

ANALYSIS OF AUTOMATIC SPEECH RECOGNITION TECHNOLOGIES FOR ENTERING TEXTUAL INFORMATION INTO COMPUTER-ASSISTED PUBLISHING SYSTEMS

Yurii Syniavskyi

Department of Multimedia Technologies/Institute of Printing and Media Technologies, Lviv
Polytechnic National University, Lviv, Ukraine
ORCID 0009-0006-0262-1988

Zoryana Selmenska

Department of Multimedia Technologies/Institute of Printing and Media Technologies, Lviv
Polytechnic National University, Lviv, Ukraine
ORCID 0000-0002-9514-7923

Abstract: This paper analyses modern ASR technologies for publishing systems. It proposes a methodology for assessing the quality of Ukrainian speech recognition and a model for integrating ASR into the pre-press preparation of publications

Keywords: automatic speech recognition (ASR), desktop publishing system, text information, pre-press preparation, neural networks, Transformer, Whisper, WER, CER..

Problem statement

The digital transformation of editorial and publishing processes aims to reduce routine operations, automate content entry and processing, and integrate intelligent technologies into computer-assisted publishing systems. Entering large volumes of textual information remains a labor-intensive stage of publication preparation, particularly when preparing authors' manuscripts, interviews, transcripts, minutes, educational publications, audiobooks, multimedia content, and textual versions of audio and video materials.

Conventional keyboard input provides a high degree of control over text, but its productivity is limited by the operator's typing speed. Automatic speech recognition (ASR) offers an alternative by converting an acoustic speech signal into a sequence of textual characters or words.

Unlike early voice-input systems, which relied on limited vocabularies, speaker-specific training, and statistical language models, modern ASR systems are primarily based on deep neural networks and end-to-end architectures. Advances in Transformer, Conformer, and self-supervised models have extended automatic speech recognition capabilities under variations in voice, accent, speaking rate, and acoustic interference [6,13].

These technologies are of particular importance to the Ukrainian-language publishing sector. In this context, the overall accuracy of ASR is not a sufficient criterion for determining the system's suitability for professional use. Publishing texts are characterised by the presence of specialist terminology, proper nouns, foreign words, abbreviations, numerical notations, units of measurement, bibliographic data and complex syntactic structures. Errors in recognising these elements directly increase the amount of editorial and proofreading work required.

It is therefore important to investigate modern ASR technologies not only in terms of overall speech recognition accuracy but also in terms of their technological suitability for editorial and publishing workflows.

The analysis of scientific publications indicates that current research on automatic speech recognition primarily focuses on improving neural network architectures, extending language coverage, increasing transcription accuracy, and adapting models to languages with limited training data.

Radford et al. [13], Babu et al. [2], and Pratap et al. [12] primarily address architectural solutions and overall speech recognition quality. Studies by Lipianina-Honcharenko et al. [9] and Kyslyi et al.

[8] focus on developing Ukrainian-language resources and accounting for the speech characteristics of Ukrainian.

However, the reviewed studies pay insufficient attention to assessing the technological suitability of ASR systems for professional editorial and publishing workflows, particularly with respect to recognizing publishing and printing terminology, proper names, abbreviations, numerical information, and structural elements of text.

Traditional Word Error Rate (WER) and Character Error Rate (CER) metrics quantify overall transcription accuracy but do not account for the differing technological significance of errors in subsequent editing, proofreading, and technical editing of textual information.

For example, incorrectly recognizing a technical term, numerical parameter, or software name may have much more serious consequences for the quality of a publication than an error in a commonly used word.

Thus, a comprehensive ASR evaluation methodology remains to be developed that considers not only the overall accuracy of Ukrainian speech recognition but also whether the resulting textual information meets the requirements for subsequent processing in computer-assisted publishing systems.

Addressing this problem requires a specialized test corpus of Ukrainian-language publishing and printing content, a system of evaluation metrics, and a technological model for integrating ASR into the prepress preparation of print and electronic publications.

The aim of this study is to analyze modern technologies for automatic recognition of Ukrainian speech and to develop a methodological approach to assessing their technological suitability for entering and initially processing textual information in computer-assisted publishing systems, taking into account the accurate reproduction of technical terminology, proper names, abbreviations, numerical data, and structural elements of text.

Analysis of the current state of automatic speech recognition technologies

The development of automatic speech recognition technologies during 2022–2026 has been characterized by the active adoption of large multilingual neural network models, self-supervised learning, weakly supervised learning, and the adaptation of pretrained models to particular languages and subject domains. These approaches enable ASR to be used not only in conventional transcription systems but also in specialized information-processing workflows, including the preparation of textual information for print and electronic publications.

A landmark study in this area is that of Radford et al. [13], which introduced Whisper, a model based on a Transformer encoder–decoder architecture. It was trained on 680,000 hours of multilingual and multitask audio data. The approach demonstrated the possibility of automatic speech recognition under different acoustic conditions without mandatory additional training on each target dataset. [11]

Whisper was subsequently developed into the large-v3 model, trained on approximately 1 million hours of weakly labeled and 4 million hours of pseudolabeled audio data. According to the model documentation, error rates decreased by 10–20% relative to Whisper large-v2 across a broad set of languages. However, these results cannot be directly extrapolated to specialized Ukrainian-language publishing and printing speech without separate experimental confirmation (OpenAI, n.d.). [11]

Of particular interest to publishing is Whisper's ability to process multilingual audio in which Ukrainian speech is combined with English software names, international abbreviations, technical terms, and foreign proper names.

Nevertheless, using a large general-purpose model does not guarantee correct reproduction of highly specialized terminology. Furthermore, the authors of Whisper note that the model may generate text fragments absent from the source audio signal, making additional verification of transcription results necessary. [11]

Automatic speech recognition is the process of converting a digital representation of a speech signal into a text sequence. The technological development of ASR has involved a transition from template-matching methods and hidden Markov models to end-to-end neural network systems (Graves et al., 2006; Radford et al., 2023).

In classical ASR systems, acoustic, phonetic and linguistic information was processed by separate components. Hidden Markov Models (HMMs) were widely used to model the temporal structure of speech, whilst n-gram language models were used to estimate the probability of word sequences.

The development of deep learning initially led to hybrid DNN-HMM systems, and subsequently to end-to-end models, in which the mapping of the acoustic signal to a text sequence is performed within a single neural network architecture [5,13]

Key contemporary approaches include Connectionist Temporal Classification (CTC), attention-based encoder-decoder architectures and the Recurrent Neural Network Transducer (RNN-T). Further developments have centred on the Transformer and Conformer, which enable more effective modelling of long-term dependencies in the speech signal [5,6].

Self-supervised learning has emerged as a distinct area of research, enabling the use of large corpora of audio data without the need for complete manual transcription. Large pre-trained models have transformed ASR from a highly specialised technology into a versatile tool for multilingual speech processing [3,2,12].

Modern multilingual ASR models are neural network-based automatic speech recognition systems, trained on multilingual audio corpora and designed to convert spoken language in various languages into text. Their operation is based on the use of deep learning, in particular Transformer, Conformer and encoder-decoder architectures, which enable the modelling of acoustic and contextual dependencies in the speech signal [2,6,13].

The application of such models in computer publishing systems paves the way for the automation of Ukrainian-language and multilingual content input, a reduction in the labour intensity of initial text processing, and the improvement of pre-press preparation processes.

A prime example of this modern approach is Whisper — a multilingual end-to-end model based on a Transformer encoder-decoder. Unlike traditional systems, which required separate training of acoustic and language models, Whisper performs transcription within a single neural network architecture [13].

An important property of modern models is their ability to operate in zero-shot or weakly supervised settings, which is particularly relevant for languages with smaller specialized training corpora [2,12,13].

Models specifically adapted to Ukrainian speech are also promising for Ukrainian. Their advantage may lie in more accurate recognition of the phonetic, morphological, and lexical features of the language.

Accordingly, the current state of ASR is better characterized not by contrasting “parametric” and “nonparametric” methods but by the following evolutionary sequence:

DTW / template matching → GMM–HMM → DNN–HMM → CTC / RNN-T / Seq2Seq →
Transformer / Conformer → self-supervised and foundation ASR models.

The use of ASR in publishing entails specific requirements that distinguish it from everyday voice input.

ASR should produce not merely lexically intelligible text but textual material suitable for subsequent technical editing.

The main categories of potential errors are:

Lexical errors — substitution, deletion, or insertion of words.

Character errors — incorrect reproduction of individual letters and symbols.

Terminological errors — incorrect recognition of specialized publishing and printing terminology.

Onomastic errors — distortion of surnames, geographical names, and organization names.

Punctuation errors — omitted or incorrectly positioned punctuation marks.

Structural errors — incorrect identification of sentence, paragraph, heading, and list boundaries.

Normalization errors — incorrect representation of dates, numbers, percentages, units of measurement, contractions, and abbreviations.

Professional vocabulary presents a particular challenge in printing: trapping, screening, color separation, imposition, CtP, ICC profile, screen ruling, prepress, page imposition, overprint, and related terms.

Therefore, evaluating ASR solely by the overall number of correctly recognized words does not fully characterize a system's technological suitability. The need for additional ASR evaluation metrics beyond conventional WER is also discussed by Morris et al. [10].

Experimental research methodology

To compare modern ASR systems objectively, it is proposed to compile a test corpus of Ukrainian professional speech whose content reflects actual editorial and publishing tasks.

The test text should comprise several functional segments:

- general scientific text;
- publishing and printing text;
- a segment containing numerous specialized terms;
- a segment containing proper names and abbreviations;
- a segment containing numerical values, percentages, and units of measurement;
- text with complex punctuation and syntactic structure.

To compile the speech corpus and assess its representativeness, approaches to constructing multilingual test sets should be considered, particularly FLEURS, which is designed to evaluate speech recognition models across languages [4].

The experiment should be conducted under identical acoustic conditions, using the same test recording for all systems under investigation. Several additional acoustic scenarios may be prepared: a studio recording, a typical office environment, and a recording with controlled background noise.

Word Error Rate (WER), calculated using the following formula [10], is proposed as the primary accuracy metric:

$$WER = \frac{\{S + D + I\}}{\{N\}} \times 100\%,$$

where:

S is the number of substituted words;

D is the number of deleted words;

I is the number of inserted words;

N is the number of words in the reference transcription.

For Ukrainian, Character Error Rate (CER) should additionally be determined:

$$CER = \frac{\{S_c + D_c + I_c\}}{\{N_c\}} \times 100\%,$$

where S_c , D_c , and I_c are the numbers of character substitutions, deletions, and insertions, respectively;

N_c is the number of characters in the reference transcription.

For publishing workflows, however, it is proposed to supplement WER and CER with a specialized Terminology Error Rate (TER), which measures the proportion of incorrectly recognized professional terms:

$$TER = \frac{E_t}{N_t} \times 100\%,$$

where:

E_t is the number of incorrectly recognized terms;

N_t is the total number of terms being evaluated.

Additional metrics may include punctuation accuracy, proper-name recognition accuracy, and a time-efficiency coefficient.

The TER formula as presented here is the method proposed in this article for operationalizing the proportion of incorrectly recognized terms. No external source is cited for this formula because the authors define this metric for their own experiment. The abbreviation TER is explicitly defined as Terminology Error Rate, since it may have a different meaning in other areas of text evaluation.

Integrating ASR into a computer-assisted publishing system is not particularly complex. From a technological standpoint, direct conversion

speech → text → page layout

is not optimal.

A multistage workflow is preferable:

Audio signal → ASR → initial transcription → automatic normalization → terminology checking → spelling and punctuation checking → editorial verification → structured text representation → import into a computer-assisted publishing system → page layout.

In this model, ASR should be regarded not as a replacement for an editor or typesetting operator but as an intelligent module for the initial entry of textual information.

Recognition results may be transferred not only as plain text but also in a structured form, with headings, paragraphs, quotations, lists, and other logical document elements marked up.

This creates opportunities to further automate style application in Adobe InDesign, Scribus, FrameMaker, XML/HTML-oriented publishing systems, and other publication-preparation environments. Adobe InDesign, in particular, supports XML import, tag creation, and mapping XML tags to paragraph and character styles [1].

The further development of voice input in publishing systems depends less on improving ASR in isolation than on integrating several artificial intelligence technologies.

A prospective technological pipeline can be represented as follows:

ASR → NLP/LLM → editorial review → structuring → automated formatting → computer-assisted publishing system.

In such a system, ASR converts acoustic signals into text, NLP/LLM performs contextual normalization and analysis, and the computer-assisted publishing system handles the structural and typographic presentation of content.

At the same time, using generative models after ASR requires accuracy safeguards. A language model must not arbitrarily alter the factual content of an author's material. Automatic correction of formal errors must therefore be distinguished from substantive editorial intervention.

Conclusions

Modern automatic speech recognition systems have reached a level of development that justifies considering them as a promising component of computer-assisted publishing systems. Unlike traditional systems, which were based on pattern matching, hidden Markov models and specialised dictionaries, modern ASR solutions utilise end-to-end neural network architectures, such as Transformer and Conformer, self-supervised learning and large multilingual models [2,12,13].

In the publishing industry, it is advisable to evaluate the effectiveness of ASR not only by the general metrics of WER and CER. The accuracy of reproducing professional terminology, proper nouns, numerical information, abbreviations, punctuation and the logical structure of the text are also important.

A methodology for experimental research into ASR for Ukrainian-language publishing and printing content is proposed, and a specialised metric, TER, is introduced to assess recognition errors in professional terminology.

A promising area of development is the integration of ASR with NLP technologies and large language models, followed by the transfer of structured and verified text to a desktop publishing system. This approach enables automatic speech recognition to be regarded as a component of an intelligent technological process for the preparation of printed and electronic publications.

References:

- 1) Adobe. (2023). Tag content for XML in InDesign. Adobe Help Center. Available at: <https://helpx.adobe.com/indesign/using/tagging-content-xml.html>
- 2) Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., & Auli, M. (2022). XLS-R: Self-supervised cross-lingual speech representation learning at scale. *Proceedings of Interspeech 2022*, 2278–2282. doi:10.21437/Interspeech.2022-143
- 3) Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 12449–12460. Available at: https://papers.neurips.cc/paper_files/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html
- 4) Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., & Bapna, A. (2023). FLEURS: Few-shot learning evaluation of universal representations of speech. *2022 IEEE Spoken Language Technology Workshop (SLT)*. doi:10.1109/SLT54892.2023.10023141
- 5) Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning*, 369–376. doi:10.1145/1143844.1143891
- 6) Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented Transformer for speech recognition. *Proceedings of Interspeech 2020*, 5036–5040. doi:10.21437/Interspeech.2020-3015
- 7) KSE Research Group. (n.d.). Whisper-large-v3-dido-yvanchyk-v2 [Machine learning model]. Hugging Face. Available at: <https://huggingface.co/KSE-RESEARCH-Group/whisper-large-v3-dido-yvanchyk-v2>
- 8) Kyslyi, R., Orlovskiy, A., Khomenko, P., Onyshchenko, B., & Guzii, Z. (2026). Building ASR resources for the Hutsul dialect of Ukrainian. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects* (pp. 186–195). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.vardial-1.15>
- 9) Lipianina-Honcharenko, K., Bohuta, H., Ivaniush, A., & Soia, M. (2025). A dataset of real and synthetic speech in Ukrainian. *Scientific Data*, 12, Article 745. <https://doi.org/10.1038/s41597-025-05084-8>
- 10) 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. doi:10.21437/Interspeech.2004-668
- 11) Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., & Auli, M. (2024). Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97), 1–52. Available at: <https://www.jmlr.org/beta/papers/v25/23-1318.html>
- 12) Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning*, 202, 28492–28518. Available at: <https://proceedings.mlr.press/v202/radford23a>
- 13) Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning*, 202, 28492–28518. <https://proceedings.mlr.press/v202/radford23a.html>