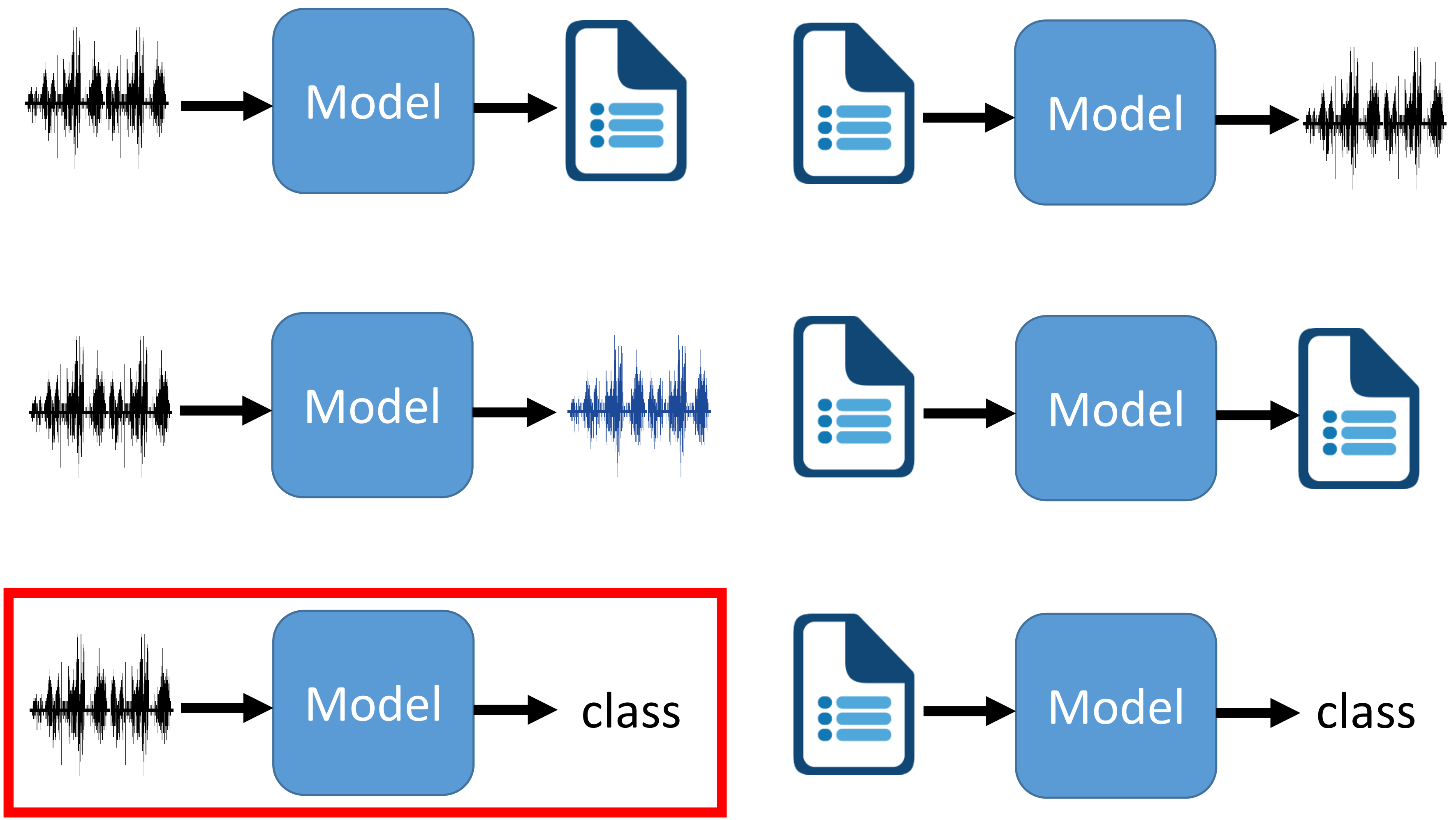
Speaker Verification

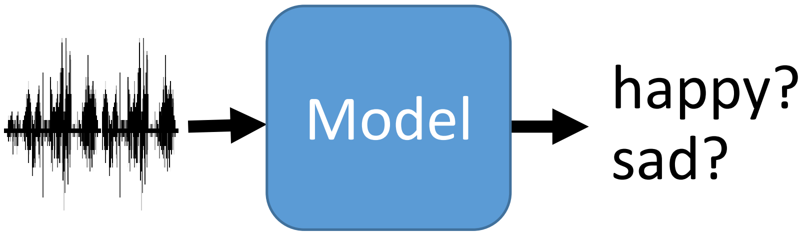
Hung-yi Lee
李宏毅

Some slides are from 袁培傑

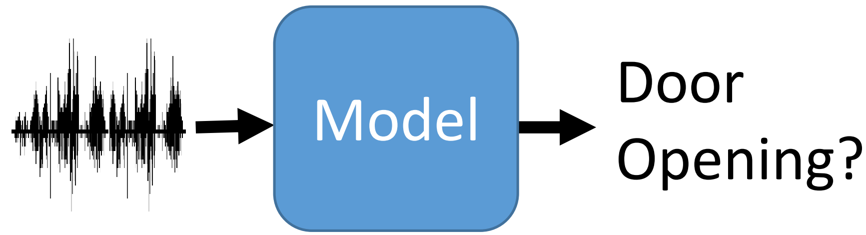
One slide for this course



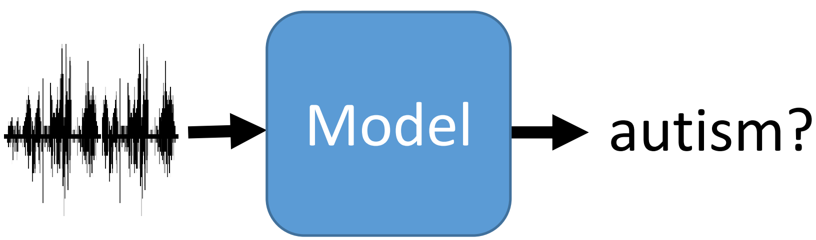
Related Tasks



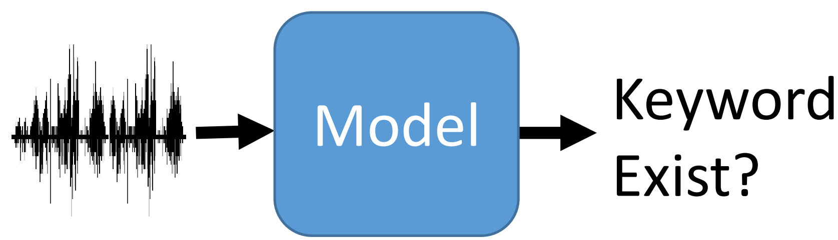
Emotion Recognition



Sound Event Detection



Autism Recognition



Keyword Spotting

We only focus on **speaker verification** today.

Outline

Task Introduction

Speaker Embedding

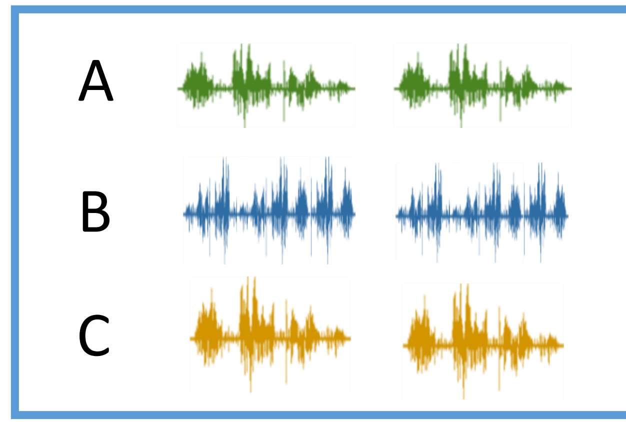
End-to-end

Task Introduction

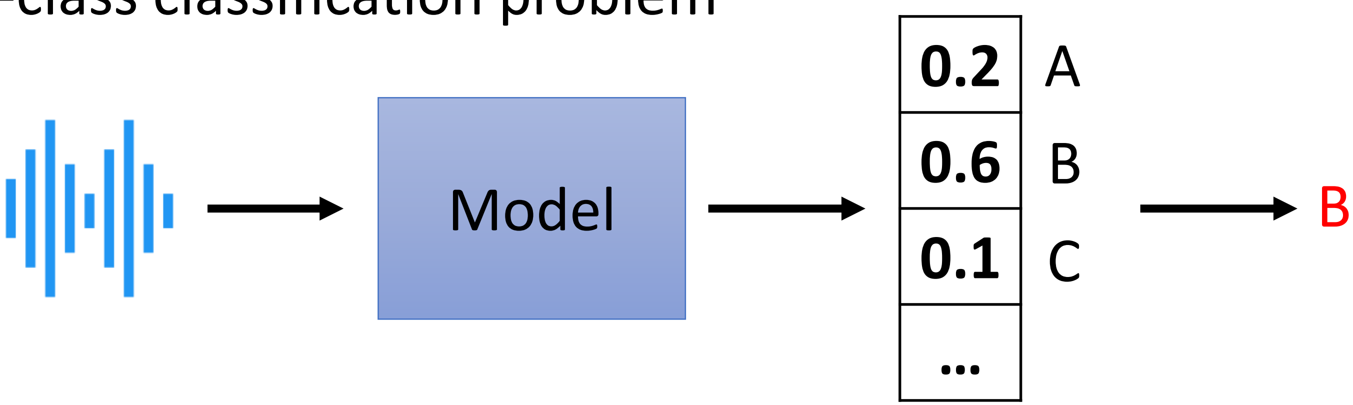
- Speaker Recognition / Identification
  - 語者識別
  - 一段語音是誰所說的
- Speaker Verification
  - 語者驗證
  - 兩段語音是否為同一人所說
- Speaker Diarization
  - 語者分段標記
  - 在一段語音中，誰在何時說話

Task Introduction

- Speaker Recognition / Identification
  - 語者識別
  - 一段語音是誰所說的



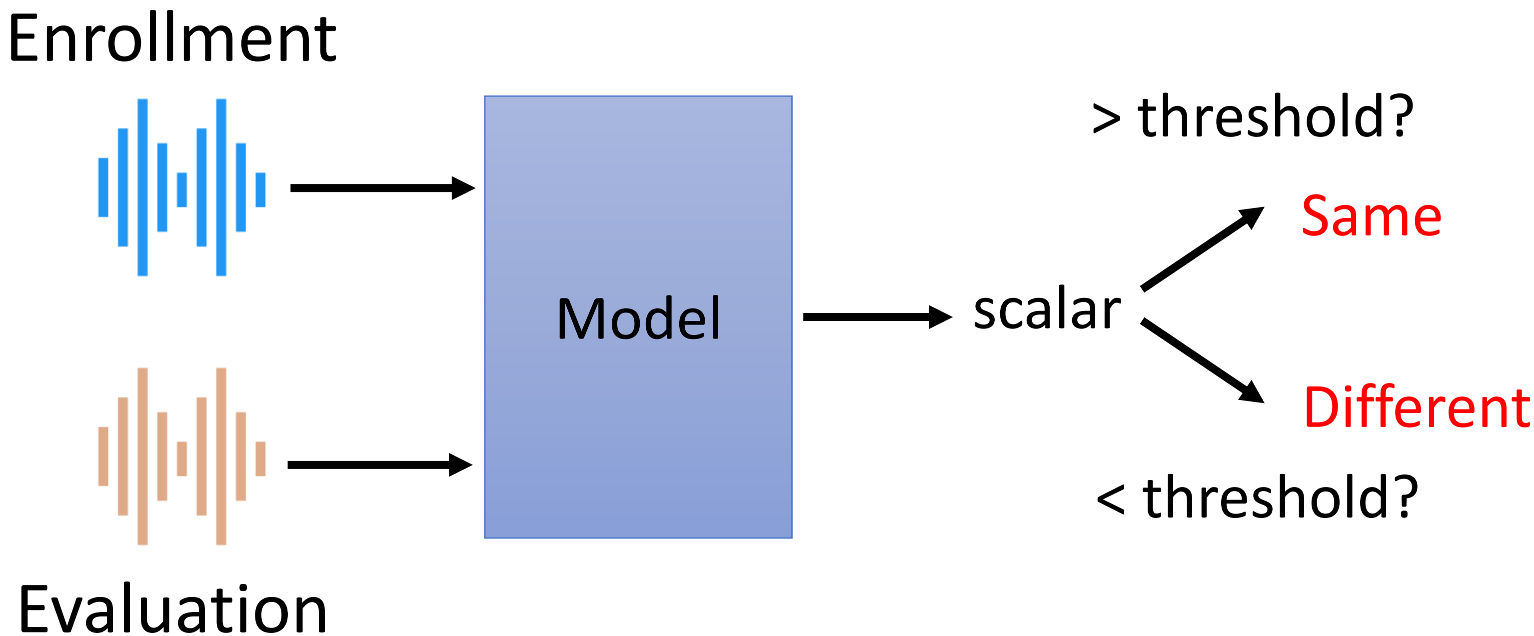
A multi-class classification problem



Task Introduction

- Speaker Recognition / Identification
  - 語者識別
  - 一段語音是誰所說的
- Speaker Verification
  - 語者驗證
  - 兩段語音是否為同一人所說

Speaker Verification



Application: 銀行客服

Equal Error Rate (EER)

False Negative (FN) Rate

同一語者被判斷成不同語者

False Positive (FP) Rate

不同語者被判斷成同一語者

threshold 1.0	
TP 0	FP 0
FN 100	TN 100

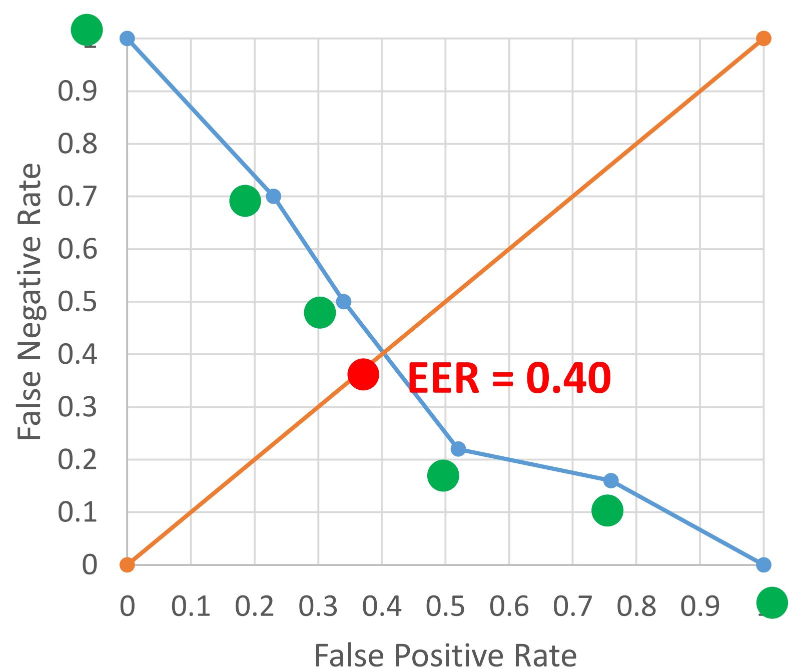
threshold 0.8	
TP 30	FP 23
FN 70	TN 77

threshold 0.6	
TP 50	FP 34
FN 50	TN 67

threshold 0.4	
TP 78	FP 52
FN 22	TN 48

threshold 0.2	
TP 84	FP 76
FN 16	TN 24

threshold 0.0	
TP 100	FP 100
FN 0	TN 0



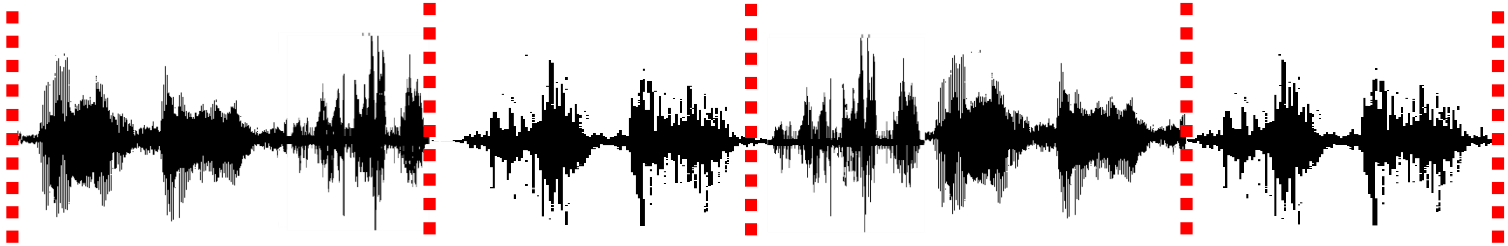
Task Introduction

- Speaker Recognition / Identification
  - 語者識別
  - 一段語音是誰所說的
- Speaker Verification
  - 語者驗證
  - 兩段語音是否為同一人所說
- Speaker Diarization
  - 語者分段標記
  - 在一段語音中，誰在何時說話

diarize: to write down your future arrangements, meetings, etc. in a diary

Speaker Diarization

Record of meeting, record of telephone conversion, etc.



no. 1 no. 2 no. 1 no. 3

Step 1: Segmentation

Step 2: Clustering

The number of speakers can be known or unknown.

Task Introduction

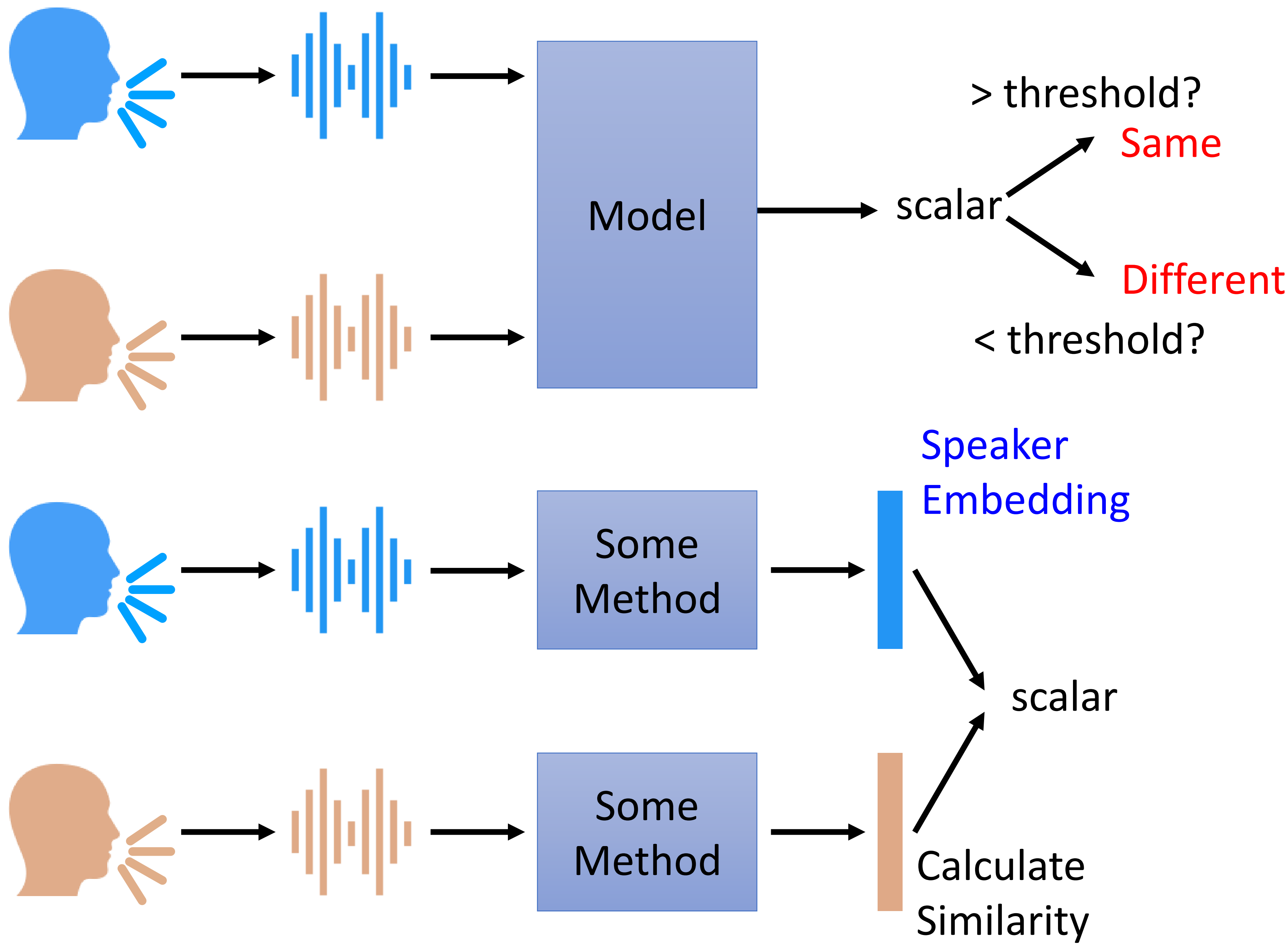
- Speaker Recognition / Identification
  - 語者識別
  - 一段語音是誰所說的
- Speaker Verification
  - 語者驗證
  - 兩段語音是否為同一人所說
- Speaker Diarization
  - 語者分段標記
  - 在一段語音中，誰在何時說話

Outline

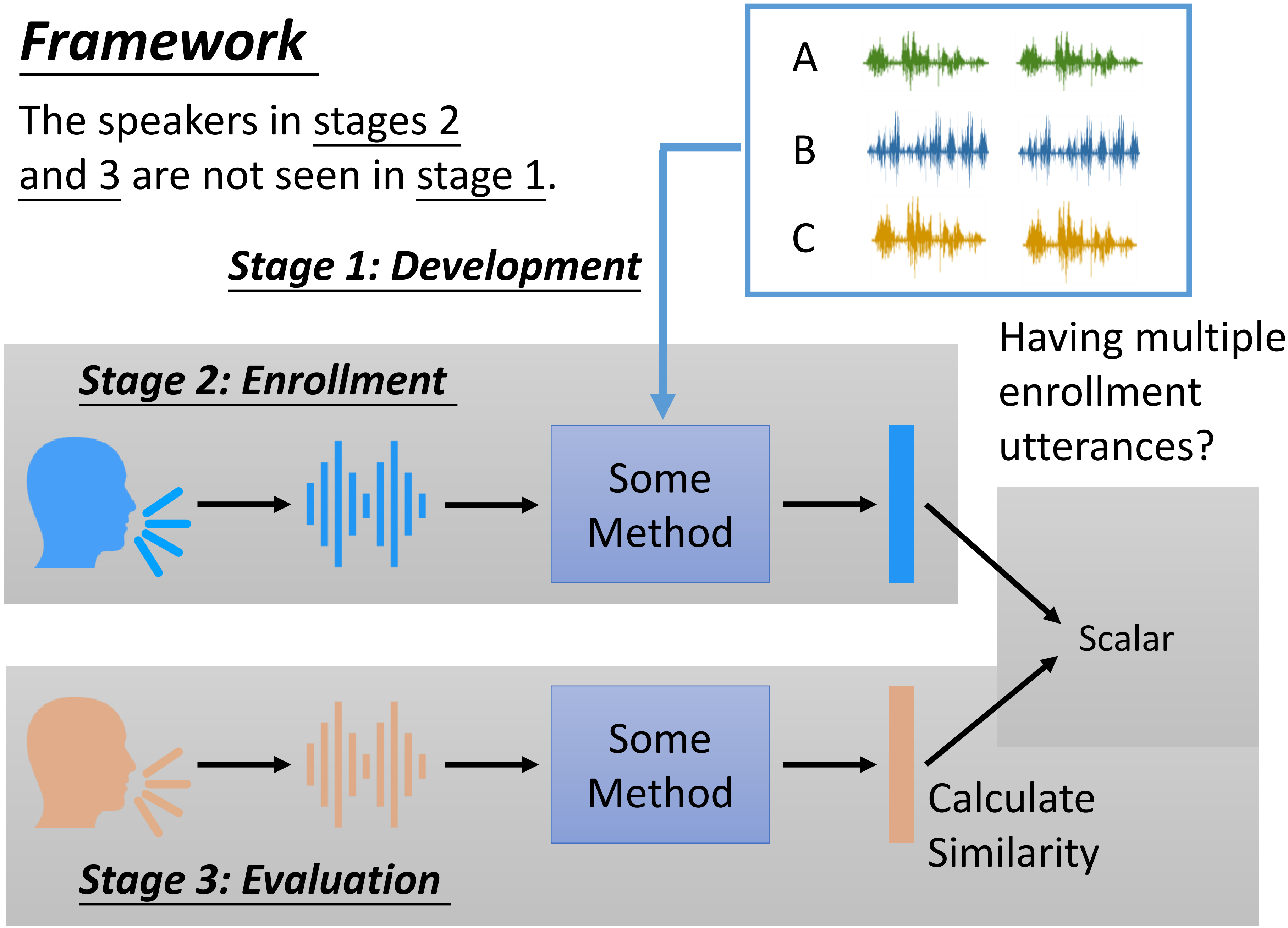
Task Introduction

Speaker Embedding

End-to-end



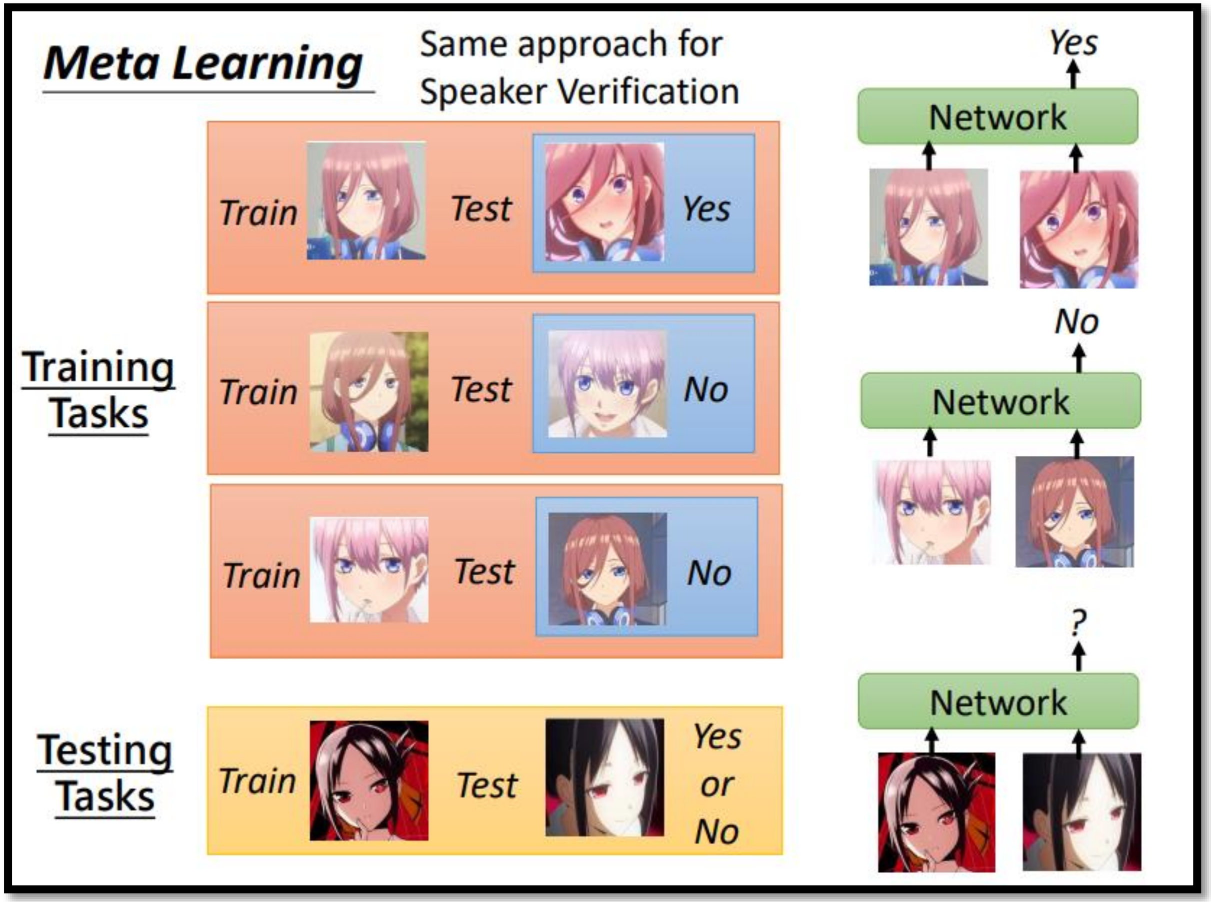
Model
> threshold?
Same
scalar
Different
< threshold?
Speaker
Embedding
Some
Method
Some
Method
scalar
Calculate
Similarity



Framework
The speakers in stages 2 and 3 are not seen in stage 1.
Stage 1: Development
A
B
C
Stage 2: Enrollment
Some Method
Having multiple enrollment utterances?
Scalar
Some Method
Calculate Similarity
Stage 3: Evaluation

Metric-based meta learning

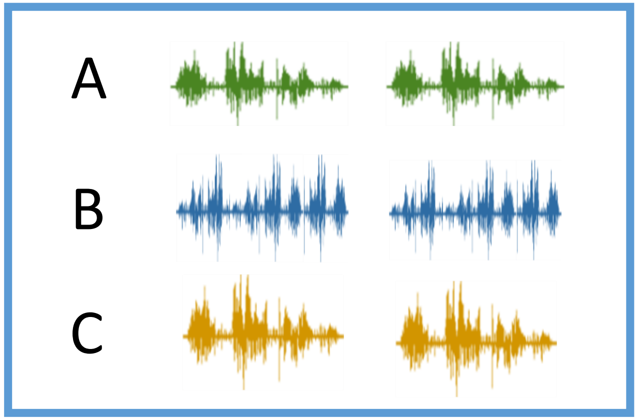
- https://youtu.be/yyKaACh_j3M



Framework

The speakers in stages 2 and 3 are not seen in stage 1.

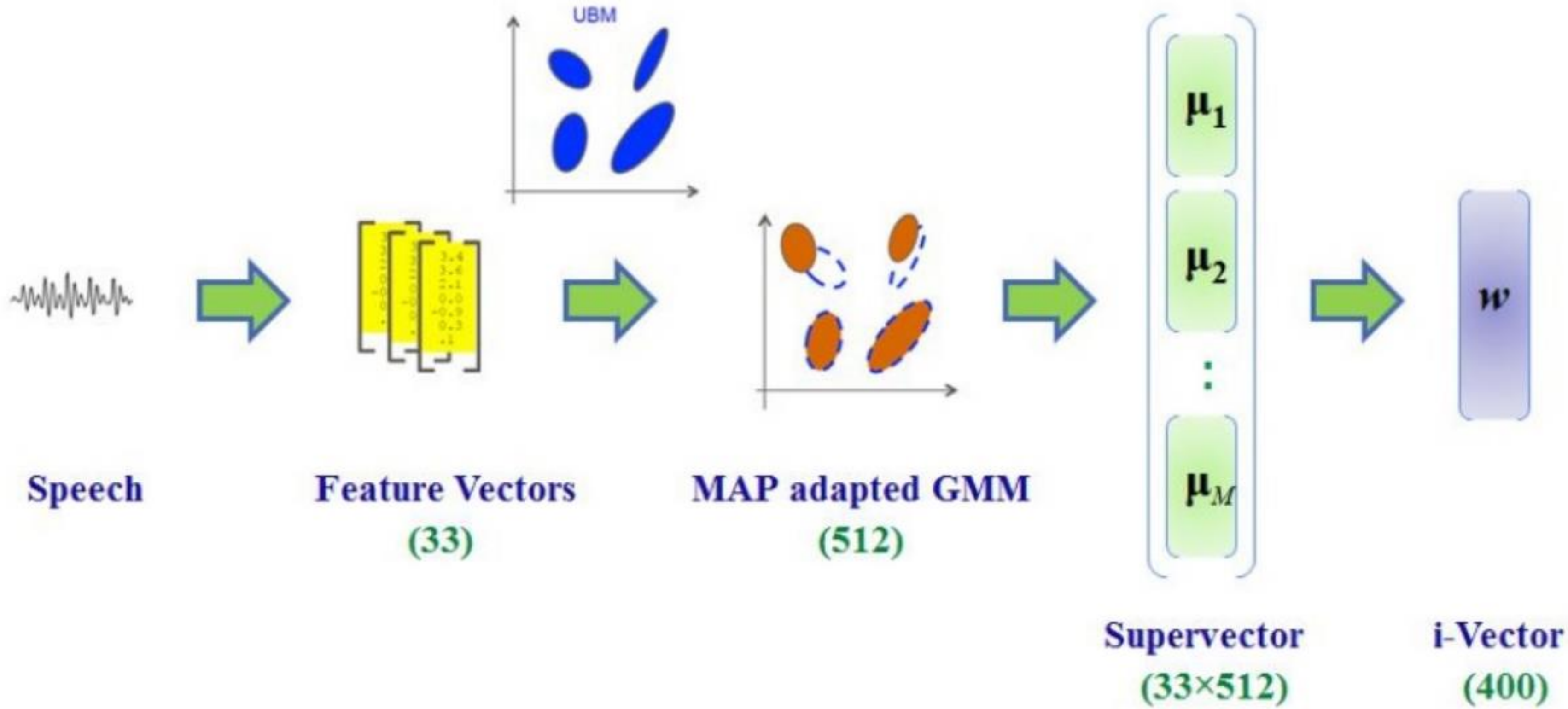
Stage 1: Development



- ~~Google's Dataset (private)~~ [Wan, et al., ICASSP'18]
  - ~~36M utterances, 18000 speakers~~
- VoxCeleb [Nagrani, et al., INTERSPEECH'17]
  - 0.15M utterances, 1251 speakers
- VoxCeleb2 [Chung, et al., INTERSPEECH'18]
  - 1.12M utterances, 6112 speakers

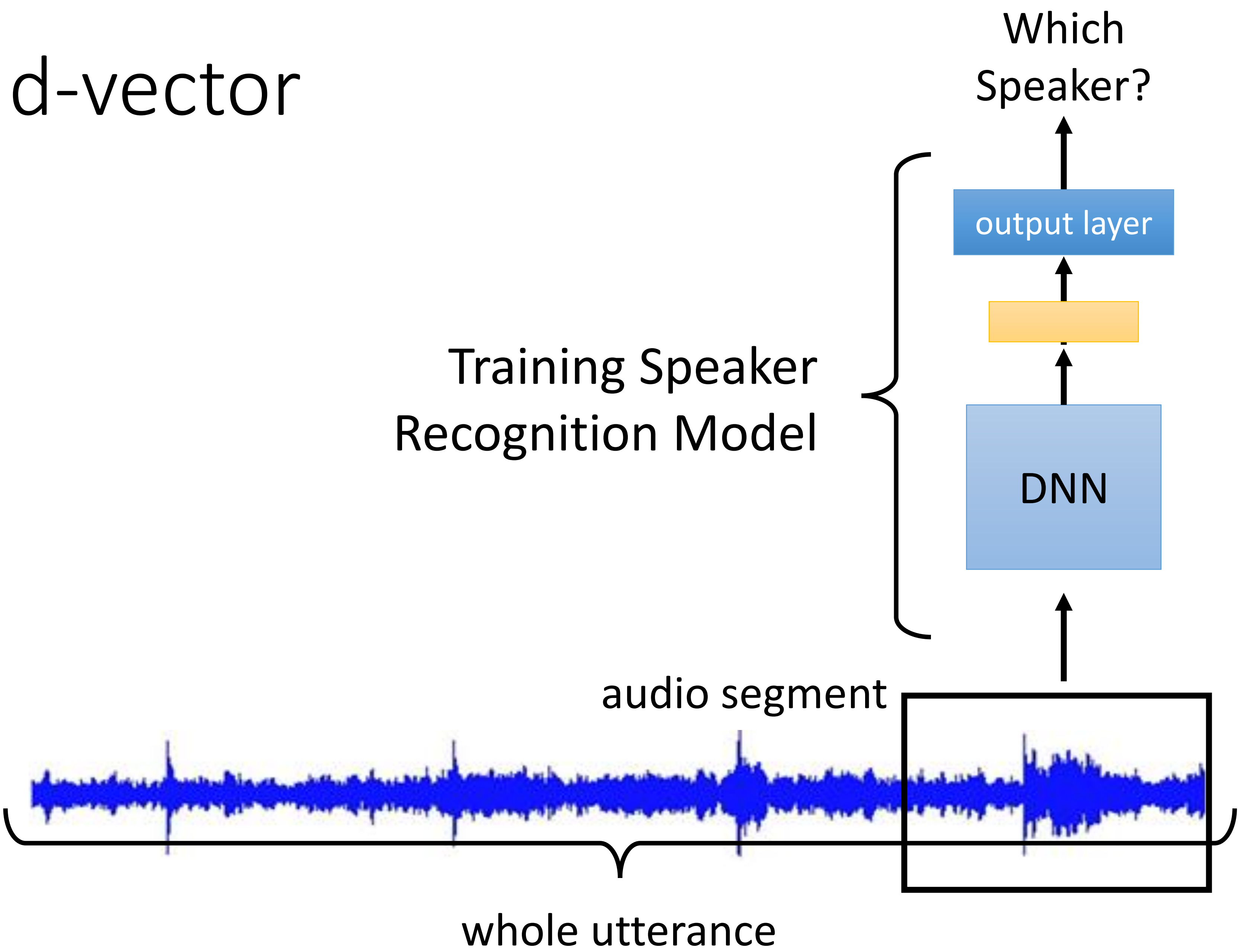
i-vector

"i" means "identity"



Source of image: https://www.slideshare.net/xavigiro/speaker-id-d3l3-deep-learning-for-speech-and-language-upc-2017

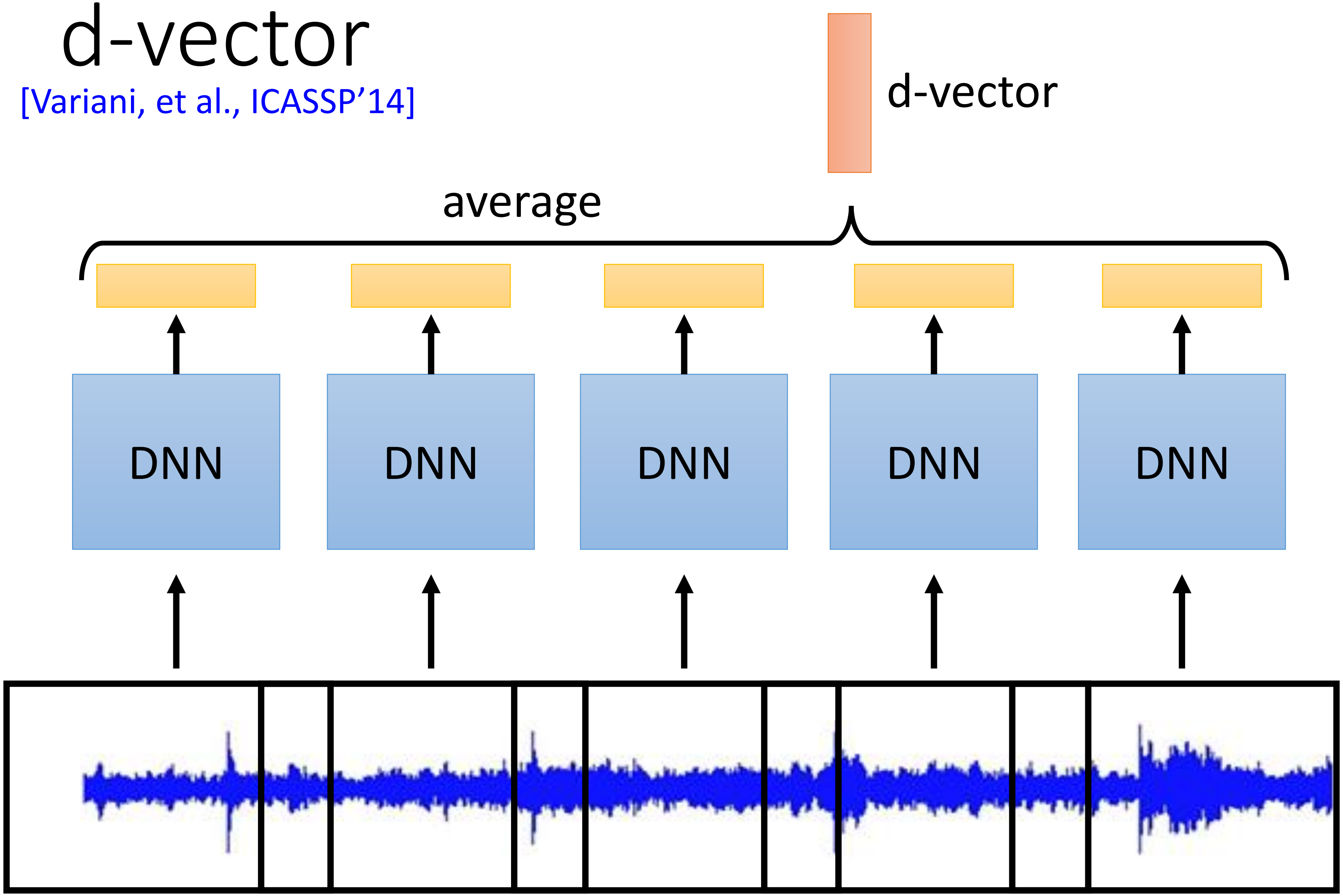
d-vector

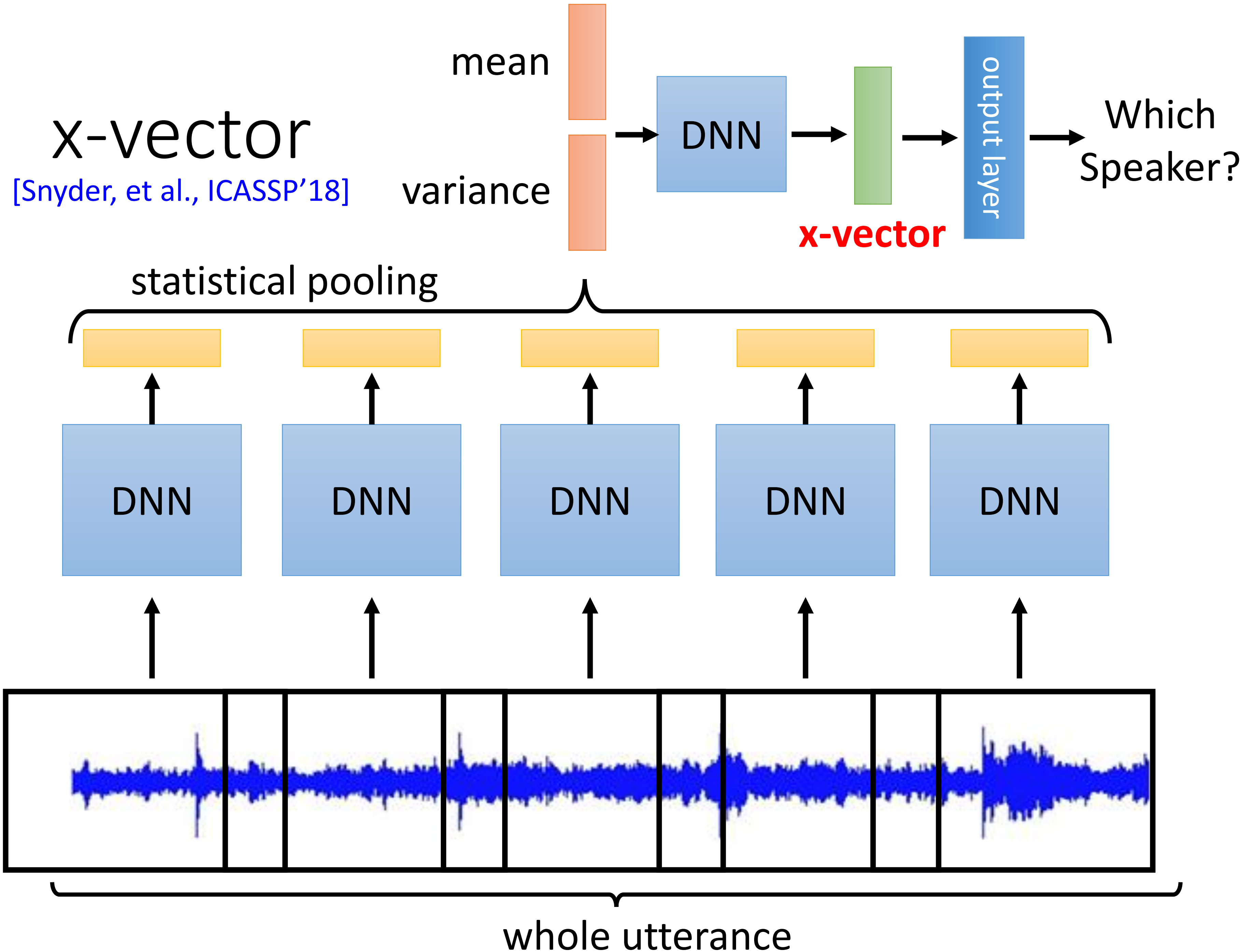

Which
Speaker?
output layer
Training Speaker
Recognition Model
DNN
audio segment
whole utterance

d-vector and i-vector are only comparable

d-vector

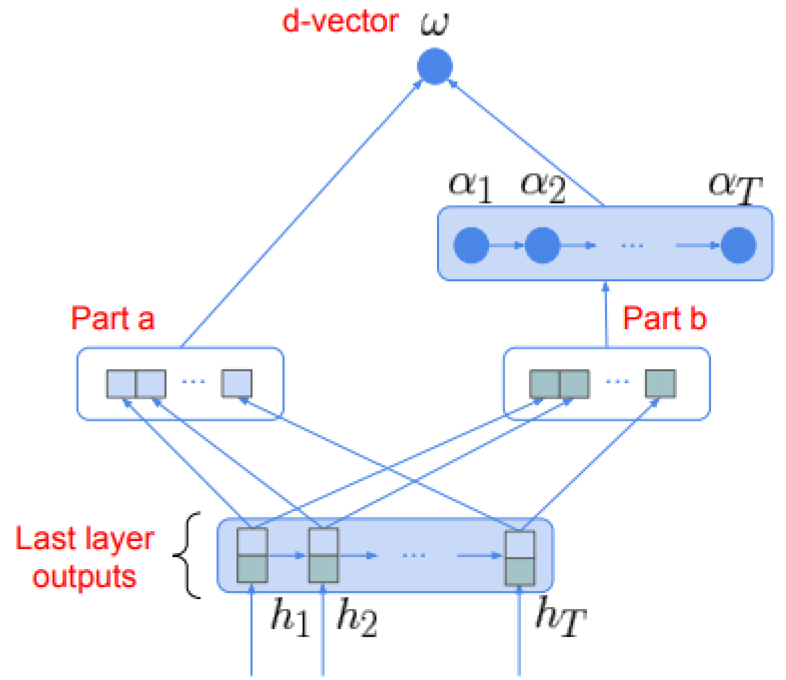
[Variani, et al., ICASSP'14]




x-vector
[Snyder, et al., ICASSP'18]
mean
variance
DNN
x-vector
output layer
Which Speaker?
statistical pooling
DNN
DNN
DNN
DNN
DNN
whole utterance

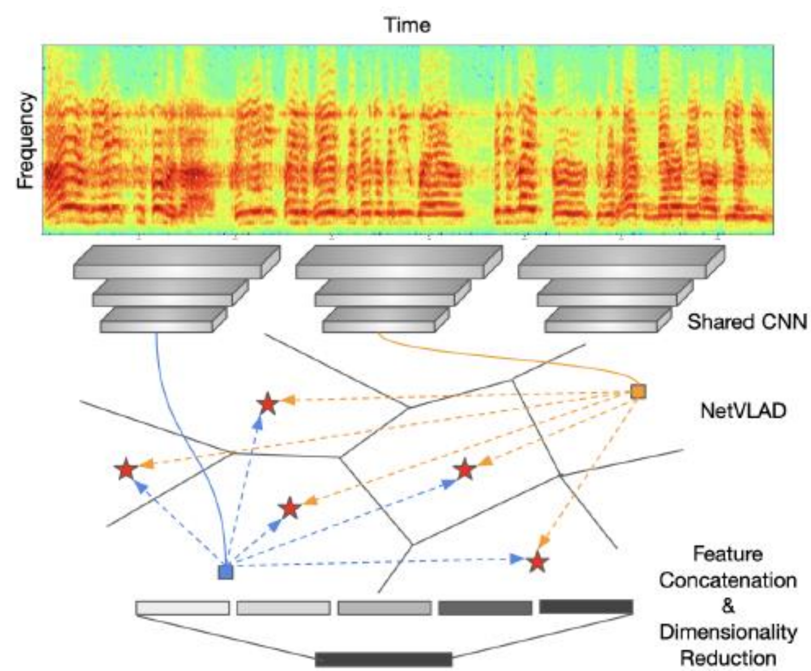
Attention Mechanism

[Chowdhury, et al., ICASSP'18]



NetVLAD

[Xie, et al., ICASSP'19]



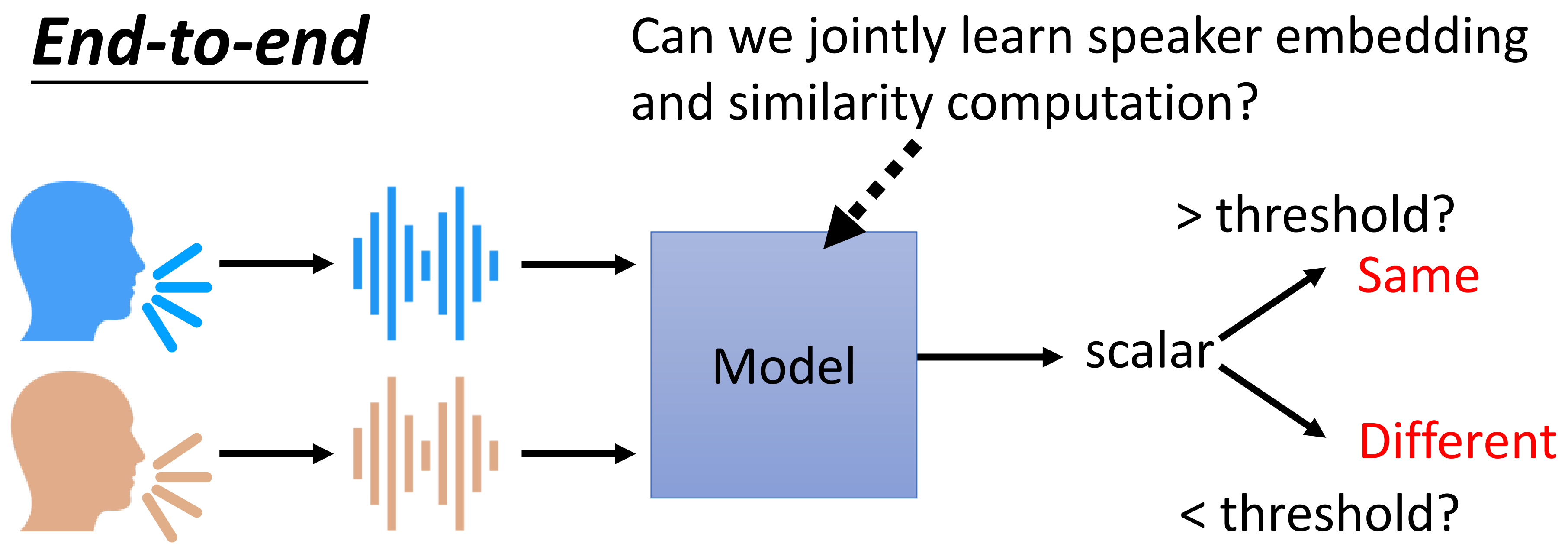
VLAD = Vector of Locally Aggregated Descriptors

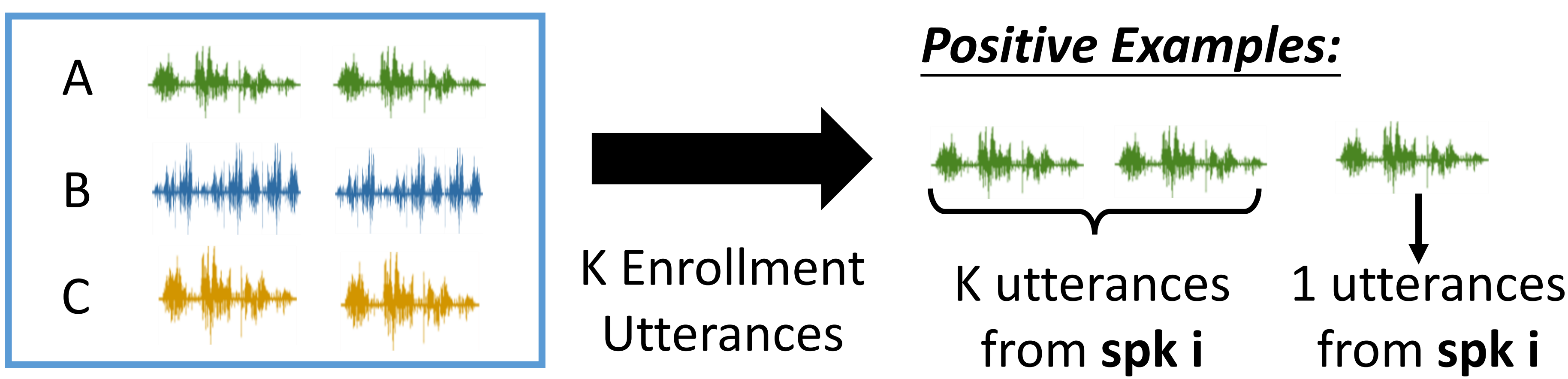
Outline

Task Introduction

Speaker Embedding

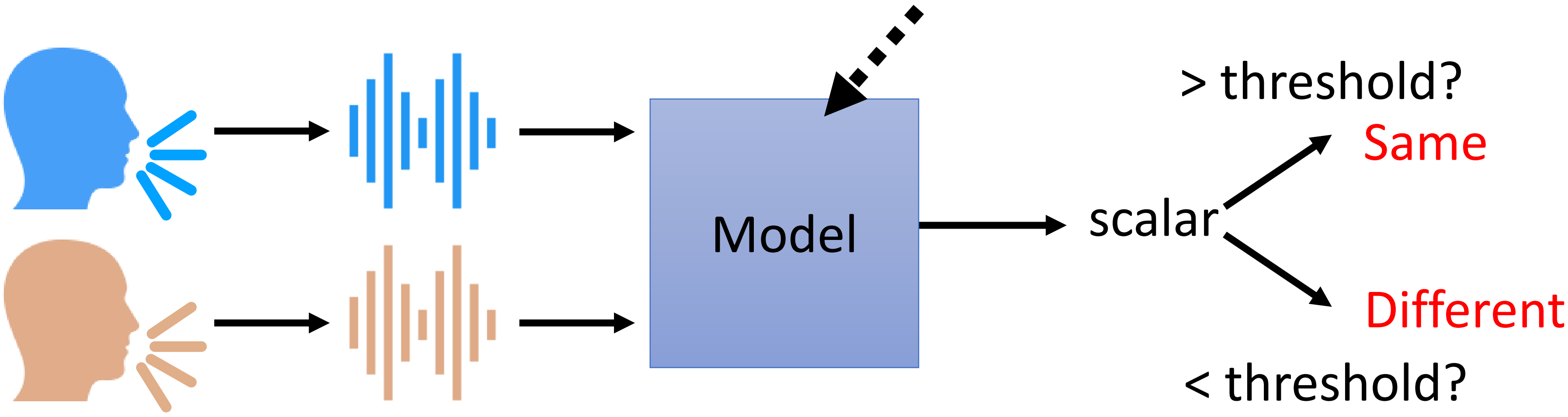
End-to-end

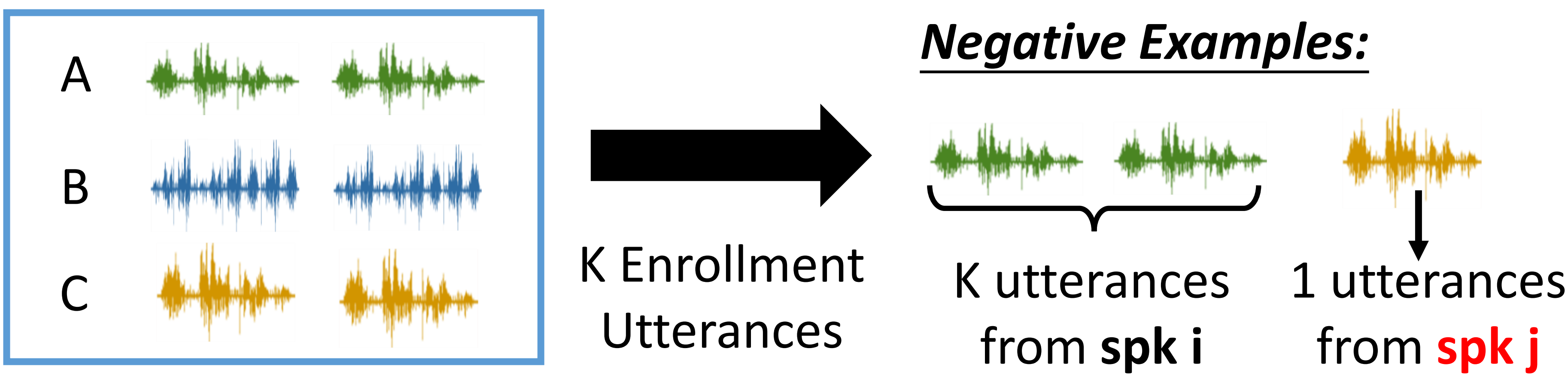

End-to-end
Can we jointly learn speaker embedding and similarity computation?
Model
scalar
> threshold?
Same
Different
< threshold?


A
B
C
K Enrollment Utterances
Positive Examples:
K utterances from spk i
1 utterances from spk i

End-to-end

Can we jointly learn speaker embedding and similarity computation?





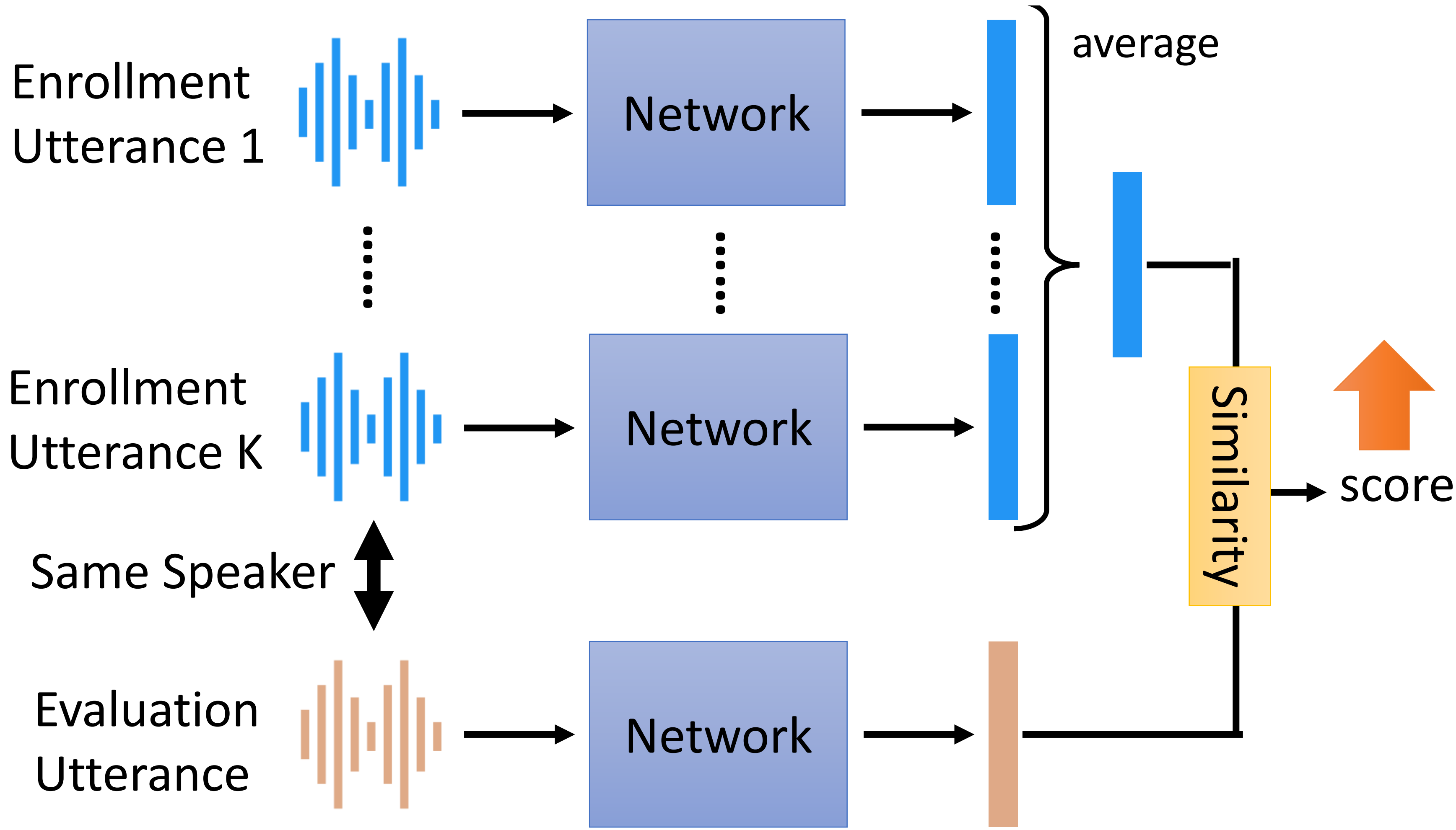
Also refer to generalized end-to-end (GE2E) [Wan, et al., ICASSP'18]

End-to-end

[Heigold, et al., ICASSP'16]

Table 1: Data set statistics.

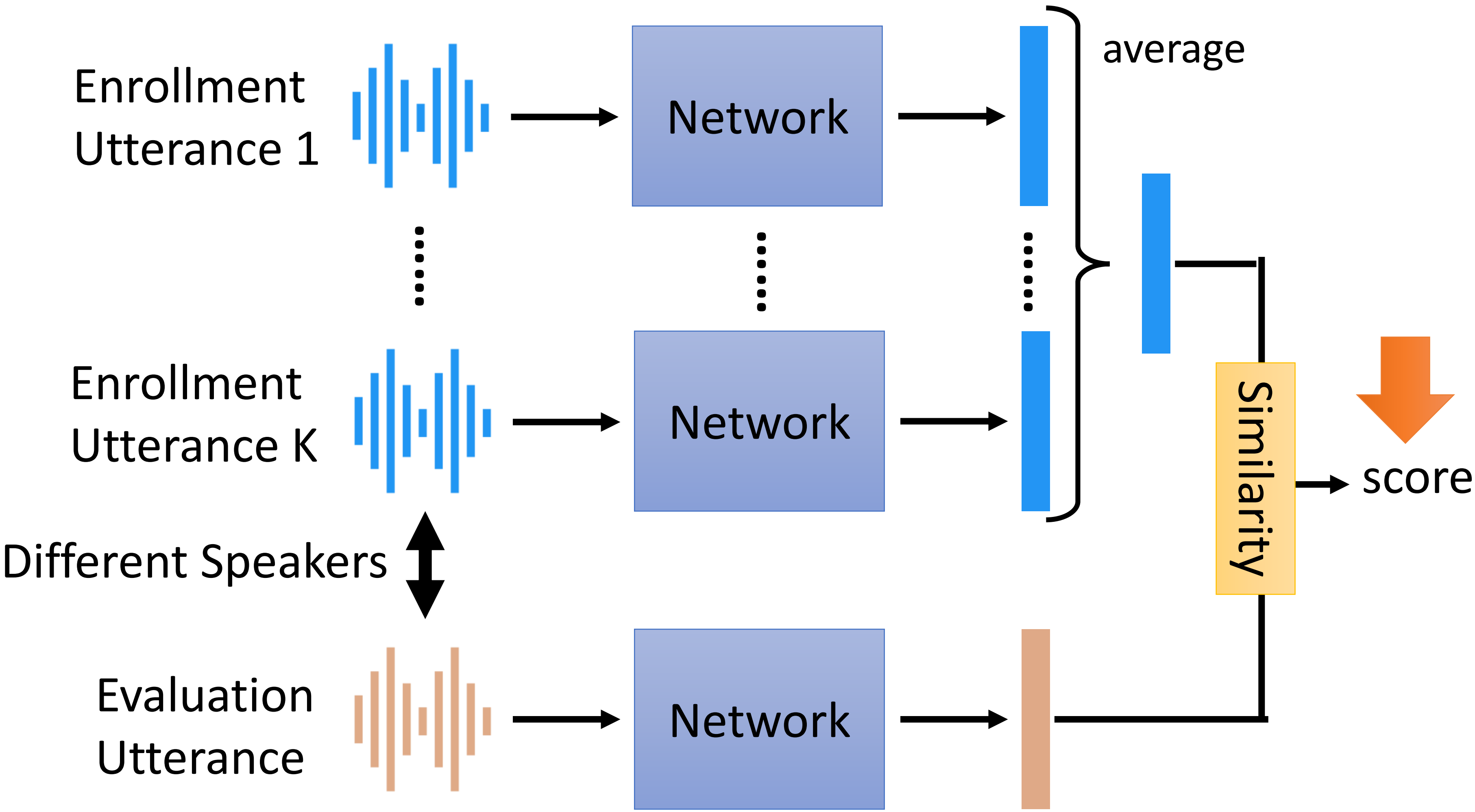
	#utterances (#augmented)	#speakers	#utts / spk
train_2M	2M (9M)	4k	>500
train_22M	22M (73M)	80k	>150
enrollment	18k	3k	1-9
evaluation	20k	3k	3-5



End-to-end

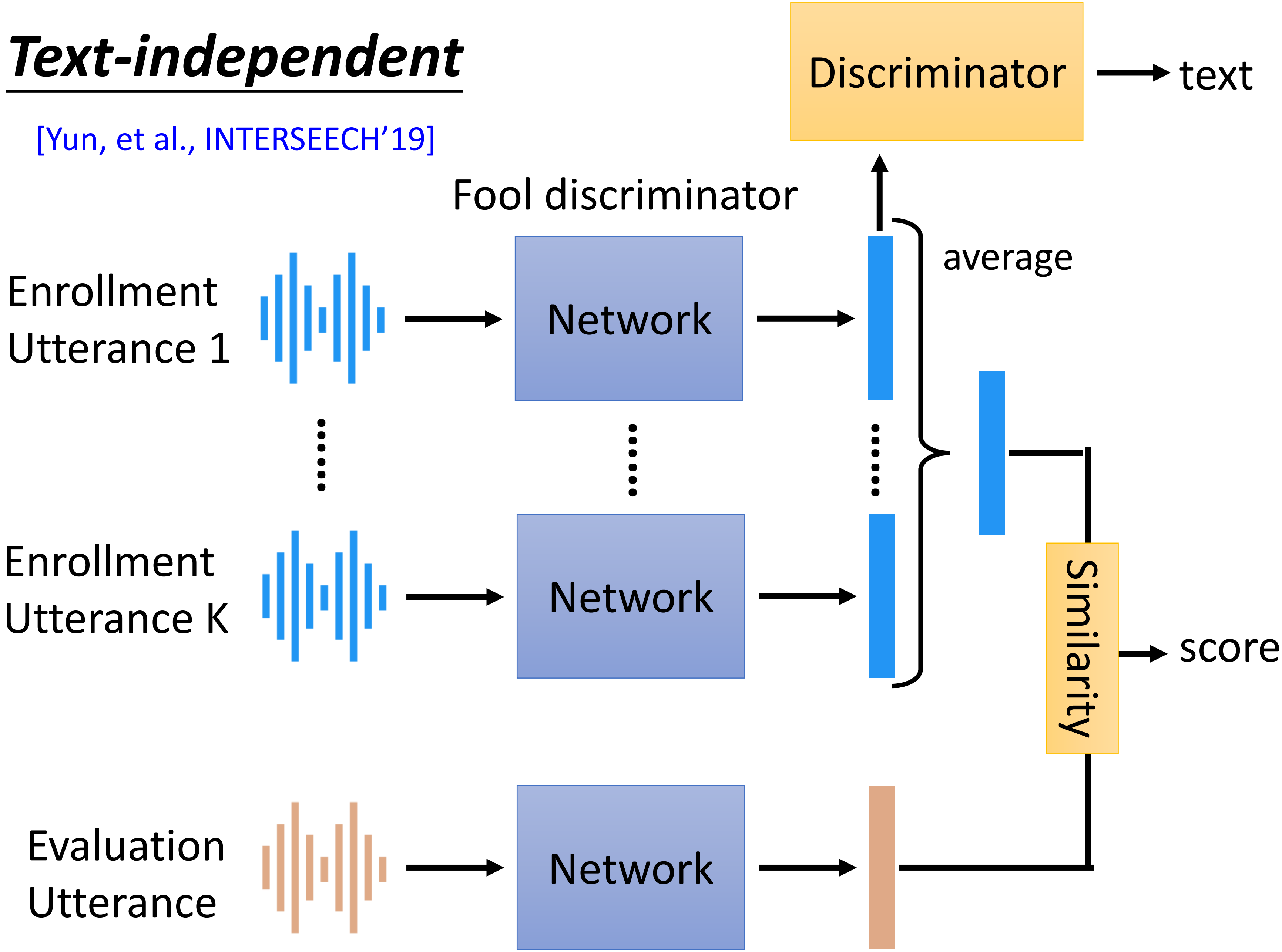
[Heigold, et al., ICASSP'16]

Text-dependent v.s. *Text-independent*



Text-independent

[Yun, et al., INTERSEECH’19]



Concluding Remarks

Task Introduction

Speaker Embedding

End-to-end

Reference

- [Variani, et al., ICASSP’14] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, Javier Gonzalez-Dominguez, Deep neural networks for small footprint text-dependent speaker verification, ICASSP, 2014
- [Heigold, et al., ICASSP’16] Georg Heigold, Ignacio Moreno, Samy Bengio, Noam Shazeer, End-to-End Text-Dependent Speaker Verification, ICASSP, 2016
- [Snyder, et al., ICASSP’18] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, Sanjeev Khudanpur, X-Vectors: Robust DNN Embeddings for Speaker Recognition, ICASSP, 2018
- [Wan, et al., ICASSP’18] Li Wan, Quan Wang, Alan Papir, Ignacio Lopez Moreno, Generalized End-to-End Loss for Speaker Verification, ICASSP, 2018
- [Yun, et al., INTERSEECH’19] Sungrack Yun, Janghoon Cho, Jungyun Eum, Wonil Chang, Kyuwoong Hwang, An End-to-End Text-independent Speaker Verification Framework with a Keyword Adversarial Network, INTERSPEECH, 2019

Reference

- [Nagrani, et al., INTERSPEECH’17] Arsha Nagrani, Joon Son Chung, Andrew Zisserman, VoxCeleb: a large-scale speaker identification dataset, INTERSPEECH, 2017.
- [Chung, et al., INTERSPEECH’18] Joon Son Chung, Arsha Nagrani, Andrew Zisserman, VoxCeleb2: Deep Speaker Recognition, INTERSPEECH, 2018
- [Xie, et al., ICASSP’19] Weidi Xie, Arsha Nagrani, Joon Son Chung, Andrew Zisserman, Utterance-level Aggregation For Speaker Recognition In The Wild, ICASSP, 2019
- [Chowdhury, et al., ICASSP’18] F A Rezaur Rahman Chowdhury, Quan Wang, Ignacio Lopez Moreno, Li Wan, Attention-Based Models for Text-Dependent Speaker Verification, ICASSP, 2018