Politecnico di Torino
Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering
XXXVI cycle

Supervisor
Sandro Cumani

Speaker verification and multi-modal identity recognition

PhD Candidate: Salvatore Sarni

1. Context

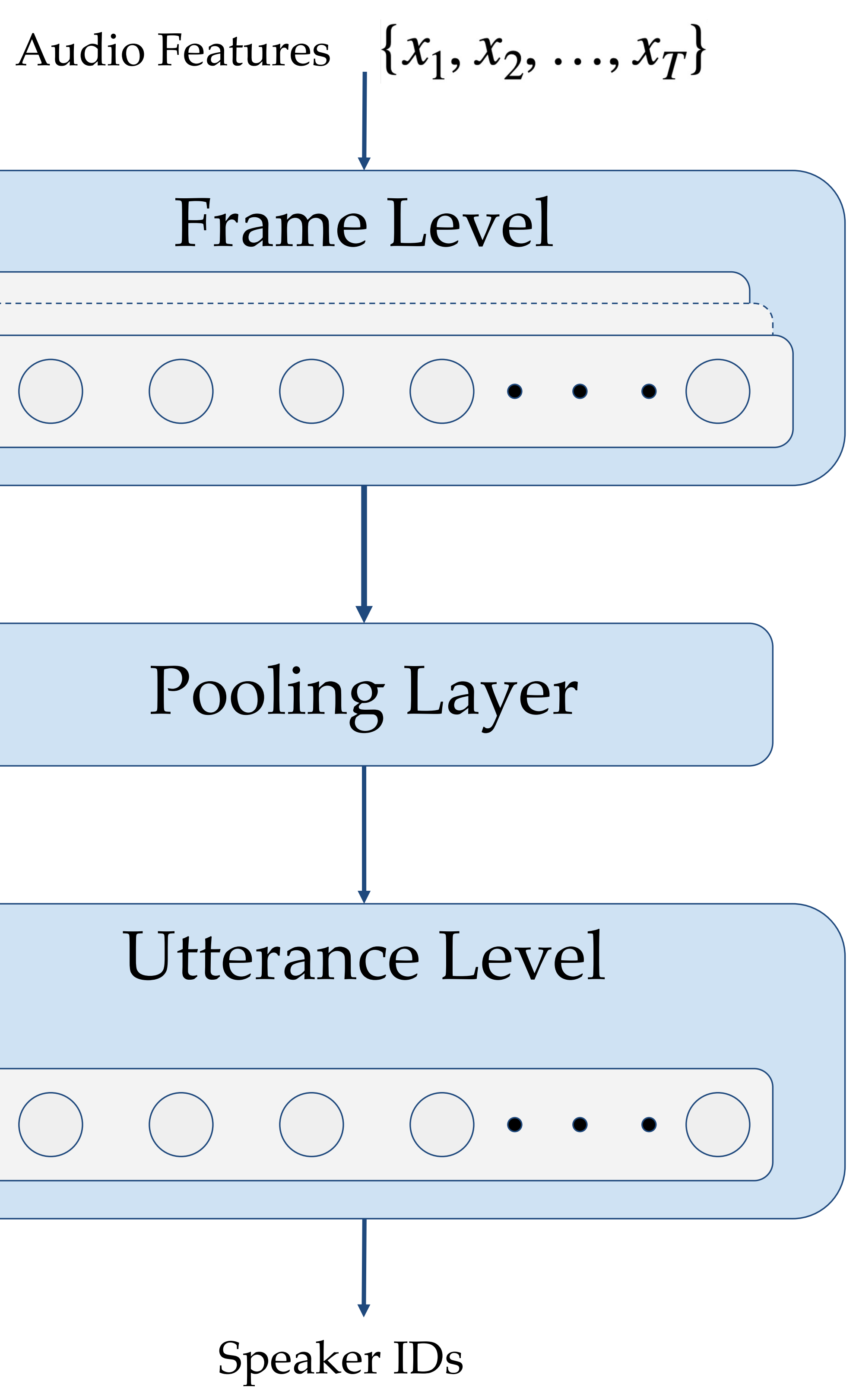
Both online and offline services require some level of identity recognition. Nowadays, phone unlocking uses face verification, and virtual assistants (e.g. Alexa, Siri) perform speaker verification before answering.

2. Goal

Deep learning has become state-of-the-art in voice and face verification systems, often employing similar solutions. The main goal is to improve the accuracy of the current frontend and backend systems. Multi-modal and cross-modal recognition is also explored.

3. Frontend

Audio comes from different sources and with different duration. A low-dimension and fixed-length vector are needed to embed the most useful information to verify and identify the speaker.



Audio Features $\{x_1, x_2, \ldots, x_T\}$

Frame Level

Pooling Layer

Utterance Level

Speaker IDs

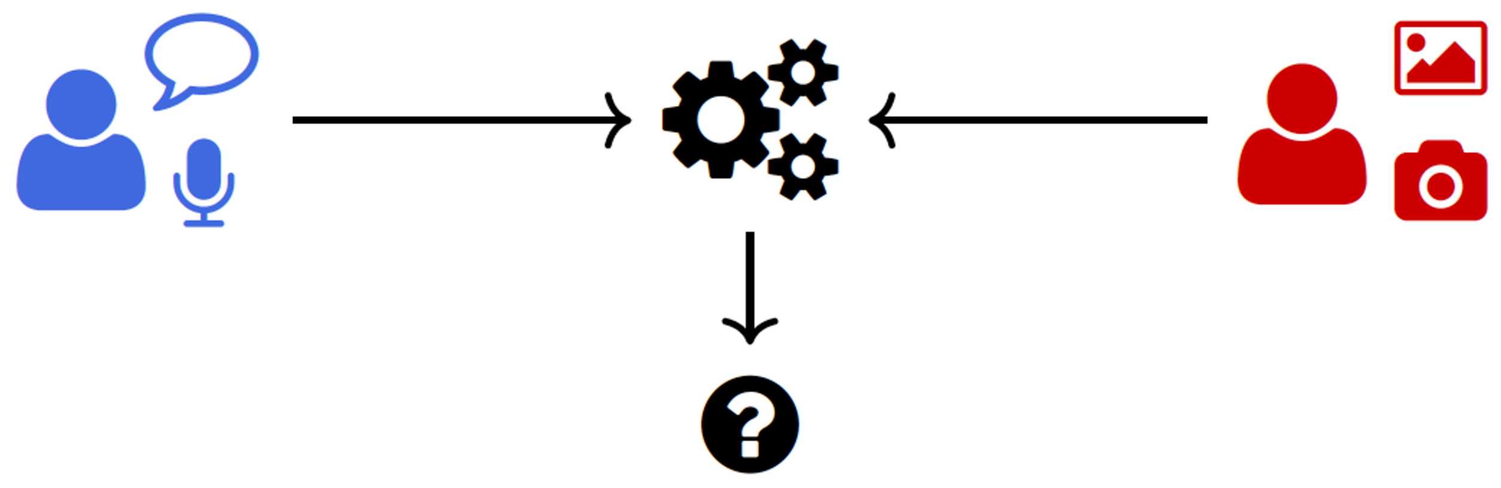
- Frame Level: single frame characteristics, different architectures
  - Time-Delay, ECAPA, ResNet, Transformers
- Pooling: aggregate frame-level information
  - standard pooling and Attention Mechanisms
- Utterance Level: embedding extraction

4. Backend

Two embeddings are compared to assess if they share the same identity. The comparison happens through a distance measure or with a probabilistic model. Side information, such as the duration of the audio, can be used to improve the performance of the backend models[1, 2]

5. Multi & Cross-Modal

Combining face and voices



From pre-trained to GAN-inspired solutions.

6. Language Recognition

NIST Language Recognition Evaluation 2022 language detection challenge. Fixed condition track with low-resource test languages.

7. References

1. Cumani, S. & Sarni, S. (2021). A Generative Model for Duration-Dependent Score Calibration. In *Interspeech* (pp. 4598-4602)
2. Cumani, S. & Sarni, S. (2022). Impostor score statistics as quality measures for the calibration od speaker verification systems. In *Proc. The Speaker and Language Recognition Workshop* (Odyssey 2022) (pp. 25-32)