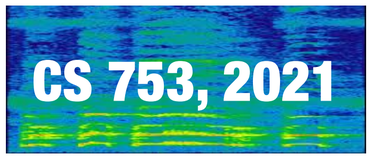
WFSTs in ASR

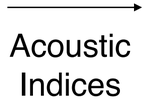
Lecture 4b



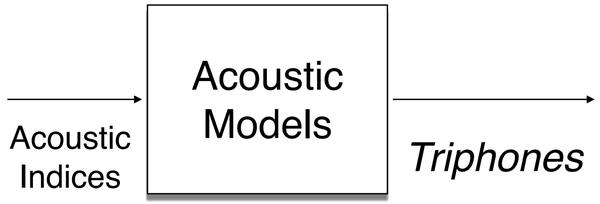
Instructor: Preethi Jyothi, IITB

WFST-based ASR System

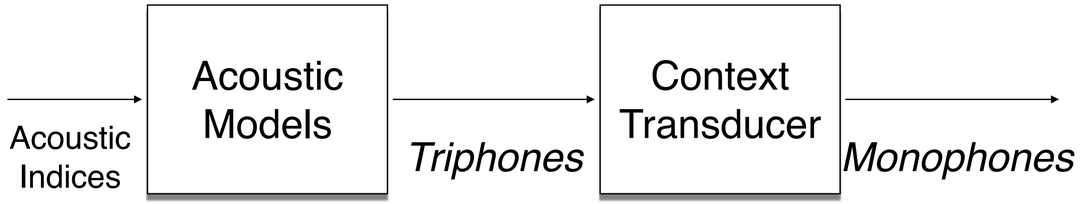
WFST-based ASR System



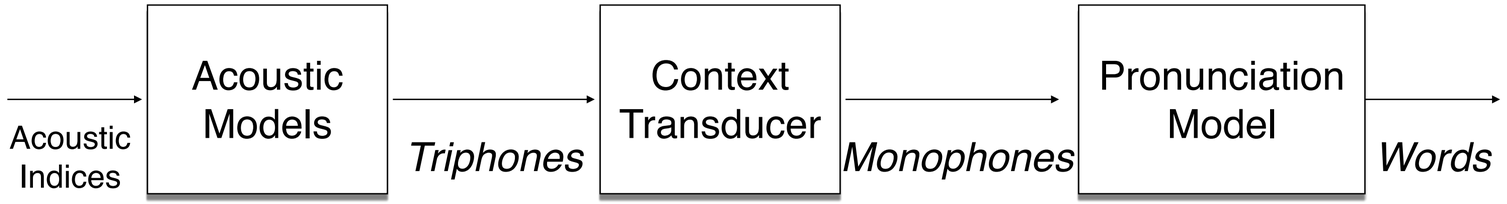
WFST-based ASR System



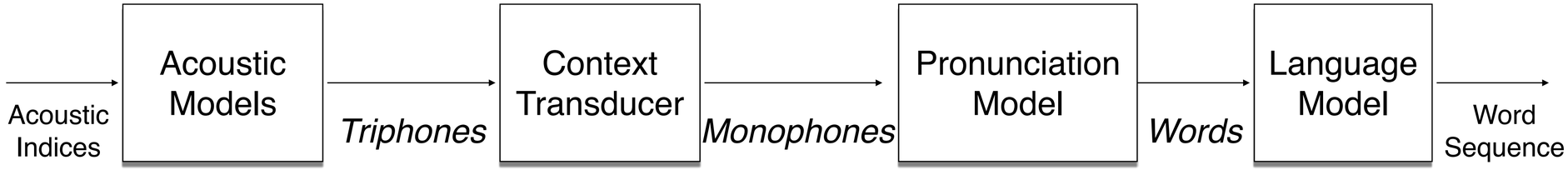
WFST-based ASR System



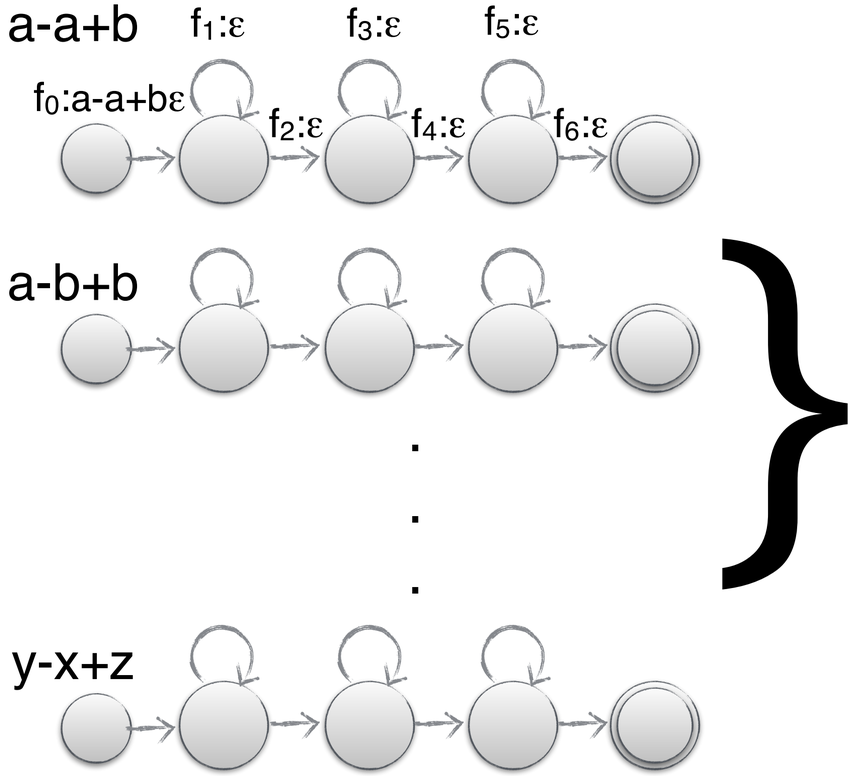
WFST-based ASR System



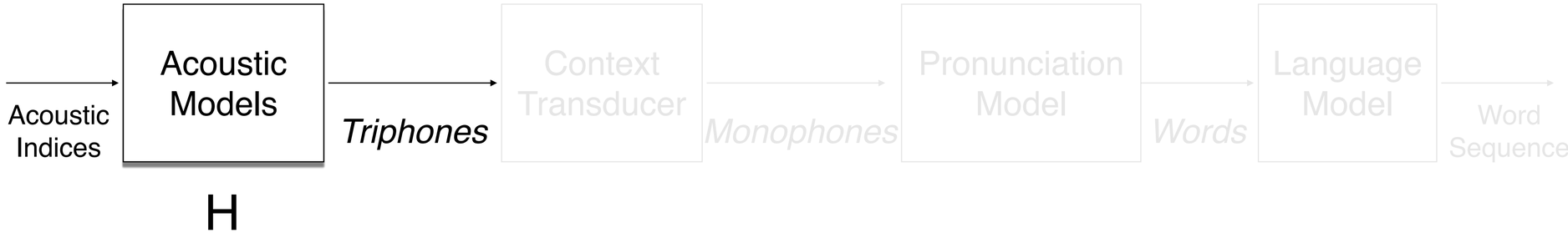
WFST-based ASR System

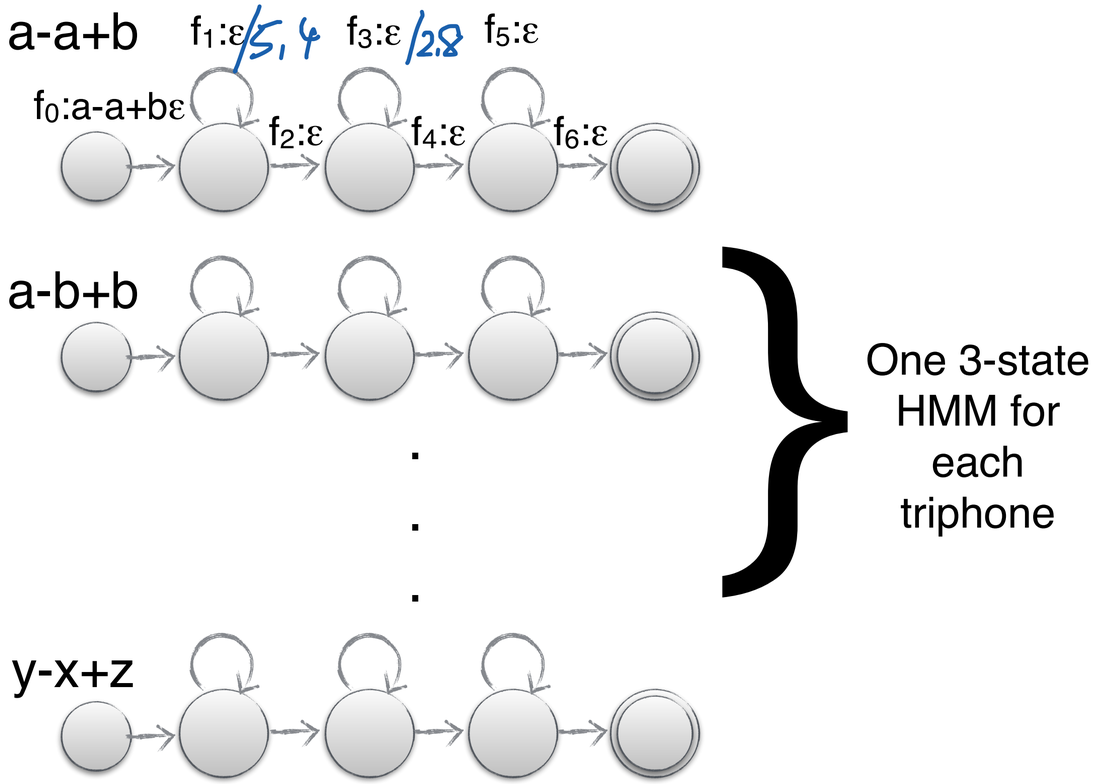


WFST-based ASR System



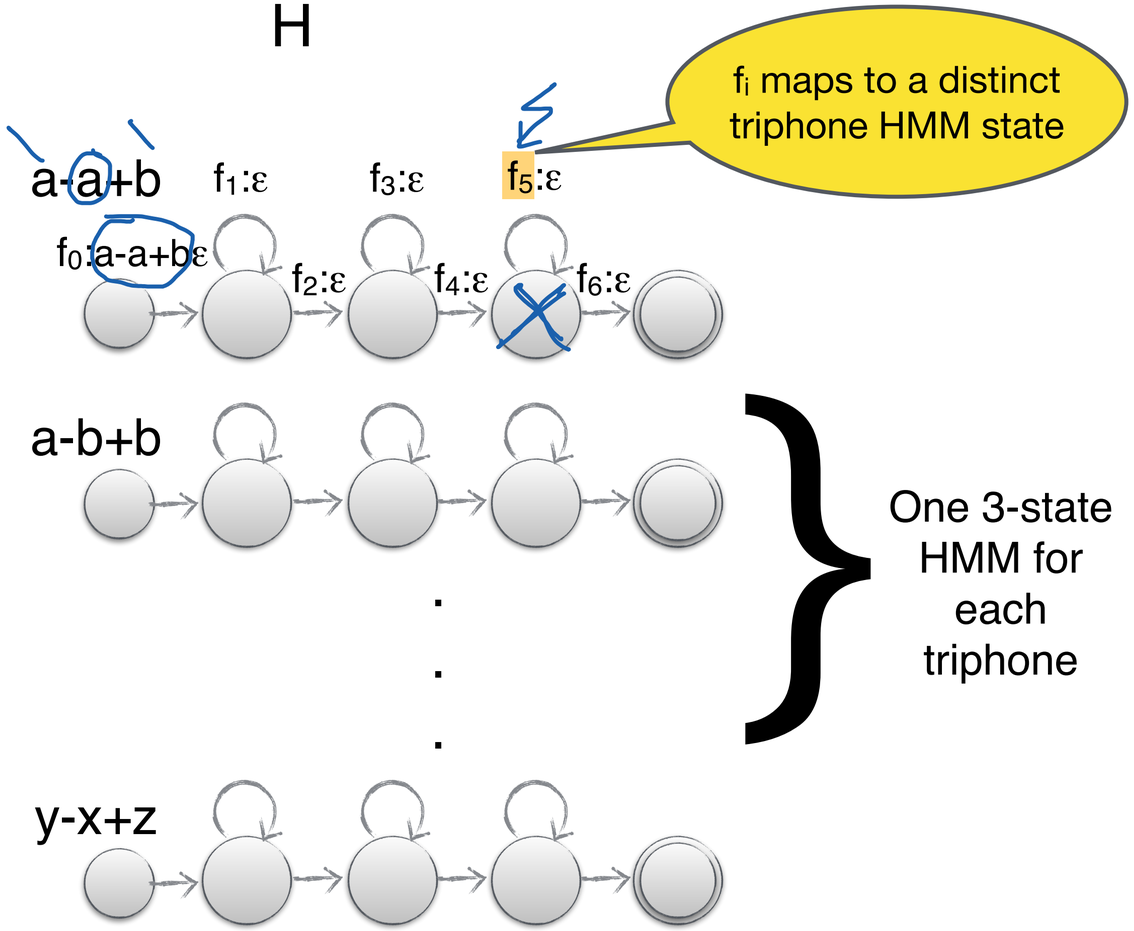
WFST-based ASR System



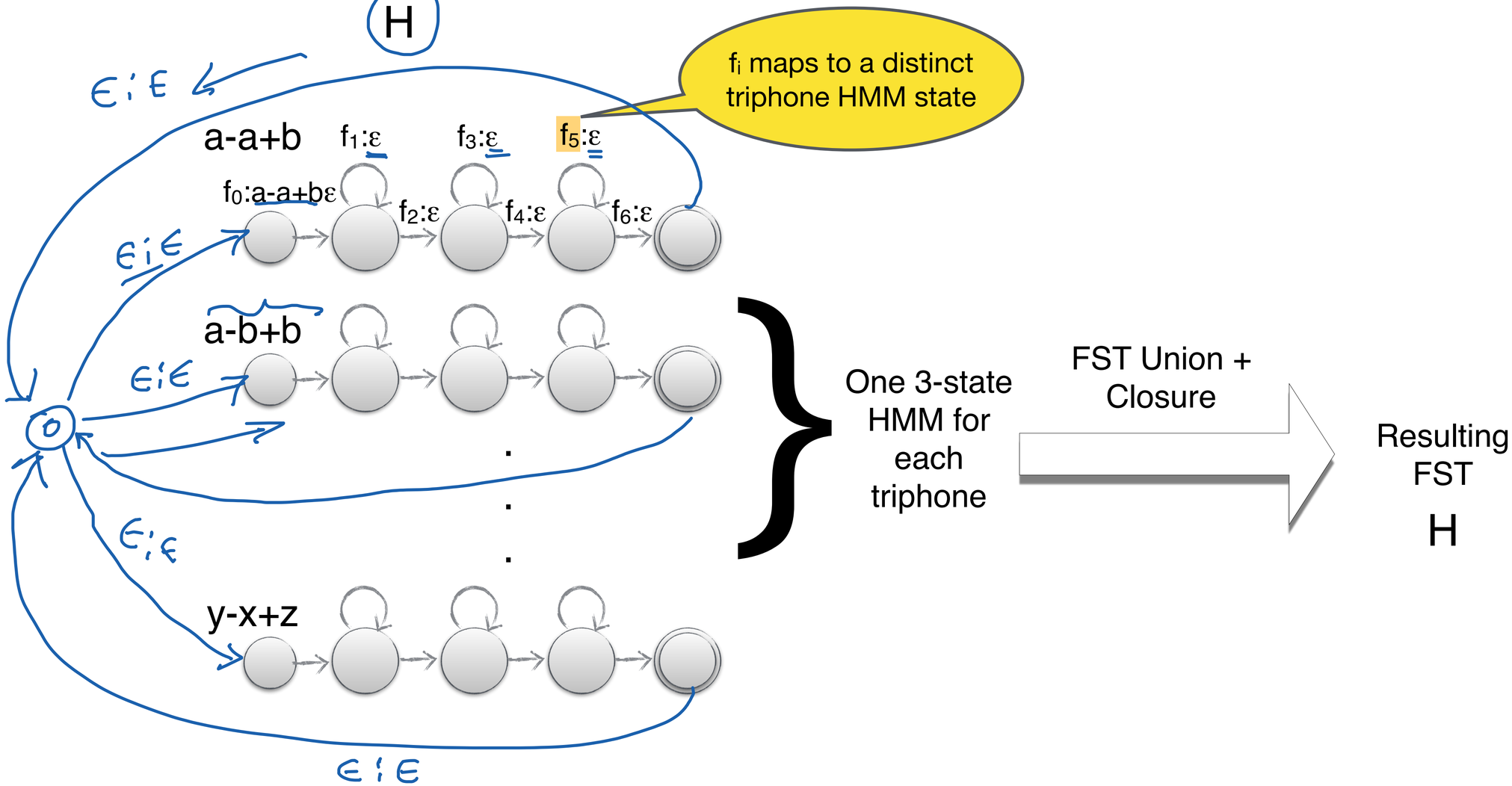


WFST-based ASR System

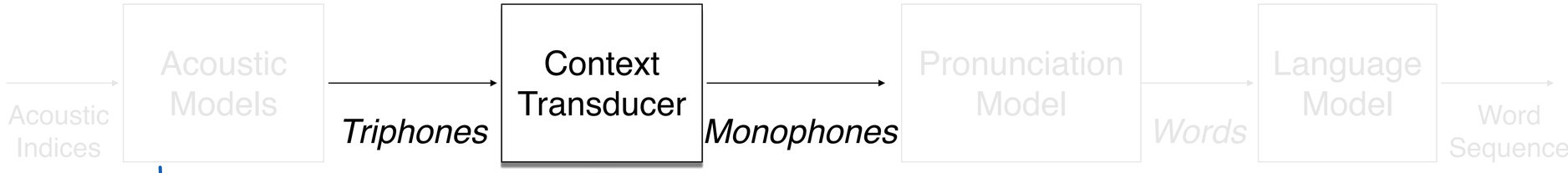
H



WFST-based ASR System



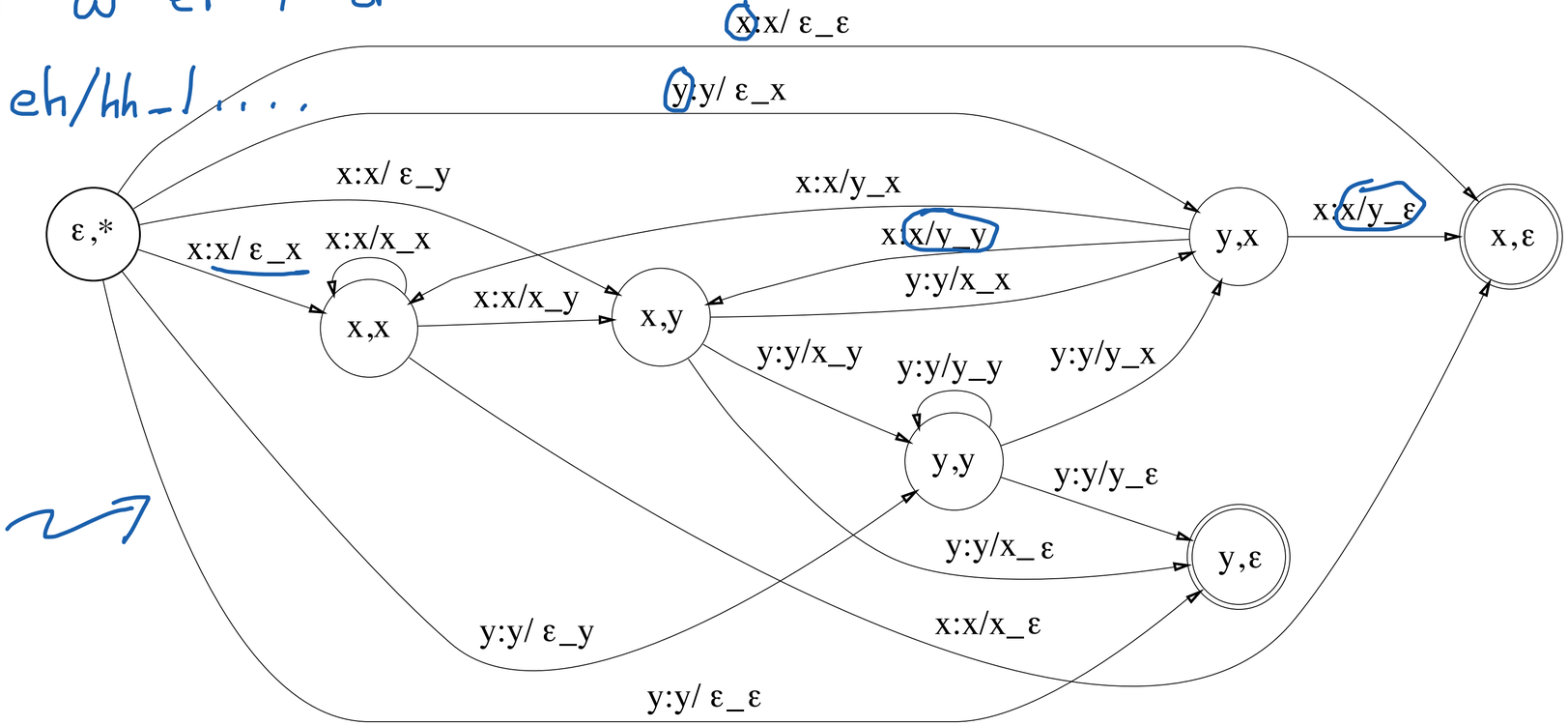
WFST-based ASR System



hello world
hh eh l ow w er l d
hh/ϵ_eh eh/hh_l ...

C

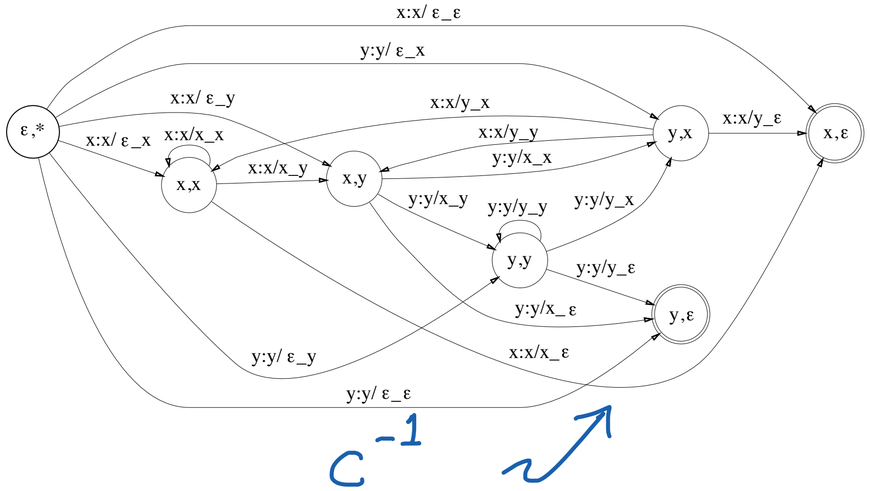
i:o

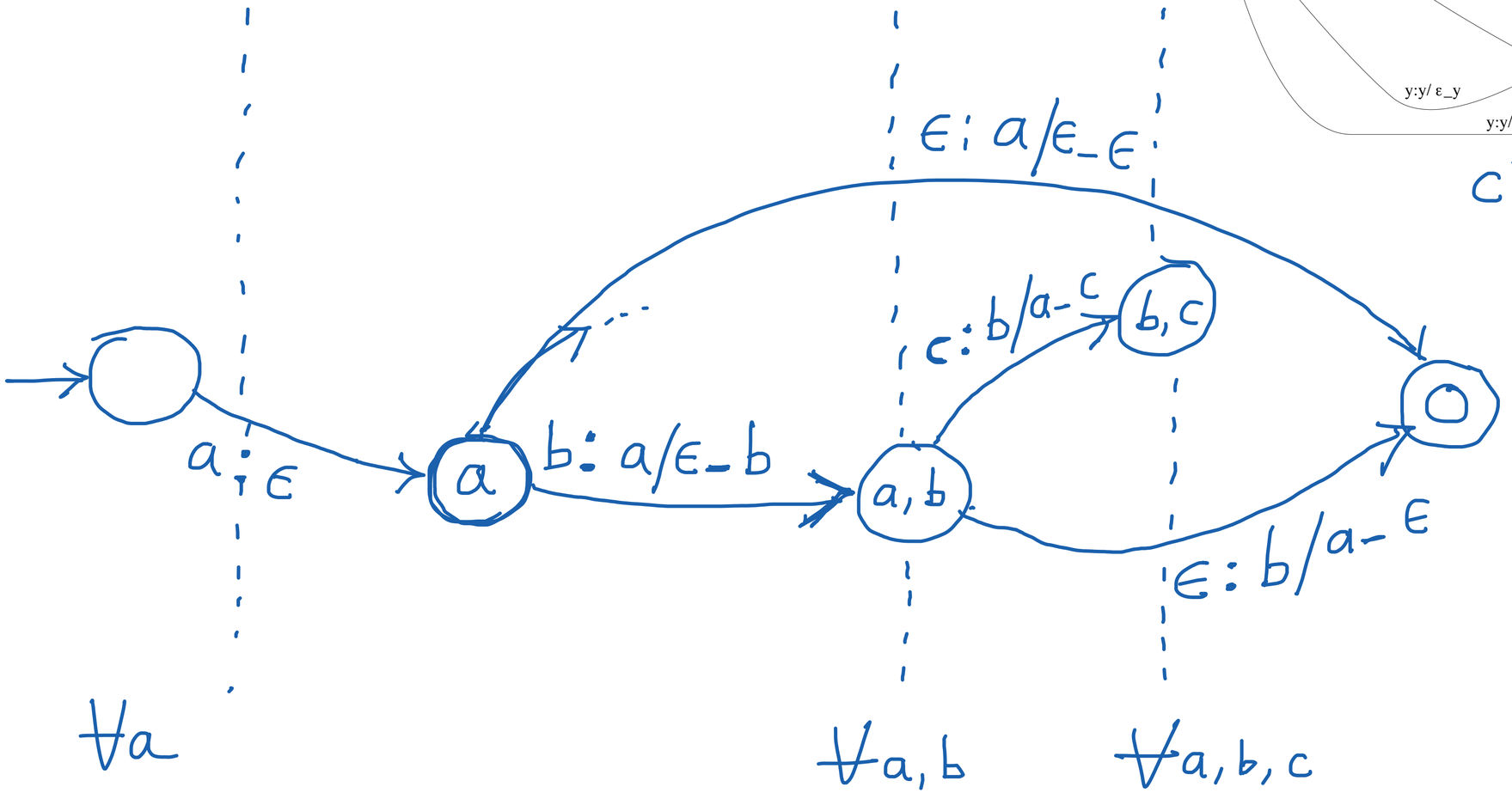


C^{-1}: Arc labels: “monophone : phone / left-context_right-context”

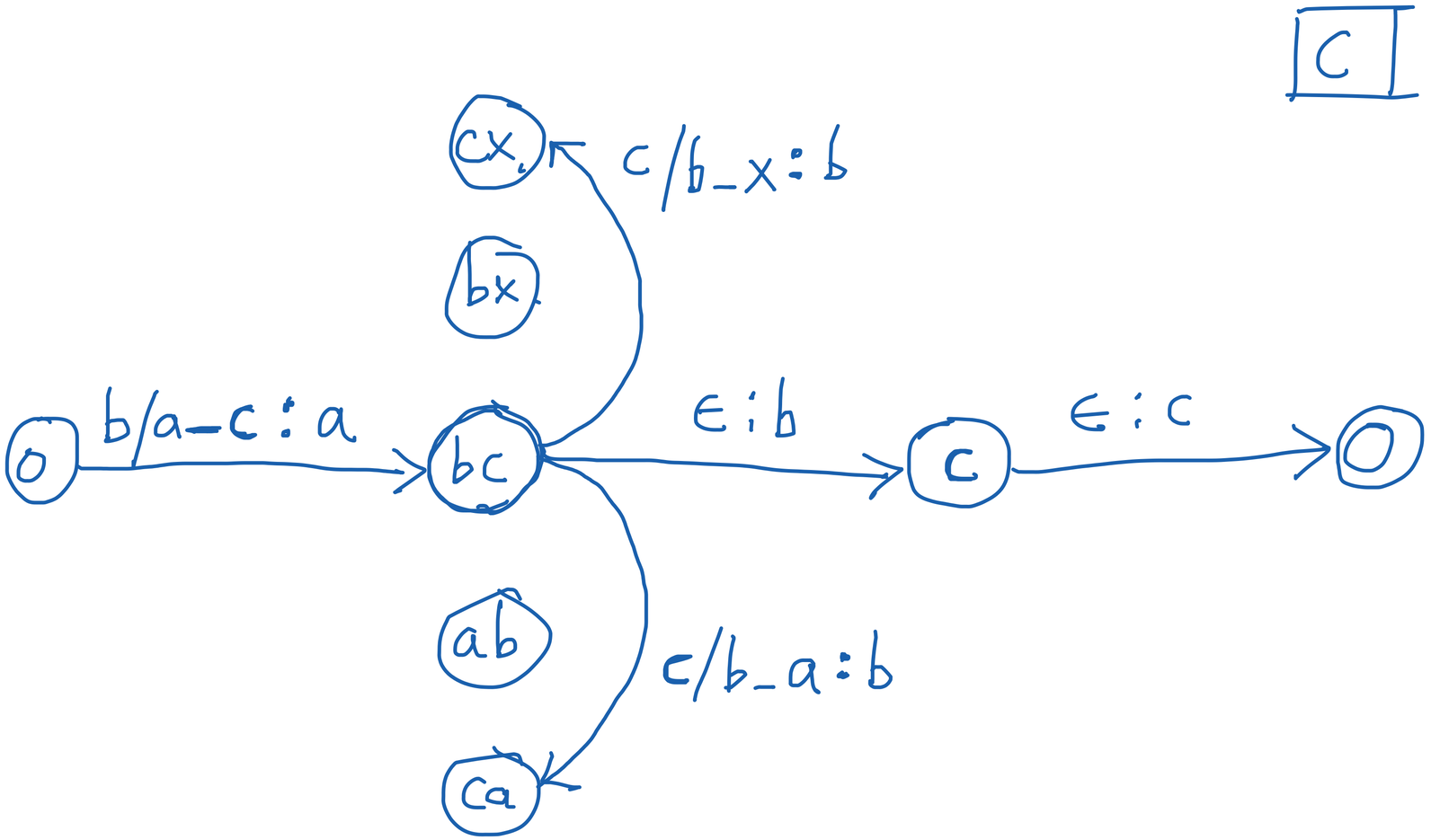
Figure reproduced from “Weighted Finite State Transducers in Speech Recognition”, Mohri et al., 2002

C: Context Dependency Transducer





C: Context Dependency Transducer



WFST-based ASR System

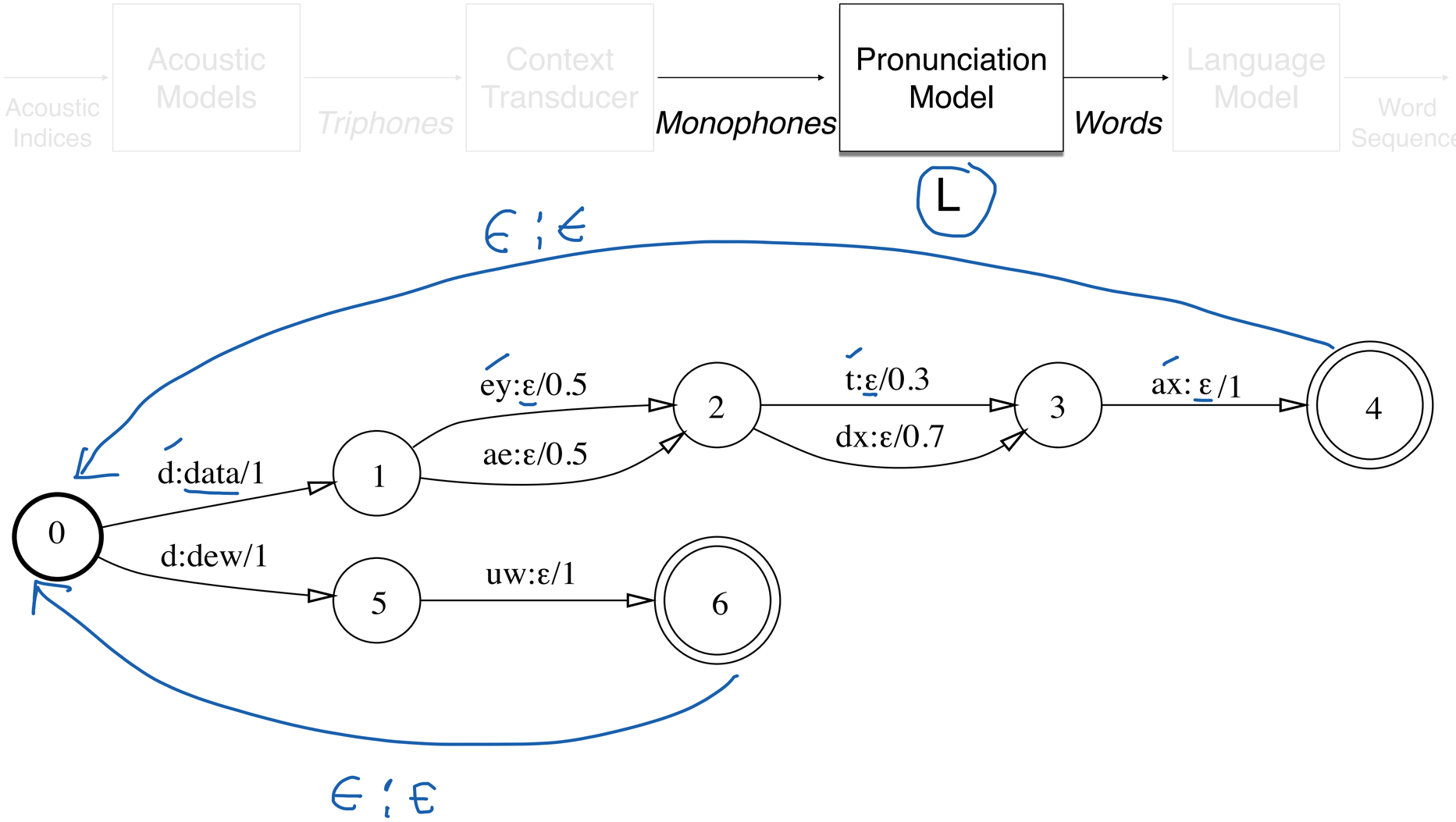
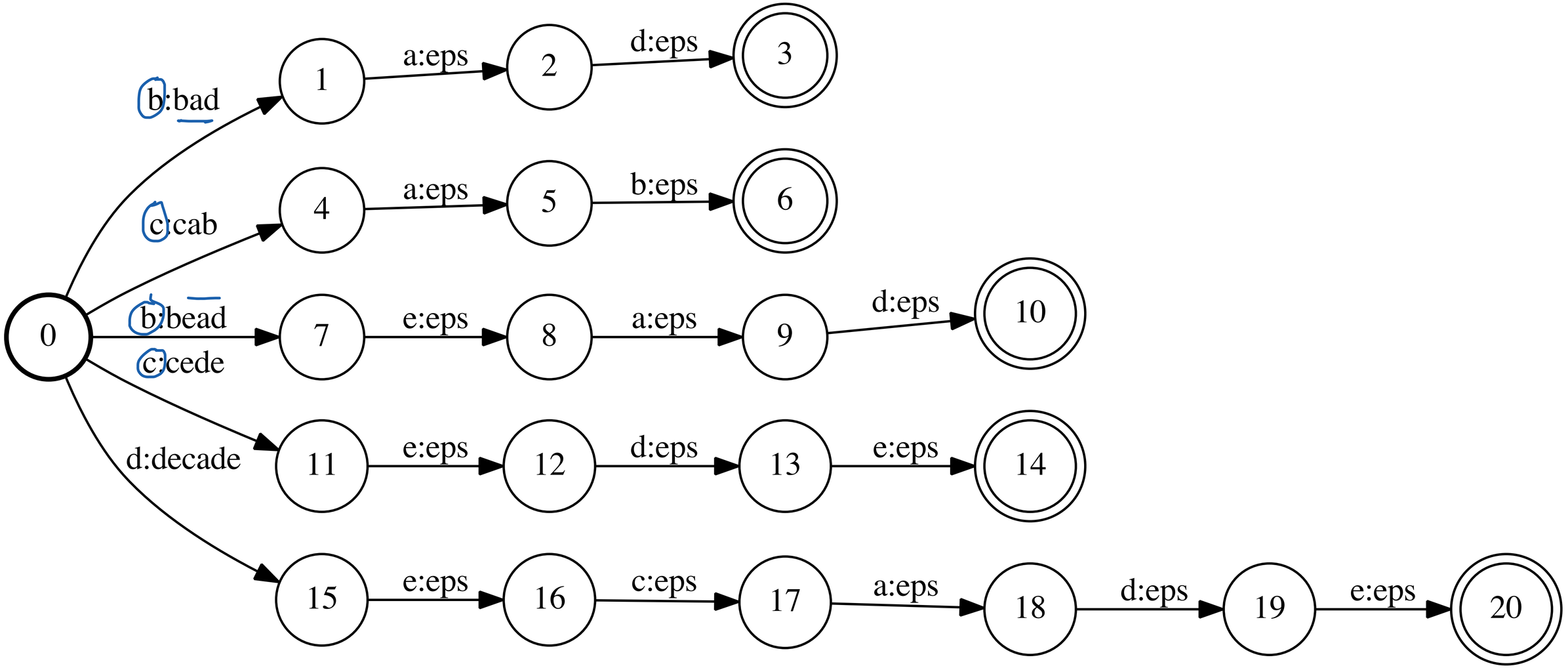


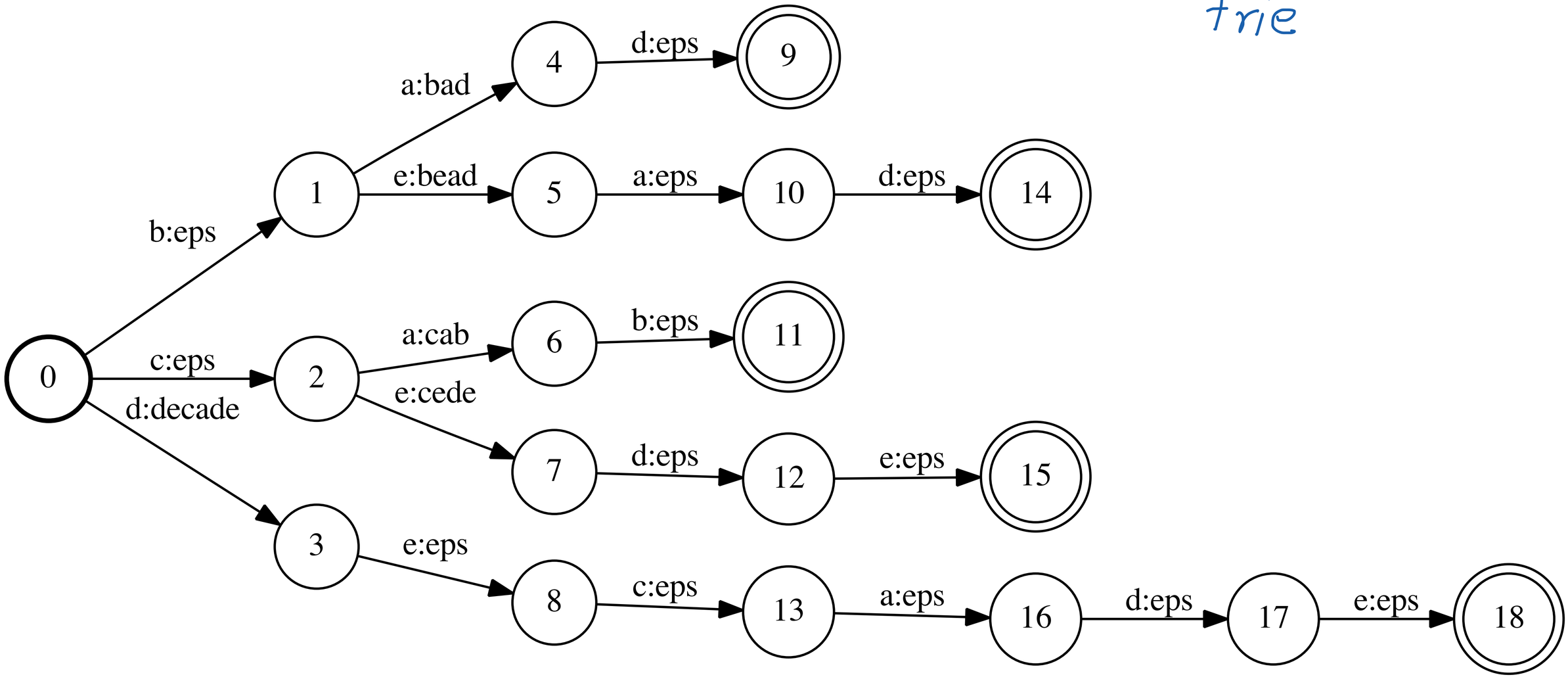
Figure reproduced from "Weighted Finite State Transducers in Speech Recognition", Mohri et al., 2002

Example: Dictionary WFST

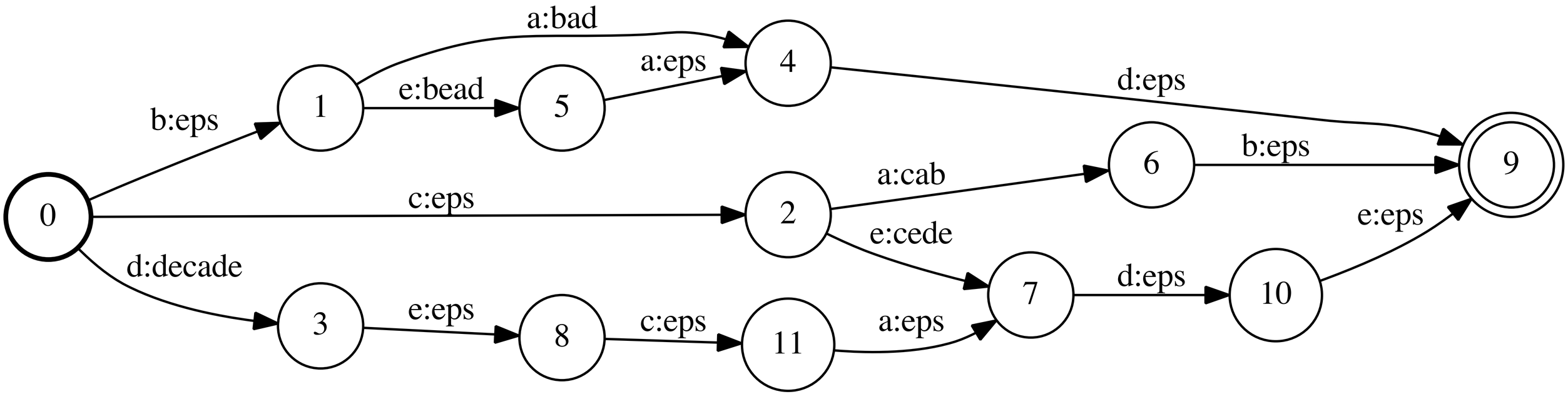


Determinized Dictionary WFST

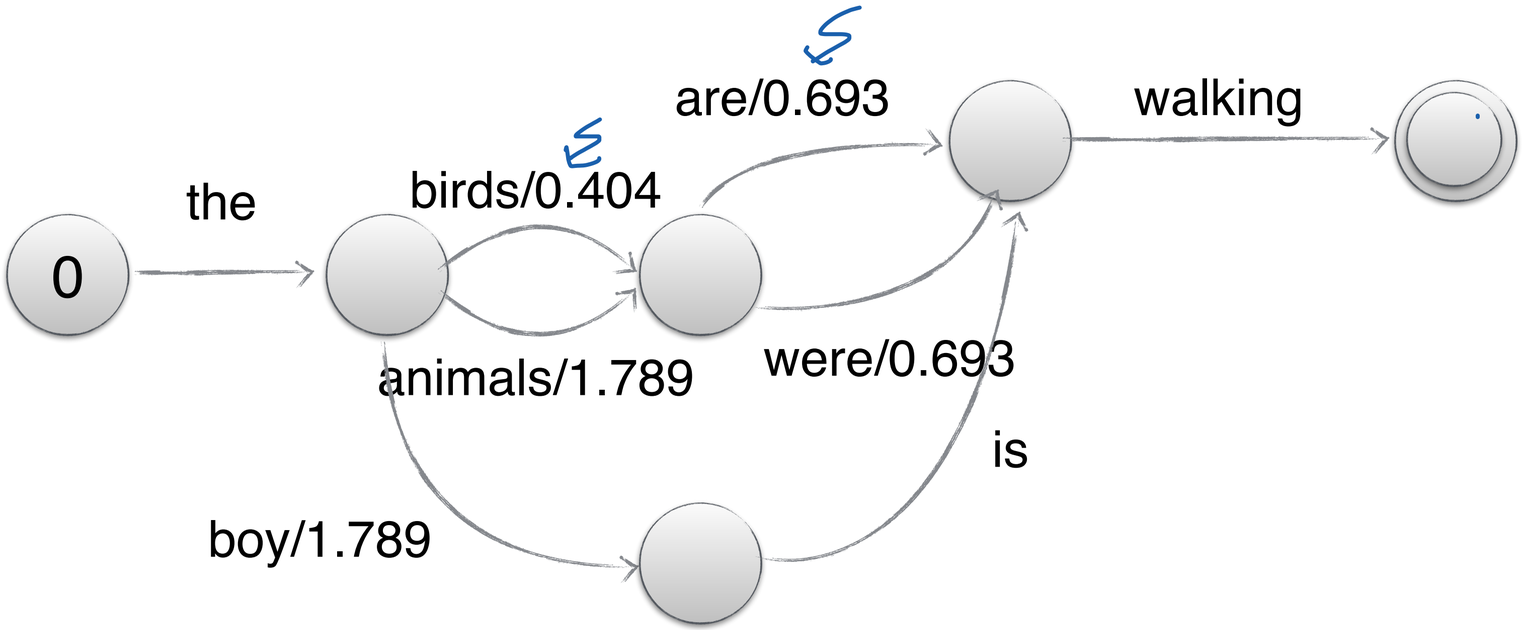
trie



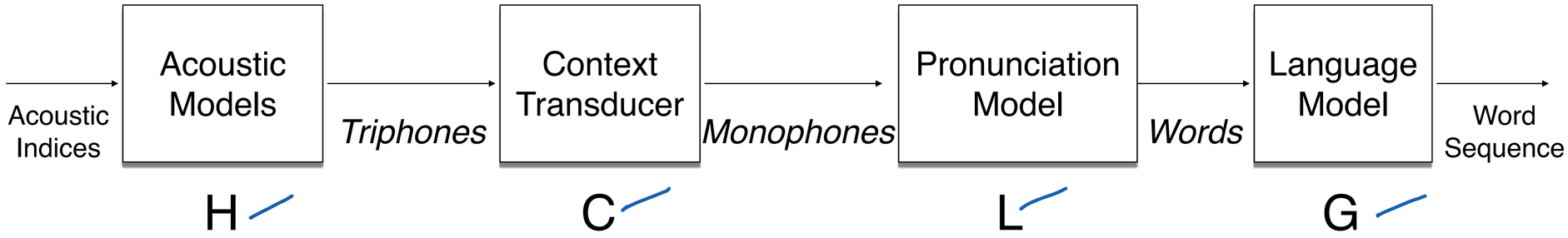
Minimized Dictionary WFST



WFST-based ASR System



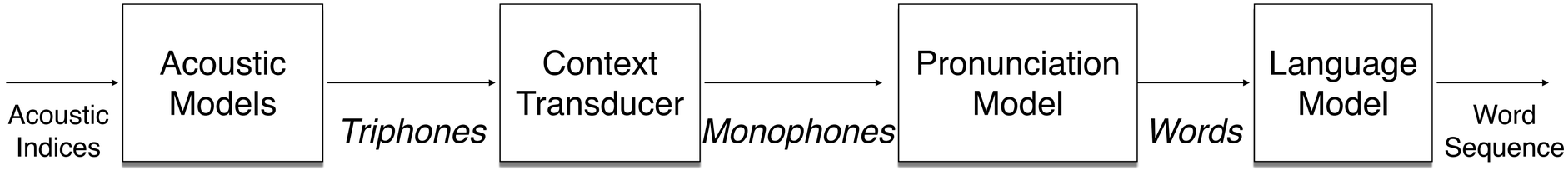
Decoding



Carefully construct a decoding graph D using optimization algorithms:

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

Decoding



H C L G

Carefully construct a decoding graph D using optimization algorithms:

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

homophones
E.g. read vs. reed
new vs. knew

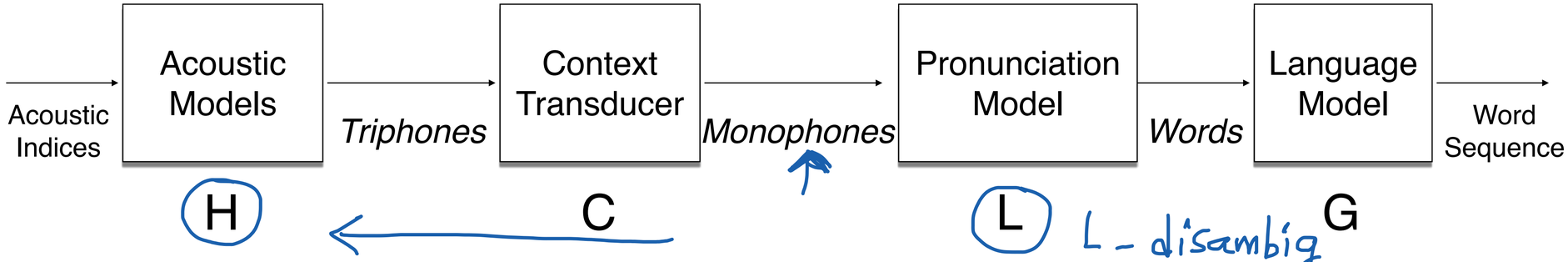
Need to add *disambiguation symbols* to L in order to make L ∘ G determinizable!

(auxilliary)

When are disambiguation symbols needed?

1. When you deal with homophones i.e., same pronunciation, different words.
2. When pronunciation of a word is a prefix of another. caf caterpillar

Decoding



Carefully construct a decoding graph D using optimization algorithms:

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

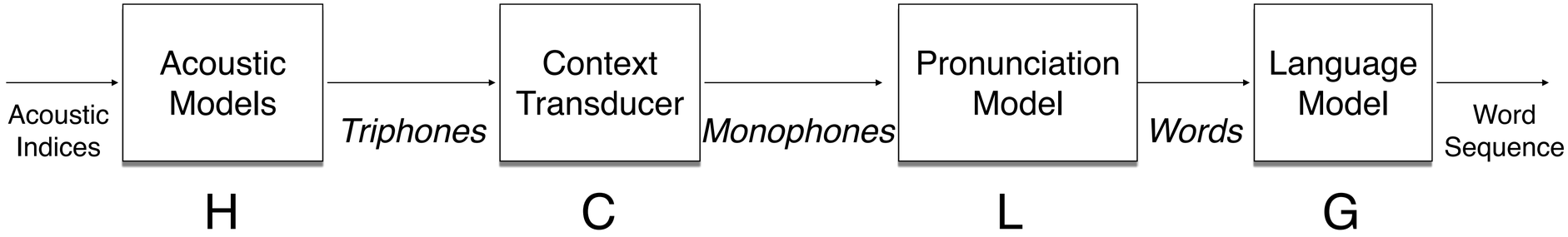
Need to add *disambiguation symbols* to L in order to make L ∘ G determinizable!

When are disambiguation symbols needed?

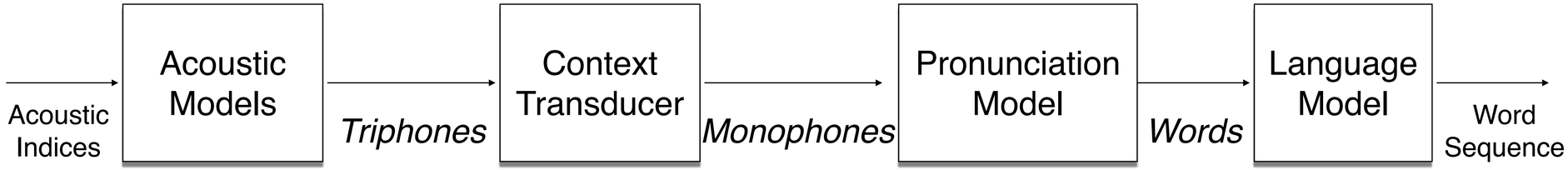
1. When you deal with homophones i.e., same pronunciation, different words.
2. When pronunciation of a word is a prefix of another.

book	b uh k #1
books	b uh k s
right	r ay t #1
write	r ay t #2
⋮	

Decoding


Acoustic Indices
Acoustic Models
Triphones
Context Transducer
Monophones
Pronunciation Model
Words
Language Model
Word Sequence
H
C
L
G

Decoding

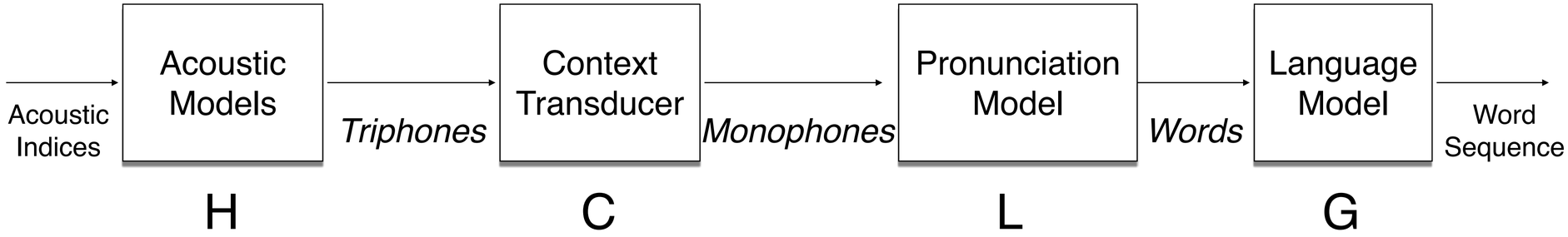


H C L G

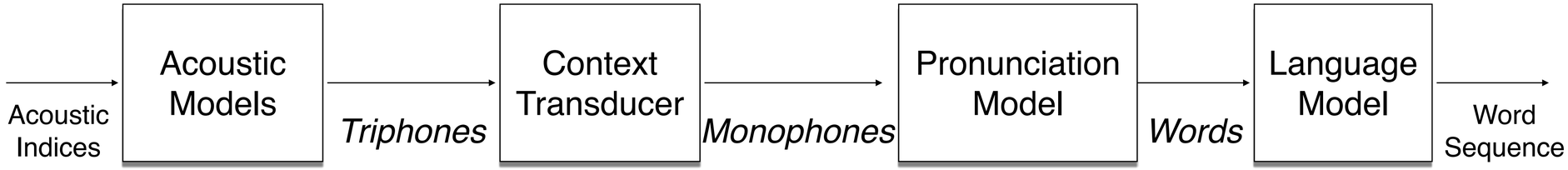
Carefully construct a decoding graph D using optimization algorithms:

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

Decoding


Acoustic Indices
Acoustic Models
Triphones
Context Transducer
Monophones
Pronunciation Model
Words
Language Model
Word Sequence
H
C
L
G

Decoding

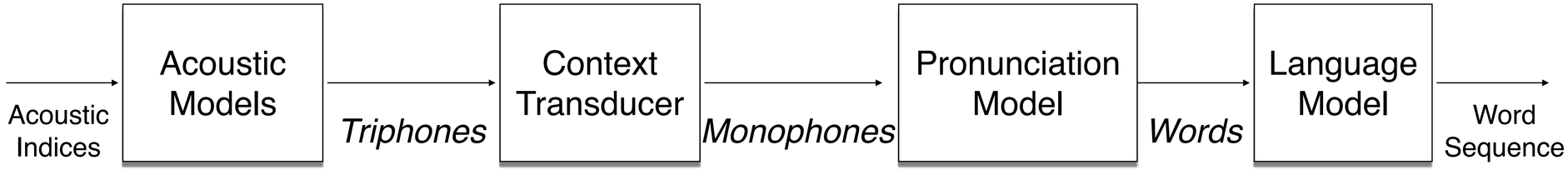


H C L G

Carefully construct a decoding graph D using optimization algorithms:

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

Decoding



H C L G

Carefully construct a decoding graph D using optimization algorithms:

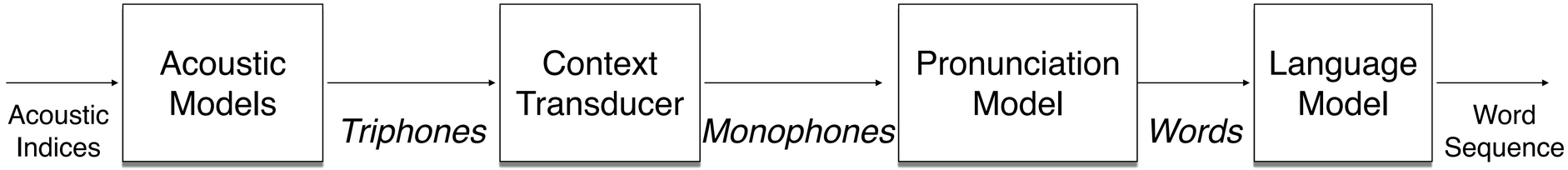
$$D = \min(\det(H \circ \det(C \circ \det(L \circ G))))$$

$O_1, \ldots, O_T$

Given a test utterance *O, how do I decode it?*

Assuming ample compute, first construct the following machine X from O.

Decoding

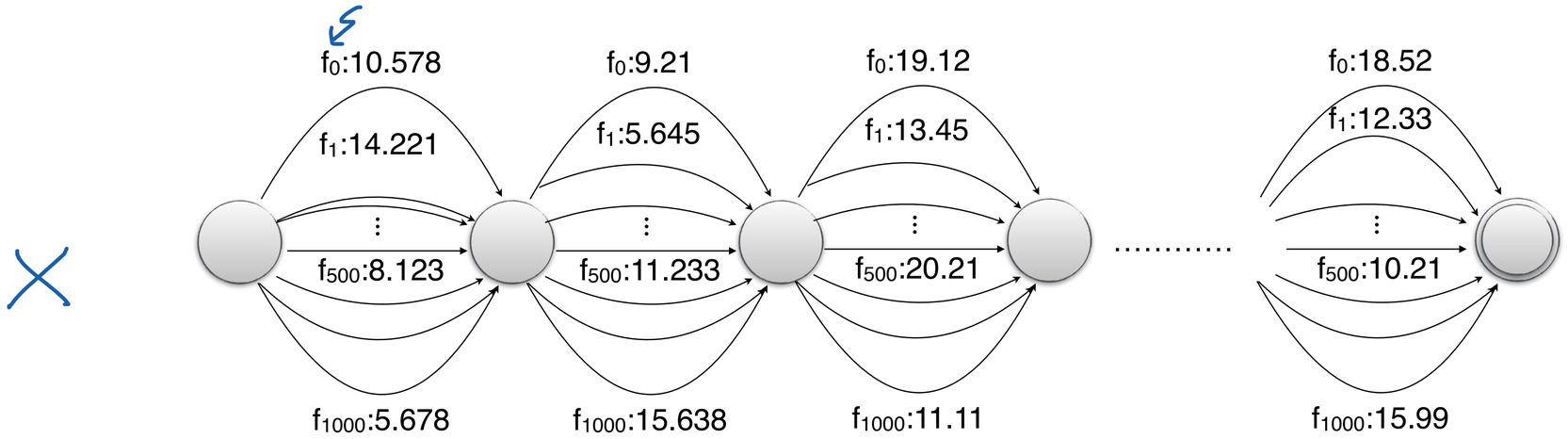


Carefully construct a decoding graph D using optimization algorithms:

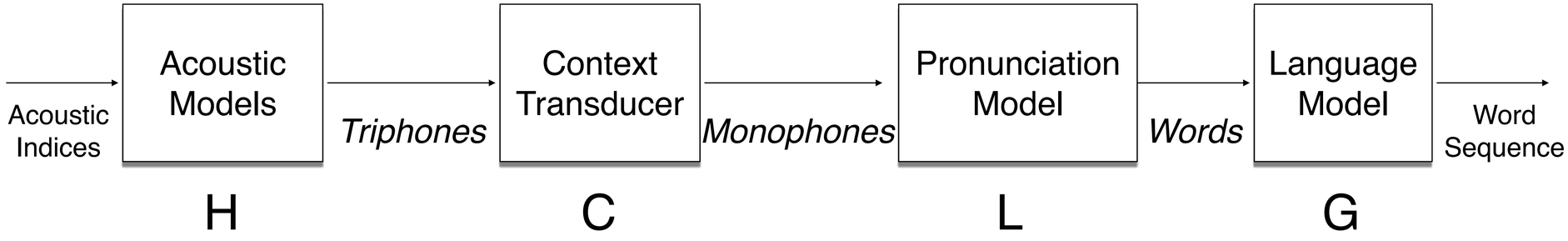
$$D = \min(\det(H \circ \det(C \circ \det(L \circ G))))$$

Given a test utterance *O, how do I decode it?*

Assuming ample compute, first construct the following machine X from O.



Decoding

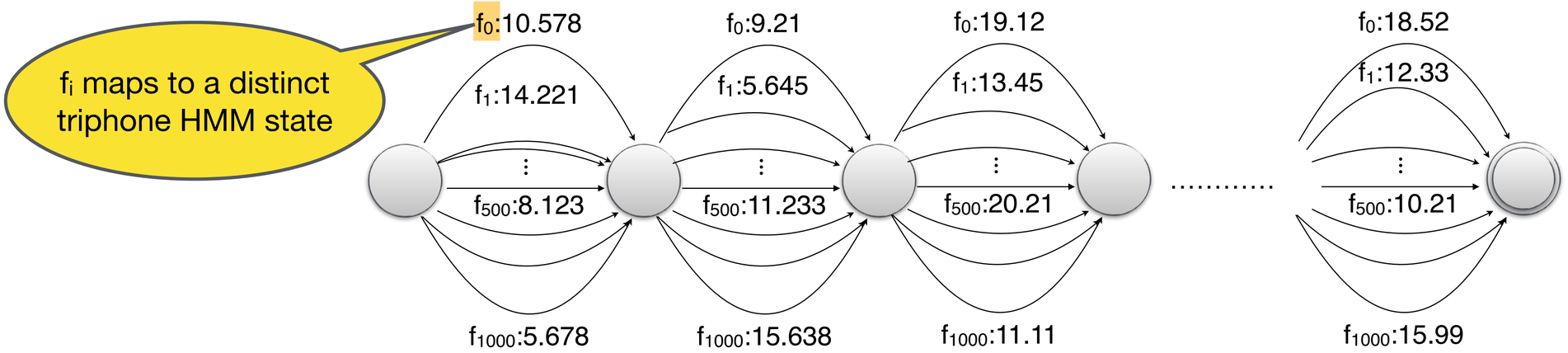


Carefully construct a decoding graph D using optimization algorithms:

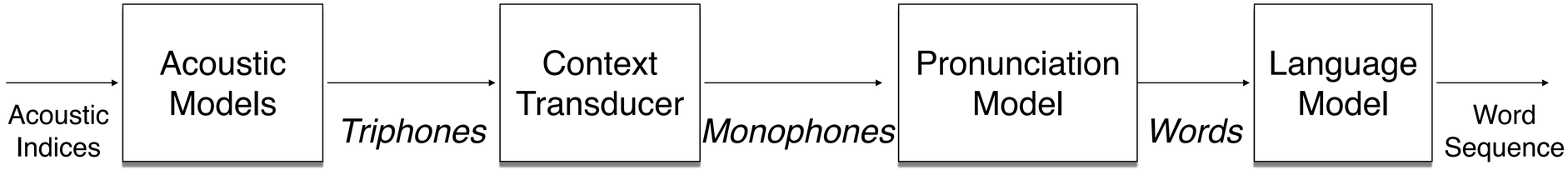
$$D = \min(\det(H \circ \det(C \circ \det(L \circ G))))$$

Given a test utterance *O, how do I decode it?*

Assuming ample compute, first construct the following machine X from O.



Decoding

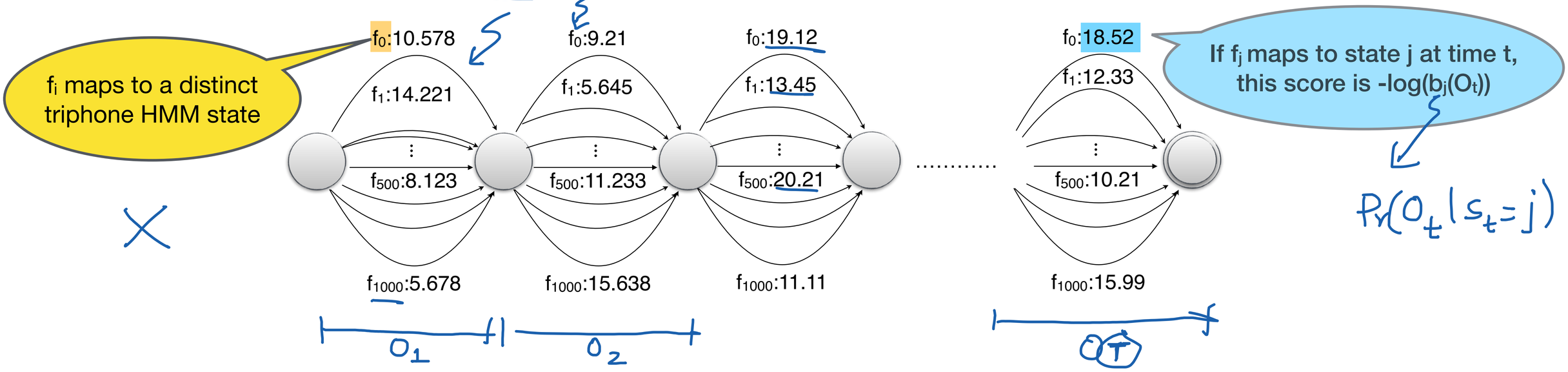


Carefully construct a decoding graph D using optimization algorithms:

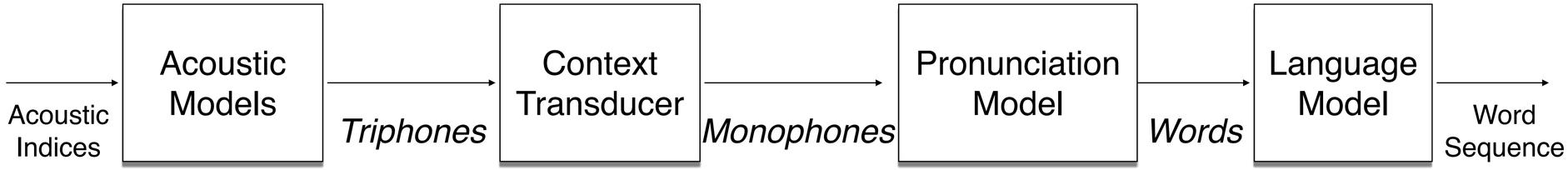
$$D = \min(\det(H \circ \det(C \circ \det(L \circ G))))$$

Given a test utterance *O, how do I decode it?*

Assuming ample compute, first construct the following machine X from O.



Decoding



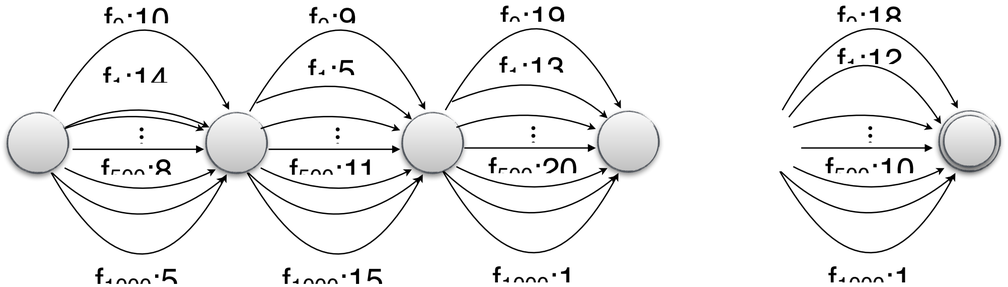
H C L G

Carefully construct a decoding graph D using optimization algorithms:

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

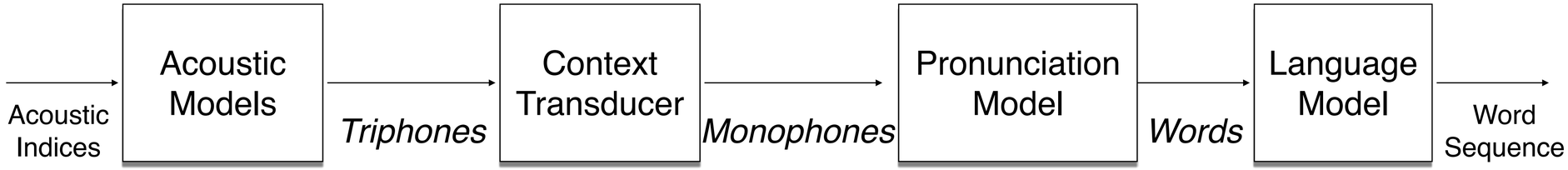
Given a test utterance *O, how do I decode it?*

Assuming ample compute, first construct the following machine X from O.



X

Decoding



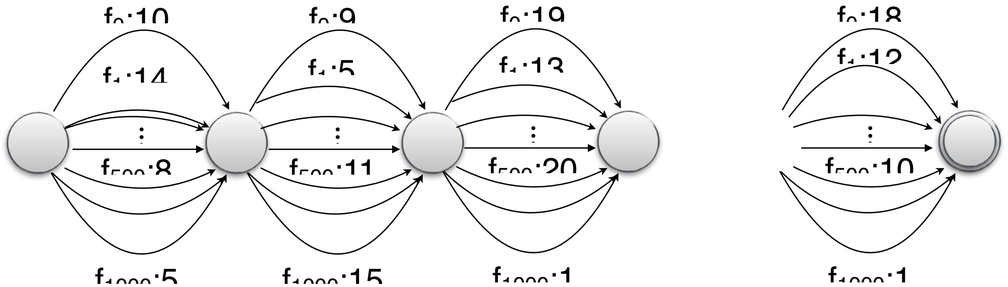
H C L G

Carefully construct a decoding graph D using optimization algorithms:

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

Given a test utterance *O, how do I decode it?*

Assuming ample compute, first construct the following machine X from O.

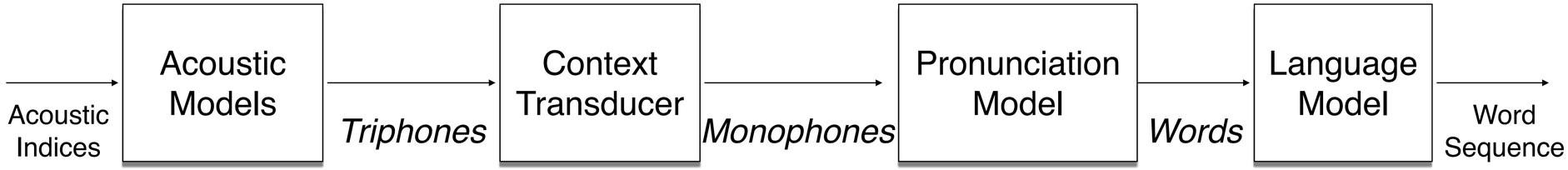


X

$$W^* = \underset{W=out[\pi]}{\arg\min} X \circ D$$

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

Decoding



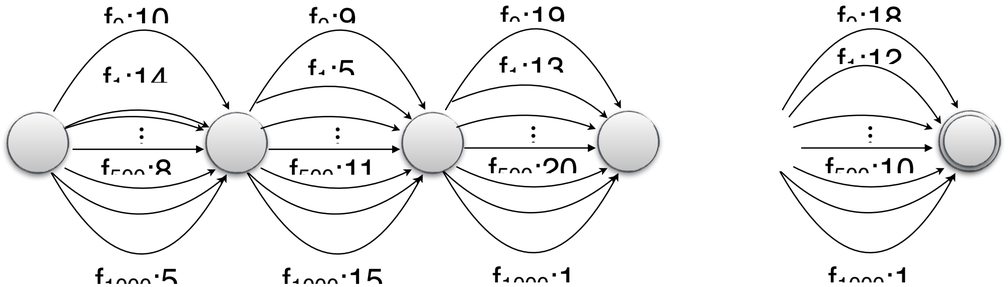
H C L G

Carefully construct a decoding graph D using optimization algorithms:

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

Given a test utterance *O, how do I decode it?*

Assuming ample compute, first construct the following machine X from O.



X

$$W^* = \underset{W=out[\pi]}{\arg\min} \; X \circ D$$

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

X is never typically constructed;
D is traversed dynamically using approximate search algorithms
(discussed later in the semester)

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

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

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