Audio-Visual Speech Recognition

Mathilde Bateson & Vivien Cabannes

Problem

Recognize audio-visual speeches.

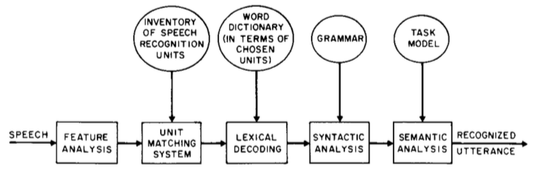
Motivations

- Graphical Model Application
- Stream Collaboration

Dataset

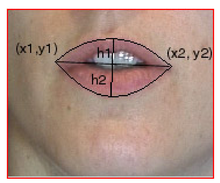
- Word samples
- Uncluttered conditions
- Small dictionary

Speech Recognizer



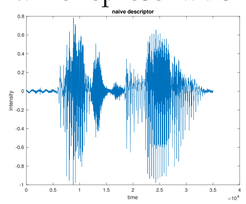
Descriptors

Video Descriptors

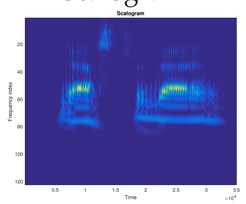


Audio Descriptors

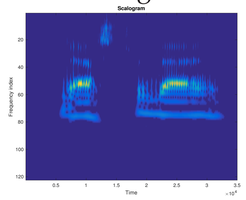
Naive representation



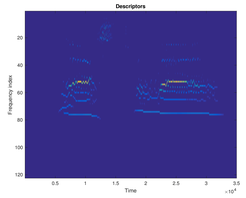
Scalogram



Removing noise



Low dimensional

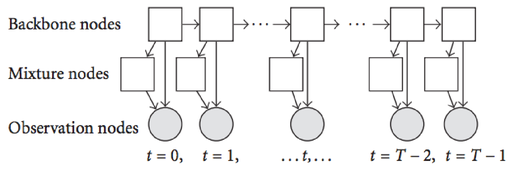


Model

Generative model

- Output: $\arg\max_w p(w|o)$
- Modeling $o|w$: hidden Markov model

Hidden Markov model



- Hidden state: stage in the word
  - $q_{t+1} \in \{q_t, q_t + 1\}$ *(left-to-right)*
- Modeling $o|q$: Gaussian mixture model
  - $o_t|(q_t = i, c_t = m) \sim \mathcal{N}(\mu_{i,m}, U_{i,m})$

Fitting parameters

- Maximum likelihood estimator
- Expectation-Maximization relaxation

Difficulties

- Exponential vanishing
- Gaussian without density

Our Results

Experimental results

Predictions precision	video	audio
train	30%	too slow
test	10%	too slow
random	1.5%	too slow

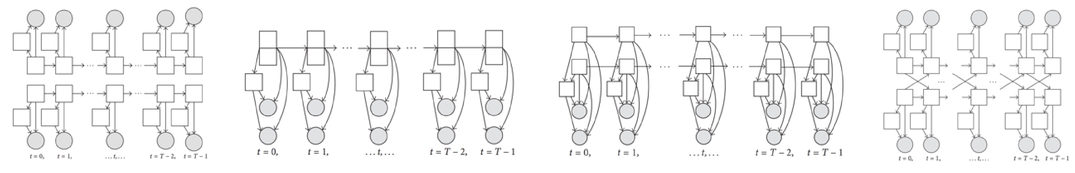
Stream Collaboration

Dynamic Bayesian Networks

- Asynchrony between the audio and visual modalities is intrinsic to human speech (ex. the movement of the lips precedes or follows the actual production of sound).
- Allow asynchrony between audio and visual streams (and defines some synchronization points)
- While preserving the natural dependency over time of the acoustic and visual features of speech.

Ideas & Models

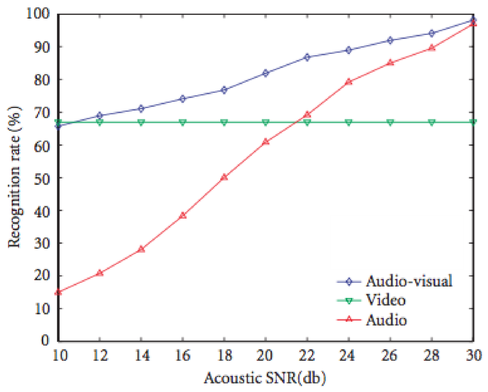
Stream collaboration	IHMM	PHMM	FHMM	CHMM
Transition probabilities	ind	joint	ind	joint
Observation likelihood	ind	joint	joint	ind

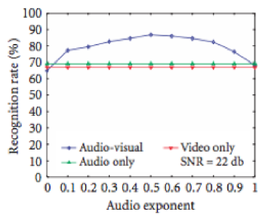


Independent HMM — Product HMM — Factorial HMM — Coupled HMM

Litterature results

Lip reading: robustness to noise





Audio exponent

How to weight stream?
Confidence exponents:

$$p(o_a, o_v) \propto p(o_a)^{\lambda} p(o_v)^{1-\lambda}$$

Recognition rate

SNR (db)	30	20	10
MS-HMM (%)	98.6	79.2	67.8
F-HMM (%)	97.8	78.6	66.8
C-HMM (%)	98.1	81.9	65.7

Future Directions & Conclusion

Algorithm acceleration

- E-step via approximate inference
- Reducing descriptors time frames

Breaking the framework

- Use silent detection
- Toward discriminative models

This project attempted to cast audio-visual speech recognition towards graphical modeling. It can be done relatively smoothly, and allows efficient stream collaboration.

References & Source code

[1] L. Rabiner. A Tutorial on Hidden Markov Models and selected Applications in Speech Recognition. In *Proceedings of the IEEE 1999*

[2] A. Nefian et al. Dynamic Bayesian Networks for Audio-Visual Speech Recognition. In *EURASIP Journal on Applied Signal Processing 2002*

The full code is provided on GitHub at:

`https://github.com/VivienCabannes/`