American Journal of Smart Technology and Solutions (AJSTS)

ISSN: 2837-0295 (ONLINE)

VOLUME 5 ISSUE 1 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Smart Technology and Solutions (AJSTS)

Volume 5 Issue 1, Year 2026
ISSN: 2837-0295 (Online)
DOI: https://doi.org/10.54536/ajsts.v5i1.7326
https://journals.e-palli.com/home/index.php/ajsts

A Hybrid RWHISYMP Speech-to-Text Noise Suppression Model: Integration of the Whisper Base Model, RNNoise, and SympSpell Algorithms

Mariann F. Bragas[1*], Laurence D. Ganda[1], Leonila R. Juanatas MIT[1], Charisse S. Ronquillo MIT[1]

Article Information

Received: January 07, 2026

Accepted: March 25, 2026

Published: June 16, 2026

Keywords

RNNoise, RWhiSymp, Speech-to-Text, SymSpell, Whisper Base Model

ABSTRACT

Deaf and Hard-of-Hearing (DHH) individuals face difficulty in accessing spoken information without the use of an interpreter and using methods such as lip reading and writing are inadequate. While Automatic Speech Recognition (ASR) technologies offer real-time transcription. Noise interference is a prevalent issue and can lead to transcription inaccuracies. This study introduces the RWhiSymp, a hybrid speech-to-text noise suppression model that integrates three components: RNNoise for noise suppression, Whisper Base Model for ASR, and SymSpell for spelling correction. The integrated system is designed to minimize the Word Error Rate (WER) by suppressing background noise, leading to improved accuracy and correcting misspelled words. The evaluation results shows that the RWhiSymp reduced WER by 2.66% in high-noise conditions at 60-80dB and 2.17% in low-noise conditions at 10-30dB. A spectrogram of the audio using RNNoise shows its effectiveness in reducing noise while preserving speech clarity. Misspelled words are corrected using SymSpell. User evaluation was conducted with DHH participants reported high satisfaction across effectiveness, productivity, and accessibility, with overall ratings interpreted as Very Satisfied. The findings indicate that RWhiSymp offers a practical, real-time, and accessible solution that empowers DHH individuals by enhancing their ability to engage in spoken communication. This research highlights the value of hybrid ASR pipelines in assistive technologies and provides a foundation for future work in speech recognition, noise suppression, and natural language processing for accessibility.

INTRODUCTION

Communication is essential in our daily lives, yet many Deaf and Hard-of-Hearing (DHH) individuals face difficulty in accessing spoken language, turning it into a barrier rather than a bridge. While the DHH community uses communication methods such as written text and lip reading, these methods can be ineffective and inadequate in real-time conversation when interacting with non-signer individuals (National Association of the Deaf, 2021). The use of interpreters helps close those gaps; however, they are not always readily available (World Federation of the Deaf, 2020). With advancements in technology, it brought numerous transformative changes particularly in Speech-to-Text (STT) also known as Automatic Speech Recognition (ASR), the process of converting spoken words into a transcribed text (Kizito *et al.,*2024).

However, noise interference remains a significant problem in speech recognition (Radford *et al.,*2022). Consequently, it increases the word error rate (WER) and introduces spelling errors in transcriptions.

Over the years, state-of-the-art algorithms have been developed to address these challenges. To begin with, RNNoise by Valin (2018) is a noise-suppression algorithm that removes background noise from audio while minimizing distortion of the speech of interest (Valin, 2018). Furthermore, Whisper is an automatic speech recognition (ASR) system trained on 680,000 hours of multilingual, multitask supervised data collected from the web, enabling robust transcription across diverse conditions. In addition, because transcription errors are common, spelling correction is crucial. The SymSpell is a symmetric-delete spelling correction algorithm that reduces the complexity of edit-candidate generation and dictionary lookup for a given Damerau–Levenshtein distance.

LITERATURE REVIEW

Deaf and hard of hearing (DHH) individuals in the Philippines face significant communication barriers that restrict their access to spoken information and contribute to social isolation. According to the Philippine Statistics Authority (PSA), there are an estimated 1,784,690 Filipinos experiencing hearing difficulty in 2020. Improvements in Natural Language Processing (NLP) make it possible for computers to read and respond to human language. NLP is a branch of Artificial Intelligence (AI) that applies machine learning in processing and understanding human language as written or verbal speech (Quarteroni, 2018).

The communication application BridgeApp developed by Samonte *et al.,*(2019) operates on Android technology. However, it lacks the ability to handle spontaneous or open-domain interactions thus requiring future development for dynamic vocabulary and context-independent communication. Shezi *et al.,* (2020) developed the real-time communication mobile STT prototype known as DeafChat. The application enables

[1] College of Engineering and Technology Education, Holy Trinity College of General Santos City, General Santos City, Philippines
[*] Corresponding author's e-mail: mariannbragas2004@gmail.com

real-time communication between users who have hearing impairments. The study recommended the adoption of a more responsive speech recognition engine with expanded accent support.

In recent years, robust speech recognition has achieved a big leap forward with the emergence of large-scale weakly supervised learning. This enables models to learn from huge amounts of noisy and diverse data. A pioneering work by Radford *et al.,*(2022) presents Whisper Model, an automatic speech recognition (ASR) system trained on 680,000 hours of both multitask and multilingual data with the aim of demonstrating high-performance over a broad set of benchmarks without the use of data-specific fine-tune models. Whisper consists of six different models. Precisely, Whisper Base Model has 74 million parameters and is supported by the English language.

Word Error Rate(WER) is the evaluation metric used to assess the transcription quality of the output of Automatic Speech Recognition (ASR) systems (Park *et al.,*2024).

WER= (S+I+D)/N

Equation 1 WER Formula

Where S, I, D, and N represent the numbers of substitutions, insertions, deletions, and total number of words in the reference transcription, respectively.

In Whisper models, as background noise increases, the WER will also increase. This suggests that a model trained on a large and diverse dataset is still susceptible to noise. To further expand the functionality of the whisper model, Macháček *et al.,* (2023) explored the adaptation of OpenAI's Whisper model into real-time transcription, the Whisper-Streaming. A complementary study by Bevilacqua *et al.,* (2024), making offline Whisper ASR simple and convenient for real-time usage. It suggests changes for latency reduction while still delivering high transcription quality for live audio streams.

Noise interference continues to be a major problem for ASR systems (Duarte & Colcher, 2021). This ASR challenge suggests the possibility of using methods for the removal of noise interference. Valin (2017) proposes a hybrid model that combines an approach combining the traditional signal processing and deep learning. RNNoise algorithm achieves significant attenuation of the noise while maintaining the natural features of the speech. The advantage of the RNNoise is that it is designed to work in real-time, which makes it suitable for live streaming audio. It is also versatile as this can also be integrated into different hardware and software. But, the drawback is its restricted compatibility with audio formats, as it exclusively processes RAW 16-bit mono PCM files sampled at 48 kHz in machine endian.

Spelling correction algorithms, such as SymSpell, are increasingly used to refine text outputs, including those generated by speech recognition systems. SymSpell is an extremely fast spell-checking and fuzzy string matching algorithm developed by Wolf Garbe. A reliable STT system holds immense potential as an effective communication tool for the Deaf and Hard of Hearing (DHH) community. By accurately transcribing spoken language into text.

MATERIALS AND METHODS

Software Development Life Cycle

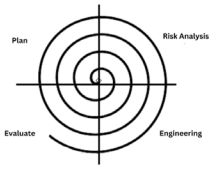

Figure 1: Spiral Model

The spiral model, created by Barry W. Boehm in 1986, is a risk-driven, iterative software development process that combines elements of both design and prototyping. It emphasizes repeated cycles (spirals) of planning, risk analysis, engineering, and evaluation, allowing teams to refine requirements, manage risks effectively, and deliver incremental improvements throughout the project lifecycle.

Plan

The researchers looked for an issue that can be mitigated by computer science. One of the marginalized communities in our country is the Deaf and Hard-of-Hearing community. The lack of assistive technologies for the DHH individuals has been an issue ever since.

Then, researchers come up with an idea, a Speech-to-Text system. Utilizing an open-source speech recognition system by Open AI called Whisper. Furthermore, a feasibility study is also conducted to evaluate the technical, economic, and operational viability of the proposed system. Integrating the Whisper Base model for speech recognition, the RNNoise algorithm for real-time noise suppression, and the SymSpell algorithm for effective spell correction is both technically feasible and practical given the existing hardware limitations and budget constraints. The integrated model is named RWhiSymp.

To successfully implement RWhiSymp, certain requirements must be considered, ensuring the system can perform efficiently in real-time environments while producing high-quality noise suppressed speech.

Risk Analysis

The risk analysis phase identifies, assesses, and mitigates risks that may affect the development process. Risk management ensures that technical and operational challenges do not hinder project success.

Engineering

For the third phase, it is composed of five activities namely: design, coding, integration, testing, and

implementation. The technical vision of the study is transformed into a working system. This phase leverages iterative development, with each iteration focusing on enhancing specific features. MATLAB will be used to test and analyze the efficiency of the RNNoise, Whisper Model, SymSpell, and RWhiSymp.

Data Flow Diagram

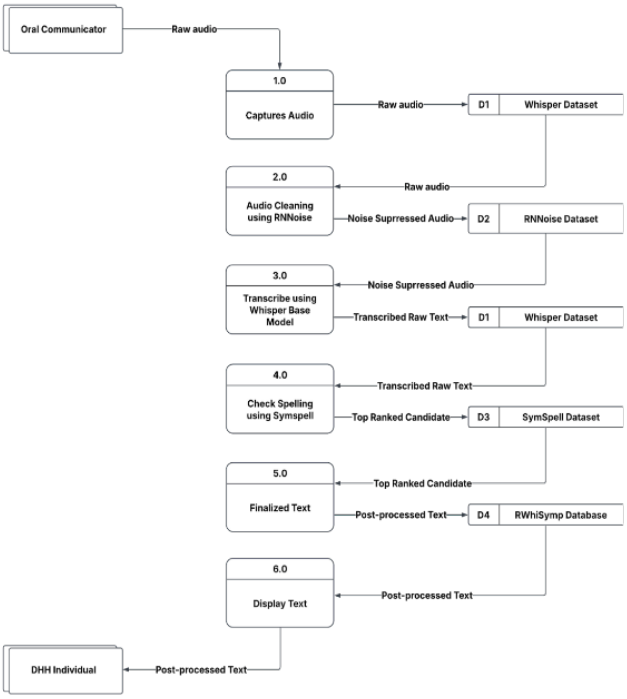

Figure 2: Level 0 Data Flow Diagram - RWhiSymp STT Noise Suppression Model

The level 0 data flow diagram describes how raw audio is transformed into a post-processed text for the DHH individual. To begin with, the Oral Communicator provides the raw audio. That raw audio will be then processed by RNNoise to remove the noise interference. The noise suppressed audio is forwarded to the Whisper Base Model for the transcription. The transcribed raw text will then be checked by the SymSpell to lookup for any OOV. The post-processed text will be stored in the RWhiSymp Database. Finally, the post-processed text is displayed, for the DHH individual to view the text.

User Interface Design

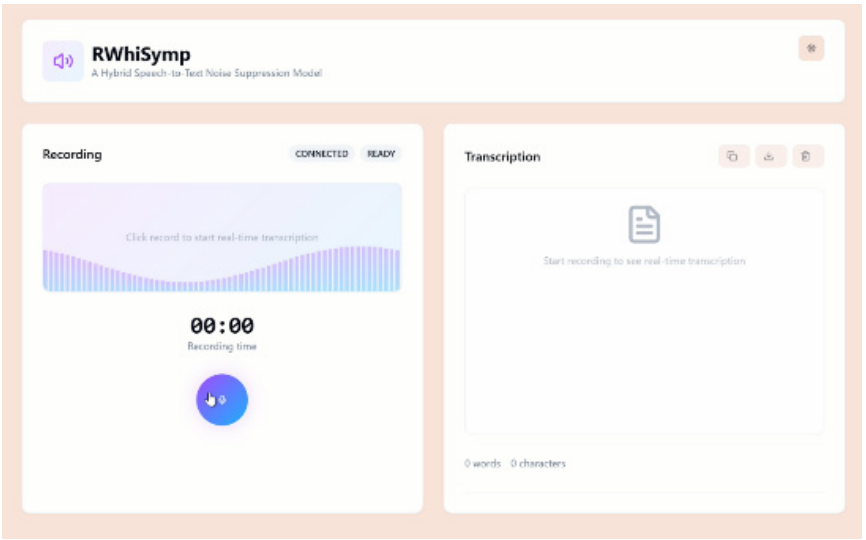

Figure 3: User Interface Design

The key features of the system

1. Microphone Button - The microphone button serves as the entry point for voice input. This feature enables users to start or stop speech recognition with a single touch.

2. Text Box - The text box displays the live transcription of the captured speech.

3. Copy Button - The copy button allows users to instantly copy the entire transcribed text to their clipboard.

4. Save Button – The save button allows users to save the transcribed text as a .txt file.

5. Delete Button - The delete button clears all content from the text box.

6. Theme Customization - This feature lets users switch between light and dark themes based on their preference or environment.

System Architecture

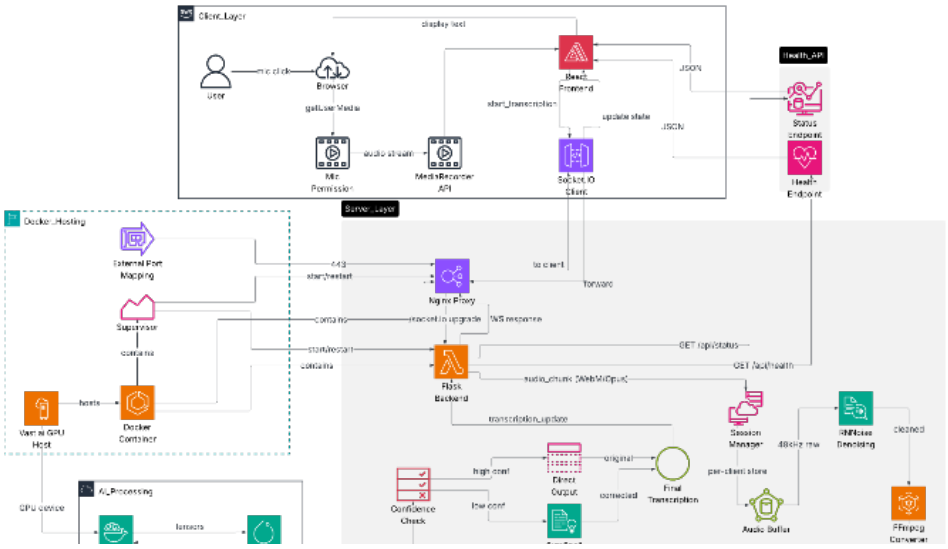

Figure 4: System Architecture

The system architecture includes a multilayered client–server design. The backend operates in a Dockerized stack (nginx + Flask SocketIO running underneath Supervisor) in a Vast.ai GPU instance. The process starts in the Client Layer. The user clicks the mic button, the browser permits capture with getUserMedia, and the MediaRecorder API splits the live stream into ~2 second WebM/Opus chunks (typically at 48 kHz). In the Edge/Proxy Layer, TLS termination occurs in the nginx, the single page React app is delivered, the WebSocket upgrade of the /socket.io takes place, and the /api requests are proxed to the Flask app. At the Server/application Layer, Flask-SocketIO opens/closes a per-client buffer at the beginning of the start-transcription and enqueues each incoming audio-chunk in a thread safe queue. In the AI/Processing Layer, the Whisper model executes with PyTorch. If the CUDA device is local at the host in Vast.ai, inference is pushed from the CPU onto the NVIDIA GPU and yields ~5–10× faster turnaround.

RESULTS AND DISCUSSION

This chapter presents the results and discussions of

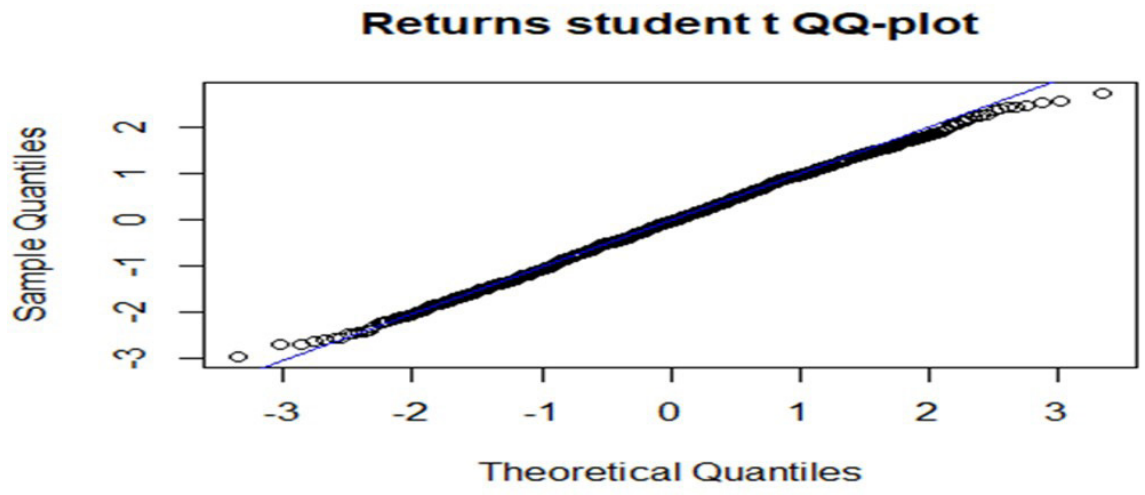

Figure 5: User Evaluation Questionnaire

the system and user evaluation. A user evaluation was conducted with a minimum of 20 participants, all of whom were Deaf or Hard-of-Hearing (DHH) individuals. The evaluation questionnaire was based on the ISO 9126-4 standard, which assesses quality through four categories: effectiveness, productivity, safety, and satisfaction.

System Evaluation

This section evaluates the systems performance based on testing the RWhiSymp and Whisper Base in real-world environments.

Table 1: WER Result

Model	S	D	I	N	WER	Description
Noise Level: 60-80dB						
Whisper Base	1328	643	337	5415	42.62	Poor
RwhiSymp Base	1333	477	354	5415	39.96	Poor
Noise Level: 10-30dB						
Whisper Base	247	32	114	3810	10.31	Moderate Low
RwhiSymp Base	240	23	47	3810	8.14	Low

In a high-noise environment (60-80dB), the RWhiSymp model reduced the WER from 42.62% to 39.96%, a 2.66 absolute drop. This shows modest enhancement by combining noise suppression and spelling correction, even under challenging environments.

In a low-noise environment (10-30dB), RWhiSymp outperforms Whisper more clearly, reducing WER from 10.31% to 8.14%, an absolute drop of 2.17%. This significant gain indicates that when the audio quality is already good and Whisper's raw transcription is reasonably close to the correct output, the SymSpell algorithm becomes highly effective. It corrects minor spelling errors, substitutions, or omissions that the model might miss, thereby refining the final transcript.

User Evaluation

The results for the user evaluation, which reflect feedback from 22 participants who are deaf and hard-of-hearing and members of the Gensan Deaf Association. The evaluation focused on assessing the quality, clarity, and accuracy of transcriptions generated by the Whisper Base model under varying noise conditions. These insights provide valuable user-centered validation of the model's performance, particularly in terms of accessibility and intelligibility for the Deaf and Hard-of-Hearing community.

Table 2: Scale Indication

Mean Range	Verbal Description
4.50 - 5.00	Extremely Satisfied
3.50 - 4.49	Very Satisfied
2.50 - 3.49	Somewhat Satisfied
1.50 - 2.49	Not so Satisfied
1.00 - 1.49	Not at All Satisfied

The table 2 presents the scale used to interpret participants' satisfaction ratings in the evaluation. The scale is based on a five-point Likert format, where mean scores fall within defined ranges and are assigned corresponding verbal descriptions. Scores from 4.50 to 5.00 indicate that participants were Extremely Satisfied, while scores from 1.00 to 1.49 reflect that they were Not at All Satisfied. This scale provides a standardized framework for summarizing and interpreting user feedback across various levels of satisfaction.

Table 3: Effectiveness Result

	Item	Mean	Verbal Description
A1	The transcription appears quickly with minimal delay.	4.09	Very Satisfied
A2	The app responds smoothly without noticeable lag or freezes.	4.00	Very Satisfied
A3	The system remains stable during continuous use (no crashes or interruptions).	3.91	Very Satisfied
A4	The transcription keeps up well with the normal pace of speech.	3.91	Very Satisfied
A5	The transcription is accurate and matches the spoken information.	4.09	Very Satisfied
Total Mean:		**4.00**	**Very Satisfied**

The table 3 shows the effectiveness of the speech-to-text system for DHH individuals. In item A1, it achieved a mean of 4.09 indicating that the users are very satisfied with how quickly transcribed text is displayed. For item A2, it achieved a mean of 4.00 indicating that the users were very satisfied with the smoothness of the system's

response and encountered only a few lags. In item A3, it achieved a mean of 3.91 indicating that the users are very satisfied with the system's stability. For item A4, it achieved a mean of 3.91 indicating that the users are very satisfied with how the system can keep pace with normal speech. Lastly, for item A5, it achieved a mean of 4.09 indicating that the users are very satisfied with the transcription accuracy of the system. Overall, the effectiveness category achieved a total mean of 4.00, interpreted as Very Satisfied. This indicates that users can complete their tasks quickly and smoothly with the use of the system.

Table 4: Productivity Result

	Item	**Mean**	**Verbal Description**
B1	The punctuation marks and line breaks help me to follow the flow of the message.	4.23	Very Satisfied
B2	The transcribed text remains accurate even with background noise.	4.18	Very Satisfied
B3	Misspellings or unclear words are corrected well enough for easy understanding.	4.00	Very Satisfied
B4	Using the system reduces the effort I spend compared to lip-reading or writing.	4.23	Very Satisfied
B5	The system helps me participate in conversations more effectively (e.g., meetings, classes, daily interactions).	4.32	Very Satisfied
Total Mean:		**4.19**	**Very Satisfied**

The table 4 shows the results for the Productivity category. In item B1, it achieved a mean of 4.23 indicating that the users are very satisfied with the punctuation and line breaks that helped them follow the flow of the message. For item B2, it achieved a mean of 4.18, indicating that users are very satisfied with how the system maintains accurate transcription despite background noise. For item B3, it achieved a mean of 4.00 indicating that the users are very satisfied with the spelling correction that keeps the text understandable. For item B4, it achieved a mean of 4.23 indicating that the users are very satisfied with the efficiency of the system compared to lip-reading or writing. Lastly, for item B5, it achieved a mean of 4.32 indicating that the users are very satisfied with the system's assistance to help them participate in conversation with the hearing individuals. Overall, the productivity category achieved a total mean of 4.19, interpreted as Very Satisfied. This indicates that the system helps boost user participation, reduces effort, and makes communication efficient.

Table 5: Safety Result

	Item	**Mean**	**Verbal Description**
C1	The app asks for my consent before recording or saving any audio or transcript.	4.23	Very Satisfied
C2	I can see in the app if the microphone is recording or listening, and I can turn it off anytime.	4.18	Very Satisfied
C3	The audio is used only for speech-to-text transcription, it will not be used for advertising or unrelated purposes.	4.14	Very Satisfied
C4	I can copy, save, and clear the transcribed text whenever I choose.	4.32	Very Satisfied
C5	I feel safe using this in public, with no extra privacy or safety worries.	4.27	Very Satisfied
Total Mean:		**4.23**	**Very Satisfied**

The table 5 shows the results for the Safety category. In item C1, it achieved a mean of 4.23 indicating that the users are very satisfied with how the system asks for consent before recording the audio and saving the transcribed text. For item C2, it achieved a mean of 4.18 indicating that the users are very satisfied with the visibility of the microphone status and they have the freedom to operate it. For item C3, it achieved a mean of 4.14 indicating that the users are very satisfied with the assurance/policy of not using the information for any unrelated use or not sharing it with any third party or advertisements. For item C4, it achieved a mean of 4.32, the highest mean for this category. Indicating that the users are very satisfied with the transcript management where they are able to copy, save, and delete. Lastly, for item C5, it achieved a mean of 4.27 indicating that the users are very satisfied with the safety of using the system in public areas. Overall, the safety category achieved a total mean of 4.23, interpreted as Very Satisfied. This indicates that the users feel comfortable, informed, in control, and trusted the system.

Table 6: Satisfaction Result

	Item	Mean	Verbal Description
D1	I'm satisfied with the accuracy of the transcribed text.	4.32	Very Satisfied
D2	I'm satisfied with the speed at which the captions appear.	4.36	Very Satisfied
D3	I feel confident using this app in everyday life.	4.23	Very Satisfied
D4	I would recommend this app to other Deaf or Hard-of-Hearing (DHH) users.	4.45	Very Satisfied
D5	Overall, I'm satisfied with the app's performance.	4.23	Very Satisfied
Total Mean:		**4.32**	**Very Satisfied**

The table 7 shows the result for the Satisfaction category. In item D1, it achieved a mean of 4.32 indicating that the users are very satisfied with the accuracy of the transcribed text. For item D2, it achieved a mean of 4.36 indicating that the users are very satisfied with the speed of displaying the transcribed text. For item D3, it achieved a mean of 4.23 indicating that the users are very satisfied that they feel confident using and adopting it in daily life. For item D4, it achieved a mean of 4.45, the highest mean for this category. Indicating a strong willingness to recommend the system to other DHH individuals. This signifies a strong advocacy to the DHH community. Lastly, for item D5, it achieved a mean of 4.23 indicating that the users are very satisfied with the system's overall performance. Overall, the satisfaction category achieved a total mean of 4.32, interpreted as Very Satisfied. This indicates that the users have strong acceptance and positive perceptions towards the system's performance and functionality.

Table 7: Overall Result

	Item	Mean	Verbal Description
E1	I'm satisfied with the accuracy of the transcribed text.	4.32	Very Satisfied
E2	I'm satisfied with the speed at which the captions appear.	4.36	Very Satisfied
E3	I feel confident using this app in everyday life.	4.23	Very Satisfied
E4	I would recommend this app to other Deaf or Hard-of-Hearing (DHH) users.	4.45	Very Satisfied
E5	Overall, I'm satisfied with the app's performance.	4.23	Very Satisfied
Total Mean:		**4.32**	**Very Satisfied**

The overall mean of 4.18 (Very Satisfied) indicates that Deaf and Hard-of-Hearing users perceive the speech-to-text system as reliably helpful and useful. By category, Effectiveness has a total mean of 4.00 interpreted as Very Satisfied. This shows that users can complete tasks smoothly and quickly, with captions appearing promptly and the app remaining stable. Second category, Productivity has a total mean of 4.19 interpreted as Very Satisfied. It indicates the system reduces effort versus alternatives (e.g., lip-reading or writing) and supports more active participation in conversations. Third category, Safety has a total mean of 4.23 interpreted as Very Satisfied. It reflects that users feel informed and in control, consent prompts, visible mic status, and transcript management (copy/save/clear) promote trust. Lastly, Satisfaction has a total mean of 4.32 interpreted as Very Satisfied. This signifies strong acceptance and willingness to continue using and recommending the system.

Taken together, these ratings portray a product that already meets user needs with solid speed, accuracy, usability, and privacy controls. To move from "Very" toward "Extremely Satisfied" (≥4.50), functions to improve: (1) shaving small delays during long sessions and ensuring zero-lag responsiveness (Effectiveness), (2) boosting robustness to background noise and tightening error correction (Productivity), and (3) making data-use assurances even more explicit while keeping recording indicators unmistakable (Safety).

Comments and Recommendations

1. Thank you very much, it's good RWHISYMP.
2. Nice [it's a] good speak and better, good job, thank you so much.
3. Good execution and satisfied with the speed [of] the caption. Good job. Thank you!
4. Thank you. It's good, satisfied.
5. Feedback about dark theme and light theme. But it's very satisfying. Good job. Thank you and God bless.
6. This is satisfied and good. You [are] able to learn to observe Deaf to use speech-to-text. It's really good for us.
7. Thank you. Good, satisfied with your RWHISYMP.
8. Good job and thank you so much. Developers can create more enjoyable and effective user experience.
9. Chat box. Performance app [is] very satisfied.
10. Thank you so much and God bless. Your work for mind.

The comments and recommendations of the respondents reveal strong early acceptability. Eight out of sixteen comments explicitly expressed gratitude and praised the system's performance. This signifies that the respondents liked the system and found it a useful tool that helps them understand what the hearing/oral communicator is

saying. The expressions of thanks also acknowledge the students' supportive efforts toward the community. Three additional comments noted that they are satisfied with the system's performance. Meanwhile, three respondents recommended adding a chat box feature where they can reply back to the hearing/oral communicators, highlighting the need for two-way communication. One respondent specifically appreciated the dark and light theme. Lastly, one called for a more enjoyable and effective user experience, pointing to opportunities for creating a more creative system for the DHH community.

CONCLUSION

The study assessed a speech-to-text system (RWhiSymp) combining Whisper Base, RNNoise, and SymSpell to improve transcription accuracy and usability for Deaf and Hard-of-Hearing (DHH) users. Results showed reduced Word Error Rate (WER) in both noisy (2.66% improvement) and low-noise conditions (from 10.31% to 8.14%), with RNNoise enhancing audio clarity and SymSpell correcting misspellings and out-of-vocabulary words effectively.

User evaluation with 22 DHH participants reported "Very Satisfied" ratings across effectiveness (4.00), productivity (4.19), safety (4.23), and overall satisfaction (4.32), with an overall mean of 4.18. Users noted improved communication and ease compared to methods like lip-reading, while also suggesting features such as a two-way chat function.

The study concludes that integrating RNNoise and SymSpell significantly enhances both accuracy and user experience, making the system a reliable assistive tool. Recommendations include adding two-way communication, improving noise handling, offering customization options, increasing privacy transparency, and continuing community-centered testing.

REFERENCES

Abidin, T. F., Misbullah, A., Ferdhiana, R., Aksana, M. Z., & Farsiah, L. (2020, October 28). Deep neural network for automatic speech recognition from Indonesian audio using several lexicon types. In *2020 International Conference on Electrical Engineering and Informatics (ICELTICs).* https://doi.org/10.1109/ICELTICs50595.2020.9315538

Andreyev, A. (2025). *Quantization for OpenAI's Whisper models: A comparative analysis* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2503.09905

Audah, H. A., Yuliawati, A., & Alfina, I. (2023). A Comparison Between SymSpell and a Combination of Damerau-Levenshtein Distance with the Trie Data Structure. *Institute of Electrical and Electronics Engineers,* 1–6. https://doi.org/10.1109/icaicta59291.2023.10390399

Awati, R., Sheldon, R., & Burke, J. (2025, June 3). What is signal-to-noise ratio and how is it measured? *Search Networking.* https://www.techtarget.com/searchnetworking/definition/signal-to-noise-ratio

Behera, S. K., & Mitali, M. N. (2020). Natural language processing for text and speech processing: A review paper. *International Journal of Advanced Research in Engineering and Technology (IJARET), 11*(11). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3878634

Bevilacqua, A., Saviano, P., Amirante, A., & Romano, S. P. (2024, May 6). *Whispy: Adapting STT Whisper models to real-time environments* [Preprint]. arXiv. https://arxiv.org/abs/2405.03484

Chaabi, Y., & Allah, F. A. (2021). Amazigh spell checker using Damerau-Levenshtein algorithm and N-gram. *Journal of King Saud University - Computer and Information Sciences.* https://www.sciencedirect.com/science/article/pii/S1319157821001828

Chen, M., Tan, X., Li, B., Liu, Y., Qin, T., Zhao, S., & Liu, T. . (2021). *AdaSpeech: Adaptive Text to Speech for Custom Voice*. arXiv (Cornell University).

Doumanidis, C. C., Anagnostou, C., Arvaniti, E.-S., & Papadopoulou, A. (2021, May 25). *RNNoise-Ex: Hybrid speech enhancement system based on RNN and spectral features* [Preprint]. arXiv. https://arxiv.org/abs/2105.11813

Duarte, J. C., & Colcher, S. (2024). Noise-Robust Automatic Speech Recognition: A Case Study for Communication Interference. *Journal on Interactive Systems, 15*(1), 670–681. https://doi.org/10.5753/jis.2024.4267

Elakkiya, A., Jaya Surya, K., Konduru Venkatesh, Aakash, S. (2022). Implementation of Speech to Text Conversion Using Hidden Markov Model. *Communication and Aerospace Technology* (pp. 359–363). Sixth International Conference on Electronics.

Huang, K., Wu, C., Hong, Q., Su, M., & Chen, Y. . (2019). Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. *CASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* https://doi.org/10.1109/icassp.2019.8682283.

Park, C., Chen, M., & Hain, T. (2024). *Automatic speech recognition system-independent word error rate estimation* (arXiv No. 2404.16743) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2404.16743

Lee, W. Y., Tan, J. T. A., & Kok, J. K. (2022). The struggle to fit in: A qualitative study on the sense of belonging and well-being of deaf people in Ipoh, Perak, Malaysia. *Psychological Studies, 67*(3), 385-400.

Macháček, D., Dabre, R., & Bojar, O. (2023). *Turning Whisper into real-time transcription system* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.14743

Mansur, Z., Omar, N., Tiun, S., & Alshari, E. M. (2024). A normalization model for repeated letters in social media hate speech text based on rules and spelling correction. *PLoS ONE, 19*(3), e0299652. https://doi.org/10.1371/journal.pone.0299652

Mesham, S., Bryant, C., Rei, M., & Yuan, Z. (2023, February 12). *An extended sequence tagging vocabulary for grammatical error correction.* arXiv.org. https://arxiv.org/abs/2302.05913

Murugan, S., Sankarasubbu, M., & Bakthavatchalam, T. A. (2020). SymSpell and LSTM based Spell-Checkers for Tamil. *Tamil Internet Conference*. https://www.researchgate.net/publication/349924975_SymSpell_and_LSTM_based_Spell-_Checkers_for_Tamil

Nitin. (2025, April 7). System Development Life Cycle (SDLC): Phases, Models & Best Practices. *eLuminous Technologies*. https://eluminoustechnologies.com/blog/system-development-life-cycle/

Nogales, A., Caracuel-Cayuela, J., & García-Tejedor, Á. J. (2024). Analyzing the Influence of Diverse Background Noises on Voice Transmission: A Deep Learning Approach to Noise Suppression. *Applied Sciences, 14*(2), 740. https://doi.org/10.3390/app14020740

Pascual, R., Apuyod, A., Bainto, K., Panit, M. S., & Llamado, J. (2024). Evaluating the Performance of a Commercial Speech-to-Text Application for Filipino Language as an Aid in Encoding Healthcare Data. *DLSU Research Congress 2024*. https://www.dlsu.edu.ph/wp-content/uploads/pdf/conferences/research-congress-proceedings/2024/HCT-12.pdf

Quarteroni, S. (2018). Natural Language Processing for Industry. *Informatik-Spektrum, 41*(2), 105–112. https://doi.org/10.1007/s00287-018-1094-1

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022c, December 6). *Robust speech recognition via Large-Scale Weak Supervision*. arXiv.org. https://arxiv.org/abs/2212.04356

Samonte, M., Gazmin, R., Soriano, J., & Valencia, M. (2019). BridgeApp: An assistive mobile communication application for the deaf and mute. In *2019 International Conference on ICT Convergence (ICTC)*. https://doi.org/10.1109/ICTC46691.2019.8939866

Seo, S., Kim, C., & JiKim, J.-H. (2022). Convolutional neural networks using log Mel-spectrogram separation for audio event classification with unknown devices. *IEEE Xplore*. https://ieeexplore.ieee.org/abstract/document/10251060

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., Zhao, S., Qin, T., Soong, F., & Liu, T. (2024). NaturalSpeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–12.

Trabelsi, A., Werey, L., Warichet, S., & Helbert, E. (2024). Is noise reduction improving open-source ASR transcription engines quality? In *Proceedings of ICAART (3)*, 1221–1228.

Valin, J. (2017, September 24). *A hybrid DSP/deep learning approach to real-time full-band speech enhancement* [Preprint]. arXiv. https://arxiv.org/abs/1709.08243

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). *Attention is all you need* [Preprint]. arXiv. https://arxiv.org/abs/1706.03762

Werff, L. & Heeren, W. (2007) Evaluating ASR output for information retrieval. Searching Spontaneous Conversational Speech. https://www.researchgate.net/publication/241880526_Evaluating_ASR_Output_for_Information_Retrieval/citations