A Preliminary Case Study on Long-Form In-the-Wild Audio Spoofing Detection

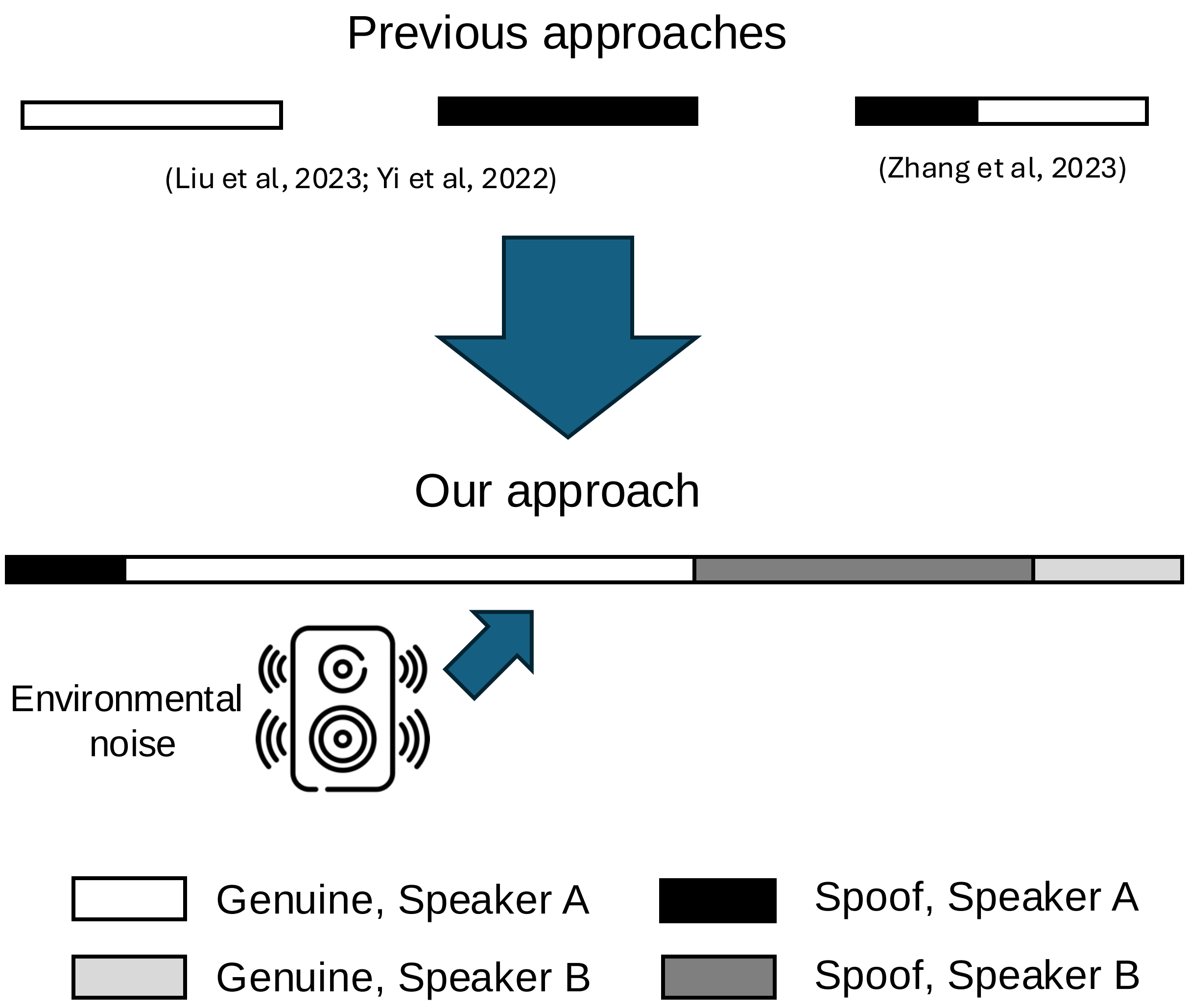
Xuechen Liu, Xin Wang, ***Junichi Yamagishi***

BIOSIG 2024

2024-09-26

Our Objective

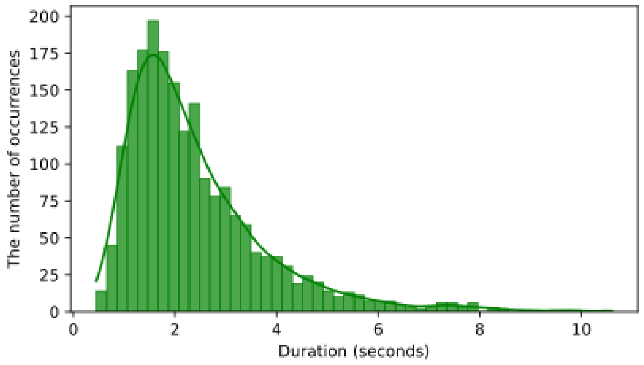
- Audio deepfakes pose a significant threat, with real-world cases of fraud and misinformation on the rise
- So far, we have focused on **short-duration (short-form), acoustically-clean and single-speaker** audio waveforms
- Longer durations and partially-spoofed audio have been addressed
- Our goal is to take a step towards realistic conditions, with **longer duration (long-form), acoustically-complex, and multi-speaker** audios

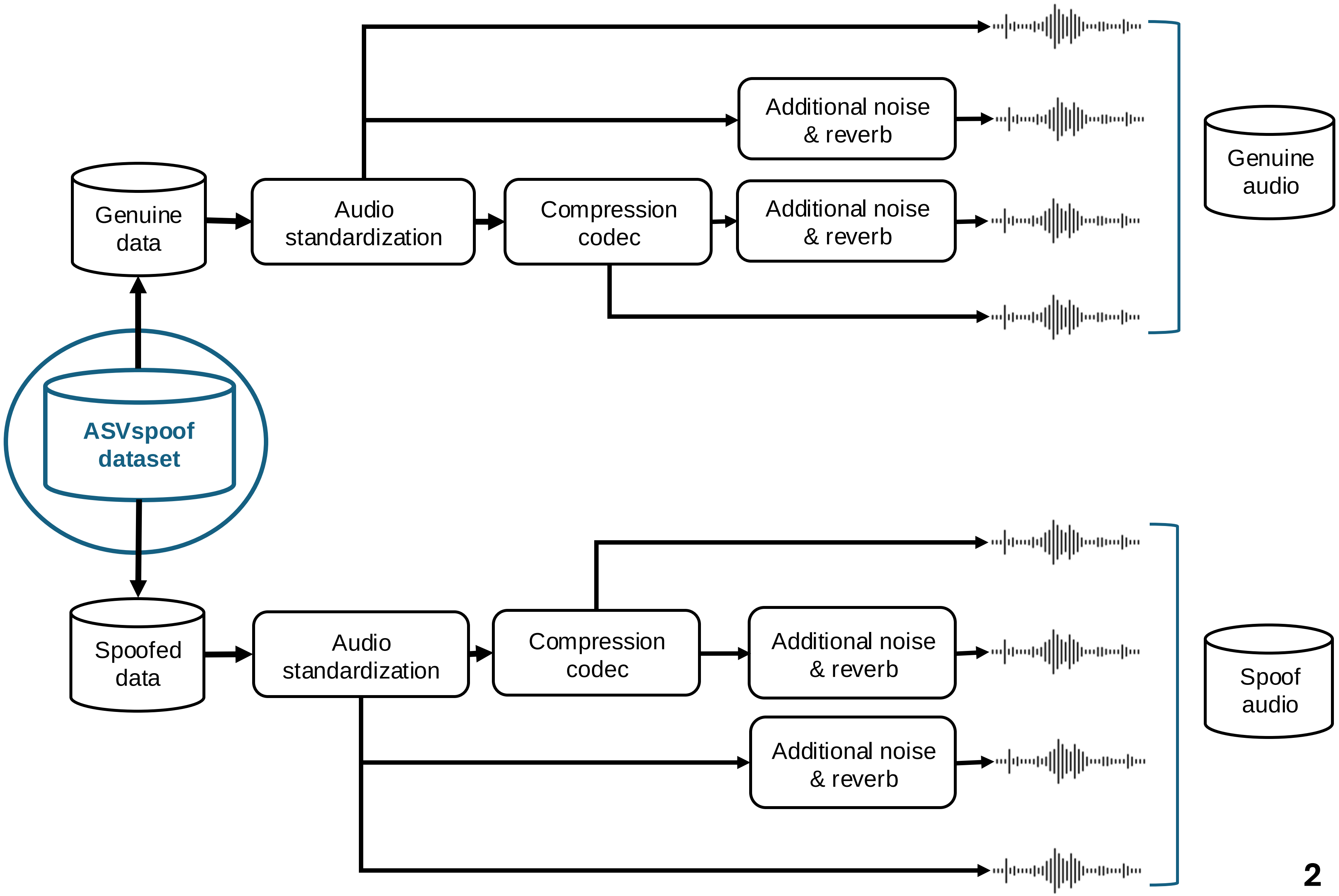


Yi, Jiangyan; Fu, Ruibo; Tao, Jianhua; Nie, Shuai et al, ADD 2022: the first Audio Deep Synthesis Detection Challenge. In: Proc. ICASSP, 2022.
Muller, Nicolas M; Kawa, Piotr; Choong, Wei Herng et al, MLAAD: The Multi-Language Audio Anti-Spoofing Dataset. IJCNN 2024.
Lin Zhang, Xin Wang, Erica Cooper, et al. The PartialSpoof Database and Countermeasures for the Detection of Short Fake Speech Segments Embedded in an Utterance, in IEEE/ACM Trans. Audio, Speech and Lang. Proc. 31 (2023), 813–825.

Long-form Spoofed Audio Generation

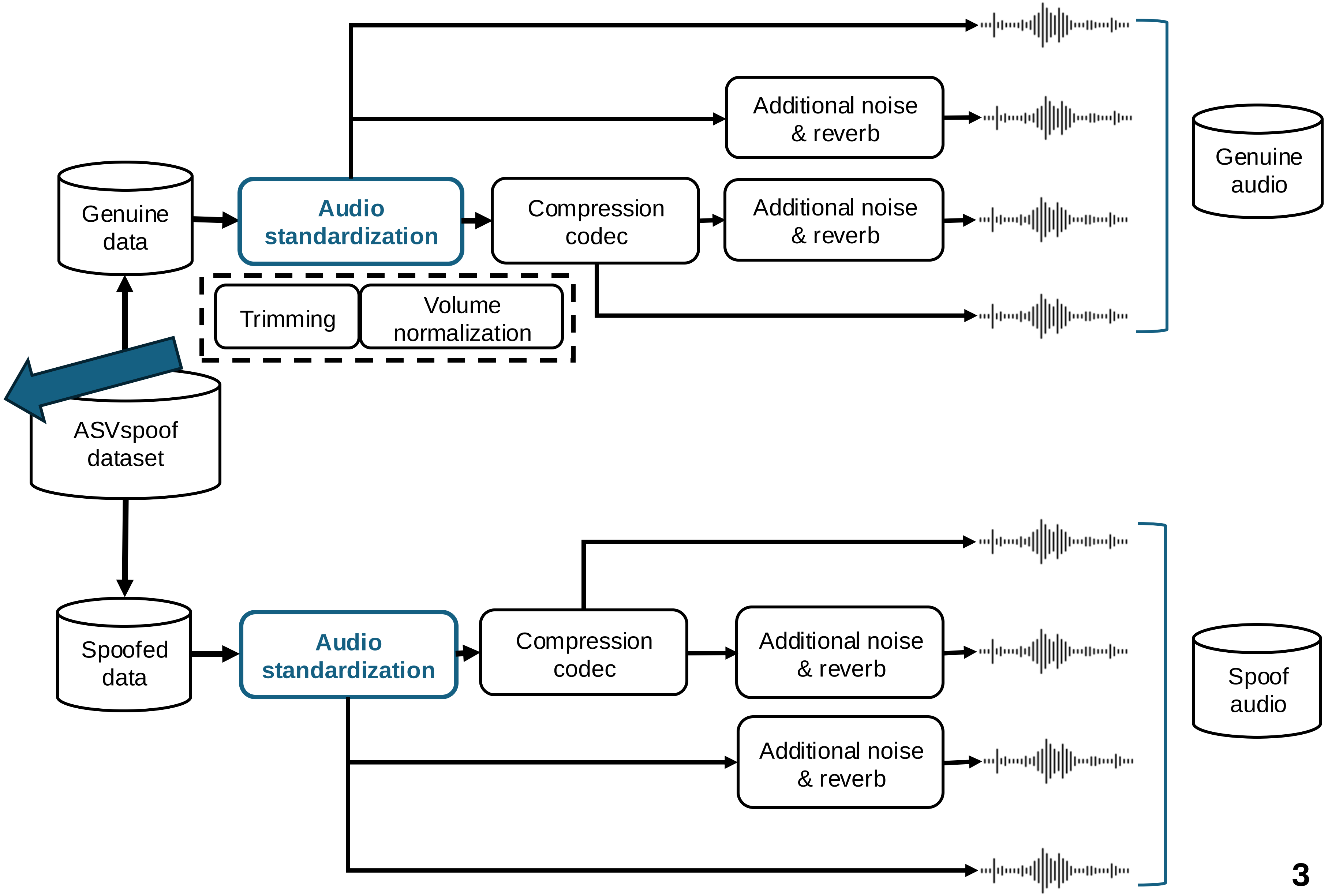
- 2-10 seconds
- Single speaker
- Clean condition



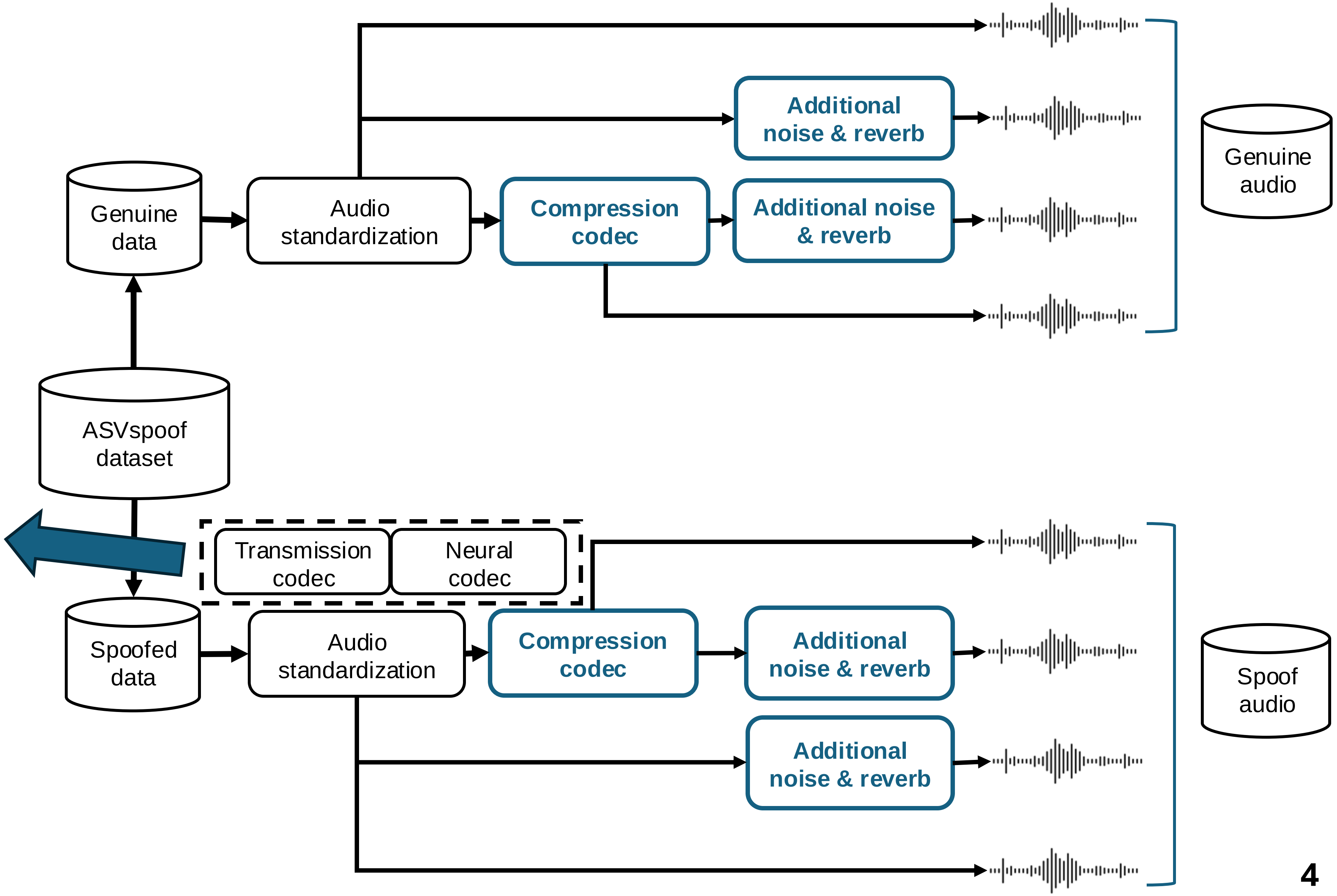


Long-form Spoofed Audio Generation

- Removing silence at beginning & end of the audio
- Normalize the speaker volume using sv56 toolkit



Long-form Spoofed Audio Generation



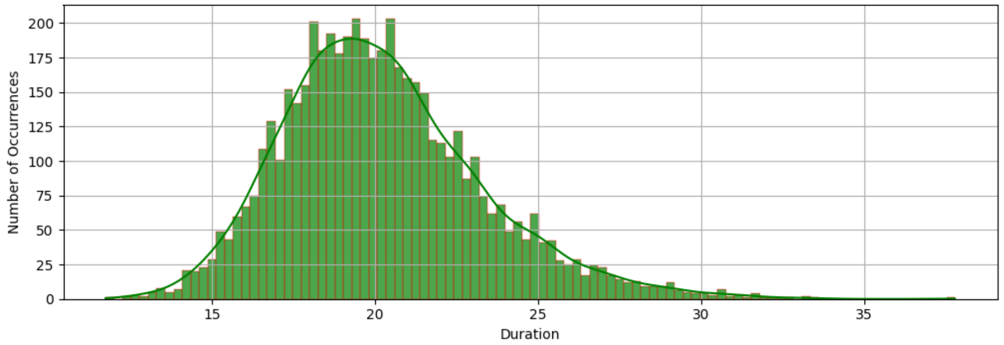
- We use various codec models as processing methods
- Transmission codec: mp3, opus, a-law……
- Neural codec: EnCodec, Soundsteam, FACodec……

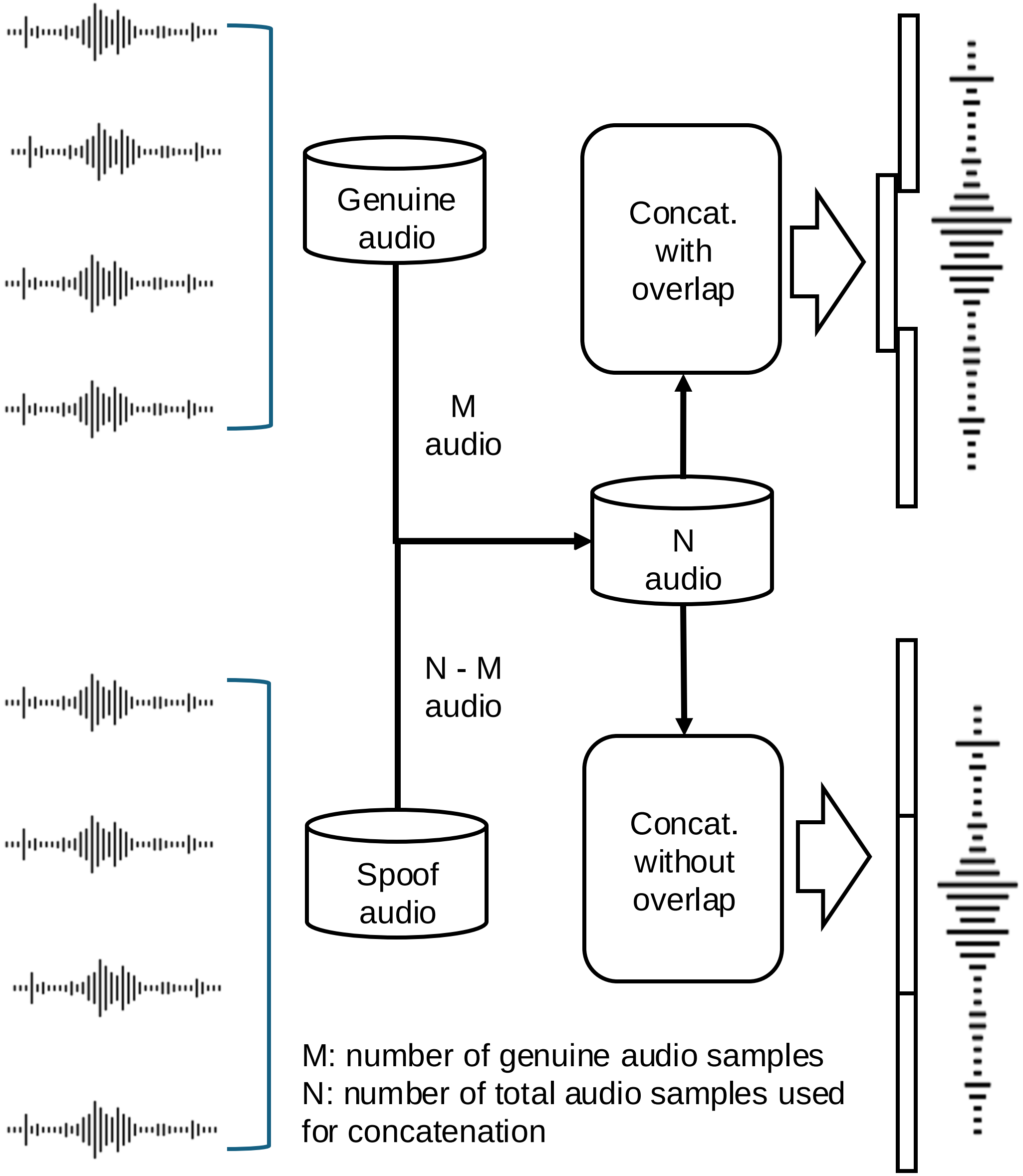
- Additions: RIR & MUSAN

NII

Long-form Spoofed Audio Generation

- Generated audio via concatenation, with optionally overlap
- Usually, 15-30 seconds long
- Multiple speakers in the same audio
- Mixed amount of genuine and spoofed content
- **Varying amount of genuine vs. spoof ratio**, so purely-genuine and purely-spoofed audios are also there



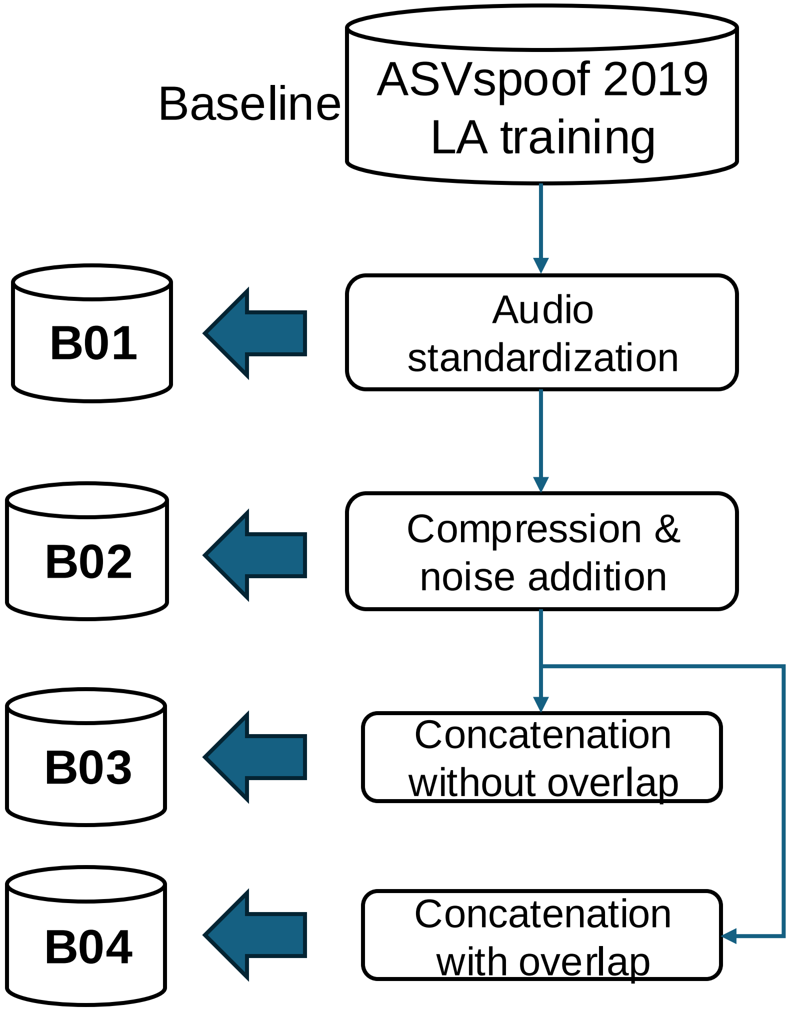


Experimental Setup

- We use state-of-the-art end-to-end audio deepfake detector called AASIST

- We developed multiple training sets
  - Baseline: The original ASVspoof 2019 LA training data (short-form, clean, single-speaker)
  - B01: Baseline + audio standardization
  - B02: B01 + audio processing
  - B03: B02 + long-form audio concatenation
  - B04: B02 + long-form audio concatenation, w/ overlap

- Evaluation audio are generated via same functions, with different placements
  - No audio process, no overlap
  - No audio process, with overlap
  - Processed, no overlap
  - Processed, with overlap



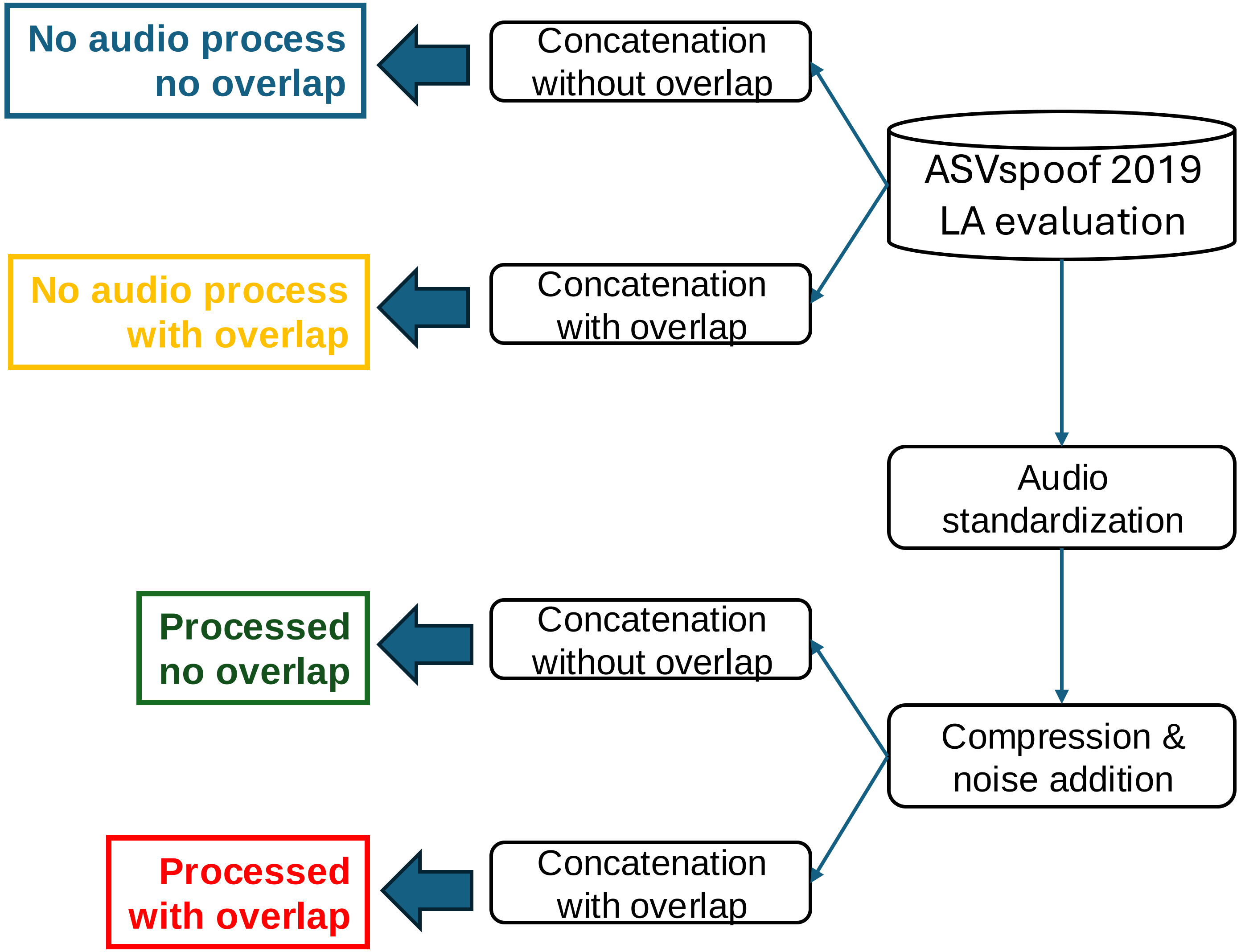
Jung, Jee-weon; Heo, Hee-Soo; Tak, Hemlata et al, AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. In: Proc. ICASSP,2022.

Experimental Setup

- We use state-of-the-art end-to-end audio deepfake detector called AASIST

- We developed multiple training sets
  - Baseline: The original ASVspoof 2019 LA training data (short-form, clean, single-speaker)
  - B01: Baseline + audio standardization
  - B02: B01 + audio processing (codec,
  - B03: B02 + long-form audio concatenation
  - B04: B02 + long-form audio concatenation, w/ overlap

- Evaluation audio are generated via same functions, with different placements
  - No audio process, no overlap
  - No audio process, with overlap
  - Processed, no overlap
  - Processed, with overlap



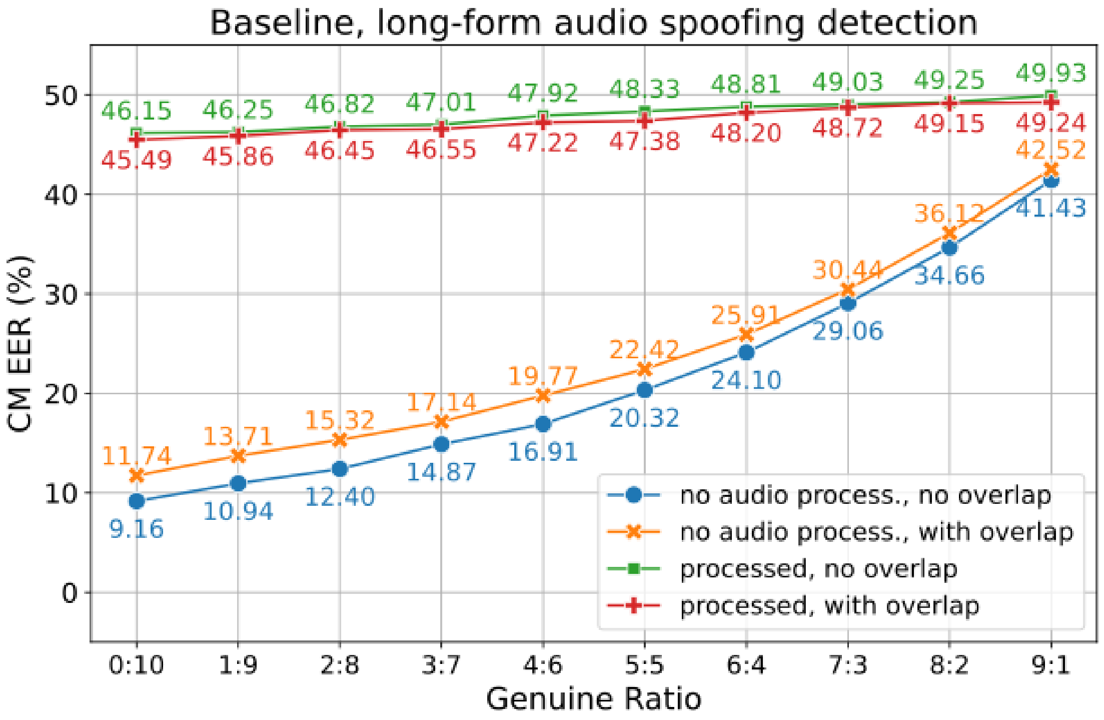
Jung, Jee-weon; Heo, Hee-Soo; Tak, Hemlata et al, AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. In: Proc. ICASSP, 2022.

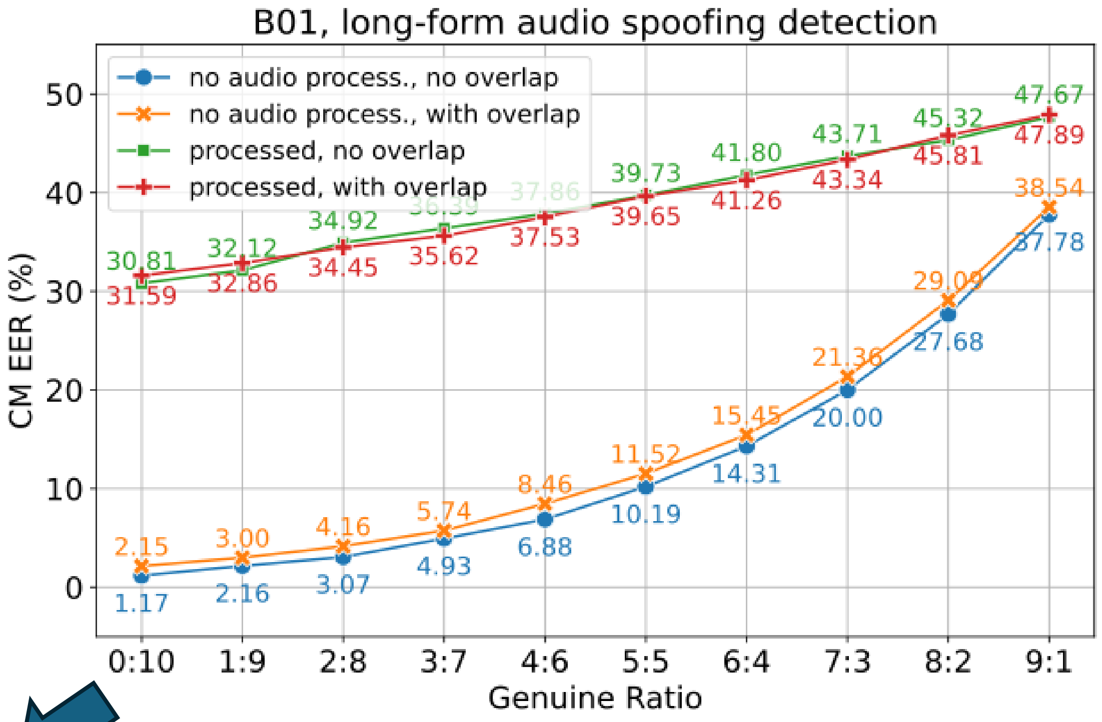
NII

Results

- Training on short, clean speech is not enough for detecting complex spoofed audios

- Performance drops as the amount of genuine content in the long-form audio increases



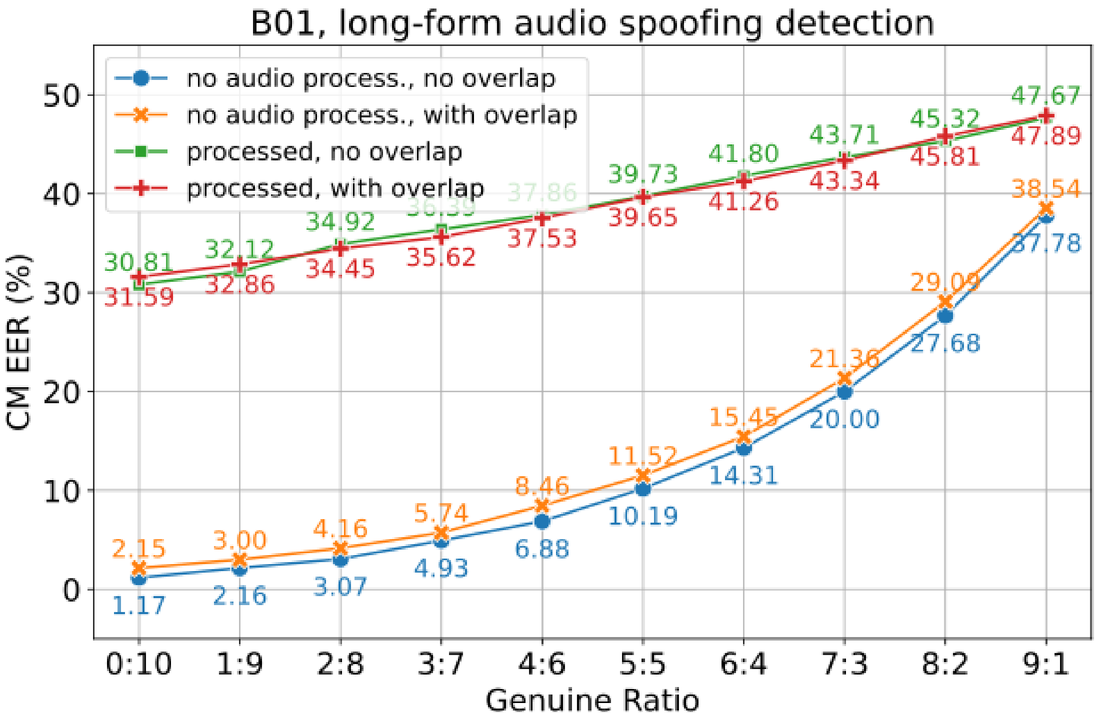


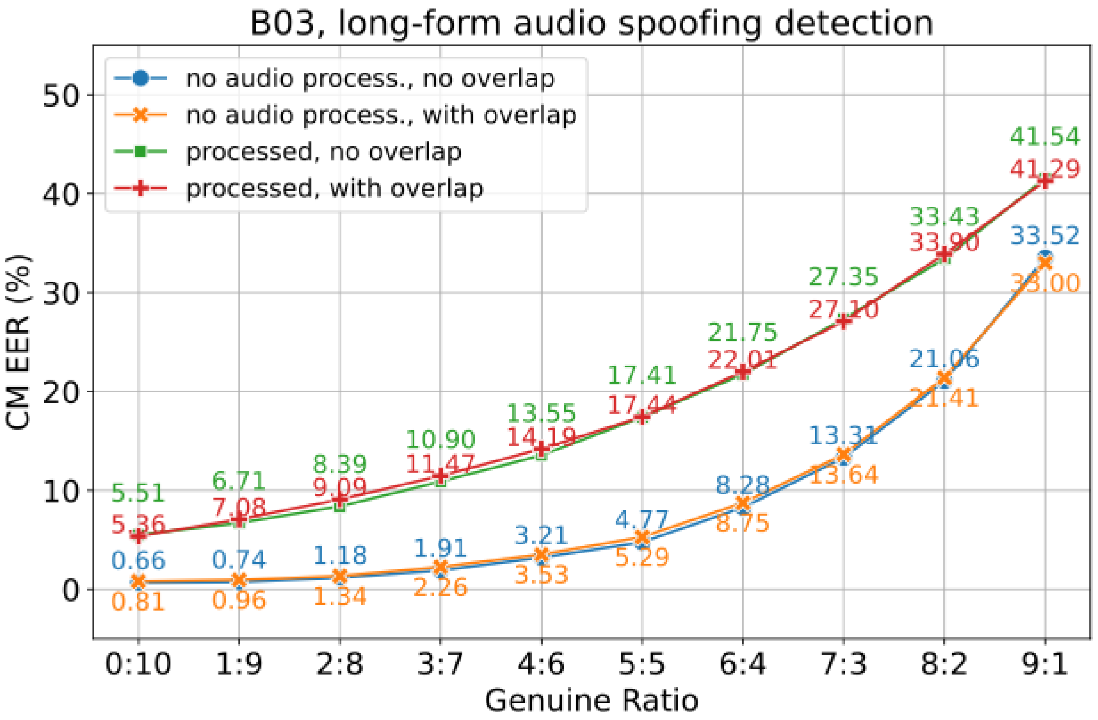
M : (N - M)

NII

Results

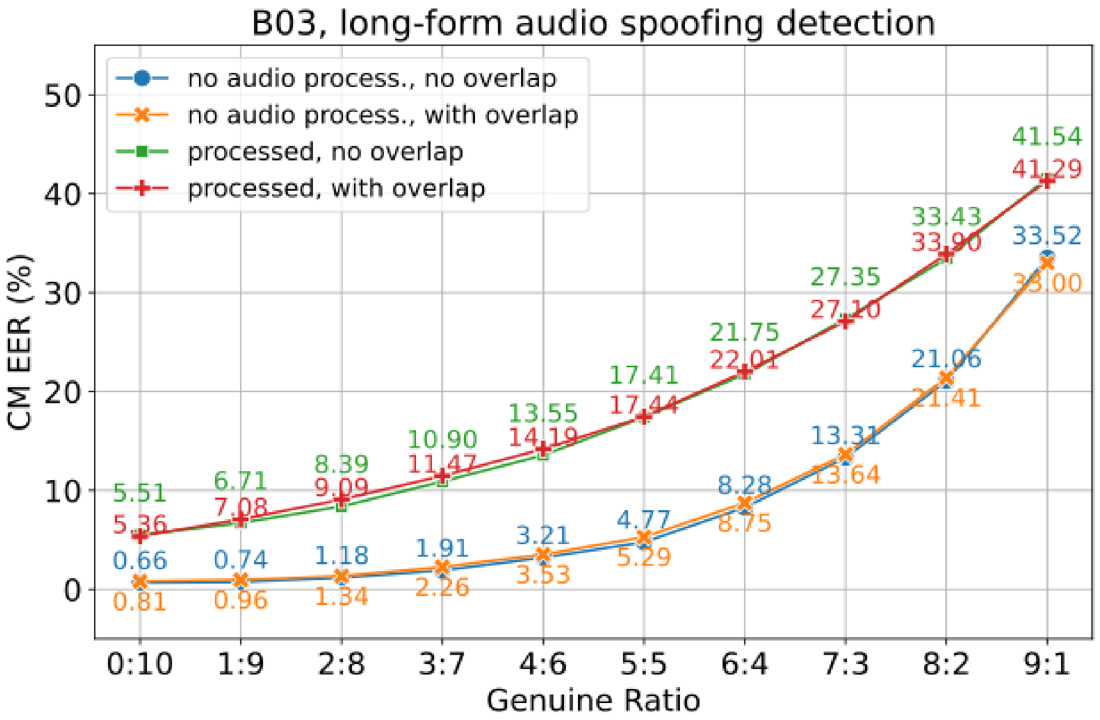
- ➢ Further improvement on training set improves performance, especially for highly-spoofed audio

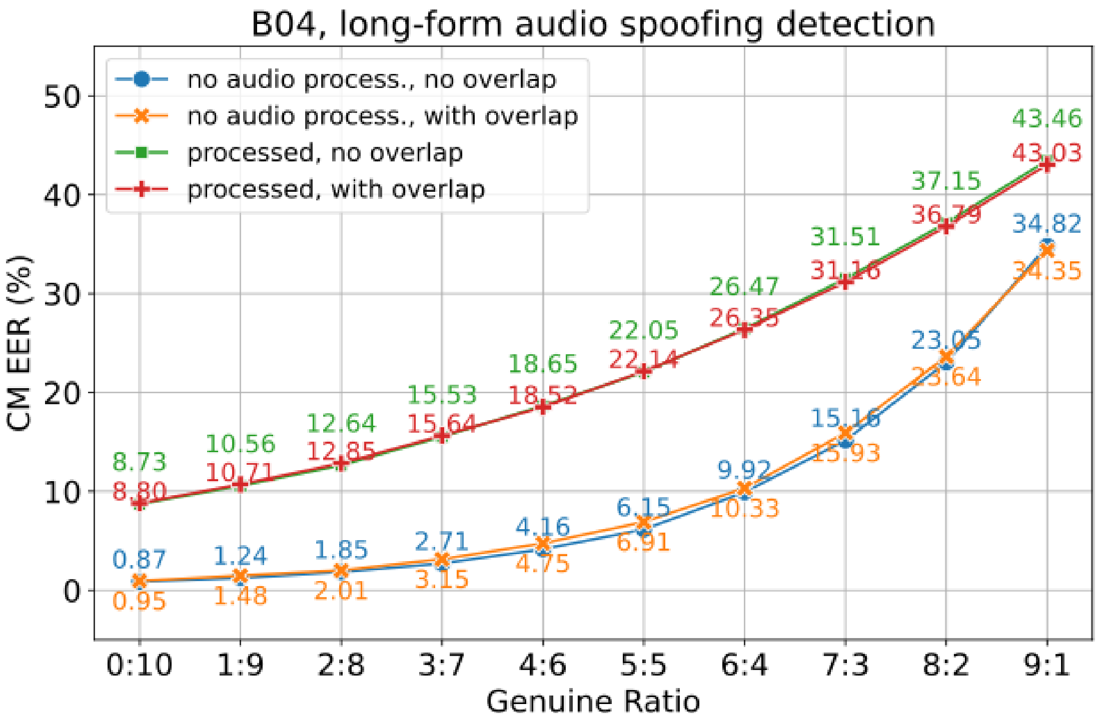




Results

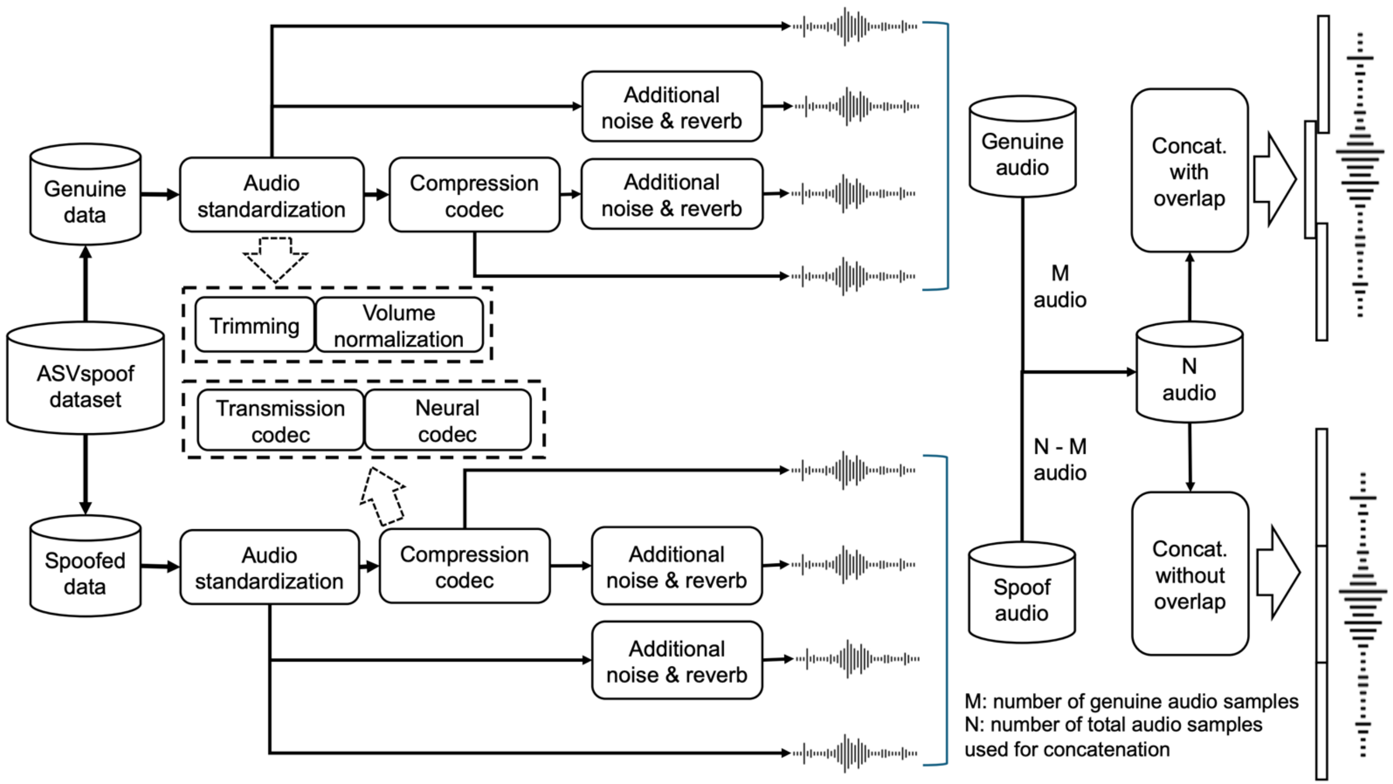
➢ Overlapping the adjacent in long-form audio in the training set slightly worsens the performance





Main Take-Aways

- Short-form detection models are insufficient for more challenging scenarios posed by audio DeepFake technologies
- We have proposed a new pipeline on generating longer duration
  - 15-30 seconds
  - Varying amount of spoofed content
  - Multi-speaker presence
- Long-form training data improves detection accuracy for complex audio spoofing task
- Future work will explore larger datasets and more complex models



Thanks for Listening!