Towards Efficient Speaker Representation Learning

APSIPA

Distinguished Lecture

2025-2026

Man-Wai MAK

Dept. of Electrical and Electronic Engineering,

The Hong Kong Polytechnic University

http://www.eie.polyu.edu.hk/~mwmak

enmwmak@polyu.edu.hk

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

APSIPA

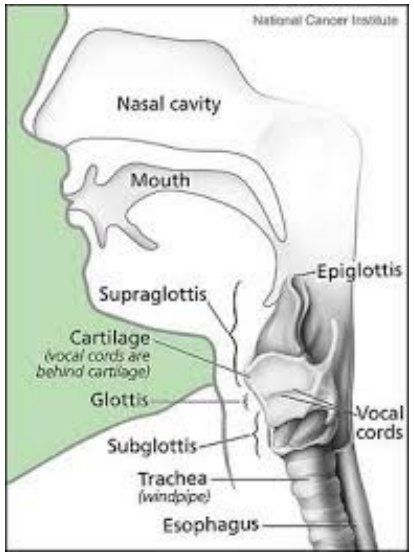
Microsoft

MSRA

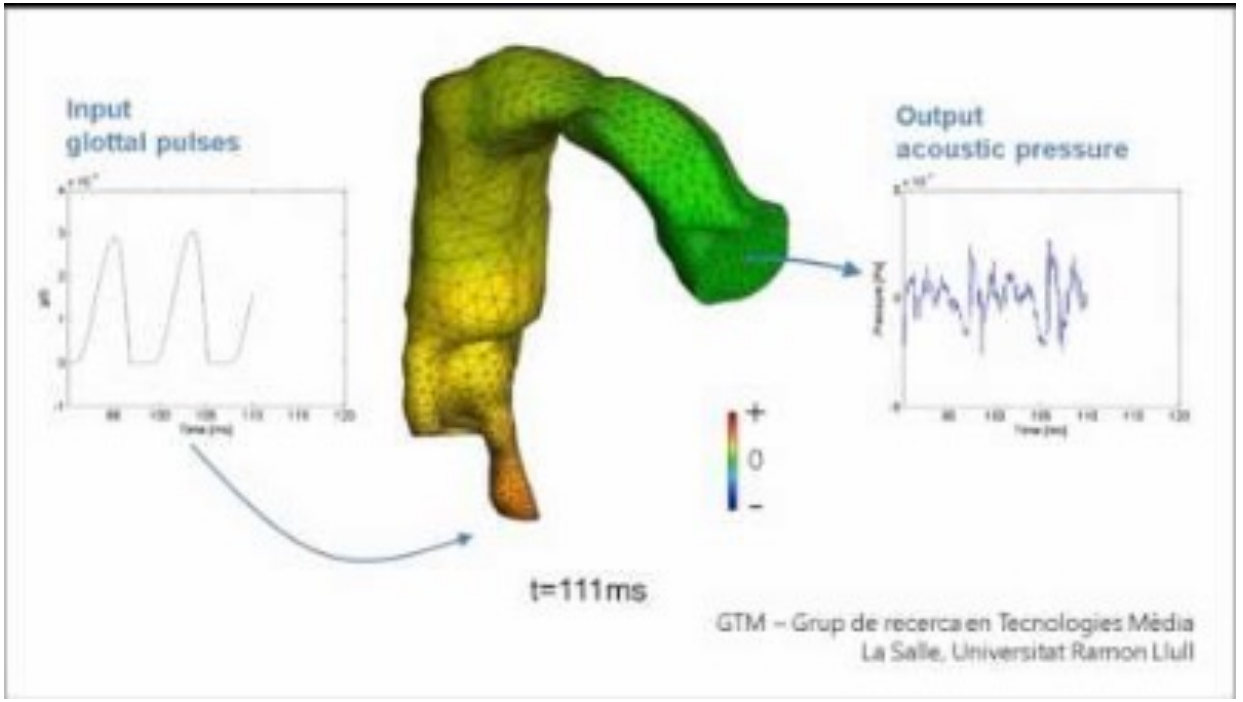
Contents

1. Speaker Representation
2. ConFusionformer: Locality-Enhanced Conformer
3. Parameter-Efficient Fine-Tuning of Pre-Trained Models
   - Speaker Prompt Tuning
   - Spectral-Aware LoRA

Voice of Individuals are Not Alike

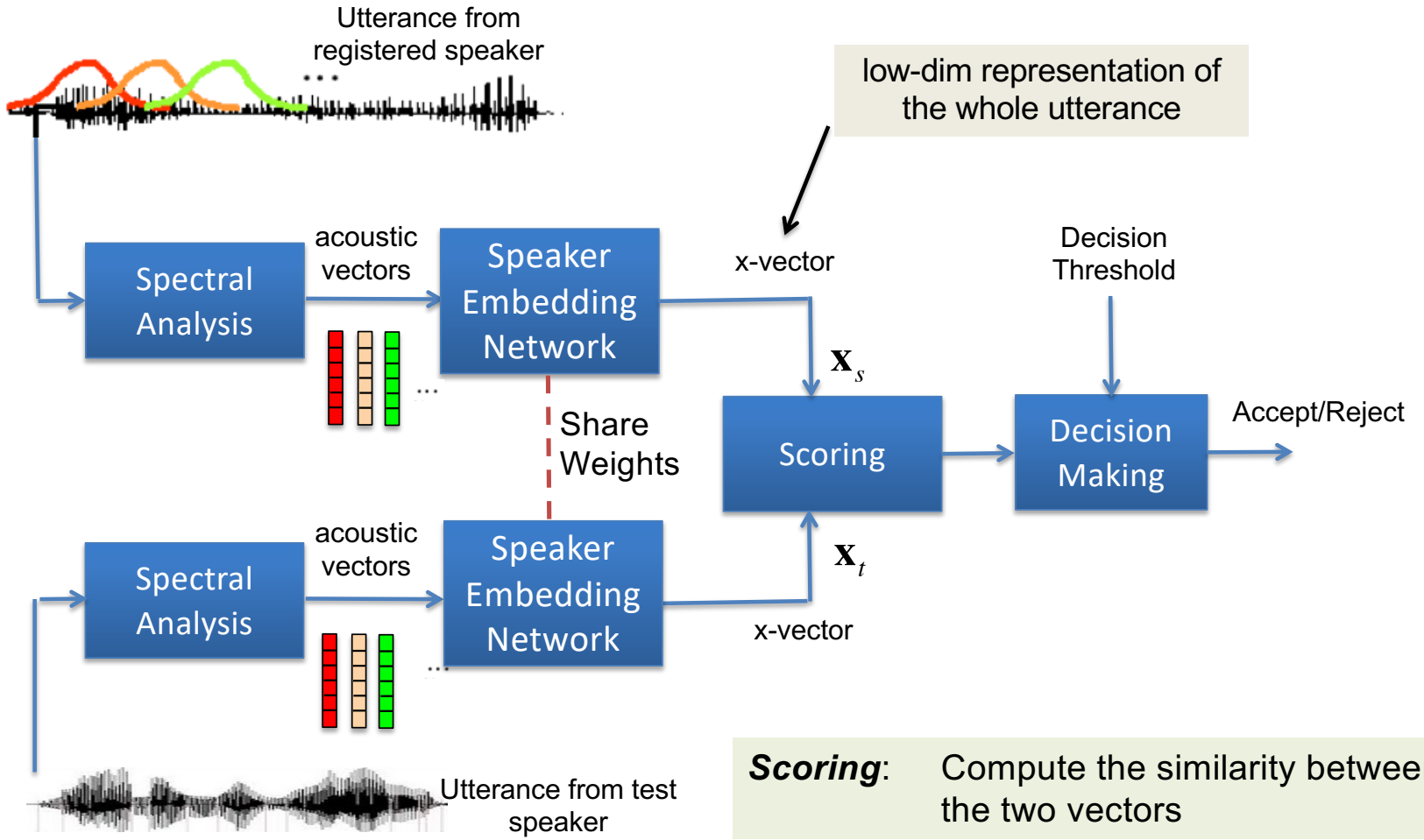


- We produce speech by moving our articulator
- But the vocal cords, vocal tract, and nasal cavity of **individuals are different**



https://www.youtube.com/watch?v=toufbFFz7Zw

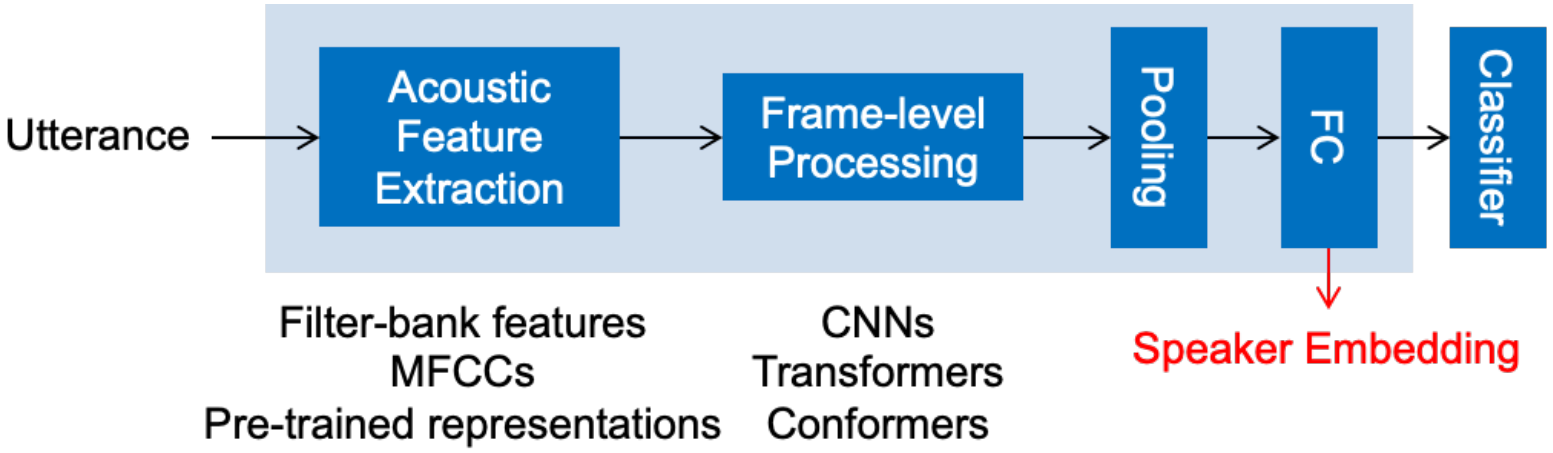
Processes of Speaker Verification



Scoring: Compute the similarity between the two vectors

Speaker Representation Learning

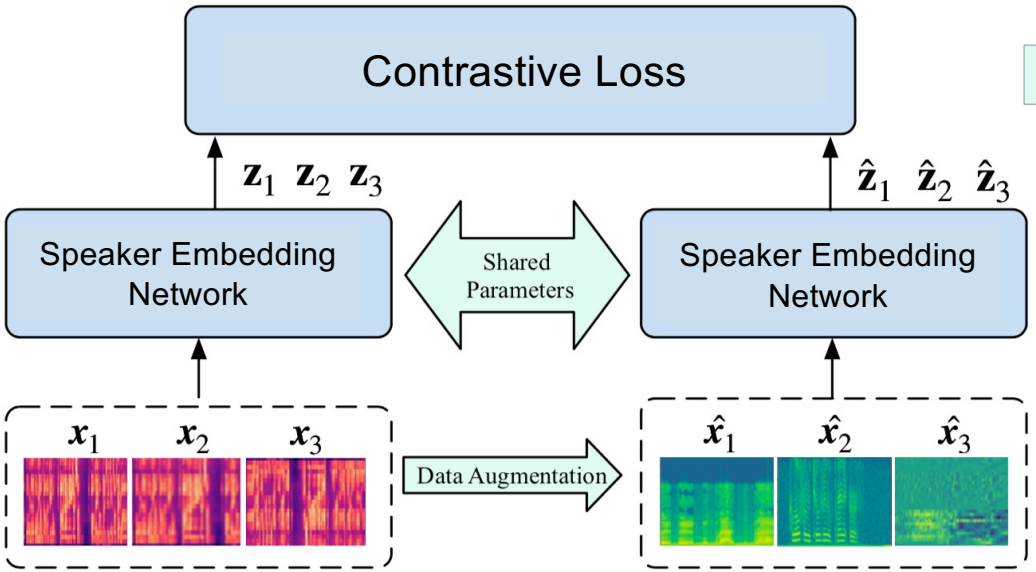
- Speaker embeddings are indispensable to a variety of speech tasks, e.g., speaker verification, speaker diarization, target speaker. extraction, speech synthesis, speaker-aware speech recognition, etc.
- Speaker embeddings are traditionally obtained by training a speaker identification network using acoustic vectors (MFCCs or filterbank) as input.

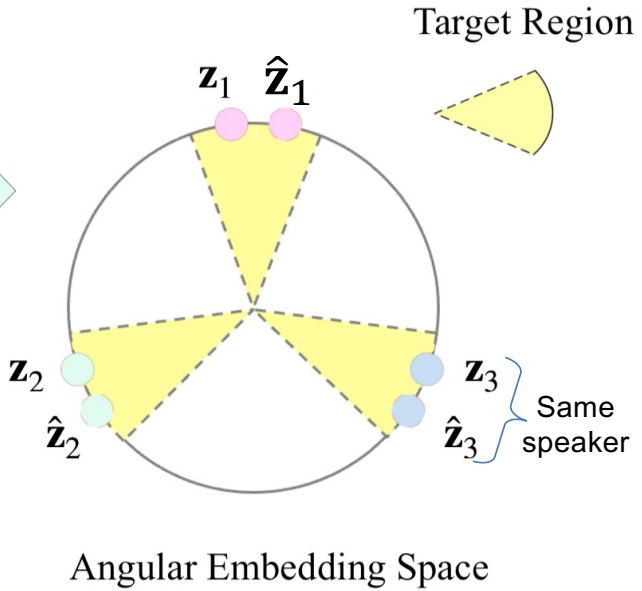


Speaker Representation Learning

We can also use supervised contrastive loss to train a speaker embedding network to find a **speaker representation space** in which vectors (embeddings) of the **same speaker** are **close** and those of **different speakers** are **far apart**

$$L_{SupCon} = \sum_{i=1}^{N} \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau)}{\sum_{a \in \mathcal{A}(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau)}$$



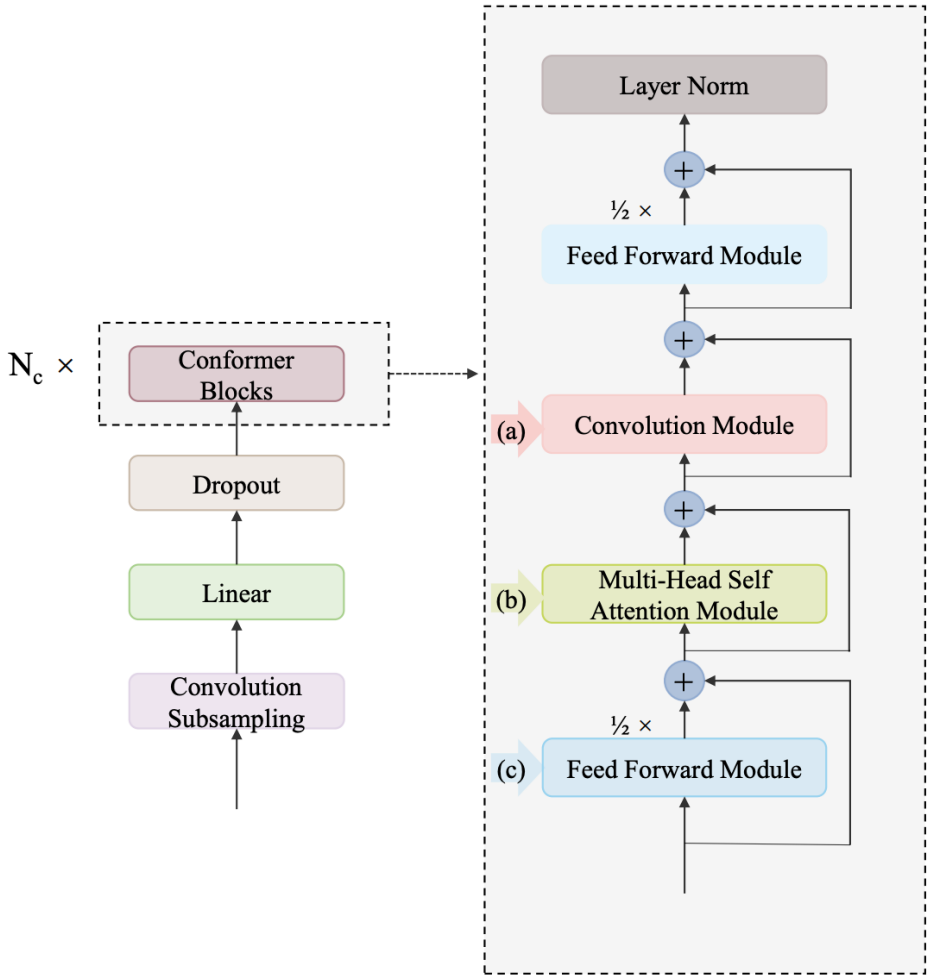


Zhe Li, Man-Wai Mak, Mert Pilanci, and Helen Meng, "Mutual Information-Enhanced Contrastive Learning with Margin for Maximal Speaker Separability", *IEEE/ACM Trans on Audio, Speech and Language Processing*, June 2025.

ConFusionformer

Youzhi Tu, Man-Wai Mak, Kong-Aik Lee, and Weiwei Lin, "ConFusionformer: Locality-enhanced Conformer Through Multi-resolution Attention Fusion for Speaker Verification", *Neurocomputing*, May 2025.

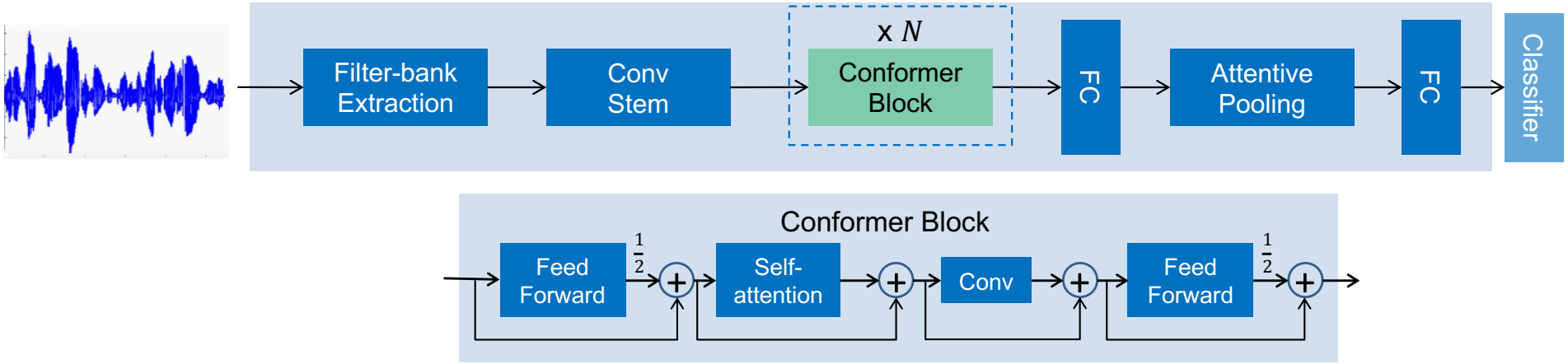
Conformers



Gulati et al. “Conformer: Convolution-augmented Transformer for Speech Recognition,” Interspeech, 2020.

Conformers for Speaker Embedding

- Conformers use self-attention to capture the **global dependencies** in speech sequences and use CNNs for **local information modeling**.
- There is still a performance gap between the Conformer-based embeddings and the CNN-based embeddings.



- Gulati, A., et al., 2020. Conformer: Convolution-augmented transformer for speech recognition, *Interspeech*, pp. 5036–5040.
- Zhang, Y., et al., 2022. MFA-Conformer: Multi-scale feature aggregation conformer for automatic speaker verification, *Interspeech*, pp. 306–310.

Motivations of ConFusionformer

- Because the **speaker characteristics** are often reflected in **local speech dynamics**, locality modeling is of vital importance in speaker embedding.
- Although the first few attention layers of a Transformer possess implicit convolutional locality, **using vanilla Transformer** layers for SV does not lead to satisfactory results in practice.
- In computer vision (CV), it is acknowledged that Transformers perform better than CNNs even without large-scale pre-training.
- The success of Transformers in NLP and CV motivates us to develop **Transformer-based speaker embedding networks** that can compete with or surpass the state-of-the-art CNN counterparts.

- R. Wang, J. Ao, L. Zhou, S. Liu, Z. Wei, T. Ko, Q. Li, Y. Zhang, Multi-view self-attention based transformer for speaker recognition, ICASSP, 2022, pp. 6732–6736.
- B. Han, Z. Chen, Y. Qian, Local information modeling with self-attention for speaker verification, *ICASSP*, 2022, pp. 6727–6731.

Motivations of ConFusionformer

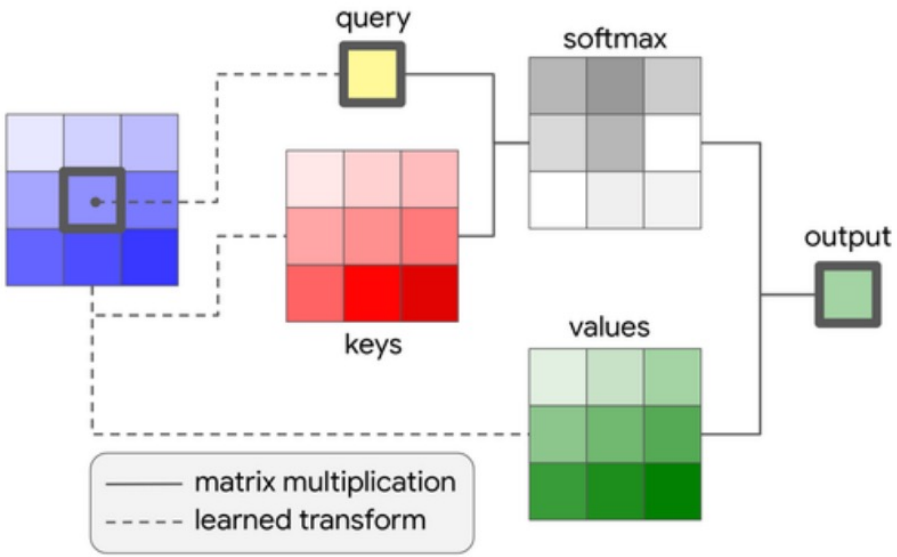
- The Conformer block is not computationally efficient
  - Given an input sequence of length T and a standard Conformer block with an encoding dimension of D, the self-attention network (SAN) consumes $\mathcal{O}(4TD^2 + T^2D)$ MAC, whereas the MAC of the feed-forward networks (FFNs) is $\mathcal{O}(16TD^2)$.
  - For a test utterance of 8s and embedding dim of 256, we have $T = 400$ and $D = 256$

    SAN: $4TD^2 + T^2D = 1.45\times10^8$ FFN: $16TD^2 = 4.19\times10^8$
  - FFNs consume a larger proportion of the computation.
- The modeling of local information is critical to learning speaker characteristics. However, using CNNs for local information modeling is not sufficient for a Conformer.
- Solution – ConFusionformer:
  - Propose an efficient architecture for the Conformer block.
  - Propose multi-resolution attention fusion for enhanced locality.

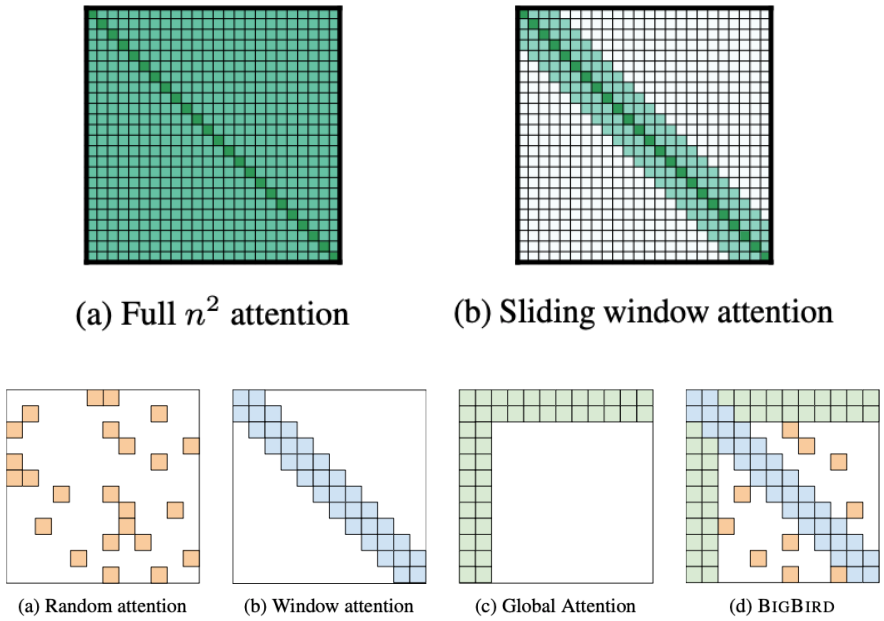
Han, B., et al, 2022. Local information modeling with self-attention for speaker verification, in: Proc. International Conference on Acoustics, Speech, and Signal Processing, pp. 6727–6731.

Previous Work on Locality Attention

$$y_{ij} = \sum_{a,b \in \mathcal{N}_k(i,j)} \mathtt{softmax}_{ab}\left(q_{ij}^{\top} k_{ab}\right) v_{ab}$$



Local Attention Layer (Ramachandran et al. 2019)

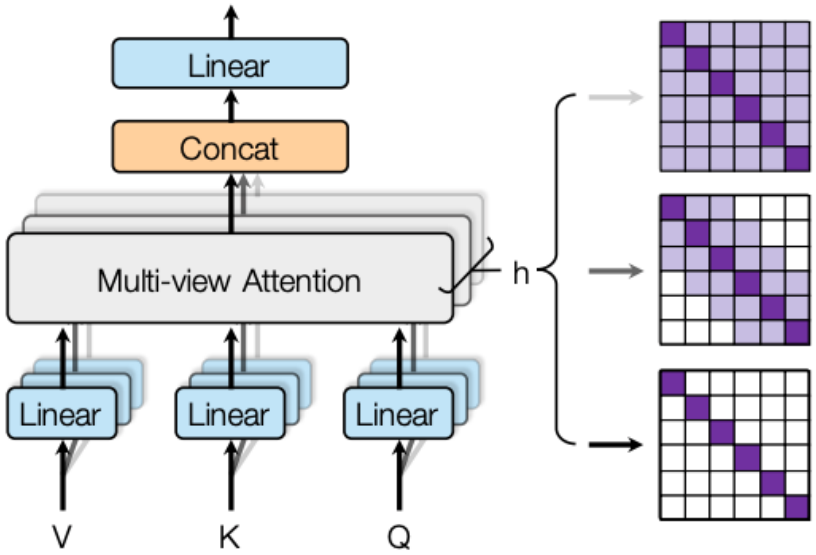


Sliding Window Attention (Beltagy et al. 2020; Zaheer et al. 2020)

Previous Work on Locality Attention

Head-wise mask matrix

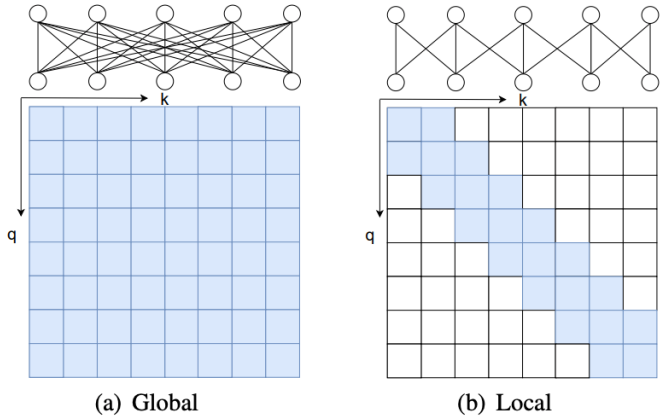
$$\text{Attention}(Q, K, V) = M \odot \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



Multi-view Attention (Wang et al. 2022)

$$o_i = \sum_{j \in T} Softmax_j(-wd_{ij}^2 + q_i k_j)v_j$$

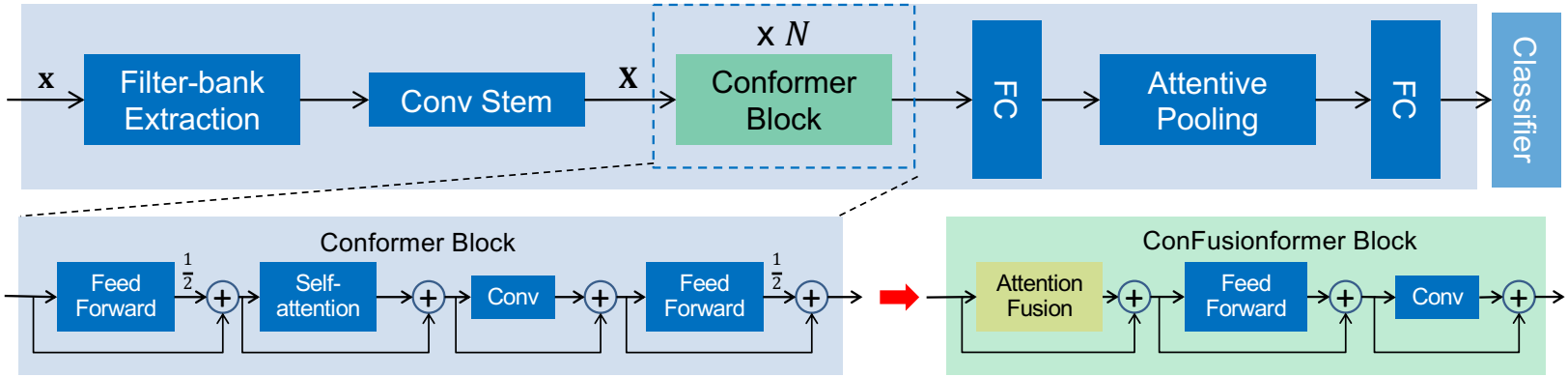
d_{ij} = distance between frames i and j



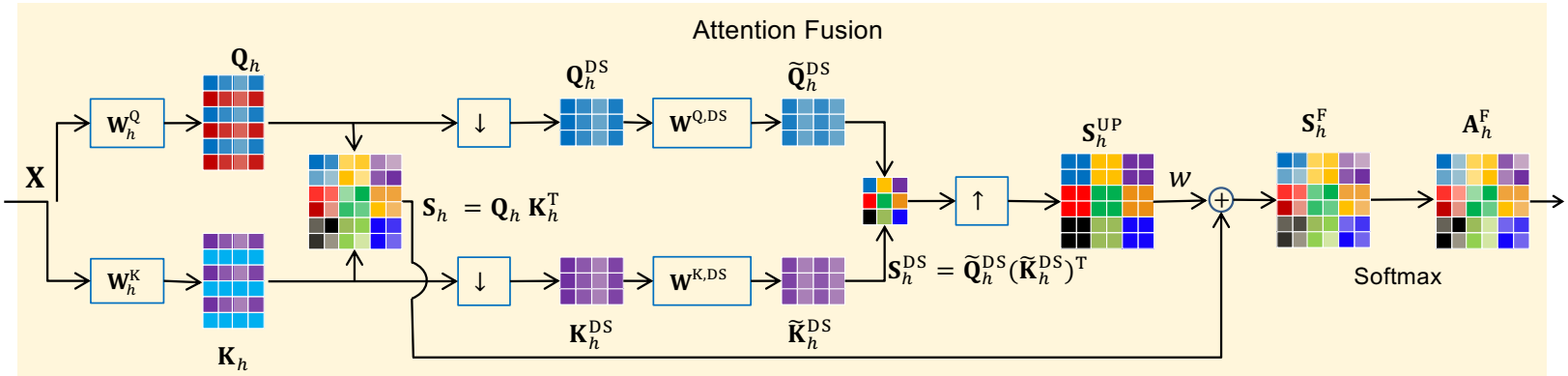
Gaussian Self-Attention (Han et al. 2022)

ConFusionformer

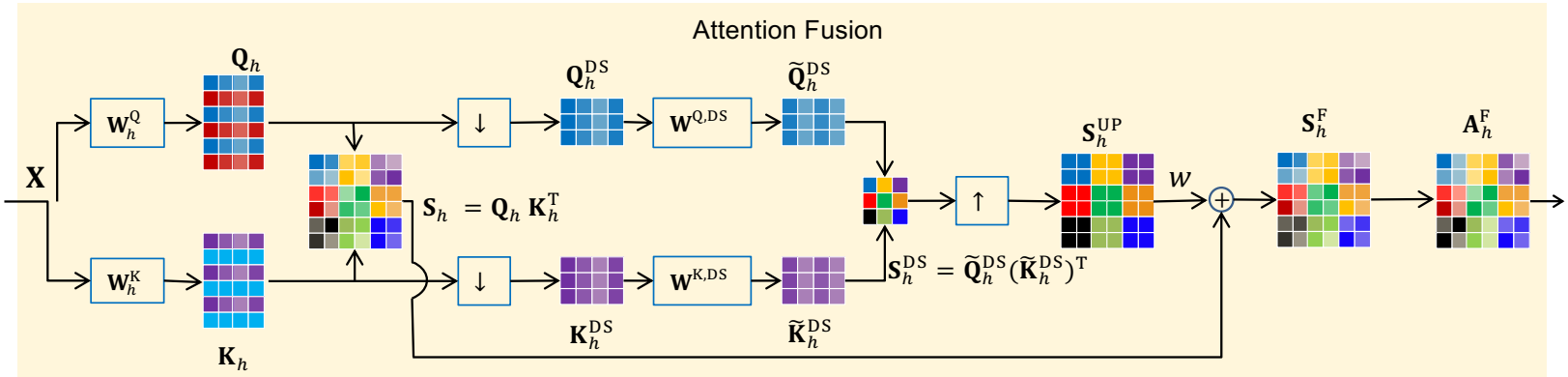
Efficient ConFusionformer blocks with smaller model size



Enhanced locality modeling through multi-resolution **attention fusion**



Multi-resolution Attention Fusion



Low-resolution attention score map creation

$\mathbf{Q}_h = \mathbf{X}\mathbf{W}_h^{\mathrm{Q}} \in \mathbb{R}^{T\times D_{\mathrm{H}}}$

$\mathbf{Q}_h^{\mathrm{DS}}[m,:] = \mathbf{Q}_h\,[r(m-1)+1,:] \in \mathbb{R}^{T_{\mathrm{DS}}\times D_{\mathrm{H}}}$

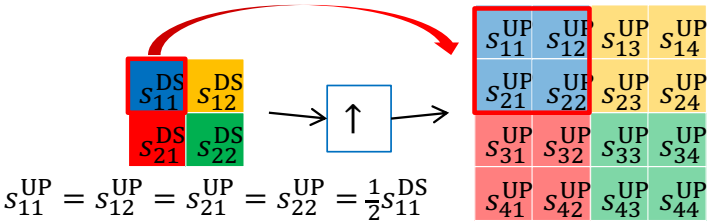
$\mathbf{S}_h = \mathbf{Q}_h\,\mathbf{K}_h^{\mathrm{T}} \in \mathbb{R}^{T\times T}$

$\widetilde{\mathbf{Q}}_h^{\mathrm{DS}} = \mathbf{Q}_h^{\mathrm{DS}}\mathbf{W}^{\mathrm{Q,DS}}$

$\mathbf{S}_h^{\mathrm{DS}}$ is a decimated version of $\mathbf{S}_h$.

$\mathbf{S}_h^{\mathrm{DS}} = \widetilde{\mathbf{Q}}_h^{\mathrm{DS}}(\widetilde{\mathbf{K}}_h^{\mathrm{DS}})^{\mathrm{T}} \in \mathbb{R}^{T_{\mathrm{DS}}\times T_{\mathrm{DS}}}$

Full-resolution attention score map restoration



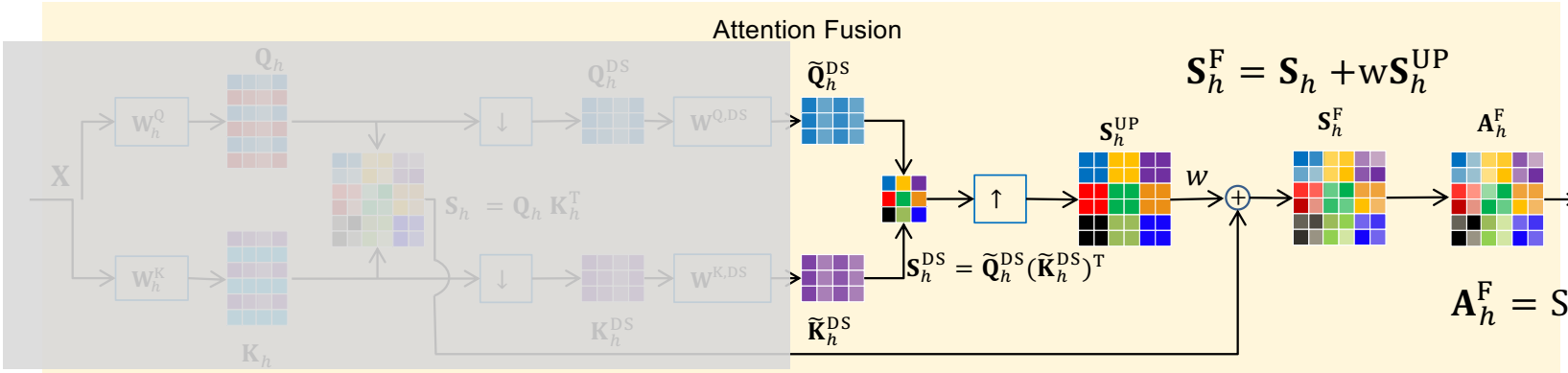
$s_{11}^{\mathrm{UP}} = s_{12}^{\mathrm{UP}} = s_{21}^{\mathrm{UP}} = s_{22}^{\mathrm{UP}} = \frac{1}{2}s_{11}^{\mathrm{DS}}$

Attention score map fusion

$\mathbf{S}_h^{\mathrm{F}} = \mathbf{S}_h + \mathrm{w}\mathbf{S}_h^{\mathrm{UP}} \quad \mathbf{A}_h^{\mathrm{F}} = \mathrm{Softmax}((\mathbf{S}_h^{\mathrm{F}} + \mathbf{B})/\sqrt{D_{\mathrm{H}}})$

T: sequence length
D_{H}: dimension of each head
r: up and downsampling rate
$T_{\mathrm{DS}} = \lceil T/r \rceil$: downsampled sequence length
h: index of head
m: index of row vectors of $\mathbf{Q}_h^{\mathrm{DS}}$
$\mathbf{B}$: relative positional attention score map, $b_{h,ij} = \boldsymbol{q}_h\,(\boldsymbol{p}_{\delta(i,j)}\mathbf{W}^{\mathrm{P}})^{\mathrm{T}}$

Locality Enhancement

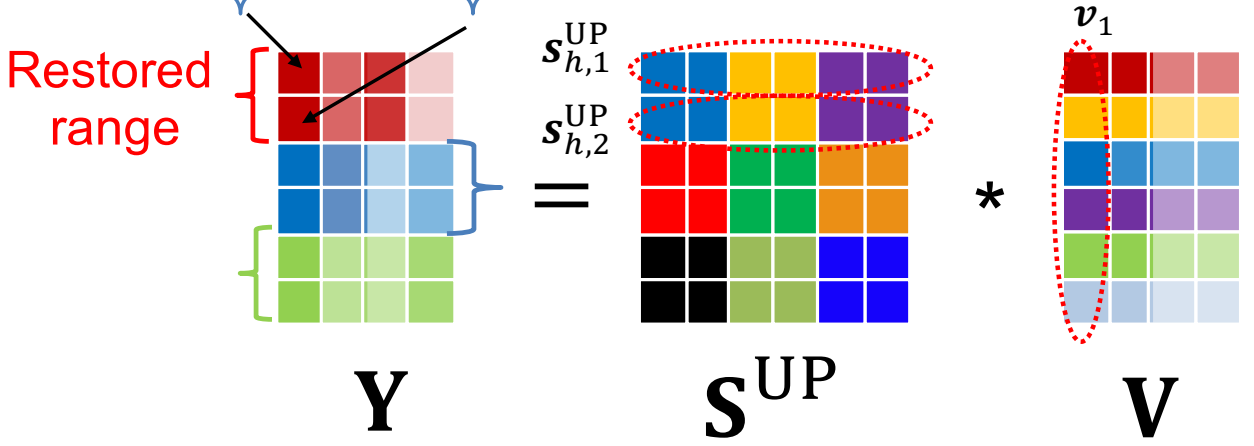


$\mathbf{Y}$ can be seen as the 1D average pooling of $\mathbf{V}$ along the time axis, weighted by $\mathbf{S}^{UP}$.

The information of the output sequence $\boldsymbol{y}$'s is local to each restored range along the temporal direction.

temporal direction

$$\left(\boldsymbol{s}_{h,1}^{\mathrm{UP}}\right) \cdot \boldsymbol{v}_1 = \left(\boldsymbol{s}_{h,2}^{\mathrm{UP}}\right) \cdot \boldsymbol{v}_1 \quad \because \boldsymbol{s}_{h,1}^{\mathrm{UP}} = \boldsymbol{s}_{h,2}^{\mathrm{UP}}$$



Experimental Setup

- Data sets

Task	Embedding training	No. of trials
VoxCeleb1	Vox2-dev, 1 million utterances from 5,994 speakers	Vox1-O: 37,611 Vox1-E: 579,818 Vox1-H: 550,894
CNCeleb1	CNCeleb2-dev, 560K utterances from 2,787speakers	3,484,292

- Acoustic feature extraction
  - 80-D filter-bank features
  - Augmentation: noise, speech, music, reverberation, speed perturbation
- Network parameters
  - Conv stem: Conv2d-Conv2d(temporal stride=2)-Conv2d
  - Conformer/ConFusionformer encoding dimension $D = 256$
  - No. of ConFusionformer blocks $N = 9$ to 12
  - Downsampling rate $\mathrm{r} = 2$
  - No. of relative positional encoding: 127

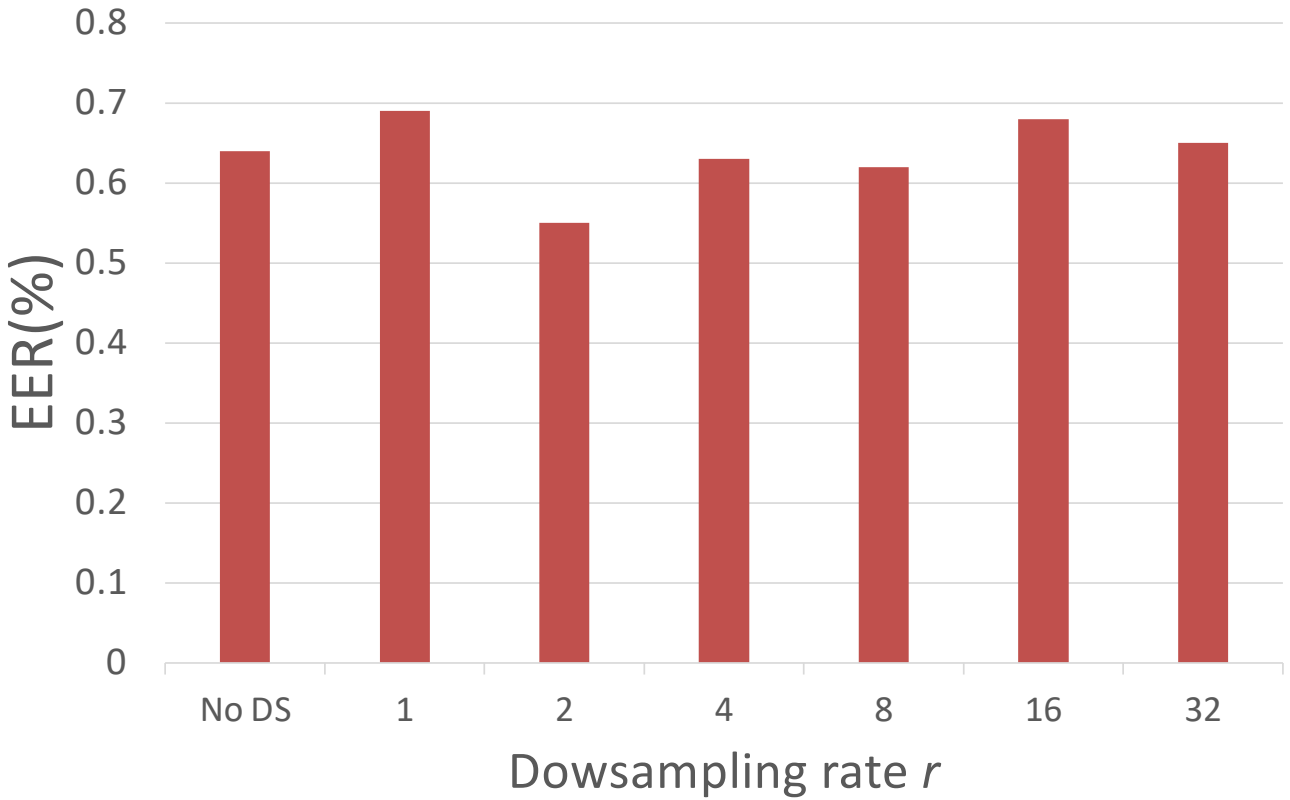
The Hong Kong Polytechnic University
香港理工大學

SV Performance

Model	# Para (M)	# Flops (G)	Vox1-O		Vox1-E		Vox1-H		CNCeleb1	
			EER (%)	minDCF	EER (%)	minDCF	EER (%)	minDCF	EER (%)	minDCF
ECAPA-TDNN	14.7	2.57	0.77	0.075	0.96	0.104	1.79	0.177	7.61	0.411
ResNet-50	11.2	9.14	0.85	0.089	0.98	0.105	1.71	0.167	7.10	0.409
ResNet-101	16.0	17.7	0.63	0.063	0.79	**0.087**	1.40	**0.134**	**6.75**	0.410
Conformer-6	11.0	2.50	0.64	0.084	0.91	0.107	1.67	0.170	7.38	0.409
Conformer-8	14.1	3.04	0.67	0.063	0.82	0.096	1.60	0.163	7.68	0.417
ConFusionformer-9	10.9	2.45	0.68	0.064	0.93	0.104	1.66	0.166	7.36	0.402
ConFusionformer-12	13.9	2.97	**0.55**	**0.050**	**0.78**	**0.087**	**1.38**	0.143	6.76	**0.379**

ConFusionformer achieves better performance under similar number of parameters and similar computation.

Downsampling Rate *r*



Nonlocality

Nonlocality measure:

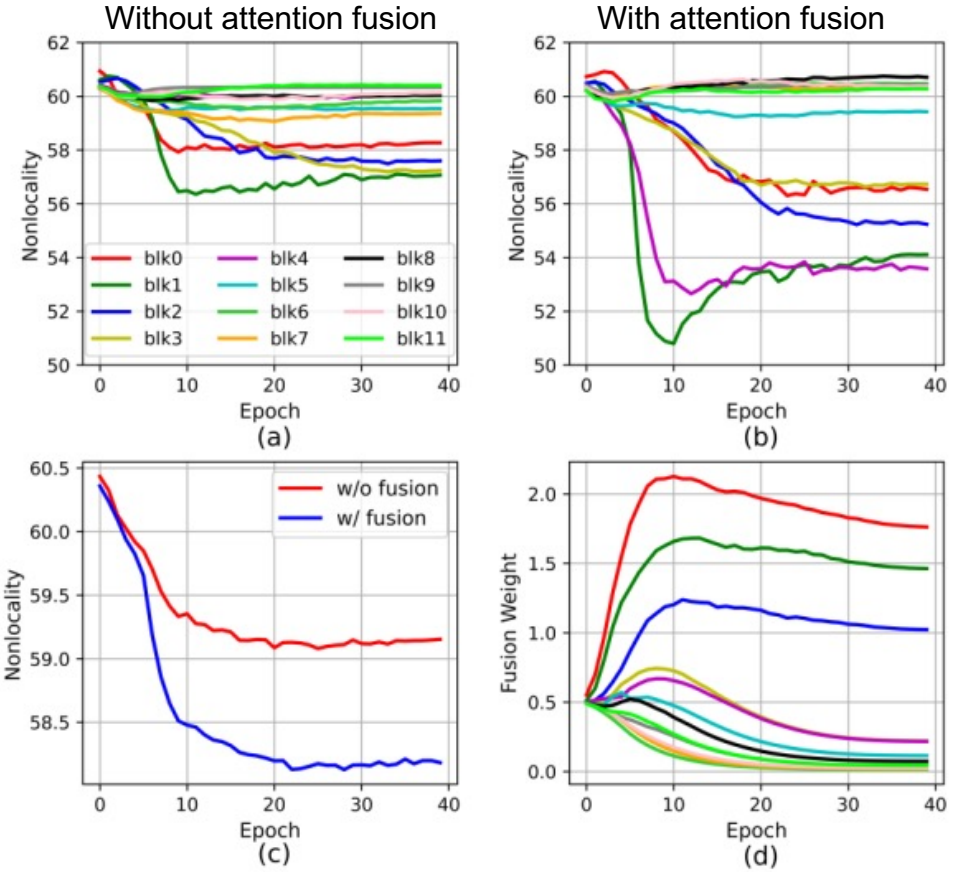
$$d_n^{\mathrm{NL}} := \frac{1}{T}\sum_{h,i,j} a_{h,ij,n}|i-j|$$

$$d^{\mathrm{NL}} = \frac{1}{N}\sum_{n} d_n^{\mathrm{NL}}$$

T: sequence length
N: No. of blocks
n: index of blocks
i, j: index of frames
$a_{h,ij,n}$: attention coefficient of Head h of Block n

The nonlocality d^{NL} indicates the average attention distance between the query frame and the key frame.

The further the query frame attends to, the larger the nonlocality, and therefore the lower the locality.



w in $\mathbf{S}_h^{\mathrm{F}} = \mathbf{S}_h + w\mathbf{S}_h^{\mathrm{UP}}$

ConFusionformers have better locality than Conformers.

d'Ascoli, S., Touvron, H., Leavitt, M., Morcos, A., Biroli, G., Sagun, L., 2021. ConViT: Improving vision transformers with soft convolutional inductive biases, in: Proc. International Conference on Machine Learning, pp. 2286– 2296.

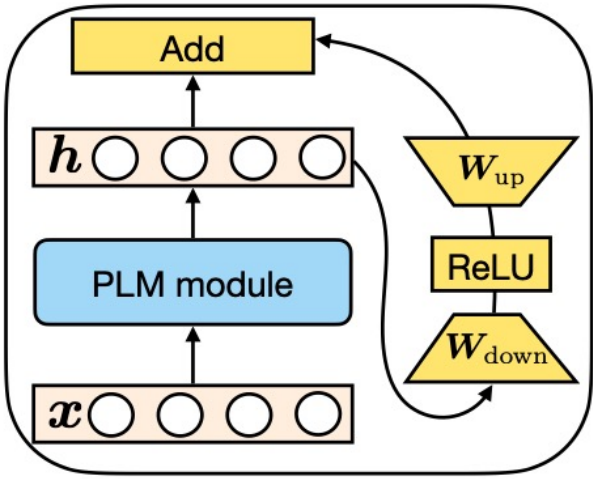
Advantages of ConFusionformer

- The ConFusionformer obtained **better computational efficiency** with fewer model parameters using modified Conformer blocks.
- Under similar computations, ConFusionformers achieved **superior performance** to Conformers on VoxCeleb1 and CNCeleb1.

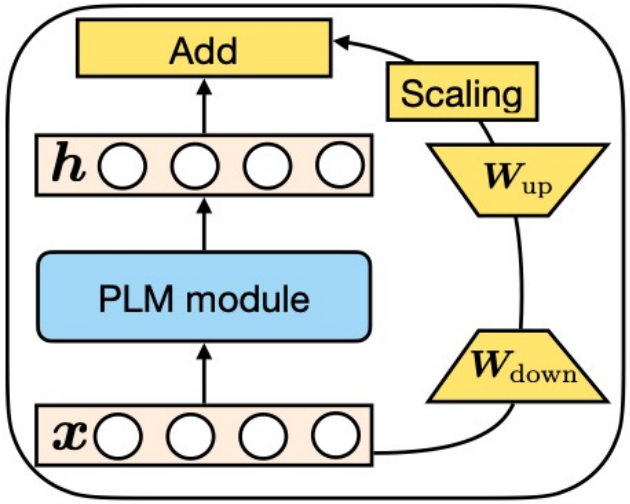
Parameter-Efficient Fine-Tuning

Fine-Tuning of Pre-Trained Models

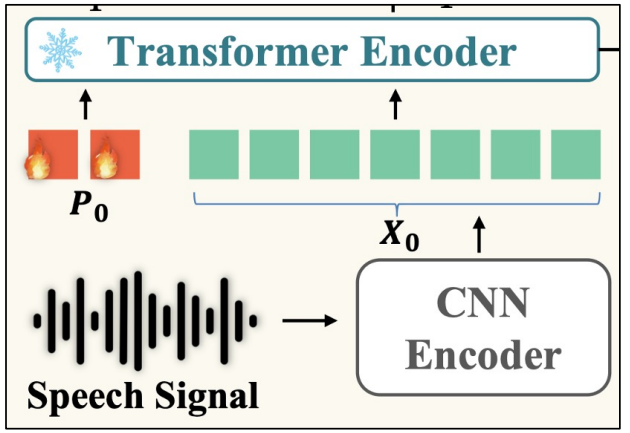
Fine-tuning a pre-trained Transformer model (PTM) for speech applications in a parameter-efficient manner offers the dual benefits of **reducing memory** and **leveraging the rich feature representations** in massive unlabeled datasets.



Adapter



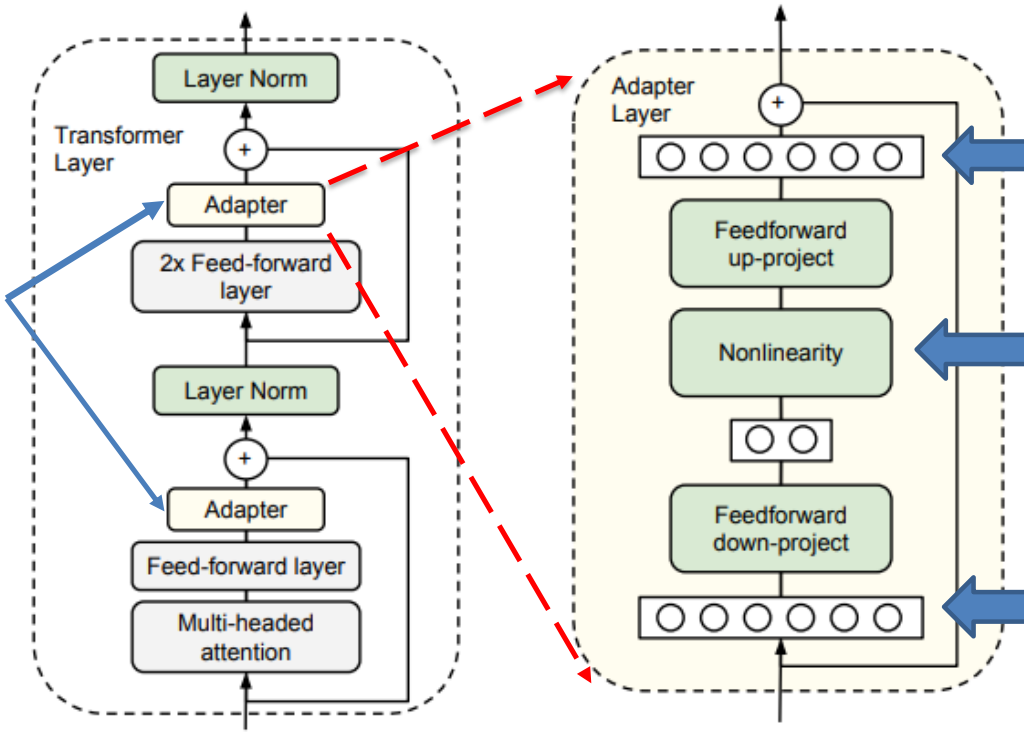
Low-Rank Adaptation (LoRA)



Prompt Tuning

Adapter

Inserted serial adapter after each of the two sub-layers in the Transformer layer



Project back to d dimensions

Apply nonlinearity

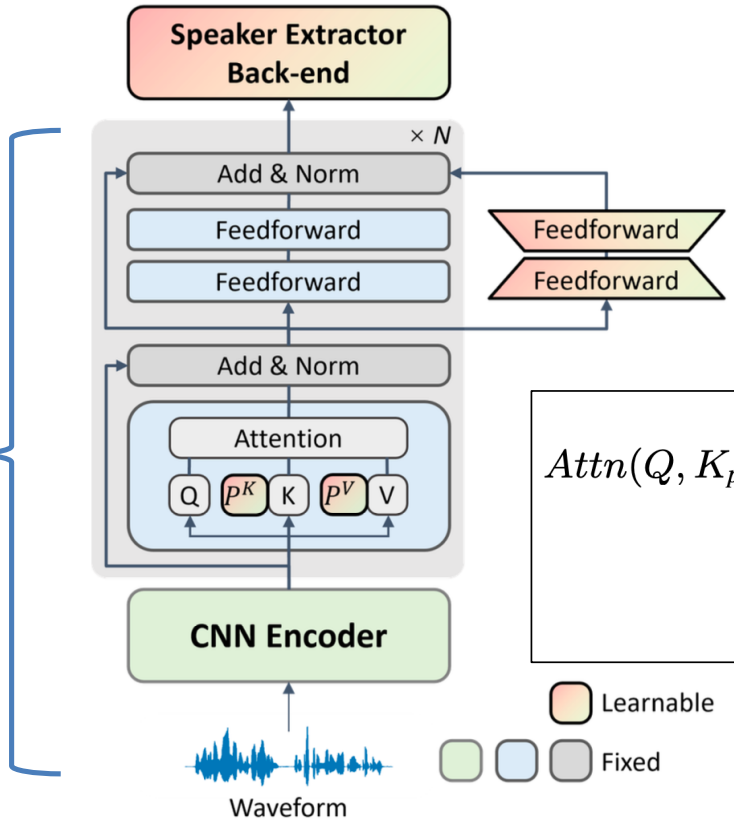
Adapters project original d-dimensional features into smaller dimension m.

Bottleneck architecture to limit the number of parameters

Adapter for SV

Downstream model:
Speaker extractor model

Self-supervised
Speech model



Mix-And-Match Adapter:
Prefix-tuning + Adapter

Prefix tuning

$$Attn(Q, K_{prefix}, V_{prefix}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}_{prefix}^{T}}{\sqrt{D_{proj}}}\right)\mathbf{V}_{prefix}$$

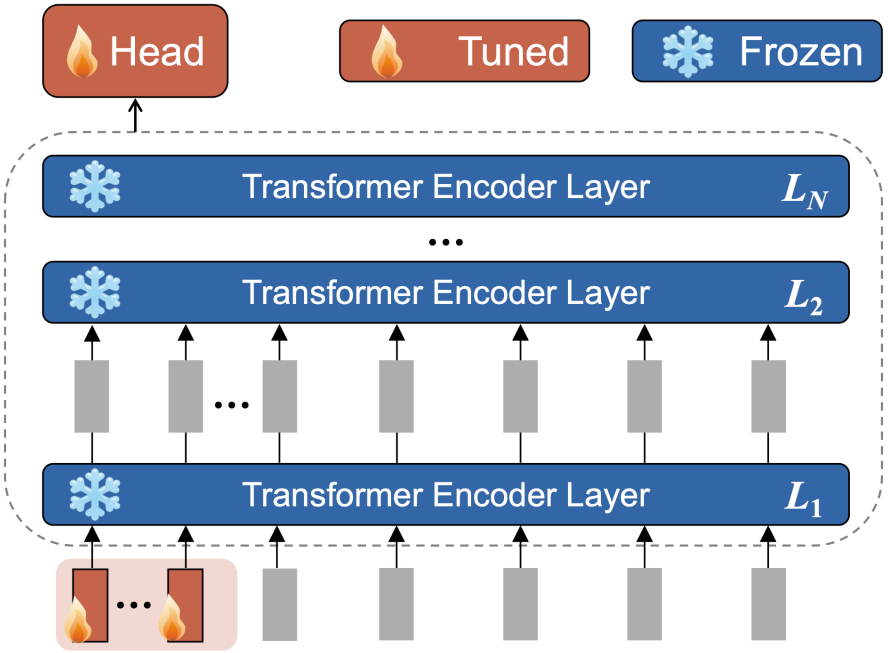
$$\mathbf{K}_{prefix} = concat(\mathbf{P}_K, \mathbf{W}_K\mathbf{H}_{in})$$

$$\mathbf{V}_{prefix} = concat(\mathbf{P}_V, \mathbf{W}_V\mathbf{H}_{in})$$

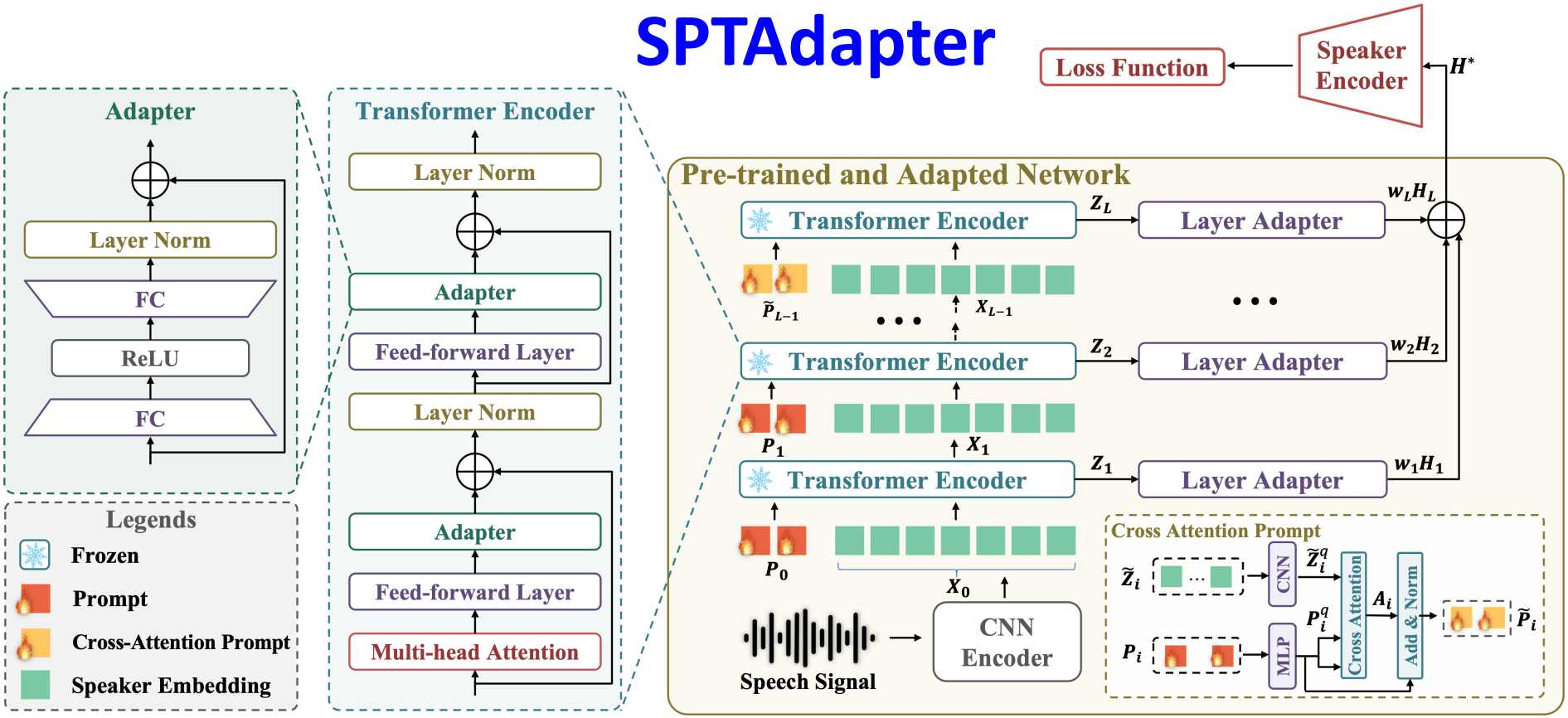
Peng, Junyi, et al. "Parameter-efficient transfer learning of pre-trained Transformer models for speaker verification using adapters." *ICASSP 2023*

Soft Prompt Tuning

- **Continuous (Soft) Prompts**
  - Soft prompts do not need to be in natural language that humans can understand
  - Special tokens (or virtual tokens) are created for the prompt to optimize in **continuous space.**



SPTAdapter



SPTAdapter combines **Adapters** and **Speaker Prompt Tuning**

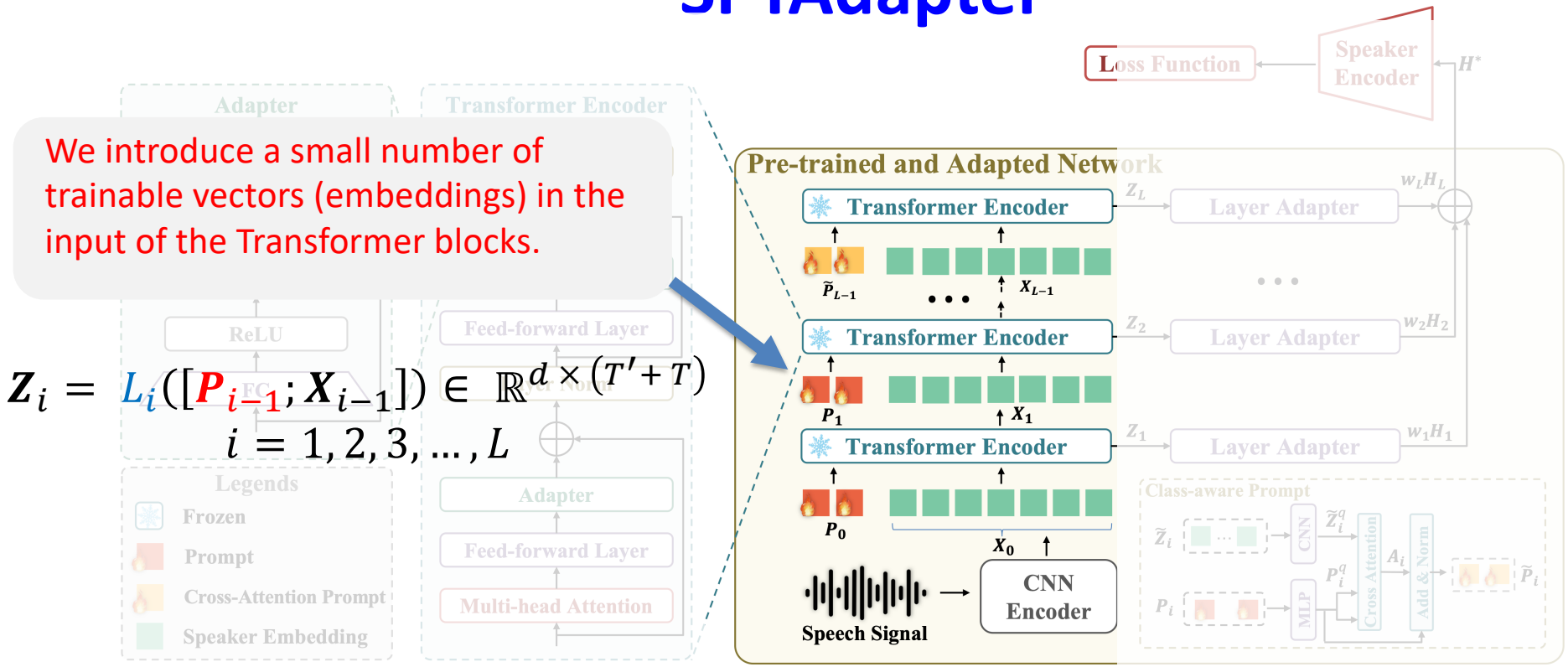
The **Cross-Attention Prompts** enhance speaker features' relevance

Li Zhe, et al. Dual Parameter-Efficient Fine-Tuning for Speaker Representation via Speaker Prompt Tuning and Adapters, in Proc. *ICASSP 2024*

SPTAdapter

We introduce a small number of trainable vectors (embeddings) in the input of the Transformer blocks.

$$\boldsymbol{Z}_i = L_i([\boldsymbol{P}_{i-1}; \boldsymbol{X}_{i-1}]) \in \mathbb{R}^{d \times (T'+T)}$$
$$i = 1, 2, 3, \ldots, L$$



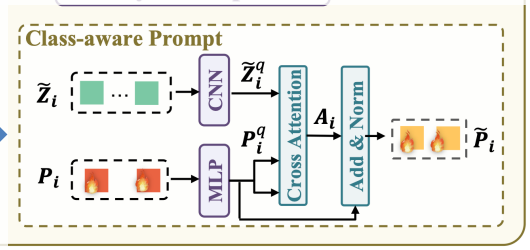
SPTAdapter

$$\boldsymbol{A}_i = \boldsymbol{W}_v \boldsymbol{P}_i^q \left[\text{Softmax}\left(\frac{\tilde{\boldsymbol{Z}}_i^{q\top} \boldsymbol{W}_q^\top \boldsymbol{W}_k \boldsymbol{P}_i^q}{\sqrt{d}} \right) \right] \in \mathbb{R}^{d \times T'}$$

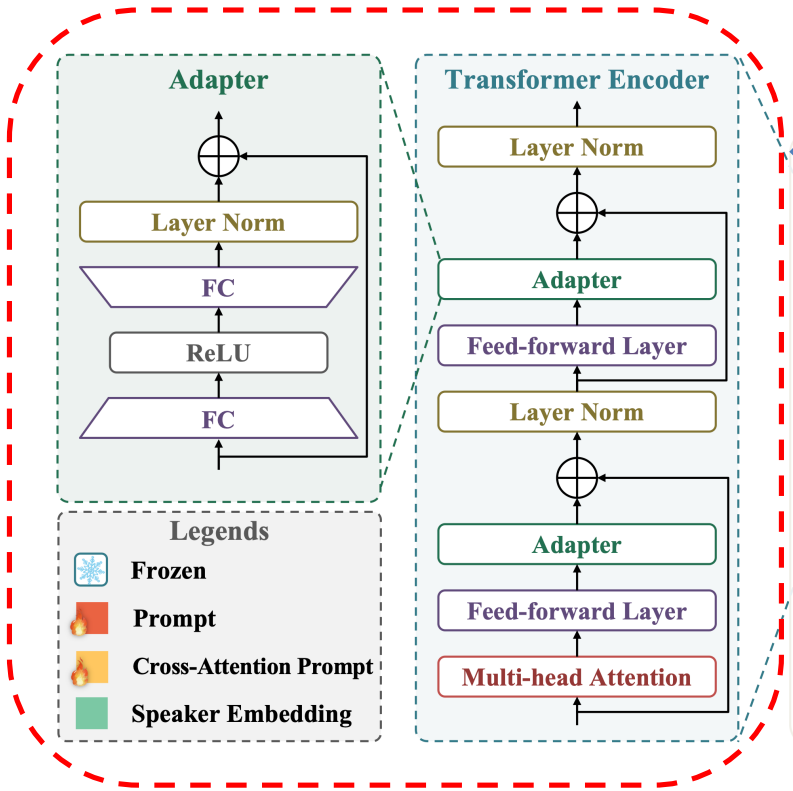
$$\tilde{\boldsymbol{P}}_i = \text{LN}\left(\boldsymbol{A}_i + \boldsymbol{P}_i^q\right) \in \mathbb{R}^{d \times T'}$$

$$\boldsymbol{P}_i \leftarrow \tilde{\boldsymbol{P}}_i$$

Cross-attention prompt tuning further enriches the speaker information in the prompts



SPTAdapter



insert adapter modules into the Transformer encoders

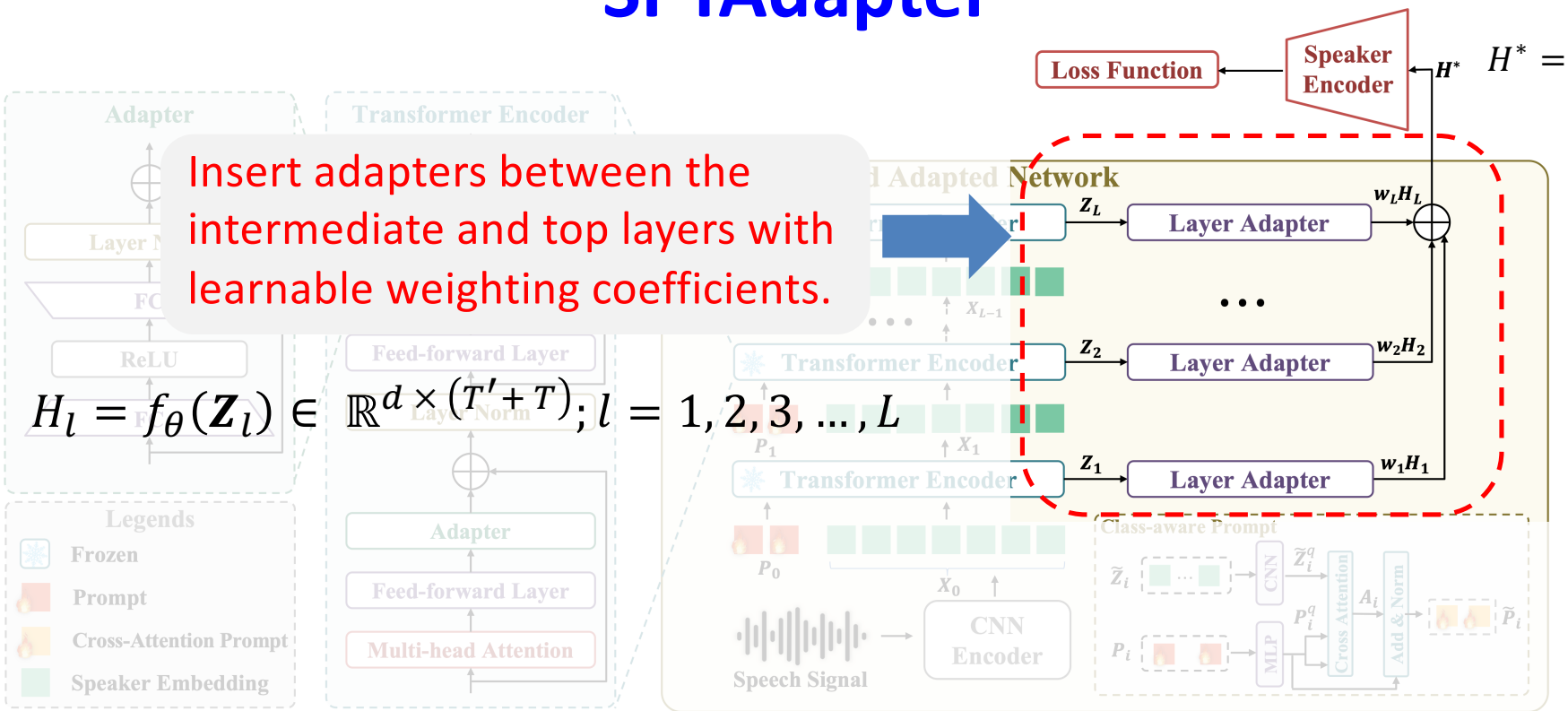
Li Zhe, et al. Dual Parameter-Efficient Fine-Tuning for Speaker Representation via Speaker Prompt Tuning and Adapters, in Proc. *ICASSP 2024*

SPTAdapter

Insert adapters between the intermediate and top layers with learnable weighting coefficients.

$$H_l = f_\theta(\boldsymbol{Z}_l) \in \mathbb{R}^{d \times (T'+T)}; l = 1, 2, 3, \ldots, L$$

$$H^* = \sum_{l=1}^{L} w_l H_l$$



Li Zhe, et al. Dual Parameter-Efficient Fine-Tuning for Speaker Representation via Speaker Prompt Tuning and Adapters, in Proc. *ICASSP 2024*

Comparing Fine-Tuning Methods

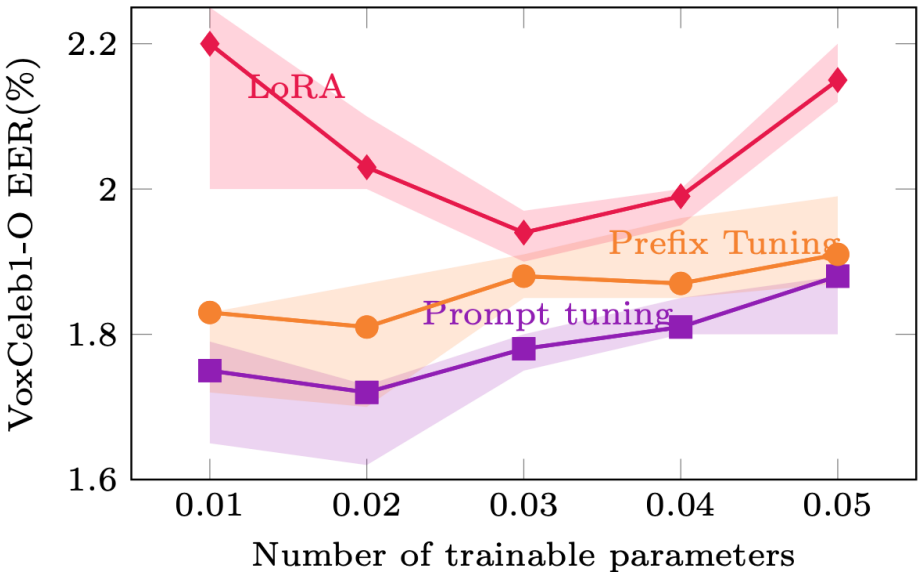


Fig . The trade-off between EER and the number of trainable parameters for various parameter-efficient transfer learning methods. The PTM is WavLM Large.

- Prompt tuning outperforms prefix tuning and LoRA.
- Pre-training can capture phonetic properties; however, they are not directly useful for distinguishing speakers.
- Too few parameters do not adapt the PTM well, while too many parameters reduce the discrimination of speaker features.

Comparing Fine-Tuning Methods

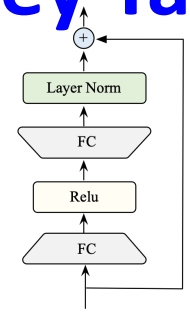
The performance on CU-MARVEL. CA: Cross-Attention Prompt Tuning

Pre-trained Transformer Model	Fine-tuning	Parames	Speaker Encoder	Evaluation metrics	
				EER(%)	minDCF
WavLM Large	Fixed	0M	ECAPA-TDNN	6.66	0.88
	Encoder Adapter	0.5M		5.58	0.81
	Layer Adapter	0.5M		5.62	0.84
	Prompt tuning	0.55M		6.42	0.88
	CA Prompt tuning	18.14M		6.26	0.85
	SPTAdapter	20.19M		**5.49**	**0.81**

Chen, Sanyuan, et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing." IEEE Journal of Selected Topics in Signal Processing 16.6 (2022): 1505-1518.
Desplanques, B., et al. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. Proc. *Interspeech 2020*

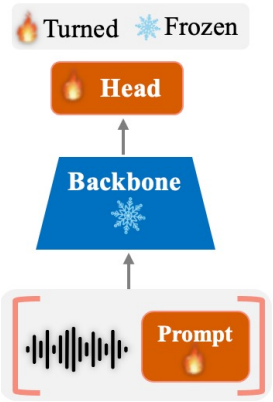
Key Takeaways

Adapters



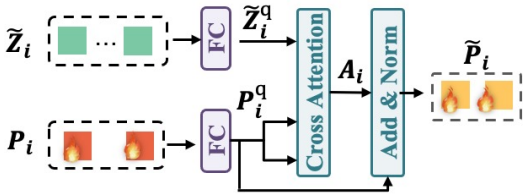
Optimizing PTM for speaker verification

Speaker Prompt Tuning



Injecting task-specific learnable parameters into Transformer

Cross attention Prompt

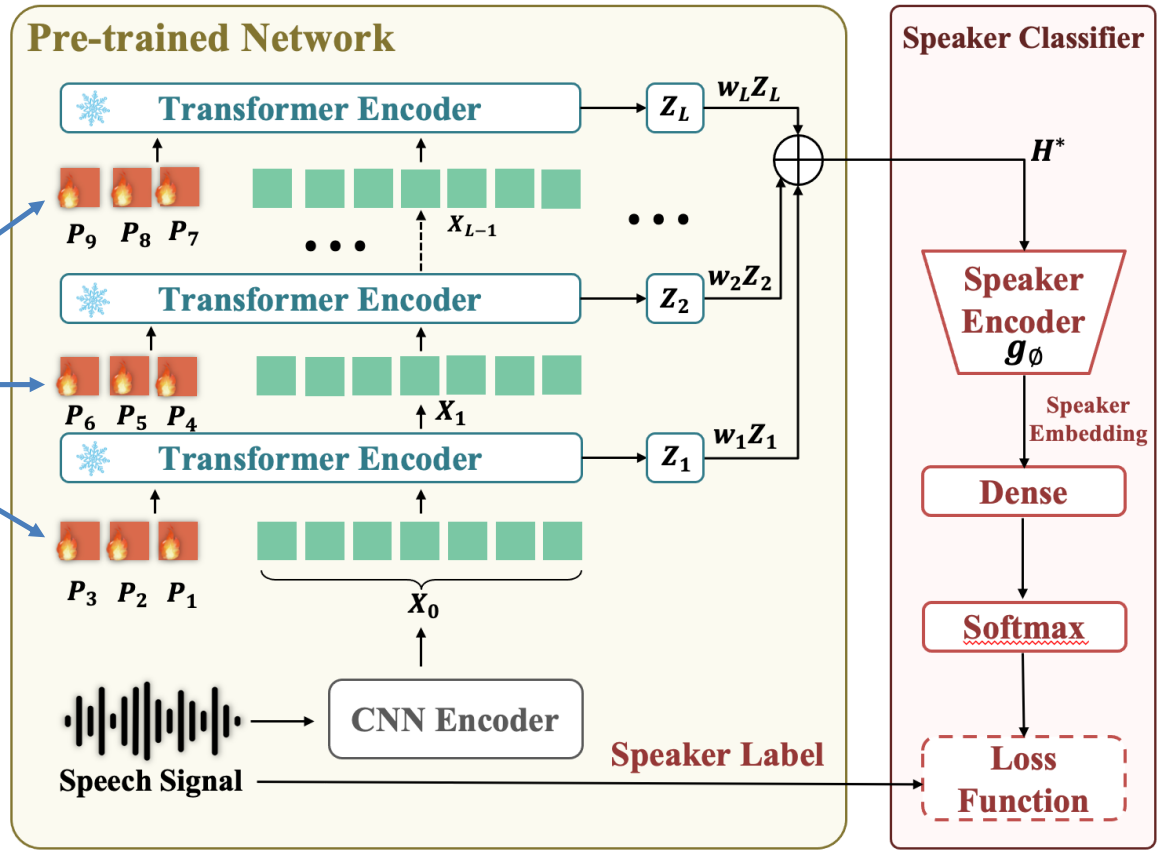


Enrich speaker information

Issues of Static Prompts

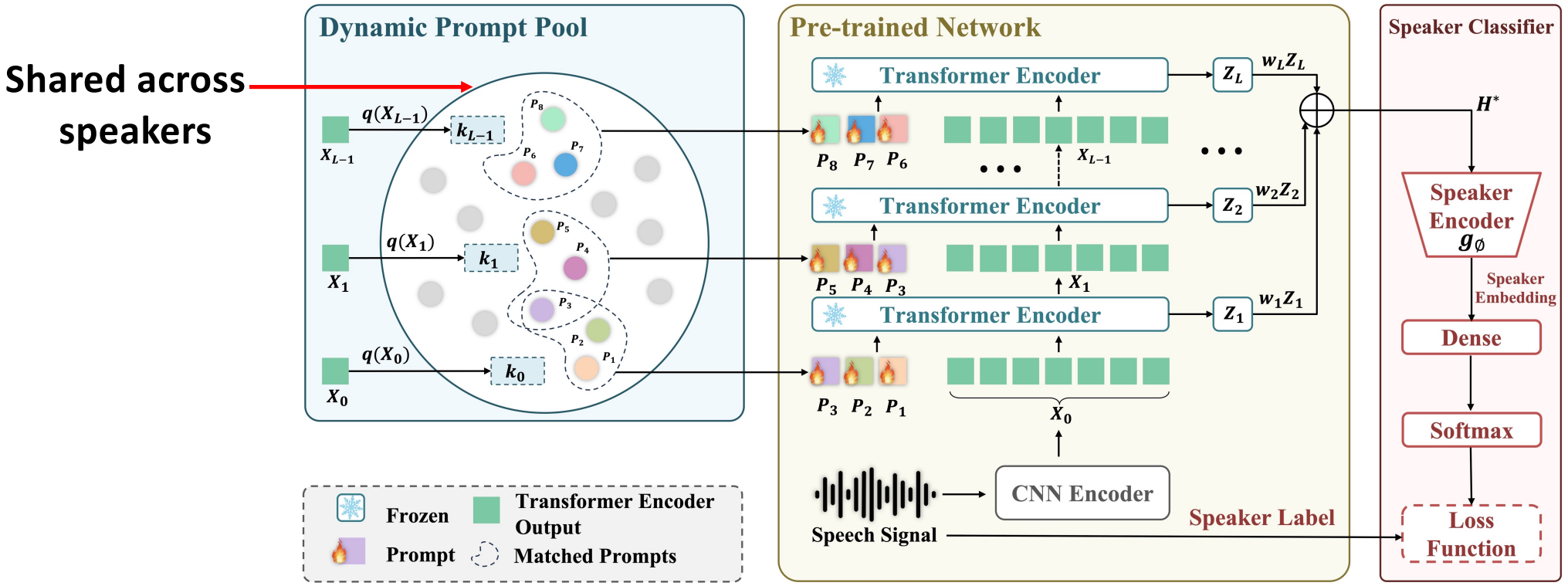
- Having a set of prompts for each training speaker leads to the static prompts overfitting the training speakers.

Speaker-Specific prompts



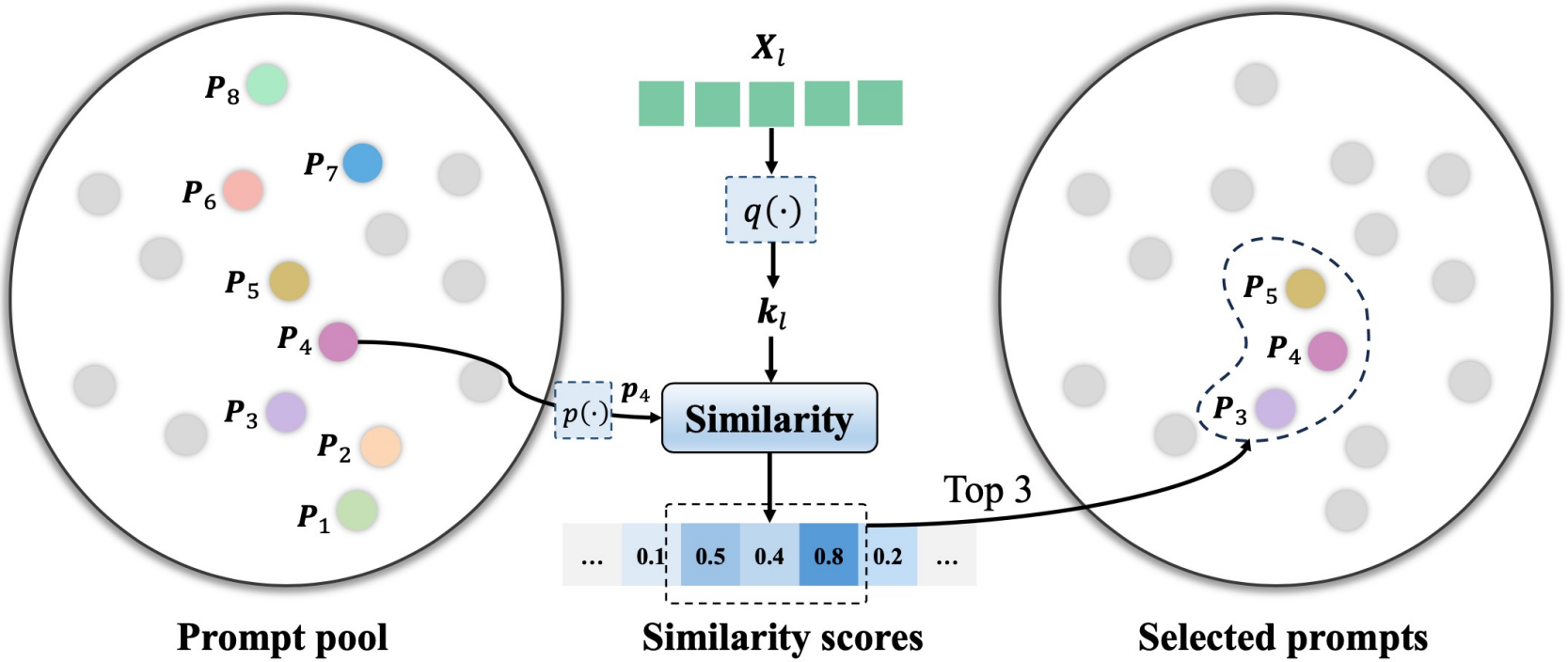
Zhe Li, Man-wai Mak, Hung-yi Lee, Helen Meng, "Parameter-efficient Fine-tuning of Speaker-Aware Dynamic Prompts for Speaker Verification", *Proc. Interspeech*, Kos Island, Greece, Sept 2024.

Extension to Dynamic Prompt Pool



Zhe Li, Man-wai Mak, Hung-yi Lee, Helen Meng, "Parameter-efficient Fine-tuning of Speaker-Aware Dynamic Prompts for Speaker Verification", *Proc. Interspeech*, Kos Island, Greece, Sept 2024.

Dynamic Prompt Search



Zhe Li, Man-wai Mak, Hung-yi Lee, Helen Meng, "Parameter-efficient Fine-tuning of Speaker-Aware Dynamic Prompts for Speaker Verification", *Proc. Interspeech*, Kos Island, Greece, Sept 2024.

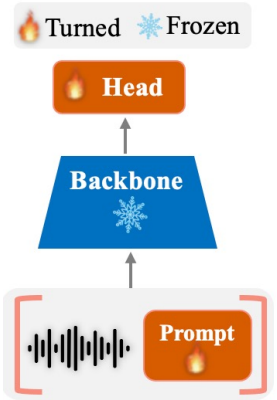
SV Performance

PTM	Fine-tuning Method	#Params	VoxCeleb1-O		CN-Celeb1		CU-MARVEL	
			EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
-	-	14.7M	2.96	0.30	12.49	0.67	7.20	0.77
HuBERT Large	Fixed	0.0M+14.7M	2.76	0.30	12.05	0.61	10.40	0.93
	Full fine-tuning	316M+14.7M	1.98	0.22	10.51	0.60	11.65	0.98
	Adapter	0.5M+14.7M	2.13	0.24	10.89	0.62	8.10	0.95
	LoRA	0.5M+14.7M	2.38	0.23	10.48	0.60	9.11	0.92
	Static prompt	0.6M+14.7M	2.26	0.23	10.69	0.59	8.31	0.88
	Dynamic prompts (Ours)	0.3M+14.7M	2.17	0.21	10.61	0.58	8.20	0.86
WavLM Large	Fixed	0.0M+14.7M	1.94	0.22	11.17	0.59	6.66	0.88
	Full fine-tuning	316M+14.7M	1.39	0.16	10.47	0.56	9.09	0.94
	Adapter	0.5M+14.7M	1.68	0.19	10.83	0.63	5.58	0.81
	LoRA	0.5M+14.7M	1.88	0.21	10.89	0.63	6.83	0.88
	Static prompt	0.6M+14.7M	1.65	0.18	10.57	0.58	6.42	0.88
	Dynamic prompts (Ours)	0.3M+14.7M	1.51	0.17	10.38	0.59	6.62	0.83

Zhe Li, Man-wai Mak, Hung-yi Lee, Helen Meng, "Parameter-efficient Fine-tuning of Speaker-Aware Dynamic Prompts for Speaker Verification", *Proc. Interspeech*, Kos Island, Greece, Sept 2024.

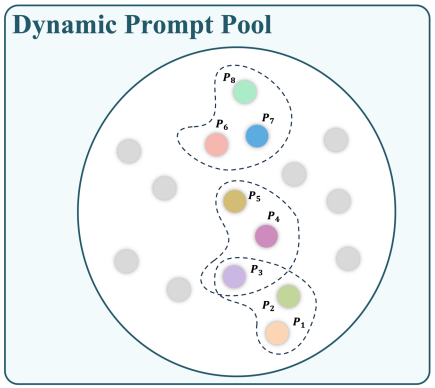
Advantages of Dynamic Prompts

Speaker Prompt Tuning



Injecting task-specific learnable parameters into Transformer layers

Dynamic Prompt pool



- Allow prompts sharing
- Avoid overfitting
- Better SV performance

Zhe Li, Man-wai Mak, Hung-yi Lee, Helen Meng, "Parameter-efficient Fine-tuning of Speaker-Aware Dynamic Prompts for Speaker Verification", *Proc. Interspeech*, Kos Island, Greece, Sept 2024.

Spectral-Aware Low-Rank Adaptation

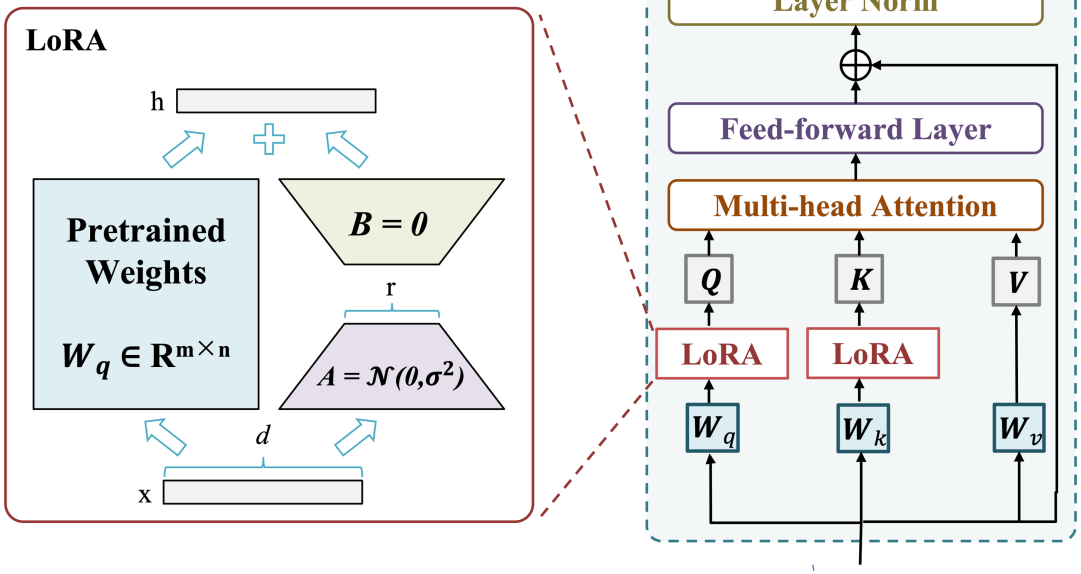
Low-Rank Adaptation (LoRA)

Simple and Effective

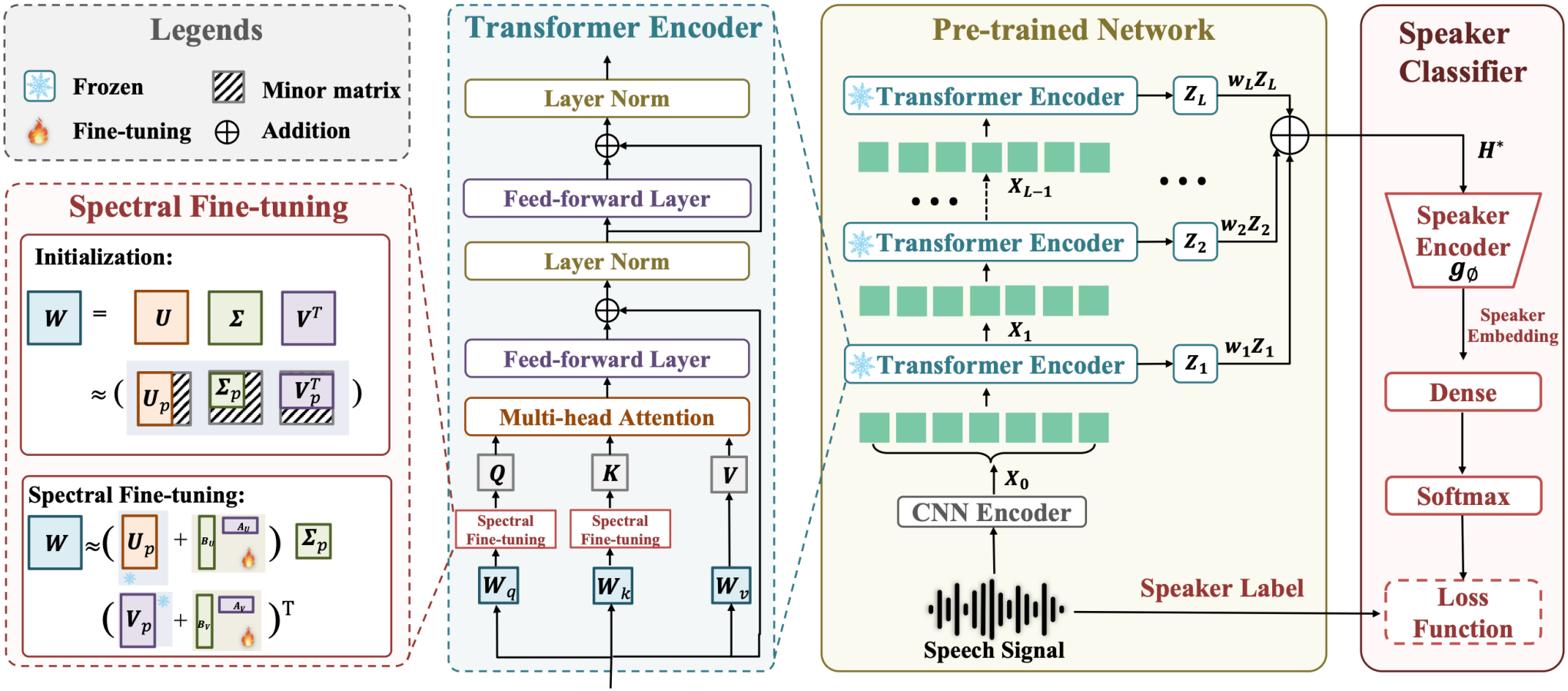
- Introduces trainable low-rank matrices to adapt pre-trained models.
- Only a small number of additional parameters are trained.
- Merged into the original weights at inference time with **zero additional latency**.

Limitations

- The **low-rank constraint** may underperform on tasks requiring high representational capacity.
- LoRA does **not explicitly utilize** top singular values and vectors of the original weight matrix, possibly missing valuable parameter space directions.



Spectral-Aware LoRA

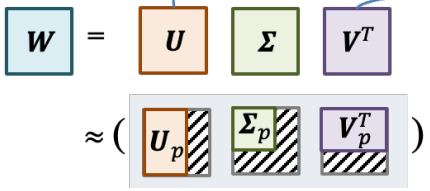


Perform SVD on W_q and W_k

$$U = [u_1, u_2, \ldots, u_m] \in \mathbb{R}^{m \times m}$$

$$V = [v_1, v_2, \ldots, v_n] \in \mathbb{R}^{n \times n}$$

$$W \in \mathbb{R}^{m \times n}$$



$$W = U\Sigma V^{\mathsf{T}} = U_p \Sigma_p V_p^{\mathsf{T}} + U_m \Sigma_m V_m^{\mathsf{T}}$$

Pre-trained weights

The first p columns contain the main information

Perform LoRA

$$\mathrm{SVD}(W) = U\Sigma V^{\top} = [U_p; U_m][\Sigma_p; \Sigma_m][V_p^{\top}; V_m^{\top}]$$

Perform LoRA

Discard

Spectral-aware Fine Tuning

Spectral-aware LoRA:

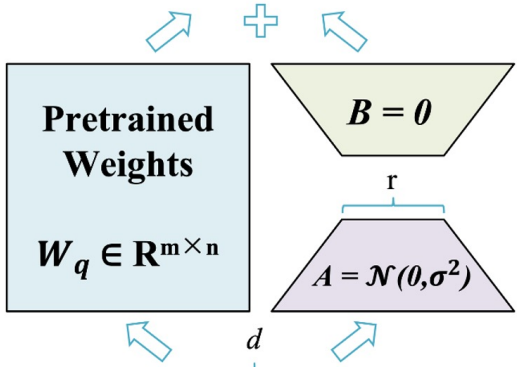
$$W \longleftarrow \left(U_p + \alpha B_U A_U \right) \Sigma_p \left(V_p + \alpha B_V A_V \right)^{\mathrm{T}}$$

Initialized to zero

Initialized using a Gaussian distribution

Standard LoRA:

$$W \leftarrow W + \alpha B A$$



Performance Comparison

Voxceleb1

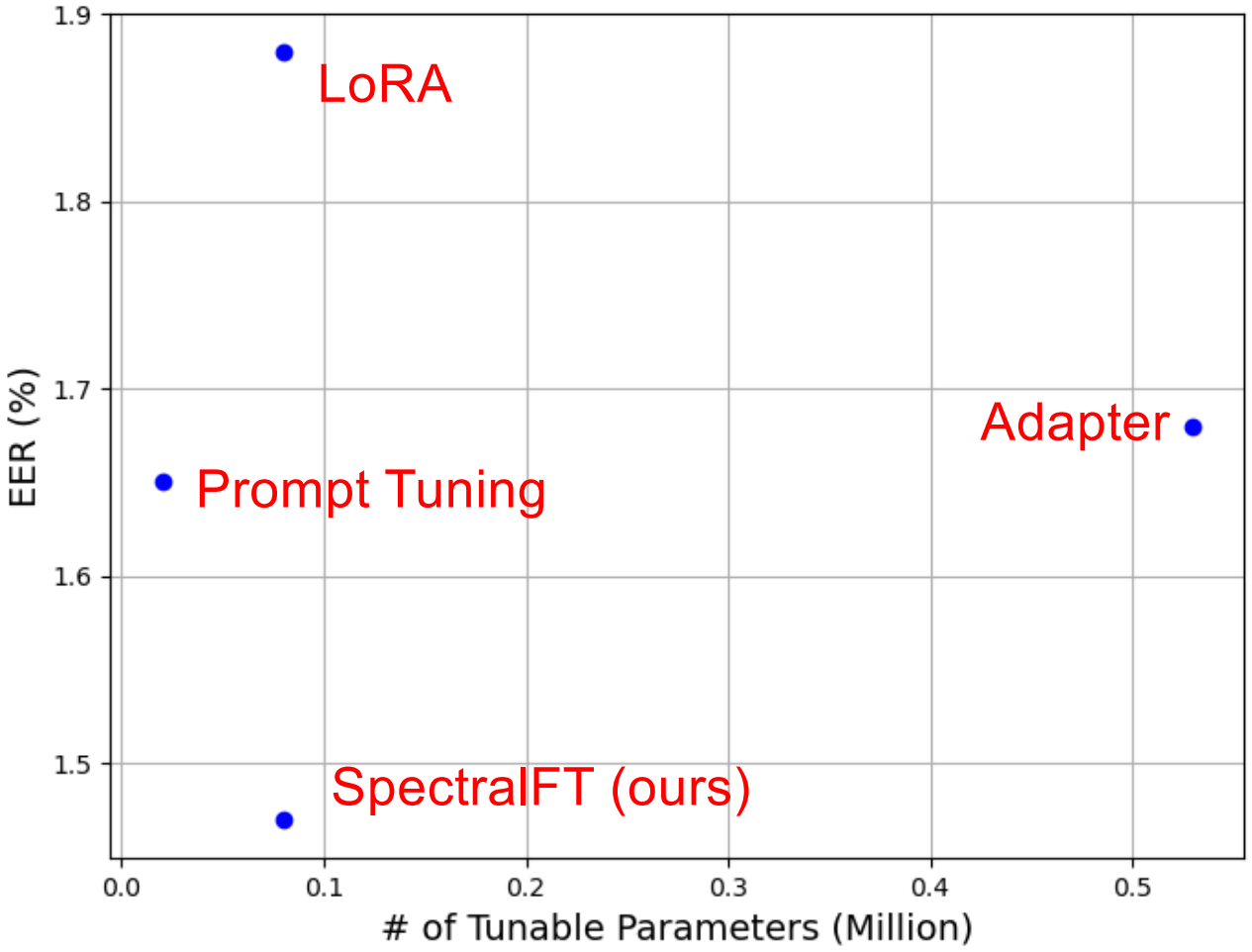
PTM	Row	Fine-tuning Method	Per-Module #Params	Trainable #Params in PTM	VoxCeleb1-O	
					EER(%)	minDCF
None	1	None	–	–	2.96	0.30
HuBERT-Large	2	None	–	–	2.76	0.30
	3	Full fine-tuning	–	315.44M	*1.98*	*0.22*
	4	Adapter [18]	0.53M	12.61M	**2.13**	0.24
	5	Static prompt tuning [18]	0.02M	0.25M	2.26	0.23
	6	LoRA (r=16, $\frac{\alpha}{r}$=0.1) [18]	0.08M	1.97M	2.38	0.23
	7	**SpectralFT (Ours)**	0.08M	1.97M	2.31	**0.22**
WavLM-Large	8	None	–	–	1.94	0.22
	9	Full fine-tuning	–	316M	*1.39*	*0.16*
	10	Adapter [18]	0.53M	12.61M	1.68	0.19
	11	Static prompt tuning [18]	0.02M	0.25M	1.65	0.18
	12	LoRA (r=16, $\frac{\alpha}{r}$=0.1) [18]	0.08M	1.97M	1.88	0.21
	13	**SpectralFT (Ours)**	0.08M	1.97M	**1.47**	**0.16**

Performance Comparison

CN-Celeb1

PTM	Row	Fine-tuning Method	Per-Module #Params	Trainable #Params in PTM	CN-Celeb1	
					EER(%)	minDCF
None	1	None	–	–	12.49	0.67
HuBERT-Large	2	None	–	–	12.05	0.61
	3	Full fine-tuning	–	315.44M	*10.51*	*0.60*
	4	Adapter [18]	0.53M	12.61M	10.89	0.62
	5	Static prompt tuning [18]	0.02M	0.25M	10.69	0.59
	6	LoRA (r=16, $\frac{\alpha}{r}$=0.1) [18]	0.08M	1.97M	10.48	0.60
	7	**SpectralFT (Ours)**	0.08M	1.97M	**10.45**	**0.58**
WavLM-Large	8	None	–	–	11.17	0.59
	9	Full fine-tuning	–	316M	*10.47*	*0.56*
	10	Adapter [18]	0.53M	12.61M	10.83	0.63
	11	Static prompt tuning [18]	0.02M	0.25M	10.57	0.58
	12	LoRA (r=16, $\frac{\alpha}{r}$=0.1) [18]	0.08M	1.97M	10.89	0.63
	13	**SpectralFT (Ours)**	0.08M	1.97M	**10.69**	**0.56**

Performance Comparison



- Pre-Trained Model: WavLMe-Large
- Test Set: Voxceleb1-O

SpectralFT uses fewer parameters but achieves good performance

Varying the Matrix Rank

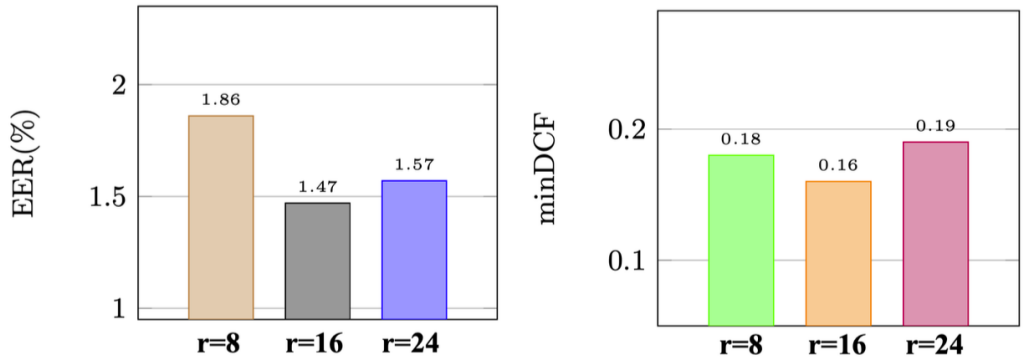
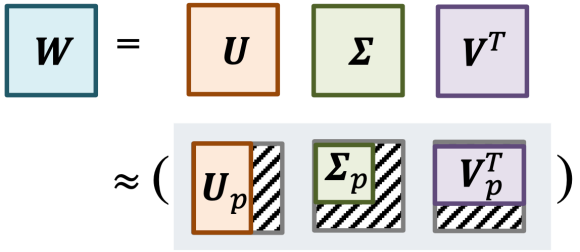


Fig. 2. Results on VoxCeleb1-O for different ranks, using WavLM-Large as the PTM.

- **Optimal Rank Selection**: SpectralFT achieves the best performance with a rank of 16.
- **Impact of Low Rank**: Insufficient rank restricts the fine-tuning subspace, preventing the model from adapting effectively to the downstream task.
- **Impact of High Rank**: Higher rank can capture more details of the downstream task, but it may lead to overfitting.

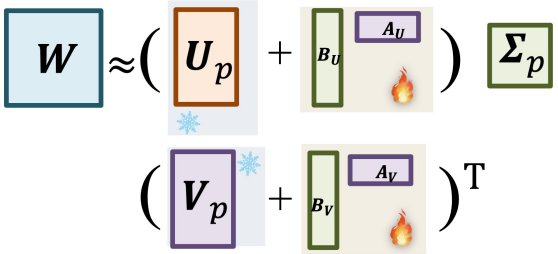
Key Takeaways

SVD



Incorporating the spectral information of pre-trained weight matrices

Spectral Fine-Tuning



Distinguishing speaker identity

Acknowledgment

- Dr. TU Youzhi (PostDoc, PolyU)
- Dr. YI Lu (PostDoc, PolyU)
- GAN Chongxin
- HUANG Zilong
- JIN Zezhong
- LI Jin
- Li Zhe
- QIN Siqing
- ZUO Lishi
- ZUO Ruichen

To Probe Further

1. Zhe Li, Man-Wai Mak, Mert Pilanci, and Helen Meng, "Mutual Information-Enhanced Contrastive Learning with Margin for Maximal Speaker Separability", *IEEE/ACM Trans on Audio, Speech and Language Processing*, June 2025.

2. Y.Z. Tu, M.W. Mak and J.T. Chien, "Contrastive Self-Supervised Speaker Embedding With Sequential Disentanglement", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, May 2024, pp. 2704-2715.

3. **Youzhi Tu, Man-Wai Mak, Kong-Aik Lee, and Weiwei Lin, "ConFusionformer: Locality-enhanced Conformer Through Multi-resolution Attention Fusion for Speaker Verification", *Neurocomputing*, May 2025**

4. Zhe Li, Man Wai Mak, Jen-Tzung Chien, Mert. Pilanci, Zezhong Jin, and Helen. Meng, "Disentangling Speaker and Content in Pre-trained Speech Models with Latent Diffusion for Robust Speaker Verification," *Proc. Interspeech*, 2025

5. **Zhe Li, Man-Wai Mak, Mert Pilanci, Hyng-yi Lee, and Helen Meng, "Spectral-Aware Low-Rank Adaptation for Speaker Verification," *Proc. ICASSP*, Hyderabad, April 2025.**

6. **Zhe Li, Man-wai Mak, Hung-yi Lee, Helen Meng, "Parameter-efficient Fine-tuning of Speaker-Aware Dynamic Prompts for Speaker Verification", *Proc. Interspeech*, Kos Island, Greece, Sept 2024.**

7. **Zhe Li, Man-Wai Mak, Helen Mei-Ling Meng, "Dual Parameter-Efficient Fine-Tuning for Speaker Representation via Speaker Prompt Tuning and Adapters", *Proc. ICASSP*, Seoul, April, 2024, pp. 10751-10755.**

8. Zhe Li, Man-Wai Mak, and Helen Mei-Ling MENG, "Discriminative Speaker Representation via Contrastive Learning with Class-Aware Attention in Angular Space", *Proc. ICASSP*, Rhodes Island, June 2023.

9. M.W. Mak and J.T. Chien, *Machine Learning for Speaker Recognition*, Cambridge University Press, 2020