

EMBRACE: Emotion-Based Representation Across Cultures & Expressions

Pravin Mote

Language Technologies Institute,
Carnegie Mellon University, Pittsburgh,
The University of Texas at Dallas,
USA
pmote@andrew.cmu.edu

Carlos Busso

Language Technologies Institute,
Carnegie Mellon University, Pittsburgh,
USA
busso@cmu.edu

Abstract

Building *speech emotion recognition* (SER) models that generalize across languages remains challenging due to limited annotations, linguistic diversity and heterogeneous label spaces. Current multilingual SER approaches either rely on fully shared models or restrict training to a small set of overlapping emotions that restrict diversity in expressiveness. We propose EMBRACE, a multilingual speech representation for language-agnostic SER. Drawing inspiration from partial-label learning paradigms in computer vision, EMBRACE adopts a unified superset emotion modeling strategy. The approach enables training on mixtures of datasets with non-overlapping emotion categories by selectively masking absent categories during optimization. To balance this complex multi-domain training, the model leverages a knowledge distillation based progressive training procedure that stabilizes multilingual optimization and mitigates catastrophic forgetting. Experiments across seven languages show that EMBRACE improves cross-corpus performance by 31.8% and in-corpus by 15.2% over strong emotional speech representations, confirming its ability to learn discriminative and transferable affective representations.

1 Introduction

Speech emotion recognition (SER) aims to automatically identify human affective states from speech (Leem et al., 2022, 2024; Lin and Busso, 2023; Wagner et al., 2023; Busso et al., 2013) and plays an important role in applications such as human-computer interaction (Cowie et al., 2001), assistive technologies (Calvo, 2015), and affect-aware conversational systems (Tu et al., 2022). In recent years, self-supervised speech models have substantially improved SER performance by providing strong acoustic representations (Wagner et al., 2023; Goncalves et al., 2024b; Reddy Naini et al., 2025). However, most SER systems are still

developed and evaluated in monolingual or single-corpus settings, where labeled data is available. In practical scenarios, emotional speech data is scarce for many languages, making it difficult to train reliable supervised models.

Multilingual SER introduces additional challenges beyond data scarcity. Emotional expression varies across languages (Feraru et al., 2015) due to phonetic, prosodic, and cultural norms in the externalization of emotions, which limits direct cross-language transfer. Moreover, existing multilingual and cross-corpus SER studies typically rely on a shared emotion taxonomy across datasets, which either reduces labels to a small overlapping subset (e.g., happiness, sadness, anger, neutral) (Mote et al., 2025b,a) or defines a new common label space such as high-low arousal or valence that can be applied uniformly across all datasets (Amiriparian et al., 2024). Aggregation of emotional labels is problematic as they oversimplify the perception of emotions (e.g., dichotomizing classes into arbitrary groups (Mariooryad and Busso, 2017), and merging dissimilar emotional classes). This simplification eliminates dataset-specific emotions and prevents models from leveraging the full diversity of available annotations. As a result, current approaches lack a mechanism to jointly learn from datasets with heterogeneous and non-overlapping emotion label spaces.

We propose EMBRACE¹, a multilingual SER framework designed to learn a generalized, language-agnostic affective representation while supporting language- and dataset-specific specialization. EMBRACE employs a *shared multilingual encoder* (SME) backbone and leverages progressive learning to train lightweight language-specific transformer blocks that act as adapters. To address heterogeneous label spaces across languages, EMBRACE incorporates an 11-class unified superset

¹<https://github.com/Lab-MSP/EMBRACE>

head that models the union of all emotion categories across datasets. This architectural choice is inspired by the multi-dataset, unified taxonomy paradigm in the computer vision tasks introduced by Zhou et al. (2022). To enable training with missing labels, EMBRACE uses masked classification and *exponential moving average* (EMA)-guided knowledge distillation, allowing each dataset to supervise only its available emotion categories without affecting the absent ones. A progressive training strategy with stage-wise freezing and consolidation further stabilizes multilingual optimization.

We evaluate EMBRACE against four strong *self-supervised learning* (SSL)-based baselines, including pretrained and fine-tuned wavLM (Chen et al., 2022), Emotion2Vec (Ma et al., 2024b) and Ex-HuBERT (Amiriparian et al., 2024). We consider two complementary settings: cross-corpus evaluation and in-corpus supervised classification. EMBRACE achieves an average relative improvement of 31.8% (cross-corpus) and 15.2% (in-corpus). The results demonstrate the effectiveness of learning a centralized multilingual affective representation and modeling heterogeneous emotion taxonomies within a single framework. Our contributions include (1) a generalized, SME backbone with language-specific adapters, (2) a mask based categorical loss, to support heterogeneous emotion label space across datasets, (3) a progressive training scheme for stable multilingual learning, and (4) a unified data-pooling strategy that integrates highly diverse speech corpora in their native form, completely bypassing the need to alter original labels, restrict data to overlapping subsets, or collapse nuanced emotional categories into oversimplified dimensions or classes.

2 Related Work

Several studies have modeled cross-lingual SER by targeting specific languages, such as knowledge transfer between English and Mandarin (Upadhyay et al., 2025a,b), by building adaptation strategies, such as low rank adaptation (Goncalves et al., 2024a), or by sharing discrete feature space adaptation (Mote et al., 2025b). Early work by Hozjan and Kačič (2003) explored context-independent SER across multiple languages using shared acoustic representations and neural classifiers. Lee (2019) showed that shared representations and regularization strategies, such as multi-task learning and normalization, can improve generalization across heterogeneous languages. Zehra et al. (2021) ex-

plored cross-corpus SER using ensemble learning, where multiple classifiers are combined to improve robustness across languages. More recent benchmarking efforts, such as EmoBox (Ma et al., 2024a), further highlight substantial performance variability when models are evaluated across mismatched corpora and languages.

Our approach aims to solve this challenge by learning speech representations that generalize across cross-lingual settings, complementing recent studies that investigate affective representations for multilingual emotion recognition. Sharma (2022) leveraged self-supervised representations such as wav2vec 2.0 with multi-task learning to share cross-lingual acoustic knowledge. Emotion2vec (Ma et al., 2024b) introduced emotion-aware pre-training with utterance and frame-level objectives to learn universal affective features. Ex-HuBERT (Amiriparian et al., 2024) extended HuBERT and fine-tuned it on EmoSet++, a set of emotional databases, to improve cross-lingual SER performance on unseen datasets.

In practice, previous approaches in SER handle heterogeneous emotion annotations by restricting training to shared categories, redefining emotions into coarse spaces such as high/low arousal or valence, or mixing samples across corpora with dataset-specific supervision. While effective for benchmarking and representation learning, these strategies limit a model’s ability to learn from the full diversity of emotion categories across datasets. To overcome similar constraints in other domain, studies have proposed partial-label learning frameworks that selectively manage missing or non-congruent annotations. For instance, Durand et al. (2019) introduced a standardized partial binary cross-entropy loss function that scales optimization based on the fraction of missing elements, allowing a model to train effectively even when only a small subset of labels is provided per image. Expanding on this concept for structural data, Zhou et al. (2022) developed a unified detector across multiple computer vision datasets by routing data through dataset-specific loss functions that dynamically mask out unseen classes from the shared taxonomy. Similarly, in the medical imaging domain, Yan et al. (2021) utilized a multi-task network with multi-expert detection heads to handle heterogeneous and partial clinical labels, ensuring missing lesion types do not introduce confusing or destructive gradients during universal training. Inspired by these cross-domain advancements, our framework

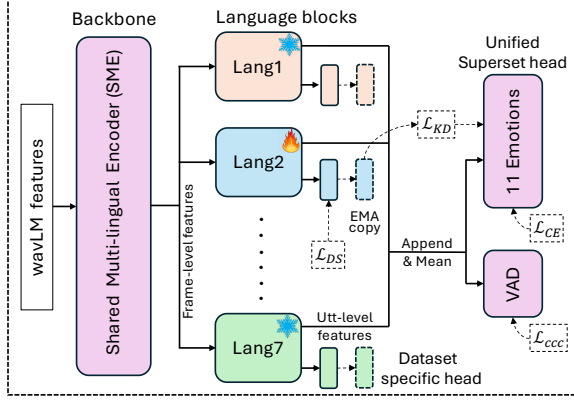


Figure 1: Architecture of EMBRACE.

translates the principles of partial-label masking into cross-lingual speech emotion architectures.

As multilingual SER models are trained sequentially across languages and corpora, they are susceptible to catastrophic forgetting, where adaptation to new languages degrades previously learned representations, motivating progressive learning strategies that explicitly preserve prior knowledge.

Progressive Neural Networks (Rusu et al., 2022) introduce a continual learning framework where each new task is learned by adding a new network column with lateral connections to previously trained columns, enabling knowledge transfer without catastrophic forgetting. This architecture demonstrates how task-specific parameters can complement shared representations, informing designs that balance shared learning and task specialization. Building on these ideas, our approach combines progressive learning with explicit handling of heterogeneous emotion annotations, enabling scalable multilingual SER that preserves previously learned knowledge while jointly learning from diverse emotion categories across languages.

3 Methodology

Figure 1 shows EMBRACE, which consists of four key components: (i) an SME backbone that provides a common representation across all languages, (ii) lightweight language-specific blocks that adapt this representation to individual languages, (iii) a unified superset head that models the union of all emotion categories across datasets, and (iv) a progressive training algorithm that combines consolidation strategy, teacher–student distillation, and a mask-based categorical loss for heterogeneous emotion labels across datasets. Together, they enable EMBRACE to jointly learn from mul-

tiple languages without restricting emotion labels to their intersection, artificially clustering classes, or suffering from catastrophic forgetting.

3.1 Shared Multilingual Encoder (SME) Backbone

The SME backbone is responsible for extracting language-agnostic affective representations from speech. This encoder should be expressive to capture the variability of emotional speech across languages. We adopt a pretrained wavLM-large-robust (Chen et al., 2022) model as the backbone, leveraging its strong self-supervised representations learned from large-scale multilingual speech data. Here, the raw audio feature extractor of wavLM is not used, as the encoder operates directly on frame-level SSL features. Given an input sequence of frame-level features, the SME backbone produces contextualized representations that capture temporal dependencies and affective cues in a shared latent space. These representations are shared across all languages and datasets, forming a stable foundation for multilingual learning and enabling a generalized affective representation transferable across languages.

3.2 Language-Specific Blocks

To complement the SME backbone, EMBRACE introduces language-specific blocks that act as lightweight adapters to model language-dependent characteristics of emotional speech. Each language is associated with a dedicated transformer block composed of a small stack of encoder layers. These blocks take the contextualized representations produced by the SME backbone as input and refine them to capture language-specific acoustic patterns, prosodic variations, and dataset-related characteristics.

The output of each transformer block is aggregated over time using mean pooling to obtain an utterance-level embedding. This pooled representation is fed into two parallel heads. The first head is the unified superset head (marked by losses $\mathcal{L}_{CE}$ and $\mathcal{L}_{ccc}$ in Figure 1), which receives inputs from all language-specific blocks by concatenating their outputs followed by mean pooling and predicts the union of emotion categories across all datasets and valence, arousal and dominance (VAD) attributes. The second classification head is a dataset-specific head (marked by loss $\mathcal{L}_{DS}$ in Figure 1) that predicts only the emotion categories available in the corresponding dataset, as described in Table 1. These

Table 1: Multilingual SER datasets used in this work. Abbreviations: A (Anger), B (Boredom), C (Contempt), D (Disgust), E (Enthusiasm), F (Fear), H (Happiness), N (Neutral), S (Sadness), U (sUrprise), R (saRcastic).

Lang.	Dataset	Domain	#Utt.	Emotion Set
English	MSP-Podcast (Busso et al., 2026)	Source	267,913	A,C,D,F,H,N,S,U
	MSP-IMPROV (Busso et al., 2017)	Target	8,438	A,H,N,S
Russian	Dusha (Kondratenko et al., 2022)	Source	184,633	A,H,N,S
	RESO (Lubenets et al., 2022)	Target	1,116	A,D,E,F,H,N,S
Mandarin	BIIC-Podcast (Upadhyay et al., 2023)	Source	59,993	A,C,D,F,H,N,S,U
	EmotionTalk (Sun et al., 2025)	Target	19,250	A,D,F,H,N,S,U
Italian	Emozionalmente (Fabio, 2023)	Source	6,902	A,D,F,H,N,S,U
	EMOVO (Costantini et al., 2014)	Target	588	A,D,F,H,N,S,U
German	Syntact (Burkhardt et al., 2022)	Source	6,001	A,B,F,H,N,S
	EmoDB (Burkhardt et al., 2005)	Target	816	A,B,D,F,H,N,S
Hindi	HindiSER (Shinde et al., 2023)	Source	3,200	A,D,F,H,N,S,R,U
	IESC (Singh and Goel, 2023)	Target	600	A,F,H,N,S
French	Café (Gournay et al., 2018)	Source	936	A,D,F,H,N,S,U
	Oréau (Leila et al., 2020)	Target	502	A,D,F,H,N,S,U

dataset-specific heads are trained using standard cross-entropy loss and provide strong supervision in its own native label space. They are not used at inference time, but play an important role in the learning process.

In addition to the online dataset-specific heads, EMBRACE maintains an *exponential moving average* (EMA) copy of each dataset-specific head, which serves as a stable teacher during training. The EMA heads evolve smoothly over time and provide reliable soft targets for knowledge distillation into the unified superset student head. Together, the language-specific blocks and dataset-specific heads enable localized adaptation and dataset-aware supervision while keeping EMBRACE stable across languages.

3.3 Unified Superset Head

At the core of EMBRACE lies a unified superset head, which enables joint modeling of heterogeneous emotion label spaces across datasets. Instead of restricting training to a small set of overlapping emotions, this head predicts over a global superset of emotional categories defined as the union of all emotions present in the training datasets. For a given utterance embedding produced by a language-specific block, the unified superset head outputs logits over all K superset emotions (with $K=11$ here).

Since each dataset contains only a subset of these emotions, direct supervision over the entire superset would introduce incorrect gradients for missing categories. To address this problem, we employ a mask-based categorical loss for het-

erogeneous emotion labels across datasets. For a dataset d with available emotion indices $C_d \subseteq \{A, B, C, D, E, F, H, N, S, U, R\}$ (abbreviations explained in Table 1), we compute the loss only over the corresponding subset of logits as shown in Eq. 1,

$$\mathcal{L}_{CE}^d = -\log \frac{\exp(z_y)}{\sum_{k \in C_d} \exp(z_k)}, y \in C_d \quad (1)$$

where, z_k denotes the logit of the k -th emotion in the unified superset head. Emotions not present in C_d are excluded from the softmax and receive no gradient updates, ensuring that missing labels do not negatively affect the training.

A teacher-student distillation mechanism between the EMA version of the corresponding dataset-specific teacher head further provides dataset-specific supervision for the unified superset student head. The student head is trained to match the teacher head’s distributions on the same subset C_d using a KL-divergence loss, as shown in Eq. 2,

$$\mathcal{L}_{KD}^d = KL(\sigma(t_d) \parallel \sigma(z_{C_d})), \quad (2)$$

where, $\sigma(\cdot)$ denotes the softmax function, t_d are the teacher logits from the dataset specific head, z_{C_d} are the student logits produced by the unified superset head, sliced to include only the emotion categories present in the dataset d . This distillation encourages the unified superset head to inherit dataset-native class structure while maintaining a shared global representation.

In addition to categorical supervision, the unified superset head incorporates *valence*, *arousal*, and *dominance* (VAD) prediction as an auxiliary multi-task objective to encourage the learning of more robust and affect-aware representations. Using emotional attributes as secondary task has been shown to increase robustness and generalization in SER (Xia and Liu, 2017; Reddy Naini and Busso, 2025). When VAD annotations are available for a dataset, the model jointly optimizes these attributes using a *concordance correlation coefficient* (ccc) loss ($\mathcal{L}_{ccc}=1-\text{ccc}$), which promotes consistency between the predicted emotional dimensions and ground-truth. This regression task is used only during training to regularize the shared representations and improve generalization across languages.

3.4 Training Strategy

Progressive Freezing and Unfreezing: EMBRACE training proceeds in multiple stages that explicitly control which components are updated at each phase, preventing catastrophic forgetting, and reducing interference across languages. Training begins with a warm-up stage in which the SME backbone and the unified superset head are frozen, while one language-specific block and its corresponding dataset-specific head are trained at a time. This strategy allows each block to adapt to its language and dataset characteristics while providing stability to the SME backbone by preventing noisy gradients from language-specific blocks. After this warm-up, training alternates between (i) language-wise adaptation and (ii) consolidation.

(i) In language-wise adaptation, the SME backbone, the active language block, and the unified superset head are jointly updated, while all other language blocks remain frozen. Optimization is restricted to a single dataset at a time, processing its batches sequentially so that the corresponding adapter can fully absorb domain-specific traits before the system switches to another language. As illustrated in Figure 1, only the specific block associated with the currently active dataset (such as Lang2) is set as trainable, while the remaining language blocks are securely locked.

(ii) In consolidation, the SME backbone and all language-specific blocks are frozen, and only the unified superset head is trained using data from all languages. In this step, batches are sampled in a round-robin manner across all datasets, with one batch from each dataset processed in turn to promote balanced global alignment in the unified su-

perset head. This alternating schedule encourages localized adaptation followed by global alignment.

Balancing Dataset and Category Imbalance: To account for differences in dataset sizes and class distributions, we apply weighting schemes at two levels. First, dataset-level weights are used to balance contributions from datasets of different sizes, preventing larger corpora from dominating. Second, category-level weights are applied within each dataset to mitigate class imbalance among emotion categories. These weights are incorporated directly into the total loss terms, allowing the model to learn balanced decision boundaries despite skewed data distributions.

Overall Objective: The total training objective for a dataset d is shown in Eq. 3,

$$\mathcal{L}_{\text{total}}^d = \lambda_{ds}\mathcal{L}_{\text{DS}}^d + \lambda_{ce}\mathcal{L}_{\text{CE}}^d + \lambda_{ccc}\mathcal{L}_{\text{ccc}}^d + \lambda_{kd}\mathcal{L}_{\text{KD}}^d, \quad (3)$$

where $\mathcal{L}_{\text{DS}}^d$ is the cross-entropy loss for the dataset-specific head, $\mathcal{L}_{\text{CE}}^d$ is the masked loss applied to the unified superset head as shown in Eq. 1, $\mathcal{L}_{\text{ccc}}^d$ is the CCC loss calculated for the VAD attributes, and $\mathcal{L}_{\text{KD}}^d$ is the KL-divergence-based distillation loss from the EMA teacher as shown in Eq. 2. The coefficients λ_{ds} , λ_{ce} , λ_{ccc} , and λ_{kd} are the hyperparameters that control the relative contribution of each term.

The proposed strategy that combines progressive freezing, mask-based supervision, and teacher-student distillation enables EMBRACE to learn multiple languages with heterogeneous label spaces in a stable manner, while maintaining a single shared classifier at inference time.

4 Experimental Settings

4.1 Multilingual Datasets

We evaluate EMBRACE in a multilingual setting spanning seven languages (English, Russian, Mandarin, Italian, German, Hindi, and French), each paired with one labeled source dataset and one unseen target dataset. Table 1 summarizes the dataset details, where the first-row for each language is a source domain and the second-row is the corresponding target domain. Importantly, the emotion label space differs across datasets, reflecting the heterogeneity commonly encountered in real-world SER benchmarks. With the exception of Section 5.2, the model is trained exclusively on the source domains and evaluated on the target domains, simulating a realistic cross-corpus scenario in which

Target Dataset	Language	wavLM	wavLM-ft	emo2vec	ExHuBERT	EMBRACE
MSP-IMPROV	English	0.174 $\pm$ 0.03	0.193 $\pm$ 0.02	0.162 $\pm$ 0.02	0.191 $\pm$ 0.02	0.233 $\pm$ 0.03*
RESERVED	Russian	0.077 $\pm$ 0.04	0.134 $\pm$ 0.06	0.062 $\pm$ 0.05	0.041 $\pm$ 0.01	0.265 $\pm$ 0.07*
EmotionTalk	Mandarin	0.162 $\pm$ 0.03	0.187 $\pm$ 0.02	0.129 $\pm$ 0.01	0.158 $\pm$ 0.01	0.225 $\pm$ 0.02*
EMOVO	Italian	0.124 $\pm$ 0.11	0.157 $\pm$ 0.08	0.051 $\pm$ 0.04	0.090 $\pm$ 0.04	0.260 $\pm$ 0.11*
EmoDB	German	0.393 $\pm$ 0.08	0.446 $\pm$ 0.07	0.248 $\pm$ 0.06	0.480 $\pm$ 0.11	0.412 $\pm$ 0.09
IESC	Hindi	0.337 $\pm$ 0.20	0.361 $\pm$ 0.15	0.326 $\pm$ 0.11	0.341 $\pm$ 0.15	0.468 $\pm$ 0.16
Oréau	French	0.143 $\pm$ 0.04	0.164 $\pm$ 0.05	0.113 $\pm$ 0.07	0.153 $\pm$ 0.06	0.158 $\pm$ 0.04

Table 2: Cross-corpus macro-F1 scores across target datasets (Mean $\pm$ SD). Bold values indicate the best-performing model. The symbols * denote statistically significant ($p < 0.05$) improvements over the second-best model based on a two-tailed paired t -test across 5 folds.

labeled data is available only for the source domain, allowing us to assess the ability of EMBRACE to generalize across datasets and languages leveraging shared affective knowledge across languages.

4.2 Implementation Details

EMBRACE is trained on 1,024-dimensional frame-level SSL representations extracted from a pre-trained wavLM-large-robust model, downloaded from huggingface and fine-tuned on the MSP-Podcast corpus (Busso et al., 2026) for the SER task. The *shared multilingual encoder* (SME) backbone comprises the 24-layer Transformer encoder from wavLM ($d_{\text{model}} = 1,024$), operating directly on input frames without raw-audio waveform processing. Each language-specific adapter block is configured as a 4-layer Transformer encoder with $n_{\text{head}} = 8$, a hidden dimension of 1,024, a feed-forward dimension of $d_{\text{ff}} = 2,048$, and a dropout rate of $p = 0.10$ with GELU activations. The unified superset head consists of a 11-way categorical classifier, while each dataset-specific teacher head uses a linear projection matching its native subset of categories $|C_d|$.

Training is conducted using PyTorch on an NVIDIA A100 GPU (80GB) with a batch size of 16. We employ the AdamW optimizer with decoupled weight decay of 1×10^{-2} and momentum parameters $(\beta_1, \beta_2) = (0.9, 0.98)$. During language-wise adaptation phases, differentiated learning rates are applied across sub-modules: 1×10^{-4} for the active language adapter and its corresponding dataset-specific teacher head, and 1×10^{-5} for both the SME backbone and the unified superset classification head. During the interleaved consolidation phases, all language-specific blocks and the backbone are locked, and only the unified prediction head is optimized with a reduced learning rate of 5×10^{-6} . Gradients are scaled and clipped to a

maximum L_2 norm of 5.0 across all training phases to stabilize optimization across different corpora.

For the multi-task and distillation objectives in Eq. 3, the loss weighting coefficients are set to $\lambda_{\text{ds}} = 1.0$, $\lambda_{\text{ce}} = 1.0$, $\lambda_{\text{ccc}} = 1.0$, and $\lambda_{\text{kd}} = 0.3$. Soft-target knowledge distillation utilizes a temperature parameter of $\tau = 2.0$ with an *exponential moving average* (EMA) momentum of $m = 0.995$ for updating the teacher heads, activated after a 1-epoch distillation warmup. To address significant disparities in corpus sizes and label distributions, optimization employs a two-tiered reweighting strategy: (i) dataset-level scaling factors to normalize gradient magnitudes across large and small corpora, and (ii) inverse class-frequency cross-entropy loss weights computed per dataset to prevent majority-class collapse on highly imbalanced targets. All target evaluations across all comparative baselines follow a 5-fold cross-validation protocol with Macro-F1 reported as the primary evaluation metric.

5 Results and Analyses

5.1 Cross-Corpus Evaluations

We first evaluate EMBRACE in a cross-corpus setting, where EMBRACE and baselines are trained exclusively on the source datasets and tested on the unseen target datasets. This setup directly measures the ability of the learned representations to generalize across datasets and languages, which is a key challenge in multilingual SER. EMBRACE is compared against four strong SSL-based baselines, (i) pretrained wavLM (Chen et al., 2022), (ii) fine-tuned wavLM on SER task using the MSP-Podcast corpus, (iii) emotion2vec (Ma et al., 2024b) and (iv) ExHuBERT (Amiriparian et al., 2024). For baselines, speech features are extracted from the four SSL models and fed to train a lightweight classifier

with three fully connected layers and the unified superset head, using the same protocol as EMBRACE. In contrast, EMBRACE is trained end-to-end as described in Section 3.4, allowing the SME backbone and language-specific blocks to adapt. This consistent setup ensures a fair comparison between EMBRACE and baselines, where differences in performance arise from the representation learning strategy rather than variations in the training conditions. To estimate the consistency in the prediction, each target dataset is randomly partitioned into five subsets, with the final mean and standard deviation calculated across these independent evaluations.

The cross-corpus results shown in Table 2 demonstrate a clear advantage of EMBRACE over all baselines. EMBRACE achieves the best performance in five out of seven datasets, consistently outperforming strong SSL feature extractors. In the remaining datasets, EmoDB and Oréau, EMBRACE performs almost on par with the strongest baselines, showing only a marginal degradation of approximately 5%. Notably, EmoDB is included in ExHuBERT’s supervised training, which may partly explain this advantage. This small gap suggests that, even in cases where pretrained representations are already highly transferable, the proposed framework remains competitive. On average, EMBRACE yields a 31.8% relative improvement over the strongest baseline (wavLM-ft) across the target datasets. This gain indicates that the improvements are not limited to isolated languages but are consistent across diverse emotional label spaces and recording conditions. Notably, in cases such as Russian or German, where certain target emotion categories are absent in the source corpus, EMBRACE benefits from complementary supervision in other languages, enabling better generalization to previously unseen emotional classes.

5.2 In-Corpus Evaluations

In addition to cross-corpus transfer, we perform an in-corpus evaluation to assess the intrinsic quality of the learned representations. For each target dataset, features are extracted using EMBRACE (trained only on source domains) and the baselines. For EMBRACE, the features correspond to the representations from the SME backbone and language-specific blocks before the unified superset head. A classifier with three fully connected layers is then trained and evaluated independently on each target dataset using a five-fold cross-validation strategy, isolating representation quality from domain adap-

tation effects.

As shown in Table 3, EMBRACE achieves the best performance in four of the seven target languages, indicating that its multilingual training produces strong emotion-discriminative embeddings even under supervised evaluation. The fine-tuned wavLM slightly outperforms EMBRACE by an average margin of 4% on EmotionTalk and EMOVO, while ExHuBERT performs better on Oréau with a margin of 8%. Nevertheless, these performance gaps remain marginal, demonstrating that EMBRACE maintains a highly competitive baseline even in target domains where it does not achieve top-tier results. Crucially, ExHuBERT had an advantage in EmoDB by including it in its supervised training. EMBRACE outperforms this baseline, demonstrating the superior generalization capability of learned representations from our multilingual synthesis framework.

On average, EMBRACE improves performance by 20.5% over fine-tuned wavLM and 15.2% over ExHuBERT, suggesting that the SME backbone combined with language-specific adaptation yields more transferable and structured emotion representations. In contrast, emo2vec shows noticeable degradation, reflecting the limitation of training with reduced common emotion taxonomies (e.g., four-class settings) and reinforcing the importance of modeling the full diversity of emotion categories rather than collapsing them into a small subset.

6 Impact of Multilingual Training

To assess the impact of multilingual training on cross-corpus evaluations, we include an additional variant, EMBRACE-It (Table 4) as an ablation study. It is trained only on the Italian source dataset with identical architecture and training procedure. This variant isolates the role of linguistic diversity by removing cross-language exposure. EMBRACE trained on all seven languages significantly outperforms EMBRACE-It, achieving a 29.5% average improvement and better results on six of seven target datasets, with only a minor drop on Oréau, highlighting the benefit of multilingual diversity for robust and transferable affective representations.

7 Progressive Optimization Trajectory

To analyze the internal optimization mechanics of EMBRACE, we monitor its micro-level behavior during both sequential training intervals and global consolidation phases. Specifically, while the

Target Dataset	Language	wavLM	wavLM-ft	emo2vec	ExHuBERT	EMBRACE
MSP-IMPROV	English	0.154 $\pm$ 0.00	0.154 $\pm$ 0.00	0.154 $\pm$ 0.00	0.159 $\pm$ 0.01	0.162 $\pm$ 0.02
RES	Russian	0.143 $\pm$ 0.03	0.196 $\pm$ 0.04	0.062 $\pm$ 0.02	0.195 $\pm$ 0.03	0.253 $\pm$ 0.01*
EmotionTalk	Mandarin	0.315 $\pm$ 0.00	0.323 $\pm$ 0.01	0.094 $\pm$ 0.00	0.230 $\pm$ 0.00	0.316 $\pm$ 0.01
EMOVO	Italian	0.187 $\pm$ 0.03	0.301 $\pm$ 0.02[‡]	0.054 $\pm$ 0.01	0.203 $\pm$ 0.06	0.280 $\pm$ 0.05
EmoDB	German	0.238 $\pm$ 0.07	0.359 $\pm$ 0.07	0.054 $\pm$ 0.03	0.483 $\pm$ 0.04 [‡]	0.494 $\pm$ 0.03[‡]
IESC	Hindi	0.323 $\pm$ 0.08	0.439 $\pm$ 0.09	0.118 $\pm$ 0.04	0.546 $\pm$ 0.04	0.572 $\pm$ 0.03[‡]
Oréau	French	0.127 $\pm$ 0.04	0.174 $\pm$ 0.04	0.066 $\pm$ 0.01	0.286 $\pm$ 0.05[‡]	0.261 $\pm$ 0.04

Table 3: In-corpus macro-F1 scores across target datasets (Mean $\pm$ SD). Bold values denote the best-performing model. * indicates statistical significance ($p < 0.05$) over the second-best model; [‡] indicates statistical significance ($p < 0.05$) over the third-best baseline under a two-tailed paired t -test across 5 folds.

Target Dataset	Language	EMBRACE-It	EMBRACE
MSP-IMPROV	English	0.150	0.233
RES	Russian	0.107	0.265
EmotionTalk	Mandarin	0.108	0.225
EMOVO	Italian	0.238	0.260
EmoDB	German	0.313	0.412
IESC	Hindi	0.326	0.468
Oréau	French	0.164	0.158

Table 4: Ablation study evaluating the impact of multilingual training on cross-corpus macro-F1 scores

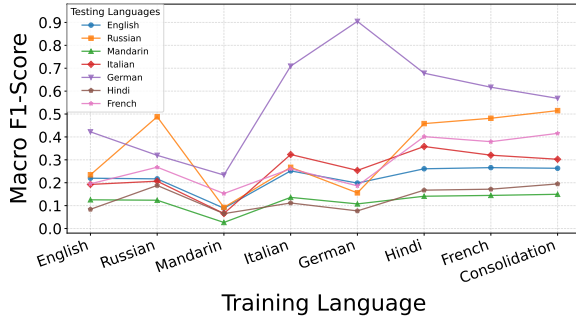


Figure 2: Progressive Language Training Trajectory

model trains sequentially on each individual source dataset, we track macro-F1 scores of the test set of each source dataset simultaneously.

7.1 Sequential Language-Wise Adaptation

The figure 2 presents the performance trajectory across a single sequential training cycle (captured during the final epoch). It clearly visualizes the tension between localized adaptation and cross-lingual gradient interference. As the model transitions sequentially from one language corpus to the next, a distinct localized pattern emerges. The macro-F1 score for the actively training language experiences a sharp spike when the model directly optimizes on its native data. However, this localized gain often comes at the expense of other inactive languages, which display noticeable drop or volatile fluctua-

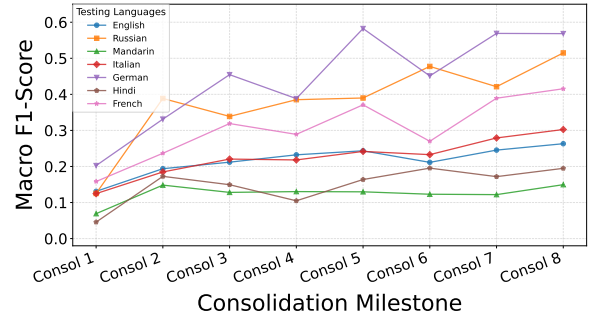


Figure 3: Macro-F1 score Along Consolidation Phases

tions during the same window. This fluctuating behavior validates our original hypothesis regarding multilingual SER that sequential training across highly heterogeneous label spaces induces catastrophic forgetting and gradient conflict. Crucially, as the sequence reaches the final "Consolidation" step at the end of the epoch, these divergent performance lines recover from their respective troughs and move upward, demonstrating how the consolidation reconciles these competing acoustic and heterogeneous label traits.

7.2 Cross-Epoch Consolidation Progression

The long-term efficacy of this balancing act is explicitly shown in Figure 3 by tracking the model's performance, strictly at the consolidation milestones across all eight training epochs. At the first consolidation milestone, the macro-F1 scores across all source languages start from a relatively low baseline, reflecting the initial chaos of forcing a shared backbone to process incompatible emotion spaces. However, from Consolidation 1 through Consolidation 8, the trajectory across all source languages exhibits a sustained, steady upward climb. Although minor localized fluctuations occur in individual epochs due to the lingering effects of the preceding sequential updates, overall the general trend points to steady improvement and long-term

stability. This progress across epochs proves that the alternating stage-wise schedule does not just prevent catastrophic forgetting; rather, it systematically uses the round-robin consolidation phase to transform fragmented, language-specific adapter knowledge into a unified, highly transferable global affective representation space.

8 Computational Complexity Analysis

EMBRACE decouples universal representation learning from language-specific specialization to balance modeling capacity and computational overhead. The architecture consists of three main components: (1) a SME backbone ($\sim 315\text{M}$ parameters based on wavLM-large), (2) dedicated language specific blocks ($N = 7$, with $\sim 12.6\text{M}$ parameters each), and (3) compact task prediction heads ($\sim 15.4\text{k}$ parameters for the unified superset emotion classification and dimensional VAD regression heads).

Parameter Footprint and Scaling: Each language specific block comprises an input linear projection followed by a stack of four Transformer encoder layers ($d_{\text{model}} = 1,024$, eight attention heads, and a feed-forward dimension of 2,048). As additional languages are added to the framework, the model size scales linearly: parameter growth is limited to $\mathcal{O}(N)$ from the lightweight language adapters, while the primary 315M-parameter acoustic backbone is shared without duplication.

Inference Complexity: For an input utterance of T acoustic frames with feature dimension $D = 1,024$, evaluating the backbone requires $\mathcal{O}(L_{\text{SME}} \cdot (TD^2 + T^2D))$ operations, where $L_{\text{SME}} = 24$ denotes the depth of the 24-layer wavLM encoder. Because our inference pipeline operates in a language-agnostic setting, representations are extracted across all $N = 7$ language specific blocks concurrently before temporal pooling and mean-aggregation across blocks. This strategy contributes an additional complexity of $\mathcal{O}(N \cdot L_{\text{col}} \cdot (TD^2 + T^2D))$, where $L_{\text{col}} = 4$ represents the 4-layer depth per language specific block. Consequently, overall inference computation scales linearly ($\mathcal{O}(N)$) with the number of supported languages.

Training Efficiency: Despite maintaining multiple language specific blocks, progressive training keeps the memory footprint bounded during optimization. In each language-wise adaptation step, only the active language adapter and the shared

backbone are updated, while the remaining $N - 1$ language specific blocks are frozen under gradient-free execution. This selective freezing avoids back-propagating through all N language specific blocks simultaneously, ensuring stable training on a single 80GB NVIDIA A100 GPU with a batch size of 16.

9 Conclusions

This paper introduced EMBRACE, a multilingual SER framework designed to learn a shared affective representation while accommodating heterogeneous emotion label spaces. EMBRACE enables joint learning across multiple languages without restricting supervision to overlapping emotion categories, by combining an SME backbone, language-specific blocks, and a unified superset head trained with a mask-based categorical loss for heterogeneous emotional labels across datasets. Progressive training with consolidation and EMA-based distillation stabilizes multilingual optimization.

Extensive experiments across seven languages demonstrate that EMBRACE consistently improves cross-corpus generalization, outperforming strong SSL-based baselines in five out of seven target datasets and achieving average gains of 31.8%. In the in-corpus evaluations, EMBRACE yields competitive and often superior performance with 15.2% gains over the best baselines, confirming the quality and transferability of the learned representations. Ablation analysis confirms that, multilingual training provides clear benefits over single-language models, highlighting the value of training on highly diverse emotional data. Finally, tracking the macro-F1 trajectories of the test set of source datasets demonstrates that, while sequential learning induces temporary localized performance trade-offs, the periodic consolidation phases successfully reconcile these conflicts to ensure steady, long-term improvement across all languages.

Overall, these findings suggest that unified superset modeling combined with progressive multilingual learning offers a practical and scalable direction for robust speech emotion recognition across languages and datasets.

Limitations

While our framework demonstrates a robust capacity for cross-lingual speech emotion recognition, it has a few distinct limitations. First, EMBRACE is dependent on precomputed features from an upstream wavLM model fine-tuned on the MSP-

Podcast corpus. Because our shared backbone operates directly on these frozen representations, any acoustic biases or emotional misalignments from the upstream encoder will naturally pass through to our system.

Second, the network introduces an inference overhead that scales with the number of language specific blocks. Since the unified superset head requires concatenating and pooling outputs from all language-specific adapters simultaneously, inference demands a full forward pass across every single language block, leading to a linear increase in parameter and memory usage as new languages are introduced.

Finally, our empirical evaluation is restricted to seven high-resource, well-structured corpora. The localized performance drops seen in specific target domains (such as Mandarin and French) highlight that highly unique recording environments and dataset-specific annotation boundaries still introduce an optimization bottleneck for generalized cross-lingual transfer.

Ethical Considerations

While EMBRACE aims to improve cross-lingual speech emotion recognition for helpful applications like accessible user interfaces, we recognize that this technology carries risks of unintended misuse. Systems that decode emotions from the human voice can unfortunately be exploited for intrusive purposes, such as unauthorized voice profiling, workplace surveillance, or manipulative targeted advertising. Additionally, because our framework is trained on existing public datasets, it may inadvertently reflect or reinforce geographic or demographic biases embedded in those original annotations. To mitigate these risks, we strongly advocate that any downstream applications deploying this framework secure explicit user consent and implement robust data privacy protections.

Artifact Licensing and Terms of Use

All 14 speech emotion recognition datasets utilized in this study are publicly or academically accessible artifacts. We have processed and evaluated these corpora in strict accordance with their original distribution terms. Specifically, datasets requiring formal research permissions (such as MSP-IMPROV) were accessed under signed Academic End-User Licenses from their respective institutional creators, while others (such as EmoDB) were used under

open-access or public domain terms for scientific exploration. Our use of these assets remains entirely consistent with their intended application as non-commercial benchmarks for speech and language modeling, ensuring full compliance with the ethical and legal frameworks established by the data creators.

Generative AI Disclosure

We utilized the Gemini large language model purely as a writing assistant to refine the language, improve clarity, and smooth out the narrative flow of our original text. The AI tool was not involved in generating new research ideas, designing experiments, all core scientific concepts and contributions remain entirely our own.

References

- S. Amiriparian, F. Packań, M. Gerczuk, and B.W. Schuller. 2024. [ExHuBERT: Enhancing HuBERT through block extension and fine-tuning on 37 emotion datasets](#). In *Interspeech 2024*, pages 2635–2639, Kos Island, Greece.
- F. Burkhardt, F. Eyben, and B. Schuller. 2022. SynAct: A synthesized database of basic emotions. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 1–9.
- F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. 2005. A database of German emotional speech. In *Interspeech*, volume 5, pages 1517–1520.
- C. Busso, M. Bulut, and S.S. Narayanan. 2013. [Toward effective automatic recognition systems of emotion in speech](#). In J. Gratch and S. Marsella, editors, *Social emotions in nature and artifact: emotions in human and human-computer interaction*, pages 110–127. Oxford University Press, New York, NY, USA.
- C. Busso, R. Lotfian, K. Sridhar, A.N. Salman, W.-C. Lin, L. Goncalves, S. Parthasarathy, A. Reddy Naini, S.-G. Leem, L. Martinez-Lucas, H.-C. Chou, and P. Mote. 2026. [The MSP-Podcast corpus](#). *IEEE Transactions on Affective Computing*, early access.
- C. Busso, S. Parthasarathy, A. Burmania, M. Abdel-Wahab, N. Sadoughi, and E. Mower Provost. 2017. [MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception](#). *IEEE Transactions on Affective Computing*, 8(1):67–80.
- R. Calvo. 2015. Affect-aware assistive technologies. *Assistive Technology for Cognition*, page 47.
- 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. 2022. [WavLM: Large-scale self-supervised pre-training for full stack speech processing](#). *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco. 2014. [EMOVO corpus: an Italian emotional speech database](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3501–3504, Reykjavik, Iceland. European Language Resources Association (ELRA).
- R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80.
- T. Durand, N. Mehrasa, and G. Mori. 2019. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- C. Fabio. 2023. [Speech emotion recognition in italian using wav2vec 2](#). *TechRxiv*, 2023(0519).
- S. Feraru, D. Schuller, and B. Schuller. 2015. [Cross-language acoustic emotion recognition: An overview and some tendencies](#). In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 125–131.
- L. Goncalves, D. Robinson, E. Richerson, and C. Busso. 2024a. [Bridging emotions across languages: Low rank adaptation for multilingual speech emotion recognition](#). In *Interspeech 2024*, pages 4688–4692, Kos Island, Greece.
- L. Goncalves, A.N. Salman, A. Reddy Naini, L. Moro-Velazquez, T. Thebaud, P. Garcia, N. Dehak, B. Sisman, and C. Busso. 2024b. [Odyssey 2024 - speech emotion recognition challenge: Dataset, baseline framework, and results](#). In *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pages 247–254, Quebec, Canada.
- P. Gournay, O. Lahaie, and R. Lefebvre. 2018. A Canadian French emotional speech dataset. In *Proceedings of the 9th ACM multimedia systems conference*, pages 399–402.
- V. Hozjan and Z. Kačič. 2003. Context-independent multilingual emotion recognition from speech signals. *International journal of speech technology*, 6(3):311–320.
- V. Kondratenko, A. Sokolov, N. Karpov, O. Kutuzov, N. Savushkin, and F. Minkin. 2022. [Large raw emotional dataset with aggregation mechanism](#).
- S. Lee. 2019. [The generalization effect for multilingual speech emotion recognition across heterogeneous languages](#). In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5881–5885.
- S.-G. Leem, D. Fulford, J.-P. Onnela, D.E. Gard, and C. Busso. 2022. [Not all features are equal: Selection of robust features for speech emotion recognition in noisy environments](#). In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, pages 6447–6451, Singapore.
- S.-G. Leem, D. Fulford, J.-P. Onnela, D.E. Gard, and C. Busso. 2024. [Selective acoustic feature enhancement for speech emotion recognition with noisy speech](#). *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 32:917–929.
- K. Leila, C. Catherine, S. Youssef, and R. Kosai. 2020. [French emotional speech database - oreau](#).
- W.-C. Lin and C. Busso. 2023. [Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling](#). *IEEE Transactions on Affective Computing*, 14(2):1215–1227.
- I. Lubenets, N. Davidchuk, and A. Amentes. 2022. [Aniemore](#).
- Z. Ma, M. Chen, H. Zhang, Z. Zheng, W. Chen, X. Li, J. Ye, X. Chen, and T. Hain. 2024a. [EmoBox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark](#). *Proc. Interspeech 2024*, 1580-1584, doi: 10.21437/Interspeech.2024-788.
- Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen. 2024b. [emotion2vec: Self-supervised pre-training for speech emotion representation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760.
- S. Mariooryad and C. Busso. 2017. [The cost of dichotomizing continuous labels for binary classification problems: Deriving a Bayesian-optimal classifier](#). *IEEE Transactions on Affective Computing*, 8(1):119–130.
- P. Mote, A. Naini, D. Robinson, E. Richerson, and C. Busso. 2025a. [Analysis of phonetic level similarities across languages in emotional speech](#). In *Proc. Interspeech 2025*, pages 4343–4347.
- P. Mote, D. Robinson, E. Richerson, and C. Busso. 2025b. [Vector quantized cross-lingual unsupervised domain adaptation for speech emotion recognition](#). In *Proc. Interspeech 2025*, pages 126–130.
- A. Reddy Naini and C. Busso. 2025. [Multi-dimensional ordinal embedding for attribute modeling in speech emotion recognition](#). In *International Conference on Affective Computing and Intelligent Interaction (ACII 2025)*, Canberra, Australia.
- A. Reddy Naini, L. Goncalves, A.N. Salman, P. Mote, I.R. Ülgen, T. Thebaud, L. Moro-Velazquez, L.P. Garcia, N. Dehak, B. Sisman, and C. Busso. 2025. [The Interspeech 2025 challenge on speech emotion recognition in naturalistic conditions](#). In *Interspeech 2025*, pages 4668–4672, Rotterdam, The Netherlands.

- A. Rusu, N. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. 2022. [Progressive neural networks](#). Preprint, arXiv:1606.04671.
- M. Sharma. 2022. Multi-lingual multi-task speech emotion recognition using wav2vec 2.0. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6907–6911. IEEE.
- A. Shinde, V. Patil, A. Shaikh, P. More, and K. Salve. 2023. Design and validation of HindiSER: speech emotion recognition dataset for hindi language. In *2023 26th Conference of the Oriental COCODA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–5. IEEE.
- Y. Singh and S. Goel. 2023. A lightweight 2D CNN based approach for speaker-independent emotion recognition from speech with new Indian Emotional Speech Corpora. *Multimedia Tools and Applications*, 82(15):23055–23073.
- H. Sun, X. Wang, J. Zhao, S. Zhao, J. Zhou, H. Wang, A. Kong, X. Yang, Y. Wang, Y. Lin, and Y. Qin. 2025. [Emotiontalk: An interactive chinese multimodal emotion dataset with rich annotations](#).
- G. Tu, J. Wen, C. Liu, D. Jiang, and E. Cambria. 2022. Context-and sentiment-aware networks for emotion recognition in conversation. *IEEE Transactions on Artificial Intelligence*, 3(5):699–708.
- S.G. Upadhyay, C. Busso, and C.-C. Lee. 2025a. Speaker style-aware phoneme anchoring for improved cross-lingual speech emotion recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2025)*, Honolulu, HI, USA.
- S.G. Upadhyay, W.-S. Chien, B.-H. Su, L. Goncalves, Y.-T. Wu, A.N. Salman, C. Busso, and C.-C. Lee. 2023. [An intelligent infrastructure toward large scale naturalistic affective speech corpora collection](#). In *International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, pages 1–8, Cambridge, MA, USA.
- S.G. Upadhyay, L. Martinez-Lucas, W. Katz, C. Busso, and C.-C. Lee. 2025b. [Phonetically-anchored domain adaptation for cross-lingual speech emotion recognition](#). *IEEE Transactions on Affective Computing*, 16(3):1631–1645.
- J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. Schuller. 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759.
- R. Xia and Y. Liu. 2017. [A multi-task learning framework for emotion recognition using 2D continuous space](#). *IEEE Transactions on Affective Computing*, 8(1):3–14.
- Ke Yan, Jinzheng Cai, Youjing Zheng, Adam P. Harrison, Dakai Jin, Youbao Tang, Yuxing Tang, Lingyun Huang, Jing Xiao, and Le Lu. 2021. [Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in ct](#). *IEEE Transactions on Medical Imaging*, 40(10):2759–2770.
- W. Zehra, A.R. Javed, Z. Jalil, H.U. Khan, and T.R. Gadekallu. 2021. [Cross corpus multi-lingual speech emotion recognition using ensemble learning](#). *Complex & Intelligent Systems*, 7(4):1845–1854.
- X. Zhou, V. Koltun, and P. Krähenbühl. 2022. Simple multi-dataset detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7571–7580.