An Automated Visual Acuity Test System Using Vietnamese Speech Recognition for Answer Selection

Pham Hoang Minh *, Nguyen Duc Thanh, and Pham Hong Duong

Institute of Materials Science, Vietnam Academy of Science and Technology, Hanoi, Vietnam
Email: minhph@ims.vast.ac.vn (P.H.M.); thanhnd@ims.vast.ac.vn (N.D.T.); duongph@ims.vast.ac.vn (P.H.D.)
*Corresponding author

***Abstract*—Visual Acuity (VA) testing traditionally requires an ophthalmologist, limiting accessibility in non-clinical settings. This paper presents an automated VA test system designed for precise vision assessment without professional supervision, which leverages the Early Treatment Diabetic Retinopathy Study (ETDRS) eye chart for higher accuracy compared to the conventional Snellen chart. The system uses a computer with an Liquid Crystal Display (LCD) monitor and incorporates automated scoring with 0.02 LogMAR precision. To facilitate remote operation, we implement speech recognition in Vietnamese via a microphone, utilizing Azure Speech API, which is enhanced with a correction function and noise classification for improved accuracy. An experiment with 80 participants (N = 80) demonstrated a speech recognition accuracy of 93.1%, with a mean response time of 4.6 s per optotype. The VA scores from our system closely matched those from standard printed ETDRS charts, with 95.6% of measurements differing by ≤0.1 LogMAR. Our automated VA test system provides a reliable, low-cost solution for vision assessment in non-clinical environments, combining high accuracy with user-friendly remote operation.**

***Keywords*—automated visual acuity test, automated Early Treatment Diabetic Retinopathy Study (ETDRS) test, speech recognition, isolated letter recognition**

I. Introduction

In recent years, the increase of vision loss, particularly myopia, has become a significant public health challenge. Independent surveys [1–3] conducted in various countries reveal that the prevalence of myopia among young people exceeds 50%. To detect and assess refractive error, individuals need to visit an eye clinic for a Visual Acuity (VA) test, where ophthalmologists use the eye chart placed 3 to 6 m away from the patient. Traditionally, the eye chart is a print sheet. However, various advanced solutions have been developed, such as specialized integrated monitor, or computer-based eye test program as Thomson software or OptoNet software. These electronic eye charts can display randomized optotypes for different test types. However, none of these systems are truly automated, as patients must still speak their answers to a clinician.

For conducting a VA test at home, several options are available, such as using a printed chart, or smartphone- based VA test apps. These apps are designed for near-distance testing, requiring users to touch the screen to submit their answers. While studies [4–7] have demonstrated the efficacy of near-distance VA testing, it cannot fully replace far-distance examinations.

For early detection of refractive errors, it is necessary to conduct the VA tests frequently at home. This requires an easy-to-use and low-cost automated VA test system, that includes an auto-scoring engine and a remote communication tool, allowing users to transmit their responses to the system from a distance. The aim of our study is to develop and evaluate such a system, incorporating speech recognition technology for answer selection.

II. Literature Review

Several prior studies proposed an automated VA test system with different input methods. The basic idea involves pressing a button to select an answer. For example, the Freiburg VA test [8] proposed a specialized response box with eight buttons for selecting the directions of Landolt C optotype. Claessens *et al*. [9] proposed a smartphone-based input method for remote web-based eye charts. While these approaches are easy to deploy, they require users to alternate their focus between the test chart and the input device, which may compromise VA measurement accuracy.

Other solutions include eye tracking technology, such as the Tobii Pro eye tracker, used by Vrabič *et al*. [10] to monitor children's eye movements during the test, and the Head-Mounted Display (HMD) approach used by Ong *et al*. [11]. However, the high cost and complexity of these devices limit their applicability. Chiu *et al*. [12] developed an automated VA system using hand tracking with a specialized sensing device for the answer selection, detecting the subject's hand gesture corresponding to the four directions of optotypes. Similarly, the study of Li and Tong [13] employed a conventional camera-based hand tracking approach. While these methods demonstrate

Manuscript received June 11, 2025; revised July 1, 2025; accepted July 17, 2025; published October 14, 2025.

promising results, they are primarily compatible with Landolt C or tumbling E chart, whereas Snellen or Sloan letters remain the clinical gold standards.

Automatic Speech Recognition (ASR) is the most natural solution for the VA testing, with several related researches:

- Ganesan and Shalini [14]: This study implemented a VA test system using the Microsoft Speech SDK within the LabView platform for English speech recognition. The system relied on a non-state-of-the-art ASR engine (in 2014), which lacked reported metrics for speech recognition accuracy or Word Error Rate (WER).
- Taufik and Hanafiah [15]: The AutoVAT system employed a custom Convolutional Neural Network (CNN) for recognizing spoken digits in English, achieving 91.4% accuracy. However, digits are non-standard optotypes for VA testing and are inherently easier for ASR systems to recognize than letters.
- Nisar *et al*. [16]: This study developed a custom ASR engine for English speech using an adaptive Mel filter bank for feature extraction and three classifiers (HMM, SVM, KNN), reporting 83.8–91.9% concordance with conventional VA scores. However, it did not provide WER or recognition accuracy metrics, limiting insights into the ASR's performance for isolated letter recognition.

Furthermore, the VA eye chart in all studies [14–16] used the Snellen format, which employs a line-by-line scoring procedure rather than assigning the score to each correctly recognized optotype. In summary, these above limitations compromise the accuracy for evaluate the quality of model.

This study introduces an automated VA testing system, that incorporates speech recognition via microphone input for answer selection. Unlike both prior studies that relied on English speech recognition, our system utilizes Vietnamese speech recognition, tailored for its application in Vietnam. Addressing limitations of prior works, our system implements three key advancements: (1) the use of Early Treatment Diabetic Retinopathy Study (ETDRS) eye chart with LogMAR-based letter-by-letter scoring; (2) a state-of-the-art Azure Speech API for speech recognition, enhanced with our novel correction function and noise classification function for improved accuracy; and (3) comprehensive performance evaluation through multivariate metrics, including VA scores, WER, and average Time per each Optotype (TpO).

III. Materials and Methods

A. The Eye Chart and the Auto-Scoring Procedure

Our system is a computer-based software solution that runs on a conventional PC equipped with a Liquid Crystal Display (LCD) monitor (see Fig. 1). We implemented the program in Python, which offers diverse libraries for both audio and image processing. The automated test is based on the ETDRS chart, which is widely regarded as the most advanced VA testing method. Numerous studies [17–19] indicate that it can provide a more precise VA score (measured in LogMAR units) compared to the traditional Snellen test. Our digital ETDRS chart randomly displays 5 optotypes per line for each size with sizes decreasing progressively. The optotypes consists of 10 Sloan letters: C, D, H, K, N, O, R, S, V, Z.

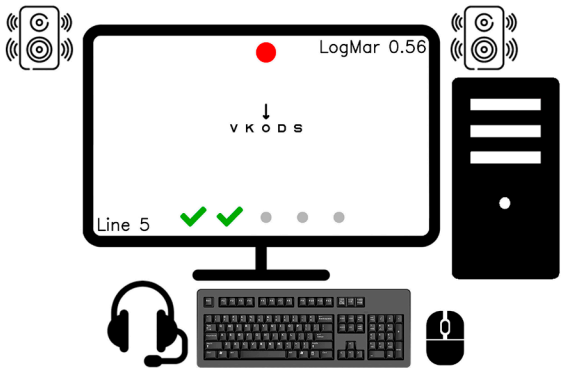

Fig. 1. System hardware components and eye chart interface.

There are 11 different sizes, ranging from largest (LogMAR = 1.0) to smallest (LogMAR = 0.0). The subject with normal vision has a LogMAR score of 0.0, which is equivalent to a 20/20 score in Snellen chart.

Our system is designed for VA testing at a distance of 3 m. This is a feasible distance for most home or clinic room in our country, where space constraints are common. In each test round, our system displays a line of 5 random optotypes. Table I provides the height of each optotype line and its corresponding LogMAR score, following the guidelines of LogMAR chart [20]. The system's calibration function allows it to calculate the optotype size (in pixels) based on the dot size (in mm) of the LCD monitor.

TABLE I. Conversion Table between the Optotype Size and Scores

Line	Optotype height (mm) (For 3 m distance)	LogMAR score	Snellen score
0	43.7	1.0	20/200
1	35.0	0.9	20/160
2	27.3	0.8	20/125
3	21.9	0.7	20/100
4	17.5	0.6	20/80
5	13.8	0.5	20/63
6	10.9	0.4	20/50
7	8.8	0.3	20/40
8	7.0	0.2	20/32
9	5.5	0.1	20/25
10	4.4	0.0	20/20

The system updates the LogMAR score each time the subject provides an answer. For each correct answer, the score decreases by 0.02 LogMAR units. If the subject correctly identifies at least 3 out of 5 optotypes on a given line, they can proceed to the next smaller line. The best achievable VA score is 0.00 LogMAR.

B. The Speech Recognition Module

For the implementation of this module, we use the Python library Speech Recognition (SR), which supports

multiple ASR engines. The speech recognition process in our program consists of two main stages:

- Capturing the subject's speech via a microphone and saving it as an audio variable.
- Sending the variable to an ASR engine which transcribes it into text and returns the result.

The detailed procedure is illustrated in Fig. 2.

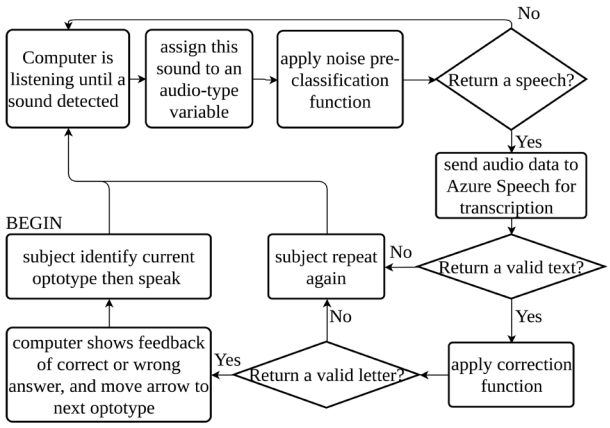

Fig. 2. Answer selection workflow using the speech recognition method.

To determine the most suitable ASR engine for our system, we tested three cloud-based engines integrated in SR library: Google Speech, Microsoft Azure Speech, and wit.ai Speech. All three engines support multi-languages, including Vietnamese, and offer free transcription services. However, wit.ai was eliminated from our project because it has a long response time in Vietnamese transcription (more than 2 s).

Vietnamese uses the same Latin alphabet as English, but differs significantly in pronunciation. For all ASR engines, transcribing isolated words is generally more difficult than transcribing sentences [21–23]. Recognizing monosyllabic letters is particularly challenging. Google Speech has the shortest response time, from 400 to 600ms. While it works well with the sentences, it performs poorly when recognizing isolated Vietnamese letters (only 30–40% success). The Azure Speech API demonstrated the best performance, with a successful rate of around 55–65% when recognizing isolated letters in Vietnamese. Furthermore, the Azure APIs indicated an acceptable response time ranged from 500 to 1200 ms.

We also explored offline ASR models inside Python SpeechRecognition library, including Vosk and OpenAI's Whisper, which support Vietnamese. However, these models demonstrated significantly lower accuracy for recognizing isolated Vietnamese letters. Kaldi, while powerful, requires extensive training data and customization for Vietnamese isolated letter recognition, which was beyond the scope of our current resources and timeline. Based on these findings, the Azure API have been integrated into our program as the primary transcription engine.

When subject utters a letter, the Azure API can return two types of errors:

- Null value: This occurs when the engine fails to transcribe the sound into any text. Causes of this error may include background noise, unclear pronunciation, or poor microphone quality. When a null value is returned, the subject needs to repeat this letter.
- Wrong word: Sometimes, the ASR transcribes the sound into an incorrect word which is a homophone, or a phonetically similar variant of the intended letter. For instance, when the subject says "C" in English, the ASR might transcribe it as "see" or "sea". In Vietnamese, this issue is exacerbated, because this language has six distinct tones, and most letters have more than one pronunciation. Therefore, the number of incorrect variants for each letter is significantly higher than in English.

To improve the accuracy of speech recognition, we implemented the correction function and the speech/noise classification function with the Azure API process.

1) Correction function

To address high rate incorrect words returned by the Azure API, we implemented a correction function to convert these errors into valid letters. This function is based on a dictionary of variants that was constructed through an experiment involving 20 participants, each with a diverse voice profile. Each participant was asked to pronounce each letter of the alphabet (using Vietnamese pronunciation) multiple times through a microphone. The Azure Speech engine then processed these recordings to generate transcription variants. A variant was added to the correction dictionary for each letter if it met three criteria:

- Phonetic similarity: The variant must share at least one similar syllable with the intended letter, based on Vietnamese pronunciation.
- Frequency Across Participants: The variant must appear at least twice and must be spoken by at least two different participants.
- Uniqueness: the variant must correspond to only one letter and should not appear as a valid transcription for any other letter

Table II summarizes the number of variants identified for the 10 Sloan letters, based on their Vietnamese pronunciation:

TABLE II. NUMBER OF VARIANTS OBTAINED FOR 10 SLOAN LETTERS IN VIETNAMESE PRONUNCIATION

Sloan letters	Number of variants
C	12
D	16
H	10
K	7
N	14
O	9
R	6
S	24
V	11
Z	14

Example: 12 identified variants of letter "C" in Vietnamese that are re-transcribed to the correct answer "C" in the correction function:

['C','CƠ', 'CỜ', 'CÒ', 'CỒ', 'XÊ','XỀ','XE', 'SÊ', 'SỀ', 'SỂ', 'SI']:

These variants are phonetically like the intended letter "C" in Vietnamese pronunciation and were mapped to the correct transcription. By utilizing this dictionary of variants, the system can effectively correct recognition errors returned by the ASR engine, thereby improving the overall accuracy of the speech recognition process. After implementing the correction function, we observed a significant improvement in the speech recognition success rate during the VA test, with accuracy exceeding 90%.

2) Speech/Noise pre-classification function

The speech recognition procedure can be disrupted by ambiance noise, and significant efforts have been made to address this issue [24, 25]. Although the Azure Speech API incorporates effective background noise suppression, we identify an additional vulnerability: when ambiance noise occurs during subject silence, the system captures and process this audio through Azure API. The result is typically a null value, then the program resumes listening. However, if subject begins speaking during the API's processing window (500–1200 ms) following prior noise capture, the program may miss speech input due to interrupted audio sampling.

To address this issue, we developed a Speech/Noise pre-classification function that executes following the audio capture. When the classifier identifies the audio segment as noise, the system: (1) bypasses the API transcription, (2) immediately reinitializes the listening phase. These two actions prevent the risk of missing the subject speech.

The classification function utilizes the Root Mean Square (RMS) energy of the captured audio, computed via Python's *audioop* library. Through empirical observation, we found two key characteristics: (1) moderate ambient noise (<50 dB) exhibits RMS value 3–6 times lowers than that of subject speech. (2) Individual speakers maintain relatively consistent volume, with speech RMS variation remaining within a 2× range throughout testing. Based on these findings, the classification works as follows (see Fig. 3)

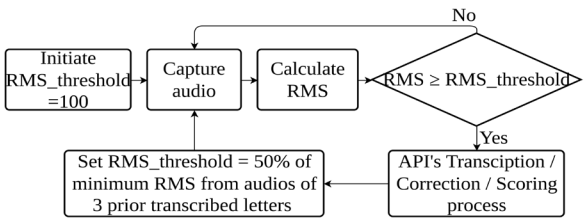

Fig. 3. Speech/Noise classification workflow.

Since rms_threshold adjust dynamically after each test iteration, the sensitivity variations of microphone don't impact the classification result. Our evaluations of three mid-range microphones (Jabra Biz 150, Yealink UH37, Logitech H540) demonstrate consistent performance, while minor sensitivity differences were evident in RMS measurement, the system maintained effective speech/noise discrimination across all devices. This eliminates the need for device-specific parameter calibration. However, the low-end microphones are not recommended as they may compromise recognition accuracy due to inferior audio capture quality.

C. Feedback by Speech From Computer

Once the system successfully recognizes an optotype through subject's speech, it immediately provides feedback by computer speech through either headphones or the PC's loudspeaker. For speech recognition method, the feedback is given in the form of: identified letter + "correct" or identified letter + "wrong", depending on the matching of the recognized letter with the letter on screen. To generate speech from the text, we utilize the Python library gTTS (Google Text-to-Speech). To ensure response accuracy, subject may say "Lại" (Vietnamese for *Undo*) when recognized letter differ from their spoken input, activating a repeat of current optotype. The auditory feedback ensures the process is more intuitive and transparent for the user

D. The Experiment Procedure

To evaluate the performance of our system, we conducted 2 separate experiments, with 80 participants (N = 80), including 40 males and 40 females, aged 18–65 years (mean age = 32). The participants were recruited from the Institute of Materials Sciences, comprising employees and internship students. These well-educated subjects were selected to ensure comprehension of the unsupervised self-testing process, which is critical for the system's intended use in non-clinical settings. For home-based VA testing, we recommend that at least one family member be a competent computer user to set up the system, while children can perform the test under parental guidance.

The participant group consisted of 44 emmetropes (without wearing refractive glasses) and 36 myopes (refractive error ranging from −0.5 to −3.0 diopters). Myopic participants were instructed to wear their corrective lenses during testing. This distribution reflects a common range of visual conditions in the general population, supporting the system's applicability to both corrected and uncorrected vision.

Participant with severe visual impairment (unable to identify optotypes at LogMAR 1.0, Snellen 20/200) were excluded from this study. The testing procedure adheres International Visual Acuity measurement standard [26] and the guidelines proposed by Elliott [27].

Each participant sat 3 m away from a 32 inches Samsung LCD screen (luminance ~190cd/m^2). We used a moderately powered computer (Intel Core i5, 8 GB RAM) running Windows 10 for the test. For speech recognition module, we used a Jabra Biz 150 microphone/headset. Other mid-range microphone could also be substituted without affecting the system's performance. The experiments took place in an office room at our institution (Institute of Materials Science).

Each subject participated in three VA experiments, first with the right eye (OD), then with the left eye (OS). The experimental setup was as follows:

- Experiment 1 (speech recognition): the subject spoke the Sloan letters in Vietnamese through microphone and the system recognized these

responses. Subjects had option to start from line 1, line 3, line 5, or the default line 7, depend on the estimated visual acuity.

- Experiment 2 (manual ETDRS control): a conventional printed ETDRS chart was placed in the same position of the LCD screen. During this test, an assistant was required to stand near the chart to point to each letter, while the subject verbally identified the letter. The assistant then manually recorded the result following standard clinical protocol.

Beyond standard VA scoring, the system quantified two addition parameters: 1) Total test duration (T) and 2) Number of attempted optotypes (Op). Otherwise, we manually recorded E_s: Instances of failed speech recognition by ASR during each test. Therefore, two important factors were calculated and analyzed:

(1) Word Error Rate (WER) = $\frac{Es}{Op}$

(2) Time per Optotype (TpO) = $\frac{T}{Op}$

IV. Result and Discussion

A. The VA Scores

Fig. 4 presents the comparative VA between the two experiments with Bland–Altman analysis [28].

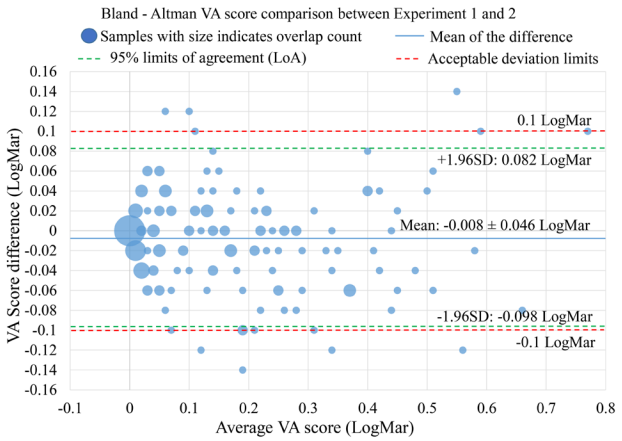

Fig. 4. VA score difference in LogMAR between Experiment 1 and Experiment 2.

Key findings include:

(1) Mean difference: The mean difference in VA score across 160 tests was −0.008 LogMAR (SD = 0.046). This mean value is close to 0, indicating no significant systematic bias between the two experiments. The small standard deviation (equivalent to a variation of 2 optotypes) reflects minor variability, which falls within acceptable clinical limits.

(2) Agreement: The 95% Limits of Agreement (LoA) ranged from −0.096 to 0.082 LogMAR, well within the acceptable deviation threshold of ±0.1 LogMAR (equivalent to a variation of 1 line). Notably, 95.6% of measurements (153/160 eyes) had the VA score difference ≤0.1 LogMAR, indicating strong concordance between two measurement methods.

(3) Identical results: In 56.3% of tests (90/160 eyes), the VA score difference was ≤0.02 LogMAR (≤1 optotype variation), further supporting the high agreement between measurements.

This result compares favorably with established benchmarks:

- Nisar *et al.* [16] indicated 83.8% to 91.9% concordance between speech recognition and conventional methods.
- Taufik and Hanafiah [15] reported 0.19 row difference compared with conventional Snellen chart
- Li and Tong [13] achieve 96.72% concordance between hand gesture recognition in the Tumbling E eye chart and conventional methods
- Cotter *et al.* [29] reported 89% agreement in test-retest study.
- Claessens *et al.* [30] found mean difference with the reference in VA assessment ranged from −0.08 to +0.10 LogMAR for digital VA tools.

Given that the visual acuity testing accepts ±1 line as clinical equivalent accuracy, our system's performance demonstrates sufficient reliability visual screening applications.

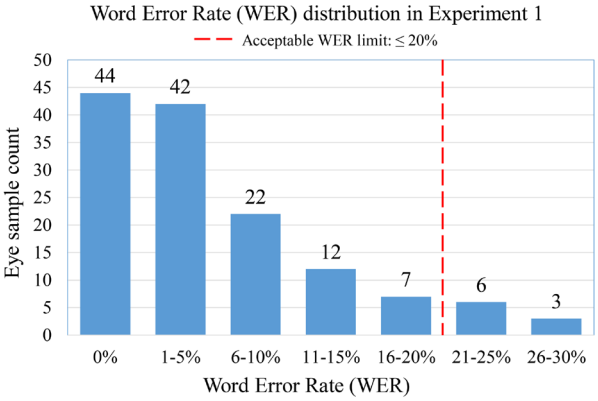

Fig. 5. The Word Error Rate (WER) distribution in Experiment 1.

B. The Speech Recognition by Error Rate

Experiment 1 demonstrated the following accuracy metrics:

(1) Error distribution: ranged from 0−6 errors per test (mean = 1.15).

(2) Word Error Rate (WER) distribution (see Fig. 5): ranged from 0–30% (mean 6.9% ± 6.6%), corresponding recognition accuracy of 93.1%.

(3) Error-free tests: 27.5% of tests (44/160) achieved perfect recognition (WER = 0).

(4) High-error tests: 5.62% of tests (9/160) exceed the acceptable threshold (WER > 20%). Note: WER = 20% represents the maximum tolerable error rate (equivalent to one incorrectly recognized letter per line)

This result is favorable compared with other studies:

(1) Ganesan and Shalini [14] and Nisar *et al*. [16]: WER metrics were not reported in the VA testing,
(2) Taufik and Hanafiah [15]: accuracy 91.4% (WER = 8.6%), using digit number input as optotype. Notably, digits are typically easier for ASR systems to recognize compared to letters.
(3) Adam and Salam [23]: 64.75% accuracy of 26 English letters for their ASR engine.
(4) Accuracy of hand tracking input technique: Chiu *et al*. [12]: 91.6%, Li and Tong [13]: 96.7%.

Our analysis of transcription outcome revealed 58% of correct transcriptions came directly from Azure Speech API. The remaining 42% return initially incorrect words but were subsequently rectified by our correction function. Otherwise, while our primary experiments were conducted in an office environment at the Institute of Materials Sciences, this setting was not acoustically controlled, with occasional ambient noises (e.g., conversations among waiting participants). In fact, more than 90% recorded concurrent noises were correctly labeled by the classification function, maintaining the uninterrupted audio sampling. These evidences proved the contribution of our correction function and classification function.

The 6.9% WER (Word Error Rate—when subject spoke clearly but ASR failed) consists of:

(1) False negatives (82% errors):
- Capture failures: a) low volume speech (RMS < threshold, misclassified as noise) or b) concurrent noise (RMS ≥ threshold, misclassified as speech) interrupt audio sampling while the subject start speaking.
- Transcription failures: a) ASR returned null for valid speech or b) correction function failed to match ASR output to a valid letter.
- The subject needs to speak again when a false negative occurs

(2) False positives (18% of errors):
- Phonetically similar Vietnamese letter pairs (D/T, A/K, /L/N).
- The correction function was unable to correct this type of error.
- The subject needs to undo this answer when a false positive occurs, then speak the letter again.

Users with unclear pronunciation may have difficulty in using this method (WER > 20%, 5.62% of tests). Therefore, two alternative methods were proposed for them in this system: (1) subject verbally responds to an assistant, then the assistant provides keyboard input. (2) the subject uses a wireless keyboard for answer input.

In summary, our speech recognition method achieved relatively high success rates (93.1%). The integrated noise classification function effectively reduces the impact of moderate ambient noise (40–50 dB), such as outside speech or street vehicle noise, while previous systems [14–16] require controlled acoustic environments (<30 dB). While extremely noisy conditions remain challenging, the system operates reliably in typical home settings.

C. The Speech Recognition by Test Time

While participants could select their starting line, total test duration is unreliable for evaluating automated VA system performance due to variable test length. Instead, the TpO is a more robust metric for assessing system efficiency.

Key findings:

(1) Automated system (Experiment 1): TpO ranged from 3.2 to 7.2 s (see TpO distribution in Fig. 6), mean TpO 4.6 ± 0.9 s.
(2) Benchmark comparison:
- Conventional clinical test (observed in Vietnamese clinic): 1.5–3 s,
- Manual ETDRS test (Experiment 2): Mean 2.4 s.

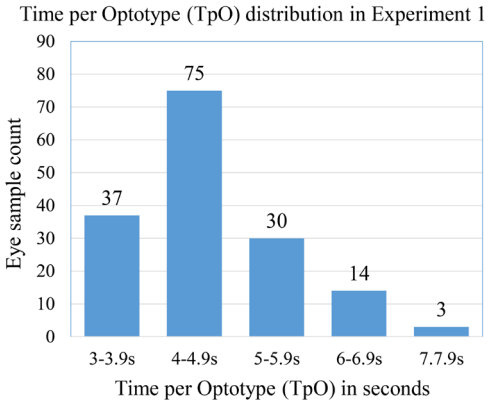

Fig. 6. The Time per Optotype (TpO) distribution in Experiment 1.

While no prior studies explicitly reported TpO benchmarks, our experiment indicated the automated system operates at approximately half the speed of conventional methods. The slower TpO could be attributed to the cloud-based Azure Speech API process, which requires more response time (500–1200 ms). Otherwise, if a speech recognition failure occurs, the subject needs to repeat the letter, causing an additional delay. The computer feedback by speech (800–1200 ms) per optotype (with format: identified letter + correctness) is another factor which slows down the test speed. However, it is necessary for subjects to verify the accuracy of speech recognition. This yields a comparative advantage: The system of Ganesan and Shalini [14] and Nisar *et al*. [16] lack auditory feedback, while the feedback of Taufik and Hanafiah [15] is binary correct/wrong only.

While the system's test speed of about half that of conventional test, this remains acceptable for home use, because:

- Eliminates need for clinical staff
- Flexible testing procedure
- Maintain accuracy through verification steps

This presents the first quantitative evaluation of timing metrics in automated VA testing, addressing a critical gap in prior works.

D. User Feedback

A usability survey was conducted with 20 participants after completing the VA tests. All of them expressed their confidence in the VA result. Qualitatively, 16 participants

found the speech recognition interface intuitive and easy to use. However, four participants reported initial difficulties in pronouncing letters clearly for recognition, attributing this to unfamiliarity with the testing procedure. They noted that their performance could be improved with more practice.

Additionally, feedback was gathered from 10 volunteers who installed the system at home for family use. These tests achieved a success rate of over 80% with speech recognition. Challenges were observed with young children (aged 6–12) and older adults (aged >70), who occasionally spoke unclearly or did not adhere to testing protocols. In such cases, an alternative method was employed: the participant spoke the letters, and a family member provided keyboard input, ensuring test completion. Volunteers reported that the system was user-friendly for home setups, with clear instructions and robust performance in typical household environments (40–50 dB ambient noise).

E. Generalization to Multi-Languages

In future work, extending the system to languages such as English would significantly broaden its applicability. To achieve it, we propose an approach in three phases:

(1) **Data collection of spoken letters:** We need to recruit native speakers pronouncing isolated alphabet letters multiple times via microphone in the target language (e.g., English). Alternatively, we will search and download an online dataset of spoken English letters. This phase ensures coverage of diverse accents and pronunciations.

(2) **Construction of a Language-Specific Correction Dictionary:** Using the collected audio data, we would process the recordings through the Azure Speech API to generate transcription variants for each letter. A correction dictionary would then be built following the three criteria outlined in our study: phonetic similarity, frequency across participants, and uniqueness to each letter. This dictionary would map wrong transcriptions to the correct letters, enhancing recognition accuracy for the target language.

(3) **Validation with Native Speakers**: The system would be validated through VA tests conducted with a diverse group of native speakers to ensure robust performance across various accents and speaking styles. While recruiting sufficient native English speakers in Vietnam may have logistical challenges, collaboration with international research partners could facilitate this process.

This approach inherits the existing framework of our system, particularly the Azure Speech API's multi-language support and our correction function methodology, making it adaptable to other languages with minimal structural changes. Future work will implement and validate this process for English and other widely spoken languages to maximize the system's global impact.

F. Comparison with Prior Work in Summary

Tables III, IV, and V indicate the comparison of our study and three prior works in 10 criteria: VA scoring, speech language, ASR engine, correction function, noise classification, auditory feedback, VA difference metrics, WER metrics, TpO metrics, and error analysis.

TABLE III. COMPARISON OF METHODOLOGY IN FOUR VA TEST SYSTEMS

System	VA scoring	Speech language	ASR engine
Ganesan and Shalini [14]	Line-by-line	English	Microsoft Speech SDK within the LabView platform
Taufik and Hanafiah [15]	Line-by-line	English	Custom Convolutional Neural network (CNN)
Nisar *et al*. [16]	Line-by-line	English	Adaptive Mel filter bank for feature extraction and three classifiers (HMM, SVM, KNN)
Our work	Letter-by-letter	Vietnamese	Microsoft Speech API

TABLE IV. COMPARISON OF FEATURES IN FOUR VA TEST SYSTEMS

System	Correction function	Noise classification	Auditory feedback
Ganesan and Shalini [14]	No	No	No
Taufik and Hanafiah [15]	No	No	Only binary Correct/Wrong
Nisar *et al*. [16]	No	No	No
Our work	Yes	Yes	Recognized letter + Correct/Wrong

TABLE V. COMPARISON OF DATA ANALYSIS IN FOUR TEST VA SYSTEMS

System	VA difference metrics	WER metrics	TpO metrics	Error analysis
Ganesan and Shalini [14]	No	No	No	No
Taufik and Hanafiah	Mean 0.19 row difference	91.4% accuracy	No	No
Nisar *et al*. [16]	83.8% to 91.9% concordance	No	No	No
Our work	95.6% measurements have difference ≤ 0.1 LogMAR	93.1% accuracy	4.6 ± 0.9 s	Yes

V. CONCLUSION

This study presents an automated Visual Acuity (VA) test system implemented as a Human-Computer Interaction (HCI) platform incorporating speech recognition technology. While speech recognition has become ubiquitous in consumer devices, no commercial VA testing system has successfully integrated this technology, due to the significant challenges in isolated letter recognition.

Key innovations compared to prior works:

(1) Accuracy enhancements:
- Improved Azure Speech API performance from 58% to 93.1% accuracy for Vietnamese letters through our novel correction algorithm
- Maintained robust performance in moderate noise environments (40–50 dB) via adaptive noise/speech classification
- Comprehensive auditory feedback for verification

(2) Established three quantitative performance metrics:

- VA score validity based on ETDRS chart, offering more accuracy than Snellen test in prior works [14–16].
- Recognition reliability (mean WER = 6.9%)
- Operational efficiency (mean TpO = 4.6 s)

(3) Provided detailed error analysis for future improvements

While the system demonstrates strong reliability, it has two notable constraints:

- Relatively slow speed than conventional method, primarily due to cloud-based speech processing latency (500–1200 ms), and auditory verification steps (800–1200 ms)
- Requires reliable Internet connection for Azure Speech and Google text-to-speech service

While this research implemented speech recognition exclusively in Vietnamese, future versions of the system will aim to support English and other languages, expanding its applicability worldwide. As Azure API supports multi-language, and almost program features are reusable, the main challenge remains in constructing a correction function dictionary for each language implemented in this system.

In summary, this system represents a significant advancement in applied AI in medical screening, offering a reliable and flexible automated solution for conducting the VA test outside clinical environments.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

All authors contributed to the study conception and design. Material preparation was performed by NDT. Methodology was performed by PHM. Formal analysis was performed by PHD. Software was developed by PHM. Investigation of experiments and data analysis were performed by PHM and NDT. Manuscript writing was performed by PHM and PHD; all authors had approved the final version.

FUNDING

This work was supported by Institute of Material Sciences, Vietnam Academy of Science and Technology (VAST) under the grant of the project CSCL 04.06/24-25.

ACKNOWLEDGMENT

The authors thank the Institute of Material Sciences, Vietnam Academy of Science and Technology (VAST) to support this work. We also extend our thank to Hanoi Eye Hospital 2 for their valuable guidance in developing the eye test procedure for our system.

REFERENCES

[1] W. Wang, L. Zhu, S. Zheng *et al*., "Survey on the progression of myopia in children and adolescents in Chongqing during COVID-19 pandemic," *Front. Public Health,* 2021. https://doi.org/10.3389/fpubh.2021.646770

[2] H. Singh, H. Singh, U. Latief *et al*., "Myopia, its prevalence, current therapeutic strategy and recent developments: A review," *Indian J. Ophthalmol.,* vol. 70, no. 8, pp. 2788–2799, 2022. https://doi.org/10.4103/ijo.ijo_2415_21

[3] J. Liang, Y. Pu, J. Chen *et al*., "Global prevalence, trend and projection of myopia in children and adolescents from 1990 to 2050: A comprehensive systematic review and meta-analysis," *British Journal of Ophthalmology*, vol. 109, pp. 362–371, 2025. https://doi.org/10.1136/bjo-2024-325427

[4] X. Han, J. Scheetz, S. Keel *et al*., "Development and validation of a smartphone-based visual acuity test (vision at home)," *Translational Vision Science & Technology*, vol. 8, no. 4, pp. 27–27, 2019. https://doi.org/10.1167/tvst.8.4.27

[5] S. Tofigh, E. Shortridge, A. Elkeeb *et al.,* "Effectiveness of a smartphone application for testing near visual acuity," *Eye*, vol. 29, no. 11, pp. 1464–1468, 2015. https://doi.org/10.1038/eye.2015.138

[6] A. A. Manzano and M. A. N. Lagamayo, "A comparison of distance visual acuity testing using a standard ETDRS chart and a tablet device," *Philipp. J. Ophthalmol.,* vol. 40, no. 2, pp. 88–92, 2015.

[7] S. K. Gupta, C. Deepa, and K. Tapas, "Validation of the smartphone-based snellen visual acuity chart for vision screening," *Optometry & Visual Performance*, vol. 11, no. 1, 2023.

[8] B. Michael, "The freiburg visual acuity test—Automatic measurement of visual acuity," *Optometry and Vision Science*, vol. 73, no. 1, pp. 49–53, 1996. https://doi.org/10.1097/00006324-199601000-00008

[9] J. Claessens, J. Van Egmond, J. Wanten *et al*., "The accuracy of a web-based visual acuity self-assessment tool performed independently by eye care patients at home: method comparison study," *JMIR Formative Research*, vol. 7, e41045, 2023.

[10] N. Vrabič, B. Juroš, and M. T. Pompe, "Automated visual acuity evaluation based on preferential looking technique and controlled with remote eye tracking," *Ophthalmic Research*, vol. 64, no. 3, pp. 389–397, 2021. https://doi.org/10.1159/000512395

[11] S. C. Ong, L. C. I. Pek, T. L. C. Chiang *et al*., "A novel automated visual acuity test using a portable head-mounted display," *Optometry and Vision Science*, vol. 97, no. 8, pp. 591–597, 2020. https://doi.org/10.1097/OPX.0000000000001551

[12] C. J. Chiu, Y. C. Tien, K. T. Feng *et al*., "Intelligent visual acuity estimation system with hand motion recognition," *IEEE Transactions on Cybernetics*, vol. 51, no. 12, pp. 6226–6239, 2021. https://doi.org/10.1109/TCYB.2020.2969520

[13] C. Li and W. Tong, "A visual acuity assessment system based on static gesture recognition and naive bayes classifier," *International Journal of Information Technologies and Systems Approach (IJITSA)*, vol. 17, no. 1, pp. 1–23, 2024. https://doi.org/10.4018/IJITSA.345926

[14] K. Ganesan and D. Shalini, "Design of customizable automated low cost eye testing system," *Journal of Clinical and Diagnostic Research: JCDR*, vol. 8, no. 3, 85, 2014.

[15] D. Taufik and N. Hanafiah, "Autovat: An automated visual acuity test using spoken digit recognition with MEL frequency cepstral coefficients and convolutional neural network," *Procedia Computer Science*, vol. 179, pp. 458–467, 2021. https://doi.org/10.1016/j.procs.2021.01.029

[16] S. Nisar, M. A. Khan, F. Algarni *et al*., "Speech recognition-based automated visual acuity testing with adaptive mel filter bank," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2991–3004, 2022. https://doi.org/10.32604/cmc.2022.020376

[17] P. K. Kaiser, "Prospective evaluation of visual acuity assessment: A comparison of snellen versus ETDRS charts in clinical practice (An AOS Thesis)," *Trans. Am. Ophthalmol. Soc.,* vol. 107, pp. 311–324, 2009.

[18] I. L. Bailey and A. J. Jackson. "Changes in the clinical measurement of visual acuity," *Journal of Physics: Conference Series*, vol. 772, 012046, 2016. https://doi.org/10.1088/1742-6596/772/1/012046

[19] S. Kalpana, J. Karthick, and S. Jayarajini, "Comparison of static visual acuity between Snellen and early treatment diabetic retinopathy study charts," *International Journal of Educational Research and Development,* vol. 2, no. 3, pp. 82–88, 2013.

[20] I. L. Bailey and J. E. Lovie-Kitchin, "Visual acuity testing. From the laboratory to the clinic," *Vision Research,* vol. 90, pp. 2–9, 2013. https://doi.org/10.1016/j.visres.2013.05.004

[21] P. Borde, A. Varpe, R. Manza *et al.,* "Recognition of isolated words using Zernike and MFCC features for audio visual speech

recognition," *Int. J. Speech Technol.,* vol. 18, no. 2, pp. 167–175, 2015. https://doi.org/10.1007/s10772-014-9257-1

[22] B. Paul, S. Phadikar, S. Bera *et al.,* "Isolated word recognition based on a hyper-tuned cross-validated CNN-BiLSTM from mel frequency cepstral coefficients," *Multimed. Tools Appl.,* vol. 84, no. 17, pp. 17309–17328, 2024. https://doi.org/10.1007/s11042-024-19750-3

[23] T. B. Adam and M. Salam, "Spoken English alphabet recognition with mel frequency cepstral coefficients and back propagation neural networks," *International Journal of Computer Applications*, no. 42, no.12, pp. 21–27, 2012.

[24] S. C. Lee, J. F. Wang, and M. H. Chen, "Threshold-based noise detection and reduction for automatic speech recognition system in human-robot interactions," *Sensors*, vol. 18, no.7, 2068, 2018. https://doi.org/10.3390/s18072068

[25] J. Chen, J. Benesty, Y. Huang *et al*., "Fundamentals of noise reduction," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, Y. A. Huang, eds. Berlin: Springer Berlin Heidelberg, 2008, pp. 843–872. http://doi.org/10.1007/978-3-540-49127-9_43

[26] Consilium Ophthalmologicum Universale, "Visual acuity measurement standard," *Visual Functions Committee*, pp. 1–15, 1988.

[27] D. B. Elliott, *Clinical Procedures in Primary Eye Care*, 3th ed. Butterworth-Heinemann, 2007, 31.

[28] J. M. Bland and D. G. Altman. "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, pp. 307–310, 1986.

[29] S. A. Cotter, R. H. Chu, D. L. Chandler *et al*., "Reliability of the electronic early treatment diabetic retinopathy study testing protocol in children 7 to <13 years old," *American Journal of Ophthalmology*, vol. 136, no. 4, pp. 655–661, 2003. https://doi.org/10.1016/S0002-9394(03)00388-X

[30] J. L. J. Claessens, J. R. Geuvers, S. M. Imhof *et al*., "Digital tools for the self-assessment of visual acuity: A systematic review," *Ophthalmology and Therapy*, vol. 10, pp. 715–730, 2021.