GRAPH ATTENTIVE FEATURE AGGREGATION FOR TEXT-INDEPENDENT SPEAKER VERIFICATION

Hye-jin Shim[1†], Jungwoo Heo[1†], Jae-han Park[2], Ga-hui Lee[2], and Ha-Jin Yu[1*]
[1]School of Computer Science, University of Seoul, [2]KT Corporation
[†]Equally contributed, [*]Corresponding author

SCAN ME!

서울시립대학교 UNIVERSITY OF SEOUL

2854

Overview

❖ Speaker Verification (SV)

- Process of verifying a person's claimed identity using their enrollment and test utterances
- 2 Steps: frame-level feature extraction, utterance-level feature aggregation

❖ Utterance-level feature aggregation

- Aggregate frame-level features into a single utterance-level feature
- Gated Recurrent Units, Learnable Dictionary Encoding, attention

❖ Research background

- Sequential information may not be the key in text-independent SV [1, 2]
- Attention cannot model each frame pair's intra relationships

❖ Proposed Method

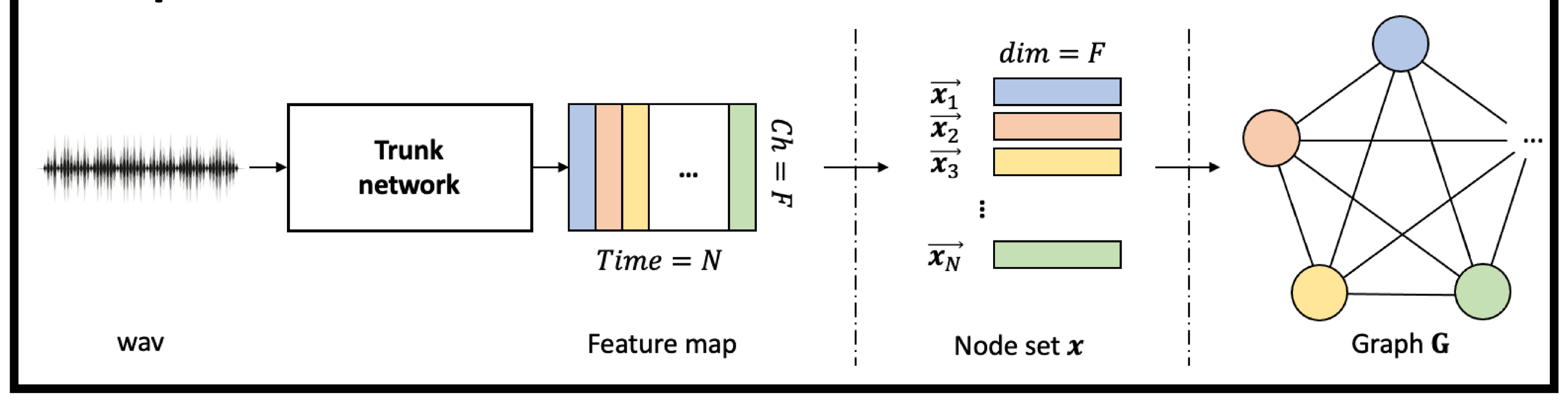
- **Graph attentive feature aggregation**
  - Improved feature aggregation method
  - Utilizing graph attention networks[3]
  - Entire frame-level features are aggregated considering their inter-relationships

❖ Our Contributions

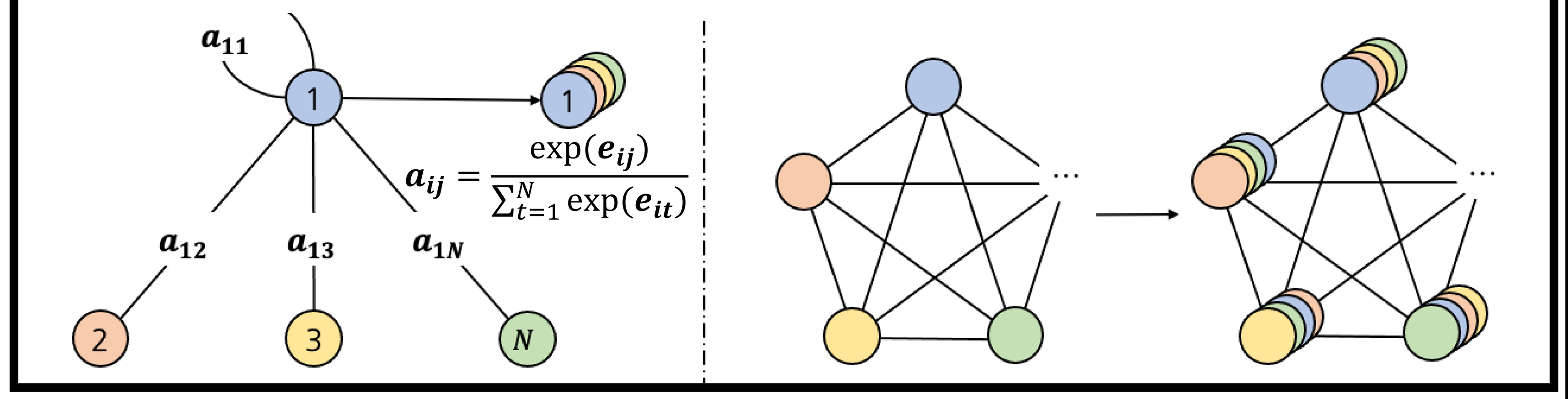
- Proposed graph attentive feature aggregation
  - First approach using GNN for feature aggregation in SV research
- Explored various readout and structure
- Validated the effectiveness of the proposed method using both spectrogram and raw wave form baselines

Proposed method

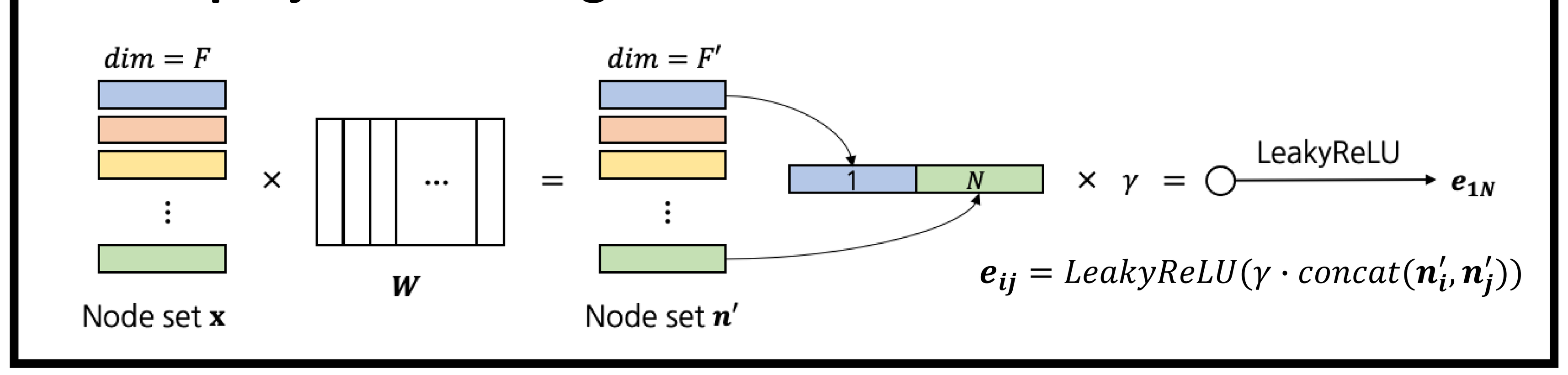
Graph formulation



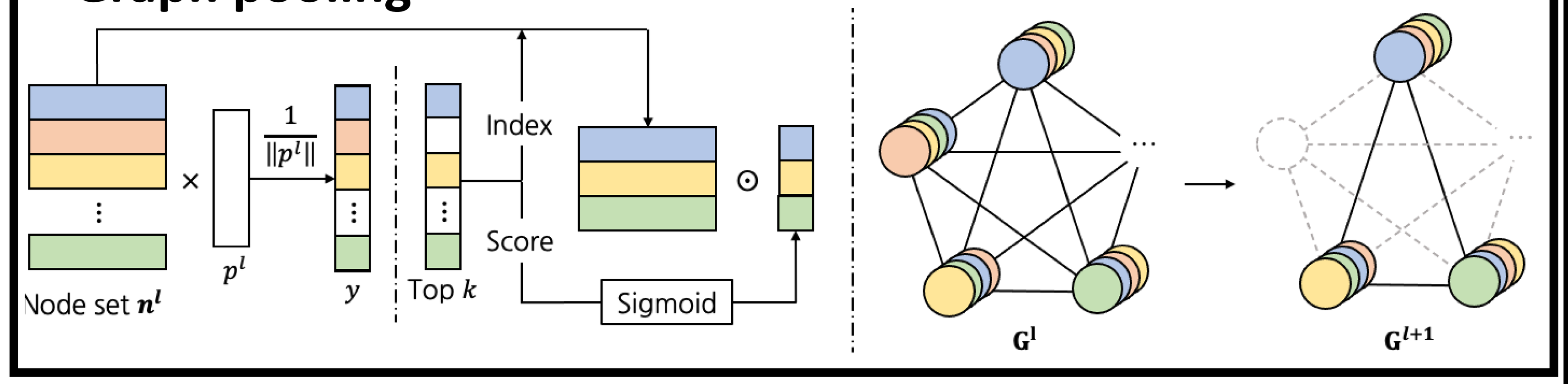
Self attention



Node projection & Edge score calculation



Graph pooling



- **Graph formulation** - formulate a graph from a feature map
- **Node projection** - Project x into F' dimensional space by matrix multiplication with learnable parameter $W \in R^{F \times F'}$
- **Edge score calculation** – calculate edge scores ($\gamma \in R^{2F' \times 1}$) Since the total number of nodes is N, $N \times N$ scores are calculated
- **Self attention** - performs self-attention on every nodes
- **Graph pooling** - reduce the original graph into a sub-graph by removing less informative nodes
- **Readout** - Combines the processed nodes into a single node

Experiments & Results

❖ Dataset

- Train: VoxCeleb2[5] developement set
- Test: VoxCeleb1[4] test set

❖ Baseline

- Used two baseline to check the effect according to the input domain
  - SE-ResNet:
    - Input: 40-dimensional mel-filterbank features
    - Modified Clova system
  - RawNet2:
    - Input: raw waveform
    - Modified original RawNet2

❖ Results

- Table 1
  - Both systems improved performance with fewer parameters than baselines
  - Proposed graph attentive feature aggregation was effective
- Table 2
  - The proposed system showed superior performance in both spectrogram and raw waveform domain
  - Our system achieved state-of-the-art performance

(Check out our paper for more results)

Table1: application of GAT

Feature extractor	Aggregation	# Params	EER (%)
SE-ResNet	SAP	6.0M	1.98
SE-ResNet	GAT	5.4M	**1.86**
RawNet2	GRU	13.2M	2.48
RawNet2	GAT	9.9M	**2.23**

Table2: performance comparison with state-of-the-art systems

	Input Feature	Front-end	Aggregation	EER (%)
Chung *et al.*	Spec-257	Thin ResNet-34	SAP	2.21
Yu *et al.*	Spec-512	ResNet-50	TAP	2.94
Liu *et al.*	MFB-40	Dense-Residual	ABP	2.54
Jung *et al.*	MFB-40	Fast ResNet-34	LDE	1.98
Kye *et al.*	MFB-40	Fast ResNet-34	CAP	1.88
Ours	MFB-40	SE-ResNet	SAP	1.98
Ours(Proposed)	MFB-40	SE-ResNet	GAT	**1.75**
Lin *et al.*	Raw waveform	wav2spk	Gating + SP	3.00
Zhu *et al.*	Raw waveform	Y-vector	SP	2.60
Jung *et al.*	Raw waveform	RawNet2	GRU	2.48
Ours(Proposed)	Raw waveform	RawNet2	GAT	**2.15**

1) K. Okabe, T. Koshinaka and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *in Proc. Interspeech*, 2018.
2) B. Desplanques, J. Thienpondt and K. Demuynck, "Ecapa- tdnn: Emphasized channel attention, propagation and aggrega- tion in tdnn based speaker verification," *in Proc. Interspeech*, 2020
3) P.Velic˘ković ,G.Cucurull,A.Casanovaetal.,"Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.
4) A.Nagrani,J.S.ChungandA.Zisserman,"Voxceleb:a large- scale speaker identification dataset," in Proc. Interspeech, 2017.
5) J. S. Chung, A. Nagrani and A. Zisserman, "Voxceleb2: Deep speaker recognition," in Proc. Interspeech, 2018.