

# Speech Technology: Frontiers and Applications

*End-to-end neural network based speaker recognition*

**Xiangang Li, Guoguo Chen**



# Outline

---

- 1 Introduction
- 2 Speaker recognition: classic methods
- 3 Speaker recognition: end-to-end approaches
- 4 Speaker recognition: related research topics
- 5 Homework

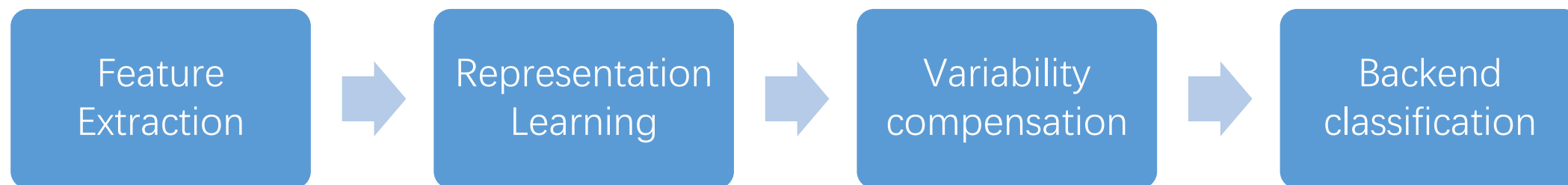
- Speech signal contains content information, and other paralinguistic speech attribute information
  - Speaker, emotion, channel, ...
- How to recognize?
  - Utterance based supervised learning with flexible duration
  - Text dependent or text independent

- Speech signal contains content information, and other paralinguistic speech attribute information
  - Speaker, emotion, channel, ...
- How to recognize?
  - Utterance based supervised learning with flexible duration
  - Text dependent or text independent
- Tasks
  - Speaker identification (closed set)
  - Speaker verification (open set)

- 1 Introduction
- 2 Speaker recognition: classic methods
- 3 Speaker recognition: end-to-end approaches
- 4 Speaker recognition: related research topics
- 5 Homework

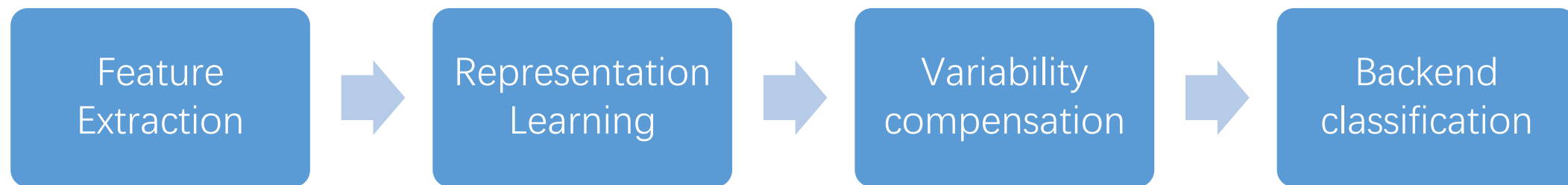
## 2 Speaker recognition: classic methods

- General framework



## 2.1 Feature extraction

- General framework



- Classic feature extraction methods
- MFCC, PLP, BNF, etc.

## 2.2 Representation Learning

- GMM-UBM
  - GMM as the generative model to model the feature representation
  - Perform model adaptation (such as MAP adaptation) on the GMM based **universal background model (UBM)**
    - Concatenating mean vector from all GMM components to get the **GMM Mean supervector**
- Factor analysis on the supervector
  - **Identity vector (i-vector)**

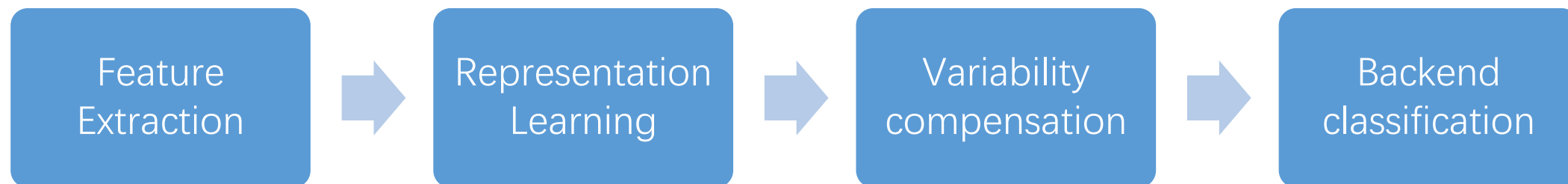
$$M = m + Tw$$

- $T$ : factor loading matrix
  - $w$ : i-vector
- DNN-ivector
  - Y. Lei, L. Ferrer, M. McLaren, et al. “A novel scheme for speaker recognition using a phonetically-aware deep neural network” , ICASSP 2014



## 2.3 Variability compensation

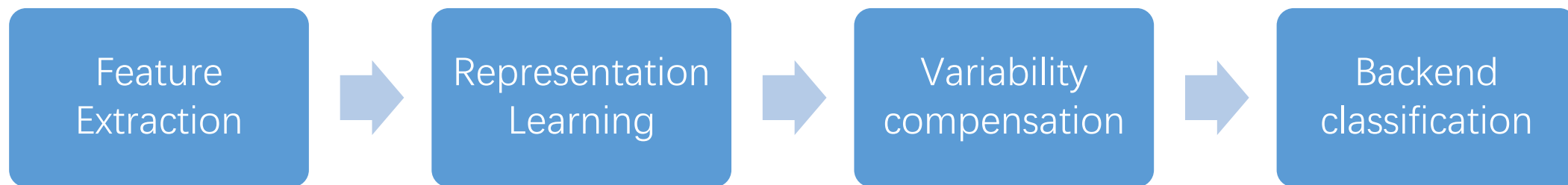
- General framework



- Classic variability compensation methods used in speaker recognition
- LDA, NDA, etc.

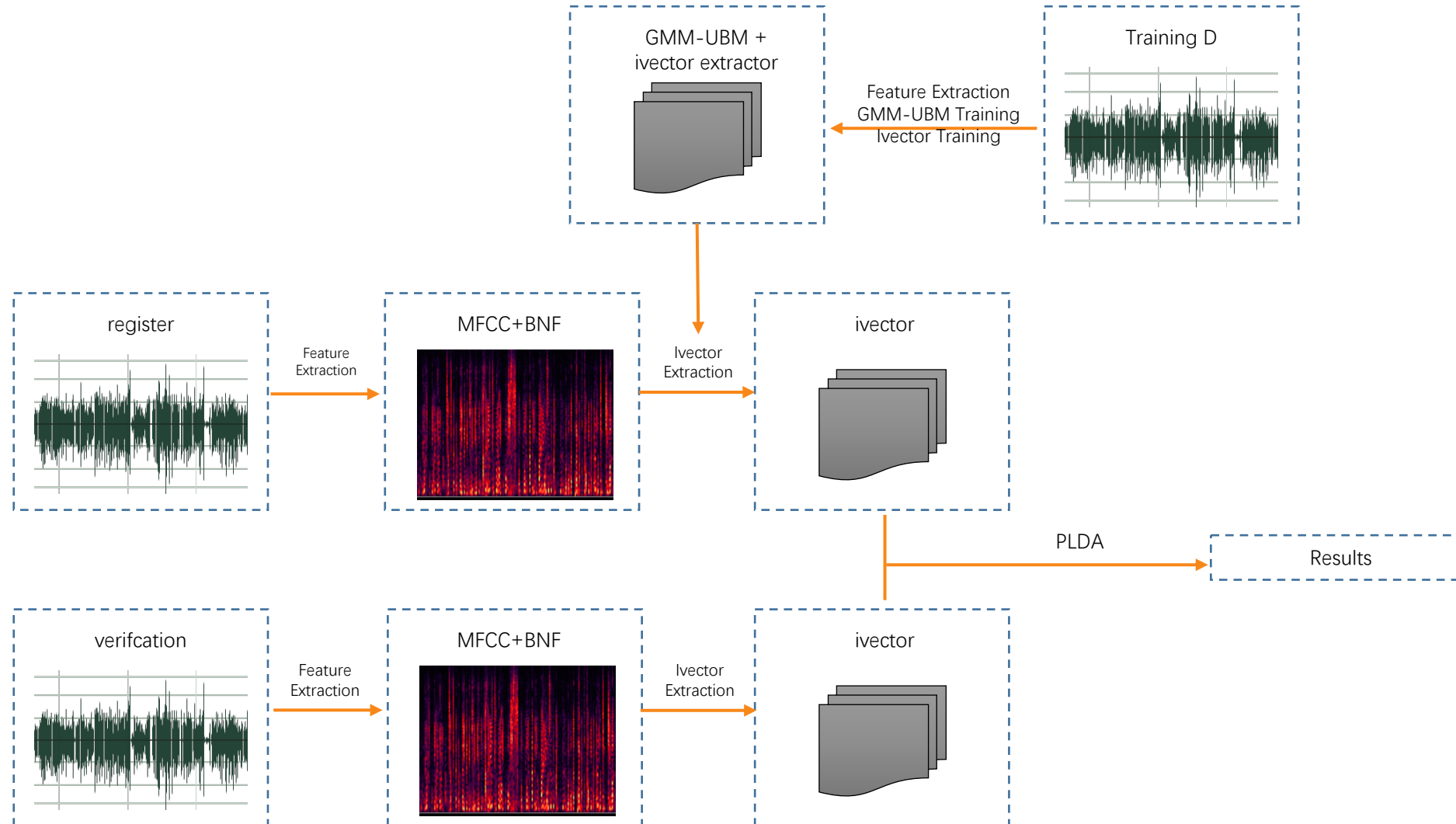
## 2.4 Backend Classification

- General framework



- Classic classification methods used in speaker recognition
- SVM, PLDA, NN, Cosine similarity, etc.

# 2 Speaker recognition: classic methods



- 1 Introduction
- 2 Speaker recognition: classic methods
- 3 Speaker recognition: end-to-end approaches**
- 4 Speaker recognition: related research topics
- 5 Homework

# 3 Speaker recognition: end-to-end approaches

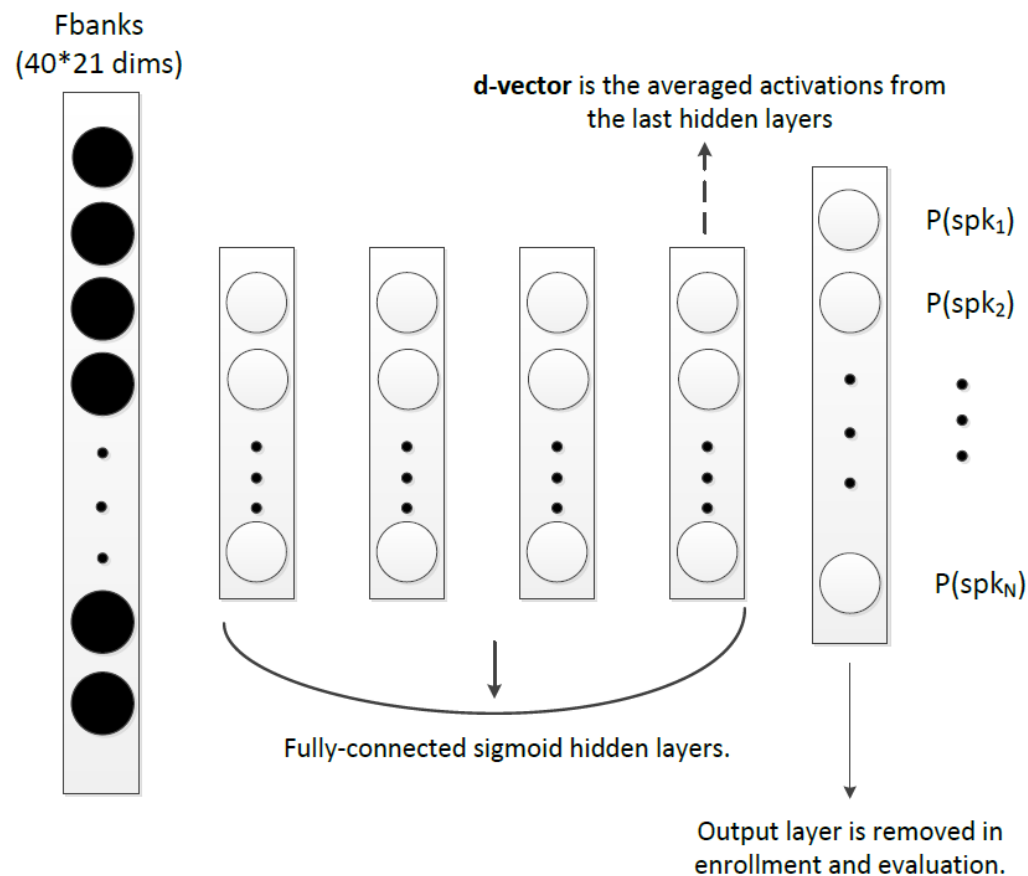
- Network pipeline



- Frame-level feature input: FBANK, MFCC etc.
- Frame-level neural extractor: representation learning
- Loss: **end-to-end optimized** with the global loss function

# D-vector

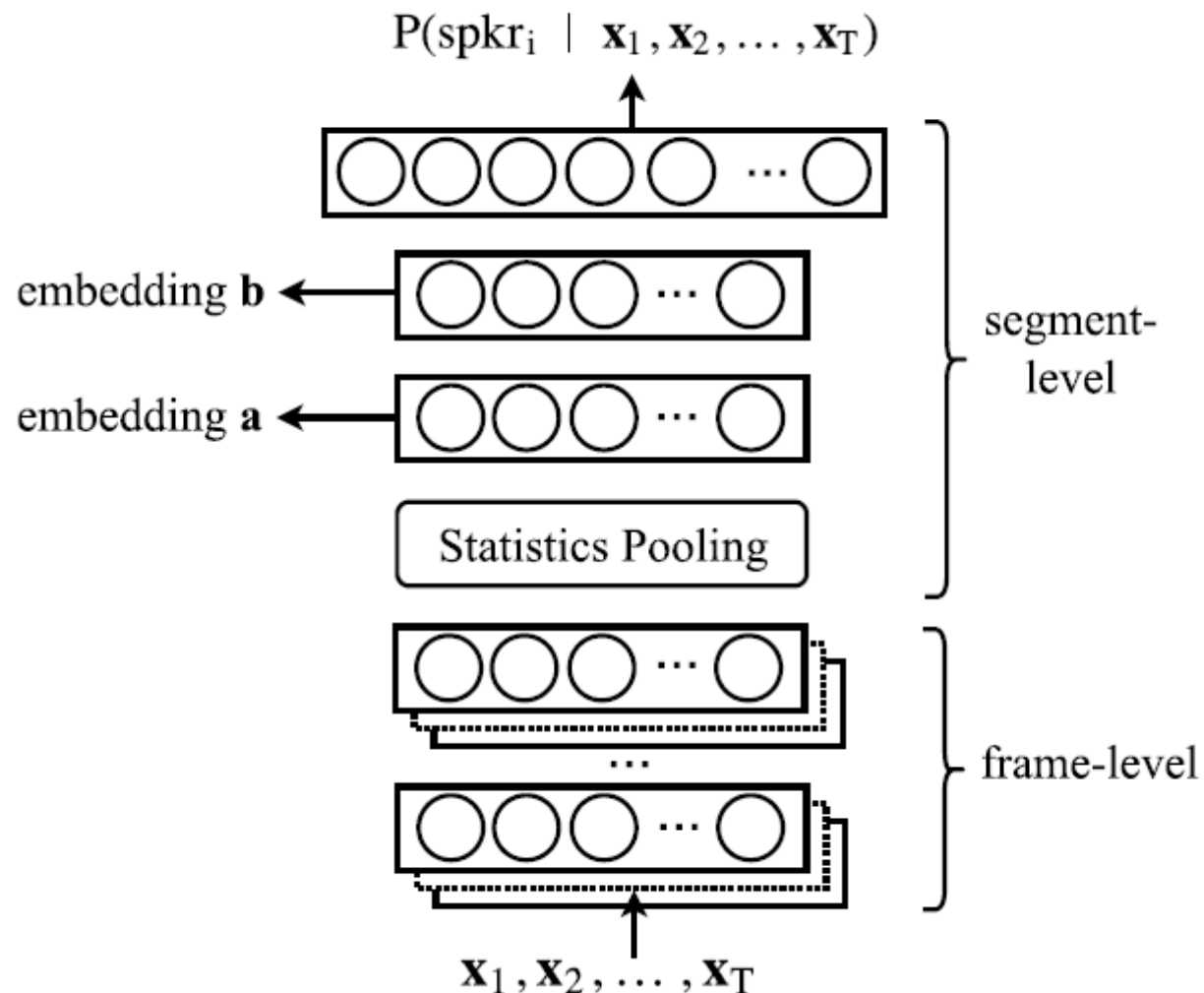
1. Ehsan Variani et al. “Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification”. ICASSP. 2014.
2. Lantian Li et al. “Deep speaker vectors for semi text-independent speaker verification”. arXiv: 1505.06427(2015).
3. Yuan Liu et al. “Deep feature for text-dependent speaker verification”. Speech Communication 2015.



- Statistics pooling

- Calculates the mean vector  $\mu$  and standard deviation vector  $\sigma$  over frame level features  $h_t (t = 1, \dots, T)$

- David Snyder, Daniel Garcia-Romero, Daniel Povey and Sanjeev Khudanpur. "Deep Neural Network Embeddings for Text-Independent Speaker Verification". Interspeech 2017.
- David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, Sanjeev Khudanpur. "X-vectors: Robust DNN Embeddings for Speaker Recognition". ICASSP 2018.
- Yingke Zhu, Tom Ko, David Snyder, Brian Mak, Daniel Povey. "Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification". Interspeech 2018.



# 3.1 Network structure

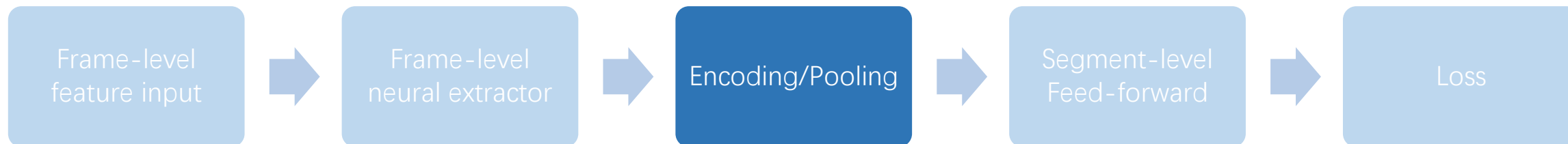


- Network structure

- FF-DNN, RNN/LSTM, CNN, TDNN



## 3.2 Encoding/Pooling



- **Conventional approaches**

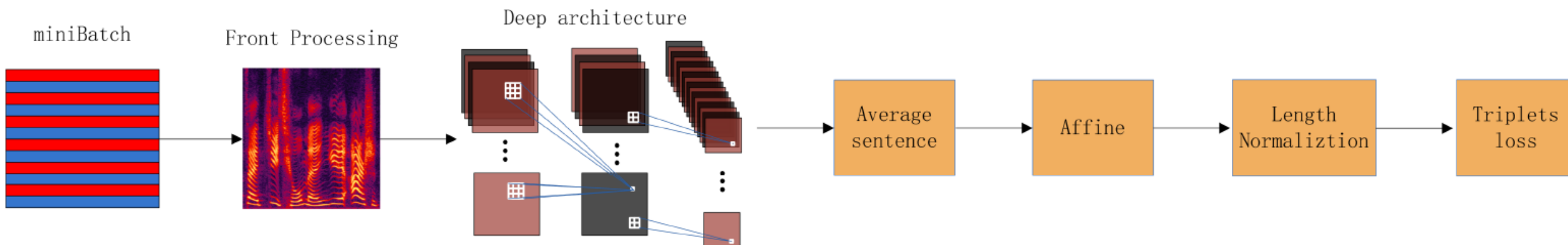
- Average: an utterance-level embedding is derived by averaging the frame-level neural network hidden layer output. (D-vector)

## 3.2 Encoding/Pooling

- Pooling layer

- Temporal pooling (mean)

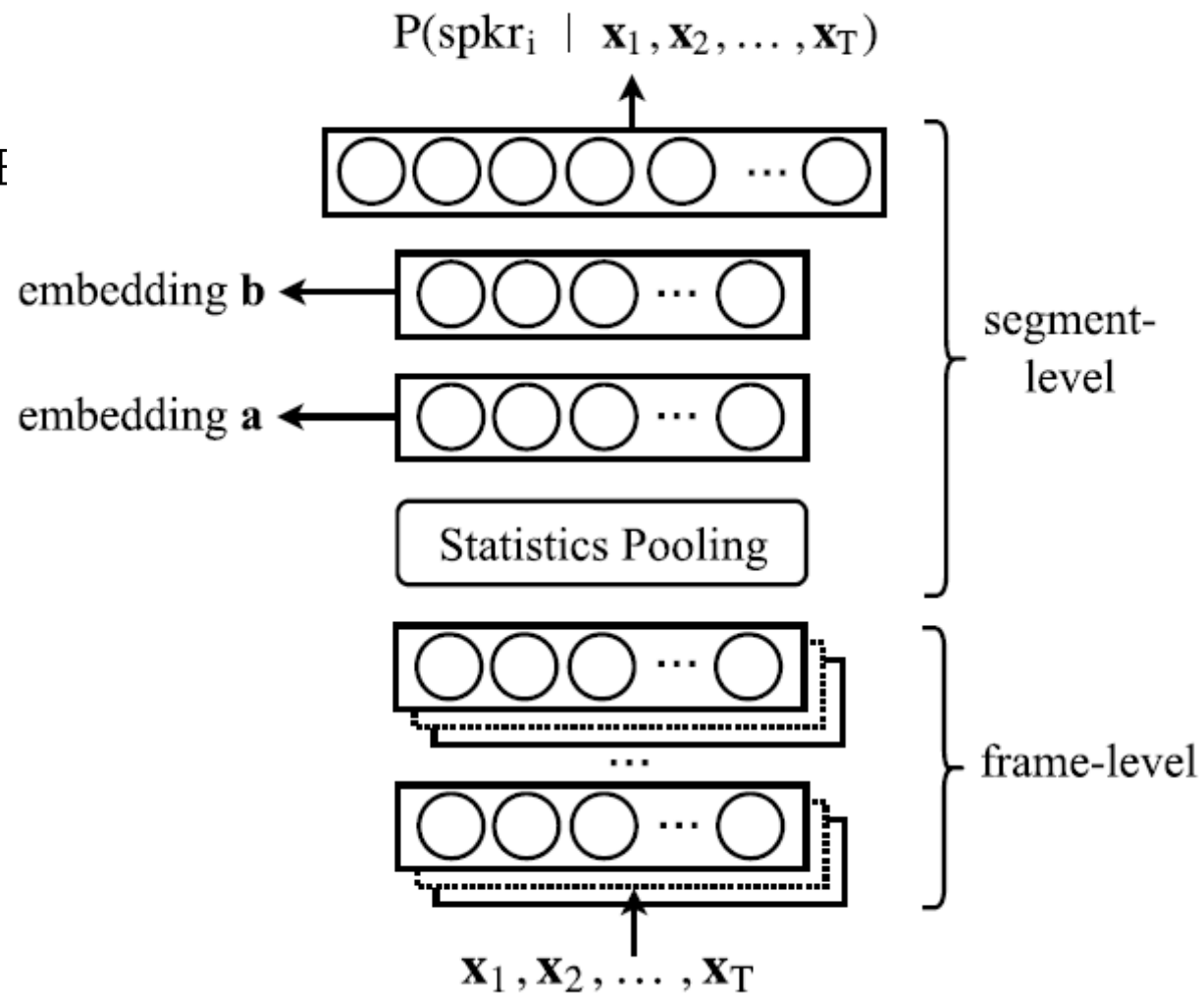
- Chao Li, et al. “Deep Speaker: an End-to-End Neural Speaker Embedding System”  
arXiv: 1705.02304



## 3.2 Encoding/Pooling

- Pooling layer

- Temporal pooling (mean)
  - Chao Li, et al. "Deep Speaker: an End-to-End Speaker Verification System" arXiv: 1705.02304
- Statistics pooling (mean + std)
  - X-vector



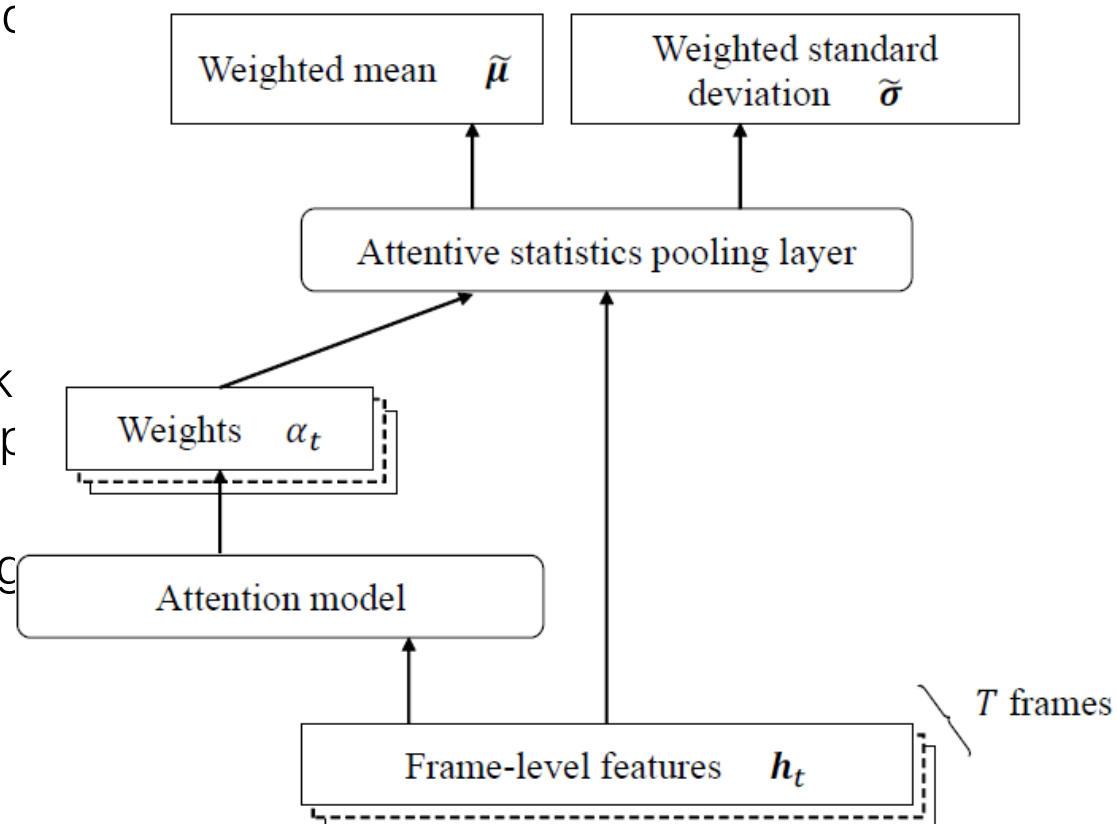
## 3.2 Encoding/Pooling

- Pooling layer
  - Temporal pooling (mean)
    - Chao Li, et al. “Deep Speaker: an End-to-End Neural Speaker Embedding System”  
arXiv: 1705.02304
  - Statistics pooling (mean + std)
    - X-vector
  - Attentive statistics pooling (mean + std)
    - Yingke Zhu, Tom Ko, David Snyder, Brian Mak, Daniel Povey. “Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification” . Interspeech 2018.
    - Koji Okabe, et al. “Attentive Statistics Pooling for Deep Speaker Embedding” .  
arXiv: 1803.10963. 2018

## 3.2 Encoding/Pooling

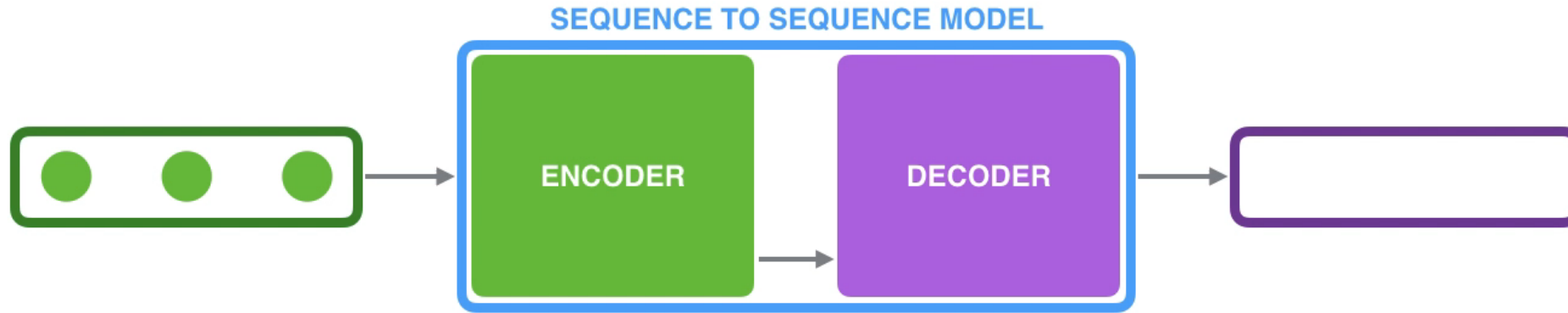
- Pooling layer

- Temporal pooling (mean)
  - Chao Li, et al. “Deep Speaker: an End-to-Enc arXiv: 1705.02304
- Statistics pooling (mean + std)
  - X-vector
- Attentive statistics pooling (mean + std)
  - Yingke Zhu, Tom Ko, David Snyder, Brian Mak Speaker Embeddings for Text-Independent Sp 2018.
  - Koji Okabe, et al. “Attentive Statistics Pooling arXiv: 1803.10963. 2018



## 3.2 Encoding/Pooling

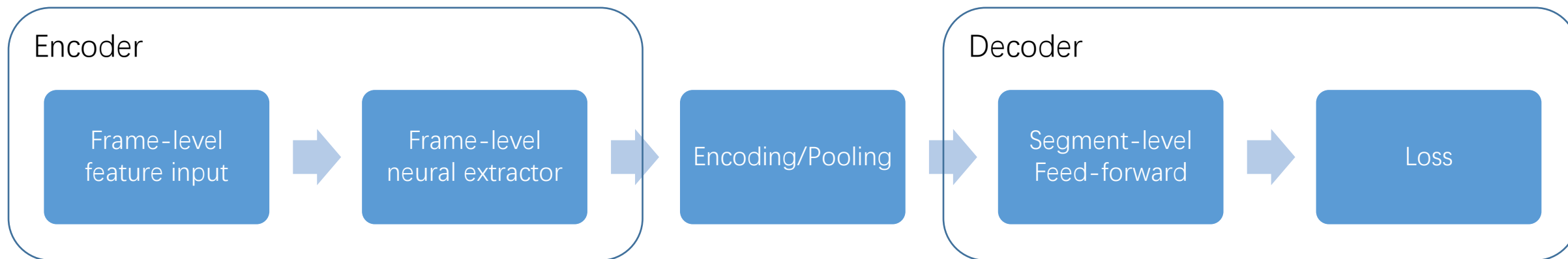
- Sequence-to-sequence modeling



- Three key components : Encoder -> Context -> Decoder
- **Speaker recognition**
  - Arbitrary-length input sequence, and 1 output target
  - Context vector -> speaker representation

## 3.2 Encoding/Pooling

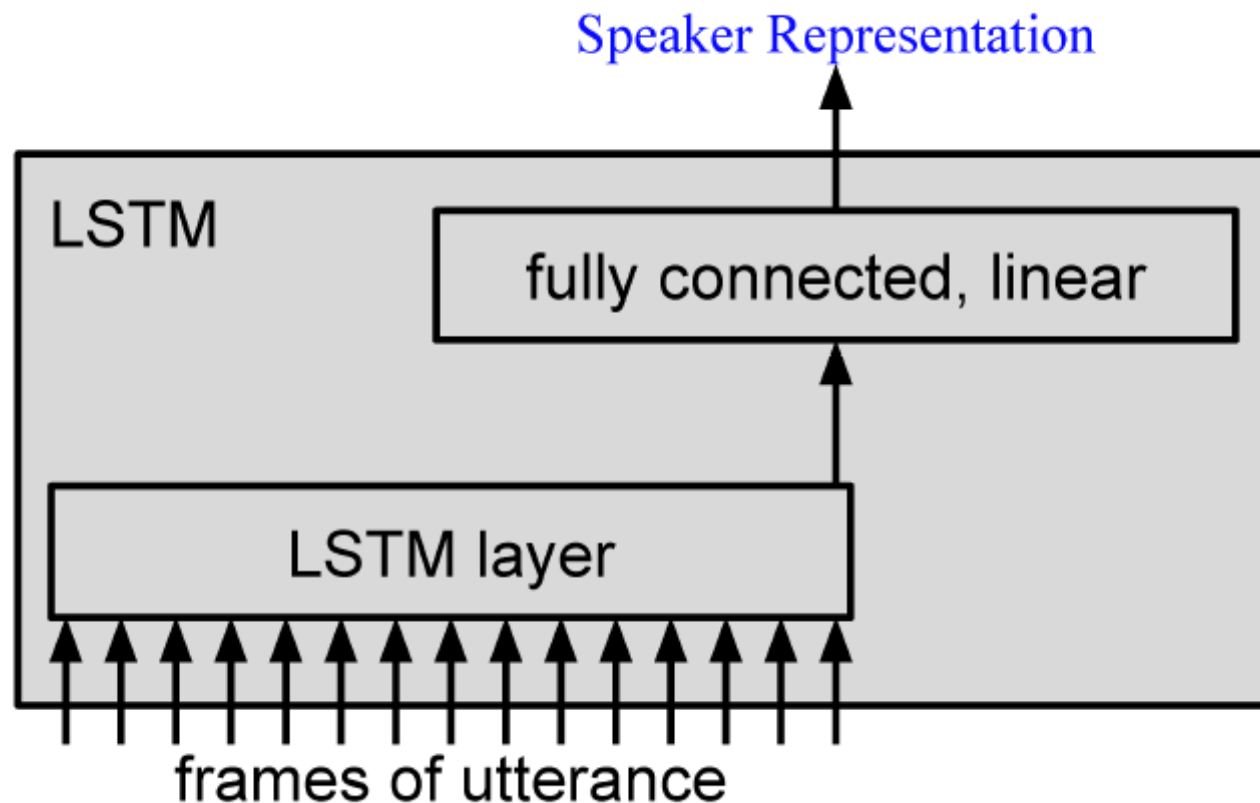
- Network pipeline



## 3.2 Encoding/Pooling

- Encoding

- RNN/LSTM encoding
- RNN/LSTM + Attention





# 3.3 Loss function

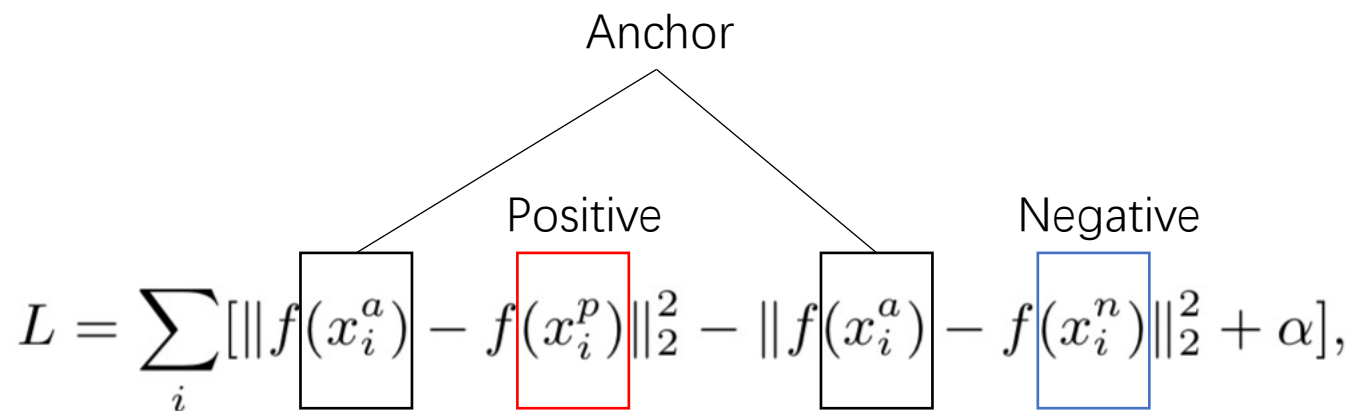


- Standard softmax loss

$$\mathcal{L}_S = \frac{1}{N} \sum_{i=1}^N -\log \frac{e^{\mathbf{w}_{y_i}^T \mathbf{x}_i + b_{y_i}}}{\sum_{j=1}^c e^{\mathbf{w}_j^T \mathbf{x}_i + b_j}},$$

# 3.3 Loss function

- Standard softmax loss
- Triplet loss
  - Chao Li, et al. “Deep Speaker: an End-to-End Neural Speaker Embedding System”  
arXiv: 1705.02304



The diagram shows the triplet loss formula with labels for the different components. The word "Anchor" is positioned above the first boxed term  $f(x_i^a)$ . The word "Positive" is positioned above the second boxed term  $f(x_i^p)$ , which is highlighted with a red border. The word "Negative" is positioned above the fourth boxed term  $f(x_i^n)$ , which is highlighted with a blue border. The formula is:

$$L = \sum_i [\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha],$$

# 3.3 Loss function

- Standard softmax loss
- Triplet loss
  - Chao Li, et al. “Deep Speaker: an End-to-End Neural Speaker Embedding System”  
arXiv: 1705.02304
- Center loss
  - Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. “A Discriminative Feature Learning Approach for Deep Face Recognition” . ECCV 2016

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_S + \lambda \mathcal{L}_C \\ &= -\sum_{i=1}^m \log \frac{e^{W_{y_i}^T \mathbf{x}_i + b_{y_i}}}{\sum_{j=1}^n e^{W_j^T \mathbf{x}_i + b_j}} + \frac{\lambda}{2} \sum_{i=1}^m \|\mathbf{x}_i - \mathbf{c}_{y_i}\|_2^2\end{aligned}$$

# 3.3 Loss function

- Standard softmax loss
- Triplet loss
  - Chao Li, et al. “Deep Speaker: an End-to-End Neural Speaker Embedding System”  
arXiv: 1705.02304
- Center loss
  - Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. “A Discriminative Feature Learning Approach for Deep Face Recognition” . ECCV 2016
- SphereFace(Angular-softmax) Loss
  - Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, Le Song.  
“SphereFace: Deep Hypersphere Embedding for Face Recognition” . CVPR 2017
- Large Margin Softmax Loss
  - Yi Liu, Liang He, Jia Liu. “Large Margin Softmax Loss for Speaker Verification” .  
arXiv:1904.03479. 2019

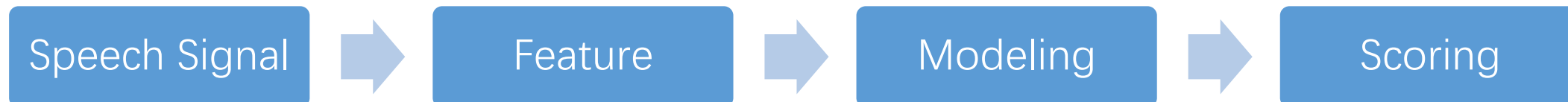
# 3.3 Loss function

- Standard softmax loss
- Triplet loss
  - Chao Li, et al. “Deep Speaker: an End-to-End Neural Speaker Embedding System”  
arXiv: 1705.02304
- Center loss
  - Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. “A Discriminative Feature Learning Approach for Deep Face Recognition” . ECCV 2016
- Other Loss from face recognition
  - ...

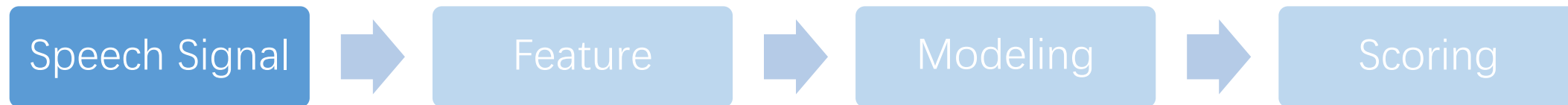
- 1 Introduction
- 2 Speaker recognition: classic methods
- 3 Speaker recognition: end-to-end approaches
- 4 Speaker recognition: related research topics
- 5 Homework

# 4.1 Robust modeling

- Room reverberation: convolutional noise
- Complex environmental noises: additional noise
- Robust modeling from different stage



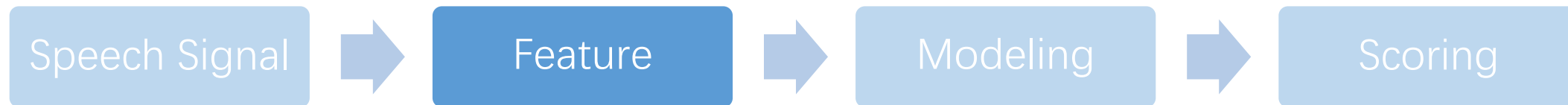
# 4.1 Robust modeling



- Signal level
  - Dereverberation
  - DNN based denoising
  - Beamforming for multi-channel speech enhancement



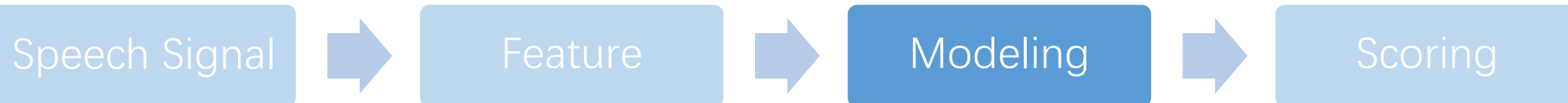
# 4.1 Robust modeling



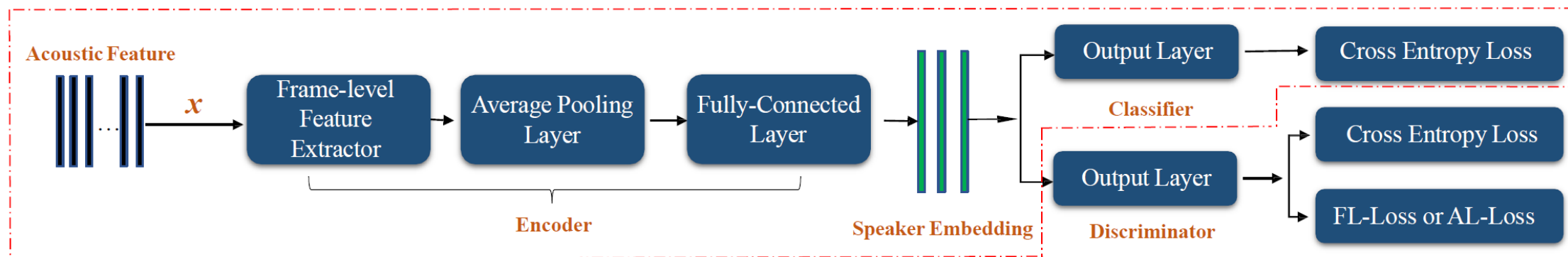
- Feature level

- Power-normalized cepstral coefficients (PNCC)
- DNN extracted features: DNN bottleneck features
- ...

# 4.1 Robust modeling

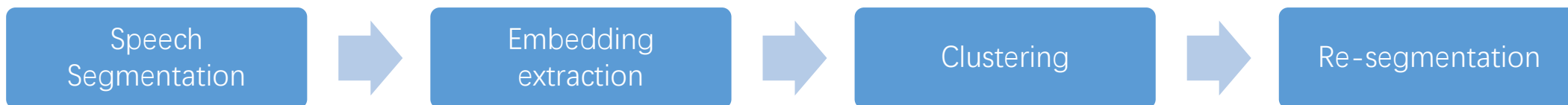


- Model level
  - Data augmentation and multi-condition training
  - Multi-task networks

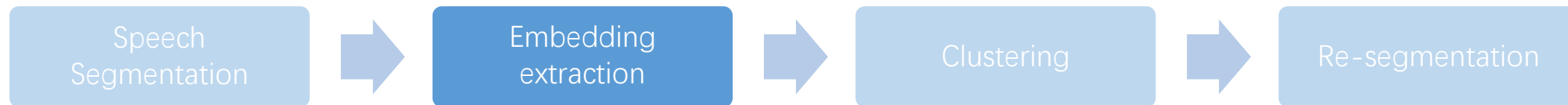


## 4.2 Speaker diarisation

- Speaker diarisation: “who spoke when”
- Classic framework
  - Speech Segmentation module: removes the non-speech parts (VAD), and divides the input utterance into small and speaker-homogeneous segments
  - Similarity measurement module: compute similarity of any two segments, where speaker-discriminative embeddings such as i-vectors, d-vectors are extracted
  - Clustering module: determines the number of speakers, and assigns speaker identities to each segment
  - Re-segmentation module: further refines the diarization results by enforcing additional constraints

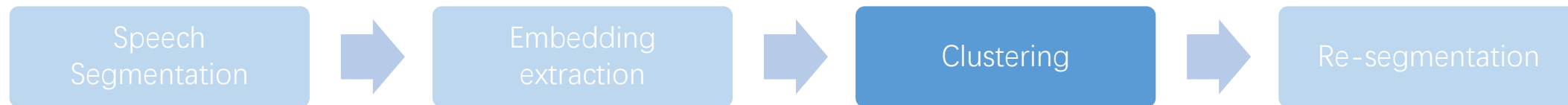


## 4.2 Speaker diarisation



- For embedding extraction module, many work has shown that performance can be significantly improved by introducing neural networks
  - D. Garcia-Remero, et al. “Speaker diarisation using deep neural network embeddings” . ICASSP 2017.
  - Q. Wang et al. “Speaker diarisation with LSTM” . ICASSP 2018.
  - G. Sell et al. “Diarisation is hard: some experiences and lessons for the JHU team in Inaugural DIHARD challenge” . Interspeech 2018
  - Q. Lin et al. “LSTM based similarity measurement with spectral clustering for speaker diarisation” . Interspeech 2019

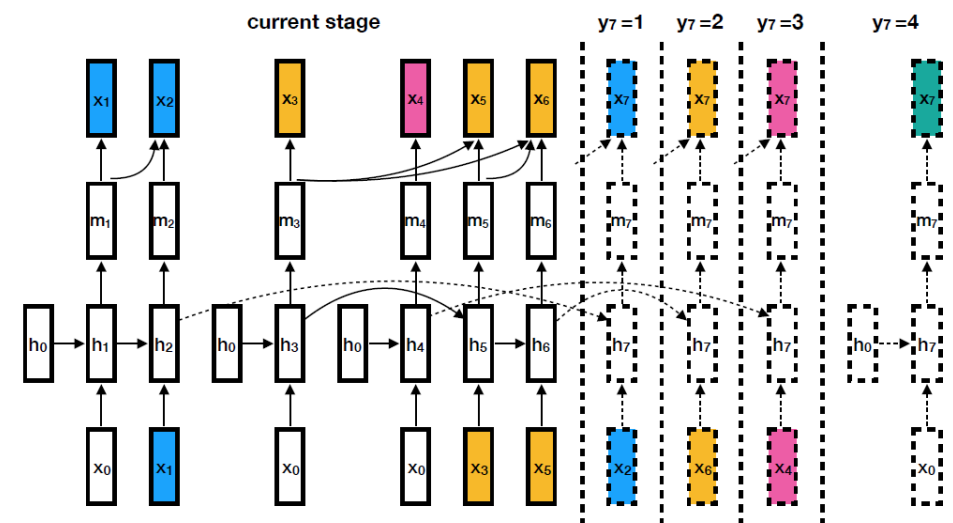
## 4.2 Speaker diarisation



- Clustering

- Classic method: Agglomerative hierarchical clustering (AHC)
- UIS-RNN

- A. Zhang, Q. Wang, Z. Zhu, J. Paisley, C. Wang, “Fully supervised speaker diarisation” . ICASSP 2019



**Fig. 2.** Generative process of UIS-RNN. Colors indicate labels for speaker segments. There are four options for  $y_7$  given  $x_{[6]}, y_{[6]}$ .

## 4.3 Anti-spoofing

- Spoofing/presentation attacks for speaker verification system
  - Replayed recording
  - Synthesis audio: TTS or voice conversion
- ASVspoof Challenge
  - <https://www.asvspoof.org/>

---

Home License Download ASVspoof 2017 ASVspoof 2015 Login

ASVspoof 2019

Automatic Speaker Verification

Spoofing And Countermeasures Challenge

*Future horizons in spoofed/fake audio detection*

18th November: the slides used during Interspeech 2019 are available to **download**.

17th July: Release of ASVspoof 2019 real PA database. In this public release, we extended the simulated data used in the challenge with a real small set of audio files (2700) recorded and replayed in 3 different labs. To download it, please visit our **Licence page**.

5th June: Database Release via **Edinburgh DataShare**.

2nd May: ASRU 2019 Special Session **description**.

7th Feb: read the ASVspoof 2019 **press alert**.

15th Jan: the ASVspoof 2019 **evaluation plan v0.4** is now available (**ChangeLog**).

- 1 Introduction
- 2 Speaker recognition: classic methods
- 3 Speaker recognition: end-to-end approaches
- 4 Speaker recognition: related research topics
- 5 Homework

# 5 Homework

- Evaluate the loss function for speaker recognition using VoxCeleb datasets
  - The VoxCeleb Speaker Recognition Challenge (VoxSRC)
    - <http://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition2020.html>



Welcome to the 2020 VoxCeleb Speaker Recognition Challenge! The goal of this challenge is to probe how well current methods can recognize speakers from speech obtained 'in the wild'. The dataset is obtained from YouTube videos of celebrity interviews, consisting of audio from both professionally edited and red carpet interviews as well as more casual conversational audio in which background noise, laughter, and other artefacts are observed in a range of recording environments.

*The VoxCeleb Speaker Recognition Challenge is scheduled to be held as normal. We are monitoring the COVID-19 situation carefully and may extend the deadlines appropriately.*

Details for the 2019 challenge can be found [here](#). Congratulations to all the winners!



# 5 Homework

- Evaluate the loss function for speaker recognition using VoxCeleb datasets
  - The VoxCeleb Speaker Recognition Challenge (VoxSRC)
    - <http://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition2020.html>
- Task
  - Run the provided baseline code for voxceleb
    - [https://github.com/clovaai/voxceleb\\_trainer](https://github.com/clovaai/voxceleb_trainer)
  - Select the first 100-speakers' data for training and testing considering the dataset is very big
    - Of course, if you have a powerful machine, you can use all data
  - A report to compare more than 3 different loss function
    - e.g. softmax, GE2E, triplet

Thanks for your Attention

