Multistream neural architectures for Cued Speech recognition using a pre-trained visual feature extractor and constrained CTC decoding

PhD Student: Sanjana Sankar
Supervisors: Denis Beautemps, Thomas Hueber
Secondment Supervisors: Jeremy Huart, Jacqueline Leybaert
CSI Members: Slim Ouni, Michèle Gouiffes

Comm4CHILD

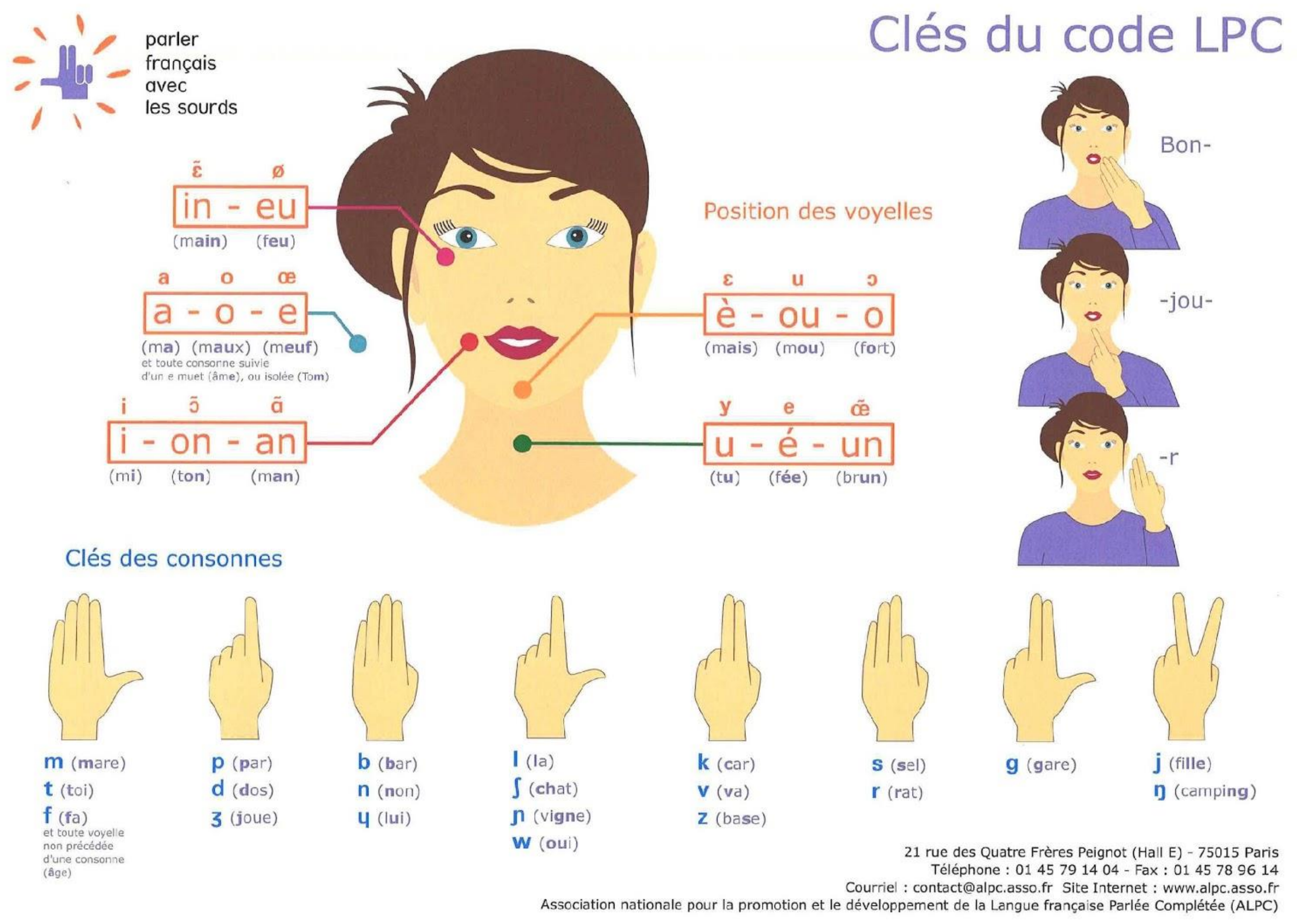
Multimodality and optimization of communication tools

ESR 10 – Automatic Recognition and Generation of Cued Speech using Deep Learning Techniques

What is Cued Speech (CS)?

❑ A visual communication tool that helps people with hearing impairment to better perceive the spoken language

❑ It encodes speech as a combination of visible hand shapes and hand positions to complement lip-reading



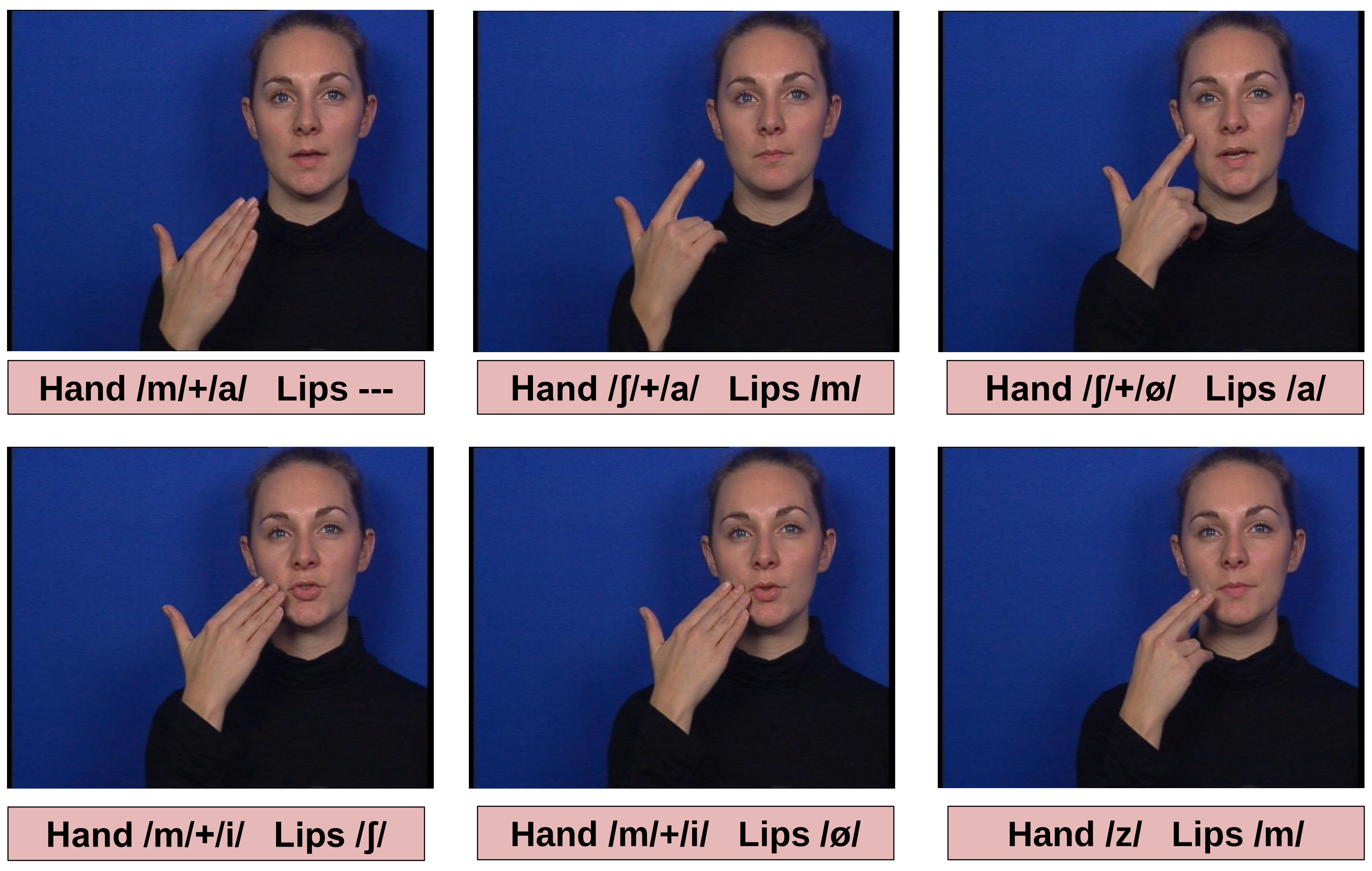
Challenges in CS Recognition

❑ Automatically learn the asynchrony between hand movement and the lips

❑ To accommodate the variability in anticipation between different speakers

❑ Limited dataset

Below is Cued Speech for the words « ma chemise »



Hand /m/+/a/ Lips ---

Hand /ʃ/+/a/ Lips /m/

Hand /ʃ/+/ø/ Lips /a/

Hand /m/+/i/ Lips /ʃ/

Hand /m/+/i/ Lips /ø/

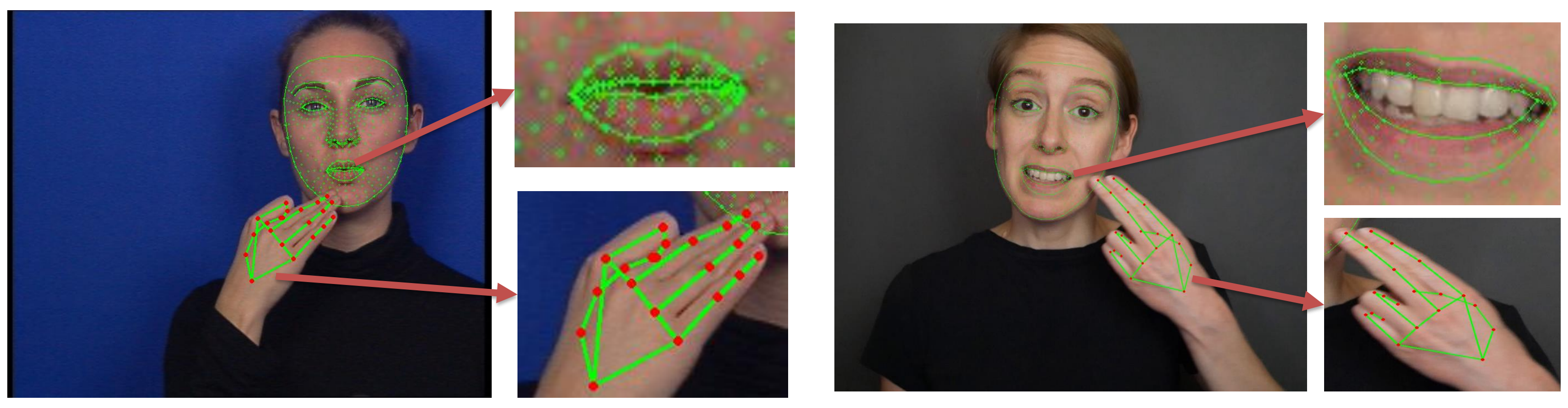
Hand /z/ Lips /m/

Feature Extraction

Extraction of primitives using pre-trained feature extractor and dimensionality reduction

❑ ***MediaPipe Hands* -** a hand and finger tracking solution by Google - infer 21 2D landmarks of a hand

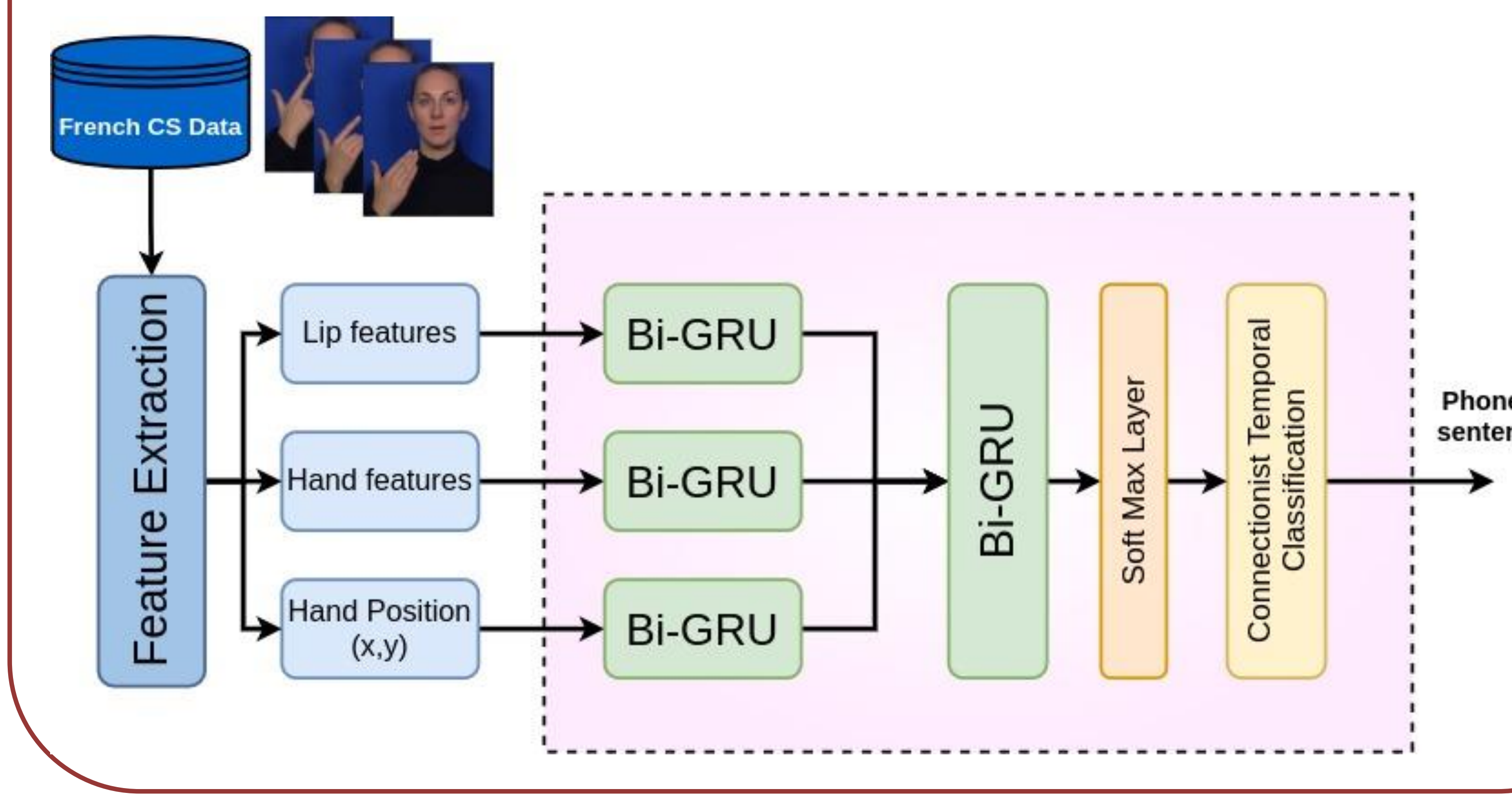
❑ ***MediaPipe Face Mesh* -** a face geometry solution to estimate 468 3D face landmarks – infer 42 2D landmarks of lips



Model Architecture

Continuous Cued Speech Recognition – aiming to transcript visual cues of speech to text

- French dataset - CS for 238 x 2 short French sentences
- Single speaker, clean environment
- **Phonetic Decoding:** Early Fusion, 3-Stream
- Phonemes recognition rate ~**71% acc.**
- **Decoding Strategies:** token passing algorithm, encoder-decoder architecture



Visible Research Output

Submitted a paper titled *"Multistream Neural Architectures for Cued Speech Recognition using a Pre-trained Visual Feature Extractor and Constrained CTC Decoding"* to ICASSP '22

Future Work

❑ Collect American-English dataset

❑ GAN-based Encoder-Decoder model for CS generation

Grenoble Images Parole Signal Automatique
UMR CNRS 5216 – Grenoble Campus
38400 Saint Martin d'Hères - FRANCE