IPTC-Amazon collaboration

Kick off meeting

(December 1st, 2022)

POLITÉCNICA

iptc

www.iptc.upm.es

Index

1. **IPTC presentation**
2. IPTC-Amazon initiative
    1. Introduction
    2. Team
    3. Main research lines
    4. Planning
    5. Application teams

Information Processing and Telecommunications Center

Technologies for creating high economic and social value

POLITÉCNICA

iptc

www.iptc.upm.es

Who We Are

ICT at Universidad Politécnica de Madrid

The **Information Processing and Telecommunications Center** was created in 2016 to bring together the expertise and resources of a number of highly competitive research groups working in the fields of **Electronics, Communications, Networks, Computing and Software**.

http://www.iptc.upm.es

IPTC in facts and figures

Yearly data

1.

180 researchers

Bringing expertise in different areas of knowledge on digital technologies and communications

2.

>110 competitive research projects

In national and international R&D and innovation competitive programmes

3.

>70 research contracts

Solving the needs of industry partners and contributing to value creation and innovation

4.

>330 journal and conference papers

High quality research outcomes challenging and advancing the state-of-the-art.

5.

>20 Ph.D. thesis

Doctoral works on relevant, state-of-the-art topics on digital technologies

iptc

What We Do

Applied and Basic Research, Innovative Engineering Solutions, Advanced Consulting Services

Center´ strategic areas

- Future telecoms
- Smart cities
- Defense and space
- Health and wellbeing
- Connected Industry
- Data engineering & digital transformation
- Mobility and transport
- Interaction and interfaces

Groups´ core technologies

- Electronics and systems
- Radio and wireless
- Multimedia technologies
- Networks and Cyber
- Statistics and Signal and Comms Theory
- Data Science

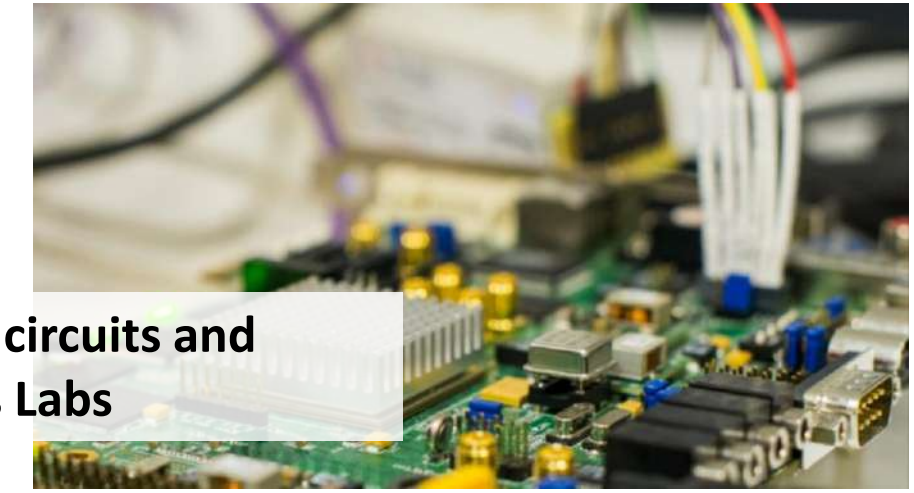
iptc

Facilities and infrastructures

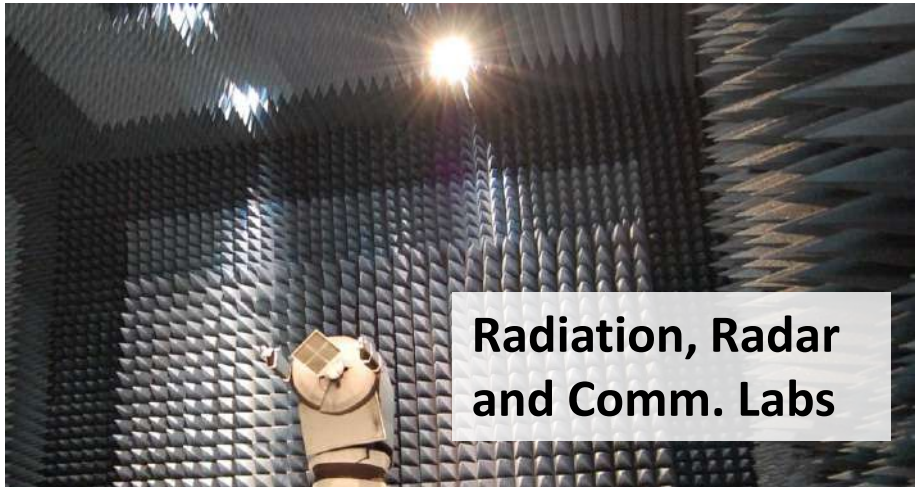
Enabling research, prototyping, user testing

Multimedia technologies Labs

Living & Experience Labs



Integrated circuits and electronics Labs

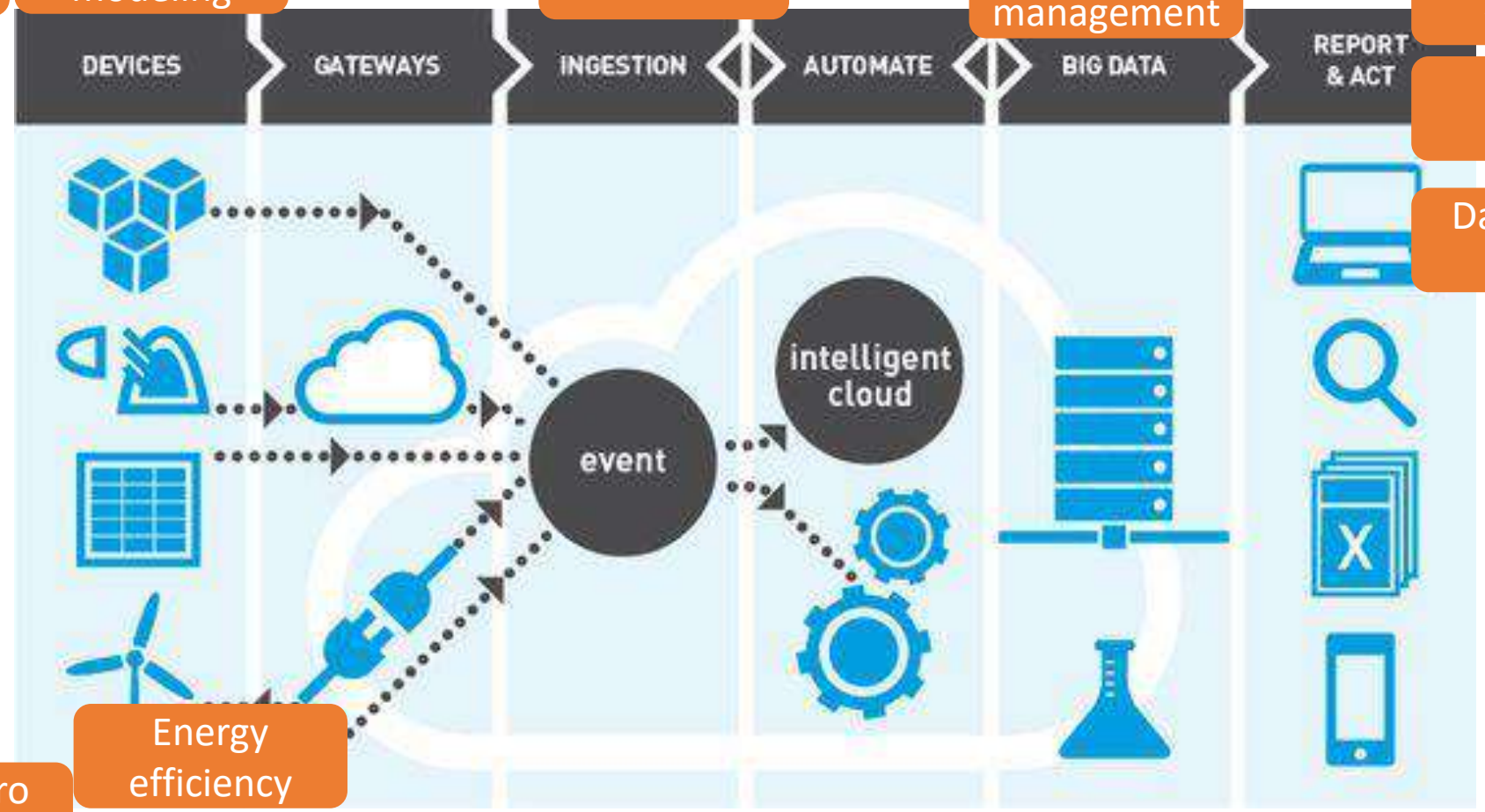


Radiation, Radar and Comm. Labs

Big Data, Data Science, Machine Learning, AI

…for people and organizations

Data-driven decision making value chain



Data acquisition

Data modeling

SW architectures

CEP

Data fusion

RT-ML

Data management

AR/VR visualization

Adaptive dashboards

Data analytics & prediction

Micro chips

Energy efficiency

Light-ML

HW acceleration

SW optimization

iptc

Research Areas in BD/ML/AI

ML and DL/ Reinforcement Learning/ Adaptive ML and Transfer Learning	**Speech, image, video and multimedia Recognition & Analytics**	**Artificial Vision and Multimodal Fusion**	**Natural Language Processing**	**Decision Making Support Techniques**
Activity Recognition (multimodal, smartphones, wearables, vision)	**Behaviour Analysis** Emotion, attention and sentiment analysis	**Recommendation systems**	**Intelligent Systems and Intelligence Augmentation**. Adaptation and learning	**Explainable and Edge AI**

Architectures for acquisition, fusion, and batch and real time processing	**Big Non-Structured Data Processing.** Non Structured Data Engineering	**Big Data Service Orchestration** (cloud and edge computing)	**Big Data Visualization and Interaction.**

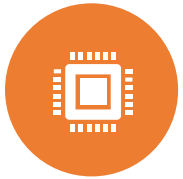
iptc

Sectors and Applications

Defense. C4ISR, Surveillance, Intelligence Data Analysis, DRIT, Hybrid Warfare, Multisensory Fusion, Situation Assessment Tools and Methods.

Smart Cities. Methodologies for data-driven management. IOT technologies.

Health and wellbeing: ML/DL on medical images . Behaviour analysis for diagnostics and treatment. Monitoring and support to diagnostics.



Cybersecurity and Cyberdefense. Big Data guided cyberdefense. Forensic applications.

Connected and 5G Industry: digital twins, prospective management and predictive maintenance.

Telecomms: Data exploitation, personal communications and mobility, conversational systems.

Creative Industries: Modelling, representation and analytics of 2D and 3D visual content. Multimedia. Neuromarketing. Natural Language Processing.

Transport: ITS, tracking and fleet management , autonomous navigation and UAV (UAV y UUV). ATC/ATM.

Decision Support: Fintech, Insurance, Data Economy.

Education

Data Visualization, Time Series Analysis, Machine Learning, Deep Learning, Reinforcement Learning, Autonomous systems for Financial Trading, Deep Learning for Audio, Music and Speech, Fundamentals of Big Data, Techniques for Decision Making, Data Analysis and Business Intelligence, Reinforcement Learning, Social Networks Analysis, etc. in:

Master in Telecommunications Engineering.
Master in Networks and Telematics Services Engineering.
Master in Statistical and Computational Processing of Information.
Master in Biomedical Engineering.
Master in Signal Theory and Communications (Speciality in Signal Processing and ML for Big Data)
Master in Electronic Systems Engineering.

Degree in Telecommunication Technologies and Services Engineering
Degree in Biomedical Engineering
Degree in Data Engineering and Systems

Moreover:

- Training pills and advanced courses for companies.
- Specialized sesions and seminars on targeted technologies and methods.
- Hackathons.
- Mentoring and active incubation of start-ups.

POLITÉCNICA

ETSIT UPM

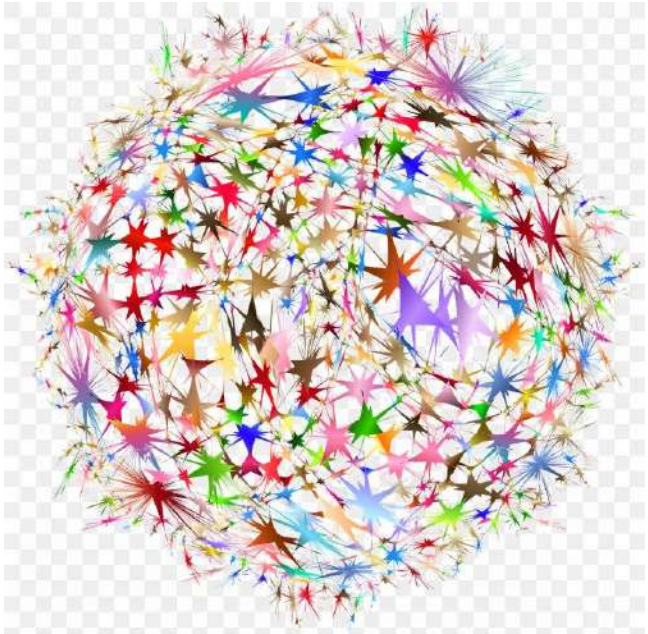
Index

1. IPTC presentation
2. **IPTC-Amazon initiative**
    1. Introduction
    2. Team
    3. Main research lines
    4. Planning

IPTC-Amazon: Introduction

This collaboration is 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: Team, IPTC advisors

- Luis Hernández Gómez (luisalfonso.hernandez@upm.es)
- Ana M. Bernardos Barbolla (anamaria.bernardos@upm.es)
- Juan Ignacio Godino Llorente (ignacio.godino@upm.es)
- Alberto Belmonte (alberto.belmonte@upm.es)
- Jose Luis Blanco (jl.blanco@upm.es)
- Mateo Cámara (mateo.camara@upm.es)
- Julián David Arias Londoño (julian.arias@upm.es)
- Manuel Gil Martín (manuel.gilmartin@upm.es)
- **Rubén San Segundo (ruben.sansegundo@upm.es)**

IPTC-Amazon: Team, Amazon advisors

- Adam Gabrys (gabrysa@amazon.pl)
- Giulia Comini (gcomini@amazon.co.uk)
- Ivan Valles (ivallesp@amazon.co.uk)
- Daniel Saez (dsaez@amazon.es)
- Andrzej Pomirski (pomirsa@amazon.com)
- Roberto Barra-Chicote (rchicote@amazon.co.uk)
- Vivek Yadav (ydvivek@amazon.com)
- **Jaime Lorenzo Trueba (truebaj@amazon.es)**

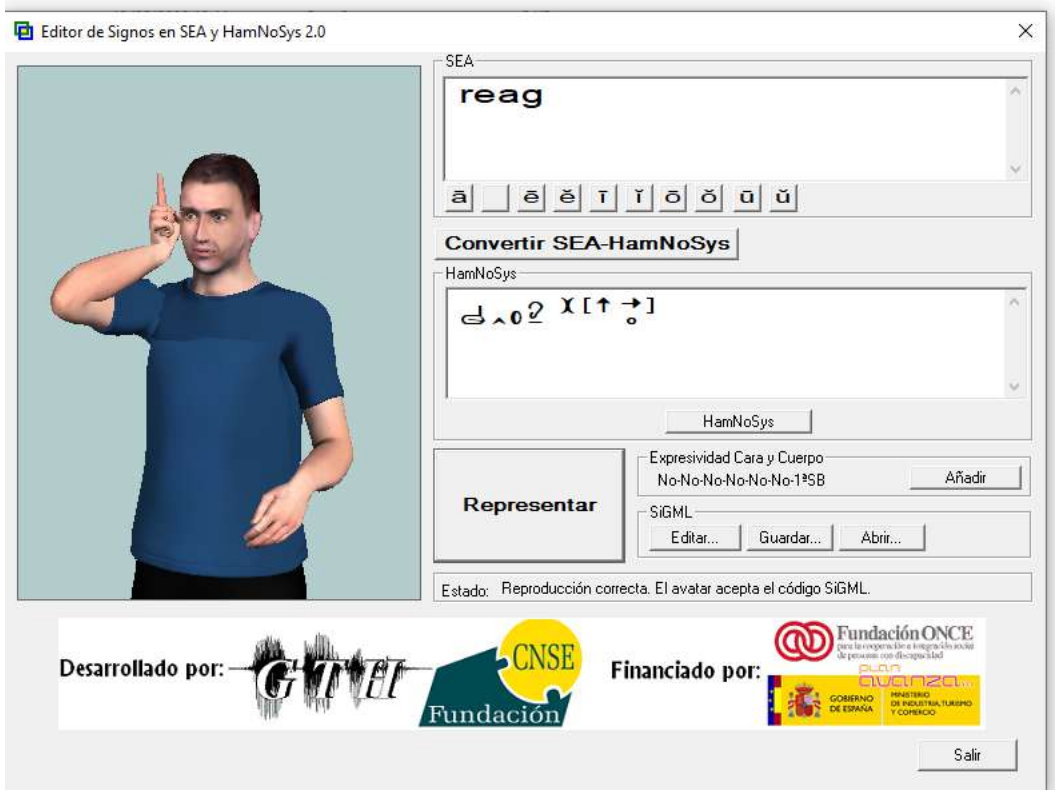
IPTC-Amazon: Main research lines

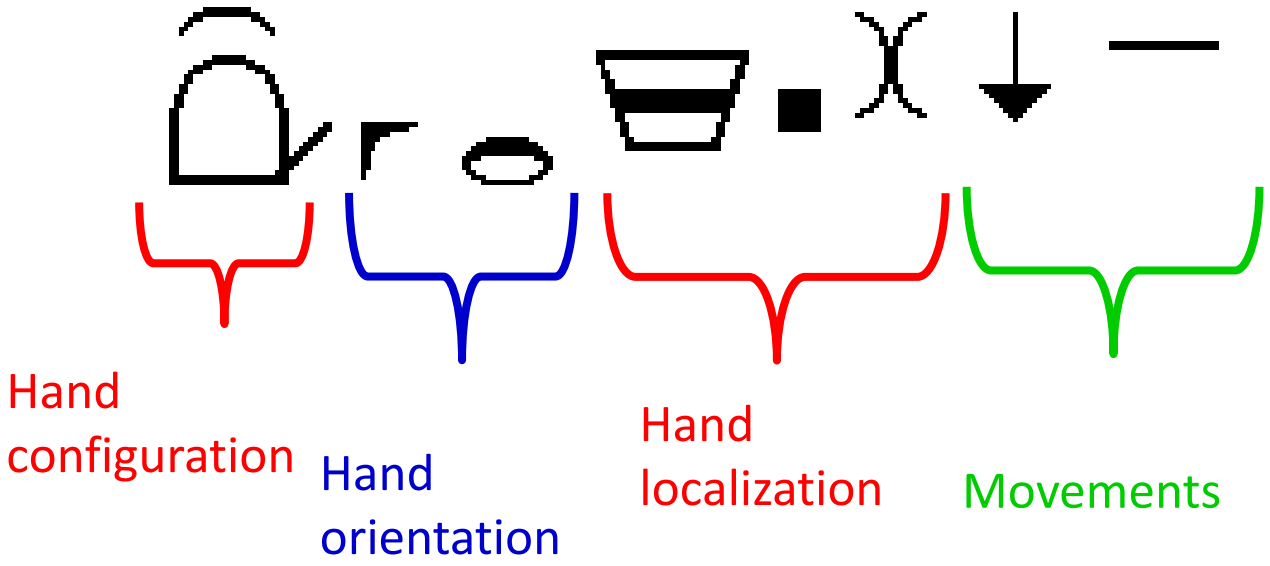
Examples of possible applications

- 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 and analysis

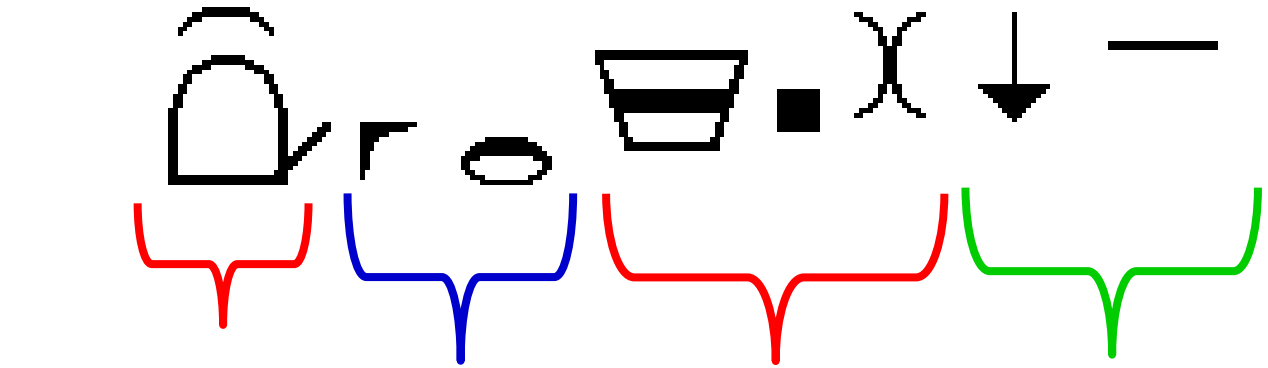
Introduction





Sign language motion generation and analysis

Introduction



Hand configuration

Hand orientation

Hand localization

Movements

Sign language motion generation and analysis

Objectives

- To generate a **dataset** containing a relevant number of signs descriptions with motion information

- To develop a **deep learning algorithm** able to associate high level sign characteristics to skeleton motion:
  - Motion generation
  - Sign classification/sign characteristics extraction

Sign language motion generation and analysis

Planning

Dataset generation:

- To obtain a **big parallel corpus** including sign characteristics and video sequences:
  - Using the first 1000 signs as references to segment videos with sign sequences using the embeddings generated from CLIP.
  - Search for specific signs in unlabeled videos.
- **Extract 2D motion characteristics from videos**: OpenPose, Mediapipe, AlphaPose, etc.

Deep learning algorithm development:

- **Motion generation** system from sign characteristics. Several deep learning strategies will be implemented and evaluated: MotionCLIP, Transformers, VAE, etc.
- Sign **characteristics detection** for sign recognition.

iptc

Speaker diarization with multimodal inputs

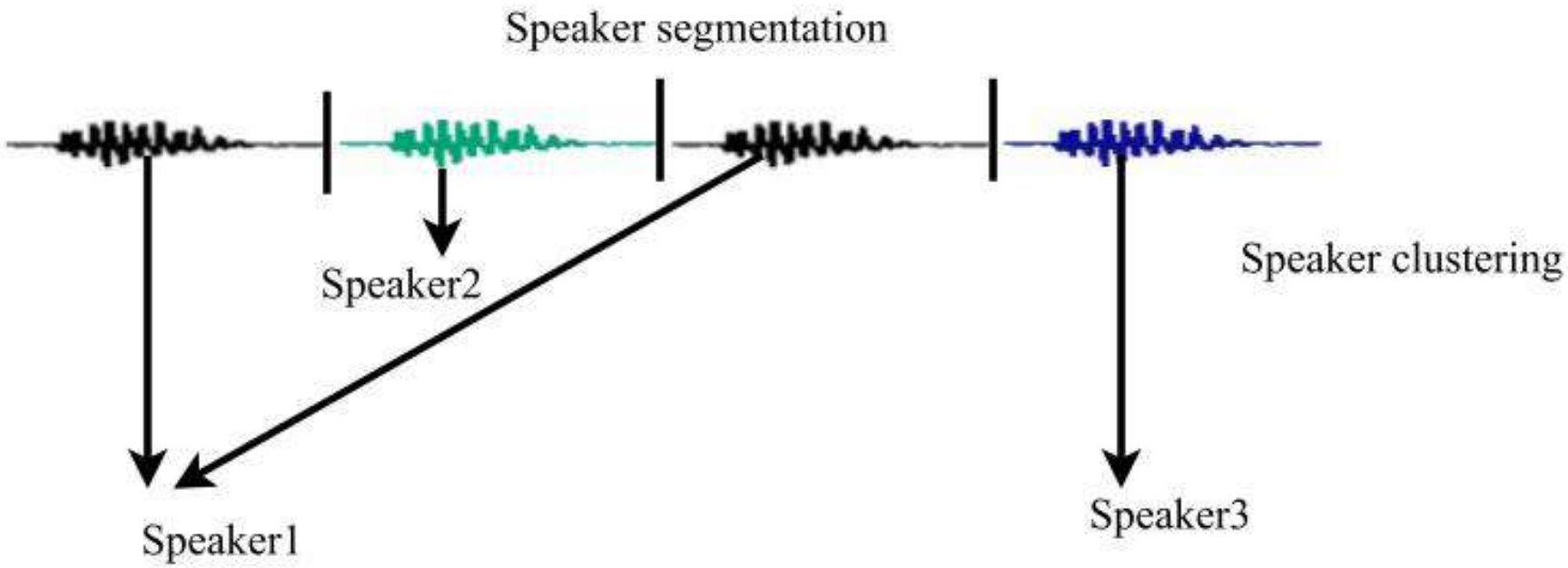
Juan Moreno
Alberto Belmonte

iptc

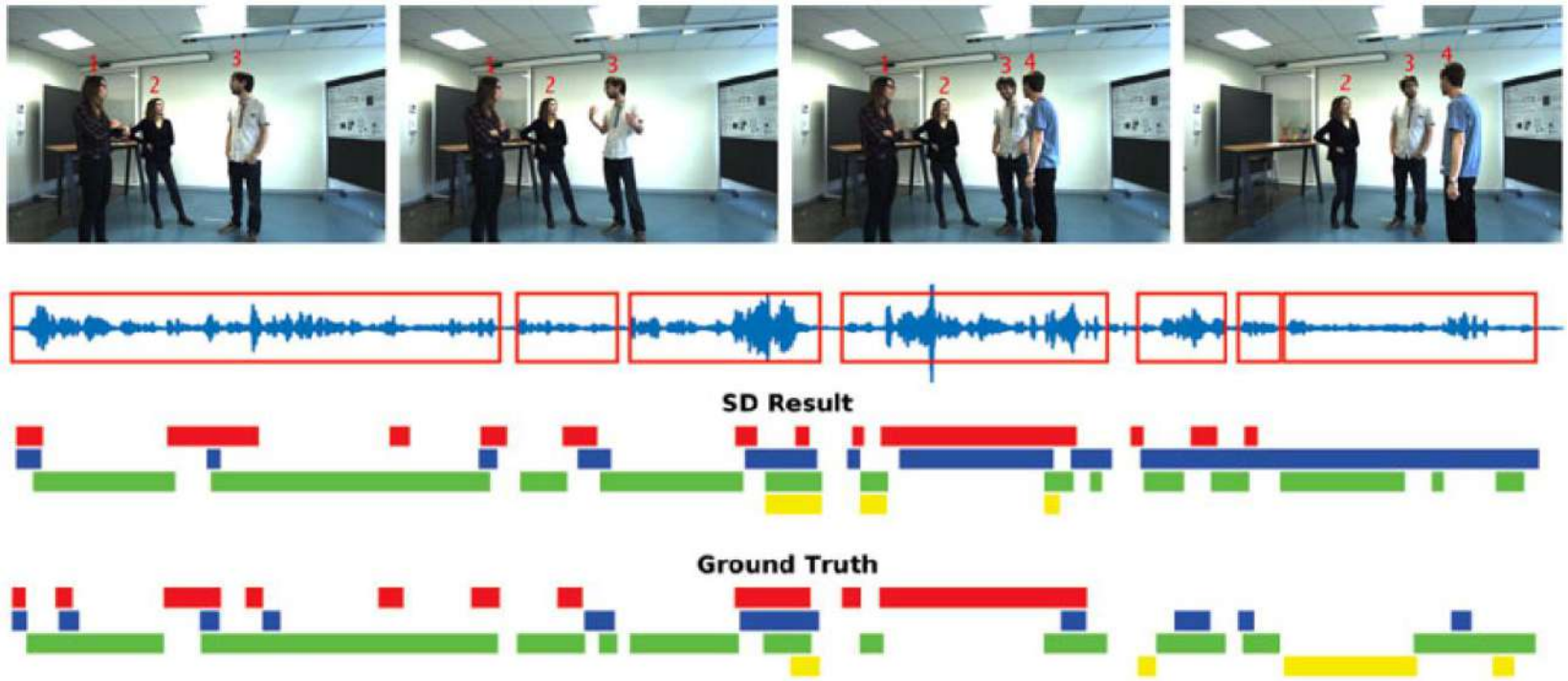
POLITÉCNICA

What is the speaker diarization?

- Answers to the question of “who spoke when”



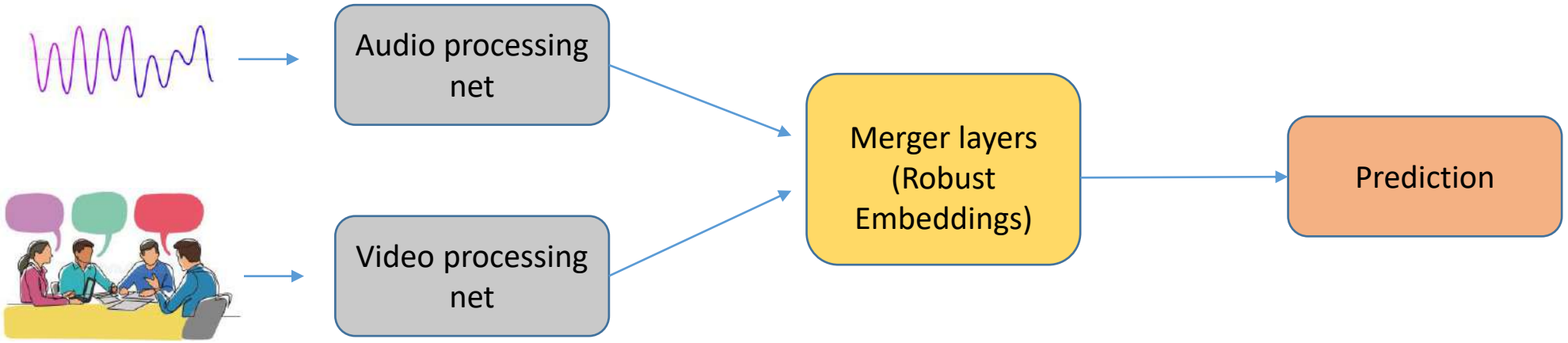
Improving the speaker diarization technique that uses only audio as input



Gebru, Israel Dejene et al. “Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2018): 1086-1099.

iptc

Who to do it? Multimodal input



- Audio Network
- Image / Video Network
- Robust features combination (embeddings)
- Predictions with multimodal embeddigs

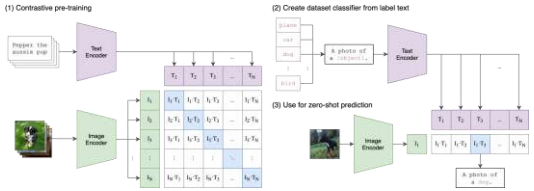
How to start?

- Paper survey (Nov 2021):
  https://arxiv.org/pdf/2101.09624.pdf

- Github review:
  https://wq2012.github.io/awesome-diarization/

GitHub

  Github CLIP:
  https://github.com/openai/CLIP
  https://github.com/moein-shariatnia/OpenAI-CLIP
  https://github.com/mlfoundations/open_clip



  Web papers with code:
  https://paperswithcode.com/task/speaker-diarization

Papers With Code

iptc

Pose and spatial movement as input for dynamic content search & generation

Introduction

Postures

Movements

Content generation

- Content search

- Posture analysis and correction:
  Physical Training & Rehabilitation
- Metaverse user avatars
- Smart space interaction

...

iptc

Pose and spatial movement as input for dynamic content search & generation

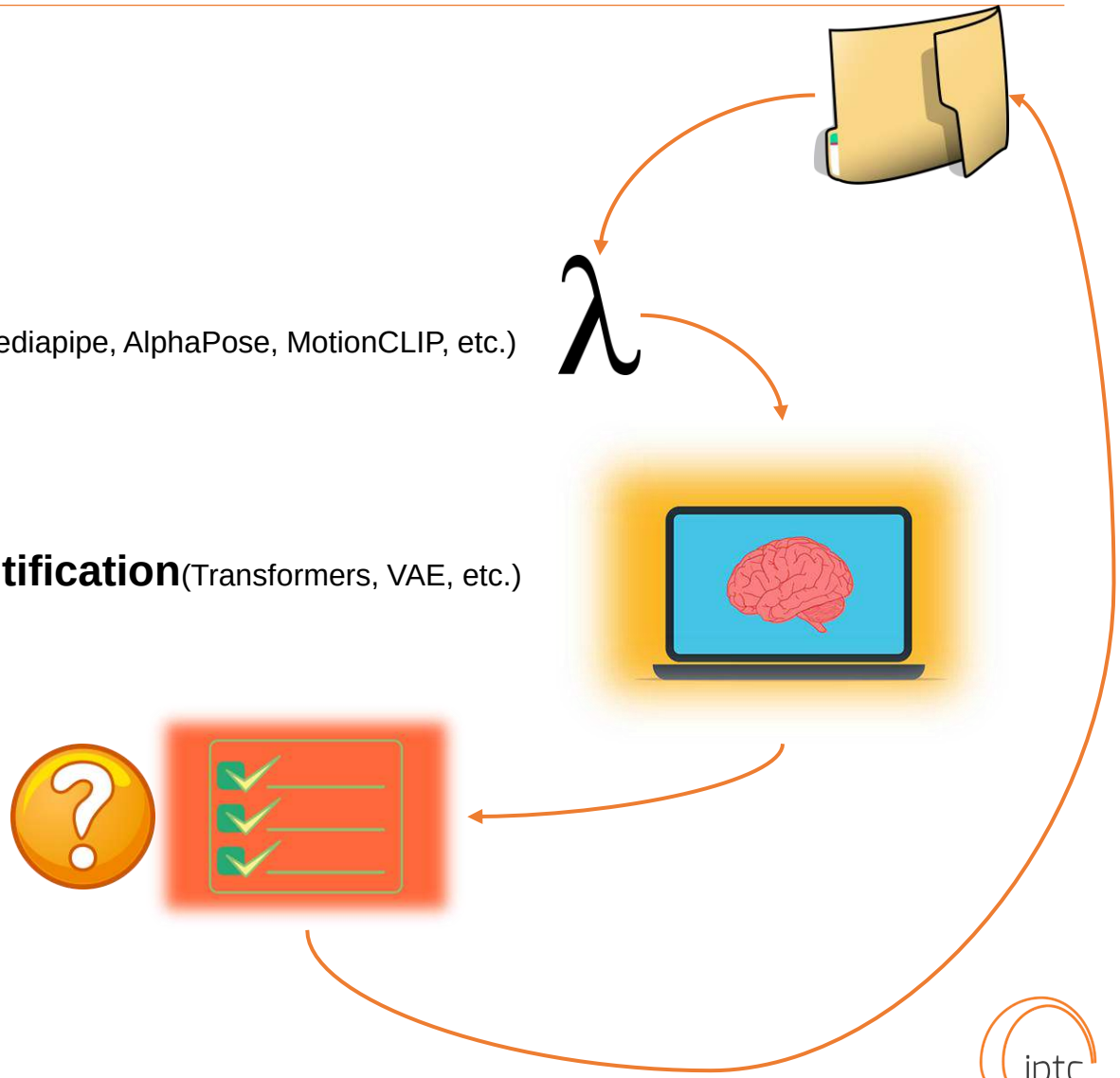
Objectives

- Posture & quality recognition component from live movement:
  - Classify poses over real physical movement.
- Define pose & execution quality features to translate them into text and image features
- Search & Generate content by using CLIP.
  - Analyze the potential of CLIP for the defined features
- Generate an integrated prototype combining the components on a camera-equipped space for the final target application.

Pose and spatial movement as input for dynamic content search & generation

Planning

1. **State of the art, dataset selection and analysis**
2. **Feature extraction & quality model definition** (OpenPose, Mediapipe, AlphaPose, MotionCLIP, etc.)
3. **AI DL Algorithm implementation for real time pose identification**(Transformers, VAE, etc.)
4. **Model evaluation**
5. **Prototype building** (MotionCLIP, OptiTrack)



iptc

Using AI to generate content

Introduction

- Platforms that generate images:

OpenAI

Dream Studio – Stable Diffusion

Midjourney

Craiyon

A comic book cover of a doctor with huge eyes

A Shiba Inu dog wearing a beret and black turtleneck

A futuristic cyborg poster hanging in a neon lit subway station

A photo of Michelangelo's sculpture of David wearing headphones djing

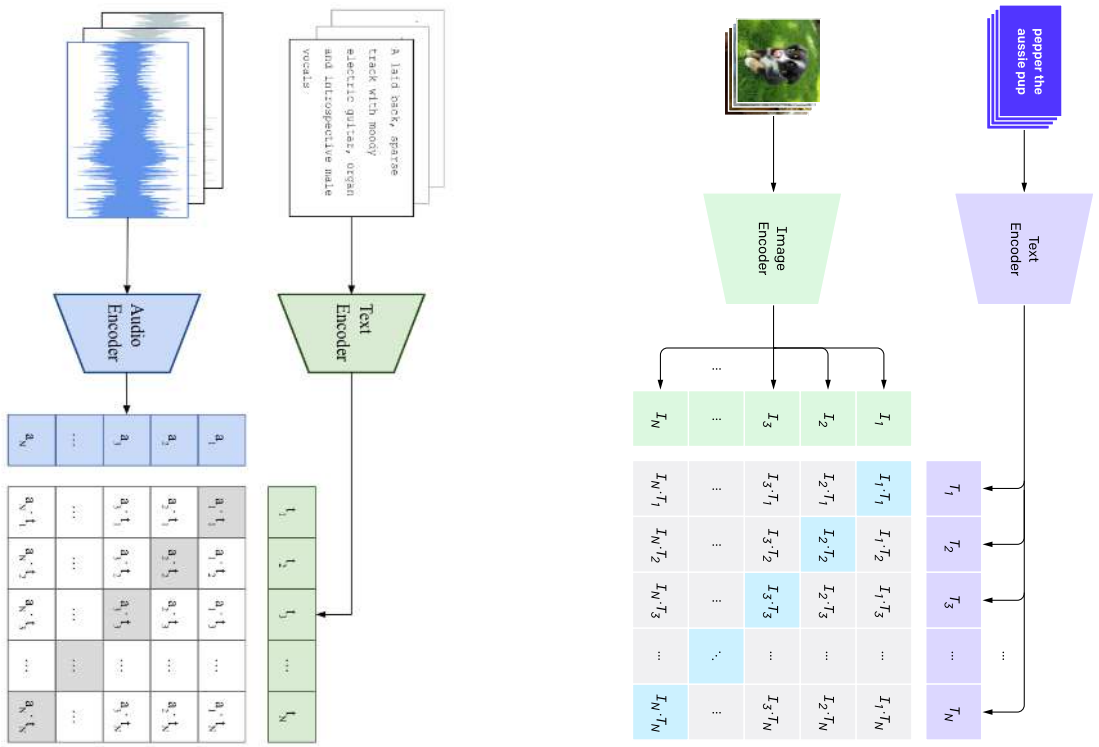
Images generated with Dall·E 2

Using AI to generate content

Introduction

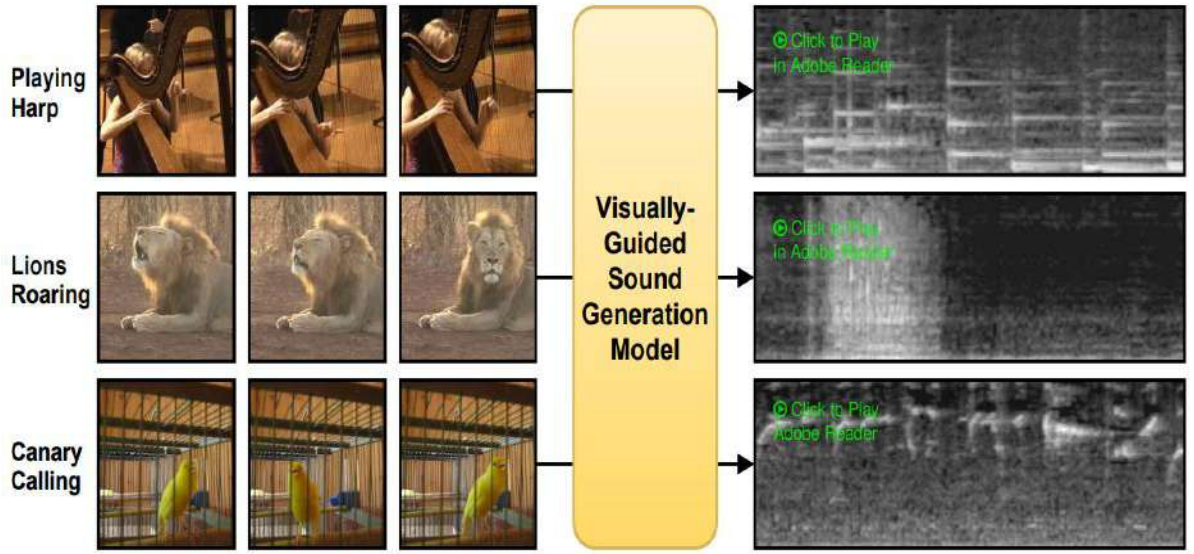
Representation correspondence

- CLIP
- CLAP



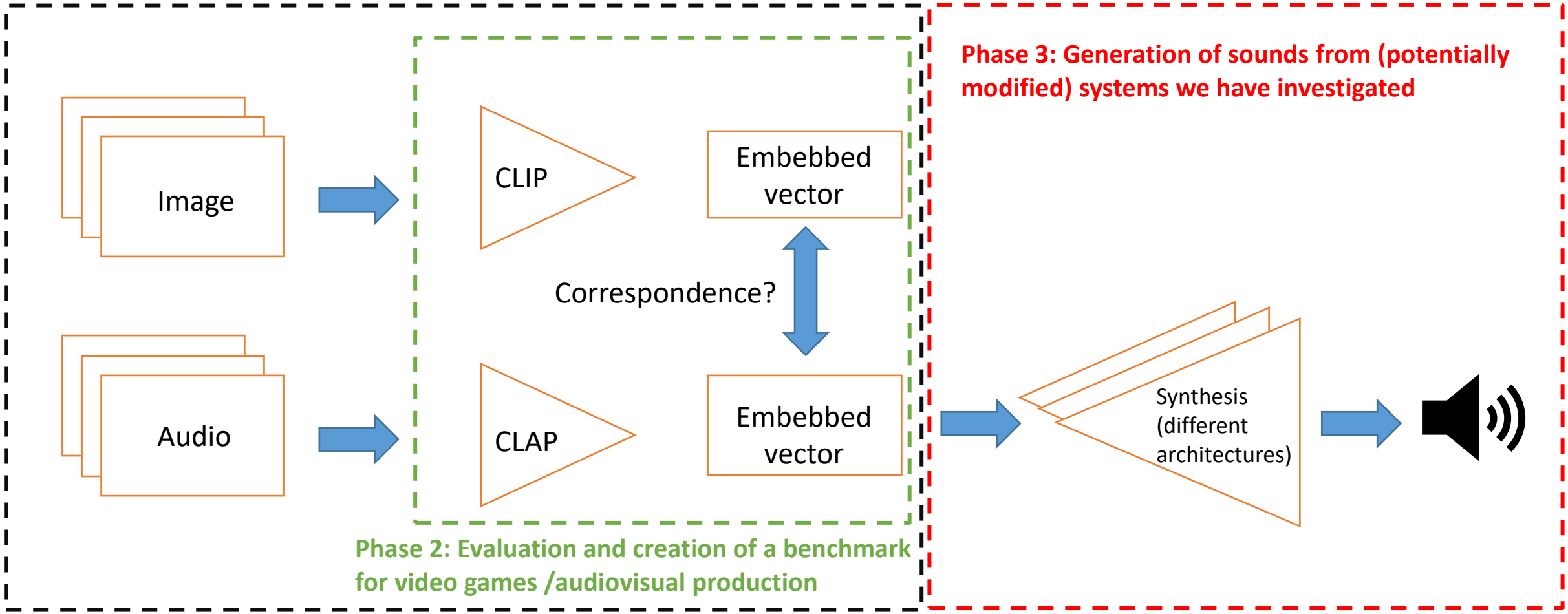
Audio generation

- IMG2WAV
- VIDEO2WAP



How can we solve the problem?

Objectives

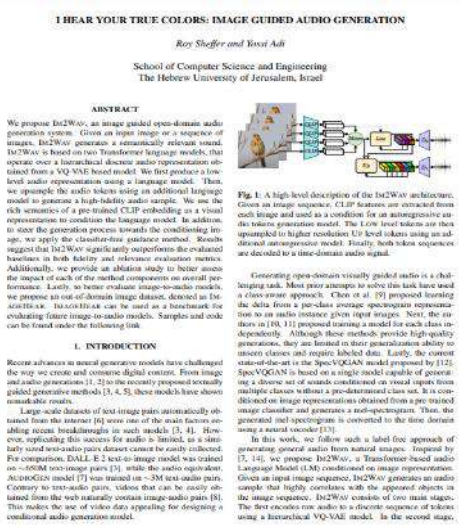


Phase 1: Analysis of existing state of the art architectures

Phase 1 - Performance evaluation

Planning

Existing state of art benchmarking



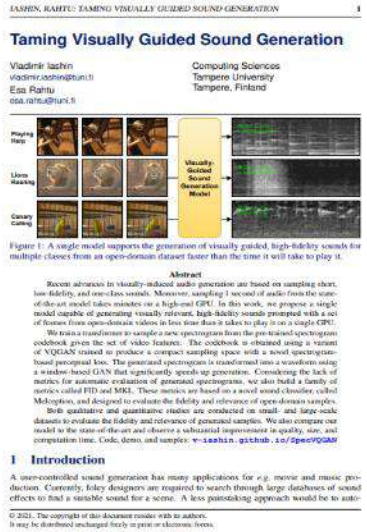
I HEAR YOUR TRUE COLORS: IMAGE GUIDED AUDIO GENERATION

Roy Sheffer and Yossi Adi

School of Computer Science and Engineering
The Hebrew University of Jerusalem, Israel

ABSTRACT

1. INTRODUCTION



IASHIN, RAHTU: TAMING VISUALLY GUIDED SOUND GENERATION

Taming Visually Guided Sound Generation

Vladimir Iashin
vladimir.iashin@tuni.fi
Esa Rahtu
esa.rahtu@tuni.fi

Computing Sciences
Tampere University
Tampere, Finland

Abstract

1 Introduction

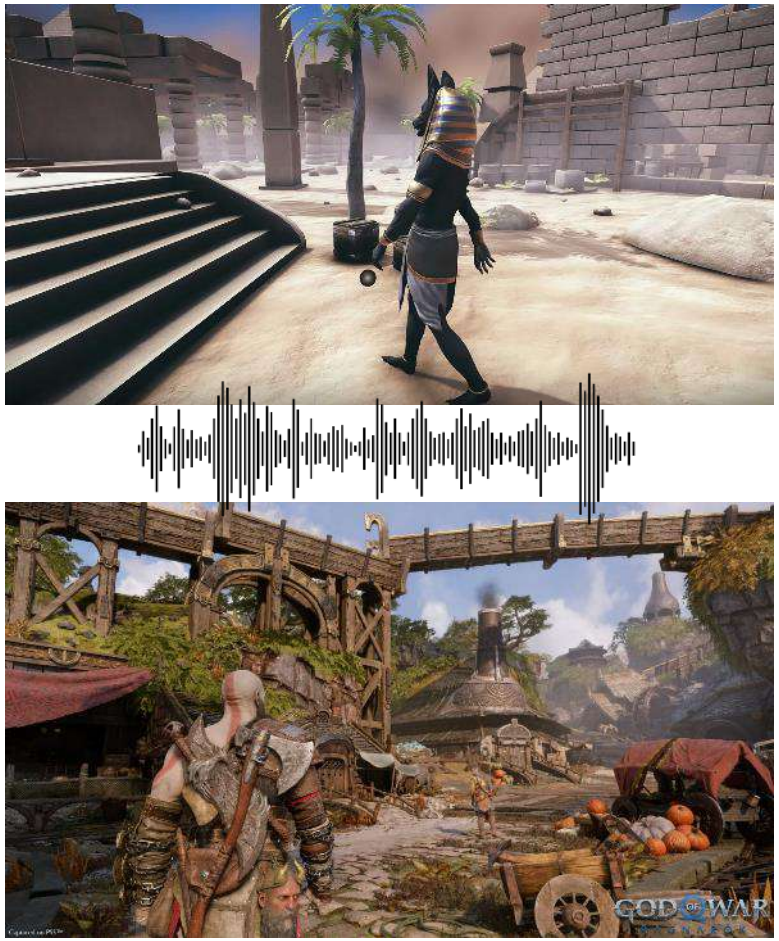
IASHIN, Vladimir; RAHTU, Esa. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, **2021**.

SHEFFER, Roy; ADI, Yossi. I Hear Your True Colors: Image Guided Audio Generation. *arXiv preprint arXiv:2211.03089*, **2022**.

DATASETS

IMAGEHEAR
VAS
VGGSound

Video games framework



Phase 2 - Audio generation using different architectures I

Planning

PHASES
Evaluation of the pre-trained net: WAV2CLIP
Entangle audios with an audio effects generator and try to connect these with GAPS VAE model system and the representation from CLIP and the VAE

Phase 2 - Audio generation using different architectures II (María)

Planning

PHASES	PRE-TRAINED NET
Evaluation of short sequences	AudioCLIP
Preparation of a scripted (text) test set	AvatarCLIP

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 will be 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
- All the students will write **detailed reports** describing all the analyses and experiments carried out.
- **Prototypes and demonstrations** to show the main research achievements.
- **Papers** submissions to international conferences or journals.

IPTC-Amazon: application 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

Rubén San Segundo
ruben.sansegundo@upm.es

Vice-director
IPTC-UPM

Information Processing and Telecommunications Center

Technologies for creating high economic and social value

POLITÉCNICA

iptc

www.iptc.upm.es