The Vicomtech Audio Deepfake Detection System based on Wav2Vec2 for the 2022 ADD Challenge

CHAL-5.6

Juan M. Martín-Doñas and Aitor Álvarez

Vicomtech Foundation, Basque Research and Technology Alliance (BRTA),
Mikeletegi 57, 20009 Donostia – San Sebastián (Spain)

Introduction

Motivation

- Recent growth of **TTS** and **VC** technologies.
- Malicious use of **deepfake** speech.
- Need for reliable **countermeasures**.
- New **challenging** scenarios: noisy environments, channel artifacts, partial deepfakes.

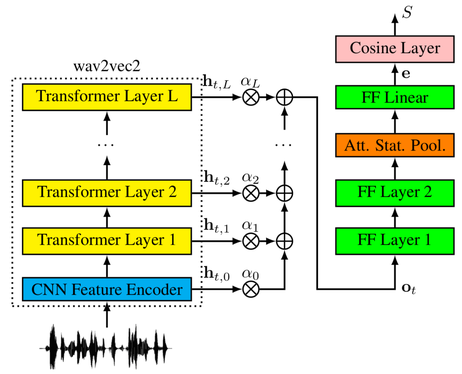
Audio Deep synthesis Detection Challenge

- **Track 1**: Low-quality fake audio detection.
- **Track 2**: Partially fake audio detection.
- Track 3: Audio fake game.

Proposed system

- **Wav2Vec2** (W2V2) feature extractor.
- **Downstream** deepfake detection model.
- **Data augmentation** techniques.
- **Winners of Track 1** and 4th in Track 2.

Wav2Vec2-based proposed system



W2V2 Feature Extractor

- Cross-lingual (**XLS**) Large models (53 and 128).
- **Self-supervised** learning with contrastive loss.
- **Pre-trained** model (freeze during training).

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

Classification Model

- Combines transformer layers $\mathbf{o}_t = \sum_{l=0}^{L} \alpha_l \mathbf{h}_{t,l}$.
- Attentive statistical temporal pooling.
- Cosine scoring $S = \cos(\mathbf{w}, \mathbf{e}) \in [-1, 1]$.
- **One-class** softmax loss function.

Experimental Framework

ADD 2022 database

- AISHELL-3 speech corpus.
- Train and dev. sets: Clean speech.
- Adaptation and test sets:

Track 1: Noises and background music.
Track 2: Partial fake manipulated audios.

ASVspoof 2021 database

- Train and dev. sets: ASVspoof 2019 LA.
- **Logical Access (LA)**: Telephonic systems.
- **Speech Deepfake (DF)**: Audio codecs.

Data augmentation techniques

- Low-pass **FIR** filtering.
- **ADD 2022**: Training using train and adap. sets.
- Track 2: Generation of **new partial deepfakes**.

Experimental Results

Results on ADD 2022 Challenge

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**

- **XLS-128** outperforms XLS-53.
- Few **adaptation data** help.
- Further improvements in Track 2 from additional partial deepfakes.
- **Narrowband FIR**: 1% EER reduction.

Results on ASVspoof 2021 Challenge

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**

Previous models

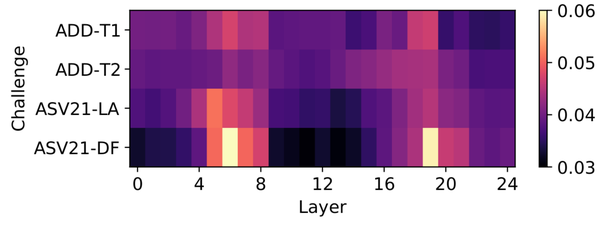
- **Supervised**: Poorly generalized in DF set.
- **W2V2**: Last transformer layer (need finetuning).

Our proposal

- W2V2 feature extractor and specialized downstream model (competitive performance).
- **FIR**: narrowband for LA, wideband for DF.

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**

Layer weights visualization



Conclusion

- **Pre-trained** W2V2 feature extractor using different transformer layers.
- Downstream spoofing classifier adapted using **data augmentation** techniques.
- Competitive results in both ASVspoof 2021 and ADD 2022.
- **Future work**: Testing new self-supervised models and other data augmentation techniques.

Contact Information

Juan M. Martín-Doñas

E-mail: jmmartin@vicomtech.org
Dept. of Speech and Natural Language Technologies,
Vicomtech Foundation, Spain