METHODS FOR IDENTIFYING SPEECH FEATURES OF WRITTEN SPEECH AND EMOTIONAL STATES OF A PERSON

Author: Anastasiia Matveeva
PhD student
ITMO University, Saint Petersburg, Russia.

Abstract: Contemporary automated speech analysis systems face several fundamental limitations, including the fragmented analysis of linguistic and emotional characteristics, a focus on a limited set of basic emotions, and insufficient adaptation to diverse language systems. This study proposes a comprehensive approach to speech feature analysis, consisting of two complementary components designed to overcome these limitations. The proposed framework comprises two complementary methods. The first method is designed to assess the level of speech activity and is based on a comprehensive analysis of 36 linguistic parameters. These parameters include quantitative features (mean utterance length, number of sentences), syntactic features (construction complexity), lexical diversity (vocabulary richness), as well as the frequency of various parts of speech. The Mann-Whitney U test is employed to classify the level of speech activity. The second method is a state-of-the-art multi-task neural network architecture based on the pre-trained RuBERT-large language model. This architecture is capable of simultaneously evaluating three fundamental parameters of emotional state: valence, arousal, and dominance. The combination of these parameters enables the identification of 26 complex emotional states. Experimental validation on heterogeneous Russian- and English-language corpora, including professional speech and spontaneous dialogues, demonstrated the superiority of the proposed methods over existing analogues. The speech activity assessment method achieved an accuracy of 92% for English and 89% for Russian. The multi-task model attained an accuracy of 85% in determining valence, 80% for arousal, and 76% for dominance. For 10 main emotional categories, the classification accuracy reached 57%. The results of this study can be applied to the development of intelligent dialogue systems and chatbots capable of adapting their communication style and emotional responses based on the user's speech activity and emotional state, thereby significantly enhancing the quality and naturalness of interaction.

Index Terms: **Speech feature analysis, Emotional States, Multi-task Learning**

1. INTRODUCTION

The analysis of speech characteristics and human emotional states represents a significant and rapidly evolving task within the fields of natural language processing and artificial intelligence. Written discourse constitutes a complex system involving the interaction of multiple levels of linguistic organization, which can be broadly categorized into two main types: static (stable) and dynamic (context-dependent) features.

Static features, which remain consistent across different contexts due to their connection to individual language habits and cognitive patterns, include:

- Lexical characteristics — word choice, part-of-speech frequency;
- Morphological features — the use of grammatical forms;
- Syntactic features — sentence structure.

These slowly evolving parameters enable the analysis of a person's speech activity. Speech activity represents a projection of fundamental linguistic parameters and demonstrates how lexical, morphological, and syntactic features are realized within a specific communicative context. High speech activity is characterized by spontaneity and a rich vocabulary, particularly emotionally charged lexicon, whereas low speech activity manifests in structural organization.

In contrast to static characteristics, semantic features are dynamic, as their interpretation is highly dependent on context and the communicative situation. For instance, the same words can convey different meanings and emotional connotations depending on the circumstances of communication.

The analysis of speech characteristics finds application in various domains—from the development of intelligent dialogue systems to psycholinguistic research. Currently, however, this analysis faces a number of significant limitations and unresolved challenges. Firstly, many existing methods examine speech features in isolation, failing to account for their deep interconnections. Secondly, most approaches dealing with semantic features are oriented towards a limited set of basic emotions, overlooking complex, multidimensional emotional states. Thirdly, there is a pronounced scarcity of methods adapted for different languages, particularly for Russian.

This article proposes a comprehensive framework for the analysis of speech characteristics, comprising two complementary components: a method for analyzing speech activity through a comprehensive assessment of 36 linguistic features, and a multitask emotion recognition model based on the RuBERT-large architecture. Its key innovation is the simultaneous analysis of three fundamental psycholinguistic parameters—valence,

arousal, and dominance—enabling the identification of complex emotional states through a system of 26 categories.

2. LITERATURE REVIEW

Contemporary research on speech characteristics and textual emotional coloring constitutes a dynamically evolving interdisciplinary field, integrating methods from computational linguistics, psycholinguistics, and artificial intelligence. This work systematizes existing approaches to text analysis across various linguistic levels.

The lexical level demonstrates the most noticeable distinctions between functional styles. As shown in studies [1], the application of machine learning methods enables the identification of specific lexical patterns in the speech of various groups, including children with autism spectrum disorder and Down syndrome. Foundational works [2] established the basis for the systematic analysis of lexical markers, while modern language models like BERT [3] achieve 89% accuracy in recognizing authorial style, despite challenges related to computational efficiency. The text data augmentation method developed in [4] allows for the preservation of individual lexical features during the processing of training datasets.

At the morphological level, key markers are grammatical forms (verb tenses, cases), which reflect the author's stylistic preferences. Academic texts are characterized by a predominance of deverbal nouns and complex adjectives, emphasizing formal exposition, whereas journalistic writing employs action verbs and modal constructions to create dynamism.

Syntactic analysis reveals substantial differences in text organization. As research [5] demonstrates, neural network architectures based on BERT and RoBERTa effectively process complex syntactic structures, achieving an F1-score of 0.78 in the analysis of dialogic sequences. However, as rightly noted by [6], automatic syntactic analysis requires careful ethical oversight to avoid reinforcing stereotypes. The work [7] provides a comprehensive review of contemporary syntactic analysis methods based on transformers, which supports our selection of BERT as the base architecture.

Semantic analysis has evolved from simplified models of basic emotions [8] to multidimensional approaches that consider valence, arousal, and dominance [9]. Modern

transformer architectures demonstrate high efficacy in recognizing complex emotional states, including sarcasm and irony.

The present study advances these directions by proposing a comprehensive framework for the analysis of speech characteristics.

3. DATASETS

As prospective datasets for identifying speech characteristics associated with high speech activity, profession-specific corpora were selected. Given the current absence of datasets explicitly categorised by high and low speech activity, a compromise approach was adopted, grounded in the hypothesis of professional determination of speech behaviour [10, 11].

This approach is consistent with the findings of [12], which demonstrated that professional status is a key sociolinguistic factor determining variability in speech behaviour.

As a proxy for high speech activity in English, public speeches by politicians were chosen. Conversely, professional athletes (ice hockey players) were selected as representatives of low speech. The National Hockey League Interviews and US 2020 Presidential Election Speeches datasets contain substantial textual data: 275 speech transcripts and 2,087 interviews, respectively.

For the Russian language, datasets compiled from discussions of two professional groups were used: Natural Language Processing (NLP) specialists, representing a profession with low speech activity, and project managers, representing high speech activity. Their professional duties directly influence the nature and intensity of their speech practices. For each dataset, 1500 messages were collected from relevant thematic forums.

For emotion recognition, a new multimodal Russian-language corpus of polylogues with comprehensive annotation across four parameters was utilised: valence (positive, neutral, negative), arousal (aroused, neutral, inhibited), dominance (high, neutral, low), and categorical emotions (26 categories, including delight, joy, embarrassment, satisfaction, etc.). The corpus volume is 662 minutes (11 hours) and 7288 utterances, attesting to its representativeness. The average audio sample length is 7 seconds, with a maximum of 15 seconds, which is significant for subsequent research into emotional states in speech.

Table 1 — Approaches to the Analysis of Stylistic Features at Different Linguistic Levels

Dataset Purpose	Language	Dataset Name	Description and Volume	Reason for Selection
Speech Activity Analysis	English	US 2020 Presidential Election Speeches	275 transcripts	**High Activity:** Public speaking, persuasive discourse
	English	National Hockey League Interviews	2,087 interviews	**Low Activity:** Brief, post-game interviews
	Russian	Project Managers (Forums)	1,500 messages	**High Activity:** Coordination, presentations
	Russian	NLP Specialists (Forums)	1,500 messages	**Low Activity:** Technica task-focused communication
Emotion Recognition	Russian	Multimodal Polylogue Corpus 275 transcripts	7,288 utterances (11 hours, 662 min.)	Comprehensive annotation for valence, arousal, dominance, and 26 emotion categories

4. METHODOLOGY

The method for identifying changes in speech characteristics based on verbal activity level is grounded in the concept of verbal activity as a key parameter reflecting an individual's degree of engagement in the communication process through the frequency and intensity of speech use. Verbal activity represents a dynamic projection of fundamental linguistic parameters and demonstrates how lexical, morphological, and syntactic features are realized in a specific communicative context. For instance, the lexical profile of high verbal activity is characterized by a prevalence of first-person pronouns and emotionally charged words, indicating spontaneity, whereas terminological precision is more typical of low activity. At the morphological level, the use of perfective aspect verbs or genitive case constructions reflects not only the speaker's cognitive predispositions but also their level of verbal activity, demonstrating a result-oriented focus or a fixation on limitations, respectively.

High verbal activity is marked by frequent and intensive speech, a rich lexical repertoire including slang and idiomatic expressions, pronounced emotional coloring, and rapid topic shifts. In contrast, low verbal activity is distinguished by restraint, lexical minimalism, and structural organization of utterances.

The core of this method for identifying speech characteristic changes based on the degree (high or low) of verbal activity is as follows: based on training data, a profile of speech characteristics typical

for individuals with high and low verbal activity is compiled. Subsequently, any utterance of interest can be assessed for its similarity to a given activity level by comparing its feature profile to the pre-established activity-level profiles. The Mann-Whitney U test [13] is employed as the criterion for evaluating this similarity. A schematic of the method is presented in Figure 1.

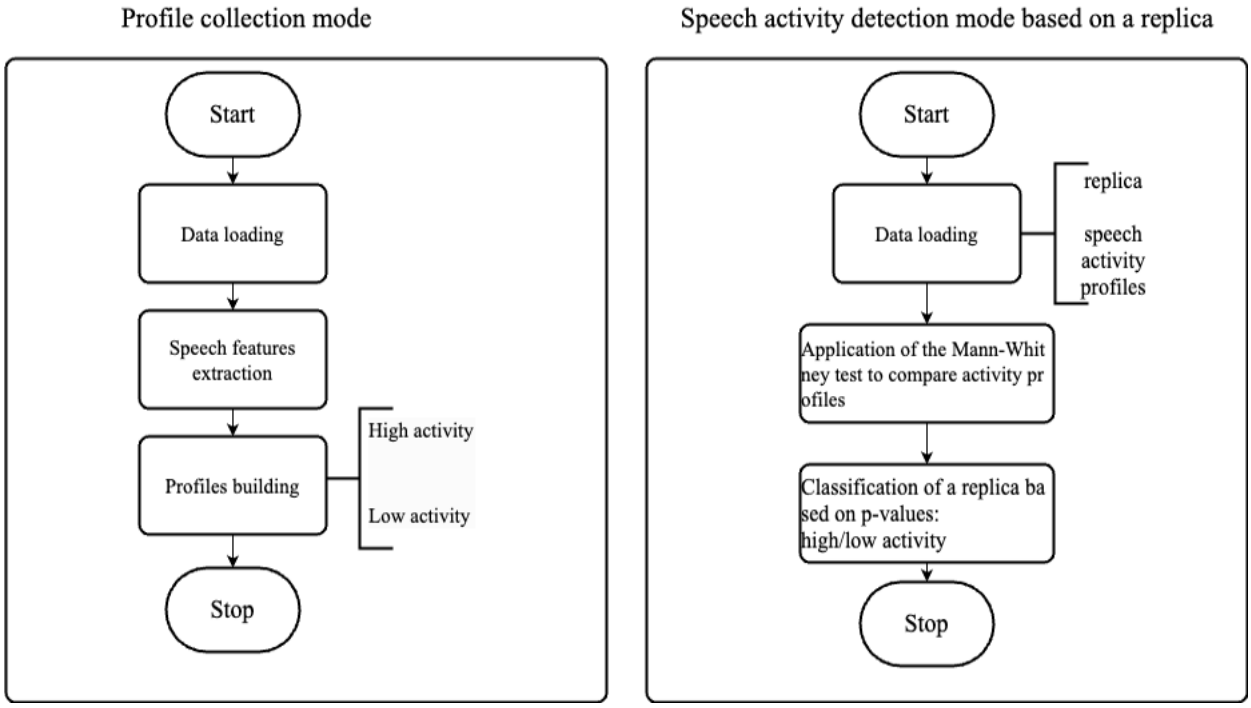

Figure 1 — A Method for Identifying Changes in Speech Characteristics under the Influence of Verbal Activity

The method implements a statistical approach to determine the degree of verbal activity (high/low) by comparing an utterance's characteristics with reference profiles formed from training data. The procedure begins with data collection, where key speech characteristics are extracted for each utterance. These include general text statistics, such as the average number of words per utterance, average number of sentences, average word length in letters, and frequency distributions of parts of speech according to Penn Treebank tags. The analysis encompasses singular and plural nouns, proper nouns, personal and interrogative pronouns, verbs in different tenses and moods, adjectives in positive, comparative, and superlative degrees, adverbs, prepositions, conjunctions, particles, modal verbs, numerals, determiners, and function words. The use of punctuation and the presence of foreign words are also considered. In total, 36 linguistic characteristics are analyzed to quantitatively assess the style, complexity, and grammatical structure of speech.

Based on the training data, two profiles are constructed—one for high and one for low verbal activity. Each profile consists of the empirical distributions of the extracted characteristics (36 in total). For a new utterance, the feature extraction steps are repeated. Then, for each of its characteristics, the Mann-Whitney U test is applied: the distribution of the feature's values in the utterance is compared to the distributions in both activity profiles. The test evaluates the probability that the samples belong to the same population.

The result is two p-values: p_1 (similarity to high activity) and p_2 (similarity to low activity). The utterance is classified according to the following rule:

- If $p_1 > \alpha$ (threshold 0.05) and $p_2 \leq \alpha$, the utterance is assigned to the high activity group.
- If $p_2 > \alpha$ and $p_1 \leq \alpha$, the utterance is classified as low activity.
- If both $p_1 > \alpha$ and $p_2 > \alpha$, the group with the maximum p-value is selected.
- If both p-values are below α, the decision is based on the maximum p-value, despite the statistical non-significance.

Multitask Model for Emotion Recognition

Semantic features of speech represent a complex set of linguistic characteristics that reflect the meaningful content of an utterance. A crucial component of semantics is emotion, which manifests through a system of psycholinguistic markers and contextual relationships. Emotional coloring does not merely complement the semantic content but significantly influences message interpretation.

This research implemented a multitask learning model for textual data classification based on three key psycholinguistic attributes—valence, arousal, and dominance—as well as a composite emotion class determined by the combination of these attributes. The model was trained and tested on a corpus compiled for this dissertation. To ensure data representativeness, a stratified 70:30 split into training and test sets was performed.

The architecture of the multitask classification model, illustrated in Figure 2, is based on the pre-trained rubert-large model, enhanced with three linear classifiers, each responsible for predicting one target attribute: valence, arousal, and dominance.

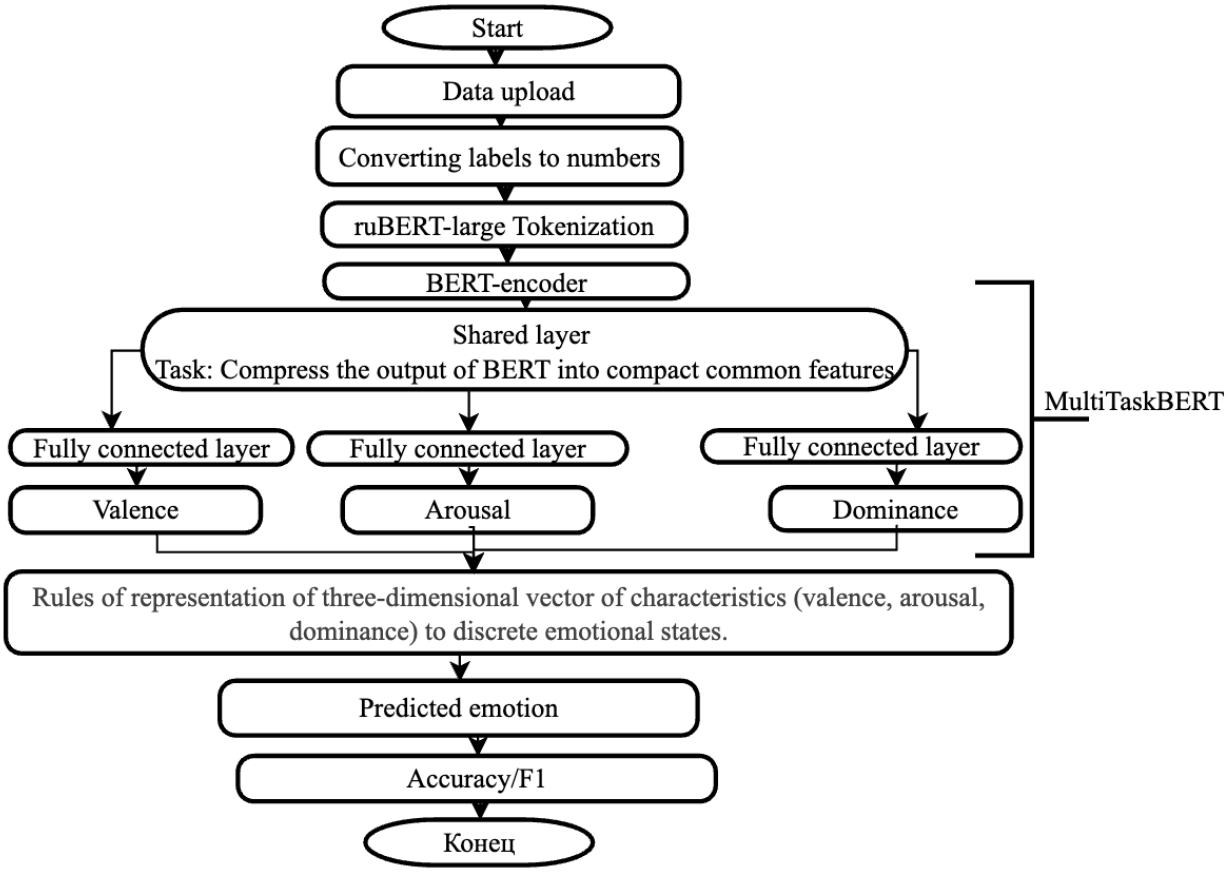

Figure 2 — A Method for Recognizing Emotional Coloring in Speech

A shared BERT encoder extracts contextual embeddings from the input text, which are then passed to the individual classifiers to generate predictions. This enables the model to learn from three correlated tasks simultaneously, improving data and computational efficiency.

For contextual embedding extraction, the input text is represented as a sequence of tokens:

$$X = [x_1, x_2, \dots, x_n]$$

where x_i is the i-th token and n is the sequence length. The RuBERT-large model converts the input tokens into contextual embeddings:

$$\mathrm{H} = \mathrm{RuBERT(X)},$$

where $H \in R^{(n \times d)}$ is the embedding matrix, and d is the embedding dimensionality; this study used a dimensionality of 1024. For classification, an aggregated text representation is used, specifically the embedding of the first token ([CLS]):

$$h_{[CLS]} = H_{[0]}$$

where $h_{[CLS]} \in R^d$

A separate linear classifier is used for each of the three attributes (valence, arousal, dominance).

5. RESULTS

Analysis of Speech Characteristic Profiles

The study analyzed speech characteristic profiles for English and Russian, revealing key distinctions between high and low verbal activity. For English, the analysis indicated that high-activity speech is characterized by a significant increase in the average number of words per utterance (17.40 vs. 10.60) and sentences (10.70 vs. 2.70), signifying greater complexity and informational density. Furthermore, high verbal activity is associated with increased use of nouns (6.70 vs. 4.30), verbs (4.70 vs. 2.90), and adjectives (3.67 vs. 0.90), reflecting lexical richness and diversity of constructions. In contrast, low verbal activity demonstrates simpler, shorter phrases with fewer parts of speech, typical of clichéd and formalized utterances.

Analogous patterns were observed for Russian, where high-activity speech also features a greater number of words per utterance (12.60 vs. 8.00) and sentences (5.64 vs. 3.30), alongside increased usage of nouns (6.70 vs. 3.41), verbs (3.60 vs. 0.50), and adjectives (4.23 vs. 2.20). These findings confirm that high verbal activity in both languages is associated with more complex and expressive speech constructions.

The validation results for the method of identifying speech characteristic changes under the influence of verbal activity demonstrated high classification accuracy for both languages. For English (Table 2), the accuracy was 0.92 for the National Hockey League Interviews dataset and 0.88 for the US 2020 Presidential Election Speeches dataset.

Table 2 — Accuracy of the Method for Identifying Changes in Speech Characteristics under the Influence of Speech Activity

Dataset	Accuracy
National Hockey League Interviews	0.92
US 2020 Presidential Election Speeches	0.88

For Russian, the method also showed high efficacy: 0.87 for NLP specialists and 0.89 for team leads (Table 3).

Table 3 — Accuracy of the Method for Identifying Changes in Speech Characteristics under Verbal Activity in Russian

Dataset	Accuracy
Natural Language Processing Specialist	0.87
Team Leads	0.89

The obtained data allow us to conclude that the method effectively distinguishes between high and low-activity speech based on linguistic features. The identified patterns confirm the hypothesis linking verbal activity with the complexity and richness of speech constructions. The differences in classification accuracy between datasets may be attributed to data-specific characteristics, as the Russian data were collected from less formal sources, yet the overall trend remained consistent.

Emotion Recognition Results

Emotion Recognition Method. The model was trained for 60 epochs using the AdamW optimizer with a learning rate of 1×10^{-5}. The loss function was the sum of cross-entropy losses for each of the three labels: L=L_1+L_2+L_3, where L_1, L_2 and L_3 are the losses for valence, arousal, and dominance, respectively. This ensured balanced learning across all target variables.

Model evaluation on the test set utilized the Accuracy and F1-score metrics for each of the three labels. Additionally, a method for determining a composite emotion class was implemented. The composite emotion class is defined based on a combination of the predicted values for valence, arousal, and dominance. Let c_v, c_a, c_d, be the predicted classes for each attribute. Then the composite class C is defined as:

$$C = f(c_v, c_a, c_d),$$

Where f is a heuristic function that maps the combination (c_v, c_a, c_d) to one of 26 possible emotional states, including the 10 most populous by number of utterances (Neutral State, Apathy, Calm Confidence, Interest, Serenity, Sadness, Fatigue, Uncertainty, Mild Satisfaction, Joy). Classification accuracy for the class was assessed by comparing predicted values with the dataset's ground truth labels.

The multi-task approach [14] enabled the model to simultaneously account for the interrelationships between valence, arousal, and dominance, which improved prediction quality for

each label. Furthermore, the use of heuristic rules to determine the composite class allowed for the creation of a complex emotion ontology without the need to collect additional labeled data for each subclass.

Table 4 — Performance Evaluation and Comparative Analysis of the Multi-Task Classification Model

Method	Accuracy V	F1 V	Accuracy A	F1 A	Accuracy D	F1 D	Accuracy E
Proposed method (trained on collected dataset)	0.85	0.82	0.80	0.76	0.76	0.75	0.57
Cross-domain evaluation on MELD dataset							
Proposed method (trained on collected dataset)	0.63	0.59	–	–	–	–	**0.54**
OPT-13B [15] (trained on IEMOCAP)	–	–	–	–	–	–	**0.46**
OPT-13B [15] (trained on MER)	–	–	–	–	–	–	**0.44**

V — Valence, A — Arousal, D — Dominance, E — Emotions

Analysis of the presented metrics leads to the following conclusions. The model demonstrates high efficacy in classifying basic psycholinguistic attributes such as valence, arousal, and dominance. For valence, the accuracy is 0.85 and the F1-score is 0.82, indicating high precision in distinguishing positive, negative, and neutral tones. Similar results are observed for arousal, where accuracy reaches 0.80 and the F1-score is 0.76. This suggests the model successfully handles the classification of emotion activation levels, although the slightly lower F1-score may be due to difficulties in distinguishing some arousal classes. For dominance, accuracy and F1-score are 0.76 and 0.75, respectively, somewhat lower than for valence and arousal. This may be due to less pronounced markers of dominance in the data or the greater complexity of classifying them.

However, performance significantly decreases for the task of classifying complex emotions into 26 classes. The accuracy for this task is 0.44 and the F1-score is 0.39. Such low metric values indicate the high complexity of the task, associated with the large number of classes and potential data imbalance. Reducing the number of classes to 10 improves accuracy to 0.57 and F1-score to 0.55, demonstrating an acceptable performance level.

To assess the robustness and generalizability of the proposed method, cross-domain validation was conducted on the Russian-translated MELD dataset. This validation minimizes the risk of model overfitting and provides a more reliable estimate of its performance on new data. The validation was performed as follows: the quality of emotion and valence classification was assessed using the model version trained on the dissertation corpus, without any additional training on the MELD dataset.

The results of this cross-domain validation demonstrated the proposed method's statistically significant superiority over the existing OPT-13B based solution in terms of the key classification accuracy metric.

6. CONCLUSION

This study presents a comprehensive solution to the key challenges in speech analysis outlined in the introduction: the fragmented analysis of speech characteristics, the limited range of recognizable emotions, and insufficient adaptation to different language systems.

The first method, aimed at identifying verbal activity, is based on a holistic analysis of 36 linguistic features, encompassing both quantitative and qualitative characteristics of speech, such as the average number of words per utterance, part-of-speech diversity, and syntactic complexity. The application of the Mann-Whitney U test for classifying speech into high and low activity levels yielded an accuracy of 92% for English and 89% for Russian, confirming the approach's cross-linguistic validity. The analysis revealed distinct patterns: high-activity speech is characterized by greater structural complexity, lexical richness, and syntactic diversity compared to the formalized and cliché-ridden speech of low activity.

The second method, implementing a multitask approach for emotion recognition based on the RuBERT-large architecture, demonstrated high efficacy in classifying three fundamental parameters of emotional states. The model achieved an accuracy of 85% in determining valence (emotional tone), 80% for arousal (level of activation), and 76% for dominance (degree of control). The combination of these parameters enabled the identification of 26 complex emotional states, including subtle nuances such as sarcasm and irony.

Although the accuracy for classifying the full set of 26 emotions was 44%, this figure increased to 57% for the 10 primary categories, surpassing the performance of existing analogues, such as OPT-13B (46%). The practical significance of this research is evident in its wide range of

potential applications—from the development of dialogue systems and emotionally intelligent assistants to the psycholinguistic analysis of professional communication and user-generated content on social media.

7. REFERENCES

1. Biber, D., Conrad, S.: Register, Genre, and Style. Cambridge Textbooks in Linguistics, Cambridge University Press (2009)
2. Ekman, P.: An argument for basic emotions. Cognition & Emotion 6, 169–200 (1992),
3. Fabien, M., Villatoro-Tello, E., Motlicek, P., Parida, S.: BertAA : BERT fine-tuning for authorship attribution. In: Bhattacharyya, P., Sharma, D.M., Sangal, R. (eds.) Title Suppressed Due to Excessive Length 13 Proceedings of the 17th International Conference on Natural Language Processing (ICON). pp. 127–137. NLP Association of India (NLPAI), Indian Institute of Technology Patna, Patna, India (Dec 2020)
4. Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., Denuyl, S.: Social biases in NLP models as barriers for persons with disabilities. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5491–5501. Association for Computational Linguistics, Online (Jul 2020).
5. Ilyin, E.P.: Emotions and feelings (2nd ed.) [PDF] (2021),
6. Makhnytkina, O., Frolova, O., Lyakso, E.: Machine learning methods for analyzing morphological and lexical characteristics of speech of boys with autism spectrum disorders and down syndrome. Vestnik NSU. Series: History and Philology 23, 39–55 (02 2024).
7. Jawahar, G., Sagot, B., Seddah, D. "What does BERT learn about the structure of language?" *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019).
8. Matveeva, A., Makhnytkina, O.: Text augmentation preserving persona speech style and vocabulary. Scientific and Technical Journal of Information Technologies, Mechanics and Optics 23, 743–749 (08 2023).
9. Mehrabian, A.: Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies. Oelgeschlager, Gunn & Hain, Cambridge (1980)
10. Savinova, M.S.: Profession as part of social status and its reflection in prosodic speech characteristics. Bulletin of Kostroma State University 15(4), 186–189 (2009)
11. Shen, W., Wu, S., Yang, Y., Quan, X.: Directed acyclic graph network for conversational emotion recognition. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1551–1560. Association for Computational Linguistics, Online (Aug 2021).
12. Pennebaker, James W. and Laura A. King. “Linguistic styles: language use as an individual difference.”*Journal of personality and social psychology* 77 6 (1999): 1296-312
13. Edwin D. Simpson; Statistical Significance Testing for Natural Language Processing. *Computational Linguistics* 2020; 46 (4): 905–908
14. Ruder, Sebastian. “An Overview of Multi-Task Learning in Deep Neural Networks.” *ArXiv* abs/1706.05098 (2017)
15. Lian, Z., Sun, L., Ren, Y., Gu, H., Sun, H., Chen, L., Liu, B., & Tao, J. (2024). MERBench: A Unified Evaluation Benchmark for Multimodal Emotion Recognition. *ArXiv, abs/2401.03429*.