International Journal of Research in Management & Social Science
Volume 13, Issue 2 April - June 2025
ISSN 2322 - 0899

ENHANCING GENERALIZATION IN SPEECH EMOTION RECOGNITION: A COMPREHENSIVE REVIEW OF DOMAIN ADAPTATION AND TRANSFER LEARNING TECHNIQUES

[1]Irfan Chaugule and [2]Dr. Satish R Sankaye
[1]Research Scholar, MGM University, DR.G.Y. Pathrikar College of Computer Science and Information Technology, Chhatrapati Sambhajinagar, Maharashtra
irfanchaugule@gmail.com

[2]MGM University, DR.G.Y. Pathrikar College of Computer Science and Information Technology, Chhatrapati Sambhajinagar, Maharashtra
Sankayesr@gmail.com

ABSTRACT
Speech Emotion Recognition (SER) is a rapidly advancing field with significant implications for human-computer interaction, affective computing, and various real-world applications such as mental health monitoring and customer service enhancement. However, a primary challenge hindering the widespread deployment of SER systems is their limited ability to generalize across diverse acoustic conditions, datasets, languages, and speaker characteristics. Models trained on specific corpora often experience substantial performance degradation when exposed to unseen data, a phenomenon attributed to dataset bias and domain shift. This paper provides a comprehensive review of Domain Adaptation (DA) and Transfer Learning (TL) techniques as pivotal solutions to address these generalization challenges in SER. We delve into the evolution of SER, highlighting the limitations of traditional approaches and the rise of deep learning. Subsequently, the paper systematically explores a wide array of DA and TL methodologies, including fine-tuning pre-trained models (e.g., Wav2Vec2, HuBERT), Parameter-Efficient Fine-Tuning (PEFT) strategies (e.g., adapters, LoRA), discrepancy-based adaptation (e.g., Maximum Mean Discrepancy), adversarial domain adaptation (e.g., DANNs), and multi-task learning. We discuss the architectural considerations, underlying principles, and application of these techniques in the SER context. Furthermore, the paper examines common benchmark datasets, cross-domain evaluation protocols, and performance metrics crucial for assessing the efficacy of DA and TL approaches. Through a synthesis of recent advancements and seminal works, this review identifies key trends, discusses current challenges such as data scarcity and the acted-versus-natural emotion gap, and outlines promising future research directions. The overarching goal is to offer researchers and practitioners a thorough understanding of how DA and TL can be leveraged to build more robust, reliable, and universally applicable SER systems.

Keywords: *Speech Emotion Recognition (SER), Domain Adaptation, Transfer Learning, Deep Learning, Generalization, Cross-Corpus SER, Cross-Lingual SER, Affective Computing, Robustness.*

1. INTRODUCTION

1.1. Background and Significance of Speech Emotion Recognition (SER)
Speech Emotion Recognition (SER) has solidified its position as a critical research area at the intersection of affective computing, signal processing, and artificial intelligence (Madanian et al., 2023 ; Schuller, 2018). Its core objective is to enable machines to automatically identify and interpret human emotional states from vocal cues embedded in speech signals (Singh et al., 2023 ; Barman & Shanthini, 2023). The significance of SER is underscored by its vast potential to transform human-computer interaction (HCI), making it more intuitive, empathetic, and natural (Cowie et al., 2001). Applications of robust SER systems are diverse and impactful, ranging from mental health assessment and patient monitoring in healthcare (Singh et al., 2023) to enhancing customer experiences in call centers (Barman & Shanthini, 2023), developing safer in-car environments by monitoring driver states (Sönmez & Varol, 2020), creating adaptive e-learning systems (Li et al., 2007, as cited in), and building more emotionally intelligent AI agents (Singh et al., 2023). As voice-based interfaces become increasingly ubiquitous, the demand for SER systems that can reliably function across varied real-world scenarios continues to grow (Bertero & Fung, 2017).

1.2. The Challenge of Generalization in SER
Despite significant progress, particularly with the advent of deep learning (DL) techniques (Latif et al., 2018 ; Schuller, 2018), a fundamental challenge plagues the practical deployment of SER systems:

Poor Generalization. Models trained on specific datasets, often recorded under controlled conditions with professional actors, tend to perform inadequately when applied to new, unseen data from different domains

ISSN 2322 - 0899

(Parry et al., 2019 ; Deng et al., 2023). This "domain shift" or "dataset mismatch" problem arises from numerous factors, including:

- **Acoustic Variability:** Differences in recording environments (e.g., studio vs. noisy real-world), microphone characteristics, and channel effects (Akçay & Oguz, 2020).
- **Speaker Diversity:** Variations in speaker age, gender, accent, dialect, and individual vocal tract characteristics (Madanian et al., 2023).
- **Linguistic and Cultural Differences:** Emotional prosody, lexical choices, and expression styles can vary significantly across languages and cultures (Deng et al., 2023 ; Latif et al., 2022).
- **Emotional Expression Styles:** Discrepancies between acted emotions (often exaggerated and clear) prevalent in many datasets and spontaneous, natural emotions (more subtle and complex) encountered in real-life interactions (Akçay & Oguz, 2020 ;).
- **Annotation Subjectivity and Scarcity:** Inherent subjectivity in labeling emotions and the limited availability of large-scale, consistently annotated emotional speech corpora further exacerbate the generalization problem (Latif et al., 2018 ;).

This lack of robustness significantly limits the real-world utility of SER systems, as they cannot be reliably deployed in the diverse and unpredictable environments for which they are intended. Addressing this generalization gap is paramount for the continued advancement and practical impact of SER technology.

1.3. Domain Adaptation and Transfer Learning as Solutions

To mitigate the challenges of domain shift and improve model generalization, **Domain Adaptation (DA)** and **Transfer Learning (TL)** have emerged as powerful and extensively researched paradigms in machine learning, with increasing application to SER (Akçay & Oguz, 2020 ; Latif et al., 2022 ;).

- **Transfer Learning (TL)** broadly refers to leveraging knowledge gained from solving one problem (source task/domain) to improve learning and performance on a different but related problem (target task/domain) (Pan & Yang, 2010, as cited in). In SER, this often involves using models pre-trained on large general speech datasets (e.g., for Automatic Speech Recognition - ASR) or even other modalities, and then fine-tuning them on smaller, specific emotion datasets.
- **Domain Adaptation (DA)** is a specific type of transfer learning where the source and target tasks are the same (e.g., emotion classification), but the data distributions of the source and target domains differ (Ganin et al., 2016). DA techniques aim to learn feature representations that are invariant to these domain differences, thereby allowing models trained on a source domain to perform well on a target domain, often with limited or no labeled data from the target domain (Akçay & Oguz, 2020).

By effectively applying DA and TL, SER models can learn more robust and generalizable representations, reducing their sensitivity to variations in datasets, languages, and recording conditions.

1.4. Objectives and Scope of the Paper

This paper aims to provide a comprehensive review and synthesis of domain adaptation and transfer learning techniques applied to Speech Emotion Recognition. The specific objectives are:

1. To provide an overview of the evolution of SER and the critical need for generalization.
2. To systematically survey various DA and TL methodologies relevant to SER, including pre-trained model fine-tuning, parameter-efficient adaptation, discrepancy-based methods, adversarial learning, and multi-task learning.
3. To discuss the application of these techniques in addressing cross-corpus, cross-lingual, and cross-condition challenges in SER.
4. To review common benchmark datasets, evaluation protocols, and performance metrics used in DA/TL for SER research.
5. To analyze the state-of-the-art, identify current challenges, and propose promising future research directions in this domain.

The scope of this paper encompasses theoretical underpinnings, architectural considerations, practical implementations, and empirical evidence from recent studies, focusing primarily on deep learning-based approaches.

ISSN 2322 - 0899

1.5. Organization of the Paper

The remainder of this paper is organized as follows: Section 2 provides a background on SER, highlighting the evolution of techniques and the persistent generalization problem. Section 3 delves into the foundational concepts of Transfer Learning. Section 4 focuses specifically on Domain Adaptation techniques. Section 5 discusses the application and impact of these DA and TL methodologies in the context of SER, covering various approaches and their effectiveness. Section 6 reviews common datasets and evaluation strategies pertinent to cross-domain SER. Section 7 presents a discussion on the current state-of-the-art, challenges, and limitations. Finally, Section 8 concludes the paper and outlines future research directions.

2. BACKGROUND: SPEECH EMOTION RECOGNITION AND THE GENERALIZATION CHALLENGE

2.1. Evolution of SER Techniques

The journey of SER began with early psychological investigations into the acoustic correlates of emotion (Blanton, 1915, as cited in Schuller et al., 2018). Computational SER gained prominence with pioneering works in the mid-1990s (Daellert et al., 1996, as cited in ; Picard, 1997). Initial systems relied on rule-based approaches or traditional machine learning models like Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), and Support Vector Machines (SVMs), typically using handcrafted acoustic features such as pitch, energy, formants, and Mel-Frequency Cepstral Coefficients (MFCCs) (Ayadi et al., 2011 ;).

The advent of deep learning (DL) marked a significant turning point, with architectures like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) – particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks – and Transformers demonstrating the ability to automatically learn hierarchical and discriminative features directly from speech data (e.g., spectrograms or raw waveforms) (Latif et al., 2018 ;). These DL models often outperform traditional methods, especially when large amounts of training data are available (Akçay & Oguz, 2020). Hybrid models, combining the strengths of CNNs for spatial/spectral feature extraction and RNNs for temporal modeling, have also become popular (Kumar et al., 2025 ; Zhao et al., 2019). More recently, self-supervised learning (SSL) models pre-trained on vast unlabeled speech datasets, such as Wav2Vec2 (Baevski et al., as cited in) and HuBERT (Hsu et al., 2021), have shown remarkable success as powerful feature extractors for SER when fine-tuned on downstream tasks.

2.2. The Pervasive Problem of Generalization

Despite these advancements, the "Achilles' heel" of SER remains its struggle with generalization. As detailed in Section 1.2, SER models often exhibit a significant performance drop when evaluated on data that differs from their training distribution (Parry et al., 2019). This issue is particularly acute in:

- **Cross-Corpus SER:** Models trained on one emotional speech corpus (e.g., IEMOCAP) perform poorly on another (e.g., RAVDESS) due to variations in recording setups, speaker demographics, annotation protocols, and emotional expression styles (Schuller et al., 2010; Deng et al., 2023). Parry et al. (2019) found that while cross-corpus training can be a promising approach, RNN-based architectures were more prone to overfitting the training corpora compared to CNNs.
- **Cross-Lingual SER:** Generalizing across different languages is even more challenging, as prosodic and lexical cues for emotions are often language-dependent (Latif et al., 2022;).
- **Cross-Condition SER:** Variations in background noise, reverberation, and communication channels in real-world scenarios can severely degrade performance.
- **Acted vs. Natural Emotions:** Most available datasets contain acted emotions, which can be acoustically distinct from spontaneous, natural emotions. Models trained on acted data often fail to generalize to real-world emotional expressions.

The inability to generalize effectively limits the practical utility of SER systems, necessitating robust solutions to bridge these domain gaps.

3. TRANSFER LEARNING IN SPEECH EMOTION RECOGNITION

Transfer Learning (TL) offers a powerful paradigm to address data scarcity and improve generalization in SER by leveraging knowledge from related tasks or domains where data is more abundant.

3.1. Core Concepts of Transfer Learning

TL involves a source domain (DS) and learning task (TS), and a target domain (DT) and learning task (TT). The goal is to improve the learning of the target predictive function $f_T(\cdot)$ in DT using the knowledge in DS and TS, where DS $\neq$ DT or TS $\neq$ TT (Pan & Yang, 2010, as cited in). In many SER applications, the target task

ISSN 2322 - 0899

TT (emotion classification) is the same as a potential source task TS, but the domains differ (DS□=DT).

Key benefits of TL include:

- Reduced need for large labeled target datasets.
- Faster training convergence on the target task.
- Improved performance on the target task by starting with a more informed model.

3.2. Fine-Tuning Pre-Trained Models

A dominant TL strategy in SER involves fine-tuning models pre-trained on large-scale datasets from related domains.

3.2.1. Pre-training on General Speech Data (ASR, Speaker ID)

Models pre-trained for Automatic Speech Recognition (ASR) or Speaker Identification (SID) on thousands of hours of speech learn rich, general-purpose acoustic representations. These representations capture fundamental characteristics of speech signals (phonetics, speaker traits, prosody) that are also relevant for emotion recognition.

- **Wav2Vec2 and HuBERT:** SSL models like Wav2Vec2 (Baevski et al., as cited in) and HuBERT (Hsu et al., 2021) are pre-trained by learning to predict masked parts of speech audio or hidden units. Fine-tuning these models on SER datasets has become a state-of-the-art approach. For instance, Li et al. proposed using transfer learning to pre-train Transformer-based and Wav2Vec2-based models for SER. Chen et al. (as cited in) proposed a model combining fine-tuned Wav2Vec2 with Neural Controlled Differential Equations (NCDE) for SER, highlighting the benefits of these pre-trained features.
- **Strategies for Fine-tuning:**
  - **Full Fine-tuning:** All layers of the pre-trained model are updated during training on the target SER dataset. This allows the model to adapt comprehensively but requires more data and computation, and risks catastrophic forgetting of the pre-trained knowledge if the target dataset is too small.
  - **Partial Fine-tuning (Layer Freezing):** Only a subset of layers (typically the top layers) are fine-tuned, while earlier layers (which learn more general features) are kept frozen. This preserves more of the pre-trained knowledge and is often more effective with smaller target datasets.

3.2.2. Pre-training on Other Modalities (e.g., Image Recognition)

Some early TL approaches in SER explored fine-tuning models pre-trained on large image datasets (e.g., ImageNet) by converting speech spectrograms into image-like inputs for CNNs (Zhang et al., 2017, as cited in). While less common now with the advent of powerful speech SSL models, this demonstrated the potential of transferring knowledge across modalities.

3.3. Parameter-Efficient Fine-Tuning (PEFT)

Fine-tuning entire large pre-trained models (like Wav2Vec2 or HuBERT, which can have hundreds of millions or billions of parameters) can be computationally expensive and memory-intensive, and may still lead to overfitting on small SER datasets. PEFT methods address this by updating only a small fraction of the model's parameters, keeping the bulk of the pre-trained weights frozen. This significantly reduces computational costs and storage requirements.

Key PEFT techniques include:

- **Adapter Modules:** Small neural network modules (adapters) are inserted between the layers of the pre-trained Transformer. Only the parameters of these adapters are trained, while the original Transformer weights remain fixed. Residual adapters, which add a small learned residual to the output of a Transformer layer, are a common variant (Xi et al., 2018). These adapters learn task-specific deviations.
- **Low-Rank Adaptation (LoRA):** LoRA injects trainable rank decomposition matrices into Transformer layers, effectively learning low-rank updates to the original weight matrices. This drastically reduces the number of trainable parameters.
- **Bottleneck Adapters (BA):** These consist of a down-projection layer (reducing dimensionality), a non-linear transformation, and an up-projection layer (restoring dimensionality), forcing information through a lower-dimensional bottleneck.
- **Weighted Sum (WS) / Weighted Gating (WG):** These methods learn weights for the outputs of different Transformer blocks (WS) or gate the hidden states (WG), allowing the model to learn the relevance of different layers or feature dimensions for the downstream task.

ISSN 2322 - 0899

PEFT methods have shown comparable or even better performance than full fine-tuning for SER while being significantly more efficient. A combination of different PEFT methods often yields the best results.

3.4. Multi-Task Learning (MTL) for Knowledge Transfer

MTL involves training a model to perform multiple related tasks simultaneously, using a shared representation. By learning SER alongside auxiliary tasks like ASR, speaker identification, or gender recognition, the model can learn features that are more robust and disentangled from task-irrelevant factors, thereby improving generalization for the primary SER task (Han et al., 2014; Kim et al., 2017).

Cai et al. (2021) proposed an MTL framework based on Wav2Vec2.0 to simultaneously perform speech-to-text recognition and emotion classification, achieving state-of-the-art performance on IEMOCAP. Ghosh et al. introduced MMER, a multimodal multi-task learning approach leveraging cross-modal self-attention and solving auxiliary tasks like ASR and supervised contrastive learning.

4. DOMAIN ADAPTATION IN SPEECH EMOTION RECOGNITION

Domain Adaptation (DA) specifically targets the problem of domain shift, where the data distribution of the source (training) domain differs from that of the target (testing) domain, even if the task remains the same (Akçay & Oguz, 2020).

4.1. Core Concepts of Domain Adaptation

The goal of DA is to learn a model f:X→Y that performs well on a target domain DT=(xiT,yiT)i=1nT by leveraging information from a source domain DS=(xjS,yjS)j=1nS, where the marginal probability distributions PS(X) and PT(X) are different, i.e., PS(X)□=PT(X) (Pan & Yang, 2010, as cited in). DA methods can be categorized based on the availability of labels in the target domain:

- **Supervised DA:** Labeled data is available for both source and target domains.
- **Semi-Supervised DA:** Labeled source data and a small amount of labeled target data, along with unlabeled target data, are available.
- **Unsupervised DA (UDA):** Labeled source data and unlabeled target data are available. This is a common and challenging scenario in SER due to the difficulty of obtaining labeled data for every new target domain.

4.2. Discrepancy-Based Domain Adaptation

These methods aim to explicitly minimize a distance metric between the source and target domain distributions in some feature space.

- **Maximum Mean Discrepancy (MMD):** MMD is a non-parametric metric that measures the distance between the means of the source and target samples mapped into a Reproducing Kernel Hilbert Space (RKHS) (Gretton et al., 2012, as cited in). By minimizing MMD between source and target feature representations, the model learns domain-invariant features. Luo and Han used MMD within a non-negative matrix factorization framework to minimize marginal and conditional distribution differences for cross-corpus SER. Liu et al. incorporated MMD loss to minimize differences between feature representations originating from the same stimuli, accounting for rater ambiguity.

4.3. Adversarial Domain Adaptation

Adversarial DA methods employ a domain discriminator network that tries to distinguish between source and target domain features, while the feature extractor network is trained to produce features that "fool" this discriminator, making them domain-indistinguishable (Ganin et al., 2016).

- **Domain-Adversarial Neural Network (DANN):** DANN introduces a domain classifier and a Gradient Reversal Layer (GRL). The GRL reverses the gradient from the domain classifier during backpropagation, forcing the feature extractor to learn domain-invariant features while still being discriminative for the main SER task. This approach has been widely used for cross-corpus and cross-lingual SER. Latif et al. used unsupervised adversarial domain adaptation for multilingual SER, aiming to learn language-invariant emotional representations.
- **Generative Adversarial Networks (GANs) for DA:** GANs can be used to generate synthetic target-domain-like data from source data or vice-versa, or to learn transformations that map features from one domain to another.

Figure 1 illustrates a general architecture for DANN.

ISSN 2322 - 0899

Figure 1: Architecture of a Domain-Adversarial Neural Network (DANN) for SER *(Conceptual Description: A diagram showing an input speech signal fed into a shared Feature Extractor. The output of the Feature Extractor branches into two paths. Path 1 leads to an Emotion Classifier (Label Predictor) which outputs the emotion label and calculates the emotion classification loss using source labels. Path 2 leads through a Gradient Reversal Layer (GRL) to a Domain Classifier, which tries to distinguish between source and target domain inputs and calculates a domain classification loss. The GRL reverses the gradient from the domain classifier to the feature extractor, training the extractor to produce domain-invariant features.)*

4.4. Reconstruction-Based Domain Adaptation

These methods, often involving autoencoders, learn a shared latent space where features from both domains are aligned. By reconstructing features or data from this shared space, the model learns representations that capture common underlying structures while discarding domain-specific variations (Akçay & Oguz, 2020). Deng et al. (2014) used autoencoder-based UDA for SER.

4.5. Instance-Based and Parameter-Based Adaptation

- **Instance Weighting/Selection:** Assigning different weights to source instances or selecting relevant source instances that are most similar to the target domain.
- **Parameter Adaptation:** Adapting the parameters of a pre-trained source model to the target domain, often with regularization to prevent catastrophic forgetting of source knowledge. PEFT methods (Section 3.3) are a form of parameter-based adaptation.

4.6. Addressing Specific Cross-Domain Challenges in SER

4.6.1. Cross-Corpus Adaptation

This involves adapting models trained on one emotional speech corpus to another. Techniques like DANN, MMD minimization, and fine-tuning SSL models are commonly applied. Parry et al. (2019) analyzed deep learning architectures for cross-corpus SER, finding CNNs to generalize better than RNNs when using cross-corpus training. Barsainyan and Singh proposed a normalized 1D CNN framework for cross-corpus SER.

4.6.2. Cross-Lingual Adaptation

Adapting SER models across different languages is particularly challenging due to variations in prosody and emotional expression. Unsupervised adversarial DA and leveraging phonetic commonalities are explored. Latif et al. demonstrated the use of adversarial training for cross-lingual SER without requiring target language labels.

4.6.3. Adaptation Under Label Space Mismatch

In real-world scenarios, the set of emotion labels might differ between source and target domains. Mathur et al. proposed AMLS (Adaptation under Mismatched Label Spaces), an end-to-end architecture using weighting schemes to separate shared and private classes, mitigating negative transfer.

4.6.4. Source-Free Domain Adaptation

A more practical scenario where the source data is not accessible during adaptation due to privacy or other constraints. Luo et al. proposed ECAN (Emotion-aware Contrastive Adaptation Network) for source-free cross-corpus SER, using nearest neighbor contrastive learning.

4.6.5. Two-Stage Adaptation Strategies

For complex shifts, like from acted to natural emotions, a two-stage strategy can be effective. First, a pre-trained model is fine-tuned on a general (e.g., acted emotion) dataset using PEFT. Then, this model is further fine-tuned on the specific target (e.g., natural emotion) dataset, potentially freezing some of the PEFT modules from the first stage to prevent catastrophic forgetting and allow adaptation to the nuances of the target domain.

5. Datasets and Evaluation for Cross-Domain SER

Evaluating the effectiveness of DA and TL techniques requires appropriate datasets and robust evaluation protocols.

5.1. Common Speech Emotion Datasets

Several publicly available datasets are commonly used for SER research, each with distinct characteristics. These variations are what necessitate DA and TL. Key datasets include:

- **IEMOCAP (Interactive Emotional Dyadic Motion Capture Database):** English, ~12 hours, 10 actors, dyadic interactions (scripted and spontaneous), categorical (e.g., happy, sad, angry, neutral, frustrated, excited) and dimensional (valence, arousal, dominance) labels. Valued for naturalistic expressions but has annotation subjectivity.

ISSN 2322 - 0899

- **RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song):** English (North American), 24 actors, ~7356 files, 8 speech emotions (calm, happy, sad, angry, fearful, surprise, disgust, neutral) at two intensity levels. High-quality acted speech.
- **EMO-DB (Berlin Database of Emotional Speech):** German, 10 actors, ~535 utterances, 7 emotions (anger, boredom, disgust, fear, happiness, sadness, neutral). High-quality acted speech recorded in anechoic conditions.
- **SAVEE (Surrey Audio-Visual Expressed Emotion):** British English, 4 male actors, 480 utterances, 7 emotions. Uses TIMIT sentences. Smaller scale, male-only speakers.
- **CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset):** English, 91 diverse actors, ~7442 clips, 6 basic emotions. Acted, diverse ethnicities.
- **MSP-IMPROV:** English, 12 actors, dyadic improvisations, natural emotional speech.

Table 1 summarizes characteristics of some of these key datasets.

Table 1: Overview of Commonly Used Speech Emotion Datasets for DA/TL Research | Dataset | Language | Speakers (M/F) | Emotions (Speech) | Size (Approx.) | Recording Type | Key Characteristics for DA/TL | |---|---|---|---|---|---|

| IEMOCAP | English | 10 (5M/5F) | Happy, Sad, Angry, Neutral, Frustrated, Excited, etc. | ~12 hrs | Acted (Scripted & Spontaneous) | Naturalistic interactions, dimensional labels, annotation subjectivity |

| RAVDESS | English (N. Am.) | 24 (12M/12F) | Calm, Happy, Sad, Angry, Fearful, Surprise, Disgust, Neutral | ~7356 clips | Acted | Balanced gender, multiple intensities, clear expressions |

| EMO-DB | German | 10 (5M/5F) | Anger, Boredom, Disgust, Fear, Happy, Sad, Neutral | ~535 utterances | Acted | High audio quality, different language from IEMOCAP/RAVDESS |

| SAVEE | English (British) | 4 (4M/0F) | Anger, Disgust, Fear, Happy, Sad, Surprise, Neutral | 480 utterances | Acted | Male-only, smaller scale, different accent |

| CREMA-D | English | 91 (48M/43F) | Happy, Sad, Angry, Fearful, Disgust, Neutral | ~7442 clips | Acted | Diverse actors, multiple intensities |

5.2. Evaluation Protocols

- **Intra-Corpus Evaluation:** Training and testing on different splits of the same dataset. Serves as a baseline but doesn't assess generalization well.
- **Cross-Corpus Evaluation:** Training on one or more source corpora and testing on an unseen target corpus. This is the standard for evaluating DA/TL effectiveness.
  - *One-vs-One:* Train on Corpus A, Test on Corpus B.
  - *Leave-One-Corpus-Out (LOCO):* Train on N-1 corpora, Test on the remaining one.
  - *Multi-Source to Single-Target:* Train on an aggregation of multiple source corpora, test on a single target corpus.
- **Cross-Lingual Evaluation:** Source and target corpora are in different languages.
- **Speaker-Independent Evaluation:** Ensuring no speaker overlap between training and testing sets is crucial for realistic performance assessment.

5.3. Performance Metrics

Common metrics for evaluating SER models, especially in DA/TL contexts with potential class imbalance, include:

- **Weighted Accuracy (WA):** Overall accuracy, giving more weight to classes with more samples. Can be misleading for imbalanced datasets.
- **Unweighted Average Recall (UAR):** The average of recall scores for each emotion class. Gives equal importance to each class, robust to imbalance.
- **F1-Score (Macro-averaged):** The unweighted average of F1-scores for each class, balancing precision and recall across classes.
- **Concordance Correlation Coefficient (CCC):** Used for dimensional emotion prediction (valence, arousal, dominance).

ISSN 2322 - 0899

- **Confusion Matrix:** To visualize misclassifications between emotion categories.

6. State-of-the-Art and Comparative Analysis

The application of DA and TL has led to significant advancements in SER generalization.

6.1. Transfer Learning Successes

- **SSL Model Fine-tuning:** Fine-tuning pre-trained SSL models like Wav2Vec2 and HuBERT consistently achieves state-of-the-art or competitive results on various SER benchmarks, even in cross-corpus settings. Their ability to learn rich representations from vast unlabeled data provides a strong foundation.
- **PEFT Efficacy:** PEFT methods like adapters and LoRA have demonstrated the ability to match or exceed the performance of full fine-tuning with significantly fewer trainable parameters, making TL more accessible and efficient. The two-stage adaptation strategy using PEFT has shown particular promise for adapting from acted to natural emotions.
- **MTL Benefits:** Jointly training SER with tasks like ASR has proven effective in learning more discriminative features.

6.2. Domain Adaptation Achievements

- **Adversarial Adaptation (DANN):** DANN-based approaches have been successful in reducing domain discrepancy for both cross-corpus and cross-lingual SER by learning domain-invariant features.
- **MMD-based Adaptation:** Minimizing MMD has also shown to be effective in aligning feature distributions across domains.
- **Combining Techniques:** Hybrid DA approaches, such as combining adversarial training with other regularization techniques or using them within more complex architectures, often yield further improvements.

6.3. Comparative Insights

- **DL vs. Traditional:** DL-based DA/TL methods generally outperform traditional machine learning approaches that rely on handcrafted features, especially in complex cross-domain scenarios.
- **Effectiveness of Pre-training:** Leveraging large pre-trained models (especially SSL speech models) as a starting point for TL or DA often provides a significant performance boost over training models from scratch on limited SER data.
- **Cross-Corpus vs. Cross-Lingual:** Cross-lingual SER remains more challenging than cross-corpus SER within the same language, highlighting the strong influence of linguistic cues on emotion expression and perception.
- **Importance of Target Domain Characteristics:** The degree of similarity between source and target domains significantly impacts adaptation success. Large discrepancies (e.g., acted vs. highly spontaneous emotions, very different languages) pose greater challenges.

While direct comparison across all studies is difficult due to variations in datasets, evaluation protocols, and specific model implementations, a clear trend emerges: DA and TL are indispensable tools for building SER systems that can generalize beyond their training data.

7. Discussion, Challenges, and Future Directions

7.1. Discussion of Key Findings

The review highlights that DA and TL are not just supplementary techniques but are becoming integral to developing robust SER systems. The ability of pre-trained SSL models to provide powerful, generalizable speech representations has been a game-changer, and PEFT methods make leveraging these large models more practical. Adversarial learning and discrepancy minimization techniques effectively address domain shift by promoting domain-invariant feature learning. However, no single DA or TL method is universally optimal; the choice often depends on data availability, the nature of domain discrepancy, and computational resources.

7.2. Persistent Challenges

Despite progress, several challenges remain:

- **Data Scarcity in Target Domains:** While UDA aims to work with unlabeled target data, performance often improves with at least some labeled target samples, which can be scarce or expensive to obtain.

- **Acted vs. Natural Emotions:** Bridging the gap between widely available acted emotion datasets and the nuances of real-world spontaneous emotions remains a significant hurdle. Generating or acquiring large-scale natural emotional speech data is difficult due to privacy and ethical concerns.
- **Label Space Mismatch:** Handling situations where emotion categories differ between source and target domains requires specialized techniques.
- **Computational Complexity:** Some advanced DA/TL methods, especially those involving large model fine-tuning or complex adversarial training, can be computationally intensive.
- **Explainability and Interpretability:** Understanding *why* certain DA/TL methods work and what features are being learned remains a challenge, especially with "black-box" deep learning models.
- **Negative Transfer:** In some cases, if the source and target domains are too dissimilar, TL or DA can lead to "negative transfer," where performance on the target domain degrades.
- **Evaluation Standardization:** Lack of standardized benchmarks and evaluation protocols for cross-domain SER makes direct comparison of different approaches difficult.
- **Fairness and Bias:** Ensuring that DA/TL techniques do not exacerbate or introduce biases related to gender, accent, or other demographic factors is crucial.

7.3. Future Research Directions

Future research in DA and TL for SER could explore several promising avenues:

- **More Sophisticated PEFT and Adapter Strategies:** Developing more efficient and effective PEFT methods tailored for SER and cross-domain adaptation.
- **Source-Free and Few-Shot Domain Adaptation:** Improving techniques that require minimal or no access to source data or only a few labeled target samples.
- **Continual Learning and Lifelong Adaptation:** Enabling SER models to continuously adapt to new domains and emotional expressions over time without catastrophic forgetting.
- **Leveraging Multimodality:** Integrating information from other modalities (e.g., text, video) in DA/TL frameworks for SER, as human emotion is inherently multimodal.
- **Causality-Inspired Domain Generalization:** Moving beyond correlation-based learning to understand causal factors in emotional expression for better generalization.
- **Personalized SER through Adaptation:** Developing techniques to quickly adapt general SER models to individual speaker characteristics and emotional expression styles.
- **Enhanced Explainability for DA/TL in SER:** Creating methods to better understand the decision-making process of adapted models and the nature of domain-invariant features.
- **Development of More Diverse and Realistic Benchmarks:** Creating new benchmark datasets and evaluation protocols that better reflect real-world complexities and facilitate fairer comparison of DA/TL methods.
- **Fair and Unbiased Adaptation:** Explicitly incorporating fairness constraints into DA/TL algorithms to ensure equitable performance across different demographic groups.

8. CONCLUSION

Domain Adaptation and Transfer Learning are indispensable for overcoming the critical challenge of generalization in Speech Emotion Recognition. The ability to leverage knowledge from existing large datasets and adapt models to new, diverse conditions is key to moving SER systems from controlled laboratory settings to robust real-world applications. This review has highlighted the significant progress made through various TL strategies, such as fine-tuning large pre-trained SSL models and employing parameter-efficient techniques, as well as diverse DA methodologies, including discrepancy-based and adversarial approaches. While challenges related to data scarcity, the acted-natural emotion gap, and computational costs persist, the ongoing research in these areas promises to yield even more powerful and versatile SER systems. By continuing to innovate in DA and TL, the field can move closer to achieving the goal of universally effective and reliable emotion recognition technology that can truly understand and respond to the rich spectrum of human emotions conveyed through speech.

ISSN 2322 - 0899

9. REFERENCES

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems, 33*, 12449-12460. (as cited in)

Barman, I. R., & Shanthini, A. (2023). Speech Emotion Recognition Using Deep Learning Algorithm. In *Emerging Trends in Signal and Data Analytics*. River Publishers.

Cai, X., Yuan, J., Zheng, R., Huang, L., & Church, K. (2021). Speech Emotion Recognition with Multi-Task Learning. *Proceedings of Interspeech 2021*, 4508-4512.

Deng, J., Xu, X., Zhang, Z., & Xu, M. (2023). Cross-Corpus Speech Emotion Recognition Based on Multi-Task Learning and Subdomain Adaptation. *Applied Sciences, 13*(2), 990.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F.,... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research, 17*(1), 2096-2030.

Ghosh, S., Arora, A., & Chaspari, T. (2023). MMER: Multimodal Multi-task Learning for Speech Emotion Recognition. *Proceedings of Interspeech 2023*.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 3451-3460.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Schuller, B. W., & Epps, J. (2022). Self Supervised Adversarial Domain Adaptation for Cross-Corpus and Cross-Language Speech Emotion Recognition. *arXiv preprint arXiv:2204.00431*.

Luo, H., & Han, J. (2019). Cross-Corpus Speech Emotion Recognition Using Semi-Supervised Transfer Non-Negative Matrix Factorization with Adaptation Regularization. *Proceedings of Interspeech 2019*.

Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning — A systematic review. *Intelligent Systems with Applications, 20*, 200266.

Mathur, A., Berthouze, N., & Lane, N. D. (2020). Unsupervised Domain Adaptation Under Label Space Mismatch for Speech Classification. *Proceedings of Interspeech 2020*.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering, 22*(10), 1345-1359. (as cited in)

Parry, J., Palaz, D., Clarke, G., Lecomte, P., Mead, R., Berger, M., & Hofer, G. (2019). Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition. *Proceedings of Interspeech 2019*, 1656-1660.

Schuller, B. (2018). Speech emotion recognition: Two decades in a nutshell. *IEEE Computational Intelligence Magazine, 13*(3), 30-38.

Xi, Y., Li, P., Song, Y., Jiang, Y., & Dai, L. (2019). Speaker to Emotion: Domain Adaptation for Speech Emotion Recognition with Residual Adapters. *Proceedings of APSIPA Annual Summit and Conference 2019*.