THE VICOMTECH AUDIO DEEPFAKE DETECTION SYSTEM BASED ON WAV2VEC2 FOR THE 2022 ADD CHALLENGE

Juan M. Martín-Doñas, Aitor Álvarez

vicomtech
MEMBER OF BASQUE RESEARCH
& TECHNOLOGY ALLIANCE

Index

Audio Deep synthesis Detection (ADD) Challenge 2022

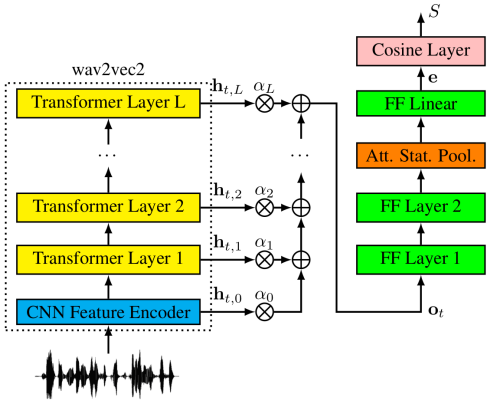
- Recent growth of deep learning based text-to-speech (**TTS**) synthesis and voice conversion (**VC**) technologies
- Generation of **deepfake speech**. Malicious use: foolish human or even automatic speaker verification systems
- Need reliable **countermersures**. Audio deepfake detection systems.
  - Example: ASVspoof series (2015-2021)
- Great improvements achieved, interest in more **challenging scenarios**:
  - Noisy and reverberant scenarios
  - Speech modified through different channels, codecs or compression algorithms
  - Partially spoofed audio

Audio Deep synthesis Detection (ADD) Challenge 2022

- **2022 Audio Deep synthesis Detection (ADD) Challenge**: Detection of deep synthesis and manipulated audios in different scenarios
  - **Track 1**: Low-quality fake audio detection
  - **Track 2**: Partially fake audio detection
  - **Track 3**: Audio fake game
- Vicomtech proposed audio deepfake detection system for Tracks 1 and 2:
  - **Wav2Vec2** pre-trained feature extractor
  - Downstream classifier model trained for deepfake detection
  - **Data augmentation** techniques to adapt the classifier
  - **Winners of Track 1** and fourth position in Track 2
  - Competitive results in ASVspoof 2021 Challenge

Wav2Vec2-based proposed system

- **W2V2 Feature Extractor**:
  - Cross-lingual Large models (**XLS**) 300M parameters
    - 53 and 128 languages
  - **Self-supervised** learning with contrastive loss
    - Masked encoded features
    - Predict quantized representations from contextualized ones
  - **Pre-trained** model
    - Freeze W2V2 parameters
    - Finetuning downstream classifier



Wav2Vec2-based proposed system

- **Classification Model**:
  - Representations from $L = 25$ transformer layers
    - $o_t = \sum_{l=0}^{L} \alpha_l h_{t,l}$
  - **Attentive** statistical temporal pooling (mean and std dev.)
  - Compute embedding e
  - **Cosine scoring** layer
    - $S = \cos(w, e) \in [-1, 1]$
  - **One-class** softmax loss function

Layer name	Output size
W2V2 features	$N \times T \times 1024 \times 25$
Temp. Norm. + Layer weight.	$N \times T \times 1024$
FF Layer (1 and 2)	$N \times T \times 128$
Att. Stat. Pool.	$N \times 256$
FF Linear	$N \times 128$
Cosine Layer	N

Experimental framework

- **ADD 2022** database:
  - Genuine and TTS/VC speech from **AISHELL-3** speech corpus
  - Training and development clean speech ($\sim$ 28K utt. each)
  - Adaptation ($\sim$ 1K) and test ($\sim$ 100K) sets for each track
    - **Track 1**: Real-world noises and background music
    - **Track 2**: Partial fake manipulation using real or synthesized audios
- **ASVspoof 2021** database:
  - Train and development sets from **ASVspoof 2019 LA** (TTS/VC)
  - Logical Access (**LA**): Transmission through real telephonic systems
  - Speech Deepfake (**DF**): Processed speech with commercial audio codecs
- Adaptation and **data augmentation** techniques:
  - Low-pass **FIR** filtering (narrowband and wideband): Frequency masking
  - ADD 2022: Training using **train and adaptation** sets
  - Track 2 (ADD): Generating **new partial deepfakes** by audio overlapping

Experimental results

- Results on **ADD 2022**:
  - XLS-128 outperforms XLS-53
  - Few adaptation data help (main improvements)
  - Generated partial deepfakes in Track 2 improve further the model performance
  - Narrowband FIR filtering reduce 1% EER in both tracks
  - Competitive system:
    - T1: 21.7% EER (1st)
    - T2: 16.6% EER (4th)

W2V2	Sets	DA	Track1	Track2
XLS-53	Train	-	32.96	38.09
	Tr.+Adap.	-	23.70	33.73
XLS-128	Train	-	32.20	45.88
	Tr.+Adap.	-	22.62	30.35
	Tr.+Adap.	FIR	**21.71**	-
	Tr.+Adap.	partial	-	17.58
	Tr.+Adap.	FIR+part.	-	**16.59**

Experimental results

- Results on **ASVspoof 2021**:
  - XLS-128 also outperforms XLS-53 model
  - Narrowband FIR for LA
    - 8 kHz bandwidth telephone channel
  - Wideband FIR for DF
    - Emulate general audio codecs

W2V2 model	Data augmentation	LA	DF
XLS-53	-	8.87	7.71
	FIR-NB	4.34	11.27
	FIR-WB	4.98	6.99
XLS-128	-	7.20	5.68
	FIR-NB	**3.54**	6.18
	FIR-WB	7.08	**4.98**

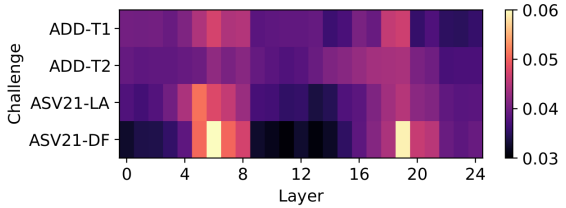
Experimental results

- Previous ASVspoof 2021 results:
  - **Ensemble** classifiers with robust neural models (LCNN, ECAPA, ResNet)
    - Poor generalization on DF set (other speech databases)
  - Previous W2V2 only used representations from **last layer** (need finetuning)
  - Our proposal uses general **W2V2 feature extractor** with specialized downstream model (using data augmentation)

System	LA	DF
LCNN+ResNet+RawNet	**1.32**	15.64
GMM+LCNN (Ensemble)	3.62	18.30
ECAPA-TDNN (Ensemble)	5.46	20.33
ResNet (Ensemble)	3.21	16.05
W2V2 (fixed)+LCNN+BLSTM	10.97	7.14
W2V2 (finetuned)+LCNN+BLSTM	7.18	5.44
Proposed system	3.54	**4.98**

Experimental results

- **Weight values** α_l for the transformer layers:
  - The information from different layers is used for deepfake detection
  - Different layer weights depending on the scenario
  - Example: For DF, special focus around layers 6 and 19



Conclusions and future work

- Our approach effectively exploits the contextualized representation from the **different transformer layers** of a pre-trained **Wav2Vec2** model
- The downstream classifier can be finetuned using these representations and adapted through adequate **data augmentation** techniques
- Our system shows competitive results in both ASVspoof 2021 (especially in **DF task**) and ADD 2022 challenges (**winner of Track 1**)
- **Future work**: Test other self-supervised models and additional data augmentation techniques

Thank you!

Contact: Juan M. Martin (jmmartin@vicomtech.org)