Integration of Asynchronous Knowledge Sources in a Novel Speech Recognition Framework

Prof. Hugo Van hamme
K.U.Leuven, dept. ESAT, Belgium

ITWR workshop
Aalborg, Denmark
4 June, 2008

Outline

• Motivation
• Training / pattern discovery with NMF
  – Input representation
  – Weak supervision
  – Multiple information sources
• Recognition
• Experiments
• Conclusions
Motivation

- ASR systems are engineered
  - Complex statistical models
  - Engineered layers
    - Sentence$\leftrightarrow$Word$\leftrightarrow$Phoneme$\leftrightarrow$State$\leftrightarrow$Spectra
  - Speech variability does not always follow this structure
    - Pronunciation variation
    - Co-articulation
- Can we automatically discover acoustic patterns?
  - Method based on matrix factorization
  - Here:
    - Only the first two layers
    - Integrate knowledge sources at different time scales
  - ACORNS project (EC – FET – STREP)

Input representation

- Input = symbolic, discrete
- Uncertainty$\Rightarrow$graph
  - Edges: symbol with probability
  - Nodes: labeled by time
- Graph types:
  - Chain: e.g. VQ of spectra
  - Chain with multiple options, e.g. soft VQ
  - Lattice: e.g. phone lattice
- Time labels of nodes
  - Fixed step: e.g. VQ of spectrum
  - Variable: change-directed segments or phone decoder
Map to a vector space: HAC

- Directed edge co-occurrence
  - Edge A comes after edge B at distance τ with joint probability
  - Accumulate joint probability in a vector
    - For all pairs A,B
    - Over complete graph
  - Note: NOT symmetric in A and B
  - Weakly describes what comes in which order
- Shift-invariant representation

- Fixed-size vector representation of a graph
  \( w_i = \text{HAC}(\Omega_i) \)

\[ \Omega_i \]

A (bi)linear generative model

- Take \( R \) graphs \( \Omega_1, \ldots, \Omega_R \)
- Cascade them ⇒ \( \Omega \)
  \[ \text{HAC}(\Omega) = \text{HAC}(\Omega_1) + \ldots + \text{HAC}(\Omega_R) \]
  \[ w = [w_1 \ldots w_R] h = Wh \]

- Why \( \approx \) ?
  - Cross-graph terms

Conclusion:
- If \( w_1 \ldots w_R \) are HACs of “words”
- \( h \) indicates presence/absence of words
- \( v \) is HAC representation observed “utterance”
- \( v \approx [w_1 \ldots w_R] h = Wh \)
- \( \approx \) noisy observation of utterance graph
Word/pattern discovery by NMF

- Take $T$ utterances containing a vocabulary $W$
  - $V_i = W h_i$ ...
- In matrix form $V = WH$
  - $V = [v_1 ... v_T]$ and $H = [h_1 ... h_R]$

- Non-negative matrix factorization:
  - $V, W$: sum of co-occurrence probs $\geq 0$
  - $H$: word activations $\geq 0$
  - $\approx$ translated as MSE (Frobenius norm) or divergence

\[
D(V || WH) = \sum_{i,j} V_{ij} \log \frac{V_{ij}}{WH_{ij}} - V_{ij} + WH_{ij}
\]

- Algorithms available:
  - Multiplicative updates used here

Weakly supervised training

- Per utterance, word identities are given
  - Order not important
  - Not all words need to be tagged
  - Weak supervision
- $[G]_{it} = \#$ times word $i$ occurs in utterance $t$

- Generative model:
  - $[W]_{ir} = 1$ iff pattern $r$ is associated with tag $i$
  - $G = WH$ with same $H$ as acoustics
- Put together:

\[
\begin{bmatrix} G \\ V \end{bmatrix} \approx \begin{bmatrix} W_k \\ W \end{bmatrix} H
\]

- Conclusion: upper rows of estimated left factor ($W$) now gives relevance of a discovered pattern to tag $\Rightarrow$ **grounding** relation
Multiple information sources

- Handle different lags $\tau_1 \ldots \tau_Q$ simultaneously
  - All sources share the same $H$
  - No explicit alignment between sources required

$$V = \begin{bmatrix} G \\ V_{\tau_1} \\ \vdots \\ V_{\tau_Q} \end{bmatrix} \approx \begin{bmatrix} W_g \\ W \\ H \end{bmatrix}$$

Recognition

- estimate pattern activations $H$ in:

$$V_{\tau_i} \approx W H$$

- compute tag activations:

$$A = W_g H$$

- $A$ predicts $G$, i.e. the number of times a word occurs in a sentence
- Activation, not decoding
  - No segmentation
  - No order
Experiments (1)

- **TIDIGITS**
  - Digit strings length 2 through 7
  - 6159 train; 6214 test
  - 11 digits; \( R = 12 \)
- **Knowledge source 1: Phone lattices**
  - Generated with HMM triphone models
  - \( \tau = 1 \)
  - \( \mathcal{V} \) has 1936 \( (=4^2) \) rows
- **Knowledge source 2 through 4: VQ-spectra**
  - MFCC’s extracted every 10ms
  - Statics, velocity, acceleration
  - \( \tau = 2, 5 \) or 10
  - Codebook size 150, 150, 100 designed by K-means
  - \( \mathcal{V} \) has 55000 rows

Experiments (2)

<table>
<thead>
<tr>
<th>segmental</th>
<th>frame synchronous</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{V}_1 )</td>
<td>( \mathcal{V}_2 )</td>
<td>( \mathcal{V}_3 )</td>
</tr>
<tr>
<td>phone</td>
<td>VQ</td>
<td>VQ</td>
</tr>
<tr>
<td>( \tau = 1 )</td>
<td>( \tau = 2 )</td>
<td>( \tau = 5 )</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

- **WER:**
  - Which \( N \) different digits are in this utterance?
- **Top part:**
  - VQ only, no phone lattice
  - More sources give better accuracy
- **Middle part:**
  - Phone only comparable error rate
- **Bottom part**
  - All sources together give best error rate
Convergence

- NMF does not always yield same result
- Scatter plot of WER vs. final divergence during training
- Non gives an invalid solution
- 3 to 5 trials and selecting lowest divergence works

Conclusions

- HAC + NMF can work in an unsupervised or weakly supervised mode
- Activation-based recognition rather than decoding/search
- Straightforward integration of features at different time scales:
  - 10ms, 20ms, 50ms, 100ms and phone level
  - Copes well with the high dimensions
- No segmentation of words
  - Like humans
- Some challenges:
  - Cyclic repetitions give almost same HAC
  - No word order
    - “bag of words” representation
    - Order can be found (IS08)
  - Will it scale up?