Speech Analysis and Linguistics
Applications: Automatic Speech Recognition & Mispronunciation Detection and Diagnosis (MDD)

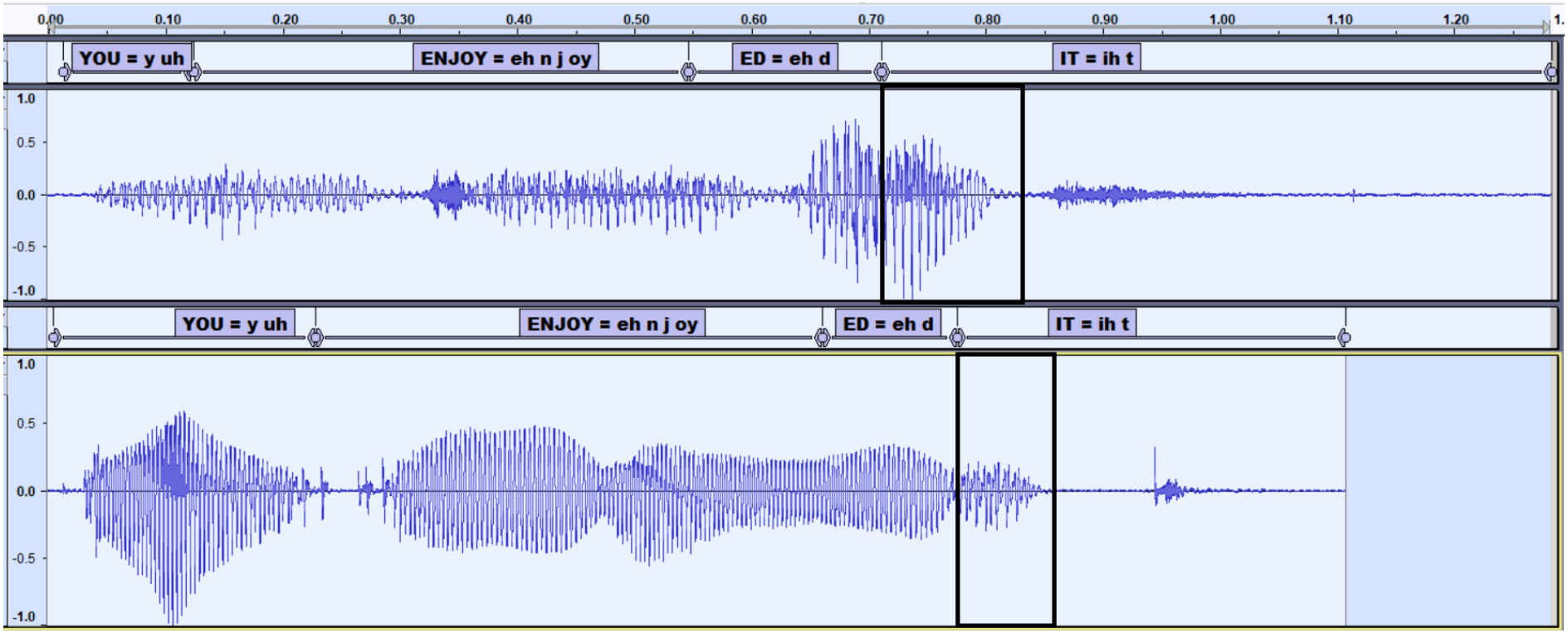
Instructor: Chiranjeevi Yarra

Speech lab, LTRC

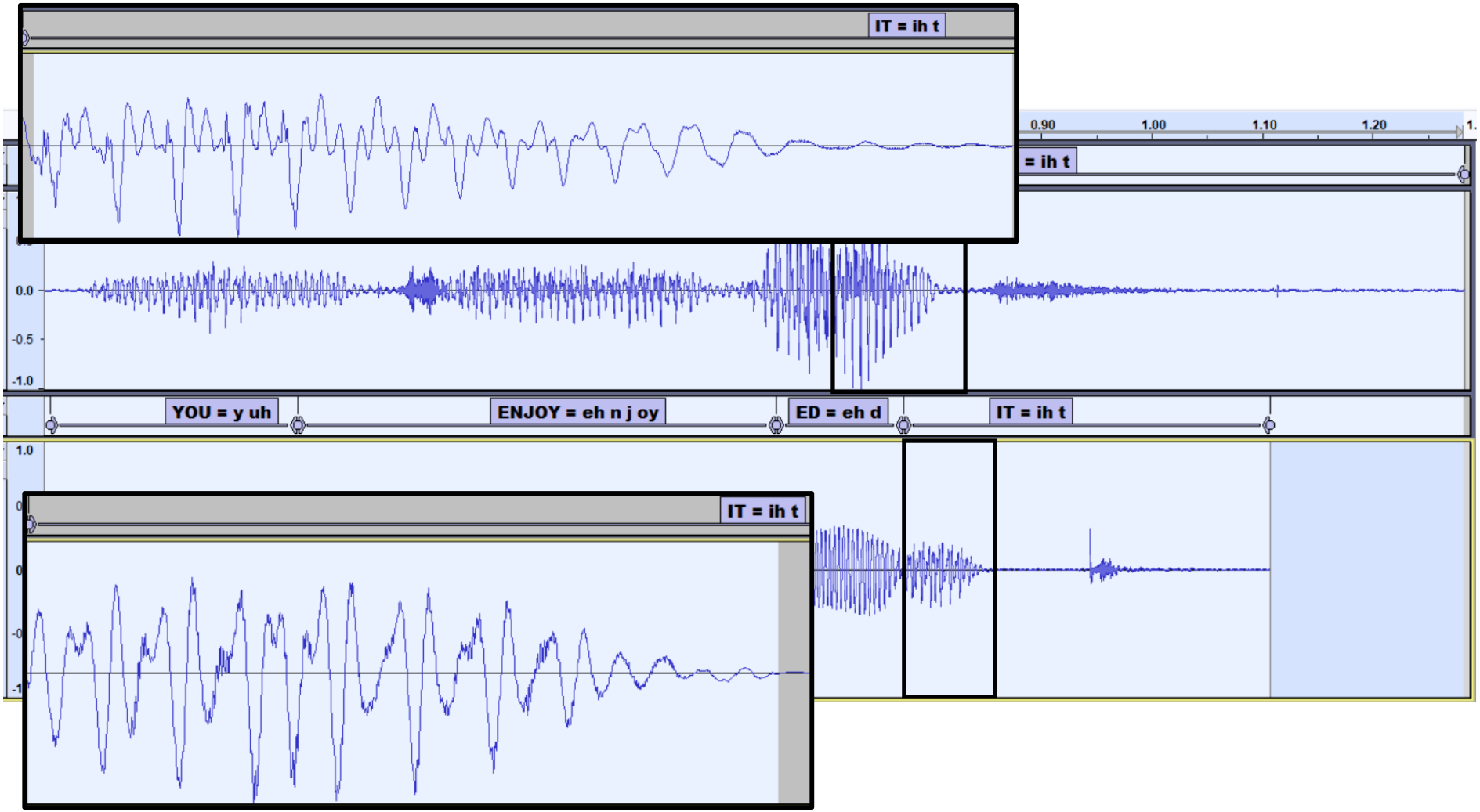
Aug 11, 2022

Outline

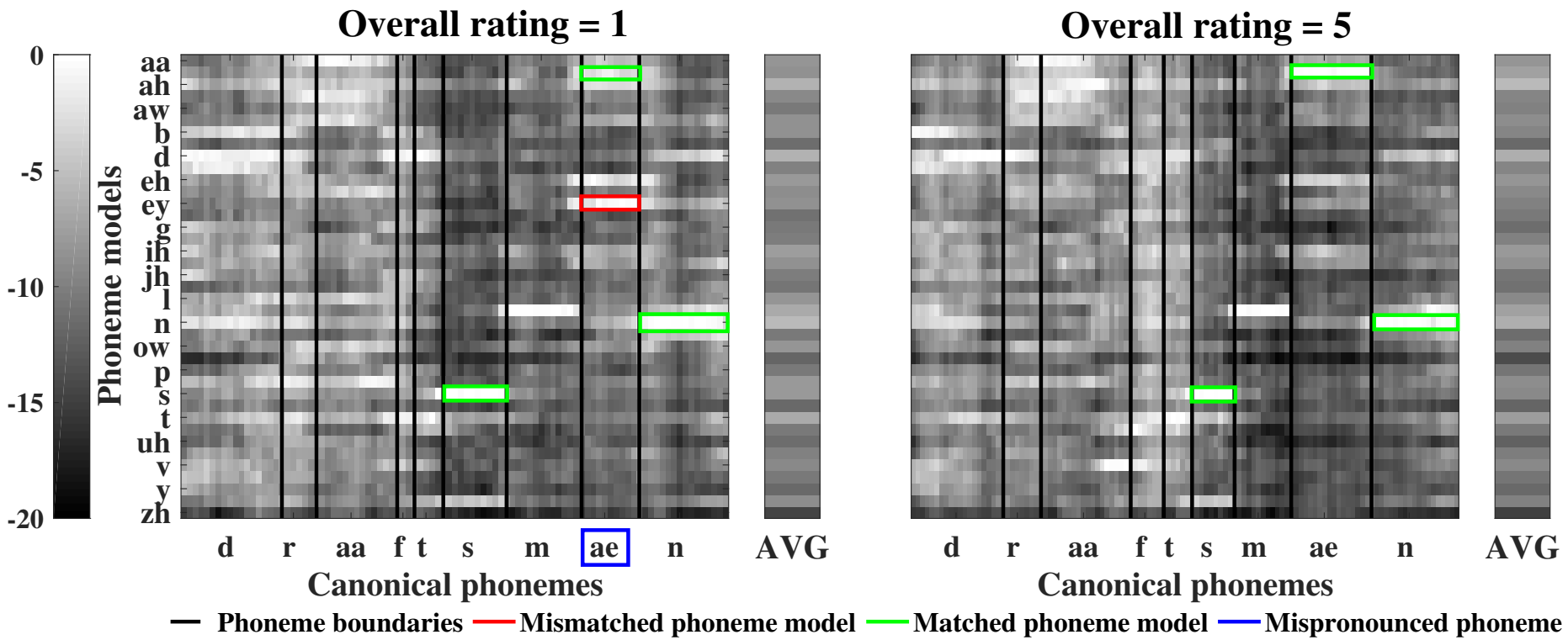
What is to be model in ASR?



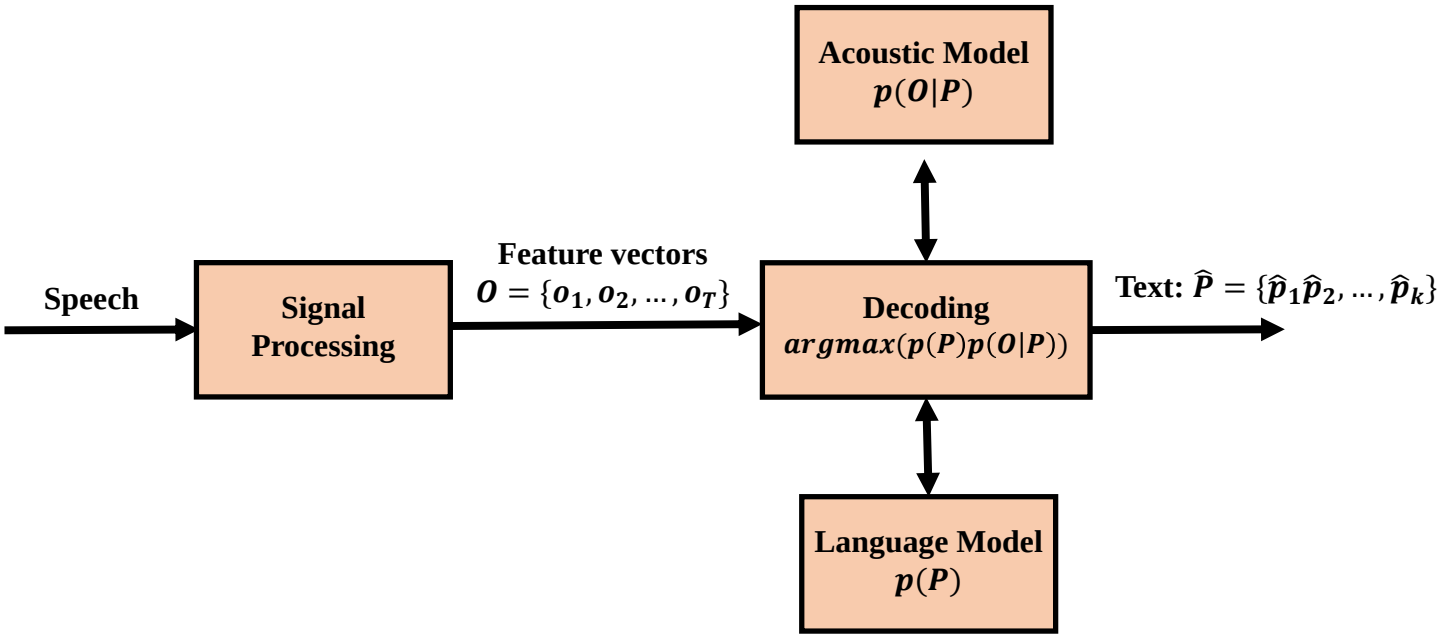
What is to be model in ASR?



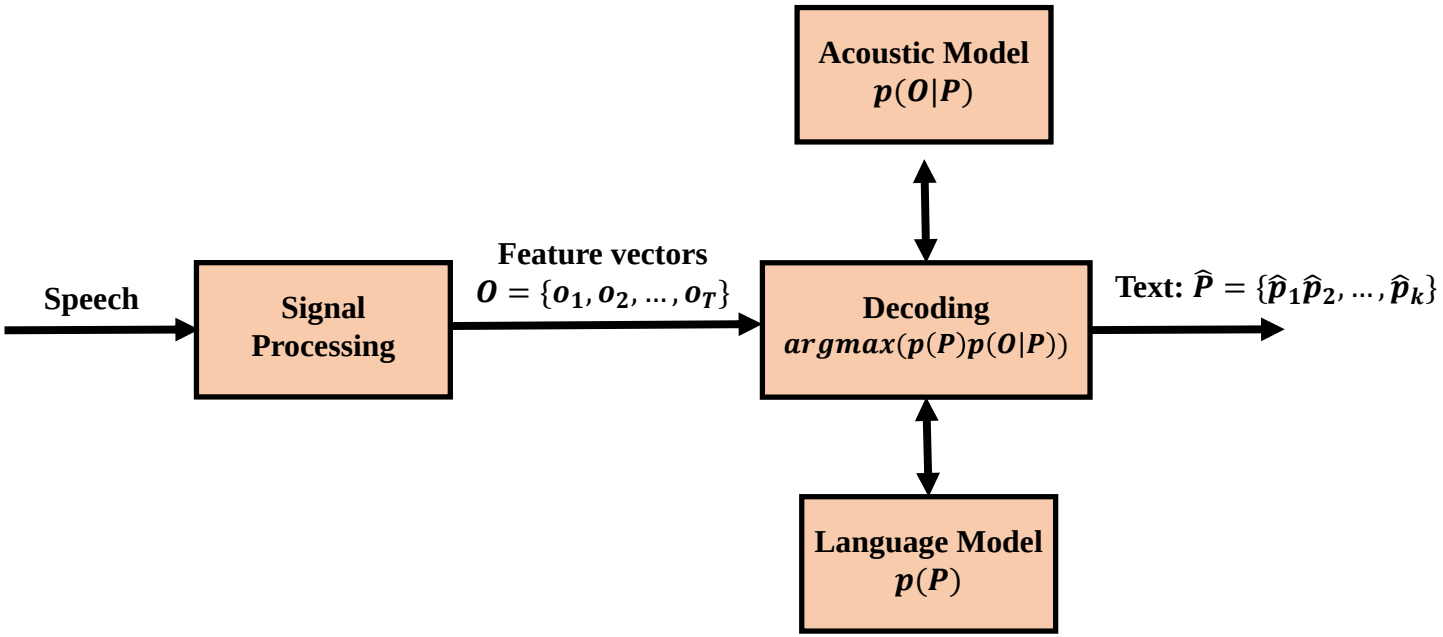
Errors in the estimation



Building blocks

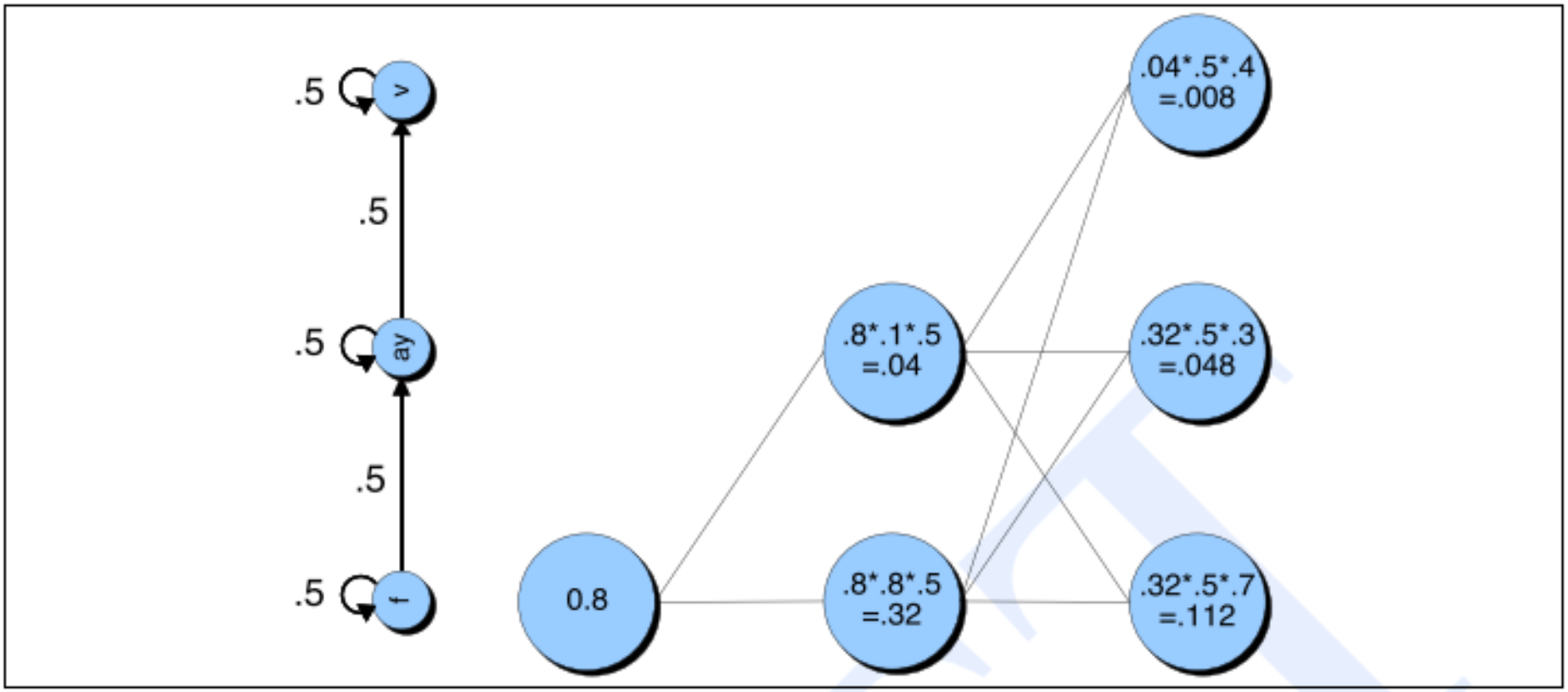


Building blocks



1. How to model $p(P)$ and $p(O|P)$?

Decoding



Decoding illustration

V	0	0	0.008	0.0072	0.00672	0.00403	0.00188	0.00161	0.000667	0.000493
AY	0	0.04	0.048	0.0448	0.0269	0.0125	0.00538	0.00167	0.000428	8.78e-05
F	0.8	0.32	0.112	0.0224	0.00448	0.000896	0.000179	4.48e-05	1.12e-05	2.8e-06
Time	1	2	3	4	5	6	7	8	9	10
	f 0.8	*f* 0.8	*f* 0.7	*f* 0.4	*f* 0.4	*f* 0.4	*f* 0.4	*f* 0.5	*f* 0.5	*f* 0.5
	ay 0.1	*ay* 0.1	*ay* 0.3	*ay* 0.8	*ay* 0.8	*ay* 0.8	*ay* 0.8	*ay* 0.6	*ay* 0.5	*ay* 0.4
B	*v* 0.6	*v* 0.6	*v* 0.4	*v* 0.3	*v* 0.3	*v* 0.3	*v* 0.3	*v* 0.6	*v* 0.8	*v* 0.9
	p 0.4	*p* 0.4	*p* 0.2	*p* 0.1	*p* 0.1	*p* 0.1	*p* 0.1	*p* 0.1	*p* 0.3	*p* 0.3
	iy 0.1	*iy* 0.1	*iy* 0.3	*iy* 0.6	*iy* 0.6	*iy* 0.6	*iy* 0.6	*iy* 0.5	*iy* 0.5	*iy* 0.4

ASR equation

$$\hat{W} = \text{argmax}\{p\,(W|O)\}$$

ASR equation

$$\hat{W} = \text{argmax}\{p\,(W|O)\}$$

$$= \text{argmax}\{p(W)p\,(O|W)\}$$

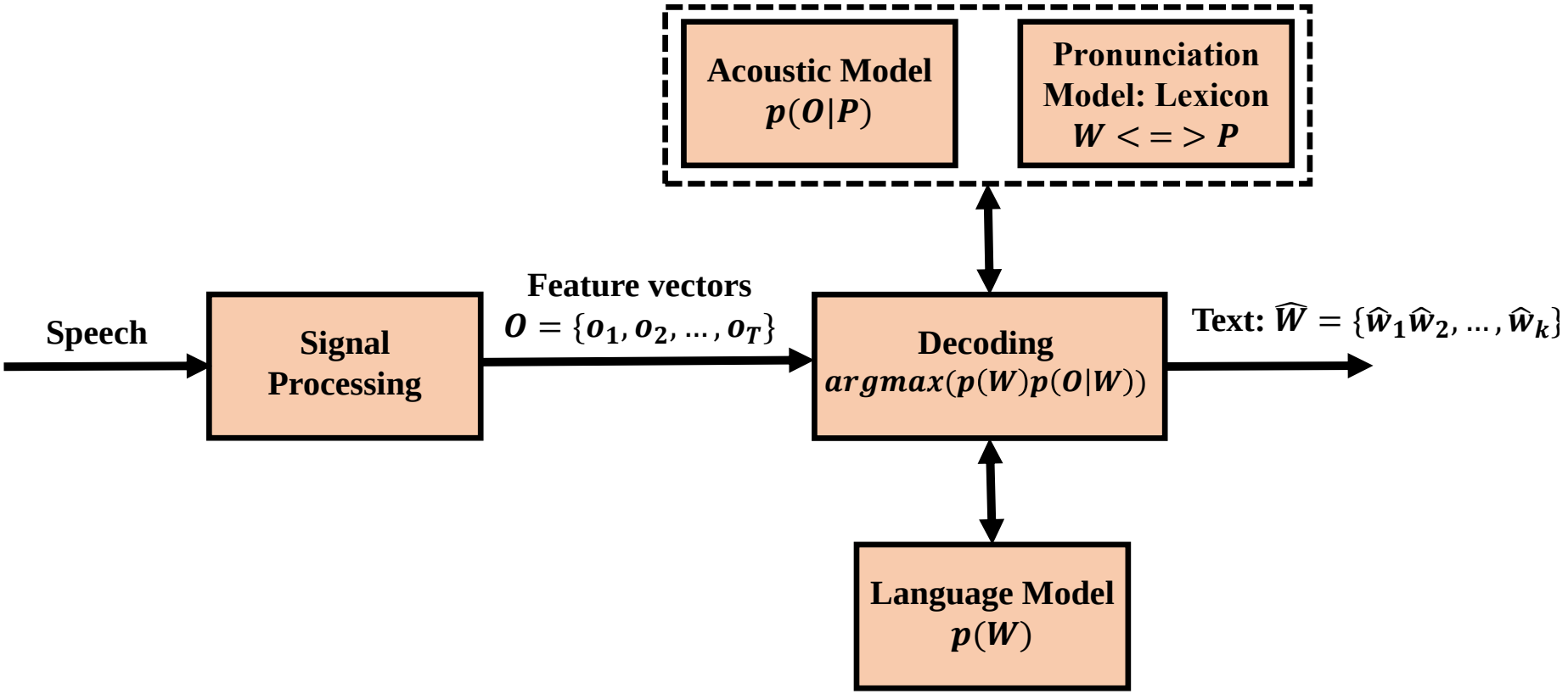
ASR equation

$$\hat{W} = \text{argmax}\{p\,(W|O)\}$$

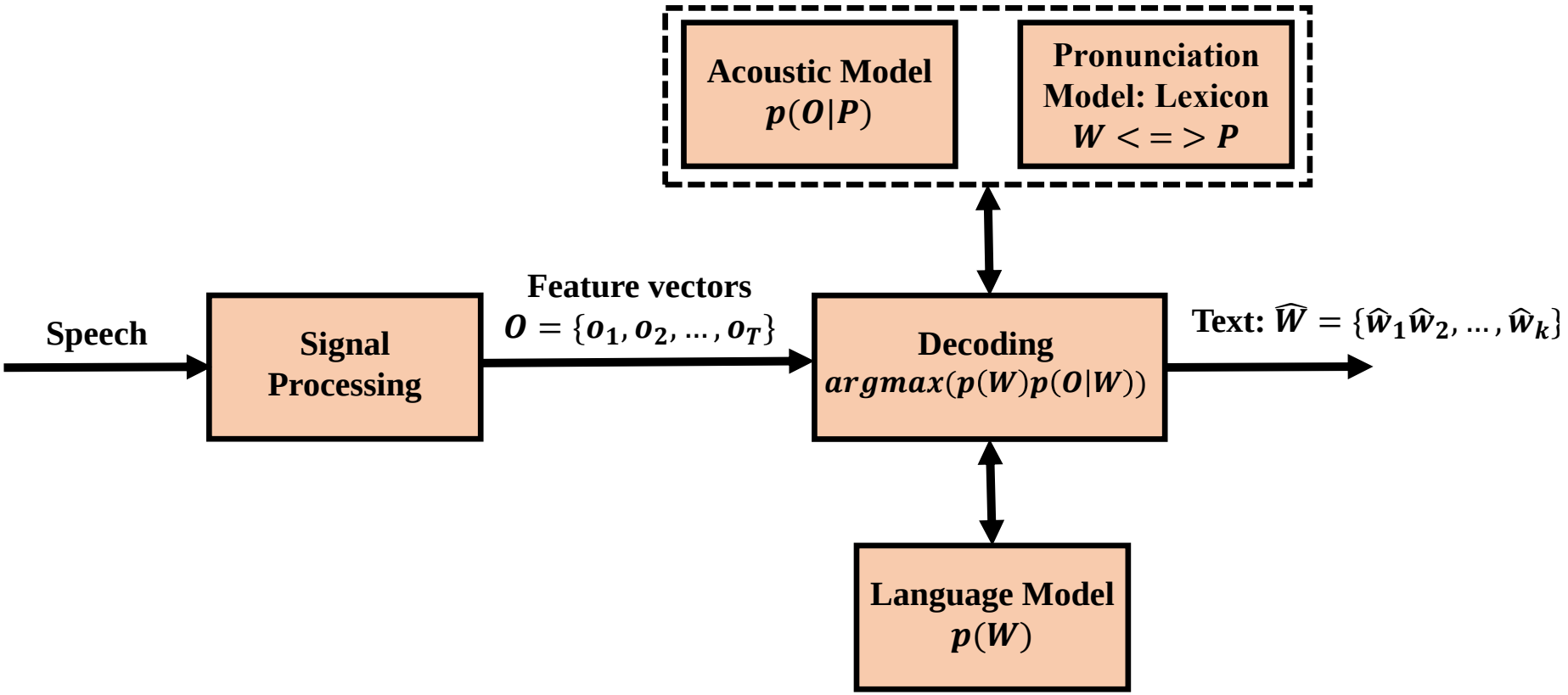
$$= \text{argmax}\{p(W)p\,(O|W)\}$$

1. What data is needed for modelling $p(W)$ and $p(O|W)$?

ASR building blocks



ASR building blocks



1. Why modelling $p(O|P)$ is efficient?

Word and phoneme sequence mapping: Lexicon?

An exemplary Lexicon

```
BRANER  B R EY1 N ER0
BRANFORD  B R AE1 N F ER0 D
BRANHAM  B R AE1 N HH AH0 M
BRANI  B R AE1 N IY0
BRANIFF  B R AE1 N IH0 F
BRANIFF'S  B R AE1 N IH0 F S
BRANIGAN  B R AE1 N IH0 G AH0 N
BRANILLO  B R AH0 N IH1 L OW0
BRANIN  B R AE1 N IH0 N
BRANISLOV  B R AE1 N IH0 S L AA2 V
BRANITZKY  B R AH0 N IH1 T S K IY1
BRANK  B R AE1 NG K
BRANK'S  B R AE1 NG K S
BRANKI  B R AE1 NG K IY0
BRANKO  B R AE1 NG K OW0
BRANKS  B R AE1 NG K S
BRANN  B R AE1 N
BRANNA  B R AE1 N AH0
BRANNAM  B R AE1 N AH0 M
```

Word and phoneme sequence mapping: Lexicon?

An exemplary Lexicon

```
BRANER  B R EY1 N ER0
BRANFORD  B R AE1 N F ER0 D
BRANHAM  B R AE1 N HH AH0 M
BRANI  B R AE1 N IY0
BRANIFF  B R AE1 N IH0 F
BRANIFF'S  B R AE1 N IH0 F S
BRANIGAN  B R AE1 N IH0 G AH0 N
BRANILLO  B R AH0 N IH1 L OW0
BRANIN  B R AE1 N IH0 N
BRANISLOV  B R AE1 N IH0 S L AA2 V
BRANITZKY  B R AH0 N IH1 T S K IY1
BRANK  B R AE1 NG K
BRANK'S  B R AE1 NG K S
BRANKI  B R AE1 NG K IY0
BRANKO  B R AE1 NG K OW0
BRANKS  B R AE1 NG K S
BRANN  B R AE1 N
BRANNA  B R AE1 N AH0
BRANNAM  B R AE1 N AH0 M
```

1. Knowing $p(O|P)$ sufficient for computing $p(O|W)$?

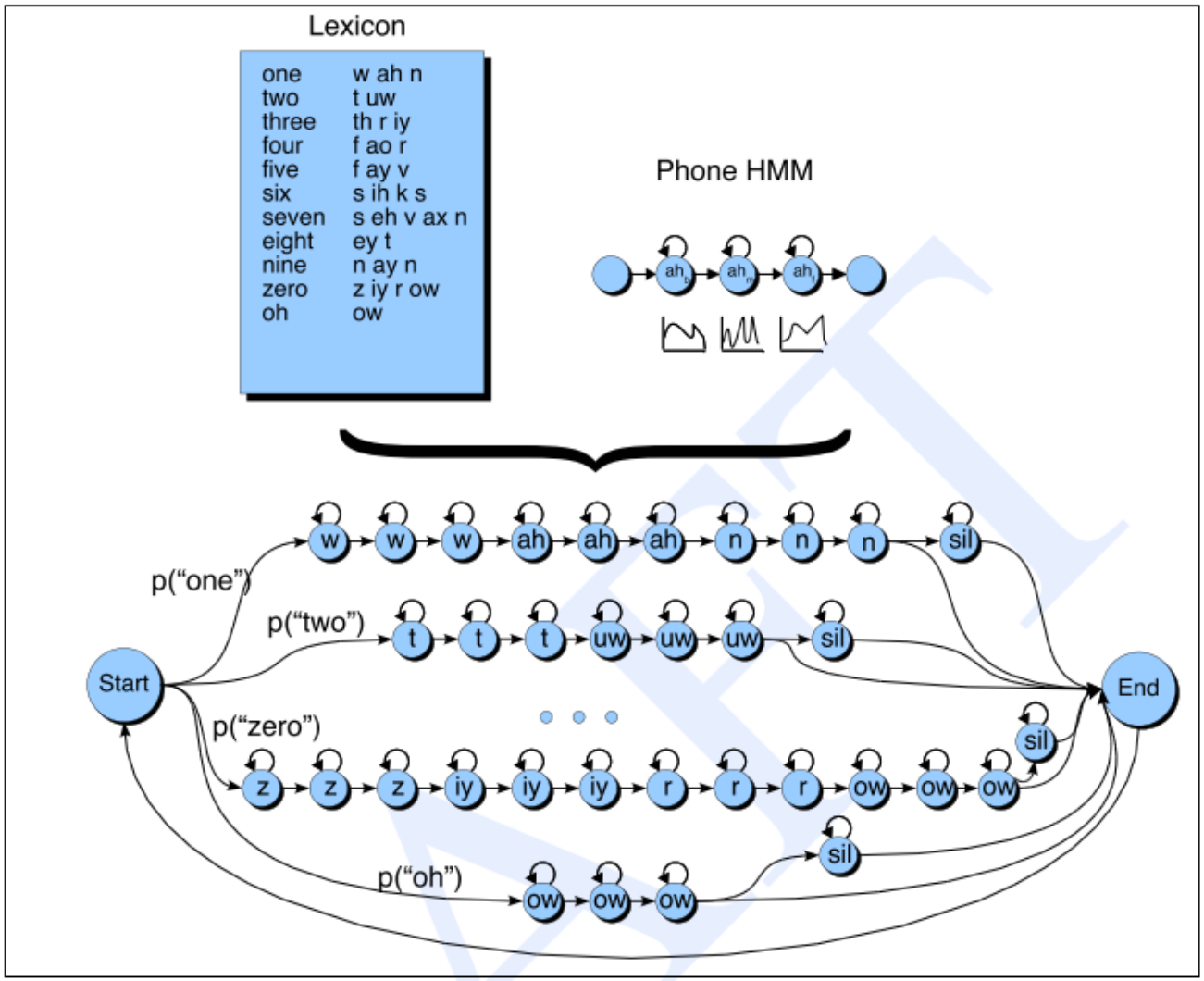
Word and phoneme sequence mapping: Lexicon?

An exemplary Lexicon

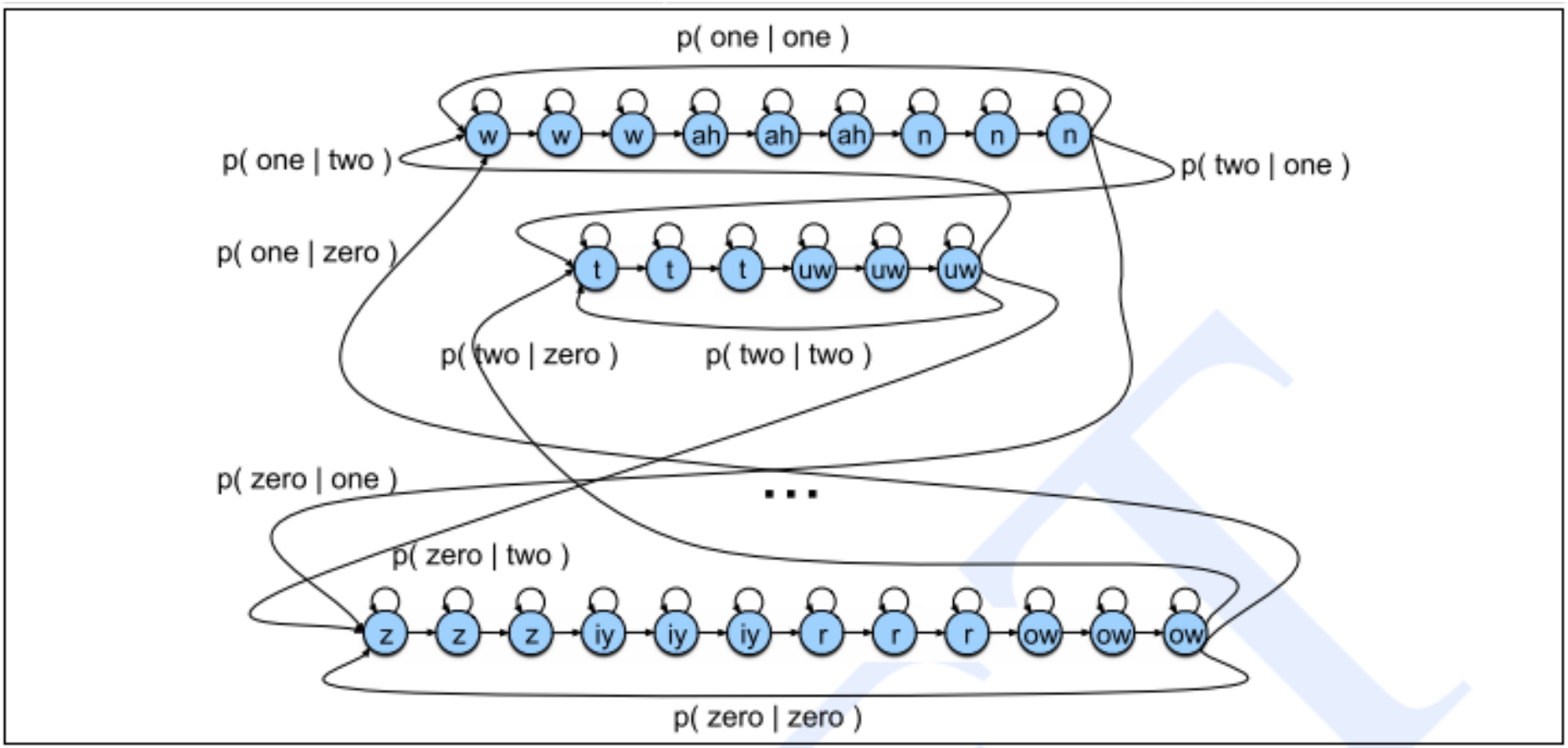
```
BRANER  B R EY1 N ER0
BRANFORD  B R AE1 N F ER0 D
BRANHAM  B R AE1 N HH AH0 M
BRANI  B R AE1 N IY0
BRANIFF  B R AE1 N IH0 F
BRANIFF'S  B R AE1 N IH0 F S
BRANIGAN  B R AE1 N IH0 G AH0 N
BRANILLO  B R AH0 N IH1 L OW0
BRANIN  B R AE1 N IH0 N
BRANISLOV  B R AE1 N IH0 S L AA2 V
BRANITZKY  B R AH0 N IH1 T S K IY1
BRANK  B R AE1 NG K
BRANK'S  B R AE1 NG K S
BRANKI  B R AE1 NG K IY0
BRANKO  B R AE1 NG K OW0
BRANKS  B R AE1 NG K S
BRANN  B R AE1 N
BRANNA  B R AE1 N AH0
BRANNAM  B R AE1 N AH0 M
```

1. Knowing $p(O|P)$ sufficient for computing $p(O|W)$?
2. Knowing W sufficient for obtaining P?

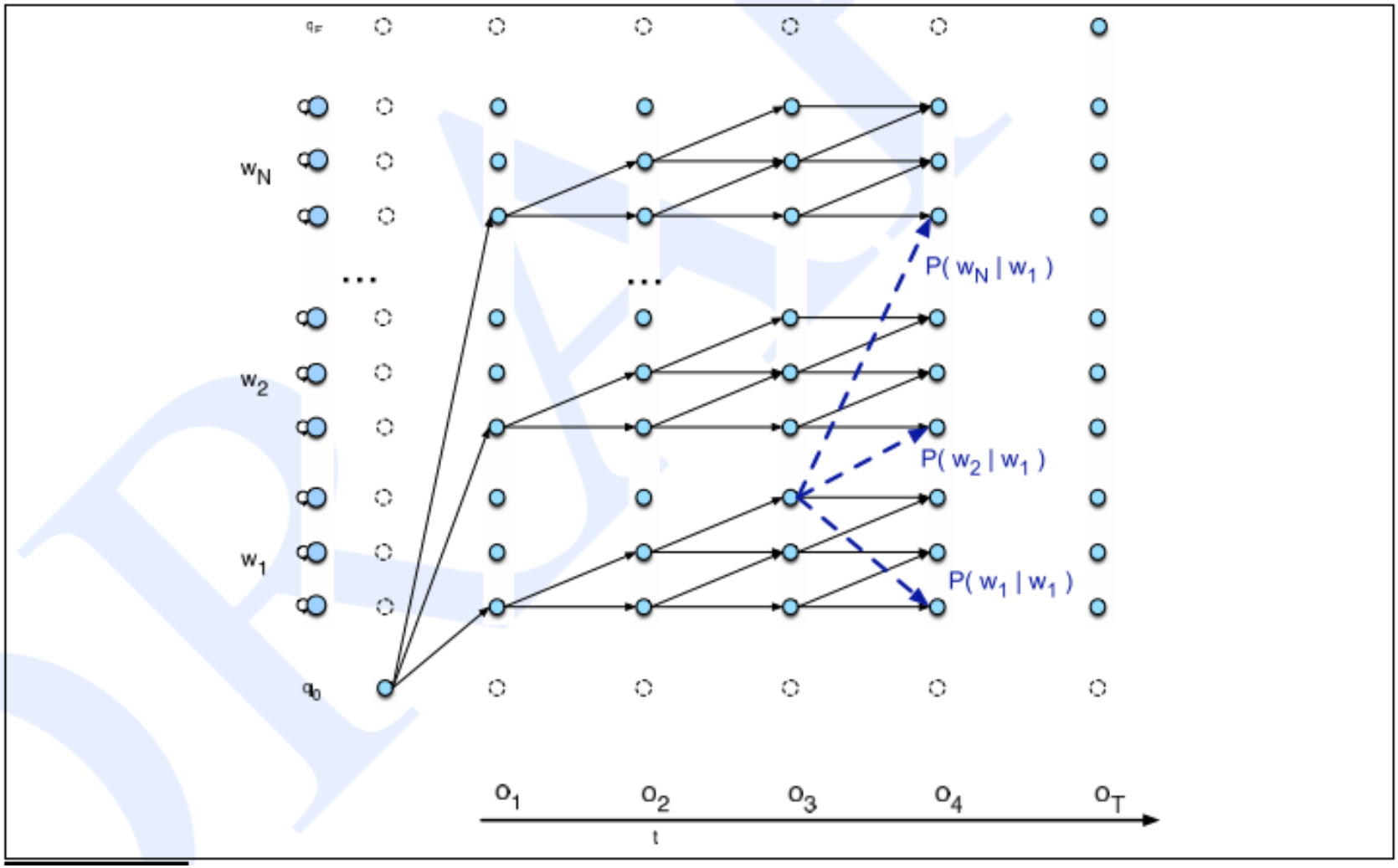
An example


Lexicon
one w ah n
two t uw
three th r iy
four f ao r
five f ay v
six s ih k s
seven s eh v ax n
eight ey t
nine n ay n
zero z iy r ow
oh ow
Phone HMM
Start
End
p("one")
p("two")
p("zero")
p("oh")
w
ah
n
t
uw
z
iy
r
ow
sil

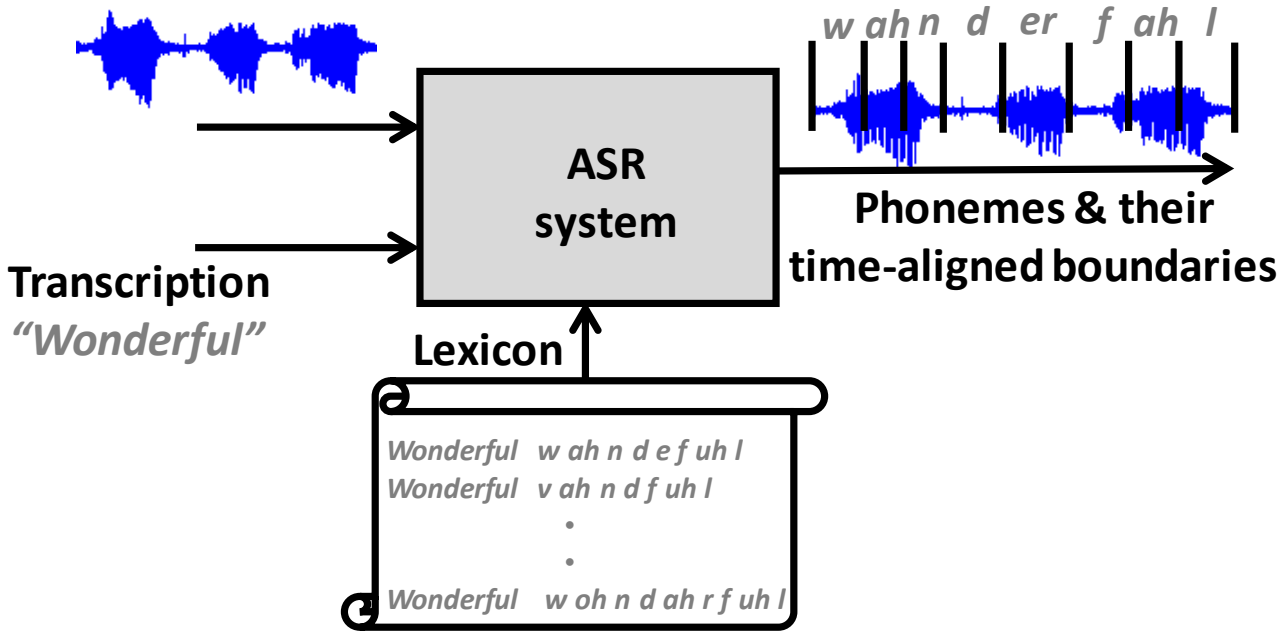
Language model for Digit Recognition



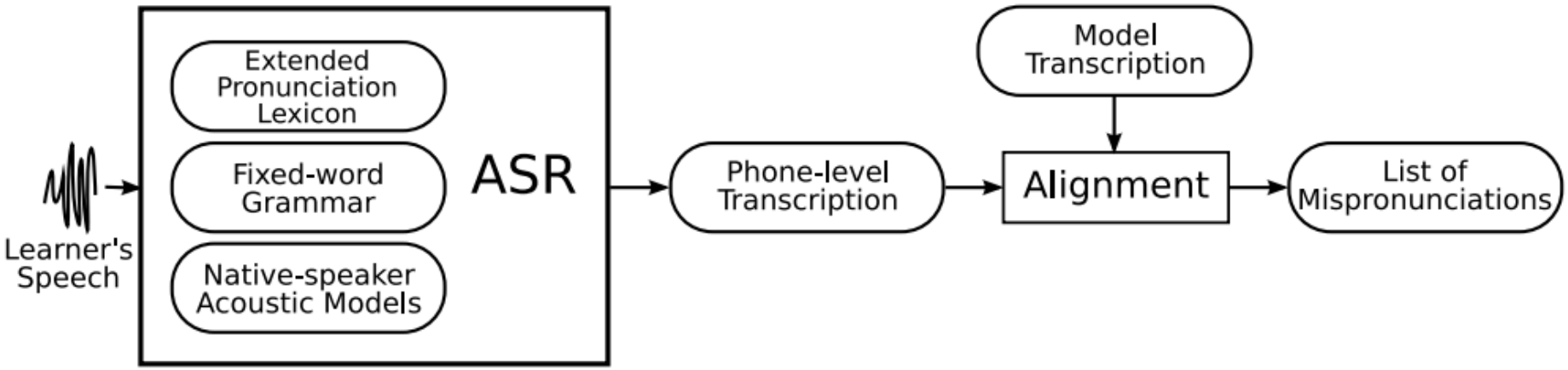
Viterbi decoding



Forced-alignment process



Mispronunciation Detection and Diagnosis (MDD)



Thank you