# JHU Summer Workshop Planning Meeting: Audio-Visual Speech Recognition

**Kate Saenko**

**Collaborators: Karen Livescu, Trevor Darrell**

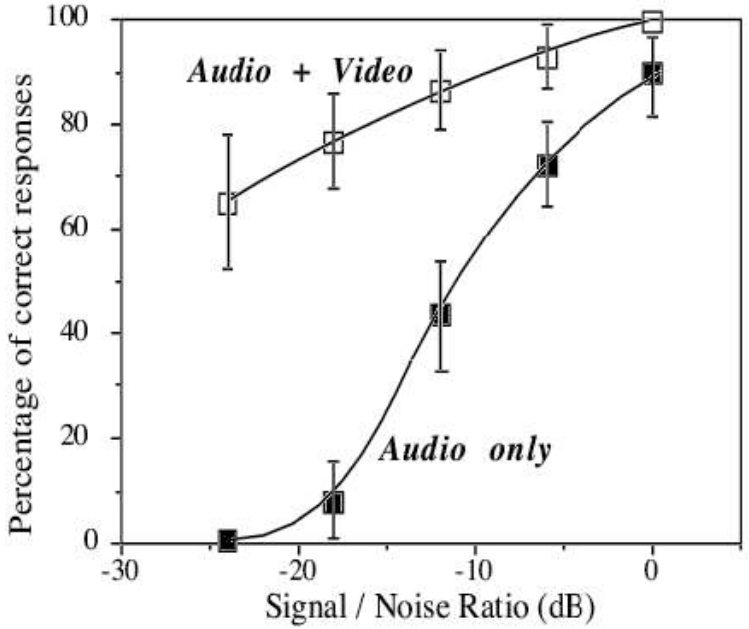
April 23, 2006

# Talk Outline

- **Audio-visual speech recognition overview**
  - * visual pre-processing
  - * visual feature types
  - * phonemes and visemes
  - * asynchrony modeling
  - * audio-visual fusion models
  - * databases
- **Webcam digits corpus**
  - * corpus description
  - * baseline models
  - * preliminary results
- **Our previous work on AF-based VSR**

# Audio-Visual Speech Recognition Overview

- **Visual signal improves recognition for humans**
- **Improves robustness of ASR**



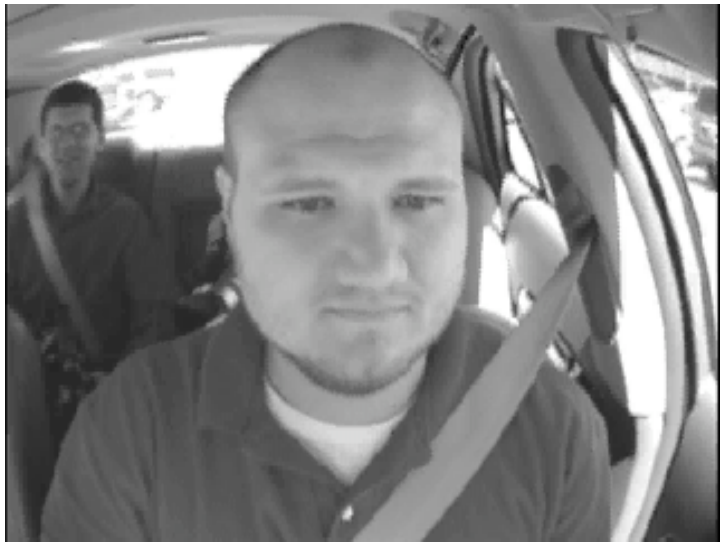
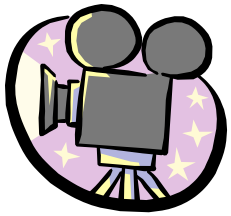
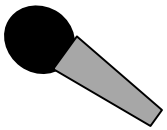



Figure from C. Benoit, "The intrinsic bimodality of speech communication and the synthesis of talking faces," 1992

# AVSR System Components





Face detection, Tracking, Visual Feature Extraction

Acoustic Feature Extraction

Audio-Visual Fusion, Speech Recognition

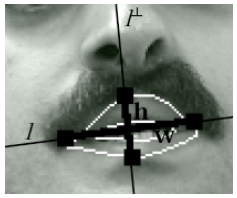
# Visual Pre-Processing Challenges

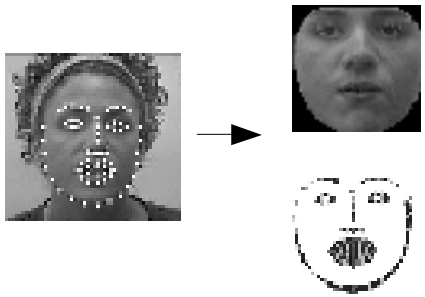
- **Face detection**
  - * rule based: eye/nose detection, symmetry
  - * sub-space: PCA, eigenfaces
  - * learning based: AdaBoost, SVM
  - * subject variability (beard, glasses)
  - * head pose, head movement
  - * lip deformation
  - * illumination

- **Region of Interest Extraction**
  - * include lips, teeth, tongue, chin, cheeks
  - * detection followed by tracking
  - * physiological subject differences
  - * speech related subject differences
  - * head pose
  - * variance of lip appearance due to illumination

# Visual Feature Extraction

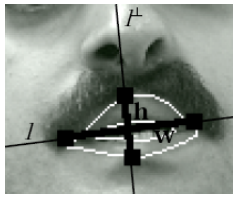
- **Model-based features**
  - * lip contour features
    - height, width, area
    - moments, Fourier Descriptors
    - contour models (deformable templates, snakes)
  - * active appearance models (Cootes et al. 98)
  - * articulatory features (Saenko et al. 04)

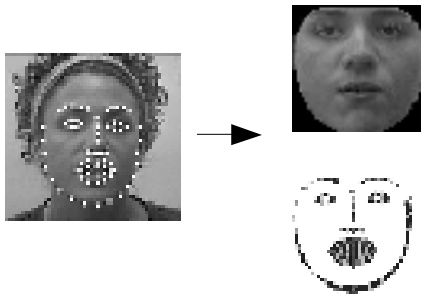


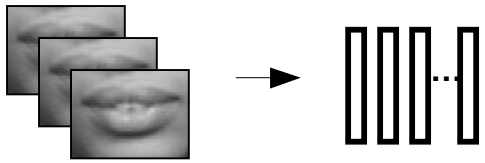
# Visual Feature Extraction

- **Model-based features**
  - * lip contour features
    - height, width, area
    - moments, Fourier Descriptors
    - contour models (deformable templates, snakes)
  - * active appearance models (Cootes et al. 98)
  - * articulatory features (Saenko et al. 04)

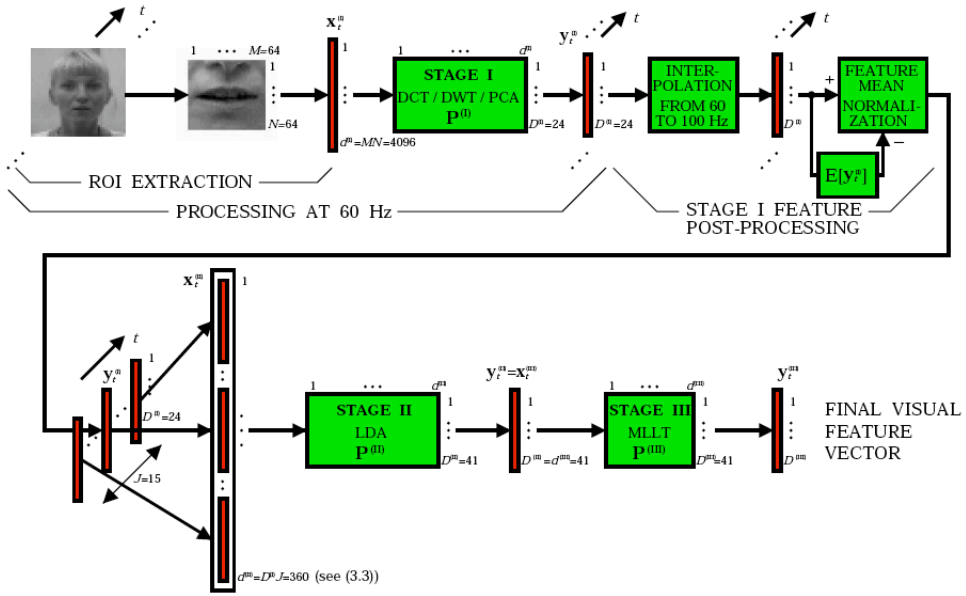


- **Appearance-based features**
  - * raw image pixels
  - * PCA/LDA (Eigenlips)
  - * DWT/DCT
  - * other linear transform



# Image Transform Features (Neti, et al. '00)




- **Remove redundant information via linear transform**
  - DCT/DWT/PCA
  - delta features
  - vector stacking to capture dynamics
- **Two further stages: LDA and MLLT**
  - separate feature space according to set of classes
  - rotate feature space to account for diagonal covariance assumption in GMMs

# Comparison of Active Appearance Models and Transform-based Features

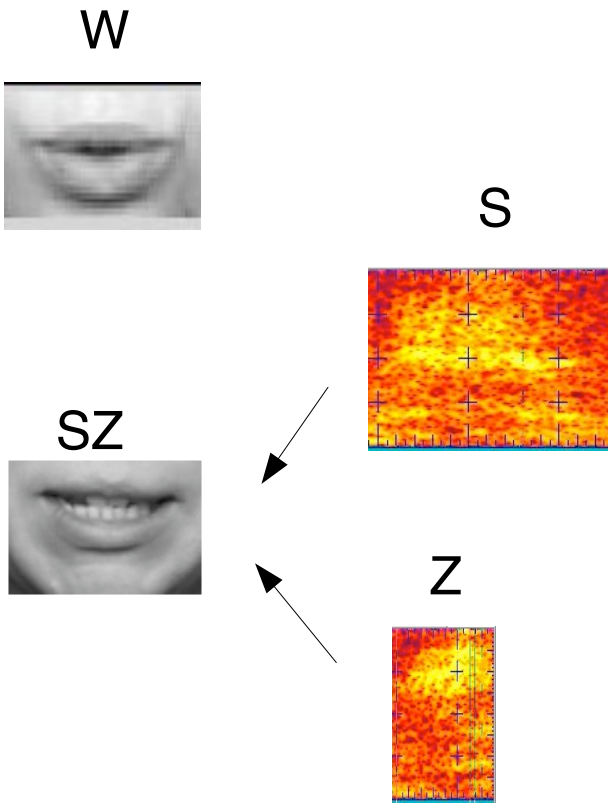
*Matthews, Potamianos, Neti and Luettin, JHU Workshop 2000*

| Modality | Parameterisation | WER % |
|---|---|---|
| Visual | DCT | 58.1 |
| | DWT 3 | 58.8 |
| | PCA | 58.8 |
| | DWT 2 | 59.4 |
| | AAM | 64.0 |
| Acoustic | MFCC (noisy audio) | 55.0 |

**Table 1:** Speaker independent, large vocabulary, continuous, audio-visual recognition word error rates (WER) for each of the proposed visual feature parameterisations, based on lattice rescoring. Audio-only (at 8.5 dB SNR), and characteristic lattice WERs are also shown.

# Audio-Visual Speech Units

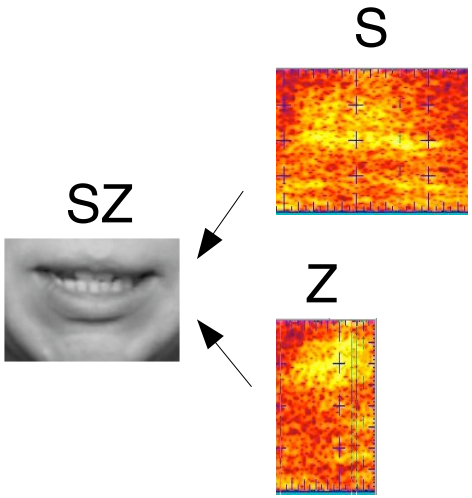
- **traditional basic units are phonemes and visemes**

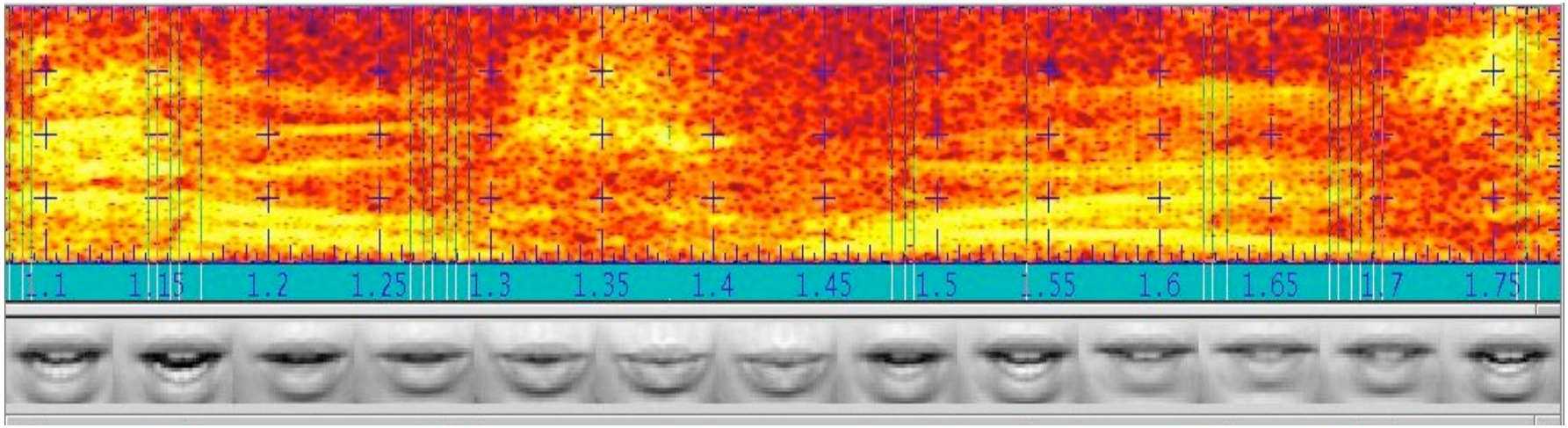



| Viseme | Phoneme |
|---|---|
| Silence | _ epi |
| A | aa ah ao aw er oy hh |
| U | uh uw ow |
| E | ae eh ey ay |
| I | ih iy ax axr |
| SZ | s z |
| TD | t d dx n en |
| SZH | ch ih sh zh |
| PB | b p m w |
| FV | f v |
| KG | g k ng |
| LR | l el r y |
| TDH | th th |

# Audio-Visual Asynchrony

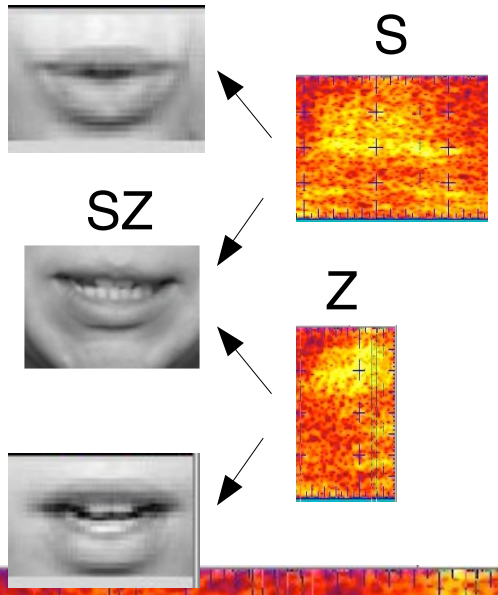
- **E.g. “housewives”**
- **asynchrony between audio and video**

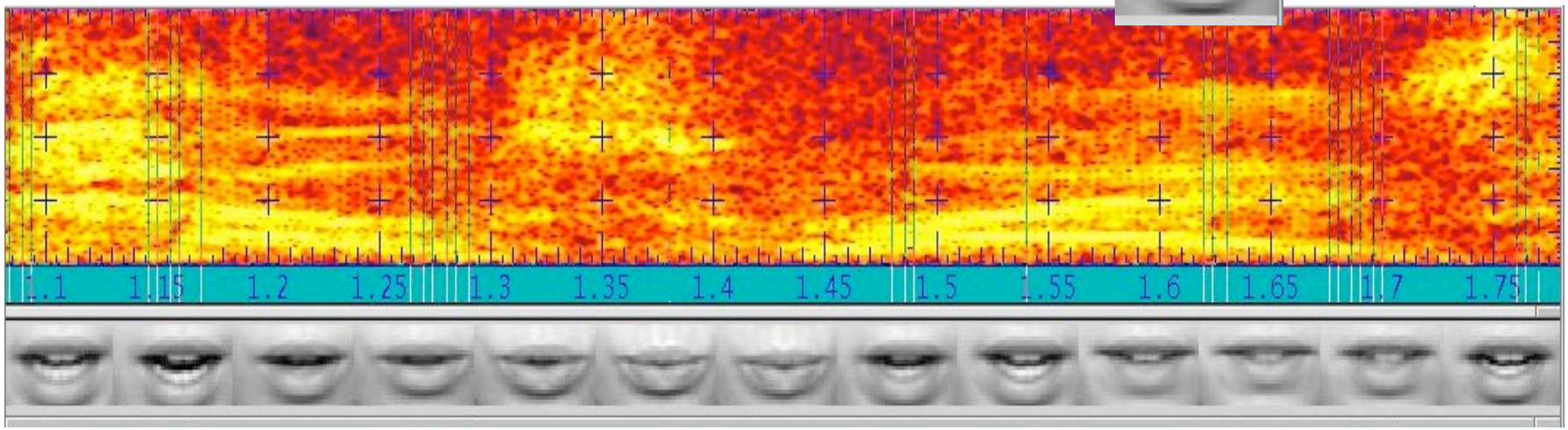


# Audio-Visual Asynchrony

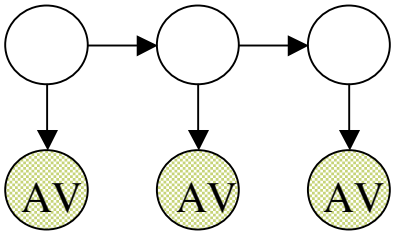
- **E.g. “housewives”**
- **asynchrony between audio and video**
- **many-to-one mapping does not hold!**



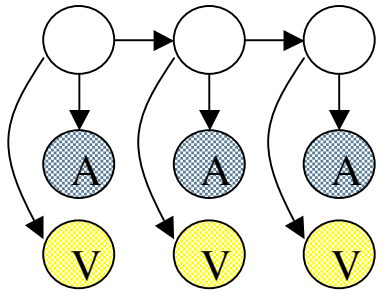



# Audio-Visual Fusion Models

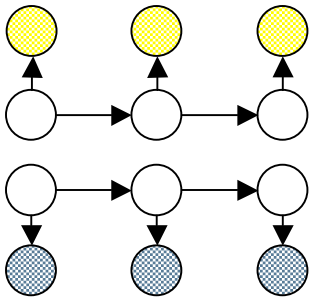
**Single observation stream (e.g. Neti et al., '00)**



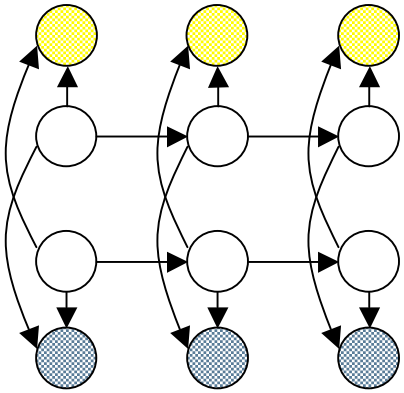

**Two observation streams (e.g. Verma et al., '99)**



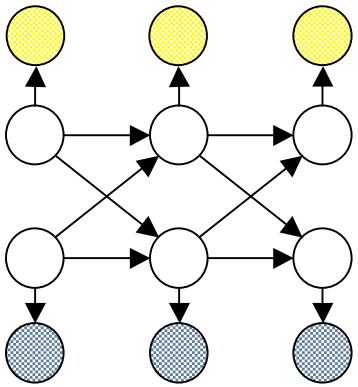

**Independent two-stream HMM (Dupont, '00)**



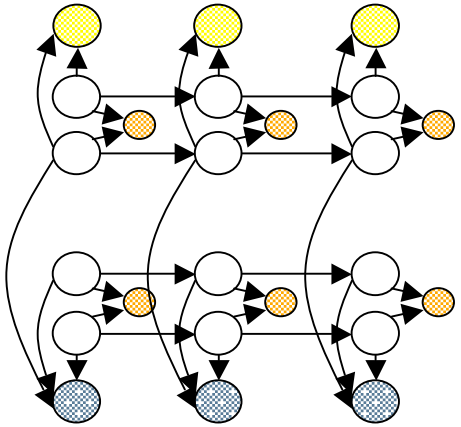
**Factorial HMM (Nefian et al, '02)**



**Coupled HMM (Liang et al, '02)**

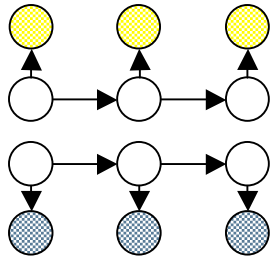


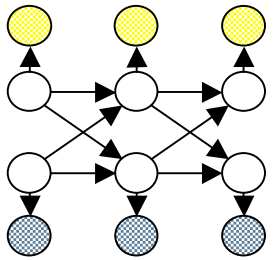
**AF-based DBN (our proposed model)**



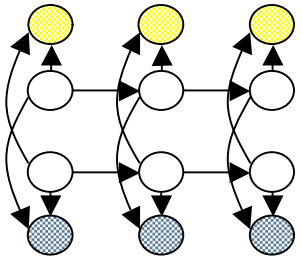
# Asynchronous Fusion Models

- **Synchronous vs. asynchronous models**
  - * Modeling asynchrony improves performance (Liang '00)
  - * But where to re-synchronize? At phone level? Word level? Sentence level?
  - * asynchrony due to context extends across units





- **What do the two state streams represent?**
  - * Phone and viseme streams
  - * Two independent (visual and acoustic) processes?
  - * But there are some dependencies between streams



# Audio-Visual Speech Corpora

| Name/Inst. | ASR Task | Noise | No. Subjects |
|---|---|---|---|
| UIUC | C digits | - | 100 |
| UIUC | I digits/letters/sents | yes, car | 100 |
| CUAVE,Clemson | C digits | - | 36 |
| IBM | C digits | - | 50 |
| Webcam, MIT | C digits | yes | 100 |
| AT&T | C letters | - | 49 |
| CMU | I words | - | 10 |
| AV-ViaVoice,IBM | 10.4k vocab | - | 290 |
| AVTIMIT, MIT | TIMIT sents | - | 230 |

**Table 1. Several English speech AV corpora.**

# Webcam Digits Corpus

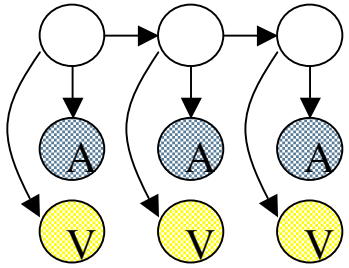
- collected at MIT
- 100 speakers (50 came back for 2nd recording)
- 26 sequences each
- 10 random digits per sequence
- three environments: “office”, “lobby”, “outside”

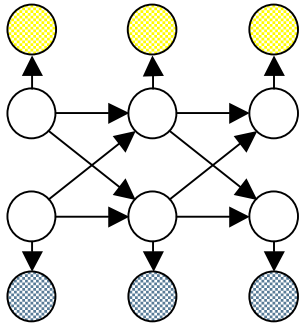
# Webcam Digits Experiments

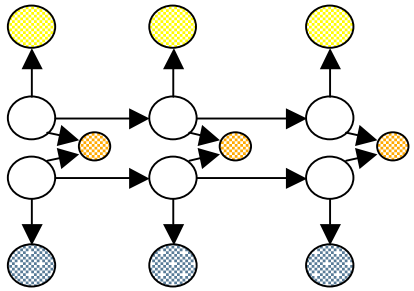

- all videos converted to raw AVI, Matlab, WAV formats
- lip tracking reasonable, failed in a few cases
- visual features: 35 DCT coefficients from 32x32 region, 100Hz
- acoustic features: 42 MFCC+delta+delta2, 100Hz
- feature mean subtraction
- split data into train/test/dev sets (~400-500utts each)

# Baseline DBN models

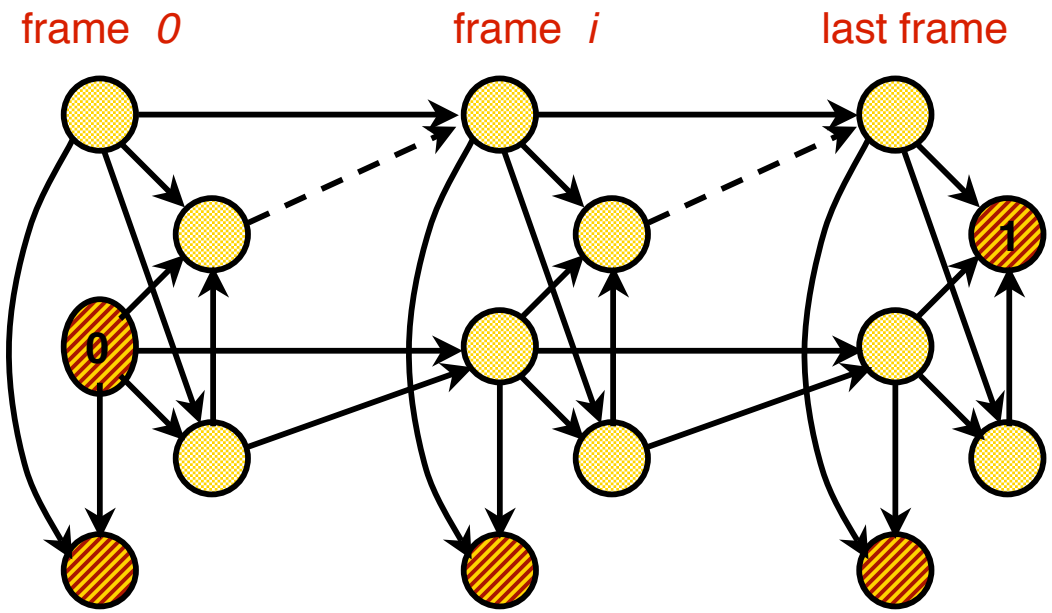
- **Audio-only models**
  - **whole-word HMM**
  - **phone-based HMM**
- **Video-only models**
  - **whole-word HMM**
  - **phone-based HMM**
- **Synchronous Audio-Visual models**
  - **whole-word MHMM**
  - **phone-based MHMM**
  - **whole-word 2-stream model with forced synchrony**
  - **phone-based 2-stream model with forced synchrony**
- **Asynchronous Audio-Visual models**
  - **whole-word/phone-based CHMM**
  - **whole-word/phone-based CHMM, cross-word async**
  - **whole-word/phone-based 2-stream with sync var. (Livescu-Glass), trained async probabilities**
    - word boundary sync
    - cross-word async





# Initial results: whole-word, single-stream HMM

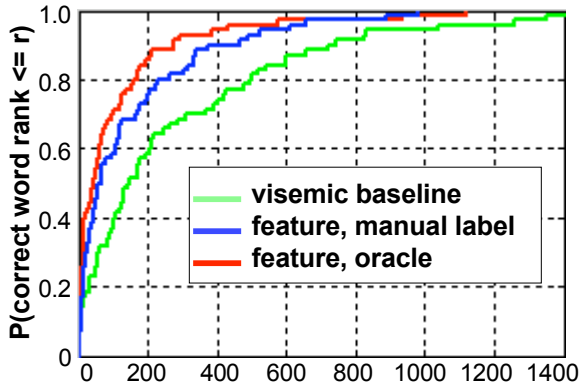



| | AHMM | VHMM |
|---|---|---|
| office | 8.1% | 80.4% |
| lobby | 33.7% | 81.9% |
| outside | 38.6% | 81.6% |

**Table 2. WER of models trained on office train set**

# Our previous work on articulatory features for *visual-only* speech recognition

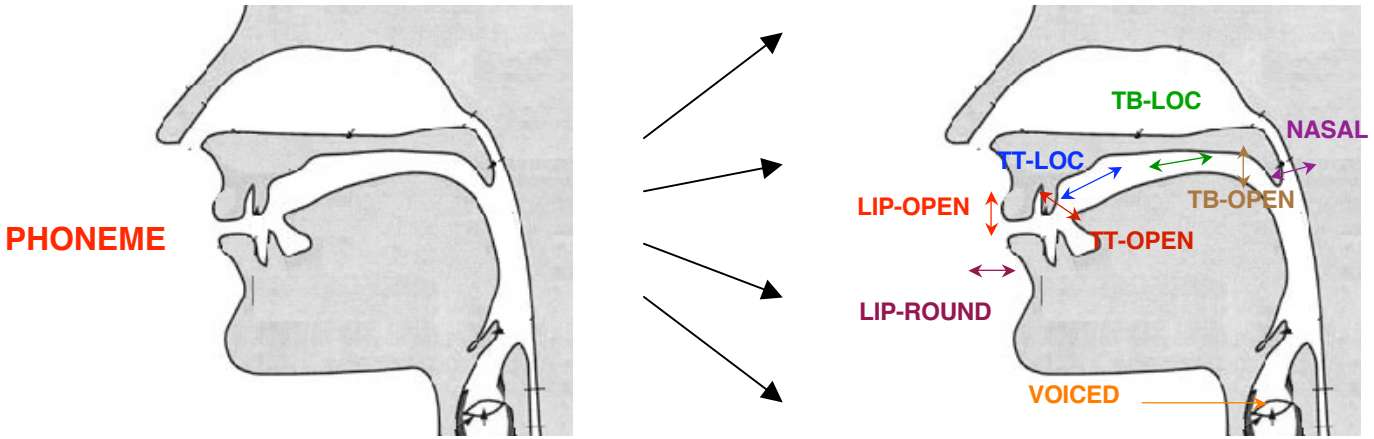
- **AVTIMIT visual-only word ranking, ICASSP '05**




- **small-vocabulary visual speech recognition, ICCV '05**
  - * results on small AV database of short phrases
  - * will be described in the rest of the slides

# Articulatory Features for VSR

- **Articulatory Features (AFs)**
  - **replace phonemes**
  - **factor phoneme/viseme state space into multiple streams**
  - **asynchrony is between articulator streams**




K. Saenko, J. Glass, and T. Darrell, “Articulatory features for robust visual speech recognition,” in *Proc. ICMI*, 2004.

K. Livescu and J. Glass, “Feature-based pronunciation modeling with trainable asynchrony probabilities,” in *Proc. ICSLP*, 2004.

# Visual Articulatory Feature Set

**lip opening**

closed/narrow/medium/wide

**lip rounding**

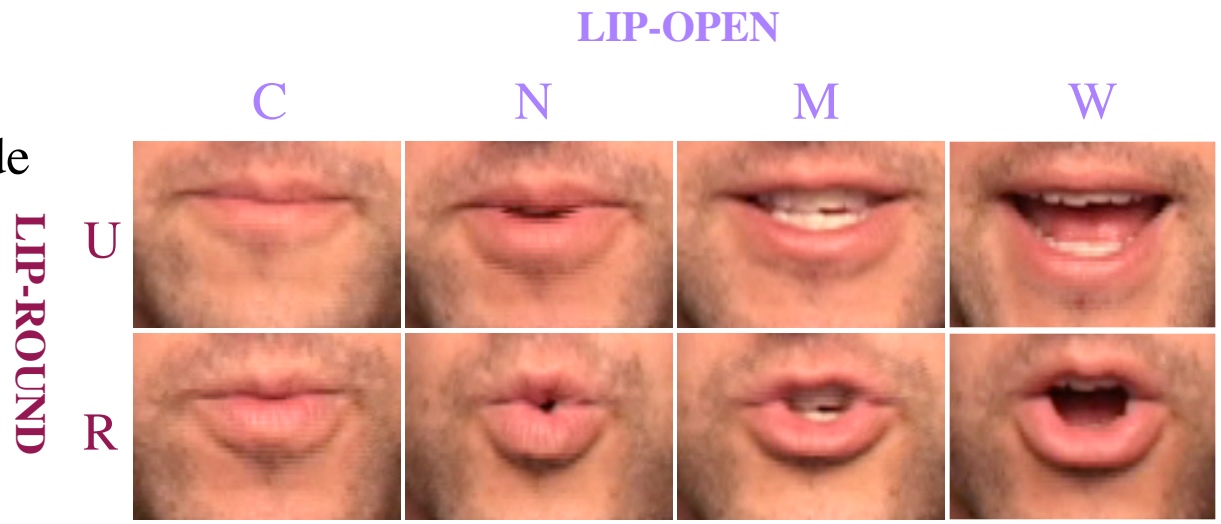
rounded/unrounded

**labio-dental**

yes/no

**teeth position**

unknown/neutral/open

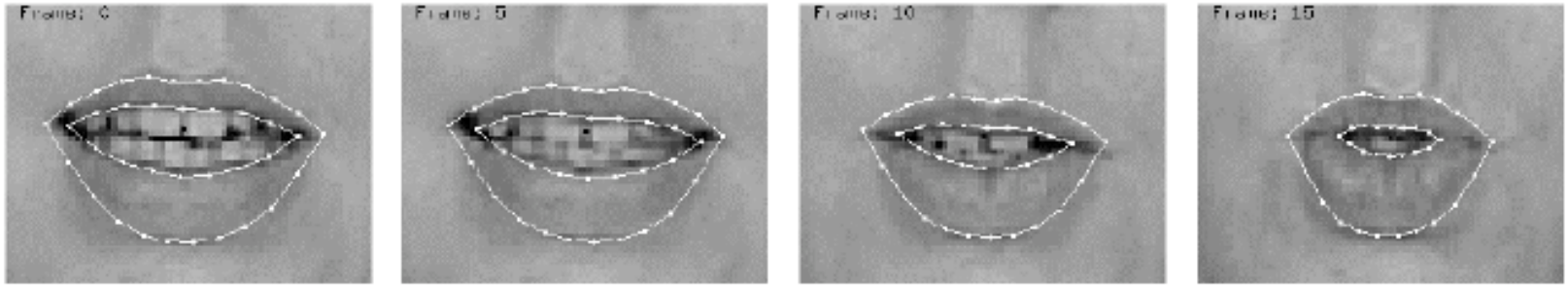



**Example**: Phone to AF mapping

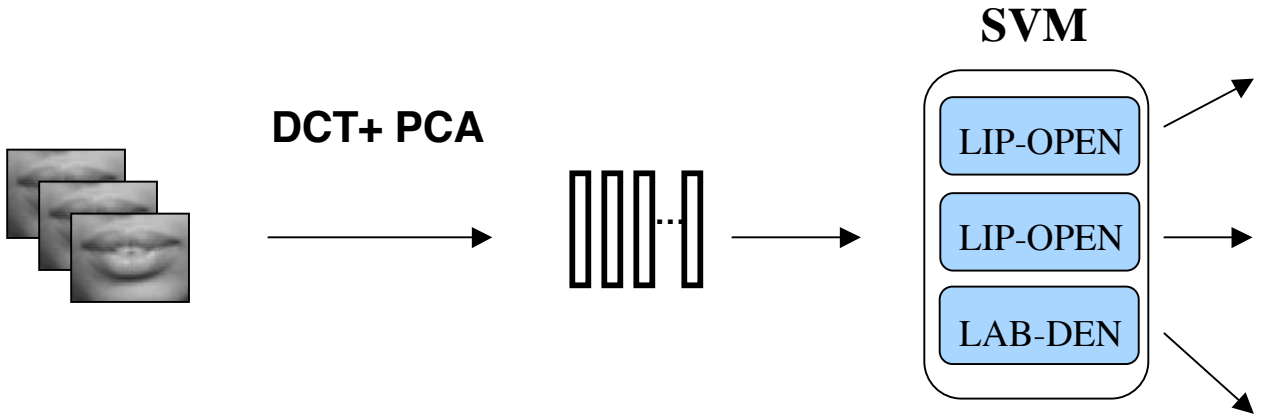
| Phone | ch | ow | z | en |
|---|---|---|---|---|
| LIP-OPEN | M | N | M | M |
| LAB-DENT | N | N | N | N |
| LIP-ROUND | R | R | U | U |

# Visual AF Classification

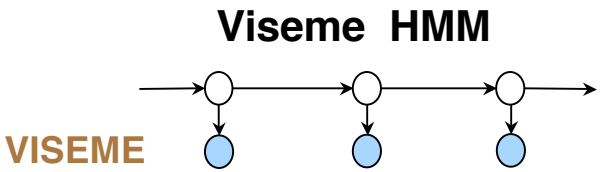
- lip contour tracking is not robust




- instead: parallel SVM classifiers, one per AF $F^i$
- train on manually labeled examples

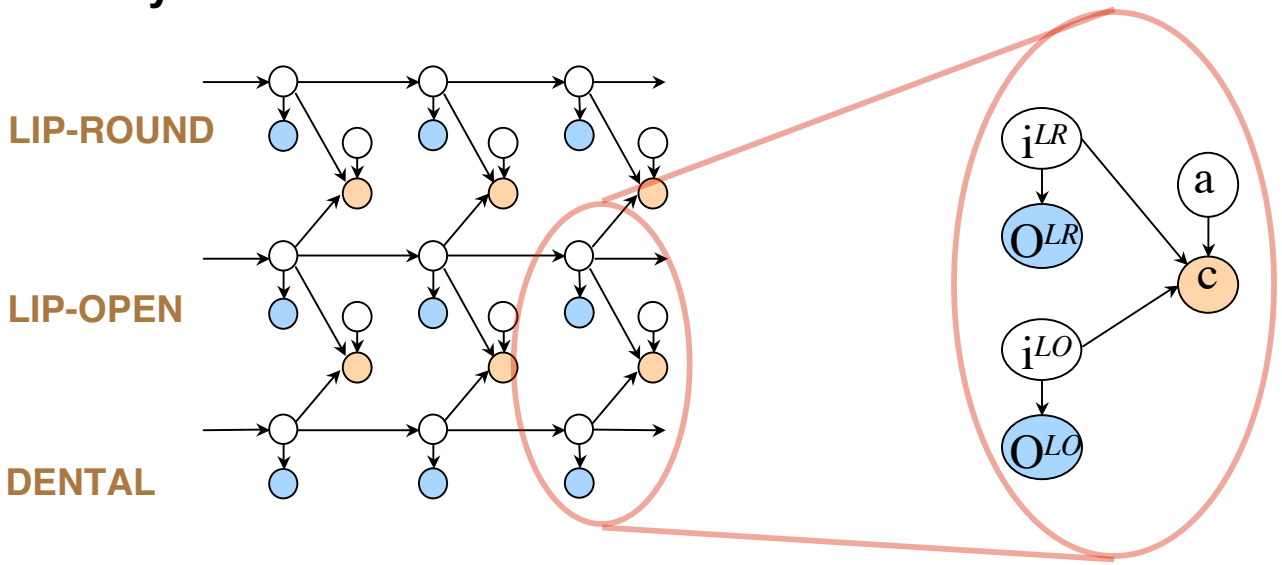


# Dynamic Bayesian Network



# Dynamic Bayesian Network

**Articulatory Feature-based DBN**




$$a = | i^{LR} - i^{LO} |$$

**asynchrony constraints:**

$$P(c \mid a, i^{LR}, i^{LO}) = \begin{cases} 1 \text{ if } a = | i^{LR} - i^{LO} | \\ 0 \text{ if } a \neq | i^{LR} - i^{LO} | \end{cases}$$

**SVM outputs → posterior probabilities → scaled likelihoods:**

$$P(O_t = o \mid F^i_t = f) \propto P(F^i_t = f \mid O_t = o) / P(F^i_t = f)$$

# Model Training

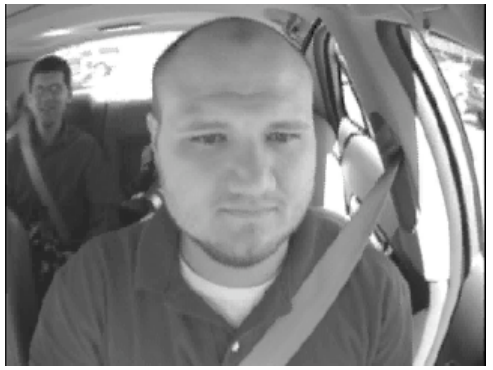
- one-vs-all SVM formulation
- run cross validation to optimise SVM parameters
- use standard DBN inference algorithms (GMTK)
- all of the parameters of the DBNs are learned simultaneously via maximum likelihood using the Expectation-Maximization (EM) algorithm

# Experiments

- investigate behavior of our model
- compare it to single-stream viseme HMM
- compare AFs to standard appearance features

# Recognition Task

- **20 commands to control a stereo system**
  - "begin scanning"
  - "browse country"
  - "CD player off"
  - ...



- **three speaking rates: slow, medium, fast**

| | Dataset | No. of speakers |
|---|---|---|
| **SVM classifiers** | phonetically-balanced TIMIT sentences | 2 |
| **DBN / HMM** | stereo system control commands | 2 |

# Experiments: AF Classification

- mapping from visemes to features

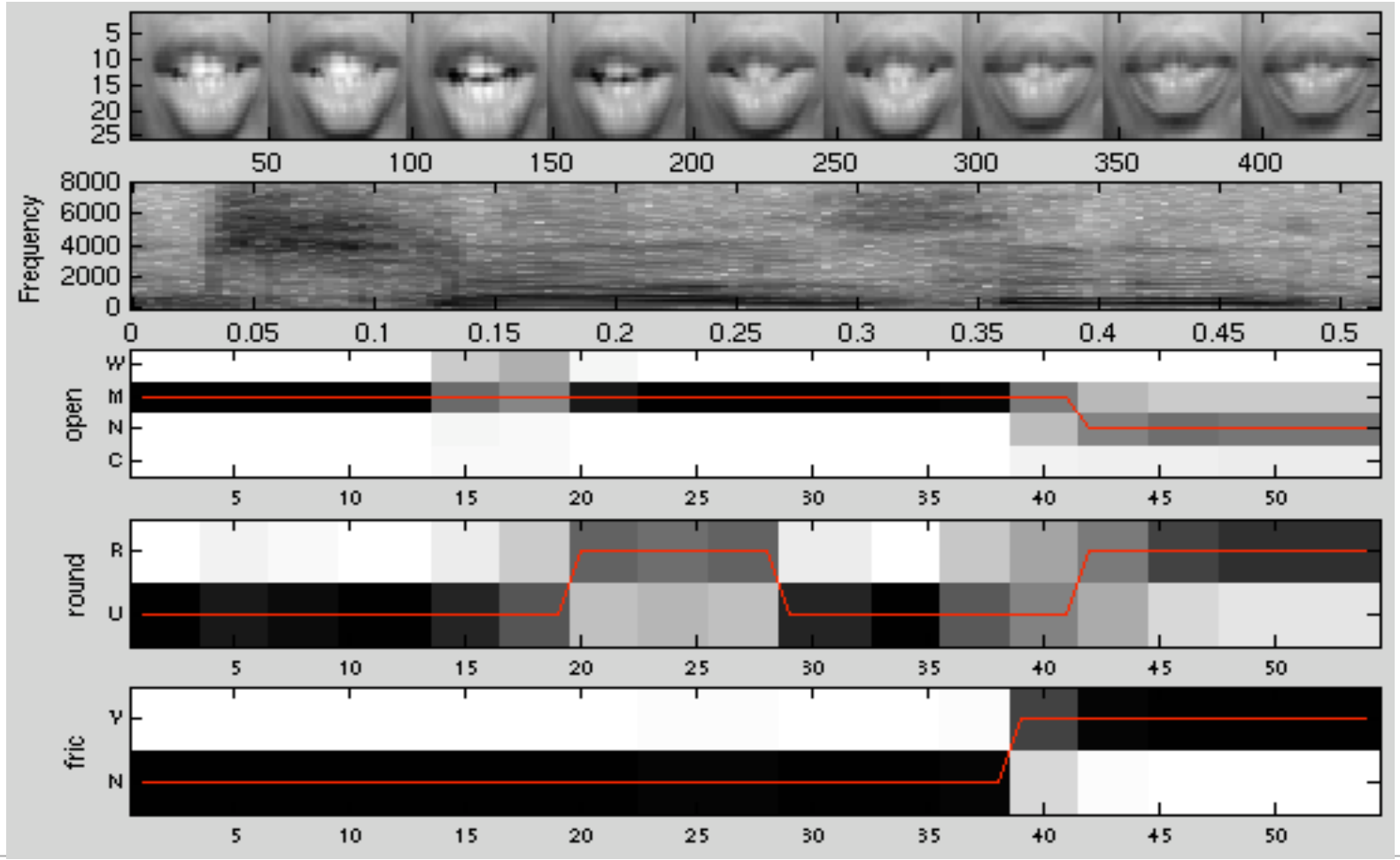
| VISEME | LIP-OPEN | LIP-ROUND | LABIO-DENTAL |
|---|---|---|---|
| 1 | closed | any | no |
| 2 | any | any | yes |
| 3 | narrow | round | no |
| 4 | medium | unround | any |
| 5 | medium | round | any |
| 6 | wide | any | any |

- average classifier accuracies

| | LIP-OPEN | LIP-ROUND | LABIO-DENTAL | VISEME |
|---|---|---|---|---|
| accuracy % | 79 | 78 | 57 | 63 |
| random % | 25 | 50 | 50 | 17 |

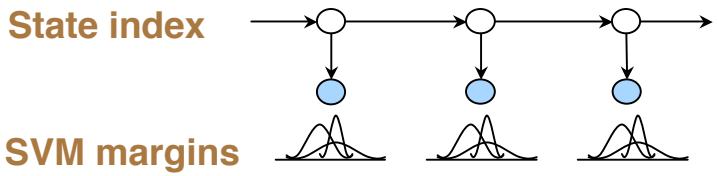
# AF Classifier Outputs

## Example : AF outputs for “chosen” (followed by “few”)

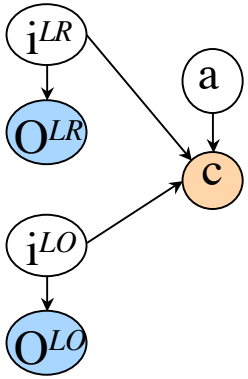


# Model Variants

- **whole word models**



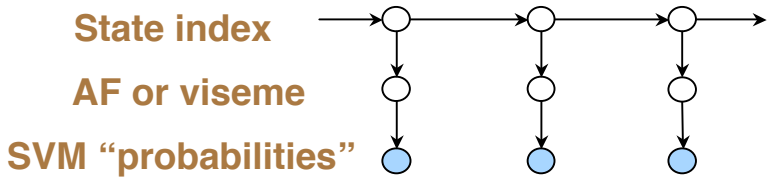

- **asynchrony in AF model**



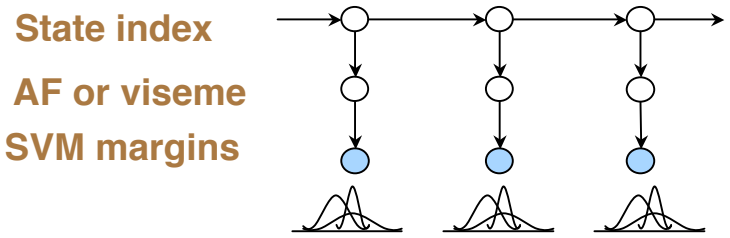

Synchronous:
a=0
Asynchronous:
a=0,1

- **dictionary-based models**

Dictionary + soft evidence (SE)*




Dictionary + Gaussian Mixtures (GM)




* model from Saenko, et al, ICASSP 05

# Experiments: dictionary-based

- **dictionary models**
- **GM vs. soft evidence (SE) observation models**

| train data speaking rate | dictionary | | |
|---|---|---|---|
| | **viseme GM** | **feature SE** | **feature GM** |
| slow, med | 10 | 7 | 13 |
| slow, fast | 13 | 13 | 21 |
| med, fast | 14 | 21 | 18 |
| average | 12.3 | 13.7 | 17.3 |
| **average%** | **30.8** | **34.2** | **43.3** |

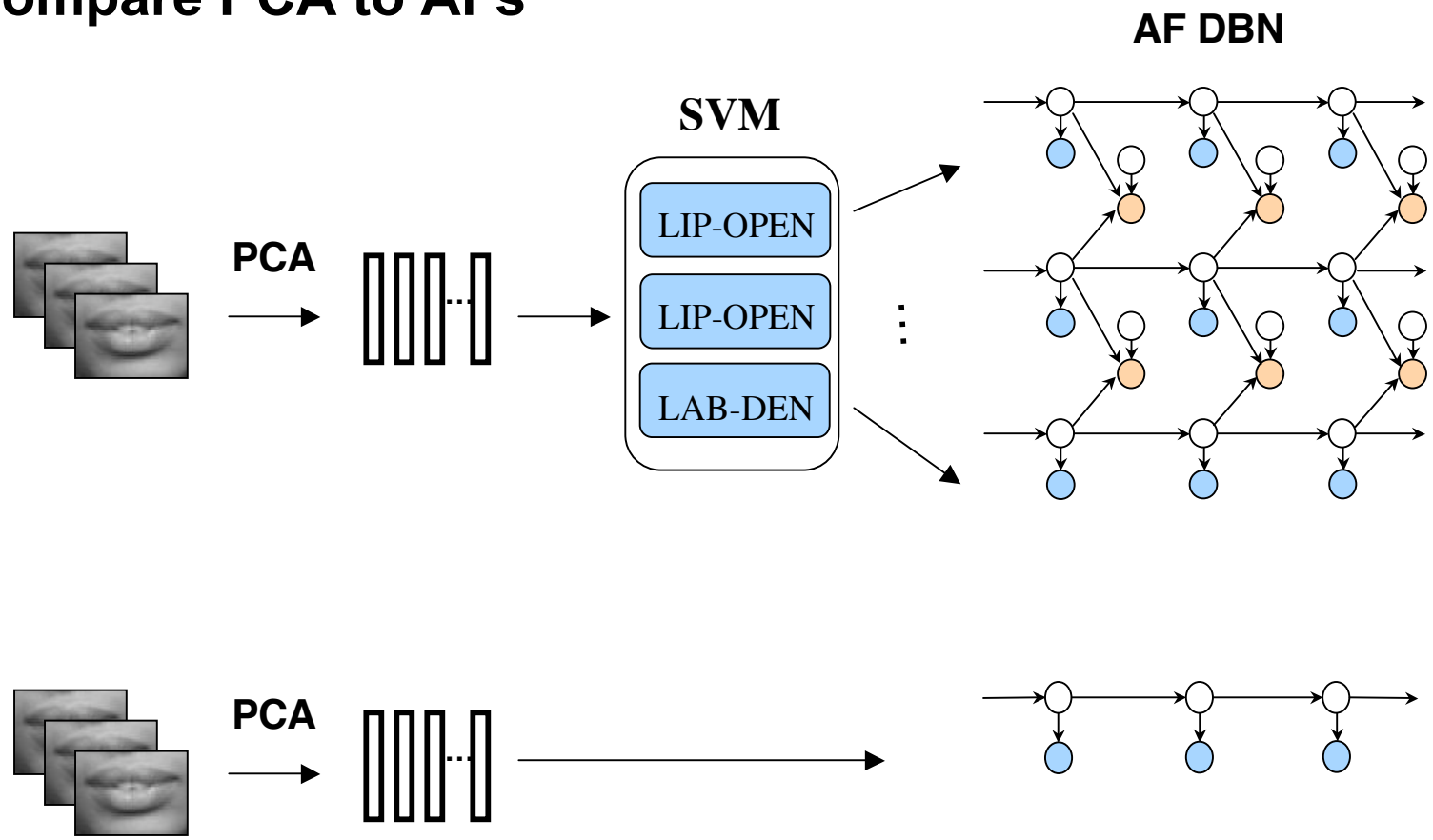
# Experiments: whole word model

- **whole-word viseme and AF models**
- **allow feature asynchrony**

| train data speaking rate | dictionary | | | whole-word | | |
|---|---|---|---|---|---|---|
| | **viseme GM** | **feature SE** | **feature GM** | **viseme GM** | **feature GM** | **async feature GM** |
| slow, med | 10 | 7 | 13 | 16 | 23 | 25 |
| slow, fast | 13 | 13 | 21 | 19 | 29 | 30 |
| med, fast | 14 | 21 | 18 | 27 | 25 | 24 |
| average | 12.3 | 13.7 | 17.3 | 20.7 | 25.7 | 26.3 |
| **average%** | **30.8** | **34.2** | **43.3** | **51.6** | **64.1** | **65.8** |

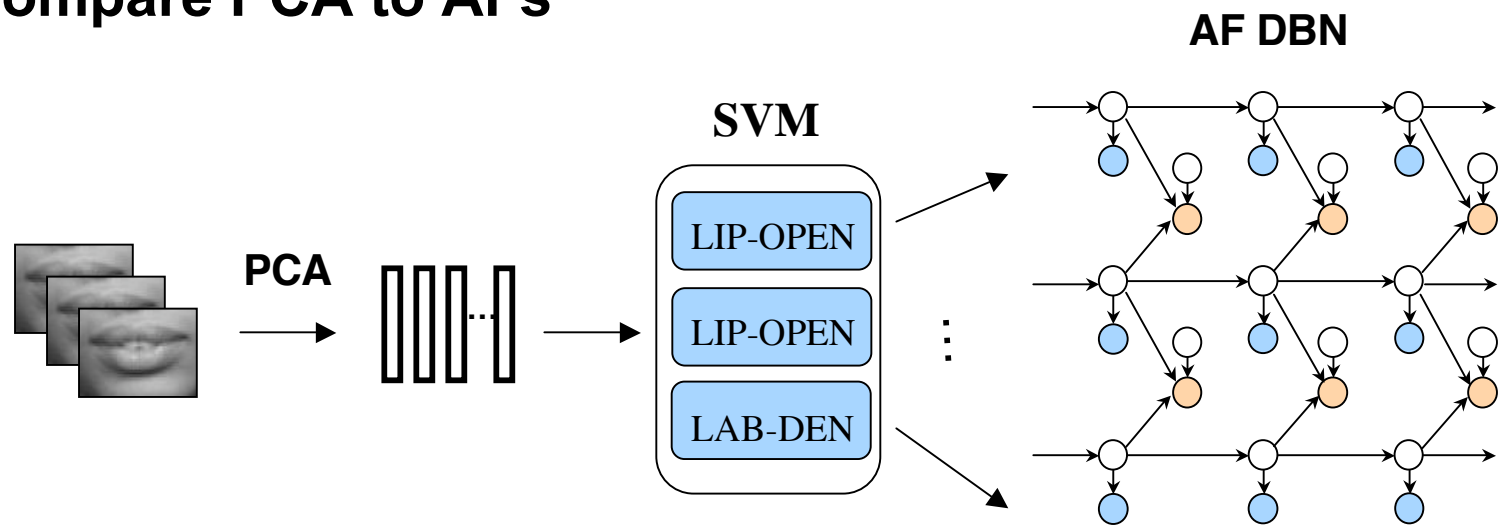
# Experiments: PCA features

- compare PCA to AFs



# Experiments: 3 Feature Set

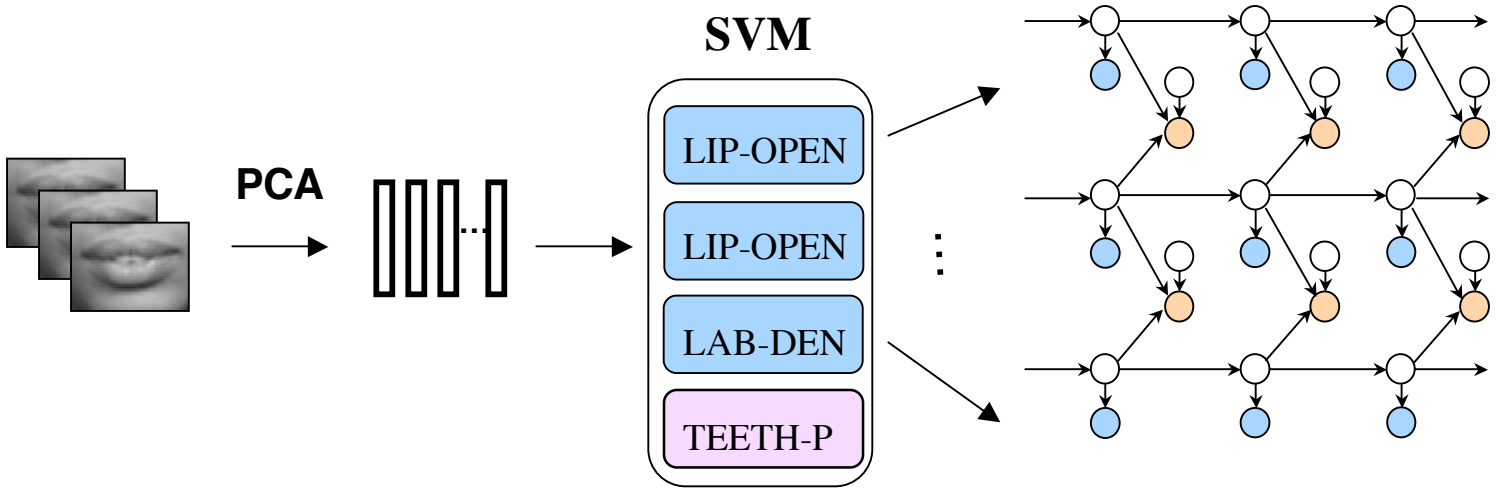
- compare PCA to AFs




| | HMM baseline | | 3 AF DBN | |
|---|---|---|---|---|
| | PCA | 6-VIS | sync | async |
| average % correct | 77.5 | 51.6 | 64.1 | 65.8 |

# Experiments: New Feature

- add 4th feature stream for Teeth Position (TP)



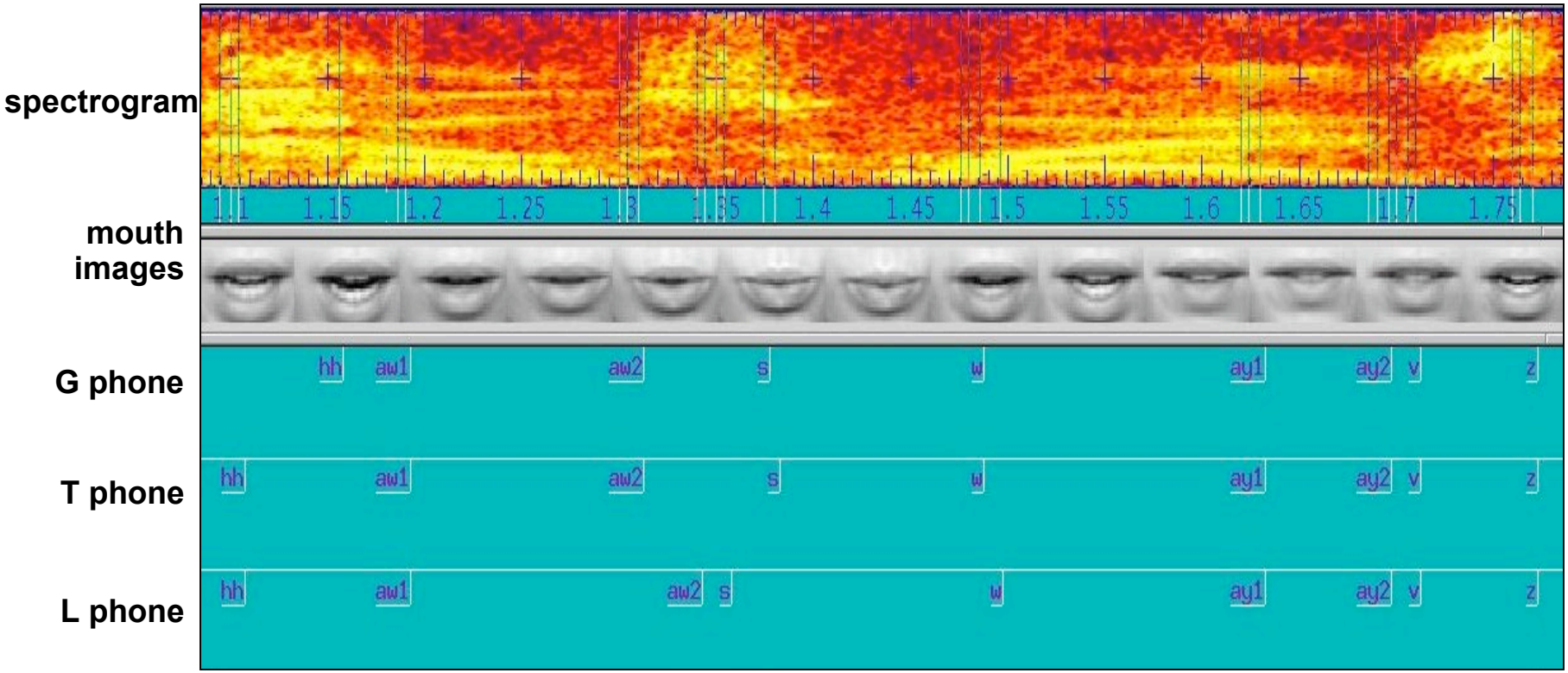

| | HMM baseline | | | 3 AF DBN | | 4 AF DBN | |
|---|---|---|---|---|---|---|---|
| | PCA | 6-VIS | 8-VIS | sync | async | sync | async |
| average % correct | 77.5 | 51.6 | 58.3 | 64.1 | 65.8 | 77.5 | 79.2 |

# Summary

- introduced AF-based model for VSR
- evaluated on command phrase task
- obtained improved performance over viseme classifiers
- with lip and teeth features, improved upon PCA HMM
- our model achieves 79% recognition

# Future (JHU Workshop 2006)

- **AF-based model for audio-visual speech!**



# Thank You!