UDC 655.41:004.91:004.8 DOI: https://doi.org/10.30837/PMW.2026.T1.335

HYBRID MULTIMODAL TEXT DIGITIZATION FOR PUBLISHING AND PRINTING

Kulchytska Kh.
PhD in Engineering, Associate Professor, Multimedia Technologies Department,
Institute of Printing Art and Media Technologies
of the Lviv Polytechnic National University
ORCID ID: 0000-0002-6184-988X

Abstract. *This paper investigates AI application in text input for the publishing sector. It establishes a classification system for digitization methods based on text complexity and defines key selection criteria. To improve the processing of complex content, the author proposes a hybrid Optical Character Recognition (OCR) and Automatic Speech Recognition (ASR) approach, alongside a specialized multimodal algorithm integrated into publishing workflows.*

Keywords: *artificial intelligence, text digitization, text complexity group, publishing system.*

Introduction

In the printing industry, artificial intelligence (AI) technology is designed to streamline routine tasks and improve the accuracy and efficiency of prepress processes, such as handwriting recognition, automatic proofreading, and text formatting. However, the most pressing need is to improve text processing efficiency, specifically by replacing manual text entry via keyboard, which requires inputting thousands of characters into a computer. Keyboard typing is a labor-intensive process that requires knowledge of foreign languages, time, attention, and a high level of skill on the part of the operator, and it causes rapid fatigue in the typist. That is why research into text input technologies for PCs via scanning, in a manner natural to humans – such as voice or even thought using neural interfaces – is relevant and promising in the printing industry and requires more detailed analysis.

Purpose and Objectives of the Study

According to experts, the implementation of AI technologies in the printing processes involved in preparing text information can increase productivity by up to 70% and reduce the number of errors by 35-40% [1, 2]. However, the choice of the optimal text recognition and digitization technology depends on many factors: the type of input data, language, volume, and complexity of the text, specifically, the presence of specialized terms, formulas, tables, font formatting, and other text highlights.

Numerous studies have been devoted to the problem of digitization and text recognition, for example, in [3]; however, most of them focus on individual aspects or technologies without providing a comprehensive analysis in the context of printing production. The study [4] examines ways to improve the efficiency of OCR

technologies, but does not account for the specific characteristics of different types of printed materials and does not provide concrete recommendations on selecting technologies based on text characteristics.

The situation is similar with voice text input. Studies conducted on this topic are also relevant to printing [5, 6], although they barely address text complexities such as tables, formulas, and highlighting, and focus more on punctuation placement. The impact of text complexities on voice input technology has been partially investigated in [7]. An analysis of the technologies shows that most text digitization systems have problems with tables, formulas, and punctuation placement, and do not support the Ukrainian language either partially or at all. Replacing the routine task of manual keyboard typing requires an analysis and comparison of the capabilities of AI-based technologies for application in printing.

The problem of integrating text complexity and input technology into the publishing and printing system remains open and requires further research.

Main Section

Text digitization technologies can be divided into several groups based on the type of input data: Optical Character Recognition (OCR) is used to digitize printed text on paper media; Intelligent Character Recognition (ICR) is used for handwritten text; Automatic Speech Recognition (ASR) is used for voice input; audio file transcription is used for pre-recorded audio and video files; Brain-Computer Interfaces (BCI) – for converting brain signals directly into text.

Examples of OCR technology: ABBYY FineReader – one of the best options, effectively recognizes Ukrainian language and punctuation marks; Tesseract OCR – a free open-source solution that also supports the Ukrainian language; Google Drive/Google Docs has built-in OCR, effective with high-quality images; online services: OnlineOCR.net, NewOCR.com.

An example of an ICR (Handwriting Recognition, OCR + AI) system is Google Lens, which recognizes handwritten text using a smartphone camera; Microsoft OneNote Ink-to-Text – converts handwriting into text.

Major ASR solutions: Google Docs (voice input) – works for free in Google Docs via Chrome; SpeechTexter (speechtexter.com) and Dictation.io – online services for voice input; Whisper (OpenAI) – a high-precision AI-based system for speech recognition with support for the Ukrainian language [8]. Audio file transcription includes speaker recognition and automatic formatting features (Whisper, Fathom, Fireflies.ai, Otter.ai, Transkriptor).

Neural interfaces use Brain-Computer Interface (BCI) technology. These are experimental systems that read the user's neural signals and convert them into text (NeuroPort, Brain-to-Text KIT, Synchron Stentrode, Facebook Meta BCI).

Classification of text by recognition complexity. To determine the optimal digitization method, the text information was classified according to the complexity of its recognition. In printing, this includes various types of formatting, the presence of

formulas, tables, and words in a foreign language. Depending on these complexities, the text information was divided into four complexity groups. Simple text with minor typographic complexities (up to 10%) – typographic complexities refer to variations in font saturation, slant, width, size, and typeface – and non-typographic formatting (such as text set in a different style, format, color, or with borders). Examples of printed materials containing text from the first difficulty group include children's books and fiction. The second group of text information contains up to 25% complexity. This is formatted text with standard elements (typographic and non-typographic highlighting, lists, simple tables), for example, a Ukrainian language textbook for elementary school students. The third group includes complex text with non-standard elements up to 50% (formulas, diagrams, multidimensional tables, words in a foreign language, specialized terminology); an example is technical literature. The fourth group includes complex text containing more than 50% of such elements (text in a foreign language, formulas, complex structured tables), such as dictionaries, a physics textbook, and scientific articles.

Research Methods and Results

The following criteria are proposed for an objective assessment of the technologies' effectiveness:

– recognition accuracy (%) – the percentage of correctly recognized words;

– processing speed (characters/min) – the average number of characters processed per unit of time;

– resource intensity (points on a scale of 1-10) – requirements for equipment, networks, energy consumption of the technology, and human resources;

– economic indicator – nominal cost ($/workstation), defined as the average subscription cost.

A mandatory requirement for text digitization technology is the software's ability to work with Ukrainian-language text and the Cyrillic alphabet, as well as its ability to adapt to various input data formats.

The technologies were tested on ten text fragments, each 1000 characters long, for each complexity group.

Recognition accuracy is the primary criterion for evaluating the performance of text recognition systems and is characterized by the World Error Rate (*WER*) [9]. The *WER* is calculated by comparing two text strings: the recognition result and the original text. This comparison is performed using a dynamic programming algorithm that calculates the Levenshtein distance [10]. The Levenshtein distance is the weighted sum of editing operations with the minimum number of word substitutions (*Z*), deletions (*D*), and insertions (*I*):

$$WER = (Z + D + I) / N , \quad (1)$$

where N is the total number of words in the phrase.

With the advancement of speech recognition technologies, the *WER* metric is increasingly approaching zero; therefore, the Word Recognition Rate (*WRR*) metric was used. *WRR* = 1 – *WER*. The recognition rate was taken as the average value across different software versions.

Based on the obtained data, a mathematical model was developed to determine the effectiveness of text input technology depending on the selected criteria.

The effectiveness of the method (*E*) was calculated using the formula:

$$E = 0.40WRR_{norm} + 0.28S_{norm} + 0.12R_{norm} + 0.18C_{norm}. \quad (2)$$

The weight coefficients for each criterion were determined using the hierarchical analysis method. To do this, the criteria were evaluated on a pairwise comparison scale, pairwise comparison matrices were obtained, and the overall weight of each criterion was determined. The weight coefficients of the criteria, taking into account the significance of each parameter for printing production, were as follows: 0.40 for accuracy (*WRR*), as the most important parameter; 0.28 for speed (S); 0.12 for resource intensity (R); 0.18 for cost (C).

The parameters were first normalized on a 0-1 scale:

$$WRR_{norm} = WRR/100;\ S_{norm} = S/S_{max},$$

where S_{max} is the maximum speed among the technologies under study.

The resource intensity and cost criteria inversely affect efficiency, therefore

$$R_{norm}=(10-R)/10;\ C_{norm}=(C_{max}-C)/C_{max},$$

where C_{max} is the maximum cost among all the technologies under study.

Example of calculating the performance of OCR (ABBYY) technology for text of the second difficulty level: normalized parameters

$$WRR_{norm}=98.5/100=0.985,\ S_{norm}=3500/4200=0.833,$$
$$R_{norm}=(10-7)/10=0.3,\ C_{norm}=(250-250)/250=0.$$
$$E = 0.40\times0.985 + 0.28\times0.833 + 0.12\times0.3 + 0.18\times0 = 0.663.$$

The calculation results are presented in the table 1.

Table 1 – Comparison of text digitization technologies for the second complexity group*

Technology	Accuracy *WRR*, (%)	Speed *S* (characters/min)	Resource consumption *R*, (points)	Cost, C ($/year)	Efficiency, *E*
OCR (ABBYY)	98,5	3500	7	165	0,663
OCR (Tesseract)	95,5	2800	5	0	0,809
ICR	90,0	1500	6	180	0,558
ACR (Google)	94,8	4200	8	150	0,755
ACR (Whisper)	96,5	4000	7	120	0,782
Transkriptor	91,0	3000	4	80	0,772

*Neural interface-based technology is currently under development and was therefore not included in the comparison

OCR technology (Tesseract) offers the best performance due to its high accuracy and low cost; however, OCR (ABBYY) provides higher accuracy and speed. For the second group of text complexity, which is the most common in terms of the volume of text information in publications, it is recommended to use ACR (Whisper) voice input technology, which is becoming competitive with OCR (Figure 1). Handwritten text recognition technology is, for now, the least effective.

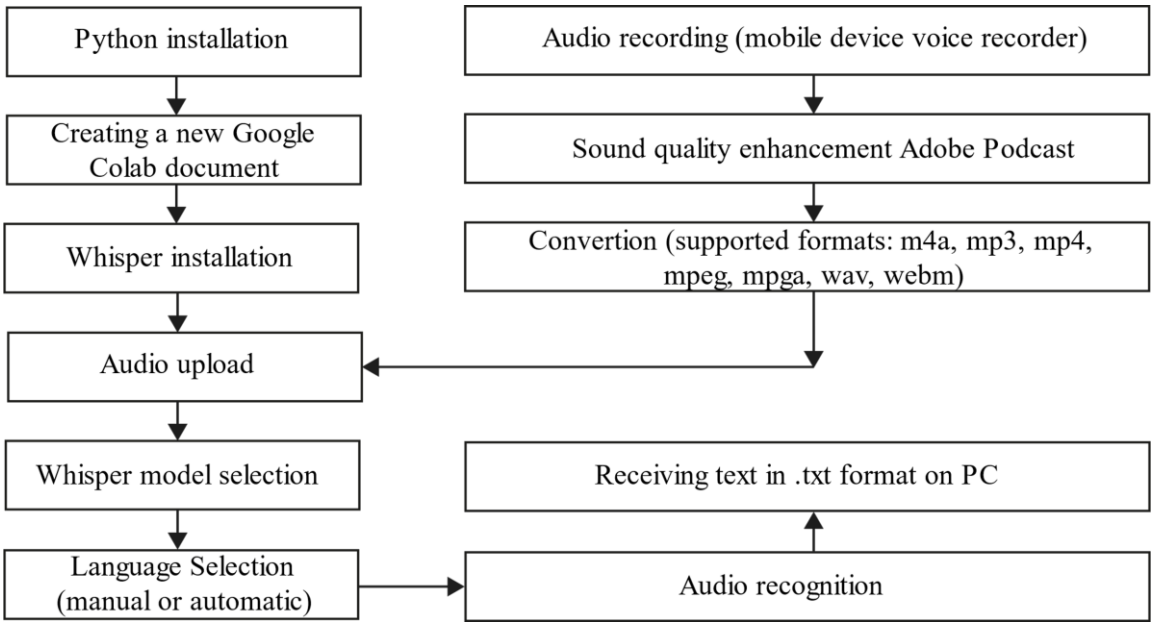

Figure 1 – Diagram of voice-to-text input using the Whisper app

Compared to manual keyboard typing, the technology for entering text into a publishing system using ACR (Whisper medium) voice input is dozens of times faster (the speed of manual typing for the second complexity group is 112 characters per minute).

The Google Colab cloud platform was selected for this study, as it provides access to the powerful GPUs required for the proper operation of the large-v2 and large-v3 models. Unlike local installations or paid applications, this method guarantees high computing power. This stage involves setting up the Python environment and deploying the Whisper library.

Data preparation involves improving audio quality as needed using Adobe Podcast to minimize noise, which is critical for reducing the WER metric. Files are also converted to compatible formats (m4a, mp3, wav, etc.), ensuring the versatility of the input data for the model.

The process is based on interacting with the Whisper model, with the ability to select parameters (language, model variant) depending on whether speed or accuracy is prioritized. The final stage is the automatic generation of text in .txt format, ready for further linguistic or quantitative analysis without additional editing.

As text complexity increases, optical character recognition (OCR) remains the best option, since AI technologies for dictating structured tables in Ukrainian are not yet available (Figure 2). Voice input can be used to fill out tables in Ukrainian using Google Assistant + Google Sheets or Microsoft Dictation (Windows 11, Excel). Table

Transformer also works with Ukrainian tables, but the interface is in English. Markdown, LaTeX, and CSV generate tables in Ukrainian and work seamlessly with the Cyrillic alphabet.

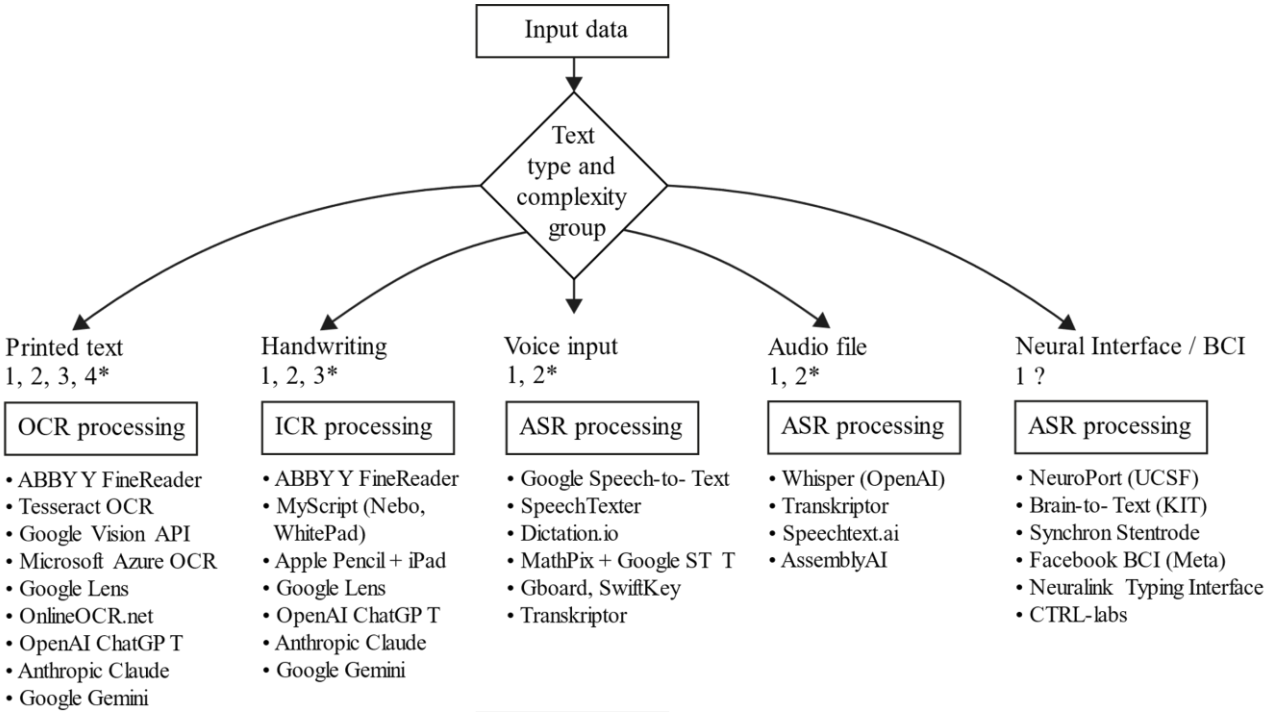

Figure 2 – Selection of text input technology for a publishing system based on the complexity group of textual information (*partial processing)

ICR technology is not suitable for complex handwritten formulas. For typesetting formulas, we recommend Overleaf (LaTeX), which supports the Ukrainian language via the babel and polyglossia packages. If formulas are typeset in Latin script, use ChatGPT (LaTeX), which generates prompts, explanations, and LaTeX code in Ukrainian and can adapt formulas to the style of Ukrainian publications.

In the printing industry, input data typically comes in various formats: text files, printed text, scans, handwritten text, and audio files. In such cases, it is recommended to use combined (multimodal) digitization technologies that combine the advantages of several technologies. The simultaneous use of different channels for inputting text information improves text accuracy; for example, combining OCR and ASR technologies.

The combined OCR and ASR method includes the following steps.

1. Parallel processing of input data using OCR and ASR.
2. Comparing recognition results and identifying discrepancies.
3. Analysis of discrepancies using a neural network module.
4. Selection of the most likely option based on contextual analysis.
5. Post-processing and correction using NLP (Natural Language Processing) tools.

Examples of several types of source texts for a single publication: a manuscript + an audio file + printed tables, or a scanned article + a translation + AI-generated missing sections. The process of digitizing such textual information for print

reproduction involves three main stages. The reading and recognition stage involves OCR technology (ABBYY FineReader, Tesseract, Google Drive OCR) from photos or scans of pages, as well as speech recognition (Google Docs Voice, OpenAI Whisper) via text dictation or audio recording followed by recognition. The second stage of post-processing and correction involves automatic punctuation correction (LanguageTool), contextual checking, and error correction. The result saving stage involves exporting to TXT, DOCX, EPUB, or PDF and integration with publishing systems.

Alternatives to keyboard typing not only speed up prepress preparation but also improve the accuracy and quality of printed materials. The choice of a specific technology also depends on the project's specifics and the qualifications of the staff.

Drawbacks and characteristics of text post-processing using AI models for printing purposes. For publishing and printing processes, it is crucial that AI does not add to or alter the text. To achieve this, rule-based prompting must be used, where clear constraints are imposed on the models, specifically prohibiting the addition of new words not present in the original, preserving sentence structure and word count, and prioritizing phonetic similarity when correcting errors. Even for the most advanced ASR systems, this approach significantly reduces the WER metric.

Many models use Generative Error Correction (GEC) technology, where an LLM can reconstruct text based on context. For example, GPT-5.5 and Claude can correct transcription errors by focusing on the acoustic or visual similarity of words, which is unacceptable for printing when high reproduction accuracy is required. In this case, when recognizing complex texts, the model can understand that a misrecognized word is phonetically close to the contextually correct term [11].

One of the biggest drawbacks of using modern models for text post-processing is the risk of degrading the quality of text that has already been correctly recognized – a phenomenon known as "overcorrection." This is particularly noticeable in high-precision systems (such as Whisper Large), where the intervention of an LLM (e.g., GPT-5.5) can lead to changes in style or the removal of necessary repetitive text fragments [12]. Despite this, Whisper is currently the gold standard for accuracy in the Ukrainian language, with a WRR of over 96%.

GPT-5.5 demonstrates a high level of overall accuracy. The model is capable of processing text, images, and audio in real time, making it ideal for a combined method (OCR + ASR). However, it may "invent" text and be less effective in specific technical contexts.

Using Claude (Opus and Sonnet series), you can digitize large volumes of text data in a single pass without losing the original structure. The model delivers superior results in data structuring, parsing legal documents, and complex texts where accurate corrections are essential.

For text information in complexity groups 1 and 2, it is advisable to use fast and inexpensive models with basic confidence filtering.

For text information in complexity groups 3 and 4, it is necessary to use Claude 4.6 or GPT-5.5 with N-best hypotheses and complex prompts, which ensure higher digitization accuracy.

The use of N-best most likely recognition variants is a key element of modern multi-stage post-processing systems, as they provide the language model with a significantly broader context than a single hypothesis.

To minimize this risk, multi-stage pipeline technologies are proposed, which include: Uncertainty Estimation using N-best hypotheses.

N-best hypotheses help LLMs avoid errors through uncertainty estimation and constrained decoding. Uncertainty estimation involves the system analyzing the probability distribution across all options in the N-best list. If the base model (ASR or OCR) outputs several alternative options with approximately equally low scores, this indicates its uncertainty. The text is sent to the LLM for correction only when the confidence of the base system (OCR or ASR) falls below a certain threshold (e.g., $\beta=0.7$). When using N-best model hypotheses, the system is provided not with a single recognition option but with several alternative options, allowing the LLM to select the most logical one without going beyond the recognized characters.

Filtering allows the system to distinguish between "reliable" segments that do not require LLM intervention and "doubtful" ones where intervention is necessary. This approach helps avoid over-correction, where the LLM might corrupt text that has already been correctly recognized. The limited decoding of N-best hypotheses serves as the "foundation" for corrections, constraining the LLM's generation space.

Special prompt rules force the LLM to select words exclusively from the N-best list. This prevents the model from using synonyms or adding unnecessary words, which is critical for preserving the accuracy of the original text. Often, the correct word is present in the N-best list but not in the top position. Using its linguistic knowledge, the LLM is capable of identifying this word as the more logical choice in the given context.

The LLM treats the N-best list as a set of building blocks. The model analyzes the differences between the options (for example, phonetically similar words) and selects the option that best fits the sentence structure and overall meaning. Since the LLM sees all the alternatives suggested by the recognizer, it can find the "middle ground" between grammatical correctness and acoustic/visual fidelity to the source text.

The use of N-best lists transforms the correction process into a form of model fusion. Research shows that combining multiple hypotheses achieves higher accuracy than any single model, since the probability that the correct word is contained in at least one of the hypotheses (the "OR" scenario) is significantly higher.

Thus, N-best hypotheses transform the LLM's task from "guessing" to an intelligent selection among options that have been pre-filtered by the base recognition system [12].

It is also important that the model simultaneously "sees" both the page scan and the OCR output. This allows for the reconstruction of table structures – the model interprets the visual boundaries of cells. For search systems in publishing houses, the LLM can automatically add synonyms to terms, which improves subsequent work with archives by 10-15% according to the MRR (Mean Reciprocal Rank) metric. For publishing systems, it is important not only to recognize text but also to identify its

location on the page (bounding boxes). This is critical for checking errors in complex layouts [13-15].

In the publishing process, it is advisable to work with two different models in parallel (for example, GPT-5.5 and Claude 4.6) for complex passages, as they have low error correlation (each model makes mistakes in “its own” cases).

The highest efficiency in the publishing process will be achieved by a combined system that dynamically distributes tasks based on complexity: simple fragments to local OCR/ASR, complex ones through a pipeline with LLM correction.

Conclusions

Modern artificial intelligence technologies offer a wide range of possibilities for digitizing textual information in printing, from optical character recognition to experimental neural interfaces. The choice of the optimal digitization method depends on the type of input data, the volume, and the complexity level of the text. The developed mathematical model facilitates this selection. For text containing significant complexities, it is recommended to use a combined method that integrates, for example, OCR, ASR, and AI technologies, thereby increasing input speed and accuracy compared to manual text entry.

Promising areas of development in the printing industry for inputting text information of varying complexity levels include multimodal systems and their integration with generative AI models. The research results are useful for selecting technology and optimizing text digitization processes in publishing houses and printing companies, as well as in the educational process when preparing text information.

References.

1. Hrozna, O.O. (2024). Tekhnolohichni innovatsii v onlain-media: rol shtuchnoho intelektu ta virtualnoi realnosti u transformatsii kontentu [Technological innovations in online media: The role of artificial intelligence and virtual reality in content transformation]. Obrii drukarstva, 1(15), 102-112. https://doi.org/10.20535/2522-1078.2024.1(15).302843.

2. MacHOUSE. (n. d.). Shtuchnyi intelekt u polihrafii: Nova era kreatyvnosti ta efektyvnosti [Artificial intelligence in printing: A new era of creativity and efficiency]. https://machouse.ua/shopblog/print-ai/.

3. Hamad, K., & Kaya, M. (2016). A detailed analysis of optical character recognition technology. IJAMEC, 4 (Special Issue), 244-249. https://dergipark.org.tr/en/download/article-file/236939.

4. Hrinkov, V., Hrinkova, G., & Hrinkov, S. (2024). Analysis of modern optical character recognition tools for character recognition and text from the image. Communication, Informatization and Cybersecurity Systems and Technologies, 6, 75-84. https://doi.org/10.58254/viti.6.2024.05.75.

5. Jones, K.S. (2001). Natural language processing: A historical review. University of Cambridge, Computer Laboratory. https://www.cl.cam.ac.uk/archive/ksj21/histdw4.pdf.

6. Evergreens. (n. d.). NLP-tekhnolohii rozpiznavannia liudskoho movlennia [NLP technologies for human speech recognition]. https://evergreens.com.ua/ua/articles/natural-language-processing.html.

7. Kulchytska, Kh., Semeniv, M., & Mazo, M. (2024). Zastosuvannia systemy rozpiznavannia audiofailiv na osnovi shtuchnoho intelektu u polihrafii [Application of AI-based audio file

recognition system in printing]. In Artificial Intelligence in Science and Education (AISE 2024). (p. 135-138). https://doi.org/10.35668/978-966-479-141-7.

8. OpenAI. (2022). Introducing Whisper. https://openai.com/research/whisper.

9. Morris, A.C., Maier, V., & Green, P. (2004). From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. Proceedings of Interspeech 2004, 2765-2768. https://doi.org/10.21437/Interspeech.2004-668.

10. Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, 6, 707-710.

11. Naderi, M., Hermann, E., Nanchen, A., Hovsepyan, S., & Magimai.-Doss, M. (2024). Towards interfacing large language models with ASR systems using confidence measures and prompting. Proceedings of Interspeech 2024, 1-5 of Sept.

12. Pu, J., Nguyen, T.-S., & Stüker, S. (2024). Multi-stage large language model correction for speech recognition (arXiv:2310.11532v2). arXiv. https://doi.org/10.48550/arXiv.2310.11532.

13. Randhawa, J. S. (2024). Claude vs GPT-4: Where it wins (and where it falls short). Dev.to. https://dev.to/jasrandhawa/claude-vs-gpt-4-where-it-wins-and-where-it-falls-short-2b0n.

14. Anonymous. (2024). Battle of the wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard. [Manuscript submitted for review]. ICLR 2025. https://openreview.net/pdf?id=77e22bf753af7c89274afed5282ade1c6881d571.

15. MindStudio Team. (2026, March 18). ChatGPT vs Claude vs Gemini: Which AI platform is best for business in 2026? MindStudio. https://www.mindstudio.ai/blog/chatgpt-vs-claude-vs-gemini-which-ai-platform-is-best-for-business-in-2026.

16. Ma, R., Qian, M., Gales, M., & Knill, K. (2024). ASR error correction using large language models. arXiv. https://doi.org/10.48550/arXiv.2409.09554.