Cross-Dataset Speech Emotion Recognition: A Comparative Study of Representation Learning, Domain Adaptation, and Mobile Constraints

Himani Thakor
Ph.D Scholar
Department of Computer Science
Veer Narmad South Gujarat University, Surat, Gujarat, India
E-Mail: ht3991214@gmail.com

Abstract:

Speech Emotion Recognition (SER) is a critical component of modern affective computing and human-computer interaction (HCI). While state-of-the-art SER models achieve high accuracy (above 80–90%) in intra-dataset settings where training and evaluation distributions match, their generalization drops drastically to 30–50% in cross-dataset scenarios. This degradation is caused by acoustic, speaker, language, and elicitation domain shifts. This paper presents an expanded comparative study of the two dominant modern paradigms for mitigating this discrepancy: Self-Supervised Learning (SSL) foundation features and Unsupervised Domain Adaptation (UDA). We analyze recent architectural breakthroughs, such as distilled mobile-efficient transformers (DistilHuBERT), adversarial dual discriminators (ADDi), and acoustic knowledge-guided transfer subspace learning (AKTLR). Finally, we categorize the major empirical phenomena that govern cross-dataset SER, including the "theatricality effect" in acted corpora, cross-lingual phonetic shifts, and multi-corpus data augmentation strategies, backed by an expanded set of primary and foundational biblio-graphic references.

1. Introduction:

Automatic Speech Emotion Recognition (SER) plays an increasingly prominent role in healthcare, robotics, customer service, and conversational agents. However, the lack of generalization in existing models remains a key bottleneck for real-world deployment. In laboratory conditions, models evaluated on stratified test splits of the same dataset routinely report excellent performance. However, evaluating these same models on unseen target databases exposes a steep performance degradation, with classification accuracies frequently collapsing to random guess rates.

This cross-dataset and cross-corpus generalizability challenge is driven by multi-faceted domain shifts, which can be categorized into four primary dimensions:

1. **Acoustic & Environmental Domain Shift**: Dissimilarities in microphone sensitivity, hardware frequency responses, background ambient noise, room acoustics, and sampling rates (e.g., studio-grade 48 kHz recordings vs. mobile-grade 8 kHz or 16 kHz streams).
2. **Speaker & Cultural Variability**: Demographic imbalances across databases (such as age, gender, and ethnic distributions), individual speaking temperaments, accent variations, and language-specific prosodic contours.

To establish a basis for comparative analysis, Section 2 surveys the standard emotional speech datasets. Section 3 contrasts foundation SSL representation models against subspace

and Adversarial domain adaptation. Section 4 examines the physical and linguistic phenomena that shape transferability, and Section 5 discusses the engineering trade-offs in mobile deployment.

2. Comparative Survey of Emotional Speech Corpora

Cross-dataset research relies on several publicly available speech databases. These databases vary significantly across multiple acoustic and phonetic dimensions, producing the domain gaps that challenge traditional machine learning models. Table 1 summarizes the characteristics of the four major corpora analyzed in this study.

Table 1: Characteristics of standard speech emotion corpora used in cross-dataset benchmarks.

Dataset	Language	Speakers	Speech ture	Na-	Utterances	Core Emotion Categories
IEMOCAP [1]	English	10	Semi-natural (dyadic)		~10,000	Anger, Happiness, Sad-ness, Neutral
RAVDESS [2]	English	24	Acted (speech & song)		1,440	Anger, Calm, Disgust, Fear, Happy
EMO-DB [3]	German	10	Acted (studio -grade)		535	Anger, Boredom, Dis-gust, Fear, Happy
CREMA-D [4]	English	91	Acted (multi -ethnic)		7,442	Anger, Disgust, Fear, Happiness, Neutral

- **IEMOCAP (Interactive Emotional Dyadic Motion Capture) [1]**: Represents semi-naturalistic interactions where actors perform in scripted and improvised dyadic scenarios. It offers rich context but contains acoustic noise and highly blended, overlapping expressions.
- **RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [2]**: A highly structured English dataset where 24 actors recite standard sentences at two different emotional intensity levels (normal and strong).
- **EMO-DB (Berlin Database of Emotional Speech) [3]**: A high-fidelity German acted database recorded in studio-grade conditions, containing seven emotional categories.
- **CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) [4]**: Consists of 7,442 clips from 91 multi-ethnic actors. It provides significant ethnic and age diversity, making it a critical asset for evaluating generalization.

3. Architectural Paradigms: SSL Features vs. Unsupervised Domain Adaptation

Modern attempts to resolve the cross-dataset performance drop relieve on two separate but complementary paradigms: Self-Supervised Learning (SSL) speech features and Unsupervised Domain Adaptation (UDA).

3.1 Self-Supervised Learning Foundation Features

Traditional speech emotion systems utilized handcrafted, low-level acoustic descriptors (LLDs) such as Mel-Frequency Cepstral Coefficients (MFCCs), fundamental frequency (F0), jitter,

shimmer, and standard feature sets like eGeMAPS. While computationally lightweight, these features fail to capture rich contextual information and are highly sensitive to microphone induced acoustic shifts.

To address this, modern models fine-tune transformer-based self-supervised foundation back bones like Wav2Vec 2.0 [5] and HuBERT [6]. Pre-trained on massive unlabeled speech corpora (such as LibriSpeech), these models construct robust contextual temporal-spectral representations. Fine-tuning the upper layers of these foundation models on SER tasks yields features that are significantly more resilient to minor acoustic and background discrepancies than shallow, handcrafted descriptors.

3.2 Distilled and Mobile-Efficient SSL Models

Although full-scale foundation models (such as Wav2Vec 2.0 Large, containing over 300M parameters) provide robust representations, their computational footprint prevents deployment on resource constrained devices like mobile phones and edge hardware. Ismail [9] investigated this limitation by validating a mobile-efficient SER system built on DistilHuBERT, a distilled and 8-bit quantified transformer.

DistilHuBERT achieves an impressive 92% parameter reduction compared to full-scale Wav2Vec 2.0 baselines, shrinking the model to a quantized footprint of only 23 MB. In Leave-One-Session-Out (LOSO) cross-validation on the IEMOCAP dataset, this quantized model retains 91% of the baseline performance, achieving an Unweighted Accuracy (UA) of 61.4% [9]. This presents a Pareto-optimal tradeoff, verifying that extreme model compression is compatible with cross dataset viability.

3.3 Unsupervised Domain Adaptation (UDA) & Subspace Learning

When target dataset labels are unavailable, aligning the marginal or conditional feature distributions of the source and target domains in a shared latent space is necessary.

3.3.1 Adversarial Dual Discriminator (ADDi & sADDi)

Latif et al. [7] proposed the Adversarial Dual Discriminator (ADDi) framework, which approaches domain alignment as a three-player adversarial game to learn generalized representations without target labels. To mitigate large domain gaps, particularly in cross-language settings, the authors introduced the self-supervised ADDi (sADDi) network. This architecture incorporates self-supervised pre-training with unlabeled data alongside synthetic data generation as an emotional pretext task, yielding highly discriminative, domain-invariant representations [7].

3.3.2 Acoustic Knowledge-Guided Subspace Learning (AKTLR)

Rather than treating domain alignment as a generic transfer learning problem, Zhao et al. [8] argued that domain-specific speech knowledge should guide the alignment process. They introduced Acoustic Knowledge Guided Transfer Linear Regression (AKTLR), built on a dual sparsity constraint mechanism.

Instead of operating on massive, unconstrained deep feature spaces that induce classifier over adaptation, AKTLR identifies a minimalistic, highly generalizable acoustic parameter set. It enforces sparsity across two levels:

$$\text{Min}||W||W\ 2,1 + \lambda\ ||W||1 \quad (1)$$

where W represents the projection matrix. This formulation optimizes for: (1) contributive acoustic parameter groups, and (2) active constituent elements within each group. Evaluated across EmoDB, eNTERFACE, and CASIA, AKTLR outperformed deep transfer learning methods, demonstrating the value of incorporating explicit acoustic domain knowledge [8].

3.3.3 Subdomain Adaptation

Standard domain-adversarial methods align the global distributions of the source and target domains. However, global alignment risks aligning different emotional categories across domains (e.g., aligning source “Anger” features with target “Happiness” features due to high arousal). Subdomain Adaptation mitigates this by aligning class-conditional distributions. By integrating deep denoising auto-encoders with joint alignment of both emotion and gender features, sub domain adaptation prevents global distribution distortions and minimizes cross class alignment errors.

4. Critical Empirical Phenomena

Understanding cross-dataset SER requires analyzing several physical, acoustic, and linguistic phenomena that consistently emerge during multi-corpus evaluations.

4.1The “Theatricality Effect” and Arousal-Valence Clustering

A primary challenge in cross-dataset validation is the elicitation mismatch between acted and naturalistic speech. Acted datasets (such as RAVDESS) feature theatrical expressions

where actors prioritize acoustic clarity over emotional subtlety. This produces severe acoustic saturation in high-energy expressions.

When a model trained on semi-natural speech (such as IEMOCAP) is evaluated on RAVDESS, the performance drops to 46.64% unweighted accuracy [9]. Analysis reveals that the theatricality of RAVDESS causes predictions to cluster primarily by arousal level rather than valence. Due to acoustic saturation, happiness predictions systematically bleed into anger predictions, while sadness predictions bleed into neutral predictions. Despite this bleed, the model maintains high arousal detection, achieving 99% recall for anger, 55% for neutral, and 27% for sadness, illustrating that arousal generalized far better than valence under acoustic saturation [9].

4.2Cross-Corpus Augmentation

Multi-corpus training has emerged as an effective methodology to regularize SER backbones. In the DistilHuBERT validation study [9], augmenting the training protocol of IEMOCAP with CREMA-D yielded significant stability improvements:

A 1.2% improvement in Weighted Accuracy (WA).

A 1.4% gain in Macro F1-score.

A substantial 32% reduction in cross-fold variance, indicating robust speaker independence.

A 5.4% F1-score improvement specifically for the highly elusive Neutral class.

4.3Cross-Lingual Degradation

Cross-language SER introduces phonetic variations that act as major domain-shifting forces. Models trained on English (IEMOCAP) experience significant degradation when evaluated on non-English datasets. However, the degradation is highly non-uniform. Under zero-shot transfer conditions, an English-trained model achieves up to 65% accuracy on German (EMO-DB) but drops to 32% accuracy when evaluated on Italian (EMOVO) [7]. This disparity highlights that acoustic-phonetic structural similarities between Germanic languages and English provide a gentler transfer boundary compared to Romance languages.

5. Engineering and Computational Trade-offs

For practical SER applications, system designers face a stark trade-off between alignment capabilities and computational complexity. Table 2 summarizes these architectural paradigms.

Table 2: Comparative analysis of architectural trade-offs in cross-dataset SER systems.

Paradigm	Model Size	Computation Type	Key Generalized Domain Benefit
Full SSL Transformer [5, 6]	>300M params	High Floating Point (FP32)	High generalizability to acoustic noise
Quantized DistilHuBERT [9]	23 MB (8-bit)	Ultra-low (Mobile)	Pareto-optimal size-to-accuracy
AKTLR (Subspace) [8]	Minimal	Low (Linear algebra)	Minimal feature set, prevents over adaptation
sADDi (Adversarial) [7]	Medium	Generative Training	Unsupervised zero-shot language transfer

While deep adversarial architectures (like sADDi) successfully construct domain-invariant spaces for zero-shot cross-language transfer, they require complex training and significant GPU resources. In contrast, quantized models like DistilHuBERT allow for immediate edge execution, but remain vulnerable to the "theatricality effect" in the absence of explicit domain adaptation.

6. Conclusion

Cross-dataset speech emotion recognition highlights the limitations of standard supervised learning under domain shifts. This comparative study demonstrates that while self-supervised foundation features (Wav2Vec 2.0, HuBERT) provide robust representations, they must be paired with either parameter-efficient quantization (DistilHuBERT) for mobile viability, or advanced distribution alignment (ADDi, AKTLR, subdomain adaptation) to mitigate acoustic and language discrepancies. Future research must focus on unifying these paradigms, developing highly compressed, quantized foundation backbones that incorporate acoustic knowledge-guided constraints to achieve robust, universal affect recognition.

References

[1] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[2] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English," *PLoS ONE*,

vol. 13, no. 5, p. e0196391, 2018.

[3] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Interspeech*, vol. 5, pp. 1517–1520, 2005.

[4] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[5] 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. 12449–12460, 2020.

[6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[7] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. Schuller, "Self-Supervised Adversarial Domain Adaptation for Cross-Corpus and Cross-Language Speech Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1045–1058, 2022.

[8] Y. Zhao, Y. Zong, H. Lian, C. Lu, J. Shi, and W. Zheng, "Towards Domain-Specific Cross-Corpus Speech Emotion Recognition Approach," *arXiv preprint arXiv:2312.06466*, 2023.

[9] S. M. Ismail, "Distilled HuBERT for Mobile Speech Emotion Recognition: A Cross-Corpus Validation Study," *arXiv preprint arXiv:2512.23435*, 2025.