Whispeak Speech Deepfake Detection Systems for the ASVspoof5 Challenge

Motivation

The increasing realism of synthetic speech poses societal challenges, including misinformation and impersonation. The ASVSpoof Challenge aims to evaluate and foster the development of countermeasures against such threats. The ASVSpoof5 edition specifically focuses on non-studio-quality data collected from audiobooks, presenting a new challenge for detection systems.

Proposed Method

The authors employ an ensemble of systems for Speech Deepfake Detection and SASV tracks:

Speech Deepfake Detection

- Models Used:
  - LFCC-LCNN
  - RawGAT-ST
  - Wav2Vec-RawGAT-ST
  - Wav2Vec-Conformer

- Data Augmentation: A wide range of techniques are applied, including:
  - Silence Removal
  - TimeStretch
  - PitchShift
  - Noise addition
  - Reverberation
  - Rawboost
  - Codec application
  - Bit Crusher
  - Gain adjustment
  - SpecAugment

- Training Details:
  - Models are trained on the ASVSpoof5 train set
  - Some models include bona fide samples from LibriSpeech 960h
  - Different probabilities of data augmentation are used
  - Cross-entropy loss with class weighting is employed

Spoofing-Aware Speaker Verification (SASV)

- ASV Model: ECAPA-TDNN trained on VoxCeleb2
- Fusion: Linear combination of CM and ASV scores

Model Fusion

- Bagging method using median of selected model predictions
- Different model combinations for closed and open conditions

The proposed approach demonstrates strong performance on the ASVSpoof5 dataset, particularly with models using Wav2Vec 2.0 frontend. However, the authors note that generalization to more diverse datasets remains a challenge.

Experimentation:

- Models trained on ASVSpoof5 train set, some with additional LibriSpeech data.
- Various data augmentation probabilities tested.
- Evaluation on ASVSpoof5 development set and other datasets for generalization testing.

Results:

- Wav2Vec frontend models showed best performance on ASVSpoof5.
- Data augmentation improved model robustness.
- Poor generalization observed on older ASVSpoof datasets and external datasets.
- Best open condition performance: EER of 4.16% and minDCF of 0.1124 on evaluation set.

Q1: Which track does this submission belong to? Track 1 or 2?
This submission belongs to both Track 1 (Speech Deepfake Detection) and Track 2 (Spoofing-Aware Speaker Verification).

Q2: Which condition does this submission belong to? Open or Closed condition?
The submission includes both Open and Closed conditions.

Q3: Which subset do the authors report the performance on? ASVspoof 5 progress set or evaluation set?
The authors report performance on both the ASVspoof5 development set and evaluation set.

Q4: What is the reported performance in terms of EER and minDCF?
For the open condition on the evaluation set, the reported performance is:

- EER: 4.16%

- minDCF: 0.1124

Q5: Can you take a look at Table 4 in this paper and guess which ID your systems belong to? You will need to compare the minDCF, actDCF, Cllr, and EER values.
T23

Q6: Provide a GitHub link to the repository if there is any

https://github.com/asvspoof- challenge/ asvspoof5/tree/main/Tool- score- fusion

---

My findings from this paper-

What They Did

- They used four different computer models to detect fake speech.

- Two models (LFCC-LCNN and RawGAT-ST) were simpler, while two others (Wav2Vec-RawGAT-ST and Wav2Vec-Conformer) were more advanced.

- They messed with the audio in different ways during training to make their system better at spotting fakes.

What They Found

- The more advanced models (using Wav2Vec) worked best on the challenge dataset.

- Messing with the audio during training (data augmentation) helped improve results.

- Their best system got an error rate of 4.16% on the final test.

Problems They Noticed

- The systems didn't work as well on other datasets, especially older ones.

- This means the detectors are good at spotting fakes in the challenge but might struggle with different types of audio.

What This Means

- We're getting better at detecting fake speech, but there's still work to do.

- Current systems are good at specific tasks but struggle to work well across all types of audio.

- More research is needed to create fake speech detectors that work well in all situations.