Multi-Channel Modeling and Integration with Applications in Speech and Multimodal Processing

Prof. Hervé Bourlard
bourlard@idiap.ch
IDIAP Research Institute, Martigny, Switzerland
http://www.idiap.ch/
Swiss Federal Institute of Technology, Lausanne (EPFL)
Outline

1. Multi-channel modeling and integration
   • Problems
   • State-of-the art solutions

2. Multi-stream and Asynchronous HMM, with application in:
   • Audio-visual speech recognition

3. Multi-layered HMMs, with applications in:
   • Speech recognition
   • Multi-modal meeting (group action) modeling
Notation

- HMMs:
  - Powerful models used to handle sequences
  - State-of-the-art models for speech recognition
  - Efficient training and decoding algorithm

- Notation:
  - $O^k = (o^k_1, \ldots, o^k_{N_k})$ is the $k$-th observation sequence
  - $O = [O^1, \ldots, O^k, \ldots, O^K]$
  - $K$ = number of channels, and $N_k$ = number of feature vectors (of dimension $d_k$) in $k$-th observation stream
  - $Q^k = \{q^k_1, \ldots, q^k_{N_k}\}$ is the set of (piecewise stationary) states modeling the $k$-th observation stream.
Joint Multi-Channel Modeling

- Training: parameters that maximise the likelihood of $L$ multi-stream sequences:

$$\theta_j^* = \arg \max_{\theta_j} \prod_{l=1}^{L} p(O_l|\theta_j).$$  \hspace{1cm} (1)

- Decoding: given a multi-stream observation sequence $O$, find the sequence $W$ of “events” (classes) maximizing the joint likelihood $p(O|W, \theta_j^*)$.

- The most well-known solution to efficiently model such a distribution is to use Hidden Markov Models (HMMs).
Multi-Channel Modeling: Problems (2)

- Examples of multiple feature sequences:
  - Extracted from the same signal, e.g., multi-rate processing, different frequency bands
  - Extracted from different modes, but associated with the same “event” (class), typically in multimodal processing (audio-visual speech recognition, multimodal meeting actions, etc).

- Asynchrony, with different frame rates \((N_k)\) and/or variable number of streams \((K)\) over time.
  - Piecewise stationarity assumption \(\Rightarrow\) exponentially complex models

- Complexity: More parameters, difficult to model/extract higher-level information
### Piecewise Stationarity Assumption

<table>
<thead>
<tr>
<th>Stream 1</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>3 d1-dimensional states</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stream 2</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>3 d2-dimensional states</td>
</tr>
<tr>
<td>Naive Integration</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Joint / Asynchronous</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>3 (d1+d2)-dimensional states</td>
</tr>
</tbody>
</table>

Multi-Stream: 3 d1- + 3 d2-dimensional states
Asynchronous: 3 (d1+d2)-dimensional states
Multi-Channel modeling: State-of-the-art

- HMM/DBN variants based on different assumptions of stream inter-dependencies, such as:
  - feature-level correlation,
  - state-level correlation, and
  - whether streams evolve synchronously or asynchronously.

- HMM/DBN variants include:
  - Early integration HMM
  - Late integration HMM
  - Coupled HMMs [Brand, 1997]
  - Multi-Rate HMMs [Cetin, 2004]
  - Multi-Stream HMMs [Bourlard et al, 1996]
  - Asynchronous HMMs [Bengio, 2002]
2) Multi-Stream and Asynchronous HMMs

Multi-stream HMM (Bourlard et al, 1996)

(1) MSHMM

\[ p(o_t^k | q_t^k) = \sum_{g=1}^{G} w_g^k \mathcal{N}(o_t^k - \mu_g^k, \sigma_g^k), \quad \forall k \]

(2) Product MSHMM

\[ p(o_t | q_t) = \prod_{k=1}^{K} p(o_t^k | q_t^k) \quad \Rightarrow \quad \text{Viterbi (1) \equiv Viterbi (2)} \]

more complex recombinations

\[ \Rightarrow \quad \text{Viterbi (1) \neq Viterbi (2)} \]
Multi-Stream and Asynchronous HMM

Multi-stream HMM:

- Assumes **partial independence** between feature streams
- While **anchor (synchronization) points** may occur anywhere, complexity becomes exponential in the number of streams.
- Thus, in practice, recombination is often done at the HMM state level.
  - Assuming that streams evolve frame synchronously.
  - Doesn’t really model the joint likelihood of feature streams.
  - **Significant success in several applications** such as noise robust speech recognition, multi-stream speech recognition, audio-visual speech recognition, etc.
Asynchronous HMMs [Bengio, 2002]

Maximise the joint likelihood $P(O^1, \ldots, O^K \mid \Theta)$ of the observation streams given the model $\Theta$

- Stream $O^k$ can emit an empty symbol ($\varepsilon$) with probability $\alpha^k$

\[
p(o^k_t \mid q_t, g) = \begin{cases} 
(1 - \alpha^k) \mathcal{N}(o^k_t - \mu^k_g, \sigma^k_g) & \text{if } o^k_t \neq \varepsilon \\
\alpha^k & \text{otherwise}
\end{cases}
\]

\[
p(o_t \mid q_t) = \sum_{g=1}^{G} w_g \prod_{k=1}^{K} p(o^k_t \mid q_t, g)
\]

Can deal with observation streams of different lengths.
Asynchronous HMMs (2)

- Enables **re-synchronization** of observation streams.
- One HMM: maximizes the **joint** likelihood of all streams, modeling **correlation** in the feature space.

\[
p(q_t | q_{t-1}) \quad P(\text{emit } o^2_s | q_t) \quad p(o^1_t, o^2_s | q_t) \quad p(o^1_t | q_t)
\]

- Forward-backward and EM training possible, involving **two hidden variables** (instead of one in standard HMMs).
Likelihood and Forward Recurrence

• Assuming two sequences $x_{1:T} = \{x_1, \ldots, x_T\}$ and $y_{1:S} = \{y_1, \ldots, y_S\}$, with $S < T$

• Forward Loop for HMMs:

$$\alpha(i, t) = p(q_t=i, x_t) = p(x_t|q_t=i) \sum_{j=1}^{N} P(q_t=i|q_{t-1}=j) \alpha(j, t-1)$$

• Forward Loop for AHMMs:

$$\alpha(i, s, t) = \epsilon(i, t)p(x_t, y_s|x_t=i) \sum_{j=1}^{N} P(q_t=i|q_{t-1}=j) \alpha(j, s-1, t-1)$$

$$+ (1 - \epsilon(i, t)) p(x_t|q_t=i) \sum_{j=1}^{N} P(q_t=i|q_{t-1}=j) \alpha(j, s, t-1)$$

• $\epsilon(i, t)$ is the probability that the system emits on sequence $y$ at time $t$ while in state $i$. 

**********
Multi-Stream vs Asynchronous HMM

- Multi-stream HMM:
  - Allows for different number of HMM states for different observation streams (e.g. allowing different models for slow and fast streams).
  - Doesn’t explicitly model correlation between streams, but does it implicitly through anchor points.
  - Full asynchrony between anchor points.

- Asynchronous HMM:
  - Assumes a single HMM, thus the same number of HMM states for each observation stream.
  - Models correlation between streams directly at the observation level, if $p(x, y|q)$ doesn’t make independence assumptions.
  - If so, actually estimates $P(X, Y|\Theta)$. 
Database

- Audio-Visual M2VTS database: 37 subjects, 185 recordings, (French) digit sequences; 5-fold cross-validation

- Audio:
  - 16 MFCC coefficients and their first derivatives, plus the derivative of the log energy: 33 features, 100Hz frame rate

- Video:
  - 12 shape features and 12 intensity features, plus first derivatives: 48 features, 25Hz frame rate
Audio-Visual Speech Recognition

Results

![Graph showing Word Error Rate against noise level for different HMM models: audio HMM, audio+video HMM, audio+video AHMM, and video HMM. The graph plots noise level on the x-axis (0db, 5db, 10db, 15db) and Word Error Rate on the y-axis. The audio HMM shows the highest error rate, while the video HMM has the lowest error rate at all noise levels.](image-url)
3) Multi-layered HMMs

- Posterior probabilities: used as **local measure/classifier** or as **transformed features**.

- **How to improve those posterior estimates:**
  1. Using contextual information, AHMM, etc
  2. Using any possible a priori knowledge (e.g., specific HMM topological constraints)
  3. Replacing \( p(q_t^i|x_t) \) by \( \gamma(i, t) = p(q_t^i|X_t, M) \), where:
     - \( X_t \) represents \( x_t \) in context (e.g., the whole utterance \( X \) in the case of ASR)
     - \( M \) is the (HMM?) model representing the prior information

- Yielding “optimal” multi-layered (hierarchical) HMM structures.
• Use MLP-generated posteriors as acoustic features
• MPL usually accommodates some acoustic context $X_{t+c}^{t-c}$, typically of 9 frames
• MLP can be used to merge features and generate a reduced set of “optimal” features (resulting of NLDA)
• Those features can be transformed (usually through KLT) to reduce the space dimension and/or to make them better suited to HMM-GMM
• Successful example: “Tandem”, integrated in full SRI CTS system
Posteriors as features: Tandem

\[ p(q^i_t | x_t) \]

MLP outputs prior to final nonlinearity
### Likelihood-based systems

- **Alpha recursion:**
  \[
  \alpha(i, t) = p(x_t^i, q_t^i) = p(x_t | q_t^i) \sum_j p(q_t^i | q_{t-1}^j) \alpha(j, t - 1)
  \]

- **Beta recursion:**
  \[
  \beta(i, t) = p(x_{t+1}^T | q_t^i) = \sum_j p(x_{t+1} | q_{t+1}^j) p(q_{t+1}^j | q_t^i) \beta(j, t + 1)
  \]

- **State Gamma:**
  \[
  \gamma(i, t) = p(q_t^i | x_1^T) = \frac{\alpha(i, t) \beta(i, t)}{\sum_j \alpha(j, T)}
  \]
HMM State Gammas as Features (2)

Posterior-based systems

• Scaled-Alpha recursion:

\[
\alpha^s(i, t) = \frac{p(x^1_t, q^i_t)}{\prod_{\tau=1}^{t} p(x_\tau)} = \frac{p(q^i_t | x_t)}{p(q^i_t)} \sum_j p(q^j_t | q^j_{t-1}) \alpha^s(j, t-1)
\]

• Scaled-Beta recursion:

\[
\beta^s(i, t) = \frac{p(x^T_{t+1} | q^i_t)}{\prod_{\tau=t+1}^{T} p(x_\tau)} = \sum_j \frac{p(q^j_{t+1} | x_{t+1})}{p(q^j_{t+1})} p(q^j_{t+1} | q^j_t \beta^s(j, t+1)
\]

• State Gamma:

\[
\gamma^s(i, t) = p(q^i_t | x^T_1) = \frac{\alpha^s(i, t) \beta^s(i, t)}{\sum_j \alpha^s(j, T)}
\]
“Gamma” posteriors as features

\[ p(q_t^i | x_t) \]

\[ p(q_t^i) \]
Multi-layered HMMs for ASR: Numbers’95

Task: Numbers’95

<table>
<thead>
<tr>
<th>Features</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP</td>
<td>6.9%</td>
</tr>
<tr>
<td>Tandem phone posteriors (alone)</td>
<td>4.9%</td>
</tr>
<tr>
<td>Gamma phone posteriors (alone)</td>
<td>4.6%</td>
</tr>
</tbody>
</table>

Table 1: Word error rate (WER) on the Numbers’95 task: 31 lexicon words, 9x39(PLP)-1200-24 MLP (resulting in 25 features after KLT), 80 CD HMM/GMM phone models, 24 dimensional Tandem features, 1206 test utterances.
DARPA CTS sub-task

Task: Conversational Telephone Speech

<table>
<thead>
<tr>
<th>Features</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP</td>
<td>44.3%</td>
</tr>
<tr>
<td>PLP+Tandem phone posteriors</td>
<td>42.5%</td>
</tr>
<tr>
<td>PLP+Gamma phone posteriors</td>
<td>41.7%</td>
</tr>
</tbody>
</table>

Table 2: Word error rate (WER) on the male part of a reduced vocabulary version of the DARPA Conversational Telephone Speech-to-text (CTS) task: 1,000 lexicon words, with multi-words and multi-pronunciations, 9x39(PLP)-1300-46 MLP (resulting in 25 features after KLT).
Multi-modal meeting (group action) modeling

IDIAP Instrumented Meeting Room

Why record a new corpus?
- Aim to address limitations of previous corpus
- Recordings will be natural
- Statistically more significant
- Cover a richer variety of research tasks
Multi-layered HMMs for meeting actions

- Event lexicon: \( V = \) (monologue1, monologue2, monologue3, monologue4, discussion, presentation, whiteboard-presentation, note-taking)

- A-V Features: 39-dimension observation vectors extracted from 12 audio and 3 visual channels, at 5 Hz.
# Meeting Actions: A-V features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Modality</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Audio</td>
<td>Visual</td>
</tr>
<tr>
<td>Seat speech activity</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>White-board speech activity</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Presentation speech activity</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Speech pitch</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Speech energy</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Speaking rate</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Head blob vertical centroid</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Hand blob horizontal centroid</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Hand blob eccentricity</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Hand blob angle</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Combined motion</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>White-board/presentation blob</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
# Recognizing Sequences of Meeting Actions (1)

<table>
<thead>
<tr>
<th>Method</th>
<th>Features</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-Layer</td>
<td>Visual Only</td>
<td>48.2</td>
</tr>
<tr>
<td></td>
<td>Audio Only</td>
<td>36.7</td>
</tr>
<tr>
<td></td>
<td>Audio Visual</td>
<td>23.7</td>
</tr>
<tr>
<td>Two-Layer</td>
<td>Visual Only</td>
<td>42.5</td>
</tr>
<tr>
<td></td>
<td>Audio Only</td>
<td>32.4</td>
</tr>
<tr>
<td></td>
<td>Audio Visual</td>
<td>16.6</td>
</tr>
<tr>
<td></td>
<td>Async HMM</td>
<td>15.1</td>
</tr>
</tbody>
</table>

AER = Action Error Rate

- A-V multi-stream better than early integration.
- Important to model correlation (interaction) between participants.
- Best results with multi-layered HMMs, using AHMM at the participant level (1st HMM layer).
Recognizing Sequences of Meeting Actions (2)
Conclusion

• Multi-channel processing is an interesting challenge, with important potential application in speech recognition, multimodal processing, etc.

• AHMM is an interesting solution towards principled modeling of multi-stream joint likelihood.

• MSHMM and AHMM gives good results in noisy speech recognition tasks, as well as multi-model group action modeling.

• Still not very efficient in space and time (but can be controlled)

• Hierarchical HMMs, and associated posterior combination, provide additional advantages and a new approach towards solving complex problem in a hierarchical approach.