MULTI-RESOLUTION MULTI-HEAD ATTENTION IN DEEP SPEAKER EMBEDDING

Zhiming Wang, Kaisheng Yao, Xiaolong Li, Shuo Fang
Ant Financial Services Group, Hangzhou, China

Outlines

- **Deep Speaker Embedding Framework**
- Pooling: an Overview
- the Proposed Pooling Methods
- Experiments and Results
- Conclusions

Deep Speaker Embedding Framework

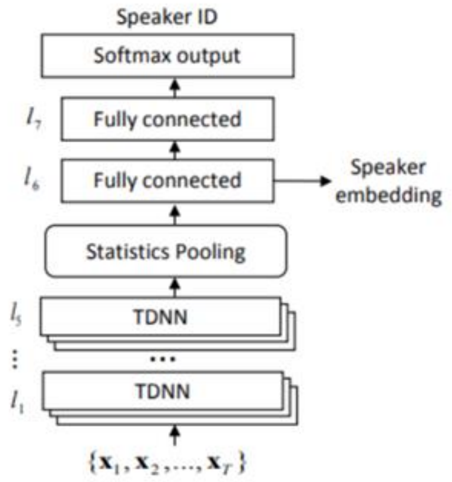
➢Deep Speaker Embedding(x-vector)

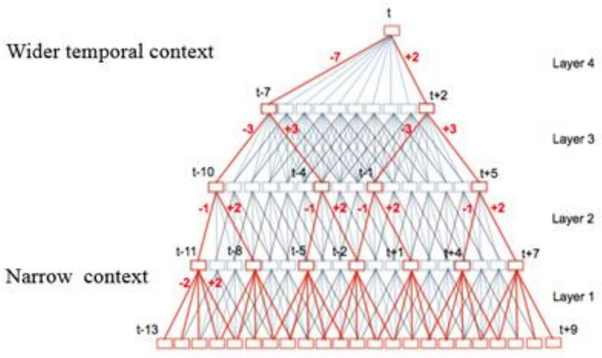
✓**DNN**: cepstral acoustic features → a sequence of encoded vector

✓**a pooling layer**: a segment-level representation(embedding); ***today's topic***

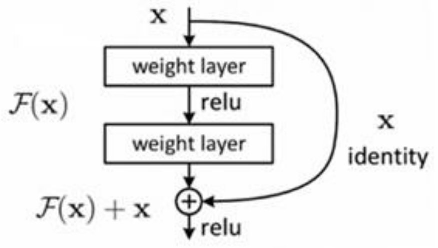
✓**a classifier**(softmax or fully connected network): project to **speaker ids**

✓**outputs at reciprocal a certain layer**： embedding feature





TDNN([3] and [4])



ResNet([17])

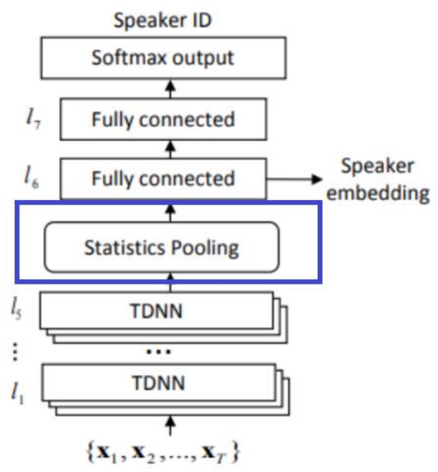
Outline

- Deep Speaker Embedding Framework
- **Pooling: an Overview**
- the Proposed Pooling Methods
- Experiments and Results
- Conclusions

Pooling(1): an Overview

- ✓Statistics pooling
- ✓Self attentive pooling
- ✓Multi-head attentive pooling(for ***increased discriminative information***)
- ✓Multi-resolution multi-head attentive pooling(for ***encouraging diversity from multiple heads***)

Pooling is Essential



Pooling(2): an Overview

➢**Statistics pooling**

✓mean + std: capture *overall information* and *dynamical variability*

➢**Self attentive pooling**

✓compute importance of each frame

$$\alpha_t = \frac{e^{(s_d(\mathbf{h}_t))}}{\sum_{i=1}^{N} e^{(s_d(\mathbf{h}_j))}}$$

$$\mathbf{e} = \sum_{t=1}^{N} \alpha_t \mathbf{h}_t$$

$$s_l^{(i)}(x) = \mathbf{V}_i^T f(\mathbf{W}_i x + \mathbf{g}_i) + b_i, \quad \textbf{MLP}$$

where $f(\cdot)$ is a non-linear activation function, $\mathbf{W}_i \in \Re^{l\times l}$, $\mathbf{g}_i \in \Re^l$, $\mathbf{V}_i \in \Re^l$ and $b_i \in \Re$ are parameters to learn.

➢**Attentive statistics pooling**

✓mean + std: using self attentive weigh α_t , NOT in average with α_t = 1/N

Pooling(3): an Overview

➢Multi-head attentive pooling([1])

✓split the encoded frame into non-overlapping homogeneous sub-vectors

✓apply attentive pooling on frame sequence of sub-vectors

✓may ignore possible correlations among different sub-vectors, especially for $\mathbf{h}_t$ with small dimensional size

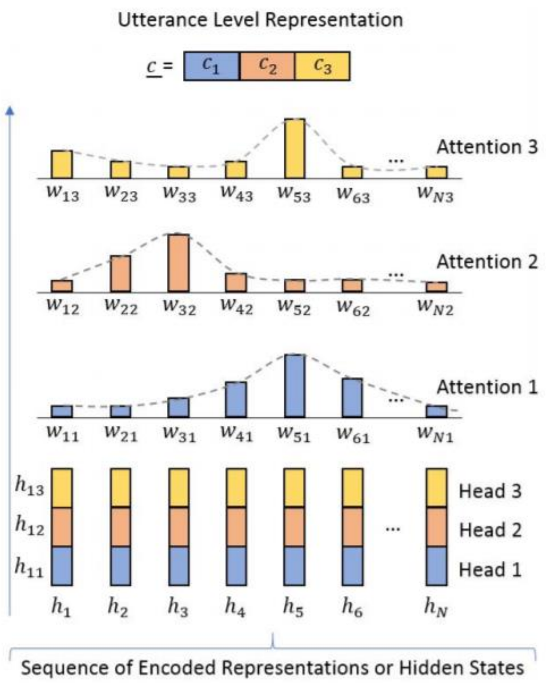
$$\mathbf{h}_t = [\mathbf{h}_t^{(1)}, \mathbf{h}_t^{(2)}, \cdots, \mathbf{h}_t^{(K)}]$$

$$\alpha_t^{(i)} = \frac{e^{(s_{d/K}^{(i)}(\mathbf{h}_t^{(i)}))}}{\sum_{j=1}^{N} e^{(s_{d/K}^{(i)}(\mathbf{h}_j^{(i)}))}}$$

$$s_l^{(i)}(x) = \mathbf{V}_i^T x,$$
$$\mathbf{V}_i \in \Re^l$$

$$\mathbf{e}^{(i)} = \sum_{t=1}^{N} \alpha_t^{(i)} \mathbf{h}_t^{(i)}$$

$$\mathbf{e} = [\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \cdots, \mathbf{e}^{(K)}]$$



Outline

- the Framework of Deep Speaker Embedding
- Pooling: an Overview
- **The Proposed Pooling Methods**
- Experiments and Results
- Conclusions

The Proposed Pooling Methods(1)

- Global multi-head attentive pooling
  - Apply K-head attention over the entire encoded sequence

$$s_l^{(i)}(x) = \mathbf{V}_i^T f(\mathbf{W}_i x + \mathbf{g}_i) + b_i$$

$$\alpha_t^{(i)} = \frac{e^{(s_d^{(i)}(\mathbf{h}_t))}}{\sum_{j=1}^{N} e^{(s_d^{(i)}(\mathbf{h}_j))}}$$

$$\mathbf{e}^{(i)} = \sum_{t=1}^{N} \alpha_t^{(i)} \mathbf{h}_t$$

$$\mathbf{e} = (\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \cdots, \mathbf{e}^{(K)})$$

The Proposed Pooling Methods(2)

➢Multi-resolution multi-head attentive pooling(for diversity)

✓ Increasing T will make $\alpha_T(z_i)$ **less sharper**, thus with **lower resolution**

✓ As T → infinity, $\alpha_T(z_i)$ = 1/N, **average pooling**: bridge between attentive and statistical pooling

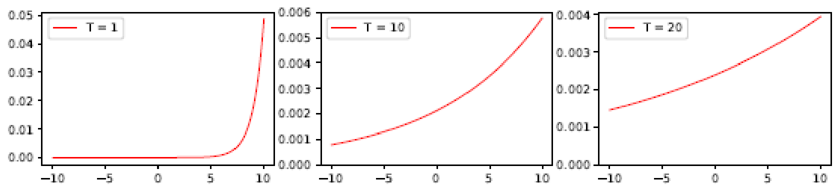


Fig. 1. The function curves of $\alpha_T(z_i) = \frac{e^{(z_i/T)}}{\sum_j e^{(z_j/T)}}$ as T varies.

$$s_l^{(i)}(x) = \mathbf{V}_i^T f(\mathbf{W}_i x + \mathbf{g}_i) + b_i$$

$$\alpha_t^{(i)} = \frac{e^{(s_d^{(i)}(\mathbf{h}_t)/T_i)}}{\sum_{j=1}^{N} e^{(s_d^{(i)}(\mathbf{h}_j)/T_i)}}$$

T_i = 1: global multi-head attention

$$\mathbf{e}^{(i)} = \sum_{t=1}^{N} \alpha_t^{(i)} \mathbf{h}_t$$

$$\mathbf{e} = (\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \cdots, \mathbf{e}^{(K)})$$

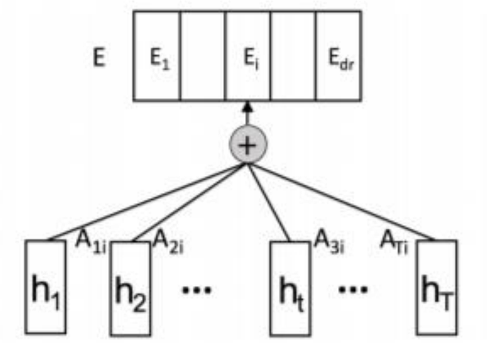
The Proposed Pooling Methods(3)

- Compared with Povey's method[6], ***using a penalty regularization in training objective to encourage diversity***
- ✓ attentive weights from different heads are orthogonal: ***a stronger requirement***
- ✓ not guarantee the approximate orthogonality of attentive weights during prediction
- ✓ our method: ***achieve diversity of extracted speech characteristics through the learned multi-resolution attentive model***

$$\mathbf{A} = softmax(g(\mathbf{H}^T\mathbf{W}_1)\mathbf{W}_2) \quad (1)$$

$$\mathbf{E} = \mathbf{HA} \quad (2)$$

$$P = \|(\mathbf{A}^T\mathbf{A} - \mathbf{I})\|_F^2 \quad (3)$$



Outline

- the Framework of Deep Speaker Embedding
- Pooling: an Overview
- the Proposed Pooling Methods
- **Experiments and Results**
- Conclusions

Experiments and Results(1)

1. dataset: VoxCeleb1
2. input features: 40-dimensional log-Mel filter banks
3. network architecture. 34-layer convolution ResNet
4. loss function: additive cosine margin softmax
5. optimizer: RAdam

$$\mathcal{L}_{CosAMS} = -\frac{1}{B}\sum_{u=1}^{B}\log\frac{e^{\eta(\cos(\theta_{<\mathbf{x}_u,\mathbf{w}_{y_u}>})-m)}}{Z_{\mathbf{x}_u}},$$

$$Z_{\mathbf{x}_u} = e^{\eta(\cos(\theta_{<\mathbf{x}_u,\mathbf{w}_{y_u}>})-m)} + \sum_{j\neq y_u} e^{\eta\cos(\theta_{<\mathbf{x}_u,\mathbf{w}_j>})},$$

Layer	Configuration
Conv1	$(3\times 3, 64)$, stride (1×1)
Res1	$[(3\times 3, 64)_2]\times 3$
Res2	$[(3\times 3, 128)_2]\times 4$
Res3	$[(3\times 3, 256)_2]\times 6$
Res4	$[(3\times 3, 512)_2]\times 3$
Conv2	$(3\times 3, 512)$, stride (1×3)
Pooling	pooling as represented in Sec.2
Linear1	output-dimension-of-pooling $\times$ 512
Linear2	512×512
Classifier	$512\times$ C, C = 1211

Experiments and Results(2)

	Front-end	Approaches	Loss	Dims	EER(%)
Nagrani et al. [15]	i-vector	PLDA	-	200	8.8
	VGG-M	adaptive average pooling	Softmax	1024	10.2
	VGG-M	adaptive average pooling	Contrastive	1024	7.8
W. Cai et al. [20]	ResNet-34	temporal average pooling	ASoftmax	128	5.27
	ResNet-34	self-attentive pooling	ASoftmax	128	4.9
	ResNet-34	learnable dictionary encoding(LDE)	ASoftmax	128	4.56
Our baselines	ResNet-34	attentive statistics pooling	CosAMS	512	4.258
	ResNet-34	multi-head attention($K = 4$), with Eq.(1)	CosAMS	512	4.385
	ResNet-34	multi-head attention($K = 4$), with Eq.(5)	CosAMS	512	4.464
Our proposals	ResNet-34	global multi-head attention($K = 4$)	CosAMS	512	**4.178**
	ResNet-34	multi-resolution multi-head attention($K = 4$)	CosAMS	512	**4.1**
	ResNet-34	global multi-head attention($K = 5$)	CosAMS	512	**4.109**
	ResNet-34	multi-resolution multi-head attention($K = 5$)	CosAMS	512	**3.982**
	ResNet-34	global multi-head attention($K = 6$)	CosAMS	512	**4.146**
	ResNet-34	multi-resolution multi-head attention($K = 6$)	CosAMS	512	**3.966**

Table 3. Results for verification on the test set of VoxCeleb1, all using the development set of VoxCeleb1 for training. “ASoftmax” represents angular softmax loss.

Experiments and Results(3)

- ➢ multi-resolution multi-head attention
- ✓ capture different views of speech characteristics
- ✓ less uncertain (lower entropy) in achieving discriminative information, representative of being more regularized

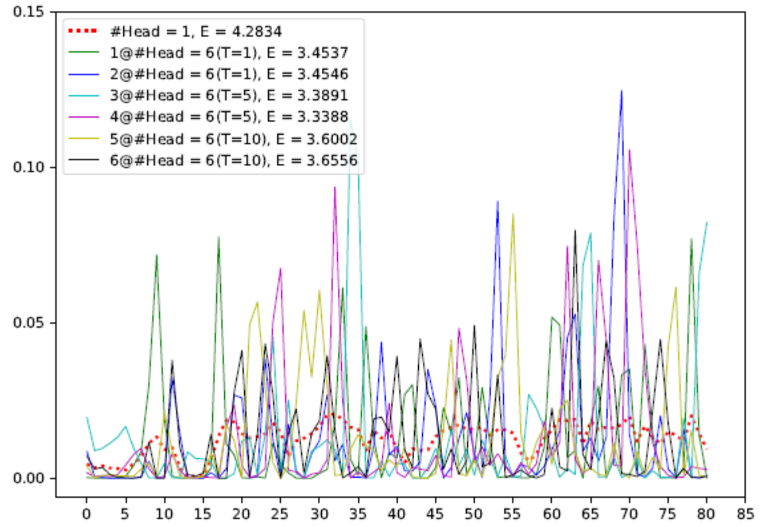


Fig. 2. The self-attentive weights from single-head vs. multi-resolution multi-head attention along the temporal axis, given the same test speech; much more attention is paid to the frames of higher weight scores. $E(= -\sum_t \alpha_t log(\alpha_t))$ is the entropy.

Outline

- ➢Deep Speaker Embedding Framework
- ➢Pooling: an Overview
- ➢The Proposed Pooling Methods
- ➢Experiments and Results
- ➢**Conclusions**

Conclusions

- ➢ proposed global and multi-resolution multi-head attention
- ➢ consistent improvement on top of that achieved with increased number of attention heads
- ➢ Why
- ✓ **analyzing speech segments as a whole**
- ✓ **multiple views from different attention heads with various resolutions**
- ✓ **improved certainty on each head**

Thank you!

Q & A