# CAU_KU team's system description for 2022 spoofing-aware speaker verification challenge

*Narin Kim*[1], *Tae-in Kang*[1], *Sunmook Choi*[2], *Junho Shin*[2], *Seungsang Oh*[2], *Il-Youp Kwak*[1]

[1]Chung-Ang University, Republic of Korea
[2]Korea University, Republic of Korea

taein@cau.ac.kr, narin@cau.ac.kr, felixchoi@korea.ac.kr, junho@korea.ac.kr, ikwak2@cau.ac.kr, seungsang@korea.ac.kr

## Abstract

This system description report describes Chung-Ang University and Korea University (CAU_KU) team's model participating in the Spoofing-Aware Speaker Verification (SASV) 2022 Challenge. To create a spoofing-aware speaker verification model, we designed the integrated model that ensembles multiple CM systems and one ASV system (ECAPA-TDNN with log-melspectrogram feature). For the CM systems, we used ResMax with CQT feature and AASIST with raw audio feature. Our integrated model achieved 5.0% SASV-EER performance.

**Index Terms**: spoofing-aware speaker verification, spoofing attack detection, deep learning

## 1. Introduction

This system description paper describes Chung-Ang University and Korea University (CAU_KU) team's participation in the Spoofing-Aware Speaker Verification (SASV) 2022 Challenge. Despite the fact that the present ASV (automated speaker verification) technology is continually evolving, it does not account for spoofing attacks. As a result, it's susceptible to spoofing attacks like speech synthesis and voice conversion, and its performance suffers as a result. The goal of the SASV competition is to create a system that can tackle the difficulties.

We created an SASV system by utilizing embeddings from both the CM and ASV systems. For the CM system building, we used ResMax [1], LCNN [2, 3], AASIST [4], and OFD (with no splits) [5] models. We experimented with changing the baseline ECAPA-TDNN [6] model by adding min and max statistics to the attentive statistics pooling layer for the ASV system. We also tried using both the ASVspoof 2019 LA dataset and Voxceleb1 for training the ECAPA-TDNN model. For the best system, we used ResMax and AASIST models for CM systems, and we used ECAPA-TDNN model for the ASV system. We integrated embeddings of those CM and ASV systems for the final model. Our integrated system achieved SASV-EER of 5.0 %, SV-EER of 9.0 %, and SPF-EER of 0.3 %.

## 2. Methodology

We used the ResMax [1] and AASIST [4] models as CM systems and the ECAPA-TDNN model [6], a baseline model, as the ASV system. The embeddings from ResMax, AASIST, and ECAPA-TDNN models were used to train an integrated model for the final prediction.

### 2.1. Feature processing for ResMax, LCNN and OFD models

We utilized CQT feature extracted using the librosa software [7] for ResMax, LCNN, and OFD models. For the CQT feature extraction, we set minimum frequency as 5, the number of frequency bins as 100, and filter scale factor as 1.

### 2.2. Data augmentation for ResMax and LCNN models

As data augmentation for ResMax and LCNN models, Mixup [8] and Frequency Feature Masking (FFM) [5] were used. Mixup's beta distribution parameter was set at 0.7, while FFM solely employed high-frequency masking.

### 2.3. CM system 1 : ResMax model

ResMax [1] is a lightweight spoofing detection system that combined the notions of skip connection (from ResNet [9]) and max feature map (from Light CNN [2, 3]). We used the same model used for ASVspoof2019 LA scenario.

### 2.4. CM system 2 : AAIST

AASIST [4] is an anti-spoofing system using integrated spectro-temporal graph attention networks. AASIST is a baseline model for the competition, and it uses the raw audio feature. We considered using the AASIST baseline model and embeddings.

### 2.5. CM system 3 : LCNN

The LCNN model has proven useful in ASVspoof 2017, 2019, and 2021 competitions [10, 3, 11, 12, 13, 14]. We used a deeper LCNN model by adding a few more layers to the Light CNN-9 model [2]. Light CNN-9 model repeats five convolution layers and four network-in-network (NIN) layers [2]. We proposed a model which iterates six convolution layers and 5 NIN layers using 32, 48, 64, 32, 32, and 32 convolution filters and 32, 48, 64, 64, and 32 NIN filters. As in the previous model, the kernel size of the first convolution layer is set to 5, and the remaining convolution layers are set to 3. Except for the first and third convolution layers, batch normalizations are followed. All NIN layers are followed by batch normalization layer. Instead of using a fully connected layer defined in the Light CNN-9 model [2], we used the global average pooling layer, batch normalization, and Dropout layer with a probability of 0.5.

### 2.6. CM system 4 : OFD

OFD model is the Overlapped Frequency-Distributed (OFD) model [5]. The model architecture described in Kwak et al. (2022) [5]. The core difference is that we used OFD with no split.

### 2.7. ASV system: ECAPA-TDNN model

#### *2.7.1. Feature processing for ECAPA-TDNN model*

The log-melspectrogram produced in 80 dimensions of the mel bin size extracted with the torchaudio module is a feature of the ECAPA-TDNN model. Log-melspectrogram features were extracted with a sample rate of 16000 after the preemphasis. The input segment is 2 seconds long, the window is 25 milliseconds long, and the hopping length is 10 milliseconds long. After that, SpecAugment was used to conduct frequency and temporal masking. Masking was applied arbitrarily in the time and frequency domains, ranging from 0 to 10 frames in the time domain and 0 to 8 frames in the frequency domain.

#### *2.7.2. Data augmentation for ECAPA-TDNN model*

The ECAPA-TDNN baseline was trained with the VoxCeleb2 dataset, and was trained by applying data augmentation with the RIR and MUSAN datasets.

#### *2.7.3. ECAPA-TDNN model*

The structure of the ECAPA-TDNN [6] extends the x-vector [15] architecture. It has a speaker encoder with a 3-layer bottleneck structure, including Res2Net and SE (Squeeze-and-Excitation) block. The 1D dilated convolution is applied to each bottleneck The size of the SE block channel is 1024. After passing through the 3-layer bottleneck, the global mean and standard deviation calculated through attentive statistics pooling are reflected in the channel-dependent frame attention. Finally, after concatenating the statistic reflecting the attention and weight passing through the bottleneck, it passes through the fully connected layer. The size of the output speaker embedding dimension is 192.

### 2.8. Integrated model

For each utterance sample, 64-, 32, 128, and 160-dimensional embeddings were extracted from the CM systems, ResMax, LCNN, OFD, and AASIST. On the other hand, 192-dimensional embedding was extracted from the ASV system, the ECAPA-TDNN model. In addition, the ASV system yielded 192-dimensional enrollment ASV embeddings for the speaker model. Concatenating all those embeddings and learning a total of 10 epochs for four layers of DNN yielded the outcome of the integrated model. The optimizer was Adam, the learning rate was 0.0001, and the learning rate scheduler's weight decay was set to 0.001.

## 3. Experiments

### 3.1. Experimental setup

VoxCeleb 1 and 2 data [16, 17] were prepared for learning ECAPA-TDNN, as well as RIR (reverb) and MUSAN (babble, noise) data sets [18, 19] for data augmentation. Pytorch was used for the entire training of the ECAPA-TDNN system. ASVspoof 2019 LA training and development data were utilized to learn ResMax. Librosa was used for data extraction, and TensorFlow and Keras were utilized for mixup augmentation, FFM augmentation, and the entire model training. Pytorch-Lightning was used for the integrated SASV system.

### 3.2. Experimental result

For the CM system building, we used ResMax, LCNN, AASIST, and OFD (with no splits) models. We experimented with changing the baseline ECAPA-TDNN model by adding min and max statistics to the attentive statistics pooling layer for the ASV system. We also tried using both the ASVspoof 2019 LA dataset and Voxceleb1 for training the ECAPA-TDNN model. Table 1 describe the performances of several systems. All models used ECAPA-TDNN baseline model as an ASV system. The 'M1' system used ResMax as a CM model, the 'M2' system used ResMax and AASIST as CM models, the 'M3' system used ResMax, AASIST and OFD as CM models,,the 'M4' system used ResMax, AASIST and LCNN as CM models, the 'M5' system used ResMax, AASIST, OFD and LCNN as CM models

Table 1: *EERs on evaluation set*

| Models | SASV-EER | SV-EER | SPF-EER |
|---|---|---|---|
| M1 | 0.06778 | 0.12123 | 0.01077 |
| M2 | 0.04959 | 0.09044 | 0.00293 |
| M3 | 0.06593 | 0.12301 | 0.00304 |
| M4 | 0.05735 | 0.09914 | 0.00316 |
| M5 | 0.06313 | 0.11390 | 0.00241 |

Among all experimental models, M2 model performed the best.

## 4. Conclusions

In order for ASV to be utilized practically, defense against spoofing is vitally important, and this SASV competition focuses on the construction of such a system. Our CAU KU team created an integrated model by utilizing ResMax and AASIST to extract the embeddings of the CM systems, and ECAPA-TDNN to extract the embeddings of the ASV system. Our proposed system achieved SASV-EER of 5.0%.

## 5. Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIP) (No. NRF-2020R1C1C1A01013020 and NRF-2017R1A2B2007216).

## 6. References

[1] I.-Y. Kwak, S. Kwag, J. Lee, J. H. Huh, C.-H. Lee, Y. Jeon, J. Hwang, and J. W. Yoon, "ResMax: Detecting Voice Spoofing Attacks with Residual Network and Max Feature Map," in *25th International Conference on Pattern Recognition (ICPR)*. IEEE Computer Society, 2021, pp. 4837–4844.

[2] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, Nov 2018.

[3] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC Antispoofing Systems for the ASVspoof2019 Challenge," in *Proc. Interspeech 2019*, 2019, pp. 1033–1037. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1768

[4] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *arXiv preprint arXiv:2110.01200*, 2021.

[5] I.-Y. Kwak, S. Choi, J. Yang, Y. Lee, and S. Oh, "Cau_ku team's submission to add 2022 challenge task 1: Low-quality fake audio detection through frequency feature masking," 2022.

[6] Desplanques, Brecht and Thienpondt, Jenthe and Demuynck, Kris, "ECAPA-TDNN : Emphasized Channel Attention, Propagation and Aggregation in TDNN based speaker verification," in *Proc. Interspeech 2020*. International Speech Communication Association (ISCA), 2020, pp. 3830–3834.

[7] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[8] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[10] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *Proc. Interspeech 2017*, 2017.

[11] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, "STC Antispoofing Systems for the ASVspoof2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 61–67.

[12] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech 2017*. Stockholm: ISCA, 2017, pp. 2–6.

[13] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. Interspeech 2019*, 2019, pp. 1008–1012. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2249

[14] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 47–54.

[15] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[16] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-950

[17] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1929

[18] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[19] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.