Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Requirement Analysis and Audio Data Collection Protocols for Developing an AI-powered Medical Scribing System at LMIC

**SAKIB ZAMAN**
United International University
**Rahat Shahrior**
United International University
**Adiba Tahira**
United International University
**Md Rasel Uddin**
United International University
**Nishat Yasmin**
United International University
**Marzia Zaman**
CMED Health Ltd
**Kaled Masukur Rahman**
United International University
**Saidur Rahman**
Model Urban PHC Clinics (Aalo Clinic)
**Mohammad Nurul Huda**
United International University
**Khondaker A. Mamun**
mamun@cse.uiu.ac.bd

United International University

**Article**

**Keywords:** Automatic Speech Recognition, Clinical NLP, Electronic Health Records, Medical Scribing System, Speaker Diarization, Speech Processing

**Posted Date:** November 24th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11504150/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Requirement Analysis and Audio Data Collection Protocols for Developing an AI-powered Medical Scribing System at LMIC

Sakib Zaman[1†], Rahat Shahrior[1†], Adiba Tahira[1], Md. Rasel Uddin[1], Nishat Yasmin[1], Marzia Zaman[1,2], Kaled Masukur Rahman[3], Saidur Rahman[4], Mohammad Nurul Huda[5], Khondaker A. Mamun[1,5,6*]

[1]Institute of Research, Innovation, Incubation & Commercialization, United International University, Dhaka, 1212, Bangladesh.
[2]CMED Health Ltd., Dhaka, 1206, Bangladesh.
[3]Department of Electrical and Electronic Engineering, United International University, Dhaka, 1212, Bangladesh.
[4]Model Urban PHC Clinics (Aalo Clinic), Bangladesh.
[5]Department of Computer Science and Engineering, United International University, Dhaka, 1212, Bangladesh.
[6]WISE Accelerator and AI Center of Excellence, Washington University of Science and Technology, Virginia, USA.

*Corresponding author(s). E-mail(s): mamun@cse.uiu.ac.bd;
†These authors contributed equally to this work.

**Abstract**

Clinical documentation for the electronic health record (EHR) carries a significant burden on physicians, specifically in low- and middle-income countries (LMICs), where high-patient turnover and limited consultation windows further push documentation to after-hours, leading to high physician burnout. In recent years medical AI scribing system has emerged as an assistive solution to pave the way for clinical documentations, although most of them are commercially built for high-resource languages, making them poorly adaptable to resource-constrained LMICs clinics due to the regional dialects, linguistic code-switching, and dominant ambient noise of busy clinical settings. In this study, we propose an end-to-end optimized framework for clinical data collection required to develop an AI-powered medical scribing system, which are validated in Bangladeshi GP clinics and can be easily adaptable to other LMIC's clinics as well. As part of this protocol, a robust recording software integrated with the EHR system was built to provide seamless audio capture within existing physician's routine workflows and secure data storage, followed by designing a customized doctor-patient consultation protocol for high-fidelity clinical recordings. We systematically evaluated six low-cost off-the-shelf recording devices across different spatial configurations through comprehensive acoustic profiling to determine the optimal choice for real-world clinical recordings. Furthermore, we conducted a quantitative analysis to obtain physicians and patient's feedback on their informed conversation recordings. Clinical recordings collected through our data collection protocol achieved the lowest WER score of only 10.27% in frontier ASR models, making our proposed framework an evidence-based high-fidelity blueprint for clinical recordings to successfully deploy clinical AI scribes in limited-resource clinical settings.

**Keywords:** Automatic Speech Recognition, Clinical NLP, Electronic Health Records, Medical Scribing System, Speaker Diarization, Speech Processing

# 1 Introduction

The administrative burden of electronic health record (EHR) documentation has become one of the most prominent challenges in contemporary healthcare facilities, significantly affecting the quality of patient care and effective healthcare delivery [1, 2]. Despite its attempt to improve patient care and streamline clinical research [3], widespread adoption of EHRs has paradoxically been associated with higher rates of physician burnout, with estimates of billions of dollars being lost each year [4, 5]. This burden is especially high in low- and middle-income countries (LMICs), where time-constrained consultation time forces physicians to manage high patient turnover and handle clinical documentation tasks after-hours, leading to massive clinician's dissatisfaction and cognitive burdens [6]. These challenges are exacerbated by the use of traditional handwritten records, which are still widely adapted in many LMIC contexts [7]. Hence, this administrative burden of manual clinical documentation has been shown to increase the likelihood of medical errors, making it a priority to have a more efficient and reliable way for clinical documentation [6, 8]. In recent years, a medical scribe industry has emerged to overcome such burdens in clinical settings, providing real-time documentation support for clinicians [9, 10]. The dominant recording paradigm in modern ambient AI scribing is ambient listening, where one or more discreet microphones passively capture all conversation that occurs within a room [11]. This approach contrasts sharply with older directed-speech systems that required clinicians to speak directly into a specific microphone, often disrupting the natural flow of interaction with patients . Ambient listening allows clinicians to maintain eye contact and build a relationship with patients, who may feel more comfortable when technology is perceived as unobtrusive [12]. However, this is not feasible in many LMIC healthcare systems due to the lack of data resources for model training and logistics to deploy AI scribes being too expensive [10]. Most commercial AI-driven scribing solutions are developed for high-resource languages, especially for English-speaking facilities, linguistic diversity, environmental noise, and infrastructure limitations make these industry-grade medical scribing solutions difficult to adapt in LMIC clinical settings [13]. For example, the clinical conversation in Bangladesh contains a mixture of medical terms written in both Bangla and English, often delivered in diverse regional dialects and difficult acoustic conditions.

At the core of AI-powered medical scribing system, the optimal choice of recording device along with its spatial configurations and setup at the clinical site significantly dictates the quality of the audio input data, and consequently, impacts the qualitative performance and feasibilities of Automatic Speech Recognition (ASR) systems. One of the biggest developments in recent scribing technology has been the use of small portable microphones, typically in the form of a lavalier or as part of a headset worn by the clinician to keep the microphone close to their mouth [14] and consequently maximize the signal to noise ratio (SNR), by reducing the amount of ambient room noise and reverberation [15]. Heidi has created their own customized wearable microphones, designed specifically for reliable audio capture in clinical settings without the use of clinicians' personal phones [16]. It is especially useful for environments with reverberations or noise, like an open-plan clinic or a busy emergency ward [17, 18]. Another novel method including the use of wearable skin-attached acoustic sensors (SAAS) can recognize conversational speech by the sound of the bone conduction, which could be useful in extremely noisy clinical environments [19]. However, implementing wearable mics comes with severe ergonomic and workflow challenges, where the user struggles to find appropriate points for positioning the microphone and to attach the device [14]. Despite capable of ensuring better sound quality for the clinician, a secondary microphone is required to capture patient's speech for real-time medical scribing at clinics, which often comes with patient's reluctance to wear the microphone for recording purposes. Alternatively, remote or plug-and-play microphones can be the ideal recording choice as an independent unit, which can be placed on the examination table [20]. Despite their flexibility in setup and installation, they still require careful positioning between the doctor and patients to ensure high-fidelity audio capture [21]. Moreover, we can add up to four "Rally Mic Pods" and two "Rally Mic Pod Hubs" to Logitech's Rally Bar system, designed for video conferencing, to create a scalable system for larger rooms or multi-provider environments [22]. Similarly, speakerphones such as Insta360 Wave feature AI-enabled noise reduction algorithms to suppress more than 300 types of environmental noise, providing better sound clarity under less controlled acoustical conditions [23]. Another device that uses AI to automatically distinguish human voices from background noise is Huawei's BP860 [24]. Although these solutions alleviate the need for users to wear their own device, they can be prone to background noise and sound reflections, most significantly they come with a relatively high-cost alternative for the clinical settings of LMIC [25, 26]. In resource-limited environments, the smartphone may be another cost-effective solution as a recording tool, considering the ubiquity of mobile phones in LMICs [27]. Although there is a great amount of variability in the quality of the microphone, its position and the possibility of interference from other phone communications

being used in parallel can hamper seamless audio capture. Besides, the internal microphone of a smartphone typically performs less well than a standalone external microphone, which can lead to poor SNR and make it more susceptible to noise and reverberation [17]. The ability to isolate the clinician's voice from ambient noise sources is one of the most important features in the unpredictable acoustic environments prevalent in many LMIC clinics [28]. Higher levels of configuration, such as speakerphone use microphone arrays and beamforming, can be placed within a suitable spatial configuration to electronically steer the microphone's reception pattern to the desired speaker while suppressing noise from other directions [29]. These systems have been shown to significantly enhance SNR and speech recognition accuracy in dynamic clinical interactions, even for speakers in motion, making them particularly well-suited for dynamic clinical interactions [30]. In addition, strategic hardware integration and placement are essential for optimizing recording setups in challenging clinical environments. Successful configurations aim at keeping the speaker's mouth as close as possible to the microphone to increase the SNR and decrease the impact of ambient noise. Some implementations include having the microphone embedded into the computer that is used in the exam rooms, providing a fixed, optimized microphone recording location. VivaLyn Labs' MedScribe, are meant for secure on-premise deployment within a hospital's network, ensuring that sensitive patient audio and transcripts are not uploaded to an external server, but can be integrated with the hospital's IT infrastructure, and configured to operate within the hospital's network [31]. Setting up microphones away from air vents or noisy equipment and placing them between the doctor and patients can be a simple yet highly effective approach to enhance the fidelity and precision of clinical audio recordings [32].

In this study, we propose an end-to-end low-cost and highly adaptable data pipeline for seamless audio capture required for AI-powered medical scribing systems, tailored to the clinical settings in LMIC with Bangladesh as a case study. This proposed pipeline aims to encompass both the linguistic and acoustic variation of the outpatient department, the inpatient department, and the emergency department. Moreover, our implementation protocol places high importance on the proper anonymization of the data and on the secure storage of the data on the premises, which is in line with local regulatory rules and ethical best practices. Our efforts emphasize the specific needs and opportunities of LMICs, with the goal of closing the digital divide in healthcare documentation and enabling context-specific and scalable solutions.

# 2 Methods

Developing an AI-powered medical scribing system requires retrieval of doctor-patient consultation audio data from clinical settings with proper medical safety protocols and ethical approvals from respective authorities (see section 2.1). Upon securing ethical approvals and ideal clinical settings, we explored a wide range of recording configurations to identify the appropriate setup (speaker distance, recording device, speaker's positioning) as part of requirement analysis and collected the doctor-patient conversation data by following an optimized consultation protocols. The raw audio data were then compressed and anonymized to protect sensitive patient's identifiable information and later stored in our encrypted and secured AIMS Lab data server, alongside their clinical metadata. Before using these data for model training, trained annotator groups and domain experts are required to curate our anonymized audio data, as demonstrated in Figure 1. Therefore, all methods were performed in accordance with the relevant guidelines and regulations.

## 2.1 Ethical Clearance and Approval for Clinical Data Collection

The development of an AI-powered medical scribing system requires a large corpus of clinical data for model training obtained through an optimized data collection protocol, which was meticulously designed to meet ethical guidelines for healthcare research. We acquired ethical and operational approvals for data collection from regulatory bodies, including the Bangladesh Medical Research Council (BMRC), the Directorate General of Health Services (DGHS), the Institutional Research Ethics Board (IREB) of United International University (UIU) and CMED Health Ltd. Our clinical partner, including Aamader Susastho under CMED Health Ltd., served as the initial clinical settings for our audio data collection, covering a diverse range of patient demographics and expert physicians from various domains, which will be followed by a future expansion of data collection in Aalo clinics.

The study was statistically designed to include a large number of patients of diverse demographics (n = 18,000). The protocol deliberately excluded very vulnerable populations to ensure the highest ethical standards in respect of patient autonomy. To be specific, patients who were legally incompetent or

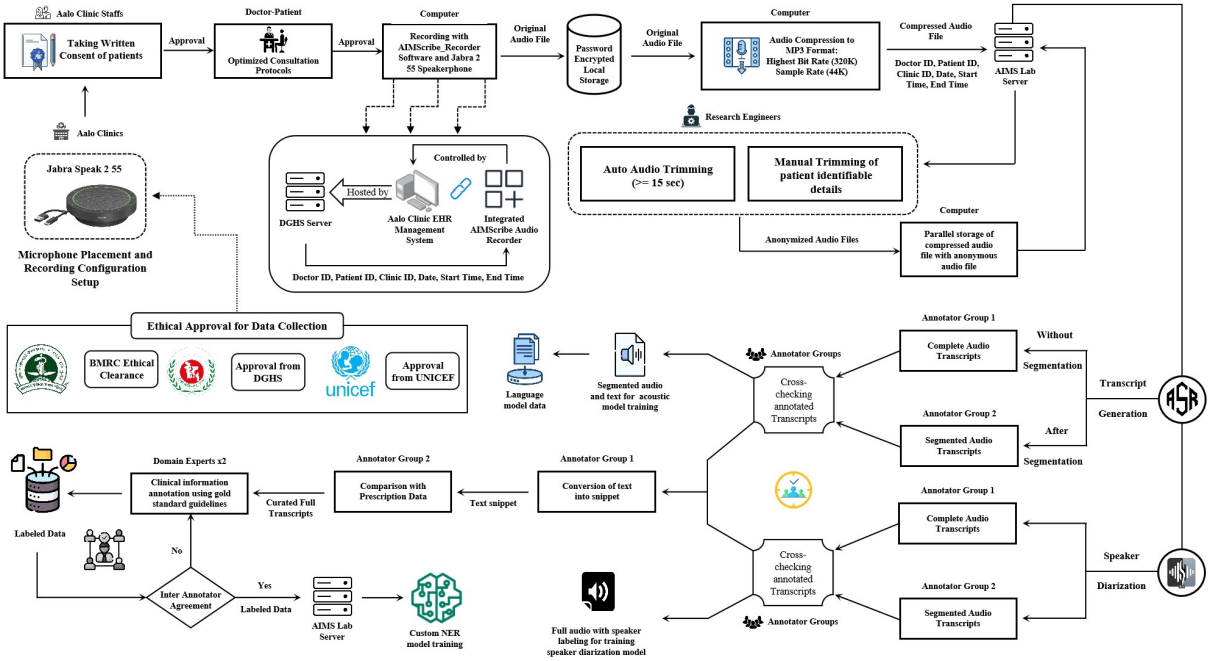


**Fig. 1** Our proposed pipeline for clinical audio data collection and requirement analysis for developing AI-powered medical scribing system.

possess a certain degree of cognitive disabilities or questionable mental aptitude were excluded from the participating cohort of subjects to ensure that no data were collected from patients incapable of fully understanding the scope of the study. Written informed consent was waived by the Institutional Research Ethics Board (IREB) of United International University (UIU). All participants were verbally communicated about the extent of withdrawing participation from data sharing at any point in the consultation while ensuring standard healthcare delivery during the consultation.

In this study, each doctor, patient and clinic were assigned unique cryptographic identifiers. All information will be preserved through a highly secure anonymization technique via de-identification rules according to HIPAA compliance and encrypted using cryptographic security. Considering clinical NLP and voice biometrics as an inherent threat to privacy, an extremely secure two-step data anonymization pipeline has been designed for data security. Physicians were asked not to request highly personal information irrelevant to the consultation and all electronic records were immediately discarded. The audio files were then moved to a secure database, which can be accessible by the research team for secondary sanitization, after recordings during the consultation periods (5 to 10 minutes). This included a manual acoustic screening to permanently drop any personally identifiable information (PII), that may have been accidentally revealed such as names, geographical addresses, etc. Subsequently, it added a probabilistic noise algorithm to the acoustic streams, masking voice biometric information and making unauthorized voice recognition mathematically infeasible. Unanonymized primary audio files and hardcopy consent forms were encrypted on local hard drives and securely kept in a locked cabinet that will only be accessible by keyed access to core researchers. All subsequent machine learning tasks utilized solely the fully anonymized corpus. The study was entirely non-invasive, involving no collection of biological specimens from the participants, thereby eliminating any physical risks. To ethically acknowledge the participants' time, a nominal honorarium were provided after the visit.

## 2.2 Exploration of Recording Devices

We explored a series of different low cost microphones, from conventional condenser microphones to modern plug-and-play USB microphones (see Table 1), so that they can be adaptable in low-resource clinical settings of LMICs.

**Table 1**: Technical specifications and configurations of selected microphones and speakerphones that are explored in terms of data collection for medical scribing

| **Metric** | AKG P120 | Fantech Leviosa MCX01 | Boya BY-MM1+ | Mirfak N4 | Rode PodMic USB | Jabra Speak2 55 |
|---|---|---|---|---|---|---|
| **Polar Pattern** | Cardioid | Cardioid | Super-Cardioid | Dual Cardioid (Front/Rear) | Cardioid | 4-Mic Beam-forming (Omni) |
| **Frequency Response** | 20 Hz – 20 kHz | 20 Hz – 20 kHz | 20 Hz – 18 kHz | 20 Hz – 20 kHz | 20 Hz – 20 kHz | 150 Hz – 7 kHz (Mic) |
| **Sensitivity** | 24 mV/Pa | -38 dB ±3 dB (12.6 mV/Pa) | -36 dB ±3 dB (15.8 mV/Pa) | -42 dB ±2 dB (Front/Rear) | 1.4 mV/Pa (-57 dB) | Not Specified |
| **Max SPL** | 150 dB (with -20 dB pad) | 110 dB | 120 dB | Not Specified | 148 dB | 85 dBspl @ 0.5m |
| **Self-Noise** | 19 dB-A | Not Specified | Not Specified | >60 dB (Signal/Noise Ratio) | 19 dB-A (Digital) / 26 dB-A (Analog) | Not Specified |
| **SNR** | 75 dB-A | 85 dB | 78 dB | 58 dB | Not Specified | Not Specified |
| **Power Requirements** | 48V Phantom | USB Bus Power (5V) | Plug-in Power (3.5mm) | USB Bus Power | USB/XLR (No Phantom Required) | Internal Battery / USB Bus Power |
| **Connectivity** | XLR | USB-C | 3.5mm TRRS/TRS | 3.5mm plug-in power | XLR/USB-C | USB-C/A, Bluetooth 5.1 |
| **Bit Depth / Sample Rate** | N/A (Analog) | 24-bit/192 kHz | N/A (Analog) | Not Specified | 24-bit/48 kHz | Up to 16-bit / 48 kHz |
| **Weight** | 453 g | 500 g | 60 g | 37g | 900 g | 280 g |
| **Special Features** | -20 dB Pad, 300 Hz Bass-Cut Filter | RGB Lighting, Pop Filter | Detachable Shock Mount | Switchable Bi-Directional Recording | APHEX® DSP, Dual XLR/USB Output | Full Duplex, AEC, Voice Level Normalization |

Among low-cost microphones with robust recording capabilities and clinical adaptability, the AKG P120 is the suitable analog option with cardioid polar pattern, which requires 48V phantom power via XLR. However, another cardioid microphone Fantech Leviosa, along with the dual cardioid Mirfak N4 ensures smooth digital integration with the use of USB-C. In case of a hybrid approach to support both analog and digital workflows, the Rode PodMic USB cardioid microphone can be an ideal choice for recording flexibility in low-resource settings, despite being heavier compared to other microphones listed here. The super-cardioid BOYA BY-MM1+ is one of the most lightweight mobile-first category microphones, which uses 3.5mm TRRS/TRS connection. Although it being a speakerphone, we included the portable Jabra Speak2 55 for our recording purpose, as it comes with four beamforming microphones, allowing it to capture sound from all directions during clinical conversations between multiple speakers. All six microphones are shown in Figure 2.

In terms of acoustic performance in handling background noise and external distortion, the AKG P120 and Rode PodMic USB can be considered as the powerhouses among the selected microphones. Their very high Max SPL limit of 150 dB and 148 dB respectively, make them capable of very loud sources without distortion. Despite having the lowest Max SPL limit (110 dB), the Fantech Leviosa comes up

with the best Signal-to-Noise ratio (85dB), indicating its suitability as the recording device to be quiet and controlled for voice-overs. The specific features of each device refer to their respective applications. The AKG P120 is made to be precise and has a 300 Hz Bass-Cut Filter to remove low-end rumble, making it ideal for medical transcription. The Rode PodMic USB microphone uses an internal APHEX® DSP to digitally process audio before it even gets to the computer. Meanwhile, Mirfak N4 and Boya BY-MM1+ are dedicated to spatial flexibility. While Mirfak's switchable Bi-directional recording is perfect for interviews, Boya's Super-Cardioid pattern is ideal for minimizing side noise in mobile applications. Unlike the AKG P120 or Rode PodMic speakerphones, which are built for high fidelity isolation (Cardioid) capture, the Jabra Speak2 55 speakerphone is a speakerphone aimed at group capture. Its 4-mic array enables it to capture speeches from multiple speakers in a room. In addition to being a light-weight option for recording with excellent portability, its frequency response is comparatively narrower within 150 Hz to 7 kHz, helping minimize rumbles at the low end and high end of the frequency spectrum to ensure clarity in speaker's speeches.




**Fig. 2** Types of microphones explored for the ideal configuration and setup design for clinical audio recording

## 2.3 Acoustic Environment Characteristics in LMIC Settings

Similarly to the choice of recording device, the acoustic environment of surrounding clinical settings, along with its characteristics, plays a crucial role in robust clinical audio recording. A successful AI scribing system depends not only on the capture of clinical audio but also collecting audio in a manner that is seamless to the clinical workflow and resilient to the acoustic challenges prevalent in many LMIC settings [33, 34]. However, the acoustic reality of many clinical settings, particularly in LMICs such as Bangladesh, poses a formidable challenge to the effectiveness of ambient listening, which are often characterized by stationary noises from fans/air conditioning, medical equipment, elevated background noise from sideway conversations, outdoor sounds, and greater sound reverberation in empty rooms with rigid wall surfaces [17, 33]. Such dominant noises and sound reverberation significantly affect the speech clarity of the participating doctor and patients, leading to poor signal-to-noise ratio (SNR) and degraded ASR systems [15]. Additionally, higher levels of background noise and reverberation negatively impacts the performance of the speaker recognition model as well as overall speech intelligibility, which in turn makes it too difficult for the diarization model to differentiate between the speeches of doctor, patient and the attendants [25]. Even in telemedicine applications in low-resource settings, handling noise to ensure speech clarity remains a crucial consideration for acoustic management [35]. Therefore, we specifically considered these internal and external factors in case of designing robust recording configuration and setups to mitigate these at the source-level for our clinical audio recording.

## 2.4 Recording Configurations and Microphone Placement Study

Strategic hardware integration and placement are essential for optimizing recording setups in these challenging environments. Successful configurations aim to minimize the distance between the sound source (clinician's mouth) and the microphone, which helps to increase the SNR and reduce the impact of ambient noise. Some implementations involve integrating microphones directly into exam room computers, providing a fixed, optimized recording point. Others, like VivaLyn Labs' MedScribe, are designed

for secure, on-premise deployment within a hospital's own network, which keeps sensitive patient audio and transcripts off-site and can be configured to run entirely within the hospital's IT infrastructure [31]. User-generated content from online forums provides anecdotal evidence of clinicians experimenting with room layouts and microphone positioning to maximize audio quality, suggesting a need for flexible and adaptable setup guides. For instance, placing microphones away from air vents or noisy equipment and using directional microphones to focus on the center of the room are simple but effective techniques to improve audio capture [32].

### 2.4.1 Single Microphone Testing between two speakers

Additionally, research into single-microphone speaker separation networks with time frequency attention shows promise for isolating voices in noisy and reverberant conditions using a single device, potentially simplifying hardware requirements [36]. For LMICs, where building acoustically treated rooms is not feasible, these technological solutions whether hardware-based (like better microphones) or software-based (like noise filtering) are critical for making ambient scribing viable.
In this scenario, a single microphone was placed between the doctor and the patient. The recording of doctor-patient conversations was collected in two separate seating configurations and three placement variations of the microphone, which was proven to be convenient in a clinical setting.

- **Equidistant Central**: In this single microphone setup, a microphone was placed in equidistance from both the doctor and the patient, where the recordings were acquired keeping both speakers at the right angle (90 degrees) or face-to-face (180 degrees). The microphone was placed on the surface of the table to cover the acoustic range of the two speakers. Except for heavy-weight microphones (AKG P120 condenser microphone and Rode PodMic USB microphone) that require the use of a boom arm for stable positioning, all other four microphones were tested in this setup.
- **Elevated Boom**: An elevated boom placement typically refers to a specific recording operational setup similar to the equidistant central, while the microphone was precisely positioned 40 centimeters above the table in our clinical environment [37]. Individual horizontal distances between the microphone and doctor-patient remained the same, regardless of their seating configurations (90-degree or 180-degree). Among the six microphones, only the AKG P120 and RodePodMic microphone was used in this setup due to their default structural setup.
- **Boundary Layer**: While the angle between both speakers was interchangeably shifted between 90 degrees to 180 degrees in this testing scenario, the placement of the microphone varied according to the seating arrangements. Although the microphone was mounted onto the surface, it was placed between the speakers according to their convenience, mostly prioritizing the patient as the primary speaker. Similarly to the equidistant central setup, these microphones were also used for audio recordings in this setup including the Fantech Leviosa MCX03, Boya BY-MM1+, Mirfak N4 and Jabra 2 55 SpeakerPhone for its surface-bounded setup.

### 2.4.2 Dual Microphones Testing between multiple speakers

For all dual-microphone setups, we exclusively used two separate AKG P120 microphones to record conversations between the doctor, patient, and an attendant (if present during the consultation) due to its specific structural and acoustic advantages for multi-speaker applications over other available choices. In contrast to the conflicts of digital clocking and latency issues often found in terms of using multiple USB microphones at once, AKG's default analog XLR input seamlessly allows for multi-channel, phase-aligned recording of all audio channels through a single, centralized audio interface. In addition, its precise cardioid polar pattern ensures the high-fidelity off-axis rejection often required for suppressing the cross-talk in spatially separated configurations such as the cross-cardioid settings and ORTF arrays. With a 300 Hz hardware-level bass-cut filter, the AKG P120 actively minimizes low-frequency ambient noise in the clinical setting. Unlike heavier options such as Rode PodMic USB (900 g), comparatively lighter AKG allows stability for secure positioning on boom arms in complex triangulated setups without sagging. Besides, AKG can handle sudden louder noises due to its max SPL level of 150 dB without any sort of distortions during multi-speaker recordings at busy clinical settings.

- **Dual Microphones for Direct Individuals (Two Speakers):** In scenarios limited to two primary participants, typically a clinician and a patient, this source-specific recording strategy can be prioritized for dual-speaker interactions. Both the doctor and patient are assigned their individual microphones in their immediate near-field. One microphone is placed directly in front of the patient

while the other microphone is mirrored in front of the doctor. This one-to-one ratio of microphone-to-speaker allowed each vocal channel to be recorded with maximum isolation as possible, as opposed to a shared acoustic zone. The configuration of two microphones being placed in proximity to the doctor and patient enables high-fidelity capture of their clinical conversations. The system allows the corresponding speakers and their respective microphones to be closer to each other, thereby minimizing the possibility of crosstalk (audio leakage from the other speaker). Such proximity of the microphone placements captured the clinical conversations at a close range, which ensured that the audio signals remained dominant over other speakers' reflections and ambient noise from clinical settings, producing higher signal-to-noise (SNR) for each individual.

- **Dual Microphones in a Cross-Cardioid Setup (Two Speakers):** The Cross-Cardioid setup uses a dual-microphone array with the microphones mounted in the middle of the consultation table and arranged in an intersecting "X" configuration. This configuration ensures that the primary lobe of each microphone is directed toward one participant while the "null" or rejection zone is oriented toward the other. By combining the microphone's polar pattern and positioning the sound source at the null point of the microphone array, this setup maximizes highest possible audio isolation between the participants. Additionally, this setup takes advantage of the off-axis rejection of the cardioid microphones to improve speaker rejection in a two-party clinical conversation. To evaluate the optimal setting for signal gain and inter-channel isolation in a clinical environment, the configuration was tested with different distances between the clinician and the mic (30–50 cm) and between the clinician and the patient (50–70 cm). The crossing of microphones helps to reduce crosstalk in high accuracy automated speech recognition (ASR) applications. Thus, the cross-cardioid placement is mainly dependent on the directionality of the microphones, and therefore can be more effective in areas with high acoustically reflective walls and ceilings, such as small rooms.
- **ORTF Inspired Dual Microphones Setup (Two Speakers):** This configuration utilizes a Dual Microphone array, where both microphones are located in the center of the table and arranged in a 110° angular displacement. The primary lobes of these two microphones point away from each other from the clinician and patient, so as to provide strong spatial cues for speaker diarization. It is inspired by the stereo techniques used in French television and radio broadcasting company, Office de Radiodiffusion Télévision Française (ORTF) for emulating the spatial perception of human hearing. This configuration is designed to deliver a coherent and naturalistic stereo soundstage of the clinical consultation, with the combination of both inter-channel intensity and arrival-time differences. The acoustic field between the two participants is balanced here with the aim of recording the dialogue with high fidelity and realistic spatial depth, crucial for the accuracy of subsequent automated speech processing and transcription.
- **Dual Microphones for Zone Coverage (Three Speakers)**: Compared to other alternatives, this setup turns out to be a more controlled one in which the two microphones are centrally placed at the right angle (one toward the doctor and the other towards the patient and attendant) to capture different voices and isolate the doctor's voice from the collective audio recordings of others. The key goal of this setup is to capture the maximum acoustic zone of audio recordings with as much clarity and vocal distinction.
- **Triangulated Dual Microphones (Three Speakers)**: With a less proximity bias and an enhanced ambient context, this triangulated setup separates speakers into groups and captures the interaction space of their multi-directional conversation. The microphones were placed at strategic angles away from each other (near corners) to establish a balanced acoustic field to cover all three main participants. The doctor sits at the front of the consultation table, and the patient and attendant sit on the other side/corners. The two microphones are installed in the corners of the table (or angled towards the corners), creating a triangular recording area. The main aim of this arrangement is to ensure that all the participants remain in the range of the pickup zone of at least one microphone and to avoid "dead zones" that can exist with a centralized and linear arrangement. This approach uses wide-angle acoustic coverage while maintaining a balanced Signal-to-Noise Ratio (SNR) between the clinician, the patient, and the attendant. This technique can be beneficial in situations where participants will be moving or speaking from different angles during the consultation.
- **Priority-based Dual Microphones Placement (Three Speakers)**: This is a special configuration of the two microphones, specially designed to prioritize patient speech and ensure patient speech is captured as best as possible. Although it is structurally similar to the triangulated configuration, the spatial logic differs from that of the triangulated configuration. Here, the second microphone is placed with intentional bias toward the patient (a minimized distance between the patient and the microphone), which improved the signal-to-noise ratio (SNR) for the patient's particular voice

structure. In case of prioritizing the capture of the patient's description regarding their symptoms, history, and concerns, and to transcribe them with greater clinical precision and fidelity, this setup can serve as the ideal choice for clinical audio recordings, despite the fidelity of the other speakers being compromised.

## 2.5 Development of Clinical Audio Recording Software

### 2.5.1 Deployment Setting and Clinical Workflow Integration

To capture clinical audio seamlessly, we have developed a recording software (AIMScribe.exe) deployed at Amader Susastho Clinics in Dhaka, Bangladesh, and integrated it with their electronic health record (EHR) platform, as demonstrated in Figure 3. The software operates as a background system-tray process without an independent graphical interface, integrating beneath the EHR platform already in routine clinical use at each site; a consultation recording is initiated, paused, and terminated through the clinician's existing workflow actions within that platform, so that deployment required no dedicated staff training.

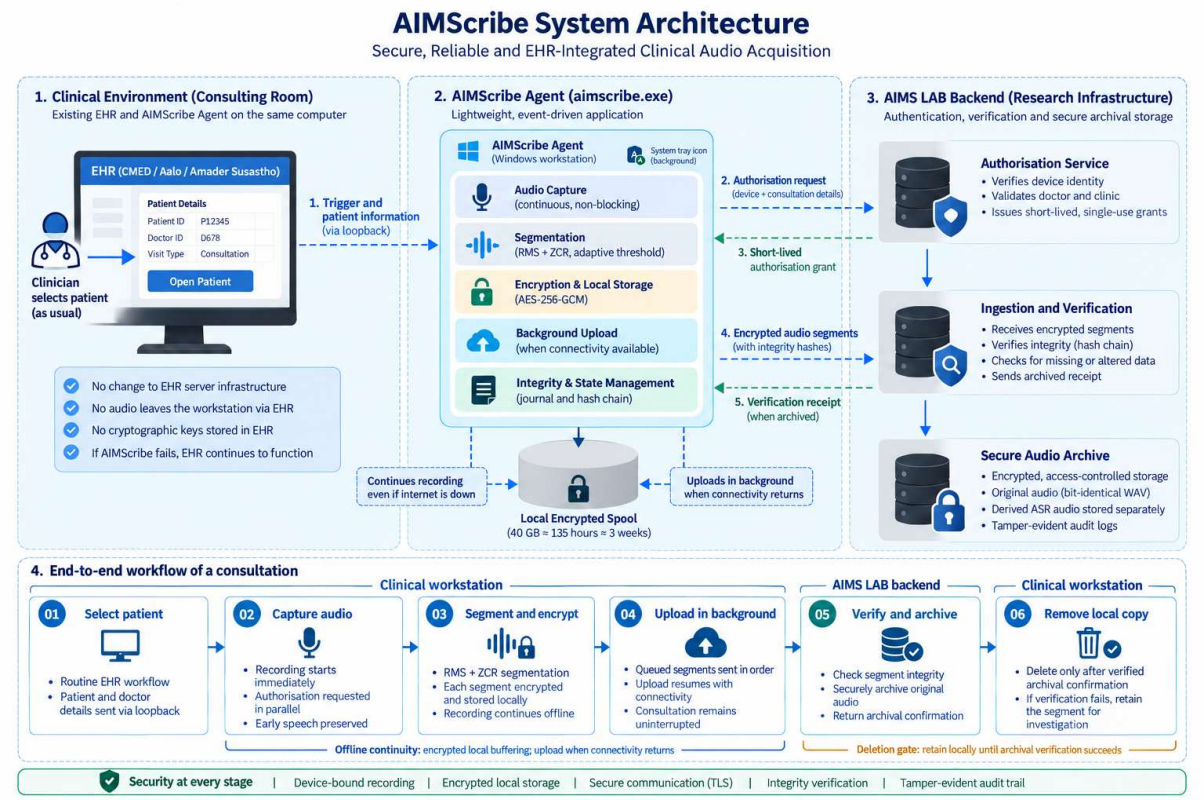


**Fig. 3** Our proposed pipeline of AIMScribe system architecture for secure clinical audio acquisition and EHR integration.

### 2.5.2 Consultation-Level Identity Attribution

Each recording is attributed to the specific clinician, hospital, patient, time, and date of that consultation. When a clinician begins a consultation, the EHR platform where the clinician is already logged in under their own account issues a short-lived, single-use, digitally signed authorization naming that clinician and patient. The acquisition agent will only record when this authorization is present, and it labels the resulting recording using exactly the identities the authorization contains. This matters because a single consulting-room computer is often shared across shifts; for example, one doctor in the morning and a different doctor in the afternoon. A system that identifies recordings by which computer made them cannot detect that handover. It would label the afternoon doctor's patients under the morning doctor's name because the machine's identity never changes. By attaching identity to each individually

authorized consultation instead of to the shared computer, this failure mode cannot occur, regardless of how many clinicians use a given device over the course of a day.

### 2.5.3 Data Protection and Integrity Assurance

Data protection was addressed through a security-by-design architecture. Audio resides in volatile memory only for the brief interval preceding local encryption; no unencrypted audio, and no identifiable session metadata, is ever committed to persistent storage at the point of capture. Each captured segment is encrypted at rest using AES-256-GCM under a device-bound key protected by the operating system's native key-protection facility, and all transmission to the central data-collection pipeline occurs over TLS (version 1.2 or higher) under a unique, per-device authentication credential. Data integrity is verified independently of transport security: the content hash of each audio segment is recomputed at three separate points along the pipeline before the segment is considered durably retained, and every recording session carries a cryptographically signed, sequentially chained record of its own lifecycle, such that any deletion, reordering, or substitution of a portion of a session's audio is independently detectable after the fact.

### 2.5.4 Acoustic Standardization and Voice-Activity Segmentation

To standardize input across acoustically heterogeneous clinical environments, the agent enforced a fixed capture format: single-channel, 16-bit pulse-code-modulated audio at 44.1 kHz and applied an adaptive voice-activity segmentation procedure in place of a fixed loudness threshold. The local noise floor is continuously re-estimated from recent short-interval energy minima, and the resulting silence threshold is further constrained relative to a running estimate of the speaker's own voice level so that boundaries are placed reliably within genuine pauses regardless of whether a given site's ambient noise floor was low or elevated. A secondary correction based on zero-crossing rate prevents low-energy unvoiced speech sounds from being misclassified as silence. Segments were nominally bounded between 30 and 60 seconds, yielding an acoustically normalized, silence-trimmed corpus without site-specific manual tuning.

### 2.5.5 Fault Tolerance, Local Buffering, and Reproducibility

Fault tolerance was addressed through a local encrypted store-and-forward buffer: should the network or central pipeline become unreachable, capture continues uninterrupted, with segments retained in encrypted form and transmitted automatically upon recovery using an exponential retry schedule. Local deletion is never triggered by transmission acknowledgement alone; a segment is removed from the clinical device only after the central pipeline returns a cryptographically signed proof that a verified, durable archival copy exists, followed by a further grace interval as a safeguard against downstream archival failure. This produces the governing operational invariant: the system degrades toward retaining audio under any failure condition, never toward discarding it. Within this design, previously sealed segments are unconditionally protected against an uncontrolled process or power interruption; audio captured after the most recently sealed segment but not yet sealed at the time of such an interruption is not recoverable under the currently deployed architecture, bounded by design to the segment ceiling described in 3.X.4. Despite these guarantees, the agent's measured steady-state memory footprint is approximately 8.3 MiB and is independent of consultation duration, and its real-time signal-analysis overhead is approximately 0.3% of a single CPU core, so as not to compete with concurrent clinical software on the same workstation. The complete source code for the acquisition module is publicly available to support methodological transparency and reproducibility, at AIMScribe.exe and AIMScribe_Backend_Render, respectively. 3.X.6 Planned Enhancement: Bounding Audio Loss on Uncontrolled Termination The residual risk described in 3.X.5 is bounded by the segmentation ceiling but not eliminated by it. To reduce this bound further without altering the segmentation, hash-chain, or encryption design described in 3.X.3 and 3.X.4, we plan to extend the acquisition agent with an intermediate durability checkpoint beneath the existing per-segment sealing step. Under this design, captured audio would be encrypted under the same per-device key and written to disk at a short, fixed interval, independent of the 30–60 second segment-boundary decision; final, chain-committed segments would then be constructed from these already-durable checkpoints rather than from an in-memory buffer, leaving the cryptographic structure of the segment-level hash chain, signature scheme, and encryption scheme entirely unchanged. On restart following an uncontrolled termination, the agent would reassemble whatever checkpoints survived and finalize a correspondingly shorter sealed segment from them, reducing the maximum audio at risk from

the current segment ceiling to the chosen checkpoint interval. This mechanism is a planned enhancement and is not part of the deployed system described in this manuscript. Its implementation, together with storage write-durability benchmarking on the deployed clinical hardware and empirical validation across repeated induced-termination trials, is intended before its description as an operational capability in future.

## 2.6 Optimized Doctor-Patient Consultation Protocol

Research study found that over half of clinicians surveyed felt that using an AI scribe altered the consultation dynamic, partly because they had to verbalize physical examination steps that were previously done silently [12]. By passively listening, the system aims to integrate into the background of the encounter, reducing the cognitive load associated with actively documenting information during the visit . This user-centered design principle is fundamental to achieving high clinician acceptance and sustained use [38]. To address real-world acoustic complexities and unstructured dialogue in Low- and Middle-Income Country (LMIC) clinical environments, we developed a standardized, six-step consultation protocol optimized for AI-powered ambient medical scribing. This framework systematically mitigates background noise, attendant interruptions, and conversational ambiguities typical of high-volume outpatient clinics. The consultation initiates with Preparation and Greetings, requiring clinicians to review preliminary vital signs and establish a private acoustic environment. During Chief Complaint Inquiry, physicians utilize an open-ended approach, permitting uninterrupted patient narrative before explicitly vocalizing a standardized summary of the primary symptoms to ensure accurate algorithmic capture. Subsequent History Taking strictly adheres to the OLDCARTS framework for present illness, sequentially addressing co-morbidities, family, and personal history.
To overcome the absence of integrated digital sensors for physical assessments, the protocol mandates verbalized running commentary during the Physical Examination, requiring clinicians to explicitly vocalize both positive and negative diagnostic findings. For Investigation and Probable Diagnosis, clinicians must verbally differentiate between provisional and definitive diagnoses while explicitly stating the name and clinical purpose of requested diagnostic tests. Finally, Medication Instruction and Advice enforces a rigid "5-Point" verbalization rule for every prescribed therapeutic: drug name, strength, frequency, duration, and specific instructions. In complex polypharmacy cases, this strict 5-point articulation is mandated primarily for essential medications, ensuring the precise acoustic capture of critical pharmacological data for natural language processing pipelines. Prior to our data collection in Aamader Susastho, we conducted focus group discussions and feedback session with healthcare experts and physicians to optimize it for large-scale adaptability. Consequently, we conducted hands-on training sessions for clinicians and participating doctors to ensure optimized data collection of high-fidelity clinical recordings required for the AIMScribe.

## 2.7 Our Clinical Study Design and Implementation

We implemented our clinical audio capture pipeline, within a live and active outpatient consultation room (dimension: 8ft*10ft) that mimics the standard real-world clinical examination room to validate its ecological validity and clinical significance. Rather than implementing the pipeline in an artificially created ideal acoustic environment, we intentionally chose to preserve the naturally occurring ambient noise conditions and acoustic artifacts, including surrounding conversations, medical equipment, operational air conditioning systems, or static noise from ceiling fans. In addition, unexpected clinical breaks arising from assistant interference or door opening events were allowed during sessions to cope with real-world clinical scenarios. This deliberate design promotes a realistic simulation of the intricacies of medical speech, allowing the downstream testing of ASR and diarization models to be more closely aligned with the diverse real-world challenges of clinical communication.
Furthermore, the acoustic environment of our clinical setting was characterized by a diverse range of transient workflow interruptions that are typically found in institutional healthcare delivery. Thereby, our audio capture pipeline explicitly considered different acoustic anomalies, such as unpredictable equipment alarm, telephonic ringings, adjacent side conversations, external corridor noise, etc. In a real-world clinical setting, the multi-speaker spatial configuration appears where the attendants (typically family members or spouses) sit within a spatial range of 1.5 to 2m alongside the patient. This multi-speaker configuration produces competing speeches and multiple overlapping acoustic fields at various angles, requiring a robust audio capture pipeline to distinctively recognize the verbal output of multiple speakers. Hence, the Jabra Speak2 55 speakerphone was chosen as the ideal recording device to overcome these

environmental and spatial challenges while preserving the natural conversation flow of the doctor and patients through its following capabilities:

- **Microphone Array & Beamforming Capabilities:** The Jabra 2 55 Speakerphone has an advanced multi-microphone array, which is optimized for intelligent beamforming. This feature enables the system to set up accurate directionality optimized for spatial speaker separation, which is essential for separation of the speaker's vocal signatures from the clinician, patient, and attendant in a consultation room.
- **Acoustic Profiling & Onboard DSP:** Jabra's hardware has a very wide, balanced frequency response curve designed to respond to the very specific frequency range of the human voice. Being coupled with powerful onboard digital signal processing (DSP) algorithms such as hardware-noise-cancellation and automatic gain control (AGC), it can aggressively block out low frequency ambient noise (e.g., HVAC systems) and normalize disparate speaker amplitudes.
- **Workflow Feasibility & Connectivity:** Its plug-and-play USB connectivity seamlessly allows for the direct capture of uncompressed digital audio into the primary recording software (AIMScribe.exe), avoiding latency and signal degradation as is common with analog-to-digital conversions.
- **Medical-Grade Compliance:** The device is also designed to meet the electromagnetic compatibility (EMC) standards that are necessary for medical device environments, so that it can function without causing interference to sensitive diagnostic devices.

During the recording phase, we collected the raw doctor-patient consultation recording without implementing any noise reduction techniques in our data preprocessing pipeline so that we can preserve the natural ambient environment of the clinical settings in our training data and consequently accustom our trained ASR model with real-world clinical settings for high-fidelity transcript generation. Although we intend to integrate the noise reduction pipeline in our pipeline during the real-world deployment of our AI-powered medical scribing system, AIMScribe in clinical settings in future.
This implementation protocol combines an uncompromised real-world clinical testing space with a structurally optimized hardware profile to deliver a fail-safe balance between high fidelity signal capture and workflow integration.

## 2.8 Acoustic Feature Extraction and Analysis

### 2.8.1 Signal-to-Noise Ratio

Instead of calculating signal-to-noise ratio (SNR) by comparing the energy level of whole speech-region to that of paused-region, we explored an in-speech minimum statistics approach to estimate the SNR. Some of the recording devices used in this research study tend to attenuate the non-speech intervals with their default noise suppression techniques. Besides, any estimator trying to quantify the noise reference from paused-regions is automatically forced to measure the suppressor's gating depth instead of the acoustic noise floor required for calculating SNR. Since noise suppression techniques differ from one recording device to another, calculated SNR through energy level of paused-region are not feasible for cross-device comparison. Therefore, we obtained an five-steps SNR estimation approach, which includes the noise-floor calculated exclusively from speech-active regions (where the suppressor necessarily remain inactive), inspired from Martin et al. who preferred tracking noise-floor from spectral minima instead of paused-regions [39].

**Step 1 - Decomposition of time-frequency:** Using 75% overlap and 32 ms analysis windows, we performed a short-time Fourier Transform on our collected audio recordings from different devices to transform them to a power spectrogram:

$$S[b,t] = |\mathbf{STFT}\{y[n]\}|^2$$

where $b$ and $t$ respectively denotes frequency bands and time frames.

**Step 2 - Restriction to speech-active frames:** We used the Silero VAD, a neural voice activity detector to find the speech-active regions. Its sole purpose was not to contributing in estimating noise floor itself rather locating specific regions where the device's suppressor was open. The use of VAD yielded a binary frame mask $m$, where the spectrogram was kept restricted within these specific frames:

$$S_{sp} = S[:, m]$$

**Step 3 - Quantification of noise-floor per-band:** Within each frequency band, the noise floor is taken as the 10th percentile of power across all speech-active frames. Syllabic and inter-phonemic energy fluctuations within connected speech expose the underlying noise floor at these low quantiles:

$$P_n[b] = \mathbf{percentile}_{10}\left(S_{sp}[b,:]\right)$$

**Step 4 - Broadband integration:** We summed up all the band-wise noise estimates to derive broadband noise power. Also, the broadband speech-plus-noise power is taken as the median across all frames:

$$P_n = \sum_b P_n[b] \qquad P_s = \mathbf{median}_t\left(\sum_b S_{sp}[b,t]\right)$$

**Step 5 - SNR Ratio with additive-noise correction:** In the final step, the noise estimates were subtracted from the total speech power to calculate the final SNR, considering the speeches remain superimposed on noise within the speech-active frames:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\max(P_s - P_n, \epsilon)}{P_n}\right)$$

where $\epsilon$ served as a small constant, integrated to avoid any undefined outcome when the noise estimate approaches the total power. The default noise suppression technique within any recording devices could no longer contaminate the result due to the exclusion of paused-regions at step 2. We verified this with the application of a simulated noise gating to a single recording within the range of 0 to 70 dB , where the estimate varied only by 0.02 dB across the full range, compared to the 0.57 dB found in the case of using a pause-based estimator.

### 2.8.2 Speech-to-Reverberation Modulation Energy Ratio (SRMR)

Reverberation was quantified using the Speech-to-Reverberation Modulation Energy Ratio (SRMR), a non-intrusive measure requiring neither a clean reference signal nor a measured room impulse response [40]. This property is essential in the present setting, as clinical consultations admit no clean reference recording and impulse-response measurement was not performed during data collection.

SRMR operates in the modulation domain, exploiting the observation that reverberation smears the temporal envelope of speech and thereby redistributes energy from low to high modulation frequencies [40]. The computation proceeds in three stages.

***Step 1 - Auditory filterbank decomposition:***

The signal is decomposed by a gammatone filterbank of $J = 23$ channels spanning 125 Hz to half the sampling rate, emulating cochlear frequency selectivity. The temporal envelope $e_j[n]$ of each channel is obtained via the Hilbert transform:

$$e_j[n] = |h_j[n] + i\,\mathcal{H}\{h_j[n]\}| \tag{1}$$

where $h_j[n]$ is the output of the $j$-th gammatone channel and $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform.

***Step 2 - Modulation filterbank analysis:***

Each envelope is passed through a secondary bank of $K = 8$ modulation filters with centre frequencies logarithmically spaced from 4 Hz to 128 Hz. The modulation energy in acoustic channel $j$ and modulation band $k$ is

$$\mathcal{E}_{j,k} = \sum_n |e_{j,k}[n]|^2 \tag{2}$$

***Step 3 - Energy ratio:***

Clean speech concentrates modulation energy in the lower bands, near the 4 Hz syllabic rate, whereas reverberation transfers energy to higher modulation bands. SRMR is the ratio of energy in the first four modulation bands to that in the upper bands:

$$\mathrm{SRMR} = \frac{\sum_{k=1}^{4}\sum_{j=1}^{J}\mathcal{E}_{j,k}}{\sum_{k=5}^{K^*}\sum_{j=1}^{J}\mathcal{E}_{j,k}} \tag{3}$$

where $K^*$ is an adaptive upper bound determined from the signal's modulation spectrum. Higher values indicate less reverberation impact.

All recordings were resampled to a fixed 16 kHz analysis rate prior to computation. This is necessary because the gammatone and modulation filterbanks are sampling-rate dependent, and published SRMR values and interpretive thresholds are established at 16 kHz [40]; computing at the native rate would place values on an incomparable scale across devices with differing native sampling rates. The measure was obtained using the SRMRpy implementation.

Two interpretive considerations apply. Because the capture devices apply beamforming and dereverberation, SRMR characterises *residual* reverberation in the delivered signal rather than the reverberation of the recording room itself; the measure therefore indexes the effectiveness of the device's spatial filtering under deployment conditions. Additionally, SRMR responds to additive noise as well as reverberation [40], and is consequently interpreted jointly with the signal-to-noise estimate rather than in isolation.

### 2.8.3 Active Speech Level and Speech Activity Factor (ITU-T P.56)

Recording level was quantified using the active speech level defined in ITU-T Recommendation P.56, Method B [41]. Unlike root-mean-square level, which decreases as the proportion of silence in a recording increases, the active speech level is computed over speech-active portions only and is therefore insensitive to pause structure. This property makes it the appropriate reference for comparing recordings that differ in conversational density, and for the cross-device comparison undertaken here, where microphones differ in output sensitivity. The implementation follows the SV-P56 reference program of the ITU-T G.191 Software Tools Library [42].

The measurement proceeds in four stages.

***Step 1 - Envelope estimation.***

The input waveform $x[n]$, normalised to the range $[-1, 1]$, is full-wave rectified and passed through two cascaded first-order exponential smoothers, as specified in P.56 Process 2:

$$p[n] = (1-g)\,|x[n]| + g\,p[n-1], \qquad q[n] = (1-g)\,p[n] + g\,q[n-1] \tag{4}$$

with smoothing coefficient

$$g = \exp\!\left(\frac{-1}{f_s\,\tau}\right), \qquad \tau = 30 \text{ ms} \tag{5}$$

The cascaded structure is required by the standard; a single smoother produces a faster-responding envelope that admits short transients the specification excludes.

***Step 2 - Threshold ladder and activity counting.***

A fixed bank of fifteen thresholds is applied, spaced in octaves relative to digital full scale:

$$c_j = 2^{-j}, \qquad j = 1, 2, \ldots, 15 \tag{6}$$

For each threshold, the number of active samples $a_j$ is counted as the union of all intervals in which $q[n] \geq c_j$, each extended by a hangover of $H = 200$ ms. Overlapping or adjacent intervals are merged before counting, so that samples falling within the hangover of multiple crossings are counted once:

$$a_j = \left| \bigcup_k \left[ n_k^{\text{start}},\ \min\!\left(N-1,\ n_k^{\text{end}} + H f_s\right) \right] \right| \tag{7}$$

where $\left[n_k^{\text{start}}, n_k^{\text{end}}\right]$ indexes the $k$-th contiguous above-threshold run.

***Step 3 - Level candidates and margin crossing.***

With total signal energy $E = \sum_n x^2[n]$, each threshold yields a candidate active level and a corresponding threshold level:

$$L_j = 10 \log_{10}\left(\frac{E}{a_j}\right), \qquad C_j = 20 \log_{10}(c_j), \qquad \Delta_j = L_j - C_j \tag{8}$$

The standard defines the active level as the point at which the candidate level exceeds its threshold by a fixed margin $M = 15.9\,\text{dB}$. The algorithm locates the first index $j$ at which $\Delta_j \leq M$, bracketing the crossing between thresholds $j-1$ and $j$, and applies the binary interpolation of the reference implementation [42] to converge on the level satisfying

$$L_{\text{ASL}} - C_{\text{interp}} = M \tag{9}$$

to a tolerance of 0.5 dB. Recordings for which no crossing exists, or for which $\Delta_1 < M$, fall outside the valid measurement range of the method and are excluded rather than assigned a value.

***Step 4 - Activity factor.***

The long-term level over the entire waveform of $N$ samples is

$$L_{\text{LT}} = 10 \log_{10}\left(\frac{E}{N}\right) \tag{10}$$

and the speech activity factor follows as the ratio of long-term to active energy:

$$\alpha = 10^{\,(L_{\text{LT}} - L_{\text{ASL}})/10}, \qquad \text{activity} = 100\,\alpha\ \% \tag{11}$$

The activity factor expresses the proportion of the recording classified as speech-active under the P.56 criterion, including the 200 ms hangover; therefore, it exceeds the proportion of strictly voiced material and should not be interpreted as a voice-activity detection rate.

The active speech level is reported in dBov relative to the digital overload point, with a full-scale sinusoid corresponding to $-3.01$ dBov. This digital level cannot be interpreted as an acoustic sound pressure level without calibrating the recording chain. Because no acoustic calibration was performed, the reported values are used to compare devices and recordings rather than to quantify absolute acoustic levels. The measure also provides a reference for normalizing audio segments prior to the prediction of perceptual quality. Preliminary testing over a 30 dB range confirmed that it accurately tracked changes in applied gain, allowing normalization to reduce the influence of differences in device output gain on perceptual quality comparisons.

### 2.8.4 Perceptual Quality Assessment (NISQA)

Perceptual speech quality was predicted using NISQA v2.0, a deep neural network trained to estimate mean opinion scores from a degraded signal alone, without reference to clean speech [43]. The model comprises a convolutional network extracting per-frame features from mel-spectrogram input, a self-attention module modeling temporal dependencies, and an attention-pooling layer producing a fixed-length utterance representation. Five independent regression heads yield an overall MOS prediction together with four interpretable quality dimensions, including Noisiness, Coloration, Discontinuity, and Loudness, each on the standard five-point MOS scale [43]. Three procedural adaptations were required for the present corpus.

***Segmentation:***

NISQA is validated on short speech samples and does not accept input of arbitrary length. Therefore, each recording was partitioned into non-overlapping windows of $T = 10\,\text{s}$. For a recording of $N$ samples at rate $f_s$, the segment set is

$$\mathcal{W} = \{y[a_i : b_i] \mid a_i = i \cdot T f_s,\ b_i = \min(N,\, a_i + T f_s),\ i = 0, \ldots, \lfloor N/(T f_s) \rfloor\} \tag{12}$$

Segments shorter than 2 s were discarded.

***Level normalization:***

NISQA does not normalise input level internally, since Loudness is among the dimensions it is designed to predict. Applied to unnormalised recordings, the model therefore reflects device output gain in addition to capture quality — an effect confirmed in preliminary analysis of the present corpus, where predicted MOS correlated with average sound pressure level at $r = 0.928$ and the Loudness dimension at $r = 0.954$. Each segment was accordingly scaled to a reference level of $-26$ dBov using the ITU-T P.56 Method B active speech level:

$$\hat{y}_i[n] = y_i[n] \cdot 10^{(L_{\text{target}} - L_{\text{ASL},i})/20} \tag{13}$$

where $L_{\text{ASL},i}$ is the active speech level of segment $i$ in dBov and $L_{\text{target}} = -26$ dBov. Where the resulting peak amplitude exceeded 0.99, the segment was rescaled to that ceiling to prevent clipping introduced by the normalisation itself.

***Occupancy gating:***

Windows dominated by silence yield depressed quality predictions that reflect the absence of speech rather than capture quality. Using the voice-activity segmentation, the speech occupancy of each window was computed as

$$\rho_i = \frac{1}{b_i - a_i} \sum_{s \in \mathcal{V}} \max(0,\ \min(b_i, s_{\text{end}}) - \max(a_i, s_{\text{start}})) \tag{14}$$

where $\mathcal{V}$ is the set of detected speech intervals. Windows with $\rho_i < 0.35$ were excluded.

***Aggregation:***

Retained segments were scored, and per-recording values reported as the median across segments, a statistic robust to individual anomalous windows:

$$\text{MOS} = \underset{i}{\text{median}}\,(\text{MOS}_i) \tag{15}$$

The across-segment standard deviation was retained as a measure of within-record perceptual stability. Model inference was executed as an isolated subprocess to prevent dependency conflicts with other components of the analysis pipeline. NISQA was trained predominantly on single-talker VoIP, telephony, and synthesised speech material; multi-speaker clinical conversation lies outside this distribution, and therefore absolute MOS values should be read as relative comparisons between devices rather than calibrated perceptual scores. Among the four dimensions, Coloration and Discontinuity are the most device-intrinsic indexing spectral distortion and temporal artifacts introduced by the capture chain and exhibited the lowest across-recording coefficients of variation in this corpus. Loudness retains sensitivity to the recording level even after normalisation and is reported for completeness rather than as a device-discriminating measure.

### 2.8.5 Speaker-Embedding Separability

Because the medical scribing system depends on correct attribution of utterances to speakers, capture devices were additionally evaluated on how well they preserve speaker-discriminative acoustic information. Two complementary measures were derived from speaker embeddings extracted per diarised turn: a between-to-within cosine-distance ratio, and the mean silhouette coefficient [44]. Both quantify cluster separation in embedding space without requiring reference speaker labels. The procedure comprises four stages.

***Step 1 - Diarisation and turn extraction:***

Speaker turns were obtained using the pyannote.audio diarisation pipeline [45], constrained to between two and three speakers to reflect the recording protocol. For multichannel recordings, the channel of highest root-mean-square energy was selected rather than downmixing, so that inter-channel phase differences could not degrade the embeddings. Turns shorter than 1.5 s were discarded, as speaker embeddings become unreliable below this duration; turns exceeding 8 s were centre-cropped to that limit. To prevent any single speaker from dominating the pairwise statistics, at most 25 turns per speaker were retained, selected at evenly spaced positions across the recording so that the sample represents the full session rather than a contiguous portion of it.

***Step 2 - Embedding extraction:***

Each retained turn was mean-centred, resampled to 16 kHz, and encoded using a pretrained ECAPA-TDNN speaker verification model [46] trained on VoxCeleb. Turns whose root-mean-square amplitude fell below $10^{-5}$ were excluded as effectively silent. Each embedding $\mathbf{e}_i \in \mathbb{R}^{192}$ was $L_2$-normalised:

$$\hat{\mathbf{e}}_i = \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|_2} \tag{16}$$

Recordings yielding fewer than two embeddings for any speaker, or fewer than six embeddings in total, were excluded from analysis as statistically insufficient.

***Step 3 - Pairwise cosine distances:***

For the normalised embedding matrix $\hat{E} \in \mathbb{R}^{N\times 192}$, the full pairwise cosine-distance matrix was computed as an inner product, exploiting the fact that $L_2$-normalisation reduces cosine distance to

$$D_{ij} = 1 - \hat{\mathbf{e}}_i^{\top}\hat{\mathbf{e}}_j \tag{17}$$

with values clipped to $[0, 2]$ and the diagonal set to zero. Letting $\ell_i$ denote the diarised speaker label of embedding $i$, the within- and between-speaker means are

$$\bar{d}_{\text{within}} = \frac{1}{|\mathcal{W}|}\sum_{(i,j)\in\mathcal{W}} D_{ij}, \qquad \bar{d}_{\text{between}} = \frac{1}{|\mathcal{B}|}\sum_{(i,j)\in\mathcal{B}} D_{ij} \tag{18}$$

where $\mathcal{W} = \{(i,j) : \ell_i = \ell_j,\ i \neq j\}$ and $\mathcal{B} = \{(i,j) : \ell_i \neq \ell_j\}$.

***Step 4 - Separability and silhouette:***

Separability is reported as the ratio of these means:

$$\text{Separability} = \frac{\bar{d}_{\text{between}}}{\bar{d}_{\text{within}}} \tag{19}$$

A value of unity indicates that embeddings of the same speaker are no more similar than embeddings of different speakers, that is, no speaker information is retained; increasing values indicate progressively cleaner separation.

The silhouette coefficient [44] provides a complementary, per-embedding view bounded to $[-1, 1]$. For each embedding $i$, let $a_i$ be its mean distance to all other embeddings of the same speaker, and $b_i$ the minimum, over all other speakers, of its mean distance to that speaker's embeddings:

$$a_i = \frac{1}{|C_{\ell_i}| - 1}\sum_{\substack{j\in C_{\ell_i}\\ j\neq i}} D_{ij}, \qquad b_i = \min_{c\neq\ell_i}\frac{1}{|C_c|}\sum_{j\in C_c} D_{ij} \tag{20}$$

where $C_c$ denotes the set of embeddings assigned to speaker $c$. The per-embedding coefficient and the reported score are then

$$s_i = \frac{b_i - a_i}{\max(a_i,\, b_i)}, \qquad S = \frac{1}{N}\sum_{i=1}^{N} s_i \tag{21}$$

with $s_i$ set to zero where the denominator vanishes. Values approaching unity indicate well-separated speaker clusters; values near zero indicate overlapping clusters; negative values indicate that embeddings lie closer to another speaker's cluster than to their own.

The two measures are reported jointly because they are sensitive to different failure modes. The ratio summarizes global cluster separation and is unbounded above, making it responsive to large between-speaker distances; the silhouette coefficient is bounded and per-embedding, and therefore reveals cases in which most turns cluster correctly while a minority are misassigned.

Two interpretive constraints apply. Speaker labels derive from automatic diarisation rather than ground truth, so both measures reflect the joint performance of the capture device and the diarisation pipeline; they are consequently interpreted as relative comparisons across devices under matched conditions rather than as absolute measures of embedding quality. Second, because the embedding model was trained predominantly on English-language VoxCeleb material [46], absolute values are expected to be compressed for the present corpus across all devices uniformly. Results are reported separately for two-speaker and three-speaker recordings, since the latter presents intrinsically greater potential for embedding overlap and pooling the conditions would confound device capability with task difficulty.

# 3 Results and Discussion

## 3.1 Acoustic performances across different recording setup and devices

### 3.1.1 Speech Quality and Intelligibility Analysis

This study validates the acoustic compatibility of commercial off-the-shelf microphones to be used for clinical audio recording within an AI-based medical scribing system in quiet and simulated loud settings (see Table 2). The main goal is to maximize transcript fidelity to ensure downstream precise extraction of clinical named entity recognition and accurate clinical documentation. For quantifying acoustic degradation in the different spatial configurations and recording setups, the relative performance shift of the Signal-to-Noise Ratio (SNR) or Speech-to-Reverberation Modulation Energy Ratio (SRMR) between simulated noise and quiet baselines (see Figure 4) was computed. The Jabra Speak2 55 Speakerphone is the best low-cost recording device choice for real-world use since it provides the best combination of omnidirectional speech capture, echo cancellation, excellent reverberation control and placement flexibility. The highest average SNR baselines were achieved by this device with 37.42 dB (quiet) and 38.44 dB (simulated) and up to 41.65 dB (3 speakers) in single microphone priority setups. However, its single microphone priority configuration for three speakers displayed SNR compression in simulated noisy environments. SRMR is highly sensitive to orientation in space. As in the 180 degree boundary layer and zone coverage configurations, Jabra 2 55 speakerphone has resulted in successful mitigation of reverberant artifacts and higher SRMR up to 13.03. However, improper 90° alignment can result in phase cancellation and intelligibility loss, specifically in the case of 90° boundary layer setup with a huge acoustic penalty of -3.94 (falling from 8.81 to 6.25). AKG P120 is ideal for studio applications in optimized multi-microphone configurations, but is highly prone to sound distortion caused by phase mismatch in stereo arrays, which should be acoustically calibrated to prevent phoneme distortion in automated medical scribing pipelines. The AKG P120 consistently performed well, with an average SNR 35.18 dB in quiet environment compared to 35.27 dB in simulated noisy environment, when tested across the advanced dual-microphone arrays and elevated boom setups. But at a spatial level, it had extremely varying performance, depending on configuration geometry. The dual zone coverage configuration for three speakers produced a fantastic reverberation suppression (SRMR is 15.91), while the Dual ORTF (180 degree) configuration had a dramatic drop in reverberation suppression (SRMR dropped from 9.36 to 5.84). Although it provides the best acoustic quality and spatial mapping, it takes up more space and requires higher booming heights for seamless deployment in physically limited facilities.

The Mirfak N4 exhibited a high equivalent SNR baseline in quiet environments from 35.64 dB to 39.39 dB, while on the boundary layer setup, it provided a SNR of 35.64 dB in the quiet environment. Its SNR resilience did not degrade under simulated noise (e.g., 35.97 dB in the 90 degree equidistant setup and 37.30 dB in the 180 degree boundary layer setup). In addition, modest positive shifts or stability in its SRMR were observed (up to 10.19 in the boundary layer setup). This means that the Mirfak N4 is capable of processing ambient chatter successfully, so it can be considered a reliable general-purpose system for two-speaker consultation, where acoustic positioning is somewhat controlled. The Fantech Leviosa was highly stable in terms of SNR under different spatial configurations, with a very stable mean SNR of 34.31 dB in a quiet environment and 34.52 dB in simulated noisy environments. Although its absolute SRMR in quiet rooms is small, it has a high acoustic enhancement in noisy rooms which makes it a suitable and affordable choice for two-speaker LMIC consultations with microphones set to balanced equidistant or boundary layer setups. It has potential as a viable cost-effective tertiary replacement option for just two speaker interactions, but it must be processed aggressively with downstream digital signal processing to remove errors in the separation of primary vocals.

Conversely, the BOYA microphone was found to be acoustically sensitive in certain spatial geometries. Its boundary layer configuration did not show a significant SNR change in the noisy environment (10.37 to 10.42 dB), while its 90 degree equidistant setup had a drastic change with its SNR dropping from 36.45 dB to 29.39 dB. This strong negative shift in SNR, as shown in Figure 4, is indicative of poor off-axis rejection. This is a problem in the context of an unconstrained, high-ambience reflection environment, such as a busy LMIC clinic, where the signal may be distorted and mispronounced, causing phonemes to be dropped and resulting in hallucination of clinical NER in subsequent downstream pipelines. The Rode PodMic was tested mainly in elevated booms, and the SNR was consistent under quiet and simulated conditions (approximately. 36.18 dB to 37.03 dB). In elevated boom configurations (180 degree) under a simulated noisy environment, its SRMR dropped to 5.96, which was a serious drawback. This steep drop suggests that the camera is sensitive to phase cancelation and room reverberation when recording multiple speakers from greater spatial angles. It is ideal for use in one-directional broadcast systems,

**Table 2** Calculation of estimated Signal-to-Noise Ratio (SNR) and Speech-to-Reverberation Modulation Energy Ratio (SRMR) across various recording configurations in both quiet and noisy environments.

| Recording Setup | Speaker Count | Microphone | Estimated SNR Values | | SRMR Values | |
|---|---|---|---|---|---|---|
| | | | Quiet | Noisy | Quiet | Noisy |
| Single Equidistant (90 Degree) | 2 | Fantech | 34.77 | 32.66 | 6.52 | 8.77 |
| | 2 | Mirfak N4 | 39.39 | 35.97 | 8.80 | 9.57 |
| | 2 | BOYA | 36.45 | 29.39 | 8.15 | 8.71 |
| | 2 | Jabra 2 55 Speakerphone | 37.45 | 35.88 | 9.66 | 5.72 |
| Single Equidistant (180 Degree) | 2 | Fantech | 34.22 | 35.64 | 7.52 | 11.82 |
| | 2 | Mirfak N4 | 35.64 | 36.66 | 7.91 | 8.90 |
| | 2 | BOYA | 34.08 | 33.9 | 8.25 | 7.75 |
| | 2 | Jabra 2 55 Speakerphone | 36.52 | 38.11 | 9.44 | 10.76 |
| Single Boundary Layer (90 Degree) | 2 | Fantech | 33.93 | 35.29 | 10.54 | 11.66 |
| | 2 | Mirfak N4 | 36.84 | 36.24 | 8.69 | 10.19 |
| | 2 | BOYA | 33.59 | 30.93 | 9.90 | 10.42 |
| | 2 | Jabra 2 55 Speakerphone | 38.89 | 35.42 | 8.81 | 6.25 |
| Single Boundary Layer (180 Degree) | 2 | Fantech | 34.32 | 34.49 | 7.93 | 8.83 |
| | 2 | Mirfak N4 | 39.19 | 37.3 | 8.82 | 8.88 |
| | 2 | BOYA | 33.58 | 30.57 | 9.31 | 10.37 |
| | 2 | Jabra 2 55 Speakerphone | 37.02 | 39.06 | 8.86 | 10.81 |
| Elevated Boom (90 Degree) | 2 | Rode PodMic | 36.4 | 36.93 | 8.07 | 8.83 |
| | 2 | AKG P120 | 33.78 | 37.36 | 10.20 | 11.47 |
| Elevated Boom (180 Degree) | 2 | Rode PodMic | 36.18 | 37.03 | 8.54 | 5.96 |
| | 2 | AKG P120 | 35.32 | 35.69 | 9.32 | 10.13 |
| Single Microphone - Zone Coverage | 3 | Jabra 2 55 Speakerphone | 37.14 | 40.52 | 9.83 | 13.03 |
| Single Microphone Priority | 3 | Jabra 2 55 Speakerphone | 37.5 | 41.65 | 9.53 | 11.50 |
| Dual Microphone Individual (180 Degree) | 2 | AKG P120 | 36.04 | 39.01 | 10.61 | 11.05 |
| Dual Microphone - Cross-Cardioid (180 Degree) | 2 | AKG P120 | 35.73 | 32.23 | 10.22 | 9.44 |
| Dual Microphone - ORTF (180 Degree) | 2 | AKG P120 | 35.75 | 32.59 | 9.36 | 5.84 |
| Dual Microphone (Zone Coverage) | 3 | AKG P120 | 39.08 | 40.05 | 13.30 | 15.91 |
| Dual Microphone (Triangulated) | 3 | AKG P120 | 31.3 | 28.45 | 12.71 | 13.97 |
| Dual Microphone (Priority) | 3 | AKG P120 | 34.44 | 36.76 | 14.54 | 15.08 |

but has poor directivity and reverberation sensitivity in multiple speaker angles, which limits its use in dynamic doctor-patient seating arrangements.

### 3.1.2 Acoustic Signal Preservation and Activity Profiling

As seen in Figure 5 the acoustic evaluation through ITU-T P.56 indicates significant differences in the preservation of the signal and gain stability of microphone setups at various ambient noise levels, with both microphone setups and devices exhibiting the differences. In terms of acoustic robustness, dedicated integrated speakerphones, specifically the Jabra Speak2 55, had optimal active speech levels ranging from -20 dBov to -25 dBov, and excellent speech activity percentages (65%–85%) in both quiet and noisy environments. Likewise, multi-capsule studio condenser configurations with double AKG P120 microphones maintained good gain characteristics (-19 dBov to -25 dBov). On the other hand, the budget dynamic and omnidirectional micro-array models (Fantech Leviosa and MirfaK N4) were severely attenuated at 180° off-axis angles (down to -52 dBov and -62 dBov respectively), and their speech activity fluctuated wildly with speech due to contamination from environmental noises. The Jabra Speak2 55 and AKG P120 dual-array configurations ensure optimal acoustic fidelity in a real-world clinical consultation environment where vertical and horizontal distances between the doctor and the patients vary.

### 3.1.3 Perceptual score analysis using NISQA Framework

The perceptual speech quality evaluation using the Non-Intrusive Speech Quality Assessment (NISQA) framework (a framework based on deep neural networks for the prediction of speech transmission quality under complex real world conditions), which has been found to be able to produce clear hierarchies of

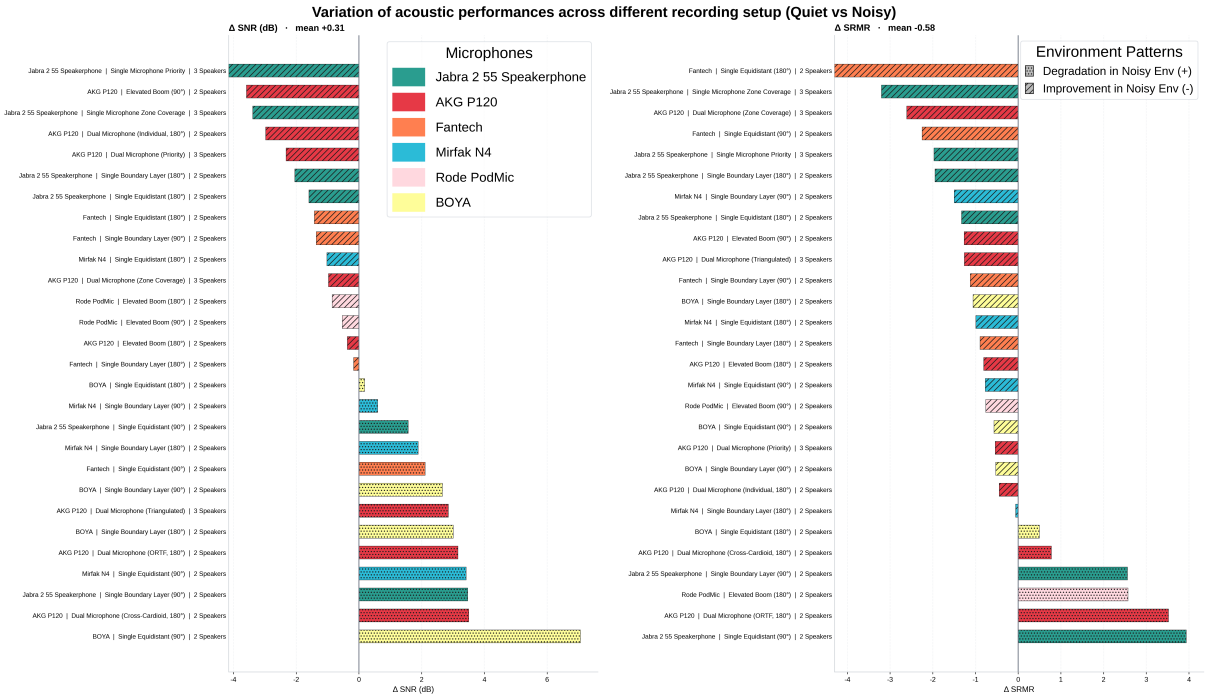


**Fig. 4** Acoustic degradation across different microphone devices and configuration regarding quiet and simulated clinical environment.

speech quality in terms of NISQA MOS, Noisiness, Coloration and Discontinuity metrics as illustrated in Figure 6. The Jabra Speak2 55 Speakerphone performed best on the perceptual speech preservation measures, delivering the best result in single microphone priority setup (MOS = 3.075, Noisiness = 4.191, Coloration = 3.520, Discontinuity = 3.577) and Zone Coverage (MOS = 2.945). Its integrated multi-element beamforming was able to effectively suppress ambient background chatter without introducing temporal gating or spectral distortion, which is important for the downstream clinical task of clinical speech recognition and named entity recognition (NER) while maintaining the clarity of the phonemes. The AKG P120 also demonstrated good spatial fidelity in dual cross-cardioid (180°) configuration (MOS = 2.790, Noisiness = 3.106), better than conventional elevated boom setup. The performance was, however, very poor in dual triangulated setup (MOS = 1.570, Discontinuity = 1.953) resulting in severe inter-capsule phase cancellation and spatial comb filtering. When the Rode PodMic was used in elevated boom (90°) configurations, it was reliable, with low discontinuity (3.314) and a reasonable MOS (2.680) despite the limited pickup angle for multi-speaker applications. In contrast, for 180° equidistant setups (MOS = 2.473), the Fantech Leviosa was fairly well, and for all noisy setups (MOS $<$ 1.800), the Mirfak N4 and BOYA were consistently poor with severe noise contamination and spectral coloration that could lead to severe hallucinations in ASR transcriptions.

### 3.1.4 Spatial Diarization and Speaker discrimination (Separability and Silhouette)

The importance of spatial transducer geometry is shown by evaluating the distinctness of speaker features in the downstream clinical cases of speech recognition and diarization. The dual-microphone setups (AKG P120 in cross-cardioid - 180° and priority configurations), and single microphone setup including Jabra Speak2 55 in Priority and Zone Coverage modes outperformed all other microphone configurations in quiet baseline conditions, producing peak speaker separability scores above 2.0 and silhouette scores between 0.48 and 0.51 (see Figure 7). Single-capsule budget microphones suffered a dramatic feature space collapse when exposed to ambient clinical acoustic noise, resulting in poor silhouette scores less than 0.20 due to noise-induced cluster overlap. In contrast, the Jabra Speak2 55 speakerphone and AKG P120 dual cross-cardioid configuration demonstrated better noise immunity, with separability scores of over 1.85 and a relatively low silhouette scores of over 0.42. As such, the Jabra Speak2 55 is the top single-unit solution for ambient deployment, and the dual AKG P120 Cross-Cardioid is the better multi-mic array for strong speaker attribution for ambient audio capture.

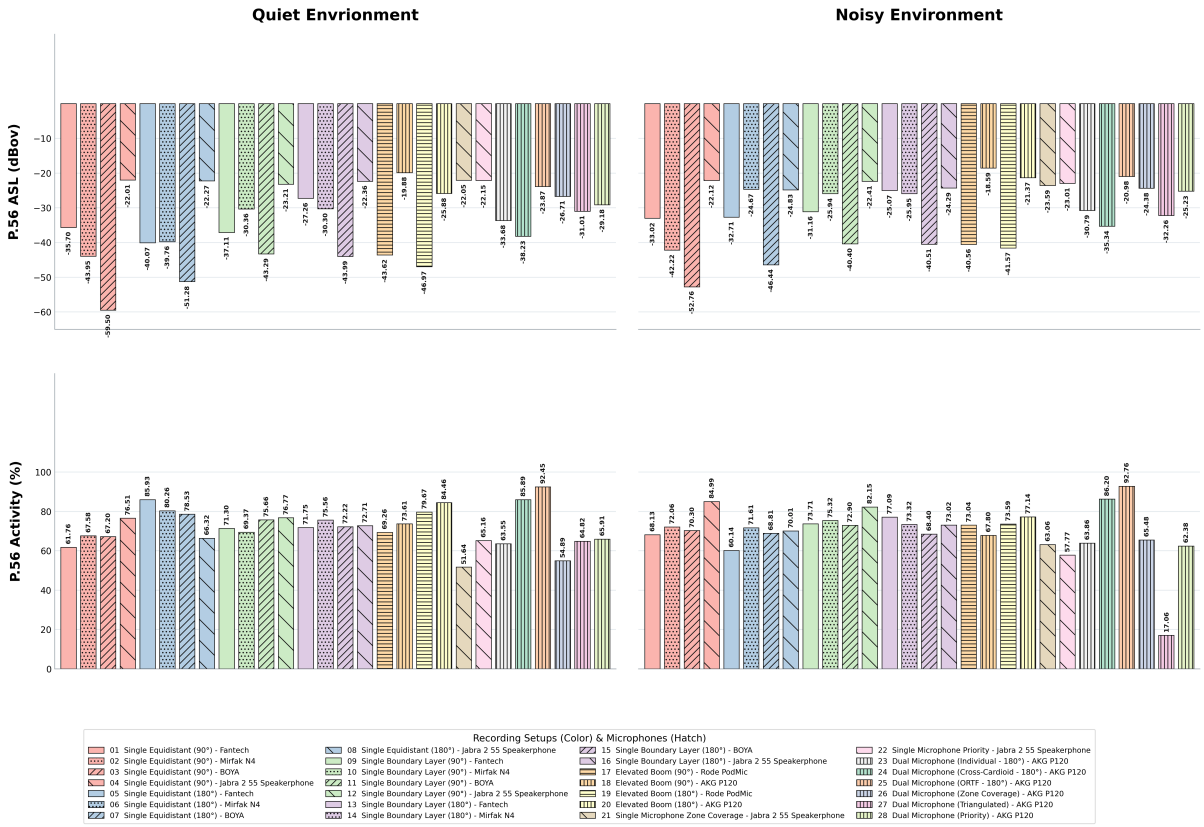


**Fig. 5** Demonstration of P.56 Active speech level and P.56 speech activity percentage across different recording setups with six different microphones in both quiet and noisy environments.

## 3.2 Doctor-patient's perception and feedback analysis

Between October and December, a total of 489 doctor-patient consultations were recorded at the clinic. Informed consent was obtained from all domain experts (physicians in this case), who participated in the evaluation and feedback analysis. When physicians asked the patient for verbal informed consent to record their consultation sessions, approximately 29.2% (143) of patients initially gave consent for recording (see Figure 8). However, 41 of these sessions had to be excluded from the final dataset because patients withdrew their consent mid-consultation. During the initial phase of data collection using the Microphone A (AKG), 29 complete consultation recordings were successfully collected. After that, a Jabra microphone was used for the recording procedure, which showed a greater patient consent rate. Interestingly, we found that when the Jabra microphone was introduced, a number of follow-up patients who had previously refused to be recorded using the AKG setup gave their consent. In the end, the team completed collection of 73 recordings using the Microphone B (Jabra) microphone during the data collection phase.

We evaluated the usability of two recording configurations using 5-point Likert scale surveys from both patients and participating physicians. The data reveals a clear difference in everyday user experience between the AKG (Microphone A) and Jabra (Microphone B) devices in terms of acceptance. The patient feedback indicates that the Jabra microphone setup is preferred on several usability measures. 75.3% (55 out of 73) of patients who were recorded with the Jabra microphone agreed or strongly agreed that they were comfortable with the microphone during their consultation. By contrast, when recorded on the AKG microphone, none of the 29 patients strongly agreed, 10.3% (3) agreed, and most (51.7%) were neutral or disagreed (37.9%). In terms of communication and privacy Jabra device was seen as less intrusive. The majority of Jabra users (68.5%) either agreed or strongly agreed that the location of their device allowed them to have open communication. Additionally, patients captured using the Jabra device showed less hesitation to share private health information. When asked whether they were satisfied with their current microphone setup and would be willing to change, the results showed that Jabra users were very satisfied with their current microphone, with 79.5% (58 out of 73) disagreeing or strongly disagreeing to change On the other hand, nearly half (48.3%) of the AKG users stated they would prefer to use a

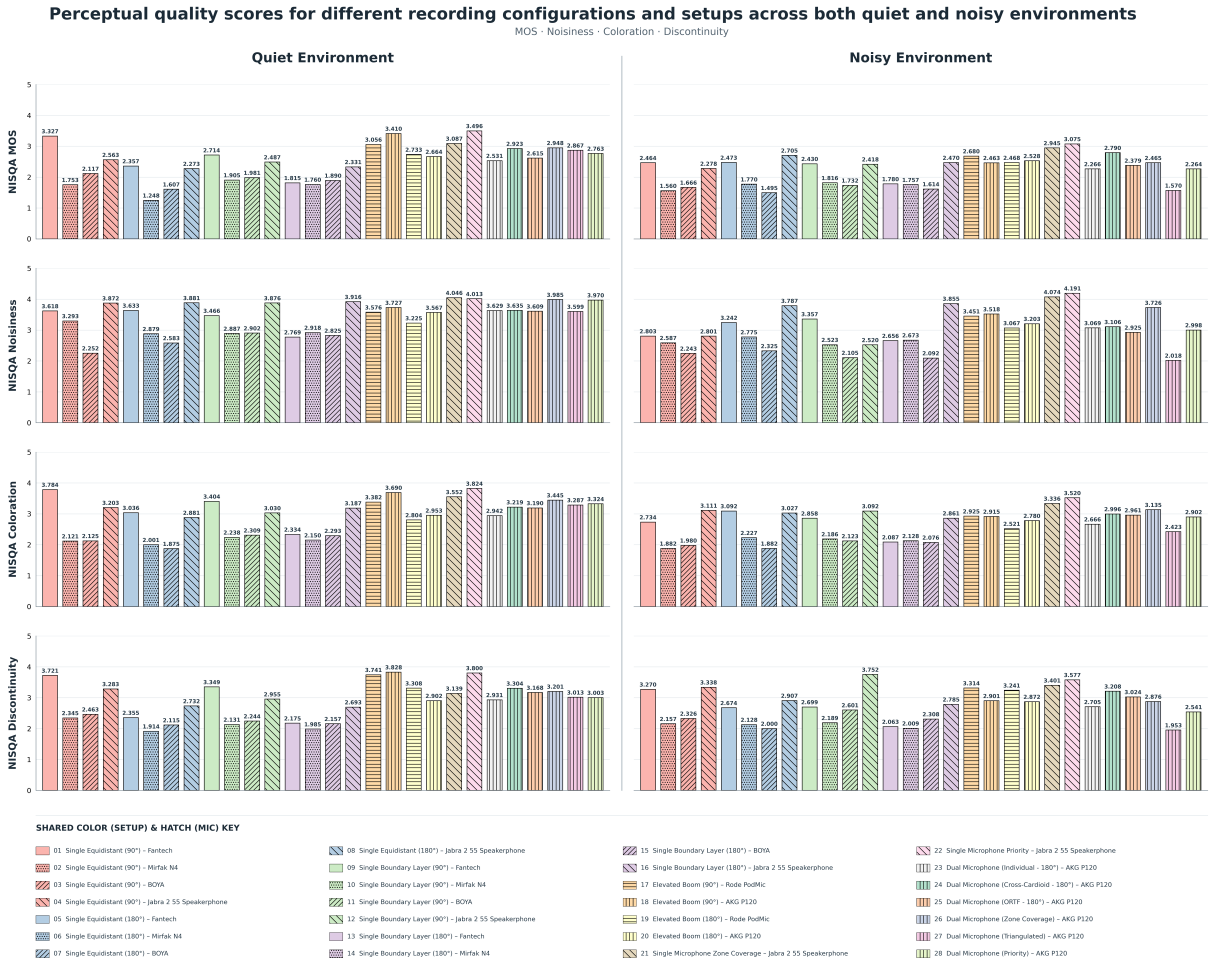


**Fig. 6** Distribution of perceptual scores including NISQA MOS, NISQA noisiness, NISQA coloration and NISQA discontinuity across different recording setups with six different microphones in both quiet and noisy environments.

different setup. The physicians who participated gave similar feedback to the patient data, with a strong preference for the Jabra microphone being integrated into the workflow. Physicians reported that the AKG microphone caused disruption with workflow. Four of six (66.7%) participants agreed or strongly agreed that the AKG system was distracting from the following of "normal procedures This distraction was especially reduced with the Jabra, as nearly all doctors either disagreed or did not say that this led difficulties to follow protocol. A primary operational obstacle faced by the physicians when using the AKG device was that the need for manual intervention while using it, with 100% of the physicians agreeing or strongly agreeing that this was a barrier. This problem was less with the Jabra device, as 5 of the 6 physicians disagreed or strongly disagreed that they needed to manually adjust it during the consultation. Physicians' comfort with the physical placement of the Jabra microphone (3 agreed, 2 neutral) versus the AKG microphone (3 strongly disagreed with the placement) was generally more favorable with the Jabra microphone. This result showed an indication that the physicians preferred different settings on the microphone when using the AKG, but this was not seen to the same extent with the Jabra device.

## 3.3 Analysis of Clinical Audio Recordings using Jabra 2 55 Speakerphone

Acoustic evaluations in 495 clinical consultation recordings demonstrate exceptional audio fidelity and background noise suppression when using the Jabra Speak2 55 speakerphone in LMIC clinical environments. The estimated Signal-to-Noise Ratio (SNR) distribution exhibits a robust mean of 32.84 dB (SD = 2.35 dB) and a median of 32.61 dB (interquartile range [IQR]: 31.30–33.86 dB), ranging from a minimum of 27.28 dB to a maximum of 42.97 dB, as listed in Table 3. The majority of collected recordings is identified as having high fidelity audio, with 80.0% (n = 396) falling in the ideal SNR tier, 12.9% (n = 64) in the excellent SNR tier see Figure 9). In contrast, only 7.1% of collected recordings were in

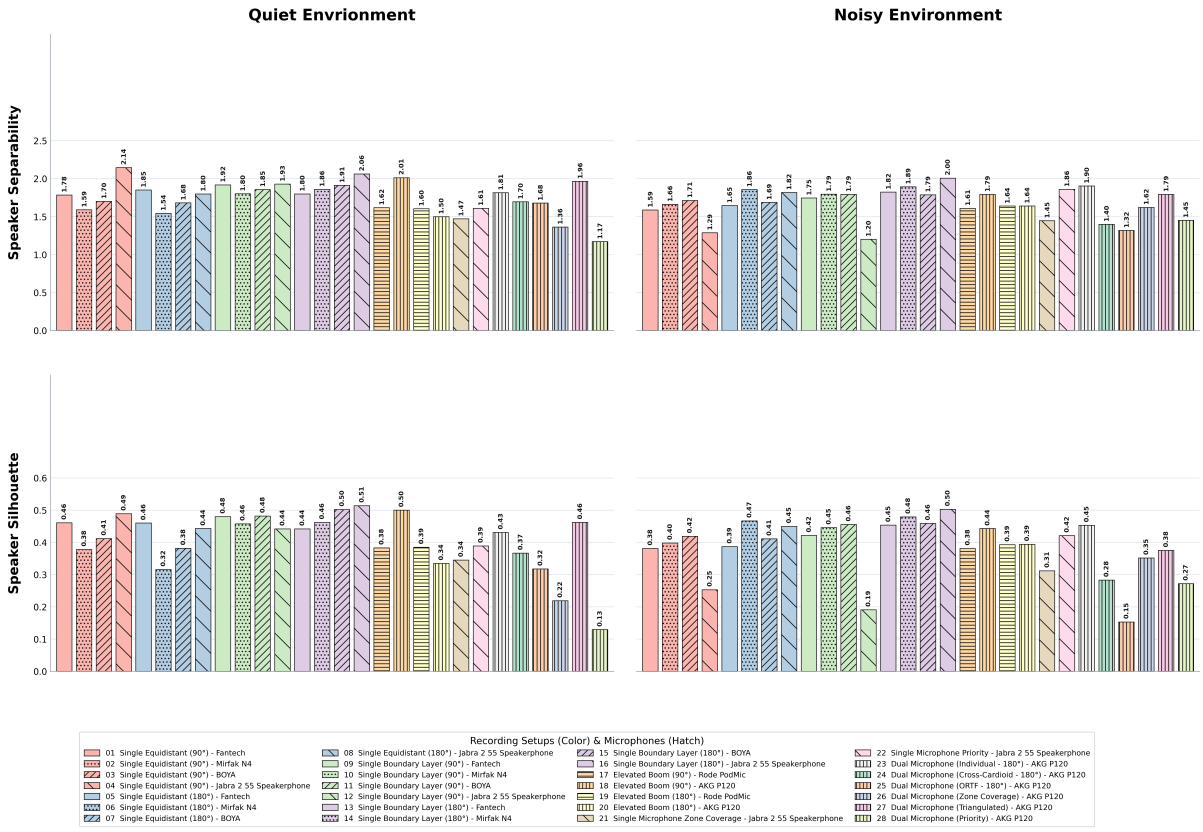


**Fig. 7** Distribution of speaker separability and speaker silhouette scores across different recording setups with six different microphones in both quiet and noisy environments.

reasonable tier, while no recordings in the poor SNR tier. These results validate the effectiveness of the directional microphone array and integrated hardware noise suppression, as downstream ASR engines in clinical NLP pipelines need to be able to achieve low WERs in the context of ambient noise in the consultation room, including background chatter and fan noise.

Using the SRMR to assess room acoustics indicates good dereverberation and preservation of spectral modulation envelopes [40]. The dataset yielded a mean SRMR of 5.19 (SD = 1.54) and a median of 4.63 (IQR: 4.02–6.16), spanning from 2.68 to 11.01. Tier composition shows distribution of clinical recordings including 13.7% in excellent (n = 68) tier, 30.1% in ideal (n = 149) tier, 56.0% in reasonable (n = 277) tier and only 0.2% Poor (n = 1). Larger SRMR values mean less reverberant energy smearing is present, enhancing the ability to retain phonetic boundary transitions and temporal speech structures that are critical to the accurate representation of speech in modern transformer-based conformer or Whisper ASR models. Absolutely an overwhelmingly favorable SNR and no Poor SRMR profiles of collected recordings show that the Jabra Speak2 55 delivers excellent front end speech capture without using expensive acoustic modifications to the room. This signal stability, which is achieved with hardware, is vital in LMICs, where the ambient noise level is often high and the architectural design lacks insulation and reflective properties. The empirical findings convincingly support the use of the Jabra Speak2 55 for large-scale clinical use, with the high preservation of the signal for downstream clinical entity recognition and automated medical scribes.

### 3.3.1 Perceptual Quality Scores

Based on clinical audio recordings obtained through the selected speakerphone, this analysis corroborates the speakerphone's strong acoustical capabilities to support AI-driven medical scribing in resource-limited LMIC clinical environments. We measured acoustic integrity in 495 clinical recordings using the NISQA framework, as demonstrated in Figure 10. The density distributions obtained show the uniform and intelligible speech capture ability. The overall predicted NISQA MOS is 2.110 with a standard deviation of 0.379, which is a very realistic baseline that is easily achievable in chaotic outpatient wards of LMICs with high ambient interference. Most importantly, the hardware has low variance and normal

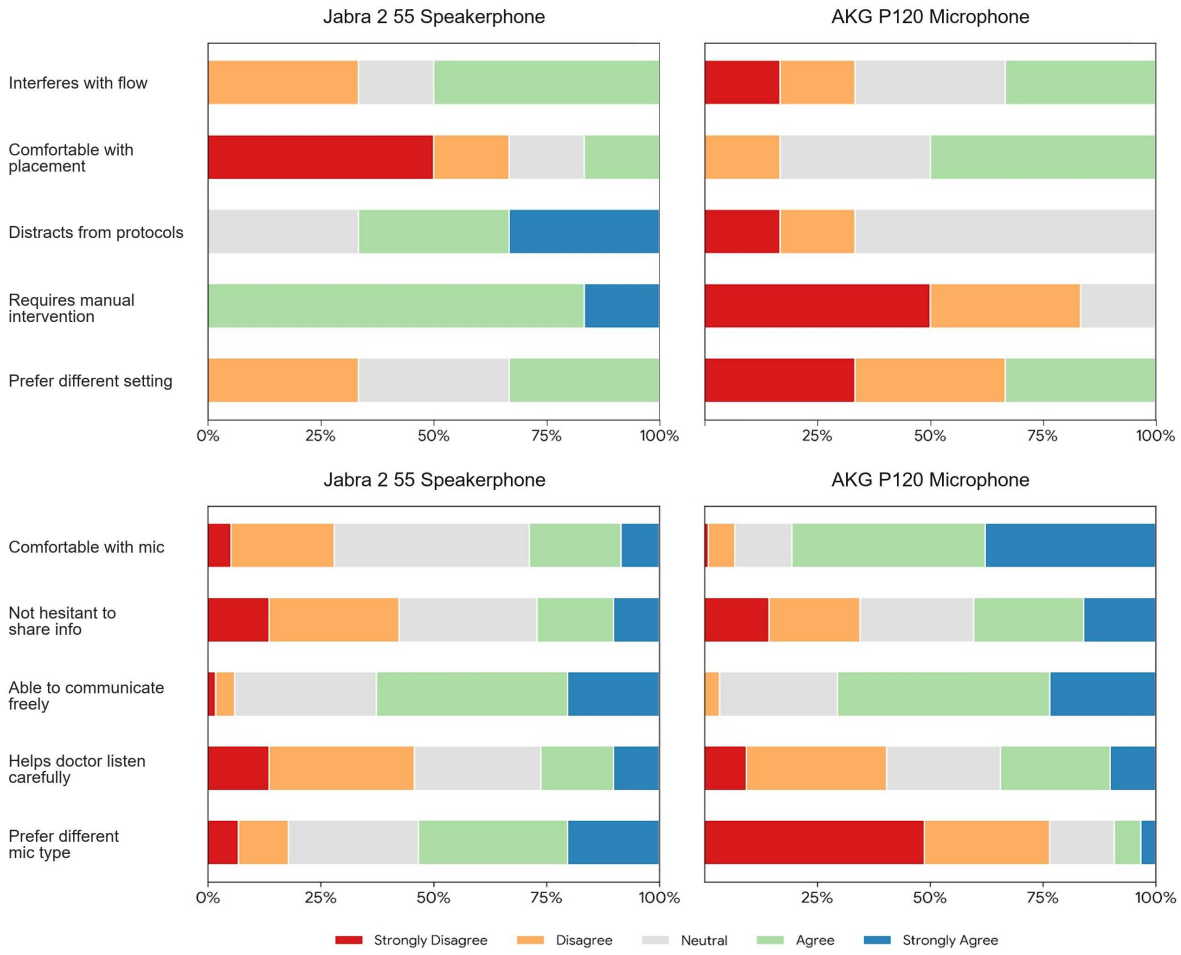


**Fig. 8** Quantitative analysis of doctor-patient perspective and feedback on the use of microphone configuration.

**Table 3** Descriptive statistics of acoustic quality metrics for 495 audio recordings

| **Acoustic Metrics** | **Mean** | **Std.** | **Max** | **Min** | **Median** |
|---|---|---|---|---|---|
| Signal-to-Ratio (dB) | 32.841 | 2.354 | 42.970 | 27.280 | 32.610 |
| Speech-to-Reverberation Modulation Energy Ratio | 5.188 | 1.536 | 11.008 | 2.681 | 4.630 |
| P.56 Active Speech Level (dBov) | -25.237 | 7.065 | -16.200 | -53.300 | -21.930 |
| P.56 Speech Activity (%) | 51.924 | 14.426 | 93.080 | 8.970 | 52.680 |
| Speaker Separability | 1.458 | 0.150 | 2.104 | 1.100 | 1.451 |
| Speaker Silhouette Score | 0.261 | 0.075 | 0.473 | 0.035 | 0.260 |
| NISQA MOS | 2.110 | 0.379 | 3.248 | 0.631 | 2.166 |
| NISQA Noisiness | 2.659 | 0.335 | 3.435 | 1.214 | 2.678 |
| NISQA Coloration | 2.585 | 0.279 | 3.558 | 1.558 | 2.606 |
| NISQA Discontinuity | 3.468 | 0.371 | 4.392 | 2.291 | 3.549 |

distribution pattern of the NISQA Noisiness (Mean: 2.659, Std: 0.335) and Coloration (Mean: 2.585, Std: 0.279). The uniform spectral fidelity validates the effectiveness of the device's built-in background suppression, providing reliable input distributions that are crucial for the fine-tuning of robust ASR pipelines.

In addition, the clinical recordings show extremely high temporal stability, as evidenced by the NISQA Discontinuity metric (Mean: 3.468, Median: 3.549). High discontinuity scores reflect the high degree of preservation of conversational flow, without any sign of audio clipping or signal dropout, even between

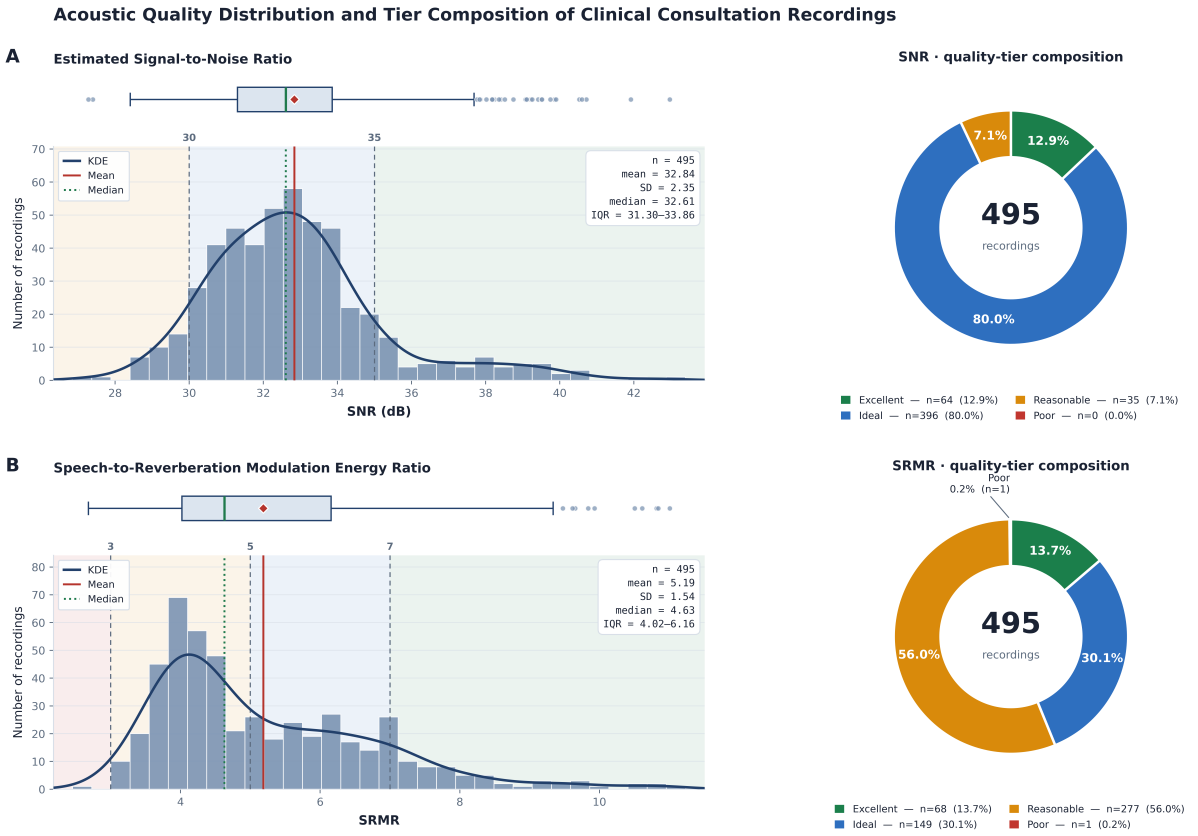


**Fig. 9** Distribution and quality-tier classification of SNR and SRMR values of 495 clinical audio recordings collected using Jabra 2 55 Speakerphone.

overlapping doctor-patient cross-talk. This continuity is crucial for maintaining the contextual semantics for downstream clinical Natural Language Processing (NLP) tasks, such as clinical Named Entity Recognition (NER) and clinical documentation. Thus, the selected hardware offers a very scalable and stable pipeline that is able to capture the high fidelity of the conversational continuity needed to support the next generation clinical AI architectures in an under-served environment.

## 3.4 Acoustic performance across different ASR models

We evaluated four ASR models on a 3-hour annotated and clinically validated dataset of multi-speaker doctor-patient conversations. Performance was quantified using word error rate (WER) and character error rate (CER), applying a normalized 3.0 WER-to-CER ratio where projected, as shown in Table 4.

**Table 4** Comparison of acoustic performances of our ASR model and different benchmark ASR models

| ASR Models | Gemini 3.5 Flash | Scribe V2 | Gemini 3 Flash | Gemini 3.1 Pro |
|---|---|---|---|---|
| **Word Error Rate (WER)** | **10.27%** | 11.85% | 13.70% | 14.38% |
| **Character Error Rate (CER)** | **5.40%** | 7.57% | 6.06% | 8.43% |

Frontier ASR models demonstrated lower overall error margins in terms of WER within a range of 10-15% and CER within a range of only 5-9%. Among all ASR models, the Gemini 3.5 Flash outperformed all other tested models in terms of acoustic fidelity with a WER of only 10.27% and CER of 5.40% respectively, establishing its structural reliability against gold-standard reference transcripts. Another specialized ASR model by ElevenLabs, Scribe_V2 achieved a WER of 11.85%, with a relatively higher

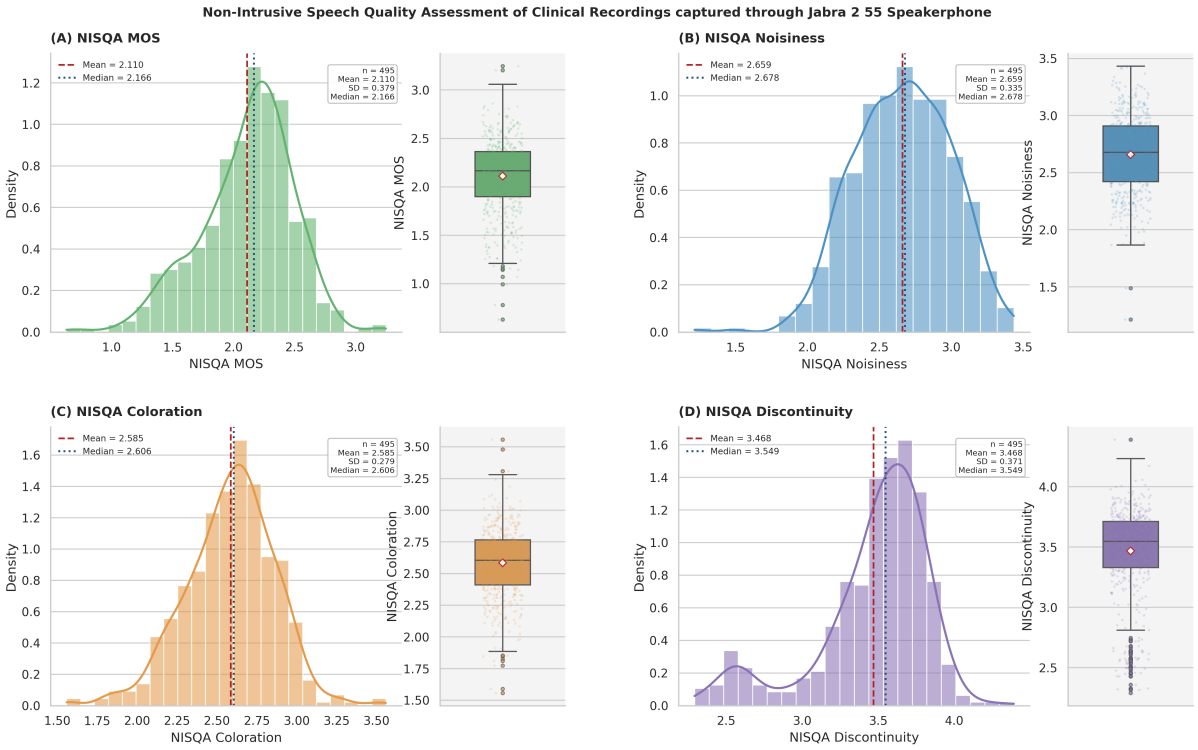


**Fig. 10** Distribution of four NISQA metrics of 495 clinical audio recordings collected using Jabra 2 55 Speakerphone, including their corresponding histograms, kernel density estimation, and boxplots.

CER of 7.57%. Gemini 3 Flash achieved a WER of 13.70%, followed by the second best CER of 6.06% only. Although Gemini 3.1 pro falls short in terms of direct comparison with the other three tested ASR models, its output still lies within suitable range of WER and CER. These findings clearly highlight discernible improvements in both WER and CER as model selection transitions toward highly scalable, and more robust frontier ASR architectures. Ultimately, coupling these frontier models with our optimized clinical data collection framework, we can ensure the high-fidelity and structural reliability required for automated clinical recordings and its corresponding documentations.

# 4 Conclusion

This study represents a framework that demonstrates the execution of a robust data collection pipeline with proper ethical compliance to develop and deploy AI-powered medical scribing system, customized for clinical settings in LMIC. Six different low-cost plug-and-play microphones were evaluated systematically in various seating positions, speaker numbers, and acoustic environments, and a compact beamforming speakerphone was determined to be the most suitable for ambient multi-speaker capture while maintaining acoustic robustness and acceptable to clinicians and patients. A standardized protocol and a secure, fault-tolerant recording architecture are also evidence that scalable, privacy-preserving data collection is possible in routine LMIC clinical workflows. Benchmarking across different ASR models revealed a substantial performance gap between the domain-specific Bengali ASR model and frontier general-purpose systems, underscoring both the promise of coupling this collection framework with stronger ASR backbone and the continued need for targeted adaptation to code-switched, low-resource clinical speech. There are additional planned extensions, such as finer-grained durability checkpointing on the recording pipeline and integrated noise reduction, that still need to be added and tested. Together, it provides a viable model for developing context-specific, resource-efficient clinical documentation systems in other language and acoustically challenging healthcare contexts outside of Bangladesh.

## Data Availability

The dataset used in this study cannot be made publicly available due to institutional, ethical and privacy restrictions. Data were not shared externally without further institutional approval. Derived de-identified artifacts, code components, or additional methodological details may be made available from the corresponding author upon reasonable request and subject to institutional approval.


## Acknowledgments

The authors would like to acknowledge the clinical and technical support of the responsible clinical authority that made this clinical study possible. The authors also thank the physicians and staff involved in the generation and maintenance of the clinical audio recordings used in this work for research purposes only.


## Author Contributions

Conceptualization, S. Zaman, R. Shahrior; Data collection, M. Zaman, A. Tahira, N. Yasmin and S. Rahman; Methodology, S. Zaman, R. Shahrior, and R. Uddin; Formal analysis, S. Zaman, R. Shahrior, A. Tahira and R. Uddin; Investigation and Supervision, M. N. Huda and K. A. Mamun; Writing original draft, S. Zaman, R. Shahrior; Writing review, K. M. Rahman, K. A. Mamun; Editing, S. Zaman, A. Tahira. All authors reviewed the manuscript.

## Declarations


### Funding

This research was supported by the Academic Transformation Fund (ATF) of the Higher Education Acceleration and Transformation (HEAT) project (Grant No.: HEAT-UIU-13863), and by the Institute for Advanced Research, United International University, Bangladesh, under the project entitled "AIM-Scribe: Automated and Intelligent Medical Scribing System for Doctor-Patient Conversation" (Project Code: UIU-IAR-01-2022-SE-24).


### Ethics approval and consent to participate

This study was conducted using clinical audio recordings of doctor-patient conversations collected from routine clinical practice. All clinical recordings used in the study were de-identified before downstream analysis, and additional post-extraction safeguards were applied to eradicate the possibility of residual protected personal information exposure. All methods were performed in accordance with the relevant guidelines and regulations. Written informed consent was waived by the Institutional Research Ethics Board (IREB) of United International University (UIU).

### Competing interests

The authors declare no competing interests.

## References


[1] Arndt, B. G. *et al.* Tethered to the ehr: primary care physician workload assessment using ehr event log data and time-motion observations. *The Annals of Family Medicine* **15**, 419–426 (2017).

[2] Kroth, P. J. *et al.* The electronic elephant in the room: physicians and the electronic health record. *JAMIA open* **1**, 49–56 (2018).

[3] Cowie, M. R. *et al.* Electronic health records to facilitate clinical research. *Clinical Research in Cardiology* **106**, 1–9 (2017).

[4] Han, S. *et al.* Estimating the attributable cost of physician burnout in the united states. *Annals of internal medicine* **170**, 784–790 (2019).

[5] Shanafelt, T., Goh, J. & Sinsky, C. The business case for investing in physician well-being. *JAMA internal medicine* **177**, 1826–1832 (2017).

[6] Shanafelt, T. D. *et al.* Burnout and medical errors among american surgeons. *Annals of surgery* **251**, 995–1000 (2010).

[7] Rodríguez-Vera, F. J., Marin, Y., Sanchez, A., Borrachero, C. & Pujol, E. Illegible handwriting in medical records. *Journal of the Royal Society of Medicine* **95**, 545–546 (2002).

[8] Balch, C. M. *et al.* Personal consequences of malpractice lawsuits on american surgeons. *Journal of the American College of Surgeons* **213**, 657–667 (2011).

[9] Gellert, G. A., Ramirez, R. & Webster, S. L. The rise of the medical scribe industry: implications for the advancement of electronic health records. *Jama* **313**, 1315–1316 (2015).

[10] Walker, K. J. *et al.* An economic evaluation of the costs of training a medical scribe to work in emergency medicine. *Emergency Medicine Journal* **33**, 865–869 (2016).

[11] Heidi Health. Ambient listening: Ai technology in healthcare. Heidi Health Blog (2026). URL https://www.heidihealth.com/en-us/blog/ambient-listening. Accessed: August 2026.

[12] Ballantyne, A., Style, R., Stubbe, M., Murton, S. & Dowell, T. Using ai scribes in new zealand primary care consultations: an exploratory survey. *The Journal of Primary Health Care* **18**, 75–81 (2025).

[13] Tran, B. D., Chen, Y., Liu, S. & Zheng, K. How does medical scribes' work inform development of speech-based clinical documentation technologies? a systematic review. *Journal of the American Medical Informatics Association* **27**, 808–817 (2020).

[14] Zolnoori, M. *et al.* Audio recording patient-nurse verbal communications in home health care settings: pilot feasibility and usability study. *JMIR Human Factors* **9**, e35325 (2022).

[15] Niu, T., Chen, Y., Qu, D. & Hu, H. Enhancing far-field speech recognition with mixer: a novel data augmentation approach. *Applied Sciences* **15**, 4073 (2025).

[16] Heidi Health. Heidi launches first hardware designed for reliable audio capture in every clinical setting. Heidi Health Blog (2026). URL https://www.heidihealth.com/blog/heidi-launches-hardware. Accessed: August 2026.

[17] Charan, G., Sivateja, G., Sarma, M. K. & Kumar, S. Unveiling the challenges of speech recognition in noisy environments: A comprehensive review of issues and solutions. *Challenges in Information, Communication and Computing Technology* 407–412 (2024).

[18] Ohde, J. W. *et al.* Barriers and opportunities of scaling ambient ai scribes for clinical documentation across diverse healthcare settings. *NPJ Digital Medicine* **9**, 369 (2026).

[19] Liu, T. *et al.* Machine learning-assisted wearable sensing systems for speech recognition and interaction. *Nature Communications* **16**, 2363 (2025).

[20] Noffs, G. *et al.* Plug-and-play microphones for recording speech and voice with smart devices. *Folia Phoniatrica et Logopaedica* **76**, 372–385 (2024).

[21] Draper, T. C. *et al.* The impact of acoustic and informational noise on ai-generated clinical summaries. *medRxiv* 2025–03 (2025).

[22] Logitech Inc. Logitech rally bar datasheet: All-in-one video bar for medium to large rooms. Official Product Documentation (2024). URL https://www.logitech.com/en-us/products/video-conferencing/room-solutions/rallybar.html. Accessed: August 2026.

[23] Joybuy Insights. Ai conference speakerphone guide 2026: Crystal-clear audio for every meeting. Joybuy Product Blog (2026). URL https://www.joybuy.nl/blog/

ai-conference-speakerphone-2026-guide/b4jnlFF6. Accessed: August 2026.

[24] Huawei Technologies Co., Ltd. Huawei pure voice and intelligent noise cancellation technology whitepaper. Corporate Technology Innovations Report (2026). URL https://www.huawei.com. Accessed: August 2026.

[25] Cesarini, V. & Costantini, G. Reverb and noise as real-world effects in speech recognition models: A study and a proposal of a feature set. *Applied Sciences* **14**, 11446 (2024).

[26] Stone, M. A., Lough, M., Wilbraham, K., Whiston, H. & Dillon, H. Toward a real-world technical test battery for remote microphone systems used with hearing prostheses. *Trends in Hearing* **27**, 23312165231182518 (2023).

[27] Hodges, E. U. *et al.* Deploying smartphone-based artificial intelligence interventions to community health workers in low-and middle-income countries: Recommendations for product developers, implementing partners, and governments. *medRxiv* 2025–12 (2025).

[28] Gordon, K. R. *et al.* Remote microphones support speech recognition in noise and reverberation for children with a language disorder. *Language, speech, and hearing services in schools* **56**, 225–233 (2025).

[29] Fora Soft. Speech recognition accuracy in noise: 2026 playbook. Fora Soft Engineering Blog (2026). URL https://www.forasoft.com/blog/article/speech-recognition-accuracy-noisy-environments. Updated August 2026. Accessed: August 2026.

[30] Luzanto, A. *et al.* Effective acoustic model-based beamforming training for static and dynamic hri applications. *Sensors* **24**, 6644 (2024).

[31] VivaLyn Labs. AI medical scribe software for Kenya healthcare. VivaLyn Labs Product Documentation (2026). URL https://www.vivalynlabs.com/ai-medical-scribe-kenya. Accessed: August 2026.

[32] Vogel, A. P. & Morgan, A. T. Factors affecting the quality of sound recording for speech and voice analysis. *International journal of speech-language pathology* **11**, 431–437 (2009).

[33] Al-Ganad, A. *et al.* Deploying medical ai in low-resource settings: a scoping review of challenges and strategies. *Frontiers in Digital Health* **8**, 1743634 (2026).

[34] Rosas-Pérez, C. *et al. Sound in clinical environments (SILENTS): Overview of a mixed methods intervention to improve acoustic comfort in a hospital ward in london (UK). Proceedings of Forum Acusticum 2025* (European Acoustics Association, 2025). URL https://doi.org/10.61782/fa.2025.0786. Presented at the 11th Convention of the European Acoustics Association.

[35] Robler, S. K., Coco, L. & Krumm, M. Telehealth solutions for assessing auditory outcomes related to noise and ototoxic exposures in clinic and research. *The Journal of the Acoustical Society of America* **152**, 1737–1754 (2022).

[36] Opochinsky, R., Moradi, M. & Gannot, S. Single-microphone speaker separation and voice activity detection in noisy and reverberant environments. *EURASIP Journal on Audio, Speech, and Music Processing* **2025**, 18 (2025).

[37] Schneider, M. *Elevation and microphone setups for 3d recordings. Proceedings of ICSA 2014 in Erlangen* (2014).

[38] Sasseville, M. *et al.* The impact of ai scribes on streamlining clinical documentation: a systematic review. *Healthcare* **13**, 1447 (2025).

[39] Martin, R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing* **9**, 504–512 (2001).

[40] Falk, T. H., Zheng, C. & Chan, W.-Y. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing* **18**, 1766–1774 (2010).

[41] International Telecommunication Union. ITU-T recommendation P.56: Objective measurement of active speech level. Recommendation P.56, ITU-T (2011).

[42] International Telecommunication Union. ITU-T recommendation G.191: Software tools for speech and audio coding standardization. Recommendation G.191, ITU-T (2019).

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

[44] Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987). URL https://doi.org/10.1016/0377-0427(87)90125-7.

[45] Bredin, H. *et al. pyannote.audio: Neural building blocks for speaker diarization. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7124–7128 (IEEE, 2020). URL https://doi.org/10.48550/arXiv.1911.01255.

[46] Desplanques, B., Thienpondt, J. & Demuynck, K. *ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. Proceedings of Interspeech 2020*, 3830–3834 (2020). URL https://doi.org/10.48550/arXiv.2005.07143.