The STAP System for CN-Celeb Speaker Recognition Challange 2022

Shiwei Jiang, Junyu Chen, Qian Liu, Yuhua Qian

Institute of Big Data Science and Industry, Shanxi University, China

June 27, 2022

Presented by Shiwei Jiang

Overview

1. Datasets & Data Augmentation
2. Models
    - Front-end backbones
    - Pooling method
    - Loss functions
    - Back-end
3. Results
4. Conclusion

Datasets & Data Augmentation

Training datasets

we only used CN-Celeb-T as our training data which contains 2,793 speakers and 632,740 utterances for both fixed track and open track in this challenge.

But we filtered the shortest samples with the duration less than 1 second and got 630,281 utterances for model training.

Data augmentation

an offline 3-fold[1] speed augmentation.
an online augmentation chain[1] for speech wave including multiple augments and each of them has probability to be activated.
Specially, we replaced MUSAN with CN-Celeb2 for noise addition

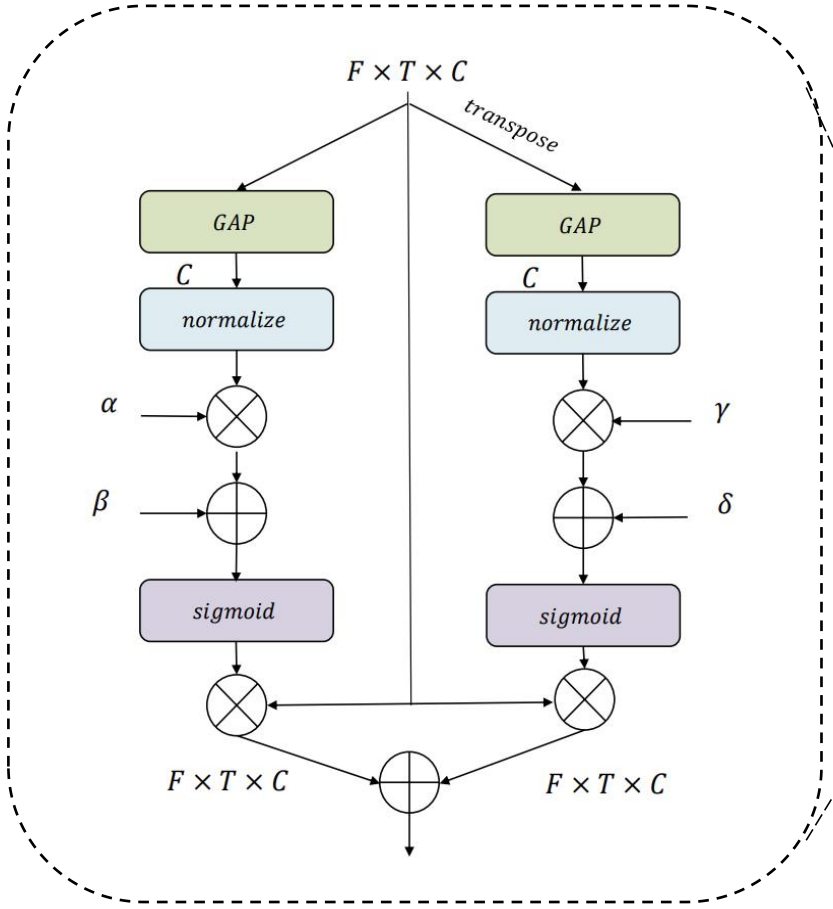
SpecAug[2] is also used for speech Fbank, and BatchAug[3] is applied for front-end training.

Feature extraction

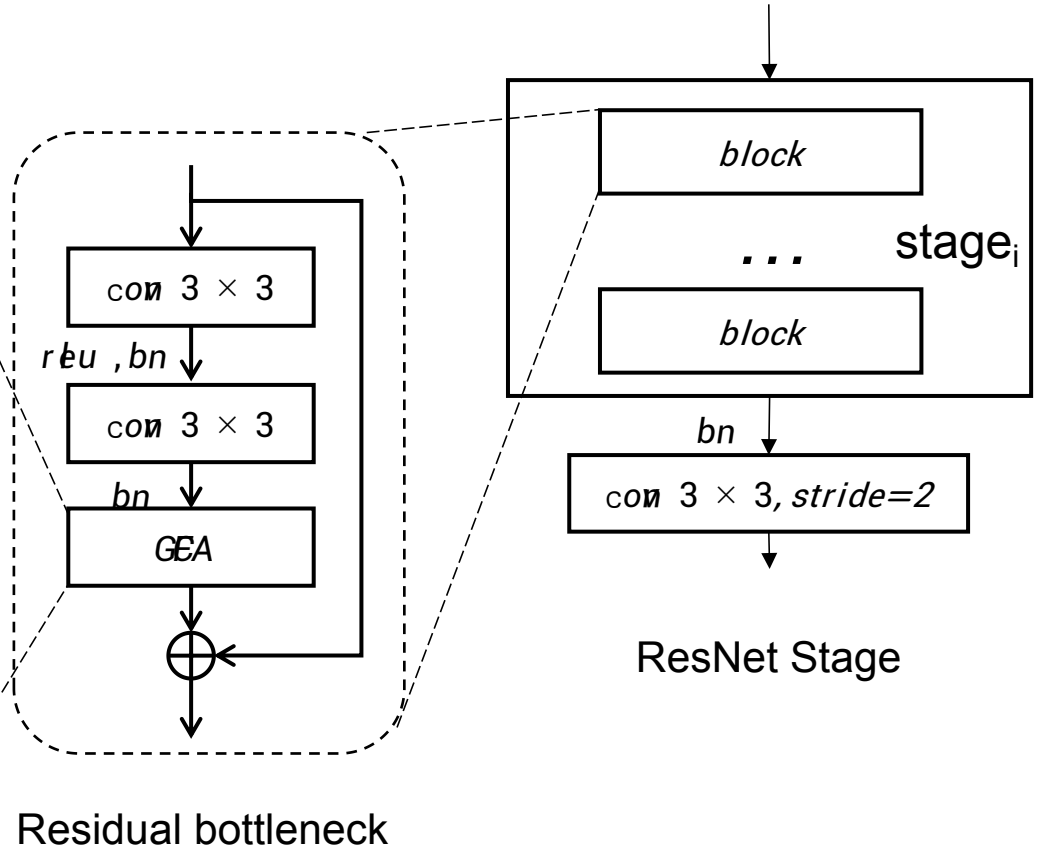
80-dimensional Fbank features
no VAD

1. Zhao Miao, Ma Yufeng, Min. Liu, and Minqiang. Xu, "The speakin system for voxceleb speaker recognition challange 2021," ArXiv, abs/2109.01989.
2. Daniel S. Park, William Chan, Yu Zhang et, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," Interspeech 2019
3. E. Hoffer, T. Ben-Nun, I. Hubara, N. Giladi et, "Augment your batch: Improving generalization through instance repetition," CVPR, 2020

ResNet-based



GCFA: Gated Channel-Frequency Attention



Residual bottleneck

ResNet Stage

The number of residual blocks of each stage is (2,2,12,2)
The number of channels of each stage is (64,96,192,384)

TDNN-based

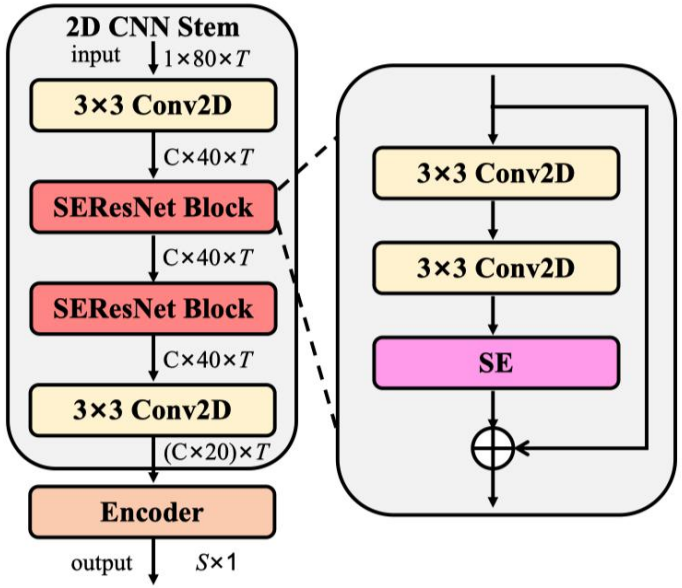
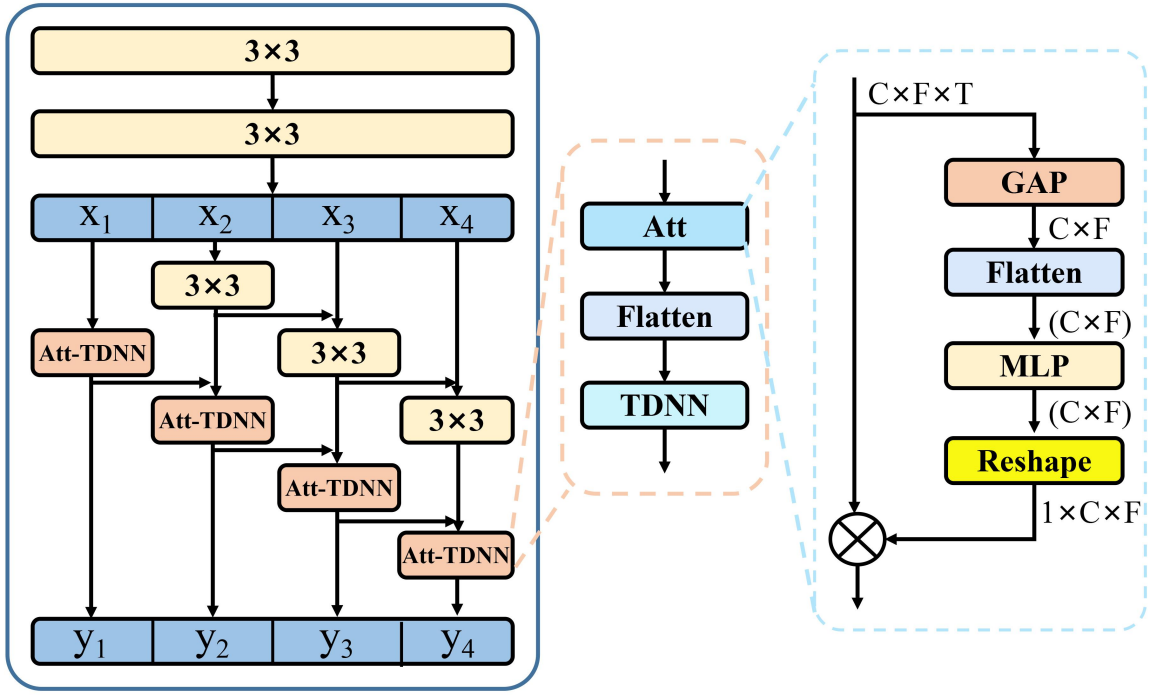


Figure 2: Structure of the CNN-Encoder

CNN-ECAPA-TDNN[1]



MFA-ECAPA[2]

1. Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "Integrating Frequency Translational Invariance in TDNNs and Frequency Positional Information in 2D ResNets to Enhance Speaker Verification," Interspeech 2021, pp. 2302 – 2306.

2. T. Liu, Rohan Kumar Das, Kong Aik Lee, and H. Li, "Mfa: Tdnn with multi-scale frequencychannel attention for text-independent speaker verification with short utterances," ICASSP 2022, pp. 7517 – 7521.

Transformer-based

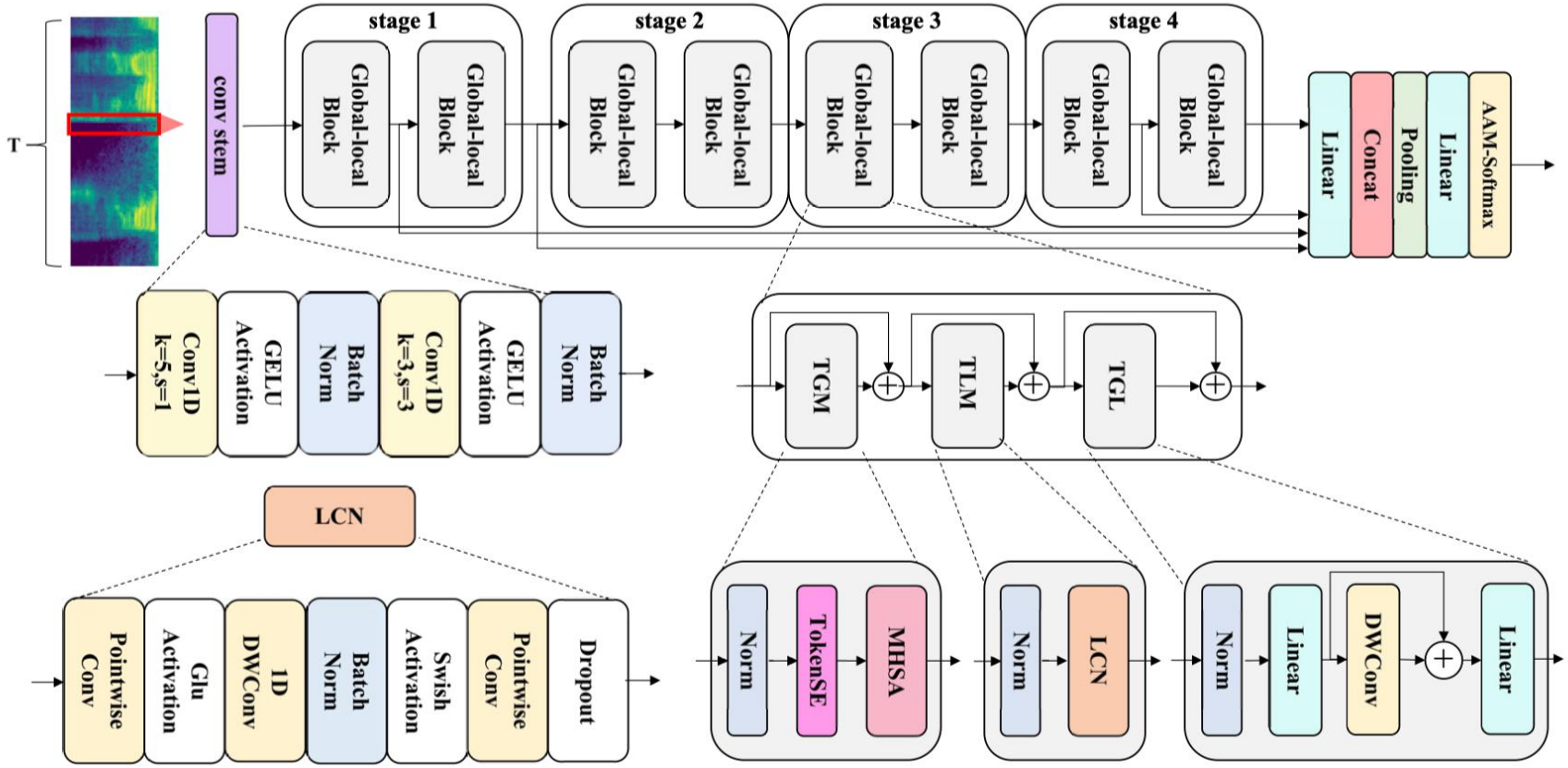


Figure 3: The overall architecture of Speaker-ViT

1. Speaker-ViT

2. CNN-Speaker-ViT(CNN-Encoder + Speaker-ViT)

Shiwei Jiang, Xinyan Liang, Yuhua Qian. “Speaker-ViT: Global and Local Vision Transformer for Speaker Verification” (under review)

Channel-independent Attentive Statistical Pooling

Suppose obtained frame-level feature $h=[h_1,h_2,...,h_T]$, with $h_t \in R^d$. We can gain a channel-independent and context-dependent scalar weight through a softmax layer

$$\alpha_{t,c} = \frac{\exp\left(\boldsymbol{v}_c^T f(\boldsymbol{W}\boldsymbol{h}_t + \boldsymbol{b}) + k_c\right)}{\exp\left(\sum_\tau^T \boldsymbol{v}_c^T f(\boldsymbol{W}\boldsymbol{h}_\tau + \boldsymbol{b}) + k_c\right)} \qquad (7)$$

where the parameters $W \in R^{R\times C}$, $b \in R^{R\times 1}$ and $v_c \in R^{R\times 1}$. The score $\alpha_{t,c}$ represents the importance of each frame given the channel c. Then, calculate the weighted mean and standard deviation of channel c and concatenate them to get pooled representation

B. Desplanques, J. Thienpondt, and K Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," Interspeech 2020, pp. 3830－3834.

Subcenter AdaFace Softmax Loss Function

The format of margin-based softmax loss function is as formula 8. And AdaFace softmax loss is defined as formula 9. Where g_{angle} and g_{add} is the space margin decided by embedding norm z_i.

We also introduced subcenter method to alleviate the effect of noisy and low-quality samples as formula 10.

s, m, h as the hyperparameters is 32, 0.2, 0.333 respectively and the number of subcenters is 2 in this challenge.

$$\mathcal{L} = -\log \frac{\exp\left(f\left(\theta_{i,y_i}, m\right)\right)}{\exp\left(f\left(\theta_{i,y_i}, m\right)\right) + \sum_{j \neq y_i}^{n} \exp\left(s\cos\theta_{i,j}\right)} \quad (8)$$

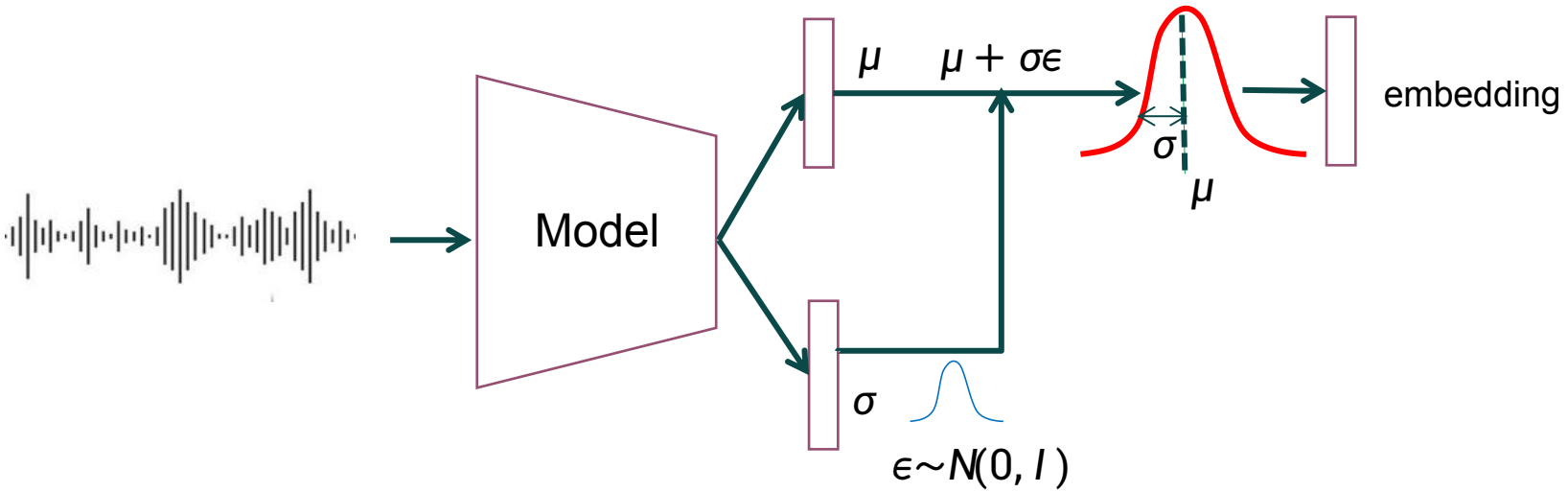
$$f\left(\theta_{i,j}, m\right)_{\text{AdaFace}} = \begin{cases} s\cos\left(\theta_{i,j} + g_{\text{angle}}\right) - g_{\text{add}} & j = y_i \\ s\cos\theta_{i,j} & j \neq y_i \end{cases} \quad (9)$$

$$g_{\text{angle}} = -m \cdot \widehat{\|\boldsymbol{z}_i\|}, \quad g_{\text{add}} = m \cdot \widehat{\|\boldsymbol{z}_i\|} + m.$$

$$\cos\left(\theta_{i,j}\right) = \max_{1 \leq k \leq K} \left(\|\boldsymbol{z}_i\| \cdot \|\boldsymbol{W}_{j,k}\|\right) \quad (10)$$

Minchul Kim, Anil K Jain, and Xiaoming Liu, "Adaface: Quality adaptive margin for face recognition," CVPR 2022.

Data Uncertainty Learning Loss Function



$$e_i = \mu_i + \epsilon\delta_i, \epsilon \in N(0, I) \quad (11)$$

$$L_{DUL} = L_{class}(e_i) + \lambda KL(N(z_i|\mu_i, \delta_i^2)||N(\epsilon|0, I)) \quad (12)$$

Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei, "Data uncertainty learning in face recognition," CVPR, 2020, pp. 5709－5718.

Attention-based back-end

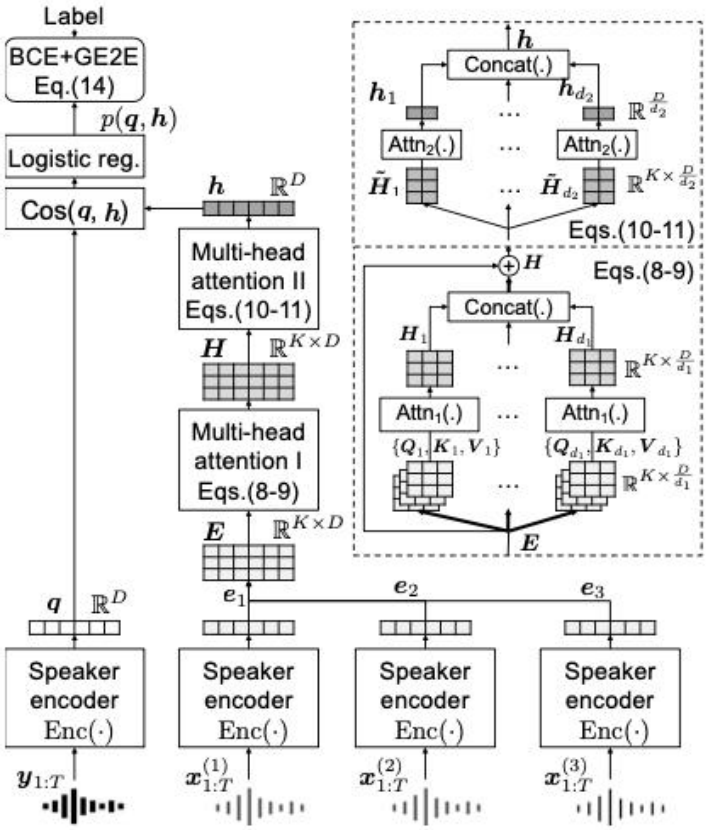


Fig. 1. Back-end architecture. Dashed boxes show detailed implementation of two multi-head attention blocks.

- Front-end with back-end finetuning
- 6 second duration
- Only GE2E loss function
- No noise addition augmentation

Chang Zeng, Xin Wang, Erica Cooper, Xiaoxiao Miao, and Junichi Yamagishi, "Attention back-end for automatic speaker verification with multiple enrollment utterances," ICASSP, 2022, pp. 6717－6721.

Ablation Study

Table 1: Ablation Study on Speaker-ViT with 2793 speakers. + here denotes stacking our methods.

Methods	CN-Celeb.E(TTA)		CN-Celeb.E(Full)	
	EER(%)	minDCF	EER(%)	minDCF
Speaker-ViT	8.178	0.4629	8.077	0.4471
+ genres noise	7.823	0.4505	7.756	0.4401
++ Subcenter-AdaFace	9.012	0.4318	8.955	0.4193
+++ back-end	7.282	0.4240	7.215	0.4136
+++ CNN-Encoder	8.679	0.4228	8.522	0.4161
++++ back-end	7.102	0.4186	7.091	0.4112

Table 2: Ablation Study on ResNet-Att with 2793 speakers. + here denotes stacking our methods.

Methods	CN-Celeb.E(TTA)		CN-Celeb.E(Full)	
	EER(%)	minDCF	EER(%)	minDCF
ResNet	7.789	0.4389	7.474	0.4300
+ GCFA	7.592	0.4238	7.535	0.4185
++ DUL loss	7.429	0.4206	7.311	0.4181

Results of Various Sub-systems and Fusion System

Table 3: Performance of Sub-Systems and Fusion System.

Sub-Systems	**CN-Celeb.E(TTA)**		**CN-Celeb.E(Full)**	
	EER(%)	**minDCF**	**EER(%)**	**minDCF**
Speaker-ViT	9.012	0.4318	8.955	0.4193
S1: Speaker-ViT + back-end	7.282	0.4240	7.215	0.4136
ResNet-Att	7.502	0.4234	7.232	0.4174
S2: ResNet-Att + back-end	7.187	0.4039	7.170	0.4023
CNN-ECAPA-TDNN	8.037	0.4137	7.829	0.4076
S3: CNN-ECAPA-TDNN + back-end	6.928	0.4048	7.006	0.4007
MFA-ECAPA	8.237	0.4162	8.127	0.4062
S4: MFA-ECAPA + back-end	**6.860**	0.3974	**6.905**	0.3932
CNN-Speaker-ViT	7.851	0.3957	7.885	0.3922
S5: CNN-Speaker-ViT + back-end	6.888	**0.3832**	6.911	**0.3771**
Fusion System	**EER(%)**		**minDCF**	
S1 $\sim$ S5 (submited)	**5.728**		**0.3399**	

Other Results

We splited the whole test trials into 22 sub-trials according to whether the genre of test utterances is included or not in the target speakers' enrollements.

minDCF for each sub-trials is shown in table 4.

“Cross” means the genre of speech is out of enrollment and “Same” is vice versa.

Table 4: Performance under Cross-Genre and Same-Genre Conditions.

	Advertisement	Drama	Entertainment	Interview	Live Broadcast	Movie	play	Recitation	Singing	Speech	Vlog
Cross	0.6902	0.9065	0.5339	0.7520	0.5745	0.9286	-	-	0.9055	0.4797	0.9560
Same	0.3333	0.2714	0.3815	0.2707	0.2223	0.5690	0.0800	-	-	0.1167	0.3523
Total	0.6762	0.3354	0.4112	0.3092	0.2389	0.5553	0.5336	0.1811	0.9055	0.1374	0.3428

Performance of our model is still poor on condition of cross-genre compared with the same-genre.

Conclusion

In this challenge, first, we modified the ResNet architecture and proposed a new attention module. Second, we introduced Speaker-ViT, a backbone based on the transformer structure. We also introduced AdaFace, DUL loss, and attention-based back-end to improve the robustness of the systems.

Then, we ensembled three different architectures to obtain good results on speaker verification of CNSRC 2022. The final result of our system was 0.3399 minDCF and 5.728% EER.

In addition, we showed the observations on the gain of each method and the performance of our system under different genres.

The End, Thanks