BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 31 No. 2. 2026

UDC 004.94:621.391.8
DOI: 10.62660/bcstu/2.2026.35

ISSN: 2306-4412
E-ISSN: 2708-6070

Analysis of methods and approaches to audio data processing

Vladyslav Radin*
Postgraduate Student
State University "Kyiv Aviation Institute"
03058, 1 Liubomyra Huzara Ave., Kyiv, Ukraine
https://orcid.org/0009-0009-1101-6888

Myroslav Riabyi
PhD in Technical Sciences, Associate Professor
State University "Kyiv Aviation Institute"
03058, 1 Liubomyra Huzara Ave., Kyiv, Ukraine
https://orcid.org/0000-0002-9651-9135

Abstract. Within the research, it is necessary to assess the automatic speech recognition systems available today, adapting them to the Ukrainian language and considering the trade-offs between accuracy, performance, and various other factors. The aim of the research was to compare the existing methods and technologies for speech recognition, with the goal of creating a system for assessing the informational impact of the analysed audio files. The study utilised traditional hidden Markov processes and Gaussian mixtures models, hybrid neural network models, such as DeepSpeech, Wav2Vec 2.0, and Whisper, cloud-based solutions, such as those by Google, Amazon, Microsoft, and IBM. Audio processing was carried out using speech detection and mel-frequency coefficients; word recognition accuracy, processing time on the CPU and GPU, response latency and instability, and the impact of recognition errors on the natural language processing pipeline with tokenisation, topic classification and sentiment analysis was assessed. As a result, it became possible to establish the dependence of the accuracy and performance of Automatic Speech Recognition systems upon the architecture of the recognition systems and the hardware upon which they are installed. Classic Hidden Markov Models were found to have the lowest performance in recognising the words spoken within audio files, with word error rates between 18 and 30%, as well as processing times of 51 to 72 seconds, indicating their limited suitability for Ukrainian language recognition. End-to-end neural network models, however, had significantly better recognition performance, with DeepSpeech models having an error rate of 22%, Wav2Vec models having an error rate of 12%, and whisper models having an error rate of only 7%. Models based upon deep learning and transformers were also found to be robust to phonetic variations of the Ukrainian language. Furthermore, the local end-to-end neural network models had significantly better processing speeds than either cloud-based solutions or classic Hidden Markov Models, with whisper models taking only around 10 seconds to process an audio file. Cloud-based recognition systems had similar accuracy to local models (between 7 and 10% error rate), but required dependence upon the network, indicating potential privacy issues for those using such systems. Thus, results of this research indicate that whisper and Wav2Vec models are the best methods to utilise for audio content analysis and information effect detection systems. The findings of this research can be applied to the development of automatic transcription systems, audio monitoring systems, voice analytics services, or the development of information influence detection modules

Keywords: speech-to-text; automatic speech recognition; Whisper; Wav2Vec; audio stream processing; word error rate; neural networks

Article's History: Received: 06.01.2026; Revised: 13.04.2026; Accepted: 18.05.2026; Published: 26.06.2026

Suggested Citation:
Radin, V., & Riabyi, M. (2026). Analysis of methods and approaches to audio data processing. *Bulletin of Cherkasy State Technological University*, 31(2), 35-47. doi: 10.62660/bcstu/2.2026.35.

*Corresponding author (RadinVladyslav@outlook.com)

INTRODUCTION

The relevance of the research is dictated by the rapid growth of audio content in the digital environment and the need to automatically process it for information analysis. Automatic speech recognition (ASR) systems are already an integral part of information analysis processes for labelling, sentiment analysis, and recognising information with high impact. The linguistic characteristics of Ukraine make the use of general ASR packages challenging. Additionally, the decision to use on-premises or cloud-based ASR solutions introduces another level of complexity to the implementation of ASR solutions. Therefore, a detailed study of the differences between classical and neural network-based ASR systems, their impact on information analysis processes, and their overall performance is of considerable scientific and practical value.

Currently, research into ASR systems focuses on neural network models, their robustness under specific computing and linguistic conditions, and their ability to learn languages with limited spoken corpora. A. Dumyn *et al*. (2025) conducted a detailed study of classical and neural network-based speech recognition systems for Ukrainian and English languages using the Word Error Rate (WER) metric. The results presented by the authors show a significant improvement in the performance of neural network models as compared to the classical methods, particularly for the Ukrainian language. A different perspective was offered by Y.V. Bai & Yu.I. Katkov (2025), who reviewed the application of neural networks for note recognition from audio signals. Although the authors' research is in the related field of music informatics, their conclusions regarding the effectiveness of deep convolutional and recurrent architectures for processing complex time-frequency patterns have a direct methodological link to ASR. The authors demonstrated that the transition from classical feature-based approaches to end-to-end training can significantly improve recognition quality. The study by D. Rakhimova *et al*. (2025) analysed the effectiveness of various ASR models for Kazakh children's speech as a typical example of a low-resource environment. A specialised corpus, adapted to the models, was created to demonstrate that the use of transfer learning and multilingual pre-training can reduce the WER by tens of percentage points compared to models trained from scratch. At the same time, a significant decline in quality regarding the acoustic variability of children's speech was highlighted.

W. Liu (2025) provided a detailed analysis of methods for improving the performance and reliability of ASR systems, including the optimisation of computational processes and the use of self-supervised learning to reduce latency and enhance recognition stability. The study by W. Bouchelligua (2025) presented the VoiceX model, which utilises an end-to-end architecture and transfer learning for ASR in low-resource languages. The authors demonstrated that combining multilingual pre-training with adaptation to a specific corpus yields high accuracy rates even on limited data. A similar approach was proposed by S. Dhahbi *et al*. (2025), who developed a neural end-to-end ASR system for low-resource languages, demonstrating the effectiveness of integrating acoustic and linguistic features to improve recognition accuracy and stability. These studies have confirmed the key role of modern architectures, transfer learning and specialised training strategies in improving the performance and reliability of ASR in challenging linguistic conditions.

A voice recognition system for automating Internet of Things (IoT) home devices, optimised for the Galician language and mobile scenarios, was proposed by I. Froiz-Miguez *et al*. (2023). The authors demonstrated that combining lightweight neural models with specialised speech processing methods can achieve high accuracy even on resource-constrained devices. In turn, the study by D.F. Mujtaba & N.R. Mahapatra (2025) conducted a systematic review of modern Application Programming Interfaces (APIs) for ASR, including an analysis of models, functional capabilities and challenges when applied to low-resource languages. The study emphasised that model selection and adaptation to language-specific characteristics significantly influence the system's accuracy and latency. A different perspective was offered by L.A. Kumar *et al*. (2024), who, in their monograph, examined the integration of ASR and machine translation for low-resource languages, demonstrating the relevance of system pipeline optimisation, use of pre-trained models, and fine-tuning to specific corpora to ensure low WER and stable real-time performance. These studies have highlighted the relevance of developing effective and adaptive ASR solutions for limited linguistic and computational conditions.

Previous studies have demonstrated some success in developing ASR systems for languages with limited resources and in integrating such systems into mobile and IoT devices. However, most studies addressed individual aspects: either on optimising models for low computational resources or on adapting them to a specific language, whilst there is a lack of a comprehensive approach that simultaneously covers multitasking efficiency, low latency and recognition accuracy for real-time applications in low-resource languages. The study aimed to conduct a comparative analysis of modern ASR systems for their application in the development of an information impact assessment system. The study set the following tasks: to analyse existing ASR models and technologies for low-resource languages with a view to identifying optimal approaches to ensure high accuracy and low latency; to identify the limitations of modern speech recognition methods in terms of their suitability for automated assessment of information impact.

MATERIALS AND METHODS

The study aimed to systematically compare classical, modern on-premises and cloud-based ASR (Speech-to-Text,

STT) systems, emphasising Ukrainian-language data, noise robustness, computational requirements and suitability for further analytical tasks. The general approach involved the creation of a representative set of models, standardising the audio input data, establishing a unified testing procedure, and developing a multi-criteria comparative evaluation scheme. In the first stage, a conceptual classification of speech recognition approaches was conducted. Traditional systems were based on a combination of hidden Markov models (HMM) and Gaussian mixture models (GMM), where HMM modelled the sequence of phonetic or sub-phonetic states, and GMM approximated the distribution of acoustic features in each state. Recognition was formalised as the problem of finding the most probable sequence of words given the observed acoustic vectors, using Viterbi or forward-backwards algorithms. Additionally, dynamic time warping methods were used to align time-warped signals, along with time-frequency analysis based on the short-time Fourier transform and derived spectral representations. Within this class, the Hidden Markov Model Toolkit (HTK) (n.d.), CMU Sphinx (n.d.) and Kaldi (n.d.) were considered as examples of the evolution of the HMM paradigm – from classical GMM-HMMs to hybrid DNN-HMM architectures.

Modern approaches to STT in this study were represented by end-to-end deep learning (DL) models, notably DeepSpeech (n.d.), Wav2Vec 2.0 (Baevski *et al.*, 2020) and Whisper (OpenAI, 2022). In these systems, acoustic and language modelling are integrated into a single neural network architecture that directly maps the input audio signal to a text sequence. DeepSpeech implements recurrent networks optimised via Connectionist Temporal Classification, which avoids the need for pre-alignment of audio and transcriptions. Wav2Vec is based on self-supervised pre-training on large datasets of unlabelled audio data, followed by fine-tuning on limited labelled corpora, which is of fundamental importance for resource-poor languages, particularly Ukrainian. Whisper is a transformer-based multilingual architecture trained on hundreds of thousands of hours of noisy web data, providing enhanced robustness to acoustic distortions and speech variability. In parallel, commercial cloud services were included: Google STT (Google Cloud, n.d.), Amazon Transcribe (n.d.), Microsoft Azure Speech Services (Azure Speech in..., n.d.), and IBM Watson Speech to Text (n.d.), which represent industry-standard STT solutions with closed models but a high level of engineering optimisation. For these, not only functional accuracy was analysed, but also architectural aspects of API usage, including dependence on network infrastructure, the ability to customise dictionaries, and integration with analytics platforms. To ensure the comparability of results, the entire audio data processing pipeline was standardised (Fig. 1).

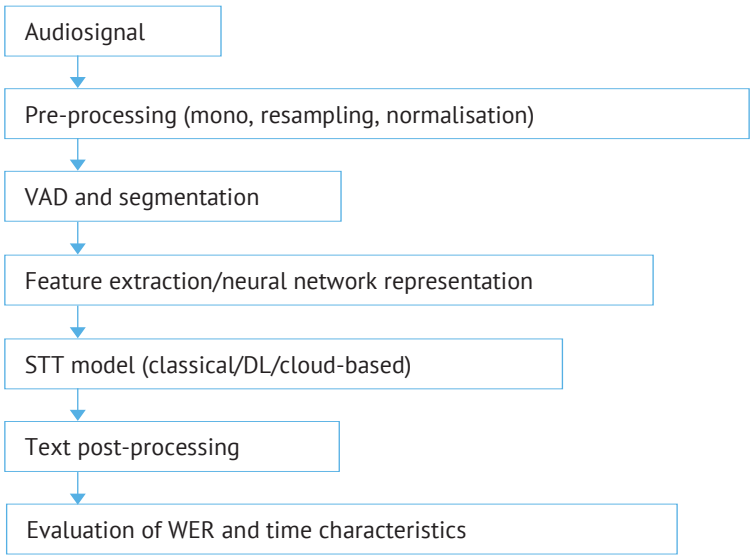

Figure 1. General logic of the experimental conveyor
Notes: VAD – Voice Activity Detection
Source: compiled by the authors

All input recordings were converted to a mono format with a standardised sampling rate, after which volume normalisation and speech segmentation were performed using Voice Activity Detection (VAD) techniques. Mel-frequency cepstral coefficients (MFCCs) were used during feature extraction depending on the requirements of the specific system. Support for the Ukrainian language was also analysed separately. The empirical part of the study used a one-minute audio recording containing 250 words of the reference audio signal. In order to simulate noise in the audio signal, white Gaussian noise with a Signal-to-Noise Ratio (SNR) of 10 dB was additively superimposed onto the audio signal across the audible frequency range. The SNR of 10 dB

was chosen as a standard level of noise in real world environments. The SNR was applied to test the robustness of the speech recognition models in the presence of noise. The main quantitative evaluation measure for assessing the performance of the speech recognition models was the WER. The WER measures the number of insertions, deletions, and substitutions of words by the speech recognition system relative to the total number of words in the reference audio recording.

The experimental design involved comparing several versions of the Whisper models (Tiny, Base, Large-v3), Wav2Vec 2.0 and DeepSpeech on Ukrainian-language and bilingual (Ukrainian-Russian) data. For each model, the WER, average audio file processing time and execution mode (CPU (central processing unit) or GPU (graphics processing unit)) were recorded, which were used to assess the trade-off between model scale, accuracy and performance. For cloud services, end-to-end latency was additionally measured, which included the time taken to upload the audio file to the server, server processing, and the client receiving the response. For each service, at least 30 repeated requests were made; jitter was defined as the standard deviation of response time, and latency variance was assessed under conditions of stable and variable network load. This was used to evaluate both the average performance and the stability of the services. A separate assessment was conducted of the impact of automatic recognition errors on subsequent natural language processing (NLP) analytical modules. A unified pipeline of tokenisation, topic classification and sentiment analysis was used. The results of automatic transcriptions were compared with a reference manual transcription, which provided a quantitative assessment of the degradation in the accuracy of NLP modules depending on the WER level.

All local experiments were conducted on a hardware platform featuring an Intel Core i7-13700H processor, an Nvidia GeForce RTX 4050 Studio graphics accelerator with 6 GB of video memory, and 16 GB of RAM. This was used for the analysis of both CPU-oriented and GPU-optimised configurations, as well as the investigation of trade-offs between processing speed and accuracy for models of varying scales. Thus, the combination of materials and methods integrates architectural analysis, a unified audio processing pipeline, a multi-criteria evaluation system, and a controlled experimental environment, creating a reproducible framework for comparing different STT paradigms and further interpreting their suitability for information analytics tasks.

RESULTS

A comparative analysis of the accuracy of ASR systems

The experimental results obtained revealed fundamentally different levels of accuracy in ASR depending on the system paradigm (classical statistical, local neural network or cloud-based). The WER value reflected the quality of acoustic and linguistic modelling, as well as the system's ability to generalise in the presence of noise and phonetic variability. Traditional systems (Hidden Markov Model Toolkit, CMU Sphinx, Kaldi) demonstrated the highest error rates: WER was 25% for HTK, around 30% for CMU Sphinx and 18% for Kaldi. These results were attributed to a cascaded processing approach, where errors accumulated in the early stages, whilst the limited training corpora of the Ukrainian language reduced the systems' robustness to noise, stress variations and timbre.

Local DL-based systems – DeepSpeech, Wav2Vec and Whisper – demonstrated a fundamentally different picture. DeepSpeech reduces WER to 22%, which already corresponds to the upper limit of the average accuracy range; however, the most telling results come from self-learning speech representation models. Wav2Vec achieves a WER of 12%, whilst Whisper Large-v3 achieves just 7%, effectively entering the category of near-perfect recognition. This difference is explained using end-to-end architectures, in which acoustic and language modelling are integrated into a single neural system, as well as the application of pre-training on massive multilingual datasets. The scale of the training data and the ability of transformer models to generate universal latent representations of speech ensure high generalisation ability even in the absence of specialised fine-tuning for a specific language. Commercial cloud services (Google STT API, Azure Speech, Amazon Transcribe, IBM Watson STT) demonstrate results comparable to the best local neural network models. Google STT API and Azure Speech achieve a WER of 7%, Amazon Transcribe around 10%, whilst IBM Watson STT demonstrates 15%. This indicates that cloud solutions utilise deep architectures of a similar class, optimised on large proprietary corpora. At the same time, despite their high accuracy, such services have inherent limitations related to network latency, data processing policies and dependence on external infrastructure, which significantly reduces their appeal for scenarios with stringent requirements for autonomy and privacy. To summarise the results, it is advisable to group the systems under investigation by WER ranges, as shown in Table 1.

Table 1. Grouping of speech recognition systems by WER

WER range	Accuracy specifications	Systems
5-7%	Very high, nearly error-free	Whisper, Google STT API, Azure Speech
8-12%	High, minor errors	Wav2Vec, Amazon Transcribe
15-22%	Average, noticeable errors	Kaldi, DeepSpeech, IBM Watson STT
⩾25%	Low, possible distortions	HTK, CMU Sphinx

Source: compiled by the authors

The presented classification clearly illustrates the technological gap between the traditional and modern approaches to neural networks. While the WER decreased from 25% to 30% to 7%, there is an actual reduction in the number of errors by a factor of 3 to 4, with a reduction in the relative number of errors by more than 70%. In the typical system for speech recognition, transcription of speech into text, classification of that text, sentiment analysis and keyword search, errors in ASR are the leading cause of errors in the remaining stages of topic modelling, sentiment analysis, etc. Any increase in the WER will exponentially increase the number of errors in the remainder of the process. Consequently, any omission of words that may convey the presence of an information effect will reduce the sensitivity of the system to those types of messages. Such errors can be represented in a cascade diagram (Fig. 2).

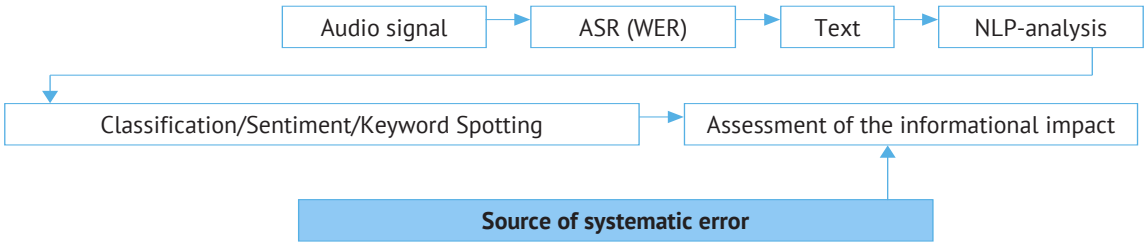

Figure 2. The process of automatic speech and information content analysis with the determination of the critical error threshold (ASR)

Source: compiled by the authors

The diagram demonstrates that the ASR module is the bottleneck in the system; any reduction in WER will enhance the reliability of the remaining components of the system. Achieving a WER of between 5 and 7% with models such as Whisper or other cloud-based offerings creates new conditions for building reliable systems for the detection of informational influences. Thus, the analysis of each of the models presented led to the conclusion that neural network models have made a radical improvement in the accuracy of ASR systems compared to classical models. The enhanced accuracy of end-to-end neural network ASR models has led to an improvement in the quality of transcription and the reliability of the remaining components of the automated analysis system. Consequently, such systems can be deployed in practice to automate the analysis of informational content.

Analysis of the computational performance and time characteristics of ASR systems

One of the key parameters in assessing the suitability of a system for practical applications is its computational performance. The computational performance of an ASR system can be reflected in the processing time of the audio signals that are processed by the system. A comparison of the CPU-oriented classical systems (HTK, CMU Sphinx, Kaldi) and the GPU-optimised DL models (DeepSpeech, Wav2Vec, Whisper) reveals the relationship between the hardware that supports an ASR system and its processing time for an audio file. Classic systems, which traditionally run on a CPU, demonstrate the longest processing times. For example, HTK required 72 seconds to transcribe a standard benchmark file, CMU Sphinx – 51 seconds, and Kaldi – 61 seconds. These figures are explained by the sequential architecture of HMM-based models, where the computation of acoustic and linguistic probabilities takes place on the CPU, without significant use of the parallel computations' characteristic of GPU. In these cases, processing time is determined not only by the complexity of the algorithms, but also by the computational throughput limitations of the central core, which significantly slows down processing of long or multi-channel audio streams.

Modern DL-based models, by contrast, demonstrate significant speed-ups thanks to optimisations for GPU computing. DeepSpeech processes a benchmark file in 38 seconds, Wav2Vec in 29 seconds, and Whisper Large-v3 in just 10 seconds, effectively delivering processing speeds faster than real time. This difference is attributed both to the parallel processing of tensor operations on the GPU and to the integration of acoustic and language modelling stages into a single end-to-end architecture, which eliminates the cascading accumulation of errors and additional intermediate computations characteristic of classical systems. Thanks to the use of a transformer architecture with multi-level self-learning blocks for speech representations, this model achieves a low WER (7%) while requiring minimal processing time. However, the use of smaller models within the Whisper library results in an increase in the WER to 10-15%. Thus, local speech recognition models with a GPU seem to be the most effective solution for tasks that require high-speed speech recognition. CPU-based solutions are still relevant for applications that require the use of stable algorithms that do not make use of GPU, or for the processing of large volumes of data that can be performed offline.

A comparison of local and cloud-based ASR systems

A comparison of local speech recognition systems to cloud-based speech recognition APIs reveals some

important differences between the two solutions. Cloud API providers, such as Google and Microsoft, offer speech recognition services that have comparable WERs to the best of the local speech recognition models. However, a closer look at specific metrics for both local and cloud-based systems reveals the true difference between the two methods of speech recognition. Results of the comparison between the local speech recognition models and the cloud APIs are presented in Table 2.

Table 2. A comparison of on-premises models and cloud services in terms of WER and processing time

Category	System	WER, %	Processing time
Local DL	Whisper Large-v3	7	~10 s
	Wav2Vec	12	~29 s
	DeepSpeech	22	~38 s
Cloud API	Google STT	7	~10-20 s
	Azure Speech	7	~5-10 s
	Amazon Transcribe	10	~10-15 s
	IBM Watson STT	15	~10-15 s

Source: compiled by the authors

Results presented show that cloud solutions can be compared with on-premises solutions in terms of the achieved performance metrics. However, there are differences between the two solutions in relation to the latency of the system. For local systems, the latency is only associated with the computational power of the system's hardware. For cloud solutions, end-to-end latency is associated with the transmission of data to third-party servers, which can introduce a stochastic component into the system's latency. In comparison with a local system whose latency is determined by the system's hardware, end-to-end latency for a cloud system also depends on the performance of the provider's servers and network. During periods of high demand for the provider's services or during network issues, end-to-step latency for cloud services can significantly increase, making those solutions unsuitable for specific applications that require high levels of real-time processing and analysis of the data, such as monitoring information flows or audio content in real time.

Cloud solutions require the transmission of audio files to third-party servers to be stored or used for training models as prescribed by the cloud service provider. This introduces additional risks for companies whose audio files contain sensitive information, such as business or personal communications, meetings, or journalistic interviews. The use of local solutions allows the company to maintain full control over its data (Jin *et al.*, 2025). For applications requiring high levels of real-time processing, on-premises solutions have a significant advantage over cloud solutions. The average latency for both solutions is similar, but the jitter (variation in latency) for cloud solutions is significantly higher. Local systems use on-premises models that are optimised for GPUs, which allows them to produce more stable results than cloud solutions (Yang *et al.*, 2024). Therefore, while the accuracy of cloud APIs and on-premises models is comparable, cloud solutions are significantly less suitable for applications that require a high level of availability, low latency, and data control. Solutions based on the Whisper and Wav2Vec models for on-premises systems, which offer a good balance of accuracy, performance, and security, present themselves as a promising solution for building autonomous systems for the analysis of audio data containing sensitive information.

Noise robustness and linguistic characteristics of Ukrainian data

An analysis of speech recognition systems in the context of the Ukrainian language and its use in a bilingual environment has demonstrated how the features of the language have a significant influence upon the accuracy of the recognition systems. The Ukrainian language contains soft consonants, which significantly increases the difficulty of recognition systems whose acoustic models do not account for these sounds. These sounds and other language features significantly complicate the process of segmenting phonemes in recognition systems whose models are based upon HMMs and other statistical features of speech. Furthermore, the language's use in a bilingual environment, especially with Russian, introduces further challenges to recognition systems based upon these classical models. These challenges result from the use of code-switching (switching between languages within the same sentence), loanwords, and accents from different regions of Russia. Additionally, the classical models have a limited training corpus that does not contain enough examples of bilingual speech to effectively recognise the sounds and linguistic features associated with those languages when spoken in conjunction. Therefore, to increase the accuracy of classical models to Ukrainian speech, it would be necessary to either expand the training corpus to include examples of speech with these features, or to perform fine-tuning of the models to account for these specific linguistic features (Bai & Katkov, 2025).

Another reason why classical acoustic models based upon HMMs have low accuracy in the recognition of the Ukrainian language is due to the shortage of training corpora that contain speech examples in that language. Furthermore, these classical models focus upon analysing statistical features of speech within limited ranges

of the speech signal, which prevents those models from generalising well to other types of speech signals. As a result of these factors, classical systems reveal errors (WER) in recognising the Ukrainian language between 25% and 30%, which is a significant error rate for information flow and keyword detection analysis systems. In contrast, neural network models such as Wav2Vec and Whisper exhibit significantly greater error tolerance for language features and noisy environments. These models have been trained on a large number of speech signal examples that include bilingual speech. The ability of these models to learn the representations of phonemes in the Ukrainian language and to understand the language's stress and soft consonants allows the models to recognise those sounds correctly. Furthermore, the exposure of these models to bilingual speech during training allows the models to recognise code-switching between Ukrainian and Russian speech, and to reduce errors that result from that code-switching. Experiments with the Wav2Vec and Whisper models reveal WER rates for Ukrainian-Russian speech signals between 7% and 12%, which is between 2 and 3 times lower than that of classical recognition systems. The results of these experiments are presented in Table 3.

Table 3. Impact of linguistic factors on WER for different classes of systems

Factor	Classical systems (HTK, CMU Sphinx)	Local DL (Wav2Vec, Whisper)	Cloud API
Soft consonants	28-30%	7-12%	7-10%
Stress variations	25-28%	7-12%	7-10%
Bilingual environment	30-32%	10-12%	7-10%
Background noise	27-31%	8-12%	8-10%

Source: compiled by the authors

Thus, the results of this study showed that in order to effectively recognise the Ukrainian language and speech of bilingual individuals, the use of language models that were pre-trained on extensive corpora of diverse and Ukrainian language texts is required. The presence of multilingual and noisy data during the training of these language models ensures that the models have developed a stable representation of speech sounds, allowing them to account for variations in the stress of syllables, the use of soft consonants, and code-switching between languages. Furthermore, the use of multilingual neural network models, especially end-to-end transformer models for ASR, has shown a low WER even in the presence of noise in the recorded audio files.

Assessment of hardware requirements and computational complexity of ASR systems

Another of the main priorities in the selection of ASR systems for various applications is the assessment of the hardware requirements of the ASR systems. The hardware requirements of a computing system can have a major impact upon the processing time of the system's audio files, and upon whether the system is able to be deployed in the targeted application of the ASR system. For instance, ASR systems based upon frameworks like HTK, CMU Sphinx, and Kaldi have low hardware requirements for their operation; they are based upon algorithms that are performed by the computer's CPU rather than relying upon the computational power of the computer's graphics card.

However, this determinism comes at the cost of significant processing time, particularly when processing long audio streams or multi-channel signals, which can be critical for real-time applications. Modern DL systems, including DeepSpeech, Wav2Vec and Whisper, exhibit a fundamentally different hardware utilisation profile. Thanks to transformer and convolutional architectures, these models make intensive use of GPUs for parallel processing of tensor operations, enabling processing times an order of magnitude lower than in classical systems, with significantly higher accuracy.

The practical implications for deploying systems on edge devices or in local on-premise environments are reflected in the choice of model, depending on the hardware platform's limitations. Local edge devices with limited GPU power effectively integrate smaller-scale models (Tiny or Base), which, although they have a slightly higher WER (10-15%), provide real-time processing and reduced power consumption. On enterprise servers or on-premise deployments with high-power graphics cards, it is advisable to use Large models, such as Whisper Large-v3 or Wav2Vec, which combine low WER with high processing speed – a critical factor for real-time analytics systems and complex streaming tasks. Options for optimising hardware requirements include a trade-off between speed and accuracy, model selection, and the integration of compression and acceleration techniques such as pruning, distillation, or quantised representations (Table 4). Such approaches make it possible to adapt large neural network models to limited resources without a significant loss of accuracy, making them suitable for both edge devices and server-based solutions where minimising response times and conserving resources are key.

A comprehensive analysis of the results of speech recognition system evaluations indicates that effective detection of the information effect is only possible through a holistic approach that simultaneously covers accuracy, performance and the specific characteristics of the input data. Classic CPU-oriented models, such as HTK, CMU Sphinx and Kaldi, whilst providing deterministic processing, demonstrate limited suitability for analytical scenarios with highly dynamic audio streams,

as their accuracy (25-30% WER) and processing time (51-72 seconds per standard file) significantly limit the capabilities of automated classification, sentiment analysis and keyword search. Such characteristics make them primarily relevant for offline analysis of archived data or research purposes.

Table 4. Hardware requirements and performance for different system classes

System/Model	CPU	GPU	Processing time (s)	WER (%)	Suitability for edge/on-premise
HTK	+	-	72	25	limited (offline)
CMU Sphinx	+	-	51	30	limited (offline)
Kaldi	+	-	61	18	limited (offline)
DeepSpeech	+	+	38	22	on-premise, highly efficient
Wav2Vec	+	+	29	12	Edge/server use
Whisper Tiny	+	+	5-8	10	edge-devices
Whisper Base	+	+	7-10	8-10	edge/on-premise
Whisper Large-v3	+	+	10	7	on-premise, real-time analytics

Source: compiled by the authors

Local DL models, notably Wav2Vec and Whisper, demonstrate an optimal balance of performance and accuracy. Wav2Vec achieves a WER of around 12% whilst taking 29 seconds to process the audio file. Whisper Large-v3 has a minimum WER of 7% but takes only 10 seconds to process the same file, making it more suitable for applications requiring faster response times, the analysis of larger volumes of audio data, and which aim to minimise errors within the analytical process. Each of these platforms has been selected for its abilities to recognise multiple languages, phonetic variations within the Ukrainian language, and for its ability to process noisy audio files. Furthermore, the models that are optimised for GPUs can be deployed both at the edge and on-premise. Finally, whilst commercial cloud services like Google STT, Azure Speech and Amazon Transcribe also report WER values of between 7 and 10%, their latency to respond to requests, their potential for performance errors during peak times, and their data policies regarding the processing of personal data can present obstacles to their implementation within systems requiring rapid responses from the service and which require control over the data being analysed. A generalised comparison of system classes based on key parameters is presented in Table 5.

Table 5. A general comparison of system classes based on key parameters

System class	Accuracy (WER, %)	Processing time	Hardware requirements	Noise suppression capabilities	Suitability for analytical tasks
Traditional CPU	25-30	51-72 s	CPU	Average	limited, offline
Local DL (Wav2Vec)	12	29 s	CPU+GPU	High	edge/on-premise, real-time
Local DL (Whisper Large-v3)	7	10 s	CPU+GPU	Very high	edge/on-premise, real-time analytics
Cloud API	7-10	~5-20 s (depending on the network)	-	High	analytics, but infrastructure-dependent

Source: compiled by the authors

Based on the analysis, a range of practical conclusions can be formed. For tasks involving information extraction and automated analysis of text derived from audio, it is advisable to base the system on local neural network models, such as Whisper and Wav2Vec, which ensure error minimisation, noise robustness and the ability to integrate into confidential environments. The choice of a specific model must balance model size and performance: large models deliver the lowest WER but require significant computational resources, whilst smaller models can be used for rapid data processing on edge devices at the cost of a slight increase in WER. Key constraints for system implementation include hardware infrastructure, the availability of language and acoustic corpora, and the scalability of processing for large streams of audio data. Thus, the integration of local multilingual models into analytical systems provides an optimal balance between performance, accuracy and data control, creating prospects for further research in the field of automated detection of information effects and systems for real-time monitoring of speech streams.

DISCUSSION

An analysis of academic publications from 2022 to 2025 has revealed active development in ASR methods, particularly for multilingual and low-resource languages. The paper by D.A. Zimenkov & I.A. Tereikovskyi (2025) investigates the challenges of multilingual ASR systems, specifically the limited amount of annotated data,

phonetic variability in speech, the impact of acoustic noise, and differences in the intonational and rhythmic characteristics of languages. The authors applied transformer architectures and transfer learning methods, which reduce WER and increase the models' robustness to acoustic distortions and noise. Compared to the study in question, which tested local and cloud-based systems on a Ukrainian corpus with the addition of white noise at an SNR of 10 dB, the authors' approach demonstrated a similarly enhanced resistance to noise. At the same time, J.L.K.E. Fendji *et al.* (2022) showed that integrating acoustic and linguistic modelling into a single architecture simplified the signal processing pipeline and improved recognition accuracy. The study also observed that end-to-end models (DeepSpeech, Wav2Vec 2.0, Whisper) outperformed classical HMM-GMM systems in terms of accuracy and noise robustness; however, the authors' results highlighted the importance of customising dictionaries for specialised terminology. In turn, current experiments with the Ukrainian language have shown that pre-trained models may require additional fine-tuning for locally specific corpora.

A comprehensive review of models for regional languages was conducted by S. Shaikh *et al.* (2024). The authors' experiments demonstrated that self-supervised learning and fine-tuning substantially improved accuracy for low-resource languages. A similar trend was observed in the current study: Wav2Vec 2.0, which utilises self-supervised learning on large corpora, demonstrated improved accuracy in Ukrainian language recognition compared to other models. A. Senyk (2025) highlighted the importance of fine-tuning models for high-quality language recognition and generation. In the present study, local modifications and fine-tuning of models on Ukrainian data improved accuracy and reduced WER compared to "clean" pre-trained models. Meanwhile, C. Chen's (2025) thesis conducted an in-depth study of strategies for adapting large ASR models to specific linguistic and acoustic conditions, including fine-tuning and transfer learning methods. The author demonstrated that adaptation significantly reduced WER in low-resource scenarios and improved the models' robustness to noise. Current experiments have shown that additional local fine-tuning on Ukrainian corpora yielded noticeably better accuracy for short recordings, which can be explained by linguistic features and the size of the training dataset.

The accuracy and adaptability of end-to-end ASR models in complex linguistic contexts were evaluated by P. Jain & A. Bhowmick (2025). They demonstrated that self-supervised learning on large unlabelled corpora, followed by fine-tuning, significantly improved model performance, particularly in languages with limited resources. Similar results were also obtained in the study presented in this paper. Wav2Vec 2.0, which was trained on multilingual data and later trained on Ukrainian recordings, also exhibited a significant reduction in WER compared to the baseline models, confirming the universality of these self-supervised learning approaches. Another Turkish language recognition system was also implemented by H. Polat *et al.* (2024). The authors found that the use of Low-Rank Adaptation (LoRA) allowed their system to significantly reduce the computational costs needed to train the model while maintaining accuracy, and that fine-tuning the model on Turkish audio corpora improved the model's noise robustness. Current results correlate with this conclusion: the application of Whisper with fine-tuning on Ukrainian data improved accuracy and robustness to noise distortions. The significance of standardised data processing pipelines and multi-criteria evaluation schemes for comparing models across different linguistic and acoustic contexts was highlighted by M. Trenton (2025) in a comprehensive review of multilingual and multimodal ASR systems. The study adopted a similar approach, creating a unified audio processing pipeline that incorporates volume normalisation, VAD segmentation, and a multi-criteria evaluation of accuracy, noise robustness and computational requirements.

M. Si *et al.* (2024) proposed the Lexical Error Guard method, demonstrating that integrating an LLM into post-processing reduced the number of insertions, deletions and word substitutions compared to baseline algorithms, thereby improving the overall ASR accuracy. Compared to the presented study, which employed Wav2Vec 2.0 and Whisper for the Ukrainian language and evaluated accuracy on test audio recordings with white noise added at a SNR of 10 dB, the results of Lexical Error Guard demonstrated a similar reduction in WER; however, this approach addressed local fine-tuning of models without additional integration of large language models for post-processing, which was reflected in slightly higher error rates for complex words and numerical entities. A. Bhandari & G. Harit (2026) demonstrated that correction algorithms significantly improved accuracy in short phrases and reduced errors associated with phonetically similar words. A similar effect was observed in the presented study: additional text normalisation and punctuation restoration rules reduced the WER for short entries, confirming the universality of post-processing approaches for low-resource languages.

Wav2Vec 2.0 was the subject of research by G. Shekoufandeh (2023) and Q. Li (2024). G. Shekoufandeh found that age and speech characteristics had a significant impact on model accuracy, whilst additional fine-tuning on specific corpora substantially improved the results. Meanwhile, Q. Li supplemented these findings and demonstrated that retraining on low-resource language corpora, considering phonetic differences between related languages, can reduce WER and improve word segmentation accuracy. Compared to the study in question, which utilised fine-tuning of Wav2Vec 2.0 and Whisper on Ukrainian corpora, the authors' results

confirmed the effectiveness of the fine-tuning approach for low-resource languages, but also demonstrated that success depended on the degree of phonetic similarity between the source and target languages.

V. Delić *et al*. (2019) conducted a retrospective review of the development of TTS and STT technologies for South Slavic languages with limited resources. They emphasised the importance of multilingual and multimodal architectures, standardised data processing pipelines, and multi-criteria performance evaluation schemes for improving the accuracy and robustness of models. S. Nithya *et al*. (2025) applied multimodal approaches and specialised data normalisation strategies to NLP, which significantly improved performance in noisy environments. Compared to the present study, which evaluated local and cloud-based STT models for the Ukrainian language, the study determined that such multimodal approaches provide increased robustness, but they require significantly greater computational resources and a more complex data processing pipeline, whereas the current experiments addressed the optimisation of local models with controlled resources. At the same time, Y. Li *et al*. (2025) noted that for low-resource dialects, the availability of well-structured training corpora and effective text post-processing to correct typical errors is critical.

A study by S. Saranya *et al*. (2025) showed that combining ASR with an automatic abstract summarisation system reduced the error rate whilst increasing the practical benefit for the end user. The authors' results demonstrated expanded practical applications of STT, but emphasised that additional modules can introduce delays and require extra computational resources, which explains the partial lack of correlation in processing speed between studies. These results were expanded upon by K. Fatehi (2023), who showed that models pre-trained on large unlabelled audio corpora and then fine-tuned on limited labelled data demonstrated a significant reduction in WER and increased robustness to noise. Compared to the presented study, the author's approach confirmed the effectiveness of self-supervised learning for low-resource languages; however, it required significantly greater computational resources and more complex audio pre-processing strategies.

S.M.H. Amiri (2025) investigated multilingual NLP systems and speech recognition for global communication, integrating various ASR models into multi-domain scenarios. The author demonstrated that combining different architectures and using language-specific adaptation improved accuracy for languages with limited data, whilst reducing the need for large annotated corpora. In turn, S.M. Zahid & S.A. Qazi (2025) showed that variations in pitch and speech rate during training significantly improve the models' ability to generalise to new recordings and reduce the number of common errors. Compared to the presented study, the results of the studies reviewed indicated a potential path to improving accuracy through generative data augmentation methods, which explains why current models demonstrated a higher WER in some scenarios. Y. Yang *et al*. (2025) presented GigaSpeech 2 – a large multilingual corpus for low-resource languages with automated collection, transcription and post-processing of audio data. The study demonstrated that large and diverse corpora improved performance of end-to-end ASR models, particularly in noisy and multilingual environments. Compared to the study in question, the results by Y. Yang *et al*. confirmed the importance of data scale and diversity, which explains the difference in the scalability of systems and their performance on large and heterogeneous corpora. Thus, the analysis of the literature confirmed the key role of self-supervised learning, multilingual adaptation, generative data augmentation and large corpora in improving the accuracy and robustness of ASR in resource-constrained settings. A comparison with the current study showed that the local and cloud-based Wav2Vec 2.0 and Whisper models delivered stable performance for the Ukrainian language, but differences in data scale, the use of self-supervised learning and multilingual architectures accounted for some discrepancies in accuracy and processing speed metrics.

CONCLUSIONS

The study found that ASR performance depended significantly on the model architecture and the computational pipeline. Classic statistical systems based on HMM (HTK, CMU Sphinx, Kaldi) demonstrated the worst performance, with WERs ranging from 18-30% and processing times of 51-72 seconds for the reference audio file. End-to-end neural network models achieved significantly higher accuracy: DeepSpeech achieved a WER of ≈22%, Wav2Vec – ≈12%, and Whisper Large-v3 – ≈7%. Thus, the transition from traditional cascade approaches to modern transformer architectures has reduced the number of errors by a factor of 3-4. It was also confirmed that even a slight increase in WER leads to a noticeable deterioration in the results of subsequent NLP tasks (classification, sentiment analysis, keyword search), which highlights the critical role of the ASR module in the entire analytical pipeline. Experimental evaluation under noise conditions and with Ukrainian code-switching showed that classical acoustic models achieved a WER of 25-32%, whilst multilingual transformer architectures achieved 7-12%. Performance analysis revealed a significant advantage for GPU-optimised local models: Whisper Large-v3 processed the reference audio file in approximately 10 seconds, Wav2Vec in 29 seconds, whilst classical systems required over 60 seconds. Cloud-based ASR services demonstrated comparable accuracy (7-10% WER) but were characterised by variable response times due to network delays and load.

In summary, the study's findings confirmed that modern local end-to-end neural network models, particularly Whisper and Wav2Vec, strike an optimal balance between accuracy, speed and data control. Achieving

a WER of 5-7% created qualitatively new conditions for building reliable systems for analysing audio content and detecting information effects, as the improvement in ASR accuracy directly enhanced all subsequent processing stages. The study demonstrated that the transition to deep end-to-end architectures is a key prerequisite for the practical use of such systems in real-world multilingual and noisy information environments. The main limitation of this study is the dependence of the results on the availability of representative Ukrainian and bilingual audio corpora, as well as on hardware resources, which limit the scale of the experiments and the comprehensiveness of model evaluation in various real-world scenarios. Further research should expand specialised Ukrainian and code-switching corpora, integrating adaptive fine-tuning and edge-level optimisations, as well as conducting an in-depth analysis of the impact of ASR errors on the final metrics of NLP components in information effect detection systems.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Amazon Transcribe. (n.d.). Retrieved from https://aws.amazon.com/transcribe.

[2] Amiri, S.M.H. (2025). Beyond language barriers: Multilingual NLP and voice recognition for global connectivity. *International Journal of Science and Research Archive*, 15(2), 406-419. doi: 10.30574/ijsra.2025.15.2.1346.

[3] Azure Speech in Foundry Tools. (n.d.). Retrieved from https://azure.microsoft.com/en-us/products/ai-services/ai-speech.

[4] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th international conference on neural information processing systems* (pp. 12449-12460). Red Hook: Curran Associates Inc. doi: 10.5555/3495724.3496768.

[5] Bai, Y.V., & Katkov, Yu.I. (2025). Review of research on non-audio recognition using neural networks. *Scientific Notes of the State University of Information and Communication Technologies*, 1(7), 106-115. doi: 10.31673/2786-8362.2025.018603.

[6] Bhandari, A., & Harit, G. (2026). Post-ASR correction for low-resource Rajasthani language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 25(3), article number 15. doi: 10.1145/3793254.

[7] Bouchelligua, W. (2025). VoiceX: An end-to-end speech recognition model for low-resource languages using transfer learning. *Arabian Journal for Science and Engineering*, 51, 11893-11920. doi: 10.1007/s13369-025-10859-7.

[8] Chen, C. (2025). *Advancing speech-to-text adaptation for large speech models*. Singapore: Nanyang Technological University. doi: 10.32657/10356/201854.

[9] CMU Sphinx. (n.d.). Retrieved from https://github.com/cmusphinx.

[10] DeepSpeech. (n.d.). Retrieved from https://github.com/mozilla/DeepSpeech.

[11] Delić, V., Perić, Z., Sečujski, M., Jakovljević, N., Nikolić, J., Mišković, D., Simić, N., Suzić, S., & Delić, T. (2019). Speech technology progress based on new machine learning paradigm. *Computational Intelligence and Neuroscience*, 2019(1), article number 4368036. doi: 10.1155/2019/4368036.

[12] Dhahbi, S., Saleem, N., Bourouis, S., Berrima, M., & Verdú, E. (2025). End-to-end neural automatic speech recognition system for low resource languages. *Egyptian Informatics Journal*, 29, article number 100615. doi: 10.1016/j.eij.2025.100615.

[13] Dumyn, A., Basystiuk, O., Dumyn, I., & Shakhovska, N. (2025). Performance evaluation of automatic speech recognition systems for Ukrainian and English languages. In P. Štarchoň, S. Fedushko & K. Gubíniova (Eds.), *Developments in information and knowledge management systems for business applications* (pp. 423-443). Cham: Springer. doi: 10.1007/978-3-031-95955-4_23.

[14] Fatehi, K. (2023). *Self-supervised learning for automatic speech recognition in low-resource environments*. Nottingham: University of Nottingham.

[15] Fendji, J.L.K.E., Tala, D.C.M., Yenke, B.O., & Atemkeng, M. (2022). Automatic speech recognition using limited vocabulary: A survey. *Applied Artificial Intelligence*, 36(1), article number 2095039. doi: 10.1080/08839514.2022.2095039.

[16] Froiz-Miguez, I., Fraga-Lamas, P., & Fernández-CaraméS, T.M. (2023). Design, implementation, and practical evaluation of a voice recognition based IoT home automation system for low-resource languages and resource-constrained edge IoT devices: A system for Galician and mobile opportunistic scenarios. *IEEE Access*, 11, 63623-63649. doi: 10.1109/ACCESS.2023.3286391.

[17] Google Cloud. (n.d.). *Cloud Speech-to-Text overview*. Retrieved from https://cloud.google.com/speech-to-text/docs/speech-to-text-requests.

[18] Hidden Markov Model Toolkit. (n.d.). Retrieved from https://github.com/mxochicale/htk.

[19] IBM Watson Speech to Text. (n.d.). Retrieved from https://www.ibm.com/products/speech-to-text.

[20] Jain, P., & Bhowmick, A. (2025). Comparative performance analysis of end-to-end ASR models on Indo-Aryan and Dravidian languages within India's linguistic landscape. *EURASIP Journal on Audio, Speech, and Music Processing*, 2025(1), article number 10. doi: 10.1186/s13636-025-00395-5.

[21] Jin, W., Cao, Y., Su, J., Wang, D., Zhang, Y., Xue, M., Hao, J., Dong, J.S., & Yang, Y. (2025). Whispering under the eaves: Protecting user privacy against commercial and LLM-powered automatic speech recognition systems. *ArXiv*. doi: 10.48550/arXiv.2504.00858.

[22] Kaldi. (n.d.). Retrieved from https://github.com/kaldi-asr/kaldi.
[23] Kumar, L.A., Renuka, D.K., Chakravarthi, B.R., & Mandl, T. (2024). *Automatic speech recognition and translation for low resource languages*. Hoboken: John Wiley & Sons. doi: 10.1002/9781394214624.
[24] Li, Q. (2024). *Fine-tuning Cantonese based on Wav2vec 2.0 XLRS model that pretrained on Mandarin Chinese to improve ASR performance*. Groningen: University of Groningen.
[25] Li, Y., Wang, Y., Hoi, L.M., Yang, D., & Im, S.-K. (2025). A review on speech recognition approaches and challenges for Portuguese: Exploring the feasibility of fine-tuning large-scale end-to-end models. *EURASIP Journal on Audio, Speech, and Music Processing*, 2025(1), article number 3. doi: 10.1186/s13636-024-00388-w.
[26] Liu, W. (2025). *Enhancing efficiency and reliability in automatic speech recognition systems*. Shatin: The Chinese University of Hong Kong.
[27] Mujtaba, D.F., & Mahapatra, N.R. (2025). Automatic speech recognition APIs: Models, features, and challenges. In H.R. Arabnia, L. Deligiannidis, F. Shenavarmasouleh, S. Amirian & F.G. Mohammadi (Eds.), *11th international conference: Computational science and computational intelligence* (pp. 33-48). Cham: Springer. doi: 10.1007/978-3-031-94940-1_3.
[28] Nithya, S., Sengar, A.S., Jayalakshmi, K., Kokulavani, K., Joel, P.J., & Srinivasulu, M. (2025). Multilingual and robust speech recognition: Leveraging advanced machine learning for accurate and realtime natural language processing. In *Proceedings of the 1st international conference on research and development in information, communication, and computing technologies* (pp. 513-518). Nagapattinam: SciTePress. doi: 10.5220/0013885700004919.
[29] OpenAI. (2022). *Whisper*. Retrieved from https://openai.com/uk-UA/index/whisper.
[30] Polat, H., Turan, A.K., Koçak, C., & Ulaş, H.B. (2024). Implementation of a whisper architecture-based Turkish automatic speech recognition (ASR) system and evaluation of the effect of fine-tuning with a low-rank adaptation (LoRA) adapter on its performance. *Electronics*, 13(21), article number 4227. doi: 10.3390/electronics13214227.
[31] Rakhimova, D., Duisenbekkyzy, Z., & Adali, E. (2025). Investigation of ASR models for low-resource Kazakh child speech: Corpus development, model adaptation, and evaluation. *Applied Sciences*, 15(16), article number 8989. doi: 10.3390/app15168989.
[32] Saranya, S., Bharathi, B., Dhanya, S.G., & Krishnakumar, A. (2025). Real-time continuous Tamil dialect speech recognition and summarization. *Circuits, Systems, and Signal Processing*, 44(4), 2855-2881. doi: 10.1007/s00034-024-02950-5.
[33] Senyk, A. (2025). *Implementing text-to-speech synthesis with voice cloning for the Ukrainian language*. Lviv: Ukrainian Catholic University.
[34] Shaikh, S., Pereira, K.W., Sahay, S., Lopes, A., & Parshionikar, S. (2024). An extensive review: Models for regional language speech recognition. In *2024 4th Asian conference on innovation in technology* (pp. 1-8). Pimari Chinchwad: IEEE. doi: 10.1109/ASIANCON62057.2024.10837903.
[35] Shekoufandeh, G. (2023) *Evaluation of wav2vec 2.0 speech recognition for the elderly Frisian population*. Groningen: University of Groningen.
[36] Si, M., Cobas, O., & Fababeir, M. (2024). Lexical error guard: Leveraging large language models for enhanced ASR error correction. *Machine Learning and Knowledge Extraction*, 6(4), 2435-2446. doi: 10.3390/make6040120.
[37] Trenton, M. (2025). A comprehensive survey on multilingual and multimodal automatic speech recognition systems. *Journal of Computer Technology and Software*, 4(10).
[38] Yang, B., Hu, Q., Xie, W., Wang, X., Luo, W., & Zhang, Q. (2024). PDAssess: A privacy-preserving free-speech based Parkinson's disease daily assessment system. In *Proceedings of the 21st ACM conference on embedded networked sensor systems* (pp. 251-264). New York: Association for Computing Machinery. doi: 10.1145/3625687.3625805.
[39] Yang, Y., *et al*. (2025). GigaSpeech 2: An evolving, large-scale and multi-domain ASR corpus for low-resource languages with automated crawling, transcription and refinement. In *Proceedings of the 63rd annual meeting of the association for computational linguistics* (Vol. 1; pp. 2673-2686). Vienna: Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.135.
[40] Zahid, S.M., & Qazi, S.A. (2025). Pitch-speed feature space data augmentation for automatic speech recognition improvement in low-resource scenario. *IEEE Access,* 13, 86998-87014. doi: 10.1109/ACCESS.2025.3569579.
[41] Zimenkov, D.A., & Tereikovskyi, I.A. (2025). Analysis of problems in speech recognition methods. *Scientific Notes of the V.I. Vernadsky TNU. Series: Technical Sciences*, 36(75(3)), 198-205. doi: 10.32782/2663-5941/2025.3.2/27.

Аналіз методів та підходів до задачі обробки аудіоданих

Владислав Радін

Аспірант

Державний університет «Київський авіаційний інститут»
03058, просп. Любомира Гузара, 1, м. Київ, Україна
https://orcid.org/0009-0009-1101-6888

Мирослав Рябий

Кандидат технічних наук, доцент
Державний університет «Київський авіаційний інститут»
03058, просп. Любомира Гузара, 1, м. Київ, Україна
https://orcid.org/0000-0002-9651-9135

Анотація. Актуальність дослідження визначається потребою комплексної оцінки сучасних систем автоматичного розпізнавання мовлення в умовах української мовної специфіки з урахуванням компромісу між точністю, продуктивністю, апаратними вимогами та впливом якості транскрипції на подальші інформаційно-аналітичні процеси. Мета дослідження полягала в порівнянні наявних методів і технологій розпізнавання мовлення для побудови системи оцінки інформаційного впливу. У дослідженні використовувалися традиційні моделі прихованих марковських процесів та гауссові суміші, гібридні нейронні мережі, локальні нейромережеві моделі DeepSpeech, Wav2Vec 2.0 і Whisper, хмарні сервіси розпізнавання мовлення від Google, Amazon, Microsoft та IBM. Здійснювалася обробка аудіо з використанням виявлення мовлення та мел-частотних коефіцієнтів, фіксувалися точність розпізнавання слів, час обробки на процесорі та графічному процесорі, затримка та нестабільність відповіді, а також оцінювався вплив помилок розпізнавання на конвеєр обробки природної мови з токенізацією, тематичною класифікацією та аналізом настрою. У ході дослідження встановлено істотну залежність точності та продуктивності Automatic Speech Recognition від архітектури моделей і апаратної підтримки. Класичні Hidden Markov Models продемонстрували найнижчу ефективність із Word Error Rate 18-30 % та часом обробки 51-72 с, що підтвердило їх обмежену придатність для української мови в умовах шуму й двомовного мовлення. Локальні end-to-end нейромережеві моделі забезпечили принципово кращі показники: DeepSpeech досяг 22 % Word Error Rate, Wav2Vec – близько 12 %, а Whisper Large-v3 – приблизно 7 %, що відповідає зменшенню кількості помилок у 3-4 рази. Було показано, що багатомовні трансформерні архітектури характеризуються підвищеною стійкістю до фонетичних варіацій української мови та код-свічингу. Оптимізовані локальні моделі забезпечили обробку майже в реальному часі (Whisper Large-v3 ≈10 с на файл), суттєво перевершивши класичні підходи за швидкодією. Хмарні сервіси продемонстрували співставну точність (7-10 % Word Error Rate), однак виявили залежність від мережевих затримок і обмеження щодо конфіденційності. Узагальнено, Whisper і Wav2Vec визначено як оптимальну основу для систем аналізу аудіоконтенту та виявлення інформаційного ефекту, оскільки вони забезпечують найкращий баланс між точністю, продуктивністю та контролем над даними. Отримані результати можуть бути застосовані у розробці систем автоматичної транскрипції, моніторингу та аналізу аудіопотоків, розробці сервісів голосової аналітики та модулів раннього виявлення інформаційних впливів

Ключові слова: speech-to-text; автоматичне розпізнавання мовлення; Whisper; Wav2Vec; обробка аудіопотоків; Word Error Rate; нейронні мережі