WFSTs in ASR
&
Basics of Speech Production

Lecture 6

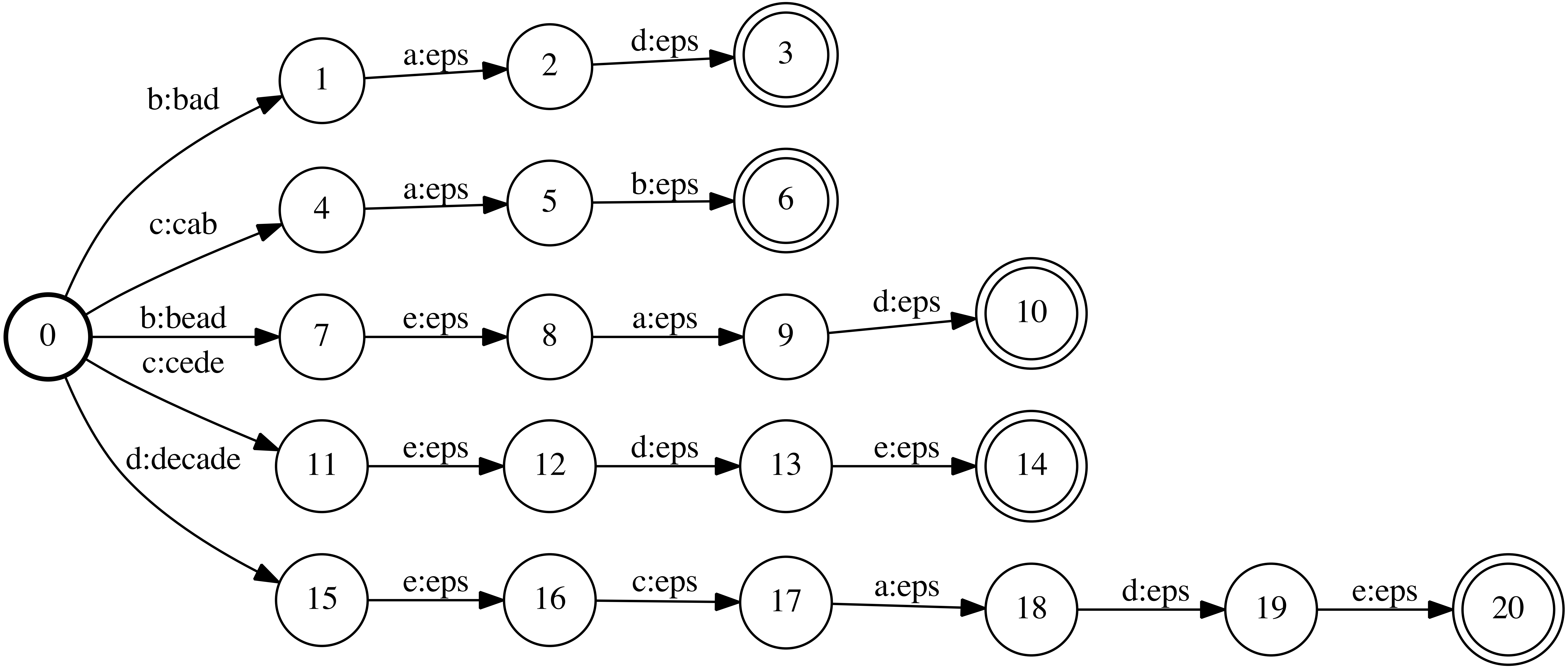
CS 753

Instructor: Preethi Jyothi

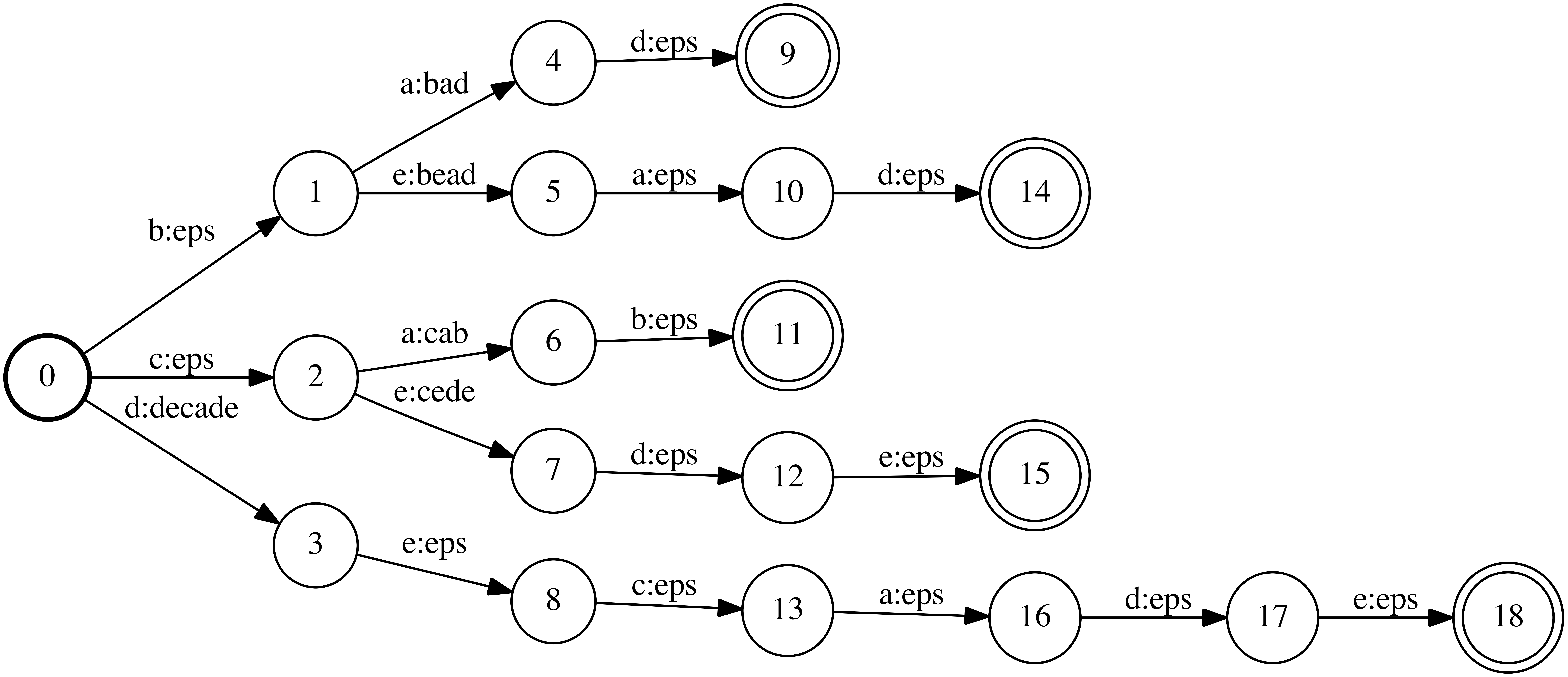
Determinization/Minimization: Recap

- 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)

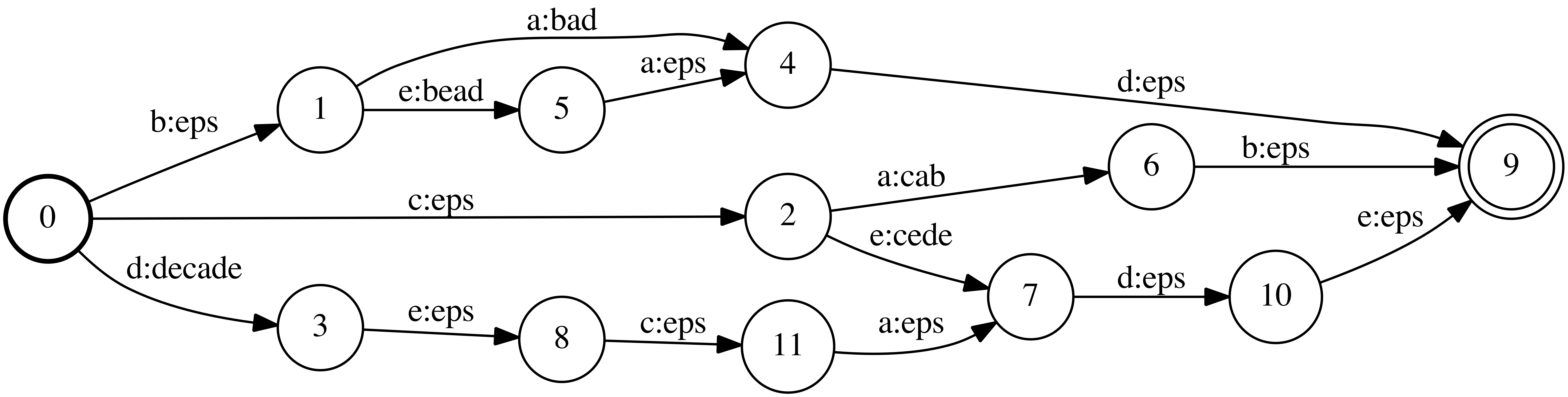
Example: Dictionary WFST



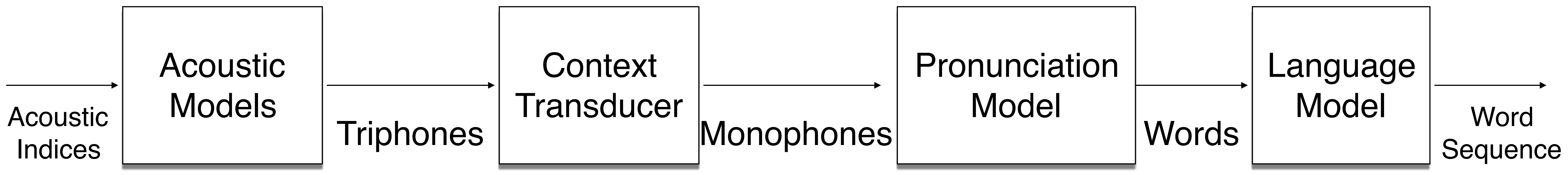
Determinized Dictionary WFST



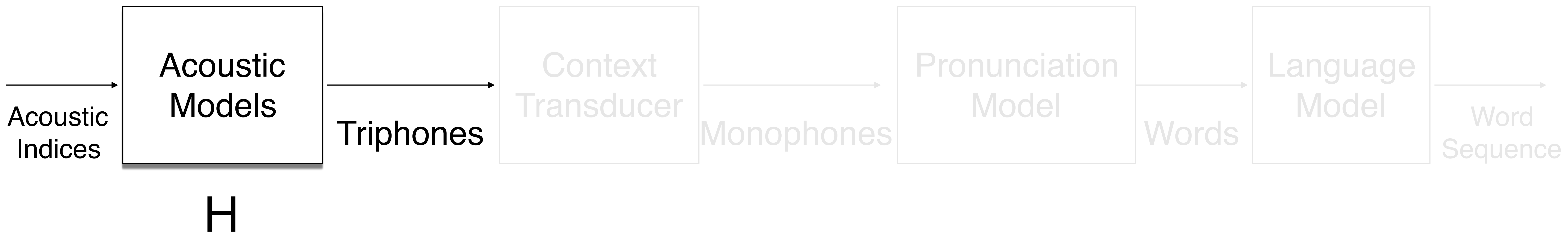
Minimized Dictionary WFST

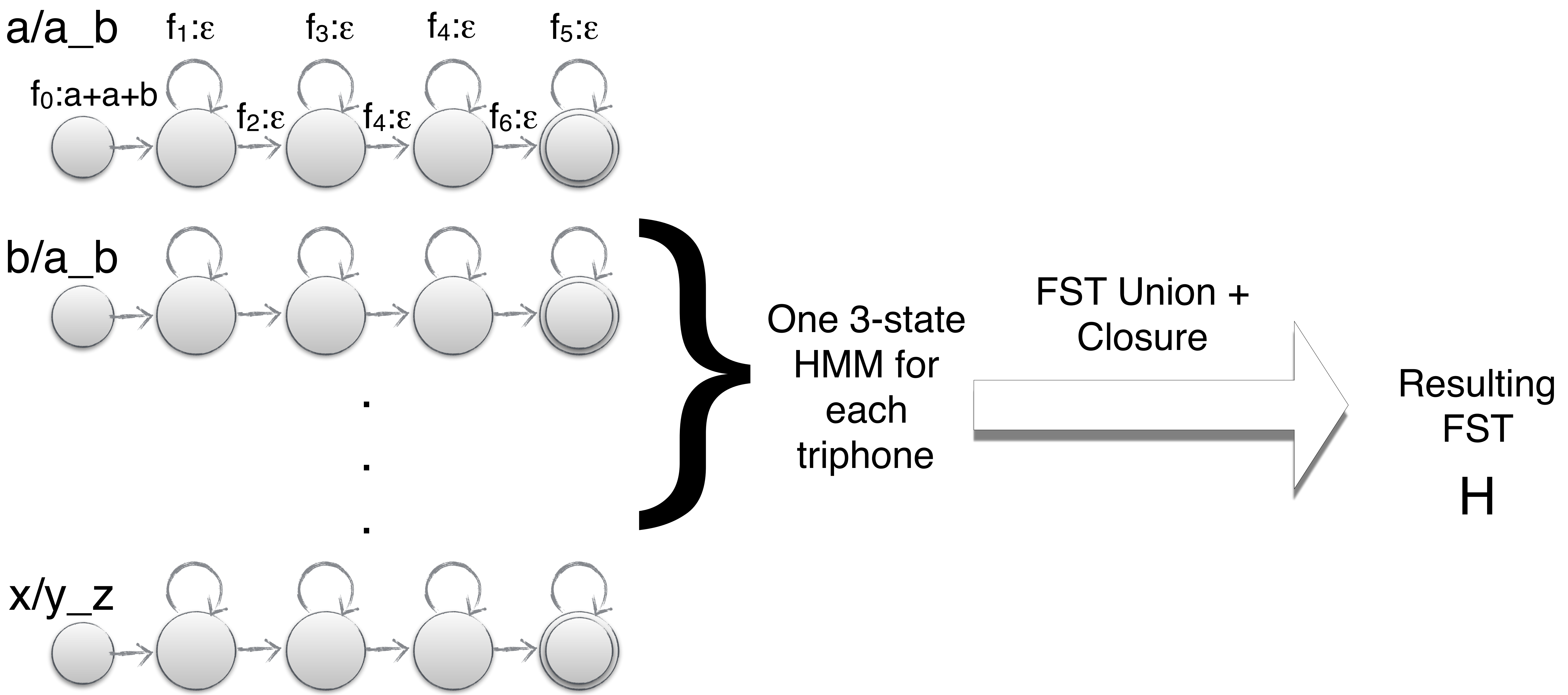


WFST-based ASR System



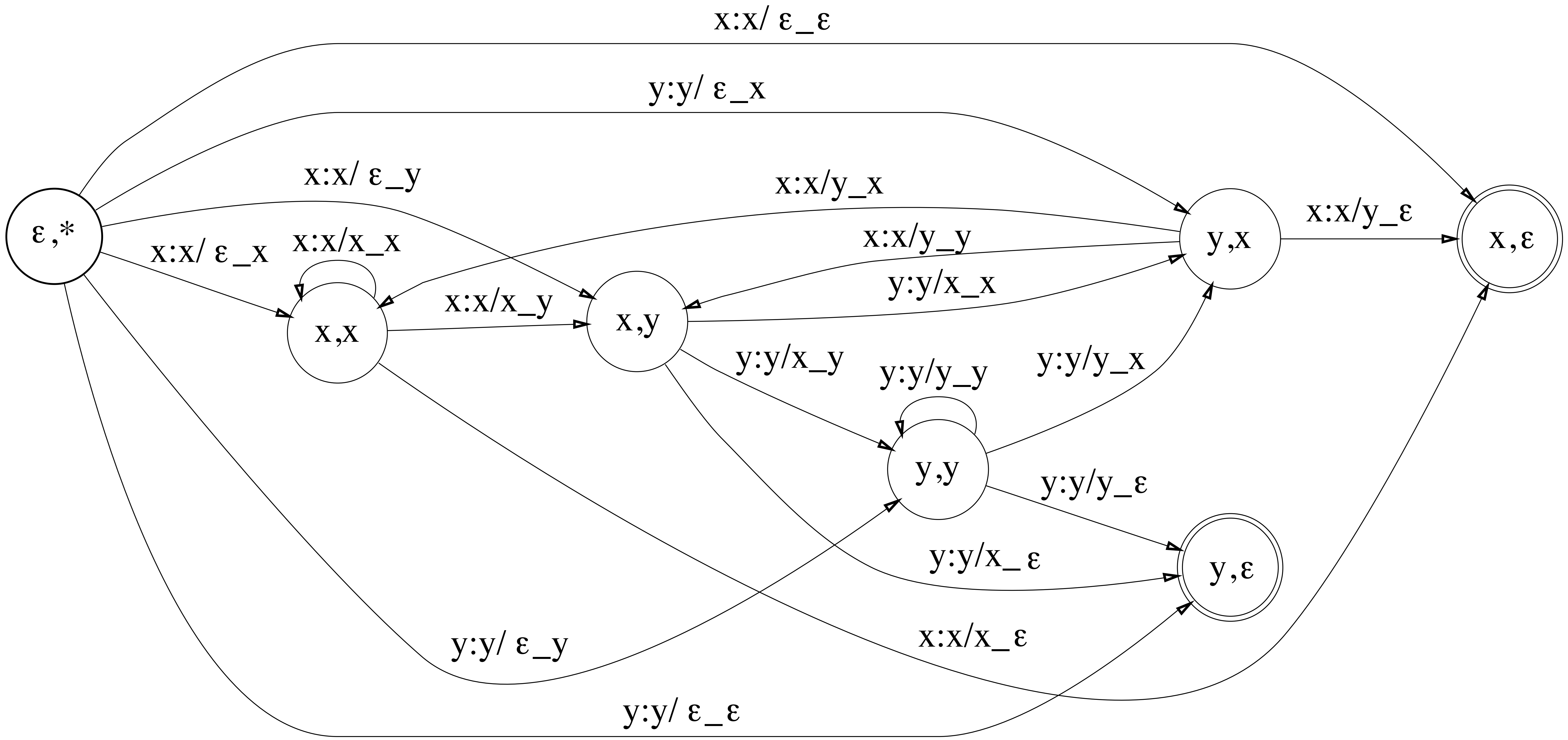
WFST-based ASR System





WFST-based ASR System

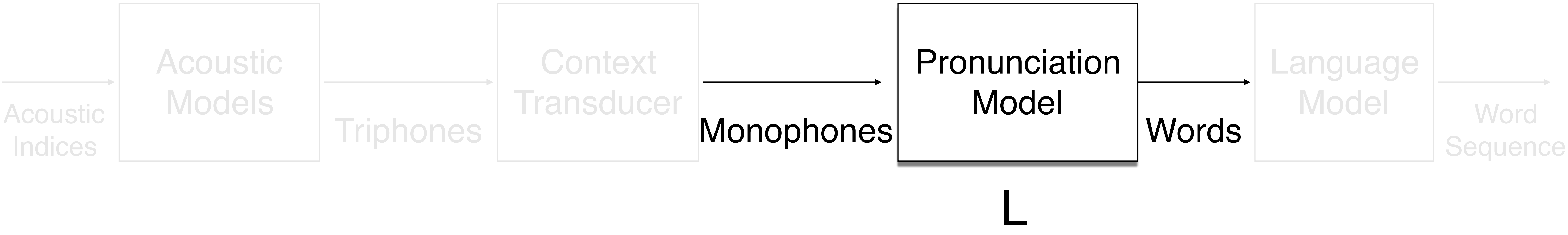
C



C^{-1}: Arc labels: “monophone : phone / left-context_right-context”

Figure reproduced from “Weighted Finite State Transducers in Speech Recognition”, Mohri et al., 2002

WFST-based ASR System



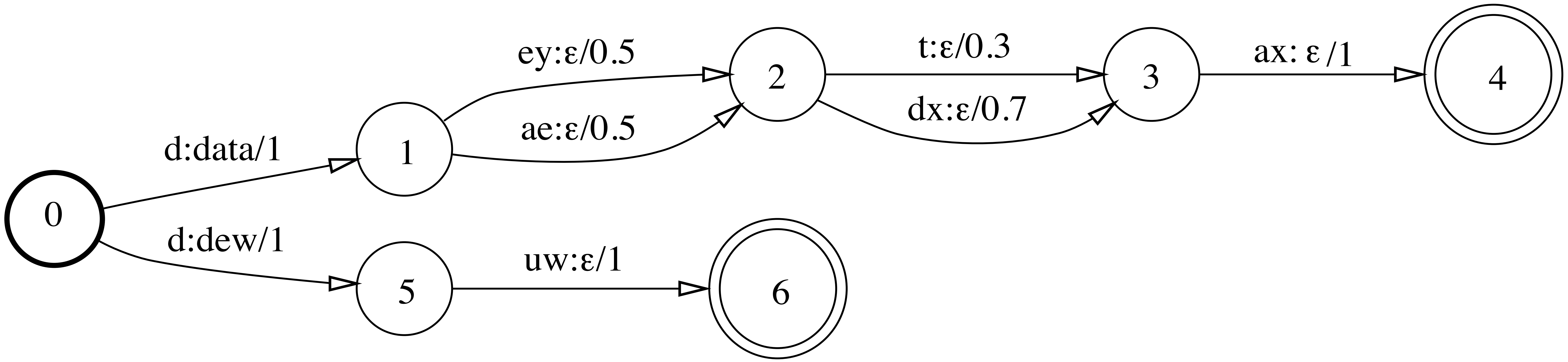
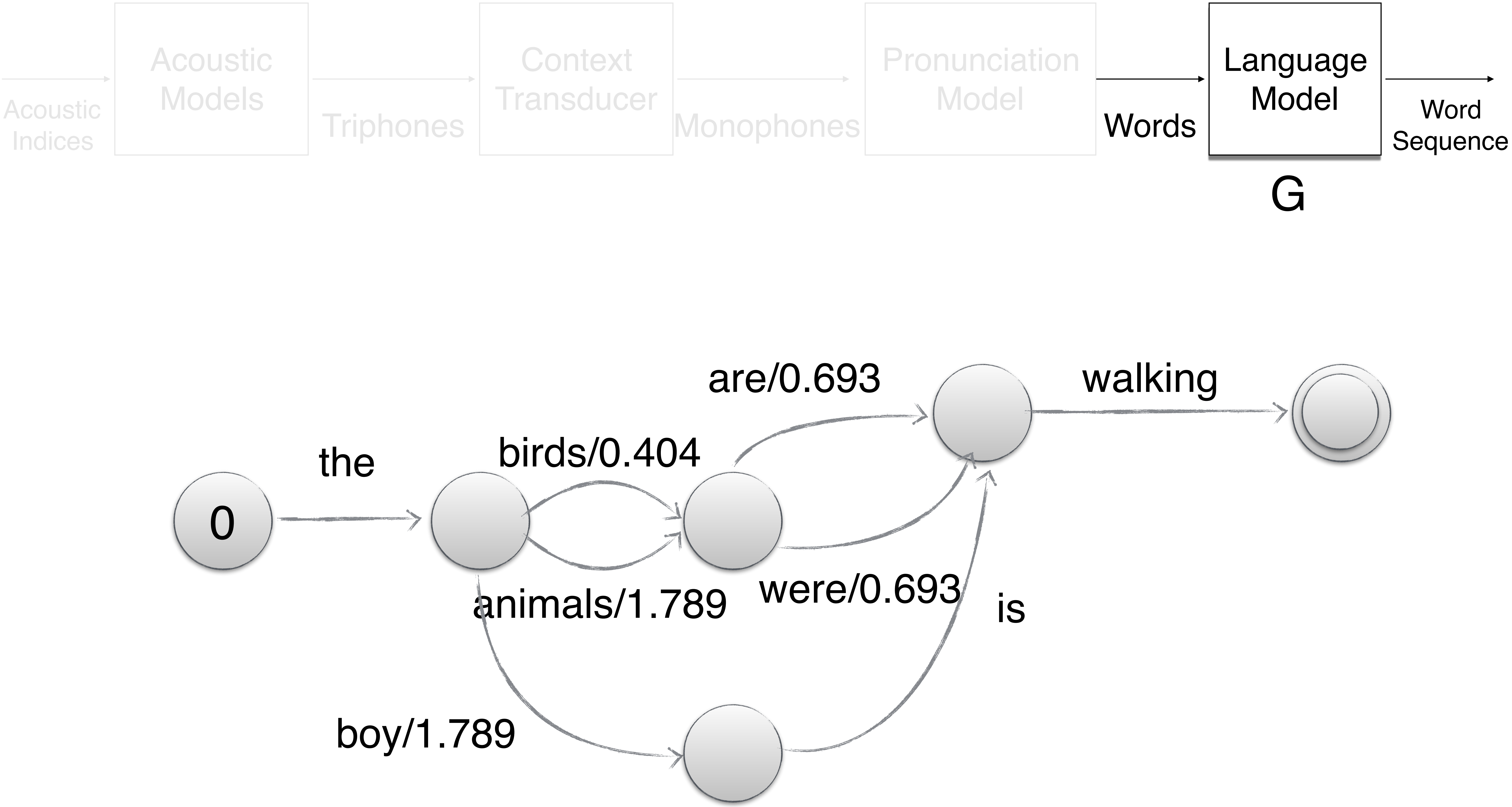
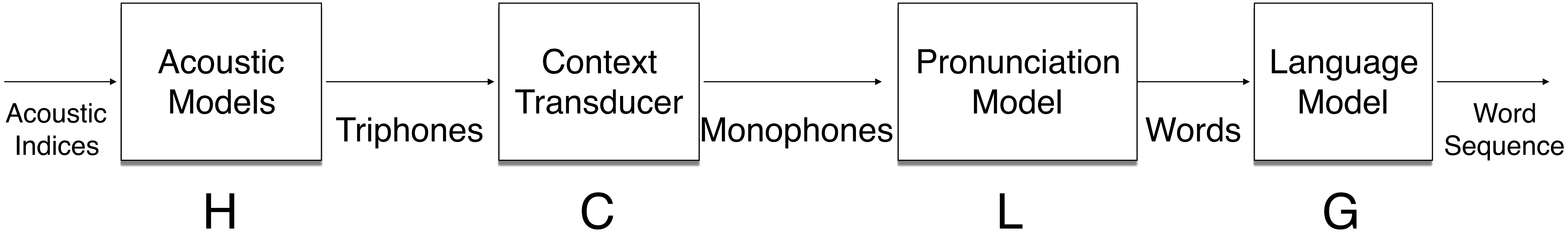


Figure reproduced from "Weighted Finite State Transducers in Speech Recognition", Mohri et al., 2002

WFST-based ASR System



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:

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

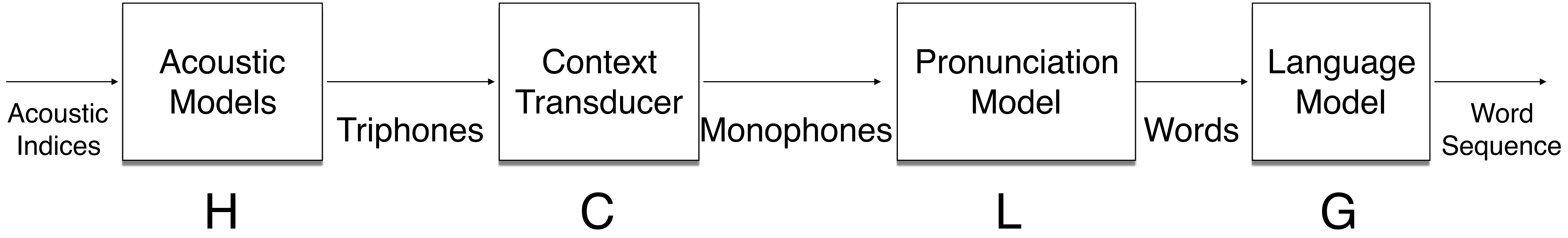
Decode test utterance O by aligning acceptor X (corresponding to O) with $H \circ C \circ L \circ G$:

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

where π is a path in the composed FST, out[π] is the output label sequence of π

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

Constructing the Decoding Graph

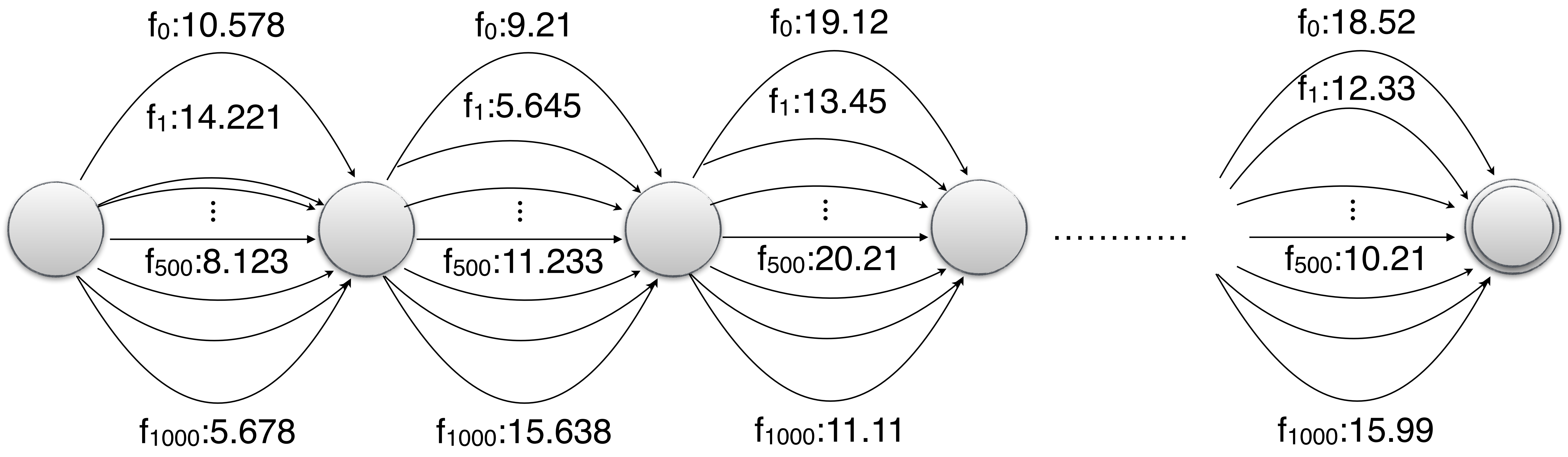


Decode test utterance O by aligning acceptor X (corresponding to O) with H ∘ C ∘ L ∘ G:

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

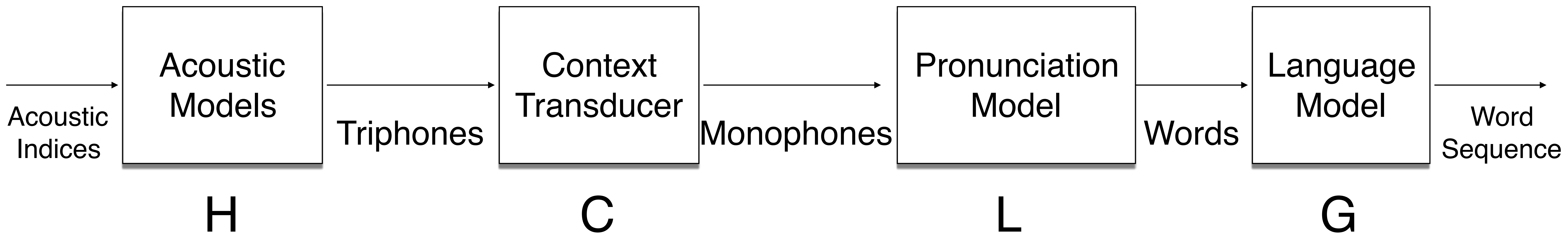
where π is a path in the composed FST, out[π] is the output label sequence of π

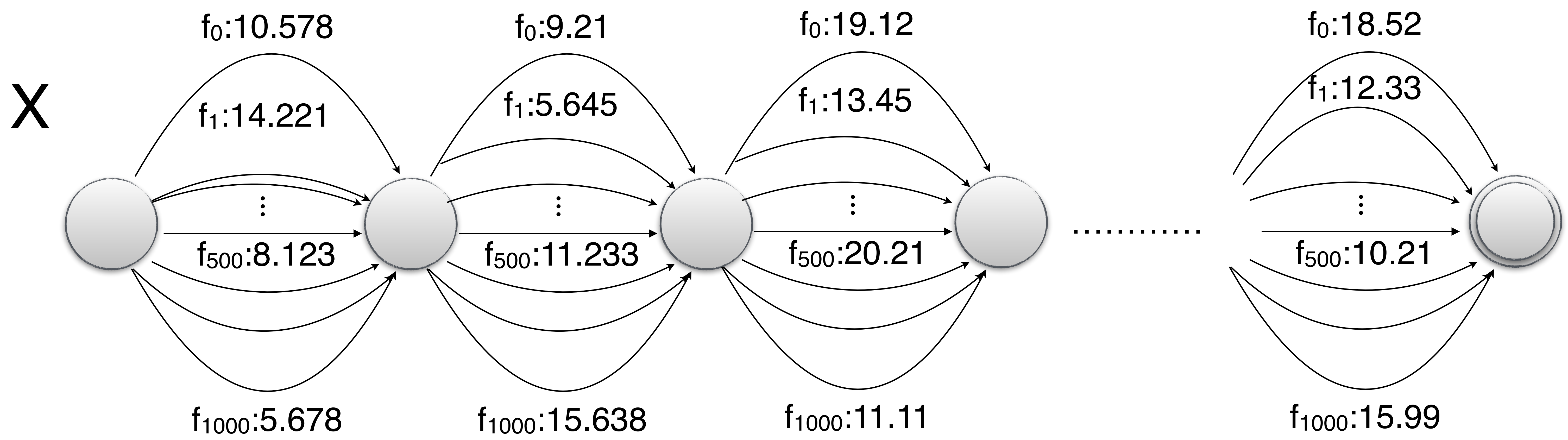
Structure of X (derived from O):



"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 ○ C ○ L ○ 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)

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

network	x real-time
$C \circ L \circ G$	12.5
$C \circ det(L \circ G)$	1.2
$det(H \circ C \circ L \circ G)$	1.0
$push(min(F))$	0.7

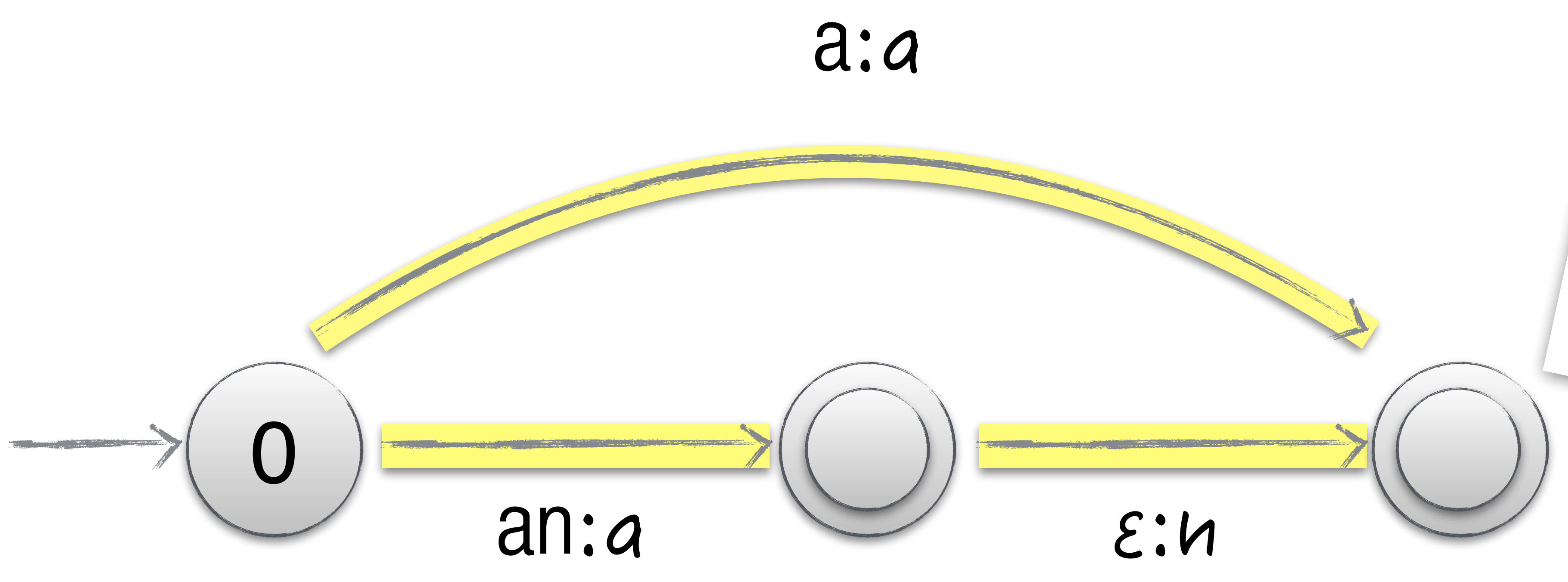
Tables from http://www.openfst.org/twiki/pub/FST/FstHltTutorial/tutorial_part3.pdf

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)



"0" label is reserved for epsilon

```
0    1    an      a
1    2    <eps>   n
0    2    a       a
1
2
```

A.txt

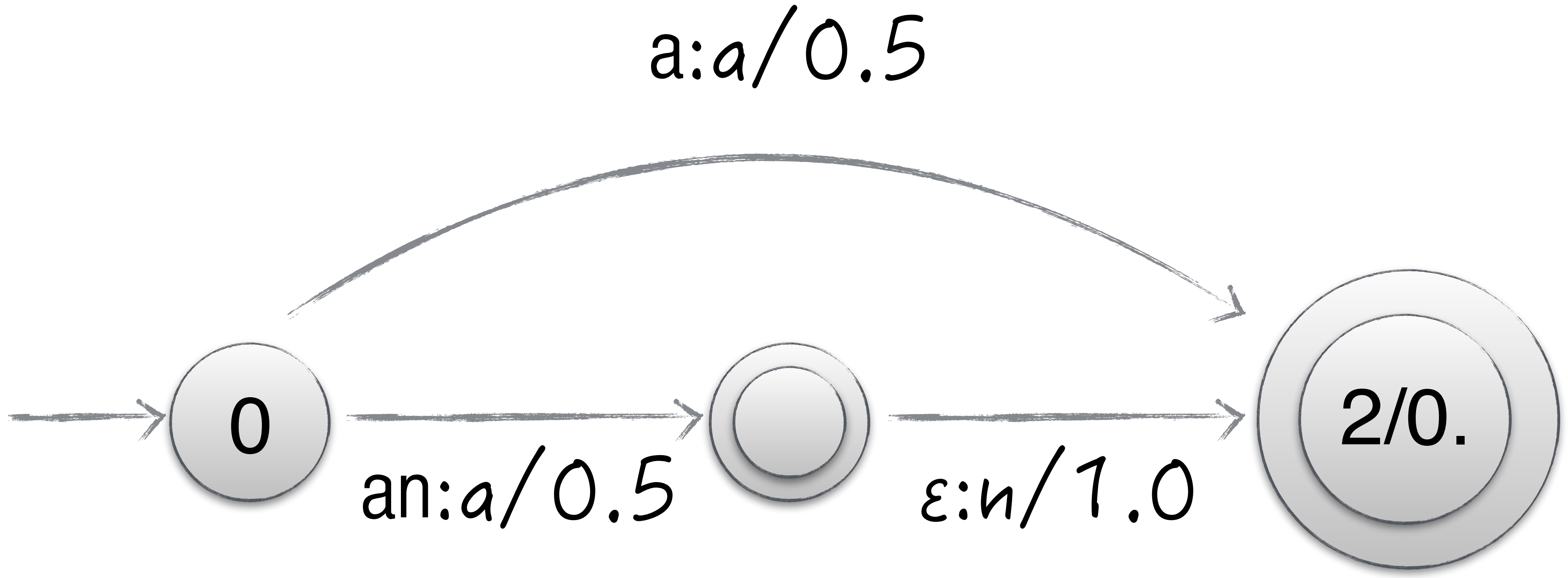
```
<eps>   0
an      1
a       2
```

Input alphabet (in.txt)

```
<eps>   0
a       1
n       2
```

Output alphabet (out.txt)

Quick Intro to OpenFst (www.openfst.org)



0	1	an	a	0.5
1	2	<eps>	n	1.0
0	2	a	a	0.5
1				
2	0.1			

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
```

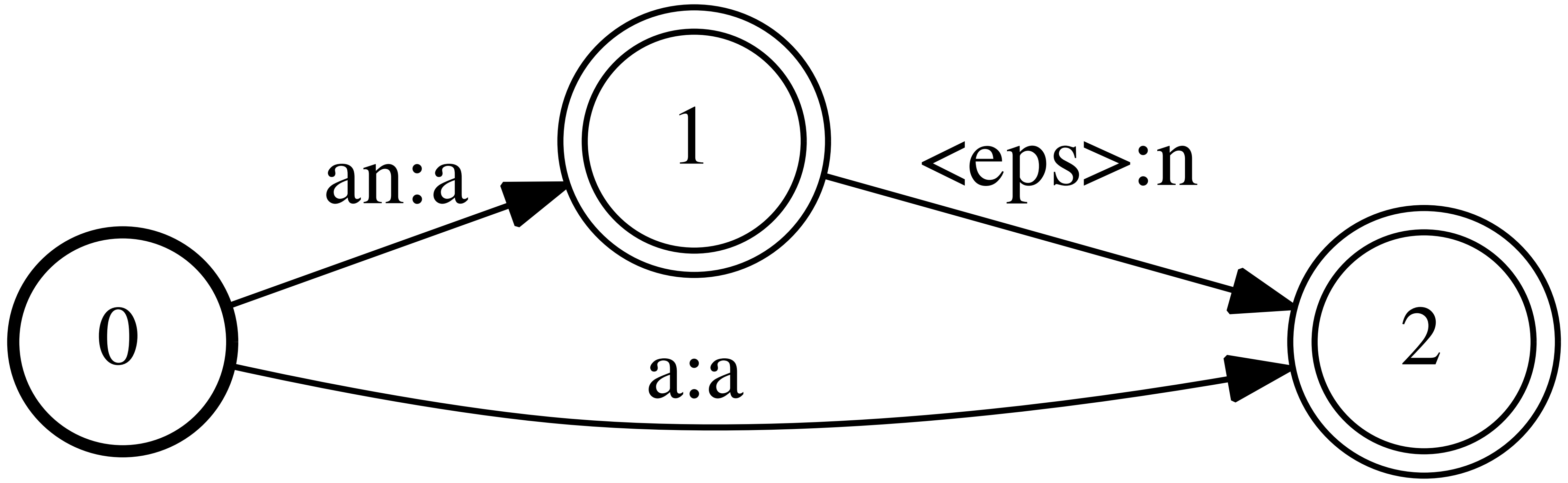
- OpenFST requirement: One or both of the input FSTs should be appropriately sorted before composition

```
fstarcsort —-sort_type=olabel A.fst |\
     fstcompose - B.fst AB.fst
```

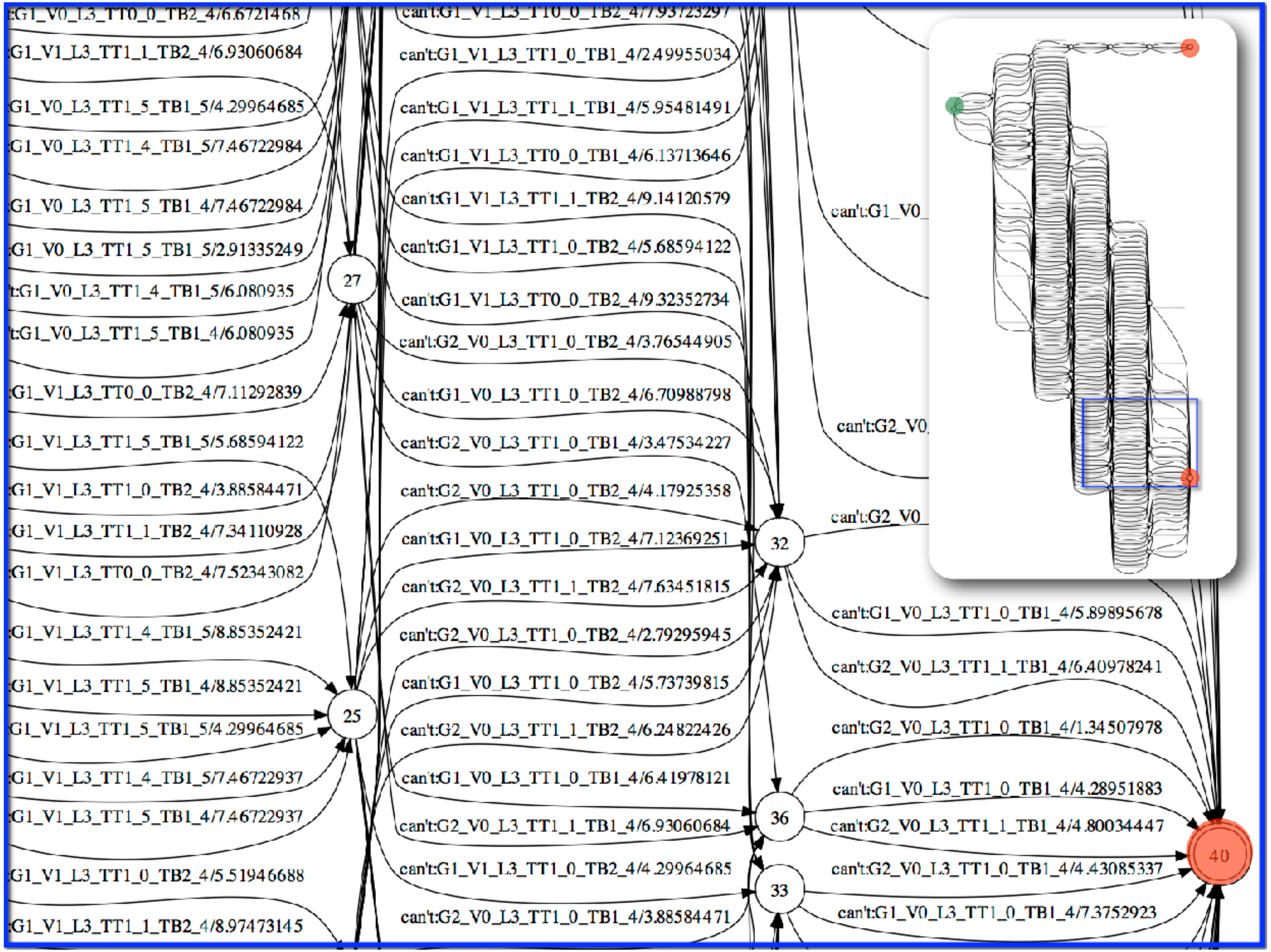
Drawing FSTs

Small FSTs can be visualized easily using the draw tool:

```
fstdraw --isymbols=in.txt --osymbols=out.txt A.fst |\
dot -Tpdf > A.pdf
```

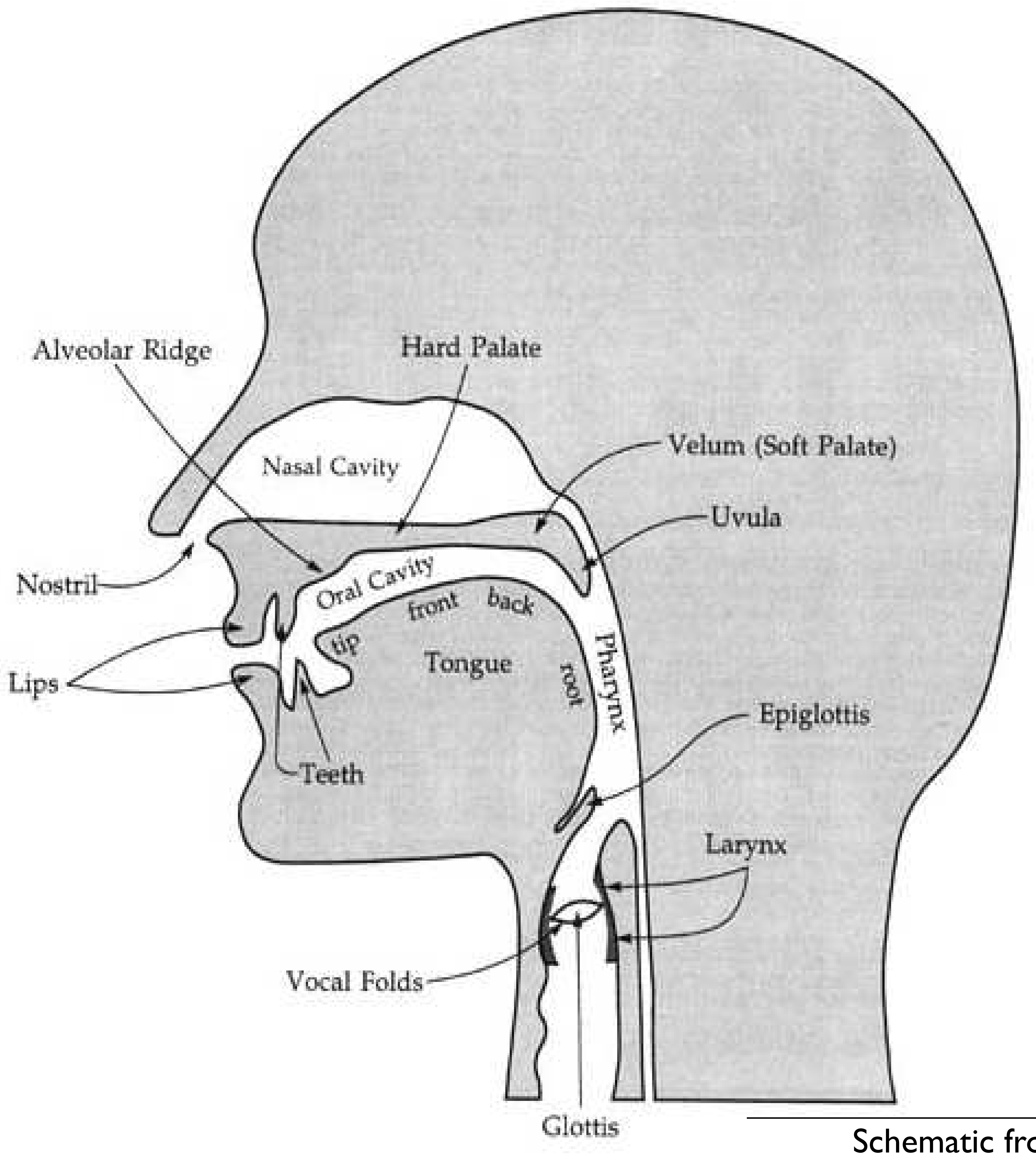


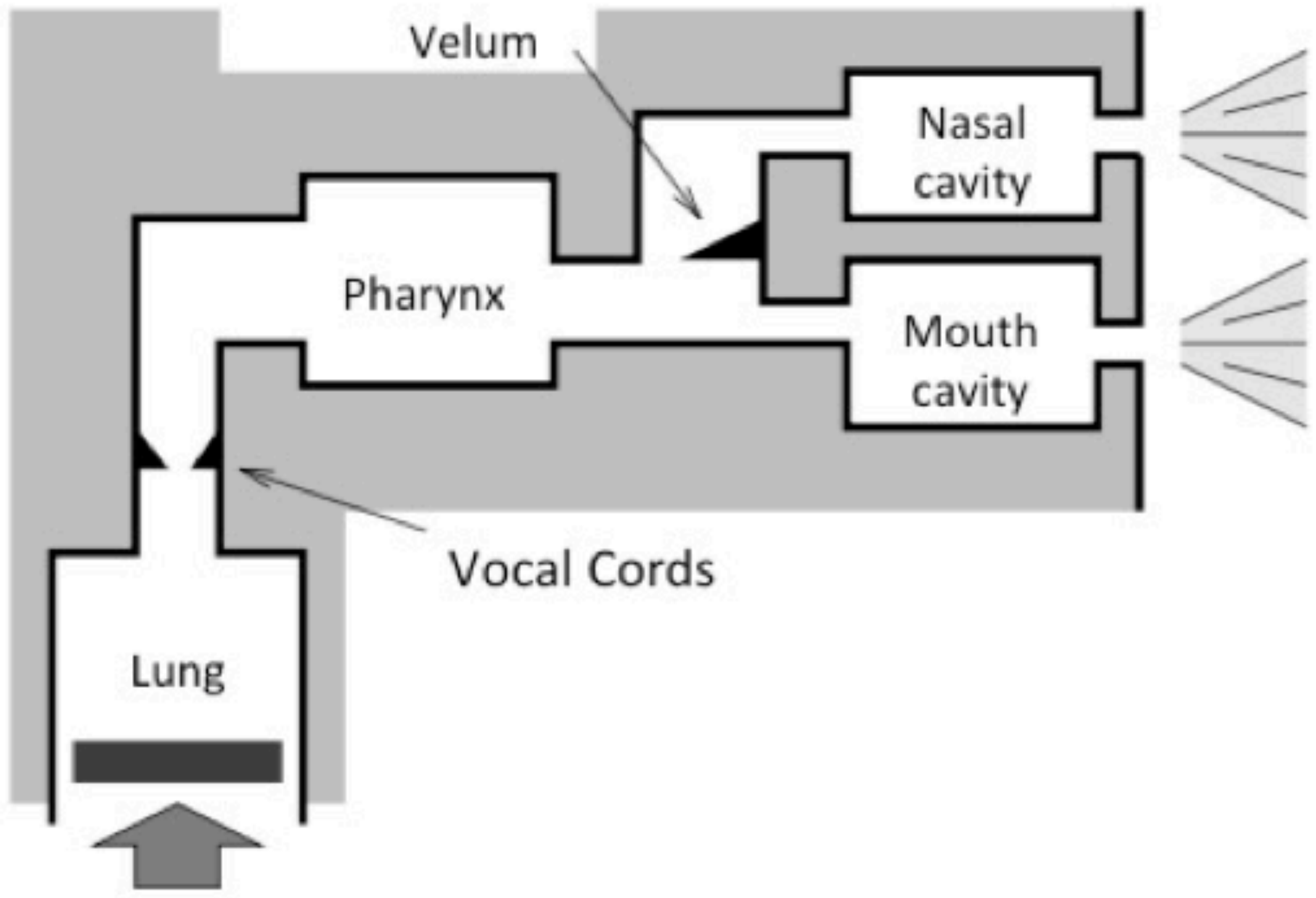
FSTs can get very large!



Basics of Speech Production

Speech Production





Schematic representation of the vocal organs

Schematic from L.Rabiner and B.-H.Juang , Fundamentals of speech recognition, 1993
Figure from http://www.phon.ucl.ac.uk/courses/spsci/iss/week6.php

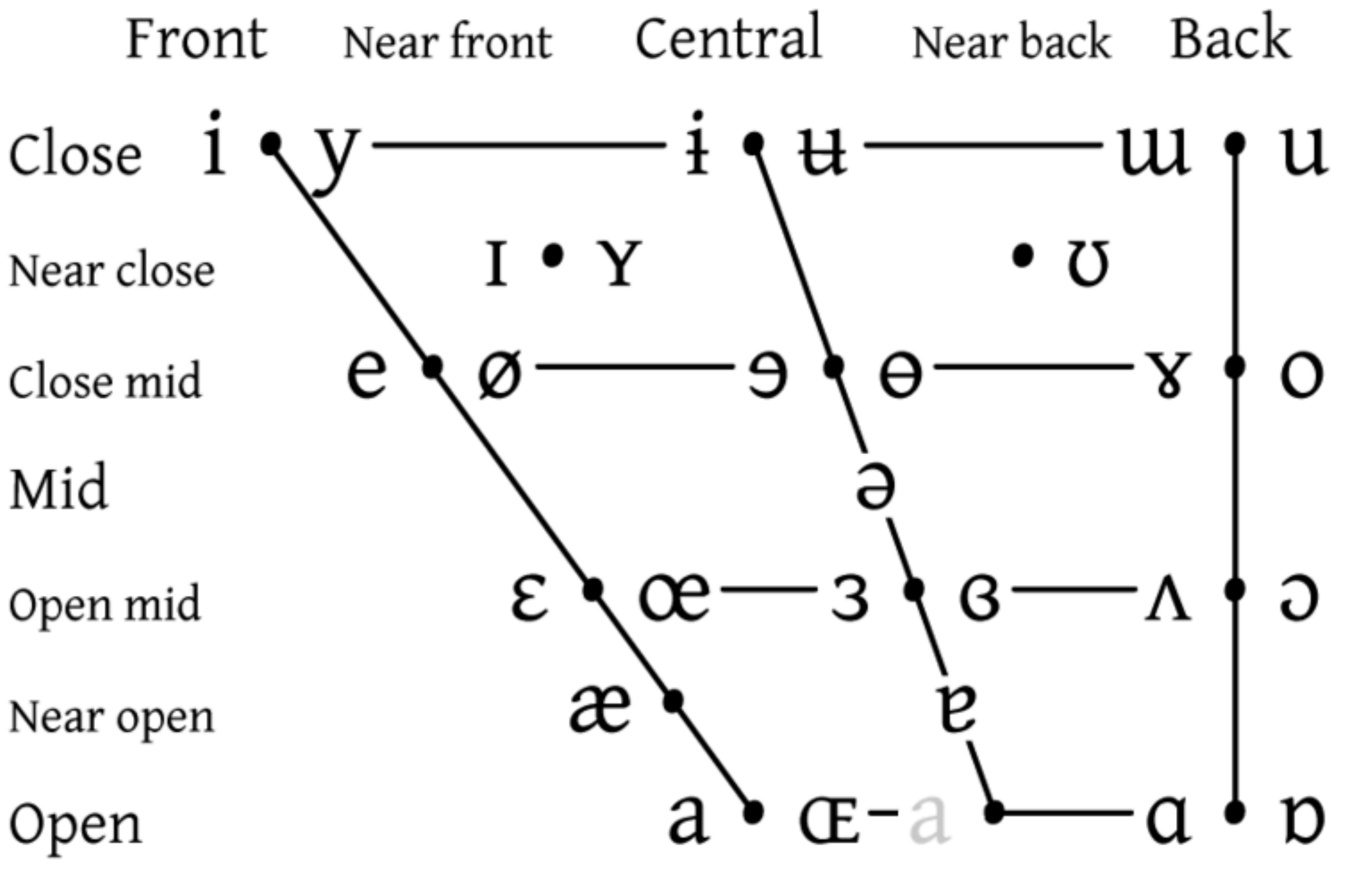
Sound units

- Phones are acoustically distinct units of speech
- Phonemes are abstract linguistic units that impart different meanings in a given language
  - Minimal pair: <u>p</u>an vs. <u>b</u>an
- 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.

Image from https://en.wikipedia.org/wiki/File:IPA_vowel_chart_2005.png

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

Vowel	F1(Hz)	F2(Hz)	F3(Hz)
iː	280	2620	3380
ɪ	360	2220	2960
e	600	2060	2840
æ	800	1760	2500
ʌ	760	1320	2500
ɑː	740	1180	2640
ɒ	560	920	2560
ɔː	480	760	2620
ʊ	380	940	2300
uː	320	920	2200
ɜː	560	1480	2520

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

Image from: https://www.phon.ucl.ac.uk/courses/spsci/iss/week5.php

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)

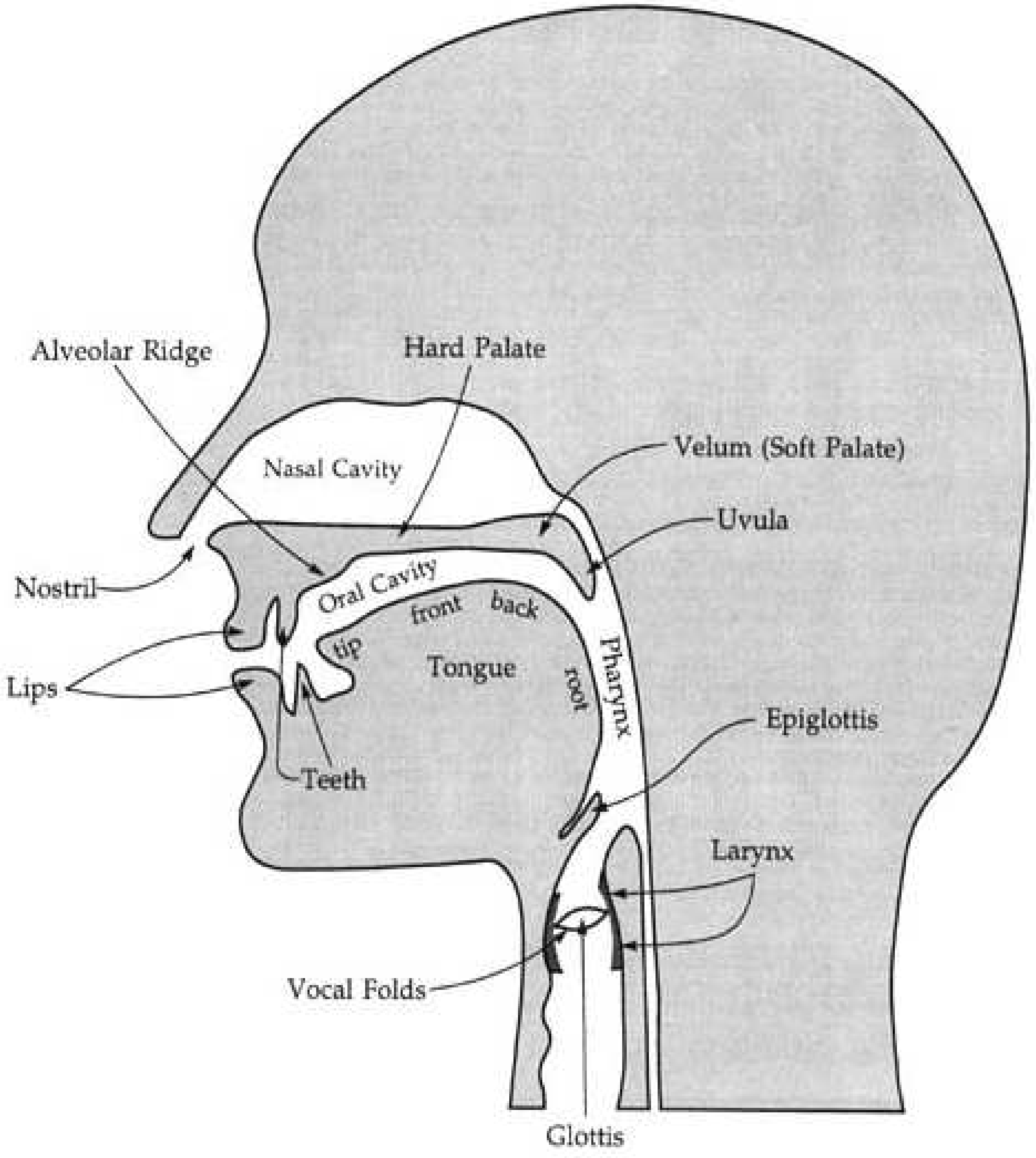
Voiced/Unvoiced Sounds

- 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)

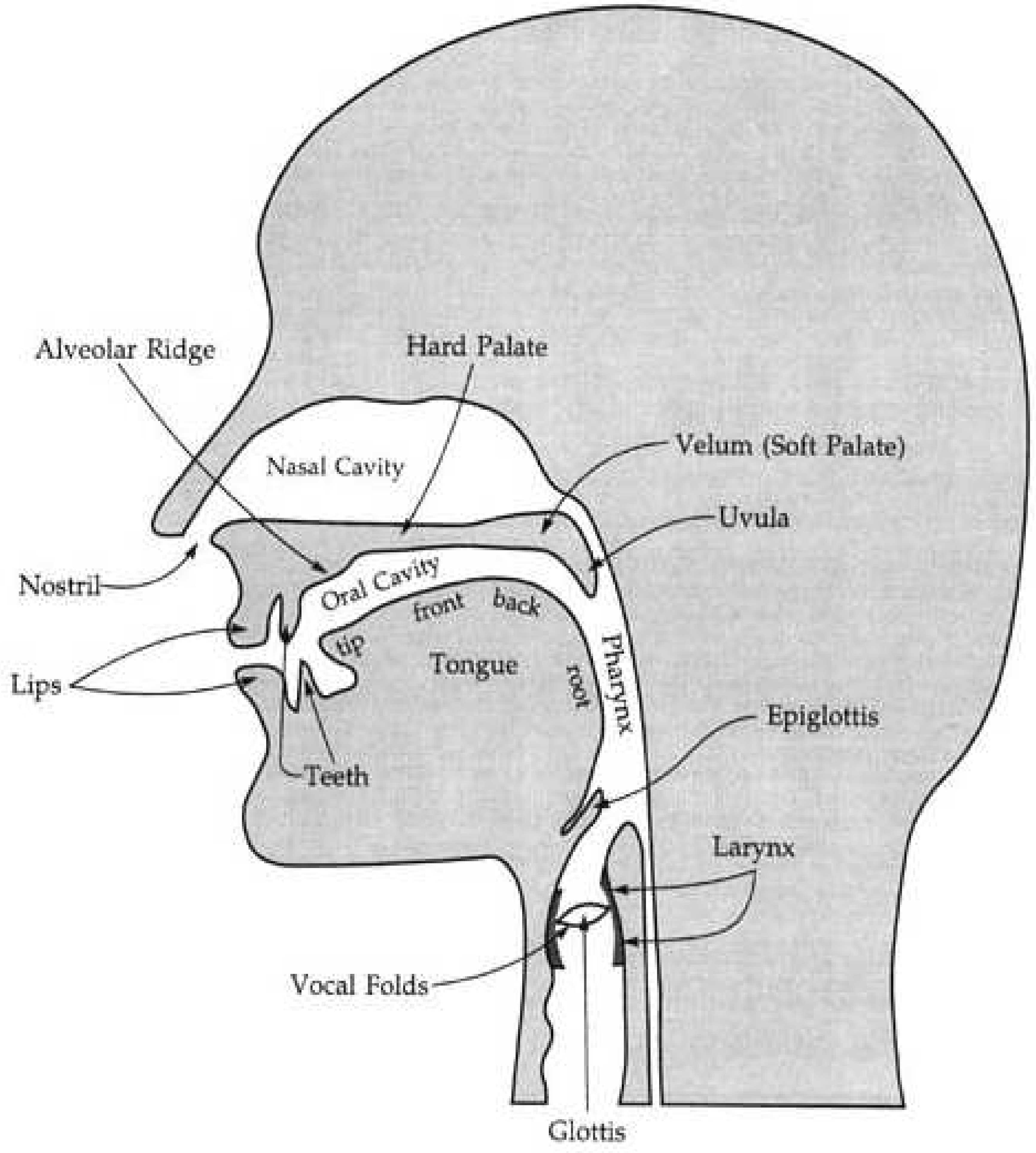
- "Consonants are made by restricting or blocking the airflow in some way, and may be voiced or unvoiced." (J&M, Ch. 7)
- Consonants can be labeled depending on
  - where the constriction is made
  - how the constriction is made

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)
  [ɵ] (<u>th</u>ought), [δ] (<u>th</u>is)

Manner of articulation



- 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