Bengali Audio Deepfake Detection

Zia Ul Hasan Abdullah (2005037)

Prithu Anan (2005045)

Problem Statement

The Threat

- **Generative AI** (VITS, HiFi-GAN) can clone voices with near-perfect fidelity, creating 'Deepfake Audio'.

The Gap

- **Low-Resource Dilemma:** Bengali lacks massive forensic datasets. [1]
- **Linguistic Specificity:** Detectors fail on Bengali nuances (e.g., retroflex stops, aspirates).
- **Algorithmic Bias:** Detectors trained on one generator (e.g., VITS) fail against others (e.g., Diffusion).

The Goal

- Build a detector that is **robust** (handles noise/compression) and **algorithm-agnostic** (detects any fake).

Current Technology (SOTA)

Traditional Methods

- **MFCCs + GMM/SVM:** Rely on spectral features. Fail on neural vocoders like VITS which reconstruct high-frequency spectra perfectly. [1]

Deep Learning Baselines

- **RawNet2**: End-to-end 1D-CNN operating on raw waveforms. Good, but lacks global context. [2]
- **Wav2Vec 2.0 (XLSR)**: Self-supervised model pre-trained on 53 languages. Excellent for capturing linguistic anomalies. [3]
- **AASIST (Graph Attention Networks)**: Current SOTA. Models audio as a graph to detect inconsistencies between spectral and temporal domains. [4]

Dataset Strategy

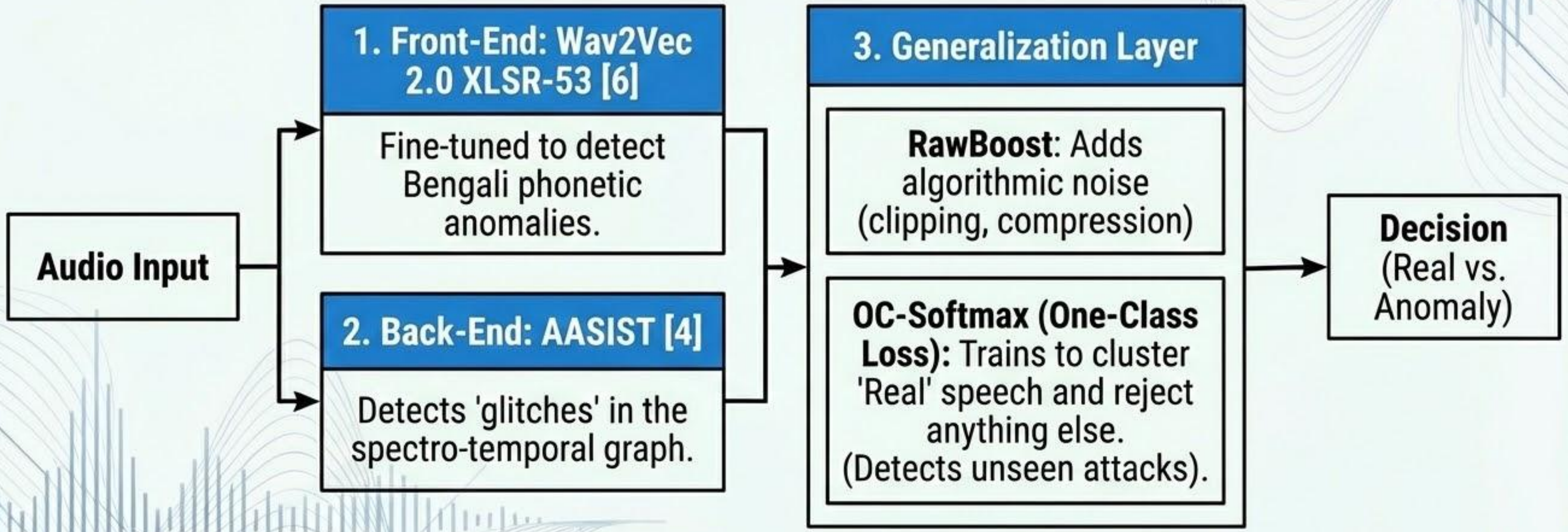
Core Dataset: BanglaFake [5]

- **Size:** 25,520 samples (12k Real / 13k Fake).
- **Source:** Real (SUST Corpus, Common Voice) vs. Fake (VITS trained on SUST).
- **Challenge:** "Matched Condition" attack—fake audio mimics the exact acoustic environment of real audio.

Proposed Expansion ('Bengali-Voice-Guard')

- To prevent overfitting to VITS, we augment with:
  - **Crikk**: Proprietary/Black-box commercial synthesis.
  - **Orpheus:** Transformer-based TTS (different artifact distribution than VITS).

Proposed Architecture (XLSR-AASIST-OC)



Evaluation Metrics

Primary Security Metric

- **Equal Error Rate (EER):** The point where False Acceptance Rate = False Rejection Rate.
- Target: < 5%. [7]

Cost-Sensitive Metric

- **min-tDCF:** Penalizes false acceptances (letting a deepfake through) more heavily than false rejections. [8]

Visual Validation

- **t-SNE Plots:** Must show clear separation between Real and Fake clusters (unlike the overlap in raw MFCCs).
- **Attention Maps:** Visualizing which which frequency bands the model focuses on to ensure it isn't overfitting to silence or background noise. [9]

Conclusion

Summary

We propose the first comprehensive, robust deepfake detection framework for Bengali.

Key Innovation

Moving from binary classification (Real vs. Fake) to One-Class Learning (Real vs. Anomaly) to ensure **future-proofing**.

Impact

Protects the Bengali information ecosystem against misinformation and fraud.

References

[1] Ayan, N. S., Dipto, M. A., et al. (2026). *Detecting Bangla DeepFake Audio: A Dual Approach Using LSTM and WaveNet*. Lecture Notes in Networks and Systems.

[2] Tak, H., et al. (2021). *RawBoost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing*. ICASSP.

[3] Baevski, A., et al. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. NeurIPS.

[4] Jung, J. W., et al. (2022). *AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks*. ICASSP.

[5] Fahad, I. A., Asif, K., & Sikder, S. (2025). *BanglaFake: Constructing and Evaluating a Specialized Bengali Deepfake Audio Dataset*. arXiv:2505.10885.

[6] Conneau, A., et al. (2020). *Unsupervised Cross-lingual Representation Learning for Speech Recognition (XLSR)*. Interspeech.

[7] [8] [9] Yamagishi, J., et al. (2021). *ASVspoof 2021: the 4th Automatic Speaker Verification Spoofing and Countermeasures Challenge*.

Thank You