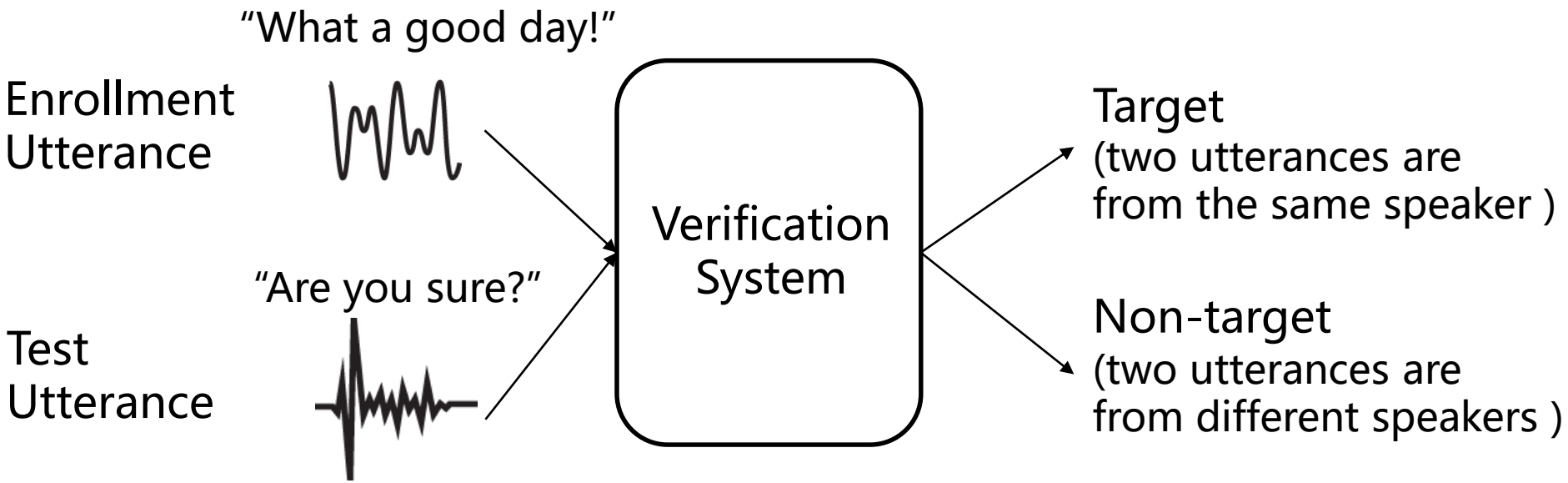
Angular Softmax Loss for End-to-end Speaker Verification

Yutian Li, Feng Gao, Zhijian Ou, Jiasong Sun

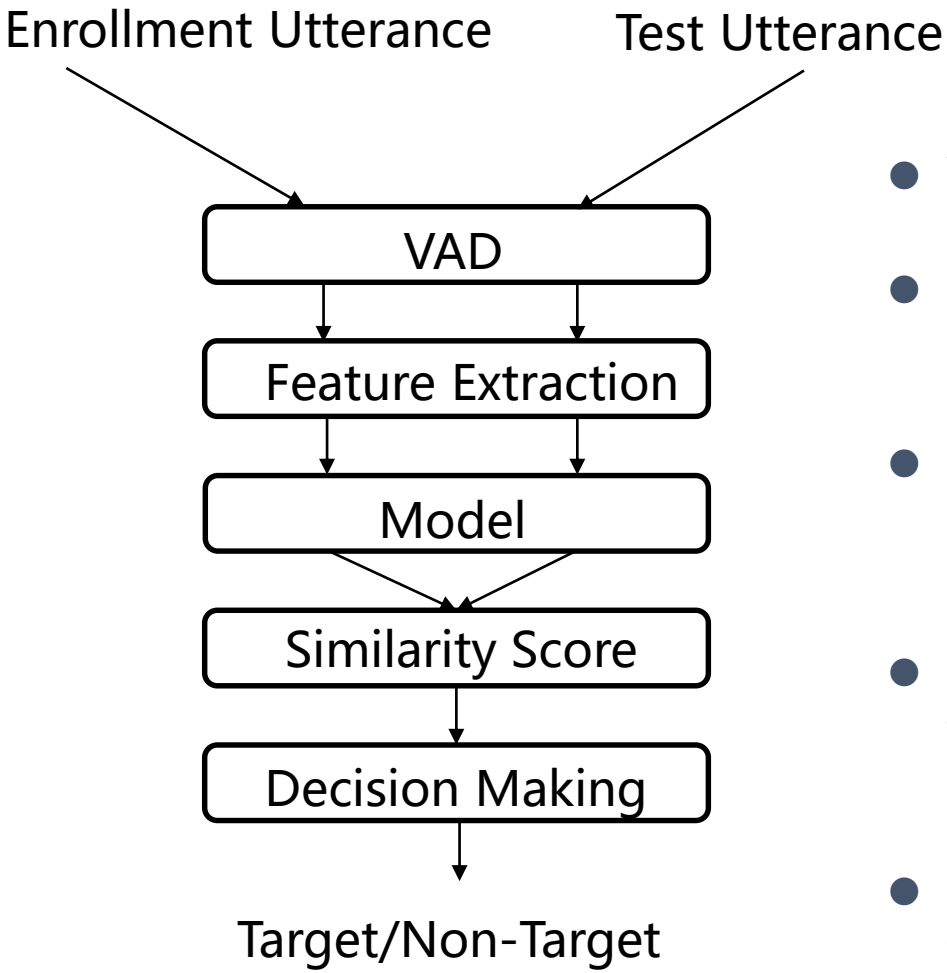
Speech Processing and Machine Intelligence (SPMI) Lab
Tsinghua University

CONTENT

What is text-independent speaker verification?



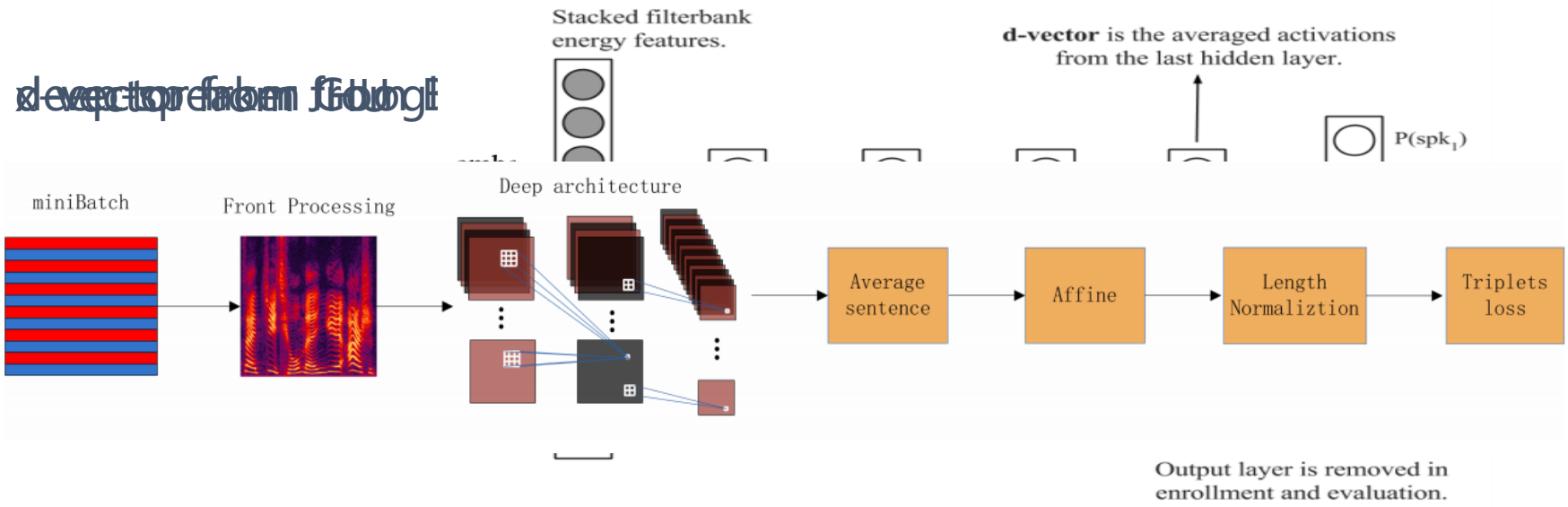
Speaker verification system pipeline



- VAD: Cut out non-speech audio;
- Feature Extraction: Extract MFCC or other types of acoustic features;
- Model: Project a variable-length feature sequence to a fixed-length embedding, e.g. i-vector, x-vector;
- Similarity Score: Usually calculated with PLDA (Probabilistic Linear Discriminant Analysis);
- Decision Making: Use a threshold to determine whether the output is target or not.

Model

- Traditionally: i-vector
- Recently: end-to-end model
  - ➢ One research direction: explore effective network architectures, e.g.

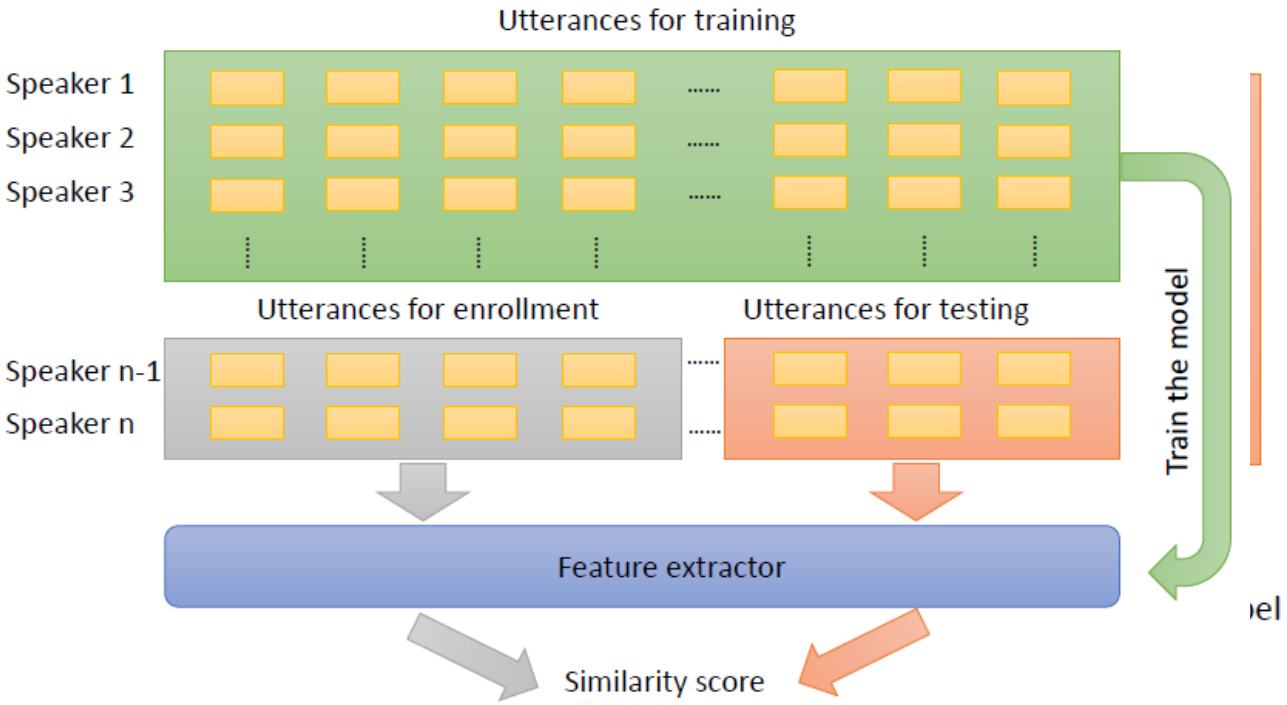


Variani, Ehsan, et al. "Deep neural networks for small footprint text-dependent speaker verification." *ICASSP*. Vol. 14. 2014.

Li, Chao, et al. "Deep speaker: an end-to-end neural speaker embedding system." *arXiv preprint arXiv:1705.02304* (2017).

Model

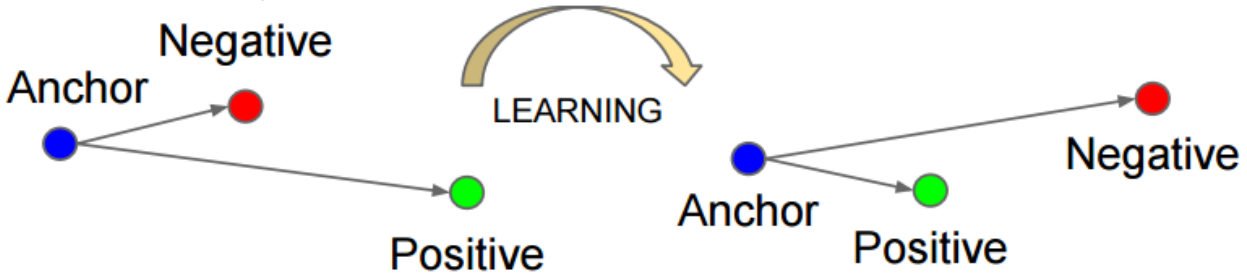
- Recently: end-to-end model
  - ➢ Another research direction: explore different criteria (loss functions)
- The softmax loss is more suitable for classification tasks
- Verification is different from classification



Losses for end-to-end model

- Triplet Loss: Make the Anchor-Positive pair closer and the Anchor-Negative pair farther.

$$\sum_{i}^{N}\left[\|f(x_i^a)-f(x_i^p)\|_2^2-\|f(x_i^a)-f(x_i^n)\|_2^2+\alpha\right]_+$$



- Requires a careful triplet selection procedure
  - ➢ both time-consuming and performance-sensitive.
- Training with triplet loss remains to be a difficult task
  - ➢ Efforts: generating triplets online from within a mini-batch (F. Schroff, et al), doing softmax pre-training (Li, Chao, et al).

F. Schroff, et al. “Facenet: A unified embedding for face recognition and clustering,” in CVPR, 2015.

Li, Chao, et al. "Deep speaker: an end-to-end neural speaker embedding system." *arXiv preprint* (2017).

CONTENT

Brief introduction for Angular Softmax Loss

- Recently proposed to improve softmax loss in face verification problem (Liu Weiyang, et al.).
- Introduces a margin between the target class and the non-target class into the softmax loss
- Drive end-to-end training of neural networks to learn angularly discriminative features
- Outperforms Softmax Loss, Triplet Loss ,Center Loss and Contrastive Loss.

Liu, Weiyang, et al. "Sphereface: Deep hypersphere embedding for face recognition." *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

A motivating binary classification example

- Softmax calculates the probability for two classes as

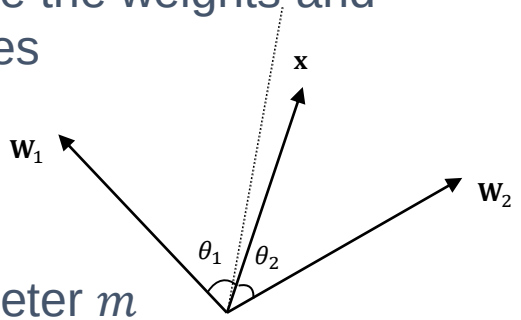
$$p_1 = \frac{\exp(\mathbf{W}_1^T\mathbf{x} + b_1)}{\exp(\mathbf{W}_1^T\mathbf{x} + b_1) + \exp(\mathbf{W}_2^T\mathbf{x} + b_2)} \qquad p_2 = \frac{\exp(\mathbf{W}_2^T\mathbf{x} + b_2)}{\exp(\mathbf{W}_1^T\mathbf{x} + b_1) + \exp(\mathbf{W}_2^T\mathbf{x} + b_2)}$$

- The decision boundary produced by Softmax Loss is

$$(\mathbf{W}_1 - \mathbf{W}_2)\mathbf{x} + b_1 - b_2 = 0$$

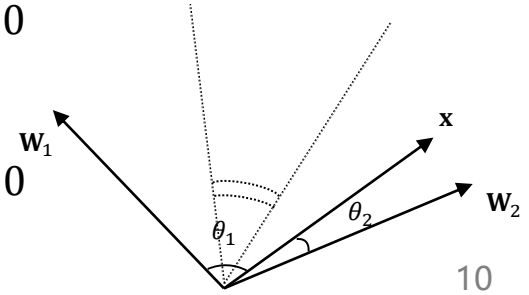
- To achieve angular decision boundary, we normalize the weights and zero out the biases. The decision boundary becomes

$$\|\mathbf{x}\|(\cos(\theta_1) - \cos(\theta_2)) = 0$$



- Further introduce angular margin, an integer parameter m

$$\Rightarrow \|\mathbf{x}\|(\cos(\theta_1) - \cos(\theta_2)) > 0$$
$$\|\mathbf{x}\|(\cos(m\theta_1) - \cos(\theta_2)) > 0 \quad \textit{for class 1}$$
$$\|\mathbf{x}\|(\cos(\theta_1) - \cos(m\theta_2)) < 0 \quad \textit{for class 2}$$
$$\Rightarrow \|\mathbf{x}\|(\cos(\theta_1) - \cos(\theta_2)) < 0$$



- Larger m means stronger discriminative ability

Definition of Angular Softmax Loss

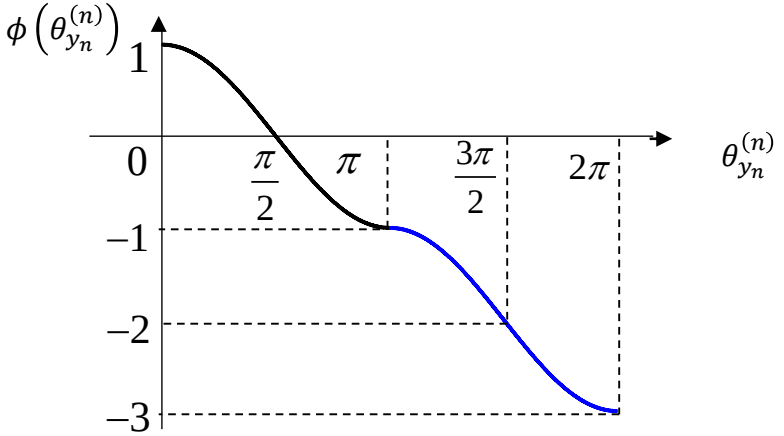
- Angular Softmax Loss, defined over training samples $\left(\mathbf{x}^{(n)}, y^{(n)}\right), n = 1, \cdots, N$

$$L_{ang} = \frac{1}{N}\sum_{n=1}^{N} -\log\left(\frac{e^{\|\mathbf{x}^{(n)}\| cos\left(m\theta_{y_n}^{(n)}\right)}}{e^{\|\mathbf{x}^{(n)}\| cos\left(m\theta_{y_n}^{(n)}\right)} + \sum_{j\neq y_n} e^{\|\mathbf{x}^{(n)}\| \cos\left(\theta_j^{(n)}\right)}}\right)$$

$$L_{ang} = \frac{1}{N}\sum_{n=1}^{N} -\log\left(\frac{e^{\|\mathbf{x}^{(n)}\| \phi\left(\theta_{y_n}^{(n)}\right)}}{e^{\|\mathbf{x}^{(n)}\| \phi\left(\theta_{y_n}^{(n)}\right)} + \sum_{j\neq y_n} e^{\|\mathbf{x}^{(n)}\| \cos\left(\theta_j^{(n)}\right)}}\right)$$

where $\phi\left(\theta_{y_n}^{(n)}\right) = (-1)^k \cos\left(m\theta_{y_n}^{(n)}\right) - 2k$ for $\theta_{y_n}^{(n)} \in \left[\frac{k\pi}{m}, \frac{(k+1)\pi}{m}\right], \quad k \in [0, m-1]$

- Examples of $\phi\left(\theta_{y_n}^{(n)}\right)$ with m=2



CONTENT

Data and Network

Data

- Training and evaluation data are both randomly chosen from the *Fisher* data.

Table: Number of speakers

	Total	Male	Female
Training	5000	2500	2500
Evaluation	1000	500	500

Network

- For comparison, our network architecture is similar to Kaldi x-vector.

utterance level layers	FC_2 (512→300)
	FC_1 (3000→512)
statistic pooling layer	mean and standard deviation
frame level layers	TDNN_5 (512×1→1500)
	TDNN_4 (512×1→512)
	TDNN_3 (512×3→512)
	TDNN_2 (512×3→512)
	TDNN_1 (23×5→512)

Experiment-1: fixed-duration enroll utterances

- Length of enrollment utterances is 3000 frames.
- Length of test utterances varies in 300, 500, 1000, 1500 frames.

EER (%)		Durations of test utterances			
Model	Loss + metric	300	500	1000	1500
i-vector	-- + PLDA	1.00	0.53	0.33	0.37
x-vector	Softmax + PLDA	1.86	0.83	0.40	0.43
Our Network	Softmax + PLDA	1.30	0.97	0.70	0.73
	Angular m=2 + Cosine	0.94	0.60	0.47	0.57
	Angular m=3 + Cosine	0.67	0.40	0.37	0.43
	Angular m=4 + Cosine	0.70	0.47	0.33	0.47
	Triplet + Cosine	2.17	1.63	1.17	1.23

1. Using triplet loss yields inferior performance

Experiment-1: fixed-duration enroll utterances

- Length of enrollment utterances is 3000 frames.
- Length of test utterances varies in 300, 500, 1000, 1500 frames.

EER (%)		Durations of test utterances			
Model	Loss + metric	300	500	1000	1500
i-vector	-- + PLDA	1.00	0.53	0.33	0.37
x-vector	Softmax + PLDA	1.86	0.83	0.40	0.43
Our Network	Softmax + PLDA	1.30	0.97	0.70	0.73
	Angular m=2 + Cosine	0.94	0.60	0.47	0.57
	Angular m=3 + Cosine	0.67	0.40	0.37	0.43
	Angular m=4 + Cosine	0.70	0.47	0.33	0.47
	Triplet + Cosine	2.17	1.63	1.17	1.23

2. For short test condition, Angular softmax performs significantly better than both Kaldi i-vector and Kaldi x-vector baseline

Experiment-1: fixed-duration enroll utterances

- Length of enrollment utterances is 3000 frames.
- Length of test utterances varies in 300, 500, 1000, 1500 frames.

EER (%)		Durations of test utterances			
Model	Loss + metric	300	500	1000	1500
i-vector	-- + PLDA	1.00	0.53	0.33	0.37
x-vector	Softmax + PLDA	1.86	0.83	0.40	0.43
Our Network	Softmax + PLDA	1.30	0.97	0.70	0.73
	Angular m=2 + Cosine	0.94	0.60	0.47	0.57
	Angular m=3 + Cosine	0.67	0.40	0.37	0.43
	Angular m=4 + Cosine	0.70	0.47	0.33	0.47
	Triplet + Cosine	2.17	1.63	1.17	1.23

3. Precluding the effect of the differences in network architecture, Angular softmax outperforms softmax significantly

Experiment-1: fixed-duration enroll utterances

- Length of enrollment utterances is 3000 frames.
- Length of test utterances varies in 300, 500, 1000, 1500 frames.

EER (%)		Durations of test utterances			
Model	Loss + metric	300	500	1000	1500
i-vector	-- + PLDA	1.00	0.53	0.33	0.37
x-vector	Softmax + PLDA	1.86	0.83	0.40	0.43
Our Network	Softmax + PLDA	1.30	0.97	0.70	0.73
	Angular m=2 + Cosine	0.94	0.60	0.47	0.57
	Angular m=3 + Cosine	0.67	0.40	0.37	0.43
	Angular m=4 + Cosine	0.70	0.47	0.33	0.47
	Triplet + Cosine	2.17	1.63	1.17	1.23

4. Angular softmax with m=4 is not always the best, due to the complication of neural network training

Experiment-2: equal durations of enroll and test utterances

- Length of enrollment utterances and length of test utterances both vary in 300, 500, 1000, 1500 frames.

EER (%)		Durations of utterances			
Model	Loss + metric	300	500	1000	1500
i-vector	-- + PLDA	2.93	1.57	0.50	0.47
x-vector	Softmax + PLDA	3.17	1.63	0.63	0.63
Our Network	Softmax + PLDA	3.43	2.40	1.20	1.07
	Angular m=2 + Cosine	2.90	1.57	0.77	0.83
	Angular m=2 + PLDA	2.17	1.33	0.73	0.80
	Angular m=3 + Cosine	2.50	1.23	0.73	0.56
	Angular m=3 + PLDA	2.10	1.33	0.70	0.77
	Angular m=4 + Cosine	2.43	1.33	0.70	0.63
	Angular m=4 + PLDA	2.23	1.37	0.73	0.90

1. For short utterance condition, using PLDA back-end significantly reduce EERs of the Angular softmax systems.

Experiment-2: equal durations of enroll and test utterances

- Length of enrollment utterances and length of test utterances both vary in 300, 500, 1000, 1500 frames.

EER (%)		Durations of utterances			
Model	Loss + metric	300	500	1000	1500
i-vector	-- + PLDA	2.93	1.57	0.50	0.47
x-vector	Softmax + PLDA	3.17	1.63	0.63	0.63
Our Network	Softmax + PLDA	3.43	2.40	1.20	1.07
	Angular m=2 + Cosine	2.90	1.57	0.77	0.83
	Angular m=2 + PLDA	2.17	1.33	0.73	0.80
	Angular m=3 + Cosine	2.50	1.23	0.73	0.56
	Angular m=3 + PLDA	2.10	1.33	0.70	0.77
	Angular m=4 + Cosine	2.43	1.33	0.70	0.63
	Angular m=4 + PLDA	2.23	1.37	0.73	0.90

2. Compared with the i-vector and x-vector baseline, the EERs of Angular softmax system are the best when utterances are short.

Experiment-2: equal durations of enroll and test utterances

- Length of enrollment utterances and length of test utterances both vary in 300, 500, 1000, 1500 frames.

EER(%)		Durations of utterances			
Model	Loss + metric	300	500	1000	1500
i-vector	-- + PLDA	2.93	1.57	0.50	0.47
x-vector	Softmax + PLDA	3.17	1.63	0.63	0.63
Our Network	Softmax + PLDA	3.43	2.40	1.20	1.07
	Angular m=2 + Cosine	2.90	1.57	0.77	0.83
	Angular m=2 + PLDA	2.17	1.33	0.73	0.80
	Angular m=3 + Cosine	2.50	1.23	0.73	0.56
	Angular m=3 + PLDA	2.10	1.33	0.70	0.77
	Angular m=4 + Cosine	2.43	1.33	0.70	0.63
	Angular m=4 + PLDA	2.23	1.37	0.73	0.90

3. Angulat softmax outperforms traditional softmax significantly on all conditions.

CONTENT

Further results on SRE-18

- We use Angular Softmax Loss in SRE-18, CMN2 (Call My Net 2) test dataset.
- Training set for neural network consists of :
  - Fisher
  - Switchborad
  - Previous SRE evaluation data
  - Mixer 6
  - Voxceleb

- The result in SRE-18 CMN2 development data.

Model + metric	EER (%)
Kaldi i-vector + PLDA	15.08
Kaldi Softmax + PLDA	9.64
Angular m=2 + PLDA	9.46
Angular m=3 + PLDA	9.93

CONTENT

Conclusion and contribution

Two contributions:

- We introduce Angular Softmax Loss into end-to-end speaker verification
  - Can learn more discriminative features than softmax and triplet loss;
  - Easy and stable for usage.

- The combination of using Angular softmax in training the front-end and using PLDA in the back-end scoring further boosts the performance.

Conclusions:

- Angular Softmax performs better than Kaldi i-vector baseline for short test utterances.

- In SRE18 development dataset, Angular Softmax x-vector outperforms both Kaldi i-vector and x-vector baselines.

Thanks for your attention !

Reporter: Yutian Li

Email: yutian-l16@mails.tsinghua.edu.cn