DATA USAGE AGREEMENT FOR ACADEMIC AND RESEARCH USE OF BhashaBluff Dataset

Introduction

This research provides the BhashaBluff Dataset. Robust detection methods are essential to address the growing threat of audio deepfakes and synthetic media, which are increasingly appearing in complex, real-world conditions. Existing research has primarily focused on clean, single-language audio, overlooking critical challenges like bilingualism, background noise, and various forms of codec compression. This paper introduces the BhashaBluff dataset, with approximately 2,500 hours of audio data and over 3.8 million samples generated using 21 distinct methods. Its key innovation lies in its extensive variability across three critical dimensions: (i) bilingualism, which encompasses Hindi and English deepfakes as well as code-mixed speech; (ii) noise, with samples corrupted by four distinct environmental noise types; and (iii) compression, incorporating seven neural compression techniques to simulate real-world artifacts. Benchmark evaluations using state-of-the-art models show substantial performance degradation when faced with these challenging conditions, revealing the limited generalizability of current methods and highlighting the urgent need for more adaptive and robust algorithms. BhashaBluff serves as a crucial benchmark for developing next-generation detection systems that are effective in diverse, noisy, and compressed audio environments. The BhashaBluff dataset, along with our evaluation benchmarks, will be made publicly available for academic research to encourage further development in this domain.

Consent

The researcher(s) agree to the following conditions on the ACID Benchmark dataset:

1. The researcher(s) shall have **no rights** with respect to the BhashaBluff Database or any portion thereof and shall **not use** the Database except as expressly outlined in this Agreement.
2. **Re-identification is strictly prohibited.** All recipients agree that they will not attempt to re-identify any individual data subjects from the Dataset (e.g., speakers). Any re-identification of any individual data subject shall be immediately reported to the authors at **databases@iab-rubric.org**.
3. Subject to the terms and conditions of this agreement, the dataset is available for **academic and research use only**, with a royalty-free, nonexclusive, non-transferable license subject to the following conditions:
    - The Database must **not be copied, distributed, published, or reproduced** in any form except for creating a secure backup by the registered user. Sharing, transferring, or disclosing any part or the entirety of the dataset to third parties, in any form, is **strictly prohibited** without prior written authorization from the IAB Lab. Any individual or organization wishing to access or use the Dataset must independently register and agree to all terms and conditions outlined in this DUA.
    - The Dataset must **not be used for commercial purposes**.

4. Any violation of this DUA or other impermissible use shall be grounds for immediate termination of use of the Dataset.
5. Any work made public, whatever the form, based directly or indirectly on any part of the Database will include the following reference:
   Reference:
   Rishabh Ranjan, Mayank Vatsa, Richa Singh. BhashaBluff: A Dataset and Benchmark for Detecting Bilingual, Noisy, Compressed Deepfakes and Synthetic Audios

6. **Bibtex:**

```
None
@inproceedings{bhashabluff,
  author = {Rishabh Ranjan and Mayank Vatsa and Richa Singh},
  title = {BhashaBluff: A Dataset and Benchmark for
Detecting Bilingual, Noisy, Compressed Deepfakes
and Synthetic Audios},
  year = {2026}
}
```

I hereby agree to adhere to this license agreement's terms and conditions.

NAME and DESIGNATION (in capitals) SIGNATURE and DATE

Organization and ADDRESS