Speaker Recognition with X-vectors and Keras

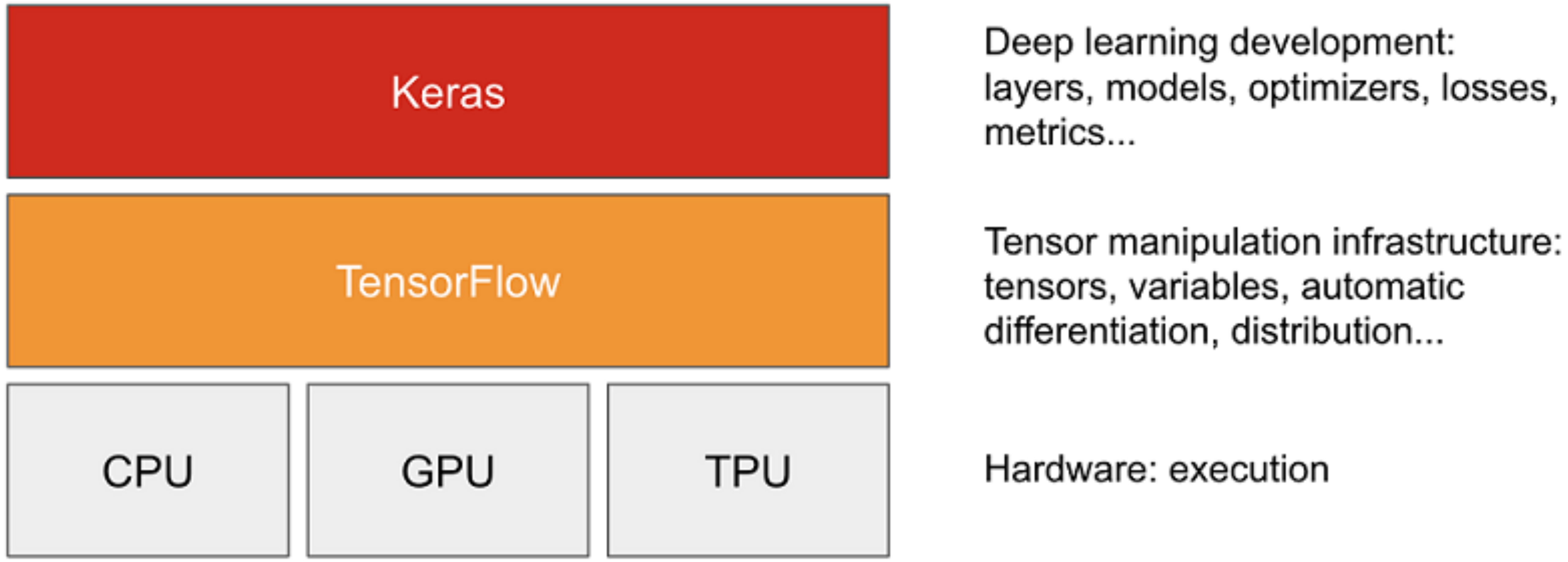
Aleksej Horvat - 10688536
16-03-2021

Research Questions

1. To what extent can speaker recognition be modelled using high-level instructions?
2. Which characteristics of Dialogue (Speech) can be used to recognise individual agents?
3. What is the accuracy of model?
4. How does the model compare to existing methods?
    1. How do we compare models

Keras

- High level framework
  - Describes model, layers, etc.
  - Less code
  - Scalable
  - Portable
- Baseline: 1D covnet as baseline
  - 98% test accuracy, score to beat



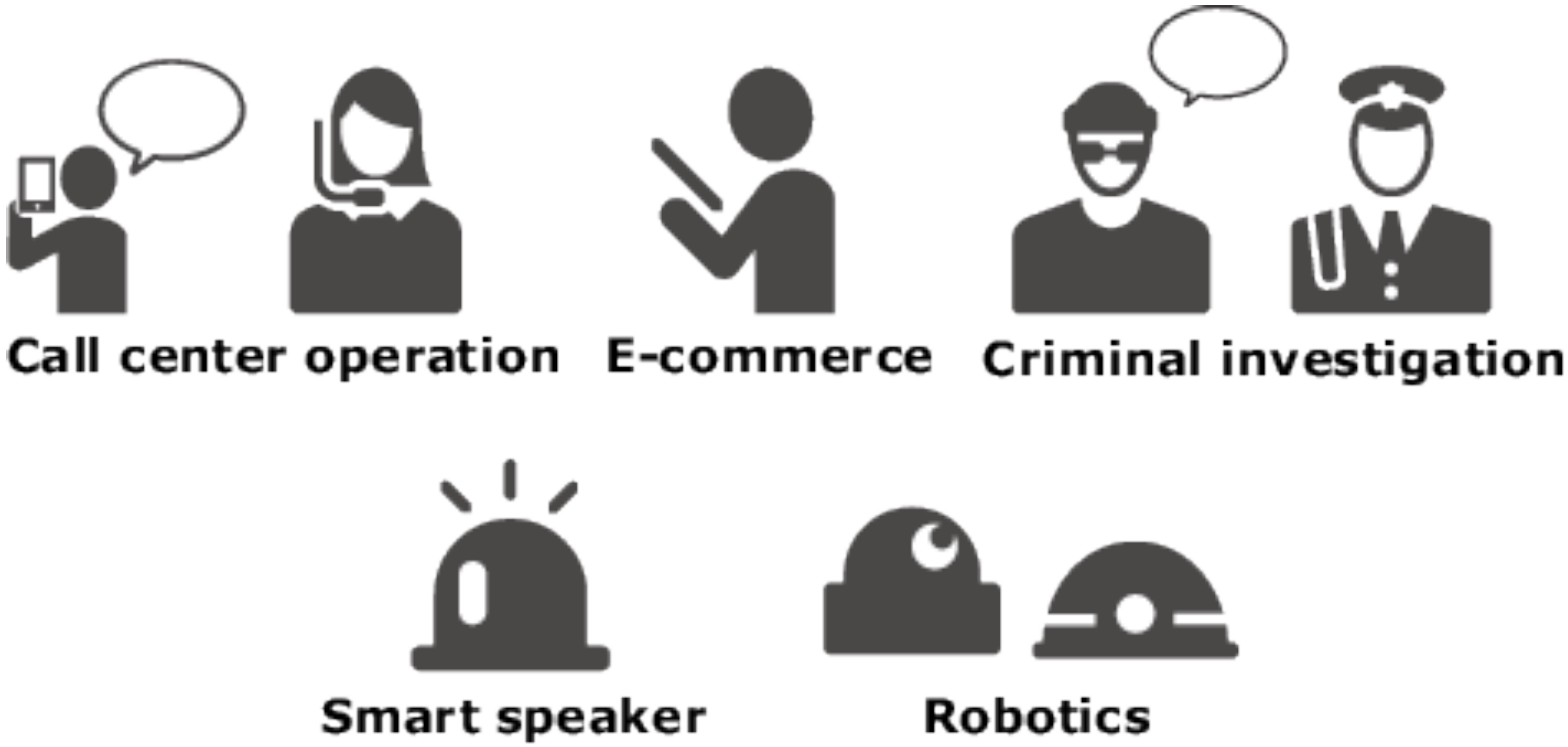
Deep learning development: layers, models, optimizers, losses, metrics...

Tensor manipulation infrastructure: tensors, variables, automatic differentiation, distribution...

Hardware: execution

Speaker Recognition Tasks

- Verification - security
- Identification
  - Personalised Responses
- Informational Retrieval

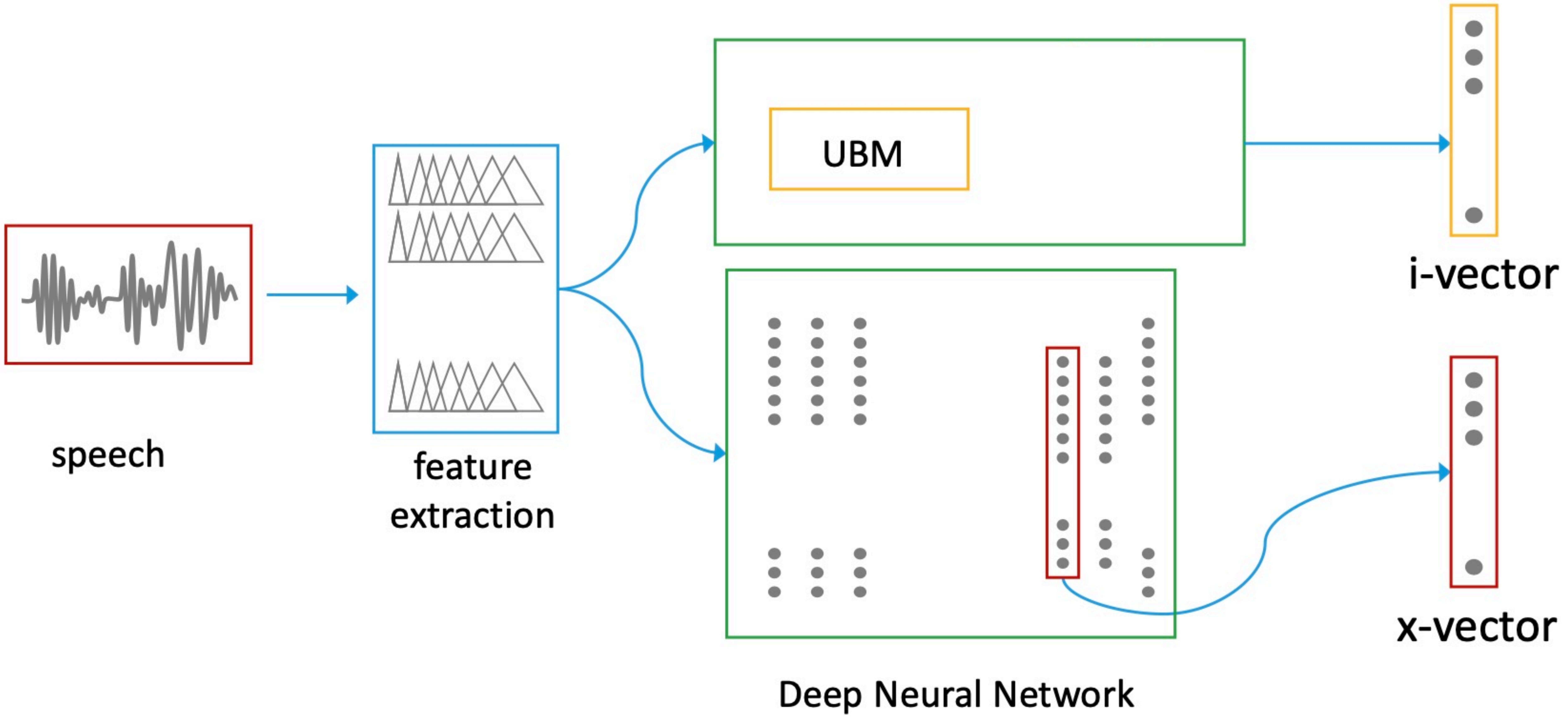


Approaches to Speaker Recognition

- Gaussian Mixture Models
- Adapted GMM-Universal Background Model

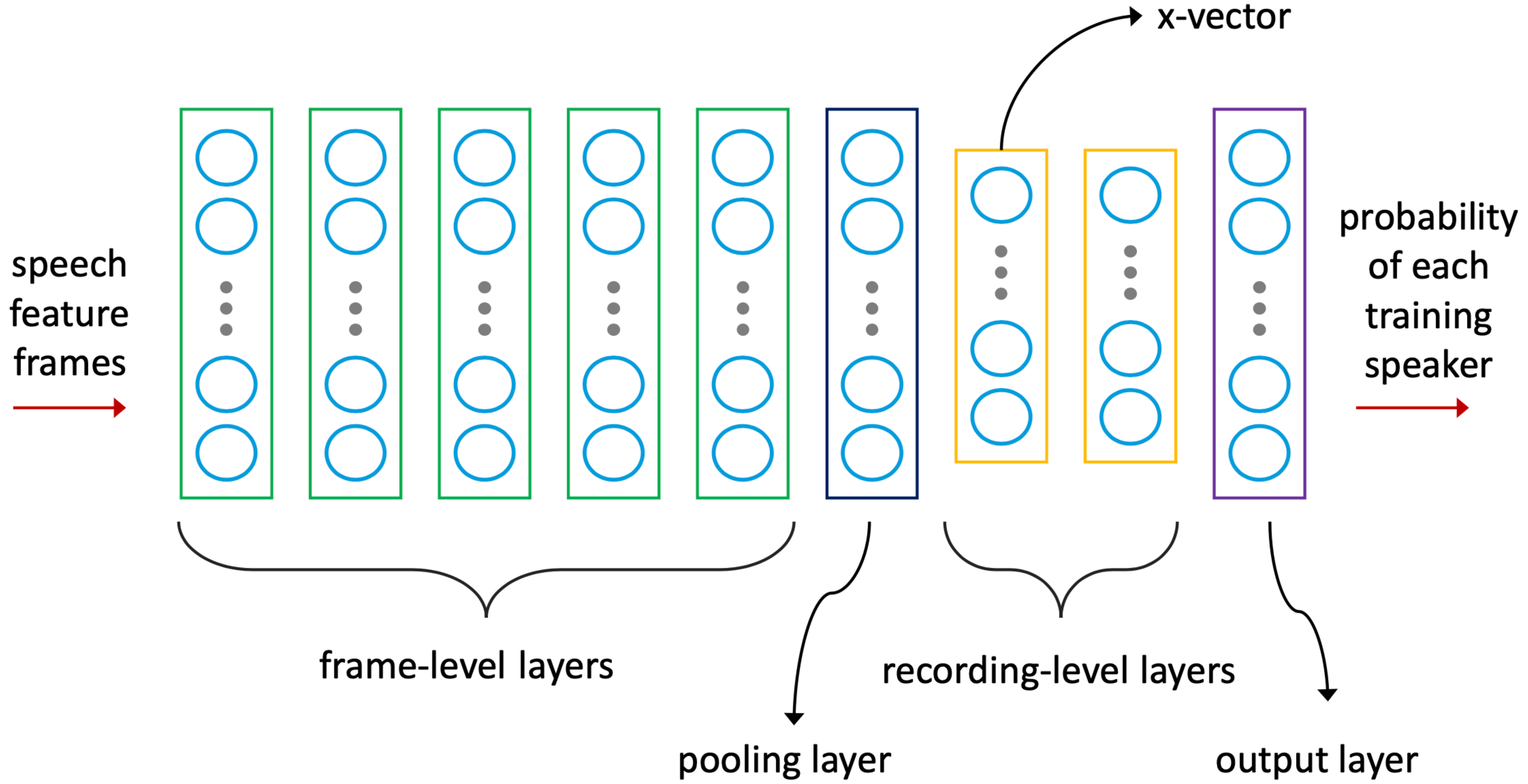
- i-vectors
- X-vectors

I-vector and x-vector Pipeline



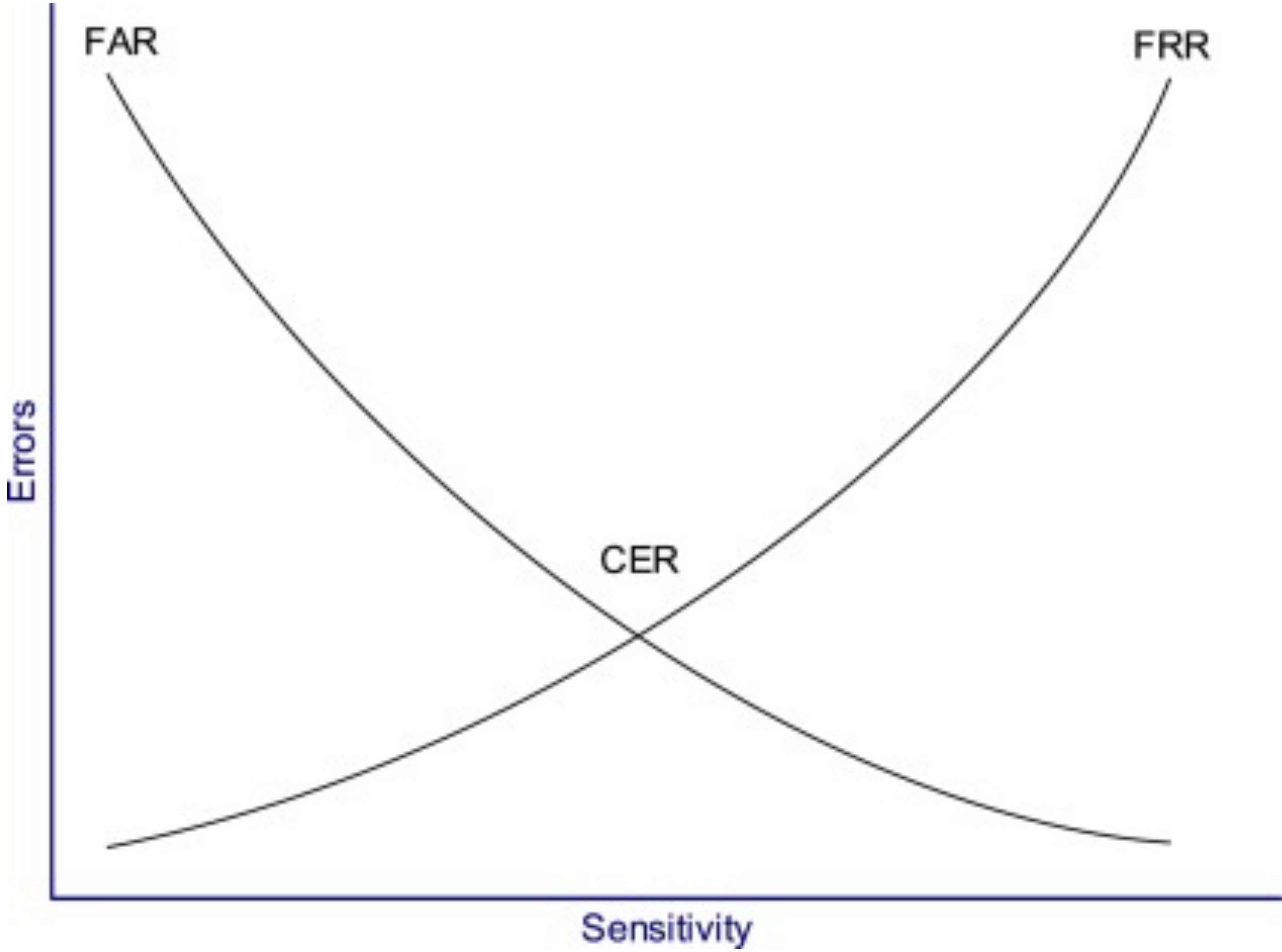
X-vector

- Map hi-dim utterances -> fixed length vectors
- Frame level layers are TDNN (temporal context)



Metrics

- Equal Error Rate (EER)
  - Single metric for comparing biometric algorithms
  - Point where FAR & FRR are euqal



Data

- Evaluated Spotify Podcasts
- Prominent Leader Speeches
  - 9,000 (1 second samples)
  - 6 unique speakers
  - Noise Samples for data augmentation
- Possible Choices
  - SITW Core
  - Vox Celeb

Results

- Data from X-VECTORS: ROBUST DNN EMBEDDINGS FOR SPEAKER RECOGNITION, 2018

	Speaker in the Wild	SRE16
Trained on VoxCeleb2	EER & (Lower is Better)	
i-vector	7.45	9.23
x-vector	**4.16**	**5.71**

Findings

- x-vector outperform i-vector models
- x-vectors can leverage larger datasets
    - Data augmentation also improves performance
- Commonality of feature extraction between I/x-vector pipelines allow direct comparison

Conclusion

1. To what extent can speaker recognition be modelled using high-level instructions?
   - Keras has robust API, full TF access
2. Which characteristics of Dialogue (Speech) can be used to recognise individual agents?

3. What is the accuracy of model?
   - Ongoing (desk research reports results around 98% accuracy)
4. How does the model compare to existing methods?
   1. How do we compare models
      - EER %
   2. X-vectors are higher performant than other common methods
   3. Can use larger datasets

Thank you

- Questions?