ASA 1pSC.15 | Cross-Dataset Generalization for Speech Emotion Recognition

Dimitra Emmanouilidou, Hannes Gamper
Microsoft Research

Motivation

- Understand the challenges in cross-dataset generalization for audio-based SER models (what causes performance to drop on unseen data).
- Explore methodologies and techniques to improve generalization capabilities of SER models across domains
- Analyze the impact of large pre-trained models on emotion representations, including potential fragmentation of emotion classes in feature space
- Examine multi-corpora learning paradigms for constructing corpus-independent class representations
- Identify strategies that make SER models more robust and dataset-agnostic

Datasets

IEMOCAP	4-class Hap/Sad/Neu/Ang , eng, per-speaker split , Train {5,123}, Valid {542}, Test {500}
CREMA-D	4-class Hap/Sad/Neu/Ang , eng, random split, Train {3,960}, Valid {1,413}, Test {1,424}
MSP-Pod	4-class Hap/Sad/Neu/Ang , eng, Train vs Test 1, Train {50,494}, Valid {6,656}, Test {10,779}
EmoDB (g)	4-class Hap/Sad/Neu/Ang , germ, rand split, Train {192}, Valid {75}, Test {72}
RAVDESS	4-class Hap/Sad/Neu/Ang , eng, by-speaker, Train {576}, Valid {144}, Test {144}

Baselines Intra-Corpus

(linear Probe)

(IEMOCAP)		AUROC	ACC	F1	AvPrec
WavLM-L	*1024*	0.78	0.46	0.45	0.55
VGGISH	*128*	0.80	0.53	0.45	0.55
CDPAM-cont	*512*	0.75	0.45	0.37	0.47
Phi4	*1024*	0.85	0.61	0.55	0.62

(MLP 128x256)

(RAVDESS)		AUROC	ACC	F1	AvPrec
WavLM-L	*1024*	0.89	0.65	0.65	0.73
VGGISH	*128*	0.81	0.52	0.52	0.57
CLAP-2013	*1024*	0.88	0.64	0.65	0.75
Phi4	*1024*	0.81	0.42	0.41	0.57

(MLP 128x256)

(IEMOCAP)		AUROC	ACC	F1	AvPrec
Mel-Spec	*128*	0.77	0.44	0.39	0.50
WavLM-L	*1024*	0.87	0.66	0.58	0.69
CLAP-2013	*1024*	0.85	0.67	0.60	0.61
VGGISH	*128*	0.81	0.55	0.50	0.59
EnCodec	*128*	0.75	0.48	0.36	0.43
CDPAM-cont	*512*	0.79	0.52	0.49	0.55
DAC-44kHz	*1024*	0.76	0.56	0.48	0.48
Phi4	*1024*	0.86	0.67	0.61	0.75

(MLP 128x256)

(CREMA-D)		AUROC	ACC	F1	AvPrec
Mel-Spec	*128*	0.79	0.53	0.45	0.49
WavLM-L	*1024*	0.90	0.71	0.63	0.67
CLAP-2013	*1024*	0.88	0.67	0.59	0.63
VGGISH	*128*	0.82	0.53	0.44	0.50
EnCodec	*128*	0.82	0.58	0.49	0.53
CDPAM-cont	*512*	0.83	0.61	0.50	0.54
DAC-44kHz	*1024*	0.82	0.56	0.46	0.51
Phi4	*1024*	0.88	0.69	0.59	0.64

(MLP 128x256)

(EMO-DB)		AUROC	ACC	F1	AvPrec
WavLM-L	*1024*	0.96	0.81	0.81	0.91
CLAP-2013	*1024*	0.96	0.86	0.86	0.94
VGGISH	*128*	0.93	0.82	0.82	0.83
Phi4	*1024*	0.93	0.77	0.77	0.84

(MLP 128x256)

(MSP-POD)		AUROC	ACC	F1	AvPrec
WavLM-L	*1024*	0.78	0.53	0.43	0.49
CLAP-2013	*1024*	0.76	0.50	0.43	0.46
VGGISH	*128*	0.71	0.34	0.34	0.39
Phi4	*1024*	0.78	0.41	0.43	0.49

Findings: Class fragmentation in Multi-Corpora Learning

Multi-corpora train | Test on:	IEMO	MSP	CREMA-D	IEMO	MSP	CREMA-D
0% augm.	-0.01	0.00	-0.11	-0.02	0.01	-0.12
	0.00	0.03	-0.09	0.01	0.03	-0.12
	0.01	0.05	-0.08	0.00	0.04	-0.12
	0.00	0.05	-0.09	-0.02	0.05	-0.12
100% augm.	-0.01	0.04	-0.07	-0.09	0.04	-0.12
	NON - SCRAMBLED			SCRAMBLED		

← Acoustic mismatch evidently bridged using: SpecAugment, TimeStretching, TimeShifting, Fading, Equalization, Pink Noise, Reverberation

(same findings for both pre-trained embeddings and from-scratch CNN architectures)

Learning from **scrambled** labels -> perform. unaffected -> class fragmentation evidence

Methods

Hyper-param tuning: realistic, minimal on BS, LR and LR drop.

Online training, continuous learning, no epoch is the same, stop at AUROC plateau.

Embed frame masking and averaging augmentations, gaussian noise perturbations.

Same-label feature mixing -> augment by joining frames of same-class samples.

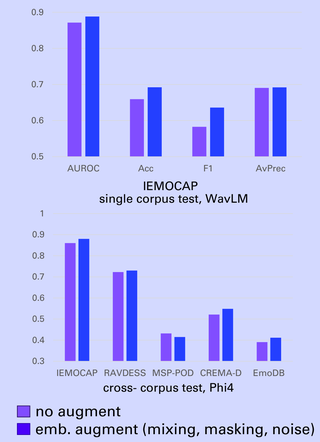
Class balancing by drawing classes at random, not files.

(test) Single Dataset Cross-corpus results

(train) Phi4, AUROC	IEMOCAP	RAVDESS	MSP-POD	CREMA-D	EmoDB
IEMOCAP	0.86	0.72	0.43	0.52	0.39
MSP-POD	0.37	0.39	0.78	0.60	0.66
CREMA-D	0.47	0.39	0.46	0.88	0.51

(test) Multi-corpora learning results

(train) AUROC	IEMOCAP	RAVDESS	MSP-POD	CREMA-D	EmoDB
WavLM 2D (i,c)	0.88	0.70	0.43	0.49	0.29
Phi4 2D (i,c)	0.87	0.68	0.43	0.47	0.45
WavLM 5D	0.83	0.71	0.36	0.59	0.31
Phi4 5D	0.86	0.73	0.40	0.51	0.40



Points of Interest

- Multi-corpora learning results in per-dataset class fragmentations despite acoustic matching.
- Models based on acoustic-centric embeddings were sensitive to parametrizations.
- Embeddings like CLAP, Phi4 were the least affected by network parameters.
- While validation perf seems unaffected, augmentations boosted within and cross-dataset test perf.
- Results on more datasets and embeddings in POMA paper.

References to checkout

- Embeds and FAD: https://github.com/microsoft/fadtk
- Phi4 multimodal : https://huggingface.co/microsoft/phi-4-gguf
- CLAP-2013: https://huggingface.co/microsoft/msclap
- EmoBox leaderboard https://emo-box.github.io/leaderboard1.html
- Emmanouilidou, D., Gamper, H., Yousefi, M. Domain mismatch and data augmentation in speech emotion recognition. MMSP at INTERSPEECH 2024: 10.21437/SMM.2024-5
- S. Braun and H. Gamper, "Multi-Label Audio Classification with a Noisy Zero-Shot Teacher," IWAENC 2024, doi: 10.1109/IWAENC61483.2024.10693989.

dimitra.emmanouilidou@microsoft.com