AISHELL Speaker Verification Challenge 2019

参赛队伍：小米AI实验室

队员：蔡国都、庄伟基、王欣、杨朔、李文凤

Contents

Data Preparation

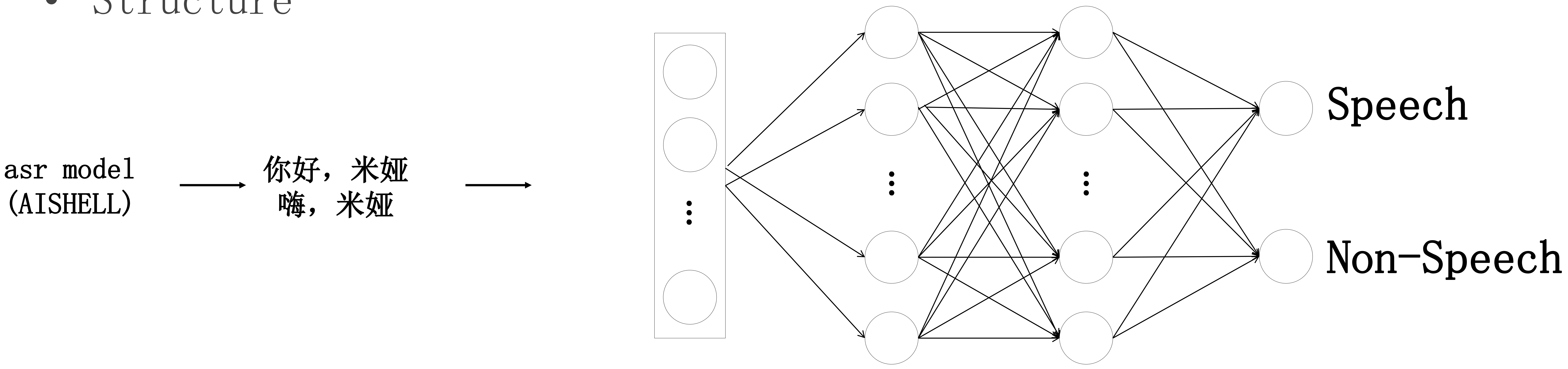
Data Preparation

- Training data
  - AISHELL-WakeUp-1: Train、Dev
  - Openslr: SLR17(MUSAN)、SLR28(RIR_NOISES)、SLR33(AISHELL)
- Downsampling
  - Hi-Fi data: 44.1kHz ⟶ 16kHz
- Data augmentation
  - MUSAN: music、babble、noise[1]
  - Volume、tempo
  - RIRS/Reverberation[2]
  - Frequency masking[3]
  - Total data: 99w ⟶ 200w

[1] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," arXiv:1510.08484 [cs], 2015.
[2] Ko, Tom , et al. "A study on data augmentation of reverberant speech for robust speech recognition." ICASSP 2017 - 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE, 2017.
[3] Daniel S. Park, et al. "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition." arXiv:1904.08779 [cs], 2019.

Data Preparation

- KW-VAD
  - Purpose
    - Energy-based kaldi vad is not perform well.
    - Train the acoustic model(aishell-1) to obtain alignment information for KW-VAD training
  - Structure



Data Preparation

- fast exp set:254 * (100/hifi,400/far)

Methods	Dev-EER	Description
baseline1	3.03	no-vad/volume/musan /reverb/tempo
Kaldi-vad	2.802	√
kw-vad	2.721	√
volume	2.845	√
volume normalization	3.056	×

Data Preparation

Methods	Dev-EER	Description
baseline2	3.242	kw-vad/no-aug
musan_noise	3.172	√
musan_music	3.171	√
musan_speech	2.94	√
Kaldi-reverb	3.171	√
sox-reverb	3.146	√
tempo	3.048	√
Frequency masking	2.86	√
fusion	2.674	√

Experimental Environment

Experimental Environment

- SGE
  - CentOS/Ubuntu
  - CPU：640 cores
  - GPU：**8*8 Tesla V100** ，vRAM:16G
- Tools
  - **Kaldi** [4]
    - wav-reverberate
    - BeamformIt
    - egs/aishell
    - sox
  - **Pytorch**
    - resnet-50

[4] Povey, Daniel, et al. "The Kaldi speech recognition toolkit." IEEE 2011 workshop on automatic speech recognition and understanding. No. CONF. IEEE Signal Processing Society, 2011.

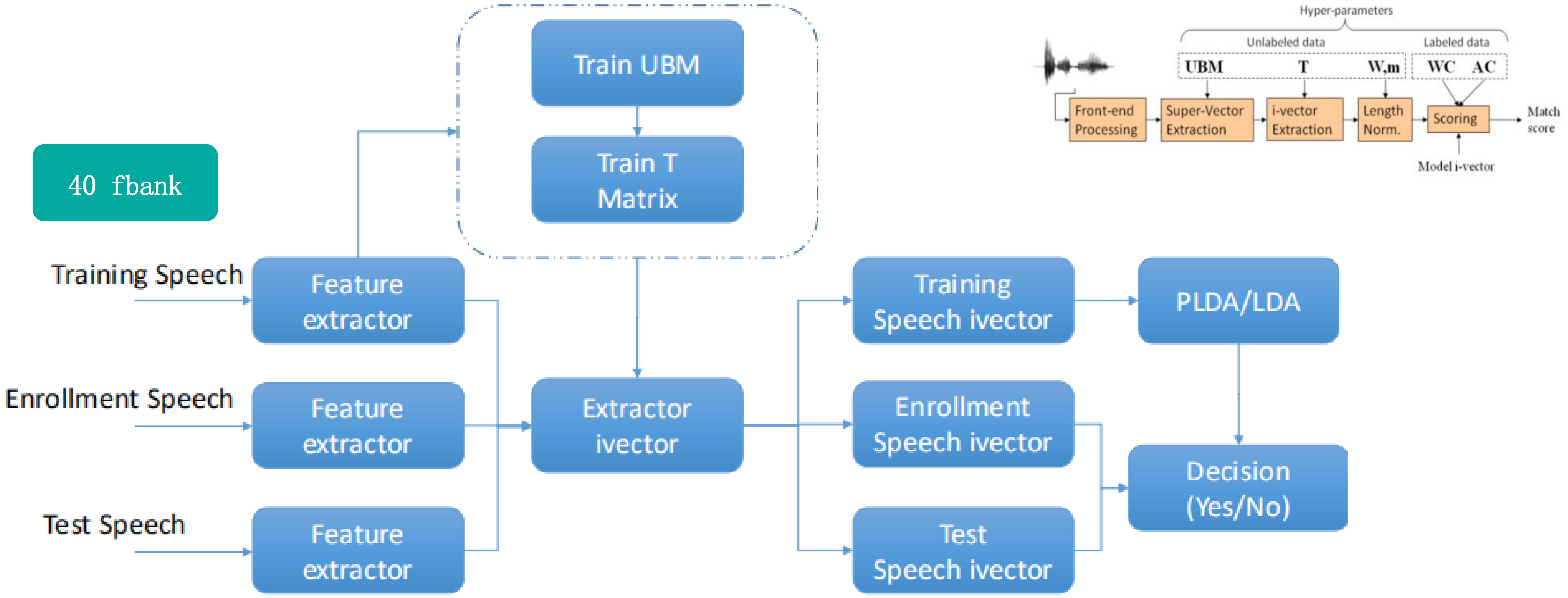
Model Training

Model Training

- i-vector [5]
- x-vector + cnn + self-attention [6][7]
- resnet + AM-Softmax [8][9][10]

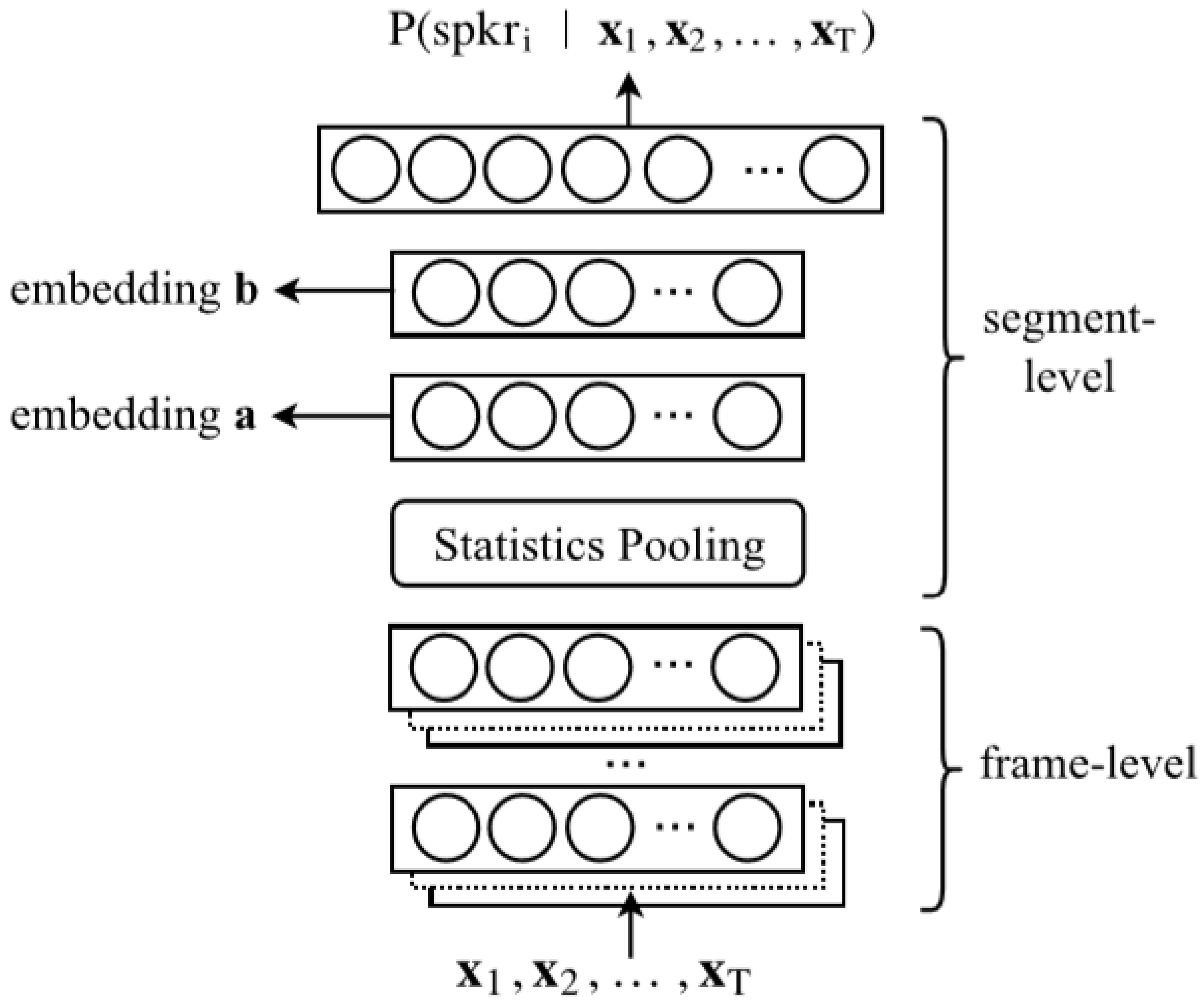
[5] Dehak, N. , et al. "Front-End Factor Analysis for Speaker Verification." IEEE Transactions on Audio, Speech and Language Processing 19.4(2011):788-798.
[6] David Snyder et al. "X-vectors: Robust DNN Embeddings for Speaker Recognition". In: Proc. of ICASSP. IEEE.2018, pp. 5329-5333.
[7] Zhu, Yingke et al. "Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification." Interspeech.2018.
[8] He, Kaiming , et al. "Deep Residual Learning for Image Recognition." (2015).
[9] Wang, Feng, et al. "Additive margin softmax for face verification." IEEE Signal Processing Letters 25.7 (2018): 926-930.
[10] Chung, Joon Son , A. Nagrani , and A. Zisserman . "VoxCeleb2: Deep Speaker Recognition." (2018).

i-vector



[5] Dehak, N. , et al. "Front-End Factor Analysis for Speaker Verification." IEEE Transactions on Audio, Speech and Language Processing 19.4(2011):788-798.

x-vector

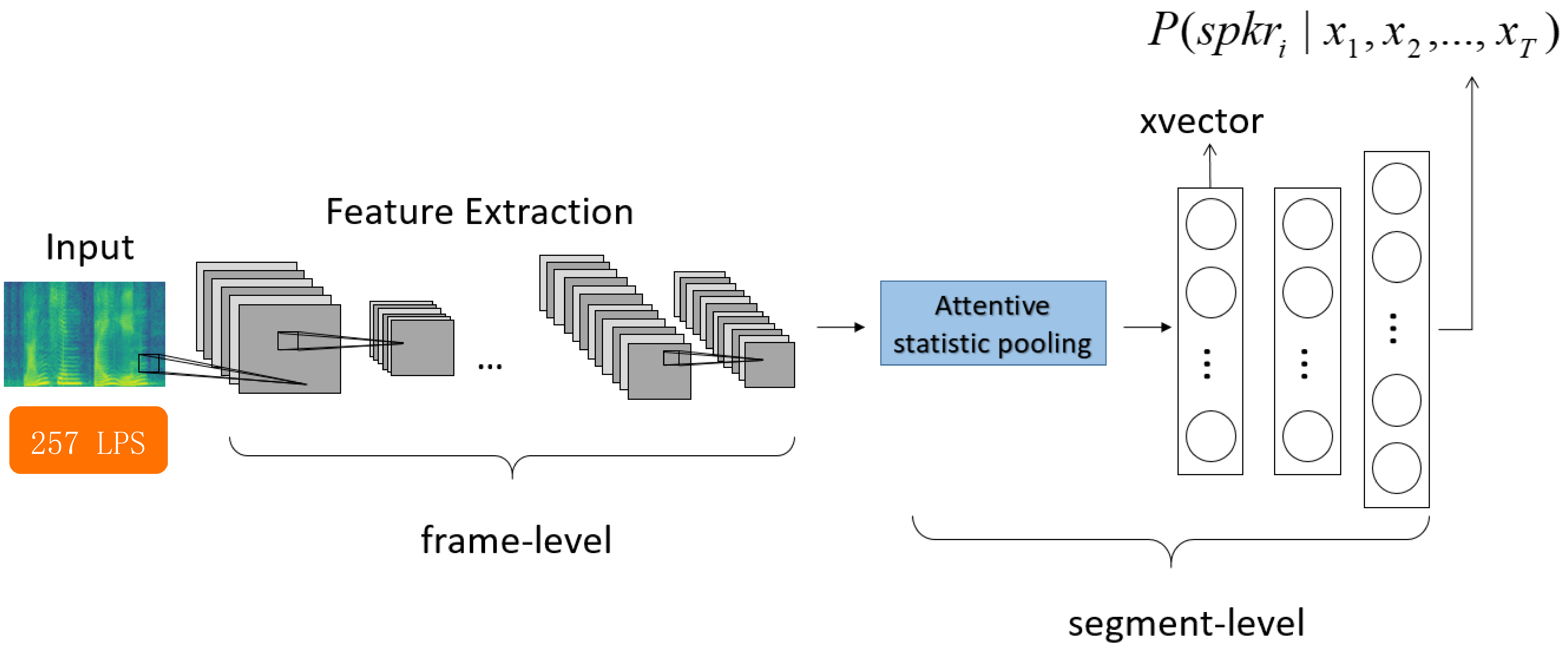


Layer	**Layer context**	**Total context**	**Input x output**
frame1	$[t-2, t+2]$	5	120x512
frame2	$\{t-2, t, t+2\}$	9	1536x512
frame3	$\{t-3, t, t+3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	1500Tx3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512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

Table 1. The embedding DNN architecture. x-vectors are extracted at layer *segment6*, before the nonlinearity. The N in the softmax layer corresponds to the number of training speakers.

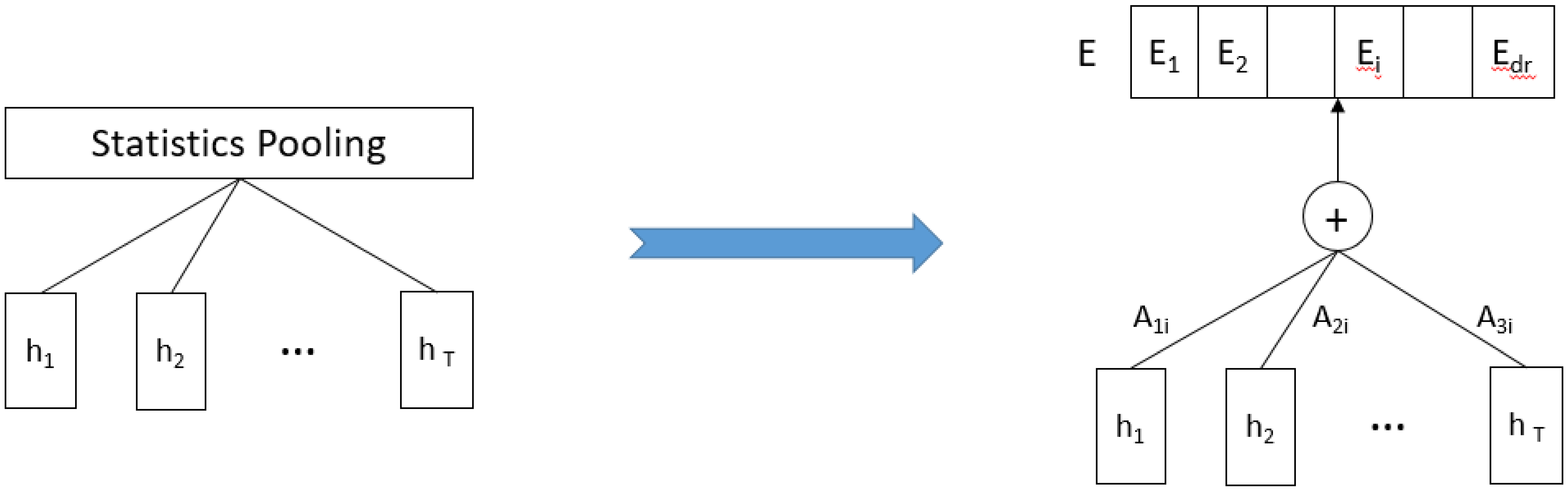
[6] David Snyder et al. “X-vectors: Robust DNN Embeddings for Speaker Recognition”. In: Proc. of ICASSP. IEEE. 2018, pp. 5329-5333.

x-vector+cnn+self-attention



[6] David Snyder et al. “X-vectors: Robust DNN Embeddings for Speaker Recognition”. In: Proc. of ICASSP. IEEE. 2018, pp. 5329-5333.
[7] Zhu, Yingke et al. “Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification.” Interspeech. 2018

x-vector+cnn+self-attention



$$A = \mathrm{softmax}(g(H^T W_1) W_2)$$

$$E = HA$$

[6] David Snyder et al. “X-vectors: Robust DNN Embeddings for Speaker Recognition”. In: Proc. of ICASSP. IEEE.2018, pp. 5329-5333.
[7] Zhu, Yingke et al. “Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification.” Interspeech.2018

x-vector+cnn+self-attention

Layer	Kernel size	Channels	Height offsets	Time offsets
conv1	3*3	64	{h-1, h, h+1}	{t-1, t, t+1}
conv2	3*3	128	{h-1, h, h+1}	{t-1, t, t+1}
conv3	3*3	128	{h-1, h, h+1}	{t-1, t, t+1}
conv4	3*3	128	{h-1, h, h+1}	{t-2, t, t+2}
conv5	3*3	64	{h-1, h, h+1}	{t-2, t, t+2}

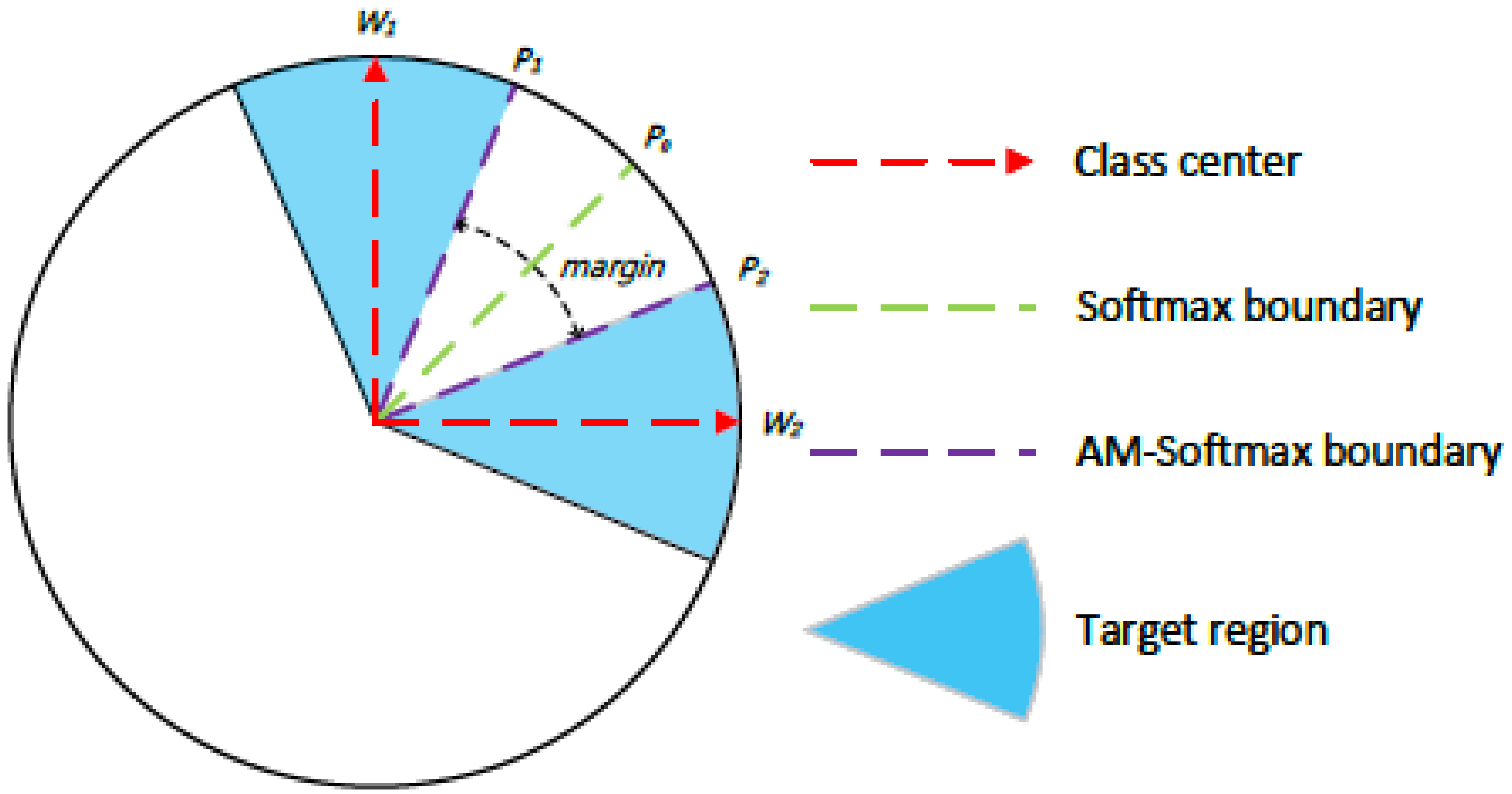
resnet50+AM-Softmax

layer name	resnet-50
conv1 pool1	7*7, 64, stride 2 3*3, max pool, stride 2
conv2_x	[1*1, 64 3*3, 64 1*1, 256] * 3, stride 2
conv3_x	[1*1, 128 3*3, 128 1*1, 512] * 4, stride 2
conv4_x	[1*1, 256 3*3, 256 1*1, 1024] * 6, stride 2
conv5_x	[1*1, 512 3*3, 512 1*1, 2048] * 3, stride 2
fc1 avg_pool fc2	9*1, 2048, stride 1 1*N, average pool, stride 1 1 * 1, 254

[8] He, Kaiming , et al. "Deep Residual Learning for Image Recognition." (2015).
[9] Wang, Feng, et al. "Additive margin softmax for face verification." IEEE Signal Processing Letters 25.7 (2018): 926-930.
[10] Chung, Joon Son , A. Nagrani , and A. Zisserman . "VoxCeleb2: Deep Speaker Recognition." (2018).

resnet50+AM-Softmax

$$\mathcal{L}_{AMS} = -\frac{1}{n}\sum_{i=1}^{n} log \frac{e^{s\cdot(cos\theta_{y_i}-m)}}{e^{s\cdot(cos\theta_{y_i}-m)} + \sum_{j=1,j\neq y_i}^{c} e^{s\cdot cos\theta_j}}$$
$$= -\frac{1}{n}\sum_{i=1}^{n} log \frac{e^{s\cdot(W_{y_i}^T \boldsymbol{f}_i - m)}}{e^{s\cdot(W_{y_i}^T \boldsymbol{f}_i - m)} + \sum_{j=1,j\neq y_i}^{c} e^{sW_j^T \boldsymbol{f}_i}}.$$



more discriminative embeddings

[9] Wang, Feng, et al. "Additive margin softmax for face verification." IEEE Signal Processing Letters 25.7 (2018): 926-930.

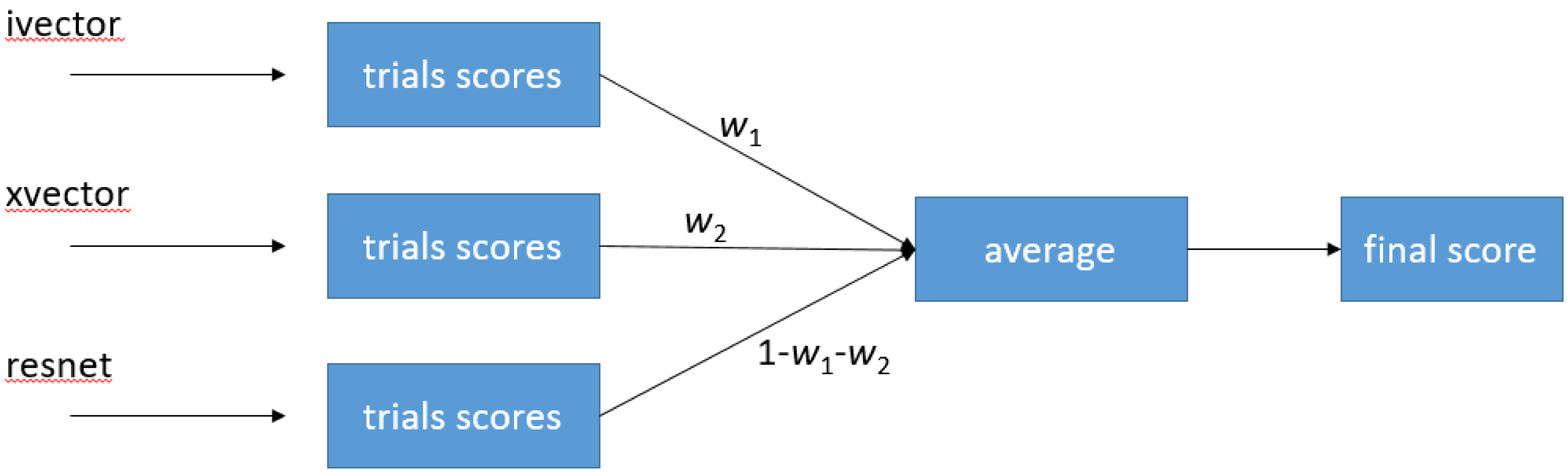
Scoring Strategy

Scoring Strategy

- PLDA [11]

$$score = \log \frac{p(\eta_1, \eta_2 \mid H_s)}{p(\eta_1 \mid H_d) p(\eta_2 \mid H_d)}$$

- Fusion



[11] Garcia-Romero, Daniel, and Carol Y. Espy-Wilson. "Analysis of i-vector length normalization in speaker recognition systems." Twelfth Annual Conference of the International Speech Communication Association. 2011.

Scoring Strategy

- Skill
  - Track1
    - Enroll：
      - downsample、noise
      - Calculate the mean of embedding for all sentences after augmentation as the vector of registered sentences
    - Eval：
      - Use 16 channels of data
      - Calculate the mean of embedding for all sentences as the vector of test sentences
  - Track2
    - Enroll/Eval:
      - Use 16 channels of data
      - Calculate the mean of embedding for all sentences as the vector of test sentences

Scoring Strategy

- Skill
  - Whitening
    - Track1: keep the same amount of Hi-Fi data and far field data.
    - Track2: use all far field data.
  - PLDA
    - All training data
    - All far field data
    - Keep the same amount of Hi-Fi data and far field data
  - Other methods:
    - Beamforming: processing the multi channel.

Results on dev

train_99w	spectrogram kw-vad cnn + x-vector + attention	PLDA	2.47	1.83
		PLDA enroll_aug	1.74	1.83
train_200w	[1]spectrogram kw-vad cnn + x-vector + attention	PLDA	1.92	1.58
		PLDA-20W	1.77	1.32
		PLDA-20W enroll_aug	**1.41**	**1.32**
train_200w	[2]spectrogram kw-vad resnet+AM-Softmax	PLDA-20W enroll_aug	**1.48**	**1.23**
train_200w	fbank kw-vad i-vector	PLDA	2.52	2.21
train_200w	[3]fbank kw-vad frequency masking i-vector	PLDA-20W enroll_aug	**2.15**	**1.87**
-	fusion[1,2,3]	weighted average	**1.33**	**1.14**

(enroll: 3 normal utts)

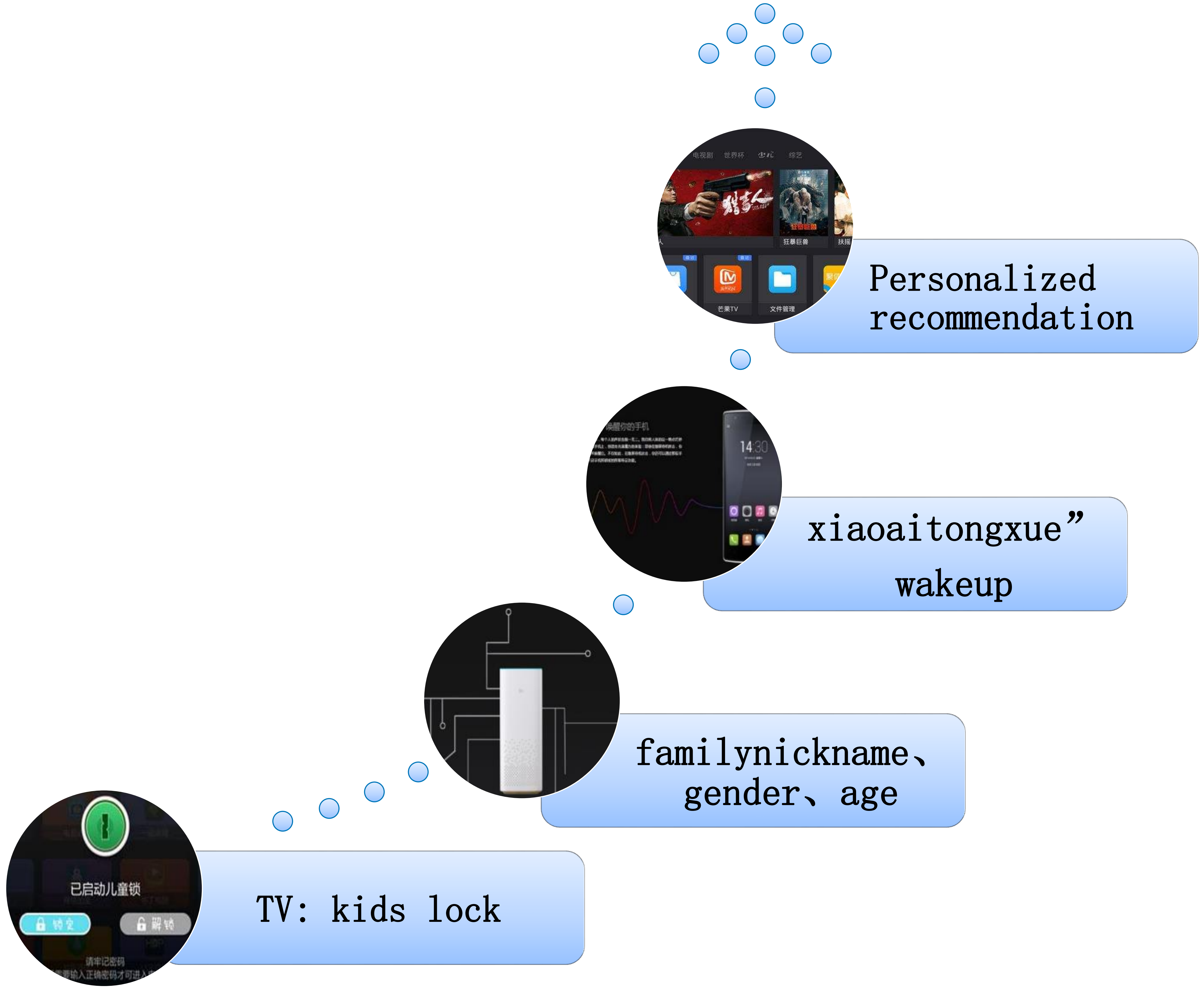
Summary

Summary

- Innovation
  - More accurate **kw-vad**
  - Rich **data augmentation** strategy
  - **Multiple-model fusion**
  - Scoring strategy
- Advantage
  - Powerful **Kaldi-based SGE queue**
  - More relevant to the Xiaomi product business
    - Mobile phone, AISoundbox, TV, etc.

Application scene in xiaomi

- phone
  - “xiaoaitongxue” wakeup
- AIoT
  - TV: kids lock
  - AISoundbox: nick name
  - gender、age
  - family members
  - Personalized recommendation



Welcome to Xiaomi

Thanks!