A Phone-Viseme Dynamic Bayesian Network for Audio-Visual Automatic Speech Recognition

Louis H. Terry    Aggelos K. Katsaggelos

Image and Video Processing Laboratory (IVPL)
Department of Electrical Engineering and Computer Science
Northwestern University

http://ivpl.eecs.northwestern.edu

Dec. 10, 2008
International Conference on Pattern Recognition

Outline

1 Introduction
   - Audio-Visual Automatic Speech Recognition
   - Research Focus: The Recognizer as a DBN

2 The Phone-Viseme Model
   - Model Description
   - Model Definition

3 Experimental Results
   - Setup
   - Results

4 Conclusion
   - Conclusion
Audio-Visual Automatic Speech Recognition

Components of an AV-ASR System

- Feature Extraction and Preprocessing
- Audio-Visual Fusion
- Recognition

![Diagram showing data flows for early, middle, and late integration as well as audio-only and video-only recognition](image-url)

**Figure:** General audio-visual automatic speech recognizer block diagram showing data flows for early, middle, and late integration as well as audio-only and video-only recognition

### Some advantages of dynamic Bayesian networks (DBNs) over hidden Markov models (HMMs):

- Ability to model the complete recognition system including:
  - Word grammar / Statistical relationships
  - Pronunciation variants
  - Model dependencies across word/sub-word/state levels
  - Internal variable switching allows for “dynamic" models
The Recognizer as a DBN

Long-Term Goal: Dynamically Adapting DBN for ASR

Build a DBN-based AV-ASR system able to:

- Dynamically change stream weights based on acoustic/visual stream reliability
- Adapt to linguistic phenomena such as Lombard speech and speech reduction
- Handle stream asynchrony across word/sub-word/state boundaries

Model Description

- Utterances consist of a sequence of words
  - Stream synchrony forced at word boundary
- Words consist of phones and visemes (sub-word units)
- Sub-Word units consist of states
- Audio observation model depends only on current audio state
- Video observation model depends only on current video state
- Observations exponentially weighted using stream weight $\in [0, 1]$. Audio weight + video weight sums to 1
  - Slightly different from paper. In paper $SW \in \{0, 1, \ldots, 10\}$ and weights sum to 10
**Model Definition**

**Graphical Model**

- **word**
  - word transition
  - phone position
  - phone transition
  - phone
  - audio state position
  - audio state transition
  - audio state
  - viseme position
  - viseme transition
  - viseme
  - video state position
  - video state transition
  - video state
  - audio observation
  - stream weight
  - video observation

---

**Setup**

**Experimental Setup**

- **Bernstein Database**
  - High quality audio and video
  - Studio setting
  - Used female speaker only
  - Training Set: first 379 sentences (80% of data)
  - Testing Set: last 94 sentences (20% of data)

- **Audio Features**
  - MFCCs + delta + acceleration

- **Video Features**
  - Group 8 MPEG-4 FAPs (outer lip)
  - Used PCA to reduce dimensionality from 10 to 2
  - Concatenated delta and acceleration values
Compared our model ("phone-viseme") to recent model ("phone-phone") that models video sub-word unit as phones, not visemes

Relevant Information/Parameters

- No grammar
- Three states per sub-word unit
- No state skipping allowed
- Word pronunciation variants used
- Monophthongs map to single viseme
- Dipthongs map to two consecutive visemes

Table: Results for both models for various values of the audio stream weight (Video stream weight = 10 - SW)
We model the complete recognizer using dynamic Bayesian network. Modeling phones and visemes helps performance. DBNs show a great deal of promise.

Ongoing work:
- Synchrony constraints at various boundaries
- Soft constraints within and across boundaries
- “Learning” the stream weights

Questions?

For more, including downloading this presentation, please visit:
http://ivpl.eecs.northwestern.edu/people/LHTerry
Thank You.