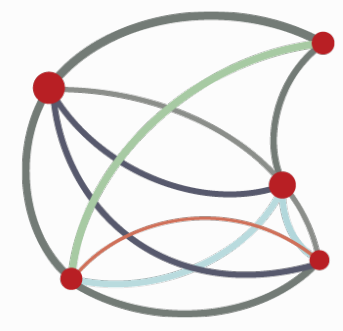


DIFFERENT CONTRIBUTIONS TO COST-EFFECTIVE TRANSCRIPTION AND TRANSLATION OF VIDEO LECTURES

Joan Albert Silvestre-Cerdà, Alfons Juan and Jorge Civera

{jsilvestre, ajuan, jcivera}@dsic.upv.es



MLLP | Machine Learning and Language Processing



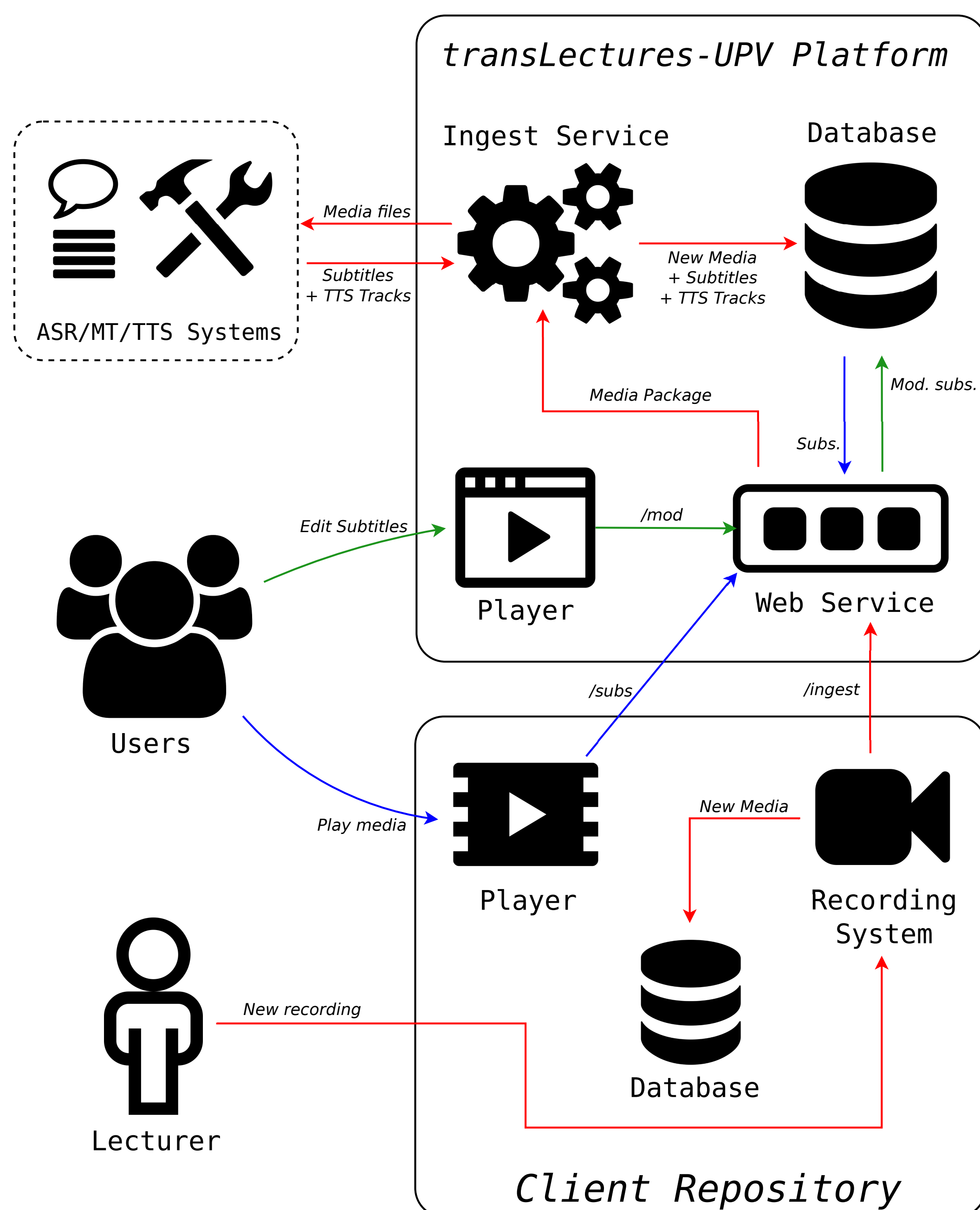
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

SCIENTIFIC AND TECHNOLOGICAL GOALS

- To design a system architecture for ASR and SMT integration.
- To propose an efficient audio segmentation approach.
- To evaluate and improve adaptation techniques for ASR and SMT.
- To propose an approach to explicit length modelling for SMT.
- To propose recommender systems using speech transcriptions.
- To assess these contributions in real-life scenarios.

THE TRANSLLECTURES-UPV PLATFORM

- **System architecture to integrate ASR and MT** technologies.
- Collaborative framework to review automatic subtitles.



- TLP is the **open source toolkit** implementing the architecture.
- Adopted in the EU projects **transLectures** and **EMMA**.
- Focus on the integration with the UPV's poliMedia video repository.
- **Currently working on production at the UPV:**
 - * 20K recordings, 4K hours, up to 40.000 users.
 - * <http://media.upv.es>

EFFICIENT AUDIO SEGMENTATION FOR SPEECH DETECTION

- The temporal cost of ASR depends on the input length.
- A prior segmentation can provide a better transcription quality.
- A **fast GMM-HMM Audio Segmentation** system is proposed.
- Excellent performance at near real-time speeds.
- **2nd place in the Albayzin 2012 Evaluations.**

ACKNOWLEDGMENTS

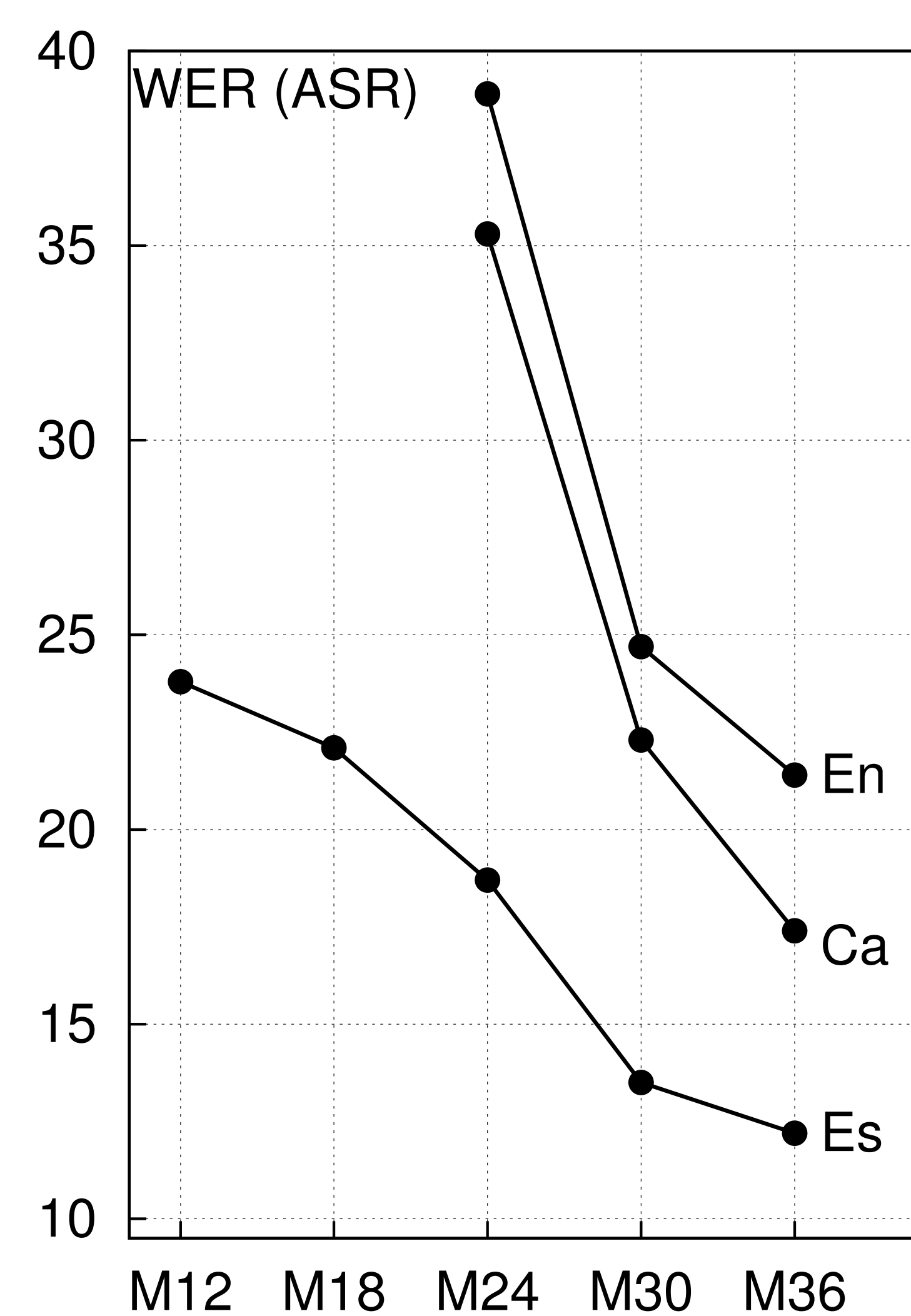
Work supported by the Spanish Government under the FPU scholarship (AP2010-4349), and under the iTrans2 (TIN2009-14511), erudito.com (TSI-020110-2009-439) and Active2Trans (TIN2012-31723) research projects. Also supported by the EC (FEDER/FSE) under the transLectures (FP7-ICT-2011-7-287755) and EMMA (ICT-PSP/2007-2013-621030) projects, and by the Spanish (MINECO/FEDER) research project MORE (TIN2015-68326-R).

LM ADAPTATION USING EXTERNAL RESOURCES FOR ASR

- ASR performance can be greatly improved using in-domain data.
- **LM adaptation by document retrieval.**
- PDF files retrieved from search queries based on lecture titles.
- **WER improvements above 10%** over solid baselines.

PLATFORM INTEGRATION AND ASSESSMENT ON POLIMEDIA

- Integration of Spanish, English and Catalan ASR systems.
- All lectures were re-transcribed (and re-translated) every 6 months.
- **Evolution of transcription quality (WER) over time:**



- **User evaluations shown savings up to 2/3 of user effort.**

EXPLICIT LENGTH MODELLING FOR SMT

- Comparative study on phrase length modelling for SMT.
- **Two novel length models for phrase-based SMT** are presented.
- Statistically significant improvements on all language pairs.
- **Work published in the Pattern Recognition journal.**

RECOMMENDER SYSTEMS FOR EDUCATIONAL PLATFORMS

- The **use of speech transcripts for recommendation** was studied.
- Deployment of a RS over the VideoLectures.NET website.
- Results after a 6-month evaluation were not conclusive.
- **Currently running on production in the VideoLectures.NET site.**

ACHIEVEMENTS AND CONTRIBUTIONS

- The transLectures-UPV Platform (TLP) for ASR & MT integration.
- A simple audio segmentation approach for fast speech detection.
- A new document retrieval technique for LM adaptation.
- An explicit phrase-length modelling approach for SMT.
- **9 publications** in international journals and conferences.
- **Technology transfers: UPV, UC3M, VL.NET, UNED, TTP:**

<http://ttp.mllp.upv.es>