Audio-Visual Speech Recognition

Readings in Advanced Intellimedia (MM2)

About the papers

- "*Audio-Visual Automatic Speech Recognition: An Overview*" published in 2004
- "*Audio-Visual Speech Recognition*" workshop in 2000
- "*Nonlinear Manifold Learning for Visual Speech Recognition*" published in 1995

Summary

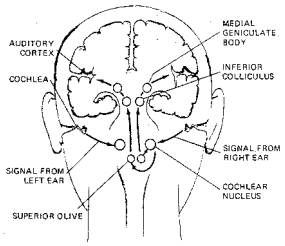
- Automatic Speech Recognition
- Audio-Visual Speech Recognition
  - Basis
  - Different methods
- Existing Databases & Software
  - OpenCV & AVCSR
- Applications
  - Home care workers' project

Automatic Speech Recognition

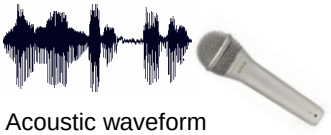
- ASR has been an active research for several decades
- What is the task?
  - Getting a computer to understand spoken language
  - By "understand" we might mean
    - React appropriately
    - Convert the input speech into another medium, e.g. text

ASR: How do humans do it?

Articulation produces sound waves which the ear conveys to the brain for processing

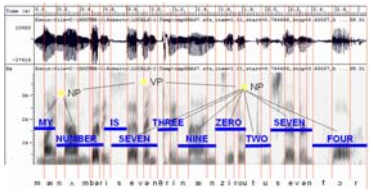


ASR: How might computers do it?



Acoustic waveform

Acoustic signal



Speech recognition

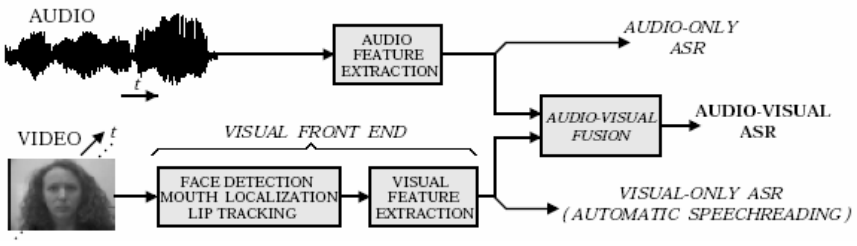
- Digitization
- Acoustic analysis of the speech signal
- Linguistic interpretation

ASR: limitations

- Performance is far from the one achieved by a human
- Limitations:
  - Ignore visual speech cues
  - Susceptible to acoustic noise
  - Examples: /m/ against /n/ and /b/ against /p/ …
- Research effort in ASR for noisy environments

AV Speech Recognition

- Human speech perception is bimodal

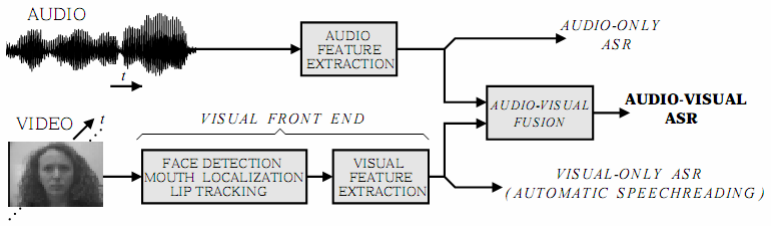


AV Speech Recognition

- Robust face detection
- Location estimation and tracking of the speaker's mouth or lips

- Two streams of features, one for each modality
- Outperform audio-only ASR

Audio-Visual SR: detection



- Face detection & Lip tracking
  - Determine a ROI (*region-of-interest*)
  - Techniques to locate ROIs
- Facial feature detectors are used

Visual features

- Appearance Based features
  - Principal components analysis
  - Discrete cosine, wavelet, and other image transforms
  - Linear discriminant analysis
- Shape Based features
  - Lip geometric features
  - Lip model features
- Appearance & Shape features

Audio-visual integration

- Grouped into 2 methods:
  - Feature fusion
  - Decision fusion
- The sequence of features can be modeled by a Hidden Markov Model (HMM): most widely used classifier
  - Markov chain – a feature is generated based on its current state, at each time step. The system transitions from one state to another.
  - We see feature, not states
  - The features corresponding to a particular state are similar. The model specifies the feature probability matrix.

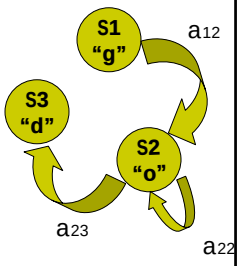


Figure: A Markov Chain with 3 states (labeled S1 to S3) with selected state transitions, such as "good".

Noise reduction

- Speech recognition in controlled situations has reached very high levels of performance
- Performance degrades in noisy situations

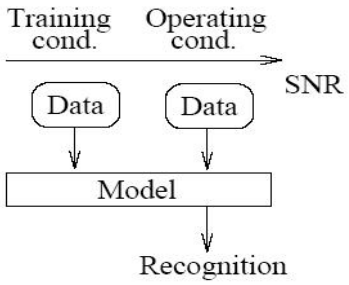
– 100% to 30% accuracy in a car (90km/h)

– 99% to 50% in a cafeteria

- A system trained with a given SNR performs worse in other SNR environments.

Noise reduction

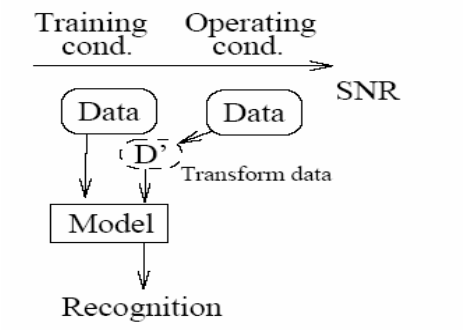
- 1. Noise Resistance



- Search for noise resistant features and robust distance measures

Noise reduction

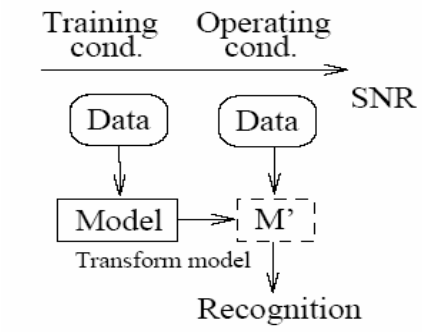
- 2. Speech Enhancement



- Remove noise from the signal

Noise reduction

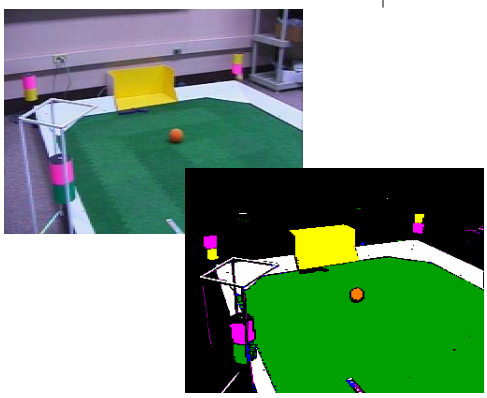
- 3. Model Compensation For Noise



- Transform speech model to accomodate noise

Softwares

- RoboRealm
- CMVision
- Gwyddion
- API (JAI, MegaWave2..)
- CMU Sphinx
- Nuance
- OpenCV



OpenCV: presentation

- Computer vision library originally developed by Intel
- Motivation: Due to the poor performance of conventional speech recognition system in noisy environment, OpenCV libraries can combine both audio and video features to achieve better speech recognition accuracy
- General ideas:
  - Video feature extraction
  - Audio feature extraction
  - Using models such as:
    - Hidden Markov Models (HMMs): Audio HMMs, Visual HMMS, Coupled HMMs
- Software: Visual C++, OpenCV image processing library, Intel AVCSR library…
- Hardware: Computer, Webcam, Microphone

Face_Detection_and_Tracking.flv

Eye_Tracking_in_real-time.flv

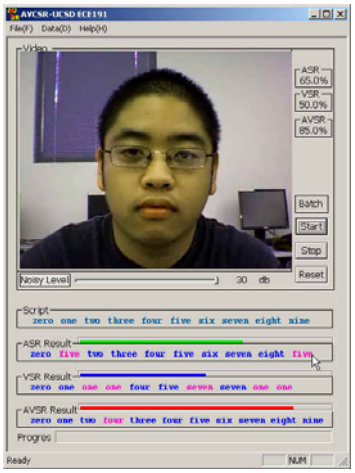
Examples on OpenCV

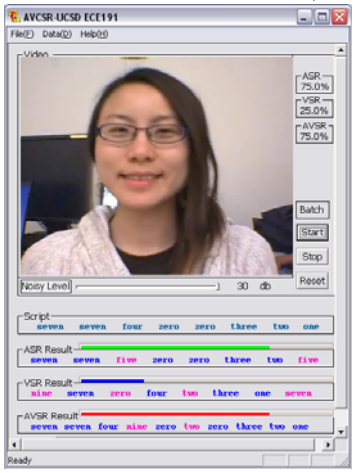
- Hough Lines
- Contours
- Squares

- Histograms
- Fit Ellipse
- Distance Transform

- Delaunay

OpenCV Results





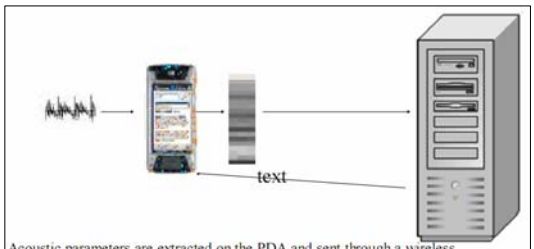
OpenCV Results

	SNR: 5dB			SNR: 30dB		
Sample Testing	**ASR**	**VSR**	**AVSR**	**ASR**	**VSR**	**AVSR**
video01	50.00%	50.00%	80.00%	65.00%	50.00%	85.00%
video02	80.00%	20.00%	100.00%	80.00%	20.00%	60.00%
video03	75.00%	25.00%	75.00%	75.00%	25.00%	75.00%

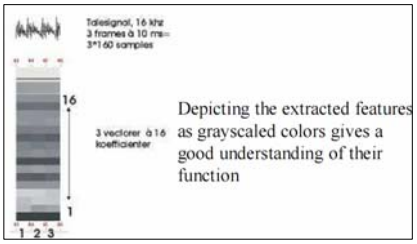
Applications

Home care workers' project

- Use a PDA to allow ASR
- Working with Distributed Speech Recognition (DSR)



Acoustic parameters are extracted on the PDA and sent through a wireless network to a remote server, where the rest of the recognition takes place



Depicting the extracted features as grayscaled colors gives a good understanding of their function

Home care workers' project

- Main problem:
  - Workers would use the system in noisy environment such as cars…
- Performance degrades in noisy situations:
  - 100% to 30% in a car (90km/h)
  - 99% to 50% in a cafeteria
- Possible surveys & discussions:
  - Remove noise from the signal to allow SR
  - Develop a AVCSR system

Thank you!

Any questions?