Asynchrony in Audio-Visual Speech Recognition (AVSR)

Virginia Estellers, Mihai Gurban, Jean-Philippe Thiran

Signal Processing Laboratory - LTS5

31 August 2009

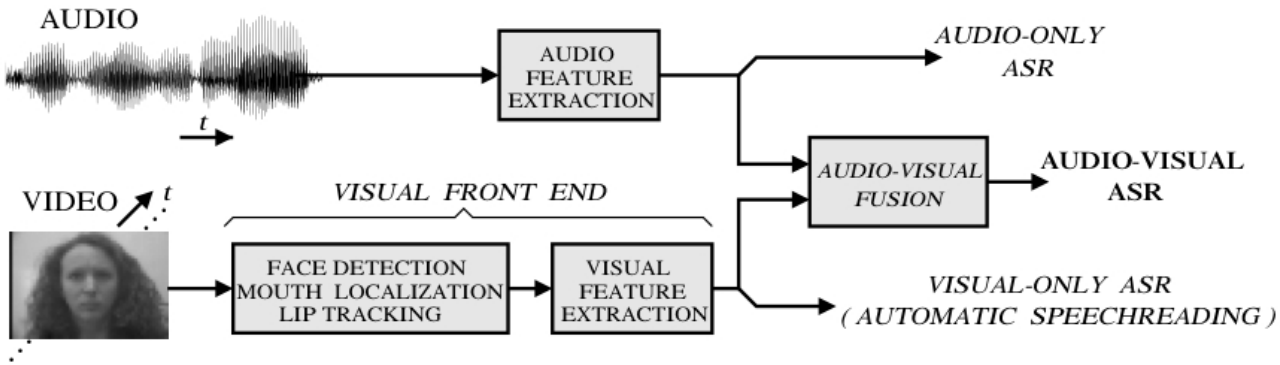
Audio-Visual Speech Recognition and asynchrony

Audio-Visual Speech Recognition

- Improve the performance of audio-only systems with the visual modality, specially in the presence of noise.

- Lip's movements and the actual production of sound are delayed between 10-100 ms.

Structure of a typical AVSR system



- Audio front-end: from audio signal to audio features designed for speech recognition.
- Visual front-end: from video signal to video features designed for speech recognition.
- Audio and video fusion.
- Pattern classifier: HMM and viterbi decoding.

Statistical models

Statistical audio models

Statistical modeling in speech recognition for the pattern classifier:

- Phoneme is the basic audio speech unit. Concatenation of phoneme models to build up words.
- Hidden Markov Models (HMMs) are the statistical models normaly used in audio.
- HMMs are a particular case of a more general model, Dynammic Bayesian Networks (DBNs).

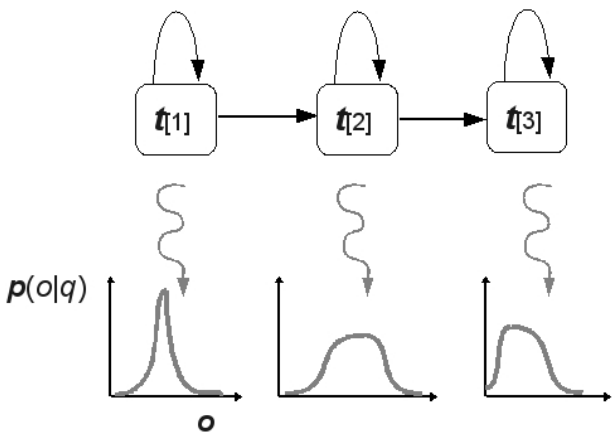


Figure: 3-state HMM

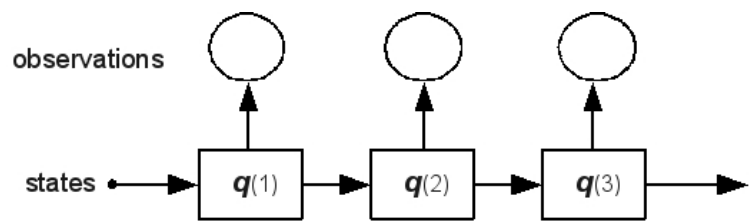


Figure: DBN representation

Extensions to audio-visual models

- Independent streams of observed features (no weights).

$$p(o_{AV}|q) = p(o_A|q)p(o_V|q)$$

- Different possibilities for state evolution of the audio and video stream account for different synchrony models:
  - MSHMM: multistream HMM.
  - IHMM: independent HMMs.
  - CHMM: coupled HMMs.

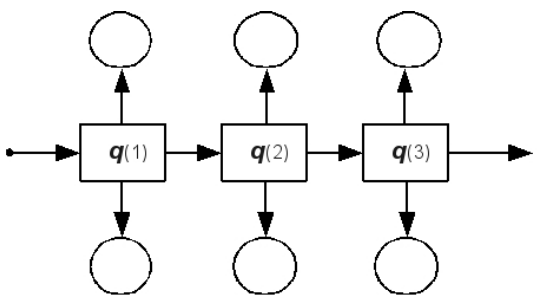


Figure: MSHMM

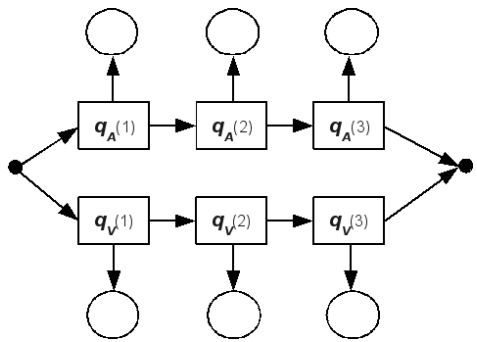


Figure: IHMM

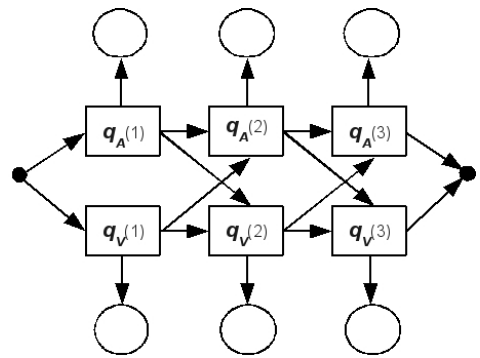


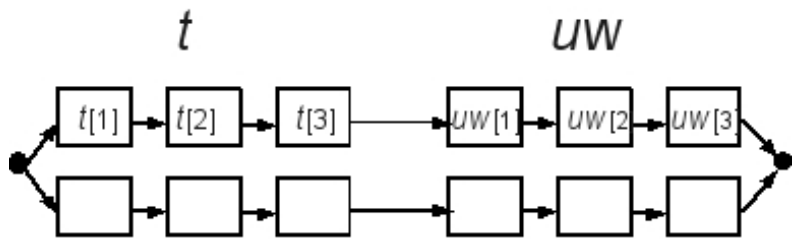
Figure: CHMM

Proposed audio-visual model

Experiments show that the audio and visual asynchrony can reach 120ms, longer than the mean duration of a phoneme.

We propose a model imposing synchrony at word boundaries.

- DBN model based on the concatenation of independent audio and video phoneme models.



- No word-DBN models. The 'pieces' of the model are independent audio and video HMMs concatenated to form words, pHMMs.

Experiments on modeling

Experimental set-up I

Experimental set-up to test different models

- Speaker independent experiments with 6-fold cross-validation, results in terms of word accuracy.
- Audio-Visual CUAVE database, consisting of 36 speakers repeating the digits in front of a camera.
- MFCC and DCT plus derivatives as audio and video features.
- In testing, we artificially add babble noise to the audio with Signal to Noise Ratios from clean to 0 dB.

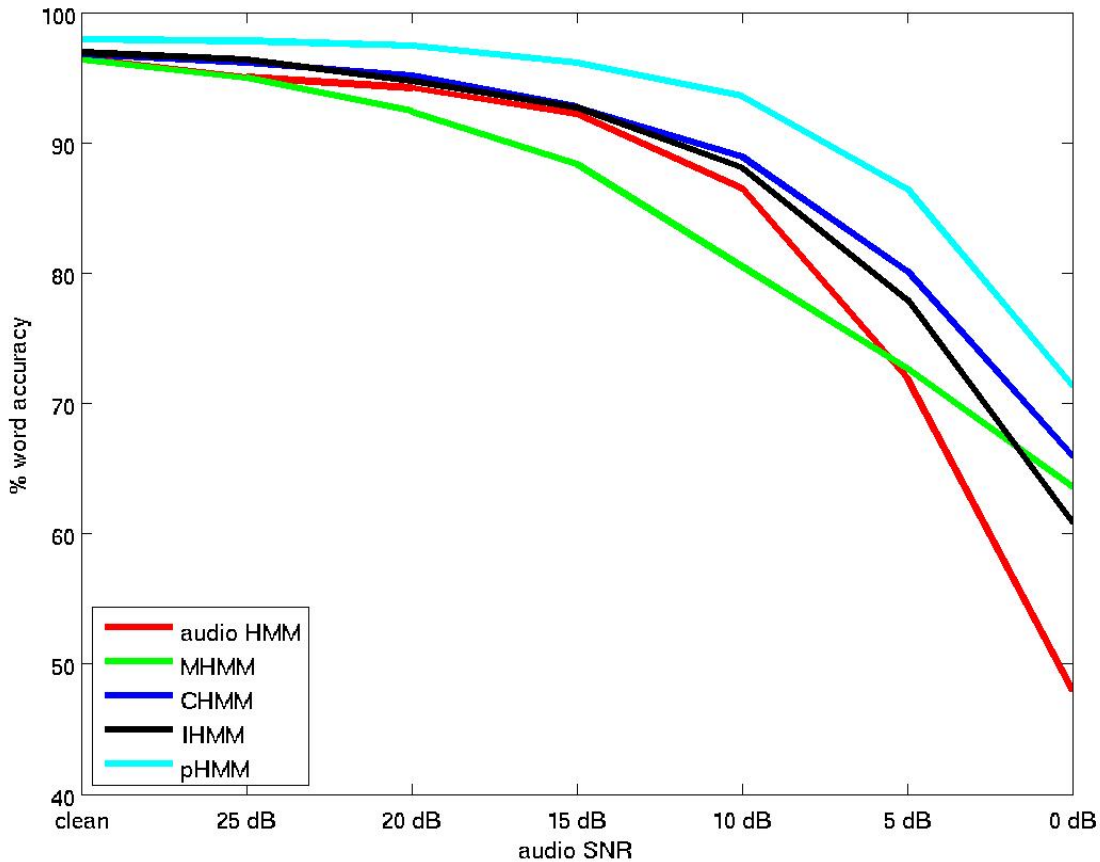
Experimental set-up II

Why testing those models?

- MSHMM: state synchrony.
- IHMM: synchrony at phone boundaries with independent time evolution of the streams.
- CHMM: synchrony at phone boundaries with non-independent time evolution of the streams.
- pHMM: synchrony at word boundaries based on phoneme models.

Results on modeling

Results audio-visual modeling



Conclusions on modeling

- Asynchrony goes beyond phoneme boundaries. We need asynchronous DBN models BUT DBNs longer to train and not much work done on stream weighting or reliability estimation.
- State evolution is coupled, presumably also in word models BUT working with word models is too constraining and context dependent phonemes need to much training data.

Feature processing to avoid asynchrony

Signal processing of the features

We want to work with:

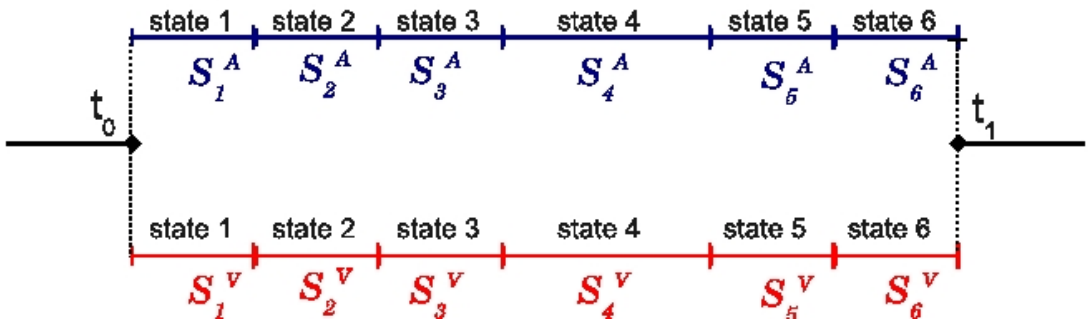
- Phoneme models allowing the large vocabulary tasks.
- Synchronous MSHMM instead of a more complex DBN model

As the synchronism is defined by the temporal evolution of the streams, we propose to process the features $o_A(t), o_V(t)$ to obtain the new streams.

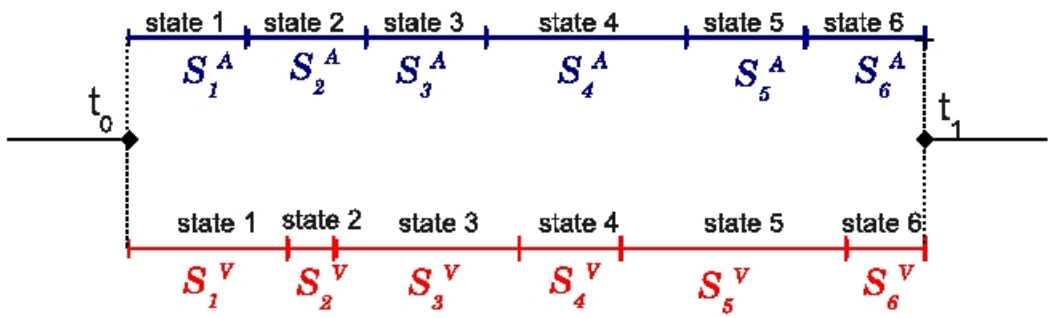
Why the asynchronous pHMM outperforms MSHMM?

Assume we detect a word between in the time interval (t_0, t_1).

- MSHMMs forces the audio and video states to evolve synchronously in time.



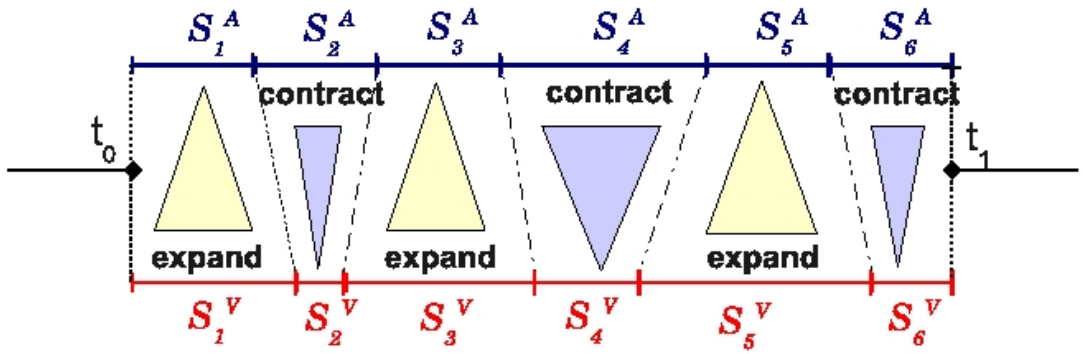
- pHMMs allow asynchronous evolution of stream states within word boundaries.



First idea

The previous representation of the models suggests to

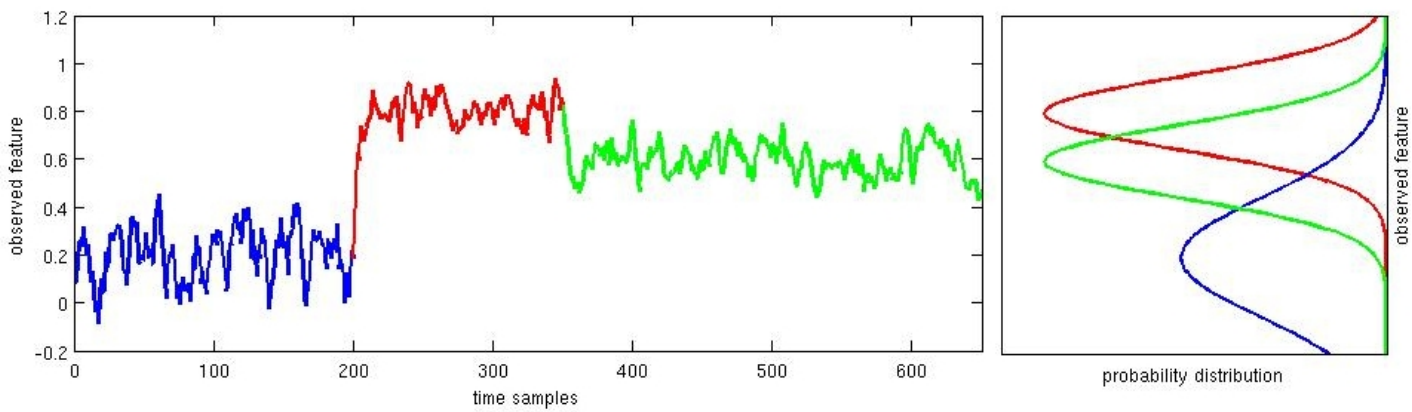
- Piece-wisely expand and contract the time.



- Limitations
  - Audio and video should agree on word boundaries.
  - The number of state partitions of both streams should be the same.

Proposed technique-I

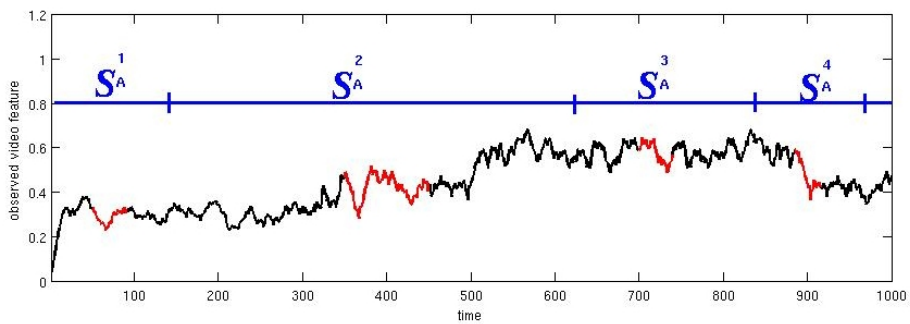
- Expand and contract the time within word boundaries t_0, t_1.
- Use the audio to detect word boundaries and adapt to it the time evolution of the video.
- Understand how the state partition of the interval is related to the features: time clustering.
  - Within any S^i the model is in the same state.
  - The HMM models the observed features of each state with gaussian mixtures, that is clusters.



Proposed technique-II

The time clustering of the new stream similar to the audio but keep the information added by the video features.

- Construct the new stream from o_V samples and the clustering in time of the audio $\mathcal{S}^A$.
- For each $S_i^A \in \mathcal{S}^A$, take the necessary samples of o_V from the central region of S_i^A and assume they come from a uniform time sampling.

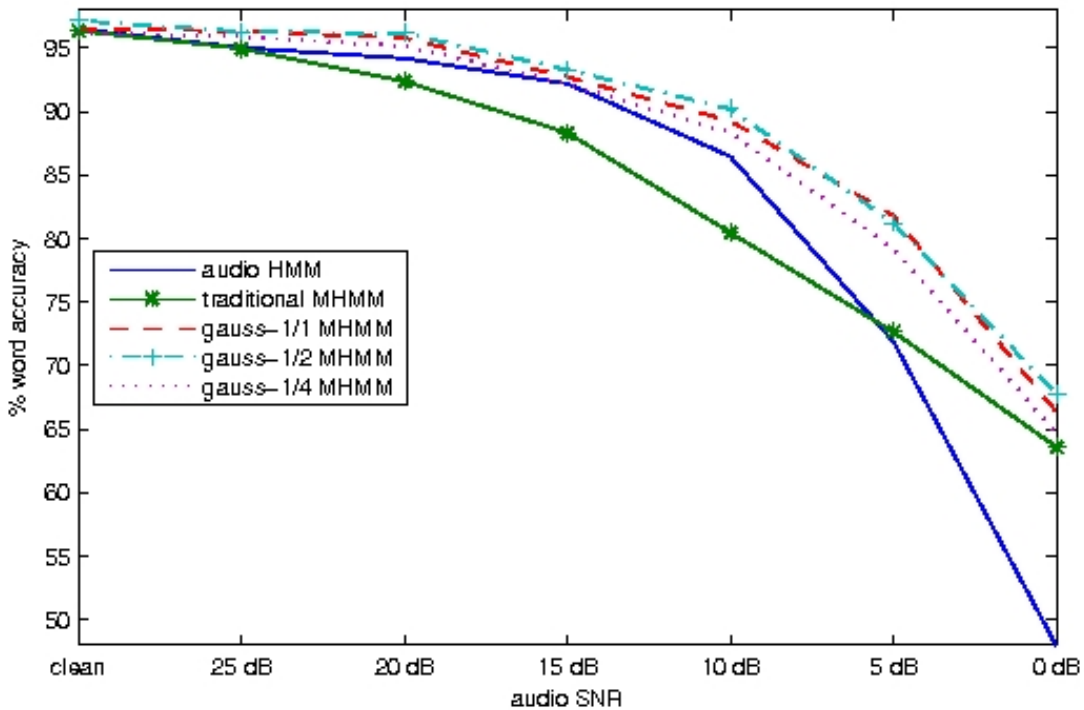


- Upsampling $o_V(t)$ and taking the necessary samples from the ones in the center of each S_i^A.

Experiments on feature processing

Experiments on feature processing I

Word accuracy with the original and the processed streams with different variants of the proposed method



Experiments on feature processing II

Proposed method in a typical MSHMM set-up: weighted likelihood combination and HTK software.

SNR	audio	MSHMM	
		original	processed
clean	99.2	99.28	99.37
25 db	99.09	98.94	99.34
20 db	97.96	97.77	98.29
15 db	95.44	94.97	95.91
10 db	89.13	87.71	90.18
5 db	78.2	72.56	78.76
0 db	59.55	53.25	60.95

Conclusions

- Imposing synchronism at state level with phoneme HMM models can lead to worse audio-visual results than an audio-only system.
- Asynchronous models at word level or an extra processing step are needed.
- It is possible to diminish the asynchrony wrapping the video stream to the audio variations defined by an audio-only system.

Thanks for your attention
Questions?