WFSTs in ASR & Basics of Speech Production

Lecture 6

Instructor: Preethi Jyothi
Determinization/Minimization: Recap

- A (W)FST is **deterministic** if:
  - Unique start state
  - No two transitions from a state share the same input label
  - No epsilon input labels

- **Minimization** finds an equivalent deterministic FST with the least number of states (and transitions)

- For a deterministic weighted automaton, weight pushing + (unweighted) automata minimization leads to a minimal weighted automaton

- Guaranteed to yield a deterministic/minimized WFSA under some technical conditions characterising the automata (e.g. twins property) and the weight semiring (allowing for weight pushing)
Example: Dictionary WFST
WFST-based ASR System

- Acoustic Models
  - Triphones
- Context Transducer
  - Monophones
- Pronunciation Model
  - Words
- Language Model
  - Word Sequence
WFST-based ASR System

**Acoustic Models**

- **Triphones**

  - H

  - a/a_b
    - \( f_0: \text{a+a+b} \)
    - \( f_1: \varepsilon \)
    - \( f_2: \varepsilon \)
    - \( f_3: \varepsilon \)
    - \( f_4: \varepsilon \)
    - \( f_5: \varepsilon \)

  - b/a_b
    - \( f_0: \text{a+a+b} \)
    - \( f_2: \varepsilon \)
    - \( f_4: \varepsilon \)
    - \( f_6: \varepsilon \)

  - x/y_z
    - \( f_0: \text{a+a+b} \)
    - \( f_4: \varepsilon \)
    - \( f_6: \varepsilon \)

**Context Transducer**

- **Monophones**

  - \( f_0: \varepsilon \)
  - \( f_2: \varepsilon \)
  - \( f_4: \varepsilon \)

**Pronunciation Model**

- **Words**

**Language Model**

- **Word Sequence**

One 3-state HMM for each triphone

FST Union + Closure

Resulting FST

H
Figure 8: Context-dependent triphone transducer.

3.1. Transducer Combination

Consider the pronunciation lexicon in Figure 2b. Suppose we form the union of this transducer with the pronunciation transducers for the remaining words in the grammar $G$ of Figure 2a and then take its Kleene closure by connecting an $\epsilon$-transition from each final state to the initial state. The resulting pronunciation lexicon $L$ would pair any sequence of words from that vocabulary to their corresponding pronunciations. Thus, $L \circ G$ gives a transducer that maps from phones to word sequences restricted to $G$.

We used composition here to implement a context-independent substitution. However, a major advantage of transducers in speech recognition is that they generalize naturally the notion of context-independent substitution of a label to the context-dependent case. The transducer of Figure 8 does not correspond to a simple substitution, since it describes the mapping from context-independent phones to context-dependent triphonic models, denoted by $\text{phone/left-context/right-context}$.

Just two hypothetical phones $x$ and $y$ are shown for simplicity. Each state encodes the knowledge of the previous and next phones. State labels in the figure are pairs $(a, b)$ of the past $a$ and the future $b$, with $\epsilon$ representing the start or end of a phone sequence and $\ast$ an unspecified future. For instance, it is easy to see that the phone sequence $xyx$ is mapped by the transducer to $\epsilon\ast\epsilon\ast\epsilon$ via the unique state sequence $(\epsilon, \ast)(x, y)(y, x)(x, \epsilon)$.

More generally, when there are $n$ context-independent phones, this triphonic construction gives a transducer with $O(n^2)$ states and $O(n^3)$ transitions. At a tetraphonic construction would give a transducer with $O(n^3)$ states and $O(n^4)$ transitions. In real applications, context-dependency transducers will benefit significantly from determinization and minimization.

C-1: Arc labels: “monophone : phone / left-context_right-context”
WFST-based ASR System

Figure reproduced from “Weighted Finite State Transducers in Speech Recognition”, Mohri et al., 2002
WFST-based ASR System

Acoustic Indices → Acoustic Models → Triphones → Context Transducer → Monophones → Pronunciation Model → Language Model

Words → G → Word Sequence

0 → the → birds/0.404 → animals/1.789 → were/0.693 → is → boy/1.789 → are/0.693 → walking

The system consists of acoustic models, triphones, context transducer, monophones, and pronunciation models. It takes acoustic indices as input and outputs a word sequence.
Constructing the Decoding Graph

Decoding graph, \( D = H \circ C \circ L \circ G \)

Construct decoding search graph using \( H \circ C \circ L \circ G \) that maps acoustic states to word sequences.

Carefully construct \( D \) using optimization algorithms:

\[
D = \min(\det(H \circ \det(C \circ \det(L \circ G))))
\]

Decode test utterance \( O \) by aligning acceptor \( X \) (corresponding to \( O \)) with \( H \circ C \circ L \circ G \):

\[
W^* = \arg \min_{W=\text{out}[\pi]} X \circ H \circ C \circ L \circ G
\]

where \( \pi \) is a path in the composed FST, \( \text{out}[\pi] \) is the output label sequence of \( \pi \)

“Weighted Finite State Transducers in Speech Recognition”, Mohri et al., Computer Speech & Language, 2002
Constructing the Decoding Graph

Decode test utterance O by aligning acceptor X (corresponding to O) with H ◦ C ◦ L ◦ G:

\[ W^* = \arg \min_{W = \text{out}[\pi]} X \circ H \circ C \circ L \circ G \]

where \( \pi \) is a path in the composed FST, out[\( \pi \)] is the output label sequence of \( \pi \)

Structure of X (derived from O):

“Weighted Finite State Transducers in Speech Recognition”, Mohri et al., Computer Speech & Language, 2002
Constructing the Decoding Graph

- Each $f_k$ maps to a distinct triphone HMM state $j$
- Weights of arcs in the $i^{th}$ chain link correspond to observation probabilities $b_j(o_i)$
- $X$ is a very large FST which is never explicitly constructed!
- $H \circ C \circ L \circ G$ is typically traversed dynamically (search algorithms will be covered later in the semester)

"Weighted Finite State Transducers in Speech Recognition", Mohri et al., Computer Speech & Language, 2002
## Impact of WFST Optimizations

### 40K NAB Evaluation Set '95 (83% word accuracy)

<table>
<thead>
<tr>
<th>network</th>
<th>states</th>
<th>transitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G$</td>
<td>1,339,664</td>
<td>3,926,010</td>
</tr>
<tr>
<td>$L \circ G$</td>
<td>8,606,729</td>
<td>11,406,721</td>
</tr>
<tr>
<td>$det(L \circ G)$</td>
<td>7,082,404</td>
<td>9,836,629</td>
</tr>
<tr>
<td>$C \circ det(L \circ G)$</td>
<td>7,273,035</td>
<td>10,201,269</td>
</tr>
<tr>
<td>$det(H \circ C \circ L \circ G)$</td>
<td>18,317,359</td>
<td>21,237,992</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>network</th>
<th>x real-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C \circ L \circ G$</td>
<td>12.5</td>
</tr>
<tr>
<td>$C \circ det(L \circ G)$</td>
<td>1.2</td>
</tr>
<tr>
<td>$det(H \circ C \circ L \circ G)$</td>
<td>1.0</td>
</tr>
<tr>
<td>$push(min(F))$</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Toolkits to work with finite-state machines

• AT&T FSM Library (no longer supported)
  http://www3.cs.stonybrook.edu/~algorith/implement/fsm/implement.shtml

• RWTH FSA Toolkit
  https://www-i6.informatik.rwth-aachen.de/~kanthak/fsa.html

• Carmel
  https://www.isi.edu/licensed-sw/carmel/

• MIT FST Toolkit
  http://people.csail.mit.edu/ilh/fst/

• OpenFST Toolkit (actively supported)
  http://www.openfst.org/twiki/bin/view/FST/WebHome
Brief Introduction to the OpenFST Toolkit
Quick Intro to OpenFst (www.openfst.org)

Input alphabet (in.txt)

Output alphabet (out.txt)

"0" label is reserved for epsilon
Quick Intro to OpenFst (www.openfst.org)

Diagram:

- State 0 with transitions:
  - an: a/0.5 to State 1
  - ε: n/1.0 to State 2
- State 1 with transition:
  - <eps>: n to State 0
- State 2

Transition Table:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>an</td>
<td>a</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>&lt;eps&gt;</td>
<td>n</td>
<td>1.0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>a</td>
<td>a</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
</tbody>
</table>
Compiling & Printing FSTs

The text FSTs need to be “compiled” into binary objects before further use with OpenFst utilities

- Command used to compile:

```
fstcompile --isymbols=in.txt --osymbols=out.txt A.txt A.fst
```

- Get back the text FST using a print command with the binary file:

```
fstprint --isymbols=in.txt --osymbols=out.txt A.fst A.txt
```
Composing FSTs

The text FSTs need to be “compiled” into binary objects before further use with OpenFst utilities

• Command used to compose:

```
fstcompose A.fst B.fst AB.fst
```

• OpenFST requirement: One or both of the input FSTs should be appropriately sorted before composition

```
fstarcsort --sort_type=olabel A.fst | \n  fstcompose - B.fst AB.fst
```
Drawing FSTs

Small FSTs can be visualized easily using the draw tool:

```
fstdraw --isymbols=in.txt --osymbols=out.txt A.fst | \ ndot -Tpdf > A.pdf
```
FSTs can get very large!
Basics of Speech Production
Speech Production

are more easily explained if a single feature value is allowed to span (what appears on the surface to be) more than one segment. Autosegmental phonology posits some relationships (or associations) between segments in different tiers, which limit the types of transformations that can occur. We will not make use of the details of this theory, other than the motivation that features inherently lie in different tiers of representation.

2.4.3 Articulatory phonology

In the late 1980s, Browman and Goldstein proposed articulatory phonology [BG86, BG92], a theory that differs from previous ones in that the basic units in the lexicon are not abstract binary features but rather articulatory gestures. A gesture is essentially an instruction to the vocal tract to produce a certain degree of constriction at a given location with a given set of articulators. For example, one gesture might be "narrow lip opening", an instruction to the lips and jaw to position themselves so as to effect a narrow opening at the lips. Figure 2-3 shows the main articulators of the vocal tract to which articulatory gestures refer. We are mainly concerned with the lips, tongue, glottis (controlling voicing), and velum (controlling nasality).

Figure 2-3: A midsagittal section showing the major articulators of the vocal tract, reproduced from [oL04].

Schematic representation of the vocal organs


Figure from http://www.phon.ucl.ac.uk/courses/spsci/iss/week6.php
Sound units

- Phones are acoustically distinct units of speech.
- Phonemes are abstract linguistic units that impart different meanings in a given language.
  - Minimal pair: pan vs. ban
- Allophones are different acoustic realisations of the same phoneme.
- Phonetics is the study of speech sounds and how they’re produced.
- Phonology is the study of patterns of sounds in different languages.
Vowels

• Sounds produced with no obstruction to the flow of air through the vocal tract

Vowel Quadrilateral

Vowels at right & left of bullets are rounded & unrounded.

Formants of vowels

- Formants are resonance frequencies of the vocal tract (denoted by F1, F2, etc.)
- F0 denotes the fundamental frequency of the periodic source (vibrating vocal folds)
- Formant locations specify certain vowel characteristics

<table>
<thead>
<tr>
<th>Vowel</th>
<th>F1(Hz)</th>
<th>F2(Hz)</th>
<th>F3(Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ĭ</td>
<td>280</td>
<td>2620</td>
<td>3380</td>
</tr>
<tr>
<td>ī</td>
<td>360</td>
<td>2220</td>
<td>2960</td>
</tr>
<tr>
<td>ē</td>
<td>600</td>
<td>2060</td>
<td>2840</td>
</tr>
<tr>
<td>æ</td>
<td>800</td>
<td>1760</td>
<td>2500</td>
</tr>
<tr>
<td>ə</td>
<td>760</td>
<td>1320</td>
<td>2500</td>
</tr>
<tr>
<td>ɐ</td>
<td>740</td>
<td>1180</td>
<td>2640</td>
</tr>
<tr>
<td>ɔ</td>
<td>560</td>
<td>920</td>
<td>2560</td>
</tr>
<tr>
<td>ɔː</td>
<td>480</td>
<td>760</td>
<td>2620</td>
</tr>
<tr>
<td>ʊ</td>
<td>380</td>
<td>940</td>
<td>2300</td>
</tr>
<tr>
<td>uː</td>
<td>320</td>
<td>920</td>
<td>2200</td>
</tr>
<tr>
<td>üː</td>
<td>560</td>
<td>1480</td>
<td>2520</td>
</tr>
</tbody>
</table>

Adult male formant frequencies in Hertz collected by J.C. Wells around 1960.
Spectrogram

- Spectrogram is a sequence of spectra stacked together in time, with amplitude of the frequency components expressed as a heat map.

- Spectrograms of certain vowels:
  http://www.phon.ucl.ac.uk/courses/spsci/iss/week5.php

- Praat (http://www.fon.hum.uva.nl/praat/) is a good toolkit to analyse speech signals (plot spectrograms, generate formants/pitch curves, etc.)
Consonants (voicing/place/manner)

• “Consonants are made by restricting or blocking the airflow in some way, and may be voiced or unvoiced.” (J&M, Ch. 7)
Voiced/Unvoiced Sounds

• Sounds made with vocal cords vibrating: voiced
  
  • E.g. /g/, /d/, etc.

• All English vowel sounds are voiced

• Sounds made without vocal cord vibration: voiceless
  
  • E.g. /k/, /t/, etc.
Consonants (voicing/place/manner)

- “Consonants are made by restricting or blocking the airflow in some way, and may be voiced or unvoiced.” (J&M, Ch. 7)

- Consonants can be labeled depending on
  - where the constriction is made
  - how the constriction is made
Place of articulation

- Bilabial (both lips) \([b],[p],[m]\), etc.
- Labiodental (with lower lip and upper teeth) \([f],[v]\), etc.
- Interdental (tip of tongue between teeth) \([\epsilon] \) (thought), \([\delta]\) (this)
2.4.3 Articulatory phonology

In the late 1980s, Browman and Goldstein proposed articulatory phonology (BG86, BG92), a theory that differs from previous ones in that the basic units in the lexicon are not abstract binary features but rather articulatory gestures. A gesture is essentially an instruction to the vocal tract to produce a certain degree of constriction at a given location with a given set of articulators. For example, one gesture might be "narrow lip opening", an instruction to the lips and jaw to position themselves so as to effect a narrow opening at the lips. Figure 2-3 shows the main articulators of the vocal tract to which articulatory gestures refer. We are mainly concerned with the lips, tongue, glottis (controlling voicing), and velum (controlling nasality).

Figure 2-3: A midsagittal section showing the major articulators of the vocal tract, reproduced from [oL04].

- Plosive/Stop (airflow completely blocked followed by a release)  
  [p], [g], [t], etc.

- Fricative (constricted airflow)  
  [f], [s], [th], etc.

- Affricate (stop + fricative)  
  [ch], [jh], etc.

- Nasal (lowering velum)  
  [n], [m], etc.

See realtime MRI productions of vowels and consonants here: http://sail.usc.edu/span/rtmri_ipa/je_2015.html