Advanced Pooling Methods for Robust Speaker Verification

Man-Wai MAK

麥文偉

Dept. of Electronic and Information Engineering,

The Hong Kong Polytechnic University, Hong Kong SAR of China

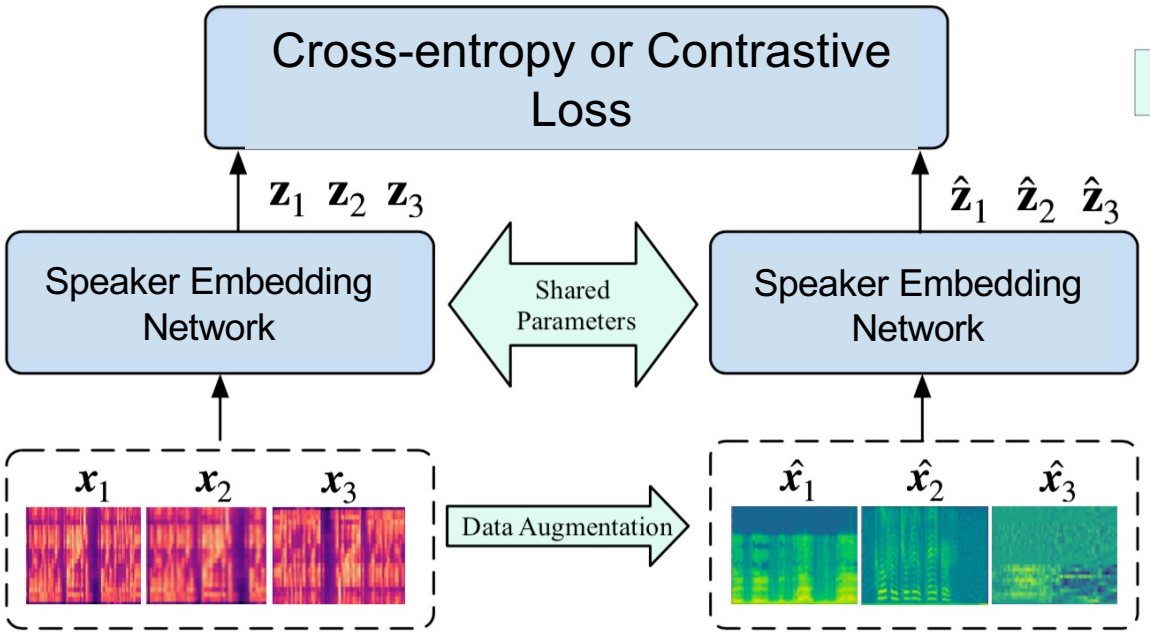
http://www.eie.polyu.edu.hk/~mwmak

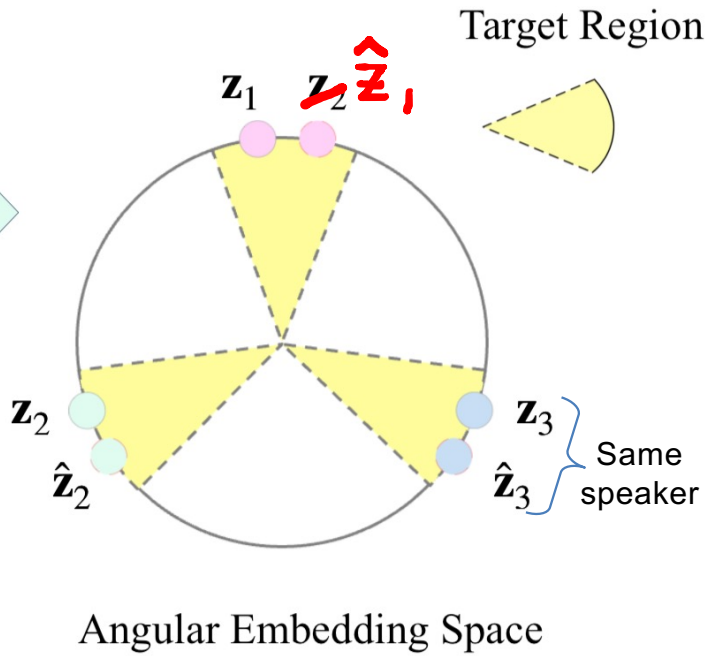
enmwmak@polyu.edu.hk

Contents

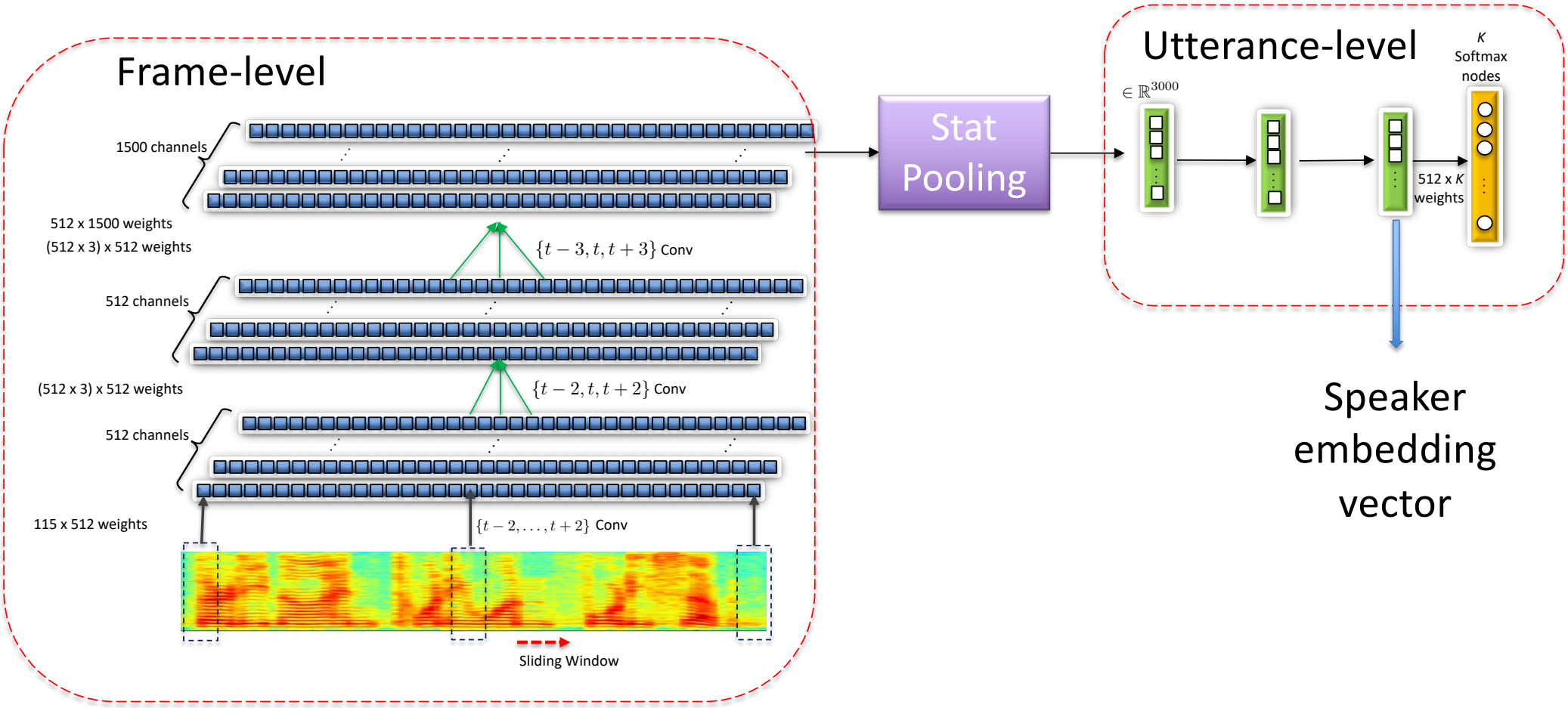
Aim of Speaker Embedding Networks

- A speaker embedding network aims to find a **speaker representation space** in which vectors (embedding) of the same speaker are close and those of different speakers are far apart.



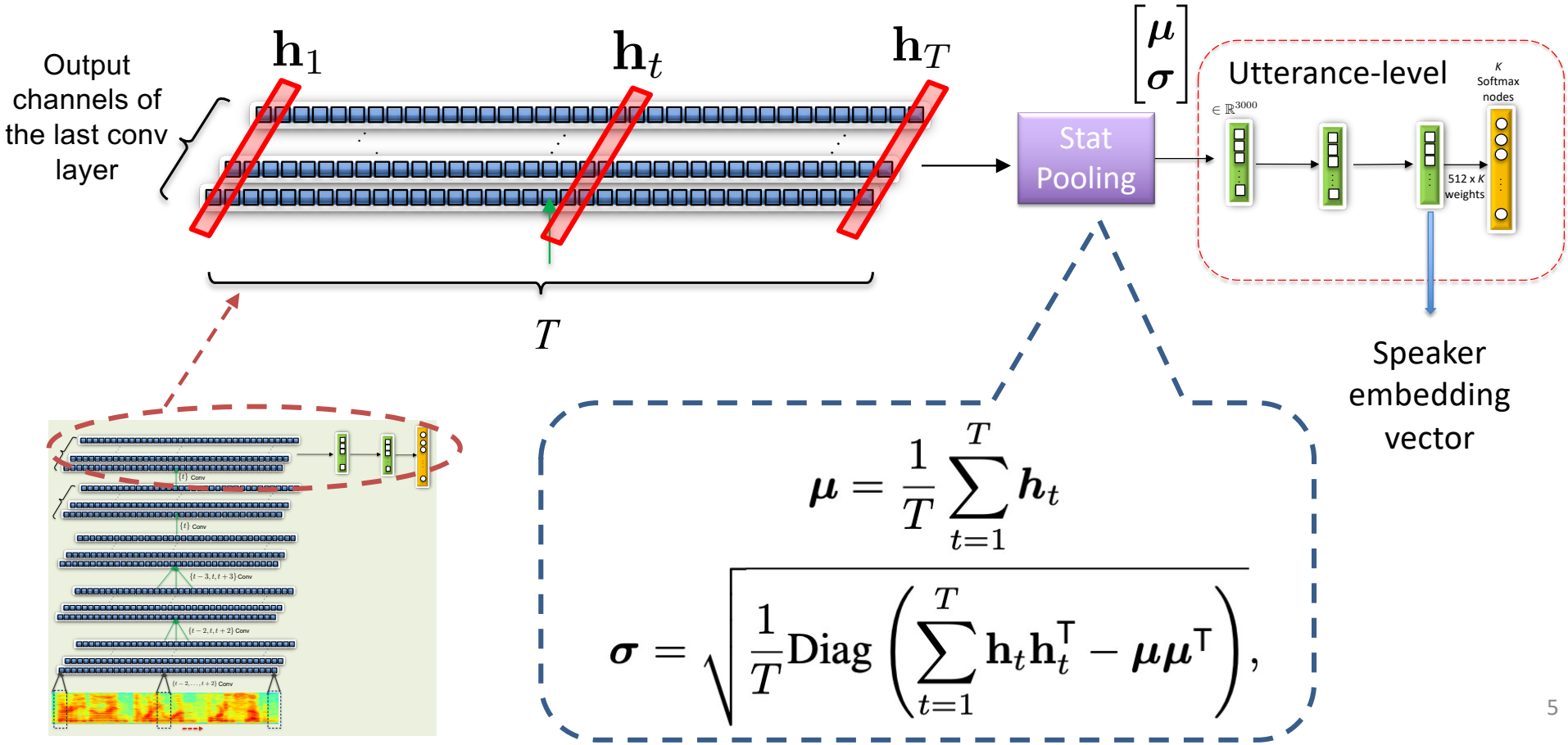


Structure of Speaker Embedding Networks



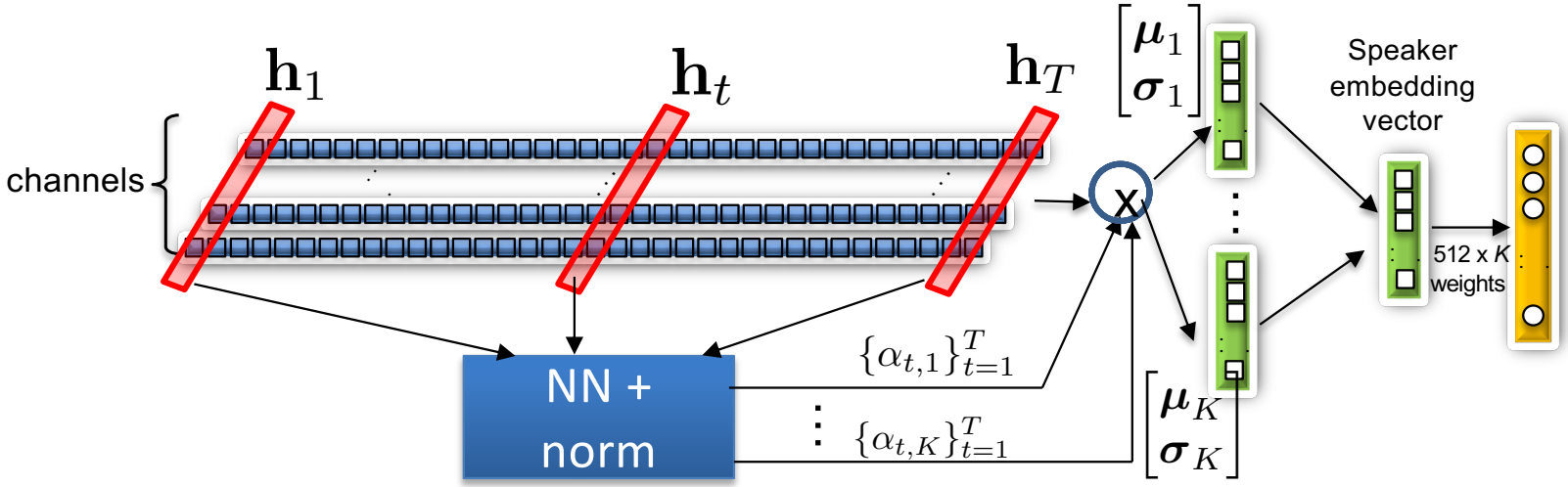
Statistics Pooling

- The statistics pooling layer concatenates the mean and the standard deviation of the activations from the last convolutional layer.



Attentive Statistics Pooling

- In attentive statistics pooling (ASP), we pay more attention to discriminative frames at the last conv layer.



$$\text{score}(\mathbf{h}_t, \mathbf{v}_k) = \mathbf{v}_k^{\mathsf{T}} f(\mathbf{W}\mathbf{h}_t + \mathbf{b})$$

$$\alpha_{t,k} = \frac{\exp\left(\text{score}(\mathbf{h}_t, \mathbf{v}_k)\right)}{\sum_{t=1}^{T} \exp\left(\text{score}(\mathbf{h}_t, \mathbf{v}_k)\right)}$$

$$k = 1, \ldots, K$$

$$\boldsymbol{\mu}_k = \sum_{t=1}^{T} \alpha_{t,k} \boldsymbol{h}_t$$

$$\boldsymbol{\sigma}_k = \sqrt{\text{Diag}\left(\sum_{t=1}^{T} \alpha_{t,k} \mathbf{h}_t \mathbf{h}_t^{\mathsf{T}} - \boldsymbol{\mu}_k \boldsymbol{\mu}_k^{\mathsf{T}}\right)}.$$

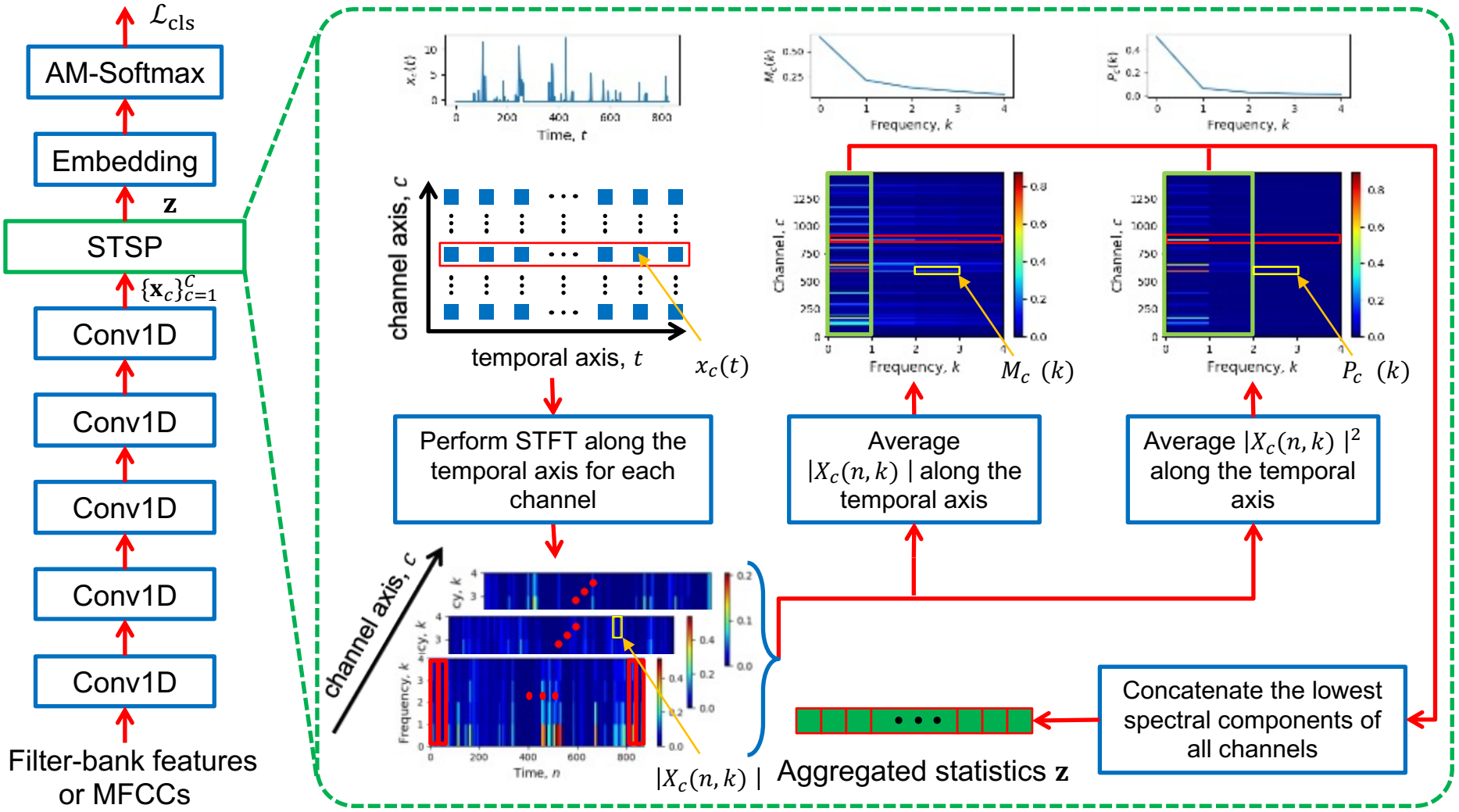
Short-Time Spectral Pooling

Motivation of STSP

- Limitation of statistics pooling
  - The temporal feature maps at the last frame-level layer is non-stationary, meaning that we should not look at the **global** statistics only.
  - From a Fourier perspective, the mean only exploits the information in the zero frequency component (DC component) in the spectral domain. The variance is sum of the spectrum over all frequencies
- **Solution**: Short-time spectral pooling (STSP)
  - Exploit the local structure of the last frame-level feature maps through short-time Fourier transform (STFT).
  - Extract multiple components (but not all) of the spectral representation as the aggregated embeddings.

Y. Z. Tu and M. W. Mak, Short-time spectral aggregation for speaker embedding", in *Proc. ICASSP'21*

Short-Time Spectral Pooling (STSP)



$$\boldsymbol{z}_c = \left(M_c(0), \sqrt{P_c(0)}, \ldots, \sqrt{P_c(R-1)}\right) \qquad \boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_c, \ldots, \boldsymbol{z}_C)$$

Attentive Short-Time Spectral Pooling

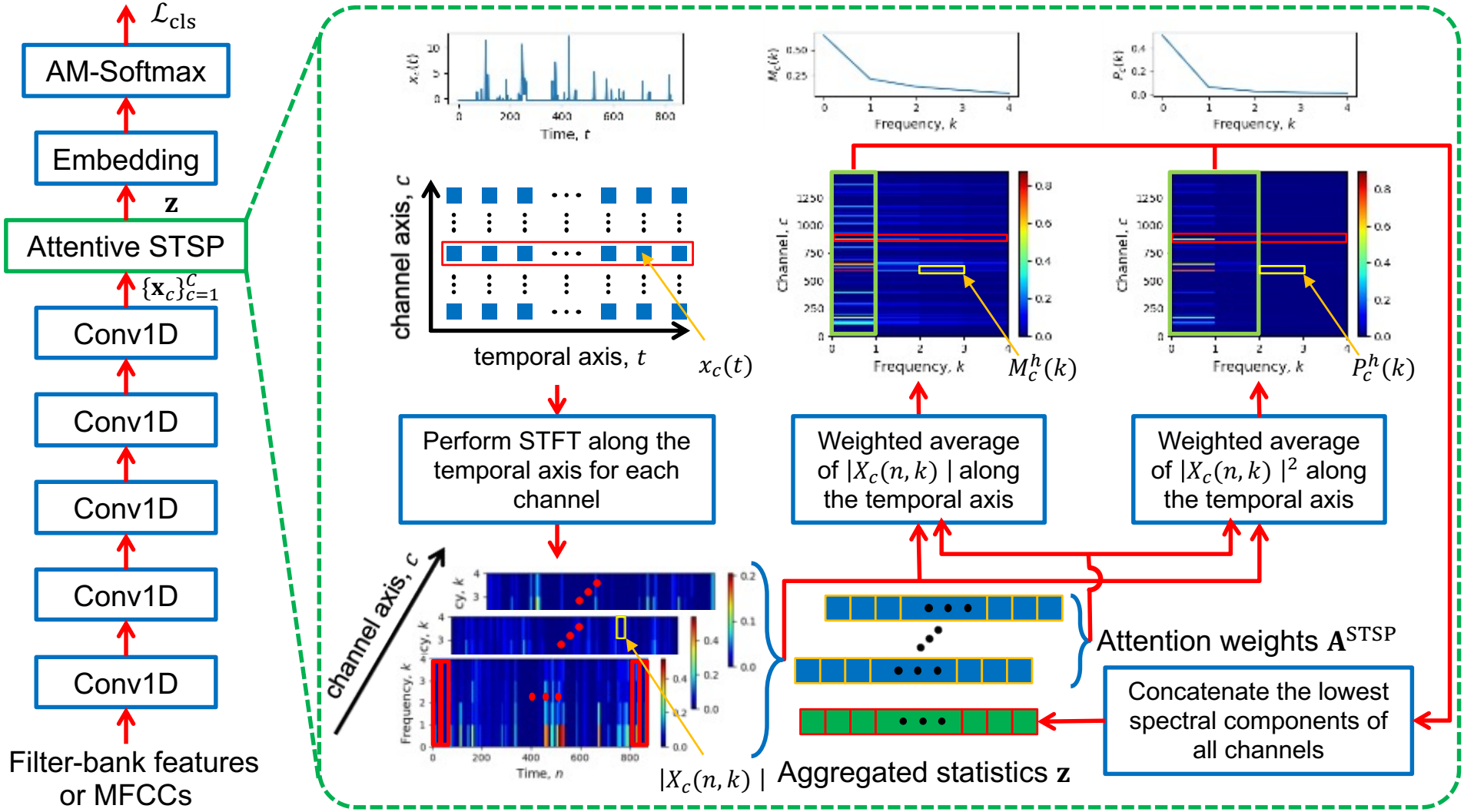
Motivation of Attentive STSP

- Limitation of STSP
  - The brute average of the spectrograms along the temporal axis ignores the importance of individual windowed segments.
  - Because phonetic information is rarely distributed uniformly across an utterance, different segments of an utterance have different speaker discriminative power.

- **Solution**: Attentive STSP
  - Apply a self-attention mechanism on the windowed segments in each spectrogram to emphasize the discriminative ones

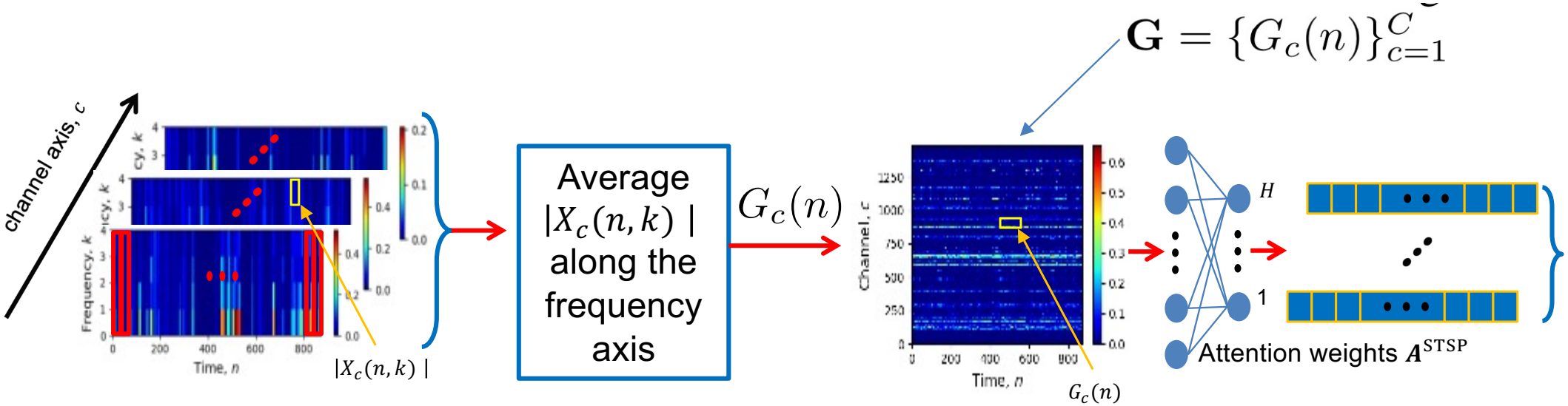
Y.Z. Tu and M.W. Mak, "Aggregating Frame-Level Information in the Spectral Domain With Self-Attention for Speaker Embedding," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 30, Feb. 2022

Multi-Head Attentive STSP



$$\boldsymbol{z}_c^h = \left(M_c^h(0), \sqrt{P_c^h(0)}, \ldots, \sqrt{P_c^h(R-1)} \right) \quad \boldsymbol{z} = (\boldsymbol{z}_1^1, \ldots, \boldsymbol{z}_c^h, \ldots, \boldsymbol{z}_C^H)$$

Computing Attention Weights



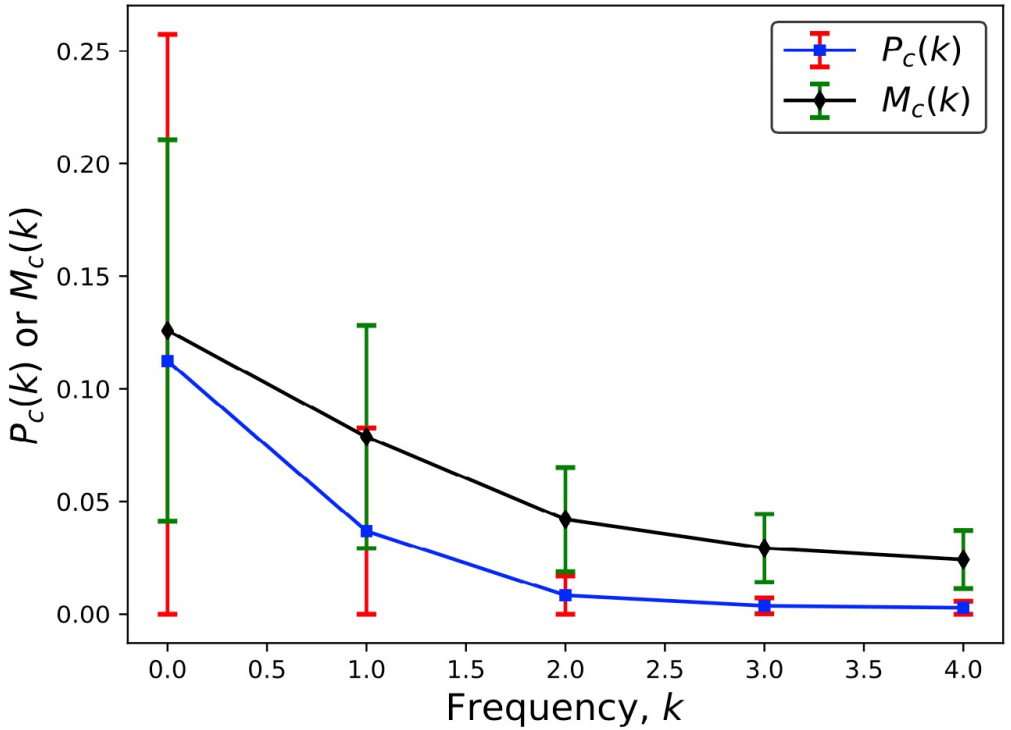
$$G_c(n) = \frac{1}{L}\sum_{k=0}^{L-1} |X_c(n,k)|, \quad n = 0, \ldots, N-1$$

$$\begin{aligned}\mathbf{A}^{\mathrm{STSP}} &= \mathrm{Softmax}\left(\tanh\left(\mathbf{G}^{\top}\mathbf{W}_1^{\mathrm{STSP}}\right)\mathbf{W}_2^{\mathrm{STSP}}\right) \\ &= \{\boldsymbol{\alpha}^h\}_{h=1}^{H}\end{aligned}$$

Property of $M_c(k)$ and $P_c(k)$

- Attentive STSP facilitates the aggregation by retaining the low spectral components only, because most of the feature energy locates at the low-frequency region.

Statistics of $M_c(k)$ and $P_c(k)$ of a randomly selected channel c over 24,220 utterances in the Voxceleb1 development set



N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in *Proc. International Conference on Machine Learning*, 2019, pp. 5301–5310.

Experimental Setup

Task	Acoustic features	Embedding training	PLDA training	Score norm cohort
VoxCeleb1 -test	40-D filter-bank features	VoxCeleb2-dev (2.09 million utterances from 5984 speakers)	VoxCeleb1-dev	N/A
VOiCES19c -eval	40-D filter-bank features	VoxCeleb1&2-dev (2.1 million utterances from 7185 speakers)	Concatenated speech with the same video session augmented with reverberation and noise	Longest two utterances of each speaker in the PLDA training data
SRE16-eval & SRE18-CMN2-eval	23-D MFCCs	SRE04-10, SWBD, Mixer6 (238,618 utterances from 5402 speakers)	clean utterances from embedding training data excluding SWBD	Unlabeled development data

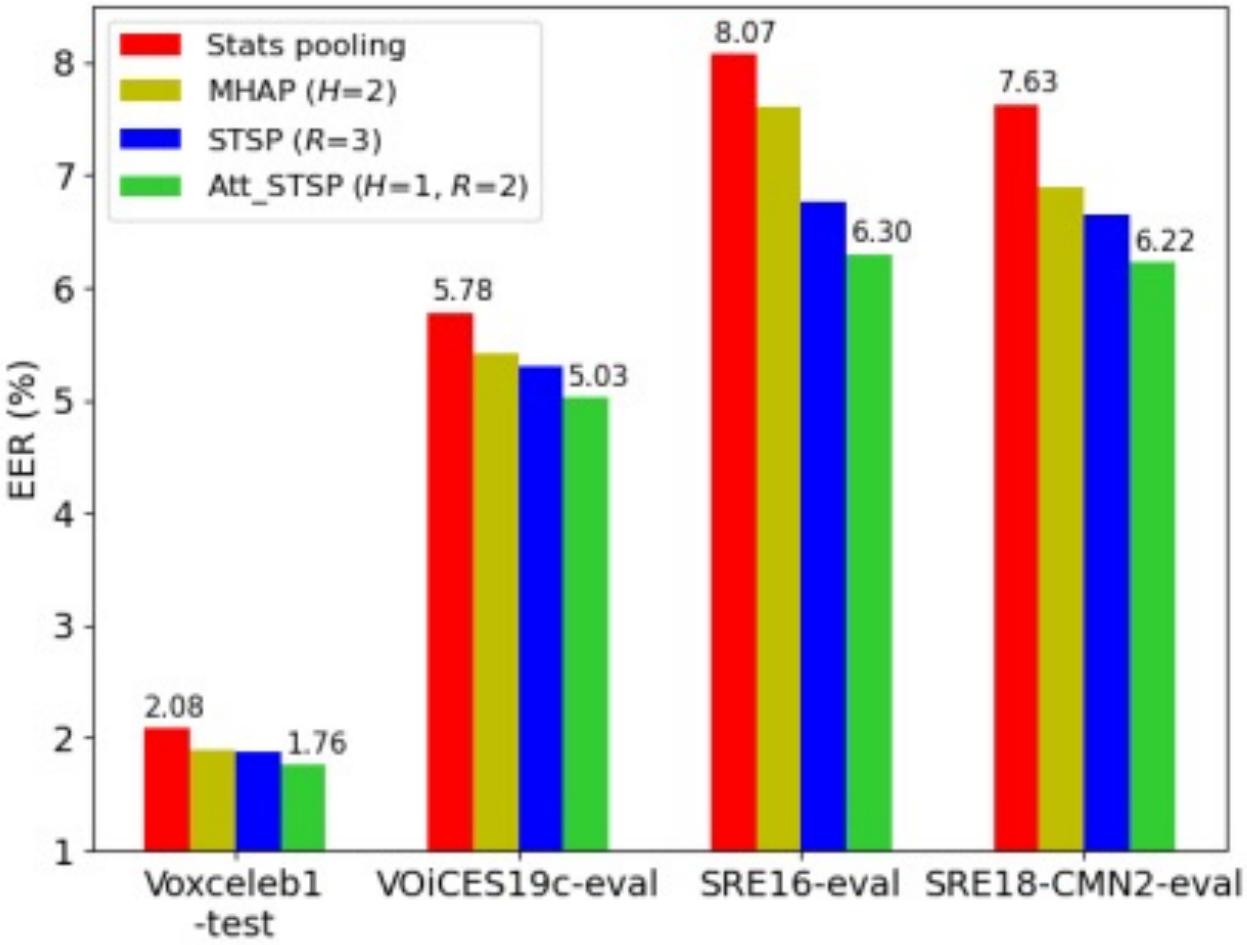
https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2
https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2

Experimental Setup

Pooling methods	Attention network config.	No. of paras of the involved emb. sys.	STFT config.
Statistics pooling	N/A	3.48 M	N/A
Multi-head attentive pooling ($H = 2$)	FC (500) + tanh + FC (2)	5.00 M	N/A
STSP ($R = 3$)	N/A	4.25 M	Rectangular window function, STFT length: 8, step size: 8
Attentive STSP ($H = 1, R = 2$)	FC (500) + tanh + FC (1)	4.61 M	

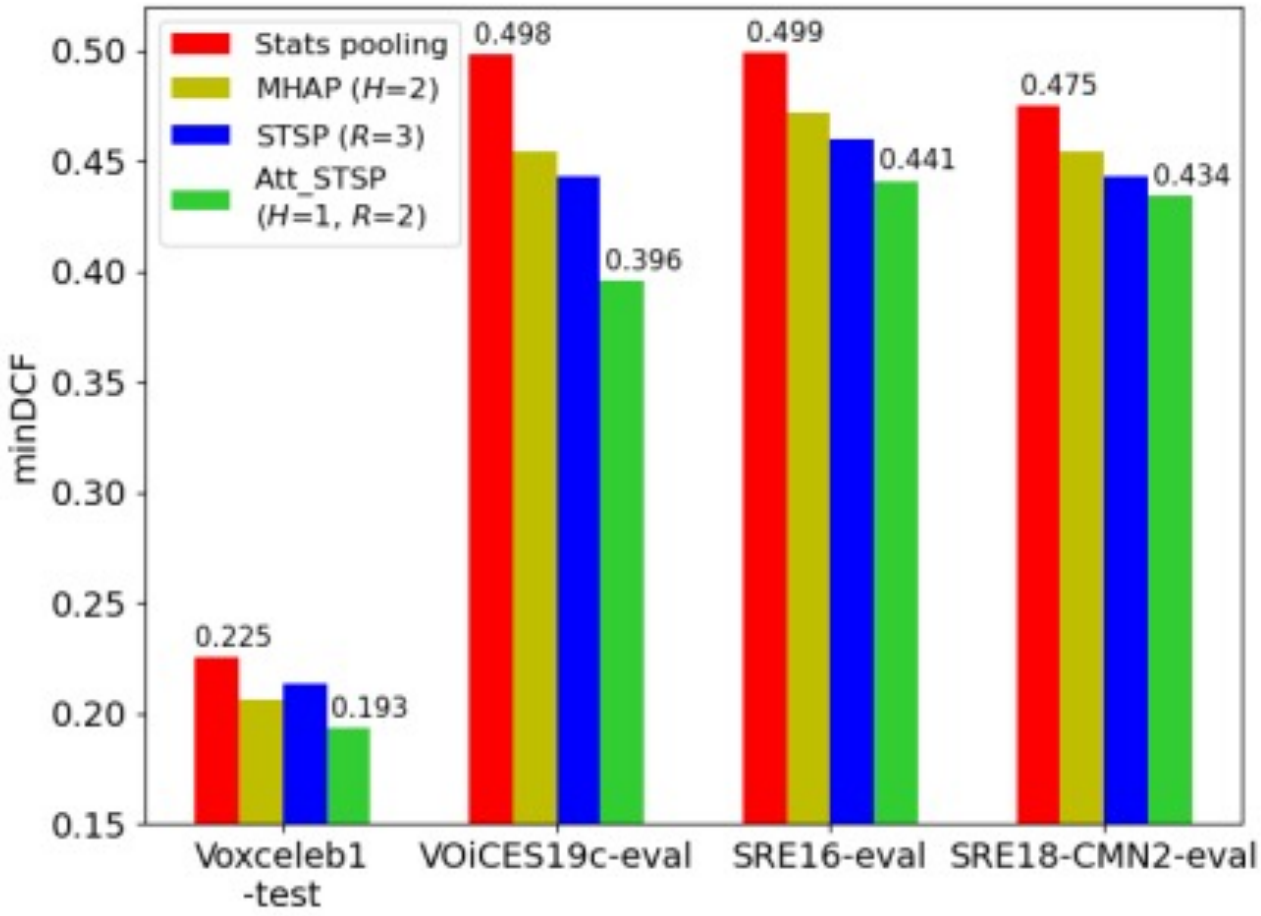
Optimizer: stochastic gradient descent (SGD) optimizer with a momentum of 0.9
Learning rate: 0.02@0, 0.05@20, 0.025@50, 0.0125@80
$H = 128$, 100 epochs

EER Performance



- Both STSP and attentive STSP **outperform statistics pooling**, which verifies that including multiple spectral components for aggregation is beneficial.
- On VoxCeleb1 and VOiCES19, multi-head attentive pooling (MHAP) and STSP perform similarly, but STSP substantially outperforms MHAP on SRE16 and SRE18-CMN2.
- Attentive STSP achieves the best performance consistently on all tasks.

Min DCF



- Both STSP and attentive STSP **outperform statistics pooling**, which verifies that including multiple spectral components for aggregation is beneficial.
- On VoxCeleb1 and VOiCES19, multi-head attentive pooling (MHAP) and STSP perform similarly, but STSP substantially outperforms MHAP on SRE16 and SRE18-CMN2.
- Attentive STSP achieves the best performance consistently on all tasks.

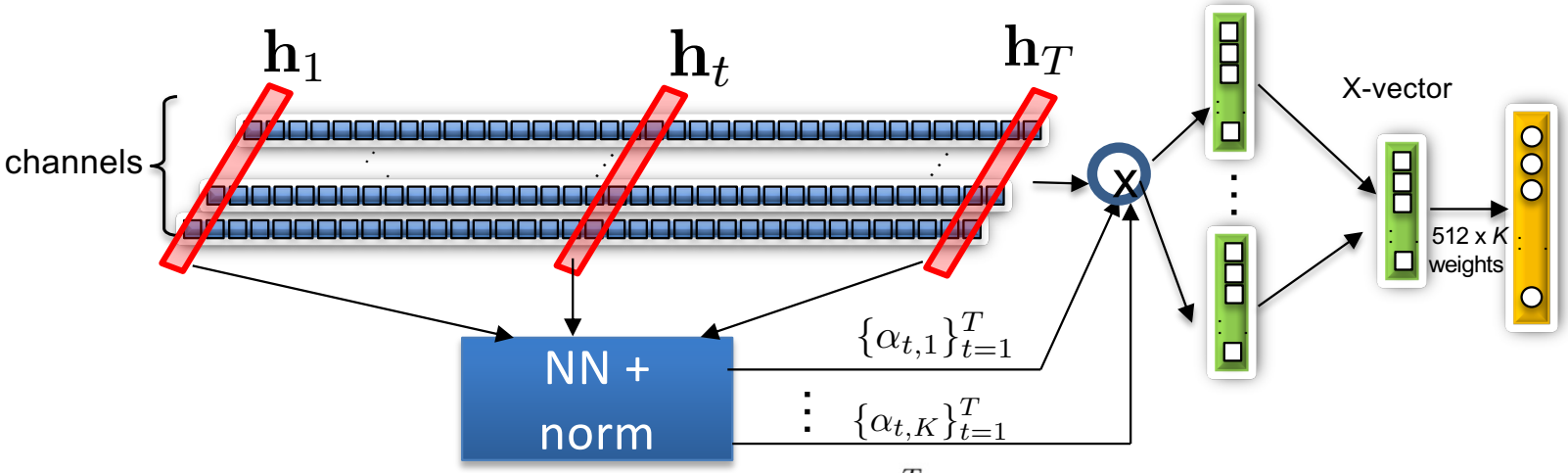
Observations

- **Attentive short-time spectral pooling** (STSP) are able to aggregate the information **beyond the DC component**, making it preserves more speaker information than statistics pooling.
- Attentive STSP exploits the **local stationarity** in the frame-level features and have better robustness against the non-stationarity in the temporal domain.
- Applying a self-attention mechanism on the windowed segments is effective to produce discriminative embeddings.

Mixture Representation Pooling

Mixture Representation Pooling

- In mixture representation pooling (MRP), the attention weights are **normalized across** the *K* attention **heads**.



$$\text{score}(\mathbf{h}_t, \mathbf{v}_k) = \mathbf{v}_k^{\mathsf{T}} f(\mathbf{W}\mathbf{h}_t + \mathbf{b})$$

$$\alpha_{t,k} = \frac{\exp\left(\text{score}(\mathbf{h}_t, \mathbf{v}_k)\right)}{\sum_{k=1}^{K} \exp\left(\text{score}(\mathbf{h}_t, \mathbf{v}_k)\right)}$$

$$N_k = \sum_{t=1}^{T} \alpha_{t,k}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{t=1}^{T} \alpha_{t,k} \mathbf{h}_t$$

$$\boldsymbol{\sigma}_k = \sqrt{\text{Diag}\left(\frac{1}{N_k} \sum_{t=1}^{T} \alpha_{t,k} \left(\mathbf{h}_t - \boldsymbol{\mu}_k\right)\left(\mathbf{h}_t - \boldsymbol{\mu}_k\right)^{\mathsf{T}}\right)}$$

W.W. Lin and M.W. Mak, "Mixture Representation Learning for Deep Speaker Embedding", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 30, Feb 2022

ASP vs. MRP

ASP

$$\text{score}(\mathbf{h}_t, \mathbf{v}_k) = \mathbf{v}_k^{\mathsf{T}} f(\mathbf{W}\mathbf{h}_t + \mathbf{b})$$

$$\alpha_{t,k} = \frac{\exp\left(\text{score}(\mathbf{h}_t, \mathbf{v}_k)\right)}{\sum_{t=1}^{T} \exp\left(\text{score}(\mathbf{h}_t, \mathbf{v}_k)\right)}$$

$$k = 1, \ldots, K$$

$$\boldsymbol{\mu}_k = \sum_{t=1}^{T} \alpha_{t,k} \mathbf{h}_t$$

$$\boldsymbol{\sigma}_k = \sqrt{\text{Diag}\left(\sum_{t=1}^{T} \alpha_{t,k} \mathbf{h}_t \mathbf{h}_t^{\mathsf{T}} - \boldsymbol{\mu}_k \boldsymbol{\mu}_k^{\mathsf{T}}\right)}$$

MRP

$$\text{score}(\mathbf{h}_t, \mathbf{v}_k) = \mathbf{v}_k^{\mathsf{T}} f(\mathbf{W}\mathbf{h}_t + \mathbf{b})$$

$$\alpha_{t,k} = \frac{\exp\left(\text{score}(\mathbf{h}_t, \mathbf{v}_k)\right)}{\sum_{k=1}^{K} \exp\left(\text{score}(\mathbf{h}_t, \mathbf{v}_k)\right)}$$

$$N_k = \sum_{t=1}^{T} \alpha_{t,k}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{t=1}^{T} \alpha_{t,k} \mathbf{h}_t$$

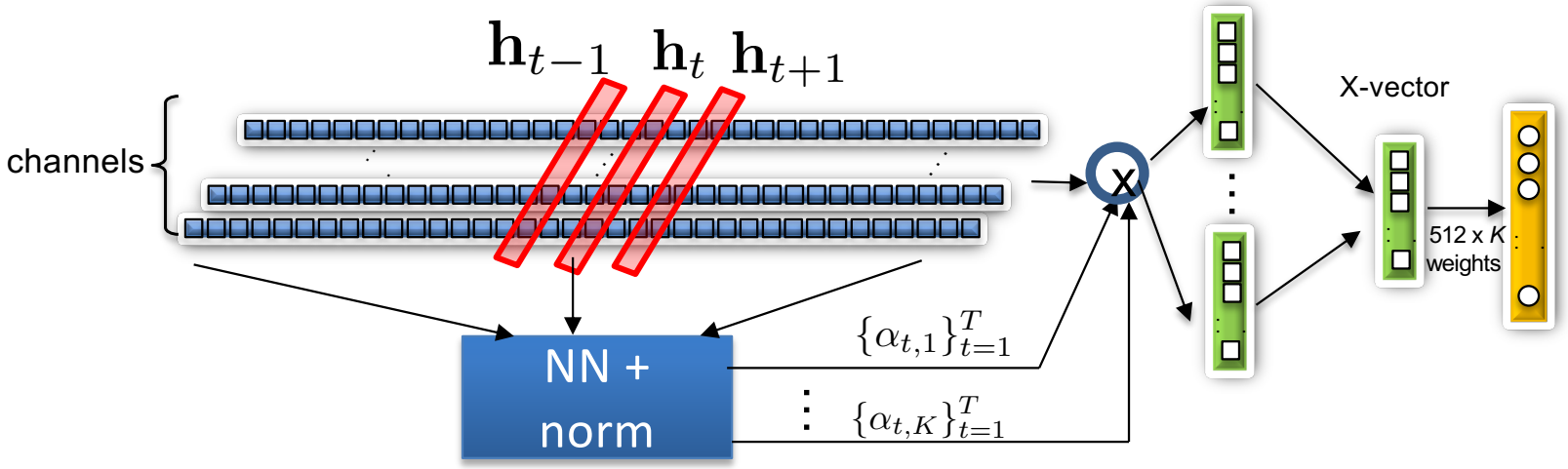
$$\boldsymbol{\sigma}_k^2 = \text{Diag}\left(\frac{1}{N_k} \sum_{t=1}^{T} \alpha_{t,k} \left(\mathbf{h}_t - \boldsymbol{\mu}_k\right)\left(\mathbf{h}_t - \boldsymbol{\mu}_k\right)^{\mathsf{T}}\right)$$

W.W. Lin and M.W. Mak, "Mixture Representation Learning for Deep Speaker Embedding", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 30, Feb 2022

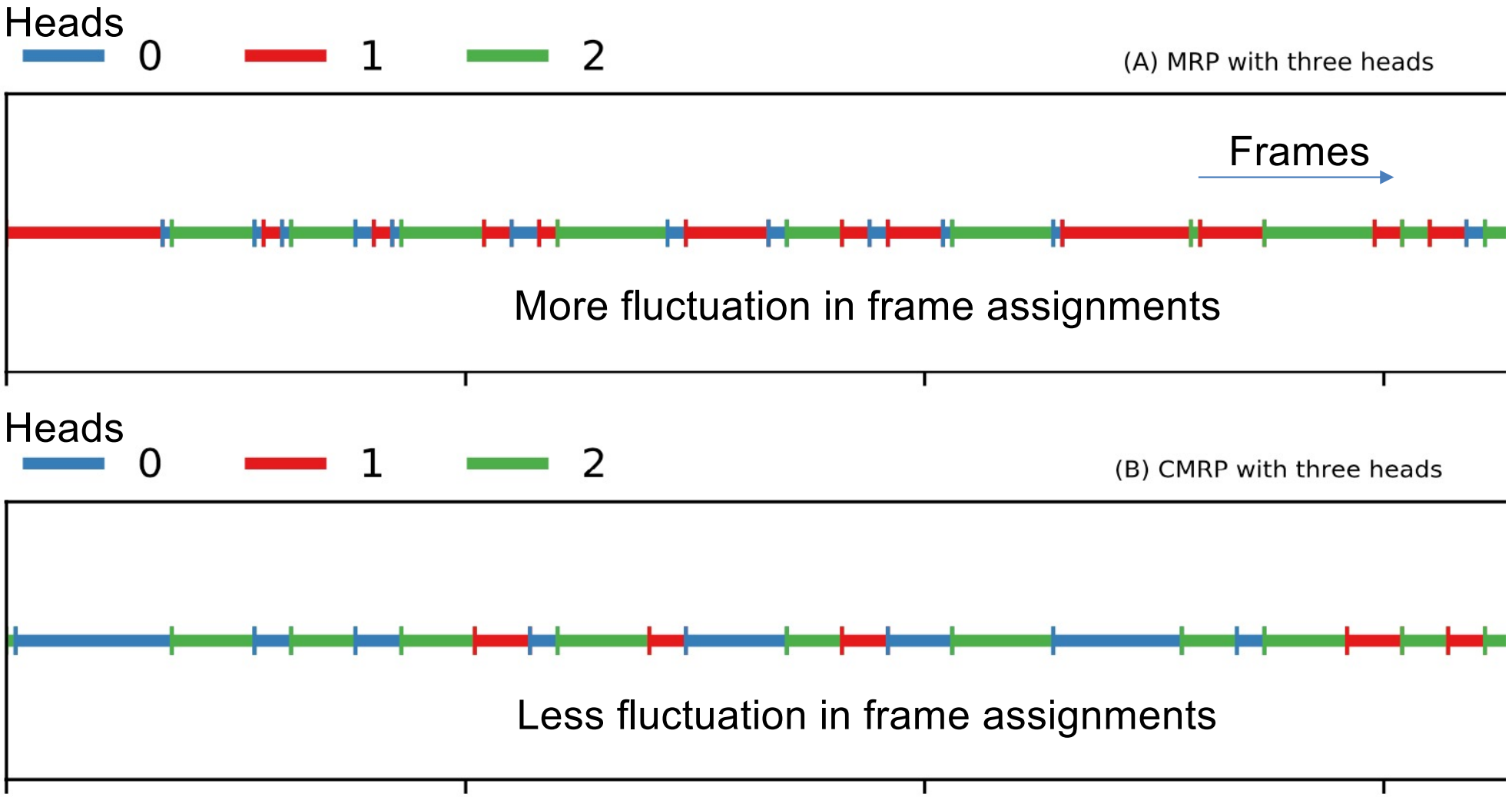
Attention with Contextual Info

- The mixture assignment should not change frequently across frames because adjacent frames are similar to each other.
- We introduce contextual information into the attention by using a block of frames adjacent to frame *t* to compute the score.

$$g(\{\mathbf{h}_{t'}\}_{t'=t-m}^{t+m}) = \frac{1}{2m+1}\sum_{t'=t-m}^{t+m}\mathbf{h}_{t'}$$



Attention with Contextual Info



Results

Model	Pooling Method	VoxCeleb1		VOiCES19-dev		VOiCES19-eval	
		EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
X-vector network	Mean & STD	2.14	0.197	2.66	0.300	6.98	0.520
Wide x-vector network	Mean & STD	2.03	0.219	2.65	0.294	6.62	0.503
Densenet121	Mean & STD	1.37	0.156	**1.53**	0.222	5.53	0.415
Densenet121	ASP	1.22	0.150	1.84	0.197	5.20	0.402
Densenet121	MRP	**1.10**	**0.131**	1.65	**0.184**	**4.77**	**0.390**

- The proposed mixture representation pooling (MRP) performs better than vanilla statistics pooling and attentive statistics pooling (ASP).
- MRP shows the most significant improvement in VOiCES19 evaluation set.

Concluding Remarks

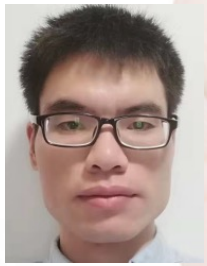
- Mixture representation pooling is inspired by Gaussian mixture models and attention mechanisms.
- Instead of normalizing frame-level features across all frames in an utterance, MRP considers each attention head as a Gaussian component of a GMM.

ASP: $\alpha_{t,k} = \frac{\exp\left(\text{score}(\mathbf{h}_t, \mathbf{v}_k)\right)}{\sum_{t=1}^{T} \exp\left(\text{score}(\mathbf{h}_t, \mathbf{v}_k)\right)}$ MRP: $\alpha_{t,k} = \frac{\exp\left(\text{score}(\mathbf{h}_t, \mathbf{v}_k)\right)}{\sum_{k=1}^{K} \exp\left(\text{score}(\mathbf{h}_t, \mathbf{v}_k)\right)}$

- The contextual information also help improves speaker embedding by reducing the fluctuation in the mixture assignments across frames.

Acknowledgment

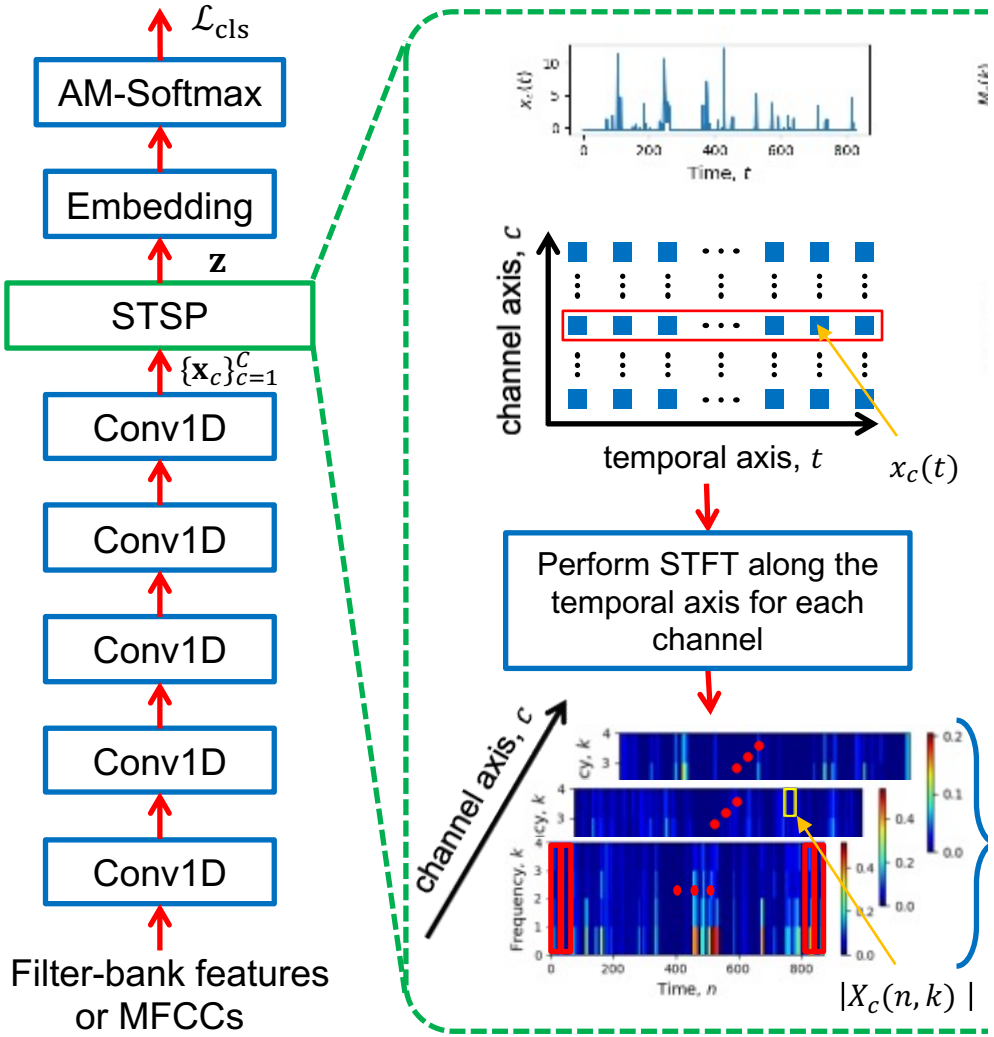
- **Youzhi TU** (my former Ph.D. student)



- **Weiwei LIN** (my former Ph.D. student, now RAP in PolyU)

References

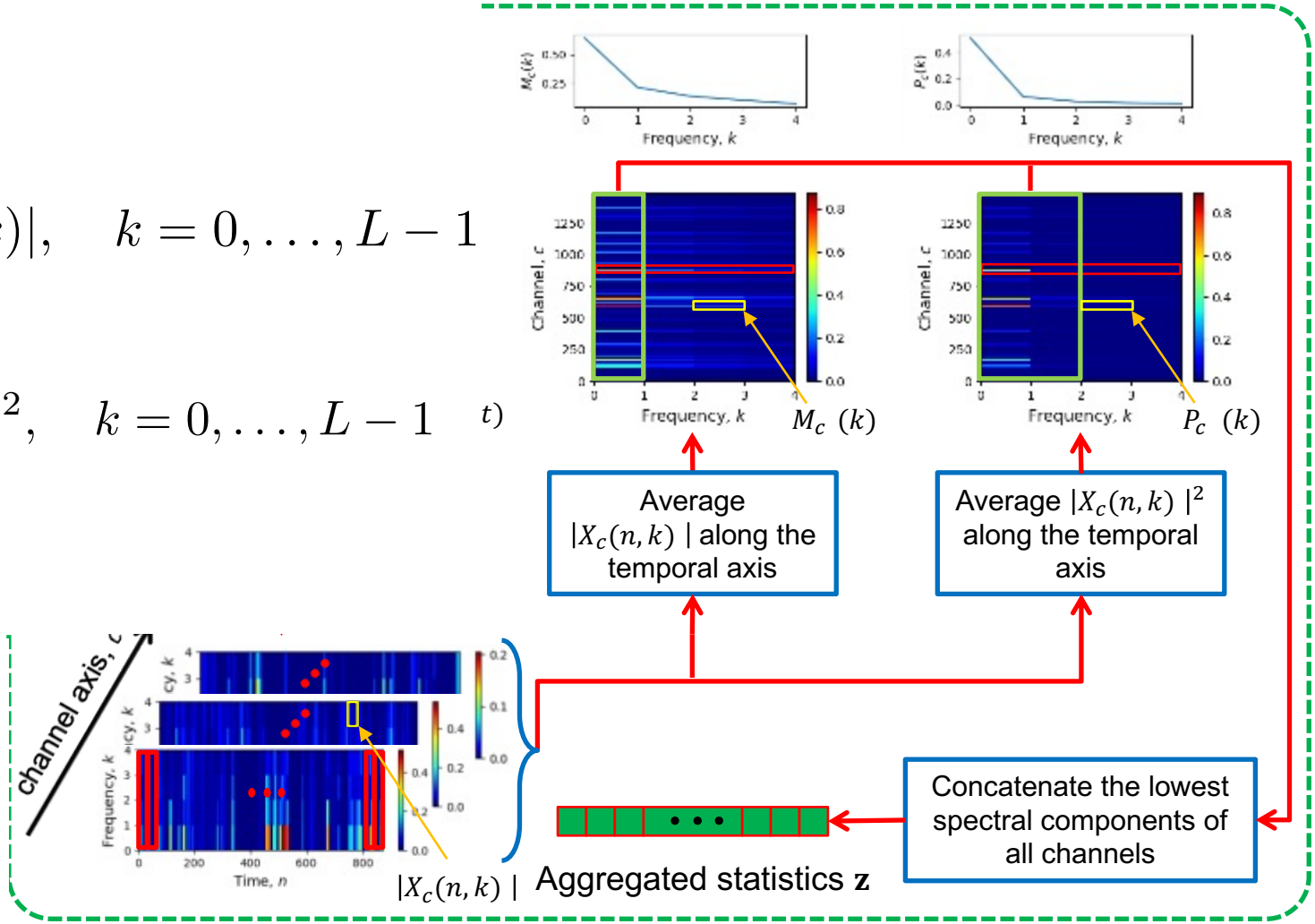
1. Y.Z. Tu, W.W. Lin, and M.W. Mak, "A Survey on Text-Dependent and Text-Independent Speaker Verification", *IEEE Access*, Sept. 2022.
2. Y.Z. Tu and M.W. Mak, "Aggregating Frame-Level Information in the Spectral Domain With Self-Attention for Speaker Embedding," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 30, Feb. 2022, pp. 944-957.
3. W.W. Lin and M.W. Mak, "Mixture Representation Learning for Deep Speaker Embedding", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 30, Feb 2022, pp. 968-978.
4. O. Rippel, J. Snoek, and R. P. Adams. "Spectral representations for convolutional neural networks." *Advances in neural information processing systems,* 28, 2015.

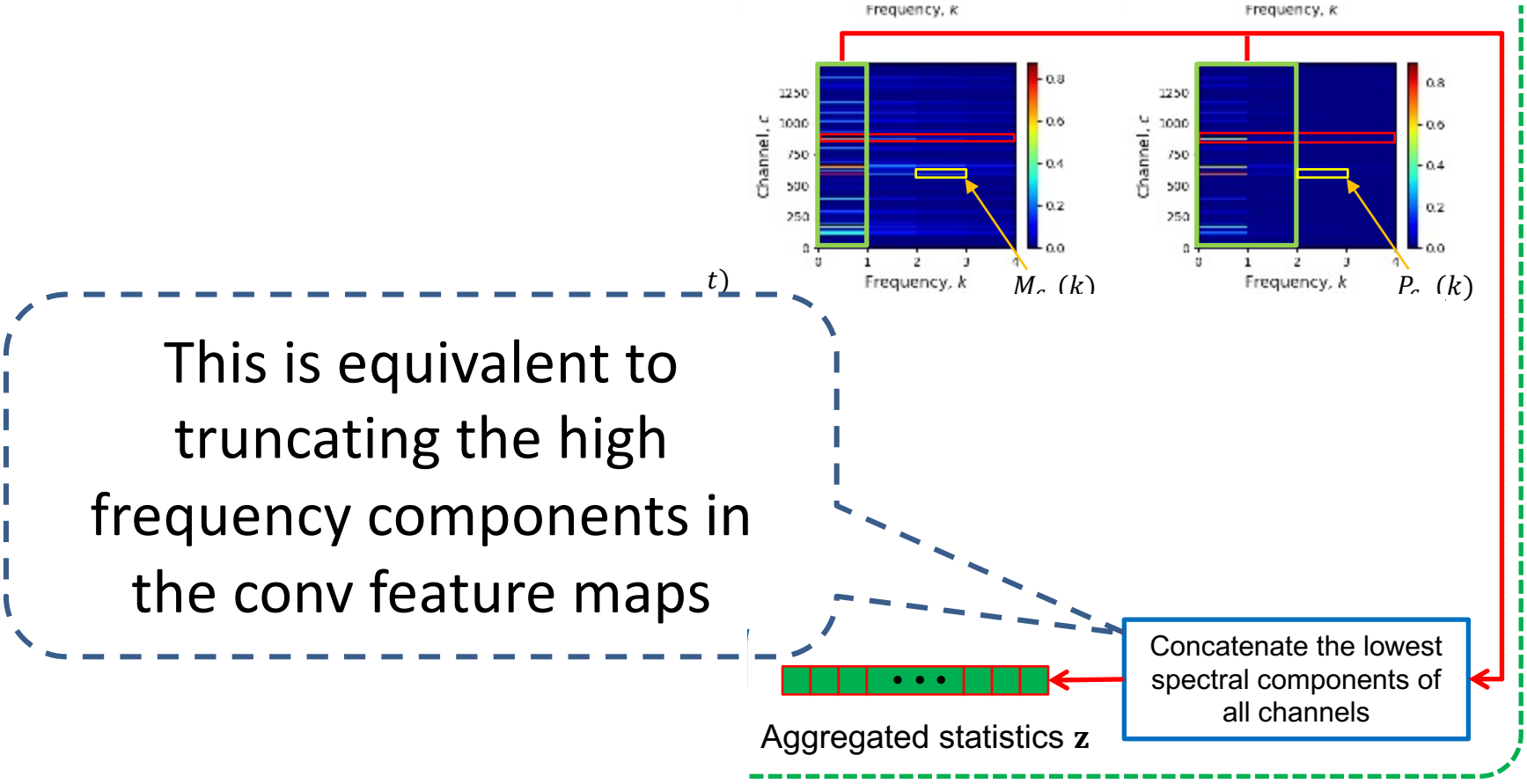


$$X_c(n,k) = \sum_{t=0}^{T-1} x_c(t) w(t - nS) e^{-\frac{j2\pi}{L} kt}$$

$$M_c(k) = \sum_{n=0}^{N-1} |X_c(n,k)|, \quad k = 0, \ldots, L-1$$

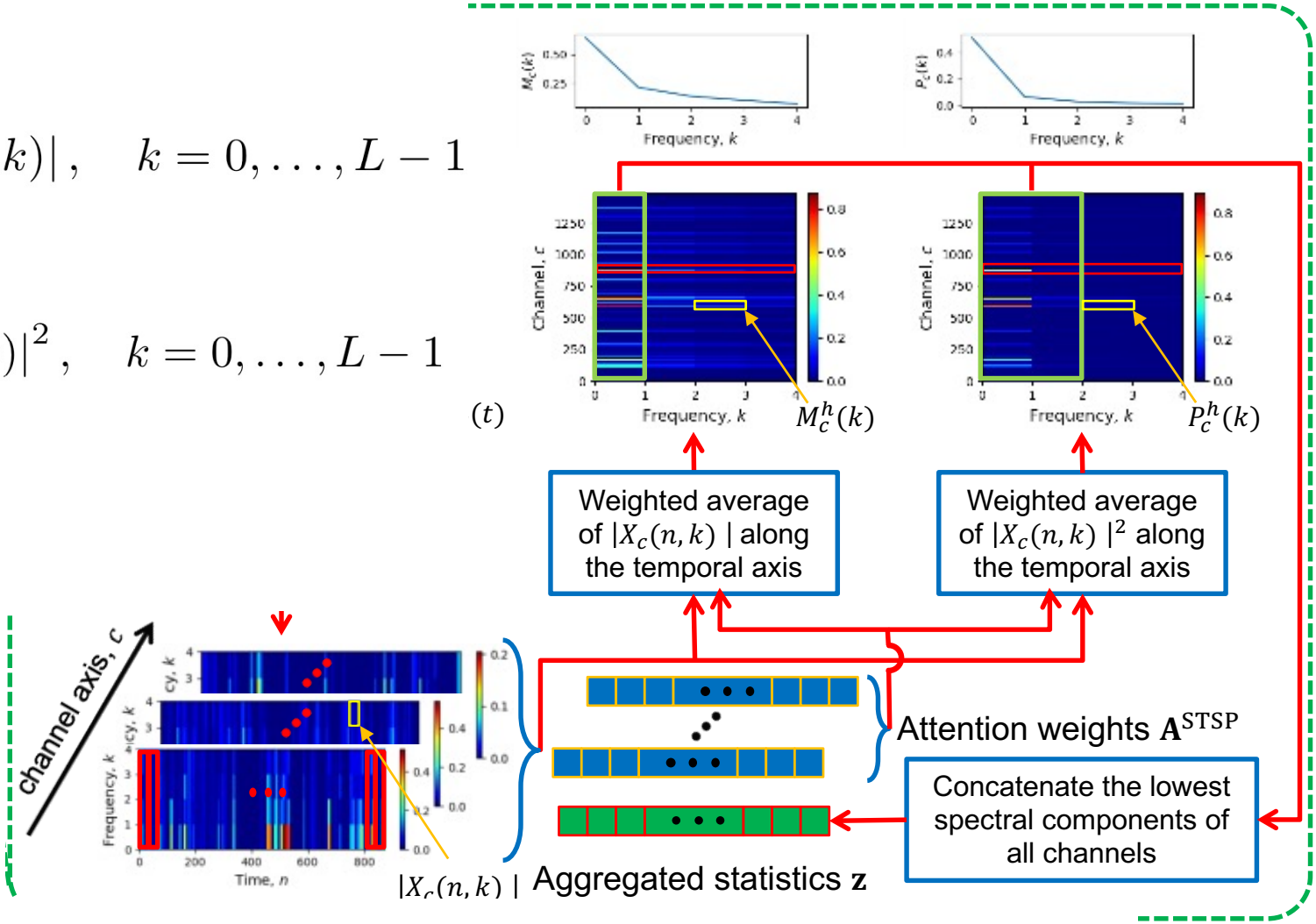
$$P_c(k) = \sum_{n=0}^{N-1} |X_c(n,k)|^2, \quad k = 0, \ldots, L-1 \qquad t)$$




Frequency, k
Channel, c
Frequency, k
$M_c(k)$
Frequency, k
Channel, c
Frequency, k
$P_c(k)$
$t)$
This is equivalent to truncating the high frequency components in the conv feature maps
Concatenate the lowest spectral components of all channels
Aggregated statistics z

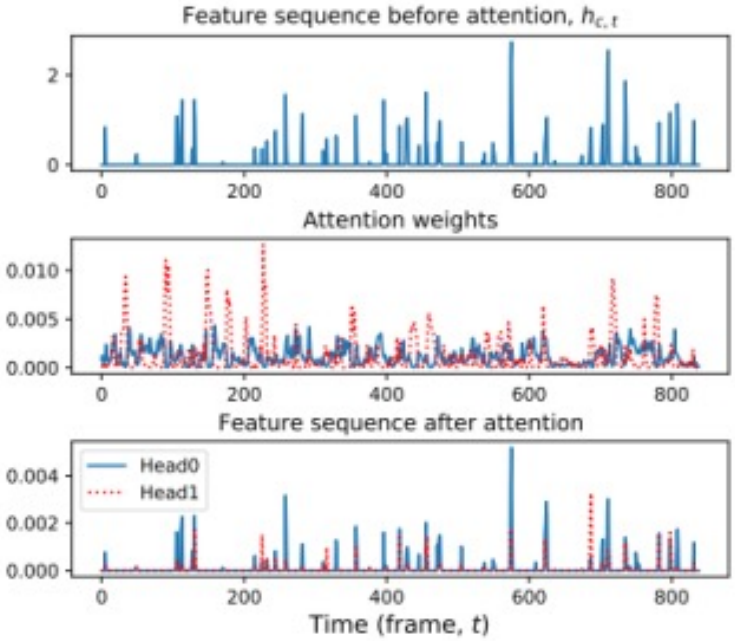
$$M_c^h(k) = \sum_{n=0}^{N-1} \alpha_n^h \left| X_c(n,k) \right|, \quad k = 0, \ldots, L-1$$

$$P_c^h(k) = \sum_{n=0}^{N-1} \alpha_n^h \left| X_c(n,k) \right|^2, \quad k = 0, \ldots, L-1$$

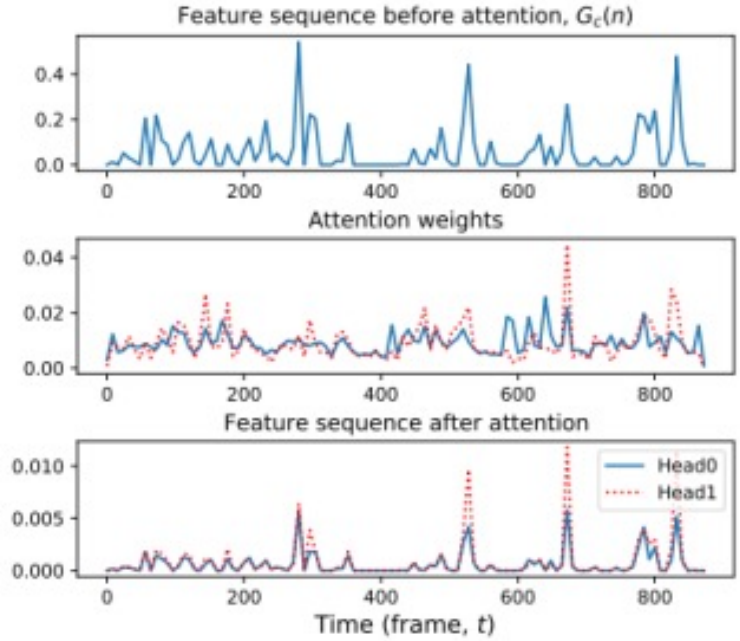


Attentive STSP vs. MHAP

Heads



MHAP with $H = 2$



Attentive STSP with $H = 2$

- In attentive STSP, the attended features by segment-level attention have less variation along the temporal axis than those in frame-level attentive pooling.

Experiments

- **Training data for DNN and PLDA**: 7302 speakers from VoxCeleb1 and VoxCeleb2.
- **Test data**: VOiCES19 evaluation set. VOiCES19 focuses on speaker verification under distracting noise and room reverberation.
- **Acoustic vectors**: 40-dim filter-bank features with mean norm
- **VAD**: Kaldi's energy-based VAD
- **DNN**: We used three models in the experiments, namely, x-vector network, wide x-vector network (channels size are doubled), and Densenet121.