# Characterizing the Fusion Strategies for Spoofing-Aware Speaker Verification

*Haibin Wu*[1,2], *Lingwei Meng*[3], *Jiawen Kang*[3], *Jinchao Li*[3], *Xu Li*[3], *Xixin Wu*[3], *Hung-yi Lee*[1], *Helen Meng*[2,3]

[1] Graduate Institute of Communication Engineering, National Taiwan University
[2] Centre for Perceptual and Interactive Intelligence, The Chinese University of Hong Kong
[3] Human-Computer Communications Laboratory, The Chinese University of Hong Kong

## Abstract

Recently, many novel techniques have been introduced to deal with spoofing attacks, and achieve promising countermeasure (CM) performances. However, these works only take the stand-alone CM models into account. Nowadays, a spoofing aware speaker verification (SASV) challenge which aims to facilitate the research of integrated CM and ASV models, arguing that jointly optimizing CM and ASV models will lead to better performance, is taking place. In this paper, we propose a novel multi-model and multi-level fusion strategy to tackle the SASV task. Compared with purely scoring fusion and embedding fusion methods, this framework first utilizes embeddings from CM models, propagating CM embeddings into a CM block to obtain a CM score. In the second-level fusion, CM score and ASV scores directly from ASV systems will be concatenated into a prediction block for the final decision. As a result, the best single fusion system has achieved the SASV-EER of 0.97% on the evaluation set. Then by ensembling the top-5 fusion systems, the final SASV-EER reached 0.89% on the evaluation set, while this number in the best baseline system is 6.37%.

**Index Terms**: anti-spoofing, speaker verification, spoofing-aware speaker verification

## 1. Introduction

Spoofing attacks and countermeasures (CM) for automatic speaker verification (ASV) have aroused keen interests from both the academia and the industry. While ASV systems aim to verify the identity of target speakers, spoofing attacks attempt to manipulate the verification results using various technologies, leading to dramatic performance degradation [1–6]. In order to ensure the robustness and security of ASV systems, CM is a necessary technique to defend or detect spoofing attacks.

The vulnerability of ASV systems was revealed in [7–9], under speech synthesis and voice conversion (VC) attacks. Currently, varies of techniques have been proposed to perform effective attacks, including audio replay [10, 11], adversarial noise [12–14], more advanced text-to-speech (TTS) and VC models [15–17]. Many works have been done to investigate state-of-the-ark CM strategies. The current solutions leverage end-to-end deep neural networks (DNNs) [18, 19], trying to distinguish artifacts and unnatural cues of spoofing speech from bona fide speech. And thanks to a series of challenges and datasets [1–4], many novel techniques were introduced to achieve promising CM performances [18–25].

However, previous works only take the stand-alone CM models into account. Recently, a spoofing aware speaker verification (SASV) challenge [26] was proposed as a special session in ISCA INTERSPEECH 2022. This challenge aims to facilitate the research of integrated CM and ASV models, arguing that jointly optimizing CM and ASV models will lead to better performance. To measure the performance of integrated models, a SASV-EER was proposed in this challenge as a primary metric, which is a variant of classic equal error rate (EER). Under this metric, the test utterances in trials belong to one of three types: impostors, target speakers, and spoofing attacks. In further, the SASV-EER can be subsetted into SV-EER (impostors vs. targets) and SPF-EER (targets vs. spoof). The former is for evaluating speaker verification performance, and the latter is for evaluating anti-spoofing performance. In this way, this metric expects the model can accept target speakers and reject any alternatives, including the impostors and spoofing attacks, which is a straightforward assessment for integrated SASV systems.

There are limited works that optimize ASV and CM models jointly. Existing methods can take the form of two categories: fusion-based solutions and integrated single model solutions. For fusion-based solutions, [27] fuses the embeddings of ASV and CM model by an integrated neural network. [28] proposed a method to model synthesis-channel subspace and perform SASV in i-vector space. Other works considered fusing the scores of CM and ASV systems, by Gaussian back-end fusion [29], cascaded/parallel fusion framework [30] and optimizing a differentiable detection cost function using reinforcement learning [31]. For integrated single model solutions, a common idea is to obtain a joint embedding representing both ASV and CM information by multi-task learning. [32] achieved this by a modified triple loss, and [33] used sequential residual convolutional blocks with Max-Feature-Map activations. In the SASV Challenge 2022, two baseline systems are presented following the above two categories, evaluated on ASVspoof 2019 dataset [34]. The details will be discussed in the below sections.

This paper described our submitted system for the SASV Challenge 2022. In order to take advantage of existing well-designed models in CM and ASV areas, we proposed a novel multi-model and multi-scale fusion framework. Compared with purely scoring fusion and embedding fusion methods, this framework first utilizes embeddings from CM models, propagating CM embeddings into a CM block to obtain a CM score. In the second-level fusion, CM score and ASV scores directly from ASV systems will be concatenated into a prediction block for the final decision. In contrast to our previous work [35] which only simply concatenates the embeddings from different CM models, we considered the potentials of pooling strategies in terms of feature aggregation, and investigated various pooling methods [36–39] when fusing embeddings across different CM models. Based on the proposed fusion framework, we presented the fusion strategies of a series of state-of-the-art CM and ASV models with different pooling strategies to boost the fusion results. As a result, the best single fusion system has achieved the SASV-EER of 0.97% on the evaluation set. Then by ensembling the top-5 fusion systems, the final SASV-EER reached 0.89% on the evaluation set, while this number in the

best baseline system from the SASV challenge is 6.37%.

This paper is organized as follows: Section 2 introduces related background. Then, the detailed methods will be presented in Section 3. Section 4 describes the experimental setups, the experimental results and analysis are discussed in Section 5. Finally, Section 6 concludes this paper.

# 2. Background

One of the motivations of this work is to leverage the successfully developed models and algorithms in standalone ASV and CM fields. In this section, the background of ASV and CM will be introduced to provide insights for further discussion.

Automatic speaker verification is a technology to verify whether a given test utterance belongs to an enrolled target speaker. Early ASV methods are based on statistical models, e.g., Gaussian Mixture Model with Universal Background Model (GMM-UBM) [40]. Further research focused on subspace methods, in which i-vecto/PLDA architecture is the most famous one [41, 42]. More recently, benefiting from the powerful representation ability of DNNs, the methods based on x-vector [37] have become the mainstream solutions. These models are trained to distinguish different speakers in training data, and then the trained models can be used as speaker feature extractors to represent input utterances into fix-length speaker vectors. Given two speaker vectors extracted from enrolling and test utterances, the distance of two speaker vectors and be regarded as a score to describe how likely they are from the same speaker. Similar to the i-vector model, back-end methods like LDA and PLDA [42] can also be used for better scoring. Currently, Many techniques have been introduced to improve ASV performance, including better structures [43–46], training scheme [47, 48] and pooling approaches [49, 50]. As a result, the current ASV models have achieved promising results on several benchmark datasets [43, 51–53].

In the meantime, spoofing attacks [10–13, 15–17] were proved to be effective techniques to fool ASV systems. Many works have investigated countermeasure techniques, also known as anti-spoofing techniques, to counter such attacks. [54–56] tried to use different input features and Gaussian Mixture Model (GMM) classifier to spotlight spoofing attacks. In further, a DNN-based solution [57] was proposed utilizing a combination of convolutional neural network (CNN) and recurrent neural network (RNN) structures. Following the development of more advanced DNN models, Resnet was introduced in [19, 22]to better grasp artifacts in attacking speech. In recent, graph attention networks (GAT) have been an emerging structure proposed by [58]. This model applies the self-attention mechanism to graph convolutional networks, with the capability to model the neighboring relationships of input representations. Improved by models such as AASIST [23], RawGAT-ST [24], the CM performance has reached a new high.

# 3. Method

## 3.1. SASV strategies

Given the enrollment utterance $x_e$ and the testing utterance $x_t$, spoofing-aware speaker verification (SASV) systems aim at telling $y_t = 1$ if $x_t$ comes from the same speaker as $x_e$, or $y_t = 0$ if $x_t$ comes from another speaker or $x_t$ is a spoofing attack. There are two typical strategies for constructing a SASV system: multi-task learning strategy and fusion-based strategy.

The multi-task learning strategy trains the models jointly with both speaker verification and anti-spoofing objectives, which is intuitive to be adopted. The two objectives share the same backbone and thus the features and embeddings, while each objective has their own predicting head and loss function. It is worth noting that speaker verification and anti-spoofing objectives are contradictory in some respects. The former drives the model to erase device and environment information to more robustly identify speakers; in contrast, the latter prompts the model to capture device and environment traces, then tells forged spoofing from authentic utterances [59]. Therefore, multi-task learning makes the system more complex and difficult to optimize, thus requiring additional supervising information, such as more training data and a particular training paradigm [32]. However, only ASVspoof dataset [60] has both speaker labels and spoofing labels to meet the requirements of the multi-task learning strategy, but its small number of speakers can barely support a generalized satisfactory speaker verification performance.

Alternatively, the fusion-based strategy has the potential to reach better SASV performance leveraging state-of-the-art CM models and ASV models trained on large-scale datasets. Considering its superiority, we propose our solutions for the SASV challenge based on the a novel multi-model and multi-level fusion strategy. Besides, the SASV Challenge 2022 also provides two baseline systems performing score-level and embedding-level fusion respectively, which will be described in the following subsection.

## 3.2. Baseline systems

The challenge organizer provides two baseline systems. Each system is based upon ECAPA-TDNN model [44] as the ASV subsystem and AASIST model [23] as the CM subsystem. The key difference between the two systems is how they fuse ASV and CM subsystem - Baseline1 adopts the score-level fusion, while Baseline2 adopts the embedding-level fusion.

For Baseline1, given a trial consisting of an enrollment utterance and a testing utterance, the ASV subsystem will yield corresponding speaker embeddings. Accordingly, the cosine similarity of the two embeddings will be calculated as the ASV score. Meanwhile, the testing utterance will be fed into the CM subsystem to derive a CM score. Subsequently, Baseline1 determines a final score by summing up the ASV and CM scores, then equal error rates (EERs) are calculated accordingly.

In comparison, Baseline2 involves the concatenation of three embeddings - one extracted from the enrollment utterance through the ASV subsystem; a second extracted in an identical fashion for the testing utterance; a third extracted from the same testing utterance through the spoofing CM subsystem. Then the concatenation of the above three embeddings is fed into a 3-layer perceptron, and the score will be predicted, indicating whether the trial belongs to the target or not.

## 3.3. Proposed fusion systems

Although achieving acceptable performances, baseline systems' fusion strategies are relatively simple and naive. Baseline1's score-level fusion does not guarantee that the ASV score and the CM score belong to a unified space and an identical magnitude. Baseline2 crudely concatenates three embeddings and throws the product into a DNN, lacking fine-grained fusion. These inadequacies motivate us to explore further possibilities of fusion strategies in SASV task.

In this paper, we proposed a multi-model and multi-level fusion strategy. In terms of the width, it employs multiple pre-

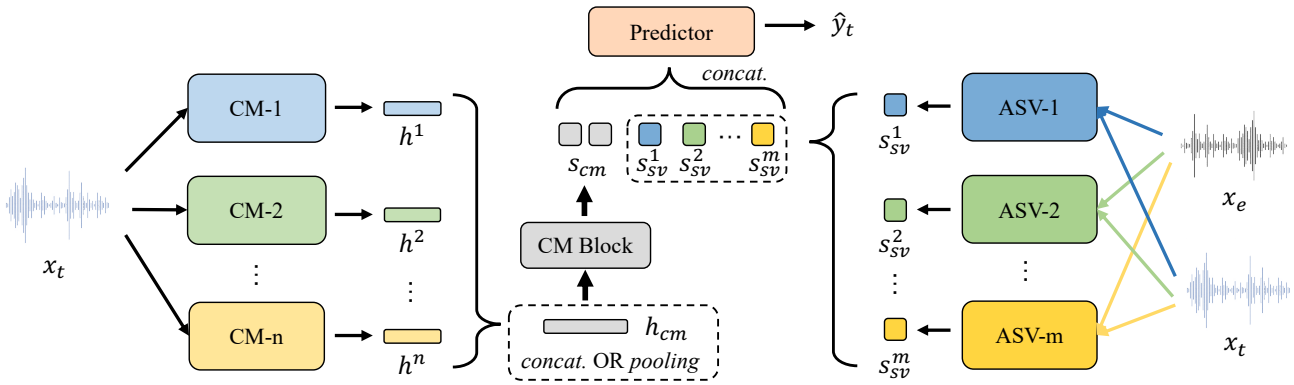
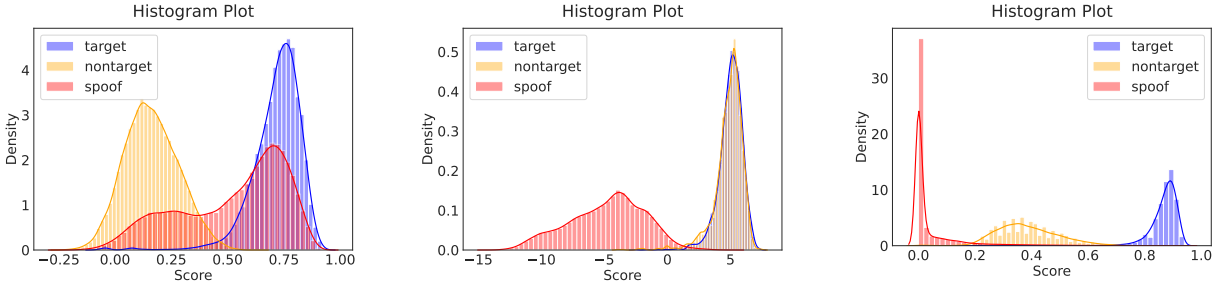



Figure 1: *The proposed multi-model & multi-level fusion framework.*




(a) ECAPA-TDNN (b) AASIST (c) Proposed top-5 ensemble system

Figure 2: *The histogram plots of the output scores predicted by ASV, CM, and the proposed top-5 ensemble system. Other proposed system variants also have similar histogram shapes as (c).*

trained ASV and CM models as plug-and-play components, where users can expand or shrink the scale of the model according to their needs. In terms of the depth of this strategy, it fuses CM embeddings to calculate a score in the first-level fusion, which is then integrated with the ASV models' outputs, and yields the final prediction in the second-level fusion.

### 3.3.1. *Overall structure*

The overall framework is shown in Figure 1, where $x_e$ and $x_t$ are the input enrollment and testing utterances, respectively; *ASV-1*, *ASV-2*, ..., *ASV-m* denote $m$ pre-trained ASV models; *CM-1*, *CM-2*, ..., *CM-n* denote $n$ pre-trained CM models. Given a trial $\{x_e, x_t\}$, a series of cosine scores $\{s_{sv}^1, s_{sv}^2, \ldots, s_{sv}^m\}$ are derived from $m$ ASV models. Given the testing utterance $x_t$, a series of CM embeddings $\{h^1, h^1, \ldots, h^n\}$ are extracted through $n$ CM models. Next comes the first level of fusion, where the $n$ embeddings are integrated into a $h_{cm}$ by concatenation or a pooling method. Our previous work [35] investigated the capacity of concatenation for SASV. In this work, we further extend the potential of concatenation by making the CM block in Figure 1 deeper, and we also adopt pooling methods to further improve the fusion performance. Further, $h_{cm}$ goes through a CM Block to better digest fused embeddings and then a 2-dimension countermeasure score $s_{cm}$ is predicted. With $s_{cm}$ and $\{s_{sv}^1, s_{sv}^2, \ldots, s_{sv}^m\}$ well prepared, they are concatenated and fed into the Predictor to yield the final prediction $\hat{y}_t$, meanwhile the second-level fusion is performed.

### 3.3.2. *Strategies in the first-level fusion*

For the first-level fusion, we attempt concatenation [35] or one of the four kinds of pooling methods to synthesize $h_{cm}$ separately, and have conducted extensive experiments accordingly. Suppose $\boldsymbol{H} = \{h^1, h^1, \ldots, h^n\}$, and the length of all the CM embeddings are projected into a same length $d_h$ by feed forward layers. The four candidate pooling methods are Temporal Average Pooling (TAP) [36], Temporal Statistics Pooling (TSP) [37], Self-attentive Pooling (SAP) [38], and Attentive Statistics Pooling (ASP) [39].

TAP is to calculate the mean value along the channels to obtain the $h_{cm}$.

TSP calculates channel-wise mean and standard deviation, then concatenates the mean vector and standard deviation vector together as $h_{cm}$.

In SAP, the self-attention mechanism takes $\boldsymbol{H}$ as input and outputs an annotation matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = softmax(tanh(\boldsymbol{H}^T \boldsymbol{W}_1)\boldsymbol{W}_2) \quad (1)$$

where $\boldsymbol{W}_1$ is a matrix of size $d_h \times d_a$; $\boldsymbol{W}_2$ is a matrix of size $d_a \times d_r$, and $d_r$ is a hyper-parameter that represents the number of attention heads; The $softmax()$ is performed column-wise. We set $d_r = 1$, therefore $\boldsymbol{A}$ degenerates into an annotation vector. Weighted by $\boldsymbol{A}$, $h_{cm}$ is calculated as the weighted mean:

$$h_{cm} = \tilde{\mu} = \boldsymbol{H}\boldsymbol{A} \quad (2)$$

For ASP, not only it calculate a attention-weighted mean as SAP do, but also it calculate a attention-weighted standard

deviation:

$$\tilde{\sigma} = \sqrt{\sum_{i=1}^{n} \alpha^i h^i \odot h^i - \tilde{\mu} \odot \tilde{\mu}} \quad (3)$$

where $\alpha^i$denotes the $i^{th}$ element of the annotation vector $\boldsymbol{A}$, $\odot$ represents the Hadamard product. By concatenating $\tilde{\mu}$ and $\tilde{\sigma}$, $h_{cm}$ is derived.

After one of the above pooling methods or concatenation, derived $h_{cm}$ goes through CM Block, which is a multi-layer perceptron, and generate a two-dimension score reflecting the possibilities the testing utterances is the target or not.

#### *3.3.3. Loss function*

Suppose the ASV and CM models parameters are well pre-trained and thus frozen, the learnable modules mainly include CM Block and the Predictor, which are multi-layer perceptrons. To prompt the model to learn to distinguish the target trials from the non-target and spoofing trials, we adopt the cross-entropy loss on $s_{cm}$ output by CM Block and $\hat{y}_t$ output by the Predictor respectively.

## 4. Experimental setup

### 4.1. Datasets

In the SASV Challenge 2022 [26], participants are restricted to utilise ASVspoof 2019 [60] and VoxCeleb2 [43] datasets for model development.

ASVspoof 2019 is a dataset used for the Third Automatic Speaker Verification Spoofing and Countermeasures Challenge in 2019 [60]. Leveraging the dataset, the creators wish to encourage further progress in automatic speaker verification and the reliability of spoofing countermeasures under the threats of the advances in text-to-speech and voice conversion technology. The dataset collects bona fide speech data captured from 107 speakers (46 males, 61 females), derived from the VCTK corpus [61]. For every speaker, spoofing trails are generated using text-to-speech, voice conversion approach. The 107 speakers are partitioned into three sub-sets, they are the training set (20 speakers with 2,580 bona fide trails and 22,800 spoofing trails), development set (20 speakers with 2,458 bona fide trails and 22,296 spoofing trails) and evaluation set (67 speakers with 7,355 bona fide trails and 63,882 spoofing trails). The way this dataset is collected and organized bridges the speaker verification and anti-spoofing tasks, which piques researchers' interest in investigating the spoofing-aware speaker verification (SASV) challenge. In the SASV Challenge 2022, development and evaluation protocols are provided, which list target, non-target and spoofing trials.

VoxCeleb2 is a benchmark dataset for speaker verification, extracted from videos uploaded to YouTube. In this work, VoxCeleb2's development set, containing over 1 million utterances for 5,994 celebrities, is used for ASV models' training.

### 4.2. Evaluation metrics

Three EERs, namely SV-EER, SPF-EER and SASV-EER are measured as the evaluation metrics, and SASV-EER is the main metric in the Challenge. Further details can be found in the SASV challenge [26].

### 4.3. Implementation details

For the ASV models, we use Resnet34 [62], ECAPA-TDNN [44] and MFA-Conformer [46]. For the countermeasure models, we use AASIST [23], AASIST-L, and RawGAT-ST [24], where AASIST-L is a light version of AASIST. The above models' parameters are pre-trained and fixed. The fusion model in Figure 1 is trained by Adam optimizer with an initial learning rate as 0.0001. We set the batch size as 32, and epoch number as 100. We use the development set of the ASVspoof 2019 dataset to select the model.

## 5. Experimental results and analysis

### 5.1. Results

As the requirements by SASV Challenge 2022, we evaluated systems on ASVspoof 2019 development and evaluation sets and reported SA-EER, SPF-EER and SASV EER, shown in Table 1. A1-A3 denote pure ASV systems; B1-B3 denote pure CM systems; C1-C2 denote two baselines provided by the Challenge organizer. D1-H8 are variants based on the proposed fusion strategy. D1-D9 denote the fusion systems using one ASV model and one CM model, e.g., 'ECAPA-TDNN + AASIST' denotes the fusion of ECAPA-TDNN as the ASV model, and AASIST as the CM model. E1-E3 denote systems fusing all three ASV models with one CM model. Note that D1-E3 ignore the first-level fusion. F1-F3 denote systems fusing one ASV model and all three CM models. G1-H8 represent systems incorporating all three ASV models and all three CM models, but with different first-level fusion strategies and different sizes of CM Block. For example, 'SV-ALL + CM-ALL-CAT-768' denotes its first-level fusion uses concatenation (abbreviated as 'CAT' in the table), and fused $h_{cm}$ is projected to 768 dimensions in CM Block's first layer. I1 is the ensemble system involving the top-5 best evaluation set SASV-EER systems in A1-H8. Figure 1 illustrates the histogram plots of three typical systems: (a) ECAPA-TDNN, a SOTA ASV system; (b) AASIST, a SOTA CM system; (c) Proposed top-5 ensemble system.

### 5.2. Observations and discussion

We have the following observations and analysis:

#### *5.2.1. Single-objective systems*

*Only using speaker verification models.* A1-A3, which are state-of-the-art ASV models, performed well on the speaker verification sub-task and achieved 1.86%, 1.38% and 1.08% SV-EERs on the evaluation set, respectively. However, they perform unacceptably on the anti-spoofing sub-task, yielding 30.75% 30.22%, 29.76% SPF-EER. Spoiled by the spoofing attacks, it is unpractical to perform SASV tasks using only the ASV models. Take ECAPA-TDNN model as an example, as shown in Figure 2 (a), the pure ASV model can separate target and non-target trials while can hardly distinguish spoofing trials from genuine trials. This phenomenon is predictable because the objective of speaker verification models tends to erase the in-congruent device and environment information to more robustly identify speakers. However, these real-world traces should have helped to defend against spoofing attacks.

*Only using anti-spoofing models.* In contrast, B1-B3, state-of-the-art CM models, can significantly discriminate spoofing utterances, but randomly guess on the speaker identification sub-task, mainly because of their speaker-unrelated objectives. They achieve SPF-EERs of 0.67%, 0.84%, 0.96% on anti-

Table 1: *Performance of all systems on the ASVspoof 2019 development and evaluation sets.*

| | System | SV-EER | | SPF-EER | | SASV-EER | |
|---|---|---|---|---|---|---|---|
| | | Dev | Eval | Dev | Eval | Dev | Eval |
| (A1) | ECAPA-TDNN | 1.64 | 1.86 | 20.28 | 30.75 | 17.37 | 23.84 |
| (A2) | MFA-Conformer | 1.61 | 1.38 | 19.94 | 30.22 | 16.91 | 23.28 |
| (A3) | Resnet34 | 1.68 | 1.08 | 17.40 | 29.76 | 14.62 | 22.69 |
| (B1) | AASIST | 46.01 | 49.24 | 0.07 | 0.67 | 15.86 | 24.38 |
| (B2) | AASIST-L | 48.30 | 49.04 | 0.13 | 0.84 | 15.72 | 24.81 |
| (B3) | RawGAT-ST | 51.25 | 49.24 | 0.34 | 0.96 | 15.96 | 24.85 |
| (C1) | Baseline1 [26] | 32.88 | 35.32 | 0.06 | 0.67 | 13.07 | 19.31 |
| (C2) | Baseline2 [26] | 12.87 | 11.48 | 0.13 | 0.78 | 4.85 | 6.37 |
| (D1) | MFA-Conformer + AASIST | 1.48 | 1.47 | 0.20 | 1.08 | 0.88 | 1.35 |
| (D2) | ECAPA-TDNN + AASIST | 1.48 | 1.58 | 0.20 | 1.06 | 0.99 | 1.42 |
| (D3) | Resnet34 + AASIST | 1.49 | 1.02 | 0.20 | 1.53 | 0.88 | 1.32 |
| (D4) | MFA-Conformer + AASIST-L | 1.68 | 1.68 | 0.15 | 2.03 | 1.21 | 1.83 |
| (D5) | ECAPA-TDNN + AASIST-L | 1.61 | 1.62 | 0.19 | 2.18 | 1.10 | 1.92 |
| (D6) | Resnet34 + AASIST-L | 1.68 | 1.69 | 0.18 | 2.01 | 1.15 | 1.84 |
| (D7) | MFA-Conformer + RawGAT-ST | 2.16 | 2.09 | 0.35 | 0.78 | 1.55 | 1.82 |
| (D8) | ECAPA-TDNN + RawGAT-ST | 2.56 | 2.17 | 0.04 | 0.78 | 1.81 | 1.94 |
| (D9) | Resnet34 + RawGAT-ST | 2.09 | 1.97 | 0.40 | 0.79 | 1.55 | 1.69 |
| (E1) | SV-ALL + AASIST | 1.42 | 1.30 | 0.27 | 1.61 | 0.81 | 1.41 |
| (E2) | SV-ALL + AASIST-L | 1.42 | 1.33 | 0.47 | 3.99 | 0.88 | 2.95 |
| (E3) | SV-ALL + RawGAT-ST | 1.82 | 1.64 | 0.40 | 0.82 | 1.28 | 1.39 |
| (F1) | MFA-Conformer + CM-ALL-CAT-256 | 1.91 | 1.66 | 0.20 | 0.64 | 1.01 | 1.30 |
| (F2) | ECAPA-TDNN + CM-ALL-CAT-256 | 1.39 | 1.73 | 0.20 | 0.74 | 0.81 | 1.40 |
| (F3) | Resnet34 + CM-ALL-CAT-256 | 1.28 | 1.12 | 0.26 | 1.43 | 0.74 | 1.32 |
| (G1) | SV-ALL + CM-ALL-CAT-256 | 1.27 | 1.20 | 0.20 | 1.15 | 0.81 | 1.17 |
| (G2) | SV-ALL + CM-ALL-CAT-512 | 1.35 | 1.15 | 0.20 | 1.12 | 0.74 | 1.14 |
| (G3) | SV-ALL + CM-ALL-CAT-768 | 1.34 | 1.12 | 0.20 | 0.99 | 0.81 | **1.08** |
| (G4) | SV-ALL + CM-ALL-CAT-1024 | 1.28 | 1.21 | 0.20 | 0.83 | 0.74 | **1.08** |
| (G5) | SV-ALL + CM-ALL-CAT-2048 | 1.35 | 1.10 | 0.23 | 1.41 | 0.74 | 1.31 |
| (H1) | SV-ALL + CM-ALL-TAP-768 | 0.08 | 0.99 | 0.02 | 1.10 | 0.51 | **1.02** |
| (H2) | SV-ALL + CM-ALL-TSP-768 | 1.21 | 1.12 | 0.20 | 1.47 | 0.74 | 1.31 |
| (H3) | SV-ALL + CM-ALL-SAP-768 | 1.15 | 1.04 | 0.17 | 0.93 | 0.54 | **0.99** |
| (H4) | SV-ALL + CM-ALL-ASP-768 | 1.18 | 1.37 | 0.15 | 1.58 | 0.67 | 1.51 |
| (H5) | SV-ALL + CM-ALL-TAP-1024 | 1.28 | 1.15 | 0.13 | 0.56 | 0.61 | **0.97** |
| (H6) | SV-ALL + CM-ALL-TSP-1024 | 1.11 | 1.12 | 0.20 | 1.77 | 0.61 | 1.43 |
| (H7) | SV-ALL + CM-ALL-SAP-1024 | 1.15 | 1.16 | 0.20 | 1.45 | 0.61 | 1.28 |
| (H8) | SV-ALL + CM-ALL-ASP-1024 | 1.51 | 1.68 | 0.40 | 1.01 | 1.08 | 1.49 |
| (I1) | Top-5 Ensemble | 1.08 | 1.01 | 0.20 | 0.71 | 0.67 | **0.89** |

spoofing, while yielding SV-EERs close to 50% for speaker verification. This phenomenon can also be observed in Figure 2 (b), where the target and non-target trials' distribution are almost totally overlapped. Affected by the awful SV performance, pure anti-spoofing models show even worse SASV performance than A1-A3.

#### *5.2.2. Baseline systems*

Compared to the above single-objective models, two baseline systems reveal superiority in the SASV challenge. Among them, Baseline2 achieves a SASV-EER of 8.75% on the evaluation set, which is better than Baseline1's 19.15%. We argue that the reason is that the simple score-level fusion used in Baseline1 does not guarantee that ASV and CM scores belong to a unified space with a consistent magnitude. As shown in Figure 2, the ASV score ranges from -1 to 1, while the CM score ranges from -20 to 15. Straightforward addition will make the ASV score submerged by the CM score. In comparison, Baseline2 uses a trainable deep neural network to digest ASV and CM embeddings better, which can contribute to distinguishing the target from non-target and spoofing trials. However, the performance of Baseline2 on the speaker verification sub-task is still not satisfactory, which motivates us to propose the multi-model & multi-level fusion strategy.

#### *5.2.3. Proposed fusion systems*

The proposed strategy fuses three SOTA ASV models (A1-A3) and three SOTA CM models (B1-B3). With different setting, a total of 28 variant fusion systems (D1-H8) are elaborately designed, and an additional ensemble system (I1) is constructed by integrating the top-5 best evaluation set SASV-EER systems from A1-H8. The best result of our previous work [35] is shown in G1. From Table 1, all the proposed systems outperform the baseline models with a large margin on the SASV-

EER metrics, while retaining universally good performances on SV-EER and SPF-EER. The proposed models perform consistently well on the speaker verification task, and quite a few can reach or even surpass the performance of the SOTA ASV models; while the models' performances on the anti-spoofing task have a more considerable variance. G1-H8 show better performance than D1-F3 in general, benefiting from the more comprehensive model fusion. Using concatenation for the first-level fusion, we attempted different CM Block sizes. From G1 to G5, as we increase the size of the CM Block, the SASV-EER of the system decreases. Until adding a layer with an output dimension of 2048 to the bottom of the CM Block, the model's performance degrades slightly because of the excessive model capacity. Since G3 or G4 have the best SASV-EER, we employ their CM Block settings and further explore the impact of replacing concatenation with different pooling methods. In H1-H8, under both CM block settings, TAP can bring more benefits to the system than other methods. Both statistics pooling methods (TSP, ASP) are less effective than TAP, SAP, and concatenation. A possible reason is that we only compute the standard deviation for up to three CM embeddings, which is not stable during training. Rather than representing useful knowledge, the standard deviation seems to be more of a noise. We argue that if more CM models enroll, the system will benefit from statistics pooling methods, and we will leave it as future work. We achieve the best individual system SASV-EER of 0.97%. Moreover, the top-5 ensemble system (I1) achieves a SASV-EER as low as 0.89%, 86% relative improvement compared to Baseline2, and 24% relative improvement compared to our previous work [35]. Figure 2 (c) shows that the distributions of the target, non-target and spoofing trials are well-separated, which verifies the effectiveness of the proposed method.

## 6. Conclusion

In this paper, a novel multi-model and multi-level fusion strategy is proposed to tackle the SASV task. The two-level fusion method can take advantage of both the state-of-the-art ASV and CM models. The best single fusion system achieves the SASV-EER of 0.97%. What's more, by ensembling the top-5 systems, the final SASV-EER reaches 0.89% on the evaluation set, which is 86% relative reduction compared to the best baseline, Baseline2, and 24% relative reduction compared to our previous work [35]. In the future work, we will introduce more CM models to investigate the potential of the proposed method.

## 7. References


[1] Z. Wu, T. Kinnunen *et al.*, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[2] T. Kinnunen, M. Sahidullah *et al.*, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," *ISCA (the International Speech Communication Association)*, 2017.

[3] M. Todisco, X. Wang *et al.*, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[4] J. Yamagishi, X. Wang *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *arXiv preprint arXiv:2109.00537*, 2021.

[5] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, and H. Li, "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP*. IEEE, 2022.

[6] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," *arXiv preprint arXiv:2103.11326*, 2021.

[7] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of hmm-based speaker verification systems against imposture using synthetic speech," in *Sixth European conference on speech communication and technology*, 1999.

[8] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in *Sixth International Conference on Spoken Language Processing*, 2000.

[9] B. L. Pellom and J. H. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 2. IEEE, 1999, pp. 837–840.

[10] Z. Wu and H. Li, "On the study of replay and voice conversion attacks to text-dependent speaker verification," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5311–5327, 2016.

[11] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *APSIPA*. IEEE, 2014, pp. 1–5.

[12] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *ICASSP*. IEEE, 2018, pp. 1962–1966.

[13] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," *arXiv preprint arXiv:2004.08849*, 2020.

[14] S. Liu, H. Wu, H.-y. Lee, and H. Meng, "Adversarial attacks on spoofing countermeasures of automatic speaker verification," in *ASRU*. IEEE, 2019, pp. 312–319.

[15] P. L. De Leon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," 2010.

[16] Z. Wu and H. Li, "Voice conversion versus speaker verification: an overview," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014.

[17] Z. Kons and H. Aronowitz, "Voice transformation-based spoofing of text-dependent speaker verification systems." in *INTERSPEECH*, 2013, pp. 945–949.

[18] J. Monteiro and J. Alam, "Development of voice spoofing detection systems for 2019 edition of automatic speaker verification and countermeasures challenge," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 1003–1010.

[19] J. Monteiro, J. Alam, and T. H. Falk, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Computer Speech & Language*, vol. 63, p. 101096, 2020.

[20] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[21] H. Wu, H.-C. Kuo, N. Zheng, K.-H. Hung, H.-y. Lee, Y. Tsao, H.-m. Wang, and H. Meng, "Partially fake audio detection by self-attention-based fake span discovery," *arXiv preprint arXiv:2202.06684*, 2022.

[22] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 132–137.

[23] 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," *arXiv preprint arXiv:2110.01200*, 2021.

[24] H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," *arXiv preprint arXiv:2107.12710*, 2021.

[25] X. Li *et al.*, "Replay and synthetic speech detection with res2net architecture," in *ICASSP*. IEEE, 2021, pp. 6354–6358.

[26] J.-w. Jung, H. Tak, H.-j. Shim, H.-S. Heo, B.-J. Lee, S.-W. Chung, H.-G. Kang, H.-J. Yu, N. Evans, and T. Kinnunen, "Sasv challenge 2022: A spoofing aware speaker verification challenge evaluation plan," *arXiv preprint arXiv:2201.10283*, 2022.

[27] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, and M. M. Doss, "On joint optimization of automatic speaker verification and anti-spoofing in the embedding space," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1579–1593, 2020.

[28] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and antispoofing in the $i$-vector space," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 821–832, 2015.

[29] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi, "Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion," in *Interspeech 2018-19th Annual Conference of the International Speech Communication Association*. ISCA, 2018.

[30] M. Sahidullah, H. Delgado, M. Todisco, H. Yu, T. Kinnunen, N. Evans, and Z.-H. Tan, "Integrated spoofing countermeasures and automatic speaker verification: An evaluation on asvspoof 2015," 2016.

[31] A. Kanervisto, V. Hautamäki, T. Kinnunen, and J. Yamagishi, "Optimizing tandem speaker verification and anti-spoofing systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 477–488, 2021.

[32] J. Li, M. Sun, and X. Zhang, "Multi-task learning of deep neural networks for joint automatic speaker verification and spoofing detection," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1517–1522.

[33] Y. Zhao, R. Togneri, and V. Sreeram, "Multi-task learning-based spoofing-robust automatic speaker verification system," *Circuits, Systems, and Signal Processing*, pp. 1–22, 2022.

[34] J. Yamagishi, M. Todisco *et al.*, "Asvspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database," 2019.

[35] H. Wu, J. Kang, L. Meng, Y. Zhang, X. Wu, Z. Wu, H. yi Lee, and H. Meng, "Tackling spoofing-aware speaker verification with multi-model fusion," 2022.

[36] L. You, W. Guo, L. Dai, and J. Du, "Multi-task learning with high-order statistics for x-vector based text-independent speaker verification," *arXiv preprint arXiv:1903.12058*, 2019.

[37] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP 2018*. IEEE, 2018, pp. 5329–5333.

[38] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification." in *Interspeech*, vol. 2018, 2018, pp. 3573–3577.

[39] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Interspeech*, 2018, pp. 2252–2256.

[40] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[41] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[42] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.

[43] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[44] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[45] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[46] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, and Z. Wu, "Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification," *arXiv preprint*, 2022.

[47] Z. Gao, Y. Song, I. V. McLoughlin, P. Li, Y. Jiang, and L.-R. Dai, "Improving aggregation and loss function for better embedding learning in end-to-end speaker verification system." in *INTERSPEECH*, 2019, pp. 361–365.

[48] J. Zhou, T. Jiang, Z. Li, L. Li, and Q. Hong, "Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function." in *Interspeech*, 2019, pp. 2883–2887.

[49] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.

[50] N. Chen, J. Villalba, and N. Dehak, "Tied mixture of factor analyzers layer to combine frame level representations in neural speaker embeddings." in *INTERSPEECH*, 2019, pp. 2948–2952.

[51] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[52] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7604–7608.

[53] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, "Cn-celeb: multi-genre speaker recognition," *Speech Communication*, 2022.

[54] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Sixteenth annual conference of the international speech communication association*, 2015.

[55] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," *ISCA (the International Speech Communication Association)*, 2015.

[56] M. Todisco, H. Delgado, and N. W. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients." in *Odyssey*, vol. 2016, 2016, pp. 283–290.

[57] C. Zhang, C. Yu, and J. H. Hansen, "An investigation of deep-learning frameworks for speaker verification antispoofing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 684–694, 2017.

[58] H. Tak, J.-w. Jung, J. Patino, M. Todisco, and N. Evans, "Graph attention networks for anti-spoofing," *arXiv preprint arXiv:2104.03654*, 2021.

[59] H.-j. Shim, J.-w. Jung, J.-h. Kim, and H.-j. Yu, "Integrated replay spoofing-aware text-independent speaker verification," *Applied Sciences*, vol. 10, no. 18, p. 6292, 2020.

[60] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[61] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.