JKTP: Jurnal Kajian Teknologi Pendidikan
Vol. 8, No. 2, May 2025: 157-169
e-ISSN: 2615-8787
DOI: 10.17977/um038v8i22025p157
https://journal-fip.um.ac.id/index.php/jktp

THE EFFECTIVENESS OF AUTOMATIC SPEECH RECOGNITION IN THE CAPCUT APPLICATION FOR DEVELOPING INCLUSIVE LEARNING MEDIA

Netaniel Giovanni[1], Wendi Nurhayat[2]
[1]*Instructional Technology Developer*
[2]*Leadership and Management Training Center, Ministry of Finance of the Republic of Indonesia*
netanielgiovanni@gmail.com

Article History
Received: 23 April 2025, Accepted: 30 May 2025, ***Published: 30 May 2025***

Abstrak
Penyandang disabilitas tuli menghadapi tantangan dalam mengakses informasi audio sehingga dibutuhkan teknologi yang mampu menghasilkan takarir otomatis yang akurat dan mudah dipahami. Namun, studi yang secara sistematis mengevaluasi akurasi teknologi Automatic Speech Recognition (ASR) dalam konteks bahasa Indonesia masih terbatas. Penelitian ini bertujuan untuk menganalisis efektivitas fitur ASR pada aplikasi CapCut dalam menghasilkan takarir otomatis untuk media pembelajaran. Dengan pendekatan evaluatif kuantitatif, sembilan video pembelajaran dipilih dari satu e-learning untuk dianalisis melalui perbandingan antara hasil transkripsi otomatis CapCut dan transkripsi manual. Data dinormalisasi sebelum dianalisis menggunakan perangkat lunak Jiwer dan OpenAI Whisper untuk menghitung nilai Word Error Rate (WER). Hasil menunjukkan rata-rata WER sebesar 3,08% yang termasuk kategori sangat baik namun analisis konten mengungkap perlunya sedikit penyuntingan manual pada istilah teknis dan struktur kalimat untuk meningkatkan keterbacaan. Dengan demikian, ASR CapCut berpotensi menjadi solusi strategis dalam pengembangan media pembelajaran yang efisien dan inklusif serta menjadi acuan dalam pengembangan sistem takarir otomatis.

Kata Kunci: Speech-to-Text; Automatic Speech Recognition; Word Error Rate; Pengembangan Media Pembelajaran.

Abstract
Deaf individuals face challenges in accessing audio information, necessitating technology that can generate accurate and easily understandable automatic captions. However, studies remain limited in systematically evaluating the accuracy of Automatic Speech Recognition (ASR) technology in the Indonesian language context. This study aims to analyze the effectiveness of the ASR feature in the CapCut application in generating automatic captions for educational media. To evaluate how well the ASR feature in the CapCut app creates automatic captions for educational videos, nine videos from one e-learning platform were chosen and compared by looking at the automatic transcriptions from CapCut alongside manual transcriptions. Jiwer software and OpenAI Whisper normalized the data before analyzing it to calculate the Word Error Rate (WER). The results show an average WER of 3.08%, categorized as excellent; however, content analysis revealed the need for minor manual editing of technical terms and sentence structure to enhance readability. CapCut's ASR feature could be a strategic solution for creating effective and accessible educational media, and it could also serve as a guide for building automatic captioning systems.

Keyword*: Speech-to-Text; Automatic Speech Recognition; Word Error Rate; Learning Media Development*

To cite this article:
Giovanni, N., Nurhayat, W. (2025). The Effectiveness Of Automatic Speech Recognition In The Capcut Application For Developing Inclusive Learning Media. *JKTP: Jurnal Kajian Teknologi Pendidikan*, 8(2), 157–169. doi: 10.17977/um038v8i22025p157

INTRODUCTION

Accessibility in education is a fundamental right for every individual, including persons with disabilities (Tanra et al., 2024). To ensure this right is fulfilled, it is essential to provide inclusive facilities, learning media, and technologies and emphasize the importance of reasonable accommodation within educational systems to foster a more inclusive and equitable learning environment for people with disabilities (UU No. 8 Tahun 2016; PP No. 13 Tahun 2020).

In practice, the use of audiovisual learning media has been shown to increase learning interest, encourage more active interaction, and improve students' cognitive achievements (Aprilliana & Efendi, 2022). However, for deaf learners, audio-based content in audiovisual media presents a significant challenge due to limited access to sound-based information (A. Putra et al., 2024). Therefore, captioning plays a critical role in enhancing the accessibility of learning for deaf individuals (Kuhn et al., 2023). Captions not only help deaf learners understand the content but also benefit other learners, for example, those in noisy environments or those with difficulties understanding spoken language. Previous studies have shown that incorporating captions in learning media can improve comprehension and enrich the learning experience across various user groups (Manu & Masan, 2020).

The need for captions in educational media has also been identified in evaluations conducted by the Leadership and Management Training Center Ministry of Finance of the Republic of Indonesia (PKM). These evaluations revealed that participants expect audiovisual learning media used in training programs to include captions in the Indonesian language. This enhancement is expected to improve accessibility for learners with disabilities, especially the deaf. The urgency to increase accessibility is further reinforced by the government's commitment to inclusivity, as reflected in the policy to open 25 special civil servant recruitment slots for persons with disabilities in the Ministry of Finance's 2024 intake.

At the same time, the appropriate use of educational technology can support the achievement of various learning domains such as cognitive, affective, and psychomotor. Instructional technology is defined as the ethical study and practice of facilitating learning and improving performance through the creation, use, and management of appropriate technological processes and resources (PermenPAN No. 28 Tahun 2017). The development of instructional technology involves processes of analysis, design, production, implementation, control, and evaluation of technology-based learning models. However, one of the main challenges in creating captions manually is the limited time and resources available to instructional technology developers.

As a solution to this challenge, the use of Artificial Intelligence (AI) Speech-to-Text (STT) systems based on Automatic Speech Recognition (ASR) has emerged as an innovation capable of automating the transcription of spoken content (Jollyta & Oktarina, 2020; Reddy et al., 2023). Speech recognition is an interdisciplinary subfield of natural language processing (NLP) that enables machines to recognize and transcribe spoken language into written text (Rong, 2024; Yu & Deng, 2015). ASR has seen rapid development and is widely applied across sectors (Ferdiansyah & Aditya, 2024; Firmansyah & Bachtiar, 2021; Salamun et al., 2022), including in the development of educational media (M. M. I. Putra et al., 2020; Syaifuddin et al., 2019).

One platform that leverages this technology is CapCut, a popular video editing application among audiovisual content creators. CapCut offers an auto-captioning feature that integrates ASR technology, allowing for automatic voice-to-text conversion. This feature not only increases efficiency in the editing process but also facilitates the production of more inclusive educational content.

However, despite its convenience, the effectiveness of ASR technology in generating accurate captions that meet the needs of inclusive education still requires thorough evaluation (Iosifova et al., 2021). In inclusive learning contexts, the accuracy level of captions becomes critical, especially for deaf learners who heavily rely on text to access audio-based information. This aligns with the principles of Universal Design for Learning (UDL), which emphasizes presenting information in multiple formats to accommodate the diverse needs of all learners, including those with sensory impairments. From a pedagogical perspective, the use of ASR technologies such as CapCut also supports the alignment between content, pedagogy, and technology as outlined in the Technological Pedagogical Content Knowledge (TPACK) framework.

Additionally, according to Mayer's Cognitive Theory of Multimedia Learning, clear textual presentation in audiovisual media can help reduce cognitive load and enhance comprehension (Mayer, 2009). Therefore, this study aims to analyze the use of ASR in the CapCut application and evaluate the accuracy level of its generated captions in developing inclusive learning media for deaf learners.

METHOD

This study employed a quantitative evaluative approach. This approach is used to measure the effectiveness, efficiency, or quality of a product or system based on numerical data and objective indicators (Creswell & Creswell, 2018). In this context, the approach was applied to assess the performance of CapCut's ASR technology using quantitative indicators and to describe the quality of the transcriptions in the context of developing inclusive learning media.

The study was conducted at the Leadership and Management Training Center (PKM) during the design phase of learning media development. In evaluation research designed to assess program effectiveness, a minimum of five samples is recommended, with the mean score serving as the primary benchmark for analysis (Alwi, 2015). The objects of this study were nine audiovisual learning videos used in the E-learning "Penguatan Integritas dan Pencegahan Korupsi", conducted in January 2025. Data were collected from the videos listed in Table 1.

Table 1. Video Source

Title	Duration	Style	Audio Source
Pencegahan	05:56	Monologue and Lecture	4
3 Lines Of Defense	05:50	Monologue and Interview	2
Latar Belakang Integritas	05:18	Monologue	2
Sejarah	04:25	Monologue and Lecture	2
Sumber Daya	03:09	Monologue	1
Overview	02:50	Monologue	1
Komitmen Pimpinan	02:20	Monologue	1
Deteksi	01:58	Monologue	1
Regulasi	01:54	Monologue	1

This study utilized the ASR feature in the CapCut version 5.7.0 application. The research design is illustrated in Figure 1.

Data collection was conducted by applying the Automatic Speech Recognition (ASR) feature in CapCut to edited audiovisual learning media. The resulting automatic transcriptions were then analyzed using the Word Error Rate (WER) method to evaluate the accuracy level of the generated captions. WER is a quantitative evaluative approach that compares the ASR-generated transcription against a manually produced reference transcription (Kuhn et al., 2024). As an industry-standard metric for evaluating speech recognition systems, WER calculates the error rate based on the number of word substitutions, deletions, and insertions relative to the total number of words in the reference transcription (Ali & Renals, 2018).

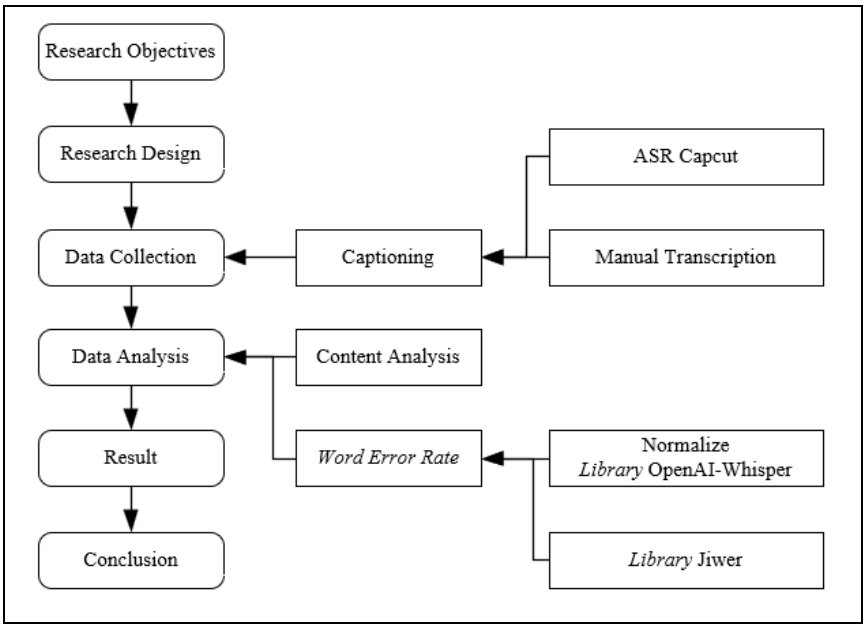

Figure 1. Research Design

The use of WER provides an objective basis for assessing the reliability of CapCut's ASR in producing accurate and readable automatic captions. The WER is calculated using the following formula:

$$WER = \frac{S + D + I}{N} \; x \; 100\% \quad (1)$$

Information:

S = number of substitutions

D = number of deletions

I = number of insertions

N = total number of words in the reference transcription.

Data analysis to obtain the WER score was performed using the OpenAI Whisper and Jiwer 3.1.0 libraries, both based on the Python programming language and implemented using Google Colab tools. OpenAI Whisper was used for text normalization, while Jiwer was employed for calculating the Word Error Rate. The use of technologies such as the Jiwer library and the OpenAI Whisper model accelerates the evaluation process while ensuring the analysis is replicable and reliable (Reddy et al., 2023).

A lower WER score indicates higher transcription accuracy (Kuhn et al., 2024), suggesting that the ASR can be effectively used in the development of audiovisual learning media. Based on established WER thresholds, transcription results with a WER between 6% and 10.99% are categorized as "Good and Usable," indicating high accuracy with only minor errors requiring minimal correction. WER scores between 11% and 20.99% are considered "Good and Acceptable," meaning the transcription contains some errors but remains understandable and requires only light revision. Scores between 21% and 30.99% fall under "Fair, Needs Improvement," as the number of errors necessitates more substantial manual corrections. Lastly, transcriptions with WER scores of ≥ 31% are classified as "Poor," indicating that the errors significantly impede comprehension and the transcription is not suitable for use without major

revisions (Microsoft Learn, 2025). The criteria used in this study to evaluate the effectiveness of the ASR feature are summarized in Table 2.

Furthermore, content analysis was conducted to examine in greater depth the differences in information, content, and errors found in the transcription results (Rukminingsih et al., 2020). This analysis provides insights into the challenges, strengths, and critical considerations for generating accurate Indonesian-language captions using CapCut's ASR feature in support of inclusive learning media development for persons with disabilities. The findings from this analysis help determine the extent to which CapCut's ASR technology can be relied upon in supporting the automatic captioning process for inclusive audiovisual learning media, particularly for individuals who are deaf or hard of hearing.

Table 2. Levels of ASR Effectiveness

WER Score (%)	Category	Description
0 – 5.99	Very Good	Very high accuracy; the transcription is nearly flawless.
6 – 10.99	Good and Usable	High accuracy with only minor errors; usable without significant corrections.
11 – 20.99	Good and Acceptable	Some errors are present, but the content is understandable with minor revisions.
21 – 30.99	Fair and Needs Improvement	A considerable number of errors requiring substantial manual corrections.
≥ 31	Poor	Too many errors; transcription is difficult to comprehend and not usable without major revisions.

In this audiovisual learning media development process, the quality of both audiovisual content and captioning was validated by three subject-matter experts, eight instructional technology developers (as media experts), and one deaf participant.

RESULT

The auto-caption feature in CapCut was utilized by selecting Indonesian as the spoken language and initiating the automatic transcription process via the "Generate" button. This streamlined workflow enabled efficient generation of captions for audiovisual learning media, as shown in Figure 2.

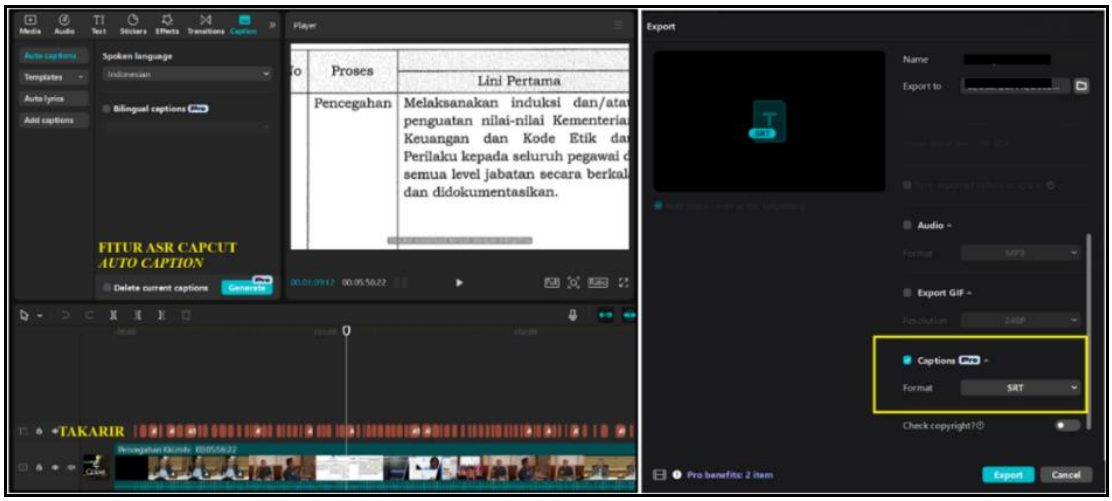

Figure 2. ASR CapCut Auto Caption Feature

To assess the accuracy of the ASR-generated captions, manual transcriptions were prepared by thoroughly reviewing each video's spoken content. These served as reference texts in evaluating the ASR outputs using the Word Error Rate (WER) metric. Any differences, such as omissions, substitutions, insertions, or incorrect words, were identified and corrected. This content

analysis also highlighted patterns of common transcription errors relevant to inclusive education, particularly for learners with hearing impairments. Table 3 summarizes several types of errors detected during this process.

Table 3. Error Detection in ASR CapCut Transcription Output

ASR CapCut Transcription Output	Actual Spoken Sentence
di Kementerian keuangan tuh punya integritas fembok	di Kementerian Keuangan tuh punya *Integrity Framework*
yang merupakan adopsi dari integrited framework OCD	yang merupakan adopsi dari *Integrity Framework* OECD
CTO meminta pendapat kepada ichen	CTO meminta pendapat kepada Itjen

A visual comparison in Figure 3 shows the contrast between ASR-generated text and manually transcribed captions. While CapCut's ASR captured much of the original speech, it exhibited issues like incorrect recognition of key phrases (e.g., "integrity framework" transcribed as "integritas fembok"), misinterpretation of institutional terms (e.g., "Itjen"), and contextual inaccuracies. These findings point to ASR limitations, especially when dealing with domain-specific or phonetically complex expressions.

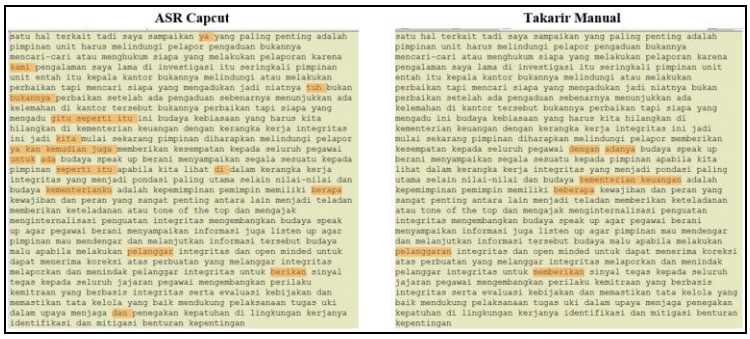

Figure 3. ASR Capcut Content Analysis

On the other hand, the manual captions had clear meaning and correct sentence structure. This shows that manual editing is still very important, especially for videos meant to help learners with hearing disabilities. These learners depend on captions to understand the content, so the text must be accurate and easy to read.

Before calculating the WER, all transcription data underwent normalization. They removed punctuation, symbols, and uppercase letters. This helped to compare the text based only on the actual words. Table 4 shows examples of how they cleaned the text before the analysis. The researchers employed Python tools to facilitate this investigation. They employed OpenAI Whisper for text refinement and utilized the Jiwer package to compute the Word Error Rate (WER). They executed this process in Google Colab, enabling them to operate online without the need for further software installation. This configuration expedited the procedure, enhanced reproducibility, and improved precision. Figure 4 illustrates the workflow of WER analysis using OpenAI Whisper and Jiwer.

The transcription data produced by CapCut's ASR and the manually created reference transcripts were normalized before the WER analysis to provide a more precise comparison. The normalization procedure involved the elimination of punctuation, symbols, and case formatting. This phase guarantees that the WER computation concentrates exclusively on word recognition accuracy, unimpeded by non-verbal textual discrepancies. By removing these apparent differences, the analysis could focus on the structure and semantics of the words and sentences generated by the ASR system. Table 4 presents examples of normalized text.

Table 4. Data Normalization for WER Calculation

Before Normalization	After Normalization
di Kementerian Keuangan tuh punya *Integrity Framework*	di kementerian keuangan tuh punya integrity framework
yang merupakan adopsi dari *Integrity Framework* OECD	yang merupakan adopsi dari integrity framework oecd
CTO meminta pendapat kepada Itjen	cto meminta pendapat kepada itjen

The calculation yielded an average Word Error Rate (WER) of 3.08% throughout the nine audiovisual learning videos. This result is categorized as "very good" according to the applied grading scale. This signifies that CapCut's ASR provides exceptionally precise transcriptions with few inaccuracies. The produced text can therefore be utilized with minimal to no manual revision. The majority of errors pertained to technical terminology, speaker enunciation, and non-verbal components, including fillers and vocal intonations. The comprehensive results of the WER analysis are presented in Table 5.

Figure 4. WER Analysis in Library OpenAI Whisper and Jiwer

The final transcription outputs, upon review, exhibited a marked enhancement in readability. Fillers like "eh," "hmm," and "gitu," which do not enhance the core meaning of the text, were eliminated to ensure clarity and focus. Furthermore, superfluous linguistic elements such as extended pauses or redundant word repetitions were eliminated to enhance the conciseness and clarity of the sentences.

Table 5. WER Score of ASR CapCut

Title	Duration	*WER Score*	Category
Pencegahan	05:56	1.33%	Very Good
3 Lines Of Defense	05:50	0.00%	Very Good
Latar Belakang Integritas	05:18	6.40%	Good and Usable
Sejarah	04:25	5.90%	Very Good
Sumber Daya	03:09	2.63%	Very Good
Overview	02:50	1.62%	Very Good
Komitmen Pimpinan	02:20	9.83%	Good and Usable
Deteksi	01:58	0.00%	Very Good
Regulasi	01:54	0.00%	Very Good
Average WER Score		3.08%	Very Good

Consequently, the captions not only communicated information with greater precision but also facilitated a more seamless reading experience for learners, especially for those with hearing impairments who depend on captions as their principal method of engaging with audiovisual

educational material. Table 6 displays instances of the final captions subsequent to the elimination of filler and non-verbal phrases.

Table 6. Final results with fillers and expressions removed for improved readability

Filler and expression	Revision
di kementerian keuangan tuh punya integrity framework gitu	di kementerian keuangan punya integrity framework
karena menurut kami satu satu apa kerangka kerja integritas	karena menurut kami suatu kerangka kerja integritas
butuh satu apa namanya piloting dulu ya kita coba dulu	butuh satu piloting dulu, kita coba dulu

The finalized media was subsequently published on the Ministry of Finance's Learning Management System.

DISCUSSION

The findings demonstrate that CapCut's ASR technology can generate transcriptions with a very high level of accuracy, as reflected in the average WER score of 3.08% across nine learning videos. This means that the transcriptions need very few corrections. This result highlights CapCut's potential to make caption production more effective and efficient. Automating the initial transcription phase with CapCut's ASR offers significant resource optimization, freeing instructional designers to focus on higher-value activities like pedagogical structuring, engagement design, and assessment strategies. This helps create inclusive learning materials, especially for students with hearing impairments who need accurate captions to understand the material. Providing high-quality captions is essential for these students to fully participate in digital learning environments, promoting equal educational opportunities.

CapCut's ability to accurately turn speech into text is in line with current research on ASR systems that use deep learning. Modern neural networks, especially end-to-end models, are very good at recognizing speech patterns and sounds (Rong, 2024). These systems learn how to convert sounds into words and can handle different speakers and sound conditions if they have been trained on them. Also, using natural language processing improves how well these systems work (Markl & Lai, 2021). NLP methods, like language modeling, help systems understand the meaning of words in context. This allows them to deal with things like different pronunciations, accents, speech speeds, and intonation, which can be hard for acoustic models to understand (Yu & Deng, 2015). The combination of deep acoustic modeling and understanding language helps artificial intelligence in CapCut to produce accurate transcriptions for lectures and conversations in educational content (Cuevas-Alonso & Tagarro, 2024).

Integrating high-accuracy ASR like CapCut's supports the main ideas of Universal Design for Learning. UDL says we should offer different ways for students to learn (Meyer et al., 2014). Accurate automated captions give another way to access spoken content, which helps students with hearing problems. But, it also helps others. Captions help students who are learning the language, those in loud places, students with trouble understanding speech, and those who learn better with both visual and auditory aids (Luchs et al., 2015). ASR can almost instantly create captions, making it easier to provide this access for everyone (Abdullah & Arief Muhsin, 2025). This promotes inclusion and helps create a fair learning environment where everyone's needs are met. This widespread access is key to good, student-centered teaching (Rizki et al., 2024).

Even though the overall accuracy is good, there are still some common mistakes. The software often gets confused with specific academic words, abbreviations used by institutions, and terms that are specific to a certain field. This problem is known in ASR research and is because the training data doesn't always include all the words and phrases used in specific fields (Hilmes et al.,

2024). When a word or phrase is not common in the data used to train the software, it is more likely to make mistakes. This means that while ASR is good at understanding general language, it needs to be checked carefully when used for specialized educational content (Hirayama et al., 2015). This is important to make sure the correct words are used, which is essential for maintaining academic integrity. If the ASR output is not checked, it could spread incorrect information specific to the subject.

Therefore, it's still important to check and edit captions to ensure they are clear, correct, and good for teaching. This means fixing mistakes in special words and making the captions easier to read. Removing extra words and sounds makes the text simpler, so students can focus on the main ideas. This editing style aligns with the Cognitive Theory of Multimedia Learning (CTML). CTML suggests that people understand information in different ways, but each way has limited capacity (Mayer, 2009). If captions are messy or wrong, it's harder to understand the visuals and audio together, which can confuse students (Iosifova et al., 2021). However, if captions are clear, short, and well-written, it helps students understand better because they don't have to process as much information. Also, when the words match the speech exactly, it helps students understand because they see and hear the same thing at the same time.

The variation in WER scores across videos highlighted the significant influence of contextual and input factors. Key variables included speech style and delivery context, with planned monologues yielding lower errors than interactive sessions featuring overlapping speech or informal diction (Emara & Shaker, 2024). Speaker articulation and vocal characteristics were also influential; clear enunciation and a steady pace improved accuracy, while mumbled speech or strong accents could increase errors (Goldrick et al., 2016). Most critically, the technical quality of the audio recording emerged as paramount. High-fidelity recordings with suitable microphones in controlled environments minimized background noise and reverberation, providing a cleaner signal. Conversely, recordings with significant ambient noise, reverberation, or issues from poor microphone placement substantially degraded performance (Michelsanti et al., 2021). These factors collectively emphasize that successful ASR implementation depends profoundly on input signal quality. Best practices in professional audio capture and post-production editing are thus critical prerequisites for maximizing ASR accuracy and ensuring accessibility (Keshet, 2018).

To leverage CapCut's ASR effectively and mitigate challenges, comprehensive technical guidelines and standardized procedures are essential (Waibel et al., 2023). These should encompass pre-recording best practices like script refinement to minimize complex jargon, speaker training for clarity, and careful selection of recording environments and equipment. Optimized recording protocols detailing noise minimization and level consistency are needed (Vogel & Morgan, 2009). Establishing a structured post-ASR workflow is equally vital, defining steps for efficient manual review prioritizing error-prone areas, systematic filler removal, punctuation correction, synchronization checks, and quality assurance (Eftekhari, 2024). Proactive collaboration between content experts and media developers during scripting is paramount. Subject Matter Experts can identify problematic terminology early, enabling simplification, while media developers advise on clear delivery and audio strategies (Mattoon, 2005). Minimizing fillers and complex sentences in the script itself eases subsequent processing. Furthermore, targeted training for instructional designers and captioning staff is crucial. Training should develop proficiency in efficiently combining ASR output with manual editing, skills for rapid error identification and correction, understanding accessibility standards, and mastering editing tools (Alonzo et al., 2022). Crucially, it should foster an understanding of ASR limitations, empowering staff to work strategically where human expertise adds the most value.

Using the Technological Pedagogical Content Knowledge framework helps clarify the skills needed to effectively integrate ASR (Mishra & Koehler, 2006). This means understanding ASR's abilities and limits, having strong knowledge of the subject to ensure correct terminology, and using good teaching methods to make captions helpful for learning (Sun, 2023). The real potential is in developing a complete TPACK, which is understanding how these areas connect specifically for ASR captioning. Teachers need help to build this knowledge, going beyond just technical skills to using ASR in a way that is good for teaching (Candra et al., 2020; Fitriyana et al., 2021).

Furthermore, integrating ASR capabilities directly within Learning Management Systems for real-time or semi-automated captioning of live and recorded content holds immense promise for scaling accessibility (Wald & Bain, 2008). Such integration, coupled with efficient human review, represents a critical step towards truly inclusive, adaptive digital learning ecosystems responsive to diverse learner needs (Srivastava et al., 2021). Future research should examine how well ASR works with more educational material, adjust ASR models using specific subject data, and assess how captions affect different learning results. Striving for smooth, correct, and educationally sound captioning is crucial for achieving true fairness in digital learning. This focus on accessibility is essential to education's purpose in the digital era.

CONCLUSION

CapCut’s ASR technology has proven to be highly effective and accurate in generating automatic captions for audiovisual learning materials. Using the WER method, supported by OpenAI Whisper and the Jiwer Python library, the study yielded an average WER of 3.08%, indicating excellent transcription accuracy and minimal need for post-editing. Content analysis revealed minor errors related to unclear pronunciation, technical vocabulary, and the presence of fillers or non-verbal expressions. These were corrected through minimal manual editing to improve comprehension. CapCut’s ASR can thus be considered a reliable tool for inclusive education, especially for deaf learners who depend on textual representations for content access. To enhance implementation, the development of technical standards for video production is necessary. Capacity-building for instructional designers will also be key to maximizing the benefits of ASR technology. Future efforts should explore integrating ASR tools into LMS platforms to provide real-time or semi-automatic captioning, thus strengthening the role of ASR in accessible and efficient digital learning environments.

REFERENCES

Abdullah, N., & Arief Muhsin, M. (2025). How Capcut Application Complete Video Assignment: A Study of Students Perception In Higher Education In Indonesia. Forum for University Scholars in Interdisciplinary Opportunities and Networking, 212–222. https://conference.ut.ac.id/index.php/fusion/article/view/4631

Ali, A., & Renals, S. (2018). Word Error Rate Estimation for Speech Recognition: e-WER. ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 2(2014), 20–24. https://doi.org/10.18653/v1/p18-2004

Alonzo, O., Shin, H. V., & Li, D. (2022). Beyond Subtitles: Captioning and Visualizing Non-speech Sounds to Improve Accessibility of User-Generated Videos. Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility. https://doi.org/10.1145/3517428.3544808

Alwi, I. (2015). Kriteria Empirik dalam Menentukan Ukuran Sampel Pada Pengujian Hipotesis Statistika dan Analisis Butir. Formatif: Jurnal Ilmiah Pendidikan MIPA, 2(2), 140–148. https://doi.org/10.30998/formatif.v2i2.95

Aprilliana, G., & Efendi, R. (2022). Penggunaan Aplikasi Capcut Untuk Meningkatkan Keterampilan Menulis Teks Iklan Pada Siswa Kelas VIII SMPN 4 Jampangtengah

Kabupaten Sukabumi. Triangulasi: Jurnal Pendidikan Kebahasaan, Kesastraan, Dan Pembelajaran, 2(2), 48–53. https://doi.org/10.55215/triangulasi.v2i2.6732

Candra, P., Soepriyanto, Y., & Praherdhiono, H. (2020). Pedagogical Knowledge (PK) Guru Dalam Pengembangan dan Implementasi Rencana Pembelajaran. JKTP: Jurnal Kajian Teknologi Pendidikan, 3(2), 166–177. https://doi.org/10.17977/um038v3i22020p166

Creswell, J. W., & Creswell, J. D. (2018). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. In SAGE Publication (Fifth Edit). SAGE Publication.

Cuevas-Alonso, M., & Tagarro, P. M. (2024). Redefining Language Education in the AI Era: Challenges, Opportunities and Perspectives. In C. Hervás-Gómez, M. D. Díaz Noguera, & F. Sánchez Vera (Eds.), The Education Revolution through Artificial Intelligence. Enhancing Skills, Safeguarding Rights, and Facilitating Human-Machine Collaboration (Issue Octaedro). Editorial Octaedro. https://doi.org/10.36006/09651-1

Eftekhari, H. (2024). Transcribing in the digital age: qualitative research practice utilizing intelligent speech recognition technology. European Journal of Cardiovascular Nursing, 23(5), 553–560. https://doi.org/10.1093/eurjcn/zvae013

Emara, I. F., & Shaker, N. H. (2024). The impact of non-native English speakers' phonological and prosodic features on automatic speech recognition accuracy. Speech Communication, 157, 103038. https://doi.org/10.1016/J.SPECOM.2024.103038

Ferdiansyah, D., & Aditya, C. S. K. (2024). Implementasi Automatic Speech Recognition Bacaan Al-Qur'an Menggunakan Metode Wav2Vec 2.0 dan OpenAI-Whisper. Jurnal Teknik Elektro Dan Komputer TRIAC, 11(1), 11–16. https://doi.org/10.21107/triac.v11i1.24332

Firmansyah, B. A., & Bachtiar, F. B. (2021). Automatic Speech Recognition Bahasa Indonesia menggunakan Unidirectional Gated Recurrent Unit. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 5(12), 5180–5187. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/10200

Fitriyana, H., Setyosari, P., & Ulfa, S. (2021). Analisis Kemampuan Technological Knowledge Calon Guru Sekolah Dasar. JKTP: Jurnal Kajian Teknologi Pendidikan, 4(4), 348–357. https://doi.org/10.17977/um038v4i42021p348

Goldrick, M., Keshet, J., Gustafson, E., Heller, J., & Needle, J. (2016). Automatic analysis of slips of the tongue: Insights into the cognitive architecture of speech production. Cognition, 149, 31–39. https://doi.org/10.1016/J.COGNITION.2016.01.002

Hilmes, B., Rossenbach, N., & Schlüter, and R. (2024). On the Effect of Purely Synthetic Training Data for Different Automatic Speech Recognition Architectures. https://doi.org/10.21437/SynData4GenAI.2024-10

Hirayama, N., Yoshino, K., Itoyama, K., Mori, S., & Okuno, H. G. (2015). Automatic Speech Recognition for Mixed Dialect Utterances by Mixing Dialect Language Models. IEEE/ACM Transactions on Audio Speech and Language Processing, 23(2), 373–382. https://doi.org/10.1109/TASLP.2014.2387414

Iosifova, O., Iosifov, I., Sokolov, V., Romanovskyi, O., & Sukaylo, I. (2021). Analysis of automatic speech recognition methods. CEUR Workshop Proceedings, 2923, 252–257. https://ceur-ws.org/Vol-2923/paper27.pdf

Jollyta, D., & Oktarina, D. (2020). Tinjauan Kasus Model Speech Recognition: Hidden Markov Model. JEPIN (Jurnal Edukasi Dan Penelitian Informatika), 6(2), 202–209. https://jurnal.untan.ac.id/index.php/jepin/article/view/39231

Keshet, J. (2018). Automatic speech recognition: A primer for speech-language pathology researchers. International Journal of Speech-Language Pathology, 20(6), 599–609. https://doi.org/10.1080/17549507.2018.1510033

Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2023). Measuring the Accuracy of Automatic Speech Recognition Solutions. ACM Transactions on Accessible Computing, 16(4), 1–23. https://doi.org/10.1145/3636513

Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2024). Measuring the Accuracy of Automatic Speech Recognition Solutions. ACM Transactions on Accessible Computing, 16(4), 1–23. https://doi.org/10.1145/3636513

Luchs, M. G., Swa, S., & Griffin, A. (2015). Design Thinking - New Product Development Essentials from the PDMA. Wiley.

Manu, G. A., & Masan, P. L. (2020). Aplikasi Text To Speech Untuk Meningkatkan Pembelajaran Bahasa Inggris Bagi Siswa Disabilitas. Jurnal Pendidikan Teknologi Informasi (JUKANTI), 3(2), 17–26. https://doi.org/10.37792/jukanti.v3i2.217

Markl, N., & Lai, C. (2021). Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation. Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, 34–40. https://aclanthology.org/2021.hcinlp-1.6/

Mattoon, J. S. (2005). Designing and Developing Technical Curriculum: Finding the Right Subject Matter Expert. Journal of STEM Teacher Education, 42(2), 61–76. https://ir.library.illinoisstate.edu/jste/vol42/iss2/5/

Mayer, R. E. (2009). Multimedia Learning. Cambridge University Press.

Meyer, A., Rose, D. H., & David Gordon. (2014). Universal Design for Learning Theory and Practice. In CAST Professional Publishing. CAST Professional Publishing.

Michelsanti, D., Tan, Z. H., Zhang, S. X., Xu, Y., Yu, M., Yu, D., & Jensen, J. (2021). An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation. IEEE/ACM Transactions on Audio Speech and Language Processing, 29, 1368–1396. https://doi.org/10.1109/TASLP.2021.3066303

Microsoft Learn. (2025). Test accuracy of a custom speech model - Speech service - Azure AI services | Microsoft Learn. https://learn.microsoft.com/en-us/azure/ai-services/speech-service/how-to-custom-speech-evaluate-data?pivots=ai-foundry-portal

Mishra, P., & Koehler, M. J. (2006). Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. Teachers College Record: The Voice of Scholarship in Education, 108(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Peraturan Menteri Pendayagunaan Aparatur Negara Dan Reformasi Birokrasi Republik Indonesia Nomor 28 Tahun 2017 Tentang Jabatan Fungsional Pengembang Teknologi Pembelajaran. https://peraturan.bpk.go.id/Details/132624/permen-pan-rb-no-28-tahun-2017

Peraturan Pemerintah Republik Indonesia Nomor 13 Tahun 2020 Tentang Akomodasi Yang Layak Untuk Peserta Didik Penyandang Disabilitas. https://peraturan.bpk.go.id/Details/132596/pp-no-13-tahun-2020

Putra, A., Eva Sri Gumilang, Lukmannul Haqim Lubay, Dian Budiana, & Gano Sumarno. (2024). Bentuk Komunikasi Guru dalam Proses Pembelajaran Pendidikan Jasmani pada Siswa Disabilitas Tunarungu di SLB Kota Bandung. Jurnal Mahasiswa Pendidikan Olahraga, 4(2), 419–429. https://doi.org/10.55081/jumper.v4i2.1655

Putra, M. M. I., Sompie, S. R. U. A., & Paturusi, S. (2020). Implementasi Speech Recognition pada Aplikasi Pembelajaran Bahasa Inggris untuk Anak. Jurnal Teknik Informatika, 15(4), 247–256. https://ejournal.unsrat.ac.id/index.php/informatika/article/view/30426

Reddy, V. M., Vaishnavi, T., & Kumar, K. P. (2023). Speech-to-Text and Text-to-Speech Recognition Using Deep Learning. Proceedings of the 2nd International Conference on Edge Computing and Applications, ICECAA 2023, 657–666. https://doi.org/10.1109/ICECAA58104.2023.10212222

Rizki, N., Asriwijiastuti, & Budiyanto. (2024). Pengembangan Media Audio Visual Animasi Gunung Berapi dalam Pembelajaran Sains bagi Penyandang Disabilitas Intelektual. GRAB KIDS: Journal of Special Education Need, 3(2), 73–76. https://doi.org/10.26740/gkjsen.v3i2.28299

Rong, Z. (2024). Application of Natural Language Processing in Virtual Experience AI Interaction Design. Journal of Intelligent Learning Systems and Applications, 16(04), 403–417. https://doi.org/10.4236/jilsa.2024.164020

Rukminingsih, Adnan, G., & Latief, M. A. (2020). Metode Penelitian Pendidikan. Penelitian Kuantitatif, Penelitian Kualitatif, Penelitian Tindakan Kelas. Erhaka Utama.

Salamun, S., Sukri, S., Amin, K., Elvitaria, L., & Trisnawati, L. (2022). Artificial Intelligence Automatic Speech Recognition (ASR) untuk pencarian potongan ayat Al-Quran. Jurnal Komputer Terapan, 8(1), 36–45. https://doi.org/10.35143/jkt.v8i1.5299

Srivastava, S., Varshney, A., Katyal, S., Kaur, R., & Gaur, V. (2021). A smart learning assistance tool for inclusive education. Journal of Intelligent \& Fuzzy Systems, 40(6), 11981–11994. https://doi.org/10.3233/JIFS-210075

Sun, W. (2023). The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: a mixed methods investigation. Frontiers in Psychology, 14(August). https://doi.org/10.3389/fpsyg.2023.1210187

Syaifuddin, M. C., Kharisma, A. P., & Akbar, M. A. (2019). Pengembangan Aplikasi Pembelajaran Pengucapan Bahasa Inggris Berbasis Android Menggunakan Automatic Speech Recognizer (ASR). Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 3(2), 1741–1748. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4558

Tanra, I., Hasudungan, A. O., Alfianto, A. S., Purnama, L. P., & Sutikno, Y. (2024). Peningkatan Pemberdayaan Pembelajaran Penyandang Disabilitas Netra melalui PeTra (Pena BerceriTra): Inovasi Teknologi untuk Aksesibilitas dan Kemandirian Literasi. Jurnal Pengabdian Masyarakat, 6(2), 76–81. https://doi.org/10.24853/jpmt.6.2.76-81

Undang-Undang Republik Indonesia Nomor 8 Tahun 2016 Tentang Penyandang Disabilitas. https://peraturan.bpk.go.id/Details/37251/uu-no-8-tahun-2016

Vogel, A. P., & Morgan, A. T. (2009). Factors affecting the quality of sound recording for speech and voice analysis. International Journal of Speech-Language Pathology, 11(6), 431–437. https://doi.org/10.3109/17549500902822189

Waibel, A., Behr, M., Yaman, D., Eyiokur, F. I., Nguyen, T. N., Mullov, C., Demirtas, M. A., Kantarci, A., Constantin, S., & Ekenel, H. K. (2023). Face-Dubbing++: LIP-Synchronous, Voice Preserving Translation Of Videos. ICASSPW 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing Workshops, Proceedings. https://doi.org/10.1109/ICASSPW59220.2023.10193719

Wald, M., & Bain, K. (2008). Universal access to communication and learning: the role of automatic speech recognition. Universal Access in the Information Society, 6(4), 435–447. https://doi.org/10.1007/s10209-007-0093-9

Yu, D., & Deng, L. (2015). Automatic Speech Recognition. In Springer. Springer London. https://doi.org/10.1007/978-1-4471-5779-3