# IPTC-Amazon collaboration

**Presentation**

**(Nov 15th, 2023)**

POLITÉCNICA

iptc

www.iptc.upm.es

# IPTC-Amazon final event: agenda

- Innovations in Text-to-Speech Technology at Amazon
- Possibilities of collaboration at Amazon
- Projects developed in 2022-23
  - Sign language motion generation from high level sign characteristics
  - Speaker diarization with multimodal inputs
  - Pose and spatial movement as input for dynamic content search & generation
  - Entangling AI-audio synthesis models and multimodal representations
  - Zero-shot sonorizing of video sequences

# Daniel Sáez-Trigueros

**Innovations in Text-to-Speech Technology at Amazon**

In this presentation, Daniel will introduce the Text-to-Speech team at Amazon and talk about some of the research projects they have published during 2023. The goal is to give an overview of the kind of problems that the team is facing and the innovative solutions that they are working on to tackle them.

**Short CV**

Daniel Sáez Trigueros received his Bachelor's degree in Telecommunication Technical Engineering from the University of Málaga in 2013, and a Ph.D. in Machine Learning from the University of Hertfordshire in 2019. His Ph.D was done in collaboration with IDscan Biometrics Ltd and focused on the development of novel face recognition techniques using machine learning.

In 2019, Daniel joined the Text-to-Speech Research team at Amazon as an Applied Scientist. Since then, he has further developed his expertise in deep learning and generative models, working on and leading different projects related to speech synthesis. In 2023, he became an advisor to master students from Technical University of Madrid as part of an IPTC-Amazon collaboration.

iptc

# Roberto Barra Chicote

**Possibilities of collaboration at Amazon**
Internships at Amazon for students, researchers and professors

**Short CV**

Roberto Barra Chicote received his Master's degree in Telecommunication at ETSIT-UPM in 2005. His Ph.D was focused on speech synthesis: Contributions to the Analysis, Design and Evaluation of Strategies for Corpus-based Emotional Speech Synthesis. His PhD obtained several national awards.

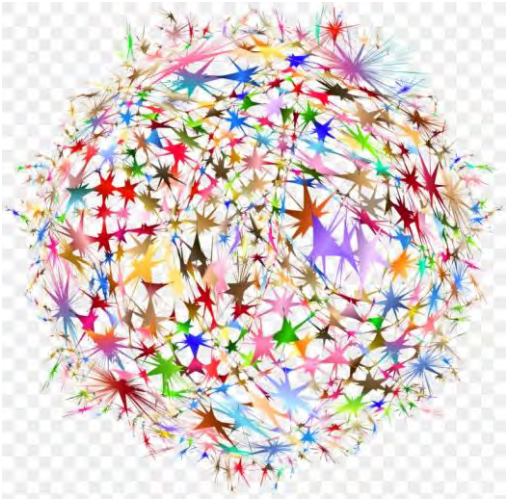
From 2009, to 2015, Roberto was assistant professor in the Speech Technology Group at ETSIT-UPM, working in several projects including speech technologies.

In 2015, Roberto joined the Text-to-Speech Research team at Amazon, and now he is Principal Scientist in this team.

iptc

# IPTC-Amazon collaboration

This collaboration has been focused on developing technologies to



**extract and combine self-supervised representations for multimedia processing.**

These technologies have a big potential in many areas such as content generation (audio, image, video, or sign language representation), classification, labelling or searching.

# IPTC-Amazon: Team, IPTC students

- Juan Moreno Galiano (juan.moreno.galiano@alumnos.upm.es)
- María Villa Monedero (maria.villa.monedero@alumnos.upm.es)
- Laura Fernández Galindo (laura.fernandez.galindo@alumnos.upm.es)
- María Sánchez Ruiz (maria.sanruiz@alumnos.upm.es)
- Andrzej Daniel Dobrzycki (daniel.dobrzycki@alumnos.upm.es)

# IPTC-Amazon: Students formation

- Introduction to **deep learning** strategies:
  - CNN, RNNs, Transformers, Adversarial and Contrastive learning, etc.
- **Tools** for deep learning:
  - CLIP or LIP and variants like WavCLIP, AudioCLIP or MotionCLIP. Also, video and audio processing tools (like OpenPose, MediaPipe, etc.).
- **Framework** for experimentation.
- **Evaluation** with several datasets.

# IPTC-Amazon: Students supervision

- Every student has been supervised by
  - one researcher from IPTC and
  - another researcher from Amazon.
  - Meetings every week (aprox.).

- Joint meetings and sessions every 1.5 or 2 months to present the last achievements.

# IPTC-Amazon: Results

- Web: provisional link (only direct access)
  - https://iptc.upm.es/education/iptc-amazon-collaboration
- **Prototypes and demonstrations** to show the main research achievements.
- **Papers** submissions to international conferences or journals.

# IPTC-Amazon: collaboration teams

- Sign language motion generation from high level sign characteristics
  - Student: María Villa Monedero
  - Advisors: Rubén San-Segundo, Manuel Gil-Martín, Andrzej Pomirski


- Speaker diarization with multimodal inputs
  - Student: Juan Moreno Galiano
  - Advisors: Alberto Belmonte, Ivan Valles

- Pose and spatial movement as input for dynamic content search & generation
  - Student: Andrzj Daniel Dobrzycki
  - Advisors: Ana Bernardos, Daniel Saez

- Entangling AI-audio synthesis models and multimodal representations
  - Student: Laura Fernández Galindo
  - Advisors: Julián David Arias Londoño, Juan Ignacio Gódino, Luis Hernández, Giulia Comini

- Zero-shot sonorizing of video sequences
  - Student: María Sánchez Ruiz
  - Advisors: Mateo Cámara, José Luis Blanco, Luis Hernández, Adam Gabrys

# IPTC-Amazon: Projects

Projects:

- Sign language motion generation from high level sign characteristics
- Speaker diarization with multimodal inputs
- Pose and spatial movement as input for dynamic content search & generation
- Entangling AI-audio synthesis models and multimodal representations
- Zero-shot sonorizing of video sequences

# Sign language motion generation from high level sign characteristics

iptc

POLITÉCNICA

# Sign language motion generation and analysis

## Introduction and Objectives

```
<hamnosys_manual>
        <hamfist/>
        <hambetween/>
        <hamfist/>
        <hamthumbacrossmod/>
        <hamextfingeru/>
        <hampalmd/>
        <hamshoulders/>
        <hamlrat/>
</hamnosys_manual>
```

Hand configurations

Landmarks

# What have we done?

## Dataset Creation

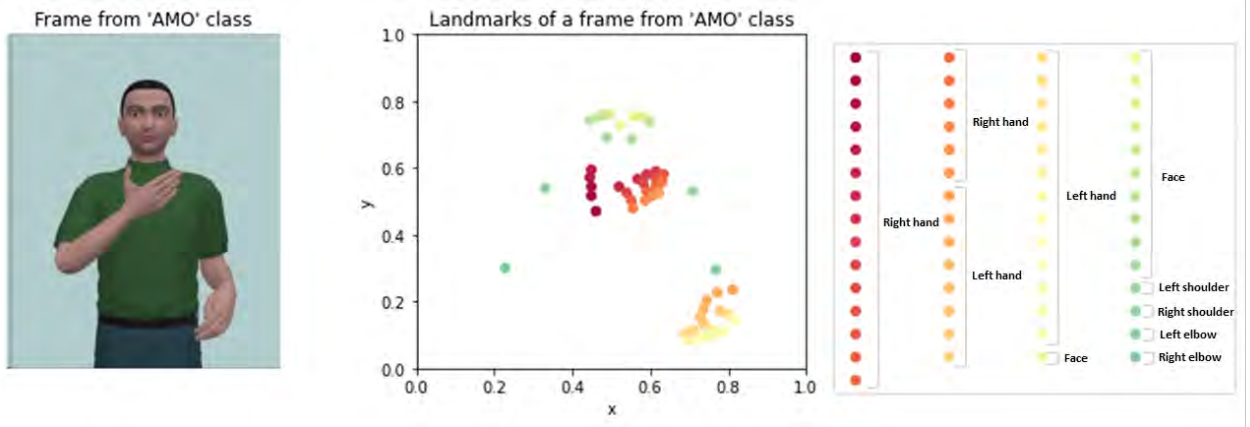
- Creation of a language sign dataset

## Transformer

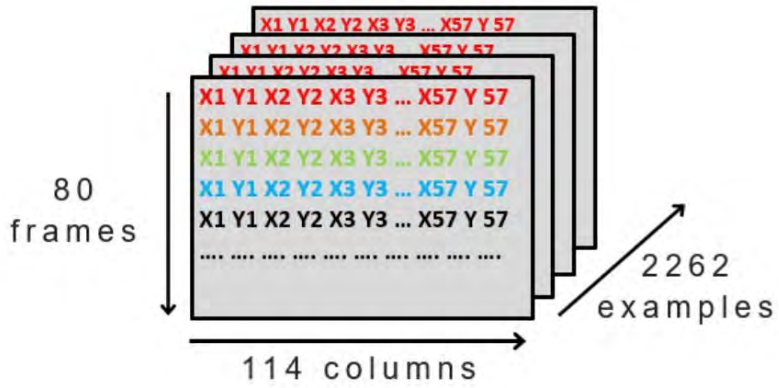
- Based on a Spanish to English translator transformer

# Dataset Creation

- 6.786 videos
- 24 frames per sign
- Three different points of view with three different avatars
- Mediapipe for extraction of coordinates.




- 3-Dimensional matrix
- 2.626 gestures
- 80 frames per gesture



# Transformer Development and Evaluation

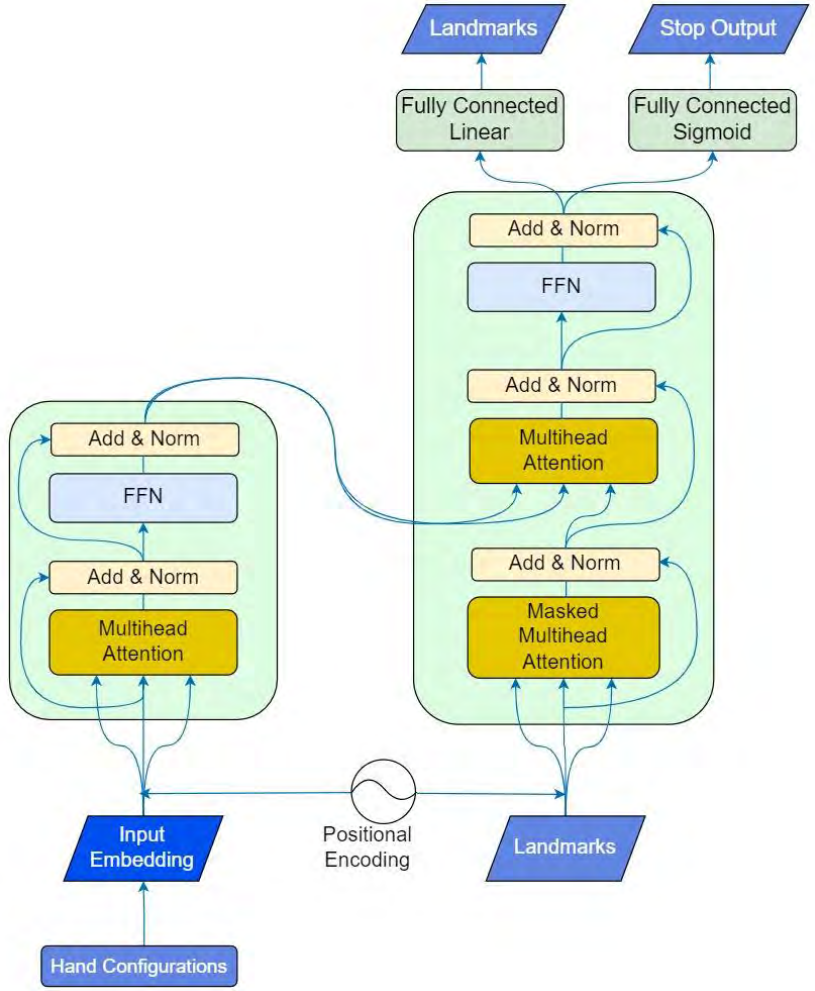
## Transformer development

### Encoder

- Sign phonemes
- *TokenandPositionEmbedding*

### Decoder

- Landmarks
- *PositionEmbedding*
- Predicts landmarks

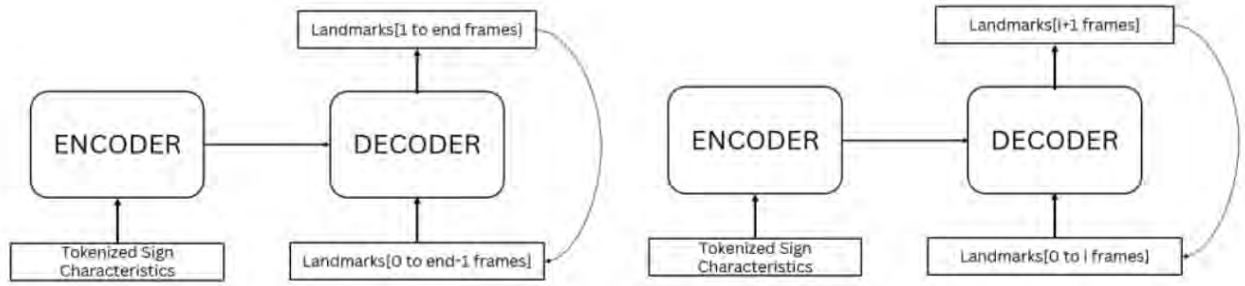


# Transformer Development and Evaluation

Strategies and Evaluation

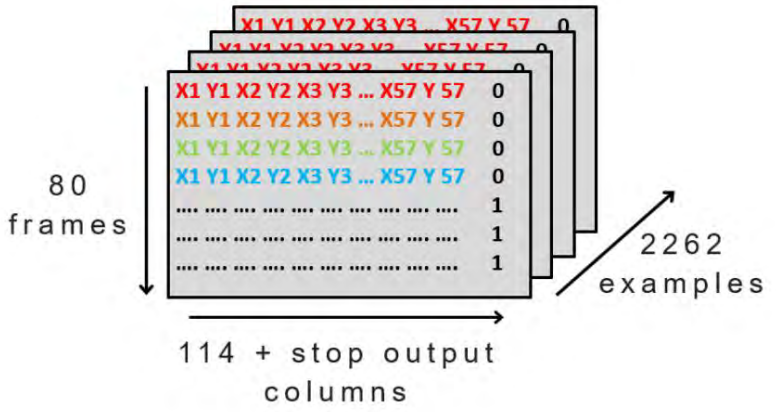
## Landmarks Generation

- Padding & Interpolation strategies
- Data augmentation
- **Mean DTW**



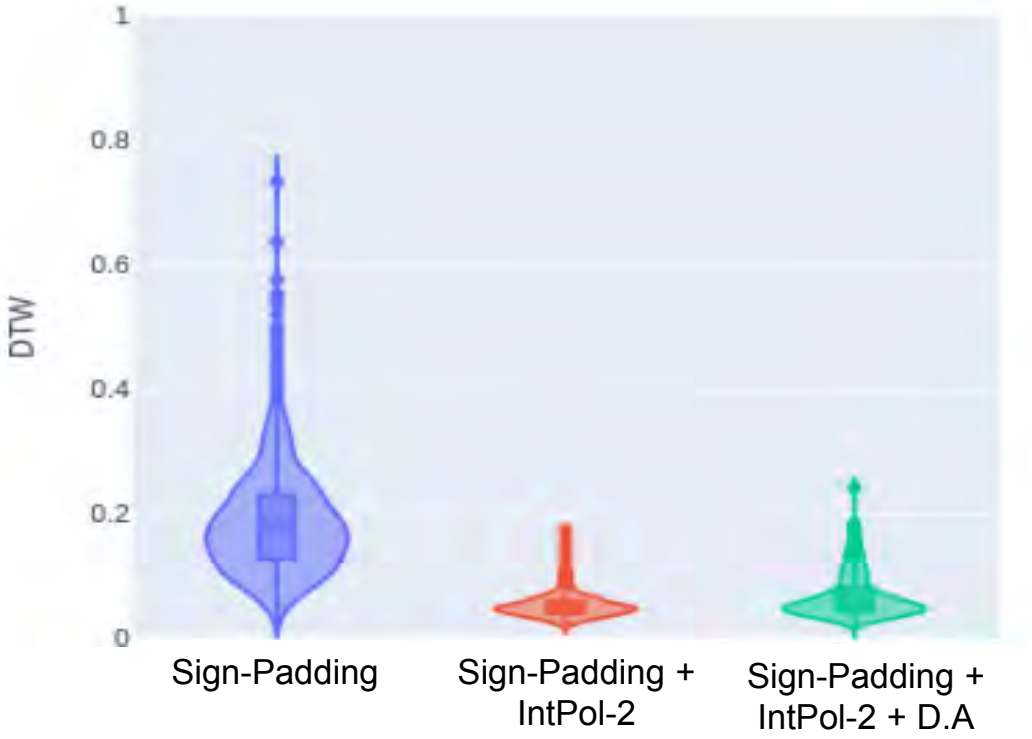

## Stop Detector

- Padding & Interpolation strategies
- Data augmentation
- **ROC Curves** and **P-values**



# Landmarks Generation

- Padding Strategies
- Interpolation Strategies
- Data Augmentation


| | Sign-Padding | Sign-Padding + IntPol-2 | Sign-Padding + IntPol2 + D.A |
|---|---|---|---|
| DTW mean | 0.1880 ± 0.0857 | **0.0523 ± 0.0177** | 0.0677 ± 0.0340 |

# Stop Detector and Final System

## STOP Detector

| | Hybrid-Padding | Hybrid-Padding + IntPol-2 | Hybrid-Padding + IntPol2 + D.A |
|---|---|---|---|
| AUC | 0.94025 | 0.94379 | **0.97653** |
| P-Value | 0.000097 | | <0.000001 |

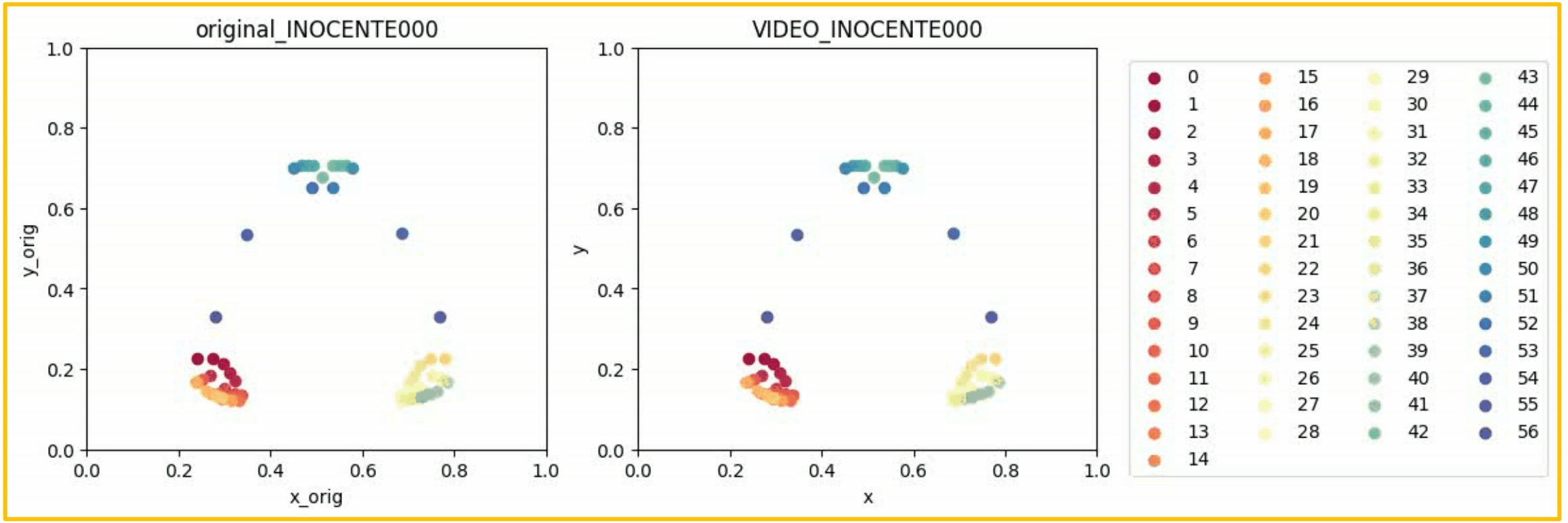
## Final System


Sign-Padding | Sign-Padding + IntPol-2 | Sign-Padding + IntPol-2 + Stop Detector

| | Sign-Padding | Sign-Padding + IntPol-2 | Sign-Padding + IntPol-2 + Stop Detector (0.5) |
|---|---|---|---|
| DTW mean | 0.1880 ± 0.0857 | **0.0523 ± 0.0177** | 0.1057 ± 0.0659 |

iptc

# Results


original_INOCENTE000
VIDEO_INOCENTE000
y_orig
x_orig
y
x
0.0
0.2
0.4
0.6
0.8
1.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
15 16 17 18 19 20 21 22 23 24 25 26 27 28
29 30 31 32 33 34 35 36 37 38 39 40 41 42
43 44 45 46 47 48 49 50 51 52 53 54 55 56

# Journal Papers

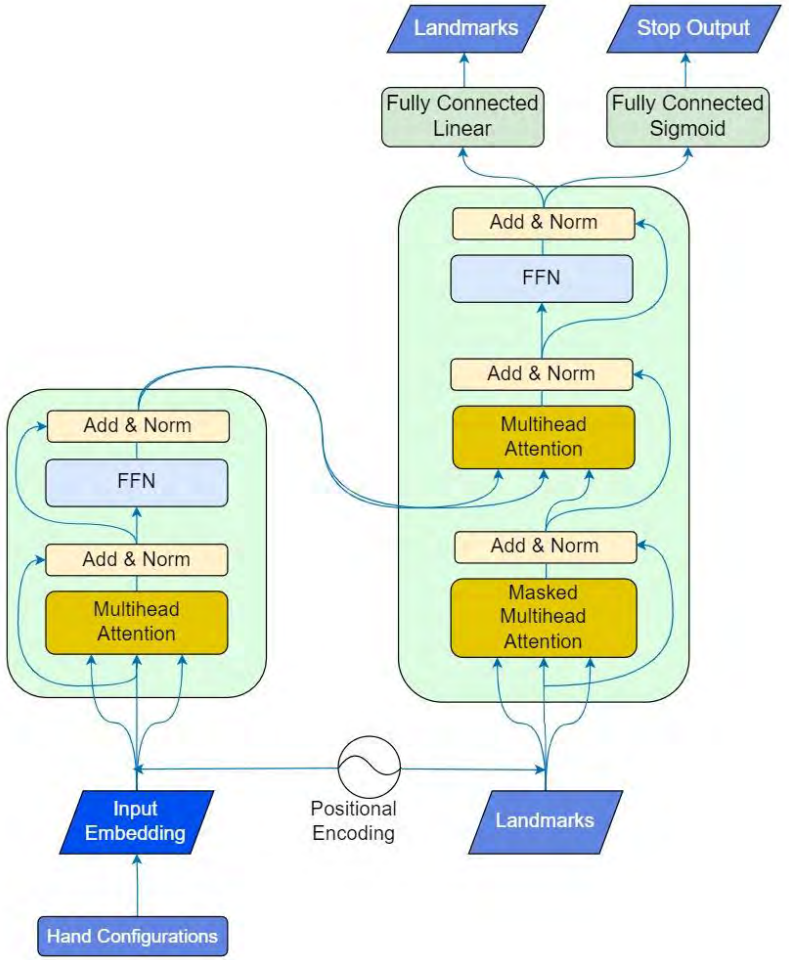
**M. Villa-Monedero**, M. Gil-Martín, D. Sáez-Trigueros, A. Pomirski, and R. San-Segundo, "*Sign Language Dataset for Automatic Motion Generation*", Journal of Imaging, In review, 2023.

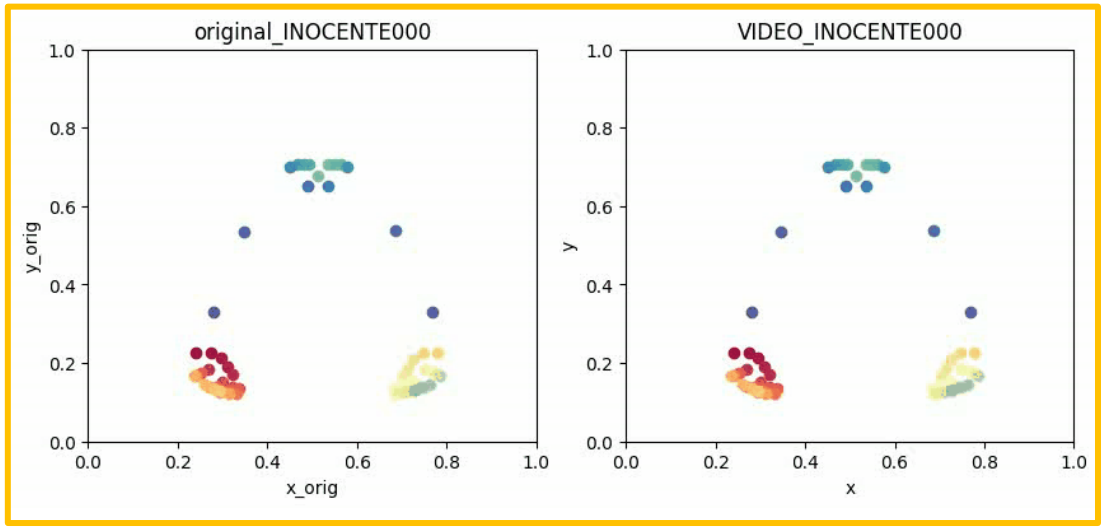
M. Gil-Martín, **M. Villa-Monedero**, A. Pomirski, D. Sáez-Trigueros, and R. San-Segundo, "*Sign Language Motion Generation from Sign Characteristics*", Sensors, In review, 2023.

iptc

# Sign language motion generation and analysis

## Conclusions and Greetings



# Speaker diarization with multimodal inputs

iptc

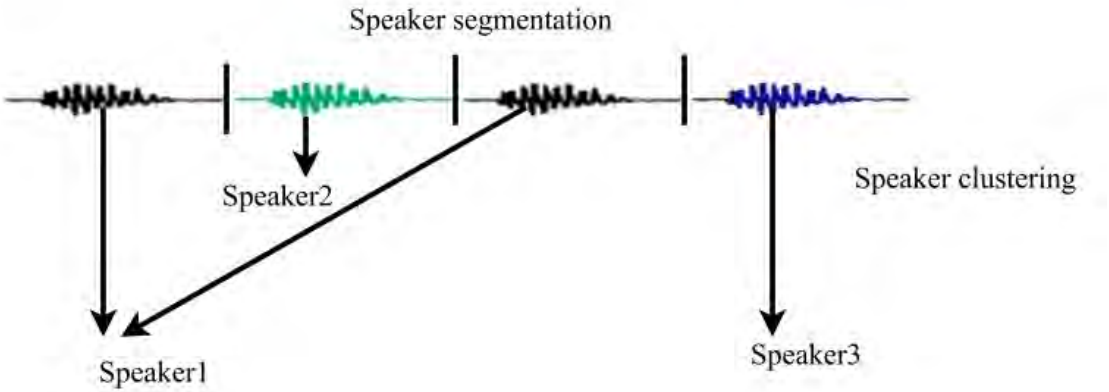
POLITÉCNICA

# Index

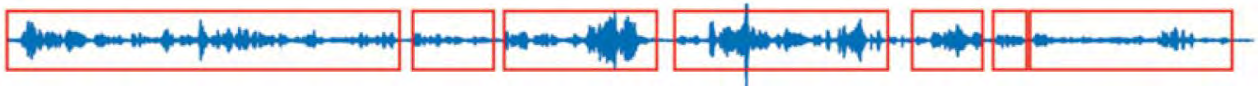
- Speaker diarization – Traditional vs Multimodal
- Our case – Tools and AI: Contrastive learning
- Dataset
- Training and testing process
- Experiments
- Results
- Future improvements

# Speaker diarization –Traditional vs Multimodal

- Answers to the question of “who spoke when”




Speaker diarization

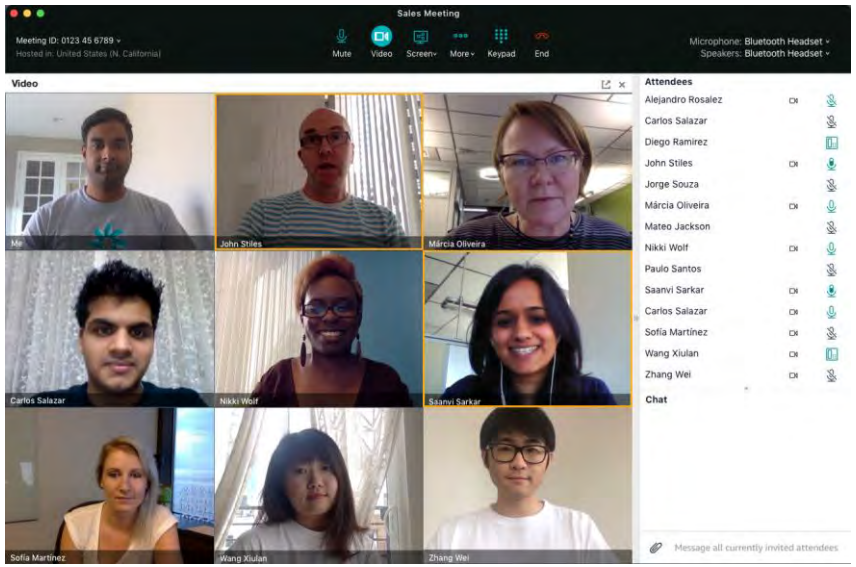


Multimodal speaker diarization

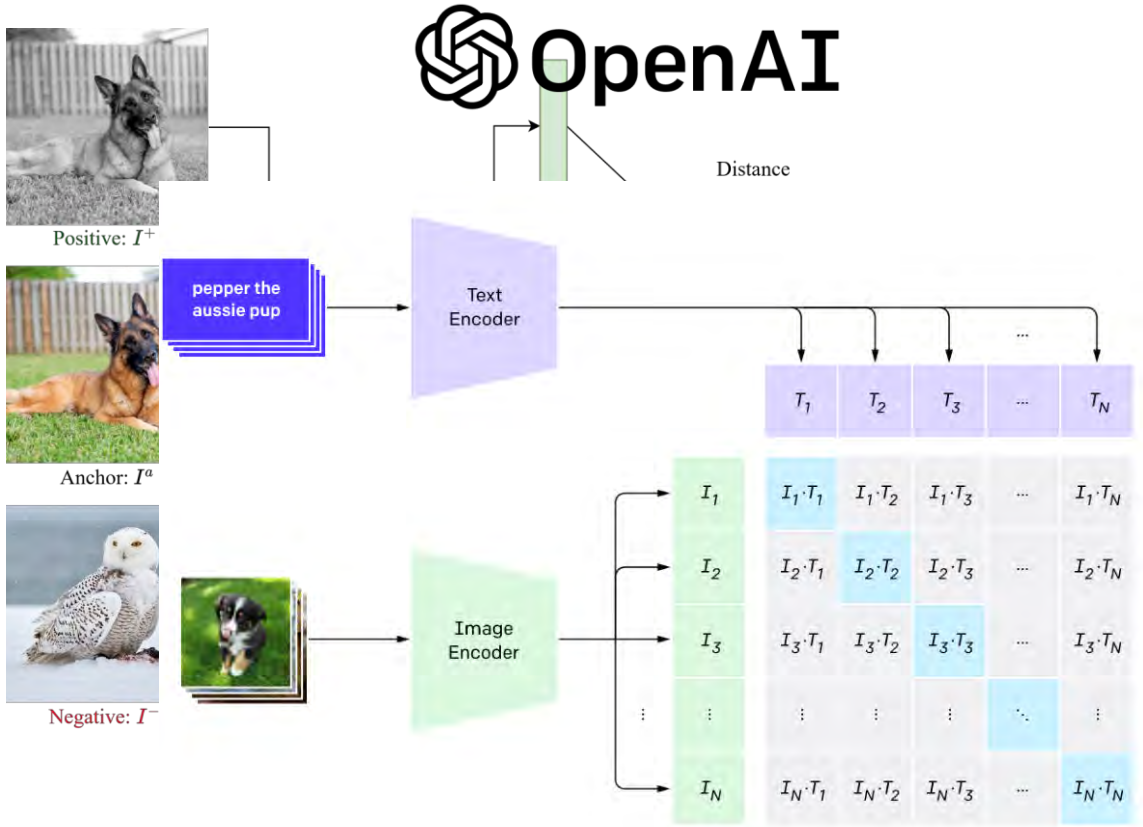
# Our case – Tools and AI: Contrastive learning

We decided to focus on a concrete use case to ease the task

A video conference



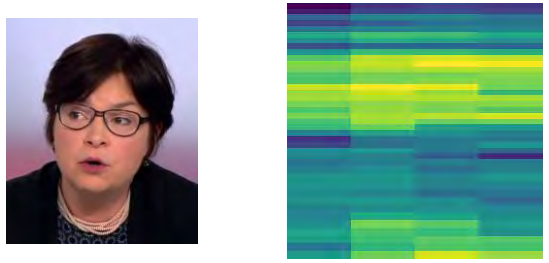

Minimize the distance between similar pairs and maximize the distance of dissimilar pairs.



# Dataset: AVSpeech

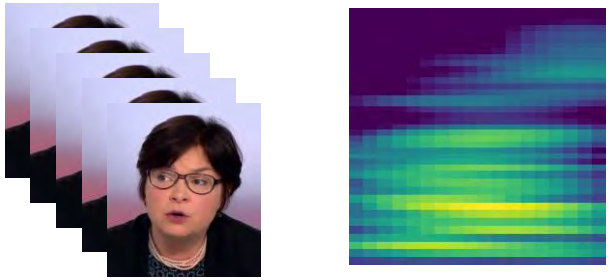

More than 200k videos of duration around 5-10 seconds

25 fps videos



31050 frames and 31050 40ms mel-spectrograms

5 fps videos



6210 5-frames sets and 6210 200ms mel-spectrograms

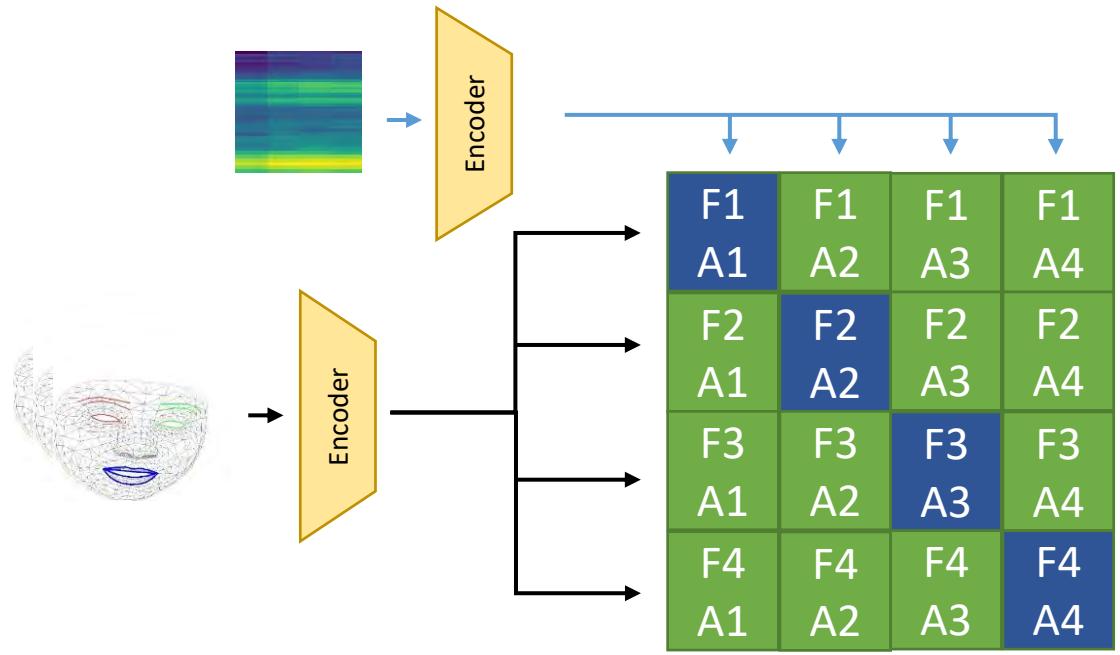
# Training and testing the network

## Training

Use all the frames and audios downloaded

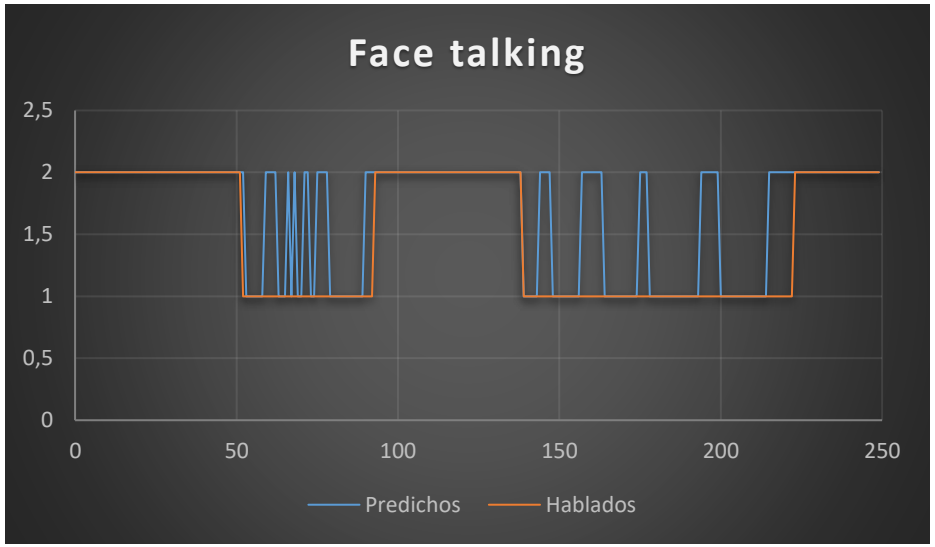
80% training data

20% validating data

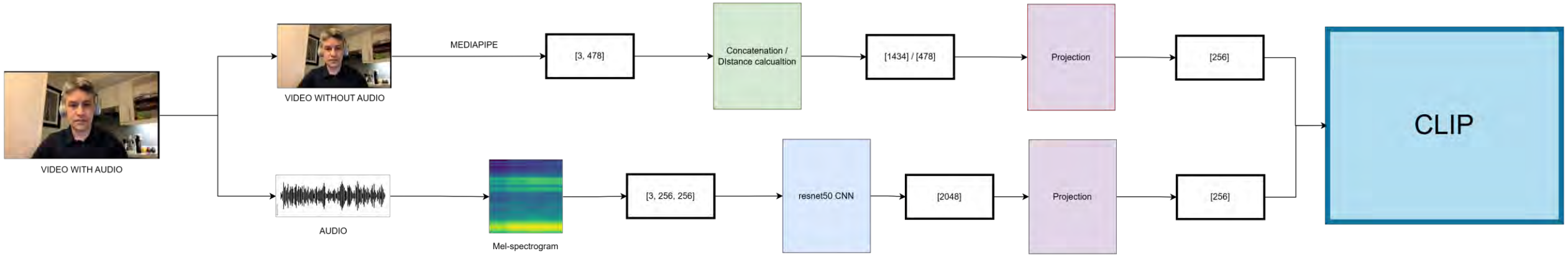



## Testing

Manually label a videoconferece video and obtain the network prediction



# Experiments




1. Inputs: Face cropped and 40 ms Mel-spectrogram image. Both CLIP branches with ResNet50.

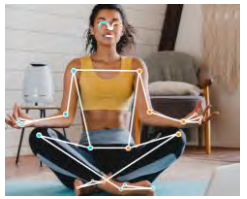
2. Inputs: Face image after MediaPipe keypoints extraction and 40 ms Mel-spectrogram image. Both CLIP branches with ResNet50.

3. Inputs: MediaPipe face keypoints (as vector) and 40 ms Mel-spectrogram image. DNN and ResNet50 in each CLIP Branch.

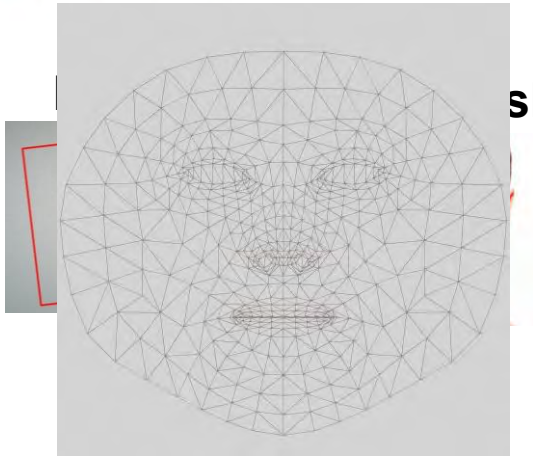
4. Inputs: MediaPipe face keypoints (as vector) normalized distance to a central point (nose) and 40 ms Mel-spectrogram image. DNN and ResNet50 in each CLIP Branch.

MediaPipe

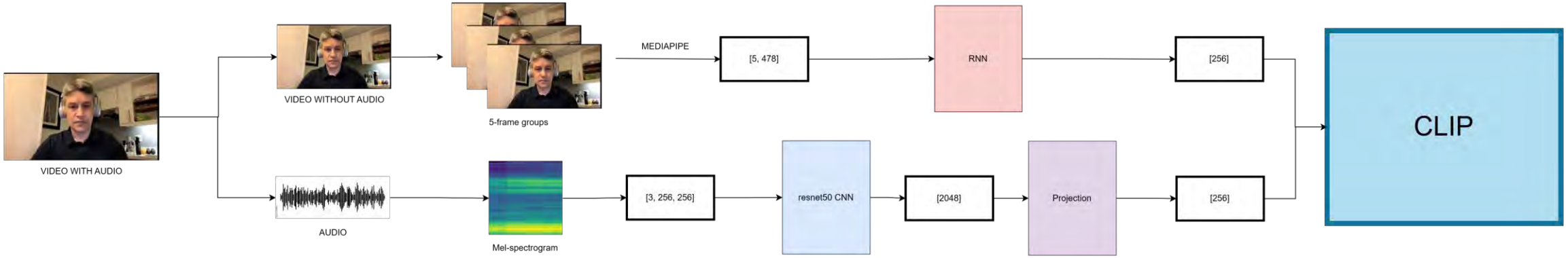
**Pose**



**sh**



# Experiments and results




5. Inputs: 5 consecutive frames MediaPipe face Keypoints normalized distances and 200 ms Mel-Spectrogram image. RNN and ResNet50 in each CLIP branch.

| Experiments | Accuracy | Comments |
|---|---|---|
| 1 - Face and 40ms Spectrogram | <20% | It is not possible to generalize the problem to real world applications, added to the bad performance of the network |
| 2 - Keypoints image and 40ms Spectrogram | <25% | Similar problem as in experiment 1 |
| 3 – Keypoints vector and 40ms Spectrogram | 43.2% | The direct use of the keypoints values with a simpler network (DNN instead CNN) is more efficient to learn |
| 4 – Keypoints distances and 40ms Spectrogram | 82% | The distances provide information about the location and movement of different parts in the face |
| 5 – 5 frames Keypoints distances and 200ms Spectrogram | 76% | Not totally finished and reduced dataset. The RNN can learn movements in the keypoints to match better with a longer spectrogram |

# Future lines

- Increase the number of data to train the network using the network proposed in experiment 5. This is the main next step, as it will lead to much better results, as with the current data the testing results are “following” the real speaking moments.
- Explore the use of Transformers instead RNNs.
- Use filters to reduce the spurious misspredictions. Moving averaging, etc.
- Change the audio encoding. Some experiments with wav2vec network, but obtaining worse results than with the CNN. However, there are more encoding networks for the audio which can be explored and evaluated.