JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7933

EVALUATION OF REAL-TIME SPEECH RECOGNITION ACCURACY IN INTERACTIVE VIDEO MEDIA FOR DEAF STUDENTS

Hilman Nuril Hadi[1*]; Adnan Zulkarnain[2];

Informatics[1],
Information Systems[2]
Universitas Bhinneka Nusantara, malang, Indonesia[1,2]
https://ubhinus.ac.id/[1,2]
hilman@ubhinus.ac.id[*], adnan.zulkarnain@ubhinus.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Deafness is a type of disability characterized by partial or complete hearing loss in one or both ears. Deaf students in higher education face several critical challenges: (1) dependence on oral communication they cannot directly access, (2) limited sign language interpreters in regular classrooms, (3) the absence of media that converts speech into real-time text while displaying the speaker's facial expressions. These conditions cause deaf students to struggle with following explanations, engaging in discussions, and participating actively in the learning process. However, individuals with hearing impairment tend to rely on visual learning, whereas the majority of instructional information is delivered through oral communication. This research aims to develop interactive media based on speech recognition and real-time video as a solution to improve communication in the learning process of deaf students. The novelty of this research lies in the integration of web-based speech recognition with a multi-actor interface (instructor, student, and general user) specifically designed for inclusive education in higher education settings, distinguishing it from conventional solutions. The method used is Research and Development (R&D) with the stages of needs analysis, system design, implementation, and functional testing and performance testing using Word Error Rate (WER). The overall average WER was 19.70%, with the range of WER being 14.05% (from the minimum of 13.22% to the maximum of 27.27%). The results showed that all system features performed as required, and an average WER indicated a good level of accuracy for interactive educational contexts.*

Keywords: *Communication, Deafness, Speech Recognition.*

Intisari— Disabilitas rungu atau tunarungu adalah salah satu jenis disabilitas dengan gangguan pendengaran sebagian atau keseluruhan pada salah satu atau kedua telinga. Penderita tunarungu mengalami keterlambatan bahasa yang berdampak pada kemampuan komunikasi, karena ketidakmampuan menginterpretasikan informasi dalam bentuk bunyi mengakibatkan terbatasnya kosakata dan pemahaman verbal. Kesulitan utama yang dihadapi mahasiswa tunarungu dalam pembelajaran perguruan tinggi meliputi: (1) ketergantungan pada komunikasi lisan yang tidak dapat diakses secara langsung, (2) keterbatasan interpreter bahasa isyarat di kelas reguler, serta (3) belum ada media yang mampu mengonversi ucapan pengajar menjadi teks secara real-time dan menampilkan mimik wajah pembicaranya. Kondisi ini menyebabkan mahasiswa tunarungu kesulitan mengikuti penjelasan, berdiskusi, dan berpartisipasi aktif dalam proses pembelajaran. Namun demikian, penderita tunarungu cenderung mengandalkan pembelajaran secara visual, sementara sebagian besar informasi pembelajaran disampaikan melalui komunikasi lisan. Penelitian ini bertujuan untuk mengembangkan media interaktif berbasis speech recognition dan real-time video sebagai solusi peningkatan komunikasi dalam proses pembelajaran mahasiswa tunarungu. Kebaruan penelitian ini terletak pada integrasi speech recognition berbasis web dengan antarmuka multi-aktor (pengajar, pelajar, dan pengguna umum) yang dirancang khusus untuk konteks pendidikan inklusif di perguruan tinggi, yang membedakannya dari solusi konvensional. Metode yang digunakan adalah *Research and Development* (R&D) dengan tahapan analisis

kebutuhan, perancangan sistem, implementasi, serta pengujian fungsional dan pengujian performa menggunakan *Word Error Rate* (WER). Rata-rata keseluruhan WER adalah 19.7%, dengan rentang WER sebesar 14.05% (minimum 13.22% sampai 27.27%). Hasil penelitian menunjukkan bahwa seluruh fitur sistem berjalan sesuai kebutuhan, dan nilai WER rata-rata menunjukkan tingkat akurasi yang cukup baik untuk konteks pendidikan intraktif.

***Kata Kunci**: Komunikasi, Ketulian, Pengenalan Ucapan.*

INTRODUCTION

Persons with disabilities in developing countries continue to face multidimensional poverty and limited access to education, healthcare, training, and employment opportunities due to persistent social and economic barriers [1]. Hearing impairment, or deafness, is one type of disability characterized by partial or complete hearing loss in one or both ears. An individual is classified as having a hearing impairment when they are unable to perceive sounds above 40 decibels (dB) for adults (>15 years) and above 30 dB for children (<15 years) [2]. Data on the distribution of individuals with hearing impairment indicate that the prevalence among the 5–14 and 15–24 age groups is relatively low, at approximately 0.8%, compared to older age groups. Nevertheless, hearing impairment within these age ranges may result in delayed language development. Language delays experienced by individuals with hearing impairment significantly affect their communication abilities, as they are unable to interpret auditory information, leading to limited vocabulary acquisition due to unprocessed sound stimuli [3], [4]. Despite these challenges, individuals with hearing impairment tend to prefer visual-based learning, in which instructional information is typically delivered through oral communication. This fundamental mismatch between the preferred learning modality of deaf individuals and the predominantly oral nature of instructional delivery represents a critical barrier to inclusive education that necessitates targeted technological intervention.

Government support and attention, as stipulated in Law of the Republic of Indonesia No. 8 of 2016, mandate the provision and/or facilitation of inclusive education and special education, requiring educational institutions to support persons with disabilities in acquiring basic skills, including communication skills. In addition, Regulation of the Minister of Research, Technology, and Higher Education No. 46 of 2017, Article 8, obliges higher education institutions to facilitate learning and assessment in accordance with the needs of students with special needs without compromising the quality of learning outcomes. The learning components referred to include learning materials, tools/media, learning processes, and/or assessment methods. Accordingly, higher education institutions are responsible for fostering an inclusive campus culture and enhancing the competencies of lecturers and educational staff in providing appropriate services to students with special needs. Based on data from the Ministry of Research, Technology, and Higher Education in 2018, a total of 74 higher education institutions in Indonesia had admitted students with disabilities.

In recent years, numerous studies have been conducted to improve educational accessibility for persons with disabilities, including individuals with hearing impairment. Speech recognition technology has been widely utilized to convert spoken language into text as a means of facilitating communication [5]. Meanwhile, advancements in video processing and digital animation have enabled the development of sign language avatars or animations to assist individuals with hearing impairment in understanding conversations or learning materials [6]. Referring to the progress in the application of speech-to-text (STT) technology, one prominent example can be found in the entertainment industry through the use of video subtitles. STT has been employed to generate captions in videos, allowing information to be conveyed not only through visual and auditory elements but also through textual transcripts of the audio content. Several studies, such as those conducted by [7], [8], [9] have developed learning media that utilize STT methods to support the learning process of individuals with hearing impairment. However, existing studies have notable limitations, as most systems were evaluated in controlled settings that do not reflect real classroom acoustic conditions, and were developed as standalone tools rather than integrated web-based platforms, limiting their scalability in higher education.

The collaboration between speech-to-text (STT) technology and visual-based methods is essential to maximize the transfer of information from instructors to students with hearing impairment. Through the integration of speech recognition and real-time video media, two-way communication between teachers and deaf students can be facilitated more effectively, enabling them to

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7933

understand learning materials through automatically transcribed text (real-time captioning) [10], [11]. Complementing this finding, a comparative evaluation of several commercial STT services revealed substantial differences in transcription accuracy across vendors, with performance degradation particularly evident under real-time streaming conditions [11]. In the Indonesian language context, prior research has explored ASR development using locally curated datasets such as Common Voice ID and TITML-IDN [12], [13]. However, these efforts have predominantly focused on general-purpose speech recognition and have not been specifically directed toward the development of inclusive learning media for deaf students in higher education. Moreover, the performance of Indonesian STT systems in domain-specific educational contexts, particularly with respect to input device variation and recognition stability across online and face-to-face learning modes, remains insufficiently studied.

Based on the background and gaps, this study aims to develop and evaluate Interactive Media Based on Speech Recognition and Real-Time Video as a solution to improve learning communication for deaf students. The main focus of this study is to analyze the accuracy of the STT system through the calculation of the Word Error Rate (WER) based on testing using two types of voice input devices. For comparative context, Google Speech-to-Text reports WER values ranging from 16.51% to 20.63% under standard conditions [14]. For domain-specific assistive systems in educational settings, WER values between 15% and 37% are generally considered acceptable [14]. The WER is the preferred metric for evaluating the performance of Speech Recognition system as it provides a direct, intuitive measure of the required corrective effort—insertion, deletion, and substitution [15], [16], [17]. In addition, this study also assesses the reliability of the system in the context of web-based online and face-to-face learning. It is hoped that the results of this study can provide an empirical contribution to improving the quality of ASR-based inclusive learning media and become a reference for further research in the field of assistive learning technology.

MATERIALS AND METHODS

Research Design

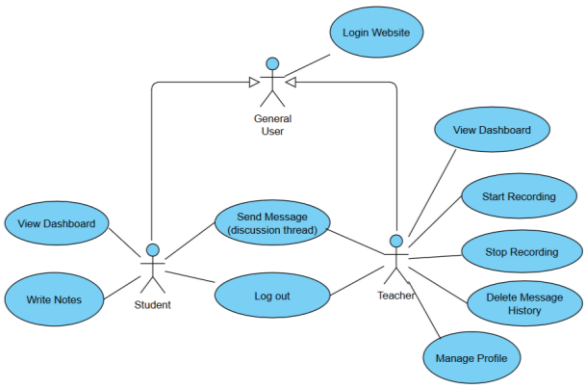
This study employs a Research and Development (R&D) approach using the Prototype Iterative development model. This model was selected because it is well suited for the development of web-based systems that require rapid feedback from end users, namely teacher and students with hearing impairment. The research process is divided into two main stages: (1) the development of interactive learning media based on speech recognition and real-time video, and (2) the evaluation of system accuracy using the Word Error Rate (WER) metric. WER method was selected as the accuracy evaluation metric in this study because it simultaneously measures three common types of errors in speech recognition, namely substitutions (incorrectly recognized words), insertions (additional words that do not exist in the reference), and deletions (missed words), thereby providing a comprehensive overview of system performance [18]. Therefore, WER is regarded as an appropriate and scientifically validated choice for evaluating the performance of the speech recognition system developed in this study.

Materials and System Environment

This stage begins with an analysis of user needs through observations of communication activities between lecturers and deaf students during online lectures. The analysis results are used to formulate the system's functional and non-functional requirements. The actors involved in the system include:

- Teachers, who interact directly with students through real-time video and speech-to-text captions,
- Deaf students, who receive automated transcriptions and can respond via text or sign language.

System requirements are then outlined in a use case (UC) diagram and functional descriptions for each actor in Figure 1. The use case diagram was deliberately selected as the primary system representation in this study, as it effectively communicates the functional requirements and interactions between actors and the system.



Source: (Research Result, 2025)
Figure 1. UC Diagram for Talk Room

The Use Case Diagram illustrates the system functionalities accessible to three main actors: the General User, Student, and Lecturer . The General User's primary interaction is limited to the Login to Application function. Upon logging in, the Student can access the Display Student Dashboard, View Recording Results (STT)—likely to review transcribed speech—Send Messages (Discussion Column), and Log out of Application. Conversely, the Lecturer has more extensive administrative and interactive roles, which include accessing the Display Teacher Dashboard, actively managing interactive sessions by being able to Conduct Conversation (Start Recording) and Stop Recording, communicating via the Send Messages (Discussion Column), maintaining system cleanliness by being able to Delete Message History, and overseeing personal data through Manage Profile Data, before ultimately being able to Log out of Application.

The system was developed using PHP and JavaScript-based web technologies, integrating the Web Speech Recognition API for speech recognition and WebRTC for real-time video transmission. The system architecture is client-server, with the client handling audio and video input, while the server processes the transcribed text and stores it in a database. The interface is responsive and accessible on both computers and mobile devices. The UI design focuses on readability of the transcribed text, with high contrast and large font sizes for deaf users [19].

Data Sources

The research data consisted of audio speech data produced by instructors during learning activities using the Indonesian language (Bahasa Indonesia). A predefined set of instructional sentences commonly used in higher education contexts was prepared as the reference (ground truth). Each participant spoke 20 sentences with an average length of 7 words. Although the dataset size is modest, this approach is consistent with prior exploratory and feasibility-oriented studies in ASR evaluation for assistive technology, where small controlled datasets have been used to establish baseline WER performance before large-scale deployment [20], [21]. The primary objective of this study is system feasibility evaluation rather than statistical generalization, and the selected sentences were carefully designed to represent diverse instructional vocabulary commonly encountered in higher education learning contexts.

Data Collection Techniques

Data collection was conducted through direct experimental testing of the proposed system during simulated learning sessions. Speech data were produced by three participants, all of whom were lecturers with ages ranging from 31 to 37 years, holding academic backgrounds in computer science. All participants were native Indonesian speakers with normal speech clarity and no reported speech or articulation disorders. The system simultaneously performed audio capture, real-time video streaming, and automatic speech-to-text transcription during each session. To evaluate the robustness of the speech recognition module, audio recordings were collected using two different input devices :

1. The laptop's internal microphone (by LENOVO ideapad flex 5)
2. Headset Plantronics (C3220)

Furthermore, data collection was carried out under two environmental conditions, consisting of **a *quiet environment*** and ***a noisy environment***. The noisy environment was conducted in a classroom learning environment during student discussion activities. The recording conditions were kept natural without any specific control over the participants' conversations. This setup was designed to assess the impact of hardware variability and acoustic conditions on transcription accuracy. Each sentence was recorded multiple times across all combinations of input devices and environmental conditions. The generated transcription outputs were stored and later compared with the predefined reference texts (ground truth). In addition to speech data acquisition, black-box testing was performed to verify the correct operation of core system functionalities, including real-time transcription display, video streaming, and text–video synchronization.

Data Analysis

Data analysis focused on evaluating the performance of the speech recognition module using the Word Error Rate (WER) metric. WER was calculated by measuring the number of substitution, deletion, and insertion errors relative to the total number of words in the reference text[16], [22]. The results were analyzed descriptively to compare transcription accuracy across different input devices and environmental conditions. Furthermore, system responsiveness was evaluated by measuring the time delay between speech input and the appearance of transcribed text to assess the real-time capability of the proposed system.

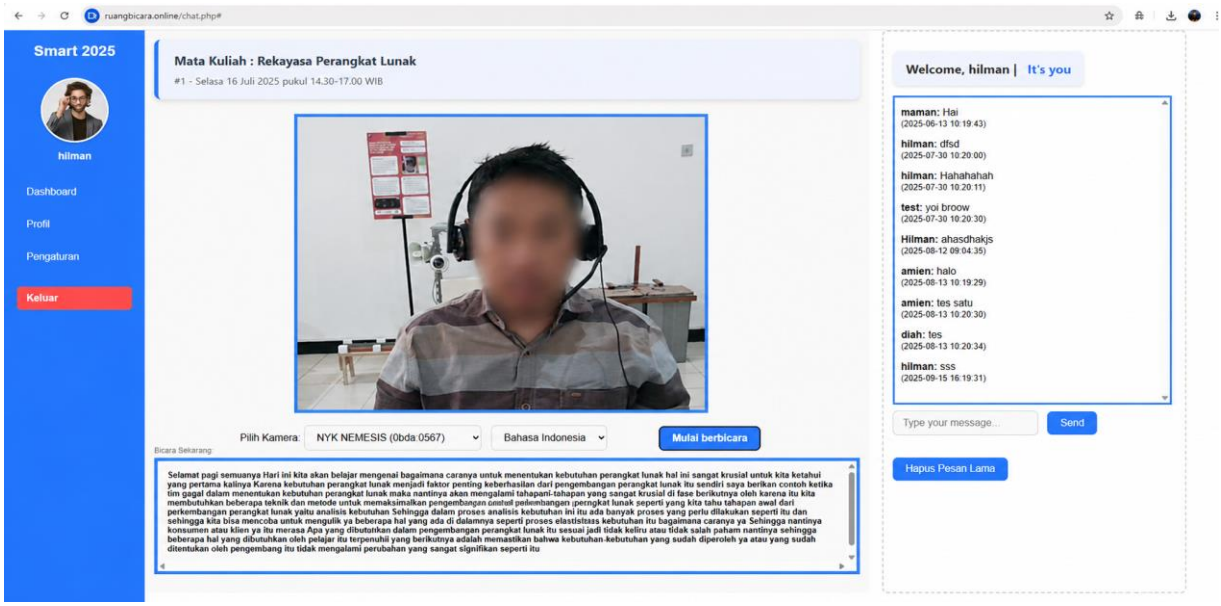
Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7933

RESULTS AND DISCUSSION

Implementation Results

The developed system is an interactive media based on speech recognition and real-time video to support communication between teachers and deaf students. Implementation utilizes modern PHP-based web technology, WebRTC for video transmission, and Google Web Speech API integration for real-time speech-to-text conversion. The results of the website implementation are represented in Figure 2.



Source: (Research Result, 2025)

Figure 2. Interface Display On Lecturer Feature

This application has two main user roles: teachers, deaf students, and general users.

1. Teachers can conduct live video communication, activate the automatic transcription feature, and correct the transcribed text.
2. Deaf students can read the text that appears directly as the teacher speaks.

We have tried to implement it in class with several deaf students, the activity is recorded in figure 3

Source: (Research Result, 2025)

Figure 3. implementation in the Learning Process

Functionality testing of the key features was systematically conducted using a black-box testing approach, focusing exclusively on external behavior and user requirements [23]. Black-box testing was deliberately selected as it evaluates system behavior from the end-user perspective, which is more aligned with the primary objective of this study, namely assessing whether the system effectively supports communication for hearing-impaired students in real learning contexts.

This approach is consistent with Ayuningtyas et al., who demonstrated that black-box testing is particularly effective for evaluating functional performance of web-based systems by focusing on input-output behavior without requiring examination of internal code structure [24]. The results showed that all key system features functioned as required in Table 1. The test results confirm that the core functionalities of the interactive system are successfully implemented and operational, validating the feasibility of the proposed design. Specifically, the successful execution of the Login and User Management feature ensures secure access for both the teacher and the student actors.

Table 1. The result of tested features

No	Tested Features	Test result
1	Login dan User Management	Running smoothly. Can log in as a teacher or student
2	Streaming Video Real-Time	running smoothly with Laptop Webcam and Webcam External
3	Speech-to-Text (Google Web Speech API)	Can capture the sound delivered
4	Automatic Captioning for Deaf Students	Can appear in the column provided
6	Text-based Chat	It's going well. Both from the teacher and student side.

Source: (Research Result, 2025)

Crucially, the Real-Time Video Streaming feature is working smoothly across different camera setups, enabling the live interaction component. The integration of the Speech-to-Text (STT) capability via the Google Web Speech API is confirmed to successfully capture audio, which directly leads to the proper function of Automatic Captioning for Deaf Students, ensuring the transcribed text appears as expected in the designated column [25], [26]. Finally, the successful operation of the Text-based Chat facility ensures that the alternative communication channel between the teacher and the student is reliable.

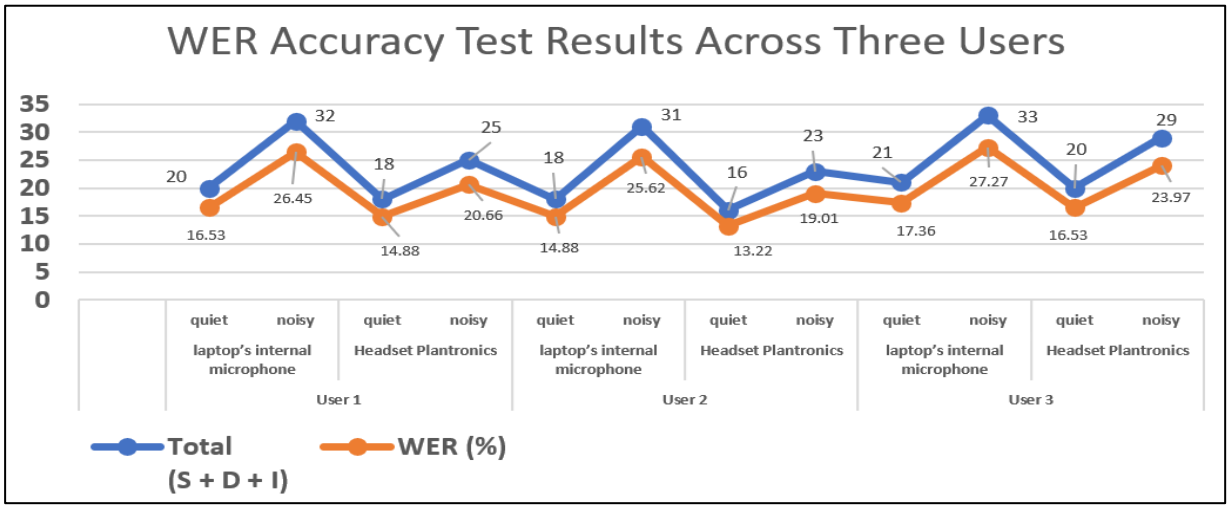
Word Error Rate (WER) Accuracy Test Results

Based on the experiments conducted by 3 users (read 121 words), are represented in Table 2 and figure 4 shows it in chart form.

Table 2. The Result of Tested WER Accuracy With 3 Users

User	Microphone Types	environmental conditions	Substitution (S)	Deletion (D)	Insertion (I)	Total (S +D + I)	WER (%)
User 1	laptop's internal microphone	*quiet*	12	5	3	20	**16.53**
		noisy	19	9	4	32	**26.45**
	Headset Plantronics	*quiet*	11	5	2	18	**14.88**
		noisy	16	5	4	25	**20.66**
User 2	laptop's internal microphone	*quiet*	10	5	3	18	**14.88**
		noisy	18	8	5	31	**25.62**
	Headset Plantronics	*quiet*	10	4	2	16	**13.22**
		noisy	16	5	2	23	**19.01**
User 3	laptop's internal microphone	*quiet*	11	7	3	21	**17.36**
		noisy	19	10	4	33	**27.27**
	Headset Plantronics	*quiet*	11	7	2	20	**16.53**
		noisy	18	5	6	29	**23.97**

Source: (Research Result, 2025)



Source: (Research Result, 2025)

Figure 4. Chart of WER Accuracy Test

Comparative Analysis

Overall, the Word Error Rate (WER) testing on the Speech-to-Text (STT) system demonstrates that accuracy is significantly influenced by both environmental conditions and the type of microphone used. A consistent pattern across all three users is that the WER invariably increased (accuracy decreased) when testing was conducted in noisy conditions compared to quiet conditions. For instance, the WER for User 1 jumped from

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7933

16.53% to 26.45% when transitioning from quiet to noisy conditions using the laptop's internal microphone. Regarding microphone performance, the Plantronics Headset demonstrated superior average performance in quiet conditions, yielding the overall lowest WER, with an average of 13.22% for User 2. Nevertheless, the laptop's internal microphone also provided competitive results in quiet environments.

The worst performance for the speech recognition system consistently occurred in noisy conditions using the laptop's internal microphone. The highest recorded WER was 27.27% for User 3, indicating that approximately one-quarter of the spoken words were incorrectly transcribed (Substitution, Deletion, or Insertion) in this most challenging environmental scenario. Although using the Plantronics Headset in noisy conditions slightly improved accuracy (the highest WER was 23.97% for User 3), the high WER in loud environments suggests that the STT system remains vulnerable to background interference. This result implies that to effectively support learning for deaf students, the application's environment should be kept as quiet as possible, or more advanced noise reduction mechanisms are necessary to maintain the required transcription accuracy.

Microphone Technology and Performance Differentiation

Regarding microphone performance, the Plantronics Headset demonstrated superior average performance in quiet conditions, yielding the overall lowest WER, with an average of 13.22% for User 2. This advantage can be attributed to several technical factors: the headset's closer proximity to the speaker's mouth, which increases the signal-to-noise ratio; its directional pickup pattern that focuses on the speaker's voice while rejecting ambient sounds; and potentially superior acoustic design optimized for voice capture. Nevertheless, the laptop's internal microphone also provided competitive results in quiet environments, with WER values ranging from 14.05% to 18.18% across the three users. This suggests that in controlled acoustic settings, built-in microphones may be adequate for basic STT applications, offering a more convenient solution without requiring additional hardware.

The performance gap between the two microphone types became more pronounced in noisy conditions. While the Plantronics Headset maintained relatively better performance with WER values between 20.66% and 23.97%, the laptop's internal microphone suffered more dramatic degradation, with WER values reaching 22.31% to 27.27%. This difference of approximately 3-4 percentage points represents a meaningful improvement in transcription quality, translating to several fewer errors per hundred words transcribed. The headset's advantage in noisy environments reinforces the importance of microphone selection for real-world deployment scenarios where perfect acoustic control cannot be guaranteed.

CONCLUSION

The implementation of the real-time interactive video media system to support learning for deaf students has been evaluated and validated through functional and accuracy testing, demonstrating its readiness for initial deployment. Functional testing confirmed that all core features operate as intended, including secure Login and User Management for both Lecturers and Students, robust Real-Time Video Streaming across different webcams, and reliable Text-based Chat. Most critically for the target users, the Speech-to-Text (STT) engine successfully captures spoken language, and the Automatic Captioning feature correctly displays the transcribed text in the provided column. This confirms that the fundamental mechanism—converting the lecturer's spoken instruction into visible text captions—is fully functional, thus establishing a viable communication bridge for deaf students.

However, the analysis of the Word Error Rate (WER) revealed a significant vulnerability in accuracy related to the environment, which must be addressed to ensure optimal learning quality. The overall average WER across all scenarios was 19.70%, but this figure masks the wide performance disparity, with the error rate soaring to a maximum of 27.27% in noisy environments. Conversely, the best performance was achieved in quiet conditions using the Headset Plantronics, resulting in a minimal WER of 13.22%. Furthermore, dedicated hardware (Headset) consistently outperformed the internal laptop microphone, yielding an average WER of 18.05% compared to 21.35%. These results underscore that while the system works, its effectiveness is highly dependent on controlling the physical acoustic environment.

For further development, it is recommended that the system be enhanced using transformer-based speech recognition models such as Wav2Vec2 or Whisper specifically trained for Indonesian, allowing the system to operate offline and be more tolerant of noise interference and variations in user accents. Furthermore, the research could be expanded to include more participants and tested

under more diverse environmental conditions to obtain more representative accuracy results.

ACKNOWLEDGMENT

The authors would like to express their deepest gratitude to the Directorate of Research and Development, Ministry of Higher Education, Science, and Technology for their invaluable support and funding through the Regular Novice Lecturer Research scheme with contract numbers: 124/C3/DT.05.00/PM/2025,119/LL7/DT.05.00/PL/2025,016/LPPM.05/UBHINUS/VI/2025. We are honored to have been part of their mission to foster innovative research that benefits society.

REFERENCES

[1] E. Lewis, S. Mitra, and J. Yap, "Do Disability Inequalities Grow with Development? Evidence from 40 Countries," *Sustainability (Switzerland)*, vol. 14, no. 9, May 2022, doi: 10.3390/su14095110.

[2] E. Juherna, D. D. Kurniawati, G. L. Sugiarti, and A. N. Falaah, "Efektifitas Penggunaan Coachlear Implant dalam Pemerolehan Bahasa Anak Tunarungu Usia 4 Tahun," *Jurnal Pelita PAUD*, vol. 6, no. 2, pp. 261–269, Jun. 2022, doi: 10.33222/pelitapaud.v6i2.1598.

[3] N. Luh Putu Sri Adnyani, N. Made Rai Wisudariani, G. Aditra Pradnyana, I. Made Ardwi Pradnyana, and N. Komang Arie Suwastini, "Multimedia English Learning Materials for Deaf or Hard of Hearing (DHH) Children," *Journal of Education Technology*, vol. 5, no. 4, pp. 571–578, Nov. 2021, doi: 10.23887/jet.v5i4.3.

[4] F. A. Nugroho and A. P. Lintangsari, "Deaf Students' Challenges in Learning English : A Literature Review," *IJDS Indonesian Journal of Disability Studies*, vol. 9, no. 02, pp. 217–224, Dec. 2022, doi: 10.21776/ub.ijds.2022.009.02.06.

[5] R. Sarkar and A. Ghosh, "Challenges faced by students with hearing impairment in higher education: A comprehensive analysis," *International Journal of Speech and Audiology*, vol. 5, no. 1, pp. 06–12, Jan. 2024, doi: 10.22271/27103846.2024.V5.I1A.43.

[6] P. A. Rodríguez-Correa, A. Valencia-Arias, O. N. Patiño-Toro, Y. Oblitas Díaz, and R. la Puente, "Benefits and development of assistive technologies for Deaf people's communication: A systematic review," *Front. Educ. (Lausanne).*, vol. Volume 8-2023, 2023, doi: 10.3389/feduc.2023.1121597.

[7] L. Pragt, P. van Hengel, D. Grob, and J. W. A. Wasmann, "Preliminary Evaluation of Automated Speech Recognition Apps for the Hearing Impaired and Deaf," *Front. Digit. Health*, vol. 4, Feb. 2022, doi: 10.3389/fdgth.2022.806076.

[8] K. K. Widiartha, K. Agustini, I. M. Tegeh, and I. W. S. Warpala, "Real Time Automated Speech Recognition Transcription and Sign Language Character Animation on Learning Media," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 13, no. 3, Dec. 2024, doi: 10.23887/janapati.v13i3.85065.

[9] L. A. Kumar, D. K. Renuka, S. L. Rose, M. C. Shunmuga priya, and I. M. Wartana, "Deep learning based assistive technology on audio visual speech recognition for hearing impaired," *International Journal of Cognitive Computing in Engineering*, vol. 3, pp. 24–30, Jun. 2022, doi: 10.1016/j.ijcce.2022.01.003.

[10] Y. Samaradivakara *et al.*, "SeEar: Tailoring Real-time AR Caption Interfaces for Deaf and Hard-of-Hearing (DHH) Students in Specialized Educational Settings," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, in CHI EA '24. New York, NY, USA: Association for Computing Machinery, 2024. doi: 10.1145/3613905.3650974.

[11] K. Kuhn, V. Kersken, B. Reuter, N. Egger, and G. Zimmermann, "Measuring the Accuracy of Automatic Speech Recognition Solutions," *ACM Trans. Access. Comput.*, vol. 16, no. 4, Jan. 2024, doi: 10.1145/3636513.

[12] P. Arisaputra and A. Zahra, "Indonesian Automatic Speech Recognition with XLSR-53," *Ingenierie des Systemes d'Information*, vol. 27, no. 6, pp. 973–982, Dec. 2022, doi: 10.18280/isi.270614.

[13] A. Adila, D. Lestari, A. Purwarianti, D. Tanaya, K. Azizah, and S. Sakti, "Enhancing Indonesian Automatic Speech Recognition: Evaluating Multilingual Models with Diverse Speech Variabilities," *Computer Science Computation and Language*, Oct. 2024, doi: 10.48550/arXiv.2410.08828.

[14] A. Ferraro, A. Galli, V. La Gatta, and M. Postiglione, "Benchmarking open source and paid services for speech to text: an analysis of quality and input variety," *Front. Big Data*, vol. 6, 2023, doi: 10.3389/fdata.2023.1210559.

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7933

[15] J. Lee and S. Watanabe, "Intermediate Loss Regularization for CTC-Based Speech Recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6224–6228. doi: 10.1109/ICASSP39728.2021.9414594.

[16] R. Yakubovskyi and Y. Morozov, "Speech Models Training Technologies Comparison Using Word Error Rate," *Advances in Cyber-Physical Systems*, vol. 8, no. 1, pp. 74–80, May 2023, doi: 10.23939/acps2023.01.074.

[17] T. Amorese, C. Greco, M. Cuciniello, R. Milo, O. Sheveleva, and N. Glackin, "Automatic speech recognition (ASR) with Whisper: Testing Performances in Different Languages," in *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, 2023. Accessed: Nov. 30, 2025. [Online]. Available: https://ceur-ws.org/Vol-3574/

[18] T. von Neumann, C. Boeddeker, K. Kinoshita, M. Delcroix, and R. Haeb-Umbach, "On Word Error Rate Definitions and Their Efficient Computation for Multi-Speaker Speech Recognition Systems," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10094784.

[19] S. Nouriska, M. C. Untoro, A. Afriansyah, M. Praseptiawan, W. Yulita, and I. F. Ashari, "USER EXPERIENCE ANSWER SYSTEM AUTOMATICALLY WITH USER CENTERED DESIGN AND USER EXPERIENCE QUESTIONNAIRE-SHORT," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 9, no. 1, pp. 81–88, Aug. 2023, doi: 10.33480/jitk.v9i1.4152.

[20] P. Roychowdhury *et al.*, "Evaluating the accuracy of speech to text applications for cochlear implant candidates during COVID-19," *Cochlear Implants Int.*, vol. 24, no. 1, pp. 1–5, Jan. 2023, doi: 10.1080/14670100.2022.2120450.

[21] L. Pragt, P. van Hengel, D. Grob, and J.-W. A. Wasmann, "Preliminary Evaluation of Automated Speech Recognition Apps for the Hearing Impaired and Deaf," *Front. Digit. Health*, vol. Volume 4-2022, 2022, doi: 10.3389/fdgth.2022.806076.

[22] A. El Hannani, R. Errattahi, F. Z. Salmam, T. Hain, and H. Ouahmane, "Evaluation of the effectiveness and efficiency of state-of-the-art features and models for automatic speech recognition error detection," *J. Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-020-00391-w.

[23] H. Kapoh, O. E. Melo, and A. A. Kimbal, "Black Box Testing in Web-based Applications: Case Study-Remedial Application at Manado State Polytechnic," *Int. J. Comput. Appl.*, vol. 174, no. 12, pp. 975–8887, Jan. 2021, doi: 10.5120/ijca2021921002.

[24] P. K. Ayuningtyas, D. Atmodjo, and P. Rachmadi, "Performance And Functional Testing With The Black Box Testing Method," *International Journal of Progressive Sciences and Technologies (IJPSAT*, vol. 39, no. 2, pp. 212–218, Jul. 2023, doi: 10.52155/ijpsat.v39.2.5471.

[25] P. Yellamma, P. R. Varun, N. C. N. L. Narayana, Y. Chowdary, P. Manikanth, and K. H. G. Sai, "Automatic and Multilingual Speech Recognition and Translation by using Google Cloud API," in *2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, 2024, pp. 566–571. doi: 10.1109/ICMCSI61536.2024.00089.

[26] D. C. Tran, D. L. Nguyen, H. S. Ha, and M. F. Hassan, "Speech Recognizing Comparisons Between Web Speech API and FPT. AI API," in *Proceedings of the 12th National Technical Seminar on Unmanned System Technology*, Springer, 2021, pp. 853–865.