# Phonetic Forced Alignment with the Montreal Forced Aligner

Tutorial: eleanorchodroff.com/mfa_tutorial.html
Please download the materials (zip file) in the first link there!

Eleanor Chodroff
University of York

# Introductions

How many of you have used a force aligner before this tutorial?

How many of you have used any version of the Montreal Forced Aligner?

How many of you have used the most recent release of the Montreal Forced Aligner (v2)?

How many of you have (seemingly) successfully installed the Montreal Forced Aligner v2?

# Introductions

About me:

Cognitive scientist
Linguist
Phonetician

Fortunate to work with many of the great minds behind the technology

—I’m also mostly a consumer of the technology

# Acknowledgements

Huge thanks to Michael McAuliffe

- Developing this aligner into something incredibly powerful
- Maintaining the aligner for the community and actively engaging with the community
- Personally answering the questions I've had in such a timely and informative manner
    - He pushed a new release for us just last night!!

# Acknowledgements

Thanks also to:

- the whole Montreal Forced Aligner team
- the JHU Kaldi team for their help back in ~2015
  - Sanjeev Khudanpur
  - Yenda Trmal
- Emily Ahn (University of Washington)
- Audiences of previous tutorials of mine

# The “Plan”

Overview to forced alignment

Overview to forced alignment with MFA

Installation + checks

Running the aligner: basic (Example 1)

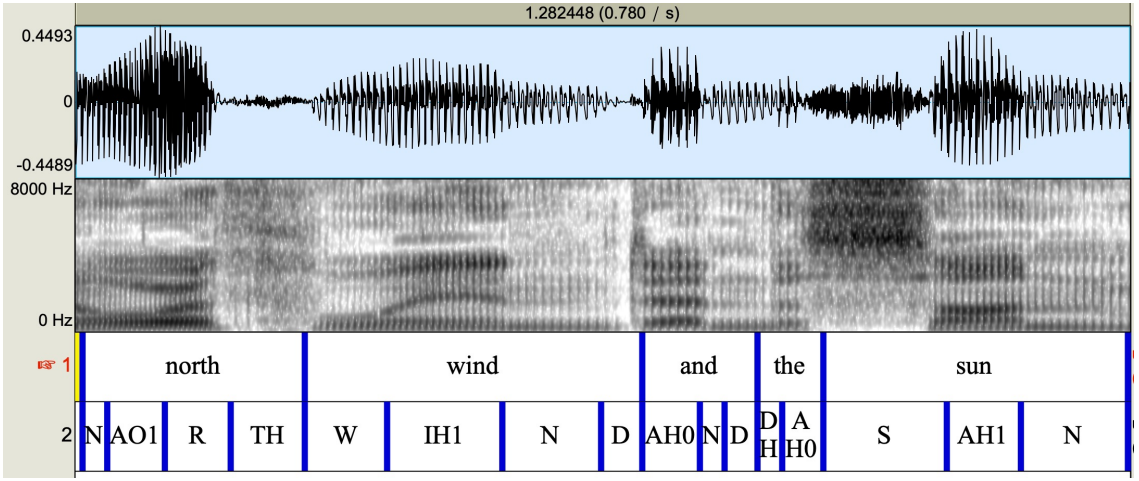
Break

Case studies (Example 2+)

# What is forced alignment?

Forced alignment: “a technique to take an orthographic transcription of an audio file and generate a time-aligned version using a pronunciation dictionary to look up phones for words”

--Montreal Forced Aligner website




aka automatic segmentation

# Benefits of automation

- Save time in the long run
- Consistency: minimize human error
- Replicability: allow others to repeat the process *identically*
- Easily correct mistakes
- Easily process large amounts of data

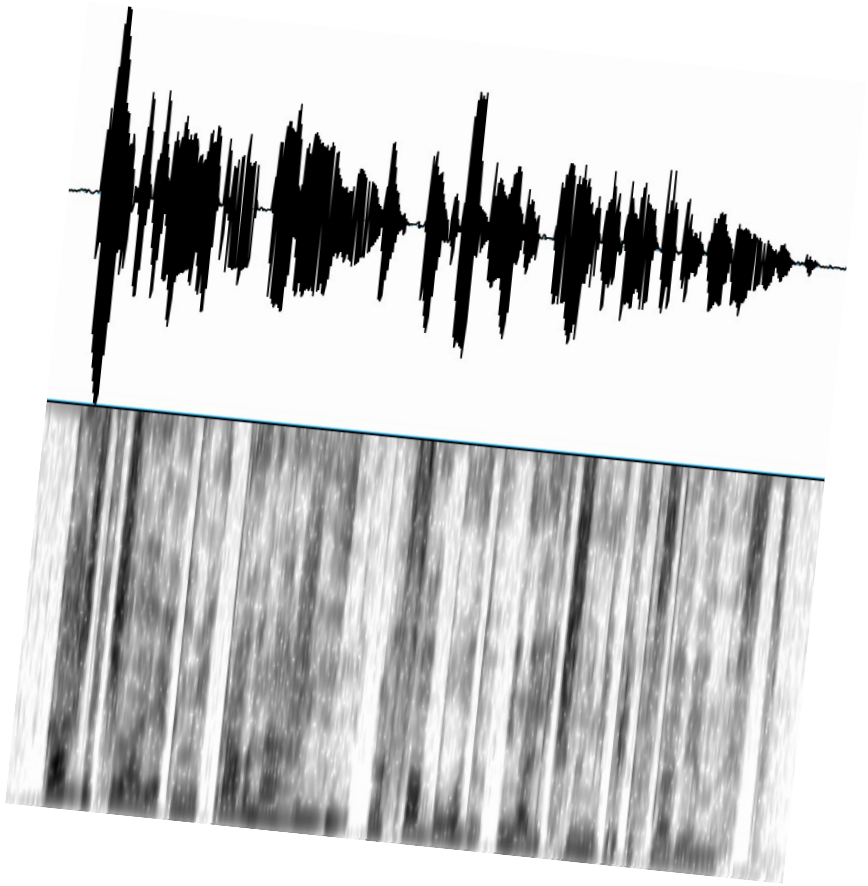
**Did you know?**

By one estimate, using phonetic forced alignment reduces time spent on segmentation by 75% (Young, 2017)

# Montreal Forced Aligner Background

- Developed by Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger at McGill University
- Currently and actively maintained by Michael McAuliffe (Amazon)
- Uses the Kaldi ASR toolkit
- Triphone GMM-HMM system with speaker adaptation
  - Minimum 30 ms phone
- English acoustic models trained on LibriSpeech
- Non-English acoustic models (except some French and German ones) are largely trained using the GlobalPhone corpus – this will be updated soon!

# Input



The North Wind and the Sun were disputing which was the stronger when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as could, but the more he blew, the more closely did the traveler fold his cloak around him, and at last the North Wind gave up the attempt. Then the Sun shined out warmly, and immediately the traveler took off his coat, and so the North Wind was obliged to confess that the Sun was the stronger of the two.

# From the text to the speech: Obtain a sequence of “phones”

```
The North Wind and the Sun were disputing which was the
stronger when a traveler came along wrapped in a warm
cloak. They agreed that the one who first succeeded in
making the traveler take his cloak off should be
considered stronger than the other. Then the North Wind
blew as hard as could, but the more he blew, the more
closely did the traveler fold his cloak around him, and
at last the North Wind gave up the attempt. Then the
Sun shined out warmly, and immediately the traveler
took off his coat, and so the North Wind was obliged to
confess that the Sun was the stronger of the two.
```

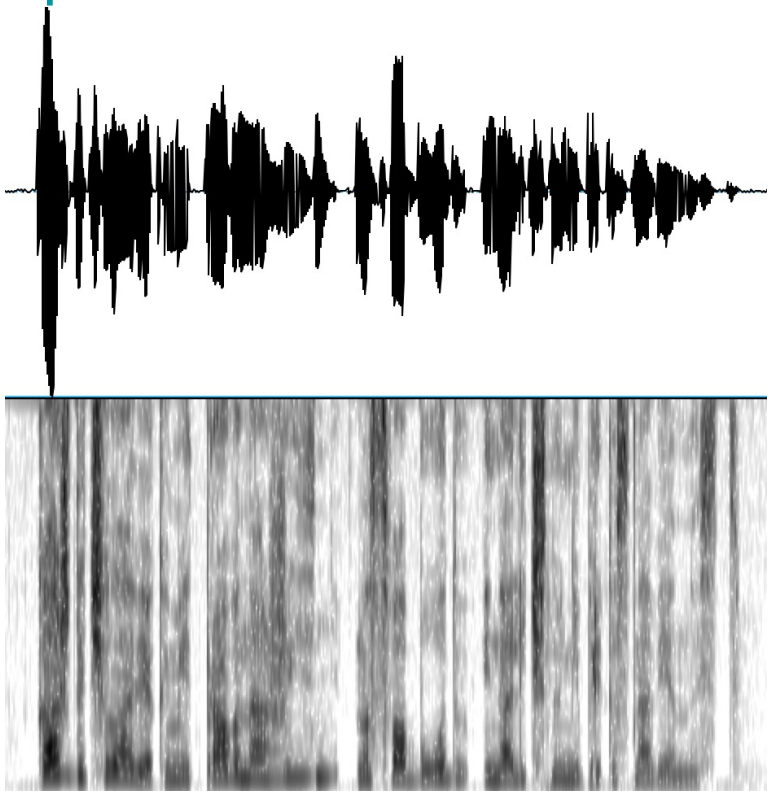
```
HALLOO'D        HH AE1 L UW0 D
HALLOOED        HH AH0 L UW1 D
HALLOOING       HH AH0 L UW1 IH0 NG
HALLOOS HH AH0 L UW1 Z
HALLORAN  HH AE1 L ER0 AH0 N
HALLORAN'S      HH AE1 L ER0 AH0 N Z
HALLOW  HH AE1 L OW0
HALLOW'D        HH AE1 L OW0 D
HALLOWAY  HH AE1 L OW0 W EY2
HALLOWAY'S      HH AE1 L OW0 W EY2 Z
HALLOWE'EN  HH AE2 L AH0 W IY1 N
HALLOWED  HH AE1 L OW0 D
HALLOWEEN  HH AE2 L AH0 W IY1 N
HALLOWELL  HH AE1 L AH0 W EH0 L
HALLOWELL'S     HH AE1 L OW0 IH0 L Z
HALLOWELLS      HH AE2 L AH0 W EH0 L Z
HALLOWING       HH AE1 L OW0 IH0 NG
HALLOWS  HH AE1 L OW0 Z
HALLS  HH AO1 L Z
HALLSTOCK       HH AE1 L S T AA2 K
HALLTON HH AO1 L T AH0 N
HALLTOWN        HH AO1 L T AW2 N
HALLUCINATED  HH AH0 L UW1 S AH0 N EY0 T IH0 D
HALLUCINATING  HH AH0 L UW1 S AH0 N EY0 T IH0 NG
HALLUCINATING  HH AH0 L UW1 S IH0 N EY0 T IH0 NG
HALLUCINATION  HH AH0 L UW2 S AH0 N EY1 SH AH0 N
HALLUCINATIONS  HH AH0 L UW2 S AH0 N EY1 SH AH0 N Z
HALLUCINATORY  HH AH0 L UW1 S AH0 N AH0 T AO2 R IY0
```

```
DH AH N AO R TH W AY N D AH N D DH AH S AH N W ER D IH S P Y UW
T IH NG W IH CH W AA Z DH AH S T R AO NG G ER W EH N EY T R AE V
AH L ER K EY M AH L AO NG R AE P T IH N EY W AO R M K L OW K DH
EY AH G R IY D DH AE T DH AH W AH N HH UW F ER S T S AH K S IY D
AH D IH N M EY K IH NG DH AH T R AE V AH L ER T EY K HH IH Z K L
OW K AO F SH UH D B IY K AH N S IH D ER D S T R AO NG G ER DH AE
N DH AH AH DH ER DH EH N DH AH N AO R TH W AY N D B L UW AE Z HH
AA R D AE Z K UH D B AH T DH AH M AO R HH IY B L UW DH AH M AO R
K L OW S L IY D IH D DH AH T R AE V AH L ER F OW L D HH IH Z K L
OW K ER AW N D HH IH M AH N D AE T L AE S T DH AH N AO R TH W AY
N D G EY V AH P DH AH AH T EH M P T DH EH N DH AH S AH N SH AY N
D AW T W AO R M L IY AH N D IH M IY D IY AH T L IY DH AH T R AE
V AH L ER T UH K AO F HH IH Z K OW T AH N D S OW DH AH N AO R TH
W AY N D W AA Z AH B L AY JH D T UW K AH N F EH S DH AE T DH AH
S AH N W AA Z DH AH S T R AO NG G ER AH V DH AH T UW
```

Pronunciation dictionary

G2P: grapheme-to-phoneme conversion

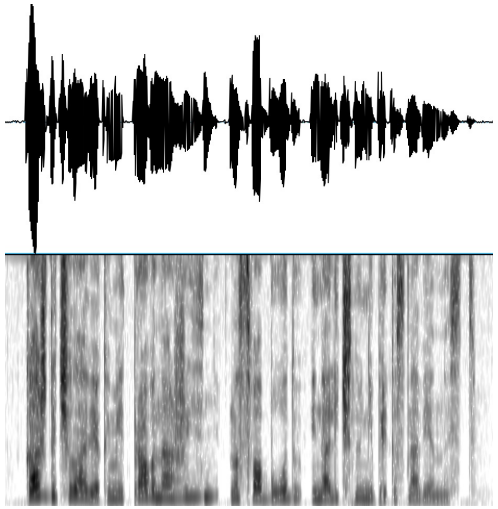
# From the speech to the text:
# Obtain a sequence of acoustic features

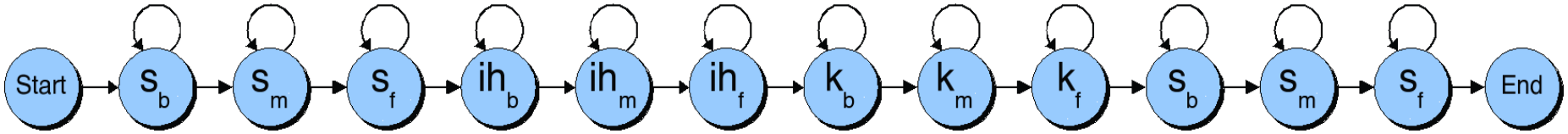


Extract MFCCs with 25 ms windows, every 10 ms

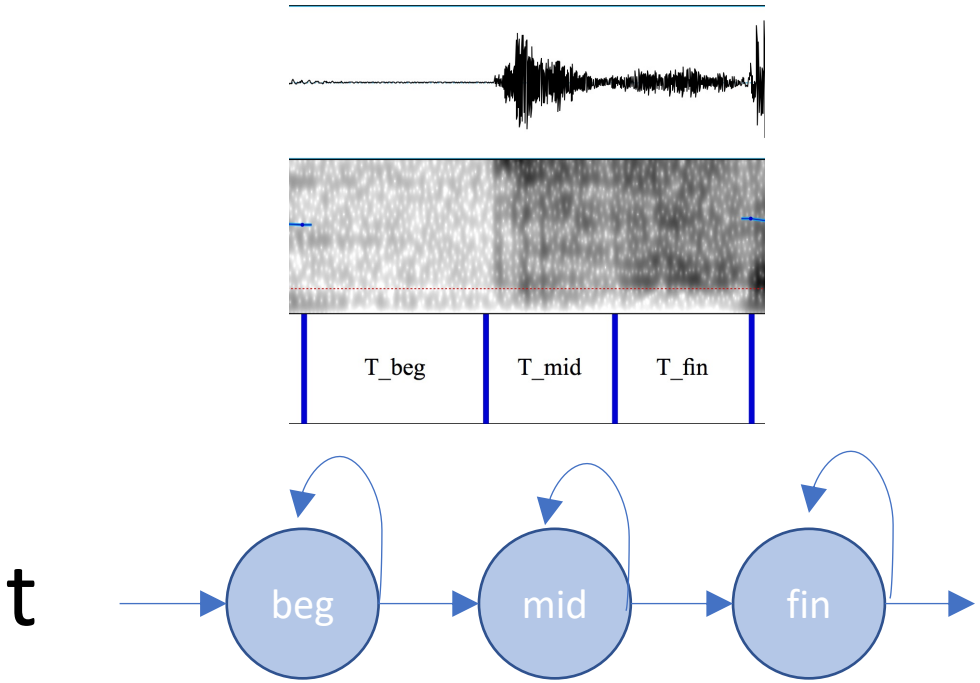
SEQUENCE OF MFCCS

# Given a sequence of phones and a sequence of features, find the best alignment…
# But how?





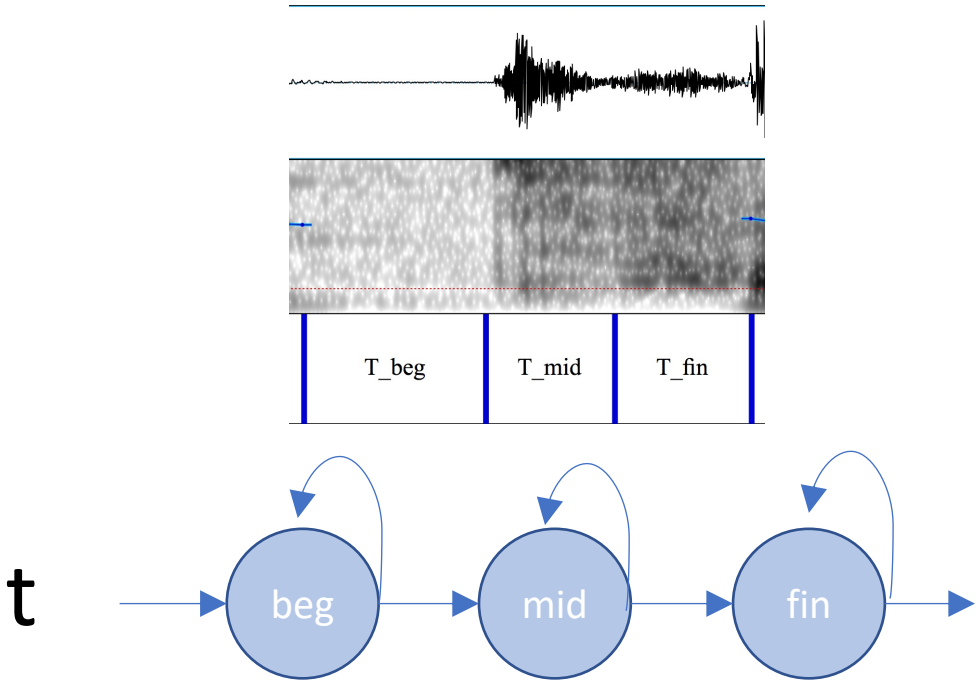
# GMM-HMM Framework+




Hidden Markov model: (loosely, and in this context,) a chain of states with transition probabilities

As we process a new frame of acoustic features, do we stay in the same state or move on to the next?

# GMM-HMM Framework+




Gaussian Mixture model: yes, that's right, Normal distributions of the acoustic features!

It basically knows the averages + SDs of each phone's acoustics
It uses this information to help decide whether to stay in the same state or move on to the next

# Training Acoustic Models

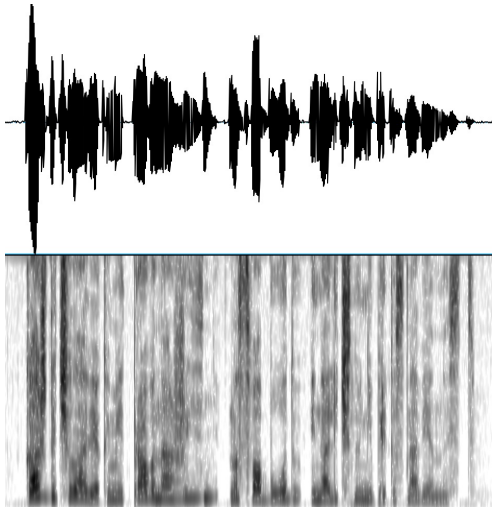
Believe it or not! Training a set of phone models ~ Aligning a sequence phones

Main difference:

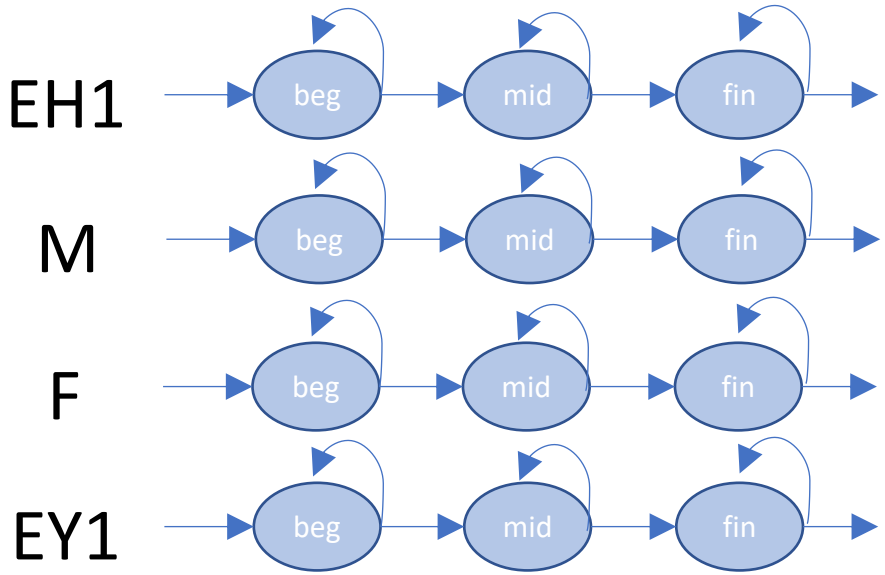
- Training: learning model parameters

- Alignment: applying model parameters

# GMM-HMM+ → Acoustic models



DH AH N AO R TH W AY N D AH N D DH AH S AH N W ER D IH S P Y UW
T IH NG W IH CH W AA Z DH AH S T R AO NG G ER W EH N EY T R AE V
AH L ER K EY M AH L AO NG R AE P T IH N EY W AO R M K L OW K DH
EY AH G R IY D DH AE T DH AH W AH N HH UW F ER S T S AH K S IY D
AH D IH N M EY K IH NG DH AH T R AE V AH L ER T EY K HH IH Z K L
OW K AO F SH UH D B IY K AH N S IH D ER D S T R AO NG G ER DH AE
N DH AH AH DH ER DH EH N DH AH N AO R TH W AY N D B L UW AE Z HH
AA R D AE Z K UH D B AH T DH AH M AO R HH IY B L UW DH AH M AO R
K L OW S L IY D IH D DH AH T R AE V AH L ER F OW L D HH IH Z K L
OW K ER AW N D HH IH M AH N D AE T L AE S T DH AH N AO R TH W AY
N D G EY V AH P DH AH AH T EH M P T DH EH N DH AH S AH N SH AY N
D AW T W AO R M L IY AH N D IH M IY D IY AH T L IY DH AH T R AE
V AH L ER T UH K AO F HH IH Z K OW T AH N D S OW DH AH N AO R TH
W AY N D W AA Z AH B L AY JH D T UW K AH N F EH S DH AE T DH AH
S AH N W AA Z DH AH S T R AO NG G ER AH V DH AH T UW

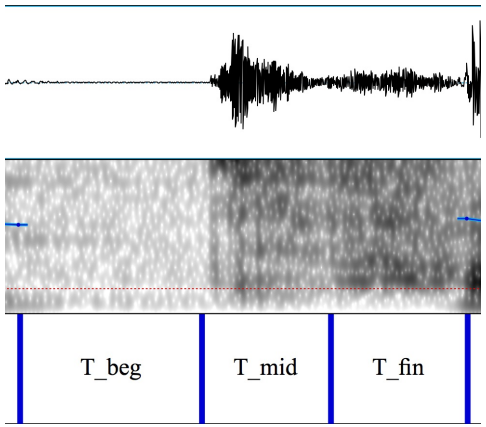



## acoustic models

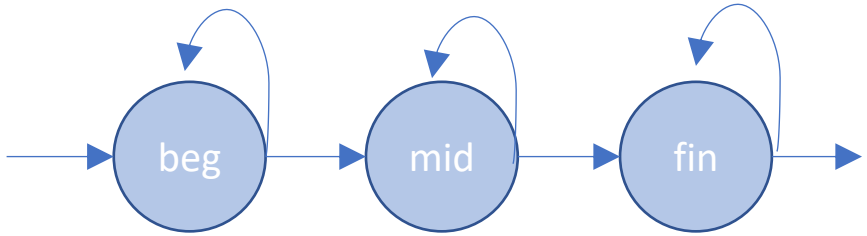
beware your assumptions

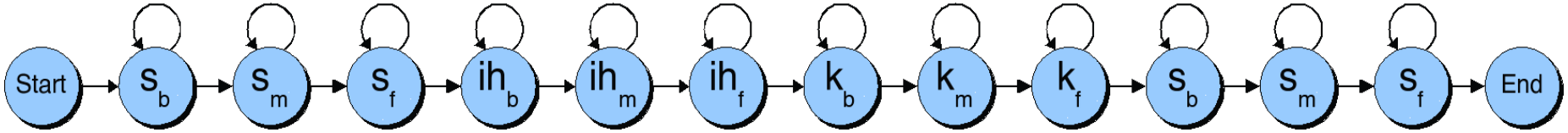
Your system is *not* assumption-free (but when is it ever, really?)

# Training/Alignment: Monophone Models




t

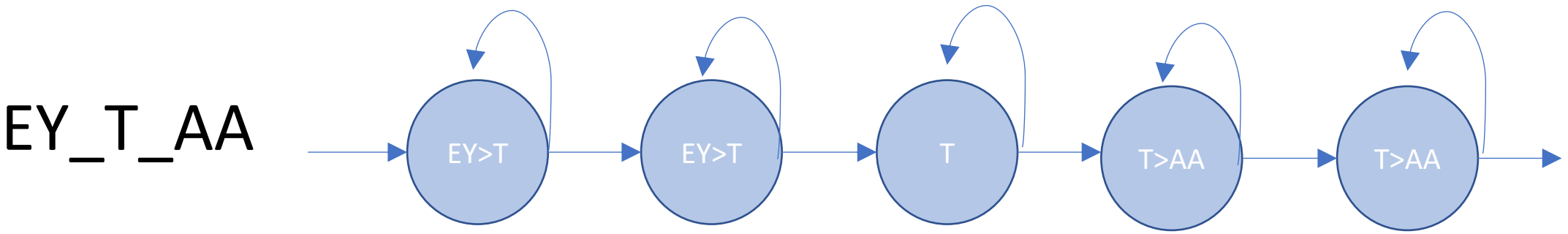


# Training/Alignment: Triphone Models

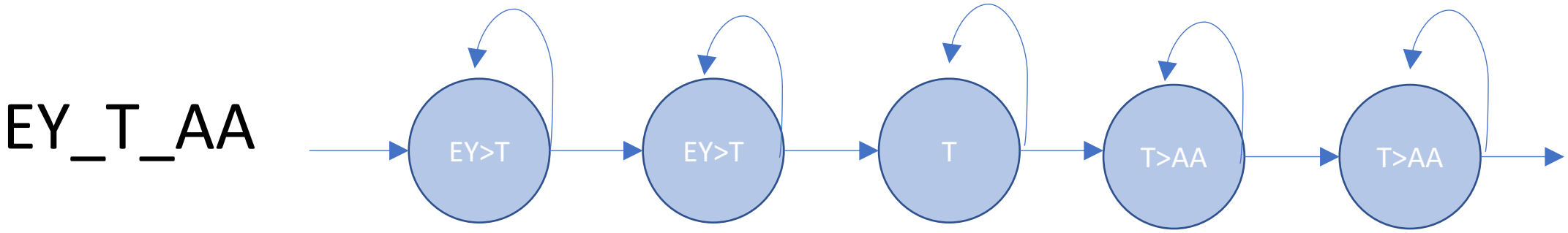
Context-dependent phones

Typically have 3 – 5 HMM states

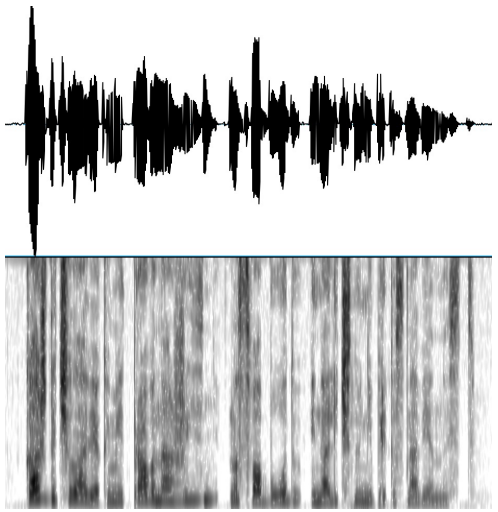


# Training/Alignment: Speaker-adapted Triphone Models

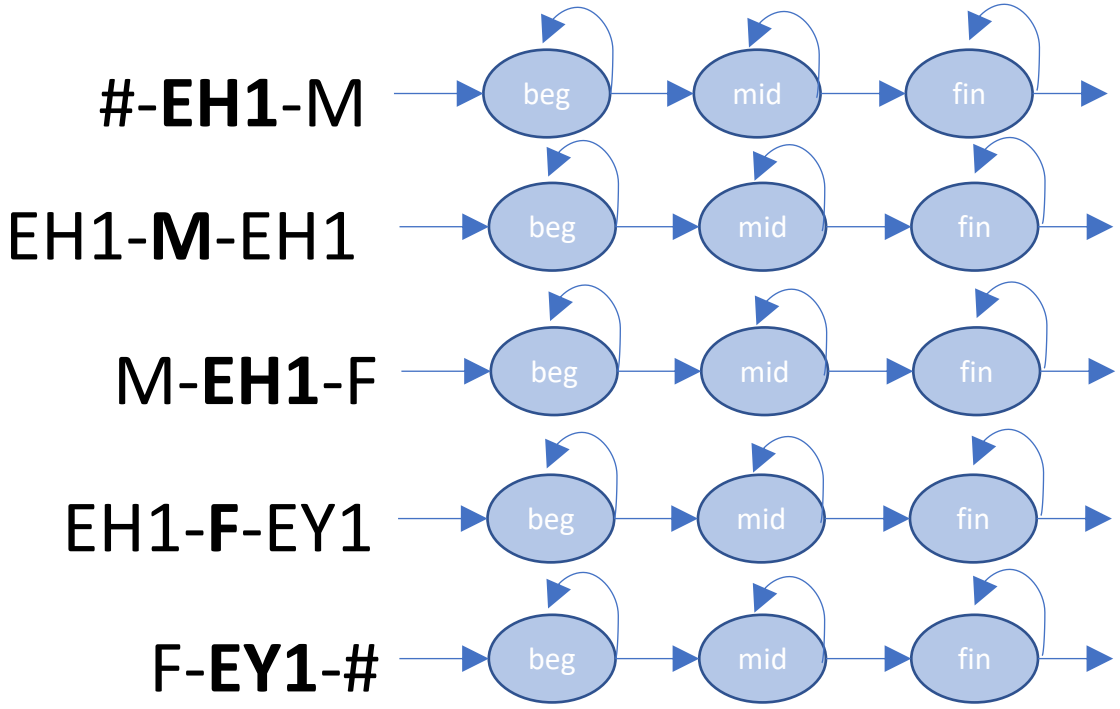
Create a transformation of the triphones for each speaker



# GMM-HMM+ → Acoustic models



DH AH N AO R TH W AY N D AH N D DH AH S AH N W ER D IH S P Y UW
T IH NG W IH CH W AA Z DH AH S T R AO NG G ER W EH N EY T R AE V
AH L ER K EY M AH L AO NG R AE P T IH N EY W AO R M K L OW K DH
EY AH G R IY D DH AE T DH AH W AH N HH UW F ER S T S AH K S IY D
AH D IH N M EY K IH NG DH AH T R AE V AH L ER T EY K HH IH Z K L
OW K AO F SH UH D B IY K AH N S IH D ER D S T R AO NG G ER DH AE
N DH AH AH DH ER DH EH N DH AH N AO R TH W AY N D B L UW AE Z HH
AA R D AE Z K UH D B AH T DH AH M AO R HH IY B L UW DH AH M AO R
K L OW S L IY D IH D DH AH T R AE V AH L ER F OW L D HH IH Z K L
OW K ER AW N D HH IH M AH N D AE T L AE S T DH AH N AO R TH W AY
N D G EY V AH P DH AH AH T EH M P T DH EH N DH AH S AH N SH AY N
D AW T W AO R M L IY AH N D IH M IY D IY AH T L IY DH AH T R AE
V AH L ER T UH K AO F HH IH Z K OW T AH N D S OW DH AH N AO R TH
W AY N D W AA Z AH B L AY JH D T UW K AH N F EH S DH AE T DH AH
S AH N W AA Z DH AH S T R AO NG G ER AH V DH AH T UW

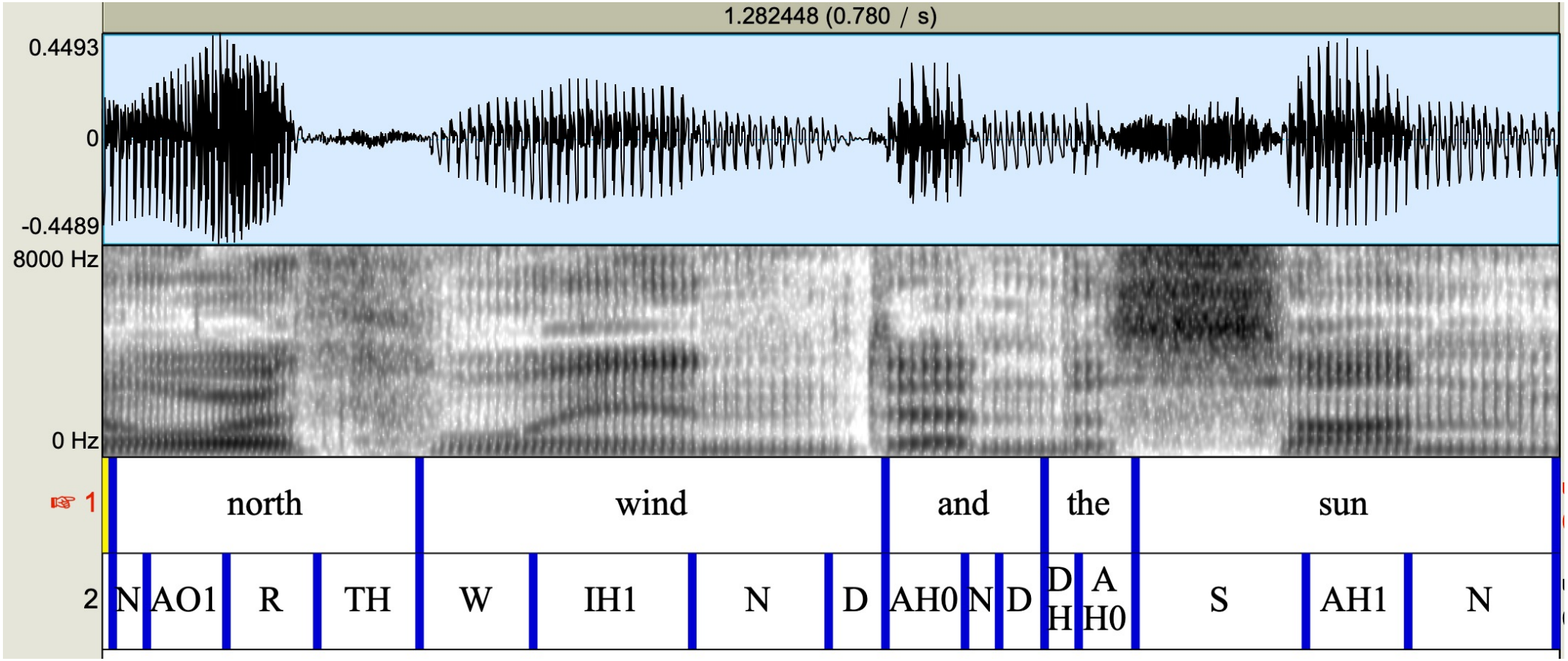



## acoustic models

beware your assumptions

Your system is *not* assumption-free (but when is it ever, really?)

# Output


1.282448 (0.780 / s)
0.4493
0
-0.4489
8000 Hz
0 Hz
1
north
wind
and
the
sun
2
N AO1 R TH W IH1 N D AH0 N D DH AH0 S AH1 N

# Forced Aligner

Composed of a:
Pronunciation dictionary
Set of acoustic models
Algorithm