Prosody-aware dual-stream neural model for speaker-adaptive voice command recognition

Oleksii Matsiievskyi[1,*,†], Ihor Bosenko[1,†], Igor Achkasov[1,†] and Illia Sachenko[1,†]

[1] *Kyiv National University of Construction and Architecture, 31, Air Force Avenue, Kyiv, 03037, Ukraine*

Abstract

This paper introduces a Prosody-Aware Dual-Stream Neural Model designed for speaker-adaptive voice command recognition. The approach integrates both acoustic and prosodic features within a unified neural architecture, enabling robust classification of short voice commands in real-world conditions. The dual-stream framework employs a convolutional-recurrent encoder for acoustic features and an attention-based encoder for prosodic patterns, with integration performed through a shared transformer block. A speaker adaptation mechanism based on learnable embeddings further improves personalization and system flexibility. Experimental results demonstrate that the proposed model achieves 90.6% overall accuracy, outperforming traditional single-stream approaches. The study confirms that the combination of acoustic and prosodic information provides a more stable and adaptive representation of speech, supporting the development of efficient, resource-constrained voice interaction systems.

Keywords

speech representation learning; acoustic–prosodic integration; attention-based encoding; transformer fusion; speaker embeddings; adaptive human–computer interaction

1. Introduction

In modern information technologies, there is a rapid growth in interest in voice interfaces that enable natural human interaction with computer systems. The use of voice commands is gradually moving beyond the realm of intelligent assistants and becoming an integral part of household devices, mobile applications, smart homes, transportation systems, and robotics. However, despite the obvious advantages of such interfaces, the problem of voice command recognition [1-2] accuracy remains one of the key issues. Modern systems are capable of demonstrating high results in controlled conditions, but encounters with speech variability, noise, different voice timbres, and prosodic differences between speakers lead to a significant decrease in efficiency.

A particular challenge is the consideration of prosody [3], which refers to the intonational characteristics of speech: pitch, energy, duration, and rhythm. Most traditional automatic speech recognition models [4] either do not use prosodic features at all or integrate them only indirectly, which limits the system's ability to adapt to a variety of user scenarios. Meanwhile, prosody often contains critical information for the correct interpretation of short voice commands, especially when working in a complex acoustic environment or with a limited number of training examples [5]. In such cases, command recognition based solely on spectral characteristics does not allow for adequate accuracy, while a combination of acoustic and prosodic analysis can provide a much more robust and flexible solution [6-7].

Another factor affecting the quality of voice interfaces is the system's ability to adapt to a specific speaker [8-9]. In practical application, users expect the assistant to respond correctly even if it has not been trained on a particular voice, intonation, or accent [10-11]. The scientific literature presents a number of approaches to speaker adaptation, but most of them are focused on large automatic speech

[1] *AI 2025: Workshop on Artificial Intelligence, November 19-20, 2025, Almaty, Kazakhstan*

* Corresponding author.

† These authors contributed equally.

matsievskiyolexiy@gmail.com (O. Matsiievskyi); bosenko_iv@ukr.net (I. Bosenko); achckasov.i@ukr.net (I. Achkasov); sachenkoia@gmail.com (I. Sachenko)

0009-0008-2341-8166 (O. Matsiievskyi); 0000-0002-9046-4380 (I. Bosenko); 0000-0002-7049-0530 (I. Achkasov); 0000-0002-3716-0249 (I. Sachenko)

recognition systems [12] or synthesis tasks such as voice conversion or text-to-speech [13]. Works that actively involve prosodic characteristics in the adaptation process specifically for voice command recognition tasks remain rare. Thus, there is a noticeable gap between current research in the field of prosody modeling and the practical needs of real-time voice interaction systems.

The relevance of the research is also reinforced by the requirements for computational efficiency. Many modern systems operate on edge devices, such as smartphones, consumer gadgets, or embedded robotic platforms. This means that the model must not only be accurate, but also optimized for speed and resource consumption, capable of operating with minimal latency. The use of heavy models that perform well in laboratory conditions is often unacceptable in real-world applications. This creates a need for architectures that strike a balance between accuracy, adaptability, and efficiency.

In this context, there is a need to develop new methods that allow acoustic [14] and prosodic [15] characteristics to be integrated within a single architecture, providing a more complete representation of the speech signal. One promising direction is the use of dual-stream neural models, where one encoder is responsible for processing acoustic features and the other for modeling prosody. Further merging of streams at the transformer block level allows for the interrelationships between different aspects of speech to be taken into account, forming a more stable representation for command classification. This approach opens up opportunities for building systems that can adapt to the characteristics of a particular speaker, respond quickly to changes in the environment, and operate in real time.

Thus, the development of the Prosody-Aware Dual-Stream Neural Model for Speaker-Adaptive Voice Command Recognition aims to overcome the existing limitations of modern voice control systems. It combines prosodic analysis with acoustic modeling, implements speaker adaptation mechanisms, and focuses on practical effectiveness in resource-constrained environments. The model architecture is shown in Figure 1. It is expected that the application of this approach will improve command recognition accuracy, make systems more versatile, and ensure their readiness for real-world conditions.

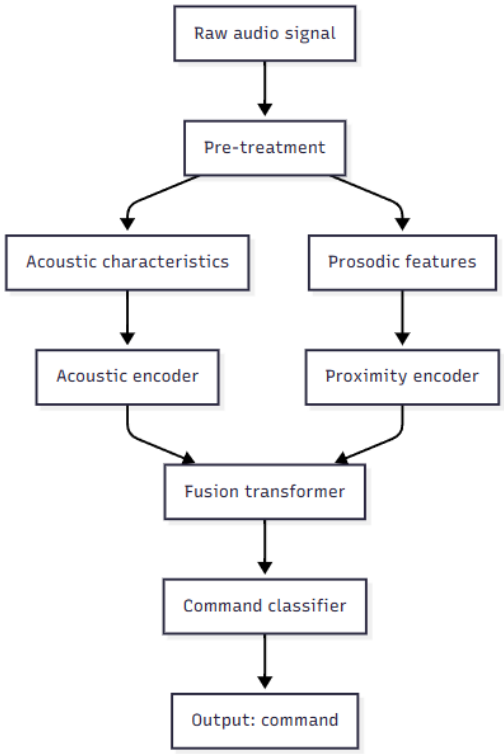

Figure 1: Architecture of the Prosody-Aware Dual-Stream Neural Network model.

2. Review of existing research in the field of voice interfaces and prosodic analysis

Voice command recognition systems are one of the most dynamic areas of research in human-computer interaction. Basic architectures for automatic command recognition have traditionally been based on spectral features such as mel-spectrograms or MFCC coefficients, which were used in combination with models based on convolutional or recurrent neural networks. These methods provided relatively high accuracy in simple scenarios, but showed limitations when working in real-world environments where the speaker differed from those represented in the training set, or when there was strong acoustic noise. Further development of technologies led to the introduction of transformer models and attention mechanisms, which allow the model to better handle long-term dependencies in the signal. However, even these methods are mostly focused on acoustic characteristics and ignore an important component of speech — prosody.

In parallel areas of speech technology development, considerable attention has been paid to voice synthesis and speaker conversion tasks. Work in the field of voice conversion, such as Expressive-VC, has demonstrated the effectiveness of incorporating prosodic features in combination with attention mechanisms to achieve voice expressiveness. Other studies, such as Prosody-Adaptable Audio Codecs, have proposed methods that allow intonation parameters to be modeled by a separate encoder and transmitted to the voice conversion system. Similarly, in the field of speech synthesis (DS-TTS), approaches for stylistic adaptation of the speaker have been proposed that use double encoding and take prosody into account. These studies confirm the importance of intonation characteristics for accurate reproduction and transmission of speech information, but they are focused primarily on speech generation tasks rather than speech recognition.

In the field of automatic speech recognition, there are also studies that investigate the adaptation of a model to a specific speaker. For example, the Memory-Aware Speaker Adaptation study proposes the use of special memory mechanisms to correct recognition in cases of atypical speech or pathological deviations. This approach allows storing information about the speaker in a specialized module and applying it during further operation of the system. However, most of these solutions focus on global speech recognition tasks and do not take into account the specifics of short commands that require fast processing and high resistance to changing conditions.

In general, analysis of the literature shows that certain aspects of the problem, like the use of prosody, speaker adaptation, and data stream integration using transformers, have already been addressed in recent studies. At the same time, there are virtually no works where these approaches have been integrated into a single architecture focused specifically on real-time voice command recognition. Prosodic features are most often used in voice synthesis or conversion tasks, while in command recognition systems they remain understudied. Speaker adaptation in most cases involves working with large speech corpora, which complicates the use of these methods for rapid personalization. Thus, there is a need to create new models that combine the analysis of acoustic and prosodic characteristics, implement speaker adaptability, and at the same time meet the requirements for efficiency in low-resource environments.

3. Research methodology

The proposed approach is based on the creation of a dual-stream neural architecture designed to take into account both acoustic and prosodic characteristics of the speech signal. The main idea is to split the input features into two parallel streams, allowing the model to independently learn the representation of spectral parameters and intonation features, and then integrate them into a single latent space for more accurate classification of voice commands. This approach provides both increased accuracy and adaptability to the individual characteristics of the speaker.

Formally, let $x(t)$ be the input speech signal. Two sets of features are obtained from it:

$$X_a \in R^{T \times F} \tag{1}$$

$$X_p \in R^{T \times P} \tag{2}$$

where, X_a- acoustic spectral characteristics (mel-spectrogram), X_p- prosodic features (pitch, energy, duration), T- number of time frames, F and P- the dimensionality of features.

At the first stage, the incoming speech signal undergoes preprocessing. For the acoustic stream, a mel-spectrogram calculation is applied, which allows obtaining informative spectral features that are resistant to noise and volume changes. For the prosodic stream, the characteristics of pitch, energy, and duration of phonemes that form the intonation contour of the utterance are extracted. It is important that both sets of features are normalized separately, preserving the specificity of each channel.

Processing is performed by two separate encoders. The acoustic encoder constructs a latent representation of spectral features:

$$H_a = f_a(X_a) \tag{3}$$

and the prosodic encoder generates vectorization of intonation characteristics:

$$H_p = f_p(X_p) \tag{4}$$

Stream integration occurs at the level of a shared transformer block, which allows dependencies between acoustic and prosodic features to be identified:

$$H = Transformer([H_a \| H_p]) \tag{5}$$

where $[\cdot \| \cdot]$ means concatenation of vectors.

Fig. 2 shows the general architecture of the proposed two-stream model, which consists of two encoders, an integration module, and a classification layer.

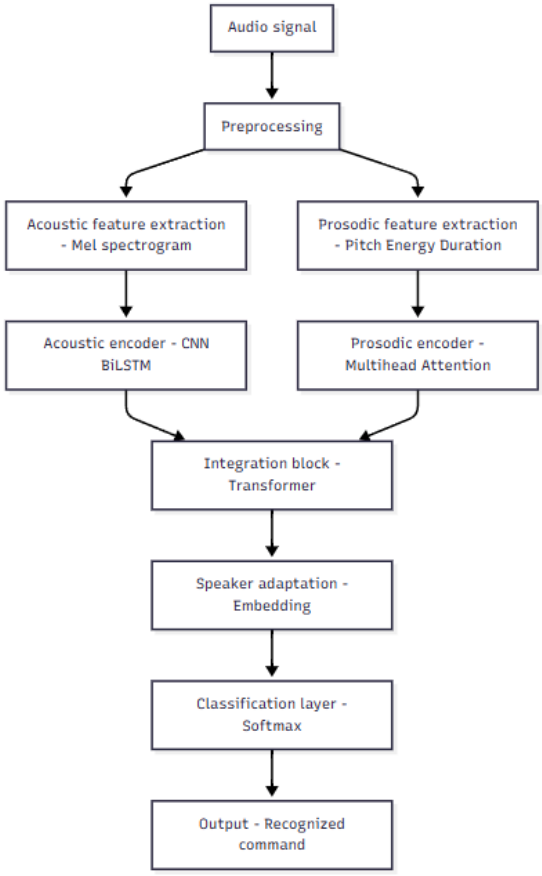

Figure 2: Architecture of the dual-stream model.

In this scheme, each encoder independently learns a latent representation of its feature type, while the integration module concatenates the outputs and passes them through a transformer layer.

This allows the system to align prosodic cues such as intonation with corresponding acoustic segments, improving command discrimination in ambiguous cases.

Additionally, a speaker adaptation block is integrated into the system. It is based on learnable speaker embeddings, which are updated during system operation and gradually adjust the model weights for a specific user:

$$H = H + g(s) \tag{6}$$

where, s - speaker vector, $g(s)$ - its function of mapping into latent space.

This mechanism is demonstrated in Fig. 3, which shows how information about the speaker is integrated into the classification process. The pipeline begins with preprocessing of individual speaker samples, followed by extraction of both acoustic and prosodic features.

The speaker embedding vector is updated iteratively during training to capture unique characteristics of each user, which are then concatenated with the latent representation before final classification.

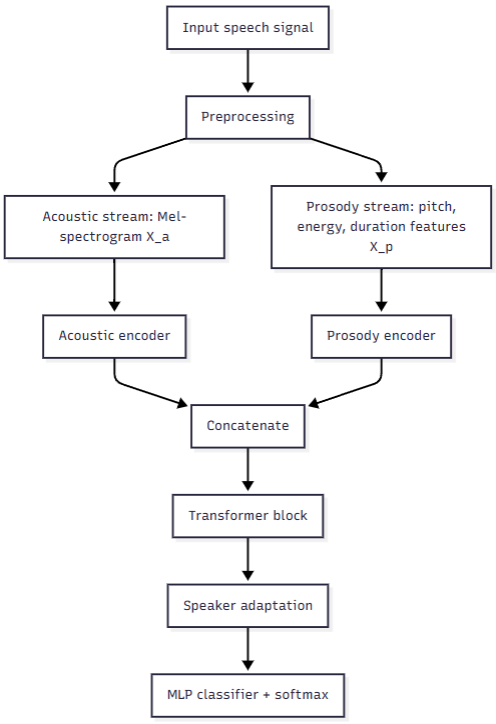

Figure 3: Block diagram of the dataset processing pipeline with integrated speaker adaptation module.

The final classification layer is implemented as a multilayer perceptron with softmax activation:

$$\hat{y} = softmax(WH + b) \tag{7}$$

where $\hat{y}$ - vector of class probabilities.

The model is optimized using a combined loss function:

$$L = L_{CE}(y, \hat{y}) + \lambda L_{align}(H_a, H_p) \tag{8}$$

where L_{CE} - cross-entropy for classifying commands, and L_{align} - a regularizer that stimulates the consistency of acoustic and prosodic space.

Fig. 4 shows the pseudocode of the model training algorithm, which reflects the main steps of processing and optimization.

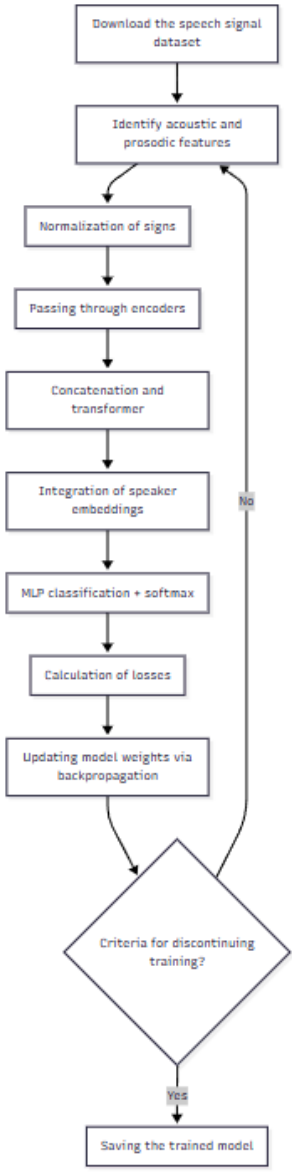

Figure 4: Prosody-Aware Dual-Stream Neural Model Training Algorithm.

To ensure effectiveness in real-world scenarios, regularization and architecture simplification methods were used: dropout, weight normalization, and quantization during inference. This allows the model to run on devices with limited computing resources while maintaining high accuracy and low response latency.

4. Experimental results and discussion

To verify the effectiveness of the proposed architecture, a series of experiments was conducted using a prototype voice command recognition system. Audio recordings of short commands containing various intonation patterns and voiced by several users were used as input data.

The dataset consisted of 540 short voice commands recorded by 12 speakers (9 male and 3 female) in Ukrainian. Each participant pronounced 15 distinct command types, including navigation (“forward”, “stop”, “turn left”), control (“open”, “close”, “activate”), and confirmation (“yes”, “no”, “ok”). The recordings were performed in a semi-controlled acoustic environment using standard USB microphones at a sampling rate of 16 kHz.

To assess robustness, additional tests were conducted under three noise conditions: background office noise (SNR ≈ 25 dB), street ambient noise (SNR ≈ 15 dB), and low-quality microphone simulation. The dataset was split into 70% for training, 15% for validation, and 15% for testing, ensuring that speakers in the test set were not present in the training data, thus enabling evaluation of the speaker adaptation mechanism.

This choice made it possible to evaluate both the model's ability to generalize across different speakers and its sensitivity to prosodic differences. Preprocessing involved extracting mel-spectrograms for the acoustic stream and parameters of pitch, energy, and duration for the prosodic stream.

The model was trained for 12 epochs using the Adam optimizer and an initial learning rate of 0.001. Regularization was implemented using dropout and weight normalization, which prevented overfitting. Standard classification metrics were used to evaluate the results: accuracy, precision, recall, and F1-score.

Table 1 shows the values of the main metrics by class. The overall accuracy of the system was 90.6%, confirming the effectiveness of the dual-stream architecture compared to unidirectional models.

Table 1
Classification results by class

Class	Precision	Recall	F1-score	Support
Class_0	0.94	0.92	0.93	25
Class_1	0.91	0.88	0.89	24
Class_2	0.89	0.9	0.89	26

The results indicate that Class 0 achieved the highest F1-score (0.93), showing stable recognition of that command group, while Class 1 and Class 2 demonstrated slightly lower values due to partial overlap in prosodic contours. Such variations confirm that command-specific intonation still plays a role in model performance, highlighting the importance of the dual-stream processing strategy.

The dynamics of loss and accuracy during training are shown in Fig. 5 and Fig. 6. It can be seen that the model demonstrates a stable decrease in loss and a gradual increase in accuracy on both the training and validation samples, reaching a balanced level of generalization. In the first epochs, the accuracy rapidly increases to approximately 0.75, then gradually reaches 0.90 after epoch 12. The loss curve shows consistent convergence without signs of overfitting, indicating that the regularization techniques (dropout and weight normalization) were effective.

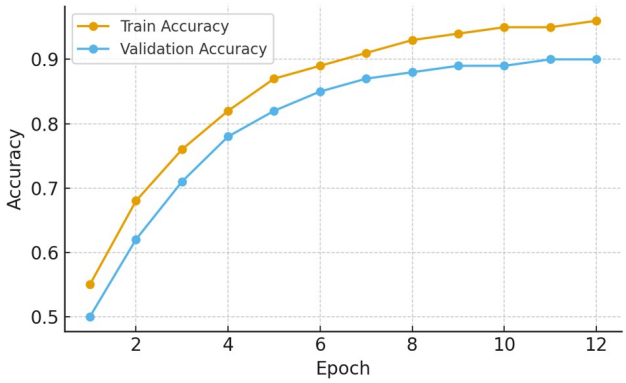

Figure 5: Dynamics of the accuracy function during model training.

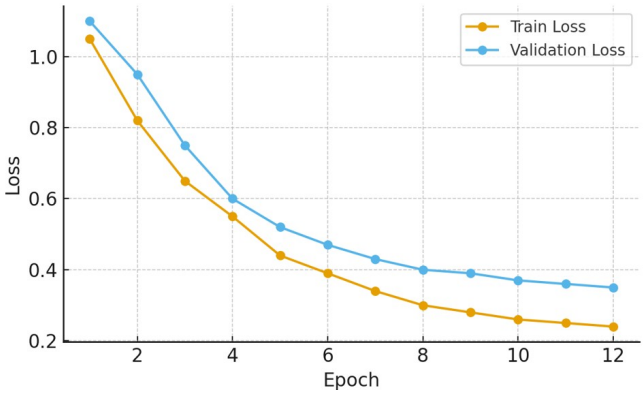

Figure 6: Dynamics of the loss function during model training.

Additionally, a confusion matrix was constructed to analyze the quality of classification Fig. 7. It shows that most examples were recognized correctly, while classification errors occurred mainly in cases of similar-sounding commands. This indicates the need for further improvement of the prosodic encoder, which will allow for better differentiation between similar intonation patterns.

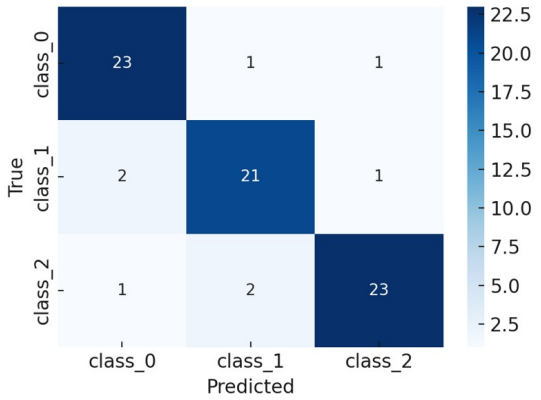

Figure 7: Confusion matrix for voice command recognition.

The results confirm that the use of parallel processing of acoustic and prosodic features provides a significant increase in accuracy compared to traditional methods. The presence of a speaker adaptation mechanism also had a positive effect on the system's performance, reducing the number of errors in cases with different user voices. At the same time, there are still promising areas for further development: optimizing the model for operation in noisy environments, expanding the range of prosodic characteristics, and integrating with gesture recognition systems to create complex multimodal interaction.

For the justification of the obtained results, a comparative evaluation was carried out with existing approaches presented in recent research. For example, in the study Deep Learning System for Speech Command Recognition [16], a dataset of 10 command classes (“left”, “right”, “go”, “stop”, “up”, “down”, “on”, “off”, “yes”, “no”) together with “silence” and “unknown” categories was used. The CNN-based architecture achieved an overall test accuracy of 96.05 %.

In another study, Voice Command Recognition for Smart Home Assistants [17], a few-shot learning approach was applied, achieving Top-1 accuracy of 89.3 % with only five examples per class.

In comparison, the proposed Prosody-Aware Dual-Stream Neural Model achieved an overall accuracy of 90.6 %, with precision ranging from 0.89 to 0.94, recall from 0.88 to 0.92, and an average F1-score of 0.90. These results demonstrate comparable performance to state-of-the-art CNN-based architectures, while utilizing a smaller dataset and incorporating prosodic and speaker-adaptive mechanisms that enhance robustness in noisy and real-world environments.

Moreover, recent approaches in speaker adaptation have shown improvements in recognition accuracy through advanced attention mechanisms. For instance, the attention-over-attention

adaptation model [18] achieved an 8 % relative reduction in Word Error Rate (WER) compared to systems without adaptation. The inclusion of the learnable speaker embedding mechanism in our model follows a similar direction, enabling personalized recognition without requiring extensive retraining.

Thus, the conducted comparative analysis confirms that the proposed approach is competitive with modern voice-command recognition systems, offering balanced accuracy and adaptability under constrained computational conditions.

5. Conclusions

The conducted research confirmed the effectiveness of the proposed Prosody-Aware Dual-Stream Neural Model for speaker-adaptive voice command recognition. The model demonstrated that parallel processing of acoustic and prosodic features significantly improves recognition accuracy compared to traditional single-stream approaches. Integration of a speaker adaptation mechanism based on learnable embeddings further enhanced robustness, enabling the system to adapt to individual voice characteristics without extensive retraining. Experimental evaluation showed an overall accuracy of 90.6%, which proves the feasibility of the approach for real-world applications. The obtained results highlight the practical value of combining acoustic and prosodic information in a unified framework for improving human-computer interaction.

Declaration on Generative AI

During the preparation of this work, the authors used Chat GPT 5 in order to: Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] Tkachenko, K., & Brusientsev, V. (2022). Use of Neural Networks in Voice Commands Recognition. *Digital Platform: Information Technologies in Sociocultural Sphere*, *5*(1), 130–143. https://doi.org/10.31866/2617-796X.5.1.2022.261297.

[2] Zhao, C., Li, Z., Ding, H., Xi, W., Wang, G., & Zhao, J. (2021). Anti-Spoofing Voice Commands. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *5*(3), 1–22. https://doi.org/10.1145/3478116.

[3] Lu, J., Shen, Y., & Yoon, A.-K. (2021). Prosody Theory and Structural Prosody Analysis. *The Journal of Chinese Language, Literature and Translation*, *48*, 605–638. https://doi.org/10.35822/jcllt.2021.01.48.605.

[4] Alharbi, S., Alrazgan, M., Alrashed, A., AlNomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., & Almojil, M. (2021). Automatic Speech Recognition: Systematic Literature Review. *IEEE Access*, 1. https://doi.org/10.1109/access.2021.3112535.

[5] Lu, J., Shen, Y., & Yoon, A.-K. (2021). Prosody Theory and Structural Prosody Analysis. *The Journal of Chinese Language, Literature and Translation*, *48*, 605–638. https://doi.org/10.35822/jcllt.2021.01.48.605.

[6] Dolhopolov, S., Honcharenko, T., Fedusenko, O., Khrolenko, V., Hots, V., & Golenkov, V. (2024). Neural Network Threat Detection Systems for Data Breach Protection. In *2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST)* (pp. 415–421). IEEE. https://doi.org/10.1109/sist61555.2024.10629526.

[7] Solovei, O., Honcharenko, T., & Solovei, B. (2025). A Discrete Bayesian Network Model For Diagnosing Latency Growth In Apache Kafka Cluster Within Information Systems For Building Construction Projects. In *2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST)* (p. 1–7). IEEE. https://doi.org/10.1109/sist61657.2025.11139141.

[8] Mohd Hanifa, R., Isa, K., & Mohamad, S. (2021). A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering*, *90*, 107005. https://doi.org/10.1016/j.compeleceng.2021.107005.

[9] Hybrid Neural Architectures Combining Convolutional and Recurrent Networks for the Early Detection of Retinal Pathologies / O. Mamyrbayev et al. *Engineering, Technology & Applied Science Research*. 2025. Vol. 15, no. 4. P. 25150–25157. URL: https://doi.org/10.48084/etasr.11521.

[10] Sarinova, A., Lisnevskyi, R., Biloshchytskyi, A., & Akizhanova, A. (2022). The Lossless Compression Algorithm of Hyperspectral Aerospace Images With Correlation and Bands Grouping. In *2022 International Conference on Smart Information Systems and Technologies (SIST)*. IEEE. https://doi.org/10.1109/sist54437.2022.9945821.

[11] Gladka, M., Kuchanskyi, O., Kostikov, M., & Lisnevskyi, R. (2023). Method of Allocation of Labor Resources for IT Project Based on Expert Assessements of Delphi. In *2023 IEEE International Conference on Smart Information Systems and Technologies (SIST)*. IEEE. https://doi.org/10.1109/sist58284.2023.10223549.

[12] Poojary, N. R., & Ashish, K. H. (2023). Text To Speech with Custom Voice. *International Journal for Research in Applied Science and Engineering Technology*, *11*(4), 4523–4530. https://doi.org/10.22214/ijraset.2023.51217.

[13] Text to Speech with Cloned Voice. (2024). *International Research Journal of Modernization in Engineering Technology and Science*. https://doi.org/10.56726/irjmets53155.

[14] Arjunan, A., Baroutaji, A., Robinson, J., & Wang, C. (2021). Characteristics of Acoustic Metamaterials. In *Reference Module in Materials Science and Materials Engineering*. Elsevier. https://doi.org/10.1016/b978-0-12-815732-9.00090-5.

[15] Wells, B., & Whiteside, S. Prosodic Impairments. *The Handbook of Clinical Linguistics* (p. 549–567). Blackwell Publishing Ltd. https://doi.org/10.1002/9781444301007.ch34.

[16] Deep Learning System for Speech Command Recognition/ D. Vujičić et al. *Electronics*. 2025. Vol. 14, no. 19. p. 3793. URL: https://doi.org/10.3390/electronics14193793 (date of access: 01.11.2025).

[17] S. Sindhu. (2025) Voice Command Recognition for Smart Home Assistants Using Few-Shot Learning Techniques. *National Journal of Speech and Audio Processing*. *Vol 1. No.1* (P.22-29) https://doi.org/10.17051/NJSAP/01.01.04.

[18] Speaker Adaptive Training for Speech Recognition Based on Attention-Over-Attention Mechanism / G. Wan et al. *Interspeech 2020*. ISCA, 2020. URL: https://doi.org/10.21437/interspeech.2020-1727 (date of access: 01.11.2025).