System Description for BIT-SR System

Xinmei Su[1], *Qingran Zhan*[1], *Chenguang Hu*[1], *Xiang Xie*[1,2]

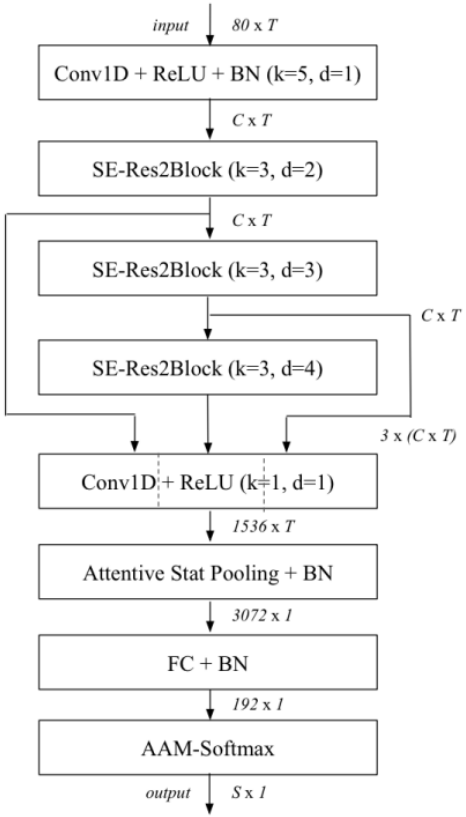
[1]School of Information and Electronics, Beijing Institute of Technology, Beijing, China
[2]Shenzhen Research Institute, Beijing Institute of Technology, Shenzhen, China
suxinmei2022@126.com

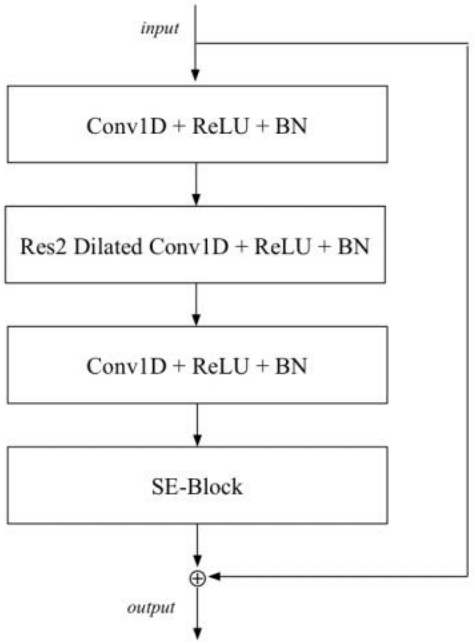
Reporter： Xinmei Su

Beijing Institute of Technology

Related work

- **ECAPA-TDNN**





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

Proposed Models

- **Overview**

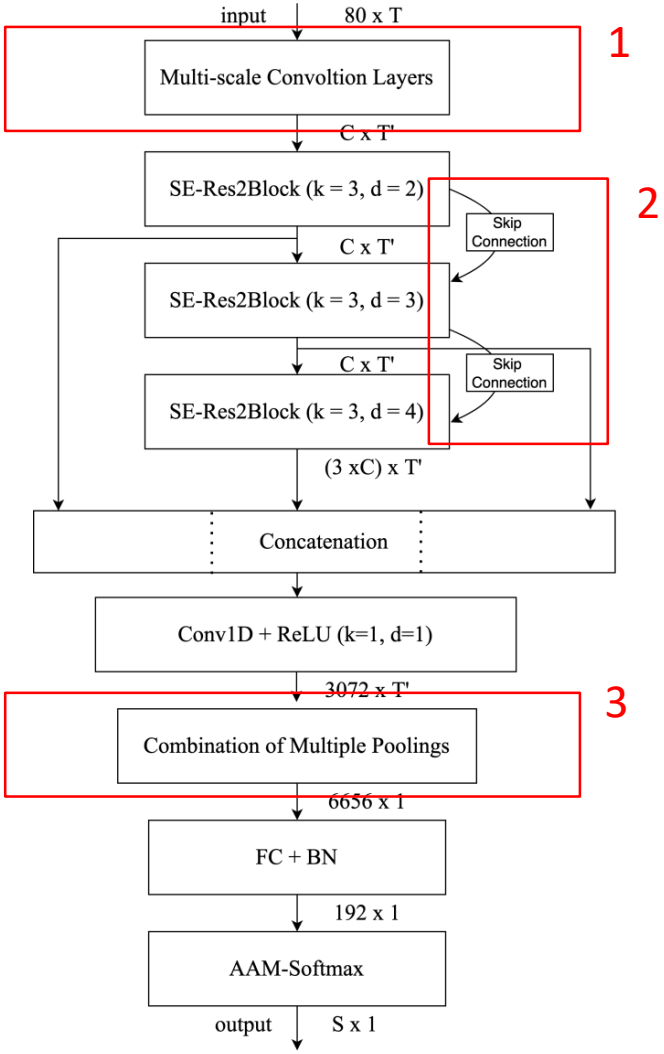


Figure 2: The whole architecture of our proposed system.

Proposed Models

- **Multi-scale Convolution Layers**

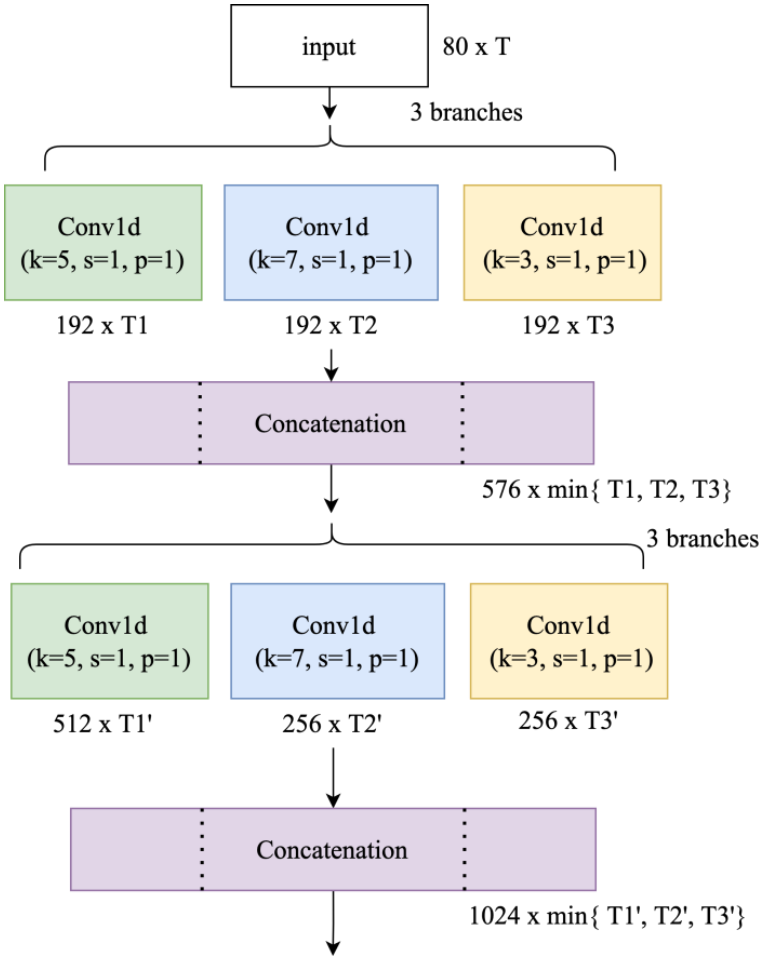


Figure 3: The multi-scale convolution structures with three different scales of kernels, where k, s, p denotes kernel size, stride and padding respectively.

Proposed Models

- **Residual Blocks in SE-Res2blocks**

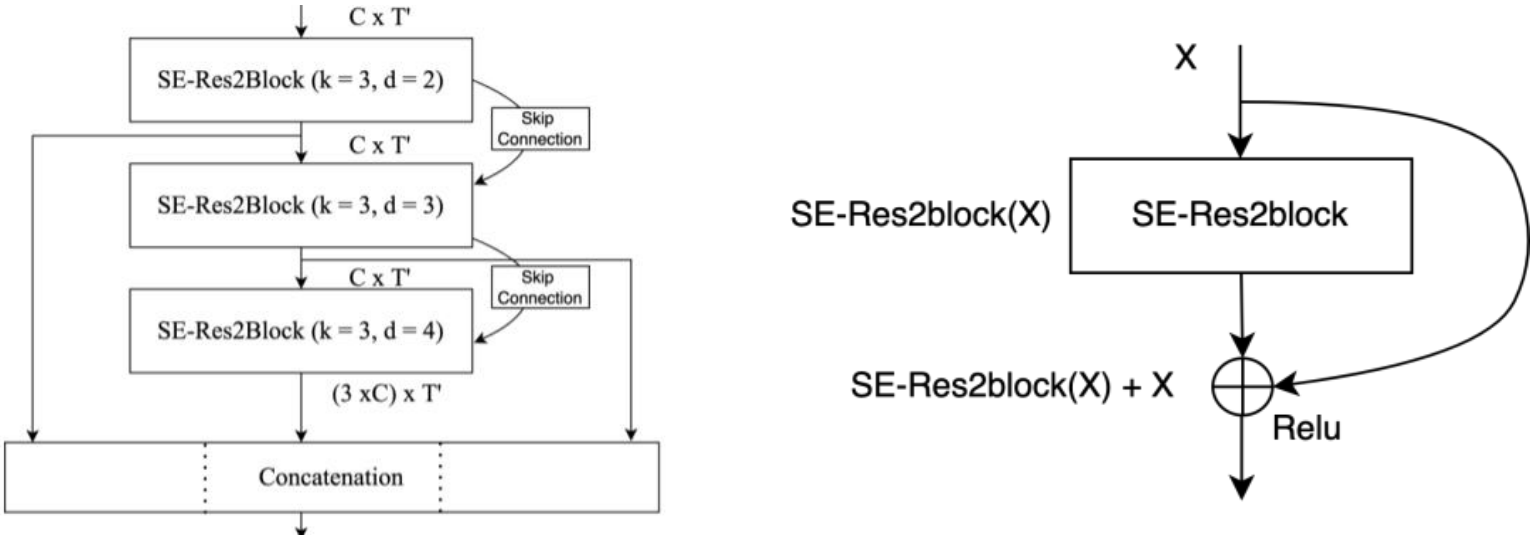


Figure 4: The residual block of the SE-Res2block.

[2] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//European conference on computer vision. Springer, Cham, 2016: 630-645.

Proposed Models

- **Combination of Multiple Poolings**

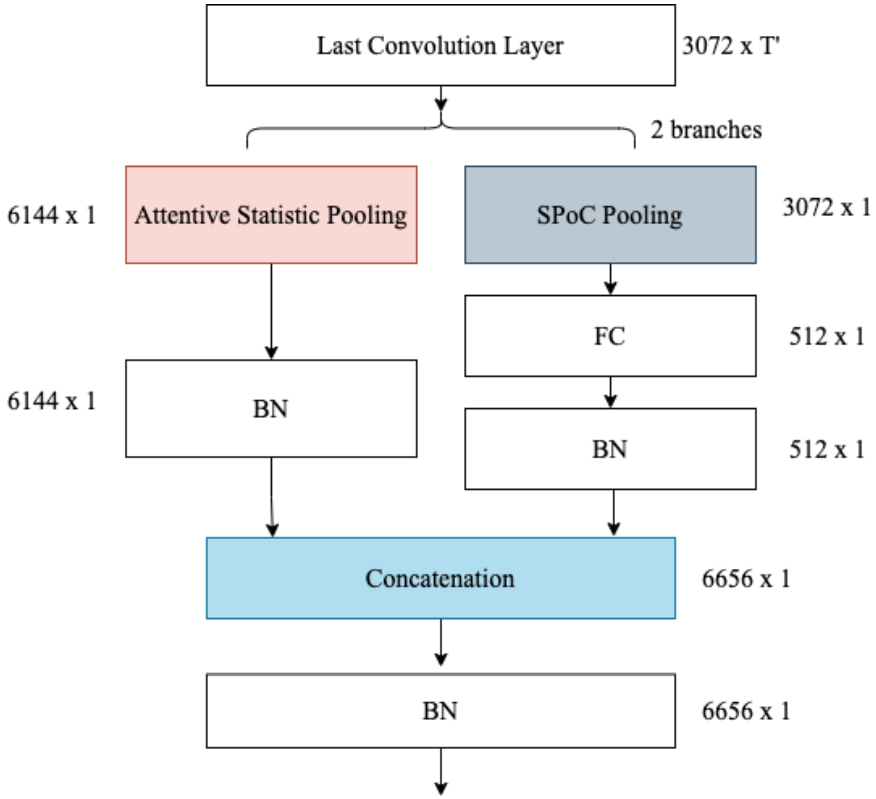


Figure 5: The combination of attentive statistic pooling and SPoC pooling, where the last convolution layer denotes the 1-D CNN after concatenation of multiple SE-Res2blocks.

Proposed Models

- **SPoC(sum poolings of convolutions)**

Given a fixed-length utterance U, the output of the hidden layer h is 2D tensor with a dimension of $C \times T$, where C denotes the number of feature maps and T denotes the length of frames. Let h_t $(C \times 1)$ and h_c $(1 \times T)$ denotes the hidden tensor of frame t and channel c. Thus, the hidden layer h can be a set of h_c defined as:

$$h = [h_1, h_2, ..., h_C]^T, c \in \{1, 2, ..., C\}$$

Then, the sum pooling method of channel c can be defined as:

$$\psi_{sumpool_c}(U) = \frac{1}{|h_c|} \sum_{t=1}^{T} h_t$$

The SPoC pooling vector is then produced by a fully-connected layer and an 1-D batch normalization layer represented as follows:

$$P_{SPoC} = f_{1d-BN}(W \cdot \psi_{sumpool_c}(U))$$

[3] Babenko A, Lempitsky V. Aggregating local deep features for image retrieval[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1269-1277.

Experiments

- **Datasets**

Table 1: *Data profile of the datasets we adopt in SR tasks. It is noted that the CN-Celeb.E in CN-Celeb1 is not used by us in experiments.*

Stages	Datasets	Speakers	Utterances	Hours
Train	CN-Celeb1	997	126532	271
	CN-Celeb2	1996	524787	1084
	Aishell	400	141600	165
Dev	SR.dev (target)	5	5	Not given
	SR.dev (pool)	Not given	20050	Not given
Test	SR.test (target)	25	25	Not given
	SR.test (pool)	Not given	500250	Not given

[4] Fan Y, Kang J W, Li L T, et al. Cn-celeb: a challenging chinese speaker recognition dataset[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7604-7608.
[5] Li L, Liu R, Kang J, et al. CN-Celeb: multi-genre speaker recognition[J]. Speech Communication, 2022, 137: 77-91.
[6]Bu H, Du J, Na X, et al. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline[C]//2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA). IEEE, 2017: 1-5.

Experiments

- **Acoustic Feature**
  - 3 seconds of fixed-length chunks
  - Input features: 80-dimensional mel-filterbanks
  - No voice activity detection (VAD) is applied during training.
  - Data augmentation recipes include:
    - Specaugment
    - Speed perturbation with rates of 0.95 and 1.05
    - Adding noise and reverberation in Room Impulse Response and Noise Database (RIRs) with the signal-noise ratio (SNR) ranges from 0 to 15
- **Training Settings**
  - Speechbrain platform with PyTorch framework
  - Speaker embeddings : 192-D vectors
  - **First** 20 epochs: the Adam optimizer is used with a learning rate of $1e^{-5}$ on the CN-Celeb1 dataset, except the CN-Celeb1 evaluation set
  - **Following** 20 epochs: Fine-tuning method of adjusting the learning rate to $1e^{-6}$ on both CN-Celeb1, CN-Celeb2 and Aishell
  - The margin and the scale parameters of the Additive Angular Margin (AAM) Loss are set to 0.2 and 30

Results

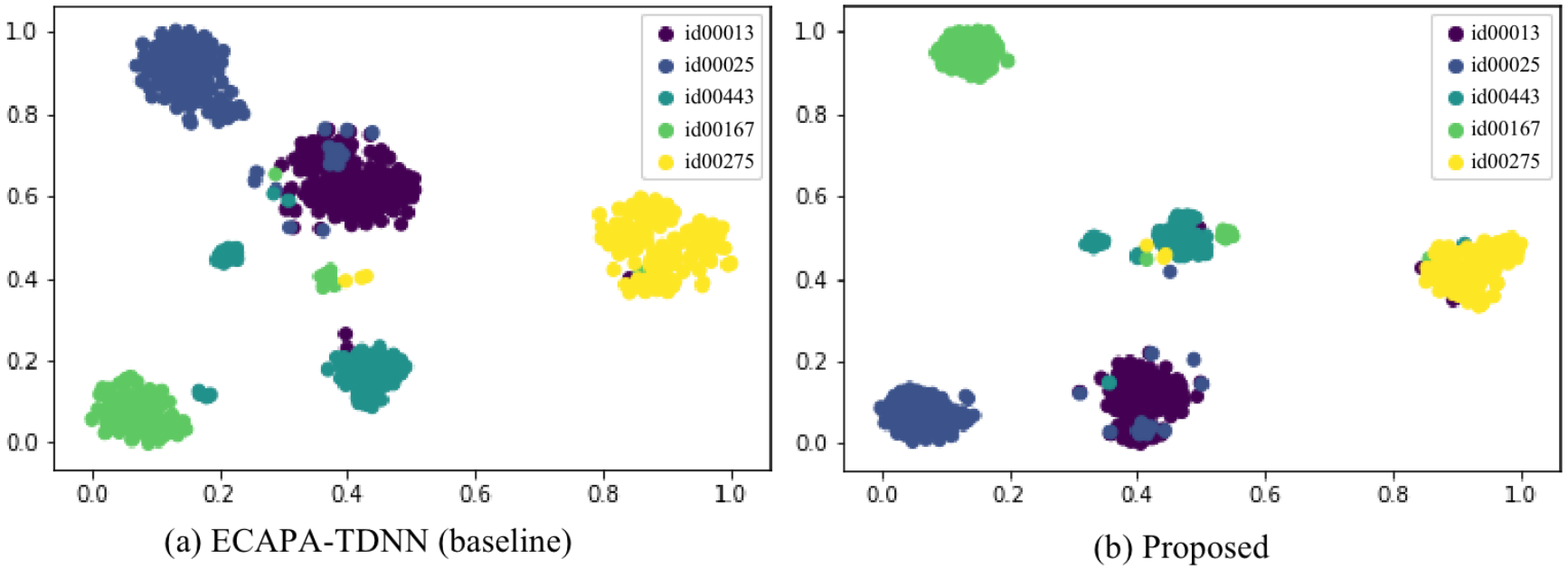
- **Comparision with the baseline system and ablation study**

Table 2: *Comparision of mAP for several different systems on dev and test set. MC denotes multi-scale convolution layers, MP denotes combination of multiple poolings and residual denotes the skip connection layers.*

System	mAP	
	Dev	Test
Baseline (ECAPA-TDNN)	0.726	0.5773
Proposed	**0.868**	**0.7085**
Proposed − residual	0.860	0.6844
Proposed − MP	0.840	0.6439
Proposed − MC	0.851	0.6755
ECAPA-2-layer-TDNN	0.765	0.6199

Results

- **Qualitative analysis**



(a) ECAPA-TDNN (baseline)

(b) Proposed

Figure 5: Visualization of the speaker embeddings in different systems. (a) represents the baseline system and (b) represents the proposed system.

[7] Wattenberg M, Viégas F, Johnson I. How to use t-SNE effectively[J]. Distill, 2016, 1(10): e2.

Conclusions

Three steps we take in our system:

- Multi-scale convolution layers are applied to substitute the single-scale TDNN layer.
- The residual layers in SE-Res2blocks can better solve the degenerate problems in deep networks.
- Concatenating the SPoC and statistic attentive pooling can provide more spatial information than single-scale pooling.
- All steps of our enhancements achieve better performances comparing to the ECAPA-TDNN baseline system.

Thanks