Modeling the Synchrony between Audio and Visual Modalities for Speaker Identification

- *Yu WANG, Zhiyong WU, Lianhong CAI, Helen M. MENG*
  王愈，吴志勇，蔡莲红，蒙美玲
  Tsinghua University,
  The Chinese University of Hong Kong

Outline

- Background and motivation
  - Synchrony between audio and visual speeches
  - Audio-visual integration
- AVCM: Audio-Visual Correlative Model
- Durational-AVCM
- Experiments

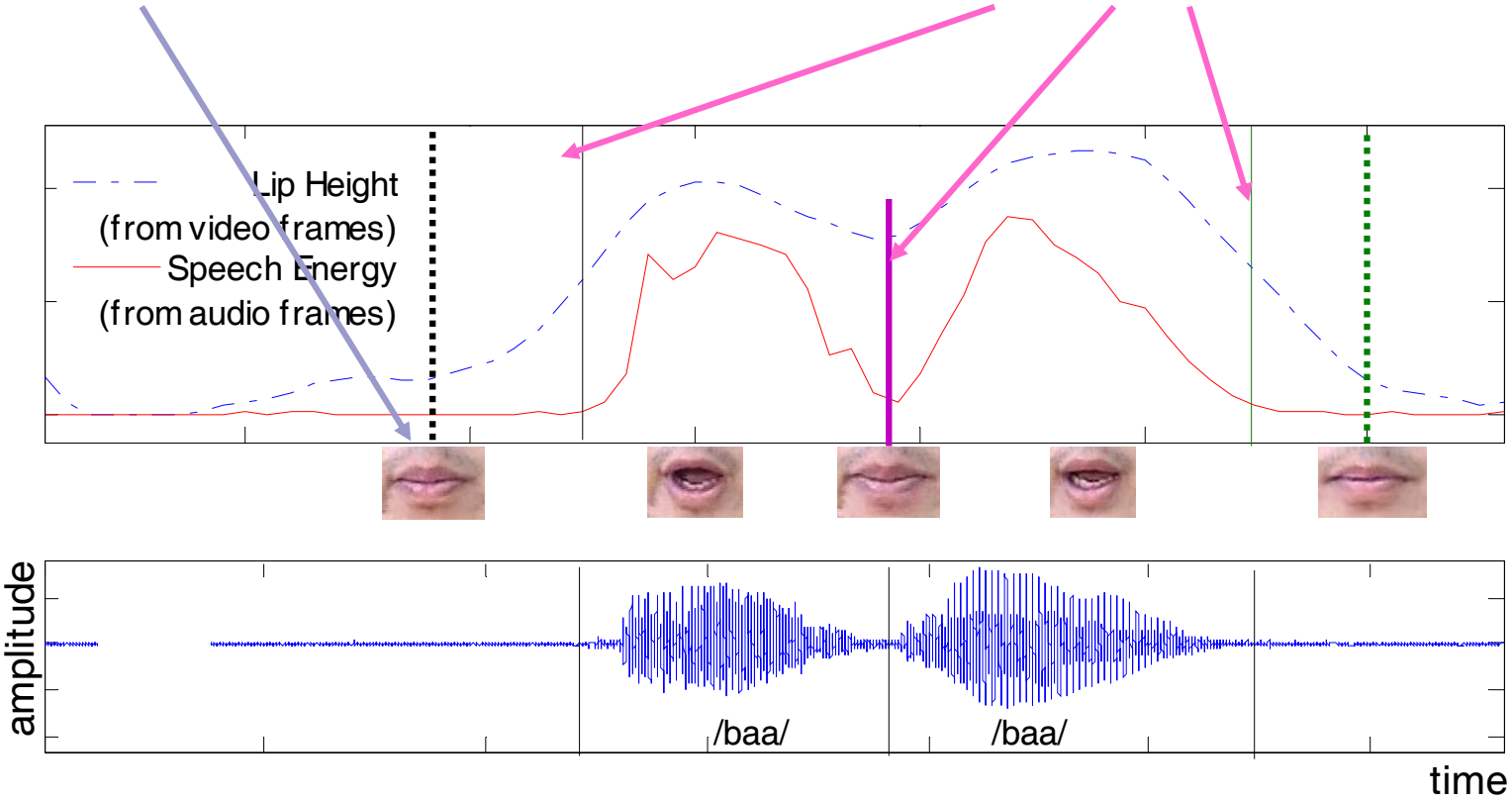
Background and Motivation

- Audio and Visual Speech
  - Most important in human-human communication
  - Most natural in human-computer interaction
- Human speech is bimodal in nature
  - Audio speech is produced by movement of the aiticulators with airflow through the vocal tract
  - Visual speech is due to the observable movement of articulators
- Two modalities are complementary
  - Macleod (1987): the contribution of visual information to the speech perception in noise will be 8-10dB*

Macleod A, 1987, et al. Quantifying the contribution of vision to speech perception in noise.

Audio-Visual Correlations

1. Inter-dependency **2. Partial temporal synchrony**



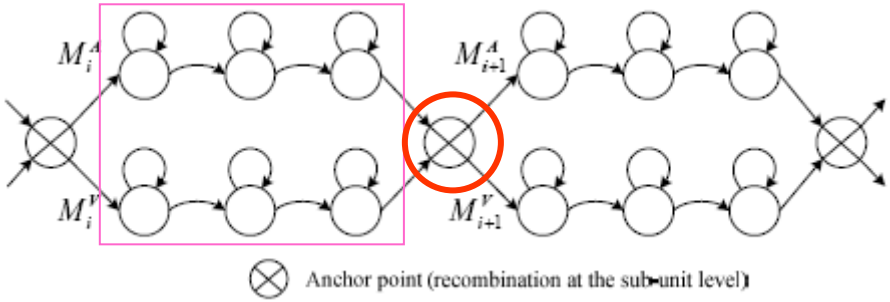
Extracted from a video with somebody uttering “baa-baa”

video leading

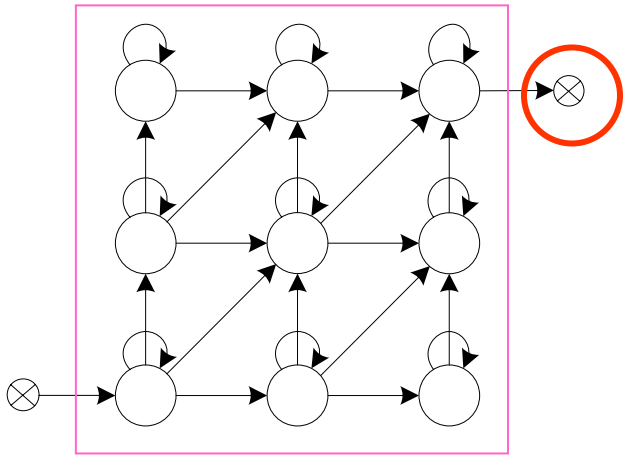
Audio-Visual Integration

- Feature-level fusion
  - Concatenating multiple features into a large vector
  - Training a single model
  - Assuming: audio-visual *in strict synchrony*
- Decision-level fusion
  - Processing audio and visual features separately
  - Building two independent models
  - Assuming: *complete independence*, ignores the audio visual correlations.
- Model-level fusion

Model-level Fusion



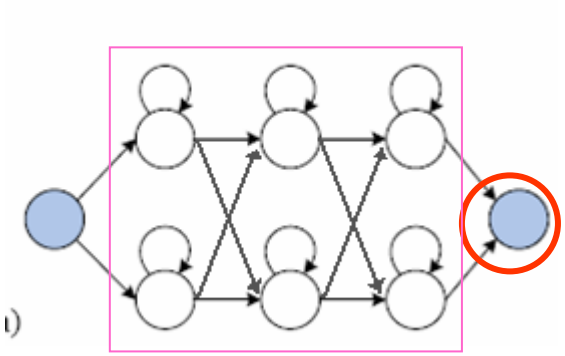
multi-stream HMM



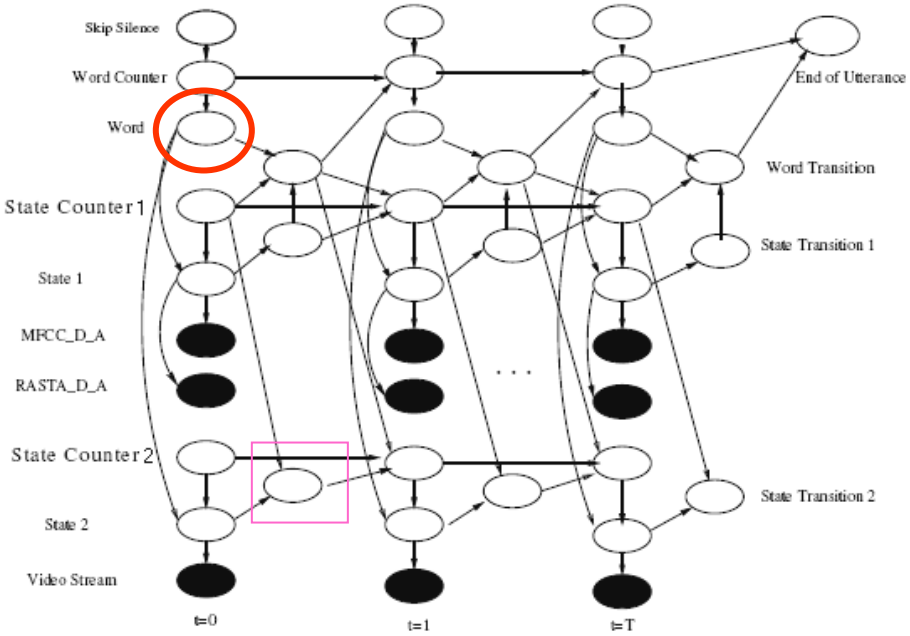
factorial HMM

- Assuming stream *independence* within unit
- Forcing *strict synchrony* at unit boundaries by introducing “anchor-points”

Model-level Fusion



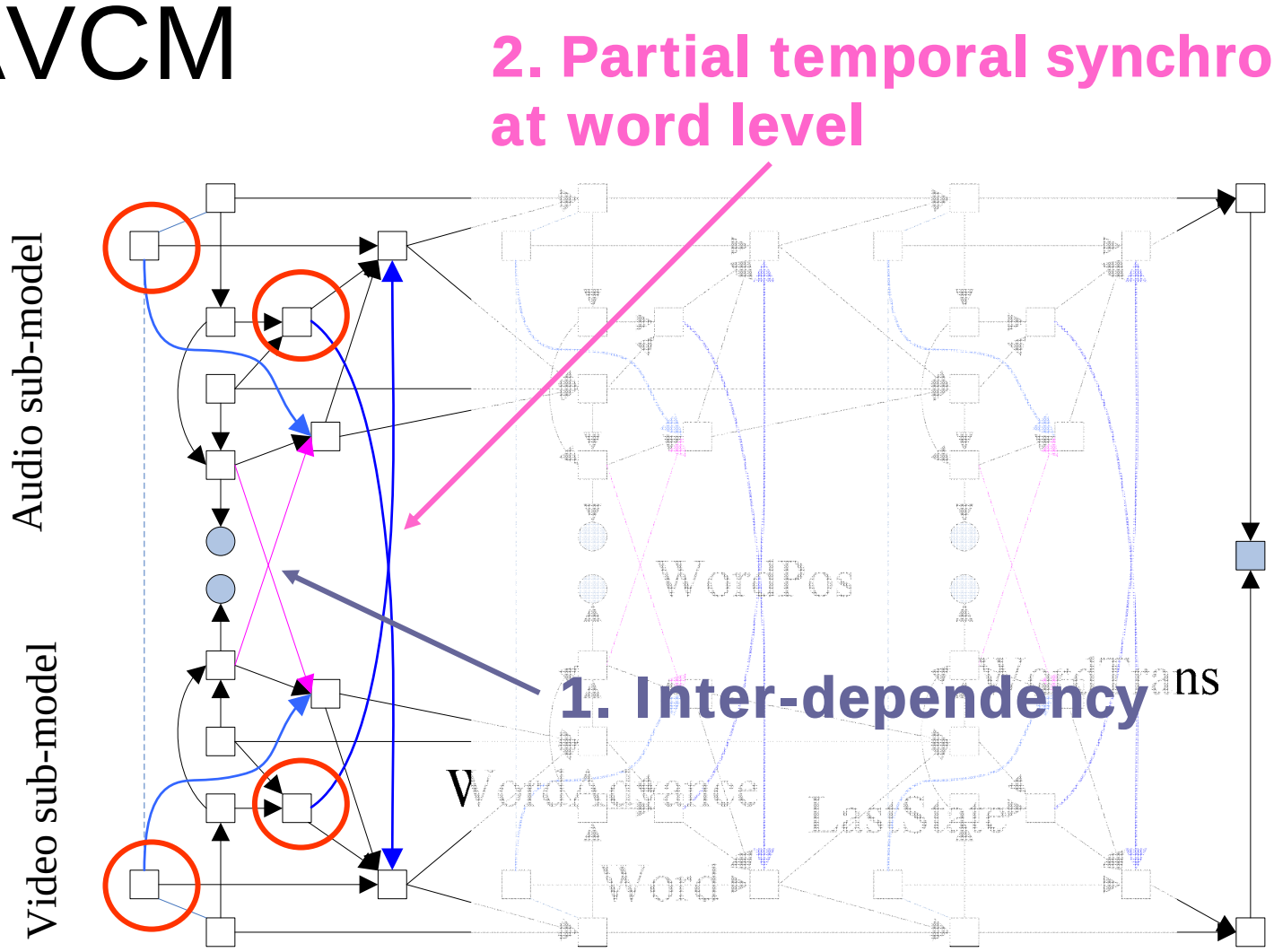
Coupled HMM



Gowdy's mixed-DBN

- Considering *inter-dependences* between streams
- Still forcing *strict synchrony* at unit boundaries by introducing "anchor-points"

AVCM



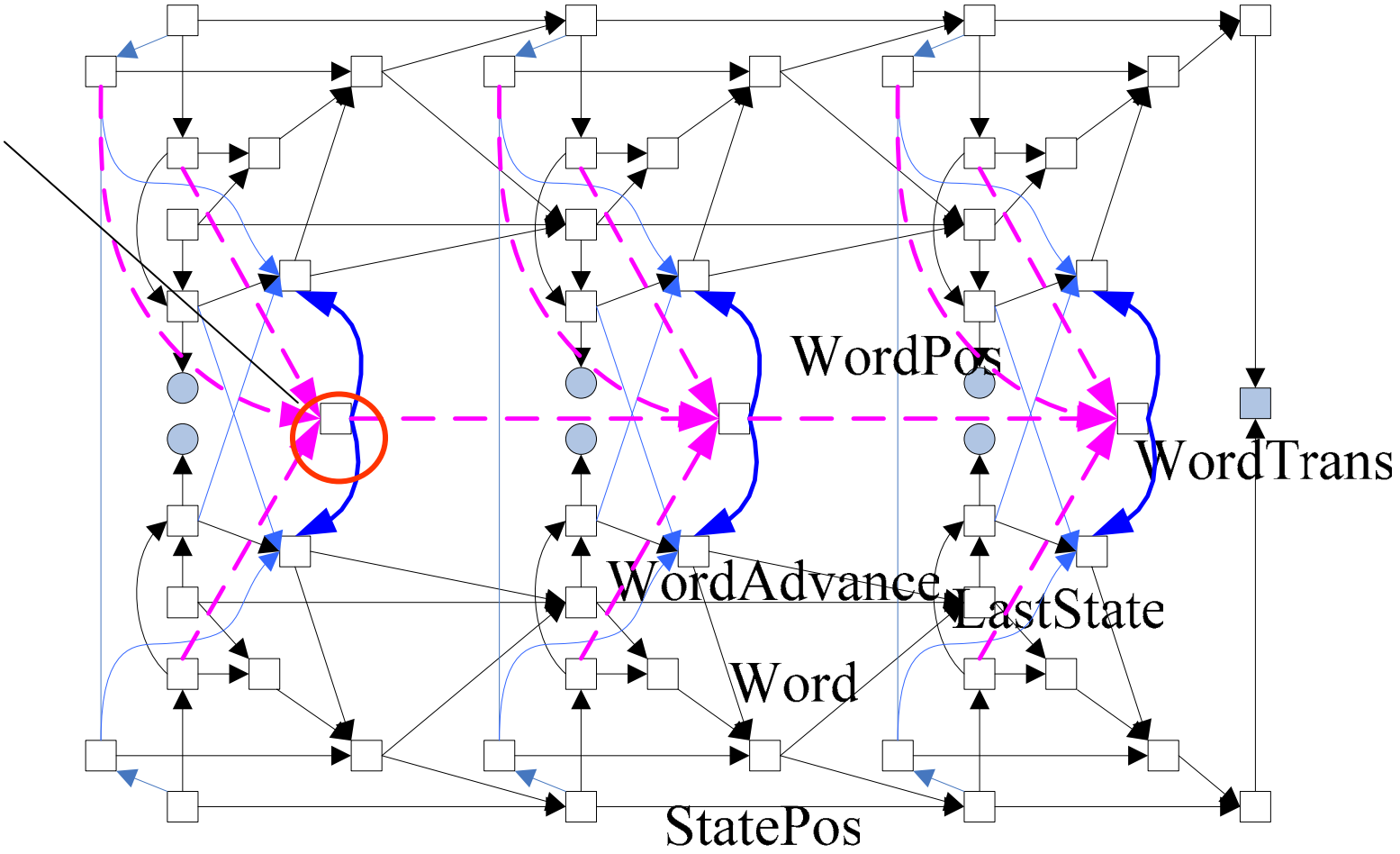
- DBN based Audio-Visual Correlative Model

AVCM

- Models the temporal relationship at word boundaries
  - For advanced stream
    - In *first state* of the new word: holds state-transition until the lagging stream catches up to the new word
- Limits the temporal asynchrony *within only one state*
- Ignores the *explicit temporal relationship* between two streams

Durational-AVCM

- Recording the time when the two streams are in asynchrony
- Controlling the transition actions of “State Trans” and further “Word Trans”



Word Advance Duration (WD)

$$P\left(WD_t = d' \mid WD_{t-1} = d, WA_t = i, W_t^A = j, W_t^V = k\right)$$

$$= \begin{cases} 1 \quad if \quad d' = -1 & and \quad i = 0 \\ 1 \quad if \quad d' = MaxAdvanceA(j,k) & and \quad i = 1 \quad and \quad d = -1 \\ 1 \quad if \quad d' = MaxAdvanceV(j,k) & and \quad i = -1 \quad and \quad d = -1 \\ 1 \quad if \quad d' = d - 1 & and \quad i \neq 0 \quad and \quad d > 0 \\ 1 \quad if \quad d' = 0 & and \quad i \neq 0 \quad and \quad d = 0 \\ 0 \quad otherwise \end{cases}$$

- *MaxAdvanceA (Word_A, Word_V)*
  - The maximum advanced time when the audio stream is in advance, the current word of audio stream and video stream is *Word_A* and *Word_V* respectively.
- *MaxAdvanceA*,*MaxAdvanceV*
  - Learnt from the training data
- *Explicitly model* the asynchrony for different word combinations

$$P\left(T_t^s = b \mid C_t^s = i, C_t^{\neg s} = j, WA_t = k, WD_t = d\right)$$

$$= \begin{cases} 1 & if \quad b=0 \quad and \quad s=A \quad and \quad k=1 \quad and \quad d=0 \\ 1 & if \quad b=0 \quad and \quad s=V \quad and \quad k=-1 \quad and \quad d=0 \\ A^s_{ii,j} & if \quad b=0 \quad and \quad s=A \quad and \quad k=1 \quad and \quad d>0 \\ 1-A^s_{ii,j} & if \quad b=1 \quad and \quad s=A \quad and \quad k=1 \quad and \quad d>0 \\ A^s_{ii,j} & if \quad b=0 \quad and \quad s=V \quad and \quad k=-1 \quad and \quad d>0 \\ 1-A^s_{ii,j} & if \quad b=1 \quad and \quad s=V \quad and \quad k=-1 \quad and \quad d>0 \\ A^s_{ii,j} & if \quad b=0 \quad and \quad s=A \quad and \quad k\neq 1 \\ 1-A^s_{ii,j} & if \quad b=1 \quad and \quad s=A \quad and \quad k\neq 1 \\ A^s_{ii,j} & if \quad b=0 \quad and \quad s=V \quad and \quad k\neq -1 \\ 1-A^s_{ii,j} & if \quad b=1 \quad and \quad s=V \quad and \quad k\neq -1 \\ 0 & otherwise \end{cases}$$

$$P\left(WT_t^s = b \mid LS_t^s = f, T_t^s = g, WA_t = i\right)$$

$$= \begin{cases} 1 & if \quad b=0 \quad and \quad f=0 \\ 1 & if \quad b=0 \quad and \quad f=1 \quad and \quad g=0 \\ 1 & if \quad b=1 \quad and \quad f=1 \quad and \quad g=1 \quad and \quad i=-1 \quad and \quad s=A \\ 1 & if \quad b=1 \quad and \quad f=1 \quad and \quad g=1 \quad and \quad i=1 \quad and \quad s=V \\ 1 & if \quad b=1 \quad and \quad f=1 \quad and \quad g=1 \quad and \quad i=0 \\ 1 & if \quad b=0 \quad and \quad f=1 \quad and \quad g=1 \quad and \quad i=-1 \quad and \quad s=V \\ 1 & if \quad b=0 \quad and \quad f=1 \quad and \quad g=1 \quad and \quad i=1 \quad and \quad s=A \\ 0 & otherwise \end{cases}$$

- If WD exceeds the *maximum advanced time*, the advanced stream should wait until the other stream catches up.
- Otherwise, the stream just proceeds at its own pace.

Experiments

- Comparison between the three models using text-prompt speaker identification experiments
  - Coupled HMM (CHMM)
  - AVCM
  - Durational-AVCM
- All the models are implemented as dynamic Bayesian networks (DBNs) with GMTK* toolkit

Bilmes, J., Zweig, G. 2002: The graphical models toolkit: An open source software system for speech and time-series processing.

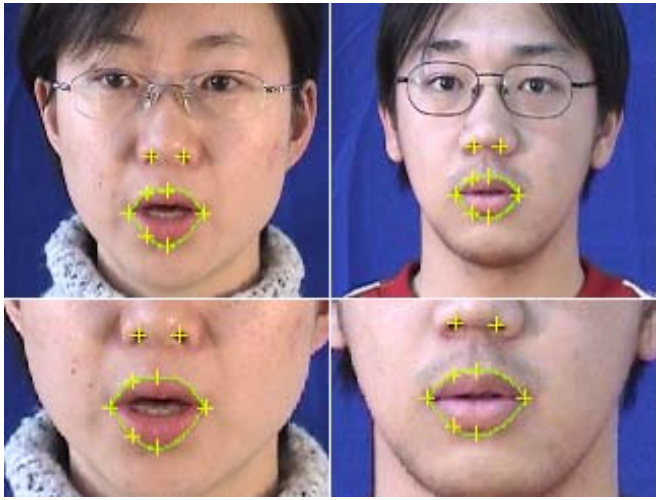
Experimental Corpora (1)

- CMU audio-visual bimodal database
  - 10 subjects (7 male, 3 female)
  - Speaking 78 isolated English words
  - Repeated 10 times

Chen, T. 2001: Audiovisual speech processing

Experimental Corpora (2)

- Homegrown audio-visual bimodal database
  - 60 subjects (38 male, 22 female, aged from 20 to 65)
  - Speaking 30 connect-digit words (digit length differs from 2 to 6) in Chinese
  - Repeated 3 times at intervals of 1 month



Front-ends and Models

- Audio Front-end
  - 13 MFCCs + 1 energy + delta
  - Frame size: 25ms, and frame rate: 11ms
- Visual Front-end
  - 1 mouth width + 1 upper lip height + 1 lower lip height + delta *
  - Frame rate: 30fps, up-sampled to 90fps (11ms)
- Model parameters
  - Audio sub-model: 5 states, GMM with 3 mixtures
  - Video sub-model: 3 states, GMM with 3 mixtures

Chen, T. 2001: Audiovisual speech processing

Experimental Setup

- Cross validation
  - 90% of all the data used for training
  - 10% of all the data used for testing
  - Repeated until all the data have been covered in the testing set
- Training: clean speech
  - SNR at 30dB
- Testing: noisy speech
  - Additive Gaussian white noise
  - SNR at 0, 10, 20, 30dB

Experimental Results (1)

- Speaker identification on CMU database

audio signal-to-noise ratio (SNR)	30dB	20dB	10dB	0dB
coupled HMM (CHMM)	100	88	79	60
AVCM	100	92	79	65
Durational-AVCM	100	93.5	82	69

Experimental Results (2)

- Speaker identification on homegrown database

audio signal-to-noise ratio (SNR)	30dB	20dB	10dB	0dB
coupled HMM (CHMM)	100	85	77	57
AVCM	100	89	78	61
Durational-AVCM	100	91	80	66

Thank you very much!

Experimental Results (1)

- Speaker identification on CMU database

audio signal-to-noise ratio (SNR)	30dB	20dB	10dB	0dB
video only	77	77	77	77
audio only	100	64	22	17
feature level fusion	99	85	30	20
decision level fusion	100	86	78	78
coupled HMM (CHMM)	100	88	79	60
AVCM	100	92	79	65
Durational-AVCM	100	93.5	82	69

Experimental Results (2)

- Speaker identification on homegrown database

audio signal-to-noise ratio (SNR)	30dB	20dB	10dB	0dB
video only	74	74	74	74
audio only	99	59	20	15
feature level fusion	99	81	26	18
decision level fusion	100	83	76	75
coupled HMM (CHMM)	100	85	77	57
AVCM	100	89	78	61
Durational-AVCM	100	91	80	66

Experiments

- Comparison with other models
  - Audio-only
  - Video-only
  - Feature level fusion
  - Decision level fusion
  - Coupled HMM (CHMM)
  - AVCM
  - Durational-AVCM
- All the models are implemented as dynamic Bayesian networks (DBNs) with GMTK* toolkit

Bilmes, J., Zweig, G. 2002: The graphical models toolkit: An open source software system for speech and time-series processing.