Layer-Wise Probing of Paralinguistic Attributes in Fine-Tuned Whisper for Kazakh Speech

Aimoldir Aldabergen
School of Sciences and Humanities, Nazarbayev University, Almaty, Kazakhstan
aimoldir.aldabergen@nu.edu.kz

Bakdaulet Kynabay
Faculty of Engineering and Natural Sciences, SDU University, Almaty, Kazakhstan
bakdaulet.kynabay@sdu.edu.kz

Shirali Kadyrov
Department of General Education, New Uzbekistan University, Tashkent, Uzbekistan
sh.kadyrov@newuu.uz (corresponding author)

Received: 20 December 2025 | Revised: 29 January 2026 and 4 February 2026 | Accepted: 7 February 2026

 DOI: https://doi.org/10.48084/etasr.17076

ABSTRACT

Large pre-trained speech models similar to Whisper are now commonly used for speech recognition and related tasks. The distribution of paralinguistic features, which include emotions and speaker characteristics across model layers, remains uncertain, particularly for low-resource languages. The current study evaluates each layer of the Kazakh-adapted Whisper encoder to determine its performance in recognizing emotional expression, speaker identity, age, and gender attributes. We extract fixed representations from every encoder layer and test them with both linear and Multilayer Perceptron (MLP) probes. The evaluation process uses accuracy, macro-averaged F1-score (Macro-F1), and balanced accuracy metrics, whereas non-parametric statistical tests evaluate the importance of changes across different layers. The experimental evaluation of KazEmoTTS focuses on emotional expression, whereas Common Voice (Kazakh) data serve for speaker identification and demographic attribute analysis. The results demonstrate that age and gender information are strongly present at all layers of the model with little change in representation across depths, yet speaker identity shows statistically significant but weak variations between layers. Emotion information appears mainly in the model's middle layers, which is the area where probing is most effective. The research findings reveal how Whisper processes Kazakh speech, allowing researchers to choose appropriate layers for paralinguistic speech applications.

Keywords-layer-wise probing; speech foundation models; Whisper; emotion recognition; speaker attributes; speaker identification; age prediction; gender prediction; low-resource languages; Kazakh speech

I. INTRODUCTION

Large-scale pre-trained speech models have become foundational components of modern speech processing systems, enabling strong downstream performance through representation reuse. While research on low-resource languages like Urdu has traditionally focused on using manually extracted prosodic features, such as pitch and intensity, to detect emotions [1], current studies increasingly explore large pre-trained models such as Whisper to identify these paralinguistic features within their hierarchical structure. However, the exact location of emotional and speaker-related attributes within these networks remains unclear. Researchers need to determine if intermediate layers generate better paralinguistic prediction features than the final layer, which is typically optimized for linguistic transcription. The internal representations of neural networks become accessible through probing-based analysis, which employs frozen layer-wise embeddings to train lightweight classifiers for quantifying linear and non-linear information accessibility. Prior work on self-supervised speech encoders suggests that paralinguistic information is distributed non-uniformly across depth and often peaks in intermediate layers, motivating systematic layer-wise evaluation rather than defaulting to top-layer features [2, 3]. Unlike authors in [4], who employed mutual information analysis to select acoustic features, we leverage the hierarchical nature of the Whisper encoder to analyze how different abstraction levels contribute to paralinguistic attribute detection.

Extensive benchmarking on models like wav2vec 2.0, HuBERT, and WavLM has demonstrated that learned weighted sums or single best layers frequently outperform the final layer, underscoring that layer selection is essential to characterize information flow [5, 6]. For instance, a multilingual Speech

Emotion Recognition (SER) study showed that a single intermediate layer outperforms both all-layer pooling and final-layer features for maximum emotion-relevant cue linear separability [7]. Similarly, analysis using Canonical Correlation Analysis (CCA) has shown that the most useful layers for non-linguistic features often reside below the top layers [6]. Authors in [8] further revealed that distinct network layers encode different signal types, and that models like WavLM and Whisper process affective information through divergent hierarchical pathways. Speaker-related features appear to follow a similar structure. Research indicates that lower layers focus on timbre and prosody, whereas intermediate layers merge speaker attributes like gender before the upper layers shift attention to language content. Although specialized encoders still outperform foundation models for pure speaker identification [9], self-supervised encoders demonstrate strong speaker information encoding abilities, often producing convex performance curves across layers [10]. Despite these advances, a gap remains regarding the Whisper model, particularly for low-resource languages. While models like wav2vec 2.0 have been extensively mapped, Whisper lacks a similarly comprehensive systematic analysis focused on paralinguistic features [8, 10]. Moreover, research in multilingual probing has predominantly focused on English, leaving major low-resource language groups like Turkic, and specifically Kazakh, underrepresented. A focused probing study of Whisper on Kazakh speech addresses this clear gap, advancing the understanding of how widely used Automatic Speech Recognition (ASR) encoder–decoders distribute paralinguistic information outside high-resource contexts.

Our previous research involved fine-tuning Whisper for Kazakh ASR, which brought major advancements in transcription accuracy [11]. This model serves as the basis for the current representation probing. In this study, individual layers of the model are evaluated by applying both linear and Multilayer Perceptron (MLP) heads to assess emotional content, speaker identity, and demographic features including age and gender. To ensure the robustness of our conclusions, we supplement performance comparisons with paired statistical significance testing across layers. Our evaluation uses two complementary Kazakh speech resources: KazEmoTTS [12] for emotion-labeled data and Common Voice (Kazakh) [13] for speaker and demographic information. Collectively, these experiments clarify how Whisper represents paralinguistic features in a low-resource Turkic language, offering guidance on feature extraction and layer selection for practical applications.

II. METHODOLOGY

A. Models and Representations

The analysis of paralinguistic and sociolinguistic information in Kazakh speech is based on the internal representations of a Transformer-based encoder derived from the Whisper architecture.

1) Encoder Adaptation and Freezing

This study is grounded in the Whisper-Small encoder architecture. We specifically utilize a version that was previously fine-tuned on a large-scale Kazakh speech dataset, as detailed in our prior work [11]. We selected this specific fine-tuned model because it represents the current state-of-the-art for Kazakh ASR, offering a more practical baseline for low-resource deployment than the original multilingual model. The process of fine-tuning modified the model's weights to better capture Kazakh phonetic and acoustic characteristics, enabling the model to generate useful representations for the target tasks.

To isolate the knowledge acquired by the encoder during pre-training and fine-tuning, the weights of the adapted Whisper encoder were strictly frozen throughout the representation extraction and subsequent probing. This approach prevents simple probing tasks from influencing the encoder's internal state, enabling an accurate assessment of the readily available information at each layer.

2) Layer-Wise Representation Extraction

The core investigation examines how different processing levels in the frozen Whisper encoder store information. The process involves obtaining hidden states from every Transformer block, which functions as an encoder component in the model. The model operates with twelve encoder layers as its fundamental structure. Twelve vector embeddings were extracted from each acoustic segment using the final hidden-state output of each layer. Since the encoder outputs a sequence of frame-level vectors, mean-pooling across the temporal dimension was applied to aggregate them into a single fixed-size embedding of dimensionality 768 for each utterance. These extracted vectors served as the input feature space for the downstream probing classifiers. The hierarchical extraction strategy enables robust comparative analysis, testing the hypothesis that lower network layers produce acoustic and phonetic features, whereas deeper layers generate complex features about speaker identity, age, gender, and emotional tone.

B. Datasets

Two independent, publicly available Kazakh speech corpora were utilized to evaluate the layer-wise representations: KazEmoTTS [12] and Common Voice (Kazakh) [13]. The data selection process involved identifying specific subsets, which were subsequently divided according to probing task requirements to prevent data leakage. The distribution of dataset labels prior to data splitting is shown in Figure 1.

- KazEmoTTS [12]: For emotion classification, a total of 1,381 audio recordings from the KazEmoTTS corpus were employed. As a controlled emotion-synthesis dataset featuring high-quality recordings annotated with six distinct emotional labels, it is well-suited for affective probing.
- Common Voice (Kazakh) [13]: A validated subset of this corpus, comprising approximately 1.04 h of speech from 31 unique speakers (806 audio recordings), was utilized. This subset supports demographic and identity probing with two gender labels and five age group labels. The validation process ensured that all selected clips met audio quality and speaker variety standards.

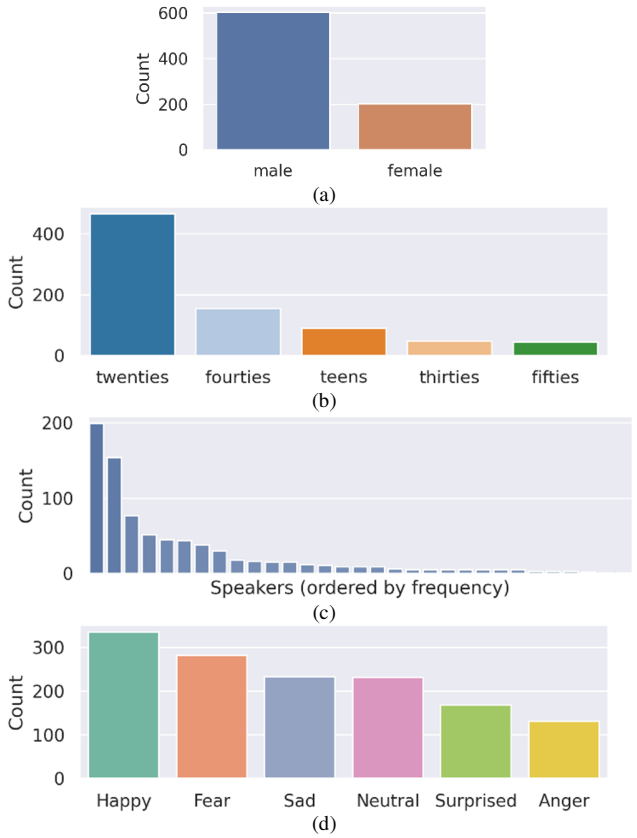

Fig. 1. Distribution of dataset labels used for probing: (a) gender, (b) age, (c) speaker identity, (d) emotional expression.

1) Experimental Setup and Data Splitting

To evaluate encoder generalization effectively, two data splitting schemes were applied based on the task characteristics. In all instances, the data were partitioned into 80% for training and 20% for testing.

- Age and gender probing: A speaker-independent split was implemented, ensuring that no individual appeared in both the training and testing sets. This stratification by age group prevents the model from memorizing specific voice characteristics, thereby forcing the learning of generalized demographic features.
- Speaker identity: A stratified utterance-level split was utilized, ensuring that all 31 speakers were represented in both the training and testing sets to evaluate recognition on unseen audio segments. It is important to note that this experimental design assesses the model's capacity for closed-set speaker identification (distinguishing among known speakers) rather than open-set generalization to unseen speakers.
- Emotion classification: The KazEmoTTS dataset was partitioned using stratified sampling to preserve balanced emotional class distributions.

C. Probing Tasks and Models

A controlled probe-learning approach was adopted to assess the amount of target information present in the frozen layer-wise embeddings. Two distinct probe architectures were trained on the embeddings derived from the encoder layers:

1. Linear probes: These were employed to determine the extent to which target attributes are linearly separable (explicitly represented) within the embeddings.
2. MLP probes: Shallow MLPs, consisting of two dense layers with ReLU activation, were used to gauge the accessibility of attributes when considering non-linear transformations.

To ensure that performance differences reflect the quality of the layer-wise representations rather than the capacity of the classifier, probe complexity was strictly constrained. By limiting the MLP to a shallow architecture, we prioritize the evaluation of information accessibility within the frozen encoder features over the learning capability of the downstream model.

The training process employed the Adam optimizer with a learning rate of 1e-4 for 50 epochs, utilizing early stopping with a patience of 5 epochs to prevent overfitting. Regularization was applied via dropout ($p = 0.1$). To ensure a fair and comparative analysis across layers and tasks, representations were extracted and learning parameters fixed prior to training, with an equivalent random seed used for all train/test splits.

Probe performance on the test sets was evaluated using standard classification metrics, including accuracy, macro-averaged F1-score (Macro-F1), and balanced accuracy. Standard accuracy serves as a baseline, whereas Macro-F1 and balanced accuracy provide a robust assessment specifically for cases of class imbalance. By treating all classes equally and accounting for both sensitivity and specificity, these metrics ensure that the reported performance reflects a fair estimate across all demographic and emotional categories.

D. Statistical Analysis

To test whether performance on probes was different according to the layer of representation and the type of probe paired, non-parametric tests were used that formally account for the repeated-measures design.

To evaluate overall layer effects, the Friedman test was used, which is preferred over other non-parametric tests in situations involving multiple related samples and for which normality assumptions are not tenable [14]. The size of these layer effects was measured using Kendall's W, which allows description of the degree of layer differences in size [15]. For those layers in which a significant overall layer effect was found using the Friedman test, subsequent comparisons of layer pairs were conducted using Wilcoxon signed-rank tests with Holm correction for family-wise error rate [14].

III. RESULTS

A. Layer-Wise Probing Performance

Figures 2(a) and 2(b) show the layer-wise probing performance for age prediction across Whisper layers using accuracy, Macro-F1, and balanced accuracy. Moderate variation is observed across layers, with no single layer dominating all metrics. Accuracy gradually increases through the layers, reaching maximum values in the latter layers, whereas Macro-F1 and balanced accuracy metrics peak around the mid-to-late layers (L7–L8). In the initial layers, the values for both Macro-F1 and balanced accuracy are relatively lower, indicating that there is limited separation based on the age attribute in the early layers. The differences between accuracy and the other two metrics indicate that improvement across classes is not uniform in the latter layers when considering raw accuracy alone. From the radar charts in Figures 2(a) and 2(b), it can be inferred that age-related information is not confined to a single layer but is well represented in the mid-layers.

Figures 2(c) and 2(d) plot layer-wise probing results for gender prediction tasks using accuracy, Macro-F1, and balanced accuracy. For almost all layers, linear and MLP probes demonstrate a ceiling effect, with accuracy and balanced accuracy approaching 1.0. Macro-F1 values demonstrate a slight decrease at a few layers (L1 and L11, among others), possibly reflecting class imbalance in gender distributions. Overall, the radar charts suggest that gender information is effectively encoded linearly at all levels, independent of the layer position.

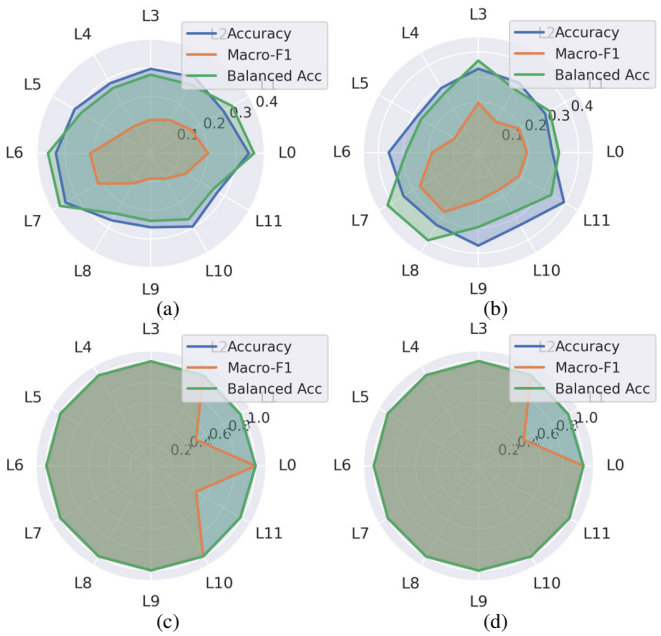

Fig. 2. Layer-wise probing performance for age and gender prediction across Whisper layers and probe types: (a) age linear, (b) age MLP, (c) gender linear, (d) gender MLP.

Figures 3(a) and 3(b) show the probing performance for speaker identification tasks using accuracy, Macro-F1, and balanced accuracy metrics in the Whisper model. Linear and MLP probes perform well in most layers, with near-ceiling accuracy and balanced accuracy in the earlier to mid-layers (approximately L2–L5). Macro-F1 values are also observed to peak in the same layers, suggesting a strong separation between speakers in these layers. Performance gradually declines in later layers, with a prominent drop in Macro-F1, suggesting a lack of class-wise robustness in these layers. Although MLP probing is not consistently better than linear probing, its curves appear slightly smoother. The radar charts indicate that speaker information is strongly present in all layers, with earlier to mid-layers showing more distinct separation than later layers.

Figures 3(c) and 3(d) depict the performance of layer-wise probing for emotion prediction in Whisper layers based on accuracy, Macro-F1, and balanced accuracy metrics. For emotion recognition, performance exhibits strong layer-wise variability compared to age, gender, or speaker identification tasks. Although early layers show relatively low performance, it steadily improves up to intermediate layers (L4–L6) before declining in deeper layers, particularly for Macro-F1. Also, it can be observed that for all layers, except possibly a few early ones, the MLP probe performs better compared to the linear probe. However, the performance maxima for both linear and MLP probes coincide at intermediate layers. The radar plots in Figures 3(c) and 3(d) reveal that emotion-related information is concentrated in the intermediate representations of the model rather than being distributed uniformly across all layers.

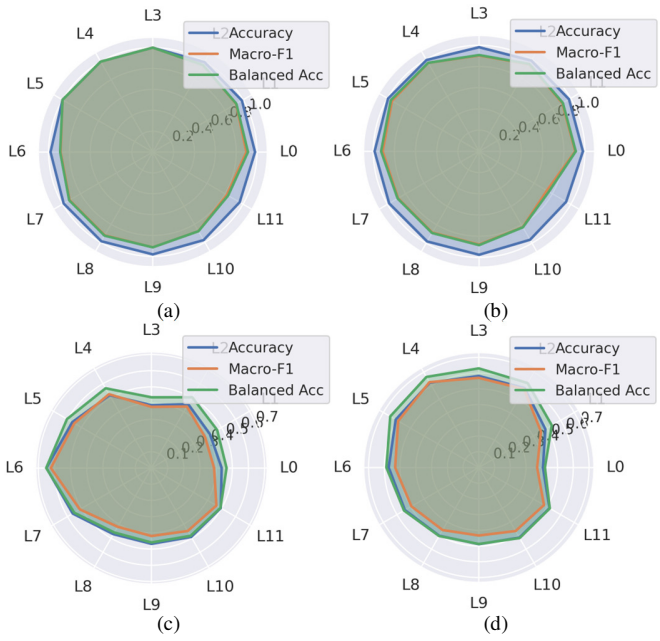

Fig. 3. Layer-wise probing performance for speaker identification and emotion prediction across Whisper layers and probe types: (a) speaker linear, (b) speaker MLP, (c) emotion linear, (d) emotion MLP.

B. Comparison of Probe Types

Figure 4 compares layer-wise Macro-F1 patterns for linear and MLP probes across all tasks. Overall, both probe types exhibit broadly similar trends, with no systematic advantage of MLP probing across all layers and attributes. For age

prediction, the MLP probe shows moderately higher Macro-F1 values in mid-to-late layers (e.g., L7–L8), whereas the linear probe exhibits more variability and lower overall performance.

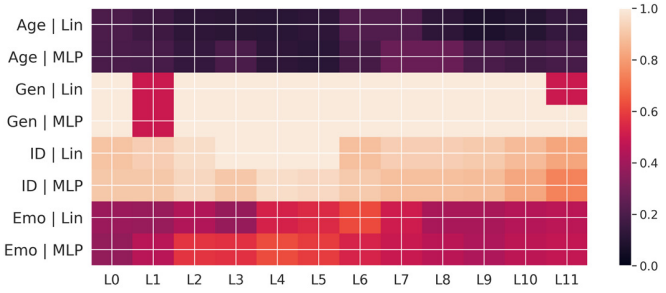

Fig. 4. Layer-wise Macro-F1 across tasks and probe types.

In contrast, for gender classification, linear and MLP probes yield nearly identical Macro-F1 profiles across layers, indicating that gender-related information is linearly accessible and stable throughout the network. For speaker identification, linear probing performs strongly across most layers, whereas the MLP probe shows slightly reduced performance in later layers, suggesting that non-linear transformations do not substantially enhance speaker-specific representations. Taken together, Figure 4 indicates that MLP probes do not fundamentally alter layer-wise trends but instead modestly amplify existing patterns for attributes that are already present in the representations.

C. Attribute-Specific Patterns

Figure 4 highlights clear distinctions between demographic attributes, such as gender and speaker identity, and emotion-related representations. Macro-F1 scores are high in most layers for both gender and speaker identity, suggesting a relatively uniform encoding of speaker-related information. Emotion recognition, on the other hand, shows clear layer dependency, with both linear and MLP probe scores peaking in layers L4–L6 and gradually decreasing in later layers. The linear probe scores, in particular, follow this pattern, consistent with statistical differences observed in post-hoc analysis for layers L4 and L6. Unlike demographic attributes, the encoding of emotion-related information seems to follow a relatively heterogeneous pattern in most layers, highlighting the importance of intermediate representations that capture prosodic and paralinguistic features.

D. Statistical Analysis

Table I summarizes the Friedman tests conducted to determine the overall layer differences for all tasks and probe types. Regarding the prediction of age and gender attributes, neither linear nor MLP probes exhibited significant layer differences, as indicated by non-significant p-values and effect sizes close to zero (Kendall's W $\approx$ 0.01–0.02). In speaker identification tasks, the omnibus layer differences were statistically significant for both probe types, although the effect sizes remained relatively small. Conversely, emotion detection tasks showed highly significant layer differences, with larger, though still moderate, effect sizes for both linear and MLP probes.

TABLE I. FRIEDMAN TESTS FOR OVERALL LAYER EFFECTS

Task	Probe	Metric	χ^2	Df	p-value	Kendall's W
Age	Linear	Accuracy	18.0	11	0.081	0.020
Age	MLP	Accuracy	10.3	11	0.504	0.012
Gender	Linear	Accuracy	10.0	11	0.530	0.011
Gender	MLP	Accuracy	11.0	11	0.443	0.012
Speaker ID	Linear	Accuracy	26.7	11	0.005	0.015
Speaker ID	MLP	Accuracy	26.6	11	0.005	0.015
Emotion	Linear	Accuracy	102.7	11	0.000	0.031
Emotion	MLP	Accuracy	88.1	11	0.000	0.027

For the linear speaker identification probe, the Friedman test revealed a statistically significant overall effect for layers ($\chi^2(11) = 26.75$, $p = 0.005$, Kendall's $W = 0.015$), as presented in Table I. However, post-hoc comparisons via Wilcoxon signed-rank tests with Holm correction did not identify any significant differences between specific layer pairs (all adjusted p-values ≥ 0.94). This pattern suggests minimal performance disparities between layers, despite the marginal global effect.

In contrast, emotion recognition demonstrated a strong layer-dependent structure. The Friedman tests indicated a significant overall effect of the representation layer for both linear and MLP probes. Given the omnibus significance, follow-up Wilcoxon signed-rank tests with Holm correction were performed. Significant differences were primarily observed between mid-layers (e.g., L6) and early or late layers (adjusted $p = 7.1 \times 10^{-9}$), as well as between early layers (e.g., L0) and mid-to-late layers for both probe types, with adjusted p-values falling below 10^{-6} in several instances. These results collectively imply a strong layer dependency for emotion, in contrast to the layer-independent behavior observed for age and gender.

E. Limitations

This study provides a detailed layer-wise mapping of the Whisper encoder, but the characteristics of the datasets impose certain constraints on generalization. First, the KazEmoTTS corpus consists of controlled, synthesized emotional speech, which often exhibits exaggerated prosodic contours and lacks the background noise and subtle linguistic nuances found in spontaneous, naturalistic communication. Consequently, the distinct separation of emotional attributes observed in the intermediate layers may be less pronounced in real-world scenarios.

Second, the datasets used are relatively small, particularly the validated Common Voice (Kazakh) subset. The speaker identification results, although statistically significant, reflect the model's ability to discriminate between a closed set of speakers within a limited audio duration. Caution should be exercised when extrapolating these findings to open-set identification tasks or larger populations. Future validation on diverse, natural speech corpora is necessary to confirm that the observed architectural biases are intrinsic to the model rather than artifacts of specific recording conditions.

IV. CONCLUSION

This study presents a layer-wise probing analysis of the Kazakh-adapted Whisper model, characterizing how affective

and speaker-centric attributes are encoded across the network. The results reveal distinct encoding strategies: demographic features (age, gender) and speaker identity exhibit robust, distributed representations across all layers, whereas affective attributes follow a convex trajectory, peaking significantly in the intermediate layers. This pattern indicates that paralinguistic emotional signals are accessible at specific abstraction levels before being suppressed in the final output layers.

The primary novelty of this work lies in providing the first systematic mapping of paralinguistic feature distribution within a foundation model for the Kazakh language. These findings confirm that middle-layer retention of affective cues is a cross-lingual phenomenon that persists even after fine-tuning in low-resource contexts. Practically, this underscores that reliance on final-layer representations, standard in many Automatic Speech Recognition (ASR) pipelines, is suboptimal for emotion recognition.

Finally, this study focuses exclusively on the representations within the fine-tuned Kazakh-adapted Whisper model. The analysis does not explicitly compare these representations with the original pre-trained multilingual Whisper baseline. While such a comparison would shed light on the dynamics of transfer learning and catastrophic forgetting during fine-tuning, it remains outside the scope of this work. Future work will therefore address this comparison, as well as extend the analysis to cross-lingual generalization and causal probing, while specifically validating these patterns on larger-scale, spontaneous speech datasets.

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.

REFERENCES

[1] S. Khan, S. A. Ali, and J. Sallar, "Analysis of Children's Prosodic Features Using Emotion Based Utterances in Urdu Language," *Engineering, Technology & Applied Science Research*, vol. 8, no. 3, pp. 2954–2957, June 2018, https://doi.org/10.48084/etasr.1902.

[2] Z. Zhu and Y. Sato, "Deep Investigation of Intermediate Representations in Self-Supervised Learning Models for Speech Emotion Recognition," in *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops*, Rhodes Island, Greece, 2023, pp. 1–5, https://doi.org/10.1109/ICASSPW59220.2023.10193018.

[3] M. Kim, J. S. Um, and H. Kim, "How Far Do SSL Speech Models Listen for Tone? Temporal Focus of Tone Representation under Low-resource Transfer." arXiv, Jan. 25, 2026, https://doi.org/10.48550/arXiv.2511.12285.

[4] H. Roubhi, A. H. Gharbi, K. Rouabah, and P. Ravier, "Mutual Information-based Feature Selection Strategy for Speech Emotion Recognition using Machine Learning Algorithms Combined with the Voting Rules Method," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 19207–19213, Feb. 2025, https://doi.org/10.48084/etasr.9066.

[5] S. Yang *et al.*, "A Large-Scale Evaluation of Speech Foundation Models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2884–2899, 2024, https://doi.org/10.1109/TASLP.2024.3389631.

[6] A. Pasad, B. Shi, and K. Livescu, "Comparative Layer-Wise Analysis of Self-Supervised Speech Models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, 2023, pp. 1–5, https://doi.org/10.1109/ICASSP49357.2023.10096149.

[7] A. Singh and A. Gupta, "Decoding Emotions: A comprehensive Multilingual Study of Speech Models for Speech Emotion Recognition." arXiv, Aug. 17, 2023, https://doi.org/10.48550/arXiv.2308.08713.

[8] S. G. Upadhyay, C. Busso, and C.-C. Lee, "A Layer-Anchoring Strategy for Enhancing Cross-Lingual Speech Emotion Recognition," presented at the Proceedings of Interspeech 2024, Kos Island, Greece, 2024, pp. 4693–4697, https://doi.org/10.21437/Interspeech.2024-469.

[9] A. Y. F. Chiu, K. C. Fung, R. T. Y. Li, J. Li, and T. Lee, "A Large-Scale Probing Analysis of Speaker-Specific Attributes in Self-Supervised Speech Representations." arXiv, Sept. 18, 2025, https://doi.org/10.48550/arXiv.2501.05310.

[10] A. Waheed, H. Atwany, B. Raj, and R. Singh, "What Do Speech Foundation Models Not Learn About Speech?" arXiv, Oct. 16, 2024, https://doi.org/10.48550/arXiv.2410.12948.

[11] B. Kynabay, A. Aldabergen, S. Kadyrov, and A. Shalkarbay-Uly, "Fine-tuning OpenAI's Whisper Model for Kazakh Speech Recognition." ResearchGate, Dec. 13, 2025, https://doi.org/10.13140/RG.2.2.29371.07205.

[12] A. Abilbekov, S. Mussakhojayeva, R. Yeshpanov, and H. A. Varol, "KazEmoTTS: A Dataset for Kazakh Emotional Text-to-Speech Synthesis," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Torino, Italia, 2024, pp. 9626–9632.

[13] R. Ardila *et al.*, "Common Voice: A Massively-Multilingual Speech Corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, 2020, pp. 4218–4222.

[14] M. Hollander, D. A. Wolfe, and N. Chicken, *Nonparametric Statistical Methods*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2013.

[15] E. Tomczak and M. Tomczak, "The need to report effect size estimates revisited. An overview of some recommended measures of effect size," *Trends in Sport Sciences*, vol. 21, no. 1, pp. 19–25, Feb. 2014.