SamaVaani: Auditing and Debiasing Multilingual Clinical ASR for Indian Languages

Subham Kumar[†] **Prakrithi Shivaprakash**[‡] **Abhishek Manoharan**[‡] **Astut Kurariya**[‡]
Diptadhi Mukherjee[*] **Prabhat Chand**[‡] **Pratima Murthy**[‡]
Koustav Rudra[†] **Lekhansh Shukla**[‡] and **Animesh Mukherjee**[†]
[†]IIT, Kharagpur, [‡]NIMHANS, Bangalore, [*]LGBRIMH, Tezpur
{kumarshubham209, prakrithishivaprakash, 12.abhishek.m, astutnamo, diptadhimukherjee}@gmail.com
prabhat@vknnimhans.in, {pratimamurthy, krudra5, drlekhansh, animeshm}@gmail.com

Abstract

Automatic Speech Recognition (ASR) is increasingly used to document clinical encounters, yet its reliability in multilingual and demographically diverse Indian healthcare context remains largely unknown. In this study, we first conduct the systematic audit of ASR performance on real-world psychiatric interview data spanning Kannada, Hindi and Indian English, comparing eight state-of-the-art models including IndicWhisper, WhisperLargeV3, Sarvam, GoogleS2T, Gemma3n, OmniLingual, Vaani, and Gemini. Our results reveal substantial variability across models and languages, with some systems performing competitively in Indian English but failing in regional speech. We further fine-tune two of the best performing open-source models, i.e., Gemma3n and OmniLingual, using various methods. With this, we uncover systematic performance gaps tied to speaker role and gender, raising concerns about equitable deployment in clinical settings, which are further mitigated by fairness-aware fine-tuning. To this end, we propose SamaVaani, a unified debiasing technique that simultaneously improves ASR performance and improves fairness across demographic groups.

1 Introduction

The field of psychiatry is highly dependent on language, with a detailed psychiatric interview the primary diagnostic tool (Sommers-Flanagan et al., 2015) rather than laboratory or radiological investigations. Subsequently, verbatim transcripts of these interviews are widely used for clinical diagnosis, academic training, qualitative research, and, more recently, for the development of AI systems in psychiatric tasks (So et al., 2024). However, producing accurate transcripts remains a major bottleneck: manual transcription is strenuous, time-intensive (5-8 hours for every hour of audio) and prone to human error (Mays and Mathias, 2019). ASR systems provide a scalable alternative albeit transcription errors (Basma et al., 2011) in psychiatric settings can significantly change clinical interpretation (Ciampelli et al., 2023; Shikino et al., 2023). Modern ASR systems, including proprietary platforms (GoogleS2T (Cloud, 2023), Microsoft Azure (Azure, 2023), Amazon Transcribe (Services, 2023)) and open-source models like Whisper (Radford et al., 2022) have improved transcription quality. Although these systems work well for standard English (American and British) and in controlled environments, performance deteriorates substantially for non-standard English, conversational and accented English (Russell et al., 2024), code-mixing and code-switching, which is common in multilingual contexts like India (Sitaram et al., 2020; Koenecke et al., 2020). In addition, Indian English and Indian regional languages remain underrepresented in global training corpora, leading to lower accuracy and greater variability (Javed et al., 2023a; Rai et al., 2024).

These difficulties are amplified with clinically distinctive speech patterns in psychiatric interviews. For instance, low tone, hesitations, and long pauses are common in depression (Alpert et al., 2001; Yang et al., 2012; Durão et al., 2025); fast and loud speech in mania (Kaczmarek-Majer et al., 2024; Anmella et al., 2024; Di Florio et al., 2021); stammering and repeating in anxiety (Silber et al., 2018; Harrigan et al., 1994; Teferra et al., 2022); and disordered agrammatical speech containing neologisms (made-up words to which the patient attaches special meaning) in schizophrenia (Covington et al., 2005; Stokes et al., 2023; Hinzen and Rosello, 2015). Beyond this,

recordings are often made under acoustically difficult conditions (in wards with noise from ceiling fans, hospital instruments, or ambulance sirens) (Baroudi et al., 2026). Another big concern is the fairness and accuracy of the transcription. These problems may be magnified in psychiatric interviews, as doctors and patients are very different in terms of education, socio-economic background, conversational role, and speaking style (Koenecke et al., 2020; Tatman, 2017). While recent advances in multilingual ASR from open-source models such as Whisper (Radford et al., 2022), IndicWhisper (AI4Bharat, 2024), OmniLingual (team et al., 2025), Vaani (ARTPARK-IISc, 2025) and commercial models like Sarvam (SarvamAI, 2025) have improved support for multilingual and regional speech in India, existing evaluations rely mainly on general-purpose datasets and do not capture the complexities of real-world psychiatric interviews. Our work addresses this gap by systematically analyzing and further improving the performance of ASR on multilingual psychiatric interactions.

Our contributions and findings: In this work, we present the first systematic audit of ASR systems on real-world multilingual psychiatric interviews in the Indian context. Leveraging a novel dataset spanning Kannada, Hindi, and Indian English, we evaluate eight state-of-the-art ASR models across languages, speaker roles, and demographic groups, and introduce a comprehensive fairness analysis grounded in WER and fine-grained error patterns. We find substantial variability in performance across models and languages, with consistently higher error rates for low-resource languages such as Kannada, as well as systematic disparities across speaker roles and gender. Therefore, we propose SamaVaani, a simple yet effective fairness-aware fine-tuning framework combining contrastive learning and CTC alignment, which significantly improves both transcription accuracy (up to $\sim$ 50% WER reduction) and fairness across demographic groups. Together, our study highlights critical limitations of current ASR systems in clinical settings and provides actionable pathways toward more equitable and robust multilingual ASR deployment in healthcare.

2 Related work

Research on ASR in Indian multilingual psychiatric settings spans *three* interconnected areas: (i) ASR for psychiatric interviews, (ii) Bias and speaker-level disparities in ASR, and (iii) ASR for Indian languages and accents. We highlight the challenges and gaps in each area that motivate the current study.

ASR for psychiatric interviews: A growing set of studies has examined the use of ASR to transcribe or analyze psychiatric interviews. Ciampelli et al. (2023) and Just et al. (2025) evaluated ASR in schizophrenia patients, comparing them with healthy controls using interviews conducted in Dutch and German respectively. Several studies (Molina et al., 2025; Hwang et al., 2025) in Spanish and French psychiatric interviews in patients with psychosis has found high WER performance. Lastly, Seyedi et al. (2023) studied ASR in psychiatric interviews in American English in patients with depression vs. controls with past history of depression, and found no difference in WER in either group.

Bias and speaker-level disparities in ASR: Prior works has identified gender differences in YouTube ASR (Tatman, 2017), higher error rates for African-American speakers in commercial ASR systems (Koenecke et al., 2020), and accent-based inequity among non-native German speakers in clinical contexts (Just et al., 2025). Biases related to age, gender, accents and low-resource languages also affect the performance of ASR (Feng et al., 2024). However, these studies are not designed for multilingual psychiatric interviews, where conversational roles are inherently unequal with clinicians typically producing longer, more structured speech and patients providing shorter and more uncertain responses.

ASR for Indian languages and accents: Recent work has focused on ASR challenges for Indian English and regional Indian languages. *Svarah* had significantly higher WER for English with Indian accent than for LibriSpeech (Javed et al., 2023b). Rai et al. (2024) observed substantial disparities in gender, region, and speech rate by analyzing 8,740 hours of NPTEL's Indian lecture in English. Considering Indic languages the large datasets **Indic-**

SUPERB and **IndicVoices** further highlight the linguistic, morphological, and prosodic diversity of Indian languages which pose challenges to ASR (Javed et al., 2022, 2024). However, they do not contain clinical conversations and analysis of error types. In this regard, the Eka Medical ASR Evaluation dataset (Care, 2024) offers valuable Indian accents and drug vocabulary with over 3,900 recordings but is limited to brief, static conversations. On the other hand, the **DISPLACE-M** dataset contains 55 hours of annotated conversational speech in the healthcare domain. However, this dataset lacks interviews of psychiatric cases.

3 Dataset

The data for this study comes from a tertiary teaching hospital dedicated to the treatment of psychiatric and neurological conditions. The hospital provides free inpatient and outpatient treatment for economically disadvantaged patients, and thus, a majority of beneficiaries are from such a background. We collected 202 audio recordings of patient and doctor/therapist interactions. These recordings were collected using Android mobile phones in mp3 format. While an attempt was made to make recordings in a quiet environment, no special arrangements were made for this. Therefore, the data represent a real-world setting where recordings are made in busy wards and outpatient department rooms. The language, duration and lexical diversity of the dataset are summarised in Table 1 and the speaker profiles are detailed in Table 2.

This dataset contains speech from 130 unique speakers, including 7 doctors/therapists and 123 patients. All conversations are between two individuals, making 202 unique doctor-patient/therapist-patient pairs.

Preprocessing: All recordings were listened to by two psychiatrists to ensure they were intelligible. It was ensured that the recordings did not contain the name of the patient or any numerical identifier like phone number, etc. However, we did not exclude segments that contained names of places, dates, etc., as we wish to evaluate if such named entities (Section 7) lead to more errors in ASR.

Annotation: As part of earlier research, transcripts in native languages were available for 103 of these recordings. For the remaining transcripts (99), two psychiatrists transcribed the recordings. The annotation guidelines for transcribing speech into text are given in the Appendix C with examples for each of the three languages.

4 Methodological details

In this section, we first note the base ASR models that we use to perform the audit. We also discuss different fine-tuning approaches to improve the WER and the fairness of the base models. Finally, we introduce the debiasing algorithm used to develop the SamaVaani framework.

4.1 ASR models

Base models: We evaluate a total of *eight* ASR models, as mentioned earlier. These include IndicWhisper (AI4Bharat, 2024), WhisperLargeV3 (OpenAI, 2024), Sarvam (SarvamAI, 2025), GoogleS2T (Cloud, 2023), Gemma3n (Team et al., 2025), OmniLingual (team et al., 2025), Vaani (ARTPARK-IISc, 2025), and Gemini (Google DeepMind / Google AI, 2025). Of these, GoogleS2T, Sarvam, and Gemini are proprietary models inferenced through APIs, while others are open-source. Recall that we have the audio files in an interview format where each of them has exactly two speakers, i.e., the patient and the doctor. We generate transcripts for each of these long-form audio files. Couple of these ASR models (Sarvam's Saarika-2.5 and Gemini) can generate transcripts with long-form audio as input while the other models (IndicWhisper, WhisperLargeV3, Vaani, Gemma3n and GoogleS2T) can only transcribe audio in 30-second chunks.

Fine-tuned models: One of the straightforward ways to improve the overall WER and fairness across demographic groups is fine-tuning the base models. For this we choose two of the best performing open-source ASR models – Gemma3n[1] and OmniLingual that have the best scores across groups (see Section 6). We perform two types of fine-tuning

[1] https://huggingface.co/google/gemma-3n-E4B-it

Characteristics	Overall N = 202*	English N = 54*	Hindi N = 78*	Kannada N = 70*	p-value#
Duration (minutes)	30.9 (6.1, 45.6)	36.5 (28.7, 51.1)	25.3 (4.5, 36.5)	27.2 (3.6, 45.2)	<0.001
Total words	4756.5 (1042, 6111.5)	5636.5 (4331, 6216.8)	3877.5 (845.8, 7135)	3637.0 (526, 4873)	<0.001
Unique words	1152.0 (407.5, 1412.2)	1269.5 (1028.8, 1404.2)	885.5 (316.2, 1185.5)	1450.5 (248.8, 1759.8)	<0.001
Moving average Type-token ratio (window = 100)	0.64 (0.61, 0.67)	0.61 (0.59, 0.62)	0.64 (0.62, 0.66)	0.69 (0.67, 0.71)	<0.001

Table 1: Dataset summary. (*) indicates median (Q1, Q3) and (#) indicates Kruskal-Wallis rank sum test.

Characteristics	Overall N = 202*	English N = 54*	Hindi N = 78*	Kannada N = 70*	p-value#
Patient's gender					
F	51 (25.2%)	2 (3.7%)	18 (23.1%)	31 (44.3%)	<0.001
M	151 (74.8%)	52 (96.3%)	60 (76.9%)	39 (55.7%)	<0.001
Doctor's gender					
F	104 (51.5%)	54 (100%)	30 (38.5%)	20 (28.6%)	<0.001
M	98 (48.5%)	0 (0%)	48 (61.5%)	50 (71.4%)	<0.001
Patient's education level					
<Graduate	132 (65.3%)	18 (33.3%)	58 (74.4%)	56 (80.0%)	<0.001
>=Graduate	70 (34.7%)	36 (66.7%)	20 (25.6%)	14 (20.0%)	<0.001

Table 2: Summary of speaker profiles. (*) Median (Q1, Q3), (#) Kruskal-Wallis rank sum test.

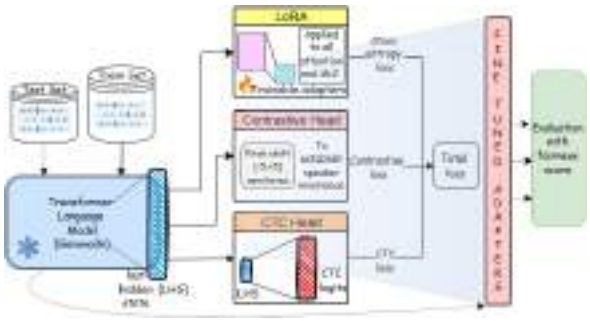

Figure 1: Architecture for proposed fine-tuning pipeline illustrating LoRA, contrastive and CTC head. (→) indicates fine-tuning flow while (⇢) shows inference on test set.

as follows.

$\text{FT}^{\text{STD.}}$: This refers to the standard LoRA fine-tuning using part of our dataset for training.

FT^{PS}: Here we double the dataset by augmenting the pitch of the audio files. The hypothesis is that this augmentation of synthetic data would result in stronger fine-tuning, and therefore better WER and fairness. For our experiments, we have used PitchShift[2] to augment our original audio by randomly selecting semitones in the range of $[-5, +5]$.

[2] https://docs.pytorch.org/audio/main/generated/torchaudio.transforms.PitchShift.html

4.2 The SamaVaani architecture

While straightforward fine-tuning can potentially improve the WER and the fairness of the models they do not specifically target the issue of debiasing. Recall that the transformer based ASR models follow the modern encoder decoder paradigm where the encoder learns the audio representation and the decoder follows the next word prediction task to generate transcription. Our approach integrates a contrastive learning and a CTC (connectionist temporal classification) (Graves et al., 2006) head in the decoding stage while fine-tuning. We only train the LoRA adapters to adapt for Indic languages, freezing the transformer layers as shown in Figure 1. The entire architecture and the associated loss functions are discussed below.

Contrastive learning: There are several studies that shows the effectiveness of improving fairness when using a contrastive learning approach (Koudounas et al., 2024; Shen et al., 2021; Wang et al., 2022; Ye et al., 2022). As in the case of FT^{PS} setup, here also we have used the PitchShift algorithm to augment our original audio by randomly selecting semitones in the range of $[-5, +5]$. Thus, in a batch of N audio samples, we have one original sample and its corresponding pitch-shifted sample constituting a positive pair, and the rest of

the $N-1$ audio samples are naturally considered as negative pairs. This forces the model to learn the same utterance regardless of the pitch (a key point of distinction between demographic groups). Further, this doubles the training data and makes the model robust to the high phonetic variance found in languages like Kannada. For an anchor representation z_a and its pitch-shifted positive pair z_p, with $N-1$ remaining negative original samples, the contrastive loss for each sample is formulated as:

$$\mathcal{L}_{CL} = -\log \frac{e^{\frac{\text{sim}(z_a, z_p)}{\tau}}}{e^{\frac{\text{sim}(z_a, z_p)}{\tau}} + \Sigma_{k=1}^{N-1} e^{\frac{\text{sim}(z_a, z_{n,k})}{\tau}}} \tag{1}$$

where $z_{n,k}$ are the other $N-1$ audio samples in the batch which are considered as negative pairs and τ is the temperature that dictates the sharpness of the probability distribution, ensuring the model is heavily penalized for any phonetic overlap between the anchor and the negative samples. The contrastive loss implemented is a variation of the NT-Xent (normalized temperature-scaled cross entropy) loss (Ågren, 2022) which is the standard objective for self-supervised frameworks like SimCLR (Chen et al., 2020). Here, we only create the pitch augmentation for the anchor file.

CTC head: In this stage, the logits from last hidden state (H) of transformer layers is projected to the vocabulary space (V). By projecting the last hidden states through the CTC head, the model receives a secondary signal that rewards correct character-level sequencing. This head is initialized from scratch and is fully trainable ($W_{CTC} = H \times V$) with CTC loss function represented as,

$$\mathcal{L}_{CTC} = -\log P(y \mid x) \tag{2}$$

where $x = (x_0, x_1, \ldots, x_T)$ denotes the input sequence of audio features, T is the time steps, $y = (y_0, y_1, ..., y_m)$ is the target label sequence text (transcription), and $P(y \mid x)$ is the total probability of correct transcription over various alignment paths. The CTC loss enables alignment-free training by allowing the model to learn mappings between input frames and output sequences without explicit frame-level labels.

Overall loss function: The weighted sum of the standard cross entropy (CE) loss, the contrastive loss (CL) and the CTC loss constitutes the final loss as follows.

$$\mathcal{L}_{total} = \alpha \times \mathcal{L}_{CE} + \beta \times \mathcal{L}_{CL} + \gamma \times \mathcal{L}_{CTC} \tag{3}$$

where α, β, and γ are the weights for each of the loss components. These weights are optimized using OPTUNA[3] over 20 trials. The best values of α, β, and γ are selected based on the WER score on the validation set.

LoRA adaptation: We train all the attention and MLP modules of the transformer based on the above loss function in eq. (3). This allows the model to repurpose the internal representations for Indic languages with less compute. In addition, it contributes toward the causal cross entropy loss for next word prediction in the speech to text task. Finally, it ensures that the model retains the ability to generate contextually and grammatically correct sentences in text from the speech in a multilingual setting.

5 Experimental setup

5.1 Evaluation metric

Performance metric: The main indicator used to assess the accuracy of ASR transcription is the word error rate (WER). Since WER is the most widely used metric for assessing ASRs and has been utilized by several researchers in the literature, we have chosen it as the evaluation metric. The WER is mathematically defined as the ratio of the sum of substitutions (S), deletions (D), and insertions (I) to the total number of words (N) in the reference transcript.

$$\mathcal{WER}\% = \frac{S + D + I}{N} \times 100 \tag{4}$$

First, we normalize both the ground truth and the generated transcripts by preprocessing the text. This preprocessing includes lowercasing, removing punctuation, and standardizing numbers to reduce superficial mismatches. Second, we use the JiWER Python library for the calculation of $\mathcal{WER}$.

Fairness metric: The fairness score ($\mathcal{FS}$) combines a couple of components, namely, the $\mathcal{WER}$ gap and the average $\mathcal{WER}$ among groups (say two groups – $\mathcal{G}_1$ and $\mathcal{G}_2$).

$$\mathcal{FS} = -\delta \times \mathcal{WER}_{avg} - \theta \times \mathcal{WER}_{gap}; \delta, \theta >= 0 \tag{5}$$

[3] https://optuna.org/

where $\mathcal{WER}_{avg}$ and $\mathcal{WER}_{gap}$ are the average $\mathcal{WER}$ of the two groups ($\mathcal{G}_1$, $\mathcal{G}_2$) and the absolute difference in $\mathcal{WER}$ between the groups ($\mathcal{G}_1$, $\mathcal{G}_2$) respectively. A lower $\mathcal{WER}_{gap}$ suggests better fairness across groups. In addition, a higher value of $\mathcal{FS}$ (ranging from -∞ to 0) indicates overall balanced performance across groups. In this study, we set $\delta = \theta = 0.5$ to assign equal importance to both units.

5.2 Implementation details

We employed GEMMA3N and OMNILINGUAL multimodal transformer-based models as the backbone for all of our experiments. We implement LoRA by injecting trainable rank 8 decomposition matrices into the query (q_{proj}), key (k_{proj}), value (v_{proj}), output (o_{proj}) and MLP projection layers ($gate_{proj}$, up_{proj}, and $down_{proj}$). The experimental configuration and the different hyperparameters are noted in Appendix A.

6 Results

Models	WER (%)	(S, D, I) %
SARVAM	34.33	(8.94 ,16.40, 36.87)
	39.03	(14.47, 6.76, 39.15)
	54.37	(27.87, 17.68, 39.03)
GOOGLES2T	74.60	(36.13, 40.22, 9.71)
	85.55	(10.09, 75.60, 0.13)
	94.90	(11.34, 84.01, 0.005)
GEMINI	**14.15**	(5.28, 18.58, 4.94)
	18.52	(11.32, 4.35, 8.74)
	35.01	(22.71, 16.86, 7.53)
WHISPERLARGEV3	46.76	(9.48, 19.81, 21.76)
	71.68	(26.23, 39.17, 9.21)
	98.55	(22.22, 76.51, 0.0014)
INDICWHISPER	-	-
	70.3	(23.44, 45.58, 4.27)
	97.05	(29.03, 68.21, 0.02)
VAANI	-	-
	44.42	(21.54, 15.60, 6.29)
	77.21	(5.56, 24.92, 1.73)
GEMMA3N	<u>40.22</u>	(14.63, 17.64, 19.54)
	48.14	(19.88, 6.45, 26.48)
	90.90	(48.57, 16.67, 30.47)
OMNILINGUAL	58.64	(23.40, 31.20, 4.00)
	<u>43.55</u>	(23.76, 12.20, 6.53)
	<u>75.35</u>	(41.08, 32.19, 2.06)

Table 3: Overall performance of ASR models across English , Hindi , Kannada . S, D, I (%) denote median insertion, deletion, and substitution components of WER. Best results among closed and open-source models are **bold** and <u>underlined</u> respectively.

This section is divided into two major parts. In the first part we audit the performance of the eight ASR models across three languages – Hindi, Kannada and Indian English. Next, we present the performance of the different fine-tuning methods as well as the debiasing algorithm SAMAVAANI proposed by us.

Audit outcomes: Table 3 reports the WER and its constituent parts, substitution (S), deletion (D), and insertion (I) in percentages, to compare ASR models in all the three languages. As we observe from the table, GEMINI achieves the best WER scores (English: 14.15%, Hindi: 18.52%, Kannada: 35.01%), demonstrating better multilingual and generalization abilities. Among the three languages, Kannada poses relatively higher challenge to the ASR models possibly due to the lack of enough pre-training data. Further, models like WHISPERLARGEV3, GOOGLES2T, and GEMMA3N show a very high WER for Kannada because their architectures and training corpora are not deeply optimised for rich Indian phonetic diversity and retroflex phonemes in Dravidian languages.

Fine-tuning and debiasing results: For fine-tuning, we choose GEMMA3N and OMNILINGUAL as both have the lowest $\mathcal{WER}$ averaged over three languages. We split the data into train, development (dev) and test folds with language as a stratification variable. Train set had 83.16 hours of audio (English 30.34, Hindi 27.56, Kannada 25.26), dev set had 9.64 hours (English 3.52, Hindi 3.02, Kannada 3.10) and test set had 10.22 hours (English 3.80, Hindi 2.96 and Kannada 3.46 hours).

Main results: We present the results from the different fine-tuned models in Table 4. The results are organized under different categories including overall $\mathcal{WER}$, language-wise $\mathcal{WER}$ and various demographic-wise $\mathcal{WER}$. Further we also present the fairness score ($\mathcal{FS}$) for each demographic group. The key observations are as follows.

1. We observe that SAMAVAANI results in a reduction of 50% in overall $\mathcal{WER}$ when compared to the BASE pre-trained model. The overall $\mathcal{WER}$ of SAMAVAANI is also substantially better than standard fine-tuning setups $\text{FT}^{\text{STD.}}$ and FT^{PS}.
2. Across all the three languages SAMAVAANI is remarkably better than the BASE model as well as $\text{FT}^{\text{STD.}}$ and FT^{PS}. Among the three languages, even SAMAVAANI struggles the most with Kannada, like all the other models.

Metric	Base		$FT^{Std.}$		FT^{PS}		FT^{CL}		FT^{CTC}		SamaVaani	
Overall WER (↓)	70.47	65.85	47.62	47.41	41.14	45.83	39.08	40.7	37.92	38.12	**35.19**	**35.29**
English WER (↓)	57.92	47.02	25.60	26.02	24.13	24.13	23.17	24.00	22.22	22.44	**20.43**	**20.54**
Hindi WER (↓)	58.33	63.44	44.34	43.95	38.80	42.31	37.25	38.23	36.36	36.00	**33.08**	**32.96**
Kannada WER (↓)	84.48	90.36	75.00	75.00	61.11	72.36	58.65	60.94	57.82	56.41	**52.84**	**52.88**
Male (M) (↓)	69.23	76.55	53.33	53.19	44.23	50.00	41.77	42.31	40.71	39.76	**37.25**	**37.25**
Female (F) (↓)	71.43	74.62	48.53	48.37	42.04	42.86	39.41	42.31	38.33	39.08	**34.83**	**35.48**
$\mathcal{FS}$: M vs F (↑)	-36.27	-43.76	-27.87	-27.80	-22.66	-26.79	-21.48	-22.46	-20.95	-20.05	**-19.23**	**-19.07**
Patient (P) (↓)	76.47	84.44	53.85	53.85	47.62	49.04	45.45	47.49	44.71	43.40	**41.67**	**41.18**
Doctor (D) (↓)	63.16	66.92	50.00	50.00	41.54	46.18	39.24	41.46	37.60	38.03	**34.40**	**34.69**
$\mathcal{FS}$: P vs D (↑)	-41.56	-51.60	-27.88	-27.88	-25.33	-26.48	-24.28	-25.25	-24.13	-23.04	**-22.65**	**-22.21**
>=Graduate (↓)	63.64	75.23	28.45	28.28	26.66	28.45	25.16	27.27	25.16	26.43	**22.28**	**22.42**
<Graduate (↓)	80.00	83.44	70.43	70.47	61.53	64.80	59.17	62.17	58.82	57.14	**53.64**	**53.39**
$\mathcal{FS}$: >=G vs <G (↑)	-44.09	-48.07	-45.71	-39.81	-39.49	-38.90	-38.09	-36.81	-37.83	-36.25	**-34.66**	**-34.44**

Table 4: Comparison of WER scores of Gemma3n and OmniLingual across different fine-tuning techniques and different demographic attributes. Lower WER (↓) and higher $\mathcal{FS}$ (↑) indicates better performance. FT^{CL}: An ablation of SamaVaani where only the contrastive loss is used. FT^{CTC}: An ablation of SamaVaani where only the CTC loss is used. **Bold** indicates the best score obtained from one model.

3. For all the demographic groups SamaVaani is not only better in terms of the group-wise $\mathcal{WER}$ but also in terms of $\mathcal{FS}$ when compared to Base, $FT^{Std.}$ and FT^{PS}.

Ablation study: A natural question in the design of SamaVaani regards the necessity of both the CL and CTC heads. In order to check whether any one of them is as good as the combination, we present in columns 4 and 5 of Table 4 the results from two additional fine-tuning setups – (i) FT^{CL}: an ablation of SamaVaani where only the contrastive loss is used and (ii) FT^{CTC}: an ablation of SamaVaani where only the CTC loss is used. We observe that though these models outperform the standard fine-tuning setups $FT^{Std.}$ and FT^{PS} in terms of both $\mathcal{WER}$ and $\mathcal{FS}$, they are not as good as SamaVaani. This quantitatively justifies the benefit of combining the two loss terms. In the next section, we present some qualitative insights into the advantages of each of these components.

7 Discussion

In this section, we discuss representative qualitative advantages of the CL and CTC loss terms (for Gemma3n[4]) and finally discuss some error cases. We posit that while the CTC head provides essential character alignment from speech to text, the contrastive learning objective acts as a phonetic regularizer.

[4]Similar observations hold for OmniLingual.

Role of contrastive learning: Recall that we use PitchShift to obtain a pitch-shifted variant of the anchor audio. With this, the objective of contrastive learning is to recognize the semantic equivalence between the original audio and its pitch augmented variant to ignore acoustic noise and focus on phonetic content. By restricting the augmentation to a single anchor-positive pair (z_i, z_i^+) within a pool of $N-1$ negative samples (z_k), the framework creates an asymmetric learning signal that is highly effective for phonetically dense languages. We set the temperature (τ) at 0.05 to sharpen the distribution, forcing the model to be very certain about the representations of samples in the latent space. $\mathcal{FS}$ improve consistently on all demographic attributes by 13-41% compared to the base model and 15-22% compared to the standard fine-tuned model.

Role of CTC head: Standard LLMs use auto regressive decoding, which are prone to hallucinations with same words repeating sequentially, for example, if if if if... (get caught in a word repeating indefinitely). The CTC heads enforces monotonic alignment. While the generative head ensures the sentence makes sense, the CTC head acts as a "sanity check" to ensure every word/token corresponds to an actual speech in the file. This balance significantly reduces instances of "word-skipping" or adding "filler" that wasn't in the original audio. In the raw audio signal, a single phoneme (like 's' in "speech") lasts for many frames. Without a special mechanism, a model might

Type	Transcript
Ground truth	friends, relatives. Yeah, if he values ... if he values that conflicted person a very high level in in his own mind ... it is uh, good decision that to worry ... for him ... that he didn't come. If he is a normal person who had a conflict, he won't mind that. He won't mind that.
BASE	friends, celebrities. Yeah, if if if if ... (repeated 210 times more)
$FT^{STD.}$	friends, celebrities. Yeah, if if if if ... (repeated 212 times more)
FT^{CL}	friends, celebrities. Yeah, if if if if ... (repeated 212 times more)
FT^{CTC}	friends, celebrities. Yeah, if if if you value if if you value that conflicted person a very high level in in his own mind, it is a good decision that to worry for him that he didn't come. If he is a normal person who had a conflict, he won't mind that. He won't mind that.
SAMAVAANI	friends cigarettes. Yeah, if you value if if you value that conflicted person a very high level in in his own mind, it is a good decision that to worry for him that he didn't come. If he is a normal person who had a conflict, he won't mind that. He won't mind that.

Table 5: Qualitative comparison of transcriptions across models of GEMMA3N. Here the speaker is a male patient.

predict the letter 's' many times in a row. CTC solves this using a unique decoding rule involving blank tokens (ϕ). The CTC decoding algorithm follows two simple but powerful rules to convert a long sequence of frame-by-frame predictions into a clean word as follows – (i) *collapse identical consecutive tokens*: if the model predicts 'aaaa-bbbb-cccc' due to slow speech, CTC collapses them into 'abc'; (ii) *blank as a separator*: to actually output two of the same letter (like the 'll' in 'hello'), the model must predict a blank token between them (e.g., h-e-l-ϕ -l-o).

Error analysis: Table 5 shows qualitative examples as to how auto-regressive models can get stuck in a loop where they start repeating the same word sequentially. Fine-tuning with only contrastive learning also fails to get out of the repeating loop. On the other hand, incorporating a CTC head on top LoRA is able to break out of this loop and generates better text. Furthermore, SAMAVAANI generates the best transcripts compared to ground truth. Not only it can get out of the indefinite loop, it also removes the same repeating words with a few inconsistencies. Lastly, the error in the transcripts generated by SAMAVAANI is essentially divided into the following three categories.

1. *Inverse text normalization*: The generation should appear in text as spoken word-by-word. For instance, "So, no matter what time I sleep, 8:00 o'clock is when I have to wake up." Here, the time should appear in words 'eight' as it is spoken and not as it is represented.
2. *Named entity errors*: There are several instances where the model fails to correct text for named entities, especially for organization and drug names. For example, the organization name 'NIMHANS' in ground truth is being substituted by terms like 'neeman's' or '2 months'. Similarily, the drug name 'benzodiazipines' is being transcribed as 'benzodiazepam pains', suggesting a lack of acoustic robustness, where the model struggles to align the correct token sequences.
3. *Deletion errors*: There are a few missing phonemes while transcribing speech to text. Forcing monotonic alignment with the CTC head prevents the model from skipping any word/token. It forces the model to attempt a phonetic transcription of every sound.

8 Conclusion

In this study, we highlighted the disparities in ASR models, particularly in a clinical psychiatric interview setting between a patient and a doctor. We comprehensively audited eight state-of-the-art models across three linguistically diverse languages and reported multiple disparities. Next, we propose SAMAVAANI that leads to simultaneous improvement of the overall $\mathcal{WER}$ as well as the demographic fairness. The key uniqueness lies in combining the two loss functions based on contrastive learning and CTC that serve complementary roles.

While our approach is better than the other fine-tuning setups, there is still a lot of scope for improvement, especially for languages like Kannada. The values suggest that it is important to build customised ASR models for a clinical psychiatric interview setting, pre-trained from scratch.

9 Limitations

While this study poses a solid foundation towards multilingual ASR in the psychiatric interview setting, there are a couple of limitations. ***First***, the experimental setup depends on LoRA fine-tuning with rank of 8 (*low*). This setup was chosen due to our hardware constraints. With a high-end infrastructure, a full supervised fine-tuning may further improve the transcription accuracy and fairness. And ***second***, our evaluation is limited only to Indian English, Hindi and Kannada psychiatric interviews. India contains substantial linguistic diversity, and ASR behaviour may differ considerably across other Indic languages, dialects, and code-mixed settings.

10 Ethical consideration

This study was approved by the Institute Ethics Committee. The speech data were collected from a tertiary teaching hospital with a specialised addiction treatment centre offering 24-hour emergency services dedicated to the treatment of psychiatric and neurological conditions, along with inpatient and outpatient services. Written informed consent was obtained from patients to audio-record psychiatric interviews. Although deidentified, the data cannot be made public as it contains highly sensitive personal health information, which can compromise patient privacy and confidentiality.

References

AI4Bharat. 2024. Indicwhisper: Multilingual asr model for indian languages. https://ai4bharat.iitm.ac.in/areas/asr. A transformer-based encoder–decoder architecture modified from Whisper and enhanced with multilingual acoustic–textual alignment and subword tokenization tailored for Indic scripts. Accessed 2025-11-21.

Murray Alpert, Enrique R. Pouget, and Raul R. Silva. 2001. Reflections of depression in acoustic measures of the patient's speech. *Journal of Affective Disorders*, 66(1):59–69.

Gerard Anmella, Michele De Prisco, Jeremiah B. Joyce, Claudia Valenzuela-Pascual, Ariadna Mas-Musons, Vincenzo Oliva, Giovanna Fico, George Chatzisofroniou, Sanjeev Mishra, Majd Al-Soleiti, Filippo Corponi, Anna Giménez-Palomo, Laura Montejo, Meritxell González-Campos, Dina Popovic, Isabella Pacchiarotti, Marc Valentí, Myriam Cavero, Lluc Colomer, Iria Grande, Antoni Benabarre, Cristian-Daniel Llach, Joaquim Raduà, Melvin McInnis, Diego Hidalgo-Mazzei, Mark A. Frye, Andrea Murru, and Eduard Vieta. 2024. Automated Speech Analysis in Bipolar Disorder: The CALIBER Study Protocol and Preliminary Results. *Journal of Clinical Medicine*, 13(17):4997. Publisher: Multidisciplinary Digital Publishing Institute.

ARTPARK-IISc. 2025. whisper-large-v3-vaani-hindi: Fine-tuned whisper asr for hindi. https://huggingface.co/ARTPARK-IISc/whisper-large-v3-vaani-hindi. Trained on 718 hours of Hindi speech from multiple datasets including Vaani, Gramvaani, IndicVoices, Fleurs and CommonVoice. Accessed 2025-11-29.

Microsoft Azure. 2023. Azure speech-to-text documentation. https://learn.microsoft.com/azure/ai-services/speech-service/. Accessed 2025-11-21.

Séverin Baroudi, Yanis Labrak, Shashi Kumar, Joonas Kalda, Sergio Burdisso, Pawel Cyrta, Juan Ignacio Alvarez-Trejos, Petr Motlicek, Hervé Bredin, and Ricard Marxer. 2026. Doctor or patient? synergizing diarization and asr for code-switched hinglish medical conditions extraction. *arXiv preprint arXiv:2603.06373*.

Sarah Basma, Bridgette Lord, Lindsay M Jacks, Mohamed Rizk, and Anabel M Scaranelo. 2011. Error rates in breast imaging reports: comparison of automatic speech recognition and dictation transcription. *American Journal of Roentgenology*, 197(4):923–927.

Eka Care. 2024. The Eka Medical ASR Evaluation Dataset.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations.

S. Ciampelli, A. E. Voppel, J. N. de Boer, S. Koops, and I. E.C. Sommer. 2023. Combining automatic speech recognition with semantic natural language processing in schizophrenia. *Psychiatry Research*, 325(115252).

Google Cloud. 2023. Speech-to-text documentation. https://cloud.google.com/speech-to-text. Accessed 2025-11-21.

Michael A. Covington, Congzhou He, Cati Brown, Lilia Naci, Jonathan T. McClain, Bess Sirmon Hjortdal, Adrianne Bassett, and Jerome Brown. 2005. Schizophrenia and the structure of language: The linguist's view. *Schizophrenia Research*, 77(1):85–98.

Arianna Di Florio et al. 2021. Acoustic and natural language markers for bipolar disorder: A pilot, mHealth cross-sectional study. *JMIR mHealth and uHealth*. PMC12017610.

Margarida Durão et al. 2025. Speech analysis for detecting depression in older adults: a systematic review. *npj Mental Health Research*. PMC12740927.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567.

Google DeepMind / Google AI. 2025. Gemini 2.5 pro — model card. `https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro`. Multimodal reasoning model with 1M-token context, supporting text, audio, image, video and code input. Accessed 2025-11-29.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 369–376. ACM.

Jinni A. Harrigan, Karen Wilson, and Robert Rosenthal. 1994. Detecting state and trait anxiety from auditory and visual cues: A meta-analysis. *Personality and Social Psychology Bulletin*, 20(3):346–357. Cited in: Moniz et al. (2023), Computer Speech & Language, for the role of repetitions, stutters, and filled pauses in detecting state anxiety.

Wolfram Hinzen and Joana Rosello. 2015. The linguistics of schizophrenia: thought disturbance as language pathology across positive symptoms. *Frontiers in Psychology*, 6:971.

Haeeul Hwang, Eric Jordan, Deok-Hee Kim-Dufor, Christophe Lemey, and Motasem Alrahabi. 2025. Evaluating asr in a clinical context: What whisper misses. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 374–378.

Tahir Javed, Kaushal Santosh Bhogale, Abhigyan Raman, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. IndicSUPERB: A Speech Processing Universal Performance Benchmark for Indian languages. ArXiv:2208.11761 [cs].

Tahir Javed, Sakshi Joshi, Vignesh Nagarajan, Sai Sundaresan, Janki Nawale, Abhigyan Raman, Kaushal Bhogale, Pratyush Kumar, and Mitesh M. Khapra. 2023a. Svarah: Evaluating English ASR Systems on Indian Accents. ArXiv:2305.15760 [cs].

Tahir Javed, Sakshi Joshi, Vignesh Nagarajan, Sai Sundaresan, Janki Nawale, Abhigyan Raman, Kaushal Bhogale, Pratyush Kumar, and Mitesh M. Khapra. 2023b. Svarah: Evaluating english asr systems on indian accents.

Tahir Javed, Janki Atul Nawale, Eldho Ittan George, Sakshi Joshi, Kaushal Santosh Bhogale, Deovrat Mehendale, Ishvinder Virender Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Sharad Gandhi, Ambujavalli R, Manickam K. M, C. Venkata Vaijayanthi, Krishnan Srinivasa Raghavan Karunganni, Pratyush Kumar, and Mitesh M. Khapra. 2024. IndicVoices: Towards building an Inclusive Multilingual Speech Dataset for Indian Languages. ArXiv:2403.01926 [cs].

Sandra Anna Just, Brita Elvevåg, Shrankhla Pandey, Ivan Nenchev, Anna-Lena Bröcker, Christiane Montag, and Sarah E Morgan. 2025. Moving beyond word error rate to evaluate automatic speech recognition in clinical samples: Lessons from research into schizophrenia-spectrum disorders. *Psychiatry Research*, 352:116690.

Katarzyna Kaczmarek-Majer et al. 2024. Acoustic features from speech as markers of depressive and manic symptoms in bipolar disorder: A prospective study. *Acta Psychiatrica Scandinavica*, 150(5):374–386.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689.

Alkis Koudounas, Flavio Giobergia, Eliana Pastor, and Elena Baralis. 2024. A contrastive learning approach to mitigate bias in speech models. *arXiv preprint arXiv:2406.14686*.

James A Mays and Patrick C Mathias. 2019. Measuring the rate of manual transcription error in outpatient point-of-care testing. *Journal of the American Medical Informatics Association*, 26(3):269–272.

José Tomás García Molina, Pablo A Gaspar, and Alicia Figueroa-Barra. 2025. Automatic speech recognition in psychiatric interviews: a rocket to diagnostic support in psychosis. *Revista Colombiana de Psiquiatría (English ed.)*, 54(4):624–631.

OpenAI. 2024. Whisper large-v3 model card. `https://platform.openai.com/docs/models/whisper`. A 1.55B-parameter transformer encoder–decoder ASR model using windowed inference on 30-second log-Mel spectrogram segments. Accessed 2025-11-21.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*. Accessed 2025-11-29.

Anand Kumar Rai, Siddharth D Jaiswal, and Animesh Mukherjee. 2024. A Deep Dive into the Disparity of Word Error Rates across Thousands of NPTEL MOOC Videos. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:1302–1314.

Sam O'Connor Russell, Iona Gessinger, Anna Krason, Gabriella Vigliocco, and Naomi Harte. 2024. What automatic speech recognition can and cannot do for conversational speech transcription. *Research Methods in Applied Linguistics*, 3(3):100163.

SarvamAI. 2025. Sarvam ai speech-to-text api documentation. `https://docs.sarvam.ai/api-reference-docs/api-guides-tutorials/speech-to-text/overview`. Accessed 2025-11-21. Saarika / Saaras ASR models support long-form, multilingual speech recognition with windowed inference and log-Mel input features.

Amazon Web Services. 2023. Amazon transcribe documentation. `https://docs.aws.amazon.com/transcribe/`. Accessed 2025-11-21.

Salman Seyedi, Emily Griner, Lisette Corbin, Zifan Jiang, Kailey Roberts, Luca Iacobelli, Aaron Milloy, Mina Boazak, Ali Bahrami Rad, Ahmed Abbasi, et al. 2023. Using hipaa (health insurance portability and accountability act)–compliant transcription services for virtual psychiatric interviews: Pilot comparison study. *JMIR Mental Health*, 10:e48517.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2021. Contrastive learning for fair representations.

Kiyoshi Shikino, Tomoko Tsukamoto, Kazutaka Noda, Yoshiyuki Ohira, Daiki Yokokawa, Yuta Hirose, Eri Sato, Tsutomu Mito, Takahiro Ota, Yota Katsuyama, Takanori Uehara, and Masatomi Ikusaka. 2023. Do clinical interview transcripts generated by speech recognition software improve clinical reasoning performance in mock patient encounters? A prospective observational study. *BMC Medical Education*, 23(1):272.

Michael H. Silber, Philip M. Becker, Mark J. Buchfuhrer, Christopher J. Earley, William G. Ondo, Arthur S. Walters, and John W. Winkelman. 2018. The Appropriate Use of Opioids in the Treatment of Refractory Restless Legs Syndrome. *Mayo Clinic Proceedings*, 93(1):59–67.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2020. A survey of code-switched speech and language processing.

Jae-hee So, Joonhwan Chang, Eunji Kim, Junho Na, JiYeon Choi, Jy-yong Sohn, Byung-Hoon Kim, Sang Hui Chu, et al. 2024. Aligning large language models for enhancing psychiatric interviews through symptom delineation and summarization: pilot study. *JMIR Formative Research*, 8(1):e58418.

John Sommers-Flanagan, Waganesh Abeje Zeleke, and ME Hood. 2015. The clinical interview. *The encyclopedia of clinical psychology. John Wiley & Sons, Inc, Hoboken*, pages 1–9.

Peter R. A. Stokes et al. 2023. Linguistic findings in persons with schizophrenia—a review of the current literature. *Frontiers in Psychology*, 14:1287706.

Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri,

Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages.

Bazen Gashaw Teferra, Sophie Borwein, Danielle D. DeSouza, William Simpson, Ludovic Rheault, and Jonathan Rose. 2022. Acoustic and linguistic features of impromptu speech and their association with anxiety: Validation study. *JMIR Mental Health*, 9(7):e36828. PMC9308078.

Yiming Wang, Jinyu Li, Heming Wang, Yao Qian, Chengyi Wang, and Yu Wu. 2022. Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7097–7101.

Yi Yang, Carol Fairbairn, and Jeffrey F. Cohn. 2012. Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing*, 4(2):142–150.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5099–5113, Seattle, United States. Association for Computational Linguistics.

Wilhelm Ågren. 2022. The nt-xent loss upper bound.

A Implementation details & hyperparameters

We implement LoRA by injecting trainable rank 8 decomposition matrices into the query (q_{proj}), key (k_{proj}), value (v_{proj}), output (o_{proj}) and MLP projection layers ($gate_{proj}$, up_{proj}, and $down_{proj}$) on the base ASR models. The experimental configuration and the different hyperparameters are noted in Table 6.

The loss coefficients (α, β, γ) for each of the three components in the total loss function for both the fine-tuned models, GEMMA3N and OMNILINGUAL are optimized through OPTUNA[5] over 20 trials. The exact coefficient values for GEMMA3N and OMNILINGUAL are (0.4135, 0.2186, 0.3679) and (0.4385, 0.2480, 0.3135) respectively.

B Acoustic features and fairness analysis of dataset

In this section, we analyze the acoustic features of the dataset used in this stdy. We evaluate several key features, which indicates how good the voice quality of the data is. Performance evaluations indicate that males, patients, and Kannada speakers are structurally disadvantaged subgroups which are disproportionately affected by higher Word Error Rates

[5] https://optuna.org/

Configuration	Value
LoRA configuration	
Rank (r)	8
Scaling factor (α_{LoRA})	16
Dropout	0.09
Target modules	$q_{\text{proj}}, k_{\text{proj}}, v_{\text{proj}}$, $o_{\text{proj}}, gate_{\text{proj}}$, $up_{\text{proj}}, down_{\text{proj}}$
Base model quantization	4-bit
Optimization	
Optimizer	AdamW (8-bit)
Base learning rate	5×10^{-5}
Learning rate schedule	Cosine
Learning rate warmup period	10% (for 3 epochs)
Max sequence length	1024 tokens
Temperature (τ)	0.05
Training infrastructure	
Per-device batch size	1
Gradient accumulation steps	8
Effective global batch size	32
GPUs	4 × NVIDIA A6000 (48 GB each)
Validation & early stopping	
Validation metric	Median WER
Early stopping patience	3 evaluation steps

Table 6: Common experimental configuration and hyperparameters.

($\mathcal{WER}$). Importantly, the performance gap stems from acoustic differences, not data size. Although males represent 65% of the dataset, their speech leads to worse WERs because of a naturally lower fundamental frequency ($r = 0.84$) and a lower voice quality, indicated by lower Harmonics-to-Noise Ratio (HNR) and higher amplitude instability (shimmer). Likewise, the speech of patients is a reflection of clinical realities such as psychomotor symptoms or emotional distress, which manifest acoustically as lower pitch and degraded voice quality. We depict the differences in pitch and voice quality of the speech data among gender and speaker role respectively from Figures 2 to 5. In addition, the illustrate the speech intelligibility analysis in Figure 6.

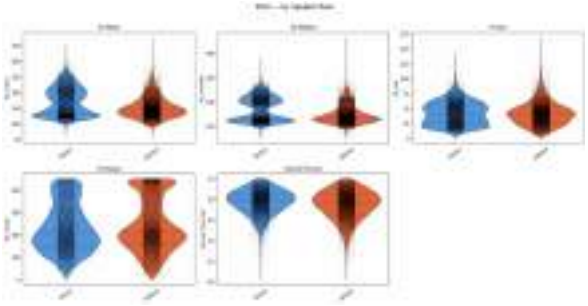

Figure 2: Pitch comparison based on the role of the speaker, i.e, doctor or patient.

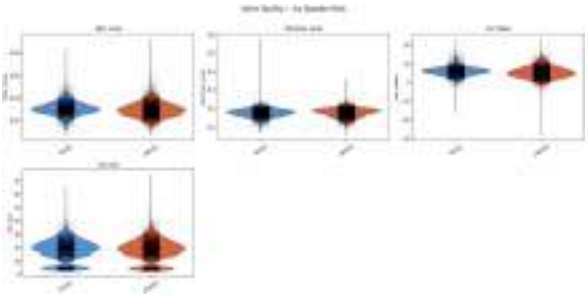

Figure 3: Voice quality comparison based on the role of the speaker, i.e, doctor or patient.

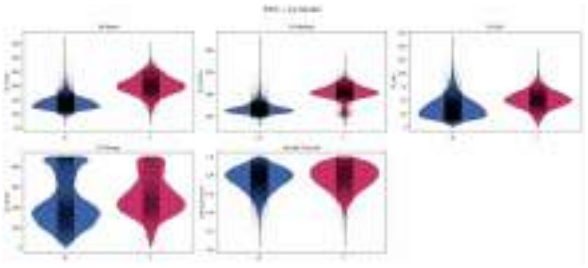

Figure 4: Pitch comparison based on the gender of the speaker, i.e, male or female.

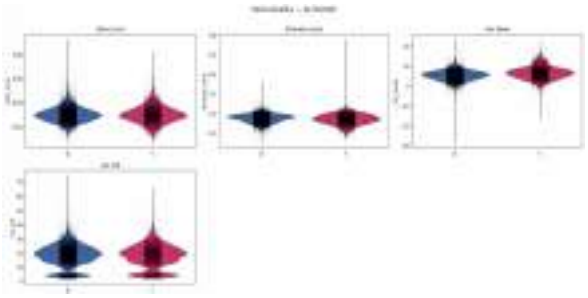

Figure 5: Voice quality comparison based on the gender of the speaker, i.e, male or female.

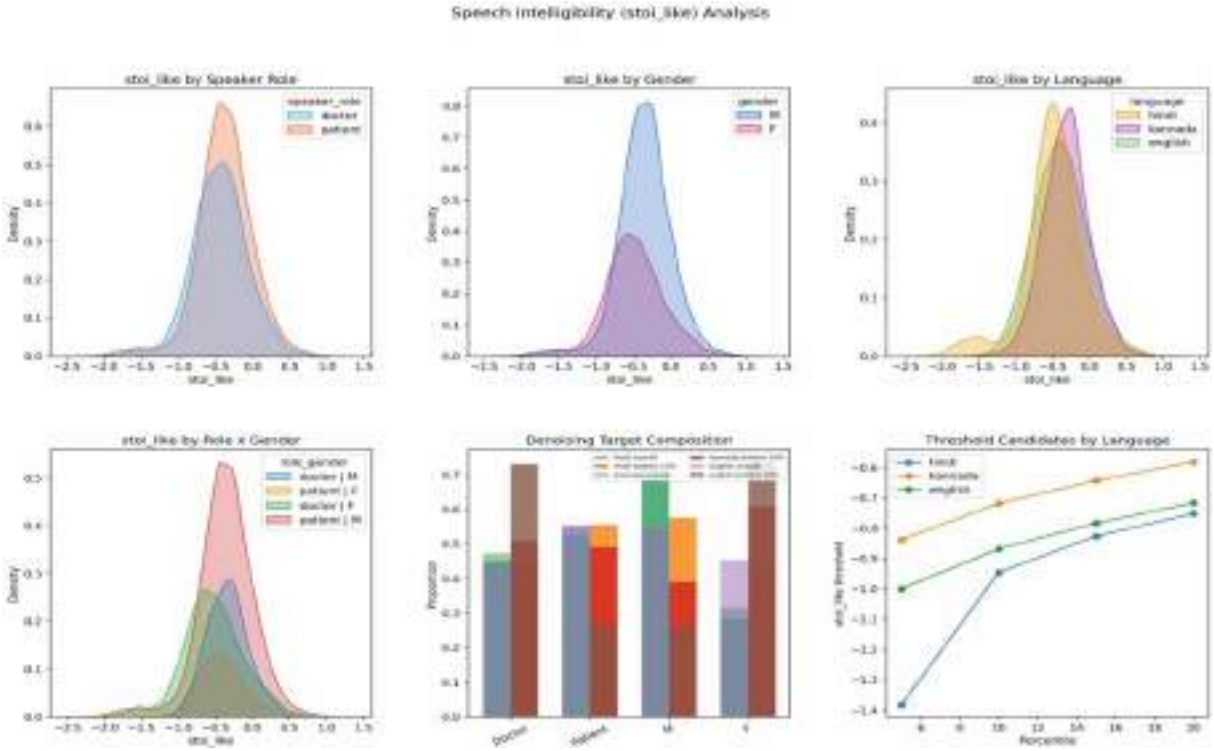

Figure 6: Speech intelligibility report of the speech data used in this study.

C Annotation guidelines

In this section, we detail the exact transcription instructions given to annotators with examples for each language.

Transcription instructions to annotators:
Your task is to carefully listen to the provided audio file and create a **.txt** or **.srt** file.
About the recordings:
• These are audio-recorded clinical interviews between doctors and patients.
• Accuracy of transcription is very important for clinical and research purposes.
• Please follow the following guidance material to ensure accurate transcriptions.
Carefully read the instructions below on how to do the transcriptions and their examples for each of the three language, namely, English, Hindi, and Kannada (in order).
Speaker diarisation:
Clearly distinguish between the speakers. Use the following labels at the beginning of each new turn of speech.
• For English → [Doctor:] [Patient:]
• For Hindi → [डॉक्टर:] [मरीज़:]
• For Kannada → [ವೈದ್ಯ:] [ರೋಗಿ:]
Use of punctuation:
Allowed punctuations are full-stop, question mark, comma, ellipsis, em-dash and exclamation marks.

Punctuation	**Guidance**
Full Stop (.)	Use for the end of a sentence.
Comma (,)	Use for short pauses and separating multiple items or phrases.
Ellipsis (...)	Use for long pauses.
Em-dash (–)	Use for interruptions or cut-offs. English example, Doctor: So what brings you h– Hindi example, चिकित्सक: तुम यहाँ क्यों – Kannada example, ಡಾಕ್ಟರ್: ನೀವು ಯಾಕೆ ಇಲ್ಲಿಗೆ –-
Question Mark (?)	Use for direct questions.
Exclamation (!)	Use a single mark to indicate strong emotions. English example, It just feels so bad I can't even tell you! Hindi example, मैं आपको बता नहीं सकता कि मैं कितना परेशान हूँ! Kannada example, ನಂಗೆ ಎಷ್ಟು ಬೇಜಾರ್ ಆಗುತ್ತೆ ಅಂದ್ರೆ ಹೇಳೋಕ್ಕೆ ಆಗಲ್ಲ!

Strict verbatim transcription:
Preserve all dysfluencies like pauses, filler words, stammers etc. Do not correct if there are grammatical errors in the speech itself. See table below.

Guidance	English examples	Hindi examples	Kannada examples
Capture filler words and phrases	Um, Yeah, Hmm, Mmmm, etc	अच्छा, हाँ, हम्म, हा, आ, etc	ಉಮ್, ಅಹ್, ಹಮ್ಮ್, ಹಾ, ಆ, etc
Include all incomplete phrases as they are said	I... I yesterday... no, the day before, I had gone there.	मैं...मैं कल...नहीं, परसों वहाँ गया था।	ನಾನು... ನಾನು ನಿ-ನ್ನೆ... ಇಲ್ಲ, ಮೊನ್ನೆ ಅಲ್ಲಿಗೆ ಹೋಗಿದ್ದೆ
Include stutters and stammers	I-I-I get scared.	मैं-मैं-मैं डर जाता हूं|	ನ-ನ-ನನಗೆ ಭಯ ಆಗು-ತ್ತೆ
Use ellipsis (three dots) for long pauses or incomplete sentences	I don't know… I feel very scared.	मुझे नहीं पता... मुझे बहुत डर लग रहा है|	ನನಗೆ ಗೊತ್ತಿಲ್ಲ... ತುಂ-ಬಾ ಭಯ ಆಗುತ್ತೆ
Transcribe as spoken, do not change informal to formal style (applies for languages like Kannada where spoken/colloquial and written styles are very different). Do not correct grammar or polish the output in any manner.	I am afraid → I am feeling scared	मुझे डर है → मुझे डर लग रहा है	ನಂಗೆ ಭಯ ಆಗ್ತಿದೆ → ನನಗೆ ಭಯ ಆಗುತ್ತಿದೆ
There can be frequent language mixing in these audios. You can expect English, Hindi, Telugu, Tamil and Malayalam. Write these words in their native script as of speech.	main samajh gaya, gottaytu	मैं समझ गया, गोथायतू	ಠೀಕ್ ಹೈ, ನಂಗೆ ಅರ್ಥ ಆಯಿತು [theek hai, nange artha aytu]
If there are numbers, type them out in words and not as numerals.	Incorrect: I took 10 tablets that day. Correct: I took ten tablets that day.	Incorrect: मैंने 10 गोलियाँ लीं / मैंने १० गोलियाँ लीं। Correct: मैंने दस गोलियाँ ले लीं	Incorrect: ನಾನು ಅವತ್ತು 10 ಮಾತ್ರೆಗಳು ತಗೊಂಡೆ / ನಾನು ಅವತ್ತು ೧೦ ಮಾತ್ರೆಗಳು ತಗೊಂಡೆ Correct: ನಾನು ಅವ-ತ್ತು ಹತ್ತು ಮಾತ್ರೆಗಳು ತಗೊಂಡೆ

Non-speech occurrences:
Five relevant non-speech sounds must be noted. Use square brackets [. . .] for these annotations. For example, when the discussion is about depression, sounds of the speaker crying becomes important. See the indicators below. Do not use any other indicators.

Type	English examples	Hindi examples	Kannada examples
Three Special tokens are allowed for emotional expressions. They must be in square brackets and are the only ones allowed.	[laughs], [cries], and [shouts]	[हंसने की आवाज़], [रोने की आवाज़] and [चिल्लाना]	[ನಗುವಿನ ಸದ್ದು], [ಅಳುವಿನ ಸದ್ದು] and [ಕೂಗುತ್ತಾ]
One special token is allowed for unclear speech	[unclear]	[अस्पष्ट]	[ಅಸ್ಪಷ್ಟ]
One special token is allowed for other noises **[noise]**. Use this for all non-human background sounds like phone ringing or buzzing, ambulance siren, car honks, furniture scraping, other mechanical sounds.	[00:03:16] Patient: I had come then, but [noise] could not meet him.	[00:03:16] मरीज़: मैं तब आया था, लेकिन [noise] मिला नहीं	[00:03:16] ರೋಗಿ: ಆವಾಗ್ಲೇ ಬಂದಿದ್ದೆ, ಆದ್ರೆ [noise] ಸಿಗ್ಲಿಲ್ಲ.

Timestamping and General Formatting:
Use uniform font size and line spacing. The document must be in .txt or .srt format. The transcript must be timestamped in [HH:MM:SS] format (hours: minutes: seconds). See the example below:

English examples	Hindi examples	Kannada examples
[00:00:05] Doctor: Please come in, have a seat, and tell me, how are you? [00:00:18] Patient: Hello, Doctor. I haven't been feeling well for the past one or two weeks and I feel sad all the time. [00:01:05] Patient: Yes, Doctor. I used to enjoy talking to my friends and gardening, but now I feel like just sitting alone all the time.	[00:00:05] डॉक्टर: आइए, बैठिए और बताइए कि आप कैसे हैं? [00:00:18] मरीज़: नमस्कार डॉक्टर साहब. पिछले एक-दो सप्ताह से मेरी तबीयत ठीक नहीं है और हमेशा उदास रहता हूं. [00:01:05] मरीज़: हाँ डॉक्टर, पहले मुझे अपने दोस्तों से बात करना और बागवानी करना अच्छा लगता था, लेकिन अब हर समय अकेले बैठे रहने का मन करता है.	[00:00:05] ವೈದ್ಯರು: ಬನ್ನಿ, ಕೂತ್ಕೊಳ್ಳಿ. ಹೇಳಿ, ಈವಾಗ ಹೇಗಿದ್ದೀರ? [00:00:18] ರೋಗಿ: ನಮಸ್ಕಾರ ಡಾಕ್ಟರ್. ಒಂದು ಎರಡು ವಾರದಿಂದ ಏನೋ ಸರಿ ಇಲ್ಲ. ಯಾವಾಗಲೂ ಬೇಜಾರು... ಒಂಥರಾ ಅನ್ಸುತ್ತೆ. [00:01:05] ರೋಗಿ: ಹೌದು ಡಾಕ್ಟರ್. ಫ್ರೆಂಡ್ಸ್ ಜೊತೆ ಮಾತಾಡೋದು, ಗಿಡ ನೋಡಿಕೊಳ್ಳೋದು ಅಂದ್ರೆ ಇಷ್ಟ. ಆದ್ರೆ ಈಗ ಏನೂ ಮಾಡೋಕೆ ಇಷ್ಟ ಆಗಲ್ಲ. ಸುಮ್ಮನೆ ಕೂತಿರ್ತೀನಿ ಅಷ್ಟೇ.

A speaker's entire turn must be kept in one line. Do not press ENTER (start a new line) in the middle of their speech even if it contains multiple sentences. Use line breaks between separate speakers. See example below.

English examples	Hindi examples	Kannada examples
Correct Doctor: When was the last time you slept well? Patient: I can't remember. It's been many months. **Incorrect** Doctor: When was the last time you slept well? Patient: I can't remember. It's been many months.	**Correct** चिकित्सक: पिछली बार आपको अच्छी नींद कब आई थी? मरीज़: मुझे याद नहीं. लगभग एक महीना हो गया. **Incorrect** चिकित्सक: पिछली बार आपको कब अच्छी नींद आई थी? मरीज़: मुझे याद नहीं. लगभग एक महीना हो गया.	**Correct** ವೈದ್ಯರು: ನೀವು ಕೊನೆಯದಾಗಿ ಯಾವಾಗ ಚೆನ್ನಾಗಿ ನಿದ್ದೆ ಮಾಡಿದ್ದೀರಿ? ರೋಗಿ: ನನಗೆ ನೆನಪಿಲ್ಲ. ಸುಮಾರು ಒಂದು ತಿಂಗಳಾಯ್ತು. **Incorrect** ವೈದ್ಯರು: ನೀವು ಕೊನೆಯದಾಗಿ ಯಾವಾಗ ಚೆನ್ನಾಗಿ ನಿದ್ದೆ ಮಾಡಿದ್ದೀರಿ? ರೋಗಿ: ನನಗೆ ನೆನಪಿಲ್ಲ. ಸುಮಾರು ಒಂದು ತಿಂಗಳಾಯ್ತು.