-contrastive adaptation stage. The baseline is trained exclusively on English emotion-annotated datasets and evaluated on out-of-domain LRL datasets in a zero-shot manner.

3.3. Training Details

All models are implemented using PyTorch v2.5.1 and trained using a batch size of 64 with early stopping for computational efficiency, AdamW optimizer with a learning rate of 1e-5. For all experiments, a training epoch lasts for 100 batches. We use Whisper.small.en via transformers 4.39.3. We repeat all experiments five times and report the mean and standard deviation of our metrics. During BYOL training, we linearly scale down the BYOL loss factor λ from 0.8 to 0.2. All trainings are executed on 2 Nvidia A5500 and 2 Nvidia L40 GPUs.

4. Results and Discussion

For all three LRL datasets, models leveraging contrastive learning and BYOL as a pre-training stage outperformed the baseline. Table 1 reports Accuracy, Macro F1, and UAR for the three LRL datasets and a held-out HRL dataset. Results are averaged over five runs with the standard deviation in the footnote and with the best-performing models in bold.

The largest improvement in CL was observed on the German dataset, with nearly a 20% increase across all metrics. Notably, the Speaker-Contrastive model achieved 90% zero-shot accuracy on EmoDB without direct training, marking the best-known performance on this dataset in an out-of-domain setting. This may be due to the dataset's more distinguishable emotional content. However, we also hypothesize that German's linguistic proximity to English may contribute, as our models undergo final fine-tuning on English speech-annotated data, potentially biasing them toward languages closer to English.

In contrast, the BYOL approach demonstrated its strength in other scenarios. On the URDU dataset, BYOL achieved the highest accuracy (65.8%) and Macro F1 (65.3%) scores among the three methods, outperforming both the baseline and the contrastive model. Similarly, for SUBESCO, BYOL also surpassed the baseline across all metrics, though the contrastive approach still led overall improvements. These results suggest that BYOL effectively learns robust latent representations in low-resource settings, providing a viable alternative when data is scarce. Yet, it is important to note that BYOL exhibit larger standard deviations, particularly on URDU and EmoDB, probably due to the small size of each cross-validation fold (< 100 samples).

The smallest gains in both approaches were observed on the English RAVDESS dataset, held out from training alongside

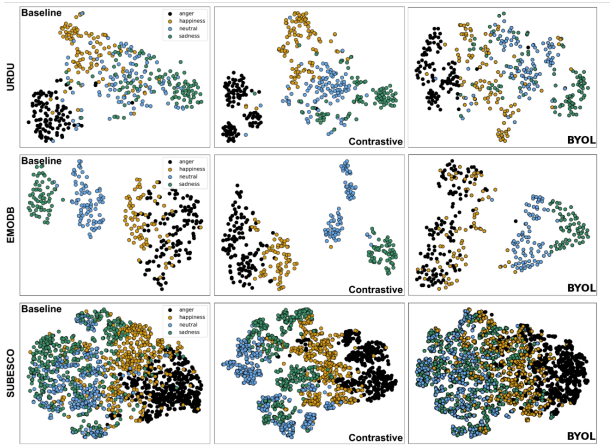

Figure 3: *T-SNE plots of learned embeddings on LRLs. Left: Baseline model after training. Middle: Speaker-contrastive model after training. Right: BYOL model after training.*

URDU, EmoDB, and SUBESCO. Performance on RAVDESS remained nearly identical to the baseline, showing no measurable improvement but also no deterioration. This suggests that despite our approach's modifications, the models retain their discriminatory ability in HRLs.

4.1. Gender Bias in Urdu SER

When we analyzed common failures of our models on the URDU dataset, we observed a strong gender bias in the CL model. The baseline model correctly predicts 74% of female and 54% of male speakers. While it achieved a 10% improvement, the CL model regressed on female speech (41%) while improving for males (73%). Nevertheless, the BYOL approach has 55% correctness on female speakers and 66% correctness on male speakers. We hypothesize that BYOL mitigates gender bias, as it does not use speaker driven positive/negative pair construction during training. This imbalance, obscured by the dataset's 86% male dominance, highlights a fairness issue for future work and improvement in CL approach.

No similar bias was observed in other datasets, where the speaker-contrastive model improved accuracy equally across genders. We suspect that this issue stems from random speaker sampling during contrastive adaptation and the male-skewed Urdu Common Voice dataset. We believe a robust and fair SER model should minimize gender imbalances, and future work will focus on mitigating this issue.

4.2. Interpreting Models Embeddings

Understanding model behavior, as seen in the hidden gender bias example, is challenging. One common interpretability method is T-SNE visualizations, which helps analyze learned embedding spaces. Figure 3 compares embeddings from the baseline, CL models, and BYOL models across the three LRL SER datasets. The CL and BYOL models exhibit more distinct emotion class boundaries, compared to the baseline model, supporting our empirical findings of improved performance. Comparing the T-SNE embeddings of CL and BYOL, CL has more dense clusters and creates more space between emotion classes. Notably, stronger clustering appears between "anger"/"happiness" and "neutral"/"sad". These pairs share similar arousal and dominance characteristics but differ in valence, which often depends on lexical content rather than speech patterns. In other words, distinguishing neutral from sad relies more on what is said than on how it is said.

Truth\Pred	anger	happiness	neutral	sadness
anger	+20	-16	-4	0
happiness	0	-1	-1	+1
neutral	0	-1	-1	+2
sadness	0	0	-26	+26

Table 2: *Confusion matrix for CL on EMODB, showing the mean performance difference between contrastive and baseline models. Higher diagonal values indicate better performance.*

This lexical dependence poses challenges in multilingual transfer learning, where valence cues may not transfer across languages. A word or sound conveying negative sentiment in one language may carry a different emotional meaning in another. For example, the sound "ach" denotes negative valence in English but is strongly positive in German. Our speaker-contrastive adaptation stage mitigates this issue by encouraging the encoder to prioritize speaker-distinguishing speech features rather than potentially misleading lexical cues. As shown in Table 2, performance gains mainly result from fewer "anger"/"happiness" and "neutral"/"sad" misclassifications.

4.3. Impact of Source Data in Self-Supervised Training

We examined the impact of different data sources on self-supervised training in BYOL and CL. Our findings show that CL benefits from a larger, diverse LRL dataset, even if unlabeled, while BYOL performs better when trained on the target LRL dataset, especially when its distribution aligns with non-labeled LRL samples used in training. Specifically, we compared BYOL training with (i) the Common Voice subset + HRL and (ii) the target LRL dataset + HRL. While CL performance improved with the Common Voice subset, BYOL performed better when trained directly on the target LRL dataset. Conversely, using the target LRL dataset instead of Common Voice in CL led to a performance drop. We hypothesize that this difference arises from how each method utilizes data. Since the Common Voice dataset is 10 times larger than the target LRL dataset, CL benefits from its greater speaker diversity, improving speaker-contrastive adaptation. In contrast, BYOL does not explicitly leverage speaker information, making the additional variety in the non-target LRL dataset less beneficial.

5. Conclusion and Limitations

In this work, we show that Contrastive Learning and and Bootstrap Your Own Latent enhance Speech Emotion Recognition performance in low-resource settings, with clearer emotion class separation in learned embeddings. Our contributions include *i)* effective improvement in performance (F1) in low-resource settings: 10.6% in Urdu, 15.2% in German, 13.9% in Bangla; *ii)* providing deeper insights of model behavior through model analysis and interpretation, also identifying challenges in LRL emotion recognition. Despite these advances, our limitations include a focus on a limited number of languages, and potential biases introduced by pre-training data.

Future work will explore linguistic distance in cross-lingual transfer, data augmentation to address gender imbalance, and modifications to CL and BYOL to reduce sensitivity to speaker-dependent features, improving embedding robustness. By addressing these challenges, we aim to advance fair and effective emotion recognition in diverse linguistic settings.

6. Acknowledgements

This research is supported in part by the Defense Advanced Research Projects Agency (DARPA), via the CCU Program contract HR001122C0034, and the National Science Foundation via ARNI Columbia 2025 Research Project. The views, opinions and/or findings expressed are those of the authors solely.

7. References

[1] M. Sheikhan, M. Bejani, and D. Gharavian, "Modular neural-svm scheme for speech emotion recognition using anova feature selection method," *Neural Computing and Applications*, vol. 23, pp. 215–227, 2013.

[2] M. J. Alam, Y. Attabi, P. Dumouchel, P. Kenny, and D. D. O'Shaughnessy, "Amplitude modulation features for emotion recognition from speech." in *INTERSPEECH*, 2013, pp. 2420–2424.

[3] P. R. Chaudhari and J. S. R. Alex, "Selection of features for emotion recognition from speech," *Indian Journal of Science and Technology*, vol. 9, no. 39, pp. 1–5, 2016.

[4] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 2. Ieee, 2003, pp. II–1.

[5] S. Yildirim, Y. Kaya, and F. Kılıç, "A modified feature selection method based on metaheuristic algorithms for speech emotion recognition," *Applied Acoustics*, vol. 173, p. 107721, 2021.

[6] Z.-T. Liu, A. Rehman, M. Wu, W.-H. Cao, and M. Hao, "Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence," *Information Sciences*, vol. 563, pp. 309–325, 2021.

[7] S. Langari, H. Marvi, and M. Zahedi, "Efficient speech emotion recognition using modified feature extraction," *Informatics in Medicine Unlocked*, vol. 20, p. 100424, 2020.

[8] M. Jain, S. Narayan, P. Balaji, A. Bhowmick, R. K. Muthu *et al.*, "Speech emotion recognition using support vector machine," *arXiv preprint arXiv:2002.07590*, 2020.

[9] H.-C. Lin, Y.-C. Lin, H.-C. Chou, and H.-y. Lee, "Improving speech emotion recognition in under-resourced languages via speech-to-speech translation with bootstrapping data selection," *arXiv preprint arXiv:2409.10985*, 2024.

[10] Z. Wu, Z. Gong, L. Ai, P. Shi, K. Donbekci, and J. Hirschberg, "Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances," 2024. [Online]. Available: https://arxiv.org/abs/2407.21315

[11] V. Hozjan and Z. Kačič, "Context-independent multilingual emotion recognition from speech signals," *International journal of speech technology*, vol. 6, pp. 311–320, 2003.

[12] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1458–1468, 2013.

[13] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu, and Y. Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Communication*, vol. 83, pp. 34–41, 2016.

[14] Y. Zhao, J. Wang, Y. Zong, W. Zheng, H. Lian, and L. Zhao, "Deep implicit distribution alignment networks for cross-corpus speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] M. Li, B. Yang, J. Levy, A. Stolcke, V. Rozgic, S. Matsoukas, C. Papayiannis, D. Bone, and C. Wang, "Contrastive unsupervised learning for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6329–6333.

[16] V. S. Alaparthi, T. R. Pasam, D. A. Inagandla, J. Prakash, and P. K. Singh, "Scser: Supervised contrastive learning for speech emotion recognition using transformers," in *15th international conference on human system interaction*. IEEE, 2022, pp. 1–7.

[17] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[18] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Byol for audio: Self-supervised learning for general-purpose audio representation," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[20] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv [eess.AS]*, Apr. 2019.

[21] I. R. Ulgen, Z. Du, C. Busso, and B. Sisman, "Revealing emotional clusters in speaker embeddings: A contrastive learning strategy for speech emotion recognition," *arXiv [eess.AS]*, Jan. 2024.

[22] O. C. Phukan, A. B. Buduru, and R. Sharma, "A comparative study of pre-trained speech and audio embeddings for speech emotion recognition," *arXiv [eess.AS]*, Apr. 2023.

[23] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," *arXiv [eess.AS]*, May 2020.

[24] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," *Neural Inf Process Syst*, vol. 29, pp. 1849–1857, Dec. 2016.

[25] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," https://ieeexplore.ieee.org/document/8003425, accessed: 2024-2-5.

[26] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[27] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Interspeech 2005*, vol. 5. ISCA: ISCA, Sep. 2005, pp. 1517–1520.

[28] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross lingual speech emotion recognition: Urdu vs. western languages," *arXiv [cs.CL]*, Dec. 2018.

[29] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "SUST bangla emotional speech corpus (SUBESCO): An audio-only emotional speech corpus for bangla," *PLoS One*, vol. 16, no. 4, p. e0250173, Apr. 2021.

[30] Z. Gong, M. Yao, X. Hu, X. Zhu, and J. Hirschberg, "A mapping on current classifying categories of emotions used in multimodal models for emotion recognition," in *LAW-XVIII*. ACL, Mar. 2024, pp. 19–28.

[31] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv [cs.CL]*, Dec. 2019.

[32] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," Technical report, OpenAI, 2022. URL https://cdn.openai. com/papers/whisper. pdf, Tech. Rep., 2022.

[33] M. Osman, T. Nadeem, and G. Khoriba, "Towards generalizable SER: Soft labeling and data augmentation for modeling temporal emotion shifts in large-scale multilingual speech," *arXiv [cs.CL]*, Nov. 2023.