HI-MIA声纹识别实战

第3节- 模型实现

讲师：覃晓逸

课程目录：

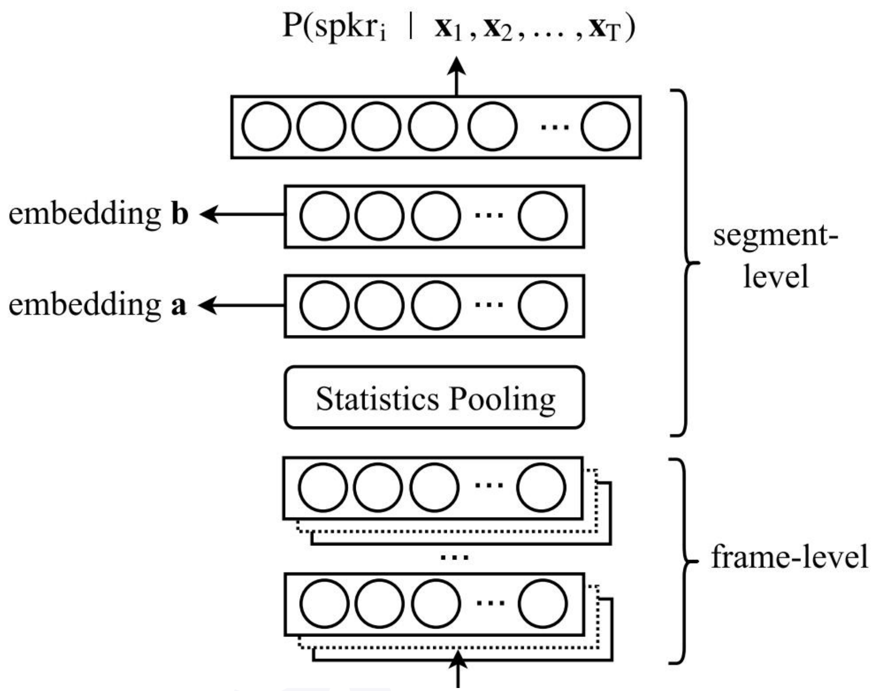
1 前端建模的实现

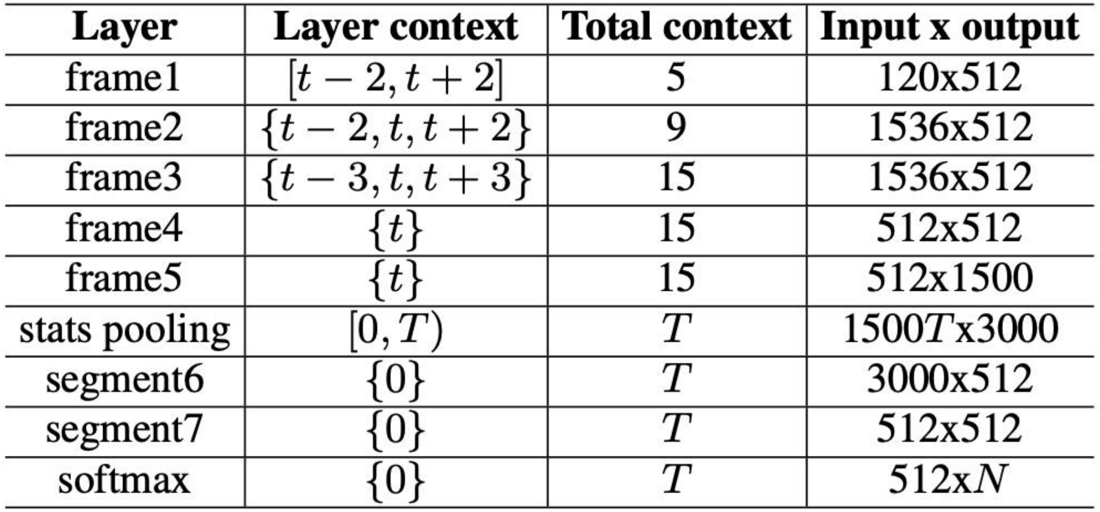
2 编码层的实现

3 分类器的实现

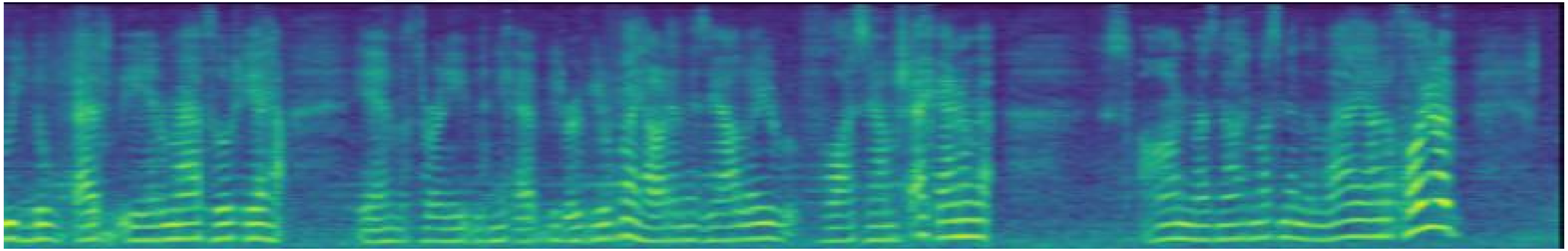
4 总结

1 X-Vector的实现


$P(\mathrm{spkr}_i \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$
embedding **b**
embedding **a**
Statistics Pooling
segment-level
frame-level



Layer	**Layer context**	**Total context**	**Input x output**
frame1	$[t-2, t+2]$	5	120x512
frame2	$\{t-2, t, t+2\}$	9	1536x512
frame3	$\{t-3, t, t+3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	1500Tx3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512xN



24 x 100

1 X-Vector的实现

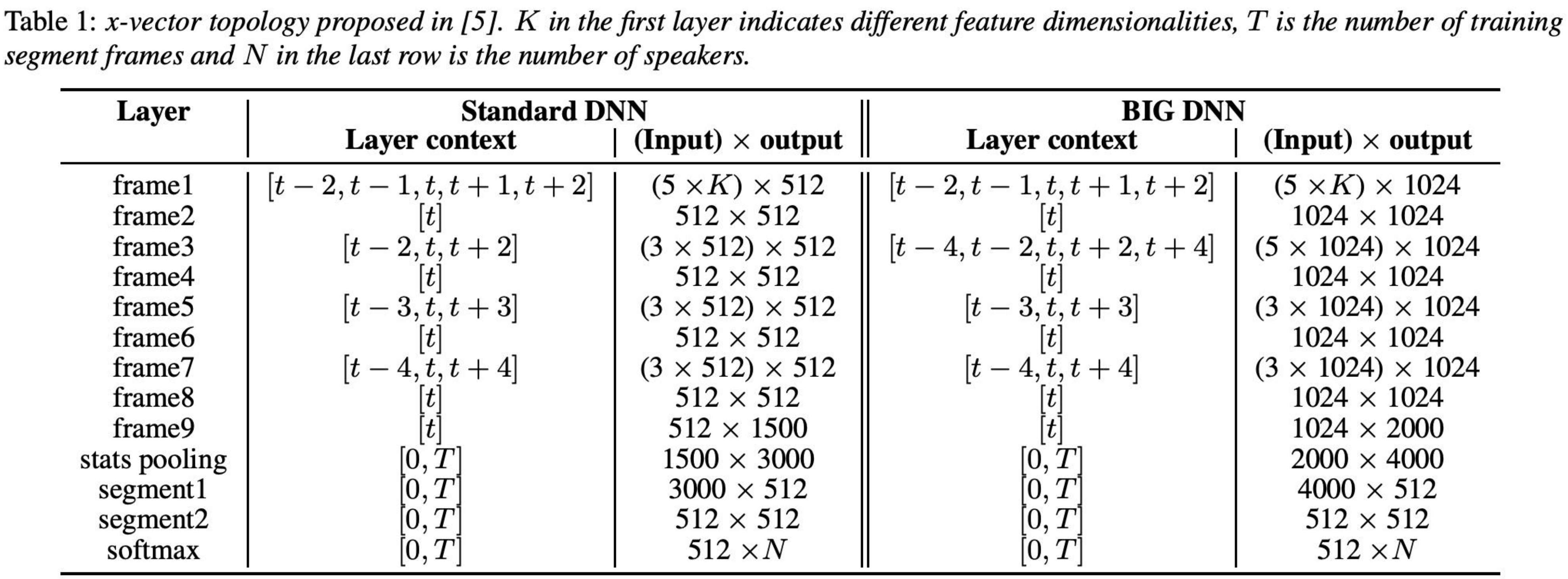
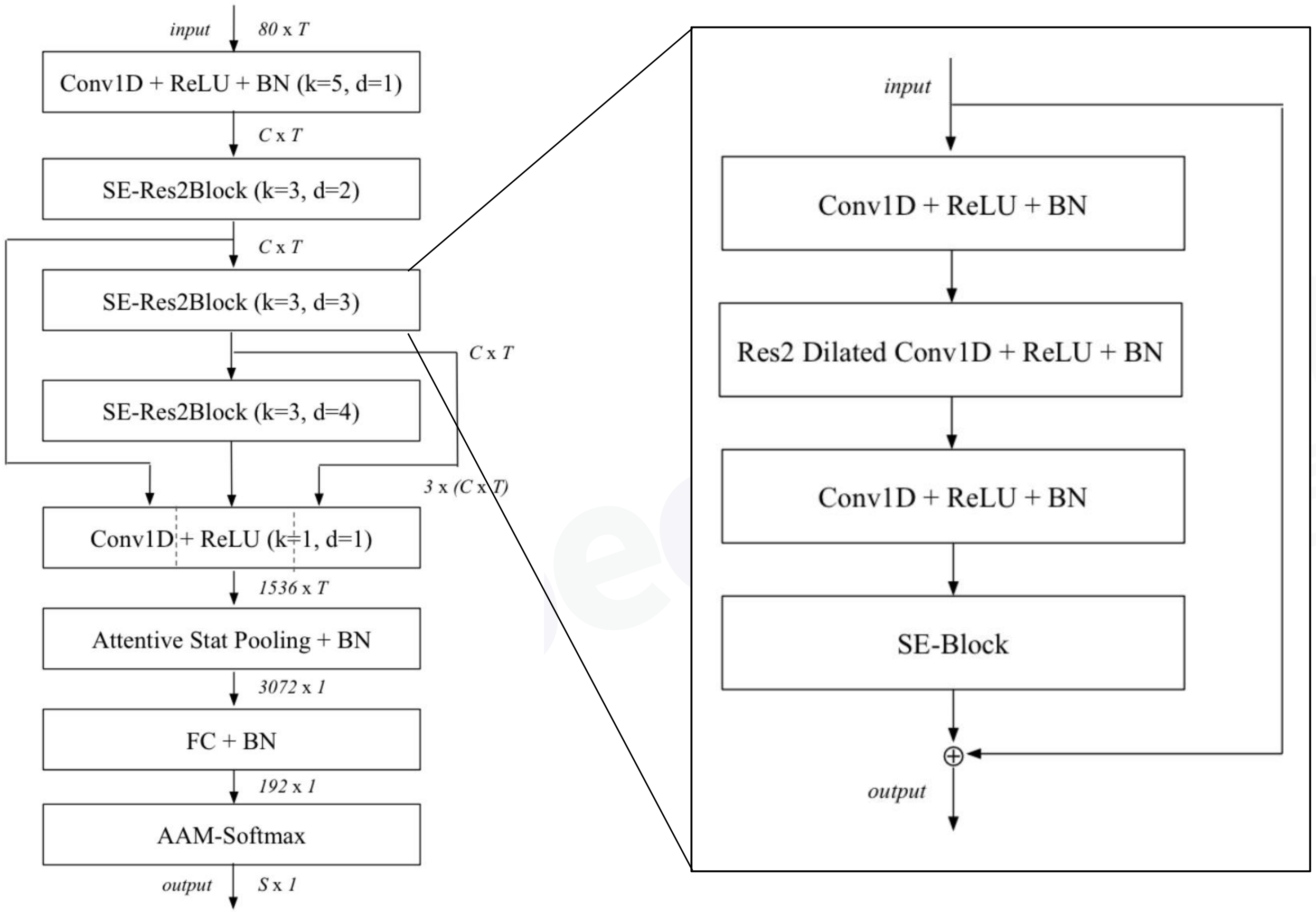


Table 1: *x-vector topology proposed in [5]. K in the first layer indicates different feature dimensionalities, T is the number of training segment frames and N in the last row is the number of speakers.*

Layer	**Standard DNN**		**BIG DNN**	
	Layer context	**(Input) × output**	**Layer context**	**(Input) × output**
frame1	$[t-2, t-1, t, t+1, t+2]$	$(5 \times K) \times 512$	$[t-2, t-1, t, t+1, t+2]$	$(5 \times K) \times 1024$
frame2	$[t]$	512×512	$[t]$	1024×1024
frame3	$[t-2, t, t+2]$	$(3 \times 512) \times 512$	$[t-4, t-2, t, t+2, t+4]$	$(5 \times 1024) \times 1024$
frame4	$[t]$	512×512	$[t]$	1024×1024
frame5	$[t-3, t, t+3]$	$(3 \times 512) \times 512$	$[t-3, t, t+3]$	$(3 \times 1024) \times 1024$
frame6	$[t]$	512×512	$[t]$	1024×1024
frame7	$[t-4, t, t+4]$	$(3 \times 512) \times 512$	$[t-4, t, t+4]$	$(3 \times 1024) \times 1024$
frame8	$[t]$	512×512	$[t]$	1024×1024
frame9	$[t]$	512×1500	$[t]$	1024×2000
stats pooling	$[0, T]$	1500×3000	$[0, T]$	2000×4000
segment1	$[0, T]$	3000×512	$[0, T]$	4000×512
segment2	$[0, T]$	512×512	$[0, T]$	512×512
softmax	$[0, T]$	$512 \times N$	$[0, T]$	$512 \times N$

[1] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, Oldřich Plchot, ”BUT System Description to VoxCeleb Speaker Recognition Challenge 2019”
”

1 X-Vector的实现



[2] **Brecht Desplanques, Jenthe Thienpondt, Kris Demuynck, ” ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification”**

4 总结

前端模型：

TDNN， ResNet， SE-ResNet， ECAPA-TDNN（作业）

编码层：

StatsPool， ASP(作业)

分类器：

Softmax， AAMSoftmax

课程问题可随时联系班主任