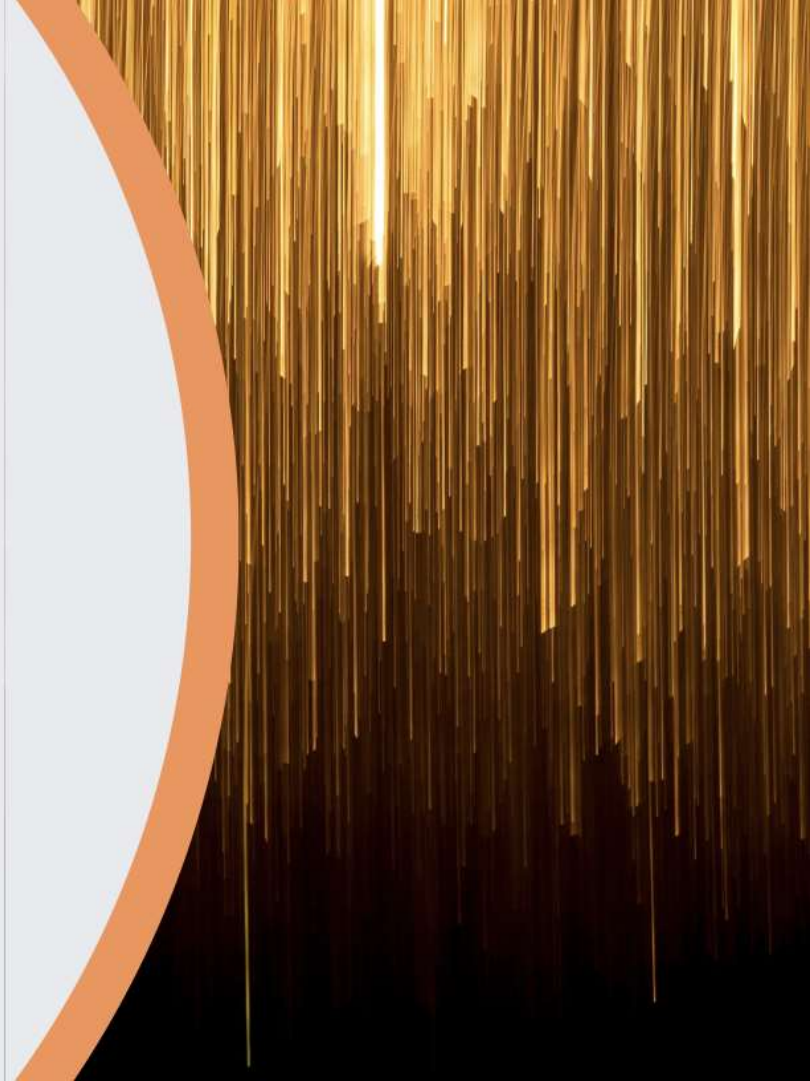
Dhwani-X: Kannada Speech Denoising & Transcription

Innovative pipeline for accurate Kannada speech transcription in noisy real-world environments. Presented by ByteBenders team, specializing in overcoming language resource constraints using intelligent preprocessing and robust models.

ByteBenders - Nikhil Y N and Nishitha Mahesh



Transcribing Kannada Speech in Noisy Environments

Resource scarcity and real-world noise degrade ASR accuracy without specialized preprocessing

Transcription becomes unreliable without **specialized preprocessing**

Noisy backgrounds (traffic, construction, crowds) severely degrade ASR accuracy

Limited data and pre-trained models for Kannada compared to English

Kannada ASR Challenges That Matter

Key language-specific obstacles that degrade transcription accuracy

Limited transcribed data for Kannada compared to English, reducing model training coverage

Acoustic uniqueness masked by noise — regional phonetics overwhelmed in real-world audio

Frequent code-mixing with English that confuses ASR pronunciation and language models

Noisy environments demand intelligent preprocessing to make transcriptions usable

Clean Before You Transcribe

Two-phase pipeline: intelligent denoising then ASR for noisy Kannada speech

Phase 1 — Intelligent denoising: noise classification, audio segmentation, targeted enhancement to improve clarity

Phase 2 — Transcription: enhanced audio fed to ASR model for more accurate Kannada transcription

Benefit: ASR performs **significantly better** when fed enhanced audio versus raw noisy input

Approach avoids direct use of noisy audio, prioritizing clarity before modeling

Evaluation Metrics & Methodology

Key metrics for audio quality, intelligibility, transcription accuracy and performance

SNR improvement — measures noise reduction magnitude (signal vs noise level)

PESQ — perceptual audio quality score reflecting subjective listening quality

STOI — speech intelligibility index for objective intelligibility assessment

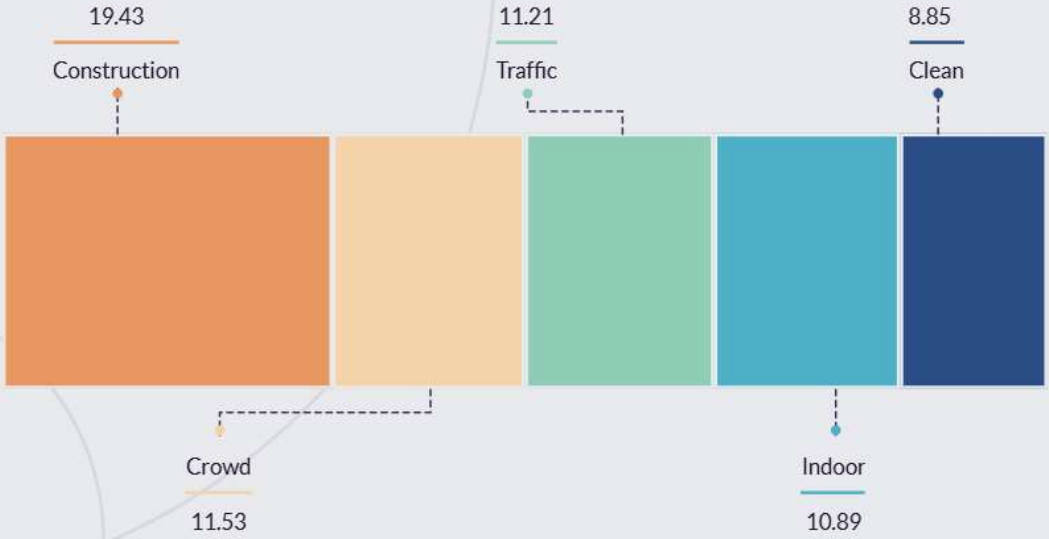
WER — Word Error Rate measuring transcription accuracy at word level

CER — Character Error Rate measuring fine-grained transcription errors

Real-time factor / Latency / Processing time — runtime performance and responsiveness

WER Improvements Across Noise Conditions

2.5-minute Kannada monologue — pipeline vs baseline



Pipeline reduces **WER** across all noise types; improvements range **10.6%–19.7%**

Clean: **8.85%** → **7.91%** (↓10.6%)

Traffic: **11.21%** → **9.17%** (↓18.2%)

Indoor: **10.89%** → **8.85%** (↓18.7%)

Crowd: **11.53%** → **9.49%** (↓17.7%)

Construction: **19.43%** → **15.61%** (↓19.7%) — enables usable transcription in harshest noise

Noise-Condition Performance Highlights

Per-condition gains, strengths and trade-offs for transcription quality

Traffic noise: pipeline filters low-frequency rumble; **18%** improvement in word boundary detection

Indoor: echo and reverb reduced effectively; **19%** overall gain

Crowd noise: adapts to varied non-stationary sounds; **18%** improvement; occasional over-processing

Construction noise: toughest condition; almost **20% WER** reduction, converts unusable to usable transcripts

Clean audio benefits from correction of informal language and loanwords

Audio Quality Metrics: SNR vs. Perceptual Measures

SNR changes minimal; PESQ/STOI better reflect intelligibility gains

Environment	Original SNR (dB)	Cleaned SNR (dB)	Change (dB)
Traffic	14.88	14.65	-0.22
Indoor	19.94	19.84	-0.09
Crowd	17.39	17.23	-0.16
Construction	10.66	10.28	-0.38

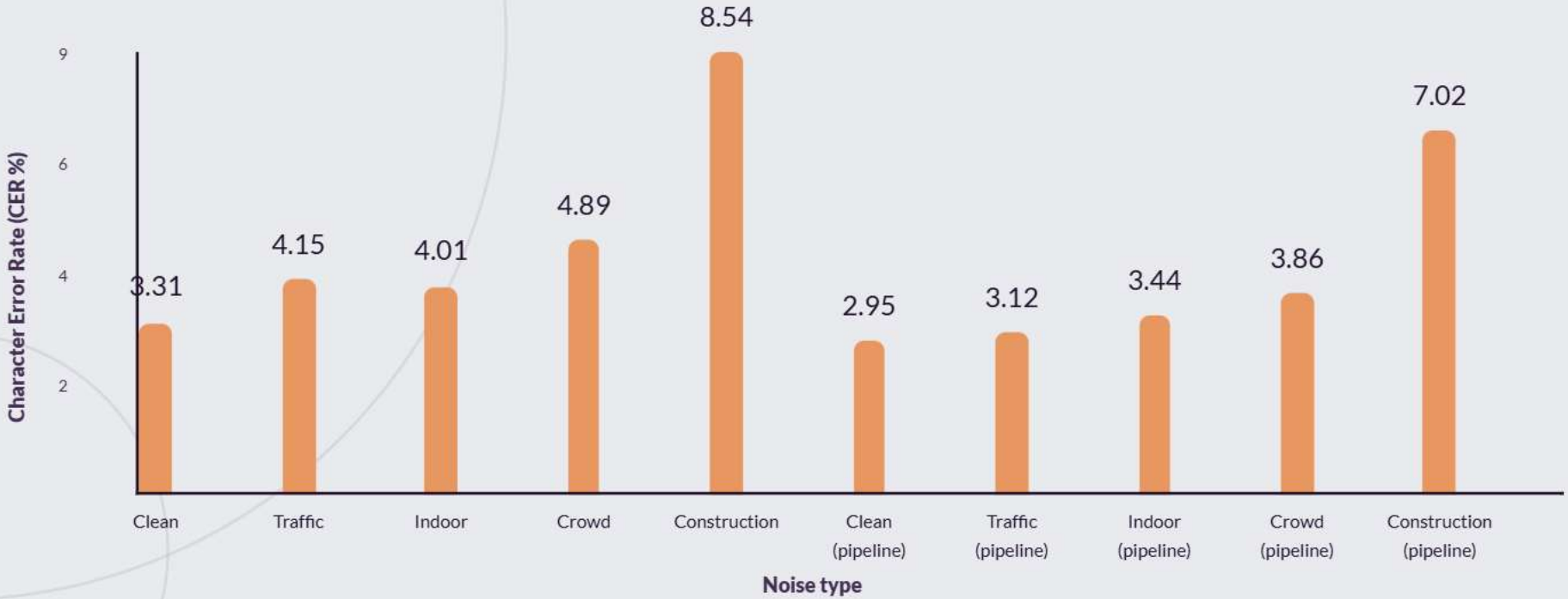
SNR improvements are minimal or slightly negative; denoiser trades raw signal strength for ASR-friendly noise removal

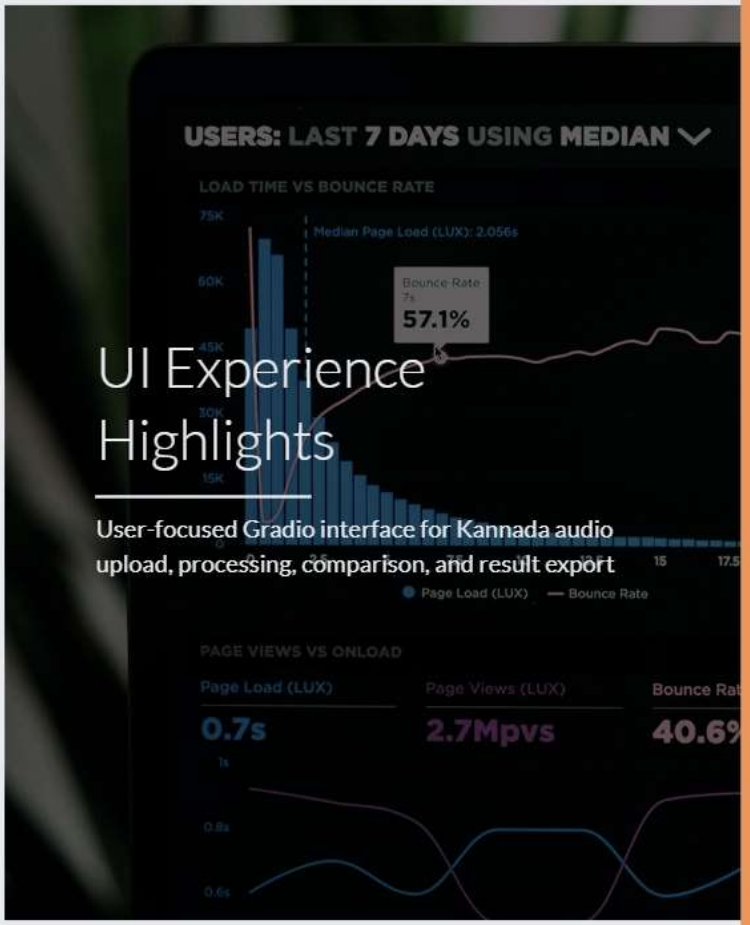
PESQ and STOI better capture **speech intelligibility** improvements than SNR

Interpretation: small negative SNR changes do not imply worse ASR performance when perceptual metrics improve

CER Improvements by Noise Type

Pipeline reduces character errors significantly across environments





UI Experience Highlights

User-focused Gradio interface for Kannada audio upload, processing, comparison, and result export

Upload any **Kannada** audio easily

Track progress through **processing stages**

Compare original and **denoised audio** side-by-side

View **baseline vs pipeline transcriptions**

Access detailed **metrics** and download results

Visualize detected **noise types**, speech segments, and processing times

Current Limitations and Impact

Key constraints that limit applicability, speed, and robustness

No real-time UI; live pipeline is command-line only

Single-speaker support; no multi-speaker diarization

Language restricted to Kannada despite architecture being language-agnostic

Requires GPU for acceptable speed

Incomplete noise adaptation despite Yamnet detection

Addressing these will broaden applicability and performance