

On efficacy of Meta-Learning for Domain Generalization in Speech Emotion Recognition

Raeshak King Gandhi · Vasileios Tsouvalas · Nirvana Meratnia

Eindhoven University of Technology

PerFail · IEEE PerCom Workshops 2023 · pp. 421-426

01 Motivation

The problem

Speech emotion recognition (SER) accuracy often drops across speakers, languages and corpora. Can training across source domains improve performance on a target corpus that is entirely unavailable during training?

Contributions

- **Domain-shift weighting** adjusts the meta-test loss using an energy-score gap.
- **Seven transfer scenarios** compare conventional training, meta-learning and feature representations.
- **Negative-result analysis** identifies persistent barriers to practical cross-corpus emotion recognition.

Target stays unseen

Training uses source corpora only. The target corpus is reserved for evaluation, distinguishing domain generalization (DG) from adaptation using target-domain samples.

Why audio is difficult

Emotion corpora are small and differ in speakers, language and class coverage. These differences make cross-corpus comparisons difficult and can obscure shared emotional features.

Scope

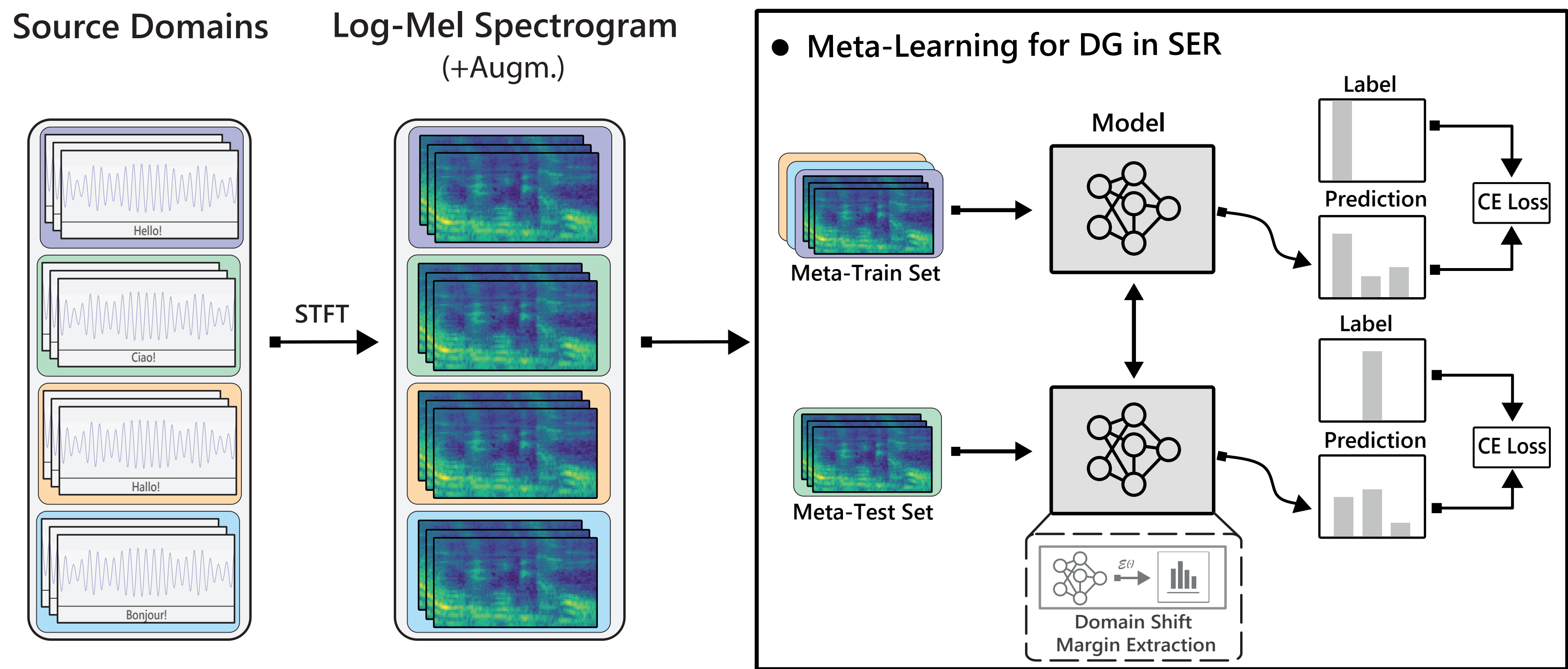
Gains remain uneven and insufficient for practical use. Small corpora, missing classes and unreported trial dispersion limit the strength of the comparisons.

Notation

domain-shift aware meta-learning for domain generalization (*DA-MLDG*) · short-time Fourier transform (STFT) · cross-entropy (CE)

02 DA-MLDG

Weight source-domain shifts to test unseen-corpus transfer



Technical core

$$\mathcal{L} = \mathcal{L}_{\theta}(S') + \gamma \mathcal{L}_{\theta'}(\mathcal{V})$$

S' and $\mathcal{V}$ are disjoint source subsets. θ' follows a meta-train update; γ controls the weight of its meta-test loss.

$$\gamma = |\mathcal{E}(S', p_{\theta'}) - \mathcal{E}(\mathcal{V}, p_{\theta'})|$$

$\mathcal{E}$ is the absolute mean log-sum-exp of temperature-scaled logits. A larger source-domain gap gives the meta-test loss more weight.



Paper



Code

03 Evaluation

4 / 7

above supervised Deep-All

Log-mel transfer scenarios

2 / 7

highest among four methods

Log-mel scenarios 4 and 5

Cross-domain emotion recognition

Scenario / target	Deep-All	RSC	MLDG	DA-MLDG (Ours)
1 / IEMOCAP	26.95	32.36	27.61	31.35
2 / RAVDESS	43.40	44.78	38.86	39.41
3 / TESS	47.03	43.77	46.29	46.80
5 / AESDD	60.65	56.69	60.45	62.26
6 / ShEMO	31.54	40.70	37.87	38.79

Held-out target accuracy (%), ten-trial mean; log-mel inputs with a VoxCeleb-pretrained ResNet-18.

Transfer with pretrained audio embeddings

Scenario / target	Deep-All (%)	DA-MLDG (%) (Ours)
1 / IEMOCAP	28.57	32.11
2 / RAVDESS	42.61	39.82
5 / AESDD	61.17	64.11
6 / ShEMO	33.16	40.22

Same source/target scenarios with TRILLsson embeddings; ten-trial mean target accuracy (%).

Takeaway

Transfer improves unevenly: *DA-MLDG* beats supervised Deep-All in four of seven log-mel scenarios and leads all four methods in two. With pretrained embeddings, it beats the supervised baseline in **only three of seven** scenarios.