MUSCLE WP5 Showcase:

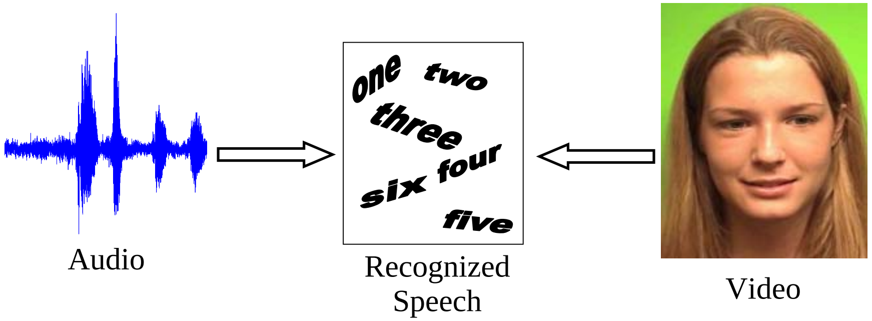
Real-Time Audio-Visual Automatic Speech Recognition Demonstrator

TSI-TUC (leader)
ICCS-NTUA
INRIA-TEXMEX

Groups and Researchers Involved

- TSI-TUC
  - A. Potamianos (showcase leader)
  - M. Perakakis
  - E. Sanchez-Soto
- ICCS-NTUA
  - P. Maragos (group leader)
  - G. Papandreou (visual/fusion)
  - A. Katsamanis (audio/fusion)
  - V. Pitsikalis (audio/fusion)
- INRIA-TEXMEX:
  - P. Gros (group leader)
  - G. Gravier (fusion)

Audio-Visual Automatic Speech Recognition



- Audio-only Automatic Speech Recognition (ASR) degrades under noise
- Use video for lip-reading to boost ASR performance

Showcase Main Points

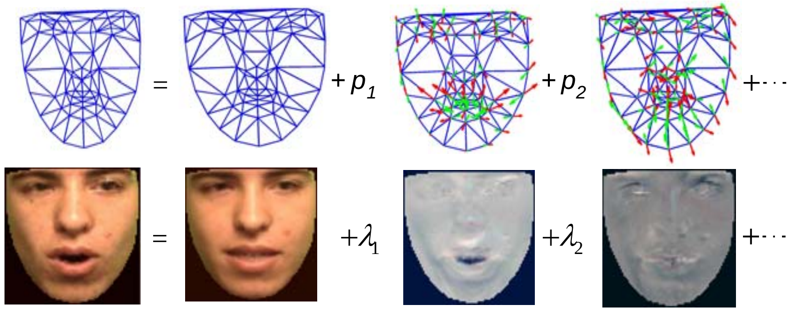
- Shortcomings of current AV-ASR systems
  - Research-level set-ups
  - videos shot under carefully controlled conditions
  - processing is performed off-line
- Goal: build a proof-of-concept *practically deployable* laptop-based AV-ASR prototype which:
  - uses low-end consumer microphone and camera to capture the speaker
  - performs visual/audio feature extraction, as well as speech recognition on the laptop in *real-time*
  - is robust to failures of a single modality, such as visual occlusion of the speaker's face

Tasks

- T1: Visual Front-end
  - ❑ Face detector (DONE)
  - ❑ Face tracking and feature extraction (DONE)
  - ❑ Optimization for real-time performance (IN PROGRESS)
- T2: Audio-Visual Recognition Model and Fusion
  - ❑ Advanced baseline audio front-end (DONE)
  - ❑ HMM-based recognition back-end (DONE)
  - ❑ Model training on audio-visual corpora (DONE)
  - ❑ Adaptive audio-visual fusion (IN PROGRESS)
- T3: System Integration
  - ❑ Laptop-based system (IN PROGRESS)
  - ❑ Usable for live AV-ASR demonstrations (IN PROGRESS)
- *Project duration:* December 2006 – June 2007

Visual Front-End

- **Analyze face expression and appearance**
- **Real-time feature extraction algorithms**
- **Excellent performance in AV-ASR experiments**



Feature Fusion

- Goal:
  - ❑ Adaptive fusion heterogeneous information streams
- Stream weights improve recognition performance
- Test alternative techniques for stream weight computation
  - ❑ Minimum classification error
  - ❑ Feature measurement uncertainty compensation
  - ❑ Previous work by all three partners
- Stream weight adaptation
  - ❑ Depending on auditory SNR
  - ❑ Either static or fully dynamic

Audio-Visual
Shocase

Visual Front-End

Audio-Only ASR Live Demo

- Real-Time continuous digits ASR
- Model Training on the WSJ database

Tasks

- T1: Visual Front-end
  - ❑ Face detector (DONE)
  - ❑ Face tracking and feature extraction (DONE)
  - ❑ Optimization for real-time performance (IN PROGRESS)
- T2: Audio-Visual Recognition Model and Fusion
  - ❑ Advanced baseline audio front-end (DONE)
  - ❑ HMM-based recognition back-end (DONE)
  - ❑ Model training on audio-visual corpora (DONE)
  - ❑ Adaptive audio-visual fusion (IN PROGRESS)
- T3: System Integration
  - ❑ Laptop-based system (IN PROGRESS)
  - ❑ Usable for live AV-ASR demonstrations (IN PROGRESS)
- *Project duration:* December 2006 – June 2007

Audio-Visual Speech Recognition Demo

AV

A