UNIVERSITAT POLITÈCNICA DE VALÈNCIA

VRAIN Valencian Research Institute for Artificial Intelligence

Automatic domain-adapted subtitling at CERN

MLLP Research Team

mllp@upv.es

www.mllp.upv.es

MLLP Machine Learning and Language Processing

Introduction

- CERN provides live (streaming) collaboration services.
- Meetings, keynotes and conferences are recorded and archived.
- CERN also produces clips on the Videos platform.
- Specific needs for:
  - accessibility,
  - lowering language barriers,
  - indexation and searchability.

Introduction

- Solution: accurate-enough automatic subtitles for:
  - Offline (recorded) multimedia material,
  - Streaming (live) webcast and videoconference meetings.
- CERN multimedia material is very specific (narrow-domain):
  - Speakers of various nationalities with strong accents (non-native).
  - Terminology from the high energy particle physics field.
  - Very heterogeneous acoustic conditions.
- A domain-adapted solution is crucial for accurate subtitling.

Introduction

- CERN’s multimedia production:
  - 30K hours of backlog (all-time).
  - 1.7K hours of new multimedia content every year.
  - 1.3K hours of live videoconferences or webcasts every year.
- On-premises solution, avoiding variable costs.
- Taking advantage of new data.
- MLLP was contacted (2020) to explore possible solutions.

MLLP research group

Machine Learning and Language Processing (MLLP),
Valencian Research Institute on Artificial Intelligence (VRAIN),
Universitat Politècnica de València (UPV).

Members:

- 15 researchers (5 lecturers, 2 postdocs, 3 PhD students).

Areas:

- Automatic Speech Recognition (ASR),
- Speaker Diarization (SD),
- Machine Translation (MT),
- Speech Translation (ST),
- Text-to-Speech (TTS).

Competitive R&D Projects:

transLectures
Transcription and Translation of Video Lectures
2011–14 (FP7)

EMMA
EUROPEAN MULTIPLE MOOC AGGREGATOR
2014–16 (FP7)

X5GON
2017–20
(H2020)

INTERACT EUROPE
2022–26*
(EU4Health)

+ other related European, Spanish and Valencian projects

Technology transfer contracts:

à
2020-2023,
2025-2028*

CERN
2020,
2022-2024

+ other: EP, CdT EU, AppTek, JSI, HPI, ULisboa. . .

Automatic Speech Recognition (ASR)

- Internal software for multilingual streaming (live) transcription.
- Multiple European languages supported: Ca, Es, En, Fr, De. . .
- Adaptation of the technology to each organization's needs.
- Cloud service or on-premises deployment.

Competitions

Winner of *2018 RTVE Speech-to-Text Challenge*

Winner of *2021 RTVE Speech-to-Text Challenge*

Selected papers

LHCP-ASR: An English Speech Corpus of High-Energy Particle Physics Talks for Narrow-Domain ASR Benchmarking (J. Santamaría et al. 2025 [submitted])

Live Streaming Speech Recognition Using Deep Bidirectional LSTM Acoustic Models and Interpolated Language Models (J. Jorge et al. 2021)

Machine Translation (MT)

- Simultaneous streaming machine translation.
- Competitive translation quality in European language pairs.
- MT systems deployed for any pair of languages on demand.

Competitions

Winner of IWSLT 2022 *Speech-to-Speech Translation*

2nd place in the IWSLT 2022 *Simultaneous Speech Translation*

Selected papers

Segmentation-Free Streaming Machine Translation (J. Iranzo et al. 2024)

Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates (J. Iranzo et al. 2020)

Text to Speech (TTS)

- Multilingual streaming text-to-speech.
- Cross-lingual automatic dubbing.
- Supported languages: Ca, Es, En, Fr, De.

Competitions

2nd place in the 2021 *Blizzard Speech Synthesis Challenge*

Selected papers

Towards cross-lingual voice cloning in higher education (A. Pérez et al. 2021)

Towards simultaneous machine interpretation (A. Pérez et al. 2021)

UPV-CERN Pilot project (2020)

- **Period**: June 2020 - November 2020 (5 months).
- **Budget**: 5K Euros.
- **Objectives**:
  - Identify in-domain data for training/adaptation and evaluation.
  - Report baseline transcription and translation quality measures.
  - Explore and assess domain-adaptation techniques for ASR.
- **Results**:
  - Definition of training and evaluation datasets for ASR.
  - Promising results on domain adaptation.

UPV-CERN Tender project (2022-24)

- **Period**: February 2022 - August 2024 (30 months).
- **Budget**: 139K Euros.
- **Objectives**:
  - Develop domain-adapted (live) subtitling systems for CERN.
  - On-premises deployment of the complete solution.
  - Ad-hoc solution to integrate live subtitling into Zoom.
  - Auto-training solution for continuous improvement of systems.
- **Results**:
  - State-of-the-art in-domain ASR and MT systems deployed.
  - More than 30K hours of backlog videos subtitled.
  - Successful integration with Zoom.
  - Auto-training module developed and deployed.

In-domain data sources

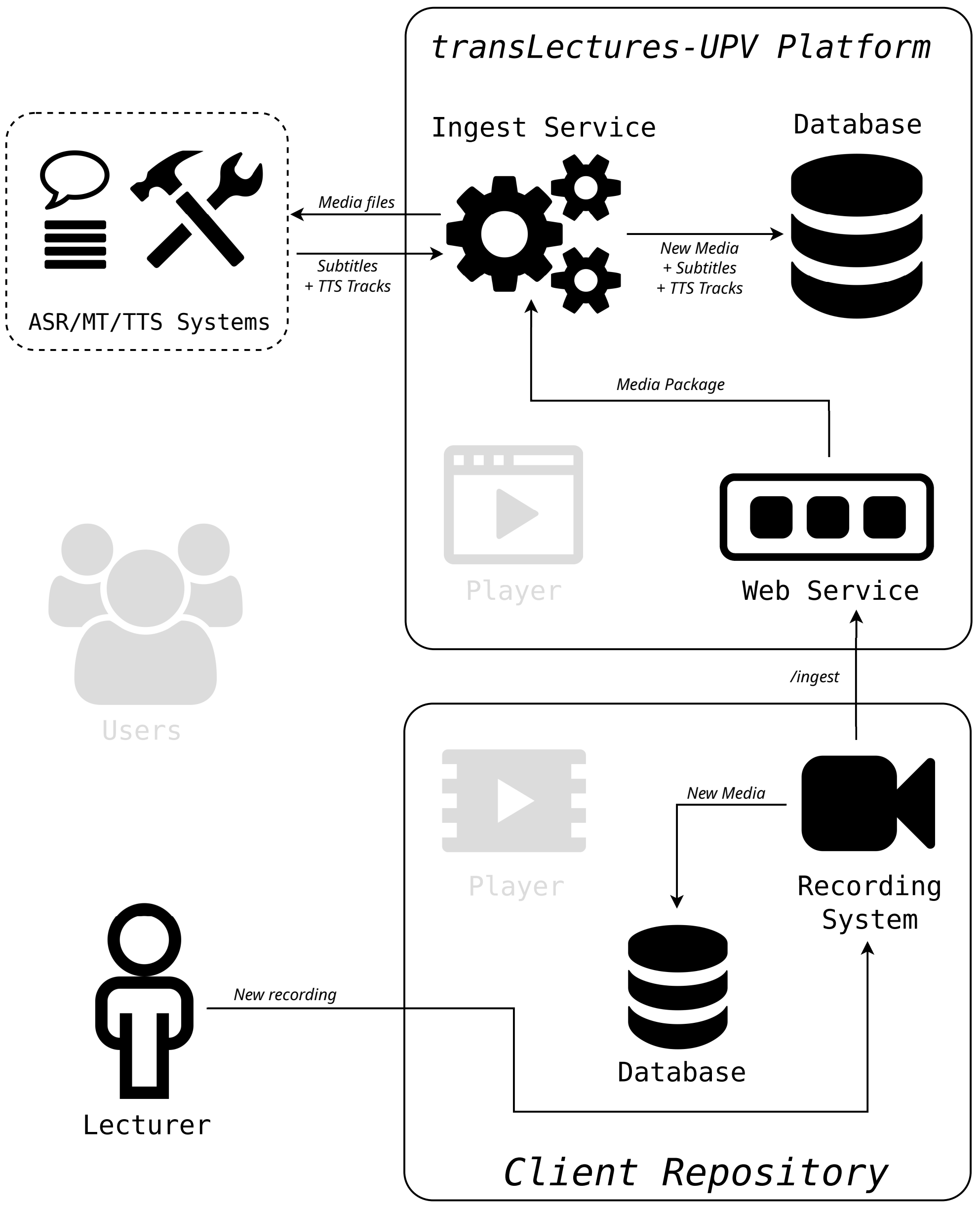
- **CERN Opencast**: heterogeneous set of conferences, seminars...
- **LHCP**: recordings from the 2020-2022 LHCP conferences.
- **e-learning**: short formative video tutorials for CERN workers.
- **Digital Memory**: audio recordings from (non-)technical meetings.
- **CERN Document Server (CDS)**: +550K records of papers, theses.
- **CERN News**: CERN news since 1993 in French and English.

Deployment of the solution

- On-premises (OpenStack) installation of MLLP's software.
- Off-line (recorded videos) subtitling:
  - *TLP: The transLectures-UPV Platform.* (UPV).
  - Database, API, Ingest Service, Front-end, Subtitle editor (Player).
- Live (streaming) subtitling:
  - *TT-Streaming: RPC API for subtitling live audio streams* (UPV).
- ASR Systems (software + models).
  - *TLK* and *pyTLK* (UPV).
  - *Fairseq* (Meta).
- MT Systems (software + models):
  - *Fairseq* (Meta).

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Off-line subtitling deployment



transLectures-UPV Platform

Ingest Service

Database

ASR/MT/TTS Systems

Media files

Subtitles
+ TTS Tracks

New Media
+ Subtitles
+ TTS Tracks

Media Package

Player

Web Service

Users

/ingest

New Media

Player

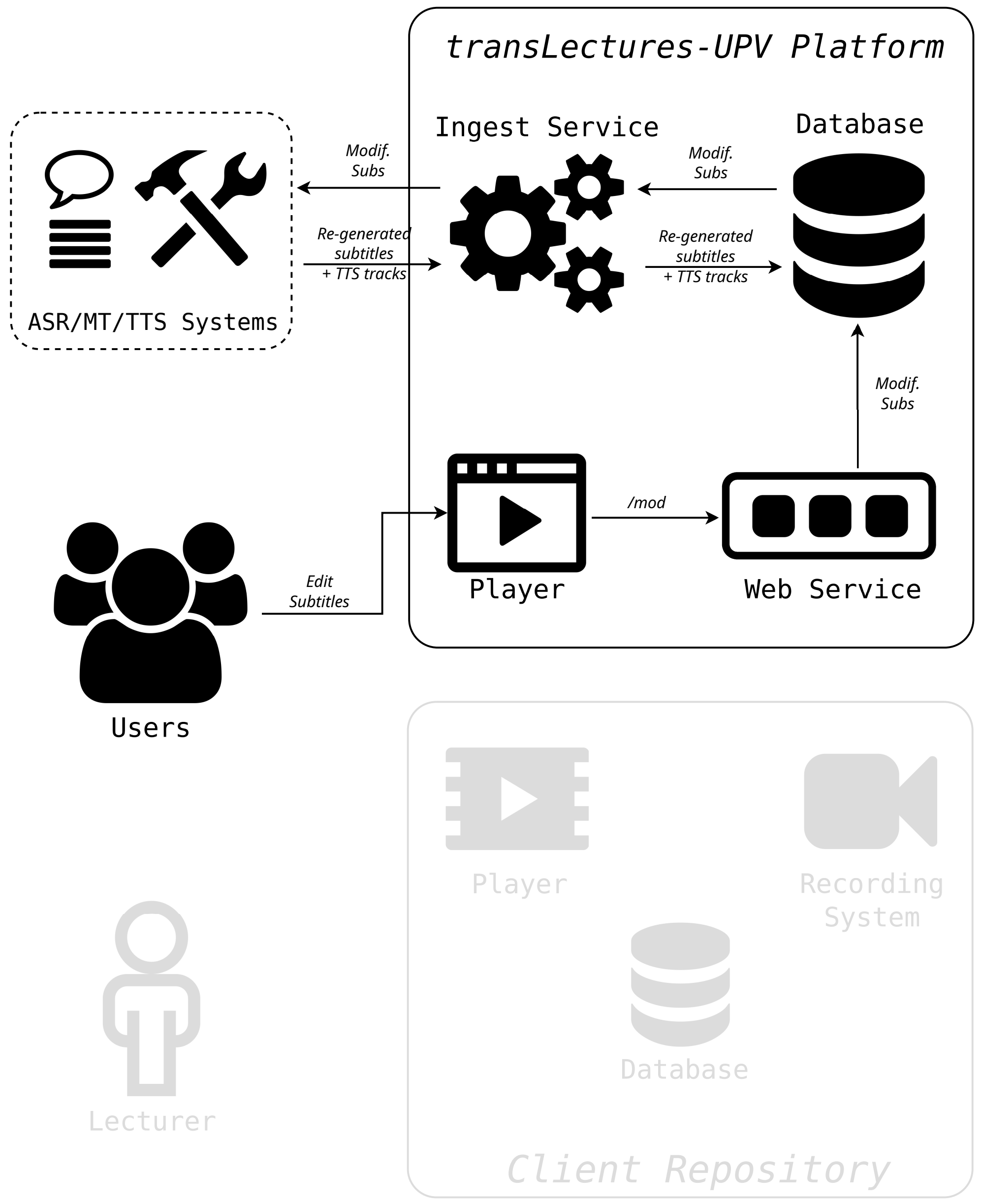
Recording
System

Database

New recording

Lecturer

Client Repository



transLectures-UPV Platform

Ingest Service

Database

ASR/MT/TTS Systems

Modif.
Subs

Re-generated
subtitles
+ TTS tracks

Modif.
Subs

Re-generated
subtitles
+ TTS tracks

Modif.
Subs

/mod

Player

Web Service

Edit
Subtitles

Users

Player

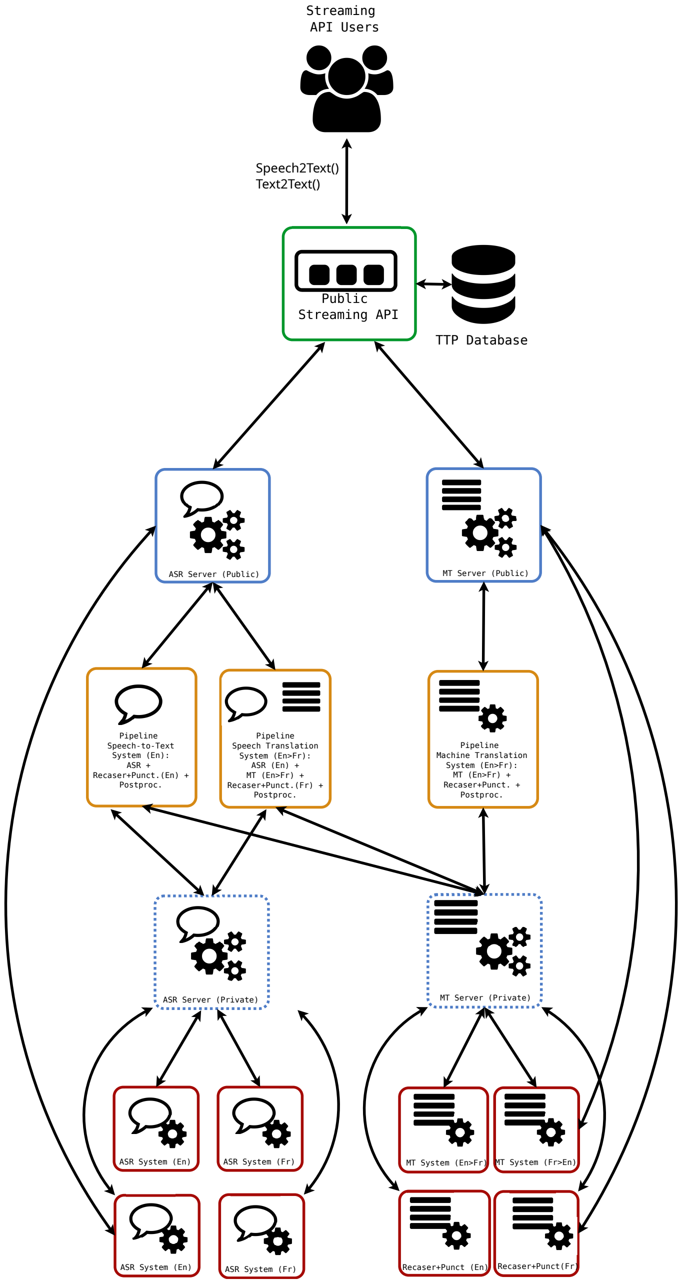
Recording
System

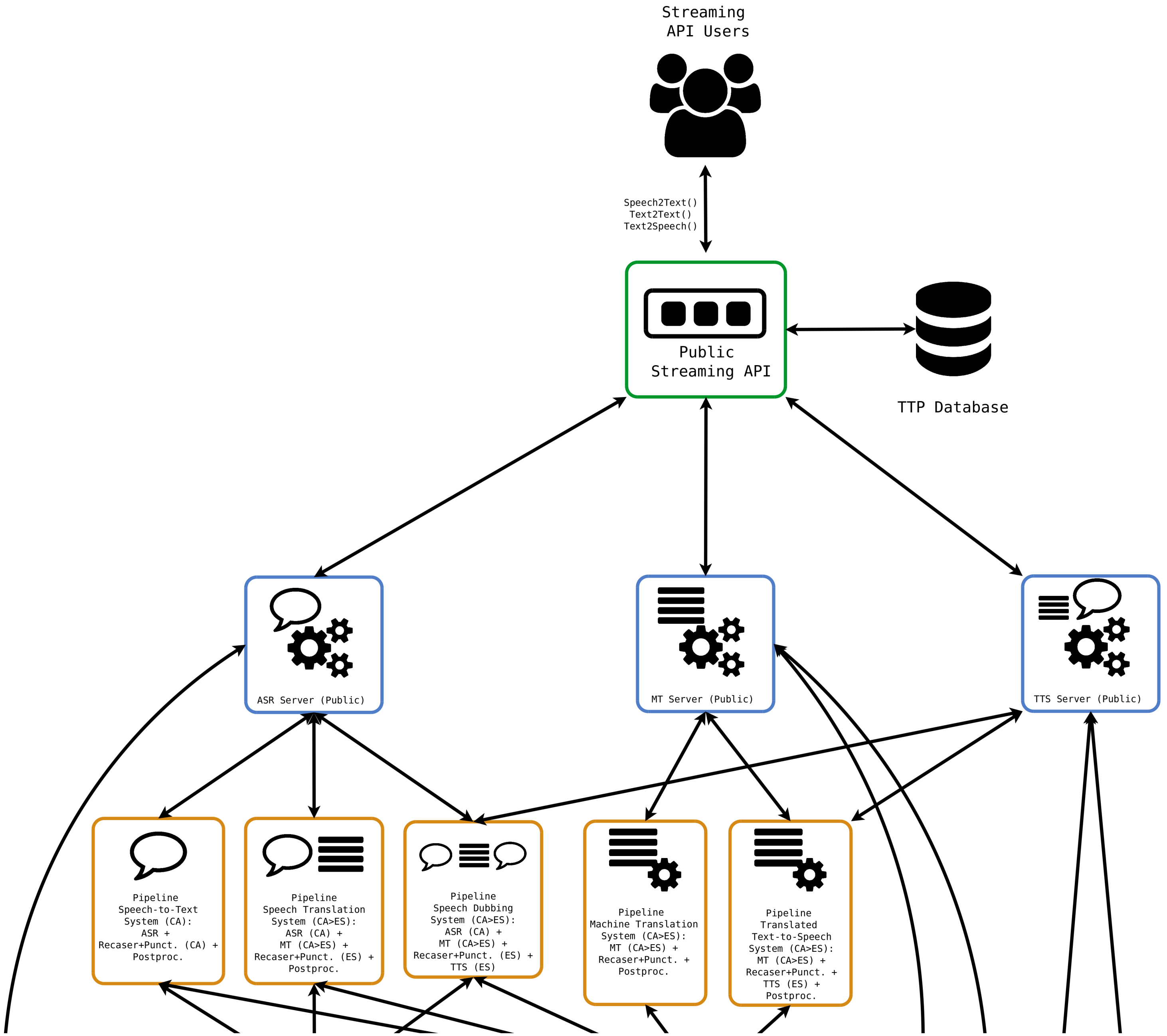
Database

Lecturer

Client Repository

Streaming subtitling deployment





Streaming API Users

Speech2Text()
Text2Text()
Text2Speech()

Public Streaming API

TTP Database

ASR Server (Public)

MT Server (Public)

TTS Server (Public)

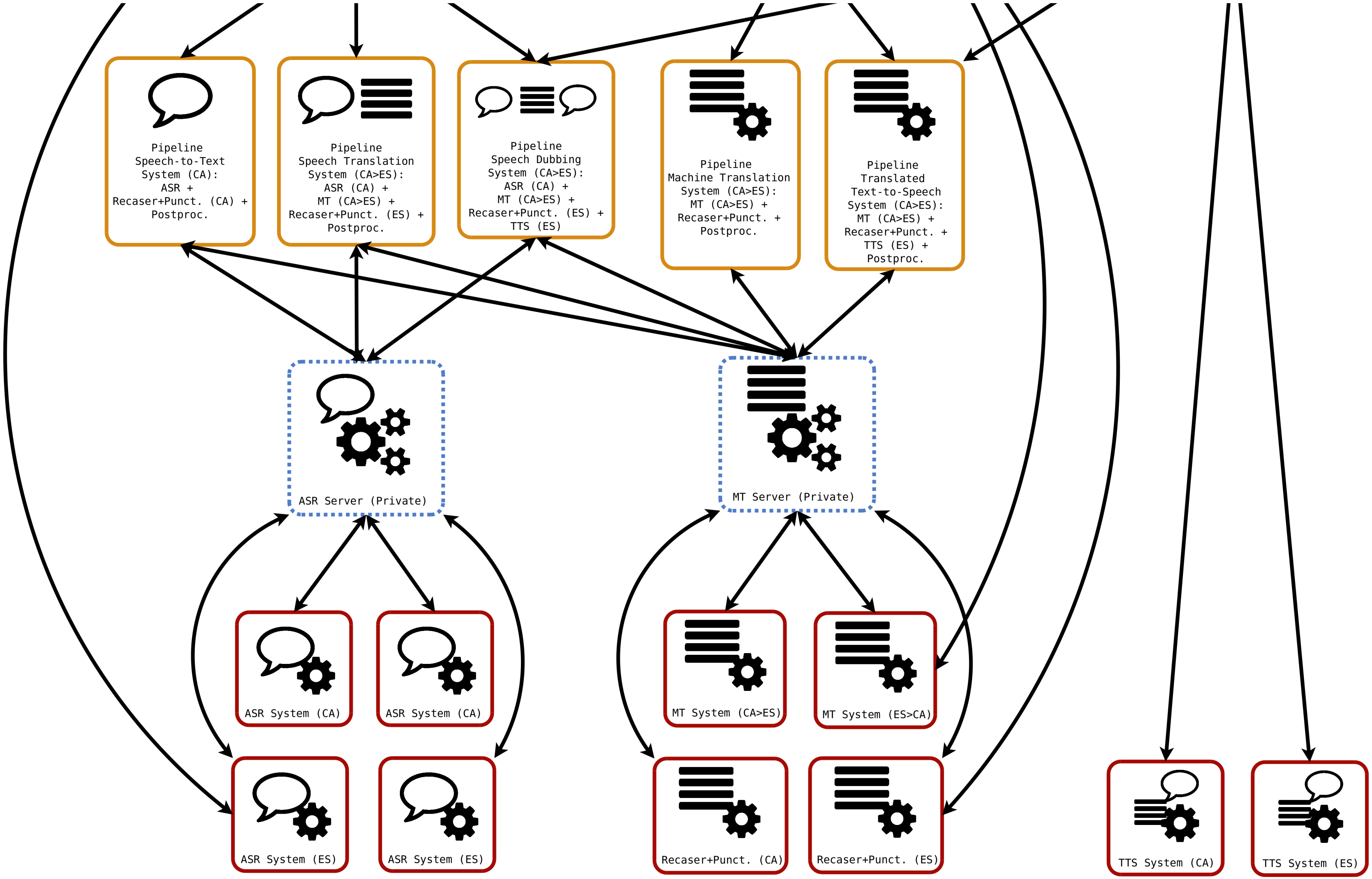
Pipeline Speech-to-Text System (CA): ASR + Recaser+Punct. (CA) + Postproc.

Pipeline Speech Translation System (CA>ES): ASR (CA) + MT (CA>ES) + Recaser+Punct. (ES) + Postproc.

Pipeline Speech Dubbing System (CA>ES): ASR (CA) + MT (CA>ES) + Recaser+Punct. (ES) + TTS (ES)

Pipeline Machine Translation System (CA>ES): MT (CA>ES) + Recaser+Punct. + Postproc.

Pipeline Translated Text-to-Speech System (CA>ES): MT (CA>ES) + Recaser+Punct. + TTS (ES) + Postproc.



Pipeline
Speech-to-Text
System (CA):
ASR +
Recaser+Punct. (CA) +
Postproc.

Pipeline
Speech Translation
System (CA>ES):
ASR (CA) +
MT (CA>ES) +
Recaser+Punct. (ES) +
Postproc.

Pipeline
Speech Dubbing
System (CA>ES):
ASR (CA) +
MT (CA>ES) +
Recaser+Punct. (ES) +
TTS (ES)

Pipeline
Machine Translation
System (CA>ES):
MT (CA>ES) +
Recaser+Punct. +
Postproc.

Pipeline
Translated
Text-to-Speech
System (CA>ES):
MT (CA>ES) +
Recaser+Punct. +
TTS (ES) +
Postproc.

ASR Server (Private)

MT Server (Private)

ASR System (CA)

ASR System (CA)

MT System (CA>ES)

MT System (ES>CA)

ASR System (ES)

ASR System (ES)

Recaser+Punct. (CA)

Recaser+Punct. (ES)

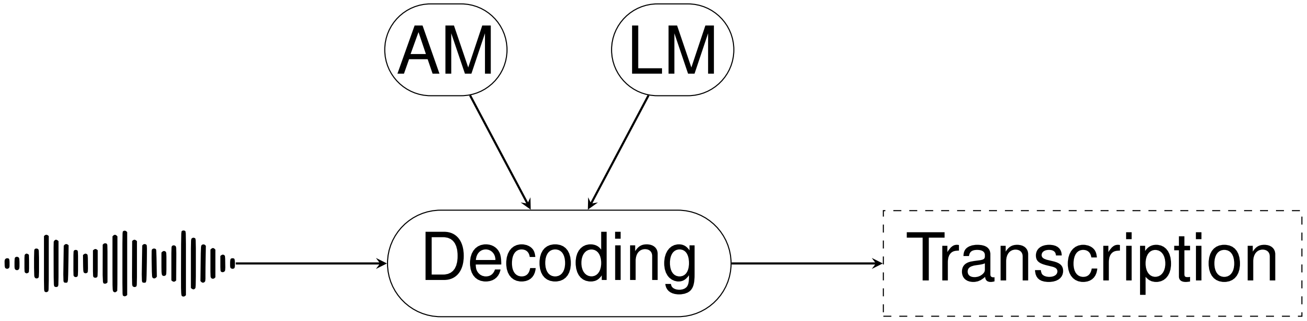
TTS System (CA)

TTS System (ES)

System auto-training

- Automatically enhance ASR systems, on a regular basis.
- Make them learn from newly produced resources and materials.
- Upgraded ASR models are able to recognize novel terminology.
- Scientific and engineering challenge.
- Main steps:
  - Gather and prepare new data,
  - Model training and assessment (sanity checks),
  - Upgraded system construction and deployment (Docker).
- Ad-hoc dockerised solution.

ASR : Hybrid Architecture



- Given an input audio, ASR systems output verbatim transcriptions
- Speech preprocessed into digital signal
- Two independent models: Acoustic Model and Language Model
- Combined for decoding

ASR: Training data

- **Opencast**
  - 494 videos, for a total of 430 hours
  - Training (domain adaptation)
- **LHCP 2020 conference**
  - 27 videos, total length of 12 hours
  - Evaluation dataset
- **CERN Document Server**
  - 543K documents for a total of 1.1G words

ASR : Assessment (WER)

- **Word Error Rate (WER%)**: incorrectly transcribed words

$$WER = \frac{I + D + S}{R} \cdot 100$$

- I: Insertions, D: Deletions, S: Substitutions, R: Reference Words

REF:	GeV	to	TeV	scale	FIP	particles.	dark	scalars	dark		photons
HYP:	GeV	to	TV	scale		particles.	dark	scalars	dark	chocolate	photons

- This hypothesis has 30% of WER

ASR : Assessment (WER)

- **Word Error Rate (WER%)**: incorrectly transcribed words

$$WER = \frac{I + D + S}{R} \cdot 100$$

- I: Insertions, D: Deletions, S: Substitutions, R: Reference Words

- WER $\leq$ 30% → profitable for indexing and semantic representation
- WER $\leq$ 20% → usable for subtitling
- WER $\leq$ 10% → high-quality transcriptions
- WER $\leq$ 5% → human-quality transcriptions

ASR : Challenges of (live) streaming

- Working with an unbounded speech signal
- Cannot process full context of the signal
- Only a few tenths of a seconds of future context can be considered
  - Typically 500 ms
- Real-Time Factor $<$ 1 necessary but not sufficient condition
- Trade-off between quality and latency

ASR : LHCP 2020

- Need of dev/test set for tuning/evaluation of ASR systems
- Manual transcriptions by 5 CERN volunteers
- Transcription process followed specific guidelines

Set	#videos	Duration (h)
Dev	14	5.8
Test	15	5.9

ASR : Baseline system

- **Acoustic Model (AM)** based on BLSTM-DNN
  - 6K hours of transcribed general-purpose audio
- **Language Model (LM)** interpolation:
  - 4-gram LM - 18G words from general-purpose text
  - Transformer LM - Subset of 1G words
- This system scored 24% WER on LHCP-2020-test

ASR : LM-adapted system

- **Adapted Transformer Language Model (TLM)**:
  - Replace the general-purpose TLM
  - 1G words of in-domain content published before 2020
  - Closed vocabulary of 250K words
  - Relative improvement of 18% w.r.t. baseline system

	Test
Baseline	24.0
LM-adapted	19.7

ASR : LM- and AM-adapted system

- **Fine-tuned Acoustic Model**:
  - Replace the general-purpose AM
  - 423 hours of in-domain pseudo-labelled acoustic data
  - Relative improvement of 17% w.r.t. previous system

	Test
Baseline	24.0
LM-adapted	19.7
LM- and AM-adapted	16.3

ASR : Massive pseudo-labelling of speech data

- About 12K hours of in-domain videos (Opencast)
- Automatic transcribed with LM- and AM-adapted system
- Filtering process based on phoneme-length heuristics
- Reduced to 9K hours of in-domain speech data

ASR : Fully Adapted ASR system

- **Same adapted TLM used by previous systems**
- **New Acoustic Model based on Conformer architecture**:
  - 9K hours of in-domain pseudo-labelled acoustic data
  - Cumulative relative improvement of 43% w.r.t. the baseline

	Test
Baseline	24.0
LM-adapted	19.7
LM- and AM-adapted	16.3
Fully Adapted system	13.6

ASR : LHCP 2022

- Second evaluation task to double-check quality through time
- 43 videos from Plenary Talks, for a total of 18.2 hours
- Manual transcriptions by 8 ASR researchers (not experts)
- Same guidelines as LHCP 2020, with new rule: `<UNK>`
- Average revision effort of 8.1 Real-Time Factor (RTF)

Set	#videos	Duration (h)
Dev	11	4.8
Test	32	13.4

ASR : Comparison with Whisper

- **OpenAI’s Whisper**, with 680K hours of general-purpose data
  - Medium (769M parameters)
  - Turbo (809M parameters)
- **Our Fully Adapted system**, with 9K hours of in-domain data
  - AM + LM for a total of 538M parameters

ASR System	*LHCP-2020*	*LHCP-2022*
Whisper-turbo	15.9	17.7
Whisper-medium	15.4	16.7
Our Fully Adapted system	**13.6**	**15.0**

ASR : LHCP-ASR paper

- **Release of LHCP-ASR dataset** (LHCP 2020–2022 editions)
- Describes its creation and provides reference WERs
- Two evaluation partitions for a total of 30 hours of verbatim data
- 205 hours of automatic trancriptions for training/adaptation
- **Submitted to InterSpeech '25 (right yesterday!)**

Auto-training: Steps

- Acquisition of Δ-dataset from CDS and CERN News
- Text data extraction and cleaning
- Data partition
- System vocabulary extension
- Transformer LM finetuning and assessment
- ASR system sanity check
- System dockerisation and deployment

Auto-training: Data acquisition

- Acquisition of Δ-dataset from CDS and CERN News
- Specific crawlers for both data sources
- Automatically collects all the available data given a period
- Minimum amount of data threshold

Auto-training: Experimental setup

- Baseline ASR system: trained with data until June 2020
- Updated system, using data from July 2020 to December 2023
- Δ-dataset of 5 months, January 2024 to May 2024

Subset	Words
Δ-train	9.7M
Δ-dev	37.9K
Δ-test	37.2K

Auto-training: Evaluation

- OOV%: Percentage of words not present in system vocabulary
- PPL: Perplexity of the Language Model

ASR systems	Train data	Δ-test		LHCP2020		
	up to year	OOV%	PPL	OOV%	PPL	WER
Baseline	2020	4.2	119	1.7	63	13.8
Updated	2023	2.8	85	1.5	59	13.5
Auto-trained	2024	2.5	72	1.5	60	13.7

- OOV and PPL improves on Δ-test
- LHCP2020's PPL slightly degrades as LM is biased to future
- WER computed solely for sanity checking purposes

Machine Translation

- Tasks:
  - Translation of transcriptions obtained from pre-recorded videos
  - Simultaneous translation of live speech from real-time transcription
- Overview for both tasks:
  - Datasets
  - Automatic evaluation
  - System description
  - Experimental results

CERN Evaluation Datasets

- Parallel texts available in French and English for MT evaluation
- **CERN News**: News available in the official website[1] on a variety of topics, such as Physics, Accelerators, Experiments, Engineering, Computing, Knowledge sharing, At CERN.
- **CERN Theses**: Parallel thesis abstracts with significantly more technical vocabulary combined with mathematical expressions

Datasets	Sentence pairs	English words	French words
CERN News	1799	44K	50K
CERN Theses	911	23K	25K

[1] https://home.cern/news

Quality evaluation in MT

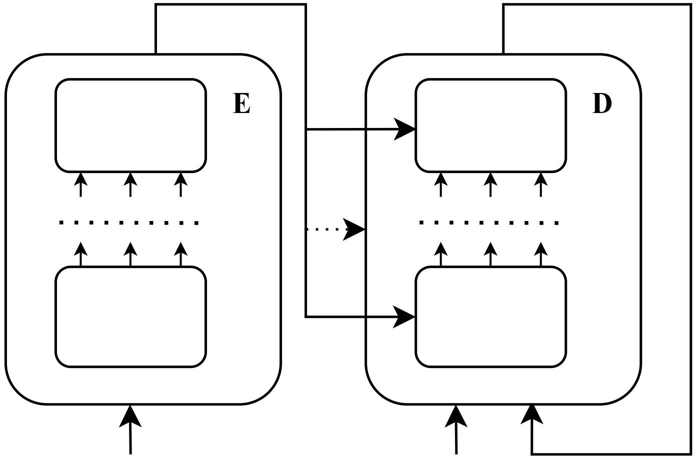
- Automatic evaluation compares machine and human translation
- Evaluation provides a score to measure translation quality
- Automatic evaluation is an open problem

Source	ATLAS récompense ses meilleures thèses 2021.
Ref	ATLAS celebrates its 2021 Thesis Award winners.
Auto 1	ATLAS rewards its best 2021 theses.
Auto 2	ATLAS honors its top theses of 2021.

- Selected evaluation score: Bilingual Evaluation Understudy (BLEU)
- BLEU: Degree of overlap between machine and human translation.
- BLEU: The higher the better. $\geqslant$40 indicates good quality.

MT systems

- Encoder-decoder Transformer architecture trained from scratch



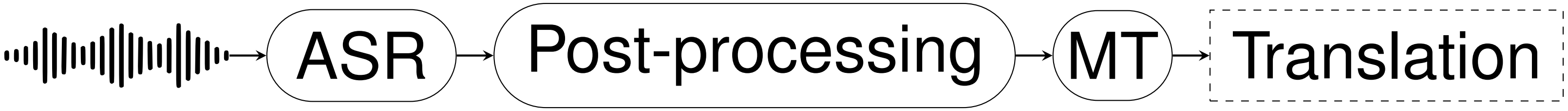
- Two systems:
  - 6-layer BIG variant (0.3B parameters)
  - 12-layer variant with pre-layer normalization (0.6B parameters)
- Comparison with NLLB: pre-trained encoder-decoder multilingual MT models.

MT results

		BLEU	
System	# params	CERN News	CERN Theses
CERN-Sep22 (6-layer)	0.3B	38.8	40.9
CERN-Nov23 (12-layer)	0.6B	**40.2** (+5.4%)	**43.0** (+7.2%)
NLLB	0.6B	36.6 (−9.0%)	39.1 (−9.1%)
	1.3B	38.3 (−4.7%)	40.7 (−5.3%)
	3.3B	39.0 (−3.0%)	40.6 (−5.6%)

Simultaneous Speech Translation

- Streaming-ready cascade-based architecture for speech translation



- Challenges:
  - ASR output may contain transcription errors
  - MT system starts translating before full sentence is available
  - Latency is bounded to keep pace with image video

Evaluation in simultaneous ST

- Trade-off between translation quality and latency
- Translation quality measured with BLEU
- Two alternative ways to measure system latency:
  - Average Lagging: Number of words the translation is behind
  - Translation Lag: Time elapsed between utterance and translation
- Out-of-domain evaluation datasets:
  - Europarl-ST: European Parliament debates
  - MuST-C: TED talks

Simultaneous ST systems

- Transformer-based architecture
- Adaptation for real-time streaming scenario
- Prefix-based training simulating limited access to future words
  - Conventional training:

Source	ATLAS récompense ses meilleures thèses
Target	ATLAS celebrates its Thesis Award winners

  - Prefix training:

Source	ATLAS récompense ses meilleures
Target	ATLAS celebrates its Thesis

Simultaneous ST systems

- History-aware model to exploit previous context

Source	... ses meilleures thèses six jeunes scientifiques
Target	... its Thesis Award winners six young

- Memory mechanism to avoid desynchronization

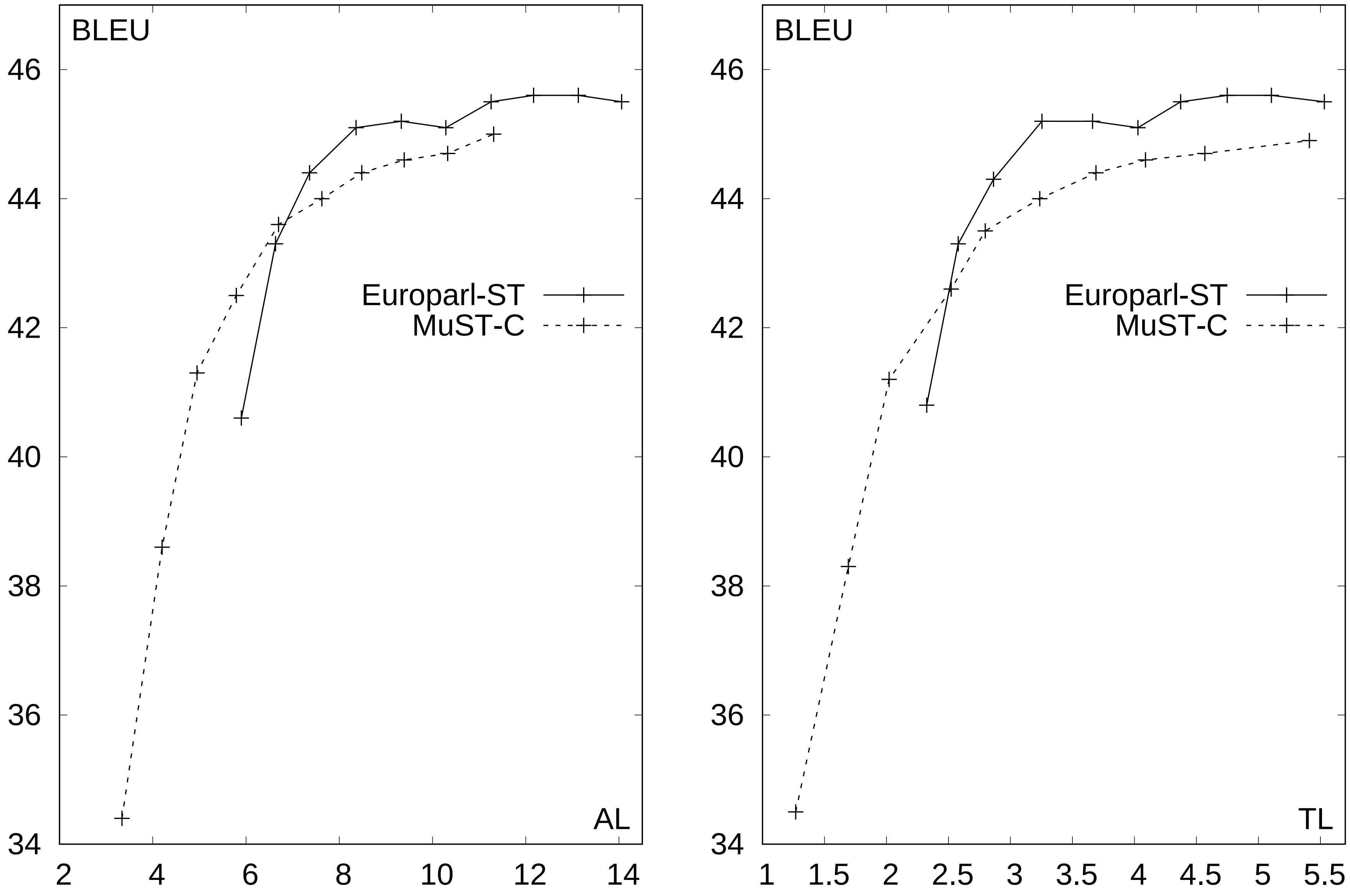
Source	... ses meilleures thèses	Six jeunes scientifiques
Target	... its Thesis Award winners [SEP]	Six young

- Adjustable latency in terms of words behind the input sentence

- This simultaneous ST system was recently published![1]

[1] J. Iranzo et al. Segmentation-Free Streaming Machine Translation. In Transactions of ACL, 2024.

Simultaneous ST results



Conclusions

- Offline and live automatic subtitles
- State-of-the-art in-domain ASR and MT systems deployed
- Close collaboration with CERN IT for on-premises deployment
- More than 30K hours already transcribed and translated
- LHCP-ASR dataset upcoming public release, including:
  - LHCP 2020 and 2022 talks: 235 hours (30h manually transcribed)
  - Papers, thesis and news: 1.5G words
- Auto-training for continuous ASR system upgrades