Investigating Sibilant Fricative Representation in Bangla Telemedicine Speech: A Cost-Aware Sampling Rate Optimization Study

Prajat Paul * , Mohamed Mehfoud Bouh , Manan Vinod Shah , Forhad Hossain , Ashir Ahmed

Posted Date: 17 March 2026

doi: 10.20944/preprints202603.1320.v1

Keywords: automatic speech recognition (ASR); bangla language; sampling rate; telehealth; lowresource language (LRL); sibilant fricatives; word error rate (WER); bandwidth optimization; speech signal processing

 doi:10.20944/preprints202603.1320.v1

Article

Investigating Sibilant Fricative Representation in Bangla Telemedicine Speech: A Cost-Aware Sampling Rate Optimization Study

Prajat Paul [1,*], Mohamed Mehfoud Bouh [1], Manan Vinod Shah [1], Forhad Hossain [2] and Ashir Ahmed [1]

[1] Faculty of Information Science and Electrical Engineering, Kyushu University

[2] Faculty of Liberal Arts, Sophia University

* Correspondence: paul.prajat.922@s.kyushu-u.ac.jp

Abstract

Automatic speech recognition has advanced rapidly for high-resource languages, yet performance remains limited for low-resource languages such as Bangla, particularly in telehealth settings. Most systems rely on a standardized 16 kHz sampling rate, a design choice despite evidence that Bangla contains sibilant fricatives and other phonetic cues with substantial high-frequency energy that may be suppressed under bandwidth and latency constraints. This study evaluates audio sampling rate as a controllable signal-level parameter for Bangla telehealth ASR to identify an empirically grounded operating range balancing transcription accuracy, execution time, and network bandwidth. Twenty real-world Bangla doctor–patient consultations recorded at 32 kHz were deterministically resampled to 55 configurations between 8 kHz and 32 kHz and transcribed using a fixed cloud-based ASR system. Session-level Word Error Rate, execution latency, payload bandwidth, and high-frequency phonetic content were analyzed using a composite sibilant-likelihood score. WER decreased from 0.338 at 8 kHz to a local minimum of 0.232 at 18.75 kHz, with gains plateauing beyond this range despite substantial bandwidth increases. Elbow-point, Pareto frontier, weighted scoring, and Minimum Acceptable Trade-off analyses converged on an optimal region between 17.25 and 18.75 kHz, demonstrating that sampling-rate optimization improves ASR accuracy without proportional resource costs in telehealth settings.

Keywords: automatic speech recognition (ASR); bangla language; sampling rate; telehealth; low-resource language (LRL); sibilant fricatives; word error rate (WER); bandwidth optimization; speech signal processing

1. Introduction

Speech recognition systems have achieved substantial progress in recent years, driven by advances in deep learning architectures and the availability of large, well-structured speech corpora for high-resource languages. These systems now power a wide range of applications, from digital assistants to automated documentation tools. However, this progress has not been uniformed across the world's languages. Many languages remain technologically disadvantaged due to limited digitized materials, lack of standardized phonetic resources, and insufficient annotated datasets. Such languages are commonly categorized as Low-Resource Languages (LRLs), reflecting their restricted representation in speech and language technologies [1]. The scarcity of high-quality linguistic resources for LRLs limits the reliability, adaptability, and scalability of ASR systems designed for them. Beyond data and model availability, many LRL ASR systems also inherit signal-processing assumptions, such as fixed sampling rates, that were optimized for high-resource languages and are rarely re-examined in low-resource deployment contexts.

 doi:10.20944/preprints202603.1320.v1

Recent Bangla ASR research reflects ongoing efforts to address data scarcity, dialectal variation, and deployment robustness through complementary strategies. Foundational supervised corpora established speaker-diverse Bangladeshi Bangla resources but offered limited explicit modeling of regional variation [2]. Subsequent work introduced dialect-aware lexical datasets with division-level metadata, enabling evaluation beyond Dhaka-centric generalization [3]. Model-centric studies demonstrate that fine-tuning self-supervised architectures such as Wav2Vec 2.0 on larger curated corpora substantially reduces WER/CER, albeit with increased computational demands [4]. To scale data, pseudo-labeling approaches have produced large domain-agnostic corpora (~20k h), while revealing persistent degradation on conversational and telephony speech [5]. These challenges are formalized by out-of-distribution benchmarks explicitly quantifying performance drops under domain and style shifts, including telemedicine [6]. Finally, decoding-level language-model rescoring yields significant accuracy gains but introduces accuracy–latency trade-offs critical for real-world Bangla ASR deployment [7]. Collectively, these studies indicate that progress in Bangla ASR hinges on representative data, scalable annotation, and resource-aware modeling, as improvements on controlled speech often fail to generalize to real-world conversations. The lack of linguistic resources in that aspect creates a noticeable gap in the speech recognition accuracy. Notably, most prior Bangla ASR studies implicitly assume fixed front-end signal representations, with limited examination of how inherited sampling-rate choices may interact with data scarcity and conversational speech characteristics.

Automatic speech recognition has been increasingly adopted in healthcare as a scalable mechanism for converting spoken clinical interactions into text, enabling downstream analysis, documentation, and decision support. In mental health research, ASR has been used to transcribe patient speech at scale to support Natural Language Processing (NLP)–based clinical assessment and longitudinal monitoring, particularly where manual transcription is impractical or privacy sensitive [8]. More broadly, multiple studies have positioned ASR as an upstream infrastructure for clinical language processing, enabling automated transcription of conversational medical speech for documentation, information extraction, and system-level learning across diverse clinical settings [9–11]. In the context of structured clinical documentation, ASR-driven systems have been proposed to assist with anamnesis creation by transforming doctor–patient conversations into editable medical records, improving efficiency and standardization [12]. Comparative evaluations of commercial ASR engines on conversational clinical speech have further established baseline performance characteristics to guide deployment in healthcare workflows [13]. More recently, specialty-specific studies, such as in dentistry and orthodontics, have demonstrated that ASR can support the generation of detailed clinical records and that coupling ASR with language models can further enhance the usability of transcribed medical text [14]. Clinical conversations are acoustically challenging due to overlapping speech, spontaneous disfluencies, and background noise, making recognition performance particularly sensitive to front-end signal representation choices such as sampling rate and bandwidth.

Most current ASR systems, including those developed for Bangla, continue to rely on the standardized 16 kHz sampling rate. While convenient for compatibility, this rate limits the captured spectrum to 8 kHz and fails to preserve the full acoustic detail of several Bangla phoneme classes, especially the language's rich fricative inventory. Sibilant fricatives are a class of consonant sounds produced by directing turbulent airflow against the teeth or alveolar ridge, generating high-intensity, high-frequency noise concentrated in the upper spectral bands. Bangla includes multiple sibilant fricatives, /s/, /ʃ/, and dialect-specific /ɕ/-like variants, whose acoustic signatures often extend beyond the 8 kHz boundary. In several regional accents, sibilants produce sharper turbulence and enhanced high-frequency energy, making them especially vulnerable under 16 kHz sampling. When these upper-band cues are truncated at recording, ASR systems struggle to separate acoustically similar fricatives in conversational and medical speech. Sibilant fricatives such as /s/, /ʃ/, and dialectal /ɕ/ exhibit distinctive high-frequency spectral patterns that differ from non-sibilant frication and ambient noise: unlike background noise, which distributes energy broadly and irregularly, sibilants

 doi:10.20944/preprints202603.1320.v1

contain structured cues such as narrow spectral peaks, high spectral moments, and stable turbulence patterns across speakers and tokens, as shown in acoustic-phonetic studies reporting higher spectral peaks, sharper energy concentrations, and greater spectral kurtosis than other fricatives and noise [15,16]. Because the 16 kHz Nyquist boundary removes much of this extended high-frequency structure, critical information is lost. Studies on extended high-frequency cues further show that preserving spectral information above 8 kHz improves phoneme separability under noise and reduces omission or substitution errors in ASR [17]. These observations are consistent with broader findings that sibilants concentrate substantial energy in higher frequencies and require spectral measures beyond narrowband limits; fricatives are turbulence-based but shaped by vocal-tract filtering, producing wider spectral distributions than vowels or sonorants. Diagnostic measures such as FM, AmpD, Fh, AmpRange, and HighLevelD often require spectral content up to at least 15 kHz [18]. Some discrimination algorithms, such as DFT-slope separation of [s] and [ʃ], require analyzable ranges up to 8 kHz and therefore need at least a 16 kHz sampling rate [19]. Evidence from 44.1 kHz conversational corpora and laboratory or modeling studies spanning 15–20 kHz indicates that extended high-frequency information can alter cue weighting and improve discrimination or perceptual-model accuracy in a context-dependent manner, while population-response modeling further shows that structured, multi-channel spectral profiles predict perceptual confusions more effectively than raw spectra, underscoring the importance of organized high-frequency acoustic cues [20–23]. Although fricatives are broadband and noise-like, their structure is not random. Vocal-tract filtering produces systematic spectral patterns that differ from unstructured noise. Spectral-peak measures, amplitude differences, high-frequency maxima, and dynamic cues capture this structure and help separate sibilants from background noise [18]. ASR front-end strategies leverage this by separately processing low- and high-frequency bands; appending de-noised high-frequency filter-bank energies improves wideband ASR under noise. Frequency-filtered band energies and sampling-rate conversion methods maintain performance across variable sampling conditions [24].

Speech recognition accuracy and robustness are strongly shaped by sampled bandwidth, front-end feature design, and training–test sampling mismatches, and prior studies have systematically quantified these dependencies while proposing practical remedies. Rather than a single universally optimal operating point, earlier work demonstrates that preferred sampling rates vary with task and feature representation; for instance, Linear Predictive Cepstral Coefficients (LPCC) and Mel-Frequency Cepstral Coefficients (MFCC) front ends were found to perform best at 12 kHz and 14 kHz, respectively, in a Hidden Markov Model (HMM)-based recognizer [25]. Nonetheless, 16 kHz wideband audio has become the dominant baseline in practice because it preserves spectral information up to 8 kHz and facilitates stable cross-device processing through subband-based descriptors [24]. Despite this convention, sampling rate is typically treated as a fixed design choice, and only coarse operating points are evaluated, leaving the effects of fine-grained rate variation, particularly above 16 kHz, largely unexplored in realistic deployment scenarios. Corpus-based evaluations show that sibilant discrimination remains robust in conversational speech only when sufficient high-frequency spectral detail is preserved, underscoring how bandwidth and sampling-rate choices directly affect the recoverability of fricative cues critical for ASR [26]. Sampling-rate transformation (SRT) techniques have shown that systems trained at 16 kHz can process 11 kHz test data with markedly reduced word error rates, e.g., from 29.89% to 18.17% without retraining, underscoring the contribution of preserved bandwidth to recognition accuracy [27]. In noisy real-world conditions such as in-car speech, 16 kHz consistently outperforms 8 kHz in robustness [28], while the explicit inclusion of de-noised high-frequency energy has been reported to yield up to 13.96% relative improvement for noisy wideband ASR [24]. It also further indicates that frequency-filtered and subband-energy representations allow high-rate systems to remain compatible with lower-rate inputs while retaining much of the performance benefit. Collectively, these findings support 16 kHz as a reliable and efficient baseline, yet they also point to a broader conclusion: higher sampling rates provide additional acoustic information that can be exploited when extended high-frequency cues, such as fricative turbulence, whispery articulation, or bandwidth-extension effects

 doi:10.20944/preprints202603.1320.v1

are relevant to the recognition task [29]. Notably, while prior work establishes the importance of bandwidth and high-frequency content, it largely treats sampling rate as a fixed design choice or evaluates only coarse operating points, leaving open the question of how incremental increases beyond 16 kHz at recording affect recognition quality under realistic deployment constraints.

Increasing the sampling rate beyond 16 kHz provides a direct acoustic advantage by capturing a wider range of high-frequency energy, thereby preserving sibilant-related cues that are critical for phoneme discrimination in Bangla speech. At the same time, higher sampling rates generate larger audio payloads, increasing network bandwidth demands and processing latency for cloud-based ASR services. These trade-offs are especially consequential in low-resource settings where infrastructure support is limited. In regions such as Bangladesh and India, variable network connectivity and constrained deployment environments amplify the practical impact of sampling-rate decisions on real-time speech applications. While the expansion and diversification of Bangla speech corpora remain the most fundamental long-term pathway for improving ASR performance, this study argues that avoidable technical constraints, such as suboptimal sampling-rate selection at recording due to operational constraints, should be systematically examined and optimized. By isolating and addressing signal-level limitations, the present work aims to clarify the possible existence of high frequency components that bears the likeliness of getting clipped by following the standardized sampling rate in recording settings to comply with minimization of resource utilization. In addition, the goal is to find an optimum value or range of values of sampling rate to maintain at audio recording to avoid the exclusion of high frequency components and improve the speech recognition accuracy.

2. Materials and Methods

2.1. Study Design

This study adopts a comparative experimental design to examine how audio sampling rate influences Bangla ASR performance under realistic telehealth constraints. The design explicitly considers three interrelated dimensions: transcription accuracy, computational execution time, and network bandwidth requirements. Rather than treating sampling rate as a fixed engineering parameter, the study models it as a controllable variable whose effects can be systematically evaluated across a dense range of configurations. This approach enables an analysis that connects acoustic signal fidelity with downstream recognition behavior and deployment-level resource costs.

2.2. Speech Data Collection and Recording Protocol

Speech data were collected through a collaboration with Grameen Communications, Bangladesh, as part of routine health checkup activities conducted using the Portable Health Clinic system [30]. Twenty complete doctor–patient consultations were recorded in Dhaka, Bangladesh. The recordings were done at a sampling rate of 32 kHz. This recording configuration reflects a realistic telemedicine environment while also providing sufficient acoustic bandwidth to support controlled resampling experiments. The consultations ranged in duration from approximately 44 seconds to 7 minutes and 30 seconds and consisted primarily of Bangla speech along with medical terminologies of which the English version is commonly used in Bangla. Each interaction followed a typical outpatient consultation flow, including greetings, history-taking, symptom description, preliminary assessment, and clinical recommendations. The conversational and dialogic nature of the recordings introduces natural variability such as turn-taking, overlapping speech, and spontaneous disfluencies, which are essential for evaluating ASR performance under ecologically valid conditions. The use of 32 kHz recordings as the source material establishes an upper bound on available acoustic information and allows all lower sampling rates to be derived through deterministic down sampling. By avoiding independent recordings at different sampling rates, the design ensures that linguistic content, speaker behavior, background noise, and conversational structure remain constant across conditions. This isolates sampling rate as the primary experimental factor and prevents confounding

 doi:10.20944/preprints202603.1320.v1

effects that could arise from differences in recording hardware or environment. All audio recordings were retained in their raw waveform form to preserve the full acoustic signal. Prior to analysis, segments containing personal identifiers or sensitive information were removed to ensure patient privacy and ethical compliance. No additional signal enhancement, denoising, or normalization was applied, allowing the ASR system to operate on audio that closely reflects real-world telehealth inputs.

2.3. Audio Preprocessing and ASR Evaluation Pipeline

Each recording was segmented into non-overlapping 15-second windows using a Python-based procedure to satisfy the input constraints of the Wit.ai, an ASR system from Meta, while preserving natural conversational dynamics, including overlapping speech and speaker interactions. Audio segments originating from the native 32 kHz recordings were resampled to target sampling rates between 8 kHz and 32 kHz using a non-uniform grid designed to balance analytical resolution and computational feasibility. Sampling rates from 8–12 kHz was evaluated at 500 Hz intervals to capture broad effects of high-frequency loss, while finer 250 Hz increments were applied in the 12–20 kHz range to resolve transitional behavior where performance sensitivity was expected to be highest. Above 20 kHz, sampling rates were evaluated at 1 kHz intervals to reduce redundancy in regions where theoretical gains are limited, and resource costs increase. All resampling was performed using identical signal-processing routines to ensure consistency across conditions, enabling detailed characterization of sampling-rate sensitivity while supporting practical evaluation of deployable operating ranges. Each resampled audio segment was submitted to the Wit.ai ASR service for transcription using a consistent API configuration. Segment-level transcripts were generated independently and later concatenated to reconstruct a full-session hypothesis transcript for each consultation. To mitigate the effects of transient network variability or service instability, the system architecture incorporated caching mechanisms that prevent redundant ASR calls for previously processed segments while preserving segment-level metadata such as processing time and payload size. Manual reference transcripts were prepared for each complete consultation by listening to the full audio recordings. These transcripts serve as ground truth for evaluation and reflect the intended lexical content of the conversations rather than segment-level approximations. After ASR processing, reconstructed hypothesis transcripts were compared against the corresponding reference transcripts using word error rate as the primary accuracy metric. WER was computed using the Jiwer Python library, which provides standardized alignment and scoring procedures suitable for conversational speech.

2.4. Computational Latency and Bandwidth Cost Measurement

Execution time was measured for each automatic speech recognition request by recording the elapsed duration between submission of an audio segment to the ASR API and receipt of the corresponding transcription, thereby capturing end-to-end latency under realistic usage conditions, including local preprocessing, network transmission, remote inference, and response delivery. Timing was performed at the segment level to account for variability arising from segment duration, network conditions, and service-side processing behavior, and execution times were subsequently aggregated across segments and consultations using robust summary statistics to characterize the computational cost associated with each sampling rate configuration. Network bandwidth requirements were estimated by recording the file size of each uploaded uncompressed audio segment and normalizing it by segment duration to obtain an effective data transmission rate. This approach directly associates bandwidth consumption with sampling rate and segment length, enabling consistent comparison across configurations while providing a conservative approximation of communication cost relevant to network-constrained telehealth deployments, where high-fidelity audio transmission is often required.

 doi:10.20944/preprints202603.1320.v1

2.5. Acoustic Analysis for Sibilant Characterization

To examine how audio sampling rate affects the preservation of high-frequency phonetic information, an acoustic analysis was conducted with a focus on sibilant fricatives. Sibilants are known to exhibit strong turbulent energy concentrated in the upper frequency range and are therefore particularly sensitive to bandwidth limitations imposed by lower sampling rates. Analyzing their acoustic properties provides a principled way to assess whether increases in sampling rate meaningfully preserve information that is otherwise truncated under standard configurations. For each original 32 kHz recording, frame-level spectral features were extracted using Python-based audio analysis library LibROSA. Three summary metrics were first computed at the file level: mean spectral centroid, mean spectral flatness, and mean high-frequency (HF) energy ratio. Spectral centroid was used as an indicator of the distribution of energy along the frequency axis, with higher values corresponding to greater high-frequency concentration typically associated with fricative articulation. Spectral flatness quantified the degree to which the signal exhibits noise-like characteristics, which are prominent in sibilant turbulence compared to voiced or harmonic speech. The high-frequency energy ratio measured the proportion of signal energy above the 8 kHz boundary, corresponding to the Nyquist limit of 16 kHz sampling, thereby directly quantifying the extent of acoustic information that would be discarded under commonly used ASR settings. Together, these measures provide complementary views of sibilant-related acoustic structure and establish a descriptive basis for understanding how much potentially informative high-frequency content is present in the recordings prior to any down sampling.

2.6. Frame-Level Sibilant Likelihood Estimation

The composite sibilant-likelihood score was computed at the frame level to enable fine-grained acoustic analysis beyond global summary statistics, using a weighted integration of seven established acoustic features: spectral centroid, spectral flatness, high-frequency energy ratio, sibilant-band energy ratio, zero-crossing rate, spectral skewness, and energy consistency. The weighting scheme was determined heuristically based on well-established acoustic–phonetic evidence for fricative discrimination. Spectral centroid and spectral flatness were assigned the highest weights (0.20 each), consistent with findings by Jongman et al. [15], who identified spectral peak location and noise-like energy distribution as primary discriminators of sibilant articulation. The sibilant-band energy ratio, computed over the 2.5–8 kHz range corresponding to the dominant turbulence region of Bangla sibilants (/s/, /ʃ/, /ɕ/), was also assigned a high weight (0.20). The high-frequency energy ratio received a slightly lower weight (0.15) to capture extended high-frequency content beyond 8 kHz, which Monson et al. [31] demonstrated to contain perceptually relevant information often lost in standard transmission. Zero-crossing rate and spectral skewness were assigned moderate weights (0.10 each); although Forrest et al. [32] confirmed their utility in distinguishing voiceless obstruents, Kong et al. [33] showed reduced reliability under low signal-to-noise conditions, motivating a limited contribution for robustness in real-world telehealth audio. Energy consistency was assigned a minimal weight (0.05) to suppress very low-energy frames and function primarily as a noise-gating mechanism [34]. All features were normalized to a common ([0,1]) scale prior to weighting. Frames were categorized based on confidence thresholds applied to the composite score, with high-confidence frames labeled as likely sibilant, intermediate-confidence frames labeled as possible sibilant, and remaining frames evaluated using an inverted spectral criterion to identify noise-dominant cases; frames not meeting either condition were labeled as unclear. This multi-feature, weighted formulation provides an interpretable and literature-aligned estimate of sibilant presence suitable for natural conversational speech without requiring manual phonetic annotation.

$$S(f) = 0.20\,\tilde{C}(f) + 0.20\,\tilde{F}(f) + 0.20\,\widetilde{E_{sib}}(f) + 0.15\,\widetilde{E_{HF}}(f) + 0.10\,\tilde{Z}(f) + 0.10\,\tilde{K}(f) + 0.05\,\widetilde{E_{cons}}(f) \quad (1)$$

Feature Definitions:

$S(f) =$ Composite sibilant-likelihood score

 doi:10.20944/preprints202603.1320.v1

$\tilde{C}(f) =$ Normalized spectral centroid
$\tilde{F}(f) =$ Normalized spectral flatness
$\widetilde{E_{sib}}(f) =$ Normalized sibilant-band energy ratio
$\widetilde{E_{HF}}(f) =$ Normalized high-frequency energy ratio
$\tilde{Z}(f) =$ Normalized zero-crossing rate
$\tilde{K}(f) =$ Normalized spectral skewness
$\widetilde{E_{cons}}(f) =$ Normalized energy consistency

2.7. Sampling Rate–Dependent ASR Performance Analysis

Sampling rate was treated as the primary independent variable in a structured evaluation of Bangla automatic speech recognition performance across three interrelated dimensions: transcription accuracy, computational execution time, and network bandwidth requirement. Accuracy was quantified using WER, a standard metric for speech recognition accuracy. It is a ratio of the sum of substitution, deletion, and insertion errors to the total number of words in the reference transcript. Execution time and bandwidth served as resource-oriented constraints relevant to real-world deployment. Sampling rates between 8 kHz and 32 kHz were selected using a non-uniform grid to provide higher resolution in frequency ranges where performance transitions were expected, while avoiding redundant evaluation at higher rates with diminishing returns. This analytical framework enables explicit examination of accuracy–efficiency trade-offs and provides the foundation for subsequent analyses of optimal and near-optimal operating points under practical constraints.

2.8. Low-Pass Filtering Control Analysis

A low-pass filtering (LPF) control condition was introduced to isolate the contribution of extended high-frequency acoustic content from sampling-rate–related effects. Original recordings being sampled at 32 kHz and low-pass filtered at 8 kHz before segmentation and ASR processing, thereby preserving temporal resolution and payload characteristics while removing spectral content above the 16 kHz Nyquist limit. Comparing ASR performance under this condition with unfiltered higher-rate and native 16 kHz configurations enables attribution of accuracy differences specifically to preserved high-frequency phonetic cues rather than to data rate or processing artifacts. Execution time and bandwidth were measured identically to other conditions, allowing the LPF configuration to serve as a controlled baseline for interpreting sampling-rate–dependent performance gains.

2.9. Elbow-Point Detection: Identifying Diminishing Returns

To identify sampling-rate regions where further increases yield diminishing improvements in recognition accuracy, an elbow-point analysis was applied to the WER–sampling rate relationship. Rather than focusing on absolute performance, this analysis highlights inflection regions where gains attributable to increased acoustic bandwidth begin to taper, indicating a transition from signal-limited to model- or data-limited performance. This provides an interpretable criterion for distinguishing sampling rates that meaningfully preserve phonetic information from those offering marginal benefit.

2.10. Pareto Frontier Analysis: Balancing Accuracy and Bandwidth

To explicitly account for deployment constraints, a Pareto frontier analysis was conducted using WER and estimated bandwidth as competing objectives. This analysis identifies sampling rate configurations that achieve optimal trade-offs, in the sense that no alternative configuration simultaneously improves accuracy while reducing bandwidth cost. By isolating non-dominated operating points, the Pareto framework complements the elbow analysis by emphasizing efficiency under network-constrained conditions rather than accuracy trends alone.

 doi:10.20944/preprints202603.1320.v1

2.11. Composite Scoring and Minimum Acceptable Trade-Off Selection

To support holistic comparison across all configurations, a composite scoring framework was employed that integrates normalized measures of transcription accuracy, execution time, and bandwidth into a single efficiency score, with accuracy assigned the highest weight. In addition, a minimum acceptable trade-off strategy was applied by first selecting configurations whose accuracy falls within a predefined tolerance of the best-performing condition and then identifying the lowest-cost option among them. Together, these analyses enable principled selection of sampling rates that balance recognition quality with practical deployment constraints when absolute optimal accuracy is not required.

2.12. Ethical Considerations

This study was conducted in accordance with ethical principles governing research involving human participants. All individuals whose voices were included in the recordings were informed in advance that the audio data would be collected and used for research purposes, and informed consent was obtained prior to participation. Participation was voluntary, and no incentives or coercive measures were used. To protect participant privacy and confidentiality, all audio recordings were de-identified prior to any preprocessing, analysis, or transcription. Any segments containing personal identifiers or information that could potentially reveal participant identity were removed before data processing. Only anonymized audio data were retained for analysis, and no personally identifiable information was stored, processed, or shared as part of this study. Although the dataset consists of doctor–patient conversational speech, the analyses conducted are limited to technical evaluation of speech signal properties, automatic speech recognition performance, and system-level metrics such as execution time and network bandwidth. The study does not generate diagnostic information, treatment recommendations, or outputs intended to influence clinical decision-making. All findings are secondary, non-interventional, and focused exclusively on methodological and computational aspects of speech processing. As such, the research presents minimal risk to participants and is consistent with ethical guidelines for secondary analysis of anonymized clinical communication data.

3. Results

3.1. Sibilant-Related Acoustic Measures

Table 1 summarizes file-level acoustic characteristics and frame-level sibilant categorization across the 20 Bangla telehealth recordings. Mean spectral centroid values span a broad range (≈1848–3018 Hz), with higher centroid values generally co-occurring with increased spectral flatness (≈0.11–0.30) and elevated HF energy ratios above 8 kHz (≈0.05–0.12), indicating substantial variability in high-frequency acoustic content across sessions. Correspondingly, the proportion of frames classified as *likely sibilant* ranges from ~8.5% to ~32.8% (mean ≈16.6%), while *possible sibilant* frames range from ~10.5% to ~43.2% (mean ≈14.8%). Noise-dominant frames remain comparatively limited (*likely noise* mean ≈0.4%; *possible noise* mean ≈3.0%), and most frames fall into the *unclear* category (mean ≈66.1%), reflecting the conversational nature of the data with frequent phonetic transitions and mixed speech content. Collectively, these distributions indicate that sibilant-related acoustic cues are present at nontrivial levels across recordings, with measurable high-frequency energy above 8 kHz available prior to down-sampling.

Table 1. Spectral Characteristics and Sibilant Classification Outcomes for 20 Audio Files.

File	Mean Spectral Centroid (Hz)	Mean Spectral Flatness	Mean HF Energy Ratio	Likely Sibilant (%)	Possible Sibilant (%)	Likely Noise (%)	Possible Noise (%)	Unclear (%)
1.wav	1848.223	0.131	0.053	13.381	11.053	0.003	4.961	75.495

 doi:10.20944/preprints202603.1320.v1

2.wav	2052.420	0.159	0.065	15.334	13.676	0.024	13.064	70.772
3.wav	2455.571	0.223	0.086	21.260	18.793	0.123	29.383	59.295
4.wav	2279.628	0.212	0.079	22.857	14.090	0.002	0.647	63.049
5.wav	2251.891	0.191	0.070	16.377	18.411	0.004	27.373	64.520
6.wav	2006.801	0.146	0.063	14.819	12.467	0.009	0.983	72.714
7.wav	1961.270	0.178	0.067	21.365	10.505	0.003	0.107	68.130
8.wav	2182.456	0.194	0.067	17.379	18.413	0.001	24.211	63.946
9.wav	2295.956	0.184	0.064	17.086	16.954	0.028	26.026	64.986
10.wav	2089.816	0.160	0.059	13.879	15.507	0.168	20.544	69.885
11.wav	2560.026	0.200	0.075	16.072	22.242	1.646	26.995	60.316
12.wav	3018.359	0.287	0.109	25.231	27.668	0.019	43.224	45.737
13.wav	2459.080	0.217	0.078	21.203	18.540	0.001	31.632	59.265
14.wav	2945.454	0.304	0.122	32.794	22.905	0.001	4.490	44.283
15.wav	2208.959	0.147	0.058	11.967	16.210	0.173	19.653	70.579
16.wav	2028.531	0.127	0.052	10.604	13.692	0.029	15.575	74.709
17.wav	2300.926	0.139	0.053	11.054	16.990	6.150	11.756	70.871
18.wav	2156.108	0.119	0.052	9.706	15.578	1.534	13.286	73.008
19.wav	2206.857	0.126	0.046	9.906	16.646	3.797	12.088	72.233
20.wav	1905.613	0.105	0.049	8.470	11.743	0.073	11.358	78.530

Table 2 summarizes file-level aggregates of the composite sibilant-likelihood score and its weighted feature components across the 20 Bangla telehealth recordings. Composite scores cluster within a narrow range (mean ≈0.59; range ≈0.56–0.65), indicating stable behavior of the scoring framework across heterogeneous conversational sessions. The largest contributions consistently arise from spectral centroid (mean ≈0.81), spectral flatness (mean ≈0.71), HF energy ratio (mean ≈0.73), and sibilant-band energy ratio (mean ≈0.60), which together dominate the composite score in accordance with their assigned weights and normalized scaling. In contrast, zero-crossing rate, spectral skewness, and energy consistency contribute minimally (all means <0.10), reflecting their auxiliary

 doi:10.20944/preprints202603.1320.v1

role. The number of frames classified as sibilant varies substantially across recordings (≈1.8k–19.1k frames), yet this variation is not accompanied by large shifts in composite score magnitude, suggesting that the score reflects stable spectral characteristics rather than frame count alone. Across files, higher composite scores generally coincide with elevated HF energy and sibilant-band energy components, indicating the presence of nontrivial high-frequency acoustic content prior to down-sampling.

Table 2. Frame-Level Sibilant Likelihood Scores and Contributing Acoustic Components Across Recordings.

File name	Sibilant Frames Count	Sibilant Score	Centroid Component	Flatness Component	HF Energy Component	Sibilant Energy Component	ZCR Component	Skewness Component	Energy Component
1.wav	6811	0.6096	0.8241	0.6871	0.7594	0.6769	0.5199	0.0377	0.0458
2.wav	6020	0.6037	0.8148	0.7227	0.7726	0.6262	0.4803	0.0550	0.0308
3.wav	6509	0.5974	0.8128	0.7558	0.7776	0.5747	0.4498	0.0615	0.0190
4.wav	9191	0.6268	0.8120	0.7771	0.7786	0.6678	0.5286	0.0479	0.0185
5.wav	2566	0.5889	0.8270	0.7375	0.7640	0.5498	0.4431	0.0577	0.0268
6.wav	4025	0.6092	0.7780	0.7178	0.7420	0.7073	0.4763	0.0791	0.0355
7.wav	1793	0.6533	0.8230	0.7834	0.8095	0.7819	0.5041	0.0300	0.0162
8.wav	5056	0.5971	0.8140	0.7219	0.7338	0.5948	0.5142	0.0494	0.0912
9.wav	19148	0.5892	0.8197	0.7098	0.7292	0.5628	0.5090	0.0687	0.0725
10.wav	2094	0.5912	0.8188	0.7170	0.7524	0.5902	0.4507	0.0680	0.0247
11.wav	4167	0.5713	0.8149	0.6869	0.7278	0.5225	0.4434	0.0938	0.0710
12.wav	5621	0.5800	0.8033	0.7546	0.7577	0.5079	0.4170	0.0941	0.0426
13.wav	3925	0.5922	0.8209	0.7360	0.7567	0.5322	0.5096	0.0733	0.0518
14.wav	3064	0.6100	0.7594	0.7871	0.7784	0.6209	0.4811	0.1044	0.0235
15.wav	6023	0.5714	0.7957	0.6623	0.7185	0.5588	0.4675	0.0978	0.0740
16.wav	4981	0.5793	0.8046	0.6538	0.7164	0.6010	0.4765	0.0830	0.0793
17.wav	4838	0.5597	0.7885	0.6170	0.6873	0.5572	0.4716	0.1177	0.1029

 doi:10.20944/preprints202603.1320.v1

18.wav	3066	0.5552	0.7783	0.5552	0.6474	0.6009	0.5221	0.1038	0.1718
19.wav	5377	0.5568	0.8035	0.5858	0.6386	0.5941	0.4899	0.0935	0.1193
20.wav	3310	0.5741	0.8025	0.6299	0.6969	0.6146	0.4990	0.0723	0.0608

The annotated spectrograms for 9.wav and 14.wav in Figure 1 visually corroborate the acoustic trends observed in Tables 1 and 2. In both recordings, intervals of elevated sibilant-likelihood scores align with structured spectral energy concentrated within the 2.5–8 kHz sibilant band and, in several instances, extending beyond the 8 kHz Nyquist boundary of 16 kHz sampling. The higher-scoring segments in 14.wav coincide with more pronounced and clustered high-frequency energy, consistent with its higher mean spectral centroid and HF energy ratio, whereas 9.wav exhibits a denser but more evenly distributed pattern of moderate sibilant activity over time. Periods with low sibilant-likelihood scores correspond to regions dominated by lower-frequency voiced speech or silence. These visual patterns support the interpretation that nontrivial high-frequency sibilant-related acoustic information is present prior to down-sampling and is selectively captured by the proposed frame-level scoring framework.

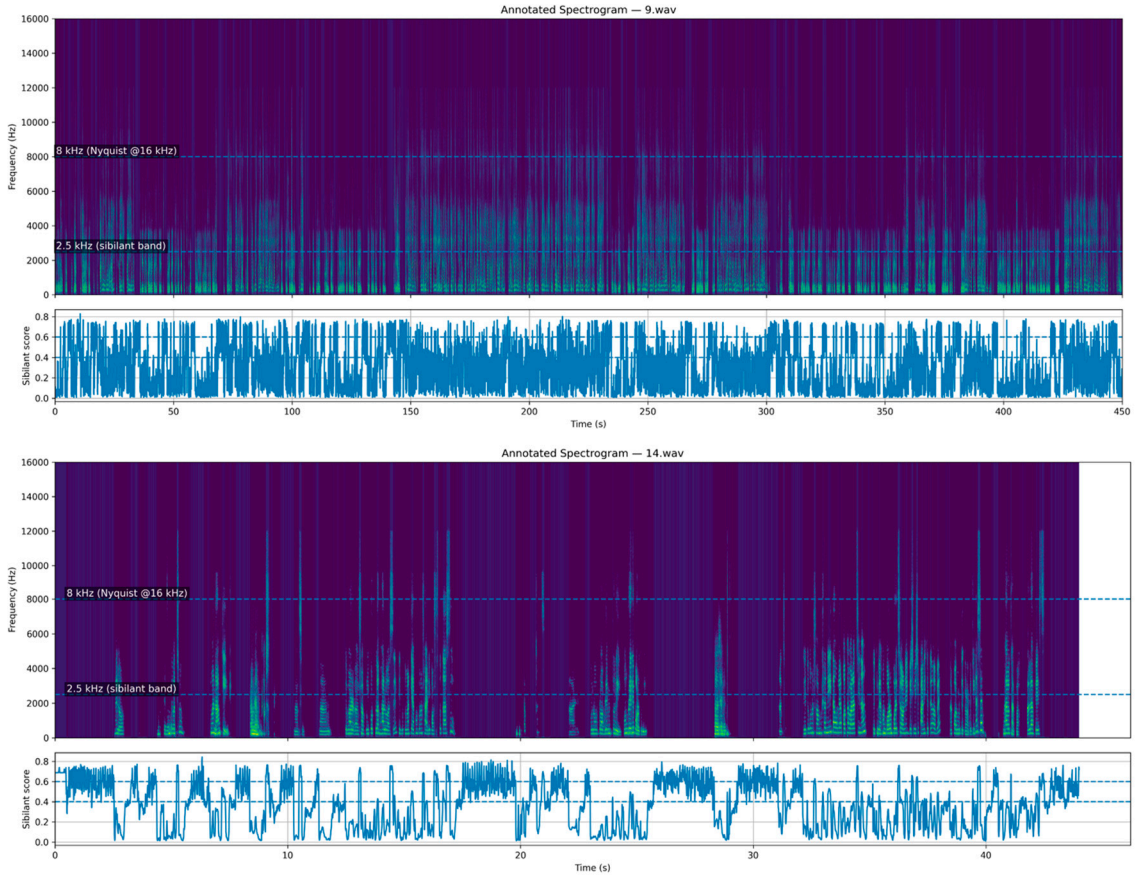

Figure 1. Annotated spectrograms and frame-level sibilant-likelihood scores for two representative Bangla telehealth recordings (9.wav and 14.wav).

 doi:10.20944/preprints202603.1320.v1

3.2. Sampling Rate Optimization

In Table 3, ASR accuracy and resource metrics obtained by deterministically resampling the original 32 kHz recordings to selected target sampling rates and evaluating all segments under an identical cloud ASR configuration is summarized. Segment-level transcripts were concatenated to form session-level hypotheses and compared with manually prepared references to compute global WER using standardized alignment. Execution latency was measured per segment as end-to-end elapsed time (submission to response) and aggregated as median and IQR per sampling rate. Estimated bandwidth was computed from uncompressed segment file size normalized by segment duration (kbps) and summarized as the median per configuration. All processing, timing, and aggregation procedures were held constant across sampling rates to isolate sampling rate as the sole experimental variable.

Table 3. Sampling rate–dependent ASR performance and resource metrics, reporting global WER, median execution latency, latency variability, and estimated payload bandwidth across selected sampling rate configurations.

Sampling Rate (Hz)	Global WER	Latency Median (sec)	Latency IQR (sec)	Payload Median (kbps)
8000	0.3383	2.0558	0.8758	132.0235
8500	0.3036	3.4907	1.4915	136.0235
15250	0.2570	3.5705	1.5301	244.0235
15500	0.2463	3.5877	1.5819	248.0235
15750	0.2492	3.6042	1.5241	252.0235
16000	0.2505	3.6562	1.5315	256.0235
16250	0.2478	3.7584	1.5453	260.0235
16500	0.2433	3.5727	1.4933	264.0235
16750	0.2437	3.8609	1.4934	268.0235
17000	0.2420	3.6006	1.5008	272.0235
17250	0.2341	3.7912	1.5355	276.0235
17500	0.2346	3.8089	1.5675	280.0235
17750	0.2411	3.7927	1.4872	284.0235
18000	0.2444	3.9073	1.5756	288.0235
18250	0.2357	3.7327	1.4722	292.0235
18500	0.2389	4.0662	1.6441	296.0235
18750	0.2320	3.5826	1.4947	297.0235
19000	0.2420	3.8670	1.5573	304.0235
19250	0.2400	3.9875	1.5398	308.0235
19500	0.2378	4.1998	1.4575	312.0235
19750	0.2333	4.1171	1.4909	318.0235
20000	0.2342	3.9819	1.5807	324.0235
20000	0.2342	3.9819	1.5807	330.0235
21000	0.2379	4.0645	1.4763	336.0235
25000	0.2317	4.1946	1.4781	400.0235
26000	0.2394	4.0909	1.5471	416.0235
27000	0.2400	4.0993	1.5345	432.0235
31000	0.2352	4.0316	1.5869	496.0235
32000	0.2312	4.1119	1.5223	512.0235

WER decreases markedly from low rates (8 kHz: ≈0.338) into the mid-range, with notable improvements between ~15–19 kHz (e.g., 15.25 kHz: ≈0.257; 18.75 kHz: ≈0.232), after which gains plateau. At higher rates (≥20 kHz), WER remains within a narrow band (≈0.231–0.240) while bandwidth rises substantially (≈324 to >512 kbps) and median latency increases modestly (~3.6–4.2 s). Latency variability (IQR) is relatively stable across rates. Together, these results indicate a mid-

 doi:10.20944/preprints202603.1320.v1

range region where accuracy improves most relative to added bandwidth and latency, followed by diminishing returns as sampling rate continues to increase.

3.3. Low-Pass Filtered High-Rate Control Condition

Table 4 presents the LPF control in which 32 kHz recordings were filtered to remove spectral content above 8 kHz while retaining the original sampling rate and payload characteristics (≈ 512 kbps). Under this condition, global WER increases to ≈ 0.252 compared with unfiltered mid-to-high sampling-rate configurations, despite comparable execution latency (median ≈ 3.41s). Because computational and network costs remain unchanged, this degradation can be attributed specifically to the removal of extended high-frequency acoustic information. The LPF control therefore isolates the contribution of high-frequency cues and reinforces the interpretation that accuracy improvements at higher sampling rates arise from preserved spectral content rather than from sampling rate or bandwidth alone.

Table 4. Low-pass–filtered control condition using 32 kHz audio with spectral content limited to ≤8 kHz, reporting word error rate, execution latency, and bandwidth to isolate the impact of extended high-frequency information on ASR performance.

Sampling Rate (Hz)	Effective Bandwidth	Global WER	Latency Median (sec)	Latency IQR sec	Payload Median kbps
32000	<= 8 kHz	0.2516	3.4115	1.6257	512.0235

3.4. Elbow-Point Detection: Identifying Diminishing Returns

The illustrated elbow-point detection in Figure 2 applied to the WER–sampling rate curve using the maximum distance–to–endpoints method, where the black curve shows WER across evaluated sampling rates and red markers denote individual configurations. Sampling rates are first ordered in ascending order, and a straight reference line is constructed between the lowest and highest sampling-rate configurations. For each intermediate point, the perpendicular distance to this line is computed, and the sampling rate corresponding to the maximum distance is identified as the elbow, which is indicated by the green dashed line. In this analysis, the elbow is detected at 17,250 Hz, indicating the point at which the rate of WER improvement begins to diminish relative to increases in sampling rate.

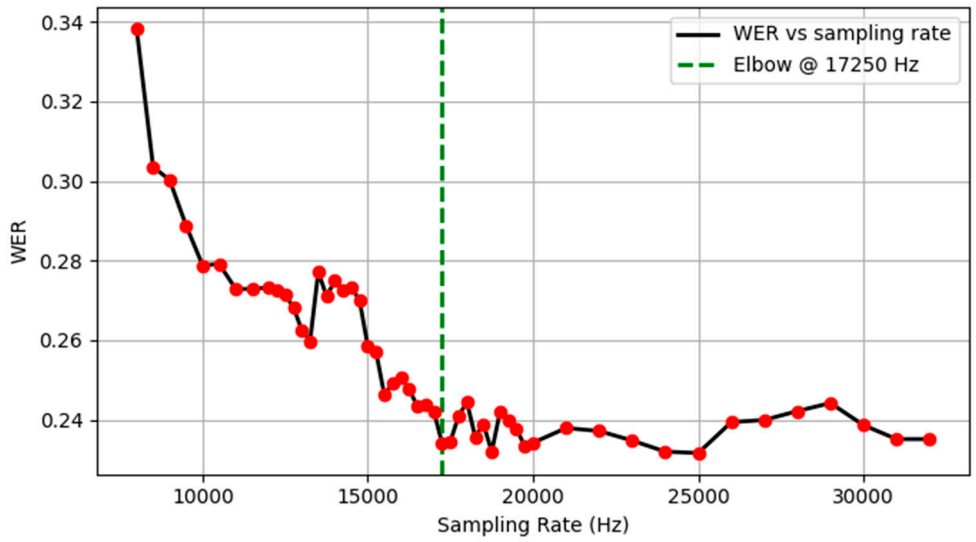

Figure 2. Elbow detection for WER as a function of sampling rate.

 doi:10.20944/preprints202603.1320.v1

3.5. Pareto Frontier Analysis: Balancing Accuracy and Bandwidth

The Pareto frontier between global word WER and median payload bandwidth illustrates the trade-off across all sampling-rate configurations, with grey circle marks representing all evaluated configurations and red circles denoting the Pareto-optimal frontier sampling rates. Each red point corresponds to a configuration for which any further reduction in WER would require an increase in bandwidth. Along the frontier, WER decreases monotonically from approximately 0.34 at ~132 kbps to ~0.23 at ~300–400 kbps, corresponding to an absolute reduction of about 0.11. In the annotated mid-bandwidth region, 17,000 Hz (≈272 kbps) achieves a WER of ~0.242, 17,250 Hz (≈276 kbps) further reduces WER to ~0.234, and 18,750 Hz (≈297 kbps) reaches ~0.232, near the minimum observed along the frontier. Beyond this point, substantial bandwidth increases (≥100 kbps) yield only marginal additional WER reductions (≤0.002), indicating diminishing returns, as shown in Figure 3.

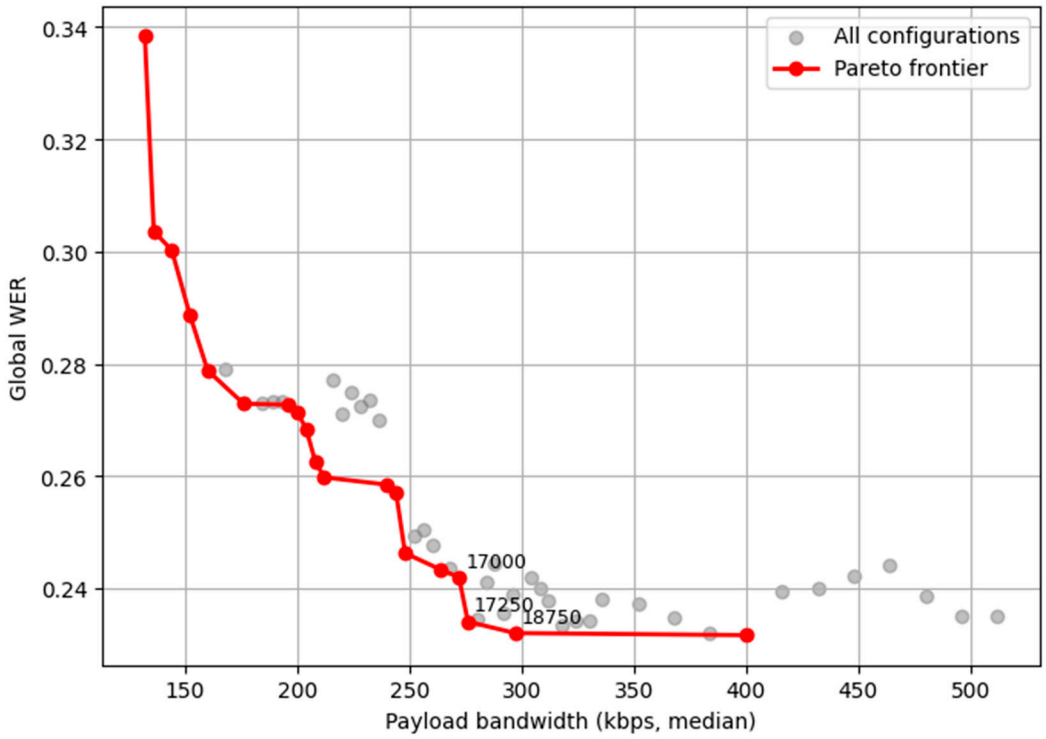

Figure 3. Pareto frontier (red line) of WER vs. bandwidth.

3.6. Weighted Scoring Model: Composite Ranking of Configurations

The weighted scoring analysis integrates normalized recognition accuracy, latency, and bandwidth into a single composite metric, computed as a weighted sum of min–max–normalized variables (lower is better), with weights of 0.60 for global WER, 0.20 for median latency, and 0.20 for median payload bandwidth, as shown in Table 5. As shown in the table, the lowest composite score is achieved at 18,750 Hz (weighted score = 0.2266), driven primarily by the lowest normalized WER (0.0035) while maintaining moderate latency (0.6883) and bandwidth (0.4342). The next two lowest scores occur at 17,250 Hz (0.2458) and 17,500 Hz (0.2526), which exhibit slightly higher normalized WERs (0.0225–0.0277) and higher latency normalization, but benefit from comparatively lower bandwidth normalization than 18,750 Hz. Configurations at 18,250 Hz and 17,000 Hz rank lower due to incremental increases in normalized WER and/or latency, despite similar bandwidth profiles. Overall, the ranking indicates that small differences in WER exert a dominant influence on the composite score relative to modest variations in latency and bandwidth.

 doi:10.20944/preprints202603.1320.v1

Table 5. Normalized performance metrics and composite weighted scores for near-optimal sampling rates, showing the relative trade-off between recognition accuracy (global WER), processing latency, and payload bandwidth under an accuracy-prioritized weighting scheme.

Sampling Rate (Hz)	Global WER	Latency Median (sec)	Payload Median (kbps)	Weighted Score
18750	0.2320	3.5826	297.0235	0.2266
17250	0.2341	3.7912	276.0235	0.2458
17500	0.2346	3.8089	280.0235	0.2526
18250	0.2357	3.7327	292.0235	0.2583
17000	0.2420	3.6006	272.0235	0.2712

3.7. Minimum Acceptable Trade-Off (MAT): Cost-Efficient Near-Optiomal Accuracy

The MAT analysis is based on a near-optimal WER criterion, where the gray curve shows WER across sampling rates, the green dotted line marks the near-optimal threshold defined as within 2% of the minimum WER, green crosses indicate configurations that satisfy this criterion, and the red star highlights the selected MAT operating point corresponding to the lowest sampling rate within the near-optimal region. The minimum observed WER across all sampling rates is 0.2316, yielding a near-optimal threshold of 0.236284, and multiple sampling rates beyond approximately 17 kHz fall below this threshold, forming a plateau of near-optimal performance, as illustrated in Figure 4. Among these, 17,250 Hz achieves a WER of 0.234054, satisfying the accuracy constraint while occurring at a substantially lower sampling rate than several higher-rate alternatives. The selected MAT point therefore represents the lowest sampling rate that meets the predefined accuracy tolerance, balancing recognition performance against increasing bandwidth and computational demands.

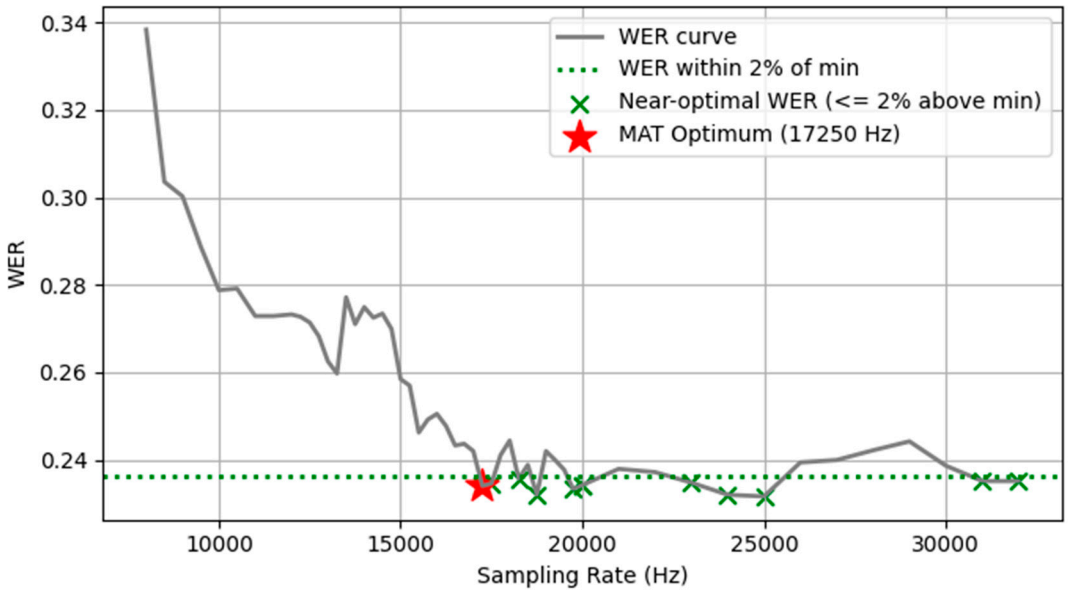

Figure 4. Near-optimal WER and MAT selection as a function of sampling rate.

4. Discussion

This study underscores the importance of audio sampling rate as a tunable optimization variable for ASR, rather than a fixed preset. In contrast to conventional approaches that fix the sample rate (e.g., at 8 or 16 kHz) and potentially confound improvements with changes in model architecture or

 doi:10.20944/preprints202603.1320.v1

training data, our experiments isolate sampling rate as the only varying factor. By focusing on Bangla medical telehealth speech and keeping the model and dataset constant, we directly quantify the impact of sampling rate on recognition performance. This controlled approach reveals that significant accuracy gains can be achieved by optimizing the sampling rate for this specific ASR scenario.

4.1. Interpretation of Sibilant Acoustic Cues

Acoustic analyses confirm that key sibilant fricative cues are present well above the traditional telephone bandwidth of 8 kHz. Spectral centroid and spectral flatness measures remain elevated in the extended high-frequency range (Table 1), and the proportion of high-frequency energy is consistently substantial (Table 2). Figure 1 further shows that energy components beyond 8 kHz persist in the audio spectrum, indicating that the original recordings carry significant high-frequency information. The composite sibilant score remains stable at approximately 0.59 across recordings, suggesting that these high-frequency sibilant-related cues are inherent to the speech data prior to down-sampling.

4.2. Sampling Rate-Dependent Accuracy Gains and Diminishing Returns

Optimizing the sampling rate yields marked improvements in ASR accuracy. WER decreases from 0.3383 at 8 kHz to 0.2320 at 18.75 kHz (Table 3), representing a substantial relative reduction. Beyond approximately 19 kHz, accuracy gains plateau, while bandwidth and latency continue to increase. Median latency rises modestly from around 3.6 s to approximately 4.2 s, and estimated bandwidth increases from roughly 132 kbps to over 500 kbps. These trends identify the 15.25–18.75 kHz region as the most efficient zone, where accuracy improves rapidly relative to added resource cost.

4.3. Isolating the Contribution of Extended High-Frequency Information

The LPF control condition isolates the role of extended high-frequency acoustic content. When 32 kHz audio is filtered to remove spectral content above 8 kHz, global WER increases to 0.2516 (Table 4), despite maintaining identical bandwidth and comparable latency. This degradation confirms that the observed accuracy gains at higher sampling rates are attributable to preserved high-frequency information rather than sampling rate or payload size alone. The LPF control therefore provides direct evidence that extended high-frequency cues contribute meaningfully to recognition performance.

4.4. Elbow Point Identification of the Accuracy-Efficiency Trade-Off

Elbow-point detection applied to the WER–sampling rate curve identifies a clear inflection at 17,250 Hz (Figure 2). This point represents the transition beyond which incremental increases in sampling rate yield diminishing improvements in WER. The elbow therefore marks a practical threshold for cost-effective accuracy gains and provides a data-driven lower bound for selecting an optimal operating region.

4.5. Pareto-Optimal Balance Between Accuracy and Bandwidth

Pareto frontier analysis further refines the optimal operating region by jointly considering WER and bandwidth (Figure 3). Along the frontier, WER decreases from approximately 0.34 at 132 kbps to around 0.23 at 300–400 kbps. Within the mid-bandwidth region, 17,000 Hz (≈272 kbps) achieves a WER of approximately 0.242, 17,250 Hz (≈276 kbps) reduces WER to approximately 0.234, and 18,750 Hz (≈297 kbps) reaches near-minimal WER at approximately 0.232. Beyond this range, bandwidth increases of more than 100 kbps result in WER reductions of 0.002 or less, clearly indicating diminishing returns.

 doi:10.20944/preprints202603.1320.v1

4.6. Composite Ranking and Minimum Acceptable Trade-Off Selection

The weighted scoring model integrates normalized WER, latency, and bandwidth into a single composite metric (Table 5). Under accuracy-prioritized weighting, 18,750 Hz achieves the lowest composite score (0.2266), driven primarily by its lowest normalized WER. The next best configurations are 17,250 Hz (0.2458) and 17,500 Hz (0.2526), which offer slightly higher WER but reduced bandwidth demands. The minimum acceptable trade-off (MAT) analysis further selects 17,250 Hz as the lowest sampling rate that remains within 2% of the minimum observed WER, identifying it as the most cost-efficient near-optimal configuration.

4.7. Recommended Sampling Rate for Bangla Medical Telehealth ASR

Based on converging evidence from acoustic analysis, WER trends, LPF controls, elbow detection, Pareto optimization, and composite scoring, 18,750 Hz is recommended as the best overall sampling rate for Bangla medical telehealth ASR. It achieves the lowest observed WER while maintaining manageable latency and bandwidth requirements. For deployments with stricter resource constraints, a sampling-rate range of 17,250–18,750 Hz is recommended, as this interval captures most of the available accuracy gains while avoiding the inefficiencies associated with higher sampling rates.

4.8. Limitations:

This study was designed to examine the effect of audio sampling rate on automatic speech recognition performance within a controlled, application-relevant telemedicine setting. All experiments employed a single cloud-based ASR system and a consistent segmentation and resampling pipeline, allowing sampling rate to be isolated as the primary variable while ensuring comparability across conditions. Although different ASR architectures or front-end configurations may yield different absolute performance levels, the present design enables robust analysis of relative trends associated with sampling-rate variation in naturalistic doctor–patient speech. Sibilant presence was estimated using a multi-feature, frame-level acoustic formulation grounded in established phonetic correlates of fricative articulation, providing an interpretable proxy for extended high-frequency information without requiring manual phonetic annotation. Execution time and network bandwidth were measured end-to-end under realistic operating conditions, capturing practical deployment behavior rather than isolated computational components. These design choices define the scope within which the findings should be interpreted while supporting a focused evaluation of sampling-rate trade-offs in comparable real-world ASR deployment contexts.

4.9. Comparison with Prior Work:

Prior research in Bangla ASR has predominantly focused on data-centric and model-centric strategies, including corpus expansion, dialect-aware annotation, self-supervised model fine-tuning, pseudo-labeling, and decoding-level language-model rescoring [2–7]. In parallel, healthcare-oriented ASR studies have emphasized system usability, transcription accuracy, and downstream clinical applications, often benchmarking commercial engines under challenging conversational conditions [8–14]. While a substantial body of acoustic–phonetic work has demonstrated the importance of high-frequency cues—particularly for sibilant fricatives—and highlighted the limitations of narrowband representations [15–24], these insights have rarely been translated into deployment-level ASR optimization for specific domains or languages. Existing sampling-rate studies typically treat sampling rate as a fixed or coarsely evaluated design parameter, inherited from telephony or benchmark conventions, with limited exploration beyond 16 kHz and little consideration of application-specific trade-offs [25–29]. In contrast, the present work complements prior efforts by isolating sampling rate as a controllable signal-level variable and systematically evaluating its fine-grained effects on recognition accuracy, latency, and bandwidth within the context of Bangla doctor–patient conversations. By grounding the analysis in domain-specific acoustic characteristics and

 doi:10.20944/preprints202603.1320.v1

explicit resource constraints, this study provides a practical signal-representation perspective that bridges acoustic theory and real-world ASR deployment in low-resource healthcare settings.

5. Conclusions

This study examined audio sampling rate as a controllable signal-level factor influencing automatic speech recognition performance for Bangla doctor–patient conversations under realistic telehealth constraints. Acoustic analyses confirmed the presence of nontrivial high-frequency sibilant-related information beyond the 8 kHz Nyquist limit of standard 16 kHz sampling, motivating systematic resampling experiments. Fine-grained evaluation across sampling rates demonstrated substantial reductions in word error rate as sampling increased into the mid-to-high range, followed by diminishing returns at higher rates despite increasing bandwidth and latency costs. Elbow-point detection, Pareto frontier analysis, weighted scoring, and minimum acceptable trade-off selection converged on a narrow operating region between 17,250 Hz and 18,750 Hz, within which most attainable accuracy gains were realized efficiently. Among these, 18,750 Hz consistently achieved near-minimal WER while maintaining moderate execution time and bandwidth, emerging as the most balanced single operating point when transcription accuracy is prioritized. These findings do not suggest a universal optimal sampling rate but demonstrate that inherited 16 kHz conventions can impose avoidable accuracy limitations in low-resource, conversational medical ASR. By isolating sampling rate from confounding factors and evaluating it under deployment-aware constraints, this work highlights signal-level optimization as a practical complement to data and model-centric advances in low-resource speech recognition.

Author Contributions: PP contributed to Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Visualization, Writing – original draft and Funding Acquisition of the manuscript. MMB contributed to Conceptualization, Methodology, and Writing – review & editing, with primary involvement in the literature review and hypothesis development. MVS contributed to Data curation, Formal analysis, and Visualization, supporting data preprocessing, analytical implementation, and graphical representation of results. Contribution of FH was focused on Data curation, Resources and Investigation. AA contributed to Conceptualization, Supervision, Methodology, Validation, Funding Acquisition, Writing – review & editing and Project Administration, providing oversight and critical evaluation across all stages of the research workflow, from idea formulation through analytical interpretation, and contributing to strengthening the overall research contribution.

Funding: This study was funded by JST BOOST (Japan Grant Number JPMJBS2406) and the APC was funded by JST Startup Ecosystem Co-creation Program for New Industry University Startups under the PARKS Startup Creation Program Student Project Step-2 (Japan Grant Number JPMJSF2317).

Data Availability Statement: The audio recordings and transcripts analyzed in this study were collected by a collaborating organization during routine health checkup activities and shared with the authors under a data-use agreement. Due to confidentiality obligations and the sensitive nature of doctor–patient conversations, these data are not publicly available and cannot be deposited in an open repository. Access may be considered on a case-by-case basis for non-commercial research, subject to approval by the data-owning organization and applicable ethical requirements.

Acknowledgments: The authors acknowledge the use of ChatGPT 5.2, an AI-based language model developed by OpenAI, as a supportive tool during the preparation of this manuscript. ChatGPT was used to assist with language refinement, structural organization, and iterative drafting of selected sections. The authors also gratefully acknowledge Grameen Communications for their contribution to the collection of the Bangla doctor–patient audio recordings during health checkup services, which formed the empirical basis of this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

 doi:10.20944/preprints202603.1320.v1

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
WER	Word Error Rate
MAT	Minimum Acceptable Trade-off
LRL	Low Resource Language
LPF	Low Pass Filter
HF	High Frequency

References

1. Magueresse A, Carles V, Heetderks E. Low-resource languages: a review of past work and future challenges. arXiv. 2020. arXiv preprint arXiv:2006.07264. 2006.
2. Kibria S, Samin AM, Kobir MH, Rahman MS, Selim MR, Iqbal MZ. Bangladeshi Bangla speech corpus for automatic speech recognition research. Speech Communication. 2022;136:84-97.
3. Aiman U, Islam MN, Chowdhury MH, Rahman MS, Habib MT, Hasan M. BRADS and BRWDS: multipurpose audio and text datasets for automatic Bangla regional speech recognition. Data in Brief. 2025;112177.
4. Hossain S, Rihan MR, Imtiaz A, Boni P, Gomes D. Enhancing Bangla local speech-to-text conversion using fine-tuning Wav2vec 2.0 with OpenSLR and self-compiled datasets through transfer learning. In: 7th IEOM Bangladesh International Conference on Industrial Engineering and Operations Management; 2024 Dec 21; Dhaka, Bangladesh. Vol 20240161. https://doi.org/10.46254/BA07
5. Nandi RN, Menon M, Muntasir T, Sarker S, Muhtaseem QS, Islam MT, Chowdhury S, Alam F. Pseudo-labeling for domain-agnostic Bangla automatic speech recognition. In: Proceedings of the First Workshop on Bangla Language Processing (BLP-2023); 2023 Dec;152-162.
6. Rakib FR, Dip SS, Alam S, Tasnim N, Shihab MI, Ansary MN, Hossen SM, Meghla MH, Mamun M, Sadeque F, Chowdhury SS. Ood-speech: a large Bengali speech recognition dataset for out-of-distribution benchmarking. arXiv preprint arXiv:2305.09688. 2023 May 15.
7. Rakib M, Hossain MI, Mohammed N, Rahman F. Bangla-wave: improving Bangla automatic speech recognition utilizing n-gram language models. In: Proceedings of the 12th International Conference on Software and Computer Applications; 2023 Feb 23;297-301.
8. Just SA, Elvevåg B, Pandey S, Nenchev I, Bröcker AL, Montag C, Morgan SE. Moving beyond word error rate to evaluate automatic speech recognition in clinical samples: lessons from research into schizophrenia-spectrum disorders. Psychiatry Research. 2025;116690.
9. Mani A, Palaskar S, Konam S. Towards understanding ASR error correction for medical conversations. In: Proceedings of the First Workshop on Natural Language Processing for Medical Conversations; 2020 Jul;7-11.
10. Klusty MA, Logan WV, Armstrong SE, Mullen AD, Leach CN, Calvert K, Talbert J, Bumgardner VC. Toward automated clinical transcriptions. AMIA Summits on Translational Science Proceedings. 2025;2025:235.
11. Salloum W, Edwards E, Ghaffarzadegan S, Suendermann-Oeft D, Miller M. Crowdsourced continuous improvement of medical speech recognition. In: AAAI Workshops; 2017.
12. Gonçalves YT, Alves JV, Sá BA, da Silva LN, de Macedo JA, da Silva TL. MedTalkAI: assisted anamnesis creation with automatic speech recognition. In: Simpósio Brasileiro de Banco de Dados (SBBD); 2024 Oct 14;83-88. SBC.
13. Kodish-Wachs J, Agassi E, Kenny P III, Overhage JM. A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. In: AMIA Annual Symposium Proceedings; 2018 Dec 5;2018:683.
14. O'Kane R, Stonehouse-Smith D, Ota LC, Patel R, Johnson N, Slipper C, Seehra J, Papageorgiou SN, Cobourne MT. Transcription accuracy of automatic speech recognition for orthodontic clinical records. Journal of Dental Research. 2025;00220345251382452.
15. Jongman A, Wayland R, Wong S. Acoustic characteristics of English fricatives. The Journal of the Acoustical Society of America. 2000;108(3):1252-1263.

 doi:10.20944/preprints202603.1320.v1

16. Maniwa K, Jongman A, Wade T. Acoustic characteristics of clearly spoken English fricatives. The Journal of the Acoustical Society of America. 2009;125(6):3962-3973.
17. Guo ZC, Chandrasekaran B. Extended high-frequency cues to phoneme recognition: insights from ASR. In: Proceedings of Interspeech 2025; 2025;1038-1042.
18. Kozierski P, Sadalla T, Drgas S, Dabrowski A, Giemacki W. Polish whispery speech recognition—minimum sampling frequency. 2017 Aug;611-615. doi:10.1109/MMAR.2017.8046898
19. Hokking R, Woraratpanya K. A hybrid of fractal code descriptor and harmonic pattern generator for improving speech recognition of different sampling rates. In: Meesad P, Sodsee S, Unger H, eds. Recent Advances in Information and Communication Technology 2017. IC2IT 2017. Advances in Intelligent Systems and Computing. Vol 566. 2018. https://doi.org/10.1007/978-3-319-60663-7_4
20. Bauerecker H, Nadeu C, Padrell J. On the advantage of frequency-filtering features for speech recognition with variable sampling frequencies: experiments with speechdatcar databases. In: INTERSPEECH 2003;869-872.
21. Liu FH, Picheny M. On variable sampling frequencies in speech recognition. IBM Thomas J. Watson Research Division; 1997.
22. Nadeu C, Tolos M. Recognition experiments with the SpeechDat-Car Aurora Spanish database using 8 kHz- and 16 kHz-sampled signals. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'01); 2001 Dec 9;135-138.
23. Ssnderson C, Paliwal KK. Effect of different sampling rates and feature vector sizes on speech recognition performance. In: Proceedings of IEEE TENCON'97; 1997 Dec 4;Vol 1:161-164.
24. Hirsch HG, Hellwig K, Dobler S. Speech recognition at multiple sampling rates. In: INTERSPEECH 2001;1837-1840.
25. Guo ZC, Chandrasekaran B. Extended high frequencies improve phoneme recognition: evidence from automatic speech recognition in spatial speech mixtures. The Journal of the Acoustical Society of America. 2025;158(4):3365-3377.
26. Roberts PJ, Reetz H, Lahiri A. Corpus-testing a fricative discriminator; or, just how invariant is this invariant? In: INTERSPEECH 2014;189-192.
27. Steiner IMA. Observations on the dynamic control of an articulatory synthesizer using speech production data. Doctoral thesis. Karlsruhe Institute of Technology; 2010. doi:10.22028/D291-23547
28. Hokking R, Woraratpanya K, Kuroki Y. Speech recognition of different sampling rates using fractal code descriptor. In: 13th International Joint Conference on Computer Science and Software Engineering (JCSSE); 2016 Jul 13;1-5.
29. Shadle CH, Chen WR, Koenig LL, Preston JL. Refining and extending measures for fricative spectra, with special attention to the high-frequency range. The Journal of the Acoustical Society of America. 2023;154(3):1932-1944.
30. Ahmed A, Inoue S, Kai E, Nakashima N, Nohara Y. Portable health clinic: a pervasive way to serve the unreached community for preventive healthcare. In: International Conference on Distributed, Ambient, and Pervasive Interactions; 2013 Jul 21;265-274.
31. Monson BB, Hunter EJ, Lotto AJ, Story BH. The perceptual significance of high-frequency energy in the human voice. Frontiers in Psychology. 2014;5:587.
32. Forrest K, Weismer G, Milenkovic P, Dougall RN. Statistical analysis of word-initial voiceless obstruents: preliminary data. The Journal of the Acoustical Society of America. 1988;84(1):115-123.
33. Kong YY, Mullangi A, Kokkinakis K. Classification of fricative consonants for speech enhancement in hearing devices. PLoS One. 2014;9(4):e95001.
34. Ramirez J, Górriz JM, Segura JC. Voice activity detection: fundamentals and speech recognition system robustness. Robust Speech Recognition and Understanding. 2007;6(9):1-22.