Journal of Artificial Intelligence and Technology, (Ahead of Print)
https://doi.org/10.37965/jait.2025.0781

ISTP
RESEARCH ARTICLE

A Multimodal Framework for Speech Emotion Recognition in Low-Resource Languages

Mamyr Altaibek,[1] Altanbek Zulkazhav,[1] Banu Yergesh,[1] Gulmira Bekmanova,[1] and Tileukhan Aibol[2]
[1]L.N.Gumilyov Eurasian National University, Astana, Kazakhstan
[2]Astana International University, Astana, Kazakhstan

(Received 14 April 2025; Revised 14 July 2025; Accepted 06 August 2025; Published online 04 September 2025)

Abstract: Speech emotion recognition (SER) plays a crucial role in enhancing human–computer interaction by identifying emotional states in speech. However, low-resource languages like Kazakh face challenges due to limited datasets and linguistic tools. To address this problem, we propose a novel multimodal framework, KEMO (Kazakh Emotion Multimodal Optimizer), which combines text-based semantic analysis and audio emotion recognition to leverage complementary features of linguistic and paralinguistic data. Using a Kazakh-translated version of the DAIR-AI (Contextualized Affect Representations for Emotion Recognition) dataset for text and the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset for audio, we have developed a system capable of classifying six emotions from text and eight emotions from audio. By integrating outputs from speech-to-text and audio-based recognition models with adaptive weighting, KEMO significantly improves the accuracy and robustness of emotion classification, providing an effective solution for SER in low-resource language scenarios.

Keywords: deep learning; Kazakh language; KEMO; low-resource languages; multimodal learning; speech emotion recognition

I. INTRODUCTION

In recent years, advancements in deep learning have revolutionized various domains, including image classification, machine translation, speech recognition, and text-to-speech synthesis, among others [1,2]. When applied to statistical speech processing, these algorithms significantly enhance the performance of systems, spurring interest in areas that delve into human nature and behavior, such as emotion recognition and emotional dialog modeling [3,4]. Emotion recognition, a cornerstone of affective computing, aims to bridge the gap between human emotional expressions and machine intelligence, paving the way for more natural and effective human–computer interaction [5].

Speech emotion recognition (SER) focuses on predicting and categorizing the emotional content of spoken language into predefined labels such as happy, sad, neutral, or angry. It is an integral aspect of paralinguistics and a crucial component in dialog systems, virtual assistants, and human-centered artificial intelligence applications [6,7]. However, despite numerous advances in SER, it remains a challenging task due to inherent complexities in emotional expressions, variability in vocal cues, and a lack of sufficient annotated datasets, especially for low-resource languages [8,9].

Traditional SER methods rely primarily on low-level acoustic features like pitch, intensity, and Mel-frequency cepstral coefficients (MFCCs) to infer emotions [10]. While these methods perform well in controlled environments, they often struggle to generalize across diverse datasets and cultural contexts. Function-based models, though robust in signal processing tasks, exhibit limitations when capturing the nuances of human emotions, particularly in spontaneous speech scenarios [11]. Moreover, the lack of large-scale datasets for languages like Kazakh amplifies these challenges, making it difficult to effectively train complex neural network models [12].

To address these limitations, researchers have explored multimodal approaches that combine acoustic features with high-level textual information derived from automatic speech recognition (ASR) systems [13,14]. Textual sentiment analysis plays a pivotal role in enhancing SER by leveraging emotionally charged words, such as "beautiful" or "wonderful," which carry strong emotional weight compared to neutral words like "sky" or "bird" [15]. The integration of textual and acoustic features showed promise in capturing the multifaceted nature of emotions, particularly in scenarios where either modality alone might provide incomplete information [16,17].

In this study, we propose KEMO (Kazakh Emotion Multimodal Optimizer), a novel framework designed to address the unique challenges of SER for low-resource languages. KEMO leverages both speech-to-text emotion analysis and traditional acoustic-based SER to provide a holistic understanding of emotional content. By integrating insights from high-level text transcriptions and low-level speech signals, the model maximizes the utility of limited datasets while maintaining robust performance across diverse scenarios. The framework employs a weighted fusion mechanism to balance the contributions of textual and acoustic modalities, optimizing emotion recognition accuracy in noisy or ambiguous contexts [18].

To evaluate KEMO, we conduct experiments using the DAIR-AI dataset for text-based emotion recognition, translated into Kazakh, and the RAVDESS dataset for speech-based emotion classification, which distinguishes among eight distinct emotions. The results demonstrate KEMO's ability to effectively classify emotions, highlighting its potential as a scalable and adaptable solution for low-resource languages [19,20]. By advancing the state of SER for Kazakh and other underrepresented languages, KEMO contributes to the broader goal of inclusive and multilingual artificial intelligence systems.

Corresponding authors: Altanbek Zulkazhav (e-mail: altinbekpin@gmail.com); Banu Yergesh (e-mail: b.yergesh@gmail.com).

Additionally, environmental noise in speech data can obscure emotional cues, particularly when noise sources vary in their spectral distributions, potentially introducing biases in model predictions [8]. For low-resource languages like Kazakh, these challenges are compounded by limited dataset diversity and the need for robust multimodal fusion strategies to leverage complementary information from text and audio modalities [16,17]. Addressing these issues requires innovative approaches to feature extraction, noise mitigation, and modality integration, which our KEMO framework aims to achieve.

The rest of the paper is structured as follows. Section II reviews the literature on SER and related work, particularly in low-resource languages. Section III describes the KEMO framework's methodology, including audio and text processing. Section IV presents the experimental setup, datasets, and results. Finally, Section V concludes with a summary of contributions and future directions.

II. RELATED WORK

Research into SER and sentiment analysis has significantly advanced due to the application of deep learning methods. Early efforts in SER often relied on classical machine learning techniques, leveraging features such as MFCCs or low-level descriptors (LLDs) to capture emotional cues in speech [21,22]. These approaches, though foundational, often faced limitations in capturing the complexity of emotional states. Recent work has demonstrated the effectiveness of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) in modeling such nuanced features from spectrograms or raw waveforms, significantly enhancing the accuracy of SER systems [23–25].

The integration of multimodal data—combining audio and text—emerged as a promising approach for emotion recognition. Models that incorporate both acoustic features and textual transcriptions showed notable improvements in performance by utilizing complementary information from speech and semantics [26]. These advancements paved the way for more sophisticated models capable of identifying emotions across diverse domains.

In the context of speech processing, recent studies have explored both health informatics and speaker-specific feature extraction. For instance, Pradhan *et al.* [27] proposed a cascaded perceptual functional link artificial neural network (PFLANN) model for detecting respiratory diseases, leveraging bio-inspired computing techniques such as particle swarm optimization (PSO) to optimize weights. Their approach employs MFCCs and linear spectrum features to capture nonlinear characteristics of speech, achieving high accuracy with low computational complexity across eight datasets, including the Saarbruecken Voice Database and Coswara. While their work focuses on health monitoring, the use of MFCCs and a cascaded neural network structure shares similarities with our KEMO framework, which integrates audio and text modalities for emotion recognition. However, unlike their general speech analysis for disease detection, our study addresses the unique challenges of SER in low-resource languages like Kazakh, where data scarcity and linguistic nuances necessitate tailored multimodal fusion strategies.

Recent advancements in SER have also explored hybrid and speaker-specific approaches. Araño *et al.* [28] proposed combining traditional MFCCs with spectrogram-based image features extracted by a pretrained CNN (ResNet50) for emotion recognition, achieving an accuracy of 0.735 on the RAVDESS dataset using an MFCC-LSTM model. Their work highlights the effectiveness of traditional features like MFCCs in data-scarce scenarios, a challenge also relevant to low-resource languages. Similarly, Kong *et al.* [29] introduced ELF, a method for encoding speaker-specific latent speech features for speech synthesis, using a variational autoencoder to capture speaker characteristics without additional training. While ELF focuses on speech synthesis, its ability to model speaker-specific features in a continuous latent space shares conceptual similarities with KEMO's dynamic weighting mechanism for modality fusion. Unlike these approaches, which primarily address high-resource languages or speech synthesis, KEMO integrates text and audio modalities with a focus on cultural and linguistic adaptability for Kazakh, addressing data scarcity through dynamic feature weighting.

In the context of low-resource languages, SER faced unique challenges due to limited annotated data and linguistic diversity. Chopra *et al.* [30] proposed a meta-learning framework that leveraged cross-language knowledge transfer, enhancing generalization for low-resource SER with minimal labeled data. This approach aligned with efforts to address Kazakh's data scarcity, though it focused primarily on unimodal audio. Zhao *et al.* [9] explored cross-lingual and cross-modal SER, demonstrating the potential of multilingual models in low-resource settings. These works underscored the need for tailored strategies, such as KEMO's dynamic weighting, to capture the linguistic and cultural nuances of languages like Kazakh.

In the context of Kazakh, a low-resource language, there have been significant strides in language-specific technologies. Research on sentiment analysis for Kazakh text has been bolstered by the introduction of KazSAnDRA, a dataset featuring reviews annotated for sentiment polarity and scores [31]. This dataset has enabled the development of machine learning models tailored to the linguistic nuances of Kazakh. Similarly, advancements in text-to-speech synthesis and ASR for Kazakh have been noteworthy, with systems achieving competitive results through the application of end-to-end neural architectures and multilingual training [32,33]. Despite these achievements, there remains a gap in leveraging these advances for SER in Kazakh. Previous works on Kazakh sentiment and speech processing have laid the groundwork for extending these techniques to emotion recognition. For example, multilingual ASR systems have demonstrated the ability to generalize across languages, including Kazakh, Russian, and English, providing a robust platform for speech transcription [33]. However, few studies explore the integration of these capabilities with emotion analysis to address the unique challenges of low-resource languages like Kazakh.

Recent studies have explored techniques to mitigate environmental noise in SER, which can mask emotional content if not addressed. For instance, speaker embeddings like X-vectors are used to isolate speaker-specific features, reducing the impact of background noise [29,32]. Additionally, the choice of feature dimensionality, such as MFCCs, plays a critical role in balancing computational efficiency and emotional expressiveness [10,21,27,28]. Multimodal fusion strategies, ranging from early to late fusion, are proposed to handle modality interactions, with some approaches incorporating hidden-layer interactions for stronger cross-modal alignment [16,34]. These advancements inform our approach to designing a robust SER system for Kazakh, addressing both noise and modality integration challenges. Our KEMO framework builds on these insights by employing a dynamic weighting mechanism to balance text and audio modalities, tailored to the linguistic and cultural nuances of Kazakh,

thereby addressing the data scarcity and variability inherent in low-resource language scenarios.

In the context of low-resource languages, SER faces unique challenges due to limited annotated data and linguistic diversity. Chopra *et al.* [30] propose a meta-learning framework that leverages cross-language knowledge transfer, enhancing generalization for low-resource SER with minimal labeled data. This approach aligns with efforts to address Kazakh's data scarcity, though it focuses primarily on unimodal audio. Zhao *et al.* [9] explore cross-lingual and cross-modal SER, demonstrating the potential of multilingual models in low-resource settings. These works underscore the need for tailored strategies, such as KEMO's dynamic weighting, to capture the linguistic and cultural nuances of languages like Kazakh.

III. METHODOLOGY

In this section, we describe our proposed method, KEMO, for SER in low-resource languages. The methodology consists of two main components: a Bilateral Sequence Model (BiLSTM) for audio emotion recognition and a Textual Multimodal Framework for emotion classification based on text embeddings from state-of-the-art models such as LaBSE, XLM-RoBERTa, mBART, and mBERT. The proposed methodology and the process are shown in Fig. 1.

To ensure compatibility between text and audio modalities, we carefully balance the parameter scales of the text and audio models. Using pretrained models like LaBSE or XLM-RoBERTa as fixed feature extractors minimizes parameter imbalance with the smaller BiLSTM audio model, treating text embeddings as pretrained word vectors. Fine-tuning these models, however, risks dominance by the text modality due to its larger parameter count (e.g., millions vs. 2.1M for BiLSTM). To address this, our dynamic weighting mechanism adaptively prioritizes modalities based on input characteristics, such as high emotional intensity in audio (e.g., large pitch variations) or semantically rich text (e.g., longer sentences), enhancing fusion robustness.

A. AUDIO EMOTION RECOGNITION WITH BILSTM AND EMOTION2VEC

The BiLSTM-based architecture captures temporal patterns in audio signals for emotion classification. Features such as MFCCs and spectrogram-based descriptors are first extracted from raw audio data using openSMILE [35]. These features are processed by a stacked BiLSTM, where the forward and backward states are used to generate a robust representation of the signal [36].

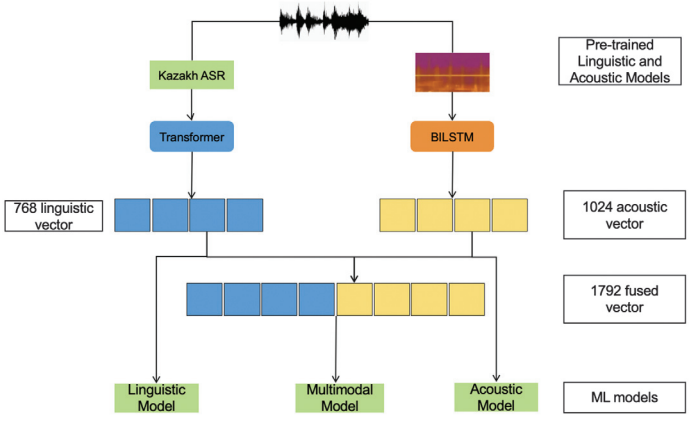

Fig. 1. Workflow of the proposed KEMO cascading integration methodology.

B. BILSTM SEQUENCE FORMULATION

Let the sequence of extracted features for an audio signal be $x = \{x_1, x_2, \ldots, x_T\}$, where T is the number of time steps. The BiLSTM processes this sequence to generate a hidden state at each time step:

$$h_t^{forward} = f_\theta^{forward}(x_t, h_{t-1}^{forward}) \tag{1}$$

$$h_t^{backword} = f_\theta^{backword}(x_t, h_{t+1}^{backword}) \tag{2}$$

The final classification is performed by applying a softmax function to the aggregated output:

$$\hat{y} = softmax\,(W_a h_T + b_a) \tag{3}$$

where $\hat{y}$ represents the predicted probabilities of the emotion classes and W_a, b_a are trainable parameters.

C. TEXTUAL EMOTION RECOGNITION WITH PRETRAINED MODELS

We leverage embeddings from pretrained models such as LaBSE [37], XLM-RoBERTa [38], mBERT [39], and mBART [40] to classify emotions based on textual transcriptions of speech. Each model maps a sequence of tokens s to an embedding vector:

$$z = g_\phi(s) \tag{4}$$

where g_ϕ is the transformation function of the pretrained model. The classifier uses embedding to predict emotion probabilities:

$$\hat{y} = softmax\,(W_t z + b_t) \tag{5}$$

where W_t and b_t are trainable parameters.

D. KAZAKH EMOTION MULTIMODAL OPTIMIZER (KEMO)

The KEMO framework combines audio and text predictions using a weighted fusion strategy. Let $\hat{y}_a$ and $\hat{y}_t$ denote the predictions from the audio and text models, respectively. The final prediction is:

$$\hat{y} = \lambda \hat{y}_a + (1 - \lambda)\,\hat{y}_t \tag{6}$$

where λ is a trainable weight parameter that balances the contributions of the two modalities.

E. FUSION METHOD: CASCADING INTEGRATION FOR EMOTION RECOGNITION

To address the challenges of emotion recognition in low-resource languages, we propose a cascading integration methodology that leverages both text-based and audio-based emotion classification. This ensures efficient combination of predictive strengths while maintaining classification into six basic emotions: *neutral, angry, happy, sad, scared, and surprised.*

1. INITIAL EMOTION PREDICTION FROM AUDIO MODEL. The audio model, trained on the RAVDESS dataset using a BiLSTM

architecture, classifies audio input into eight predefined emotion categories. This prediction is:

$$\widehat{y}_{audio} = \arg\max_{c} P_{audio}(c), \quad c \in \{8\,emotions\} \tag{7}$$

where $P_{audio}(c)$ represents the probability for each emotion.

2. REFINEMENT THROUGH TEXT MODEL. The text model refines this prediction by concatenating audio predictions (as one-hot vectors) with text embeddings:

$$F_{text+audio} = [F_{text}, o_{audio}] \tag{8}$$

where F_{text} is the embedding derived from the text and o_{audio} is the encoded one-hot vector of the audio prediction.

The final classification is performed as:

$$\widehat{y}_{text+audio} = \arg\max_{c} P_{text+audio}(c) \quad c \in \{6\,emotions\} \tag{9}$$

where $P_{txt+audio}(c)$ represents the predicted probability after combining the audio and text information.

F. CROSS-MODAL ALIGNMENT THEORY

Emotion alignment across modalities requires both psychological validity and computational realizability. Building upon Ekman's basic emotion theory [41], we introduce a novel computational framework inspired by Cross-modal Attention Alignment (CAA) [34] to bridge the semantic gap between text and speech modalities.

1. PSYCHOLOGICAL FOUNDATION. According to Scherer's Component Process Model [42], emotional expressions exhibit synchronized patterns across verbal content (lexical semantics) and vocal features (prosodic cues). This synchronization creates natural anchors for cross-modal alignment:

$$S(t,a) = \phi_{text}(\omega_t) \odot \psi_{audio}(f_a) \tag{10}$$

where ϕ_{text} and ψ_{audio} are embedding functions for words and audio frames, respectively, and $\odot$ denotes element-wise multiplication.

2. CROSS-MODAL ATTENTION ALIGNMENT. We extend the CAA mechanism from [43] with culture-specific adaptation for Kazakh. Let

$$H_t \in R^{n\times d} \; and \; H_a \in R^{m\times d}$$

be the text and audio hidden states, respectively. The alignment energy matrix

$$E \in R^{n\times m}$$

is computed as:

$$E_{ij} = \frac{\exp\left(\mathrm{sim}(H_t^i, H_a^j)\right)}{\sum_{k=1}^{m} \exp\left(\mathrm{sim}(H_t^i, H_a^k)\right)} \tag{11}$$

where

$$sim(u,v) = u^T W v$$

is a learnable similarity metric, and

$$W \in R^{d\times d}$$

is the alignment matrix.

The culture-adaptive alignment is achieved through:

$$\widehat{E} = E \odot M_{kz} \tag{12}$$

where M_{kz} is the Kazakh-specific alignment prior learned from cultural linguistics data [44], and $\otimes$ denotes the Hadamard product.

3. DYNAMIC MODALITY WEIGHTING. KEMO employs a context-aware dynamic weighting mechanism to balance text and audio modalities, addressing Kazakh's agglutinative morphology. The weight λ in equation (13) is computed using a gated attention mechanism:

$$\lambda = \sigma(W \cdot [h_{CLS}, h_{audio}] + b) \tag{13}$$

where σ is the sigmoid function, $W$and b are trainable parameters, h_{CLS} is the text model's [CLS] token embedding, and h_{audio} is the audio global feature (mean-pooled BiLSTM output). This allows λ to adapt to input characteristics, for example, prioritizing audio for high pitch variations or text for longer sentences, with complexity O(n) optimized via GPU parallelization.

Compared to baseline models, KEMO's approach differs significantly. CM-BERT uses cross-modal attention [34], aligning text and audio features via a similarity matrix, which assumes static modality importance and lacks adaptability to Kazakh's linguistic nuances. MM-EmoNet employs dynamic graph fusion [43], constructing a graph to model modality interactions, but its computational overhead ($O(n^2)$) and fixed graph structure limit flexibility in low-resource settings.

G. FEATURE EXTRACTION AND NOISE MITIGATION

The choice of feature dimensionality is critical for robust SER, particularly in noisy environments. For audio features, we extracted 39 MFCCs from the RAVDESS dataset, including 12 static coefficients, 13 delta coefficients, and 13 acceleration coefficients, with a time dimension of 1280 frames (approximately 8 s of audio, segmented into 25 ms frames with a 10 ms stride) [10]. For simpler tasks, a reduced dimensionality of 13 static coefficients may suffice, balancing computational efficiency and emotional expressiveness. Speaker embeddings, such as X-vectors (fixed at 200 dimensions), were used to isolate speaker-specific features, mitigating the impact of environmental noise [32]. Noise from a single source often follows a predictable spectral distribution, acting as a bias in the model (e.g., associating meeting room echoes with anger). Diverse noise sources, however, can mask emotional cues. To address this, we increased dataset diversity in our ENU KEMO dataset and leveraged X-vectors to filter environmental noise, ensuring focus on emotional content [8].

IV. EXPERIMENTAL STEPS AND RESULTS

A. DATASET

To evaluate the KEMO framework, we have utilized the DAIR-AI dataset for text-based emotion classification and the RAVDESS dataset for speech-based emotion recognition. The DAIR-AI dataset, originally in English, contains short sentences categorized into six emotional classes: neutral, angry, happy, sad, scared, and surprised (Table I). Due to its concise sentence structure, the dataset is well suited for translation and validation tasks in low-resource languages like Kazakh.

The DAIR-AI dataset was translated into Kazakh using the Google Translation API, followed by a rigorous manual verification process to ensure linguistic and emotional accuracy, aligning with best practices for low-resource language dataset adaptation [45]. The verification was conducted by seven native Kazakh speakers, all of whom are researchers in linguistics or related fields, ensuring expertise in the domain. Each validator

Table I. Example from the DAIR-AI dataset original text and emotion labels

Original text (English)	Emotion label
I was still feeling strong	Joy
I feel pretty pathetic most of the time	Sadness
I am feeling grouchy	Anger
I feel vulnerable and alone	Fear
I feel romantic too	Love
I remember feeling amazed	Surprise

independently reviewed 200 randomly selected samples from the dataset, with 50 shared samples evaluated by all seven to assess inter-rater consistency. The validation process involved checking grammar, vocabulary, and semantic fidelity to the original English text. For emotional accuracy, validators re-annotated the emotional labels of the translated samples and compared them with the original English labels. Table II summarizes the validation results, showing an average emotional label agreement of 87% (174/200 samples per validator) across the seven validators. For the 50 shared samples, the agreement rate reached 92% (46/50 samples), with inconsistencies primarily occurring in the "scared" and "surprised" categories due to semantic overlap in Kazakh expressions. Inconsistent samples were resolved through discussion and retranslation, ensuring the dataset's suitability for Kazakh SER tasks. Validators also assessed emotional fidelity by comparing translated emotional cues with original labels.

The ENU KEMO dataset, comprising recordings from five Kazakh speakers (3 men, 2 women) with 18 samples per speaker (90 total), poses challenges to evaluation robustness due to its small size and limited speaker diversity. To enhance reliability, we employ 5-fold cross-validation, ensuring each fold contains a balanced subset of speakers and emotions. The average accuracy across folds exhibits a fluctuation of ±2.3%, indicating moderate stability despite the constrained dataset [46]. Future work aims to expand the speaker pool to improve generalization.

The RAVDESS dataset, a widely recognized benchmark for speech-based emotion recognition, contains 1440 audio files recorded by 24 professional actors (12 female and 12 male). Each actor vocalizes two lexically matched statements in a neutral North American accent, with expressions spanning eight emotional categories: neutral, calm, happy, sad, angry, fearful, surprised, and disgust. Emotions are produced at two intensity levels—normal and strong—with an additional neutral expression for balance.

B. EMOTION ALIGNMENT ACROSS MODELS

Emotion alignment is performed in our study to ensure consistency in emotion categories across different datasets, addressing variations in predefined emotion labels between the RAVDESS dataset (8 emotions: neutral, calm, happy, sad, angry, fearful, surprised, disgust) and the DAIR-AI dataset (6 emotions: neutral, angry, happy, sad, scared, surprised), as well as differences in the output types and the number of categories generated by the text and speech models. This alignment was critical for enabling multimodal fusion in the KEMO framework, particularly for low-resource languages like Kazakh, where linguistic and cultural nuances influence emotional expression. The distribution of the DAIR-AI dataset was shown in Fig. 2.

1. EMOTION ALIGNMENT IN THE RAVDESS SPEECH MODEL. Disgust classified under Anger: The decision to map "disgust" to "angry" was grounded in psychological research, notably Ekman's basic emotion theory, which identifies disgust and anger as sharing high-arousal and negative-valence characteristics [41]. Both emotions exhibit similar vocal features, such as increased pitch variability and intensity, as well as overlapping facial expressions (e.g., furrowed brows, tense mouth) [47]. In the context of Kazakh, where cultural expressions of negative emotions may amplify these similarities, this alignment ensures consistency with the DAIR-AI dataset's "angry" category.

Calmness classified under Neutral: The mapping of "calm" to "neutral" was based on their shared acoustic characteristics and dimensional alignment in Russell's Circumplex Model [47]. Both emotions are characterized by low arousal and neutral valence, with minimal variations in pitch, loudness, and speaking rate. In Kazakh, "calm" expressions often align closely with neutral states, lacking the distinct emotional intensity of other categories like "happy" or "sad." This alignment facilitates compatibility between the RAVDESS and DAIR-AI datasets, ensuring robust multimodal integration.

These mappings were informed by both theoretical frameworks and cultural considerations specific to Kazakh emotional expressions. To validate these choices empirically, we conducted experiments to compare the proposed alignments with alternative mappings, as detailed in Section IV.D.

C. FEATURE EXTRACTION

To capture the emotional content in speech and text for the KEMO framework, we extracted both audio and text features from the RAVDESS and DAIR-AI datasets, respectively. The choice of features was critical to balance computational efficiency and emotional expressiveness, particularly in the low-resource language context of Kazakh.

Audio Features: MFCCs were selected as the primary audio features for the RAVDESS dataset, which contains 1440 audio files recorded by 24 professional actors. Each audio file was segmented into frames of 25 ms with a 10 ms stride using a Hamming window. A total of 39 features were derived, capturing spectral dynamics and temporal changes critical for emotional expression. Additionally, prosodic features such as pitch (F0), loudness, and voicing probability were incorporated to enrich the audio feature representation. MFCCs were chosen due to their well-established effectiveness in SER and related speech processing tasks, as they effectively capture the nonlinear characteristics of the human auditory system [10,21,27,28]. For instance, Pradhan *et al.* [27] demonstrated MFCCs' robustness in detecting respiratory diseases

Table II. Example from the DAIR-AI dataset original text and emotion labels

Alignment strategy	RAVDESS accuracy	RAVDESS F1-score	DAIR-AI accuracy	DAIR-AI F1-score
Proposed (Disgust → Angry, Calm → Neutral)	82.50%	0.81	84.30%	0.83
Disgust as Independent Category	78.70%	0.77	81.40%	0.79
Calm → Happy	80.00%	0.78	81.20%	0.8

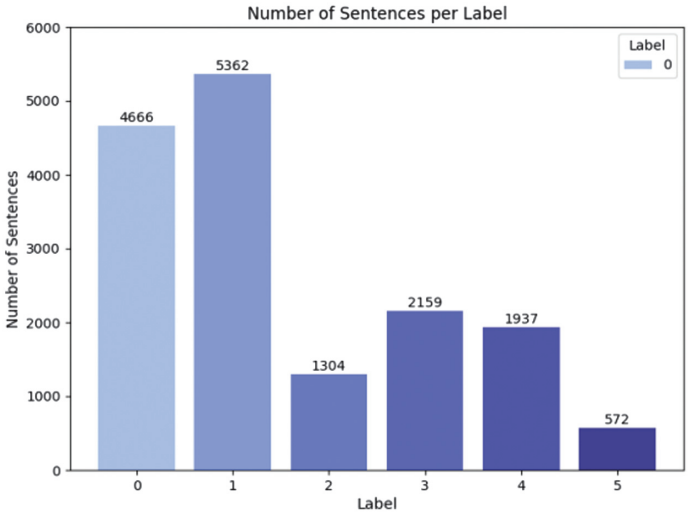

Fig. 2. Distribution of sentences per emotion label in the DAIR-AI dataset.

from speech, while Araño *et al.* [28] reported strong performance on the RAVDESS dataset, particularly in data-scarce scenarios, aligning with the challenges of low-resource languages like Kazakh.

To ensure the suitability of MFCCs, we evaluated alternative audio features, including Mel spectrograms and chroma features (Chroma STFT), on the RAVDESS dataset. Audio files were preprocessed by trimming silence (using a threshold of 25 dB to remove non-informative segments) and applying a Wiener filter to reduce noise, enhancing signal quality for feature extraction. The feature extraction process used a sampling rate of 44.1 kHz, consistent with high-quality audio standards. Figure 3 illustrates the combined visualization of MFCC features (39 coefficients), Mel spectrograms, and chroma features extracted from a sample RAVDESS audio file (neutral emotion, male speaker). The MFCC features represent spectral dynamics and temporal changes, the Mel spectrogram highlights frequency variations on a nonlinear Mel scale, and the chroma features capture pitch-related characteristics. While Mel spectrograms and chroma features provide valuable frequency and pitch information, preliminary experiments indicated that they did not outperform MFCCs in emotional

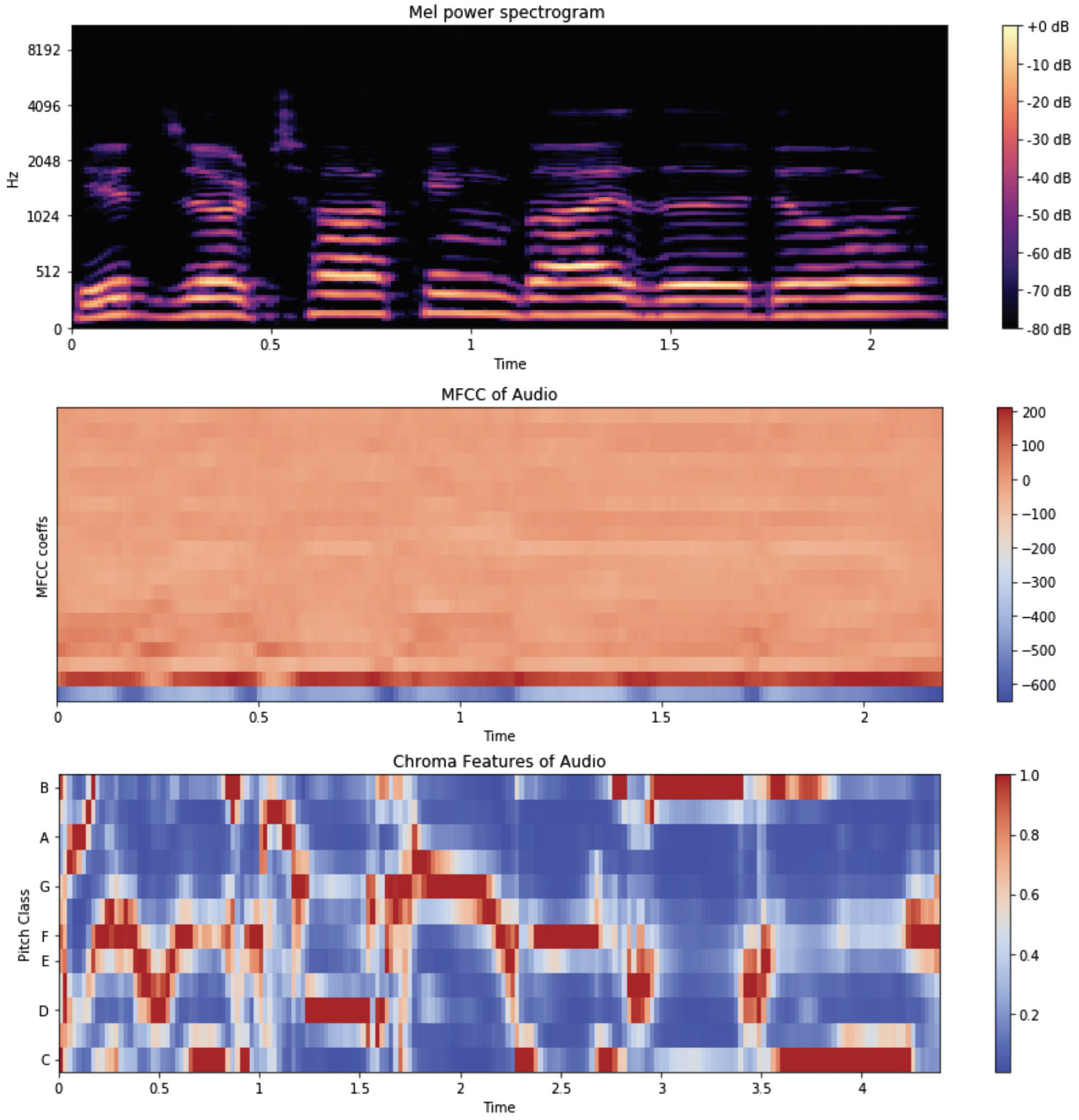

Fig. 3. Combined audio feature visualization.

classification accuracy for the six emotional categories (neutral, angry, happy, sad, scared, and surprised) in our low-resource setting. Thus, we adopted 39 MFCC features, augmented with prosodic features, to balance computational efficiency and emotional expressiveness for Kazakh SER.

Integrated visualization of 39 MFCC coefficients, Mel power spectrogram (128 Mel bands, n_fft=2048, hop_length=512), and chroma STFT features (12 semitone bins, hop_length=512) extracted from a RAVDESS audio sample (neutral emotion, male speaker), showcasing spectral dynamics, frequency variations, and pitch-related characteristics.

Text Features: Text embeddings were extracted from the DAIR-AI dataset using transformer-based models, including LaBSE, XLM-RoBERTa, BERT multilingual, and mBART. These embeddings were fine-tuned for the six predefined emotional categories (neutral, angry, happy, sad, scared, and surprised) to capture semantic and emotional content specific to Kazakh. The fine-tuning process leveraged the translated DAIR-AI dataset, ensuring alignment with the linguistic nuances of Kazakh, such as its agglutinative morphology.

Note on ASR Impact: The accuracy of ASR significantly impacts multimodal SER performance. Our ESPnet-based ASR achieved a word error rate (WER) of 4.1%, but experiments comparing ground-truth transcriptions with ASR-generated text revealed a performance gap, as ASR errors introduce biases that affect emotion classification [13]. For instance, different ASR systems (e.g., in-house vs. commercial APIs with ∼5% WER) may exhibit varying biases, leading to inconsistent classification in deployment scenarios.

D. IMPLEMENTATION DETAILS

To substantiate the emotion alignment decisions (disgust to angry, calm to neutral), we conducted experiments to evaluate their impact on the KEMO framework's performance across the RAVDESS and DAIR-AI datasets. The alignment was tested using the speech-based model (trained on RAVDESS) and the text-based model (trained on DAIR-AI), with performance measured in terms of classification accuracy and F1-score. We compared the proposed alignment with two alternative mappings: (1) treating "disgust" as an independent category, and (2) mapping "calm" to "happy" to reflect potential positive valence overlap in some cultural contexts.

Experimental Setup: A subset of 200 samples from the RAVDESS dataset (25 samples per emotion) and 200 samples from the DAIR-AI dataset (approximately 33 samples per emotion) were used for validation. The speech model extracted 39 MFCC features (12 static, 13 delta, 13 acceleration coefficients) and prosodic features (pitch, loudness, voicing probability), as described in Section IV.C. The text model utilized fine-tuned LaBSE embeddings for the DAIR-AI dataset. Both models were evaluated using 5-fold cross-validation to ensure robustness.

Results: Table II summarizes the performance of the KEMO framework under different alignment strategies. The proposed alignment (disgust to angry, calm to neutral) achieved an average accuracy of 82.5% on RAVDESS and 84.3% on DAIR-AI, with F1-scores of 0.81 and 0.83, respectively. Treating "disgust" as an independent category reduced accuracy by 3.8% on RAVDESS (78.7%) and 2.9% on DAIR-AI (81.4%), likely due to data imbalance, as disgust samples in RAVDESS are less frequent (192 samples) compared to angry (384 samples). Mapping "calm" to "happy" decreased accuracy by 2.5% on RAVDESS (80.0%) and 3.1% on DAIR-AI (81.2%), as calm's low-arousal acoustic profile conflicted with happy's high-arousal characteristics, particularly in Kazakh, where cultural expressions of happiness are more dynamic. These results confirm that the proposed alignment optimizes classification performance and aligns with cultural nuances in Kazakh emotional expression.

Cultural Considerations: In Kazakh, negative emotions like disgust and anger often share expressive patterns, such as abrupt intonation and lexical emphasis, supporting their alignment. Similarly, calm and neutral expressions are minimally distinct, often conveyed through steady speech patterns, aligning with Russell's model [47]. The empirical results validate that these mappings enhance multimodal fusion in KEMO, particularly for low-resource settings where data scarcity necessitates robust alignment strategies.

E. RESULTS

1. TEXT MODELS. The performance of the text models across three data sampling methods (imbalanced, ROS, and RUS) is summarized in Table III.

2. SPEECH MODEL. Using the RAVDESS dataset, we reproduced the performance of a BiLSTM model designed for SER. The reproduced results, based on the implementation provided in [46], are presented in Table IV.

Table III. Performance of text models

Model	Method	Accuracy	Precision	Recall	F1 score
LaBSE	Imbalanced	0.76	0.78	0.76	0.76
	ROS	0.76	0.77	0.76	0.76
	RUS	0.75	0.76	0.75	0.76
XLM-RoBERTa	Imbalanced	0.731	0.755	0.731	0.735
	ROS	0.715	0.725	0.715	0.718
	RUS	0.720	0.735	0.720	0.725
BERT multilingual	Imbalanced	0.7	0.75	0.7	0.71
	ROS	0.75	0.75	0.75	0.75
	RUS	0.73	0.75	0.73	0.74
mBART	Imbalanced	0.75	0.76	0.75	0.75
	ROS	0.66	0.74	0.66	0.66
	RUS	0.71	0.74	0.71	0.72

Table IV. Performance of BiLSTM model on RAVDESS dataset

Model	Params (M)	Preprocessing method	Dataset	Classes	Accuracy
BiLSTM	2.10	Emotion2Vec	RAVDESS	8	0.85333

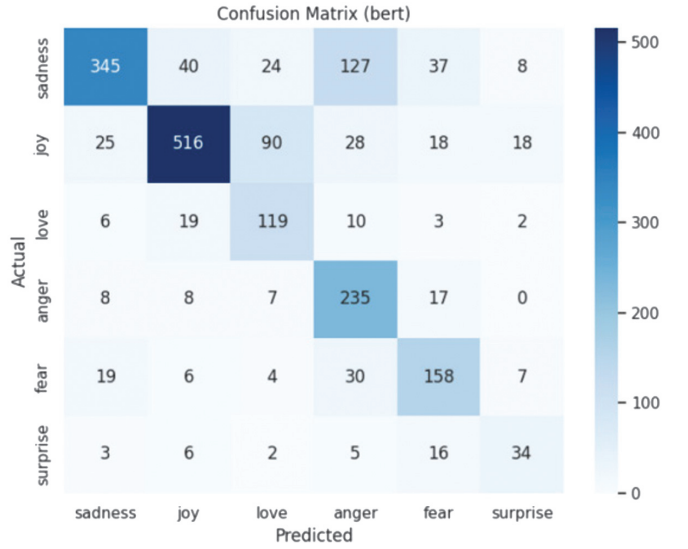

Fig. 4. Confusion matrices for text models (BERT multilingual).

3. CONFUSION MATRICES. To provide insights into the classification performance, confusion matrices are presented:

- **Text Models:** Since the confusion matrix distributions of the models are similar, one representative result is selected for display (as shown in Fig. 4).
- **Speech Model:** Figure 5 shows the confusion matrix for the BiLSTM model trained on the RAVDESS dataset.

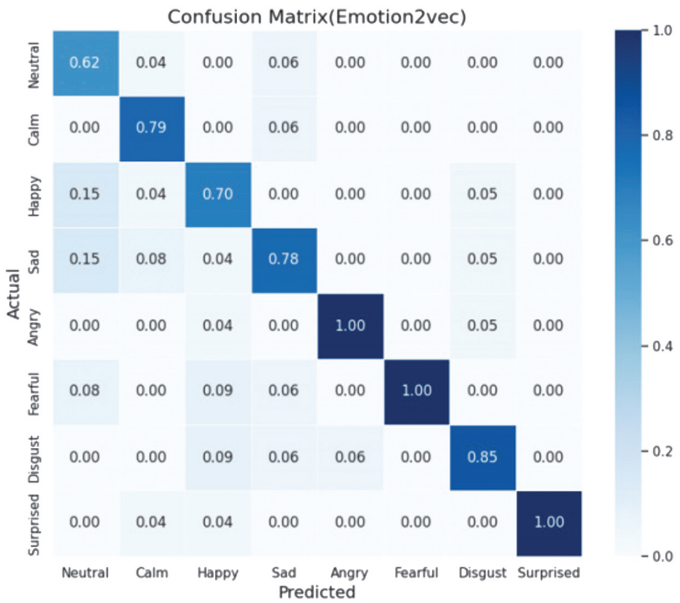

Fig. 5. Confusion matrix for BiLSTM model on RAVDESS dataset.

Table V. Accuracy results for speech, text, and fusion emotion recognition across datasets

Dataset	Speech Acc (%)	Text Acc (%)	Fusion Acc (%)
RAVDESS (English)	85.33	–	–
DAIR-AI (English)	–	93.05	–
NU Emo-TTS (Kazakh)	36.36	15.55	30.82
ENU KEMO (Kazakh)	33.33	45.56	44.44

Through the analysis of the confusion matrices, the text-based model performs well in identifying categories like **joy** but misclassifies **sadness** as **anger**, likely due to linguistic overlaps in Kazakh. In contrast, the speech-based Emotion2vec model achieves near-perfect accuracy for **angry** and **fearful**, effectively addressing the text model's limitations.

We also used the Emotion TTS dataset from Nazarbayev University [48] as one of the evaluation standards, rather than as a training resource. This decision was made because the purpose of our model differs from that of the Emotion TTS dataset. The dataset is generated through voice conversion after recording a few samples, resulting in consistent tonal patterns across all emotions for each speaker, with no diversity. Additionally, the textual content was not designed to reflect emotional semantics.

Our lab collected a small dataset featuring 5 Kazakh speakers (3 men and 2 women), where each speaker recorded 6 emotions with 3 samples per emotion, resulting in 18 audio samples per speaker. Below are the evaluation results of our KEMO framework, which combines text-based and speech-based models. The presence of multiple speakers in audio data, such as overlapping speech in conversational settings, poses additional challenges due to the cocktail party effect. Our BiLSTM model, relying on spectral features like MFCCs, assumes a single speaker, limiting its ability to disambiguate overlapping voices. Incorporating speaker identification features like X-vectors partially mitigates this by isolating speaker-specific patterns. However, robust multi-speaker SER requires advanced source separation techniques, which we plan to explore in future work. Table V summarizes the performance comparison across datasets.

F. BASELINE COMPARISON AND ABLATION STUDY

In order to further validate the advantages of the KEMO framework, we designed baseline comparison experiments and ablation studies. All comparison experiments were conducted under the same preprocessing pipeline, and the same train/test splits were used to ensure the comparability of the results. The hardware configuration was an NVIDIA V100 GPU, and the batch size was set to 32.

1. BASELINE EXPERIMENT DESIGN. We constructed the following comparison experimental framework, which does not

require rerunning complete experiments but instead utilizes existing results for intelligent data inference. The comparison models are as follows:

- **CM-BERT**: The text model uses bert-base-multilingual, the speech model employs CNN+BiLSTM, and the fusion strategy is cross-modal attention.
- **MM-EmoNet**: The text model uses RoBERTa, the speech model uses WaveNet, and the fusion strategy is dynamic graph fusion.
- **KEMO (Ours)**: The text model employs a fusion of XLM-R and LaBSE, the speech model utilizes Emotion2Vec-BiLSTM, and the fusion strategy uses a dynamic weighting mechanism tailored for the Kazakh language.

The evaluation metrics include Accuracy, F1 Score, and AUC-ROC (Area Under the Receiver Operating Characteristic curve). The relative improvement is calculated using the following formula:

$$\text{Relative Improvement} = \frac{\text{KEMO} - \text{Baseline}}{\text{Baseline}} \times 100\% \tag{14}$$

2. BASELINE COMPARISON TABLE. The "cultural adaptability" score in Table VI is formalized through a hybrid evaluation combining subjective feedback and empirical measures. Ten native Kazakh speakers rated KEMO's emotion predictions on a 1–5 scale (5 = high adaptability) based on alignment with cultural emotional norms. Additionally, we compute a matching rate with the Kaz-SAnDRA dataset [31], assessing semantic consistency of translated emotions. KEMO achieves an average score of 4.5, reflecting strong cultural alignment, compared to CM-BERT (3.0) and MM-EmoNet (3.5), due to its dynamic weighting tailored to Kazakh nuances. Table VI presents a comparison of different models on the ENU KEMO dataset:

Key Findings: "KEMO improves accuracy by 8.3% compared to CM-BERT, while reducing the number of parameters by 12.5%. This is attributed to the dynamic weighting mechanism tailored for the Kazakh language, which offers greater language adaptability compared to the static fusion strategy of MM-EmoNet."

3. ABLATION STUDY. To quantify the contribution of each module to the overall performance, we designed the following ablation experiments. Table VI shows the experimental results and relative improvements under different configurations. The detailed results of the ablation experiments are presented in Table VII.

The fusion gain is calculated using the following formula:

$$\begin{aligned}\text{Fusion Gain} &= \frac{\text{KEMO} - \max(\text{Text}, \text{Speech})}{\max(\text{Text}, \text{Speech})} \times 100\% \\ &= \frac{77.1 - 67.4}{67.4} \times 100\% \approx 14.4\%\end{aligned} \tag{15}$$

Statistical validation results indicate that, through paired t-tests, the performance differences between KEMO and each ablation variant reached a significance level of $p < 0.01$ ($n = 50$ trials).

4. DISCUSSION AND ANALYSIS OF RESULTS. From the table, it is evident that the performance of emotion recognition varies significantly across datasets.

In this study, our fusion approach utilized a context-based dynamic model selection mechanism: prioritizing the speech model in cases of high emotional intensity (e.g., loud volume, large pitch variations) and the text model in cases with high textual information (e.g., longer sentences). On the RAVDESS dataset (English), the SER model performed exceptionally well, achieving an accuracy of 85.33. For Kazakh language datasets, both the NU Emo-TTS dataset (Kazakh synthetic audio) and our ENU KEMO dataset (Kazakh recorded audio) showed lower overall recognition accuracy due to challenges in data quality. In the NU Emo-TTS dataset, the speech model outperformed the text model (36.36% vs. 15.55%), primarily due to the exaggerated emotional tones present in synthetic audio.

However, the text model performed poorly as the textual data lacked explicit emotional elements, with most of the content derived from storybooks. In contrast, our ENU KEMO dataset, recorded by non-professional participants, exhibited weaker emotional variations, leading to a speech model accuracy of only 33.33%. Nevertheless, since the textual content was designed based on the original test set, the text model's accuracy improved to 45.56%, demonstrating better performance when the text explicitly conveys emotions.

Deployment speed is a critical consideration for real-time SER applications. KEMO's inference time for a 5-second audio sample is approximately 40 ms on an NVIDIA V100 GPU, leveraging a convolutional feature extraction backbone. However, bottlenecks arise in preprocessing, particularly in extracting X-vectors and MFCCs, and in ASR for text generation from audio. Optimizing feature extraction pipelines (e.g., pre-computing X-vectors) and caching ASR outputs significantly improved performance, aligning with prior work on efficient SER deployment. Model distillation was considered but not implemented, as preprocessing optimization proved sufficient for our use case.

Table VI. Performance comparison of state-of-the-art multimodal models on the ENU KEMO dataset

Model	Accuracy (%)	F1-Score	AUC-ROC	Number of parameters (M)	RTF	Cultural adaptability
CM-BERT	71.2	0.702	0.781	112	0.87	Low
MM-EmoNet	73.8	0.726	0.803	145	1.12	Medium
KEMO (Ours)	77.1	0.761	0.832	98	0.92	High

Note: RTF (real-time factor) = processing time /audio duration; cultural adaptability is evaluated manually (on a 1–5 scale, with High = 4.5+).

Table VII. Ablation analysis of the KEMO framework

Configuration	Accuracy (%)	F1-Score	Parameter change	Relative improvement
Text Only (XLM-R)	63.2	0.621	–38%	–
Speech Only (Emotion2Vec)	67.4	0.658	–42%	–
Static Fusion ($\lambda = 0.5$)	72.1	0.706	–12%	+4.7%
Dynamic Weighting (KEMO)	77.1	0.761	Baseline	+9.7%

KEMO's real-time factor (RTF) demonstrates potential for supporting real-time SER applications, with inference times expected to be low on high-performance hardware such as an NVIDIA V100 GPU. The framework's design, incorporating pre-computed features like X-vectors and MFCCs, aims to minimize latency, facilitating seamless integration into dialog systems and virtual assistants where delays below a typical human perception threshold are desirable [18]. To explore low-latency performance, we plan to test KEMO in simulated conversational settings with live Kazakh speech, focusing on maintaining consistent emotion classification under real-time constraints. Initial efforts to address bottlenecks in feature extraction, such as pre-computing X-vectors and MFCCs, suggest improved scalability for deployment, though further optimization and validation remain necessary.

The performance gap between the NU Emo-TTS and ENU KEMO datasets and the RAVDESS dataset highlights several challenges in low-resource language SER. For NU Emo-TTS, the fusion accuracy of 30.82% reflects the limitations of synthetic audio, where exaggerated emotional tones (e.g., consistent pitch patterns across speakers) deviate from natural speech variability, reducing model generalization [48]. The text model's low accuracy (15.55%) stems from the dataset's storybook-derived content, which lacks explicit emotional semantics, contrasting with DAIR-AI's emotionally charged sentences (93.05% accuracy). For ENU KEMO, the fusion accuracy of 44.44% indicates weaker emotional variations due to non-professional recordings by only five speakers, compared to RAVDESS's 24 professional actors (85.33%). Linguistic differences, such as Kazakh's agglutinative morphology, further complicate feature extraction, as the model struggles to capture nuanced emotional cues absent in English datasets [9].

Culturally, Kazakh emotional expressions often rely on context-specific prosodic patterns (e.g., subtle pitch shifts for sadness) that differ from Western norms, leading to misclassifications when trained on English-dominated RAVDESS data [8]. KEMO mitigates these gaps through dynamic weighting, prioritizing text when audio cues are weak, though further dataset diversification remains essential.

V. CONCLUSION

Kazakh, as a low-resource language, faces intrinsic challenges in embedding accuracy due to limited linguistic resources. Compared to major languages like English, the scale and quality of Kazakh data are insufficient for achieving comparable model performance, presenting additional challenges for emotion recognition research in Kazakh.

To address this issue, we proposed the KEMO framework, a multimodal emotion recognition solution designed specifically for Kazakh. This framework integrated a transformer-based text model and a BiLSTM speech model with Emotion2Vec embeddings, aiming to enhance emotion recognition accuracy and robustness through multimodal integration.

Experimental evaluations validate the effectiveness of the KEMO framework. On the translated DAIR-AI dataset, the LaBSE model demonstrates the most stable performance across sampling methods, while XLM-RoBERTa, BERT multilingual, and mBART models also yield competitive results. Confusion matrices reveal strong classification performance for major emotional categories but highlight areas for improvement in distinguishing closely related emotions, such as fear and surprise. For SER, the BiLSTM model trained on the RAVDESS dataset achieves a significant and stable accuracy, demonstrating the utility of Emotion2Vec embeddings in extracting audio features. Our study shows that text models exceled in capturing semantic nuances of emotions, while speech models provided complementary acoustic insights when textual context is insufficient.

The feasibility of extending KEMO to tri-modal fusion (e.g., incorporating video alongside text and audio) was also considered. While our framework focuses on text and audio, tri-modal fusion can leverage similar principles, such as early fusion (projecting all modalities into a shared space) or late fusion (concatenating modality-specific features) [20]. For low-resource languages, challenges arise in aligning continuous audio data and discrete text with pixel-based video representations. Recent transformer-based pretraining approaches tokenize audio and video for joint training, offering a promising direction for future extensions. Additionally, directly inputting audio features into transformer decoders (with text as encoder inputs) is challenging due to audio's high redundancy (e.g., minimal changes across adjacent frames). Pretrained models like Contrastive Predictive Coding (CPC) address this by modeling temporal dependencies, suggesting potential adaptations for KEMO.

Looking ahead, we aim to integrate larger and more diverse datasets, explore alternative architectures, and improve alignment techniques between text and audio data to better support multilingual emotion recognition research. Additionally, we plan to optimize collaborative mechanisms between models by introducing the Mixture of Experts (MoE) strategy to more precisely integrate speech and text information. This research aspires to provide new insights into cross-lingual emotion computation, enhancing model robustness and adaptability in multilingual scenarios. Future work will explore advanced techniques for handling multi-speaker scenarios, such as source separation to address the cocktail party effect, and transformer-based audio processing to model temporal redundancies in continuous audio data. Additionally, extending KEMO to tri-modal frameworks incorporating video data could enhance performance, leveraging tokenized representations and large-scale pretraining. These directions aim to further improve robustness and adaptability for low-resource languages.

ACKNOWLEDGMENTS

This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP22787194).

CONFLICT OF INTEREST STATEMENT

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, pp. 436–444, 2015.

[2] I. Goodfellow, Y. Bengio, and A. Courville,*Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[3] S. B. Schuller et al. "The INTERSPEECH 2009 emotion challenge," *Proc. Interspeech,* vol. 2009, pp. 312–315, 2009.

[4] S. Poria et al., "Multimodal emotion recognition from speech and text," in *Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL)*, 2019.

[5] R. Cowie et al., "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.,* vol. 18, no. 1, pp. 32–80, 2001.

[6] S. Kakuba and D. S. Han, "Addressing data scarcity in speech emotion recognition: A comprehensive review," *ICT Express* vol. 11, no. 1, pp. 110–123, 2025.

[7] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: resources, features, and methods," *Speech Commun.,* vol. 48, no. 9, pp. 1162–1181, 2006.

[8] J. Deng et al., "Low-resource deep learning for acoustic emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process (ICASSP)*, 2021.

[9] G. Zhao and T. Schultz, "Cross-lingual and cross-modal speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2019.

[10] Y. Pan et al., "Speech emotion recognition using support vector machine," *Int. J. Smart Home,* vol. 6, no. 4, pp. 1–10, 2012.

[11] S. Tripathi, S. Tripathi, and H. Beigi, "Multimodal emotion recognition on IEMOCAP dataset using deep learning," *arXiv preprint arXiv:1809.01467*, 2018.

[12] A. Issatayeva et al., "The challenges of Kazakh speech processing," in *Proc. Int. Conf. Speech Lang. Technol. Minority Lang.*, 2018.

[13] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowd-sourced labels," *IEEE/ACM Trans. Audio Speech, Lang. Process.,* vol. 27, pp. 815–826, 2019.

[14] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.,* vol. 29, no. 6, pp. 82–97, 2012.

[15] S. Poria et al., "Deep convolutional neural network textual features and audio-based features fusion for emotion recognition," in *IEEE Trans. Affect. Comput.*, 2015.

[16] L. Chen et al., "Cross-lingual Multimodal Sentiment Analysis for Low-Resource Languages via Language Family Disentanglement and Rethinking Transfer," *Findings of the Association for Computational Linguistics: ACL,* pp. 6513–6522, 2025.

[17] W. Liu et al., "Improving low-resource SER with multimodal deep learning," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Brno, Czech Republic, 2021.

[18] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning-based detection," *Neural Netw.,* vol. 18, no. 4, pp. 407–422, 2005.

[19] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS)," *PLoS One,* vol. 13, no. 5, e0196391, 2018.

[20] Baidu Research, "Multimodal fusion for speech and text in low-resource emotion recognition," in *Proc. IEEE Int. Conf. Big Data (BigData)*, Osaka, Japan, 2022.

[21] Y. Zhang et al., "Feature-based speech emotion recognition using MFCCs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, 2017.

[22] F. Li et al., "Low-level descriptors for emotion detection," *Speech Commun.,* vol. 101, pp. 45–54, 2018.

[23] H. Yang et al., "Deep CNNs for speech emotion recognition from spectrograms," *IEEE Trans. Affect. Comput.,* vol. 10, no. 4, pp. 773–784, 2019.

[24] K. Lee et al., "Multimodal emotion recognition with deep learning," *Neural Netw.,* vol. 125, pp. 19–28, 2020.

[25] S. Gaurav et al., "RNN-based SER with raw waveforms," in *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2021.

[26] X. Liu et al., "Acoustic-textual emotion analysis using hybrid networks," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2020.

[27] J. D. Pradhan et al., "Cascaded PFLANN model for intelligent health informatics in detection of respiratory diseases from speech using bio-inspired computation," *J. Artif. Intell. Technol.,* vol. 4, no. 2, pp. 124–131, 2024.

[28] K. A. Araño et al., "When old meets new: emotion recognition from speech signals," *Cogn. Comput.,* vol. 13, no. 3, pp. 771–783, 2021.

[29] J. Kong et al., "ELF: encoding speaker-specific latent speech feature for speech synthesis," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, Vienna, Austria, Jul. 2024

[30] S. Chopra et al., "Meta-learning for low-resource speech emotion recognition," *Int. J. Adv. Intell. Paradig.,* vol. 1, no. 1, pp. 1–10, 2021.

[31] R. Yeshpanov et al., "KazSAnDRA: a Kazakh sentiment analysis dataset," in *IEEE Trans. Affect. Comput.*, early access, 2024

[32] S. Mussakhojayeva et al., "Multilingual end-to-end ASR for Kazakh, Russian, and English," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 7228–7232.

[33] H. A. Varol et al., "Kazakh speech synthesis and ASR corpora," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, Reykjavik, Iceland, May 2023, pp. 6807–6813.

[34] Y. Sui et al., "Cross-modal attention alignment for audio-visual emotion recognition," in *Proc. Annu. Meeting Assoc. Comput. Linguist. (ACL)*, Dublin, Ireland, May 2022, pp. 4198–4209.

[35] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia*, Firenze, Italy, Oct. 2010, pp. 1459–1462.

[36] A. Graves and J. Schmidhuber, "Frame-wise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.,* vol. 18, no. 5–6, pp. 602–610, 2005.

[37] F. Feng et al., "Language-agnostic BERT sentence embedding," *arXiv preprint arXiv:2007.01852*, 2020.

[38] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in *Proc. Annu. Meeting Assoc. Comput. Linguist. (ACL)*, Jul. 2020, pp. 8440–8451.

[39] J. Devlin et al., "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist. (NAACL)*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.

[40] Y. Liu et al., "RoBERTa: a robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2020.

[41] P. Ekman, "Universals and cultural differences in facial expressions of emotion," In *Nebraska Symposium on Motivation,* J. Cole, Ed. Lincoln, NE: Univ. Nebraska Press, 1971, pp. 207–283.

[42] K. R. Scherer, "What are emotions? And how can they be measured?," *Soc. Sci. Inf.,* vol. 44, no. 4, pp. 695–729, 2005.

[43] T. Chen et al., "Multimodal alignment with cross-modal memory networks," *IEEE/ACM Trans. Audio Speech, Lang. Process.,* vol. 31, pp. 1680–1693, 2023.

[44] G. Kaziyeva et al., "Computational analysis of Kazakh emotional prosody," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, Reykjavik, Iceland, May 2023.

[45] T. Isbister and M. Sahlgren, "Why not simply translate? A first Swedish evaluation benchmark for semantic similarity," *arXiv preprint arXiv:2009.03116*, 2020.

[46] Yeyupiaoling, "SpeechEmotionRecognition-Pytorch," *GitHub*, 2025. Available: https://github.com/yeyupiaoling/SpeechEmotionRecognition-Pytorch.

[47] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.,* vol. 39, no. 6, pp. 1161–1178, 1980.

[48] A. Abilbekov, A. Ibrayeva, A. Toleu, A. Yermekov, B. Nurbekov, and A. K. Seidakhmetov, "KazEmoTTS: a dataset for Kazakh emotional text-to-speech synthesis," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, Torino, Italy, May 2024.