Robust Audio-Visual Speech Recognition in Noisy Clinical Environments

Akshita Abrol*, Ridwan Arefeen†, Haotong Yu†, Alexi George†,
Kelvin Zhenghao Li‡, Zhengkui Wang*, Rong Tong*
* Singapore Institute of Technology, Singapore, {akshita.abrol,zhengkui.wang,tong.rong}@singaporetech.edu.sg
† Singapore Institute of Technology, Singapore, {2403754,2302705,2400788}@sit.singaporetech.edu.sg
‡ Tan Tock Seng Hospital, Singapore, kelvin_li@ttsh.com.sg

***Abstract*—Visual acuity (VA) testing is a foundational clinical procedure for assessing vision clarity. Traditionally, VA testing requires patients to read optotypes aloud from a distance, with an assistant present to record the responses. Speech-enabled automation has been explored, however these solutions are highly susceptible to performance degradation from ambient noise and speech crosstalk: a common challenge in real-world, multi-lane clinical environments. In this work, we introduce a novel multimodal pipeline system that combines audio and visual speech recognition to enable robust and accurate VA transcription even under noisy conditions. By integrating synchronized microphone and camera inputs, our architecture employs advanced noise and crosstalk detection, and utilizes a dynamic decision module to intelligently select between audio-based ASR and lip-reading outputs based on real-time environmental analysis. To further improve accuracy under crosstalk conditions, we introduce a lip-guided audio masking mechanism that suppresses non-target speech by detecting when the patient is actively speaking. This multimodal pipeline represents a significant advancement for automated vision screening, offering enhanced transcription accuracy and reliability where single modality systems fail. Preliminary results demonstrate that the addition of lip-reading achieves a 33% and 16% relative reduction in word error rate (WER) compared to the audio-only baseline on the three-speaker and two-speaker crosstalk visual acuity test sets, respectively.**

I. Introduction

Visual acuity (VA) is a fundamental component of eye examinations for quantifying the level of vision. It provides quantitative measure of an individual's ability to distinguish symbols or letters at a standardized distance, typically between 4 and 6 meters. Traditionally, this test involves a clinician instructing patients to read letters from a chart, such as the Snellen chart. To enhance efficiency in busy eye clinics, automated solutions have been developed, including speech-enabled systems that utilize automatic speech recognition (ASR) to interpret patient responses and reduce the need for constant human supervision. However, these systems typically assume a single-speaker, noise-free environment, an assumption that fails in practical deployments with simultaneous screenings in adjacent lanes. Automating this workflow can reduce human error and resource usage, robust systems are needed to function reliably in the acoustic complexity typical of clinical settings.

In real-world clinical environments, overlapping speech and ambient noise frequently lead to significant transcription errors, severely compromising the accuracy and reliability of ASR-based systems. To overcome these challenges, we present a multimodal pipeline approach that augments conventional audio-based ASR with visual speech recognition (VSR), thereby leveraging lip-reading to sustain accurate performance in noisy conditions. This system is designed to adaptively fuse between audio and visual modalities based on real-time environmental analysis, enhancing robustness and workflow integration in clinical settings.

Our previous work [1] introduced a novel automatic visual acuity test system employing automatic speech recognition and image recognition and this laid the foundation for understanding the ASR challenges in clinical environments. Building upon our preliminary system, this study presents an enhanced multimodal pipeline specifically addressing the challenge of crosstalk and noise through real-time dynamic modality selection and lip-guided audio masking."

II. Related Work

Automatic Speech Recognition in multi-speaker environments has long been a key research area, with numerous models and techniques designed to separate speech from overlapping speakers. Yet, traditional approaches like beamforming and microphone arrays often struggle in dynamic settings where real-time separation is crucial. This issue is particularly pronounced in clinical ASR, where crosstalk, which is the interference from simultaneous speech sources, poses a major challenge. Such challenges are frequently encountered in real-world, multi-lane clinical environments where simultaneous screenings in adjacent lanes result in overlapping speech and ambient noise.

Early applications of automatic speech recognition (ASR) in visual acuity testing include the work of Lazaro et al. [2], who evaluated using Hidden Markov Models (HMM) classifiers for speech recognition. Nisar et al. [3] evaluated Support Vector Machines (SVM) and K-Nearest Neighbors as classifers. However, these approaches primarily target isolated (individual) character recognition and exhibit limited performance in terms of speech recognition accuracy.

Recent advances in multimodal speech enhancement have improved noise robustness by leveraging both audio and visual information. For example, Jeon et.al. [4] combined deep neural

979-8-3315-7206-8/25/$31.00

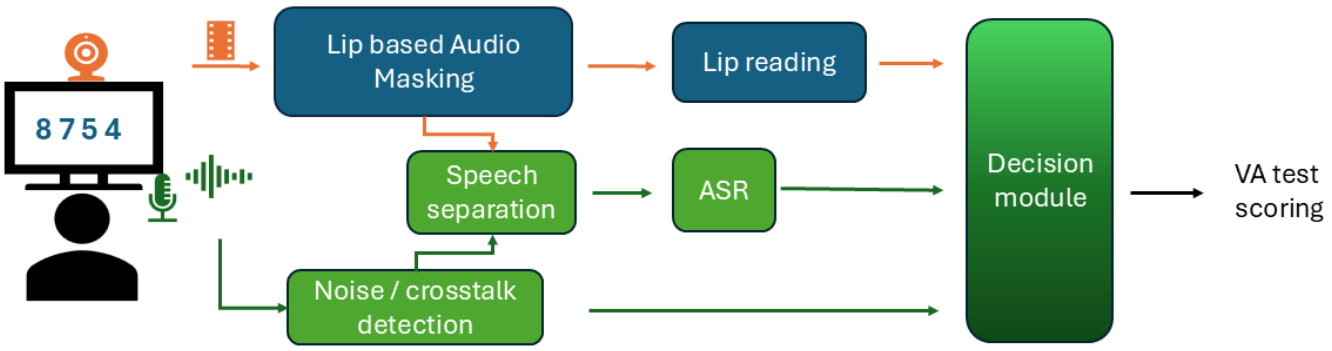

Fig. 1. Proposed audio visual pipeline system for automatic visual acuity screening

network-based visual speech recognition with cloud-based audio recognition, achieving a 6.7% gain in accuracy over audio-only models in noisy settings. The work in [5] introduced an audio-visual speech enhancement method based on conditional variational auto-encoders (CVAE), which demonstrated notable improvements, especially in highly corrupted speech signals. Likewise, Hong et.al. [6] developed the Visual Context-driven Audio Feature Enhancement module (V-CAFE), which enhances noisy audio by capturing lip motion transitions to generate effective noise reduction masks, thereby improving noise robustness in audio-visual ASR.

To address noise and interference collectively, recent research has explored unified frameworks. Graph temporal classification methods have shown improvements in multi-speaker ASR with crosstalk and noise [7], while target speaker extraction models trained on diverse conditions have demonstrated notable gains for real-world ASR tasks [8]. Moreover, state-of-the-art speech separation models such as SepFormer [9] and MossFormer2 [10] have advanced the handling of noise and overlapping speech.

Despite these research advances, the deployment of robust, real-time ASR in busy clinical environments continues to be challenging. General-purpose models like Whisper [11] perform well under controlled laboratory conditions but often struggle in the presence of overlapping speech and high background noise typical in clinical settings.

To address these limitations, this work introduces a novel multimodal pipeline solution designed specifically for the clinical visual acuity testing workflow. Our system leverages both audio and video modalities, utilizing lip-reading to maintain reliable performance in acoustically challenging and multi-speaker environments. By dynamically integrating information from both sources, we move beyond audio-only or vision-only techniques and demonstrate substantial gains in transcription robustness and accuracy, marking a significant step towards scalable, automated VA screening in real-world clinical settings.

III. System Architecture

Figure 1 presents an overview of the proposed multimodal pipeline system for automatic visual acuity test.

The proposed workflow begins with the concurrent capture of both audio and video streams at each test station, where the patient is prompted to read the displayed visual acuity chart. The video feed is processed to extract the relevant region of interest (e.g., the patient's mouth) and then analyzed using a lip-reading engine to predict the spoken sequence visually. Simultaneously, the raw audio is analyzed by a noise and crosstalk detection module, which assesses background noise and detects the presence of overlapping speech. If crosstalk is detected, the audio is first processed by a masking algorithm that suppresses segments where the target speaker is not speaking, based on lip movement analysis. The masked audio is then passed to the speech separation module to isolate the target speaker's voice before being forwarded to the automatic speech recognition engine. Otherwise, the audio may be sent directly to the ASR.

Both the visual-based recognition result (derived from lip-reading) and the audio-based transcription (generated by the ASR) are sent to a central decision module, which also incorporates noise and crosstalk detection analysis. This module assesses the reliability of each transcription stream in real time, dynamically selecting the most dependable result based on prevailing environmental conditions and system confidence levels. The selected transcription is subsequently used for automated visual acuity test scoring. A detailed breakdown of each component is provided below:

A. Audio-Visual Input Processing

The system simultaneously receives mixed audio input from the patient reading optotypes, which may include background speech from other patients or caregivers, as well as ambient noise. Concurrently, the video feed captures the patient's lip movements for visual analysis.

B. Noise and Crosstalk Detection

A dedicated module continuously monitors the audio input to assess environmental conditions. This module incorporates Mean Opinion Score (MOS)-based speech quality prediction [12] and speaker diarisation [13] to detect overlapping speakers, which is indicative of crosstalk.

For speech quality assessment, we adopt an instrumental Mean Opinion Score (MOS) prediction model NISQA [12]. Unlike traditional subjective listening tests, NISQA uses deep learning to estimate overall perceived speech quality and key dimensions (Noisiness, Coloration, Discontinuity, Loudness)

from the degraded audio alone, making it suitable for real-time, single-ended quality monitoring without clean reference signals. Under VA test scenario, a higher MOS score indicates the speech is clear and less noisy.

Speaker diarisation is used to segment the audio stream according to the active speaker(s), enabling the detection of overlapping speech. To achieve accurate diarisation, especially in the presence of overlapping speakers, we employ a system utilizes an end-to-end neural diarisation approach [13]. This method directly predicts speaker activity configurations per frame, effectively identifying simultaneous speakers and improving performance on short, noisy audio segments typically encountered in clinical environments. The speaker diarisation model outputs the number of speakers present in the given speech. If more than one speaker is detected, that means the speech is cross talked.

These MOS score and speaker diarisation output are utilized by the subsequent decision module to suggest which modality output to use.

C. Lip-Based Audio Masking

To mitigate interference from non-target speakers in crosstalk scenarios, a lip-activity-based audio masking technique is employed. This module leverages visual cues from lip movement to determine target speaker activity.

Facial landmarks are extracted from video frames using the MediaPipe FaceMesh model [14]. Specifically, the vertical distance between the upper and lower inner lips (landmarks 13 and 14) is computed per frame, serving as a proxy for mouth movement. Frames are classified as speech-active if this distance exceeds a predefined threshold; otherwise, they are deemed silent. This generates a binary temporal mask of speech activity.

A synchronized audio signal is segmented into intervals aligned with video frames. Each segment is then either retained or suppressed according to the lip motion mask. Segments lacking lip motion are zeroed out, effectively silencing non-speech regions.

By eliminating irrelevant acoustic regions prior to separation, this approach enhances the model's ability to isolate the target speaker's voice, particularly in multi-speaker environments. Furthermore, this preprocessing step substantially reduces the audio data load for the subsequent speech separation module, leading to significant computational resource savings and improved efficiency.

D. Speech Separation Module

If noise is detected in the audio input by the detection module, the input speech is then processed by state-of-the-art speech separation models.

In our previous work, we have evaluated the effectiveness of several models on speech separation for VA settings [15]. These models are designed to divide the mixed audio into distinct audio streams, with the aim of isolating individual speakers and reducing crosstalk. In this work, we utilize MossFormer2 [10] as it outperforms others on the VA test sets.

E. Dual Recognition Engines

ASR Engine: A Whisper-based model fine-tuned on Singaporean English and optotype characters. It provides fast and accurate transcription in clean speech conditions.

Lip-reading Engine: A visual speech recognition (VSR) model trained on a curated dataset of patients articulating optotypes. It extracts visual embeddings of lip movements and maps them to spoken characters.

F. Decision Module

The Decision Module is a pivotal component of the system architecture. It operates by intelligently assessing the reliability of different transcription streams in real-time, dynamically selecting the most dependable result based on current environmental conditions and internal system confidence levels.

The decision module's logic can be expressed as:

$$\mathcal{T} = \begin{cases} \mathcal{T}_{ASR} & \text{if } C_{MOS\_high} \text{ or } C_{Single\_speaker} \\ \mathcal{T}_{VSR} & \text{if } C_{Adverse\_noise} \text{ or } C_{Crosstalk} \end{cases} \quad (1)$$

where $\mathcal{T}$ is the final decision, and $\mathcal{T}_{ASR}$, $\mathcal{T}_{VSR}$ are outputs from ASR and lip-reading respectively. The final decision is made according to the following two conditions:

- Preference for ASR in Optimal Conditions: when the Mean Opinion Score is assessed as high C_{MOS_high} , or speaker diarisation confirms the presence of a single, distinct speaker $C_{Single\_speaker}$, the system prioritizes the output from the ASR engine: $\mathcal{T}_{ASR}$.
- Switching to lip-reading in Adverse Conditions: conversely, if the noise and crosstalk detection module identifies significant ambient noise $C_{Adverse\_noise}$ or overlapping speech (crosstalk) $C_{Crosstalk}$, the system intelligently switches to utilizing the Lip-reading engine's output: $\mathcal{T}_{VSR}$.

This multi-step framework is integral to creating a robust and efficient pipeline that directly addresses the challenges posed by noisy, real-world clinical environments, thereby significantly improving the reliability of automated visual acuity assessments.

IV. Experimental Setup

A. ASR Model Fine-Tuning

The Whisper model [11] has demonstrated strong performance across a range of transcription tasks; however, its deployment in real-time applications necessitates minimizing latency without significantly sacrificing accuracy. To address this, we employed the Whisper-small variant and utilized the Faster-Whisper library [1] for optimized inference. Faster-Whisper leverages CTranslate2 for accelerated decoding on both CPUs and GPUs, significantly improving processing speed while preserving recognition performance.

[1]https://github.com/SYSTRAN/faster-whisper

TABLE I
VISUAL AND AUDIO DATASETS

Name	speaker	condition	speech content
VA-clean	1	quiet	412 digit sequences
VA-2speaker	2	noisy	
VA-3speaker	3		
AVDigits	1	quiet	540 single digits
AVDigits-2speaker	2	noisy	
AVDigits-3speaker	3		

We fine-tuned the Whisper-small model using Low-Rank Adaptation (LoRA) [16] techniques to efficiently update a subset of model parameters, thereby reducing memory footprint and training time. Using the Parameter-Efficient Fine-Tuning (PEFT) library, LoRA adapters were applied to the attention layers with a rank of 32, alpha of 64, and a dropout rate of 0.05, enabling efficient domain adaptation while keeping the majority of the model weights frozen. Fine-tuning was conducted on a custom clinical dataset simulating real-world visual acuity testing scenarios, where patients read optotypes aloud amid background speech and ambient clinical noise. Special emphasis was placed on adapting the model to the Singapore English context, accounting for local accents, and pronunciation variations.

The fine-tuning objective was to optimize transcription of short, discrete utterances (letters and numbers), which are characteristic of visual acuity assessments, while minimizing inference latency. These optimizations are critical for supporting live patient interactions, ensuring accurate and timely transcription in complex, multi-speaker environments.

B. Lip-reading engine

For lip-reading, we employ the open-source AVSR model described in [17]. This system utilizes a two-stage process that exploits large-scale, unlabeled audio-visual datasets in conjunction with publicly available pre-trained automatic speech recognition models. In the first stage, audio from unlabeled video datasets such as VoxCeleb2 and AVSpeech is transcribed automatically using pre-trained ASR models, thereby augmenting the training corpus with pseudo-labels. Combined with labeled datasets like LRS2 and LRS3, this expanded dataset supports the training of audio-only, visual-only, and audio-visual recognition models within a unified architecture.

The model features a visual front-end based on a modified 3D and 2D ResNet-18, and an audio encoder using 1D ResNet-18, both followed by Conformer blocks. Their outputs are fused via a multi-layer perceptron, then processed through a projection layer and a Transformer decoder for joint CTC/attention-based sequence prediction.

In our experiments, we utilize only the visual-only recognition component, as the ASR module's performance does not surpass that of the Whisper model.

C. Dataset Description

In our experiments, we utilized two distinct categories of datasets to evaluate the performance of the proposed multimodal pipeline. Table I summarizes the details of these datasets.

VA Dataset: To assess the system's performance in visual acuity (VA) testing, we collected audio and video data from a cohort of 60 speakers. Among them, 31 speakers provided only audio recordings, while 29 speakers contributed both audio and video data. The data collection was conducted to replicate real-world clinical scenarios, with participants reading optotypes aloud in environments featuring varying levels of background noise and crosstalk from caregivers or other patients. Consequently, the dataset contains recordings with overlapping speech sources and diverse noise conditions, mirroring the acoustic complexity typically encountered in busy clinical settings.

For this study, we utilize only the subset containing both audio and video data from the 29 speakers. To simulate noisy environments, we artificially augmented the original audio recordings by imposing speech from one or two additional speakers. The interfering speakers' speech was randomly sampled from the 31 single-speaker audio-only recordings and mixed with the target speaker's audio at randomized SNRs to simulate realistic crosstalk. Through this process, we generated two crosstalk datasets: *VA-2speaker* and *VA-3speaker*, corresponding to mixtures containing speech from two and three speakers, respectively.

AVDigits: We employed the AV Digits Database [18], an open-source audiovisual dataset comprising normal, whispered, and silent speech. The dataset consists of two sections: digits and short phrases. For this work, we focus exclusively on the digits section with normal speech. In this subset, participants were asked to recite the digits 0 through 9 in English, in random order, five times each. The dataset comprises recordings from 53 participants (41 males and 12 females).

The original AV Digits recordings were collected in a quiet environment without significant background noise. To simulate realistic crosstalk scenarios, we created noisy versions of the dataset by augmenting the clean audio with one or two additional speaker tracks using the same procedure as with the VA dataset. This resulted in two additional test sets: *AVDigits-2speaker* and *AVDigits-3speaker*, representing environments with two and three simultaneous speakers, respectively.

D. Evaluation Metric

The performance of the proposed system was evaluated using the word error rate (WER) metric, which quantifies the discrepancy between the predicted transcription and the ground truth. Specifically, WER is calculated as the sum of substitutions, deletions, and insertions required to transform the predicted transcription into the reference, normalized by the total number of words in the ground truth. This metric provides a comprehensive evaluation of transcription accuracy, allowing for objective comparison of system performance under various testing conditions.

V. RESULTS AND DISCUSSION

A. Performance of ASR in clean and noisy environment

To establish a baseline and evaluate the inherent capabilities and limitations of an audio-only system, we conducted ASR

TABLE II
ASR PERFORMANCE ON VA TEST SETS

Dataset/model	Pretrained	Fine-tuned	Error reduction
VA-clean	0.026	0.013	50%
VA-2speaker	0.441	0.298	32%
VA-3speaker	0.756	0.480	37%

evaluations across various datasets. This analysis was designed to quantify the performance of the Whisper ASR model, both in its original pretrained form and after fine-tuning for our specific application, under conditions ranging from clean speech to those affected by noise and crosstalk.

Table II summarizes the performance of the ASR models on both clean and crosstalk speech. For our experiments, we used the small variant of the Whisper model for both pretrained and fine-tuned versions.

As the number of interfering speakers increases, ASR performance degrades significantly. For example, the word error rate for the original Whisper model increases from 0.026 under clean conditions to 0.441 and 0.756 in environments with overlapping speech, indicating the significant challenges posed by crosstalk. This result highlights the limitations faced by audio-only systems in realistic and practical settings.

Importantly, the fine-tuned Whisper small model consistently outperforms the original pretrained model across all test conditions. Specifically, we observed reductions in ASR error rates of 50%, 32%, and 37% on the VA-clean, VA-2speaker, and VA-3speaker datasets, respectively, when using the fine-tuned model. Nevertheless, the error rates in crosstalk conditions remain relatively high, motivating the exploration and adoption of a multimodal pipeline to further enhance system robustness.

B. Performance of proposed pipeline

Table III presents the comparison of state-of-the-art models for single modality (ASR only and lip-reading only) systems with the proposed multimodal pipeline system. Specifically the performance of fine-tuned ASR system (MossFormer2/ASR), lip-reading model (AVSR/lip-reading), and proposed multimodal pipeline on noisy and crosstalk test sets.

It is worth to note that the performance of the lip-reading model remains consistent across the 2-speaker and 3-speaker test sets, since the video data for the target speaker is unchanged in these scenarios.

When comparing the two single modality systems, the fine-tuned ASR model consistently outperforms the lip-reading model on all datasets. The relatively low accuracy of the lip-reading system can be attributed to the fact that most of its training data consists of videos containing long sentences, whereas digit sequences can easily be confused with similar-looking words. For example, *'three two'* versus *'treat'*, the lip movements are similar in Singapore English.

Despite its modest standalone accuracy, the lip-reading model contributes significantly within the multimodal pipeline. Specifically, on the two-speaker test sets, the word error rate was reduced from 0.30 to 0.25 on the VA dataset, and from 0.62 to 0.31 on the AVDigits dataset. For the three-speaker test sets, the WER decreased from 0.48 to 0.32 on the VA dataset and from 0.88 to 0.47 on the AVDigits-3speaker set. These results highlight that the contribution of lip-reading becomes increasingly significant under more challenging conditions. For example, the addition of lip-reading resulted in a 33% WER reduction on the AV-3speaker set, compared to a 16% reduction on the AV-2speaker test set. A similar trend is observed for the AVDigits datasets.

TABLE III
PERFORMANCE OF SINGLE MODALITY (ASR ONLY AND LIP-READING ONLY) VS PROPOSED MULTIMODAL PIPELINE

Dataset/WER	ASR	Lip-reading	Pipeline
VA-2speaker	0.30	0.71	0.25
VA-3speaker	0.48	0.71	0.33
AVDigits-2speaker	0.62	0.77	0.31
AVDigits-3speaker	0.88	0.77	0.46

ASR: audio only system, MossFormer2/ASR; Lip-reading: vision only system, AVSR; Pipeline: proposed multimodal pipeline

The lip-reading model is particularly effective in accurately detecting the start and end times of the target speaker, which greatly improves audio segmentation. Furthermore, the lip-reading component demonstrates increased robustness in extremely noisy environments where audio-based models alone often fail. By leveraging these complementary strengths, the multimodal pipeline consistently outperforms either single-modality system, underscoring the benefits of integrating visual cues with audio information in acoustically challenging environments.

It was also observed that the AVDigits test set exhibited lower accuracy, even though the target speakers only read single digits. One contributing factor is the very short duration of these audio segments. When noisy speech from additional speakers is added, the overall length of the mixed audio increases. This often results in a significant rise in insertion errors, which adversely affects recognition accuracy.

C. Performance of decision making module

Table IV provides a breakdown of the contributions from the ASR and lip-reading subsystems to the final decision-making process within the proposed pipeline system. The table reports the number of sentences for which each subsystem made the final decision, as well as the corresponding percentage relative to the total number of test cases.

The results show that the lip-reading module contributes to the final decision in 10% of cases for the VA-2speaker test set and 23% for VA-3speaker, while its contributions are 5% and 9% for AVDigits-2speaker and AVDigits-3speaker test sets, respectively. In both test sets, the proportion of decisions influenced by the lip-reading subsystem increases notably as the level of background noise and number of speakers rises.

This trend highlights the growing importance of integrating visual information under extremely noisy conditions, where audio-only models are more likely to fail. These findings supports the effectiveness of the multimodal approach, demonstrating that introducing visual cues through lip-reading provides critical support to the decision-making process, particularly in challenging multi-speaker environments.

TABLE IV
DECISION MAKING IN PROPOSED PIPELINE

Dataset/Decision	by ASR	by Lip-reading
VA-2speaker	369(90%)	43(10%)
VA-3speaker	317(77%)	95(23%)
AVDigits-2speaker	514(95%)	26(5%)
AVDigits-3speaker	490(91%)	50(9%)

VI. CONCLUSION AND FUTURE WORK

This work presents a novel multimodal system designed to deliver robust and accurate visual acuity testing through speech, even in the presence of noise and multiple speakers as commonly encountered in real-world clinical environments. Traditional automated VA solutions relying solely on Automatic Speech Recognition have been highly susceptible to performance degradation from ambient noise and crosstalk. To address these challenges, our system integrates synchronized microphone and camera inputs, incorporates advanced noise and crosstalk detection, and employs a dynamic decision module that intelligently selects between audio-based ASR and lip-reading outputs based on real-time environmental conditions. This multimodal approach marks a significant advancement in automated vision screening, providing improved transcription accuracy and reliability where single-modality systems are insufficient.

While the current system demonstrates strong performance and resilience, future work will focus on optimizing the end-to-end latency of the multimodal pipeline and investigating more sophisticated fusion strategies within the dynamic decision module. Additionally, we plan to expand our clinical dataset to include a broader diversity of accents and noise profiles, further enhancing the system's generalizability and robustness across varied clinical settings.

ACKNOWLEDGMENT

This work is supported by Singapore Ministry of Education (MOE) Ignition grant R-IE2-A405-00006.

REFERENCES

[1] B. P. Yap, M. K. L. Tan, Z. Li, and R. Tong, "Speech enabled visual acuity test," in *Proceedings of Interspeech 2024*, 2024.

[2] J. B. Lazaro, R. G. Garcia, A. L. Generalo, M. A. Halili, and M. S. Montebon, "Speech recognition for control of optotype characters of the snellen chart using logmar transformation," in *AIP Conference Proceedings*, 2018.

[3] S. Nisar, M. A. Khan, F. Algarni, A. Wakeel, M. I. Uddin, and I. Ullah, "Speech recognition-based automated visual acuity testing with adaptive mel filter bank," *Computers, Materials & Continua*, no. 2, pp. 2991–3004, 2022.

[4] S. Jeon and M. S. Kim, "Noise-robust multimodal audio-visual speech recognition system for speech-based interaction applications," *Sensors*, vol. 22, no. 20, p. 7738, Oct. 2022.

[5] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoders," in *Proceedings of the IEEE*, 2020.

[6] J. Hong, M. Kim, D. Yoo, and Y. M. Ro, "Visual context-driven audio feature enhancement for robust end-to-end audio-visual speech recognition," *arXiv preprint arXiv:2207.06020*, Jul. 2022.

[7] X. Chang, N. Moritz, T. Hori, S. Watanabe, and J. Le Roux, "Extended graph temporal classification for multi-speaker end-to-end asr," in *ICASSP 2022*, IEEE, 2022, pp. 1–5.

[8] Y. Liu, X. Liu, X. Miao, and J. Yamagishi, "Libri2vox dataset: Target speaker extraction with diverse speaker conditions and synthetic data," *arXiv preprint arXiv:2412.12512*, 2024.

[9] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021*, IEEE, 2021, pp. 21–25.

[10] S. Zhao, Y. Ma, C. Ni, *et al.*, "Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation," in *ICASSP 2024*, IEEE, 2024, pp. 10 356–10 360.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, PMLR, 2023, pp. 28 492–28 518.

[12] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *arXiv preprint arXiv:2104.09494*, 2021.

[13] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proc. INTERSPEECH 2023*, 2023.

[14] C. Lugaresi, J. Tang, H. Nash, *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[15] A. Abrol, R. Arefeen, K. Z. Li, Z. Wang, and R. Tong, "Real-time speech recognition for noisy multi-speaker clinical environments," in *Accepted by IALP 2025*, 2025.

[16] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, vol. 1, 2022, p. 3.

[17] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *ICASSP*, 2023, pp. 1–5.

[18] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-only recognition of normal, whispered and silent speech," in *2018 ICASSP*, IEEE, 2018, pp. 6219–6223.