The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)
31 August 2024, Kos, Greece

Whispeak Speech Deepfake Detection Systems for the ASVspoof5 Challenge

Pierre Falez, Tony Marteau

Whispeak, France

pfalez@whispeak.io, tmarteau@whispeak.io

Abstract

In this paper, we present the system submitted by Whispeak for the ASVSpoof5 Speech Deepfake Detection and SASV Challenge. We use an ensemble of systems, consisting of LFCC-LCNN, RawGAT-ST, Wav2Vec-RawGAT-ST and Wav2Vec-Conformer for the Speech Deepfake Detection tracks. We use a linear fusion of an ECAPA-TDNN with the previous ensemble for SASV tracks. A dozen data augmentation techniques are applied during training in order to improve the robustness of the models on the ASVSpoof5 dataset. We also test our models on external datasets and show that models are not yet able to generalize well on out-of-domain data. The final system gives an EER of 4.16% and a minDCF of 0.1124 on the track 1 evaluation set in open condition.

1. Introduction

Generative deep learning and especially the speech synthesis is an active field of research. Today, the synthetic speech is realistic and natural to the point where it becomes difficult to differentiate them to bona fide speech. Therefore, speech synthesis and ease of access to these technologies is an increasing society problem. For example, misinformation, or impersonation. To protect society and automatic speaker verification (ASV) systems, it is essential to progress in the speech synthesis detection.

For several years, ASVSpoof Challenge is a good support to evaluate the progress of research and foster the development of countermeasures (CM). The main purpose of this challenge is to evaluate generalization capacity of detectors over unknown attacks by providing multiple datasets. First editions of ASVSpoof Challenge (2015 [1] and 2017 [2]) focused on the development of stand-alone CM against synthetic speech or replay speech. Then 2019 edition [3] of the challenge combined two past editions with a track on synthetic speech detection (LA) and a track on replay speech detection (PA). This edition also introduces an evaluation more ASV-centric. Finally, the last edition in 2021 [4], proposed three independent tracks to evaluate the detection of replay speech (PA), the detection of synthetic speech (DF) and the detection of synthetic speech pass through codecs and communication channels (LA).

In the ASVSpoof 5 edition, unlike the past editions, the challenge is based on non-studio-quality data collected in audiobook datasets [5]. The challenge is organized in two tracks: Track1 to evaluate Speech Deepfake detection systems and Track2 to evaluate Spoofing-Aware Speaker Verification (SASV) systems.

This paper presents the Whispeak submissions on tracks 1 and 2 with a focus on the different spoofing detection systems used in Section 2. Then in Section 3, we present the method of applying the data augmentation. Section 4 focus on the details of the model training procedure. Finally, all spoofing detection systems are compared with the development set provided by the organizers and other datasets with the aim of measuring their generalizability in Section 5

2. System Overview

2.1. Speech Deepfake Detection

We have chosen four models that have shown good performance on the ASVSpoof21 dataset [4]. Two of them do not use transfer learning in order to comply with the closed condition, the other two uses pre-trained models in order to offer better generalizations in the open condition. The first model is LFCC-LCNN [6], which use linear frequency cepstral coefficients (LFCCs) as features. LFCC are extracted with 1024 FFT, a window length of 20ms, a stride of 10ms and 20 filters. Delta and delta-delta of the filter are concatenated, resulting in 60 channels. LCNN9 [7] followed by a Bi-LSTM and a linear layer is used as a classifier. The second model is a RawGAT-ST [8], which use a frozen sincconv [9] that extract 70 filters followed by a Graph Neural Network. The third model is a Wav2Vec-RawGAT-ST [10]. It uses a pre-trained Wav2Vec 2.0 [11] as a frontend extractor, followed by a classifier similar to the one used in the RawGAT-ST. Unlike the original model, we have used the Wav2Vec 2.0 Large trained on LibriSpeech 960h [12] instead of the XLS version to comply with the challenge rules. The last model is a Wav2Vec-Conformer [13], which also use a Wav2Vec 2.0 as frontend, but use an adapted Conformer as a classifier. We use the version improved in [14] which shows a better capacity for generalisation.

2.2. Spoofing-Aware Speaker Verification

For the second track, we choose to use independent CM and ASV models, in order to reuse CM models of the first track. As ASV models, we use an ECAPA-TDNN [15] trained on VoxCeleb2, in a similar way than [16]. In order to comply with closed condition rules, we do not use babble data augmentation. A first training with AMM-Softmax is done with $\alpha = 0.2$ on 2-second segments, and then, a fine-tuning is done with $\alpha = 0.5$ on 6-second segments. We use multiple data augmentation (Noise addition with the MUSAN [17] noise dataset, Reverberation with the RIRS_NOISES [18] dataset, codec application and SpecAugment [19]).

3. Data Augmentation

In order to make CM models more robust, we applied various techniques of data augmentation:

- Silence Removal: we used an energy-based voice activity detection (VAD) to retrieve speech frames. All the frames where the VAD measurement is below a threshold is set to zero.

10.21437/ASVspoof.2024-5

Model	p_{DA}	ASVSpoof19 LA Test		ASVSpoof21 LA Eval		ASVSpoof21 DF Eval		ASVSpoof5 Dev	
		EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF
LFCC-LCNN	0.0	34.33%	0.5000	37.20%	0.5000	36.06%	0.5000	22.34%	0.3544
	0.01	26.18%	0.3684	30.74%	0.4337	25.69%	0.3580	12.84%	0.1835
	0.05	33.14%	0.4667	34.34%	0.4815	28.01%	0.4014	12.11%	0.1726
	0.1	35.45%	0.5000	36.06%	0.5000	28.63%	0.4110	12.95%	0.1844
RawGAT-ST	0.0	33.80%	0.4859	32.92%	0.4761	36.63%	0.3863	16.25%	0.2278
	0.01	21.82%	0.3153	22.41%	0.3222	25.25%	0.3248	16.38%	0.2364
	0.05	19.17%	0.2762	16.73%	0.2365	27.22%	0.3580	13.71%	0.1980
	0.1	25.22%	0.3622	25.19%	0.3651	27.43%	0.3859	13.64%	0.1971
Wav2Vec-RawGAT-ST	0.0	**15.25%**	**0.2198**	**15.68%**	**0.2259**	**13.30%**	**0.1868**	1.82%	0.0257
	0.05	21.56%	0.3093	18.31%	0.2628	14.18%	0.2001	1.04%	0.0148
	0.1	19.17%	0.2762	23.24%	0.3368	14.82%	0.2130	**0.76%**	**0.0109**
	0.2	19.35%	0.2779	16.63%	0.2381	14.18%	0.2031	0.83%	0.0114
Wav2Vec-Conformer	0.0	29.87%	0.4259	40.48%	0.4628	19.40%	0.2809	5.44%	0.0788
	0.05	17.90%	0.2563	17.63%	0.2371	14.33%	0.1926	0.93%	0.0129
	0.1	22.80%	0.3233	20.92%	0.2918	15.64%	0.2189	1.12%	0.0157
	0.2	19.83%	0.2830	18.85%	0.2643	16.03%	0.2231	1.15%	0.0164

Table 1: Models EER ↓ and minDCF ↓ on different ASVSpoof datasets.

- TimeStretch: change the speed with a ratio from 0.7 to 1.3 without changing the pitch.
- PitchShift: change the pitch in a range from -10 to 10 semitones without changing the tempo.
- Noise: a random sample from MUSAN [17] noise dataset is added with a random SNR from 0dB to 15dB.
- Reverberation: a random room impulse response from RIRS_NOISES [18] dataset is applied.
- Rawboost [20]: ISD, SSI and LnL are each applied independently.
- 16Khz Codec: a codec is randomly selected between mp3, aac, opus, vorbis, mu-law and G722 with a random compression level.
- 8Khz Codec: audio is resampled at 8khz, then a random codec is randomly selected between A-law, Mu-law, AMR-NB and GSM. Finally, the audio is resampled back to 16khz.
- Bit Crusher: reduce the bit depth randomly from 5 to 14 bits.
- Gain: apply a gain to the audio in the range 0.25 to 2.0.
- SpecAugment [19]: A mask ranging from 100 samples to 1000 samples is applied on time and 5% of frequency are masked.

Each of these data augmentation is applied online in series. Each of them has a probability p_{DA} to be applied. In this way, different combinations of data augmentation are used on the data.

4. Training Details

All CM models are trained on the ASVSpoof5 train set, except for a few models trained with the addition of bona fide samples from LibriSpeech 960h [21] (only in open condition). We use by default the same hyperparameters as in the original works. We have adapted some parameters to take into account the specifics of ASVSpoof5. Models are trained four times with an ADAM optimizer and different p_{DA} probabilities. LFCC-LCNN uses a learning rate of 3^{-4} and no weight decay. RawGAT-ST uses a learning rate and a weight decay of 1^{-4}. Models with Wav2Vec frontend use a learning rate of 1^{-6} and a weight decay of 1^{-4}. LFCC-LCNN and RawGAT-ST are trained for 40 epochs with an exponential learning rate scheduler (γ is fixed at 0.85) and a batch size of 64. Wav2Vec-RawGAT-ST and Wav2Vec-Conformer are trained for 200,000 updates with a constant learning rate and a batch size of 20. In each configuration, we use a cross-entropy loss with a weight of 0.9 for bonafide class and a weight of 0.1 for spoof class, in order to respect the distribution of classes in ASVSpoof5. All the models use a randomly cropped 4-second segment of audio. For shorted audio, a circular padding is used to reach the required dimension. Each model is trained on a single Nvidia L40S.

For each model, the final weights are obtained by averaging the 5 checkpoints that achieved the best minDCF on the development dataset [22] during training.

5. Results

Predictions are made using unified feature maps techniques [25]. We extract crops of 4-second with an offset of 2-second and compute the final prediction by averaging model output of all the crops.

As specified in the challenge guidelines, our models are evaluated using the EER and minDCF metrics for track 1 and the tandem-EER (t-EER) and agnostic-DCF (a-DCF) metrics for track 2. For minDCF and a-DCF, the costs provided by organizers are respectively $C_{miss} = 1$, $C_{fa} = 10$ and $C_{miss} = 1$, $C_{fa} = 10$, $C_{fa,spoof} = 10$.

Table 1 shows that using Wav2Vec frontend greatly improve results on the ASVSpoof5 development dataset. p_{DA} has an impact on performances of the models. For example, adding data augmentation to the Wav2Vec-Conformer improve EER from 5.44% ($p_{DA} = 0.0$) to 0.93% ($p_{DA} = 0.05$) on the development set. Wav2Vec-RawGAT-ST is the model which gives the best overall performances. For the closed condition, LFCC-LCNN tends to perform better than RawGAT-ST.

Model	p_{DA}	In-The-Wild [23] EER	In-The-Wild [23] minDCF	Fake-Or-Real [24] EER	Fake-Or-Real [24] minDCF
LFCC-LCNN	0.0	32.64%	0.5000	25.85%	0.4732
	0.01	32.31%	0.4386	18.28%	0.2496
	0.05	27.16%	0.3742	24.95%	0.3235
	0.1	22.98%	0.3133	24.06%	0.3216
RawGAT-ST	0.0	26.86%	0.2853	**9.75%**	**0.1203**
	0.01	25.30%	0.3146	18.39%	0.2431
	0.05	26.33%	0.3051	25.53%	0.3464
	0.1	26.39%	0.3044	0.2257	0.2991
Wav2Vec-RawGAT-ST	0.0	16.30%	0.2180	16.92%	0.1966
	0.05	21.91%	0.2782	21.56%	0.2768
	0.1	19.75%	0.2575	20.41%	0.2512
	0.2	19.90%	0.2588	19.16%	0.2333
Wav2Vec-Conformer	0.0	**14.25%**	**0.2044**	9.80%	0.1296
	0.05	14.49%	0.2061	24.47%	0.3215
	0.1	18.17%	0.2524	34.55%	0.4049
	0.2	16.23%	0.2336	27.69%	0.3187

Table 2: Models EER ↓ and minDCF ↓ on other datasets.

Model	LibriSpeech	EER	minDCF
LFCC-LCNN	✗	12.11%	0.1726
	✓	14.83%	0.2148
Wav2Vec-RawGAT-ST	✗	1.04%	0.0148
	✓	1.69%	0.0241

Table 3: Models EER ↓ and minDCF ↓ on ASVSpoof5 Track1 development set.

However, all the models perform poorly on older ASVSpoof datasets (ASVSpoof19 LA Test, ASVSpoof21 LA Eval and ASVSpoof21 DF Eval). One possible reason is that other datasets are based on studio-quality samples (from VCTK [26]), which is out-of-domain compared to ASVSpoof5 which uses lower quality data, obtained from audiobooks on LibriVox. This shows that detection models are not yet able to generalize to all types of data. Table 2 confirms this observation. Accuracy on other datasets is much worse than on ASVSpoof5. This time, Wav2Vec-Conformer without DA seems to be the configuration that generalizes best.

Table 3 shows that adding bona fide from LibriSpeech during training degrade the performance of the single models. However, such models seem relevant for fusion, since the LFCC-LCNN trained with LibriSpeech was chosen as part of the best models combination.

We use bagging method to do the model fusion. Score merging is done by using the median of the prediction of selected models. For track 1 in closed condition, all models that comply with the closed rules were selected (LFCC-LCNN with $p_{DA} = (0.01, 0.05, 0.1)$ and RawGAT-ST with $p_{DA} = (0.01, 0.05, 0.1)$). For track 1 in open condition, models were selected by finding the combination that gives the best minDCF on the development set (LFCC-LCNN trained on ASVSpoof5 and Librispeech, Wav2Vec-Conformer with $p_{DA} = 0.1$ and Wav2Vec-RawGAT-ST with $p_{DA} = (0.1, 0.2)$). Table 4 shows the final result of the ensemble of models for the Speech Deepfake Detection track.

For SASV track, we used the same ensemble of models for the CM part. ASV part use the single ECAPA-TDNN model. The fusion between CM and ASV score is realized in the same way as the script provided with the baseline[1], by estimating log-likelihood ratios of the two systems on development set and then linearly combining predictions. Table 5 show final submission for the track 2.

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

Condition	Development Set EER	Development Set minDCF	Evaluation Set EER	Evaluation Set minDCF
Closed	12.23%	0.1772	20.13%	0.5312
Open	0.58%	0.0082	4.16%	0.1124

Table 4: Fusion of models EER ↓ and minDCF ↓ on ASVSpoof5 Track 1 development and evaluation partitions.

Condition	Development Set t-EER	Development Set aDCF	Evaluation Set t-EER	Evaluation Set aDCF
Closed	5.49%	0.1767	49.34%	0.4513
Open	1.54%	0.4732	4.63%	0.1492

Table 5: Fusion of models t-EER ↓ and aDCF ↓ on ASVSpoof5 Track 2 development and evaluation partitions.

6. Conclusion

We have presented the Whispeak systems for the different track and condition of the ASVSpoof5 Challenge. Models that perform best on the ASVSpoof5 dataset use Wav2Vec 2.0 frontend. We show that using a wide range of data augmentation applied in series with a defined probability improve the robustness of the models when tested on the ASVSpoof5 dataset. However, more work is needed to create systems capable of generalizing on more diverse data. Models trained on ASVSpoof5 have a significantly lower performance when tested on alternative datasets.

7. References

[1] Zhizheng Wu, Tomi Kinnunen, Nicholas W. D. Evans, Junichi Yamagishi, Cemal Hanilçi, Md. Sahidullah, and Aleksandr Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Int. Speech Conf. (INTERSPEECH)*, Dresden, Germany, Sept. 2015, pp. 2037–2041.

[2] Tomi Kinnunen, Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas W. D. Evans, Junichi Yamagishi, and Kong-Aik Lee, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," in *Int. Speech Conf. (INTERSPEECH)*, Stockholm, Sweden, Aug. 2017, pp. 2–6.

[3] Massimiliano Todisco, Xin Wang, Ville Vestman, Md. Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas W. D. Evans, Tomi H. Kinnunen, and Kong Aik Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Int. Speech Conf. (INTERSPEECH)*, Graz, Austria, Sept. 2019, pp. 1008–1012.

[4] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Online, Sept. 2021, pp. 47–54.

[5] Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi, "ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," in *ASVspoof Workshop 2024 (accepted)*, 2024.

[6] Xin Wang and Junichi Yamagishi, "A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection," in *Int. Speech Conf. (INTERSPEECH)*, Brno, Czechia, Aug. 2021, pp. 4259–4263.

[7] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 11, pp. 2884–2896, 2018.

[8] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Online and Singapore, May 2022, pp. 6367–6371.

[9] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *IEEE Spoken Language Technology Workshop (SLT-W)*, Athens, Greece, Dec. 2018, pp. 1021–1028.

[10] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Odyssey: The Speaker and Language Recognition Workshop*, Beijing, China, June 2022, pp. 112–119.

[11] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Neural Information Processing Systems (NeurIPS)*, Online, Dec. 2020, pp. 12449–12460.

[12] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, Apr. 2015, pp. 5206–5210.

[13] Eros Roselló Casado, Alejandro Gómez Alanís, Ángel Manuel Gómez García, Antonio Miguel Peinado Herreros, et al., "A conformer-based classifier for variable-length utterance processing in anti-spoofing," in *Int. Speech Conf. (INTERSPEECH)*, Dublin, Ireland, Aug. 2023.

[14] Duc-Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu-Thi Luong, Kong Aik Lee, and Eng Siong Chng, "Temporal-channel modeling in multi-head self-attention for synthetic speech detection," *arXiv preprint arXiv:2406.17376*, 2024.

[15] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Int. Speech Conf. (INTERSPEECH)*, Online and Shanghai, China, Oct. 2020, pp. 3830–3834.

[16] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn based speaker verification," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, June 2021, pp. 5814–5818.

[17] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[18] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 5220–5224.

[19] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Int. Speech Conf. (INTERSPEECH)*, Graz, Austria, Sept. 2019, pp. 2613–2617.

[20] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Online and Singapore, May 2022, pp. 6382–6386.

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Apr. 2015, pp. 5206–5210.

[22] Yingbo Gao, Christian Herold, Zijian Yang, and Hermann Ney, "Revisiting checkpoint averaging for neural machine translation," in *Findings of the Association for Computational Linguistics (AACL-IJCNLP)*, Online, Nov. 2022, pp. 188–196.

[23] Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger, "Does audio deepfake detection generalize?," in *Int. Speech Conf. (INTERSPEECH)*, Incheon, Korea, Sept. 2022, pp. 2783–2787.

[24] Ricardo Reimao and Vassilios Tzerpos, "For: A dataset for synthetic speech detection," in *Int. Conf. on Speech Technology and Human-Computer Dialogue (SpeD)*, Timisoara, Romania, Oct. 2019, pp. 1–10.

[25] Cheng-I Lai, Nanxin Chen, Jesús Villalba, and Najim Dehak, "Assert: Anti-spoofing with squeeze-excitation and residual networks," in *Int. Speech Conf. (INTERSPEECH)*, Graz, Austria, Sept. 2019, pp. 1013–1017.

[26] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al., "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, vol. 6, pp. 15, 2017.