THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Short-time Spectral Aggregation for Speaker Embedding

Youzhi TU and Man-Wai MAK

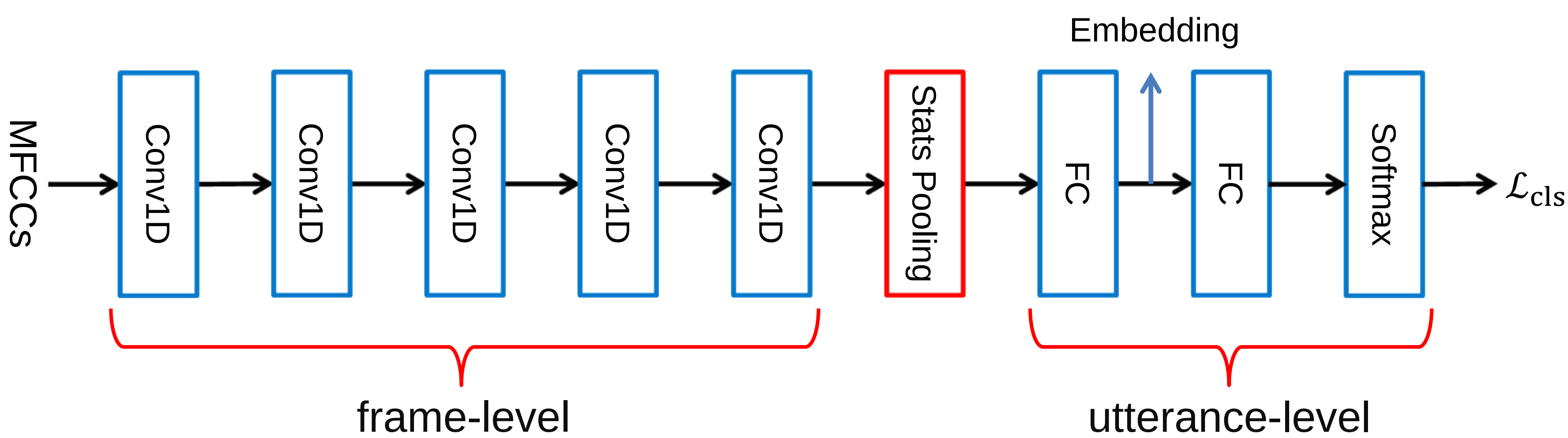
Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong SAR of China

ICASSP’21
6-11 June 2021

Contents

Speaker Embedding Networks

- X-vector extractor is a popular baseline
  - Frame-level layers: Time-delay neural networks (TDNNs), ResNets, DenseNets, Res2Nets, etc.
  - Pooling layer: Aggregate frame-level information
  - Utterance-level layers: Fully-connected (FC) layers

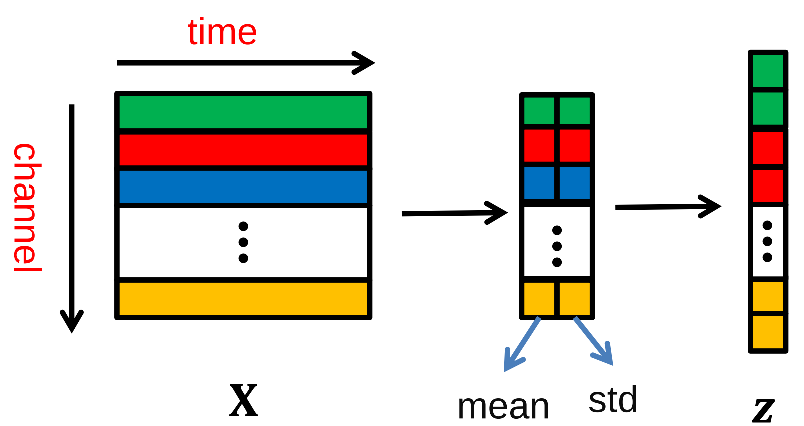


Each layer except the pooling layer is followed by a batch normalization layer and an ReLU layer.

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

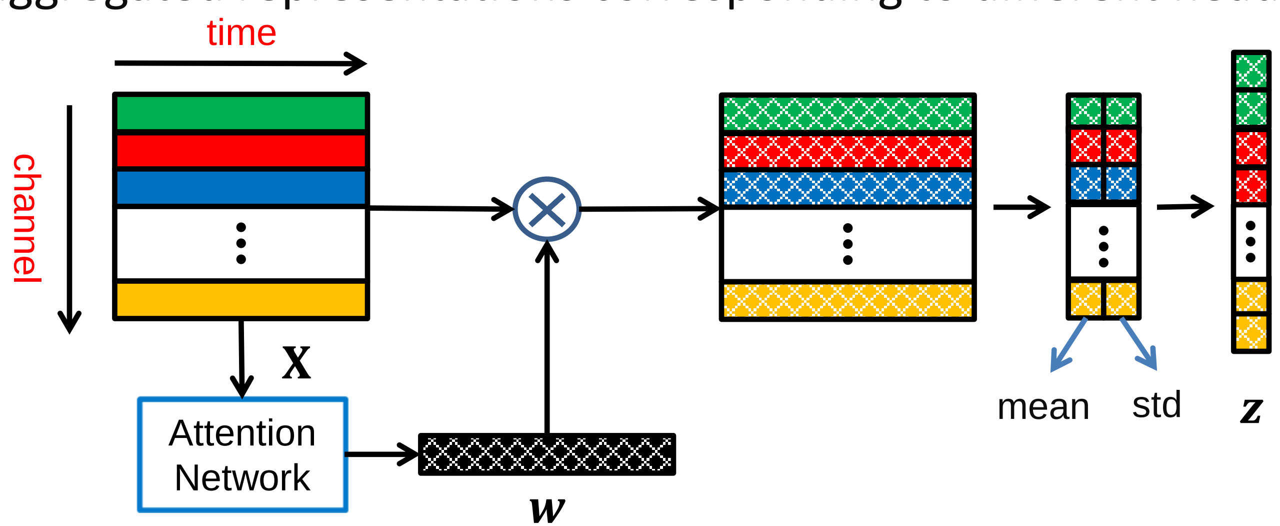
Pooling Methods

- Input: Temporal feature maps $\mathbf{X} \in \mathbb{R}^{C \times T}$ at the output of the last frame-level layer, C and T are the number of channels and frames, respectively
- Output: Aggregated representation $\boldsymbol{z}$ at utterance-level
- Statistics pooling: $\boldsymbol{z}$ is the concatenation of channel-wise mean and standard deviation (std)



Pooling Methods

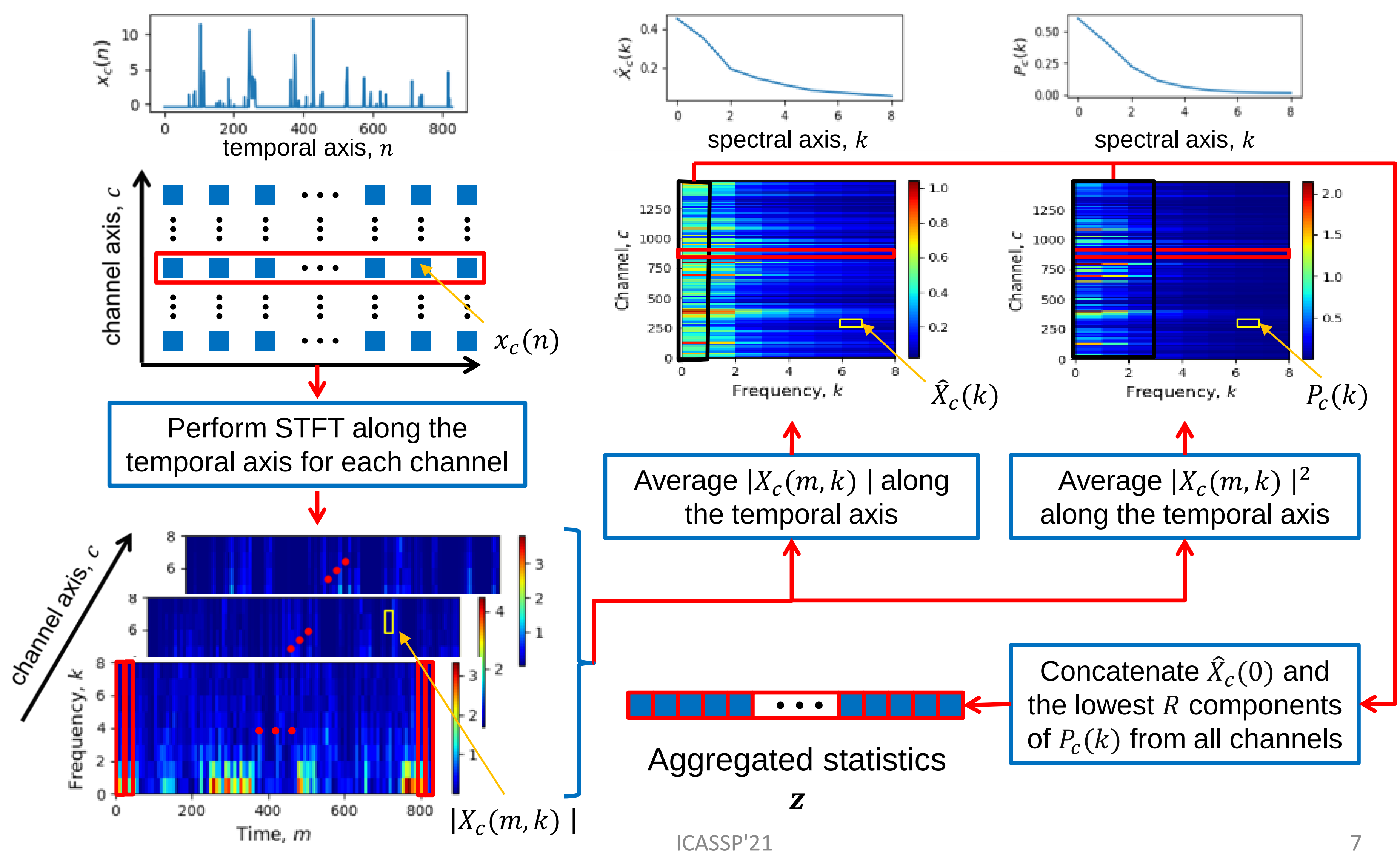
- Attentive pooling (AP): Attend to discriminative frames
  - The aggregated representation $\boldsymbol{z}$ is the concatenation of weighted channel-wise mean and standard deviation
  - The attention weight vector (for a single head) $\boldsymbol{w} \in \mathbb{R}^{1\times T}$ is learned from an attention network and applied to the features of each channel
  - For multi-head attentive pooling, $\boldsymbol{z}$ is the concatenation of the aggregated representations corresponding to different heads



Motivation

- Limitation of statistics pooling
  - Using means and standard deviations is not enough to preserve sufficient speaker information for statistics pooling
  - From a Fourier perspective, statistics pooling only exploits the information in the 0-th frequency component (DC component) in the spectral domain
- Solution
  - Extract multiple spectral components of the spectral representation (besides the DC component) as aggregated embeddings

Short-time Spectral Pooling (STSP)



Relation to Statistics Pooling

- Short-time Fourier transform (STFT) of the c-th channel feature $\mathbf{x}_c = \{x_c(n)\}_{n=0}^{N-1}$ (N is the number of frames)

$$X_c(m,k) = \sum_{n=0}^{N-1} x_c(n)\omega(n-mS)e^{-j\frac{2\pi}{L}kn}, \quad k = [0, L-1]$$

$\omega(\cdot)$: window function, L: STFT length, S: step size of the sliding window

m: index of windowed segments, k: index of spectral components

- When we use $\omega(n) = 1$ (rectangular window) and $S = L = 1$ (the step size and STFT length are both 1), we have

$$\hat{X}_c(0) = 1/M \sum_{m=0}^{M-1} X_c(m,0) = 1/N \sum_{n=0}^{N-1} x_c(n) \triangleq \mathrm{mean}(\mathbf{x}_c),$$

$$P_c(0) = 1/M \sum_{m=0}^{M-1} |X_c(m,0)|^2 = 1/N \sum_{n=0}^{N-1} [x_c(n)]^2 \triangleq \mathrm{var}(\mathbf{x}_c) + [\mathrm{mean}(\mathbf{x}_c)]^2.$$

- Under above conditions, using means and stds for statistics pooling is an analogy to using the DC components $\hat{X}_c(0)$ and $P_c(0)$ for STSP
- Because STSP uses more frequency components of $P_c(k)$ ($k > 0$) for aggregation, it can preserve more information than statistics pooling

Experiments

- Compare statistics pooling, attentive pooling and STSP on VoxCeleb1-test, VOiCES19-dev and VOiCES19-eval
- Speaker embedding network training
  - 40-dimensional filter bank features
  - VoxCeleb1&2-dev for VOiCES19 (2,105,949 utterances from 7,185 speakers) and VoxCeleb1-dev for VoxCeleb1 (2,092,009 utterances from 5,984 speakers)
  - Baseline: Standard x-vector network
  - Attention network: FC (500) + ReLU + $\mathrm{FC}\ (H)$, H is the number of heads
  - STSP: Rectangular window function, STFT length and window step size were 16

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

Experiments

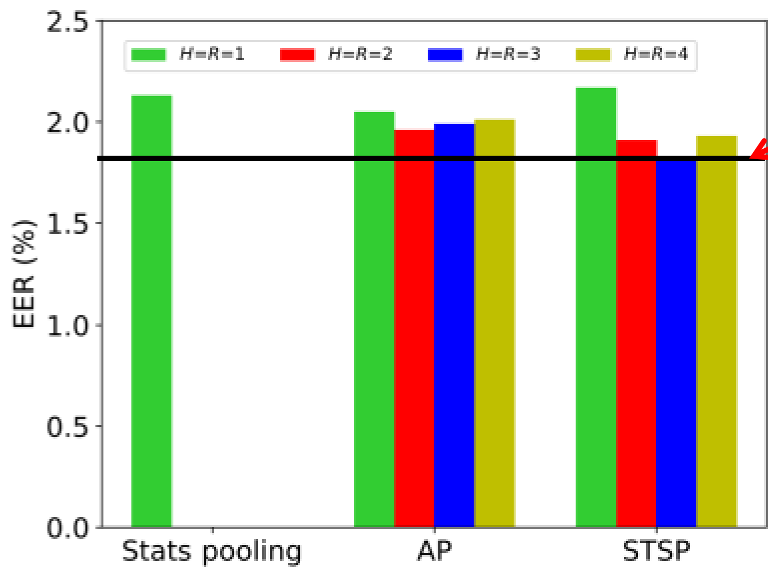
- PLDA training
  - VoxCeleb1: Clean VoxCeleb1-dev (1,240,651 utterances)
  - VOiCES19: Concatenated speech with the same video session augmented with reverberation and noise (334,776 utterances)
  - Pre-processing: Center + LDA (200 for Voxceleb1 and 150 for VOiCES19) + whitening + length normalization
- Score normalization (only for VOiCES19)
  - Cohort: Longest two utterances of each speaker in the PLDA training data

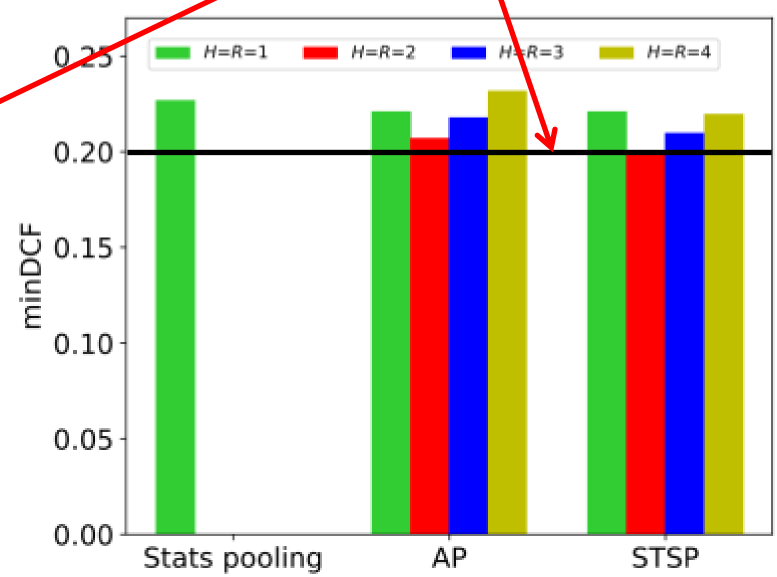
Results on Voxceleb1-test

H: Number of heads in attentive pooling

R: Number of spectral components of $P_c(k)$ in STSP

	Stats pooling	AP (H=1)	AP (H=2)	AP (H=3)	AP (H=4)	STSP (R=1)	STSP (R=2)	STSP (R=3)	STSP (R=4)
EER	2.13	2.05	**1.96**	1.99	2.01	2.17	1.91	**1.82**	1.93
minDCF	0.227	0.221	**0.207**	0.218	0.232	0.221	**0.199**	0.210	0.220



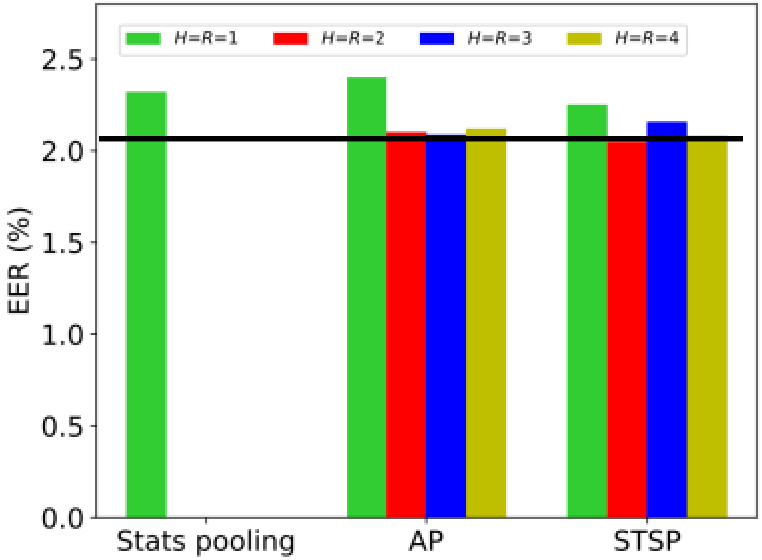


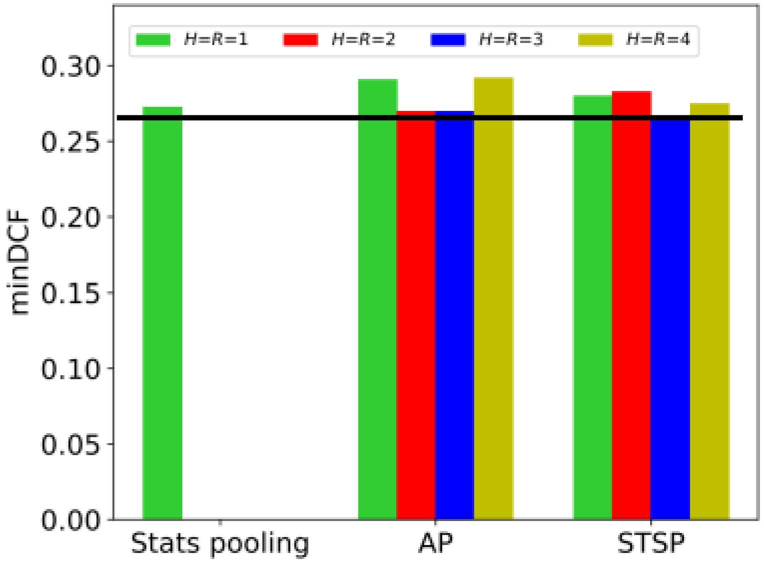
Results on VOiCES19-dev

H: Number of heads in attentive pooling

R: Number of spectral components of $P_c(k)$ in STSP

	Stats pooling	AP (H=1)	AP (H=2)	AP (H=3)	AP (H=4)	STSP (R=1)	STSP (R=2)	STSP (R=3)	STSP (R=4)
EER	2.32	2.40	2.10	**2.09**	2.12	2.25	**2.05**	2.16	2.08
minDCF	0.273	0.291	**0.270**	**0.270**	0.292	0.280	0.283	**0.266**	0.275





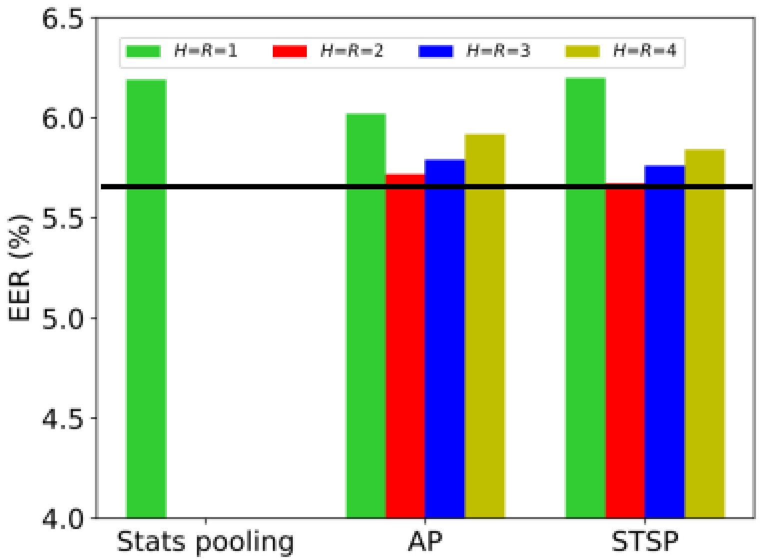
THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

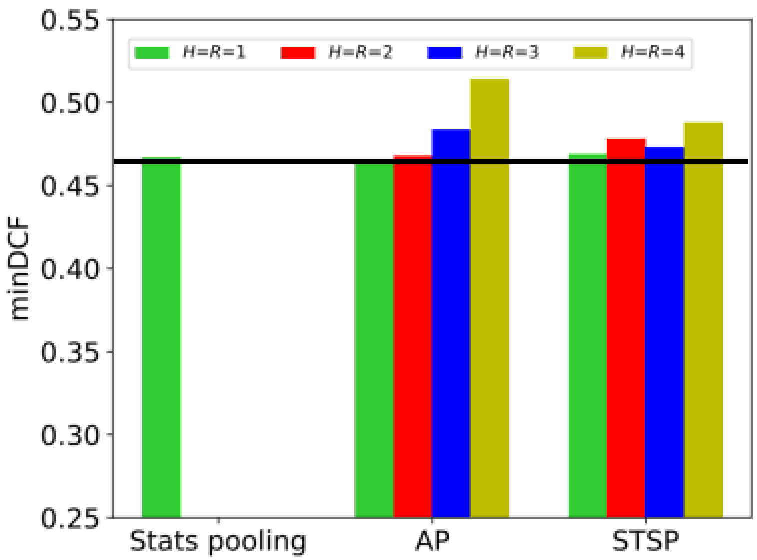
Results on VOiCES19-eval

H: Number of heads in attentive pooling

R: Number of spectral components of $P_c(k)$ in STSP

	Stats pooling	AP (H=1)	AP (H=2)	AP (H=3)	AP (H=4)	STSP (R=1)	STSP (R=2)	STSP (R=3)	STSP (R=4)
EER	6.19	6.02	**5.72**	5.79	5.92	6.20	**5.67**	5.76	5.84
minDCF	0.467	**0.465**	0.468	0.484	0.514	**0.469**	0.478	0.473	0.488





Conclusions

- Proposed a new pooling method for speaker embedding from a Fourier perspective
- STSP is able to aggregate the information in higher frequency components (besides the DC component), making it preserve more speaker information than statistics pooling
- Generally, STSP outperforms attentive pooling and statistics pooling on Voxceleb1 and VOiCES19

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

References

1. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in Proc. International Conference on Acoustics, Speech, and Signal Processing, 2018, pp. 5329–5333.
2. Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, “Self-attentive speaker embeddings for text-independent speaker verification,” in Proc. Annual Conference of the International Speech Communication Association, 2018, pp. 3573–3577.
3. O. Rippel, J. Snoek, and R. P. Adams, “Spectral representations for convolutional neural networks,” in Advances in Neural Information Processing Systems, 2015, pp. 2449–2457.

Thank you!