RawNeXt: Speaker verification system for variable-duration utterances with deep layer aggregation and extended dynamic scaling policies

Paper #3386

Ju-ho Kim, Hye-jin Shim, Jungwoo Heo, and Ha-Jin Yu

School of Computer Science, University of Seoul, South Korea

서울시립대학교 UNIVERSITY OF SEOUL

Overview

❖Speaker verification (SV): The task of determining whether the identity of an anonymous voice matches the target speaker

❖Problems: Variable-duration input utterance degrades the reliability of SV system
- Insufficient speaker-specific information of short utterance
- SV systems operating in a fixed way with manually designed layers

❖Proposed model: **RawNeXt**
- **Apply deep layer aggregation**: Enhance speaker information by iteratively and hierarchically aggregating features
- **Propose extended dynamic scaling policy**: Process features according to the length of the utterance
- 28.7% and 28.4% relative improvement compared to baseline for full-length result and mean result of 1,2, and 5s lengths for the VoxCeleb1 evaluation set

Baseline architecture with raw waveform

❖Input feature of models: Raw waveform
1. Data-driven manner on less-processed data can extract discriminative representations suitable for SV tasks
2. Minimal hyper-parameter search of acoustic feature pre-processing

❖DNN architecture: A variant of ResNeXt[1]
- Contain the grouped convolutional layers (Number of group: 32)
- Input: Raw waveform (59,049 sample)
- Output: Speaker embedding (512 dim)

❖Detailed architecture

Level	Block structure	# Blocks	Output
Convs	Conv(3, 3, 128)	1	2,187×128
	Conv(3, 1, 128) Maxpool(3)	2	
Stage 0	Conv(1,1,256) Conv(3,1,256), C=32	2	729×256
Stage 1	Conv(1,1,256) Maxpool(3)	4	243×256
Stage 2	Conv(1,1,512) Conv(3,1,512), C=32	4	81×512
Stage 3	Conv(1,1,512) Maxpool(3)	2	27×512
Pooling	ASP	1	1,024
Embedding	FC(512)	1	512

Experiments & Results

❖ Experiment configurations
- Training dataset: VoxCeleb2
  - 6,112 speakers
- Evaluation dataset: VoxCeleb1
  - 40 speakers
- Batch size: 320
- Test utterance duration : 1s, 2s, 5s and full length
- Performance comparison : Equal error rate (EER)
- Optimizer: AMSGrad
- Training epoch: 80
- Weight decay: 10^{-4}
- Learning rate (LR): $10^{-3} \rightarrow 10^{-7}$

❖ Exp1: Comparison with recently proposed SV system for variable-duration utterances
- Proposed RawNeXt outperforms other models for all test conditions
- Compared to baseline, 28.7% improvement for full-length test / 28.4% improvement for mean result of 1,2, and 5-sec lengths
- RawNeXt demonstrates superior generalization and robustness to variable-length utterances

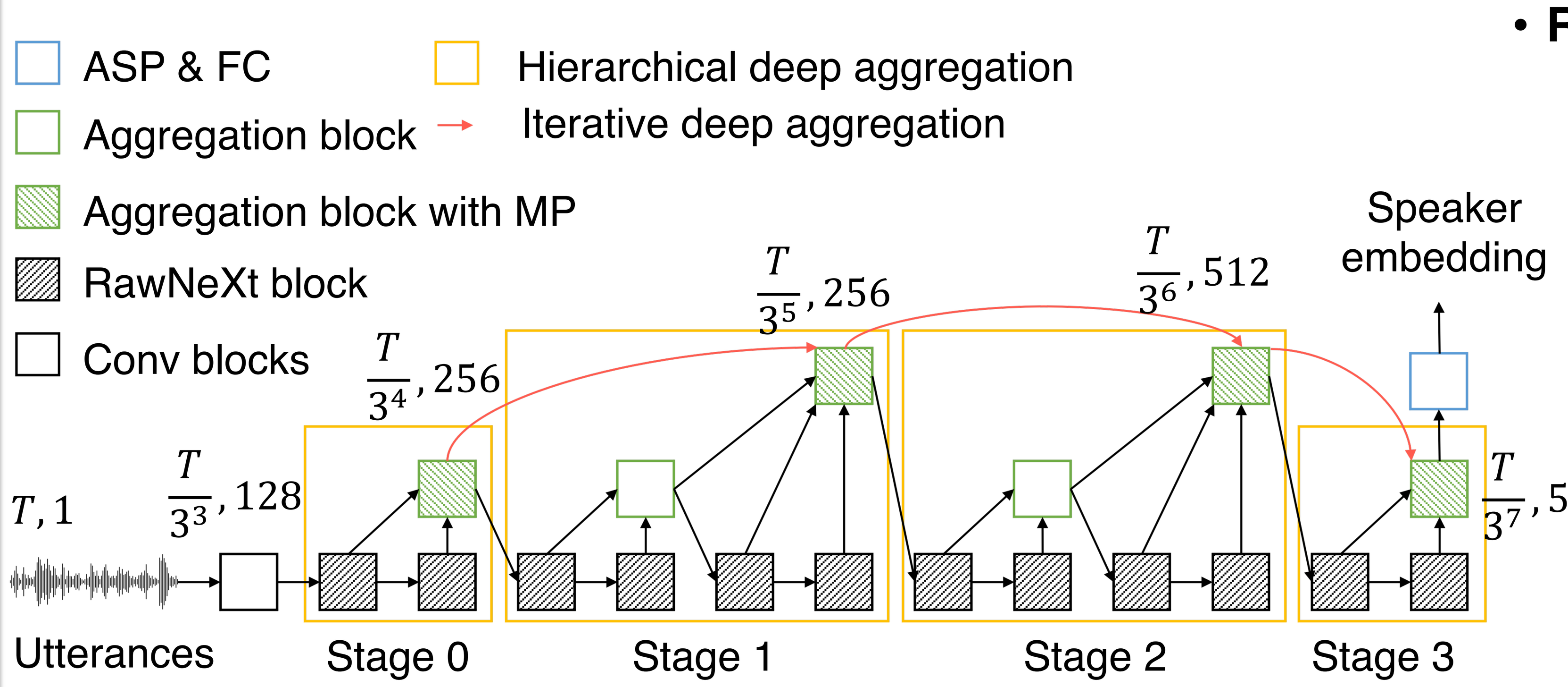
Model	Input Feature	Loss Function	Test utterance length (EER, %)			
			1s	2s	5s	full
MESA+FPM[4]	MFB-64	A-Softmax	5.92	3.38	2.17	1.98
ResNet34[5]	MFB-40	Softmax+PN	4.49	2.88	2.04	1.91
ResNeXt	Waveform	Softmax	6.12	3.68	2.45	2.16
RawNeXt	Waveform	Softmax	**4.47**	**2.58**	**1.72**	**1.54**

RawNeXt with Deep Layer Aggregation & Extended Dynamic Scaling Policy

❖Combining features of multiple layers for variable-duration SV
- Yield context-rich representations by merging intermediate features of various time scales

1.Deep layer aggregation (DLA)[2]
- Apply to derive speaker embeddings by fusing features in a more iterative and hierarchical manner for utterances of various lengths
- Iterative deep aggregation module: Enrich temporal context information by merging the different time resolution of features
- Hierarchical deep aggregation module: Enhance spectral context information by combining the feature channels of different levels
- Aggregation block: Learn to select important information from the multiple inputs and project it into a single output

- **RawNeXt structure**



- f_i^r :1d convolutional layer of the i-th path in the r resolution branch
- D: Downsampling function(average pooling layer)
- U_i^r: Upsampling function (transposed convolutional layer)
- σ: Batch normalization layer & rectified linear unit

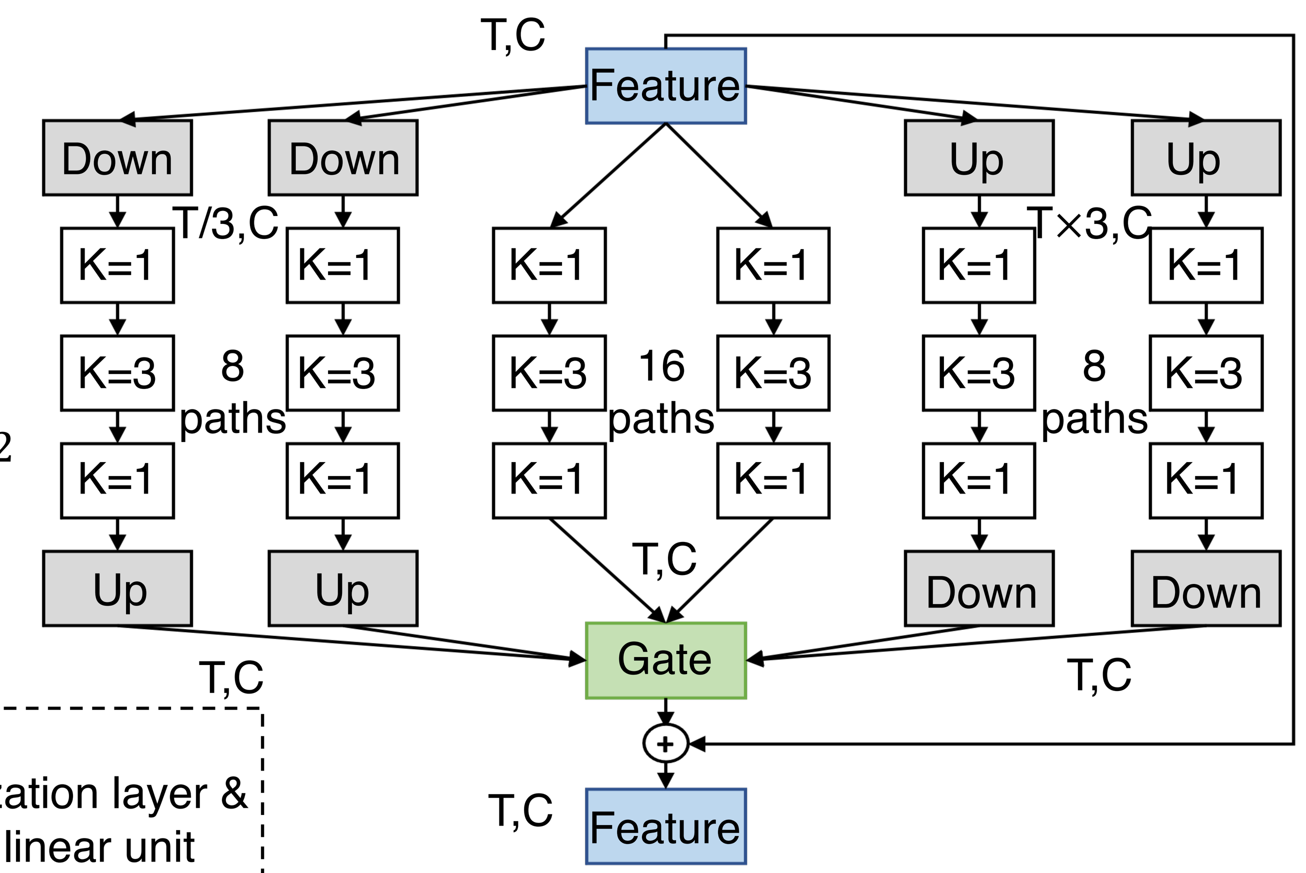
❖Elastic[3]: Processing images with various scales in vision tasks
- Learn a scaling policy from data by combining the features output by the original path and downsampling path of each block

2.Extend dynamic scaling policy (EDSP)
- Propose for utterance of arbitrary lengths based on Elastic
- Utilize three resolution branches and a gate module
- Low, original, and high resolution branches: Feature extraction with receptive fields of different sizes

$$F^l(x)=\sum_{i=1}^{8} U_i^l(f_i^l(D(x))),\quad F^o(x)=\sum_{i=1}^{16} f_i^o(x),\quad F^h(x)=\sum_{i=1}^{8} D(f_i^h(U_i^h(x)))$$

- Gate module: Selectively merge the activation of each branch according to the length of input utterance by using self-attention mechanism
- RawNeXt block with skip-path: $B(x)=\sigma(Gate(F^l(x),F^o(x),F^h(x))+x)$

- **RawNeXt block architecture**



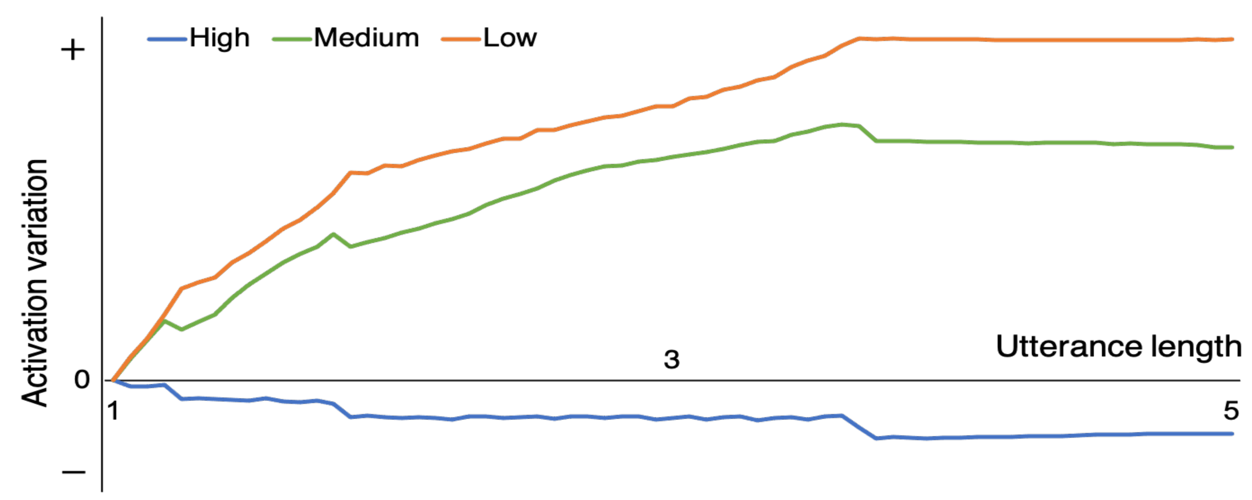
❖ Exp2: Ablation experiments of RawNeXt components
- #1: ResNext (Baseline), #7: RawNeXt (Proposed)
- Performance improves as each module is applied
- The motivations of each method are well aligned with the goal of variable-duration utterance SV

Model	D	E	G	U	Test utterance length (EER, %)			
					1s	2s	5s	full
#1	x	x	x	x	6.12	3.68	2.45	2.16
#2	o	x	x	x	4.82	2.98	2.08	1.93
#3	x	o	x	x	5.39	3.18	2.16	1.95
#4	o	o	x	x	4.66	2.94	2.13	1.94
#5	o	o	o	x	4.67	3.01	2.08	1.88
#6	o	o	x	o	4.65	2.81	1.94	1.82
#7	o	o	o	o	**4.47**	**2.58**	**1.72**	**1.54**

❖ Exp3: Variation score for mean activation of each resolution path according to the input utterance length
- Score at each r resolution branch by differences of mean activations between L and a 1-second utterance

$$S_L^r=\frac{1}{TC}\left(\sum_{t=1}^{T}\sum_{c=1}^{C}x_{Ltc}^r-\sum_{t=1}^{T}\sum_{c=1}^{C}x_{1tc}^r\right)$$

- RawNeXt extracts speaker information with appropriate resolutions by dynamically applying scaling policies according to the length of the utterance



1. S. Xie et. al., Aggregated residual transformations for deep neural networks, CVPR 2017.
2. F. Yu et. al., Deep layer aggregation, CVPR 2018.
3. H. Wang et.al., Elastic: Improving cnns with dynamic scaling policies, CVPR 2019.
4. Y. Jung et. al., Improving multi-scale aggregation using feature pyramid module for robust speaker verification of variable-duration utterances, Interspeech 2020.
5. S. Kye et. al., Supervised attention for speaker recognition, SLT 2021.