Down with Sound
The Story of Silent Speech

Professor Bruce Denby
Université Pierre et Marie Curie
Sigma Lab, ESPCI-ParisTech
Paris, France

Speech Communication

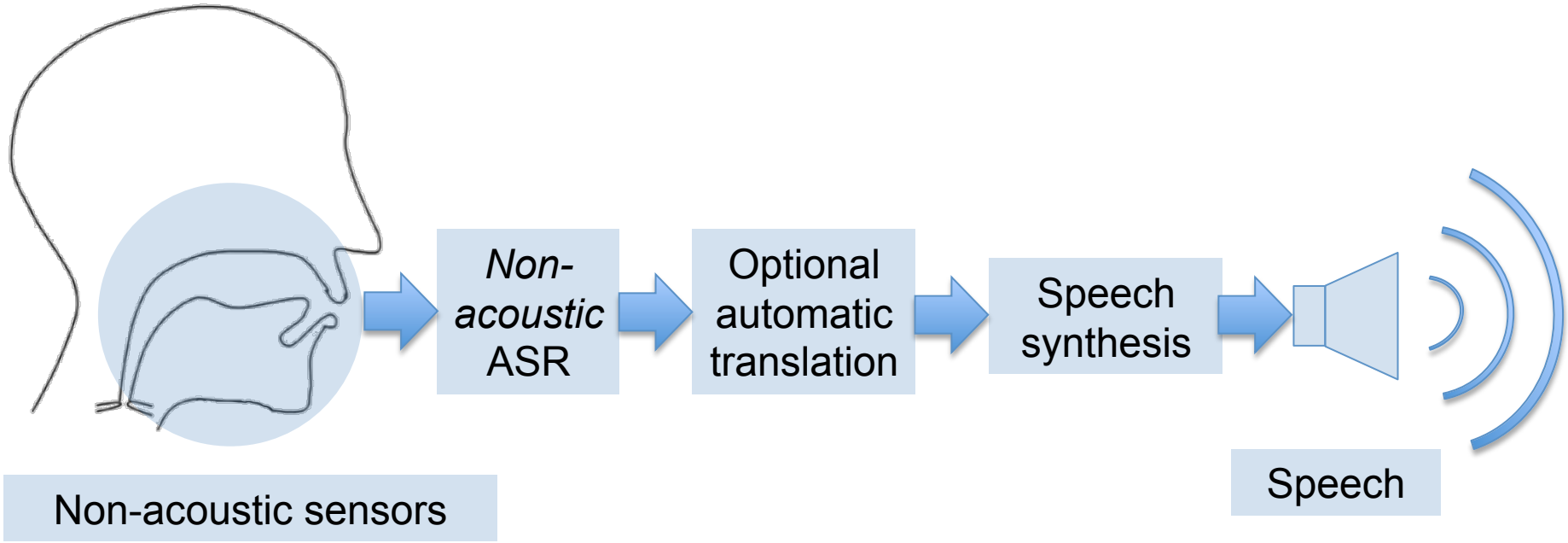
- Speech has always been the most natural and spontaneous modality for communication between human beings (and machines…?)

But speech has certain problems

- Noise sensitivity
  - ASR performance degrades very rapidly in the presence of background noise
- Airborn nature of speech signal
  - Interference with other communications
  - Security problems
- Inaccessibility for certain populations
  - Laryngectomy
  - Paralysis, pulmonary insufficiency, etc.
- Language dependence…

The *Silent Speech Interface* (SSI) Idea

- Use non-acoustic sensors to augment or completely replace the acoustic speech signal



SSI Application Areas

- In medicine
  - Give back original voice to voice-handicapped persons: laryngectomy or other pathologies
- In telecommunications
  - Telephone securely without disturbing others
  - Silent Man-Machine Interface (data entry, etc.)
  - Communicate verbally even in very noisy environments
  - Automatic translation

SSI – *"a hot new area"*

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 52 (2010) iii

SPEECH COMMUNICATION

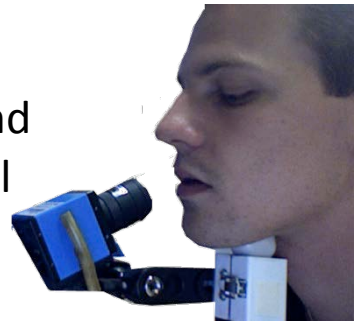
www.elsevier.com/locate/specom

Contents

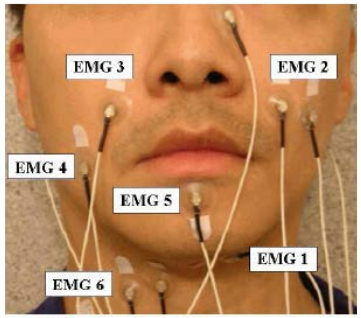
Special Issue: "Silent Speech" Interfaces
Guest Editors: B. Denby, T. Schultz and K. Honda

Variety of Approaches

Ultrasound + Optical

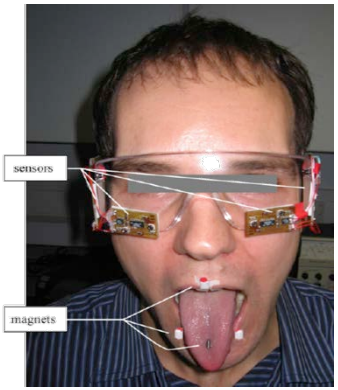


EMG



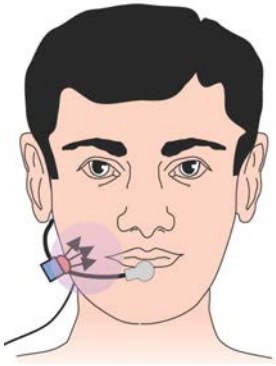
NAM

EMA

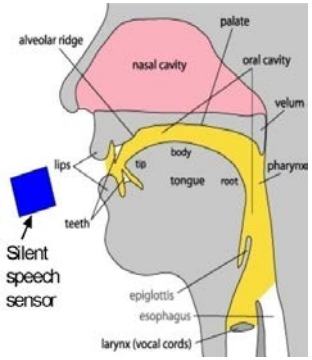


Vibration Sensors

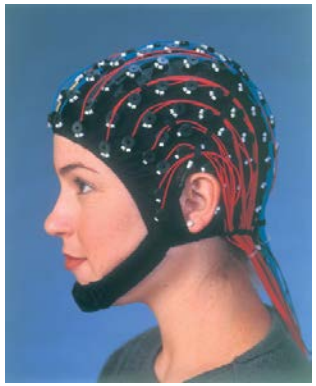
Radar



Low Freq. Ultrasound In Air



BCI EEG

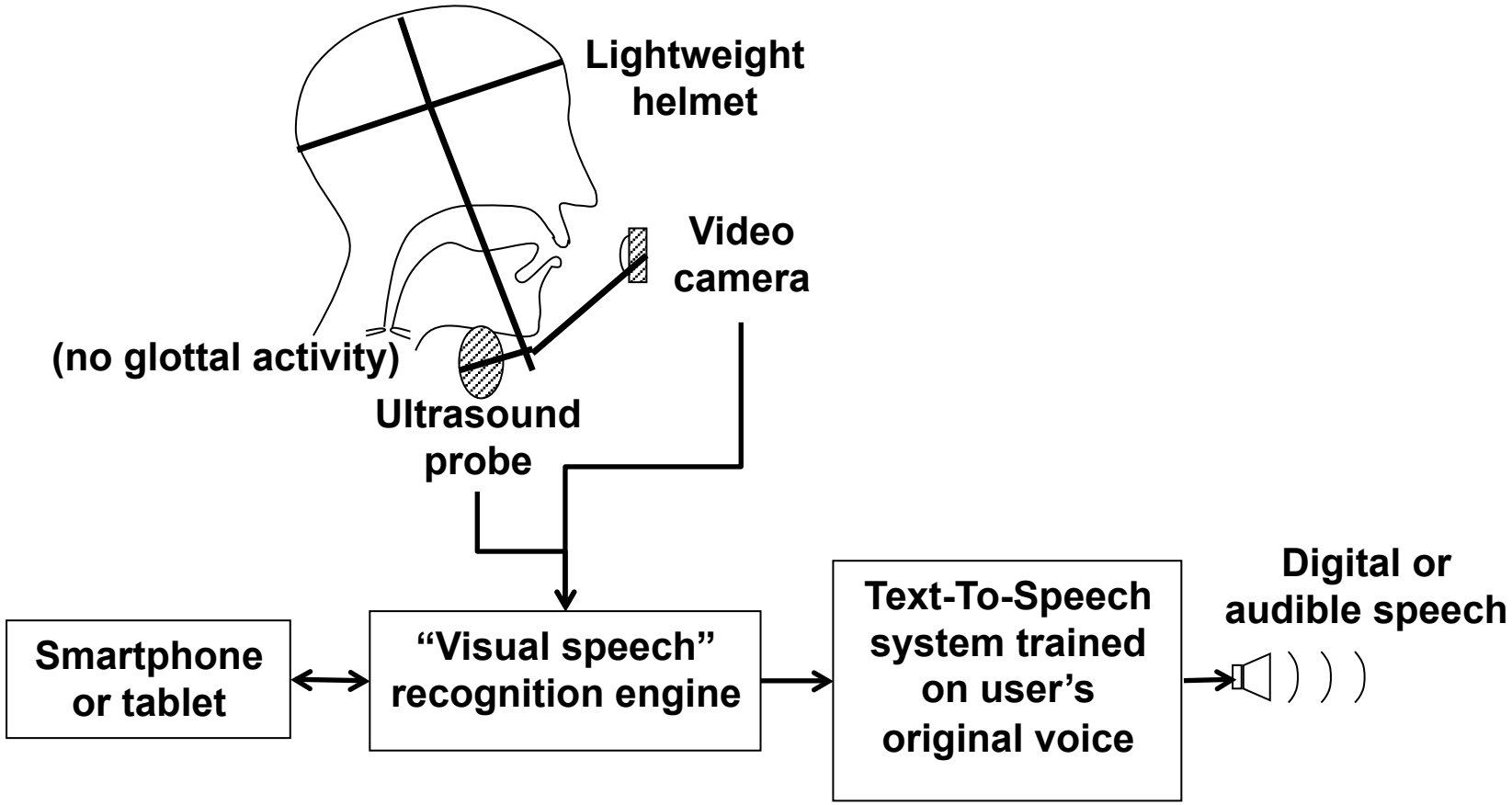


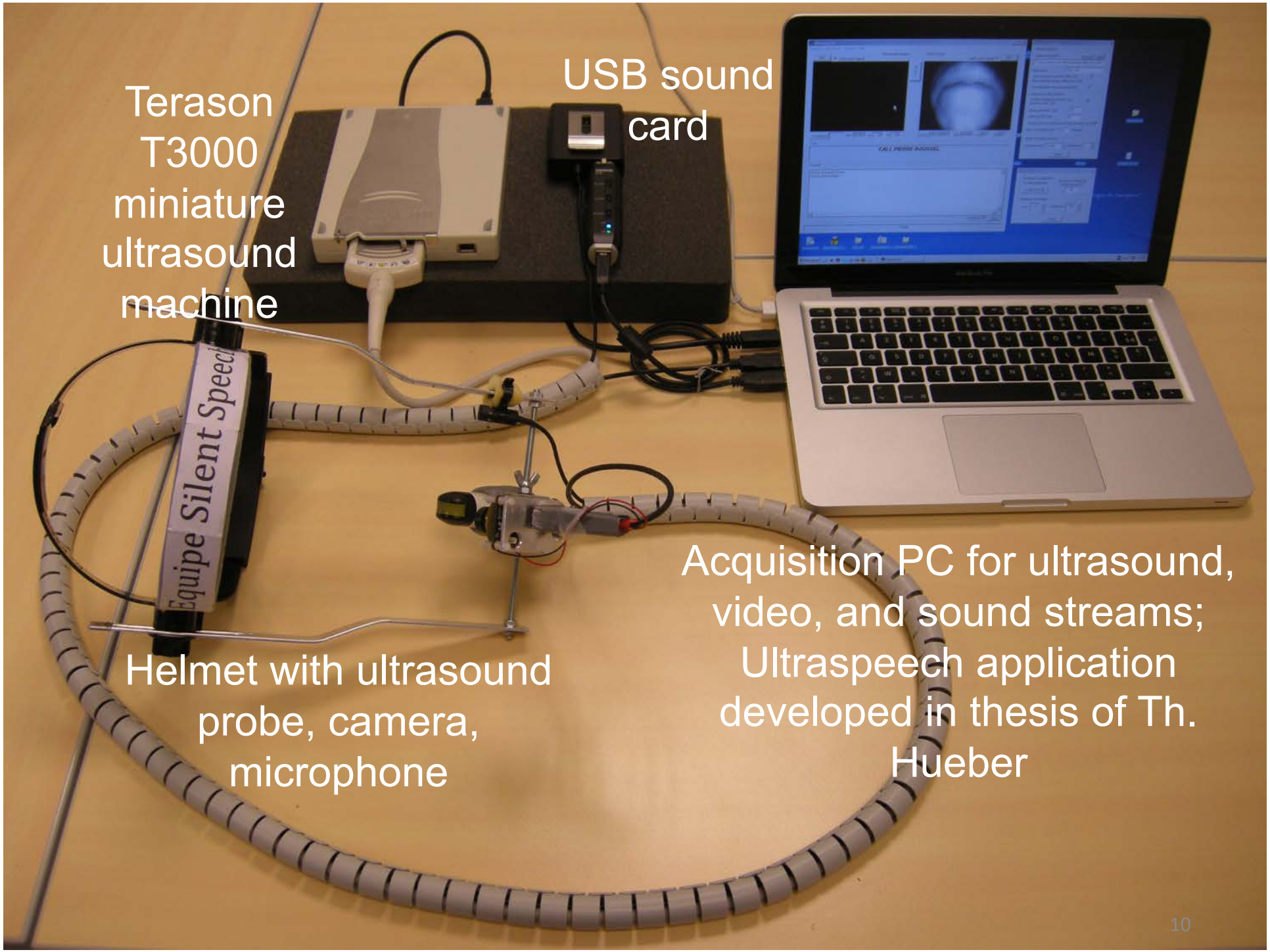
BCI cortical

Ultrasound data acquisition

- a) Classic cart-based multi-probe machine
- b) Siemens Acuson P10 ultra-miniaturised
- c) Terason solution – support PC/Windows
- Price: 20 000 € (small, 30 Hz) to 200 000 € (standard, high speed and/or 3D)

SIGMA Lab SSI: Ouisper Overview

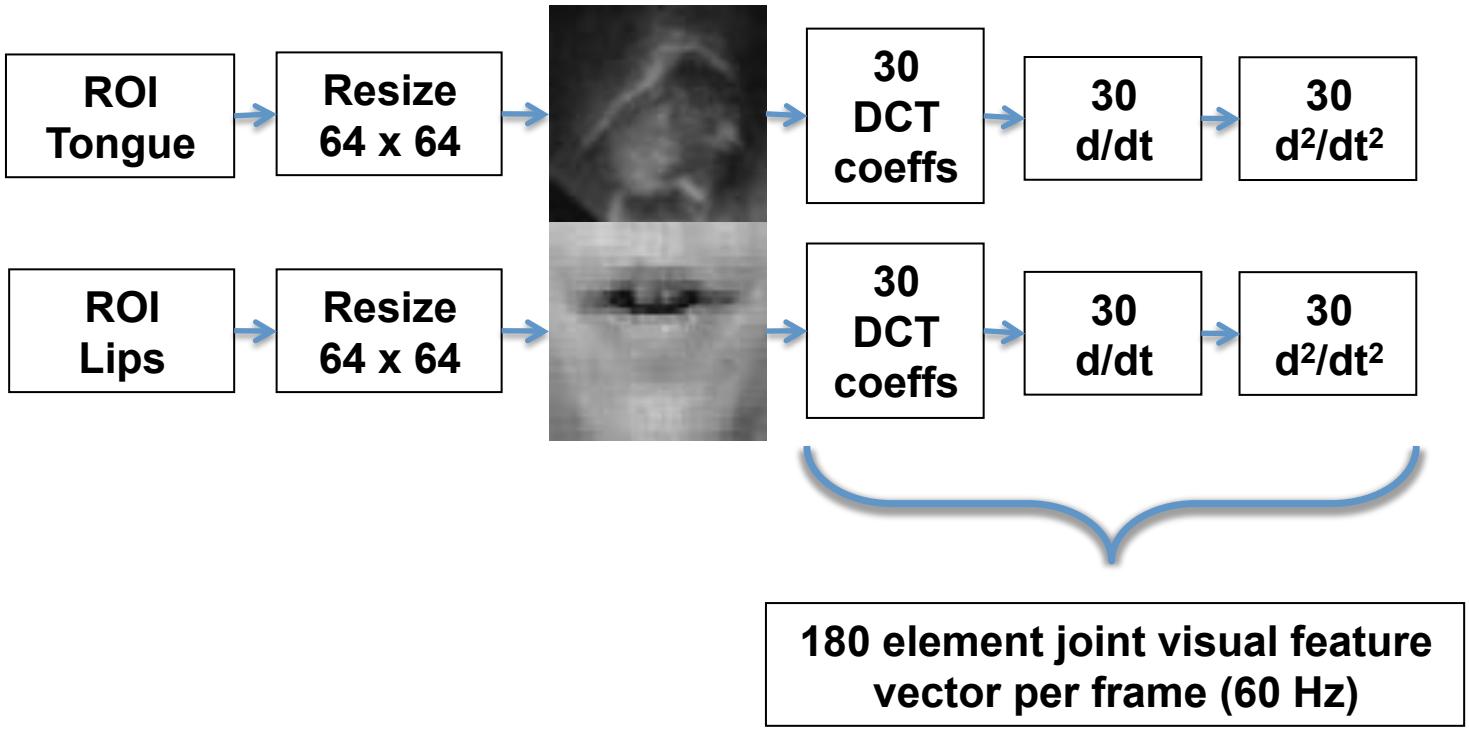



Terason T3000 miniature ultrasound machine
USB sound card
Equipe Silent Speech
Helmet with ultrasound probe, camera, microphone
Acquisition PC for ultrasound, video, and sound streams; Ultraspeech application developed in thesis of Th. Hueber

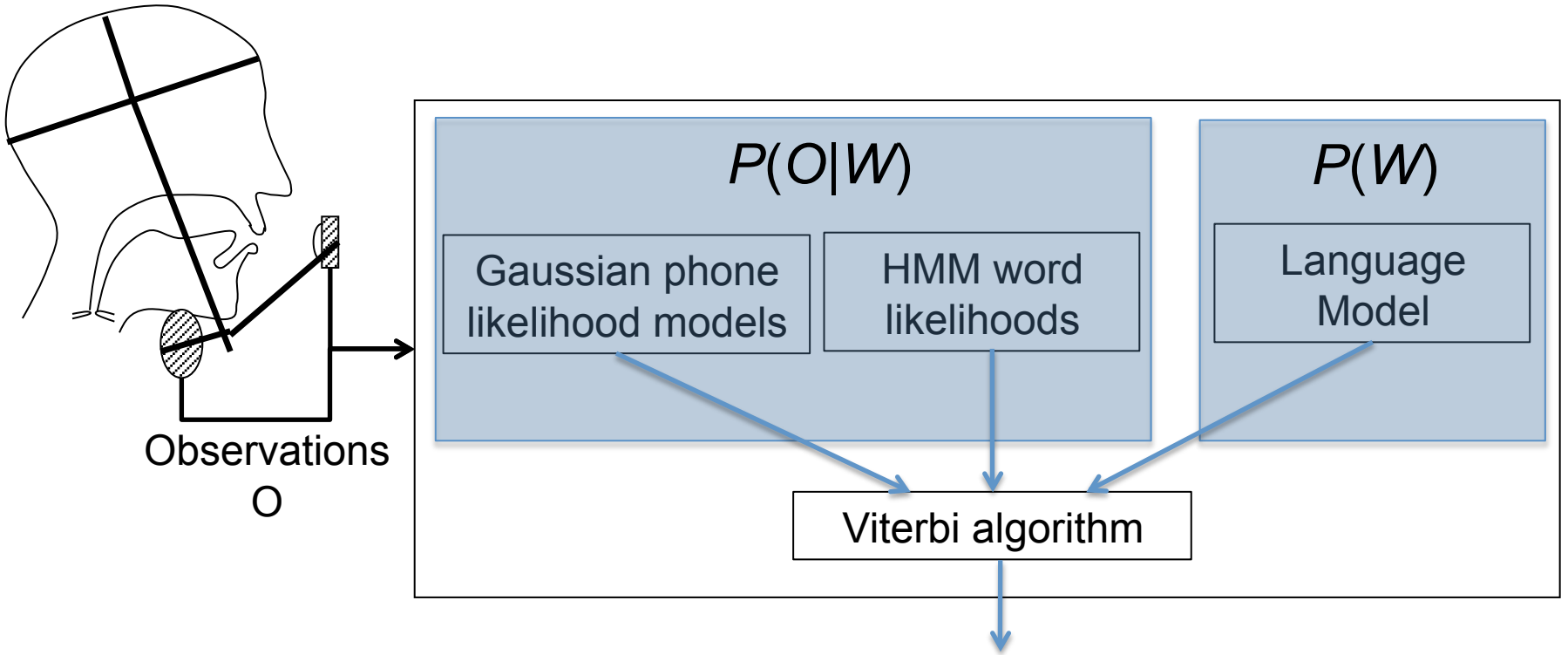
Example: laryngectomy voice recovery with phrasebook (French)

Example of captured data (English)

Visual Feature Extraction: selected method for Ouisper SSI



Continuous Non-Acoustic Speech Recognition: Overall Decoding Strategy



Estimated sequence:

$$\hat{W} = \underset{W \in dictionary}{\arg\max}\, P(O \mid W) P(W)$$

Recognition Results for English

- DARPA TIMIT corpus used for training triphone HMMs
- 3000 sentences, 2 months of recording
- Accuracy = (N – D – S – I) / N

Test Set + Language Model	Recognition Accuracy (%)	
	Word Level	Phone Level
Visual WSJ0 5k + WSJ0 5k NVP bigram	84	93
Visual Gigaword 20k + Gigaword 20k bigram	77	90
Visual Gigaword 20k + Gigaword 20k trigram	86	94

Presented at Interspeech, Sept. 2011, Florence, Italy

Real Time Recognition with Julius

- HMMs implemented with HTK toolkit
- This is slow
- Julius HMM code is available
- Developed in Japan
- Uses 2-path search
  - Initial search with reduced number of Gaussians in GMM
  - Full search on reduced set of paths
- Allows faster than real time recognition even on visual speech

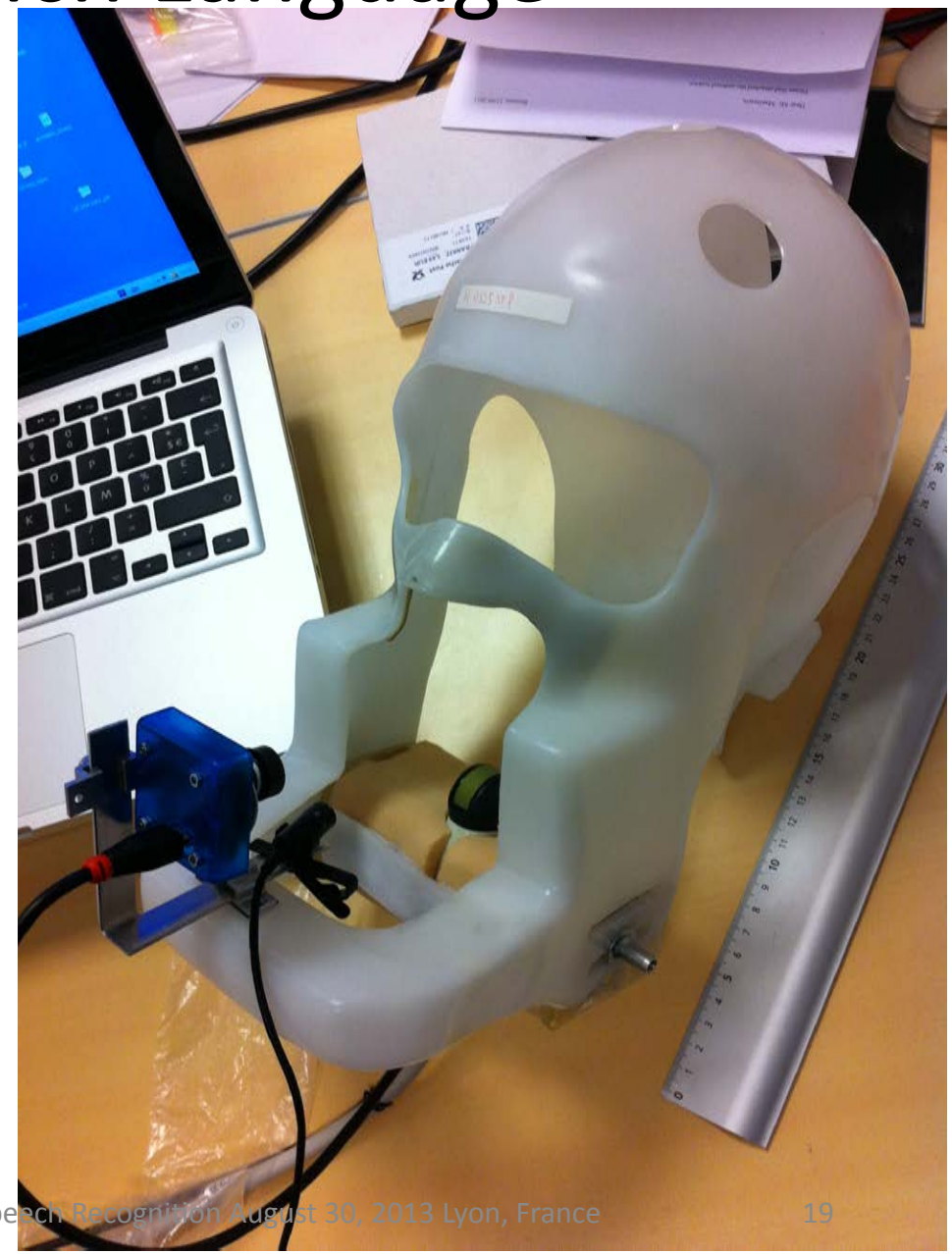
Synthesis Step: Chosen Method for Ouisper SSI

- Train Text-To-Speech (TTS) on speaker's voice
- Use OpenMary open source TTS tools
- Record 1500 triphone-rich sentences in sound cabin (microphone only)
- TTS resides on a server accessible via Internet
- Laryngectomy case:
  - Ideally pre-op voice if good quality recordings exist
  - If not, we have found a voice imitator is a good solution

Demo at ISSP, June 2011, Montréal

Move to French Language

- Individually thermoformed plastic helmets
- French is hard!
  - Corpora, LM, tools less available
  - Liaisons, variants, nasality, etc.
  - New synthesis module!
- Recording 3 speakers one of whom is post laryngectomy



Recognition Results for French

Table 2. *Word and phone level recognition accuracy in % for the 3 speakers and 2 LMs, with 95% c.l. intervals.*

Triphone enriched Polyvar training set and Polyvar Language Model (train set excluded)

Speaker	Recognition Accuracy in %			
	Word Level		**Phone Level**	
	Word Loop bigram	Polyvar bigram	Word Loop bigram	Polyvar bigram
PRR	10.1±2.4	**64.0**±3.8	62.1±2.0	**77.9**±1.7
LCB	12.4±3.2	**65.7**±4.5	65.0±2.4	**78.7**±2.1
JCD*	29.0±4.3	**65.4**±4.5	71.3±2.3	**74.6**±2.2

*post-laryngectomy speaker

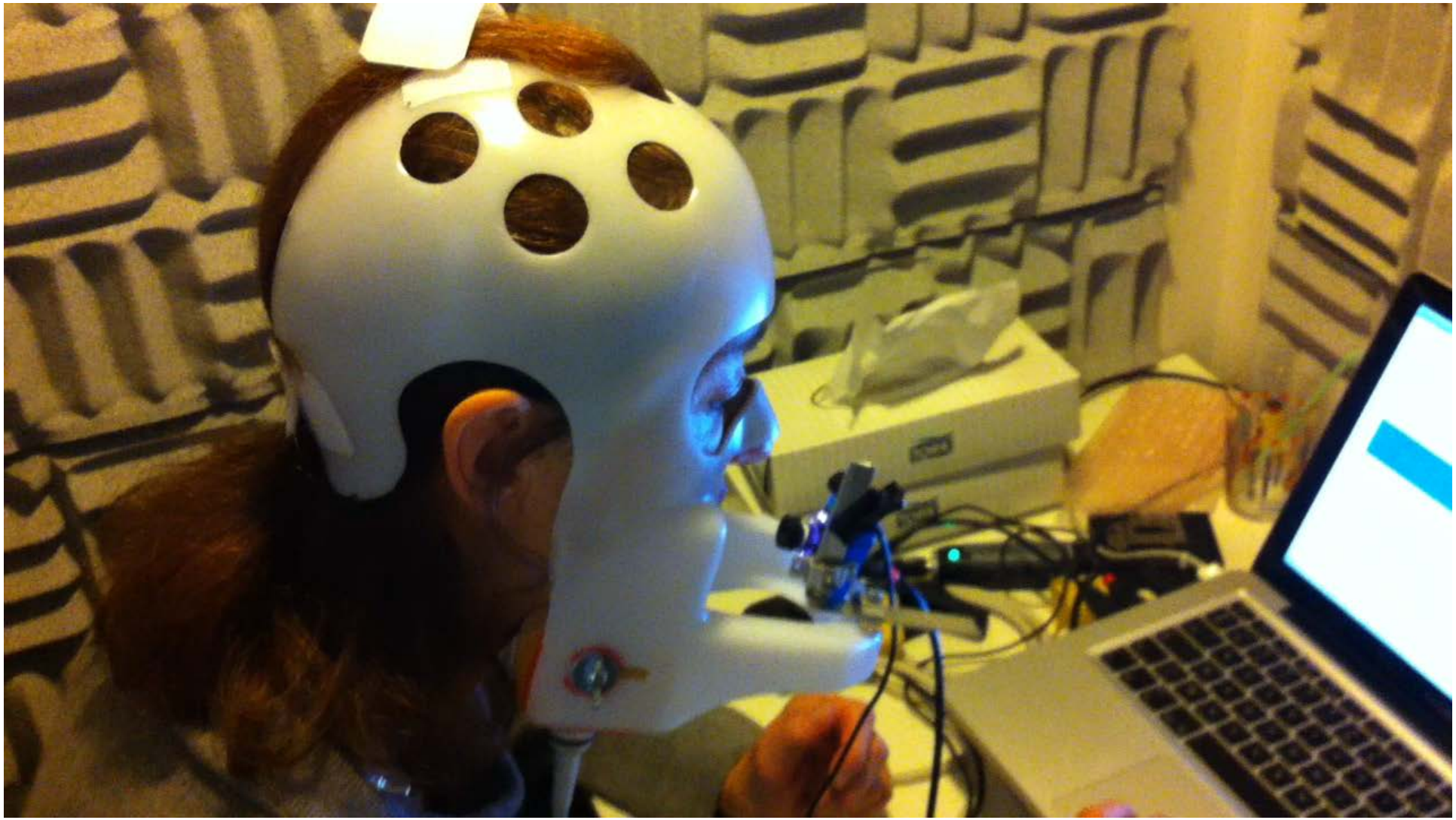
Presented at I2MTC, Minneapolis, USA, May 2013

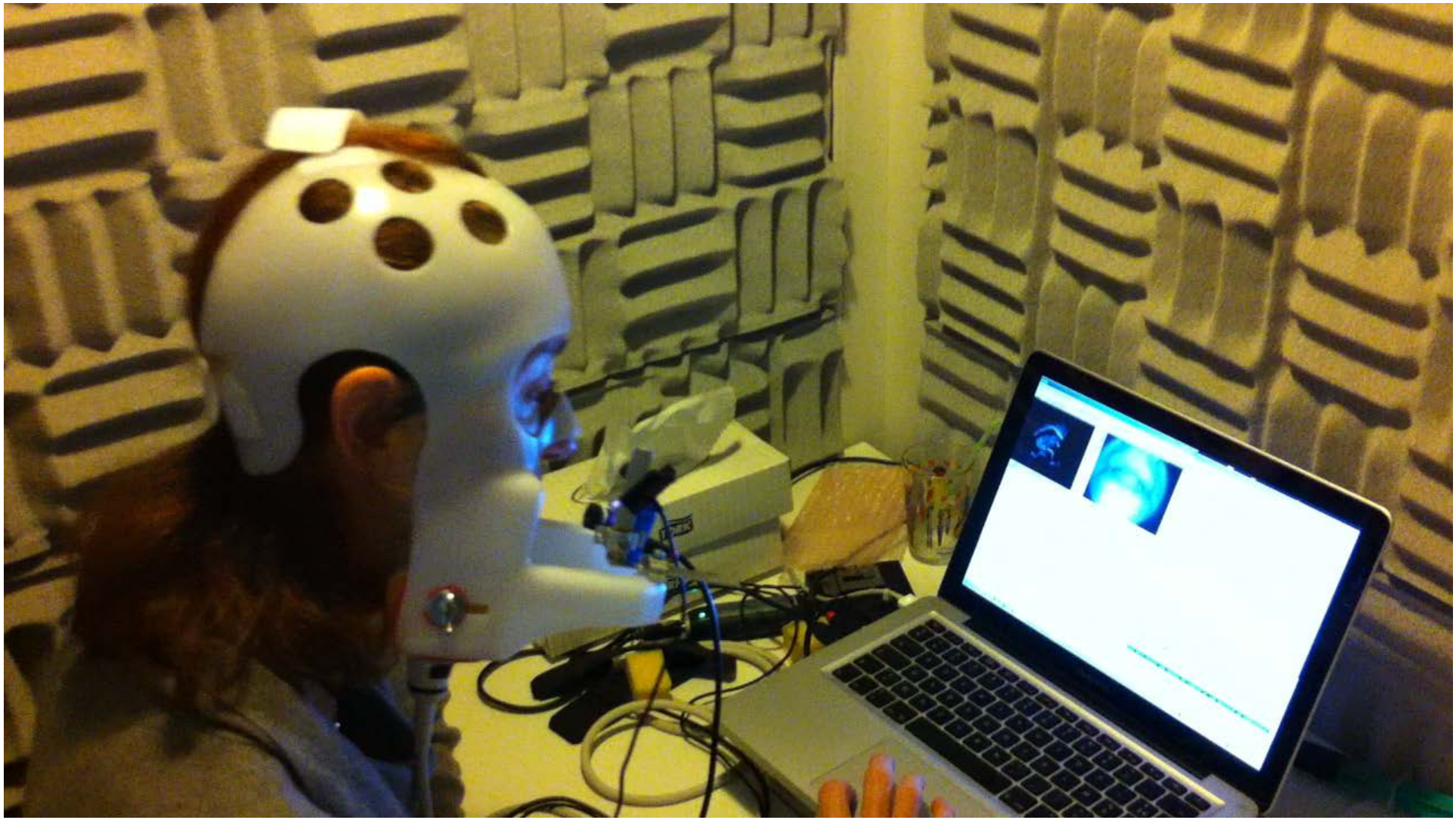
SIGMA

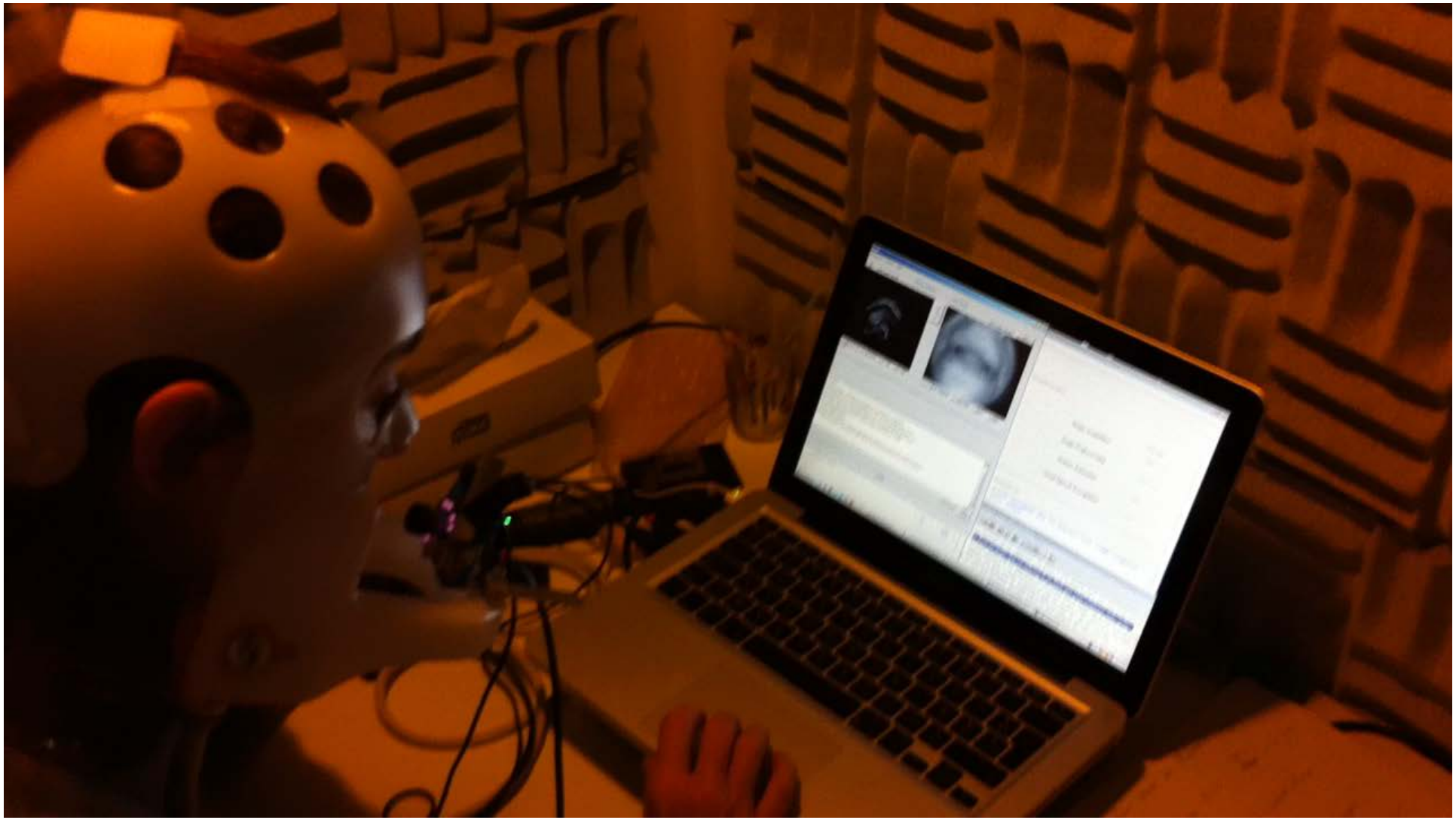
UPMC SORBONNE UNIVERSITÉS

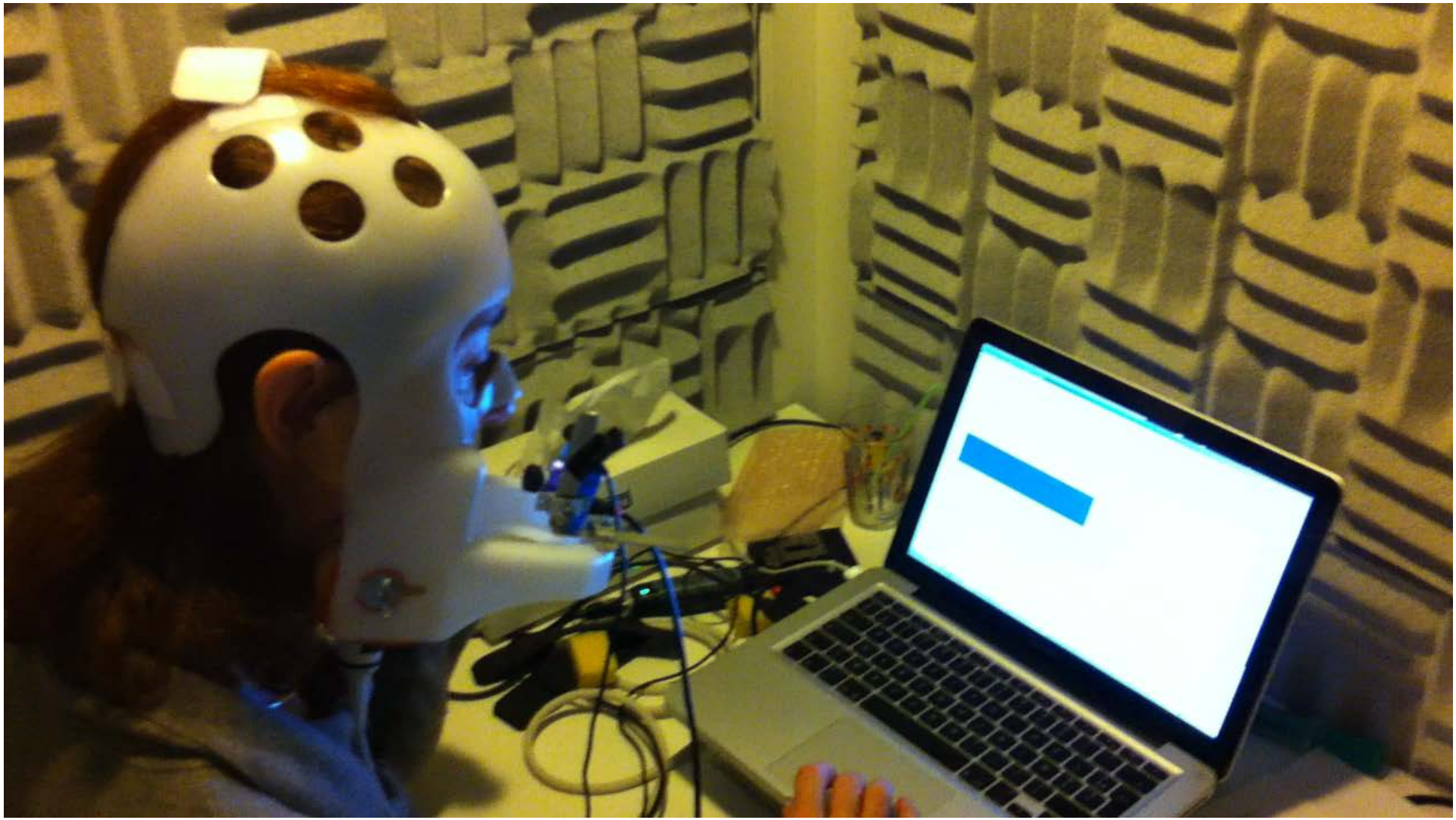
Real Time Tests in French
(text output only, no synthesis)

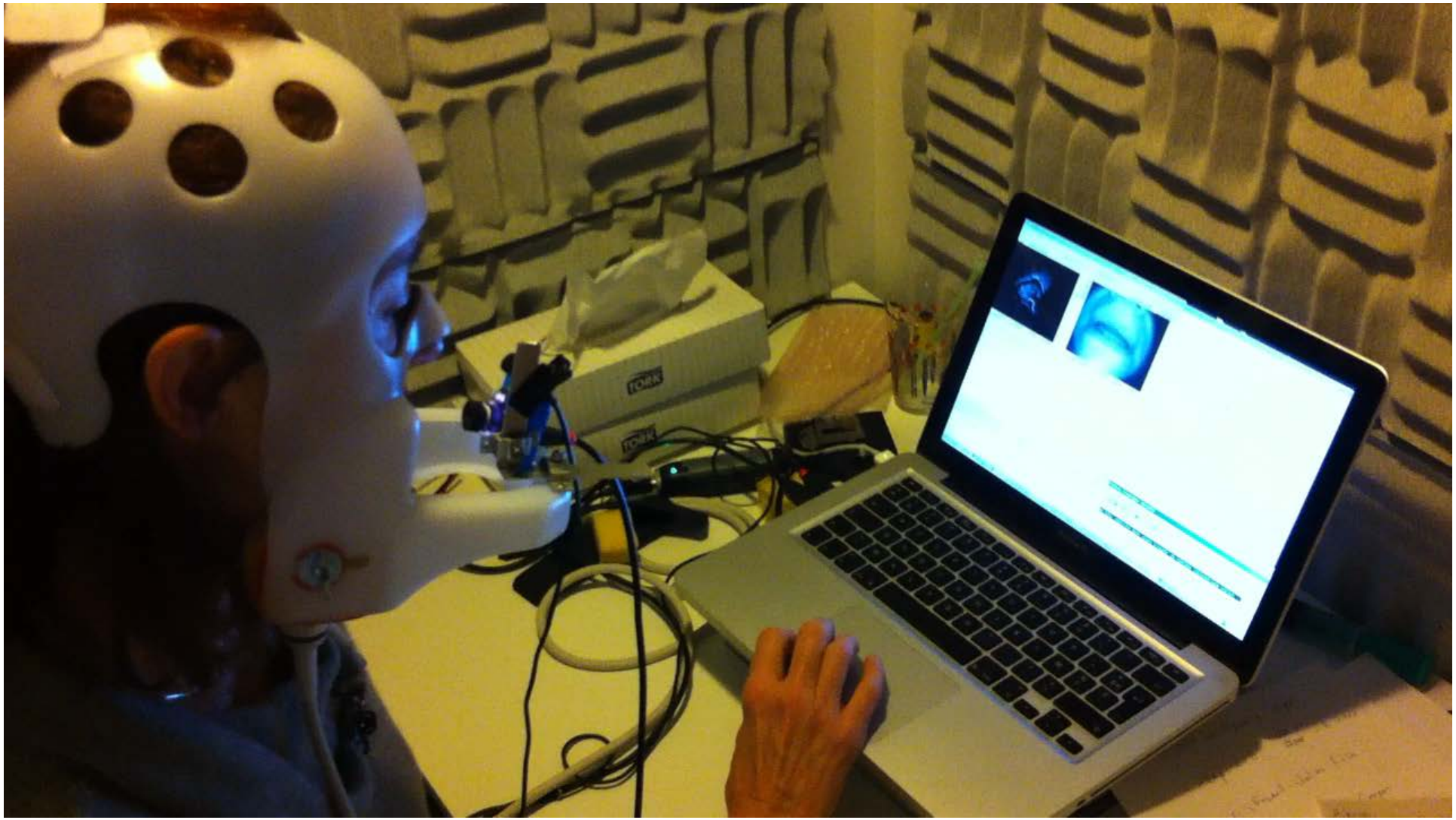

TORK
TORK











SIGMA

UPMC SORBONNE UNIVERSITÉS

ANNOUNCEMENT

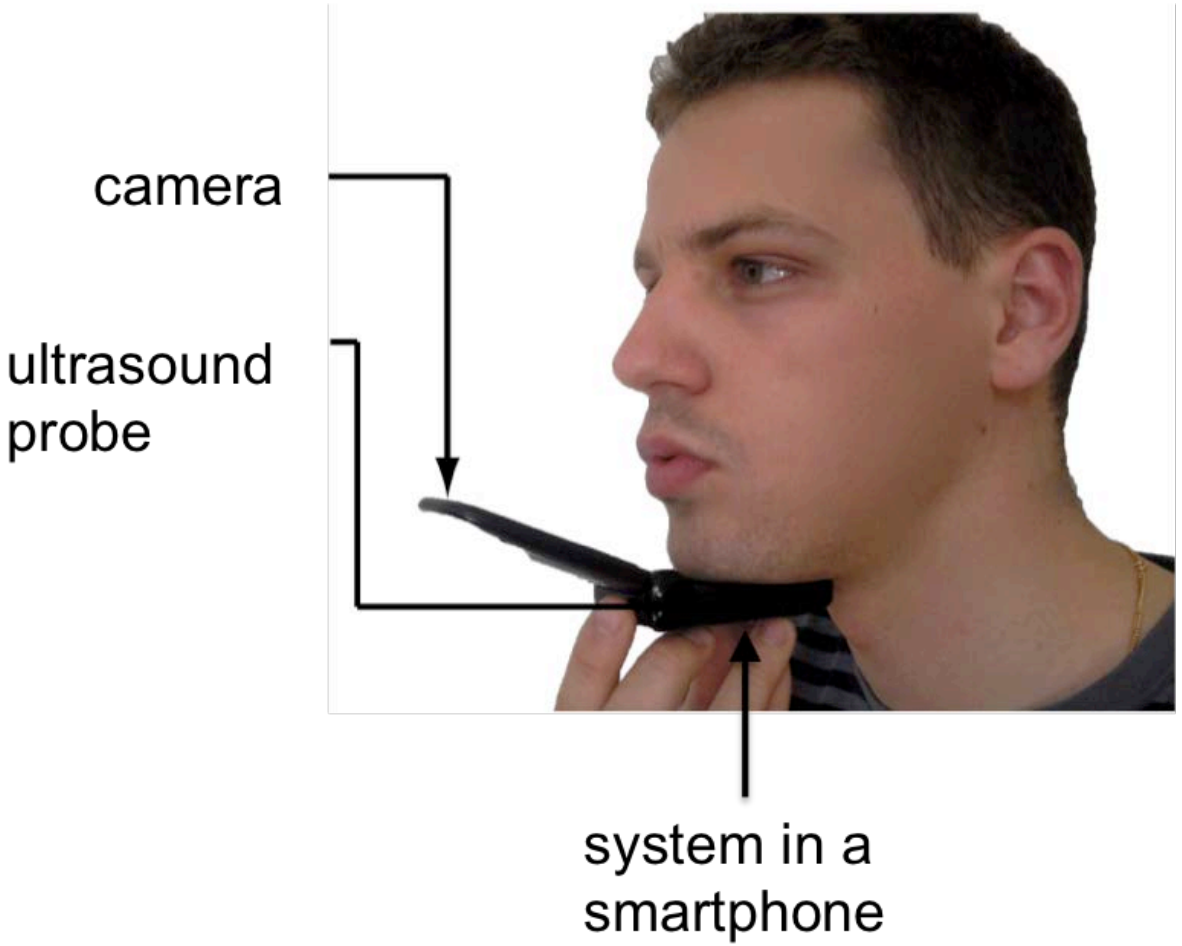
Silent Speech Challenge

- English multimodal silent speech data are available online for others to try!!
- Includes raw single speaker ultrasound and lip images plus text of sentences pronounced (no sound: silent speech!)
- Hundreds of Gbytes of data!
- Includes TIMIT training/development corpus plus WSJ0 test corpus (100 sentences) and language model.
- Benchmark: 93% phone recognition, 84% word recognition on WSJ0 set
- Available by anonymous ftp at ftp.espci.fr (login anonymous, password your email address)
- Go to pub, sigma, then TIMIT_Training or WSJ05K_Test, then to Ultraspeech Capture

More Recently

- Specially manufactured miniature US probe

Perspective



SIGMA

UPMC SORBONNE UNIVERSITÉS

SPINOFF

i-Treasures Project

- FP7 IP project
- 12 Partners
  - CERTH, UOM, AUTH (Greece)
  - UPMC, ARMINES/ENSMP, LPP/CNRS (France)
  - UCL (UK)
  - CNR (Italy)
  - UMONS, Acapela (Belgium)
  - Turkish Telecom
  - University of Maryland (observer)
- Propose novel multisensory technologies to safeguard and transmit ICH (Intangible Cultural Heritage)

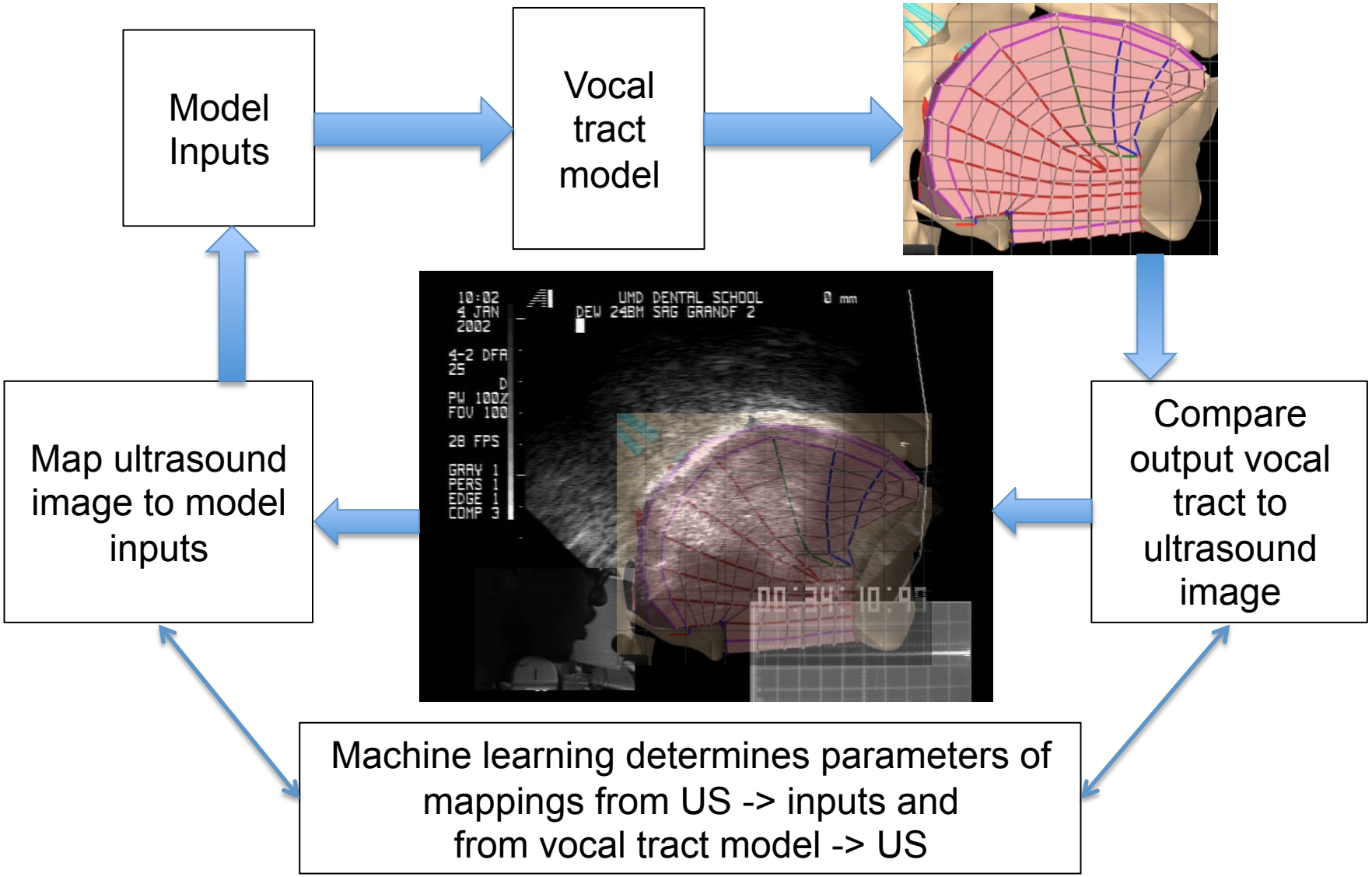
i-Treasures

- Four use cases in different ICH domains:
  - Rare traditional songs
  - Rare dance interactions
  - Traditional craftsmanship
  - Contemporary music composition
- Capture and analyze Intangible Cultural Heritage:
  - Facial expression analysis
  - Body and gesture recognition
  - Vocal tract modeling
  - Sound processing
  - Electroencephalography

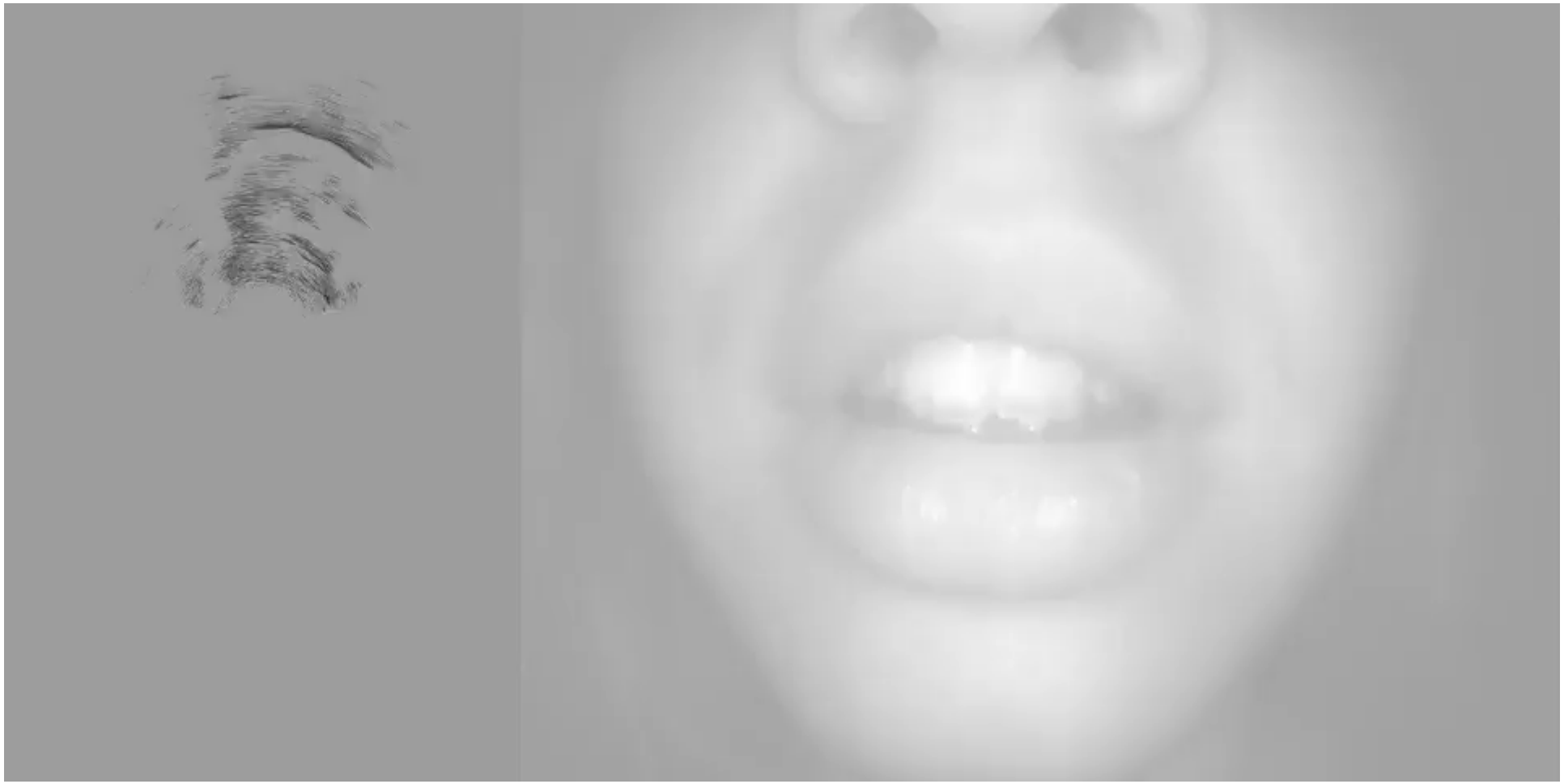
Rare traditional song use case

- Capture and analyze vocal tract of expert singers and use to create an independent 'avatar' of the singer
- Student displays expert avatar and tries to match his own to it in real time, in order to learn the technique
- Targeted rare singing techniques:
  - Corsican cantu in paghjella
  - Sardinian canto a tenore
  - Byzantine singing
  - Human beat box

Possible Strategy



First recording of singing data



SIGMA

UPMC SORBONNE UNIVERSITÉS

The End
Thank you!