Automatic Speech Recognition (CS753)

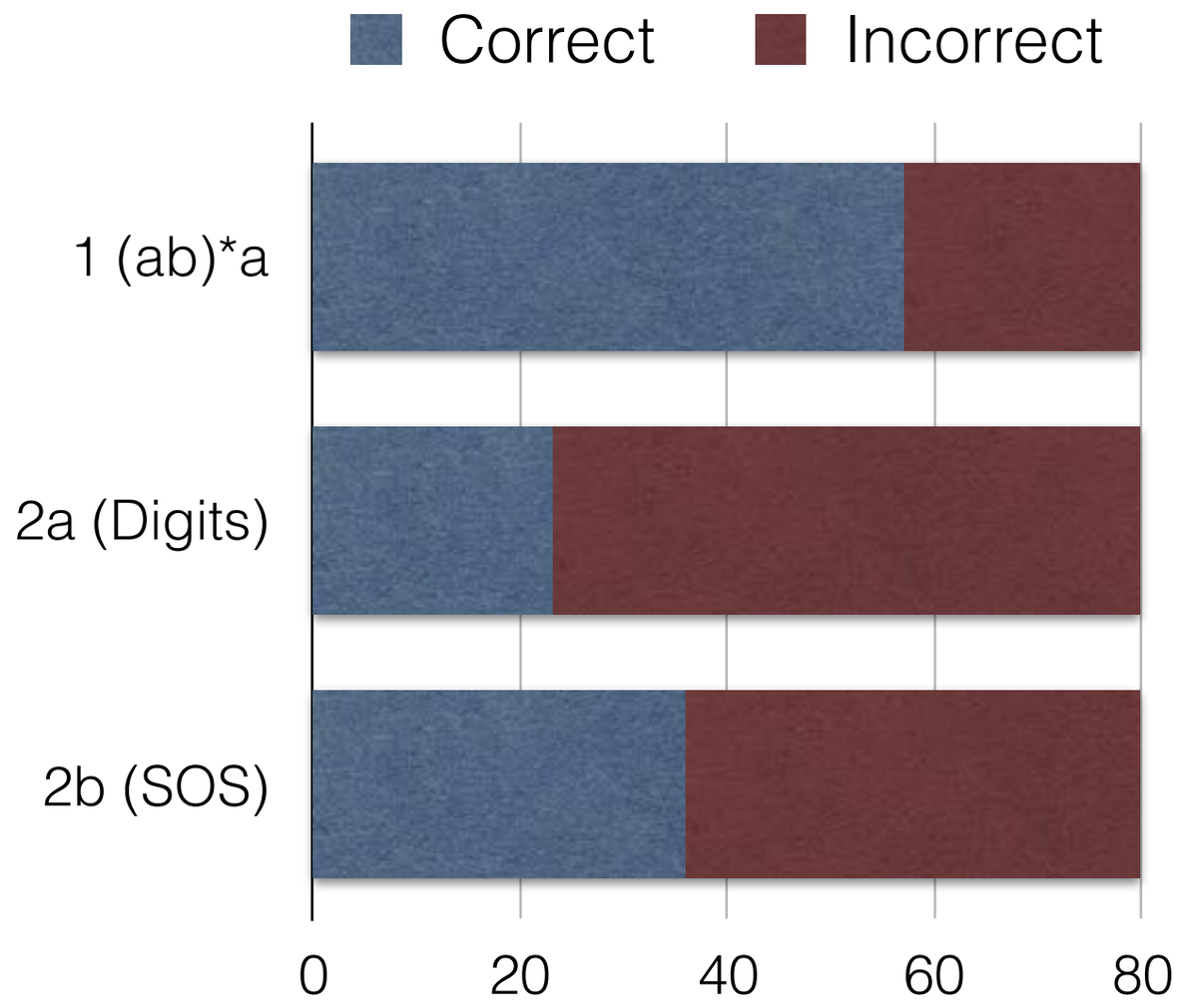
Lecture 4: WFSTs in ASR + Basics of Speech Production

Instructor: Preethi Jyothi

Lecture 4

Quiz-1 Postmortem

- Common Mistakes:
  - Output vocabulary for 2(a) used complete words “ZERO”, etc. rather than letters.
  - 2(b) No self-loops on start/final state in the “SOS” machine.
  - 2(b) All states marked as final.



Project Proposal

- Start brainstorming!

- Discuss potential ideas with me during my office hours (Thur, 5.30 pm to 6.30 pm) or schedule a meeting

- Once decided, send me a (plain ASCII) email specifying:

  - Title of the project
  - Full names of all project members
  - A 300-400 word abstract of the proposed project

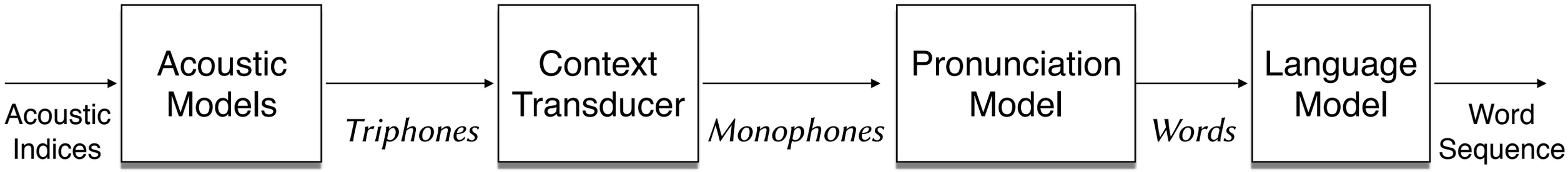
- Email due by 11.59 pm on Jan 30th.

Determinization/Minimization: Recap

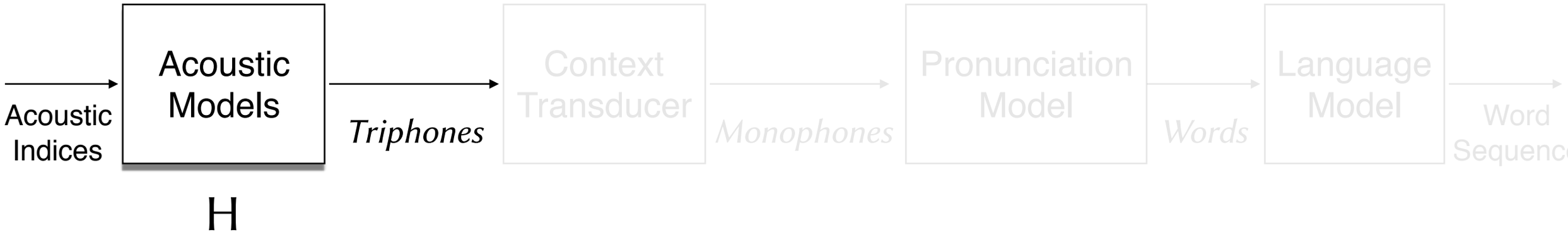
- A (W)FST is **deterministic** if:
  - Unique start state
  - No two transitions from a state share the same input label
  - No epsilon input labels
- **Minimization** finds an equivalent deterministic FST with the least number of states (and transitions)
  - For a deterministic weighted automaton, 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)

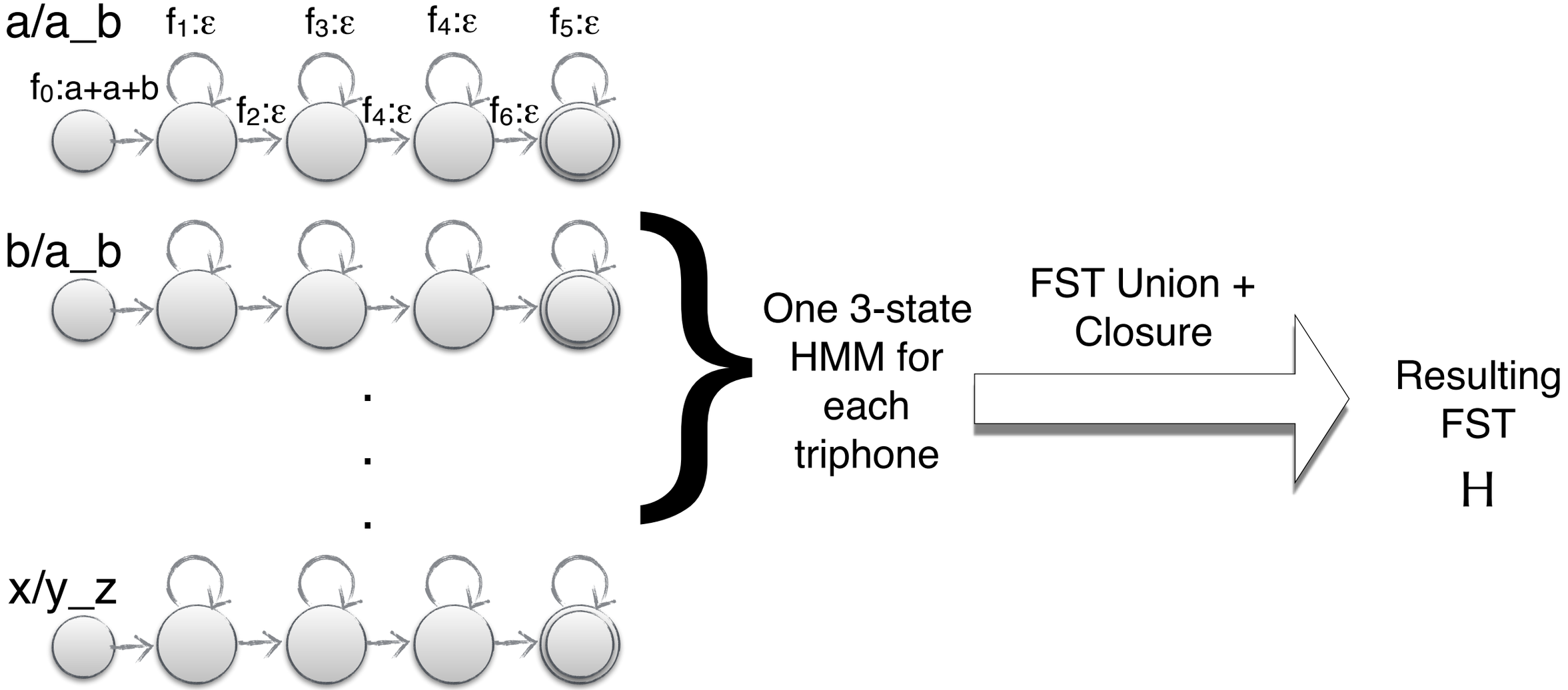
WFSTs applied to ASR

WFST-based ASR System

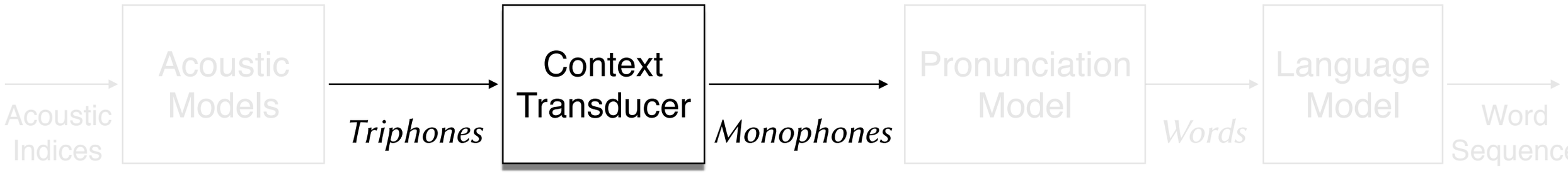


WFST-based ASR System

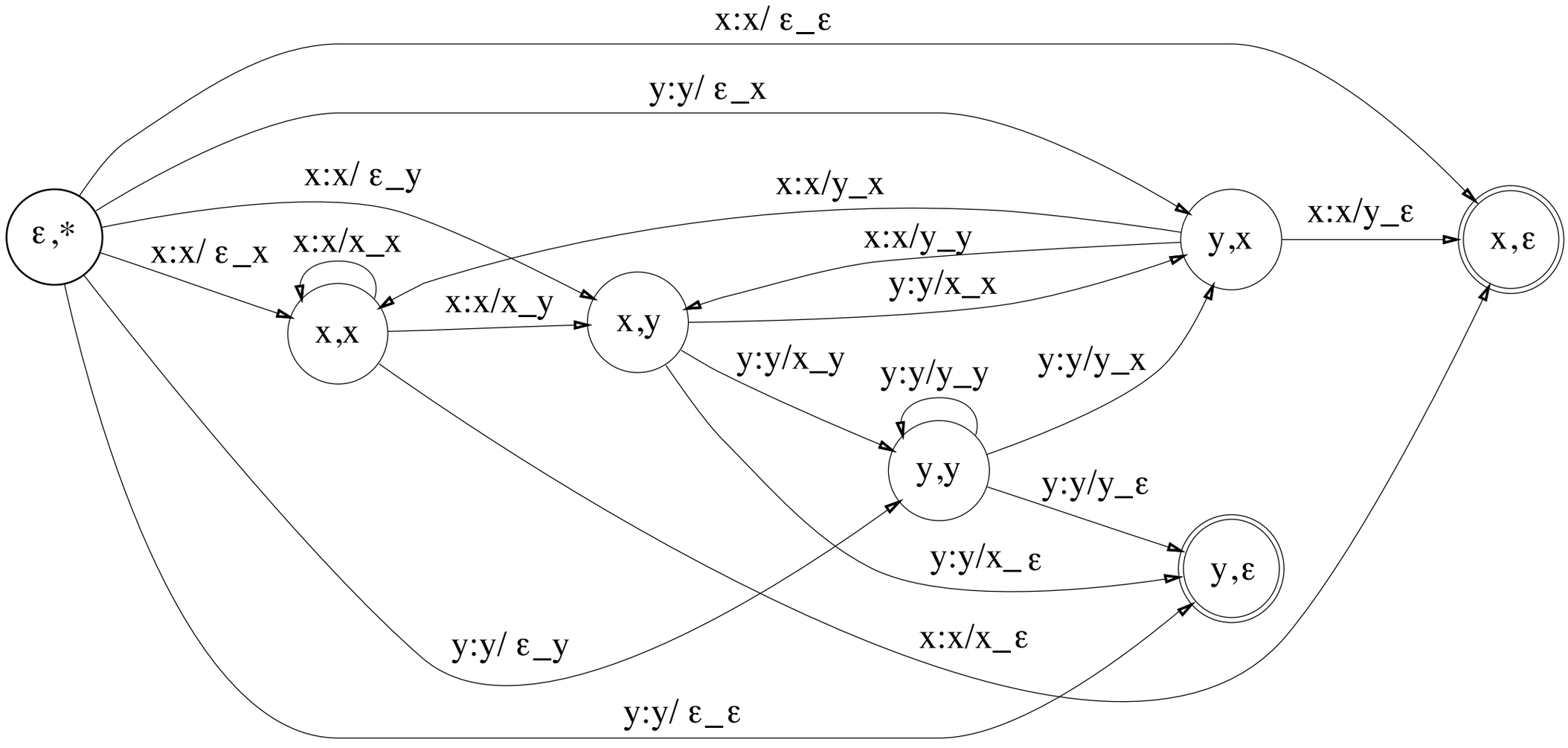




WFST-based ASR System



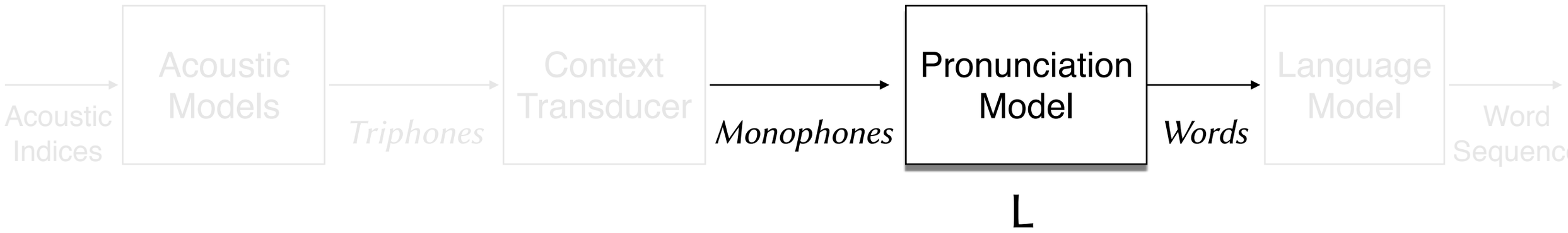
C



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



L

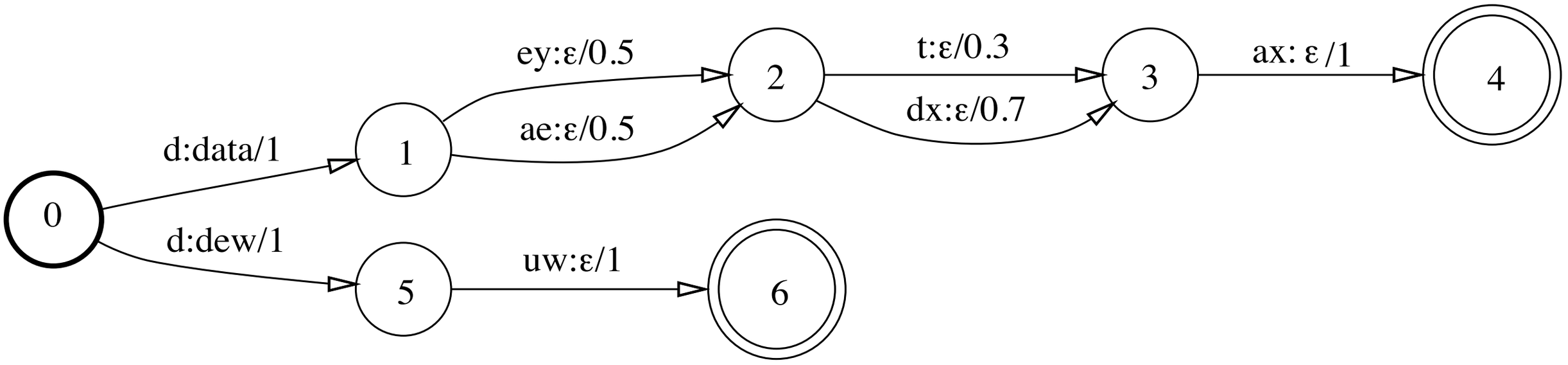
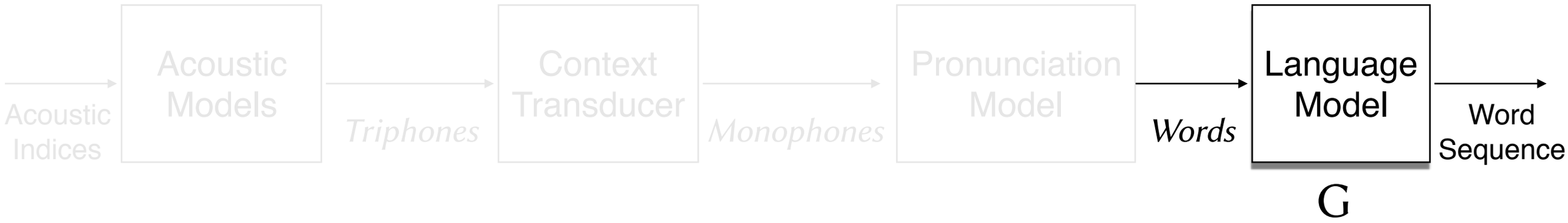
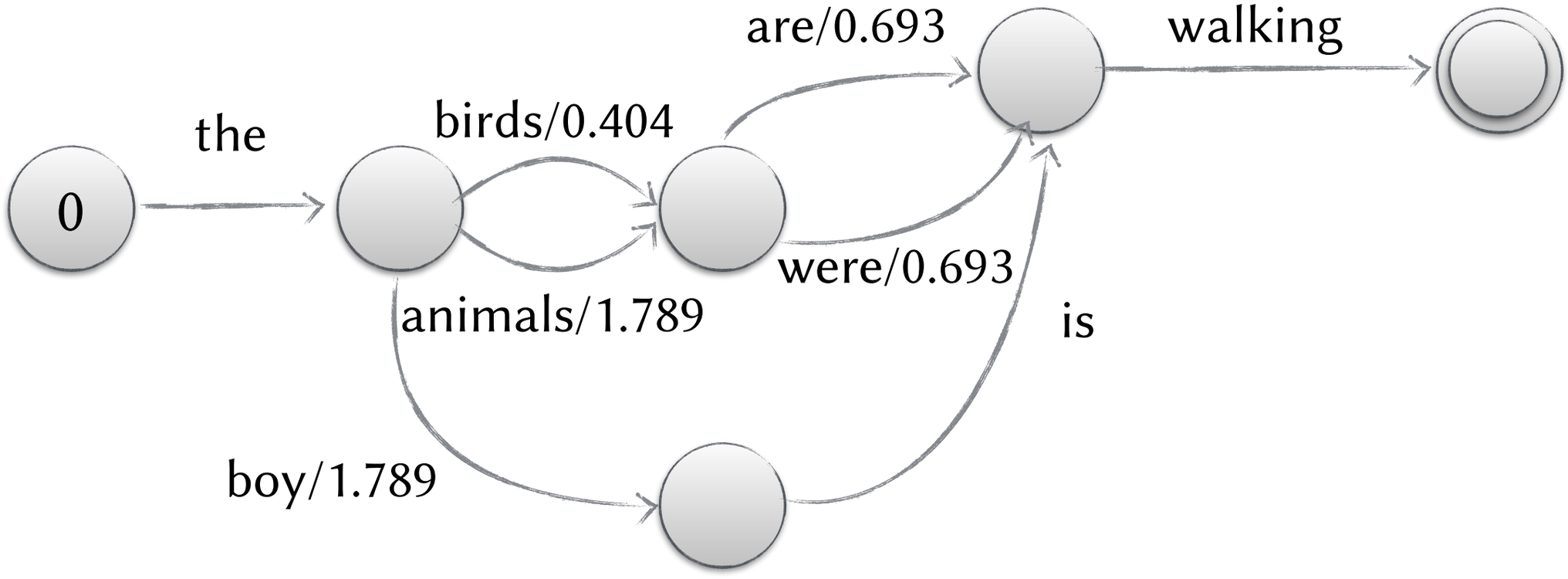


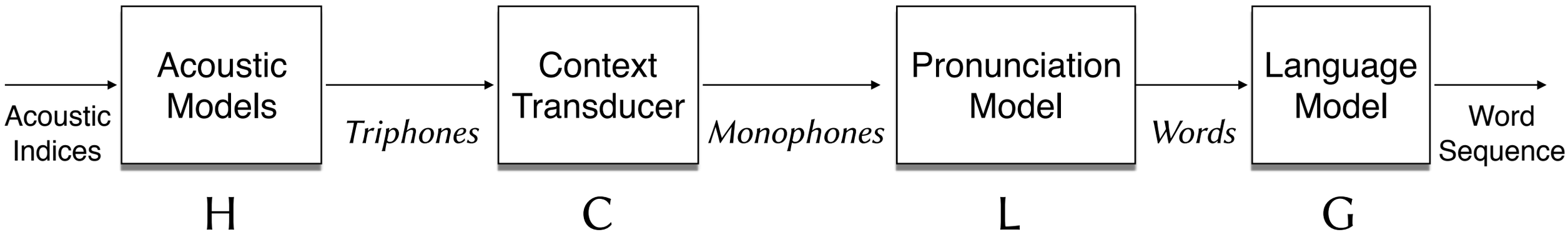
Figure reproduced from "Weighted Finite State Transducers in Speech Recognition", Mohri et al., 2002

WFST-based ASR System





Constructing the Decoding Graph



Decoding graph, D = H ∘ C ∘ L ∘ G

Construct decoding search graph using H ∘ C ∘ L ∘ G that maps acoustic states to word sequences

Carefully construct D using optimization algorithms:

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

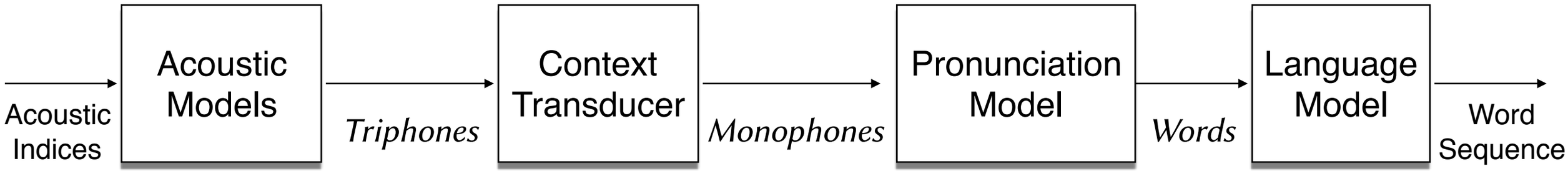
Decode test utterance *O* by aligning acceptor X (corresponding to *O*) with H ∘ C ∘ L ∘ G:

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

where π is a path in the composed FST, $out[\pi]$ 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

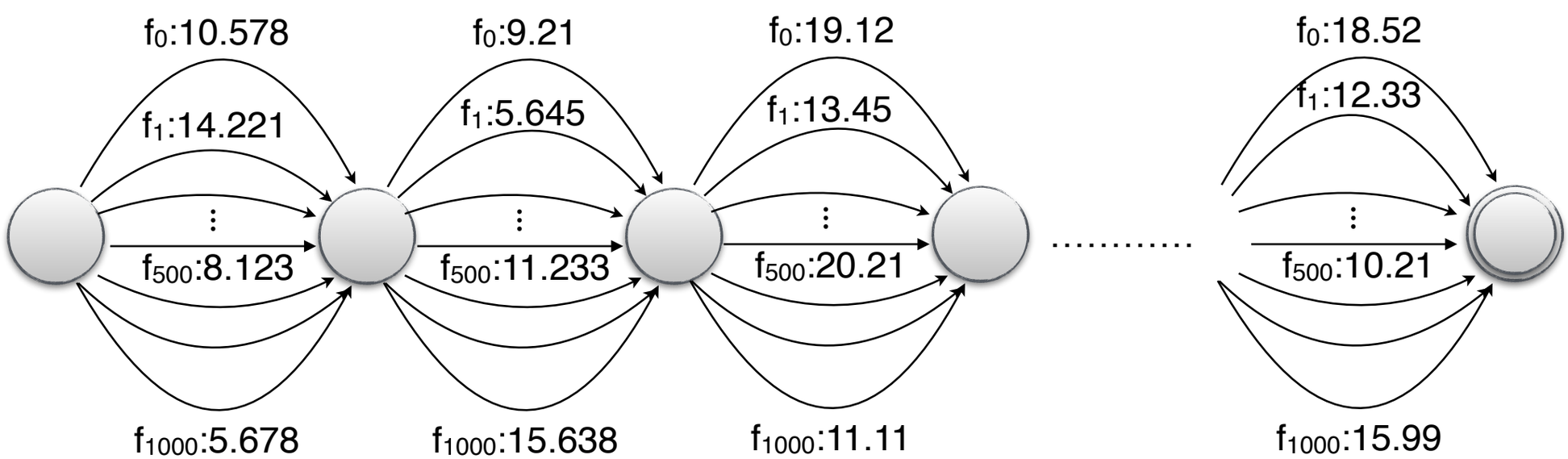


Decode test utterance O by aligning acceptor X (corresponding to O) with H ∘ C ∘ L ∘ G:

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

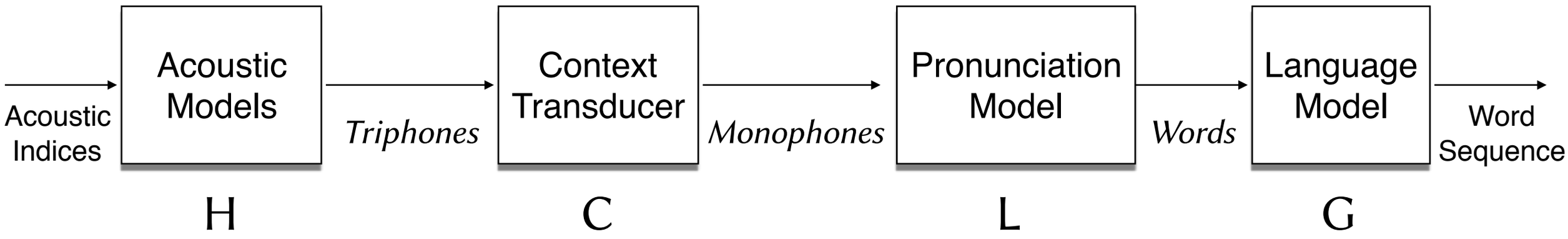
where π is a path in the composed FST, $out[\pi]$ is the output label sequence of π

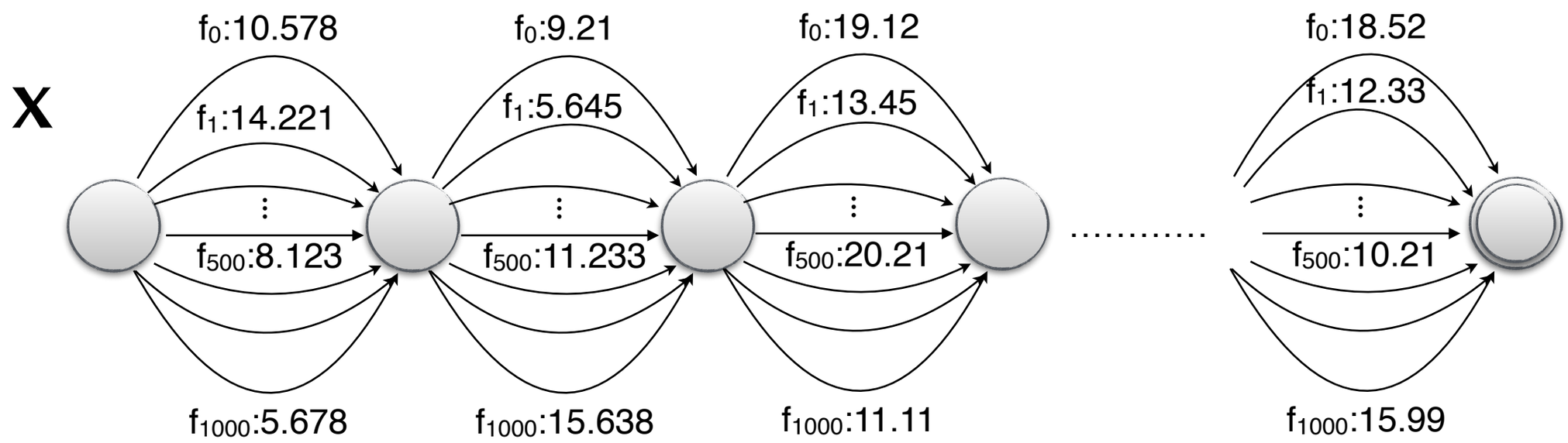
Structure of X (derived from O):



"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 $b_j(o_i)$ (discussed in the next lecture)
- 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)

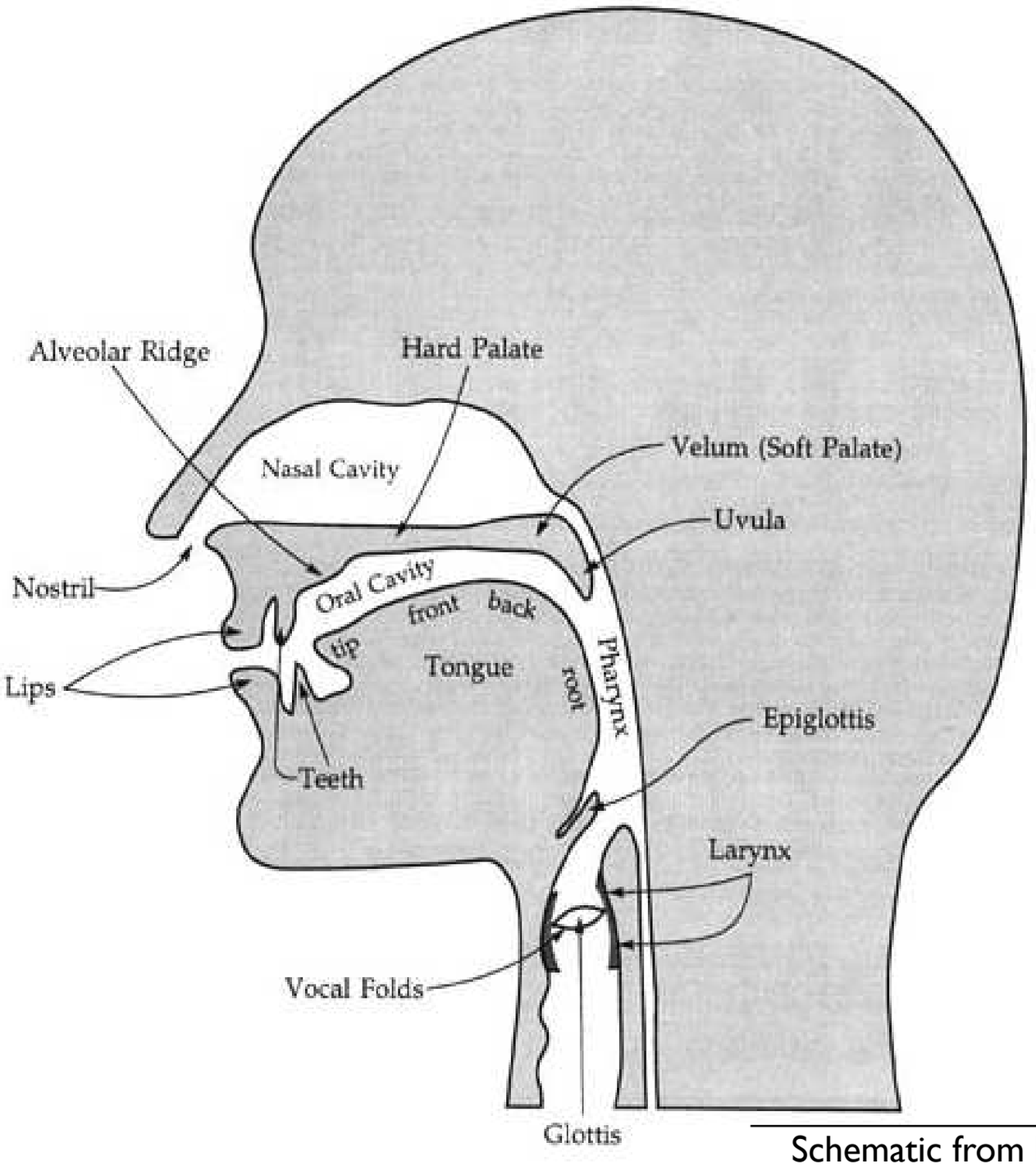
network	states	transitions
G	1,339,664	3,926,010
$L \circ G$	8,606,729	11,406,721
$det(L \circ G)$	7,082,404	9,836,629
$C \circ det(L \circ G))$	7,273,035	10,201,269
$det(H \circ C \circ L \circ G)$	18,317,359	21,237,992

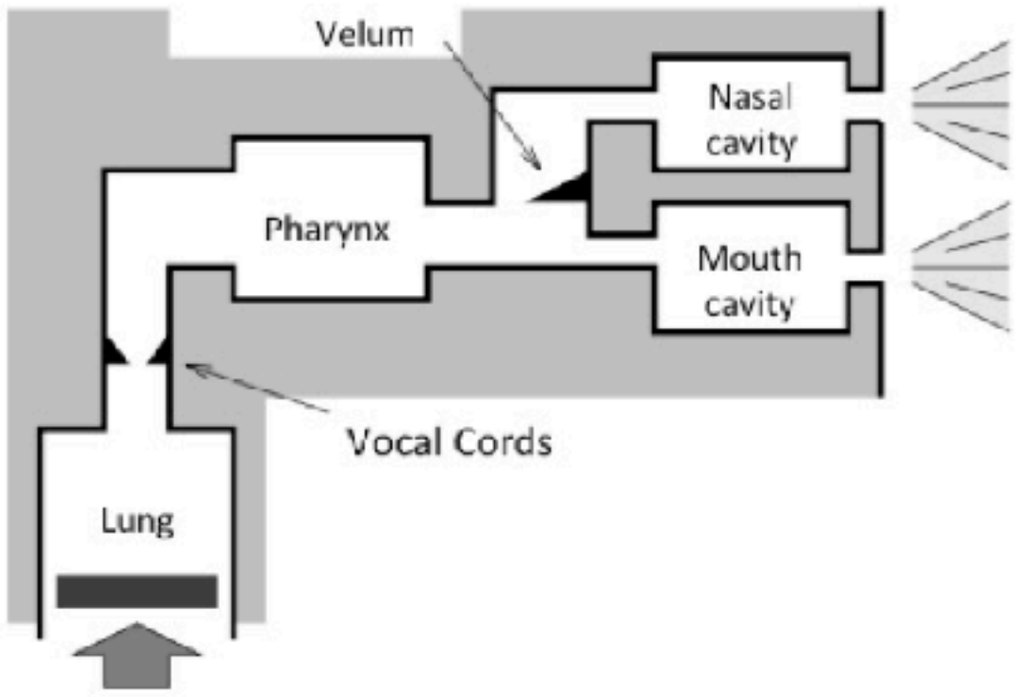
network	x real-time
$C \circ L \circ G$	12.5
$C \circ det(L \circ G)$	1.2
$det(H \circ C \circ L \circ G)$	1.0
$push(min(F))$	0.7

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

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: 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

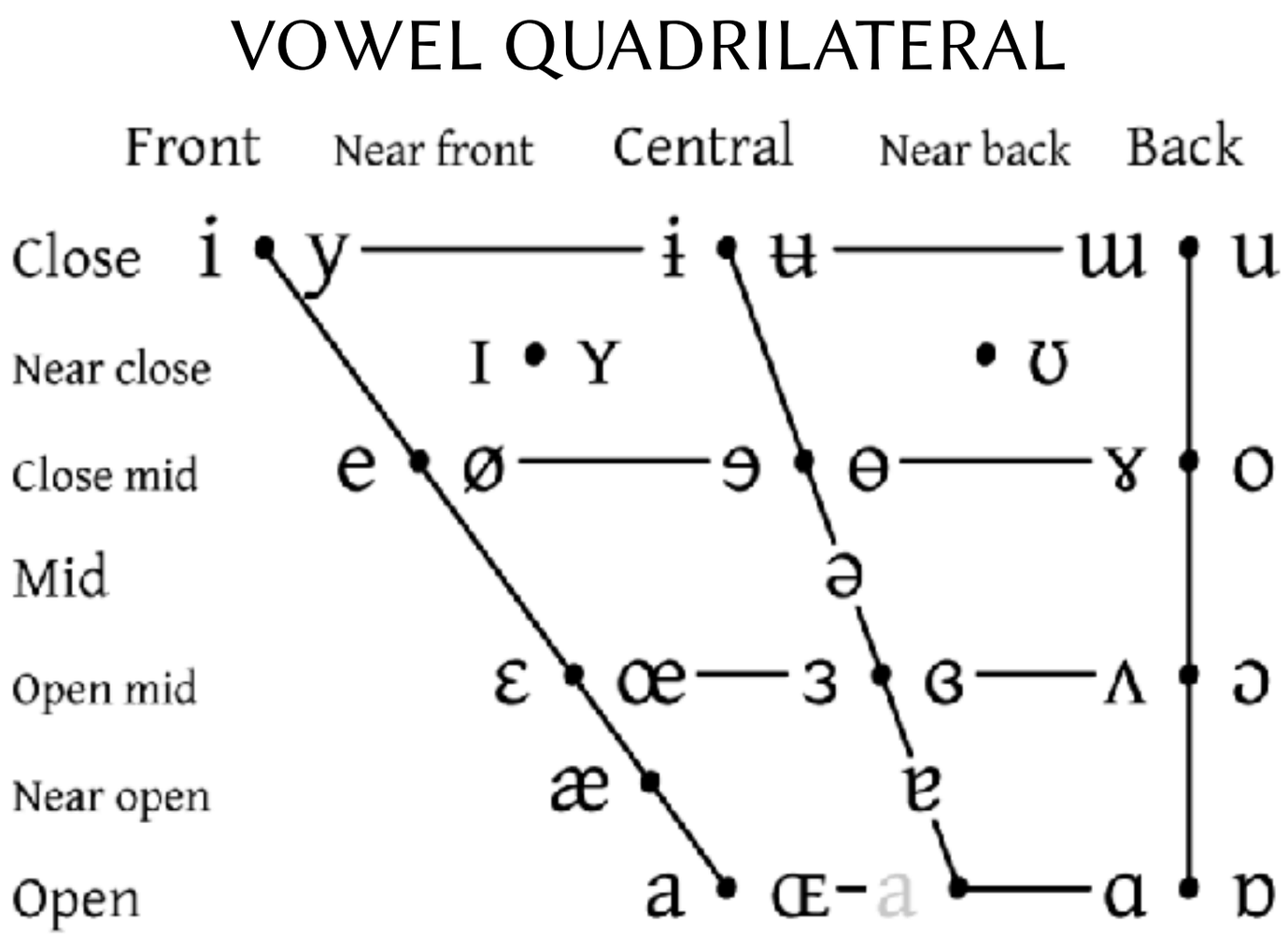


Image from https://en.wikipedia.org/wiki/File:IPA_vowel_chart_2005.png

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

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.)

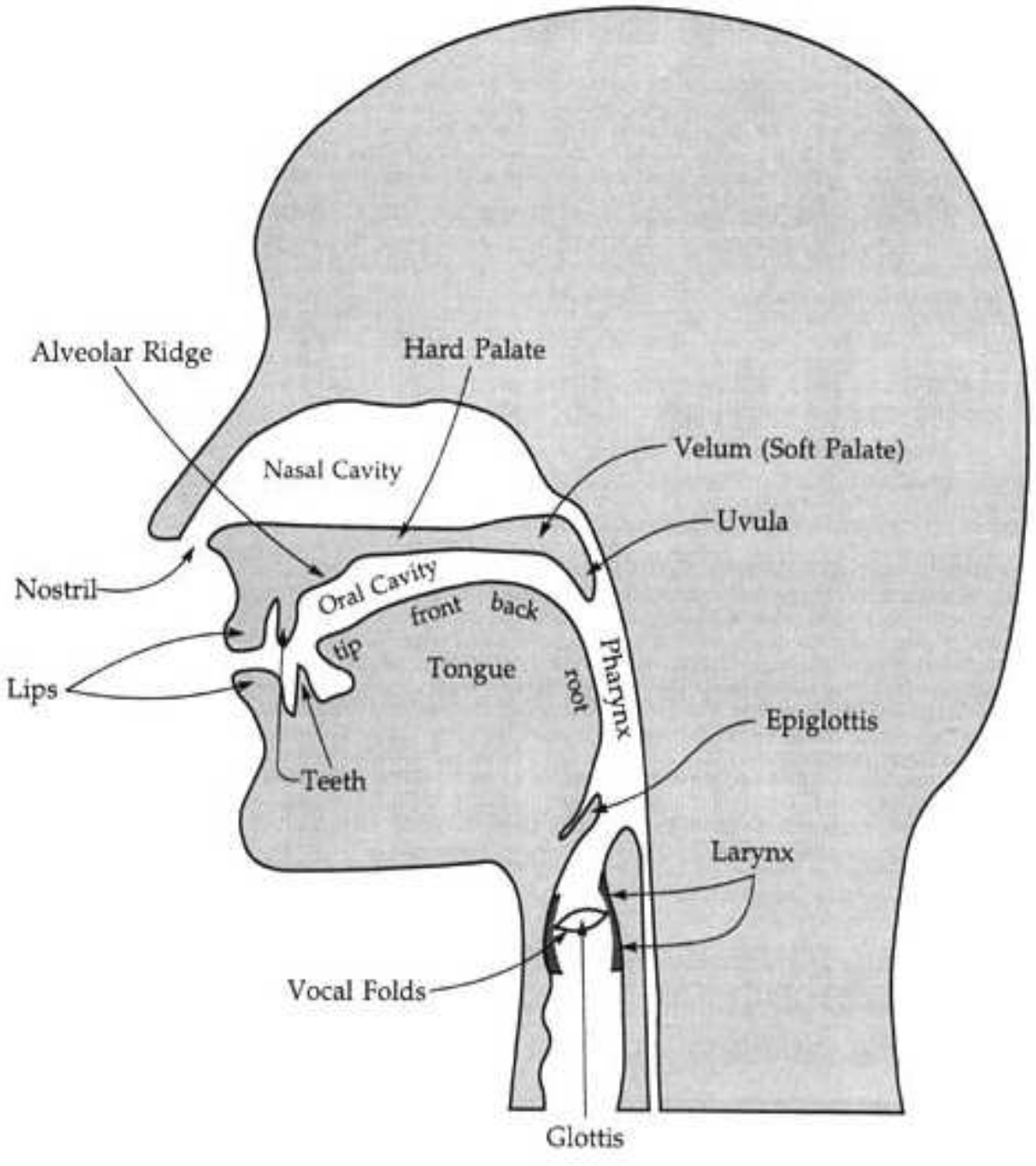
Consonants (voicing/place/manner)

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

Voiced/Unvoiced Sounds

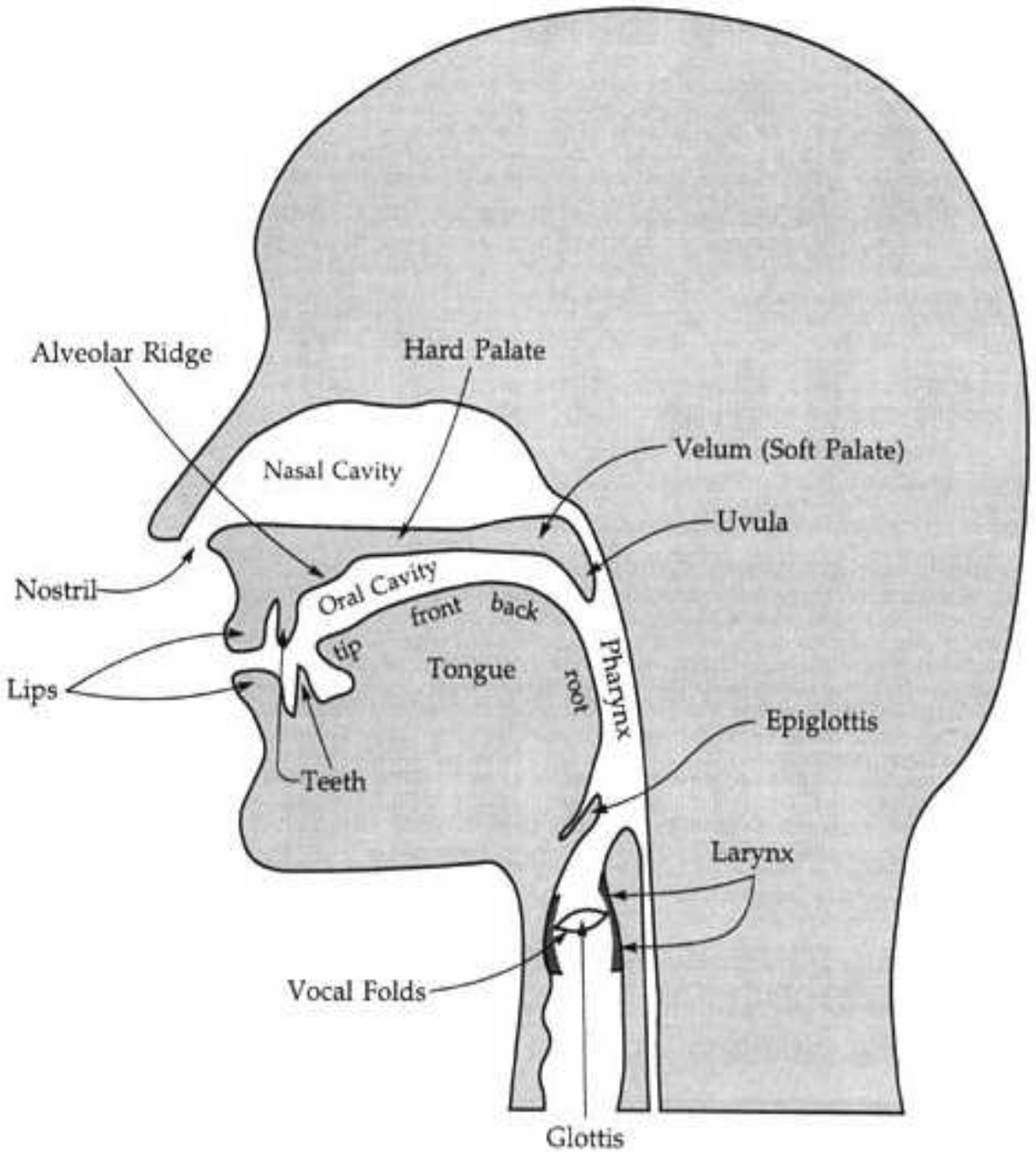
- Sounds made with vocal cords vibrating: **voiced**
  - E.g. /g/, /d/, etc.
    - All English vowel sounds are voiced
- Sounds made without vocal cord vibration: **voiceless**
  - E.g. /k/, /t/, etc.

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)

Place of articulation



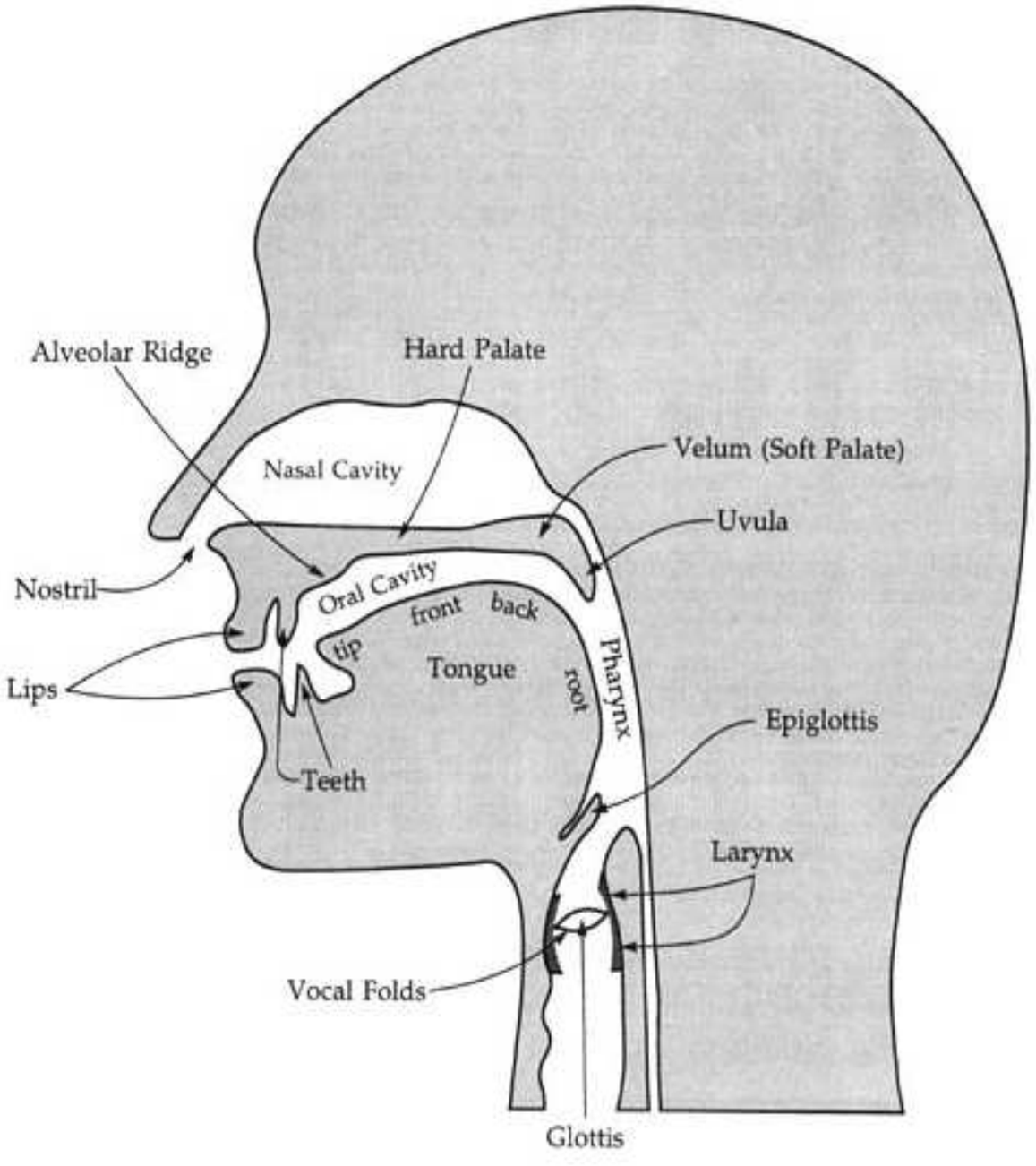
- Alveolar (tongue tip on alveolar ridge)
  [n],[t],[s],etc.

- Palatal (tongue up close to hard palate)
  [sh], [ch] (palato-alveolar)
  [y], etc.

- Velar (tongue near velum)
  [k], [g], etc.

- Glottal (produced at larynx)
  [h], glottal stops.

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