C-TriLoRA: Cross-Corpus Kazakh SER via Tri-Factor LoRA and CORAL

Bakdaulet Kynabay[1]*, Aimoldir Aldabergen[2], Shirali Kadyrov[3]
Faculty of Engineering and Natural Sciences, SDU University, Almaty, Kazakhstan[1]
School of Sciences and Humanities, Nazarbayev University, Almaty, Kazakhstan[2]
Department of General Education, New Uzbekistan University, Tashkent, Uzbekistan[3]

***Abstract*—Speech Emotion Recognition (SER) in low-resource languages deals with the scarcity of labeled corpora and the instability of learned representations when transferred across diverse recording conditions and speaker demographics. This study introduces Conditional Tri-Factor Low-Rank Adaptation (C-TRILORA), a multi-task architecture that jointly performs automatic speech recognition (ASR) and SER on Kazakh speech while generalizing reliably across corpora. The proposed model extends a pre-trained Whisper encoder–decoder backbone through three primary innovations: a Tri-LoRA routing module that disentangles lexical, emotional, and speaker latent factors; CORAL domain alignment that matches second-order statistics between source and target domains without target labels; and a gradient reversal layer (GRL) that suppresses speaker-identity information. Experimental evaluations on the KazEmoTTS and ENU KEMO datasets demonstrate that C-TRILORA achieves a competitive in-domain Macro-F1 of 86.09% and significantly outperforms standard baselines in cross-corpus conditions (41.44% Macro-F1 versus 37.34% for the Shared-Head baseline). McNemar and Wilcoxon signed-rank tests confirm that explicit factor disentanglement is essential for cross-corpus robustness. These results show that separating speech components effectively mitigates negative transfer, making C-TRILORA a practical approach for low-resource SER deployment.**

***Keywords*—Speech emotion recognition; low-resource languages; parameter-efficient fine-tuning; domain adaptation; multi-task learning; disentangled representations**

I. Introduction

Over the past decade, deep learning has transformed speech emotion recognition, yet most of this progress has concentrated on a handful of well-resourced languages such as English and Mandarin [1], [2]. Languages with limited data infrastructure remain largely overlooked. Kazakh is one such case: a Turkic language with around 18 million speakers whose emotional speech has received little systematic attention, partly because publicly available, labeled corpora for it are scarce [3]. Without adequate training data, building SER systems that work reliably in Kazakh is difficult from the outset.

Beyond data scarcity, a pervasive challenge in deploying SER systems is corpus dependency, commonly referred to as the cross-corpus generalization problem. Acoustic representations learned by deep neural networks are notoriously brittle; they tend to conflate paralinguistic affective cues with domain-specific artifacts such as recording conditions, background noise, and speaker demographics [4], [5]. When a model trained on a specific source dataset is evaluated on an unseen target dataset, performance typically degrades significantly due to this covariate shift. Existing approaches to mitigate domain shift, such as adversarial domain adaptation and self-supervised learning, often treat the acoustic signal as a monolithic entity [6], [7], [8]. They generally fail to explicitly disentangle the underlying factors of speech, namely, lexical content, emotional prosody, and speaker identity, which limits their effectiveness in highly variable, cross-corpus scenarios.

To address these limitations, recent advancements have increasingly focused on mitigating the computational costs of large pre-trained models and the performance degradation caused by domain shifts in data-scarce environments. For instance, parameter-efficient fine-tuning (PEFT) techniques, particularly Low-Rank Adaptation (LoRA), have been successfully applied to foundation models like Whisper to extract robust emotional representations with minimal trainable parameters [9]. Furthermore, researchers have recently leveraged transfer learning and transformer-based paradigms to enhance cross-lingual and cross-corpus generalization [10]. However, aligning the feature distributions between mismatched source and target datasets remains a critical bottleneck. Consequently, contemporary cross-corpus SER frameworks are increasingly adopting advanced domain adaptation strategies, such as adversarial domain generalization [11] and classification inconsistency alignment [12]. These ongoing trajectories underscore the necessity of synergizing PEFT with unsupervised domain alignment and explicit factor disentanglement.

Within speech emotion recognition specifically, self-supervised transformer representations combined with lightweight fine-tuning have improved robustness in data-scarce conditions [13], and multi-task learning coupled with subdomain adaptation has been used to bridge the feature-distribution gap between mismatched corpora [14].

Motivated by these findings, this study develops C-TRILORA, a multi-task architecture for cross-corpus Kazakh SER. Rather than applying standard fine-tuning to a pre-trained Whisper backbone [15], we introduce a Tri-LoRA routing module that splits the shared speech representation into three separate streams:lexical, emotional, and speaker; each processed by its own low-rank adapter. This separation is the core design choice, aimed directly at the corpus dependency problem. The three main contributions of the work are:

- Architectural innovation: A Tri-LoRA routing module is integrated into a pre-trained Whisper backbone, explicitly decomposing shared representations into disentangled lexical, emotional, and speaker latent

*Corresponding author

factors.

- Unsupervised domain alignment: Correlation Alignment (CORAL) is integrated directly into the emotional latent space, successfully matching second-order statistics between source (KazEmoTTS) and target (ENU KEMO) domains without requiring labeled target data [3], [15]. This is coupled with a gradient reversal layer (GRL) to suppress speaker-identity leakage.
- Validation in a low-resource context: The efficacy of this approach is demonstrated on Kazakh speech, showing that explicit factor disentanglement significantly mitigates negative transfer and outperforms standard baseline models in cross-corpus SER tasks.

II. Related Work

A. Speech Emotion Recognition in Low-Resource Languages

Modern SER systems achieve strong results on richly resourced languages, supported by large corpora and by two decades of feature-engineering and benchmark work [1], [2]. For low-resourced languages, including Kazakh, progress is constrained by the absence of large labeled emotional corpora. The KazEmoTTS corpus [3] was released primarily for emotional text-to-speech rather than recognition, and recent multimodal efforts target low-resource settings explicitly [20]. A complementary direction reduces the labelled-data requirement through transfer learning and transformer architectures [10], [8]. These studies establish that low-resource SER is feasible, but remains sensitive to the conditions under which the training data were collected.

B. Cross-Corpus SER and Unsupervised Domain Adaptation

Cross-corpus SER addresses the realistic case in which training and test data are taken from different corpora with mismatched recording environments, speakers, languages, and elicitation styles. Classical solutions align feature distributions through subspace learning or transfer factorization, while deep approaches minimise an explicit domain discrepancy. Domain-adversarial training (DANN) [7] learns domain-invariant features through a gradient reversal layer, and Deep CORAL [16] aligns the second-order statistics of source and target representations. Self-supervised adversarial domain adaptation has further been applied jointly to cross-corpus and cross-language SER [25]. Recent transformer-based methods push this further: adversarial domain-generalised transformers [11], classification inconsistency alignment networks [12], multitask transformers that jointly model auxiliary tasks for cross-corpus transfer [22], source-free contrastive adaptation networks [27], and the mining of corpus-invariant emotional acoustic features [26]. A separate line of work disentangles speaker identity from affect through adversarial speaker-invariant representation learning [28], which directly motivates the speaker-suppression component used in this work. A recurring limitation of most of these methods is that they treat the acoustic embedding as a single monolithic vector and align it as a whole, without separating the lexical, affective, and speaker components that respond differently to domain shift.

C. Parameter-Efficient Fine-Tuning of Speech Foundation Models

Large self-supervised and weakly-supervised speech models such as wav2vec 2.0 [6], HuBERT [24], WavLM [23], and Whisper [17] provide powerful general-purpose representations, but full fine-tuning is expensive and prone to catastrophic forgetting in data-scarce regimes. Compact task-adapted variants such as Vesper [29], and parameter-efficient fine-tuning: Low-Rank Adaptation (LoRA) [18] in particular – instead update a small number of parameters while freezing the backbone, and have been shown effective for SER [9]. Existing PEFT-for-SER work, however, adapts the backbone with a single shared adapter and does not couple PEFT with explicit factor disentanglement or unsupervised cross-corpus alignment.

D. Research Gap

In summary, prior cross-corpus SER methods align a monolithic embedding, PEFT-SER methods use a single shared adapter without domain alignment, and low-resource SER work rarely evaluates cross-corpus transfer at all. The present work closes this gap by combining, within a single parameter-efficient model, three elements:

- A Tri-LoRA router that explicitly disentangles lexical, emotional, and speaker factors;
- CORAL alignment applied selectively to the emotional factor;
- Adversarial speaker suppression.

The resulting combination is validated on a genuinely low-resource language (Kazakh) under cross-corpus evaluation.

III. Methods

A. Problem Formulation

The fundamental objective of this study is to learn a robust emotion classifier capable of generalizing across different speech corpora without requiring target-domain labels during training. Let Eq. (1) denote the source corpus, where x_i represents a raw speech waveform, $y_i^e \in \{0, \ldots, E-1\}$ is the emotion label, y_i^t is the word-level transcript, and y_i^s is the speaker identity. Concurrently, let Eq. (2) denote the unlabelled target corpus utilized exclusively for domain alignment. The goal is to optimize a classification function, Eq. (3), that maintains high predictive accuracy on both the source domain D_S and the unseen test partition of the target domain D_T.

$$D_S = \{(x_i, y_i^e, y_i^t, y_i^s)\}_{i=1}^{N} \tag{1}$$

$$D_T = \{x_j\}_{j=1}^{M} \tag{2}$$

$$f : x \longmapsto y^e \tag{3}$$

B. Backbone Architecture: Kazakh Whisper

The proposed Conditional Tri-Factor Low-Rank Adaptation (C-TRILORA) architecture, as shown in Fig. 1, utilizes a pre-trained encoder–decoder backbone based on the Whisper architecture, which was originally developed as a robust, large-scale multilingual speech recognition system trained on weakly

supervised data [17]. Specifically for this study, the model is initialized using a Kazakh-adapted Whisper-small checkpoint [15], which was previously fine-tuned on 335 hours of Kazakh read speech. The encoder module, denoted as $\mathcal{E}$, maps an input log-Mel spectrogram $X \in \mathbb{R}^{80\times T}$ into a sequence of hidden states $H = \mathcal{E}(X) \in \mathbb{R}^{T'\times d}$, where the hidden dimension $d = 768$. To interface with the subsequent routing modules, a mean-pooled representation $\bar{h}$ is computed across the temporal dimension, as in Eq. (4):

$$\bar{h} = \frac{1}{T'}\sum_{t=1}^{T'} H_t \tag{4}$$

Whisper was chosen over encoder-only self-supervised models such as wav2vec 2.0 [6], HuBERT [24], and WavLM [23] for three reasons. First, it is an encoder–decoder model, so it natively supports the auxiliary ASR objective that grounds the lexical factor; encoder-only backbones would require an externally attached decoder for sequence-to-sequence transcription. Second, Whisper is trained on large-scale weakly-supervised, multi-condition data [17], which yields representations that are comparatively robust to the channel and noise variation that characterises cross-corpus evaluation. Third, and decisively for the low-resource setting, a Kazakh-adapted Whisper-small checkpoint fine-tuned on 335 hours of Kazakh speech is publicly available [15]; no comparable in-language checkpoint exists for the alternative backbones, so initialising from Whisper provides a substantially stronger starting point for Kazakh.

C. Parameter-Efficient Fine-Tuning and Tri-LoRA Routing

To adapt the Whisper backbone efficiently without catastrophic forgetting or excessive computational overhead, Low-Rank Adaptation (LoRA) is applied to all attention projection matrices (W_Q, W_K, W_V, W_O) in both the encoder and the decoder [18]. For any given frozen pre-trained weight matrix W_0, the adapted weight W' is computed as in Eq. (5):

$$W' = W_0 + \Delta W = W_0 + \frac{\alpha}{r} BA \tag{5}$$

where, $A \in \mathbb{R}^{r\times d}$ and $B \in \mathbb{R}^{d\times r}$ are trainable low-rank matrices, $r = 16$ is the rank, and $\alpha = 32$ is a scaling factor. This parameter-efficient strategy reduces the trainable parameter count to approximately 0.31% of the total backbone. The routing module and the three auxiliary heads add only a small number of further trainable parameters on top of the LoRA adapters; a detailed profiling of the additional training-time and inference overhead they introduce is identified as a direction for future work.

The core architectural innovation of this framework is the Tri-LoRA Routing Module. Instead of relying on a single shared embedding, this module decomposes the pooled representation $\bar{h}$ into three semantically specialized, disentangled latent factors: the lexical factor z_{lex}, the emotional factor z_{emo}, and the speaker factor z_{spk}. Each factor is processed through an independent low-rank adapter stream gated by learned soft routing weights, ensuring that the model explicitly isolates affective cues from confounding linguistic and demographic variables.

D. Unsupervised Domain Alignment and Optimization

To directly mitigate corpus dependency and domain shift, Correlation Alignment (CORAL) is integrated into the emotional latent space [16]. The CORAL objective minimizes the Frobenius-norm distance between the covariance matrices (second-order statistics) of the source and target z_{emo} distributions. Furthermore, to suppress the leakage of speaker-identity information into the emotion representations, a Gradient Reversal Layer (GRL) is applied to z_{emo} during adversarial training [7].

The total training objective $\mathcal{L}$ is formulated as a weighted sum of five distinct loss components:

- Automatic Speech Recognition Loss ($\mathcal{L}_{\text{ASR}}$): Cross-entropy loss from the Whisper decoder to ground z_{lex} in lexical semantics.
- Speech Emotion Recognition Loss ($\mathcal{L}_{\text{SER}}$): Weighted cross-entropy for emotion classification.
- Speaker Classification Loss ($\mathcal{L}_{\text{SPK}}$): Cross-entropy to explicitly capture speaker traits in z_{spk}.
- Counterfactual Loss ($\mathcal{L}_{\text{CF}}$): A margin-based penalty enforcing the separation of lexical and emotional variations for in-batch pairs sharing transcripts but differing in emotion.
- CORAL Loss ($\mathcal{L}_{\text{CORAL}}$): The unsupervised covariance alignment term.

After training, post-hoc temperature scaling is applied to the output logits to optimize probabilistic calibration and reduce the Expected Calibration Error (ECE) [19].

E. Speech Emotion Datasets

To evaluate the proposed C-TRILORA architecture, particularly its cross-corpus generalization capabilities, this study utilizes two distinct Kazakh speech emotion corpora:

- KazEmoTTS (Source Domain): This dataset serves as the primary labeled source domain (D_S) for training. It consists of high-quality, studio-recorded Kazakh emotional speech originally developed for text-to-speech applications [3], and comprises six emotion categories (neutral, angry, happy, sad, fear/scared, and surprise).
- ENU KEMO (Target Domain): This dataset serves as the unlabeled target domain (D_T) for the unsupervised Correlation Alignment (CORAL) and is subsequently used to evaluate cross-corpus performance. ENU KEMO features different acoustic conditions and speaker demographics compared to KazEmoTTS, providing a robust testbed for covariate shift [20].

F. Experimental Setup and Evaluation Metrics

All acoustic inputs were standardized. Raw audio waveforms were resampled to 16 kHz, and 80-dimensional log-Mel spectrograms were extracted using a 25 ms window and a 10 ms hop size to match the input requirements of the Whisper encoder.

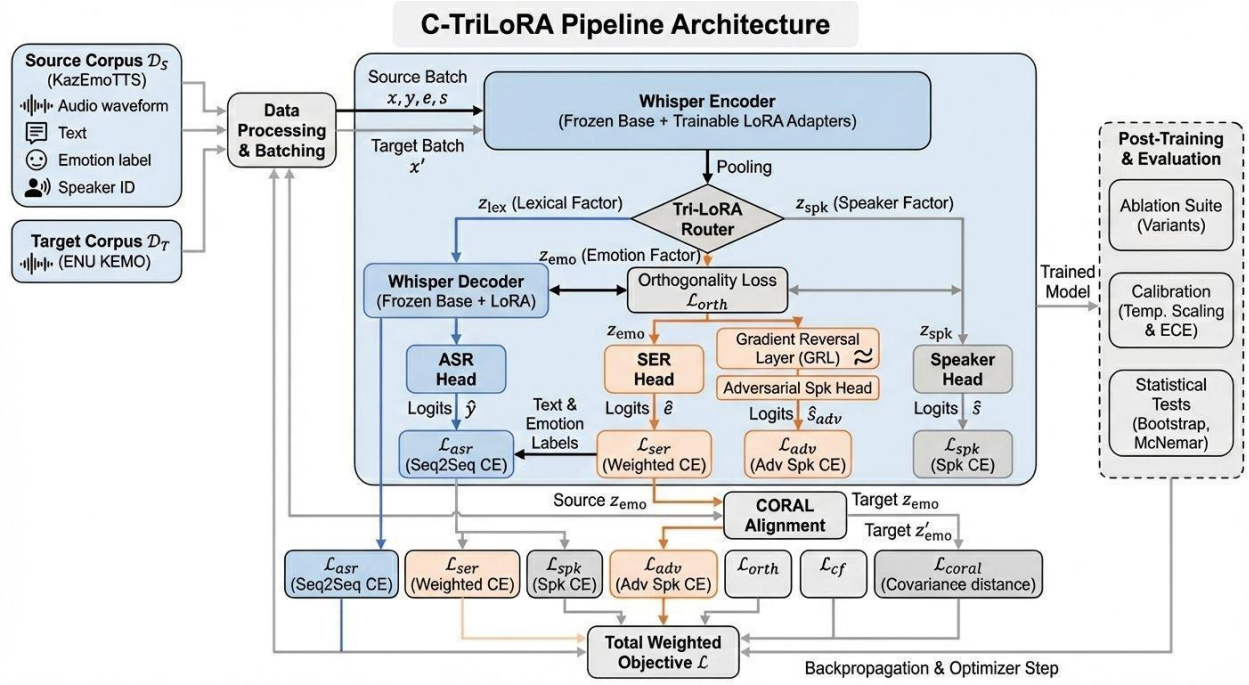


Fig. 1. The proposed C-TRILORA architecture illustrating the Tri-LoRA routing module for factor disentanglement and the CORAL module for domain alignment.

To rigorously assess the performance of the emotion classification, the primary evaluation metric used is Unweighted Average Recall (UAR). UAR is standard in speech emotion recognition tasks because it prevents majority-class bias in imbalanced datasets by calculating the accuracy independently for each emotion class and then averaging the results. Macro-F1 is reported alongside UAR throughout, as it likewise weights all emotion classes equally while being sensitive to false positives.

Additionally, to validate the statistical significance of the performance improvements yielded by the Tri-LoRA routing and CORAL components, two statistical tests were employed: *McNemar's test*, used to assess the statistical significance of differences in classification outcomes between the baseline and proposed models; and the *Wilcoxon signed-rank test*, applied to compare the paired differences in model performance across different experimental folds [21].

All models were trained using the AdamW optimizer with a learning rate of 1×10^{-4}, linear warmup over the first 10% of steps, and weight decay of 0.01, for 5 epochs with a batch size of 16 on a single NVIDIA A100 GPU (40 GB) using mixed-precision (fp16). The LoRA rank was $r = 16$ with scaling factor $\alpha = 32$. Multi-task loss weights were set empirically: $\lambda_{\text{ASR}} = 0.3$, $\lambda_{\text{SER}} = 1.0$, $\lambda_{\text{SPK}} = 0.5$, $\lambda_{\text{CF}} = 0.2$, and $\lambda_{\text{CORAL}} = 0.1$. KazEmoTTS was split 80/10/10 for training, validation, and testing; ENU KEMO was used exclusively for unsupervised domain alignment and cross-corpus evaluation.

IV. Results and Discussion

A. In-Domain and Cross-Corpus Classification Performance

The primary objective of this study was to evaluate the efficacy of the proposed C-TRILORA architecture under cross-corpus evaluation scenarios, where the model is tested on unseen target data (ENU KEMO) after being trained exclusively on the source domain (KazEmoTTS). Table I presents the classification performance across all model variants. Furthermore, the overall training stability and the convergence of the distinct multi-task loss components over the five epochs are illustrated in Fig. 2.

On the in-domain KazEmoTTS test set, the C-TRILORA model achieved strong performance with a Macro-F1 score of 86.09%. While the standard Shared Head baseline slightly outperformed the proposed model in-domain (Macro-F1 $=$ 89.69%), this is a well-documented phenomenon in representation learning. The Shared Head overfits to the joint statistics of the source domain's recording conditions and actor pool. Consequently, its representations are optimal for in-domain discrimination but highly fragile under distributional shift. Conversely, in the cross-corpus evaluation on ENU KEMO, C-TRILORA demonstrated superior robustness, achieving a Macro-F1 of 41.44% and significantly outperforming the Shared Head baseline (37.34%).

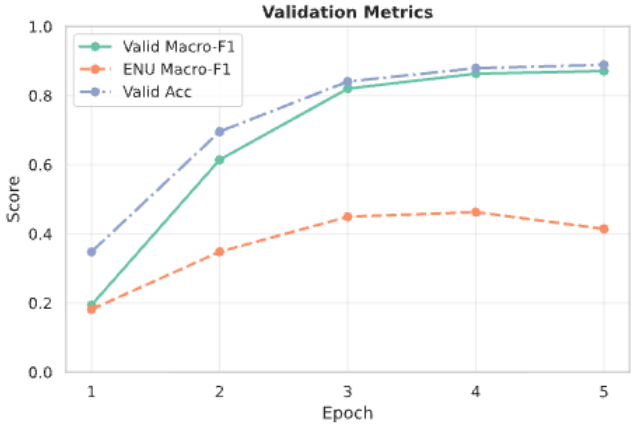
This confirms that explicitly disentangling the lexical, emotional, and speaker factors sacrifices a marginal amount of in-domain specificity in exchange for a substantial increase in cross-domain transferability. This preservation of class separability and domain-invariant clustering under covariate shift is visually corroborated by the t-SNE projections of the emotional latent space (z_{emo}) presented in Fig. 3.

B. Ablation Study and the Role of Domain Alignment

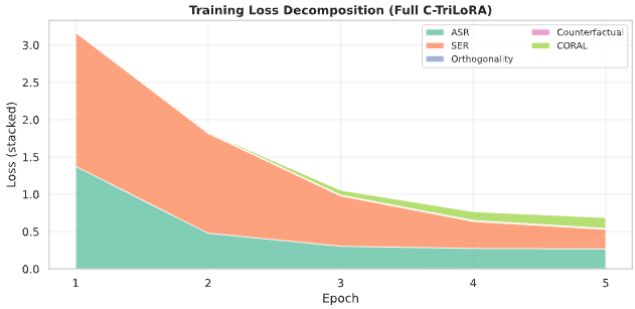
To isolate the contribution of each architectural component, an extensive ablation study was conducted. As illustrated in Fig. 4, the removal of any single component resulted

TABLE I. MACRO-F1 CLASSIFICATION PERFORMANCE OF THE C-TRILORA ARCHITECTURE AND ITS ABLATION VARIANTS.

Model Configuration	KazEmoTTS (In-Domain)	ENU KEMO (Cross-Corpus)
Shared Head (Baseline)	0.897	0.374
Full C-TRILORA (Proposed)	0.861	0.414
SER Only	0.722	0.255
No Orthogonality	0.638	0.264
No CORAL	0.615	0.211
No GRL	0.606	0.256
No Counterfactual	0.603	0.264
CORAL Only	0.578	0.239
ASR Only	0.057	0.051



(a) Validation metrics (accuracy and Macro-F1) over 5 epochs on the in-domain and cross-corpus sets.

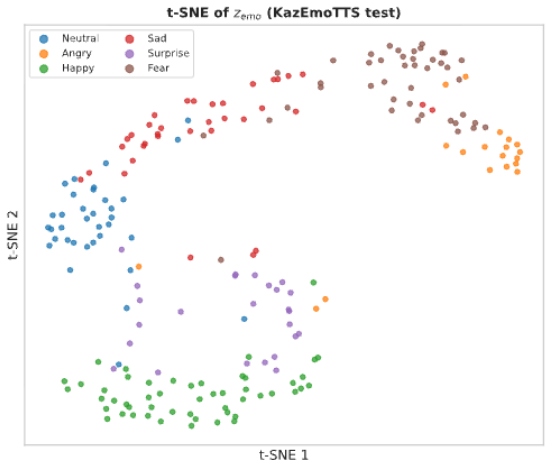


(b) Stacked decomposition of the multi-task training loss (ASR, SER, Orthogonality, Counterfactual, CORAL).

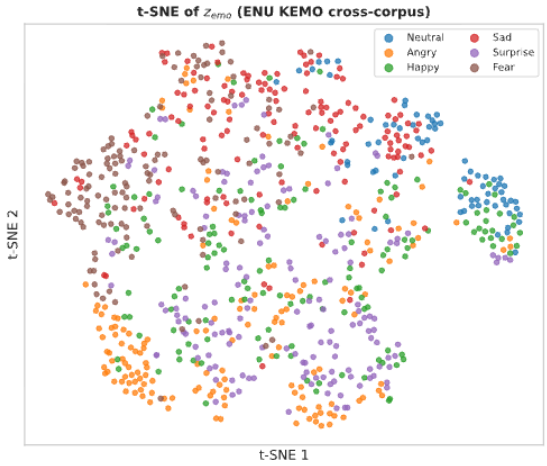
Fig. 2. Training and validation progression of the C-TRILORA architecture over 5 epochs.

in a degradation of cross-corpus performance. To rigorously validate these performance drops, statistical significance was assessed using McNemar's test. As shown in the heatmap in Fig. 5, the cross-corpus performance degradation observed in all ablation variants is statistically significant ($p < 0.05$) when compared to the full C-TRILORA architecture.

Most notably, as visualized in the waterfall chart in Fig. 6, the removal of the Correlation Alignment (CORAL) objective caused a dramatic 17.5 to 19.3 percentage point absolute drop



(a) In-domain (KazEmoTTS test)



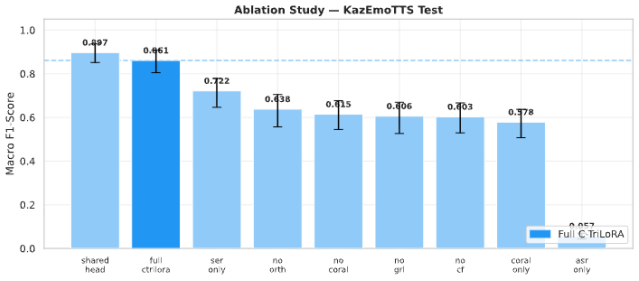
(b) Cross-corpus (ENU KEMO)

Fig. 3. t-SNE visualizations of the emotional latent space z_{emo}, illustrating the preservation of class separability under domain shift.

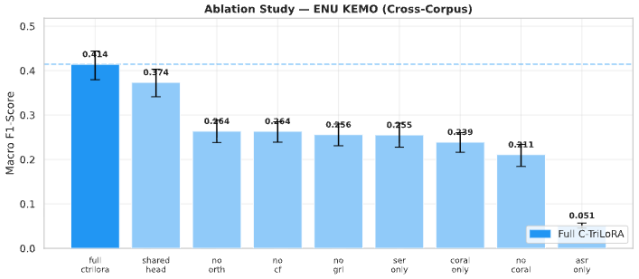
in Macro-F1 on the ENU KEMO dataset depending on the specific configuration. This finding underscores the centrality of second-order distribution matching for cross-corpus Speech Emotion Recognition (SER). This empirical result strongly aligns with established domain adaptation theory, which posits that minimizing the discrepancy in the covariance structure of source and target distributions provides a mathematical upper bound reduction on target-domain generalization error [4].

Furthermore, the CORAL-only variant only achieved a 23.91% Macro-F1. This indicates that unsupervised domain alignment alone is insufficient. It must be synergistically paired with explicit factor disentanglement to prevent negative transfer. We further note that two of these ablation configurations double as implicit comparisons to standard domain-adaptation baselines: the CORAL-only variant instantiates Deep CORAL [16], and the gradient-reversal component corresponds to domain-adversarial training (DANN) [7]. The full C-TRILORA surpasses both, indicating that explicit factor disentanglement adds value beyond these established alignment strategies; a broader empirical comparison against further

published cross-corpus systems [11], [12], [22], [27], [26] re-implemented under an identical Kazakh protocol is identified as future work.



(a) In-domain KazEmoTTS test set.



(b) Cross-corpus ENU KEMO dataset.

Fig. 4. Ablation study: Macro-F1 across model configurations.

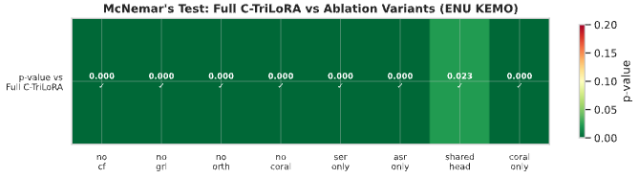


Fig. 5. Heatmap of p-values from McNemar's test, confirming the statistical significance of performance differences between the full C-TRILORA model and its ablation variants on the ENU KEMO dataset.

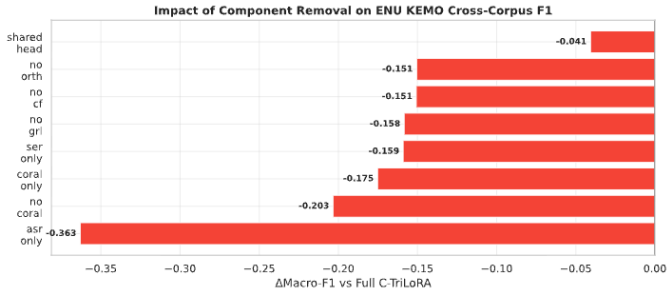
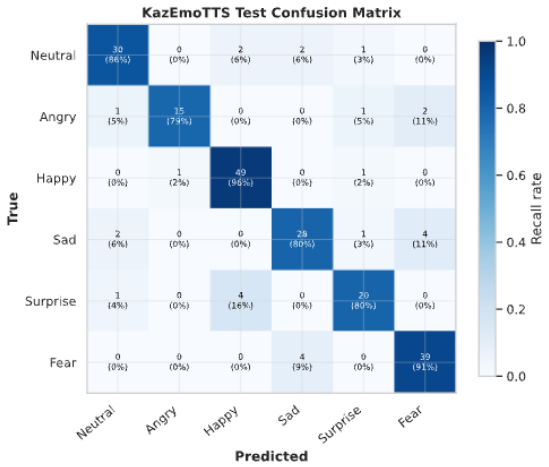


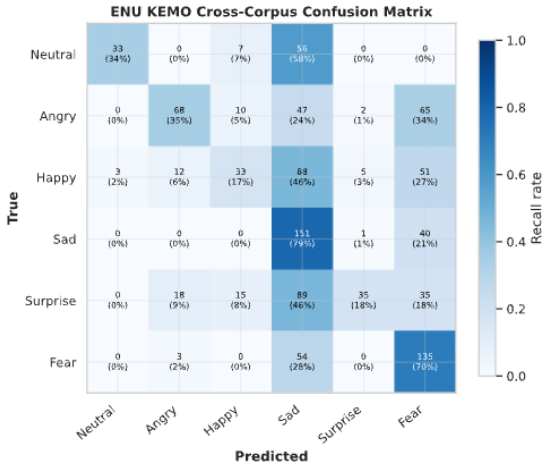
Fig. 6. Waterfall chart illustrating the absolute reduction (delta) in cross-corpus Macro-F1 on the ENU KEMO dataset when individual architectural components are removed from the full C-TRILORA model.

studies on the "valence bottleneck" in SER [5]. The confusion matrices in Fig. 7 demonstrate that high-arousal emotions with unambiguous, biologically driven prosodic signatures such as fear and neutral affect generalized relatively well across the two Kazakh corpora.

However, as highlighted by the per-class distribution in the radar plot (Fig. 8), mid-valence emotions like happiness and sadness suffered the most severe performance drops. The model exhibited systematic over-prediction of sadness and confusion between happy and sad states when tested on ENU KEMO. This aligns with broader SER literature indicating that the acoustic realizations of mid-valence emotions vary substantially across different speaker demographics, recording setups, and elicitation styles (e.g., read speech versus acted speech). Explicitly mapping these nuanced valence differences without target-domain labels remains a fundamental challenge.



(a) KazEmoTTS test set.



(b) Cross-corpus ENU KEMO.

Fig. 7. Confusion matrices of the proposed C-TRILORA model.

C. Error Analysis and Valence-Axis Shift

A detailed per-class analysis of the cross-corpus predictions reveals a systematic pattern consistent with previous theoretical

D. Probabilistic Calibration

Finally, the probabilistic calibration of the network was assessed using the Expected Calibration Error (ECE). As shown

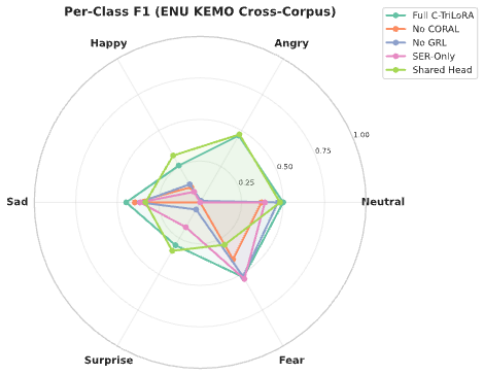
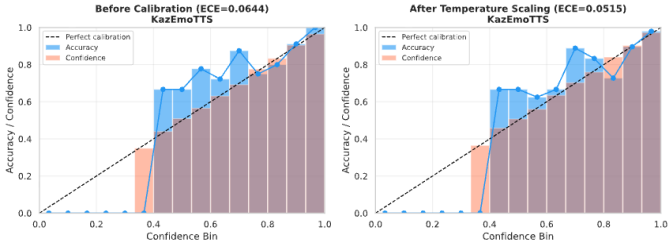


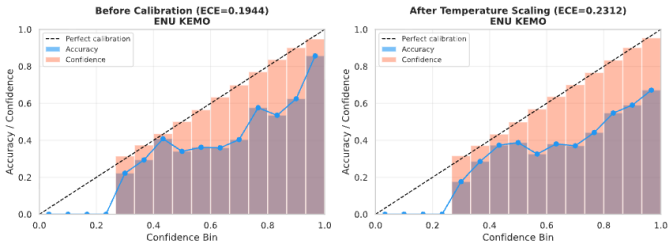
Fig. 8. Radar plot of per-class F1-scores for the ENU KEMO cross-corpus evaluation across ablation variants, demonstrating the "valence bottleneck" for mid-arousal emotions such as happiness and sadness.

in Fig. 9, before calibration, the model exhibited slight overconfidence in-domain (ECE $= 6.44\%$). Post-hoc temperature scaling successfully reduced this to 5.15%. However, under cross-corpus domain shift, the raw ECE rose substantially to 19.44%, reflecting systematic overconfidence on out-of-distribution data.

Crucially, applying the optimal in-domain temperature parameter actually worsened the cross-corpus ECE to 23.12%, confirming that standard temperature scaling does not seamlessly transfer across domains. This highlights a critical area for future research in domain-adaptive calibration for affective computing architectures.



(a) In-domain KazEmoTTS test set.



(b) Cross-corpus ENU KEMO dataset.

Fig. 9. Reliability diagrams comparing model accuracy against confidence (ECE) before and after post-hoc temperature scaling.

V. Conclusion

This study successfully addressed the challenge of cross-corpus covariate shift in low-resource speech emotion recognition through the development of the C-TRILORA architecture for Kazakh speech. The integration of a Tri-LoRA routing module separated lexical, emotional, and speaker representations, which, when combined with unsupervised CORAL domain alignment, reduced the fragility of standard acoustic models. Experimental results confirmed that this parameter-efficient approach significantly outperforms baseline models on unseen target domains by preserving domain-invariant affective features. Taken together, the results support the value of both feature disentanglement and second-order distribution matching for deploying reliable affective computing systems in low-data settings.

This study nevertheless has several limitations, which in turn define clear directions for future work. First, the evaluation is currently restricted to two Kazakh corpora; extending it to additional Kazakh datasets and, importantly, to other low-resource languages is a priority for establishing broader generality. Because the Tri-LoRA router, the CORAL alignment, and the gradient reversal layer make no language-specific assumptions (only the Whisper backbone checkpoint is language-dependent, and multilingual Whisper checkpoints are readily available), the architecture is expected to transfer to other languages, and a cross-lingual evaluation is planned. Second, beyond the internal Shared-Head baseline and the ablation variants two of which, the CORAL-only and gradient-reversal configurations, already correspond to the established Deep CORAL and DANN baselines - a comprehensive empirical comparison against the recent published cross-corpus systems surveyed in Section II, re-implemented under an identical Kazakh protocol, is an important next step. Third, the multi-task loss weights were set empirically; a systematic sensitivity analysis and joint optimisation of these weights remain to be carried out. Fourth, a detailed profiling of the computational cost: the additional training time, FLOPs, peak memory, and inference latency introduced by the routing module relative to the baseline is left for future work. Finally, the cross-corpus gains, while statistically significant, are not large in absolute terms, and the valence bottleneck for mid-arousal emotions together with domain-adaptive probabilistic calibration remain open problems for affective computing in low-data settings.

Acknowledgment

This research was funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan within the framework of the project AP23487777. The authors declare that there is no conflict of interest regarding the publication of this article and confirm that the study is free of plagiarism.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used Gemini to assist with language translation, structural formatting, and proofreading to align with journal guidelines. After using this tool, the authors closely reviewed and edited the generated text to ensure scientific accuracy and take full responsibility for the final content.

References

[1] M. B. Akcay and K. Oguz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[2] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[3] A. Abilbekov, S. Mussakhojayeva, R. Yeshpanov, and H. A. Varol, "KazEmoTTS: A dataset for Kazakh emotional text-to-speech synthesis," in *Proc. LREC-COLING*, Torino, Italy, 2024, pp. 9626–9632.

[4] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1–2, pp. 151–175, 2010.

[5] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Proc. Interspeech*, 2018, pp. 951–955.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12449–12460.

[7] Y. Ganin et al., "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[8] J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10745–10759, 2023.

[9] T. Feng and S. Narayanan, "PEFT-SER: On the use of parameter-efficient transfer learning approaches for speech emotion recognition using pre-trained speech models," in *Proc. ACII*, 2023, pp. 1–8.

[10] J. Ye, Y. Wei, X.-C. Wen, C. Ma, Z. Huang, K. Liu, and H. Shan, "Emo-DNA: Emotion decoupling and alignment learning for cross-corpus speech emotion recognition," in *Proc. 31st ACM Int. Conf. Multimedia (MM)*, 2023, doi: 10.1145/3581783.3611704.

[11] Y. Gao, L. Wang, J. Liu, J. Dang, and S. Okada, "Adversarial domain generalized transformer for cross-corpus speech emotion recognition," *IEEE Trans. Affective Computing*, vol. 15, no. 2, pp. 697–708, 2024.

[12] X. Zhou, J. Li, Q. Yu, and Q. Wu, "Classification inconsistency alignment network for cross-corpus speech emotion recognition," in *Proc. IEEE ICASSP*, 2025, pp. 1–5.

[13] S. Kakouros, T. Stafylakis, L. Mošner, and L. Burget, "Speech-based emotion recognition with self-supervised models using attentive channel-wise correlations and label smoothing," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[14] H. Fu, Z. Zhuang, Y. Wang, C. Huang, and W. Duan, "Cross-corpus speech emotion recognition based on multi-task learning and subdomain adaptation," *Entropy*, vol. 25, no. 1, p. 124, 2023.

[15] B. Kynabay, A. Aldabergen, S. Kadyrov, and A. Shalkarbay-uly, "Fine-tuning OpenAI's Whisper model for Kazakh speech recognition," 2025. [Online]. Available: https://doi.org/10.13140/RG.2.2.29371.07205

[16] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. European Conf. Computer Vision Workshops*, 2016, pp. 443–450.

[17] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28492–28518.

[18] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022.

[19] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th ICML*, 2017, pp. 1321–1330.

[20] M. Altaibek, A. Zulkhazhav, B. Yergesh, G. Bekmanova, and T. Aibol, "A multimodal framework for speech emotion recognition in low-resource languages," *Journal of Artificial Intelligence and Technology*, vol. 5, pp. 354–364, 2025.

[21] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[22] C.-S. Ahn, R. Rana, C. Busso, and J. C. Rajapakse, "Multitask transformer for cross-corpus speech emotion recognition," *IEEE Trans. Affective Computing*, 2025.

[23] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[24] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[25] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Self-supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Trans. Affective Computing*, 2022, doi: 10.1109/TAFFC.2022.3167013.

[26] H. Lian, C. Lu, Y. Zhao, S. Li, T. Qi, and Y. Zong, "Exploring corpus-invariant emotional acoustic feature for cross-corpus speech emotion recognition," *Expert Systems with Applications*, vol. 258, p. 125162, 2024.

[27] Y. Zhao, J. Wang, C. Lu, S. Li, B. W. Schuller, Y. Zong, and W. Zheng, "Emotion-aware contrastive adaptation network for source-free cross-corpus speech emotion recognition," in *Proc. IEEE ICASSP*, 2024, pp. 11846–11850.

[28] H. Li, M. Tu, J. Huang, S. Narayanan, and P. Georgiou, "Speaker-invariant affective representation learning via adversarial training," in *Proc. IEEE ICASSP*, 2020, pp. 7144–7148.

[29] W. Chen, X. Xing, P. Chen, and X. Xu, "Vesper: A compact and effective pretrained model for speech emotion recognition," *IEEE Trans. Affective Computing*, vol. 15, no. 3, pp. 1711–1724, 2024.