# Cross-Lingual Speech Emotion Recognition with Self-Supervised Models: A Confound-Controlled Comparison

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Speech emotion recognition across languages remains difficult because emotional cues interact with speaker identity, language, and recording conditions. emotion2vec, an emotion-specialized self-supervised learning (SSL) speech model, reports large gains over general-purpose SSL encoders such as HuBERT and WavLM across multiple non-English languages. This paper re-examines that claim under a confound-controlled cross-lingual evaluation. We compare HuBERT, WavLM, and emotion2vec on five emotional speech corpora spanning German, English, Mandarin, and Bangla, with an additional external test on Thai. Across speaker-independent probing, matched per-dataset evaluation, non-EmoBox generalization, and zero-shot/few-shot cross-lingual transfer, the general-purpose encoders consistently match or outperform emotion2vec. In speaker-independent four-class evaluation, emotion2vec ranks last on all five corpora, with significant gaps on four. Under cross-lingual transfer across 12 source-target language pairs, emotion2vec trails general-purpose SSL models by 12 to 18 percentage points in zero-shot transfer and remains about 10 points behind in the few-shot setting with 100 target-language examples per source-target pair. On Thai, a language outside the EmoBox fine-tuning distribution, the fine-tuned emotion2vec variant performs worse than both general SSL models and its own non-fine-tuned base version. We further show that the gap is not explained by speaker identity: after iterative null-space projection removes speaker-discriminative directions, HuBERT and WavLM remain ahead. These results suggest that emotion2vec, the emotion-specialized model whose published cross-lingual per-language results we re-examine, does not transfer across languages as reliably as general-purpose SSL. In practice, HuBERT and WavLM remain strong defaults, reaching 0.78 to 0.93 in-language accuracy on the four-class task and about 0.77 cross-lingual accuracy in the few-shot setting with only 100 target-language labels per source-target pair. Code is available at `https://anonymous.4open.science/r/ser-cross-ling-pub-536B`.

## 1 Introduction

A speech emotion recognition (SER) system trained to detect anger in English may have little to say about anger in Mandarin or Bangla, because the acoustic signature of every emotion shifts with the speaker's voice, the language being spoken, and the recording conditions. This is the cross-lingual SER problem, and the applications that need it (multilingual voice assistants, mental-health screening tools that ship across regions, call-centre analytics for global products) cannot accept a system that collapses on a new language. The dominant approach in recent years is to take a self-supervised speech model (pretrained on unlabeled audio), freeze its parameters, and train a small classifier on top using labeled emotion data (Hsu et al., 2021; Chen et al., 2022; Yang et al., 2021).

A more recent class of SSL backbone is pretrained on *emotion-rich* audio rather than general speech. To our knowledge, the only published example with cross-lingual per-language results is emotion2vec (Ma et al., 2024b). For each of nine non-English languages, the emotion2vec paper trains a separate emotion classifier on top of frozen emotion2vec and reports test-set accuracy on that language, finding gains of 25 to 45 percentage points over the same procedure with HuBERT or WavLM. The same paper introduces emotion2vec_plus

variants, downstream-tuned on the EmoBox cross-corpus benchmark (Ma et al., 2024a). These variants are now widely deployed in production SER pipelines through the FunASR ecosystem (Gao et al., 2023).

These per-language gains are striking for the setting emotion2vec tested, where each language has its own labeled training data. But cross-lingual SER in deployment often demands more: working on speakers never seen in training, working on languages outside the original fine-tuning distribution, and working when the target language has little or no labeled data (the practically important case for low-resource deployment). Under each of these conditions, emotion2vec's reported advantage proves protocol-dependent, with general-purpose SSL backbones (HuBERT, WavLM) consistently outperforming emotion2vec_base.

We conclude that for cross-lingual SER in practice, especially the low-resource case, general-purpose backbones remain the safer default. HuBERT and WavLM reach 0.78 to 0.93 in-language probe accuracy on the four-emotion task (angry, happy, neutral, sad), with emotion2vec 4 to 11 points behind on every dataset. In a low-resource setting (only 100 target-language labels per source-target pair), they reach a usable 0.77 cross-lingual accuracy, versus emotion2vec's 0.67. Our argument rests on four complementary lines of evidence:

1. **Per-language probing under speaker-independent splits** places emotion2vec_base third on every dataset, across our five corpora (German EmoDB, English RAVDESS and CREMA-D, Mandarin ESD-ZH, Bangla SUBESCO). Gaps to the best general SSL backbone are 4 to 11 percentage points and statistically significant on four of five corpora; only on SUBESCO (Bangla) is the three-way ordering within seed noise (Sec. 4.1).

2. **Per-dataset evaluation on the original benchmark datasets** places emotion2vec last by point estimate under both our speaker-independent protocol (no speaker overlap between train and test) and the published protocol (which we re-implement, Sec. 4.2). The most striking observation is on the baseline side. Our WavLM under the published protocol scores 36 to 51 percentage points higher than the WavLM baseline reported in the original paper, while our emotion2vec roughly matches theirs. The published gap therefore reflects baseline-implementation choices outside the emotion2vec model itself.

3. **Generalization to a non-EmoBox language** (Thai, Tai-Kadai family) shows that the fine-tuned emotion2vec_plus_base variant loses 7 to 10 points to general SSL and to its own non-fine-tuned counterpart, with multi-seed standard deviations under one percentage point (Sec. 4.3). The published `_plus` advantage does not transfer outside EmoBox.

4. **Cross-lingual zero-shot and few-shot transfer** across all 12 source-target language pairs places emotion2vec 12 to 18 points behind general SSL at zero-shot, and approximately 10 points behind at $K = 100$ target-language labeled examples per pair (Sec. 4.4). emotion2vec gains more raw percentage points from $K = 0$ to $K = 100$ than general SSL, but only because it starts further behind: at $K = 100$ it remains 10 points below.

A natural skeptic might ask whether emotion2vec's gap reflects how much speaker identity each model encodes rather than how well it captures emotion. If HuBERT and WavLM partly rely on speaker cues to predict emotion, removing those cues should close the gap. To test this, we apply Iterative Null-space Projection (Ravfogel et al., 2020), a technique that iteratively removes the directions in which the representation encodes speaker identity, and re-test emotion accuracy on the cleaned representation. The ranking does not change. HuBERT and WavLM still outperform emotion2vec, and all three models lose a comparable 14 to 16 percentage points of emotion accuracy after the projection (Sec. 4.5). If anything, the test favours emotion2vec, since its representations encode less linear speaker information than HuBERT or WavLM in the first place. Its lower emotion accuracy cannot be attributed to losing a speaker shortcut that only it relied on. To our knowledge this is the first application of iterative null-space projection to a frozen SSL speech encoder for SER.

We also ask a structural question: maybe emotion2vec is no more accurate but it at least organizes emotions in more language-portable ways. Cross-lingual representational similarity analysis (RSA) (Abdullah et al.,

2021) tests this by comparing how the four emotions are arranged geometrically in each language and asking whether the same arrangement carries over. All three models do preserve emotion structure across languages above chance, but no model is meaningfully better at it than the others (Appendix D).

A second diagnostic looks at which layers of each model the emotion classifier draws from. emotion2vec puts most of its weight on the convolutional input embedding (the very first layer, before any transformer block). HuBERT and WavLM, in contrast, peak at middle transformer layers in the pattern reported by prior SSL probing work (Appendix E). In other words, emotion2vec's transformer stack is contributing little to emotion prediction; the probe is relying mostly on the model's raw convolutional features.

Each finding is hardened against the obvious objections. The headline accuracy numbers are averaged over 5 random seeds rather than a single run, so the gaps cannot be dismissed as seed luck. We test both classifier head designs (the single linear layer we use by default and the two-linear ReLU head used in the emotion2vec paper) and find the choice does not account for the gaps. For the cross-lingual RSA test, where individual language-pair correlations are too noisy on their own, we use an aggregate permutation test that pools across pairs.

Another contribution of this work is methodological. The evaluation protocol we use here (speaker-independent splits, matched per-dataset re-implementation, a non-EmoBox external language, cross-lingual transfer, and INLP-based speaker confound control) is itself a reusable audit that any future emotion-aware SSL backbone claiming a cross-lingual advantage can be held to.

The rest of the paper is organized as follows. Section 2 positions our work against the closest prior work on SSL-based SER, cross-lingual SER, and post-hoc speaker confound removal. Section 3 describes the datasets, models, probing protocol, and statistical inference procedures. Section 4 reports the four arms of evidence and the supporting analyses. Section 5 discusses why the published advantage is protocol-dependent, what this means for SER practice, and concludes.

## 2 Related Work

### 2.1 SSL Backbones for SER and Emotion-Specialized Pretraining

Self-supervised speech representations have become the standard upstream for SER. HuBERT (Hsu et al., 2021) and WavLM (Chen et al., 2022) are the canonical general-purpose backbones, pretrained on speech with no emotion-specific objective and benchmarked on SER as part of SUPERB (Yang et al., 2021) and follow-up work.

To our knowledge, emotion2vec (Ma et al., 2024b) is the only published emotion-specialized SSL pretraining model with cross-lingual per-language probe results, which makes it the natural target for our evaluation. It applies online distillation on 262 hours of emotion-rich speech and reports per-language probe accuracy across nine languages in its Tables 3–4, claiming that emotion2vec_base outperforms WavLM-base by 25–45 percentage points on EmoDB, RAVDESS, and SUBESCO under random utterance-level 10-fold cross-validation. The fine-tuned `emotion2vec_plus` variants, whose teacher was supervised on the EmoBox cross-corpus benchmark (Ma et al., 2024a), are distributed through the FunASR ecosystem (Gao et al., 2023) and are a common off-the-shelf choice for SER. Other emotion-specialized SSL pretraining exists, notably Vesper (Chen et al., 2024), which adds an emotion-guided masking strategy on top of WavLM; it is evaluated only on English corpora (IEMOCAP, MELD, CREMA-D) and reports no cross-lingual per-language results, leaving emotion2vec as the one model carrying the cross-lingual claim we test.

Our work directly tests these claims under speaker-independent splits, on a non-EmoBox language, and under cross-lingual transfer, with an independent re-implementation of the probing protocol. We adopt the SUPERB-style frozen-encoder design used by Yang et al. (2021); Ma et al. (2024b), but probe the frozen features with the canonical SUPERB learnable weighted sum over all transformer layers and a single linear head, whereas emotion2vec's released downstream classifier uses the encoder's final-layer representation with a two-linear-with-ReLU head. Section 4.2 traces the resulting baseline gap to probe configuration rather than to the models themselves.

### 2.2 Cross-Lingual SER

Cross-lingual SER with SSL has previously been studied along a different axis from ours. Han et al. (2025) compares humans against *language*-specialized Wav2Vec2 variants (pretrained per language on general speech: Chinese, German, English) and WavLM-large, showing that PEFT methods (LoRA, Bottleneck Adapter) on the best layer recover most of the cross-lingual deficit. Their setup is orthogonal to ours: they test language-specialization with PEFT-tuned models, we test emotion-specialization with frozen-encoder probing. For this reason we do not include their language-specific Wav2Vec2 variants in our experiments, since that would mix two axes that we want to keep separate.

On the emotion-specialization axis, we compare HuBERT, WavLM, and emotion2vec across four EmoBox languages plus an external non-EmoBox test on Thai, using SUPERB-style weighted-sum probing with frozen encoders (Sec. 3).

Two methodological touchstones inform our analysis. Abdullah et al. (2021) introduces RSA on cross-lingual acoustic word embeddings; our adaptation to emotion-centroid distance matrices follows that template. Ma et al. (2024a) provides the EmoBox cross-corpus benchmark whose language coverage motivates our choice of Thai for the external test in Sec. 4.3.

### 2.3 Speaker Confound and Post-Hoc Representation Cleaning

Speaker invariance for SER has typically been pursued through adversarial training during model fitting. Tu et al. (2019) applies a gradient-reversal layer on top of a TDNN+Bi-LSTM encoder; Li et al. (2020) replaces gradient reversal with a max-entropy speaker loss; Yin et al. (2020) extends adversarial domain adaptation to a speaker-invariance term across multimodal cross-corpus emotion recognition; and Gat et al. (2022) performs speaker normalization on top of frozen HuBERT-base with gradient-ascent updates against a VoxCeleb-trained speaker classifier. All four require modification to the upstream encoder during training. We deliberately remain post-hoc on frozen encoders.

Two recent related lines work post-hoc on frozen encoders. Chiu et al. (2025) probes speaker, gender, and prosodic attributes across eleven SSL models without removing them, reporting that large-scale SSL models retain speaker-discriminative information even in late layers; this finding motivates why a confound-controlled probe matters, though they only diagnose speaker leakage rather than remove it. Zhu et al. (2025) does remove it, performing post-hoc speaker disentanglement of HuBERT and WavLM features for content-style downstream tasks (automatic speech recognition, voice conversion) using SHAP (SHapley Additive exPlanations) attribution-based perturbation, but for a different downstream target than SER.

The closest concurrent work to our INLP arm is SNAP (Jung et al., 2026), which projects WavLM-Large utterances onto the orthogonal complement of a top-$K$ PCA subspace fit to per-speaker centroids, evaluated on English speech deepfake detection. We share the high-level move of post-hoc orthogonal projection of a speaker subspace from frozen SSL features, but our approach differs on three methodological dimensions. First, INLP iteratively removes *all* linear speaker-discriminative directions (Ravfogel et al., 2020), while SNAP's top-$K$ PCA captures only the directions of highest centroid variance and may miss linear speaker signal lying outside that subspace. Second, we operate on a SUPERB weighted-sum representation in which the probe learns the layer mixture from data, rather than a hand-picked concatenation of WavLM-Large layers 8 and 22 (an empirical choice by SNAP). Third, we certify removal completeness by reporting pre- and post-projection linear speaker decodability against the 1/93 chance baseline, rather than silhouette score change. Our substantive setting (cross-lingual SER across four languages and 155 speakers) also differs from SNAP's English-only deepfake detection.

## 3 Methodology

### 3.1 Datasets

We use five publicly available emotional speech corpora spanning four languages from three language families (Indo-European, Sino-Tibetan, Indo-Aryan) for the cross-lingual analysis (Table 1). The corpora are among

Table 1: Emotional speech corpora used in the cross-lingual analysis. Counts are for the version used here. *ESD-ZH is subsampled from 17,500 at 100 utterances per (speaker, emotion) cell with a fixed seed.

| Dataset | Language | Utterances | Speakers |
|---|---|---|---|
| EmoDB (Burkhardt et al., 2005) | German | 535 | 10 |
| RAVDESS (Livingstone & Russo, 2018) | English | 1,440 | 24 |
| CREMA-D (Cao et al., 2014) | English | 7,442 | 91 |
| ESD-ZH* (Zhou et al., 2022) | Mandarin | 4,000 | 10 |
| SUBESCO (Sultana et al., 2021) | Bangla | 7,000 | 20 |

Table 2: Self-supervised speech encoders compared in this work. All have hidden dimension 768.

| Model | Layers | Pretraining |
|---|---|---|
| HuBERT-base-ls960 (Hsu et al., 2021) | 12 | masked-prediction on 960h LibriSpeech |
| WavLM-base (Chen et al., 2022) | 12 | masked-prediction + denoising on 960h LibriSpeech |
| emotion2vec_base (Ma et al., 2024b) | 8 | online distillation on 262h emotion-rich speech |

those on which emotion2vec's per-language advantage was originally reported (Ma et al., 2024b), so the comparison is directly testable. All five corpora are acted speech with the neutral emotion class present; this rules out corpora like ShEMO (naturalistic, would create a recording-style confound that masquerades as cross-lingual difference) and AESDD (no neutral class). Other plausible candidates were excluded for too few speakers (TESS, SAVEE) or restricted licensing (IEMOCAP, MSP-Podcast, MELD). All audio is resampled to 16 kHz mono.

The five corpora use different emotion taxonomies (7 classes in EmoDB, 8 in RAVDESS, 6 in CREMA-D, 5 in ESD-ZH, and 7 in SUBESCO), so cross-corpus comparison requires a common label set. We use the four-class set {angry, happy, neutral, sad}, the standard cross-corpus convention also used by EmoBox (Ma et al., 2024a). We therefore restrict the cross-lingual analyses (Sec. 4.1, 4.4, 4.5, and the RSA analysis of Appendix D) to this 4-class intersection. After filtering, the per-dataset 4-class subset sizes are 339 (EmoDB), 672 (RAVDESS), 4,900 (CREMA-D), 4,000 (ESD-ZH), and 4,000 (SUBESCO).

For the per-dataset evaluation (Sec. 4.2), we additionally use the full per-corpus taxonomies on EmoDB (7 classes), RAVDESS (8 classes), and SUBESCO (7 classes), matching the labels reported in Ma et al. (2024b).

To test whether emotion2vec's reported cross-lingual advantage transfers to a language not seen during EmoBox-derived fine-tuning, we add THAI-SER (Vistec-AI, 2021) (Thai, Tai-Kadai language family). THAI-SER contains 27,854 acted utterances from 200 speakers across 5 emotions with multi-annotator agreement scores. We retain utterances at agreement $\geq 0.5$ on the 4-class subset (excluding the THAI-SER-specific "frustrated" label), yielding 13,132 records. Thai is not among the 14 languages covered by EmoBox (Ma et al., 2024a), so it is an external test outside the EmoBox fine-tuning distribution for any emotion2vec variant fine-tuned on EmoBox-derived data.

### 3.2 Models

We compare three self-supervised speech encoders with hidden dimension 768 and broadly comparable parameter counts (Table 2). HuBERT and WavLM are general-purpose SSL backbones pretrained on speech with no emotion-specific objective. emotion2vec_base is, to our knowledge, the only publicly available emotion-specialized SSL backbone with cross-lingual per-language probe results, and is pretrained without supervised emotion labels. None of emotion2vec's pretraining corpora (IEMOCAP, MELD, MEAD, CMU-MOSEI, MSP-Podcast) intersect our test corpora, so emotion2vec_base is not directly trained on any of our test data.

We also include emotion2vec_plus_base, a fine-tuned variant whose pseudo-labels were generated by a teacher model (`_plus_seed`) directly fine-tuned on EmoBox (Ma et al., 2024a). We use it only as a control

on Thai (Sec. 4.3); on our other test corpora it is downstream-contaminated since EmoBox covers German, English, Mandarin, and Bangla.

All encoders are kept frozen throughout the experiments. Per-layer hidden states for HuBERT and WavLM are obtained from the official Hugging Face checkpoints (`facebook/hubert-base-ls960`, `microsoft/wavlm-base`) with the `output_hidden_states` flag enabled, yielding 13 hidden states per utterance (input embedding + 12 transformer outputs). For emotion2vec, the high-level FunASR API (Gao et al., 2023) returns by default a single mean-pooled vector across all layers; we bypass it and call the underlying model's `extract_features` method to obtain the per-layer outputs from `layer_results`, yielding 9 hidden states per utterance (input embedding + 8 transformer outputs). All hidden states are mean-pooled over the time axis and stored as $(N, L+1, 768)$ arrays per (model, dataset).

### 3.3 SUPERB-Style Weighted-Sum Probing

Following the SUPERB benchmark (Yang et al., 2021), we use a frozen-encoder probe with a learnable weighted sum over hidden states followed by a small classifier head, trained jointly. Concretely, given an utterance with $L+1$ mean-pooled hidden states $\{h_0, h_1, \ldots, h_L\}$ where $h_l \in \mathbb{R}^{768}$, the probe computes layer weights $\boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{w})$ for learnable parameters $\mathbf{w} \in \mathbb{R}^{L+1}$, the weighted-sum representation $z = \sum_{l=0}^{L} \alpha_l \, h_l \in \mathbb{R}^{768}$, and the prediction $\hat{y} = \mathrm{head}(z)$. Inputs are standardized per-feature using training-set statistics. The default head is a single linear layer $\mathbb{R}^{768} \to \mathbb{R}^{|C|}$. The emotion2vec paper (Ma et al., 2024b) instead uses a two-linear head with hidden dimension 256 and ReLU activation; we evaluate both choices in Sec. 4.2. Training uses Adam with learning rate $10^{-3}$, weight decay $10^{-4}$, and batch size 128 (256 for the larger Thai corpus). We train for 200 epochs, except for the few-shot transfer experiments where accuracy plateaus earlier in spot checks (50 epochs).

Throughout the cross-lingual analyses we use **speaker-independent splits**, where speakers are randomly partitioned with 60% in train and 40% in test, with all utterances inheriting their speaker's split. We use speaker-independent splits because accuracy on novel speakers is what matters in deployment for cross-corpus SER. Random utterance-level splits on small actor corpora leak speakers across train and test, which inflates absolute accuracy. We additionally follow Ma et al. (2024b)'s exact random utterance-level 10-fold protocol for the matched-protocol comparison in Sec. 4.2.

### 3.4 Cross-Lingual Transfer

For each (model, source language $\ell_s$, target language $\ell_t$) triple with $\ell_s \neq \ell_t$, we train a SUPERB probe on the source-language training set (60% of source-language speakers) and evaluate on the target language. Four languages give 12 ordered source-target pairs. We report two settings. **Zero-shot transfer** ($K = 0$) tests the source-trained probe directly on all target-language data. **Few-shot transfer** ($K \in \{10, 50, 100\}$) augments the training set with $K$ target-language examples sampled in a stratified way across the four emotions ($\lceil K/4 \rceil$ per class, with backfill if a class has fewer than that). The probe is retrained on this augmented set and evaluated on the held-out remainder of the target language, so the $K$ shots never overlap with the test set. Each $(\ell_s, \ell_t, K)$ cell uses 3 random samples of shots (deterministic for $K = 0$). The few-shot probes train for 50 epochs (verified to plateau by then in spot checks). Per-cell results are aggregated as mean accuracy across the 3 trials.

### 3.5 Confound-Controlled Probing via INLP

A standard concern with frozen-encoder SER probes is that test-set accuracy could partially reflect speaker identity rather than emotion. SSL embeddings are known to encode speaker information across all layers (Chiu et al., 2025), and SER probes might exploit speaker-emotion correlations in the training set. We address this by removing the speaker subspace from the SUPERB-pooled representation post-hoc, using Iterative Null-space Projection (INLP) (Ravfogel et al., 2020), and re-evaluating the emotion probe.

Intuitively, INLP repeatedly fits a linear speaker-identity classifier and projects the representation onto the orthogonal complement of the directions it uses, until speaker identity is no longer linearly decodable; we then retrain the emotion probe on this cleaned representation, so any surviving emotion accuracy cannot be

Table 3: 4-class per-language probe accuracy (mean over 5 seeds, speaker-independent splits). Best per row in bold. Per-seed standard deviations are 1.1–4.9% across all cells; full per-seed values in Table S3.

| Dataset (lang) | HuBERT | WavLM | emotion2vec_base |
|---|---|---|---|
| EmoDB (de) | 0.920 | **0.927** | 0.887 |
| RAVDESS (en) | **0.788** | 0.779 | 0.701 |
| CREMA-D (en) | 0.822 | **0.827** | 0.776 |
| ESD-ZH (zh) | **0.866** | 0.858 | 0.756 |
| SUBESCO (bn) | 0.823 | **0.833** | 0.815 |

a linear speaker shortcut. The speaker classifier is fit on 93 of the 155 speakers; the projection drives their speaker decodability to chance, and as a stricter check we also probe the 62 held-out speakers (Table 8). Full procedure in Appendix A.

### 3.6 Statistical Inference

For the headline experiments (per-language probing, Sec. 4.1; per-dataset speaker-independent, Sec. 4.2; and Thai non-EmoBox, Sec. 4.3), we retrain each (model, dataset) probe with 5 random seeds (`20260505` through `20260509`). The seed determines the speaker-split shuffle, the probe weight initialization, and the mini-batch ordering. We report mean accuracy with sample standard deviation across seeds, which captures training-time variance that test-set bootstrap alone cannot. For per-(model A, model B, dataset) accuracy comparisons in the headline tables, we additionally report paired Student $t$-test p-values across the five matched seeds (each seed produces a paired ($\text{acc}_A$, $\text{acc}_B$) observation under the same train/test split). The RSA permutation test is described in Appendix D.

## 4 Results

### 4.1 Per-Language Emotion Decodability (RQ1)

Table 3 reports 4-class accuracy on each of the five corpora under speaker-independent SUPERB-style probing (Sec. 3.3), aggregated over 5 random seeds. WavLM and HuBERT are essentially tied across all five datasets (paired $t$, all $p > 0.08$ for HuBERT-vs-WavLM on every dataset). Mean accuracy differences are 1–2 percentage points, with overlapping standard-deviation bands.

emotion2vec_base ranks third on every dataset by point estimate. The gap to the best general SSL backbone ranges from 1.8 percentage points (SUBESCO) to 11.0 (ESD-ZH). Paired $t$-tests across the five matched seeds confirm the gap is statistically significant on four of five corpora (HuBERT vs. emotion2vec on RAVDESS $p = 0.015$, EmoDB $p = 0.005$, CREMA-D $p < 0.001$, ESD-ZH $p < 0.001$; same comparisons against WavLM yield $p \leq 0.020$). On SUBESCO the gap to either general SSL is within training-noise ($p = 0.28$–$0.31$), so we report it as a three-way tie. The four significant gaps establish a real underperformance pattern.

The pattern is consistent. Emotion-specialized SSL pretraining does not provide an advantage over general-purpose SSL for cross-lingual SER on the 4-class intersection. We return to per-dataset full-taxonomy results (Sec. 4.2) and to the cross-lingual transfer setting (Sec. 4.4) where the gap widens.

### 4.2 Per-Dataset Evaluation under Speaker-Independent Splits (Claim 1)

We next match emotion2vec's published evaluation more closely. We re-run per-dataset full-taxonomy SER on EmoDB (7 classes), RAVDESS (8 classes), and SUBESCO (7 classes), the three corpora from Ma et al. (2024b)'s Table 4 where we also have the full taxonomies available. The published numbers report a 25–45 percentage-point advantage of emotion2vec over WavLM-base on these datasets (84.34 vs. 59.06 on EmoDB; 82.43 vs. 37.01 on RAVDESS; 90.91 vs. 54.50 on SUBESCO).

Table 4: 7/8-class per-dataset accuracy (mean over 5 seeds, speaker-independent splits). Best per row in bold; gap reported between best general SSL and emotion2vec_base. Per-seed standard deviations are 1.2–3.9%; full per-seed values in Table S4.

| Dataset | HuBERT | WavLM | emotion2vec_base | Gap |
|---|---|---|---|---|
| EmoDB (7) | **0.901** | 0.890 | 0.853 | −4.8 |
| RAVDESS (8) | **0.700** | 0.696 | 0.572 | −12.8 |
| SUBESCO (7) | 0.654 | **0.662** | 0.658 | −0.4 |

Table 5: 7/8-class per-dataset accuracy under Ma et al. (2024b)'s exact random utterance-level 10-fold CV protocol (mean ± std across folds). Best per row in bold. "Pub." values are the published WavLM and emotion2vec accuracies from Tables 3–4 of Ma et al. (2024b) (HuBERT-base is not reported there).

| Dataset | HuBERT (ours) | WavLM (ours) | emotion2vec (ours) | Pub. WavLM | Pub. emotion2vec |
|---|---|---|---|---|---|
| EmoDB (7) | $\mathbf{0.961 \pm 0.020}$ | $0.948 \pm 0.029$ | $0.888 \pm 0.062$ | 0.591 | 0.843 |
| RAVDESS (8) | $0.874 \pm 0.031$ | $\mathbf{0.881 \pm 0.026}$ | $0.750 \pm 0.029$ | 0.370 | 0.824 |
| SUBESCO (7) | $0.907 \pm 0.009$ | $\mathbf{0.912 \pm 0.009}$ | $0.884 \pm 0.012$ | 0.545 | 0.909 |

**Speaker-independent multi-seed result** Table 4 reports our per-dataset accuracy under speaker-independent splits, multi-seed (5 seeds), with the SUPERB protocol of Sec. 3.3.

emotion2vec_base ranks third on EmoDB and RAVDESS (paired $t$ vs. HuBERT: EmoDB $p = 0.010$, RAVDESS $p = 0.0015$; same comparisons against WavLM: $p \leq 0.011$). On SUBESCO the three models are tied within 1 point with no significant gap (paired $t$ $p \geq 0.39$).

The 25–45 point advantage of emotion2vec over general SSL reported in the original paper does not appear under speaker-independent evaluation on the two corpora with significant separation, and emotion2vec ties general SSL on the third. The RAVDESS gap is the strongest single signal. emotion2vec is 12.8 points behind HuBERT (mean across 5 seeds, paired $t$ $p = 0.0015$).

**Side-table: matched random-CV protocol** To address the obvious objection "you used a different cross-validation protocol, so the comparison is not apples-to-apples," we re-ran the same datasets and models under Ma et al. (2024b)'s exact random utterance-level 10-fold CV protocol (80/10/10 train/val/test per fold, no speaker grouping). Table 5 reports the result. Two observations stand out.

First, emotion2vec is the weakest model on every dataset under its own published protocol with our independent re-implementation. The accuracies are 88.80 vs. 96.09 (HuBERT) on EmoDB, 75.00 vs. 88.13 (WavLM) on RAVDESS, 88.41 vs. 91.24 (WavLM) on SUBESCO. Per-fold paired $t$-tests confirm the gap is statistically significant on every dataset against the best general SSL: HuBERT vs. emotion2vec $p = 1.8\times10^{-3}$ (EmoDB), $5.3\times10^{-7}$ (RAVDESS), $5.5\times10^{-4}$ (SUBESCO); WavLM vs. emotion2vec $p = 1.0\times10^{-2}$ (EmoDB), $1.7\times10^{-7}$ (RAVDESS), $1.4\times10^{-4}$ (SUBESCO). HuBERT and WavLM are statistically tied on every dataset (paired $t$ $p \geq 0.10$).

Second, our WavLM scores well above the WavLM baseline reported in the original paper under the same protocol: 94.78 vs. 59.06 on EmoDB, 88.13 vs. 37.01 on RAVDESS, and 91.24 vs. 54.50 on SUBESCO. Our emotion2vec, by contrast, roughly matches their reported value (within ±8 points; on RAVDESS our 75.0 sits below their 82.4, the one cell where the original paper extracts more from a model than we do, though our WavLM still exceeds it). A sweep of 192 WavLM-base probe configurations reproduces the published 25 to 45 point advantage in none of them and traces the gap to baseline configuration rather than to the models; the full sweep, the WavLM-base+ comparison, and a cross-check against the same group's later EmoBox benchmark are in Appendix C.

We further checked that the classifier head is not load-bearing for the gap. Sweeping the emotion2vec paper's two-linear-h256-ReLU head against our single-linear default across all nine (model, dataset) combinations

Table 6: Thai SER 4-class accuracy (mean ± std over 5 seeds, speaker-independent). Best in bold.

| Model | Accuracy |
|---|---|
| WavLM (general SSL) | $\mathbf{0.837 \pm 0.003}$ |
| HuBERT (general SSL) | $0.830 \pm 0.005$ |
| emotion2vec_base (clean) | $0.801 \pm 0.005$ |
| emotion2vec_plus_base (EmoBox-tuned) | $0.730 \pm 0.006$ |

Table 7: Cross-lingual transfer: mean accuracy across 12 ordered source-target language pairs at each $K$ (number of target-language examples added to source-language training). Each cell is the mean over 3 random shot samples (deterministic for $K = 0$); per-cell std across trials averages 1.5–2.1%. Gain is $K = 0 \rightarrow K = 100$.

| Model | $K = 0$ | $K = 10$ | $K = 50$ | $K = 100$ | Gain |
|---|---|---|---|---|---|
| HuBERT | 0.565 | 0.639 | 0.728 | 0.775 | +0.210 |
| WavLM | 0.512 | 0.630 | 0.722 | 0.771 | +0.259 |
| emotion2vec_base | 0.390 | 0.516 | 0.621 | 0.670 | +0.280 |

under speaker-independent splits changes accuracy by at most ±3 percentage points (mean −0.009, the larger head being slightly worse on average), so the head choice does not recover emotion2vec's deficit. Per-cell deltas are in Appendix B (Table S1).

Taken together, these results establish two things. The published emotion2vec advantage is protocol-dependent (it does not appear under speaker-independent splits). It is also implementation-dependent (it does not appear under their own protocol with our independent re-implementation). Neither head architecture nor cross-validation scheme is load-bearing for this conclusion.

### 4.3 Generalization to a Non-EmoBox Language: Thai (Claim 2)

A second line of evidence concerns the fine-tuned *plus* variants of emotion2vec, whose teacher (`_plus_seed`) was directly fine-tuned on EmoBox (Ma et al., 2024a). Since EmoBox covers German, English, Mandarin, and Bangla among its 14 languages, `_plus_base`'s reported per-language scores on those languages are downstream-contaminated. Thai is not in EmoBox, so it is an external test of whether the published *plus* advantage reflects genuine emotion specialization or fine-tuning to in-distribution corpora. Table 6 reports 4-class accuracy on the THAI-SER 13,132-record subset, multi-seed (5 seeds), speaker-independent.

emotion2vec_plus_base is the weakest of all four models on Thai, by a large margin. The gaps are 7.1 points below emotion2vec_base (the same architecture without fine-tuning), 9.9 points below HuBERT, and 10.7 points below WavLM. Multi-seed standard deviations are tight (0.3–0.7 percentage points), so the gap-to-std ratio for the `_plus_base`-vs-others comparison is 11–15$\sigma$. The gap is not training-noise; it is a real generalization failure.

This is the strongest single piece of evidence in our experiments that the published per-language advantage of fine-tuned emotion2vec variants reflects overfitting to the specific datasets (EmoBox) used during fine-tuning. It does not reflect a property of emotion-aware fine-tuning that transfers to genuinely novel languages. Note also the ordering of the unfine-tuned models on Thai: WavLM ≈ HuBERT > emotion2vec_base, the same ranking as on the four EmoBox languages from Sec. 4.1.

### 4.4 Cross-Lingual Transfer (Claim 3)

To test whether emotion-specialized SSL transfers better across languages with limited target-language supervision, we report few-shot transfer accuracy across all 12 source-target pairs (Sec. 3.4). Table 7 reports the mean accuracy across pairs at $K \in \{0, 10, 50, 100\}$, and Fig. 1 shows the corresponding curves.

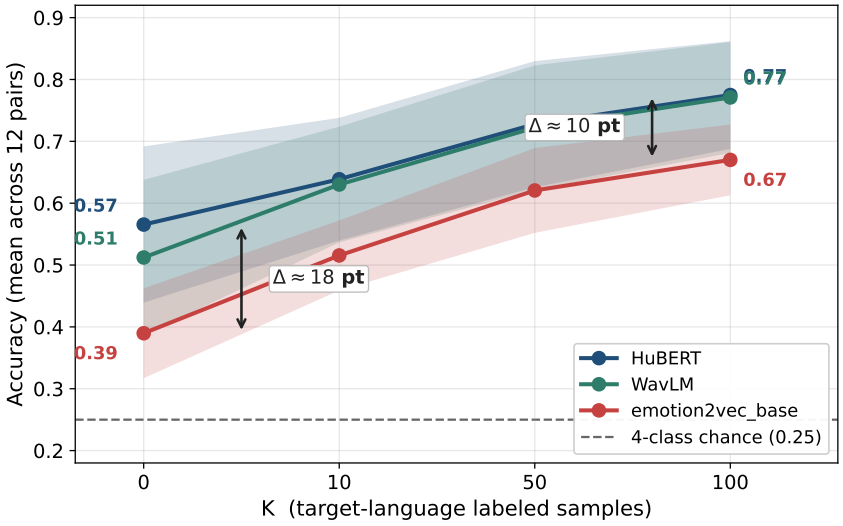


Figure 1: Cross-lingual K-shot transfer accuracy per model. Lines are means across the 12 source-target language pairs; shaded bands are ±1 std across pairs. Dashed line: 4-class chance (0.25). emotion2vec is consistently 10–18 points behind general SSL across all $K$.

This is the cleanest single signal in the paper. At zero-shot, emotion2vec is 12–18 points behind general SSL (mean 0.390 vs. 0.512 and 0.565), and it remains about 10 points behind HuBERT and WavLM at every larger $K$.

HuBERT and WavLM converge to nearly identical performance at $K = 100$ (0.775 vs. 0.771), approaching in-language probe accuracy (≈0.84–0.85, Table 3). emotion2vec at $K = 100$ is 0.670, leaving a residual cross-lingual gap. $K = 100$ target-language examples close approximately three-quarters of the cross-lingual deficit for general SSL and approximately seven-tenths for emotion2vec (0.706).

emotion2vec's absolute $K = 0 \rightarrow K = 100$ gain is in fact slightly larger than HuBERT's (+0.280 vs. +0.210), but only because it starts further behind. At $K = 100$ it covers a smaller fraction of its own in-language ceiling under the same supervision budget, and the residual gap to general SSL persists. This finding holds across all 12 pairs (no inversion) and is consistent with the per-language ranking from Sec. 4.1 and Sec. 4.2. Emotion-specialized SSL pretraining does not preserve cross-lingual emotion structure better than general SSL on these corpora.

To quantify uncertainty across language pairs rather than only across trials, we treat the 12 ordered source-target pairs as matched observations (every model is evaluated on the same pairs) and compute statistics over them. The spread across pairs is substantial but consistent: the across-pair standard deviation is 0.13 for both general SSL models and 0.07 for emotion2vec at $K = 0$, narrowing to 0.09 and 0.06 respectively at $K = 100$. A paired $t$-test across the 12 pairs confirms that emotion2vec sits below both general SSL models at every $K$ (all $p \leq 8.6 \times 10^{-4}$ against HuBERT and $p \leq 5.0 \times 10^{-4}$ against WavLM, with the smallest $p < 10^{-5}$), while HuBERT and WavLM are statistically indistinguishable across pairs at every $K$ ($p \geq 0.13$). The transfer gap is therefore not an artifact of a few easy or hard language pairs; it holds as a paired effect across all 12.

Because we retrain only a linear probe on frozen features, these few-shot numbers measure frozen-representation transfer; methods that adapt the encoder itself (the adversarial, domain-adaptation, and PEFT approaches in Sec. 2) can raise transfer accuracy further but answer a different question.

Table 8: Confound-controlled probing (155 speakers across 4 languages, 4-class emotion). INLP iteratively projects out the linear directions that predict speaker identity; columns marked *post* are measured after this projection (full procedure in Appendix A). *Emo*: emotion-probe accuracy. $Spk_{train}$ (post-INLP) and $Spk_{holdout}$ (reported both pre- and post-INLP): speaker-classifier accuracy on the 93 train-speakers and the 62 held-out speakers (70/30 within-holdout split); chance ≈ 0.011 and 0.016. *Lang.*, *Dataset*: linear probes on pre-INLP. *Iters*: INLP iterations. See Sec. 4.5 for discussion.

| Model | Emo (pre) | Lang. | Dataset | Iters | $\text{Spk}_{\text{train}}$ (post) | $\text{Spk}_{\text{holdout}}$ (pre) | $\text{Spk}_{\text{holdout}}$ (post) | Emo (post) |
|---|---|---|---|---|---|---|---|---|
| HuBERT | 0.807 | 1.000 | 1.000 | 11 | 0.053 | 0.958 | 0.667 | 0.662 |
| WavLM | 0.811 | 1.000 | 1.000 | 12 | 0.063 | 0.960 | 0.664 | 0.648 |
| emotion2vec | 0.779 | 1.000 | 1.000 | 10 | 0.051 | 0.821 | 0.455 | 0.616 |

### 4.5 Confound-Controlled Probing (RQ4)

To test whether the model differences in Sec. 4.1–4.4 are driven by speaker confound rather than emotion, we apply Iterative Null-space Projection (Sec. 3.5) to remove the speaker subspace from the SUPERB-pooled cross-lingual representation, then re-evaluate emotion decodability. As diagnostic comparators, we also report linear-probe accuracies for language and dataset on the same representation. Table 8 summarizes.

First, all three models near-perfectly encode language and dataset on the SUPERB-pooled representation; probing for either attribute reaches ≈ 1.000 accuracy (HuBERT, WavLM all-1.000; emotion2vec language 1.000, dataset 0.99982) even though the SUPERB probe was trained only for emotion. The cross-lingual representation is not language-blind.

Second, emotion accuracy drops by 14–16 percentage points across all three models after speaker INLP, but it remains well above the 4-class chance baseline of 0.25, confirming that the original emotion-decodability is not a speaker artifact.

Third, the general-vs-specialized ordering is preserved after speaker removal: HuBERT (0.662) and WavLM (0.648) both remain ahead of emotion2vec (0.616). The within-general HuBERT/WavLM ordering swaps on the point estimates (WavLM > HuBERT pre-INLP, HuBERT > WavLM post-INLP), but the swap is small relative to single-seed noise. emotion2vec is third pre and post; that is the robust finding.

Fourth, the result survives a stricter held-out test. Speaker identity is strongly encoded for the 62 held-out speakers before projection and drops by 29–37 percentage points after it, and the post-INLP emotion ranking (Table 8) is unchanged. Residual decodability on speakers the projection never saw is the expected behavior of a linear projection fit on a speaker subset rather than a shortcoming of the control (Appendix A).

emotion2vec is the most speaker-orthogonal of the three models post-INLP (0.455 vs. 0.667/0.664). As such, emotion2vec's lower emotion accuracy cannot be attributed to losing a speaker shortcut that only it relied on. Speaker confound substantially reduces absolute emotion accuracy across all three models, but does not explain the relative ordering between emotion-specialized and general SSL.

A naive reading of emotion2vec's lower speaker decodability would be favourable. An emotion-specialized model *should* discard irrelevant information like speaker identity, and emotion2vec apparently does so. But this reading does not survive joint consideration with the emotion numbers. A model that is more speaker-orthogonal *and* more emotion-accurate would be straightforwardly better. emotion2vec is more speaker-orthogonal *and* less emotion-accurate, both pre and post INLP. The two together are inconsistent with a clean story in which emotion2vec selectively keeps emotion information and discards speaker identity. Under both attributes we probed, emotion2vec carries less information than general SSL. We do not claim this generalizes to all attributes (paralinguistic dimensions we did not probe could in principle still favour emotion2vec), but it does mean the speaker-orthogonality cannot be invoked to explain away the lower emotion accuracy.

This interpretation dovetails with the layer-weight diagnostic of Appendix E, which shows that emotion2vec puts most of its probe weight on the pre-transformer convolutional input embedding for all five datasets and barely engages its transformer hierarchy. A representation that does not exploit its transformer stack

is at least consistent with reduced expressivity along multiple latent attributes, not selective emphasis on emotion.

## 5 Discussion and Conclusion

We have presented a confound-controlled cross-lingual comparison of self-supervised speech models for speech emotion recognition. Across four lines of evidence under speaker-independent SUPERB-style probing of frozen encoders, the published per-language advantage of emotion2vec, the emotion-specialized model we test, over general SSL backbones (HuBERT, WavLM) does not appear. The same ranking holds across per-language probing, matched per-dataset re-implementation, an external non-EmoBox language (Thai), and zero-shot and few-shot cross-lingual transfer, and removing the speaker subspace by INLP leaves it intact, so the gap is not a speaker artifact (Sec. 4.5).

Two factors explain the discrepancy with the published numbers. The base-model gap is implementation-dependent: under our re-implementation of the original random utterance-level 10-fold protocol all three backbones reach 88–96% on EmoDB with emotion2vec weakest, and the reported 25-to-45-point advantage appears in none of the configurations we swept, so it traces to an under-configured general-SSL baseline rather than to emotion2vec (Sec. 4.2); that protocol also leaks speakers on these small corpora, which is why we evaluate speaker-independent (Sec. 4.1). The `_plus` gains reflect EmoBox contamination, since the teacher is fine-tuned on EmoBox-derived data covering our four languages, and on Thai, outside EmoBox, the advantage disappears (Sec. 4.3); the concern is specific to this fine-tuning, since emotion2vec_base's own pretraining corpora do not overlap our test data. We do not claim the original numbers are wrong under their exact pipeline; we claim the advantage is sensitive to factors (CV scheme, baseline tuning, EmoBox overlap) a downstream user cannot control, and disappears under stricter evaluation.

For practice, general-purpose backbones remain the safer default for cross-lingual SER. Evaluate with speaker-independent splits rather than random utterance-level CV, which inflates absolute accuracy and obscures model differences; and prefer light target-language supervision to zero-shot transfer, where one hundred labeled examples per pair close roughly three-quarters of the cross-lingual gap and even ten close about a third. That the fine-tuned `_plus` variant does worse than its own base model on an external language cautions against assuming emotion-aware pretraining transfers.

Beyond the empirical finding, the evaluation protocol we use (speaker-independent splits, matched re-implementation, a non-EmoBox test, cross-lingual transfer, and INLP confound control) is a reusable audit for future emotion-aware SSL claims, and we hope its protocol-dependence framing encourages evaluation against speaker-independent splits and held-out languages by default.

## 6 Limitations

We test one emotion-specialized backbone, emotion2vec, the only such model with published cross-lingual per-language results, so our conclusions concern that model under frozen-encoder probing rather than emotion-aware pretraining as a category; methods that adapt the encoder (adversarial, domain-adaptation, PEFT; Sec. 2) answer a different question. The cross-lingual analyses use the four-emotion intersection common to all five corpora, a coverage rather than validity constraint since the full-taxonomy results (Sec. 4.2) preserve the ranking. Finally, all corpora are acted speech, so the absolute accuracies are likely optimistic for spontaneous in-the-wild speech.

## 7 Broader Impact Statement

Speech emotion recognition is deployed in sensitive settings (affective computing, mental-health screening, voice analytics) where it can encode demographic biases or be misused for surveillance. The speaker-confound analysis here addresses only linear speaker identity, not fairness across demographic groups, so cross-lingual SER systems should be treated as assistive signals under human oversight rather than autonomous decision-makers.

## References


Badr M. Abdullah, Iuliia Zaitova, Tania Avgustinova, Bernd Möbius, and Dietrich Klakow. How familiar does that sound? cross-lingual representational similarity analysis of acoustic word embeddings. In *Proc. BlackboxNLP Workshop*, pp. 407–419, 2021.

Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. A database of German emotional speech. In *Proc. Interspeech*, pp. 1517–1520, 2005.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans. on Affective Computing*, 5(4):377–390, 2014.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Weidong Chen, Xiaofen Xing, Peihao Chen, and Xiangmin Xu. Vesper: A compact and effective pretrained model for speech emotion recognition. *IEEE Transactions on Affective Computing*, 15(3):1711–1724, 2024.

Aemon Yat Fei Chiu, Kei Ching Fung, Roger Tsz Yeung Li, Jingyu Li, and Tan Lee. A large-scale probing analysis of speaker-specific attributes in self-supervised speech representations. *arXiv preprint arXiv:2501.05310*, 2025.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang. FunASR: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*, 2023.

Itai Gat, Hagai Aronowitz, Weizhong Zhu, Edmilson Morais, and Ron Hoory. Speaker normalization for self-supervised speech emotion recognition. *arXiv preprint arXiv:2202.01252*, 2022.

Zhichen Han, Tianqi Geng, Hui Feng, Jiahong Yuan, Korin Richmond, and Yuanchao Li. Cross-lingual speech emotion recognition: Humans vs. self-supervised models. In *Proc. ICASSP*, pp. 1–5, 2025.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Kyudan Jung, Jihwan Kim, Minwoo Lee, Soyoon Kim, Jeonghoon Kim, Jaegul Choo, and Cheonbok Park. SNAP: Speaker nulling for artifact projection in speech deepfake detection. *arXiv preprint arXiv:2603.20686*, 2026.

Haoqi Li, Ming Tu, Jing Huang, Shrikanth Narayanan, and Panayiotis Georgiou. Speaker-invariant affective representation learning via adversarial training. In *Proc. ICASSP*, pp. 7144–7148, 2020.

Steven R. Livingstone and Frank 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*, 13(5):e0196391, 2018.

Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, Jiaxin Ye, Xie Chen, and Thomas Hain. EmoBox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. In *Proc. Interspeech*, pp. 1580–1584, 2024a.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760, 2024b.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proc. ACL*, pp. 7237–7256, 2020.

Alexandra Saliba, Yuanchao Li, Ramon Sanabria, and Catherine Lai. Layer-wise analysis of self-supervised acoustic word embeddings: A study on speech emotion recognition. In *ICASSPW*, pp. 590–594, 2024.

Sadia Sultana, M. Shahidur Rahman, M. Reza Selim, and M. Zafar Iqbal. SUST Bangla emotional speech corpus (SUBESCO): An audio-only emotional speech corpus for Bangla. *PLOS ONE*, 16(4):e0250173, 2021.

Ming Tu, Yun Tang, Jing Huang, Xiaodong He, and Bowen Zhou. Towards adversarial learning of speaker-invariant representation for speech emotion recognition. *arXiv preprint arXiv:1903.09606*, 2019.

Vistec-AI. THAI-SER: Thai speech emotion recognition dataset. airesearch.in.th, 2021.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB: Speech processing universal PERformance benchmark. In *Proc. Interspeech*, pp. 3161–3165, 2021.

Yufeng Yin, Baiyu Huang, Yizhen Wu, and Mohammad Soleymani. Speaker-invariant adversarial domain adaptation for emotion recognition. In *Proc. ICMI*, pp. 481–490, 2020.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases and ESD. *Speech Communication*, 137:1–18, 2022.

Xiaoxu Zhu, Junhua Li, Aaron J. Li, Guangchao Yao, and Xiaojie Yu. Speaker disentanglement of speech pre-trained model based on interpretability. *arXiv preprint arXiv:2507.17851*, 2025.

## A Iterative Null-Space Projection: Full Procedure

This appendix details the INLP procedure summarized in Sec. 3.5.

We pool the four-language 4-class data (approximately 14,000 utterances, 155 unique speakers in total). For each model we train one global SUPERB probe (Sec. 3.3) on emotion targets with 80% of speakers in train. We take the weighted-sum representation $X \in \mathbb{R}^{N\times 768}$ on all utterances as the post-hoc analysis target. We then form a speaker-independent 60/40 split for the post-hoc analysis. The 60% of speakers (93 of 155) form the "speaker-train portion" on which the speaker classifier is fit. The remaining 40% (62 speakers) are reserved for the post-INLP emotion probe.

INLP proceeds iteratively. At each iteration $t$, we (i) standardize $X^{(t)}$, (ii) fit a multinomial logistic regression on the speaker-train portion of $X^{(t)}$ to predict speaker identity among the 93 train-speakers, (iii) compute the rank-revealing SVD of the classifier's weight matrix $W^{(t)} \in \mathbb{R}^{93\times 768}$ to obtain a basis $V^{(t)}$ for its row space, and (iv) update $X^{(t+1)} = X^{(t)} - X^{(t)}V^{(t)}V^{(t)\top}$, projecting onto the orthogonal complement of the speaker-discriminative subspace. For step (ii) we use `SGDClassifier` with log loss, $\alpha = 10^{-4}$, 30 epochs (lbfgs on 155 classes was prohibitively slow on our matrix sizes).

We iterate until the linear speaker-classifier accuracy on the speaker-train portion of the projected data falls below a conservative threshold of 0.056, well above the $1/93 \approx 0.011$ chance for the 93-class problem. We cap the iteration count at 12. The convergence criterion is measured on training data following Ravfogel et al. (2020), so we additionally report the post-INLP speaker accuracy in Table 8 to certify that the linear speaker-discriminative directions identified on the train-speaker subset have been removed. The number of iterations needed to reach the threshold differs slightly across models: 11 (HuBERT), 12 (WavLM), 10 (emotion2vec).

We then apply the same accumulated sequence of projections to the post-hoc 60/40 speaker-independent split (the 93 speakers used to fit the speaker classifier, plus the 62 reserved speakers). We retrain the emotion probe (a single linear layer on $X_{\text{cleaned}}$, since the weighted-sum has already been baked in) to obtain the post-INLP emotion accuracy. The 80/20 split used to train the original SUPERB probe (Sec. 3.3) is no longer in play once the weighted-sum representation $X$ has been computed. From this point all post-hoc analyses use the 60/40 speaker-independent split.

Because the speaker classifier is fit only on the 93 train-speakers, each INLP iteration removes the row-space of a $93 \times 768$ weight matrix, the directions that separate those speakers, and on the train-speakers this drives linear speaker decodability to chance (Table 8). The 62 held-out speakers are different individuals whose discriminative directions need not lie in that subspace, so a linear projection fit on a speaker subset is not expected to eliminate their speaker information: partial transfer to unseen speakers is the principled outcome, not a shortcoming of the control. We nonetheless probe the held-out speakers as a stricter check than the standard INLP report, which evaluates removal only on the data the projection was fit on. Before projection their speaker identity is strongly encoded (consistent with the SSL speaker-leakage reported by Chiu et al. (2025)), and the projection still cuts held-out decodability by 29–37 percentage points. The general-vs-specialized ranking and emotion2vec's status as the most speaker-orthogonal model hold under this stricter test, so the emotion gap cannot be a linear speaker shortcut. As diagnostic comparators, we also report linear-probe accuracies for language and dataset on the pre-INLP weighted-sum representation, computed under random 60/40 splits for that within-attribute view.

## B Head-Architecture Sweep

The emotion2vec paper uses a two-linear-layer head with hidden dimension 256 and ReLU activation, while our default is single-linear. Table S1 reports $\Delta$ accuracy (2-linear minus 1-linear) for each (model, dataset) under speaker-independent splits. All nine deltas are within $\pm 3$ percentage points. The mean across the nine combos is $-0.009$, meaning the larger head is on average slightly *worse*. The 1-vs-2-linear head choice is not load-bearing for the gap (Sec. 4.2).

Table S1: Head-architecture sweep: Δ accuracy (2-linear-h256-ReLU minus 1-linear), single seed, speaker-independent.

| Model | EmoDB | RAVDESS | SUBESCO |
|---|---|---|---|
| HuBERT | −0.009 | +0.008 | −0.002 |
| WavLM | −0.005 | −0.032 | −0.003 |
| emotion2vec | −0.014 | −0.018 | −0.004 |

## C Baseline-Configuration Sweep

To locate the source of the published baseline gap (Sec. 4.2), we re-ran the WavLM-base probe across 192 configurations, sweeping the learning rate ($10^{-3}$, $10^{-4}$), training budget (5 to 200 epochs), feature standardization, and layer aggregation (last layer or learnable weighted sum), under both our single-linear head and emotion2vec's two-linear-h256-ReLU head. The published 25 to 45 point advantage appears in none of them: WavLM-base matches or exceeds emotion2vec on EmoDB and RAVDESS at every setting, and on SUBESCO the two are close (emotion2vec leading by up to about 6 points under last-layer probing, WavLM ahead under the full weighted sum). The published low WavLM accuracy is reproducible only with a thin recipe (a low learning rate, short budget, and no standardization), which puts RAVDESS WavLM near 37 but also drives emotion2vec to near 29, far from the 82 they report, so no single configuration reproduces both published columns. The WavLM-base+ variant they report is only marginally higher (65.66, 38.89, 54.73), still far below our WavLM-base, so the choice of checkpoint does not explain the gap, and the same author group's later EmoBox benchmark (Ma et al., 2024a) independently scores WavLM-base at 62.1 on RAVDESS against 37.0 in the emotion2vec paper. Together these are consistent with a general-SSL baseline trained under a lighter configuration than the emotion2vec entries rather than with a property of the models.

## D Cross-Lingual Representational Alignment (RSA)

This appendix gives the methodology and results for the cross-lingual representational similarity analysis summarized in Sec. 1.

To compare emotion-similarity structure across languages within a single representation space, we follow representational similarity analysis (Abdullah et al., 2021). For each model, we train one *global* SUPERB probe on the pooled multilingual 4-class data (80% of speakers in train, all four languages combined). The probe's learned weighted sum projects every utterance to a single 768-D representation. By design, this is the task-adapted representation that the SUPERB probe has learned to use for emotion classification, not the raw encoder output. We are asking whether the representation each model actually deploys for SER preserves emotion structure across languages. We compute one centroid per (language, emotion) cell as the mean of utterance representations; with 4 emotions and 4 languages this yields 16 centroids per model.

Within each language we form a $4 \times 4$ cosine-distance matrix between emotion centroids. For each unordered pair of languages $(\ell_a, \ell_b)$, we compute the Spearman rank correlation between the upper triangles of the two matrices (6 entries each). Higher $\rho$ indicates that emotions are arranged similarly in the two languages' geometries. With 4 languages there are 6 unordered pairs.

For the statistical test, our primary test is an aggregate permutation test on the mean $\rho$ across the 6 language pairs. Per-pair $\rho$ values would be statistically underpowered, since the upper-triangle of a $4 \times 4$ emotion-distance matrix has only 6 entries and the permutation null on 6 ranks has standard deviation $\approx 0.45$. We therefore test the aggregate question: "Is the mean $\rho$ across the 6 language pairs greater than chance?" For each of 1,000 permutations, we shuffle utterance-level emotion labels independently within every language (preserving each language's marginal emotion distribution), recompute all 6 distance matrices and pair correlations on the SUPERB-pooled representation, and take the mean. The one-sided p-value is the fraction of null mean $\rho$ values at or above the observed mean $\rho$. Per-pair p-values are reported in the released per-pair results for diagnostic purposes only.

Table S2: Cross-lingual emotion-similarity preservation. Observed mean Spearman $\rho$ across 6 language pairs, with one-sided aggregate permutation test against shuffled labels (1,000 permutations).

| Model | Mean $\rho$ | Null mean $\pm$ std | p-value |
|---|---|---|---|
| HuBERT | $+0.514$ | $+0.000 \pm 0.185$ | 0.010 |
| WavLM | $+0.457$ | $-0.003 \pm 0.183$ | 0.022 |
| emotion2vec | $+0.457$ | $-0.014 \pm 0.176$ | 0.016 |

We complement the accuracy-based comparisons with this structural one. Even if emotion2vec is no more accurate than general SSL, does it organize emotions in more language-portable ways? Does the geometry of emotion centroids in one language match the geometry in another, after passing through the same SUPERB-pooled representation? Table S2 reports the observed mean Spearman $\rho$ across the 6 language pairs and the corresponding aggregate permutation test (1,000 permutations of utterance-level emotion labels per language).

All three models preserve cross-lingual emotion-similarity structure significantly above chance ($p \leq 0.022$ for all three), with mean $\rho$ in the 0.46–0.51 range. The three models are statistically indistinguishable on this measure. No model has a meaningful structural advantage. emotion2vec has no structural advantage on top of its accuracy disadvantage.

We deliberately do not lead with per-pair $\rho$ values. With $N = 6$ entries in each upper triangle, the per-pair permutation null has standard deviation $\approx 0.45$, so individual pair correlations of 0.4–0.7 are not distinguishable from chance under a per-pair test. As an exploratory observation, the German-Mandarin (de-zh) pair has the highest correlation in each of the three encoders ($\rho = 0.83, 0.83, 1.00$ for HuBERT, WavLM, emotion2vec respectively, with uncorrected per-pair $p \leq 0.042$). Under a per-encoder Bonferroni-6 correction across the six pairwise tests, this pair survives only for emotion2vec. Under a Bonferroni-3 correction across the three encoders for the same pair (treating de-zh as the question and asking whether multiple encoders agree on it), it survives for all three. We treat the de-zh observation as exploratory rather than as a confirmatory claim and report per-pair correlations in the released per-pair results for diagnostic purposes only.

## E Layer-Weight Diagnostic

The SUPERB probe learns one $\boldsymbol{\alpha}$ vector per (model, dataset) probe. Inspecting which layer dominates the weighted sum is informative about which layers the linear probe relies on in each encoder under this protocol. Fig. S1 shows the mean $\alpha$ curves across datasets (panel a) and the per-(model, dataset) heatmap of $\alpha$ values (panel b) for the 4-class probes from Sec. 4.1.

HuBERT and WavLM peak at middle transformer layers (argmax layer in $\{3, 4, 5, 6, 7\}$ across all five datasets), the standard pattern reported in prior SSL probing work (Saliba et al., 2024; Chiu et al., 2025) and visible as the mid-stack rise in Fig. S1a. Emotion signal accumulates as the encoder processes, peaks mid-stack, then linguistic content takes over in late layers. emotion2vec_base puts its highest weight on layer 0 (the pre-transformer convolutional input embedding) for all 5 datasets, visible as the bright row in Fig. S1b ($\alpha_0 = 0.53, 0.48, 0.43, 0.17, 0.14$ for SUBESCO, ESD-ZH, CREMA-D, RAVDESS, EmoDB respectively). The probe is learning that emotion2vec's transformer hierarchy contributes less to downstream emotion classification than its pre-transformer convolutional features. This dovetails with the accuracy results. If the emotion-aware transformer pretraining is not preserving more emotion signal than the input embedding, it is unsurprising that emotion2vec does not outperform general SSL whose transformer layers carry more discriminative information.

## F Per-Seed Probe Accuracy

This appendix reports the per-seed values behind the speaker-independent headline tables (Tables S3 and S4). Two patterns are worth noting. First, emotion2vec's deficit is consistent rather than seed-dependent: it trails

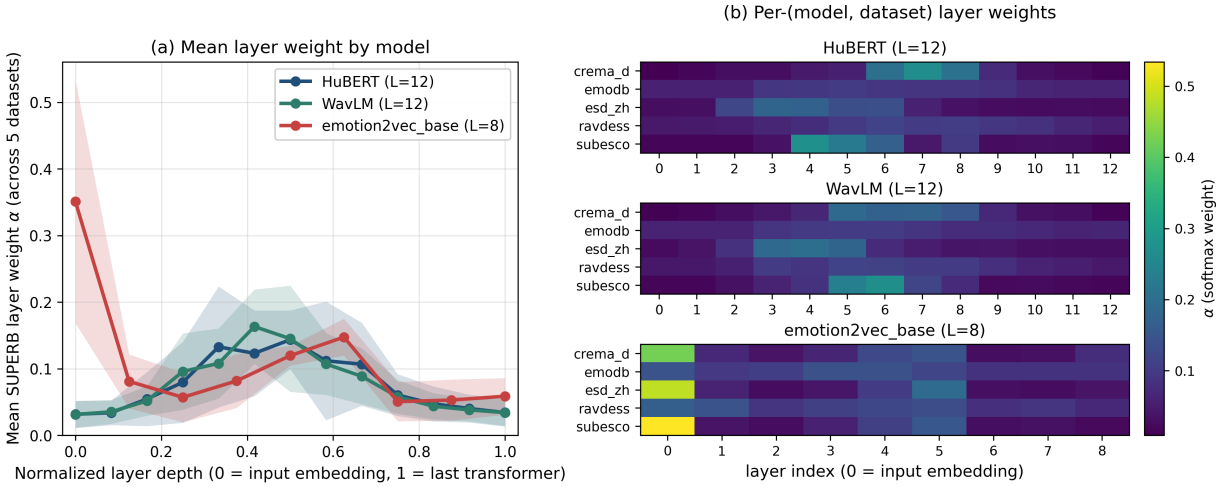


Figure S1: SUPERB-probe layer weights $\boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{w})$ for the 4-class RQ1 probes (Sec. 4.1). **(a)** Mean $\alpha$ across the five datasets vs. normalized layer depth (0 = input embedding, 1 = last transformer), with $\pm 1$ std bands. emotion2vec (red) places most of its weight on the input embedding and decays into the transformer stack; HuBERT and WavLM rise to a mid-stack peak in the canonical SSL pattern. **(b)** Per-(model, dataset) heatmap of $\alpha$ values; same colour scale across models. Layer 0 is the pre-transformer input embedding; layers 1–12 are transformer outputs for HuBERT/WavLM and layers 1–8 for emotion2vec. These weights indicate which layers the linear probe relies on under this protocol, not directly where emotion information is encoded in the network.

both general SSL models on every seed of EmoDB, RAVDESS, CREMA-D, and ESD-ZH, and is competitive only on SUBESCO, the dataset we report as a three-way tie (Sec. 4.1). Second, although accuracy swings noticeably from seed to seed, especially on SUBESCO and ESD-ZH, the three models swing together: a seed that is hard for one model is hard for all three (seed 20260506, for instance, lowers HuBERT, WavLM, and emotion2vec on SUBESCO at once). These swings therefore come from some seeds drawing harder speaker splits, not from one model being noisier than another. Because all three models are evaluated on the same split within a seed, comparing them seed by seed removes this shared swing, which is why we test the gaps with paired tests across seeds (Sec. 3.6) rather than from the seed means directly.

## G Reproducibility and Artifacts

We release artifacts to make every result reproducible and the evaluation reusable: the full analysis code (SUPERB-style probing, cross-lingual transfer, RSA, and INLP, together with the baseline-configuration sweep of Sec. 4.2); the cached per-layer embeddings for every (model, dataset) pair, so that each probe in this paper can be rerun without GPUs or the underlying audio; the exact speaker-independent split assignments and random seeds; and the per-seed tables behind every figure and table. The probing, transfer, and INLP code is packaged as a model-agnostic audit, so that a new self-supervised encoder can be dropped in and run through the same speaker-independent, cross-lingual, and confound-controlled evaluation we apply here. Links to the repository and a versioned archive of the embeddings and splits will be added in the camera-ready version.

Table S3: Per-seed 4-class probe accuracy for the per-language experiment (RQ1, Sec. 4.1); the seed means appear in Table 3. Columns 1–5 are seeds 20260505 through 20260509.

| Dataset | Model | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| EmoDB (de) | HuBERT | 0.928 | 0.946 | 0.924 | 0.906 | 0.897 |
| | WavLM | 0.942 | 0.946 | 0.924 | 0.899 | 0.923 |
| | emotion2vec | 0.906 | 0.905 | 0.879 | 0.862 | 0.880 |
| RAVDESS (en) | HuBERT | 0.761 | 0.807 | 0.764 | 0.839 | 0.768 |
| | WavLM | 0.789 | 0.757 | 0.761 | 0.818 | 0.768 |
| | emotion2vec | 0.707 | 0.732 | 0.643 | 0.689 | 0.732 |
| CREMA-D (en) | HuBERT | 0.802 | 0.840 | 0.815 | 0.831 | 0.822 |
| | WavLM | 0.815 | 0.834 | 0.825 | 0.847 | 0.812 |
| | emotion2vec | 0.760 | 0.788 | 0.770 | 0.779 | 0.783 |
| ESD-ZH (zh) | HuBERT | 0.876 | 0.907 | 0.828 | 0.902 | 0.818 |
| | WavLM | 0.890 | 0.879 | 0.866 | 0.871 | 0.782 |
| | emotion2vec | 0.785 | 0.788 | 0.711 | 0.798 | 0.698 |
| SUBESCO (bn) | HuBERT | 0.876 | 0.744 | 0.828 | 0.845 | 0.823 |
| | WavLM | 0.896 | 0.782 | 0.814 | 0.839 | 0.832 |
| | emotion2vec | 0.844 | 0.737 | 0.828 | 0.854 | 0.811 |

Table S4: Per-seed 7/8-class per-dataset probe accuracy under speaker-independent splits (Claim 1, Sec. 4.2); the seed means appear in Table 4. Columns 1–5 are seeds 20260505 through 20260509.

| Dataset | Model | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| EmoDB (7) | HuBERT | 0.894 | 0.920 | 0.915 | 0.888 | 0.887 |
| | WavLM | 0.890 | 0.903 | 0.893 | 0.893 | 0.870 |
| | emotion2vec | 0.881 | 0.876 | 0.844 | 0.846 | 0.819 |
| RAVDESS (8) | HuBERT | 0.718 | 0.693 | 0.712 | 0.707 | 0.670 |
| | WavLM | 0.732 | 0.643 | 0.693 | 0.723 | 0.688 |
| | emotion2vec | 0.603 | 0.603 | 0.530 | 0.557 | 0.565 |
| SUBESCO (7) | HuBERT | 0.703 | 0.604 | 0.653 | 0.662 | 0.648 |
| | WavLM | 0.702 | 0.613 | 0.684 | 0.659 | 0.655 |
| | emotion2vec | 0.705 | 0.603 | 0.675 | 0.668 | 0.640 |