Prototypical Networks for Small Footprint Text-independent Speaker Verification

Tom Ko, Yangbin Chen and Qing Li

Southern University of Science and Technology

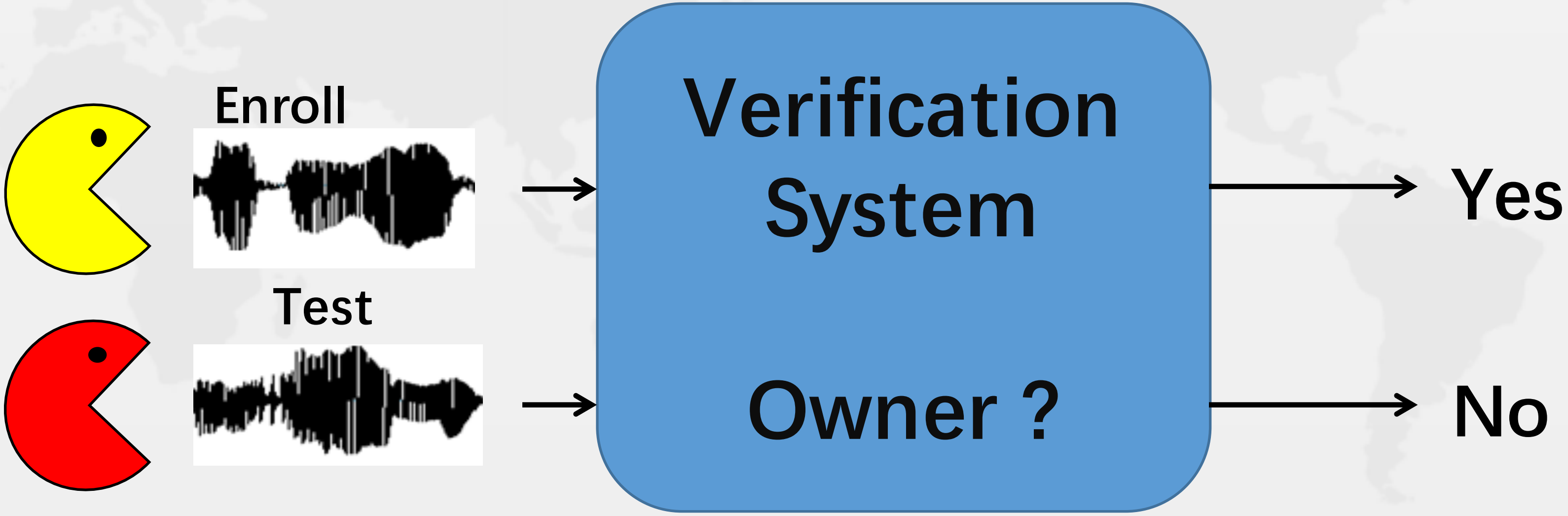
ICASSP 2020

Motivation

- There is a mismatch of the training objective between the front-end DNN and the PLDA backend in the speaker embedding approaches.
- Prototypical Networks aim at learning a non-linear mapping from the input space to an embedding space with a predefined distance metric. It tries to minimize the intra class distance and maximize the inter class distance, just like PLDA.
- It is worth to investigate the use of prototypical networks in a small footprint text-independent speaker verification task.

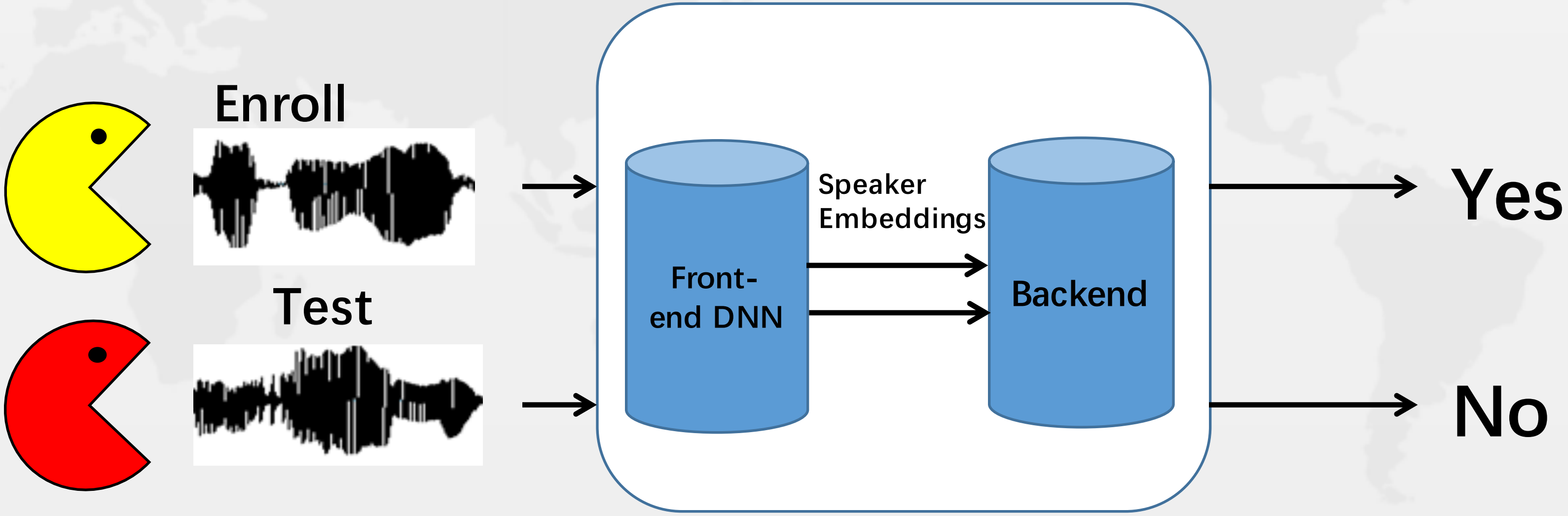
Text-independent Speaker Verification

- It needs to verify if the test speaker and enroller speaker are the same one.



The Speaker Embedding Approach

- Front-end DNN for speaker embedding extraction.
- Backend for similarity measure.

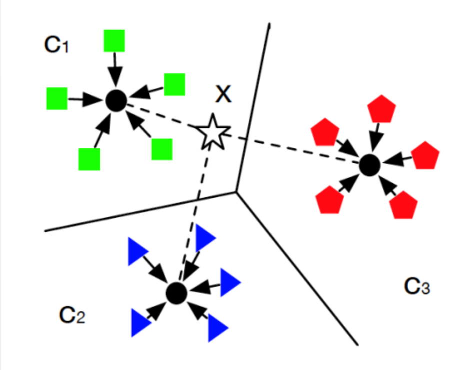


Meta-learning

- It becomes the most popular solution for solving few-shot classifications.
- Also known as 'learning to learn', aims to learn new skills or adapt to new environments rapidly with only a few examples.
- Many elegant solutions are proposed:
  - Matching Networks
  - Prototypical Networks
  - Model-agnostic Meta-learning

Prototypical Networks

- To train a model which can generalize to new classes not seen in the training set , given only a few examples per new class. Thus, it has to **learn a good representation**.
- It tends to minimize the intra class distance and maximize the inter class distance.
- The distance metric can be defined in a flexible way.

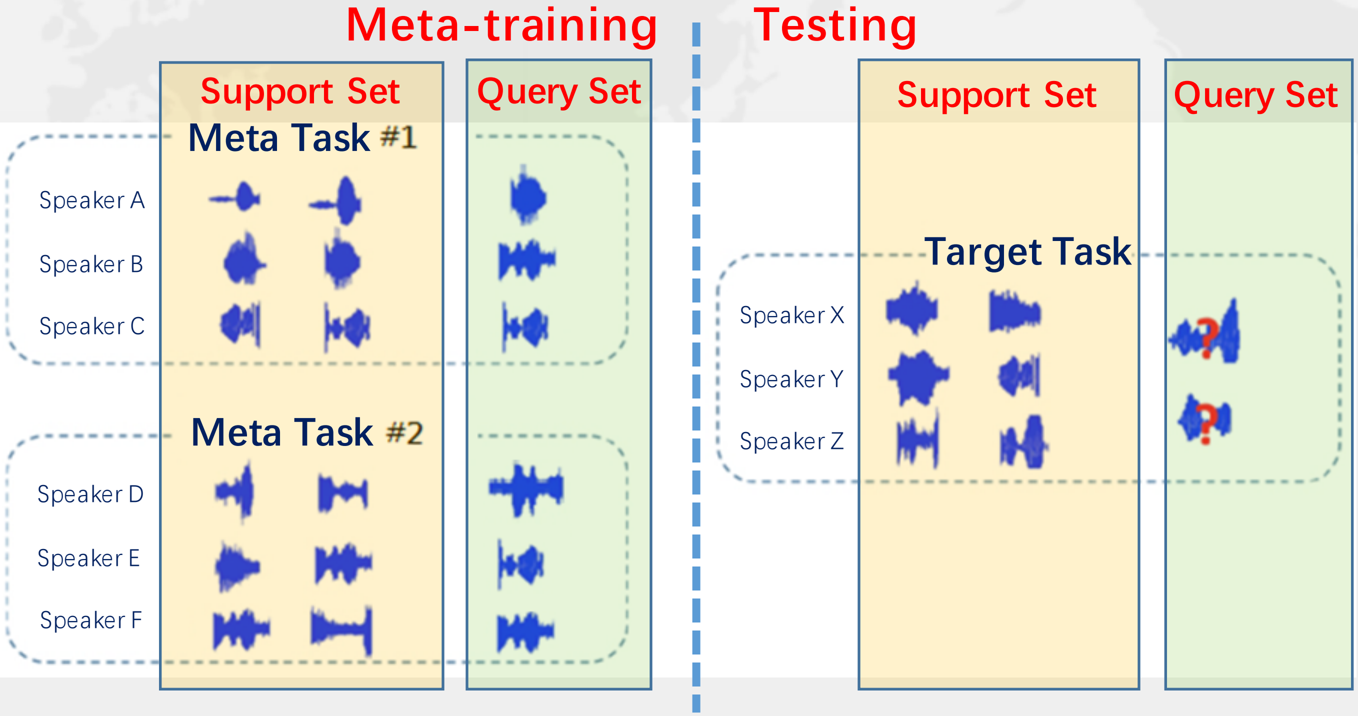


Jake Snell, Kevin Swersky, and Richard Zemel. "Prototypical networks for few-shot learning." In: *Advances in neural information processing systems*. 2017. p. 4077-4087.

SUSTech Southern University of Science and Technology

Meta-training in Few-shot Classification

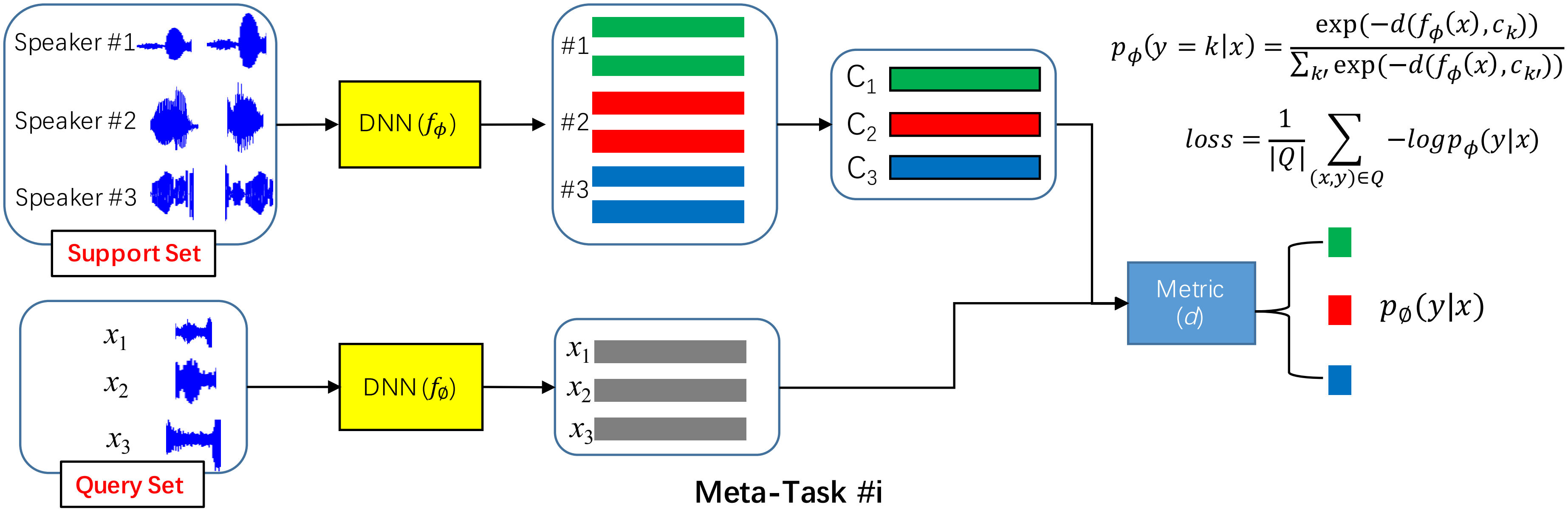
- The model is trained on a number of meta-tasks and it treats an entire task as a training example.



SUSTech Southern University of Science and Technology

Prototypical Networks as the SV Frontend

- Support sets are used for computing class centroids.



The learned DNN will be used as the frontend

Experimental Setup

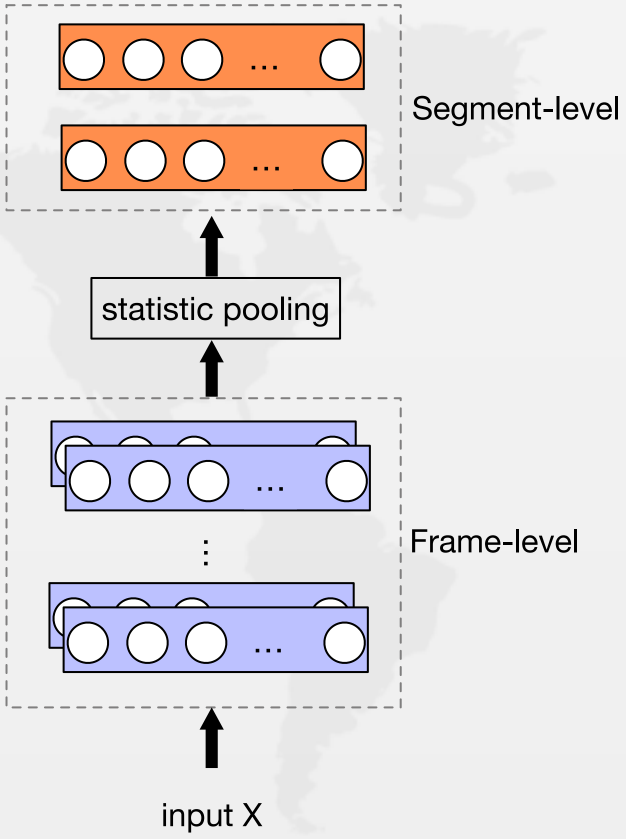
- Training data
  - SWBD dataset: 28k recordings from 2.6k speakers
  - SRE dataset: 35k recordings from 3.8k speakers

  - *4k_full*, *4k_2utt*, *2k_2utt* are sampled to compare the proposed method and the conventional one.
- Evaluation data
  - SRE10

  - Both the enrollment and test utterances are truncated to the first $T \in \{2,5,10,30\}$ seconds of speech, as determined by an energy-based VAD.

Model Structure

- We use a similar model structure as the X-vector* approach.
- Several layers are removed to fulfill the small footprint requirement.
- We compare our approach with the conventional learning approach.



*David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2018.

SUSTech Southern University of Science and Technology

Practical Implementation of Prototypical Networks

- Our work has a large number of speakers within each meta-task, which costs a high memory usage. To address this problem, we design an expectation-maximization (EM) like algorithm which save the memory cost and does not affect the performance.
- In the E step, the embeddings of the support set are extracted and the class centroids are estimated.
- In the M step, the embeddings of the query set are extracted, then the distances and the losses are estimated.

Baseline

- Conventional learning approach with different backend metrics

Table 1. *EER(%) of a conventional front-end with different backend metrics. The models are trained with 2k_2utt training set.*

Backend Metric	2s-2s	5s-5s	10s-10s	30s-30s
Euclidean	45.85	46.07	45.85	46.48
Cosine	46.14	46.00	46.02	46.76
LDA+Euclidean	41.23	34.54	29.90	23.04
LDA+Cosine	36.66	28.77	21.94	15.32
LDA+PLDA	**34.51**	**26.26**	**18.39**	**12.27**

SUSTech Southern University of Science and Technology

Results

- Prototypical networks with different backend metrics

Table 2. *EER(%) of prototypical embeddings (10-shots) on SRE10. The models are trained with the 2k_2utt training set.*

Front-end Metric	Backend Metric	2s-2s	5s-5s	10s-10s	30s-30s
Euclidean	Euclidean	40.94	34.50	30.06	26.01
Euclidean	LDA+Euclidean	43.66	38.57	33.19	27.29
Euclidean	LDA+PLDA	34.34	25.70	18.62	11.81
Cosine	Cosine	36.07	29.39	25.72	23.17
Cosine	LDA+Cosine	36.88	28.52	21.62	14.94
Cosine	LDA+PLDA	**33.42**	**24.59**	**17.37**	**10.97**

Results

- Comparing prototypical networks and baseline approach

Table 3. *EER(%) on SRE10 with various training set*

Training set	System	2s-2s	5s-5s	10s-10s	30s-30s
$2k\_2utt$	Baseline	34.51	26.26	18.39	12.27
	Cosine	**33.42**	**24.59**	**17.37**	**10.97**
$4k\_2utt$	Baseline	33.47	24.98	17.44	11.61
	Cosine	**32.17**	**22.77**	**15.46**	**9.66**
$4k\_full$	Baseline	**29.79**	21.48	13.96	**8.52**
	Cosine	30.14	**21.28**	**13.75**	8.55

Observations

- The prototypical networks are better than the conventional approach when the front-end is directly evaluated with Euclidean or Cosine distance.
- LDA brings negative impact when Euclidean distance is used while it does not bring negative impact to Cosine distance.
- When there are **limited amount of training data per speaker**, prototypical networks perform obviously better than the baseline approach. When the entire training set is used, the two approaches obtain similar performance.

SUSTech Southern University of Science and Technology

Future Work

- In this paper, we apply the prototypical networks to improve the front-end in the speaker embedding approach.
- In the future, we want to further exploit the meta-learning framework to implement an end-to-end speaker verification system.
- Improve the overall performance with data augmentation techniques.
- Explore other meta learning methods.

SUSTech Southern University of Science and Technology

Thank you!