# PAU Submission of NIST 2018 Speaker Recognition Evaluation

**Youzhi TU*, Yingke ZHU#, Man Wai MAK*, Dongpeng CHEN†, Weiwei LIN*, Brian K.W. MAK#, Zhuxin CHEN† and Weibin ZHANG†**

***Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR of China**

**#Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong SAR of China**

**†Voice AI, China**

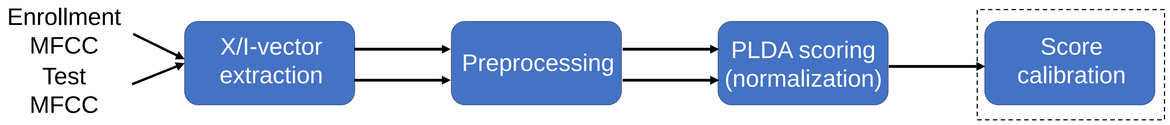
## Summary

- We show the fusion result of a GMM i-vector system and five DNN x-vector systems.
- PolyU has one i-vector system (S1) and two x-vector systems (S2 and S3) which are all trained on 8kHz speech data.
- The HKUST system uses 8kHz x-vectors for CMN2 (S4-CMN2) and 16kHz x-vectors for VAST (S4-VAST), applying the self-attention mechanism.
- VoiceAI implemented a larger 16kHz DNN (S5) to extract 1024-dimensional x-vectors for VAST data. The VAD is based on a bidirectional long short-term memory network.

## X/I-Vector/PLDA Training

- S1 (i-vector system) is based on gender-independent UBM with 2048 mixtures and 600 dimensional total variability matrix.
- The DNN in S2 was retrained using the Kaldi SRE16 x-vector recipe.
- S3 uses the pre-trained DNN from the Kaldi repository.
- In S4, the pooling-layer in the standard x-vector extractor was replaced by a self-attentive layer.
- S5 has a larger DNN architecture than the standard x-vector extractor. The dimension of the x-vector is 1024 rather than 512.
- We trained gender-independent out-of-domain PLDA models first, and then adapted them using suitable in-domain dataset for S1, S2, S3 and S4-CMN2. For S4-VAST and S5, we directly trained PLDA models using in-domain data.
- X/I-vector pre-processing: centering + LDA + length-norm
  - S1: 600 $\rightarrow$ 300
  - S2: 512 $\rightarrow$ 300
  - S3, S4: 512 $\rightarrow$ 150
  - S5: 1024 $\rightarrow$ 150

## Work Flow of Evaluation




## Datasets used for PLDA training, PLDA adaptation, calibration, and PLDA scoring

| Sys | Embedding | Sub-task | PLDA Training | PLDA adaptation | Snorm | PLDA scoring | | | Score calibration | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Adapt source | Scoring source | Centering | Adapt source | Scoring source | Centering |
| S1 | i-vector | CMN2 | sre04-12 + mx6 + aug | sre18-unlabeled | Yes | sre16-eval | sre18-dev | sre18-unlabeled | sre16-major | sre16-eval | sre16-major |
| | | VAST | | sitw-dev-enroll | | sitw-eval | | | | sitw-eval | sre18-unlabeled |
| S2 | x-vector | CMN2 | sre04-10 + mx6 + aug | sre18-unlabeled | Yes | sre16-eval | sre18-dev | sre18-unlabeled | sre16-major | sre16-eval | sre16-major |
| | | VAST | | sitw-dev-enroll | | sitw-eval | | | | sitw-eval | sre18-unlabeled |
| S3 | x-vector | CMN2 | sre04-10 + mx6 + aug | sre18-unlabeled | No | -- | sre18-dev | sre18-unlabeled | -- | sitw-eval | sitw-eval-test |
| | | VAST | | voxceleb1-enroll | | | | sre04-10+mx6+aug | | | voxceleb1-enroll |
| S4 | x-vector | CMN2 | sre04-10 + mx6 + aug | sre18-unlabeled | Yes | sre18-unlabeled | sre18-dev | sre18-unlabeled | sre16-major | sitw-eval | sitw-eval-test |
| | x-vector16k | VAST | voxceleb1-2 + aug | -- | No | -- | | sitw-eval-test | -- | sitw-dev | sre18-unlabeledU |
| S5 | x-vector16k | VAST | voxceleb1-2 | -- | No | -- | sre18-dev | -- | -- | -- | -- |

The column “Adapt source” indicates the data source for computing the S-norm parameters.
*sre18-unlabeled*: unlabeled data in SRE18. *sre18-unlabeledU*: 16kHz upsampled version of *sre18-unlabeled*.
*i-vector* and *x-vector*: speaker embedding based on 8kHz speech data. *x-vector16k*: speaker embedding based on 16kHz speech data.

## System Fusion

- We have three fused systems denoted as Fuse1, Fuse2 and Fuse3, respectively.
- Fuse1 is a linear weighted fusion of S1+S2 (S1 and S2 were first fused by Bosaris) and S4.
- Fuse2 is a linear weighted fusion of Fuse1 and S3.
- Fuse3 is formed by fusing Fuse1 and S5 using Bosaris.

| Fused systems | Individual systems for fusion | Sub-task | Fusion equation |
|---|---|---|---|
| Fuse1 | S1, S2, S4 | CMN2 | $0.4(\omega_0+\omega_1 S1+\omega_2 S2)+0.6S4$ |
| | | VAST | $0.9(\omega_0'+\omega_1' S1+\omega_2' S2)+0.1S4$ |
| Fuse2 | S1, S2, S3, S4 | CMN2 | $0.8(0.4(\omega_0+\omega_1 S1+\omega_2 S2)+0.6S4)+0.2S3$ |
| | | VAST | $0.6(0.9(\omega_0'+\omega_1' S1+\omega_2' S2)+0.1S4)+0.4S3$ |
| Fuse3 | S1, S2, S4, S5 | VAST | $\omega_{b0}+\omega_{b1}(0.9(\omega_0'+\omega_1' S1+\omega_2' S2)+0.1S4)+\omega_{b2} S5$ |

S1-S5 are scores of individual systems. $\omega_i$ and $\omega_i'$ are fusion weights computed by Bosaris.

## Performance

- **SRE18 Development**

Individual systems

| Sys | CMN2 | | | VAST | | | Both |
|---|---|---|---|---|---|---|---|
| | EER | mDCF | aDCF | EER | mDCF | aDCF | aDCF |
| S1 | 13.14 | 0.684 | 0.778 | 7.82 | 0.490 | 0.819 | 0.798 |
| S2 | 8.86 | 0.567 | 0.644 | 9.47 | 0.481 | 0.642 | 0.643 |
| S3 | 8.60 | 0.546 | 0.556 | 7.41 | 0.498 | 0.535 | 0.545 |
| S4 | 7.39 | 0.553 | 0.565 | 4.53 | 0.416 | 0.671 | 0.618 |
| S5 | -- | -- | -- | 4.94 | 0.523 | 0.576 | -- |

Fused systems

| Sys | CMN2 | | | VAST | | | Both |
|---|---|---|---|---|---|---|---|
| | EER | mDCF | aDCF | EER | mDCF | aDCF | aDCF |
| Fuse1 | 6.72 | 0.499 | 0.509 | 5.35 | 0.407 | 0.407 | 0.458 |
| Fuse2 | 6.57 | 0.496 | 0.510 | 4.94 | 0.412 | 0.486 | 0.498 |
| Fuse3 | -- | -- | -- | 5.35 | 0.449 | 0.572 | -- |

- **SRE18 Evaluation**

Individual systems

| Sys | CMN2 | | | VAST | | | Both |
|---|---|---|---|---|---|---|---|
| | EER | mDCF | aDCF | EER | mDCF | aDCF | aDCF |
| S1 | 13.70 | 0.741 | 0.771 | 18.61 | 0.682 | 0.712 | 0.741 |
| S2 | 9.02 | 0.607 | 0.647 | 13.02 | 0.608 | 0.630 | 0.639 |
| S3 | 10.05 | 0.594 | 0.597 | 13.76 | 0.636 | 0.655 | 0.626 |
| S4 | 8.47 | 0.565 | 0.568 | 18.73 | 0.737 | 0.902 | 0.735 |
| S5 | -- | -- | -- | 14.29 | 0.551 | 0.671 | -- |

Fused systems

| Sys | CMN2 | | | VAST | | | Both |
|---|---|---|---|---|---|---|---|
| | EER | mDCF | aDCF | EER | mDCF | aDCF | aDCF |
| Fuse1 | 7.68 | 0.520 | 0.523 | 13.02 | 0.591 | 0.643 | 0.583 |
| Fuse2 | **7.64** | **0.513** | **0.519** | **12.06** | 0.572 | 0.600 | 0.560 |
| Fuse3 | -- | -- | -- | 13.02 | **0.522** | **0.543** | -- |

## References

1. Yingke Zhu, Tom Ko, David Snyder, Brian Mak, and Daniel Povey, “Self-attentive speaker embeddings for text-independent speaker verification,” *Proc. Interspeech 2018*, pp. 3573–3577, 2018.
2. Carlos Busso Fei Tao, “Bimodal recurrent neural network for audiovisual voice activity detection,” *Proc. Interspeech 2017*, pp. 1938–1942, 2017.