Building an ASR using HTK CS4706

Fadi Biadsy

Mar 24th, 2010

Outline

- Speech Recognition
- Feature Extraction
- Modeling Speech
  - Hidden Markov Models (HMM): 3 basic problems
- HMM Toolkit (HTK)
  - Steps for building an ASR using HTK

Automatic Speech Recognition (ASR)

- Speech signal to text

ASR → There’s something happening when Americans…

It’s hard to recognize speech

- Contextual effects
  - Speech sounds vary within contexts
    - “How **do** you **do**?”
    - *Half and half*
    - */t/ in butter vs. bat*
- Within-speaker variability
  - Speaking rate, Intensity, F0 contour
  - Voice quality
  - Speaking Style
    - Formal vs. spontaneous register
    - Speaker State: Emotion, Sleepy, Drunk,…
- Between-speaker variability
  - Gender and age
  - Accents, Dialects, native vs. non-native
    - Scottish vs. American /r/ in some contexts
- Environment variability
  - Background noise
  - Microphone type

Outline

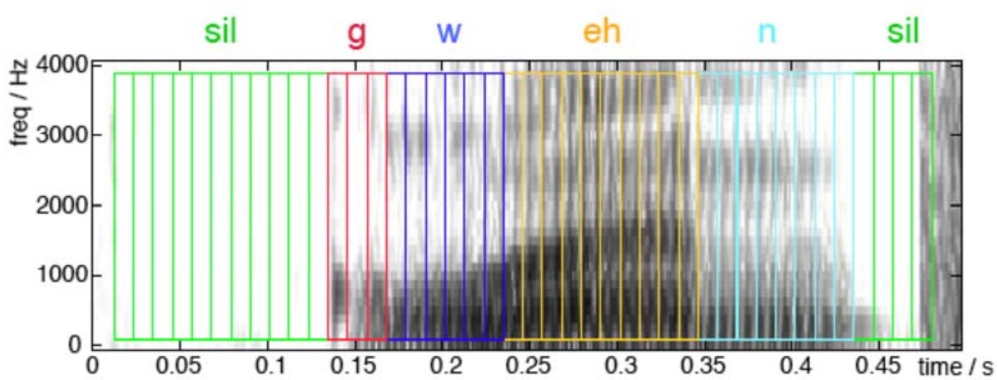
- Speech Recognition
- Feature Extraction
- Modeling Speech
  - Hidden Markov Models (HMM): 3 basic problems
- HMM Toolkit (HTK)
  - Steps for building an ASR using HTK

Feature Extraction

- Wave form?
- Spectrogram?
- Need representation of speech signal that is robust to acoustic variation but sensitive to linguistic content

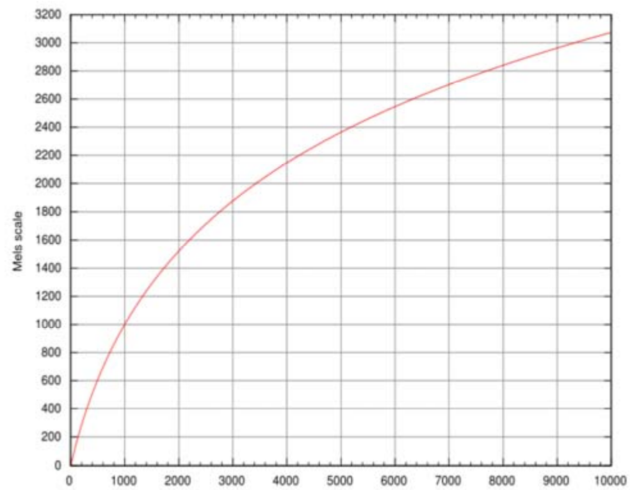
Feature Extraction

- Extract features from short frames (frame period 10ms, 25ms frame size) – a sequence of features



Feature Extraction - MFCC

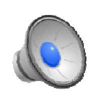
- Mel Scale: Approximate the unequal sensitivity of human hearing at different frequencies
- Based on pitch perception

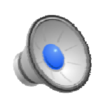


Feature Extraction - MFCC

- MFCC (Mel frequency cepstral coefficient)
  - Widely used in speech recognition

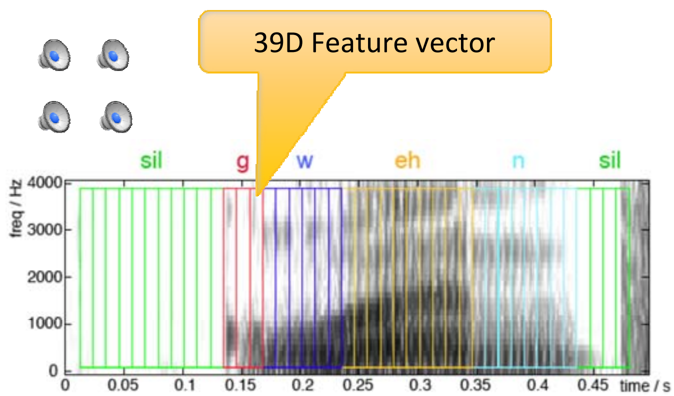
1. Take the Fourier transform of the signal ➔ spectrum
2. Map the powers of the spectrum to the mel scale and take the log
3. Discrete cosine transform of the mel log-amplitudes
4. The MFCCs are the amplitudes of the resulting spectrum





Feature Extraction - MFCC

- Extract a feature vector from each frame
  - 12 MFCC coefficients + 1 normalized energy = 13 features
  - Delta MFCC = 13
  - Delta-Delta MCC = 13
- Total: 39 features
- Inverted MFCCs:

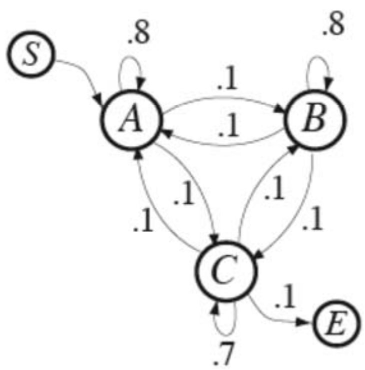


Outline

- Speech Recognition
- Feature Extraction
- Modeling Speech
  - Hidden Markov Models (HMM): 3 basic problems
- HMM Toolkit (HTK)
  - Steps for building an ASR using HTK

Markov Chain

- Weighted finite state acceptor: Future is independent of the past given the present



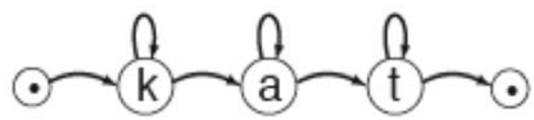
$p(q_{n+1} \mid q_n)$		q_{n+1}: S	A	B	C	E
q_n	S	0	1	0	0	0
	A	0	.8	.1	.1	0
	B	0	.1	.8	.1	0
	C	0	.1	.1	.7	.1
	E	0	0	0	0	1

S A A A A A A A A B B B B B B B B B C C C C B B B B B B C E

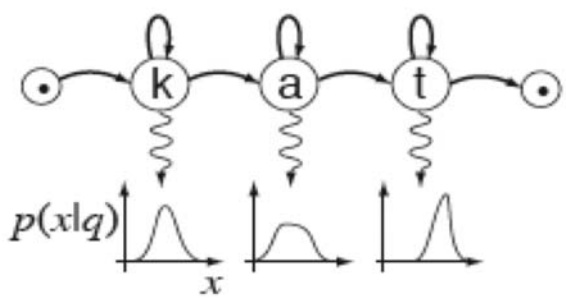
Hidden Markov Model (HMM)

- HMM is a Markov chain + emission probability function for each state

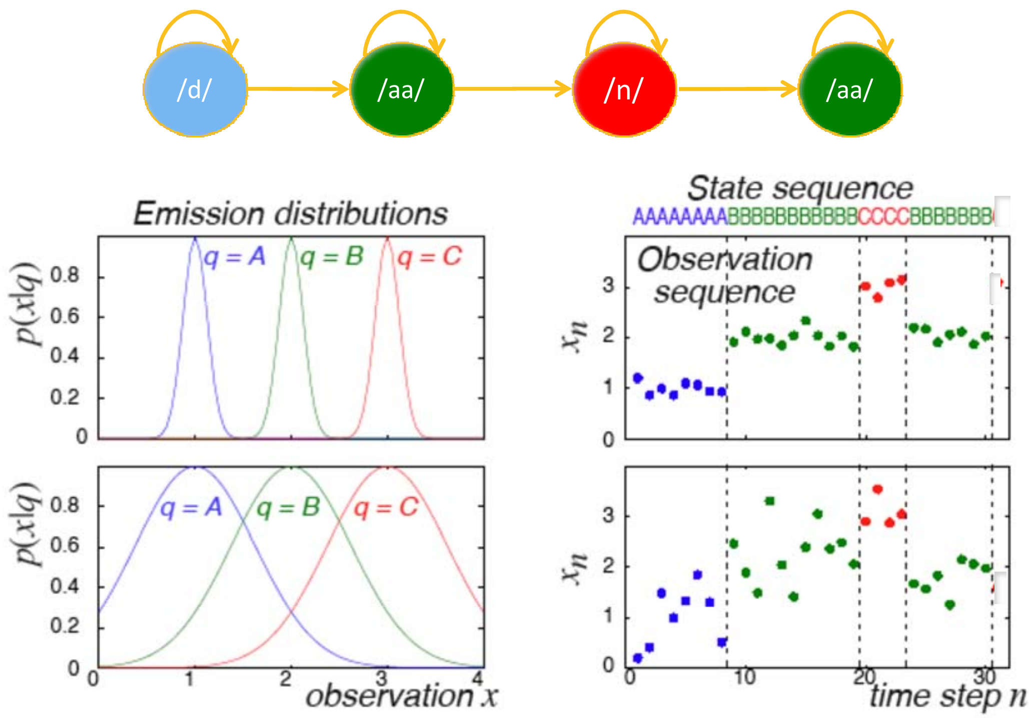
- Markov Chain



- HMM M=(A, B, Pi)
- **A = Transition Matrix**
- **B = Observation Distributions**
- **Pi = Initial state probabilities**



HMM Example



/d/
/aa/
/n/
/aa/
Emission distributions
p(x|q)
q = A
q = B
q = C
observation x
State sequence
AAAAAAAABBBBBBBBBBBBCCCCBBBBBBB
Observation sequence
x_n
time step n

HMM – 3 Basic Problems

I. Evaluation

II. Decoding

III. Training

HMM – I. Evaluation

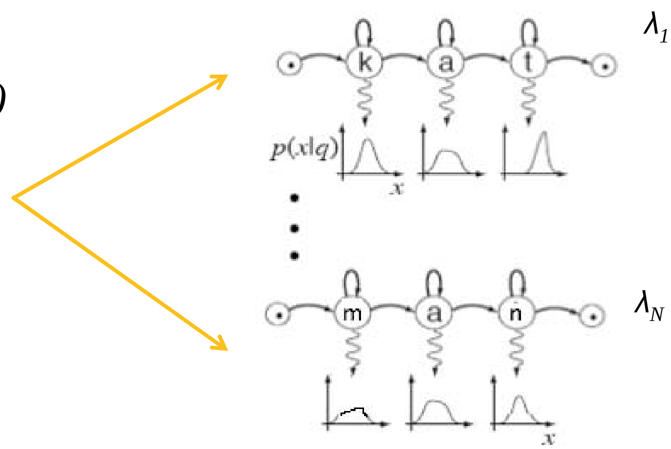
- Given an observation sequence O and a model *M,* how can we efficiently compute:

  $P(O \mid M) =$ the likelihood of O given the model?

$argmax_i\ P(\lambda_i|O)$

$\Leftrightarrow$

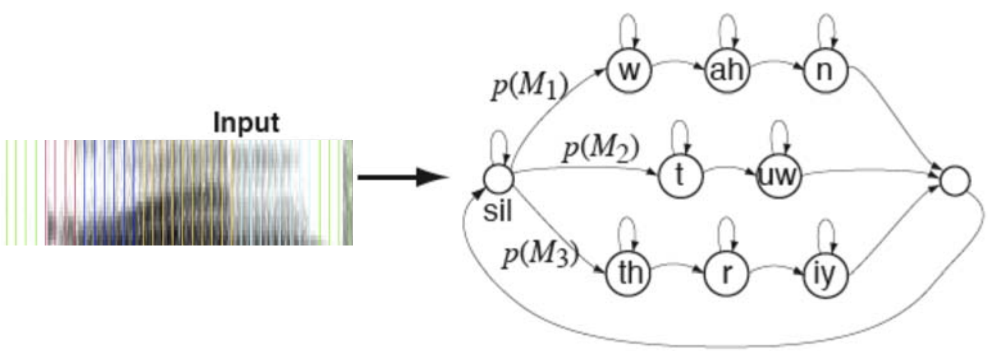
$argmax_i\ P(O|\lambda_i)\ P(\lambda_i)$



HMM – II. Decoding

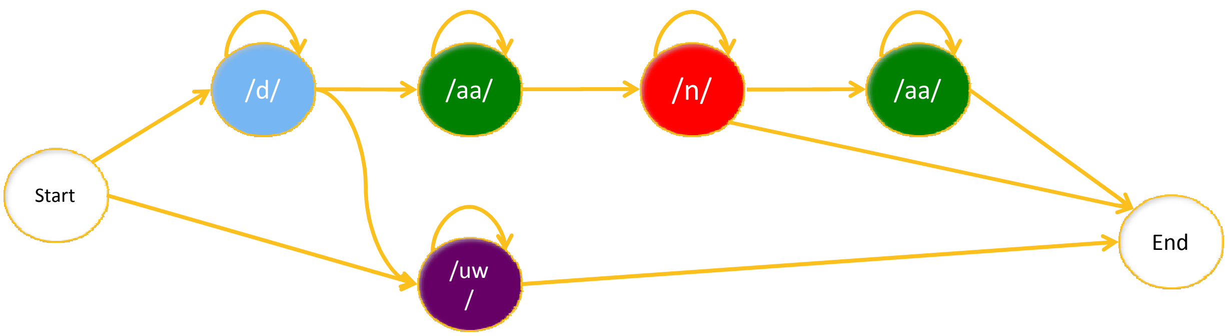
- Given an observation sequence O and a model M:

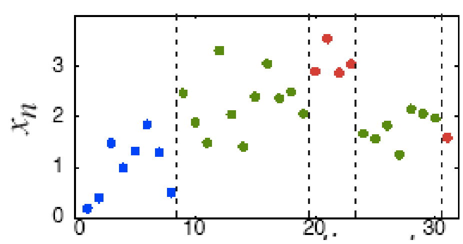
  How can we obtain the most likely state sequence Q = {q1, q2,…,qt}?



Viterbi Algorithm

- Efficient algorithm for decoding O(TN^2)

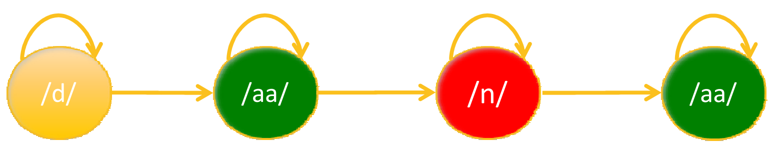




/d/ /aa/ /n/ /aa/ => dana

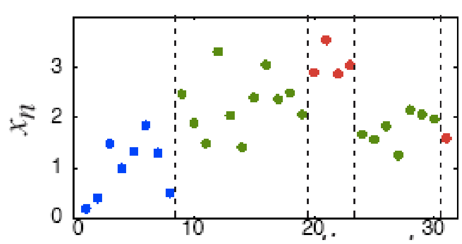
HMM – III. Training

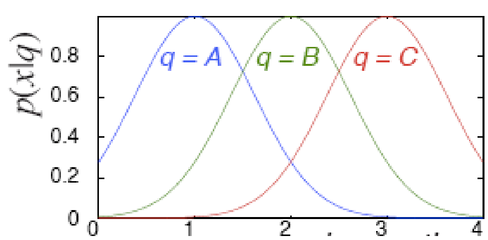
- How do we estimate the model parameters M=(***A, B, Pi***) *to maximize P(O|M)?*
  - *Baum-Welch algorithm*



dana => /d/ /aa/ /n/ /aa/

1) Transition Matrix: A
2) Emission probability distribution:





Outline

- Speech Recognition
- Feature Extraction
- Modeling Speech
  - Hidden Markov Models (HMM): 3 basic problems
- HMM Toolkit (HTK)
  - Steps for building an ASR using HTK

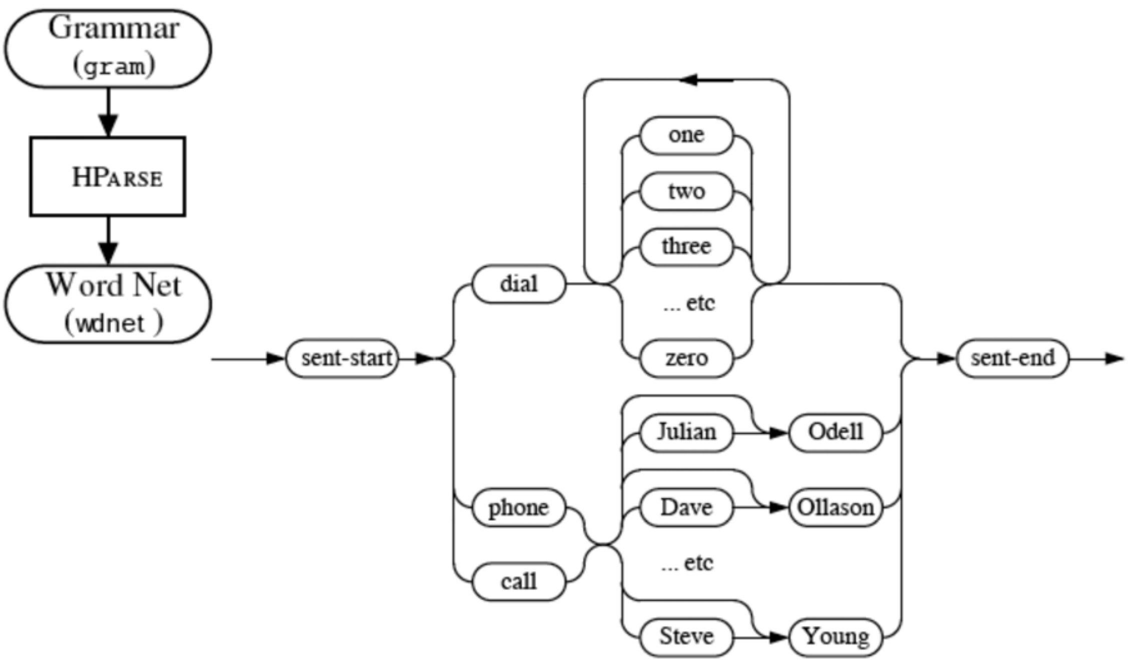
Hidden Markov Model Toolkit (HTK)

- HTK is a research toolkit for building and manipulating HMMs
- Primarily designed for building HMM-based ASR systems
- Tools, for examples:
  - Extracting MFCC features
  - HMM algorithms
  - Grammar networks
  - Speaker Adaptation
  - …

Steps for building ASR: Voice-operated interface for phone dialing

- Examples:
  - Dial three three two six five four
  - Phone Woodland
  - Call Steve Young
- Grammar:
  - $digit = ONE | TWO | THREE | FOUR | FIVE | SIX | SEVEN | EIGHT | NINE | OH | ZERO;
  - $name = [JOOP] JANSEN | [JULIAN] ODELL | [DAVE] OLLASON | [PHIL] WOODLAN
  - (SENT-START (DIAL <$digit> | (PHONE|CALL) $name) SENT-END)

Convert Grammar to Network (HParse)



Training the system

Lab files:

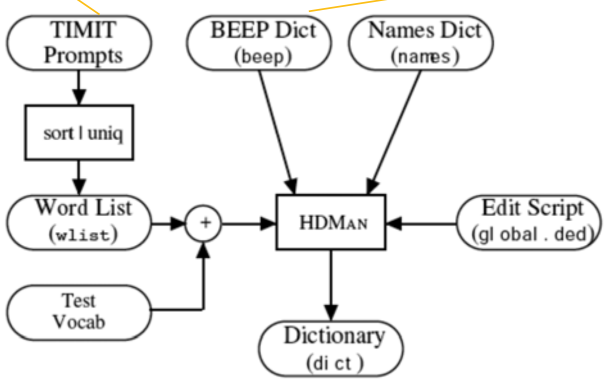
Wave files

S0001 ONE VALIDATED ACTS OF SCHOOL DISTRICTS

S0002 TWO OTHER CASES ALSO WERE UNDER ADVISEMENT

S0003 BOTH FIGURES WOULD GO HIGHER IN LATER YEARS

A	ah sp
A	ax sp
A	ey sp
CALL	k ao l sp
DIAL	d ay ax l sp
EIGHT	ey t sp
PHONE	f ow n sp
...	



Words to Phones (using HLEd)

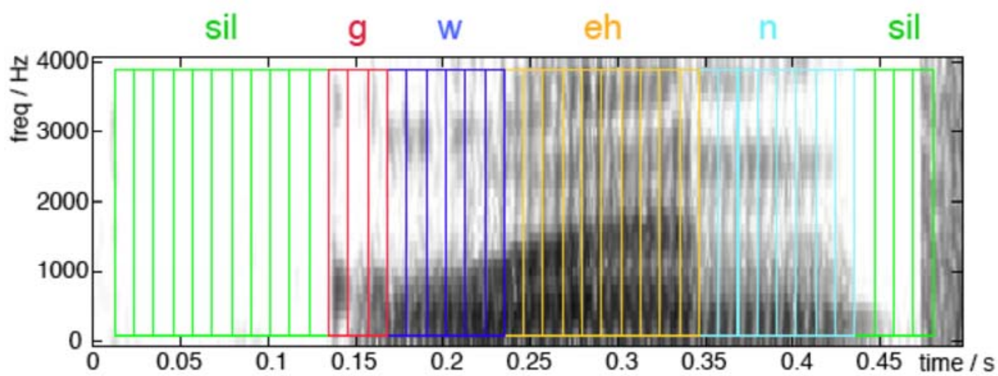
- HTK scripting language is used to generate phonetic transcription for all training data

```
#!MLF!#
"*/S0001.lab"
ONE
VALIDATED
ACTS
OF
SCHOOL
DISTRICTS
.
"*/S0002.lab"
TWO
OTHER
CASES
ALSO
WERE
UNDER
ADVISEMENT
.
"*/S0003.lab"
BOTH
FIGURES
(etc.)
```

```
#!MLF!#
"*/S0001.lab"
sil
w
ah
n
v
ae
l
ih
d
.. etc
```

Extracting MFCC (using HCopy)

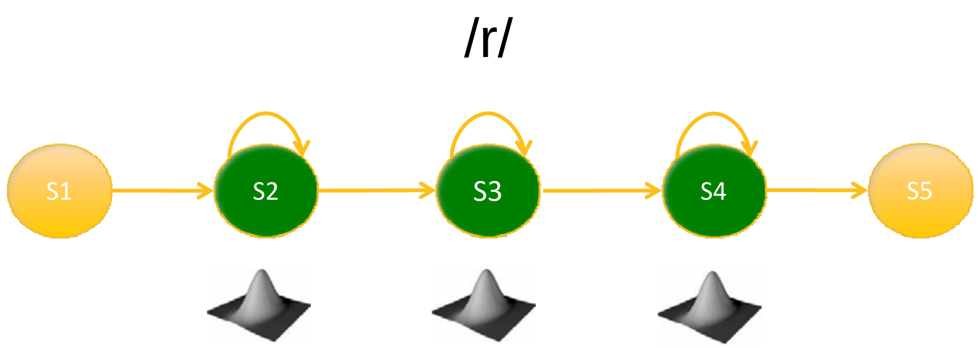
- For each wave file, extract MFCC features.



- .wav ➔ .mfc files

Specifying Monophone HMM Topology

- 5 states: 3 emitting states



- Flat Start: Mean and Variance are initialized as the global mean and variance of all the data

Training (HERest)

- For each training pair of files (mfc+lab):

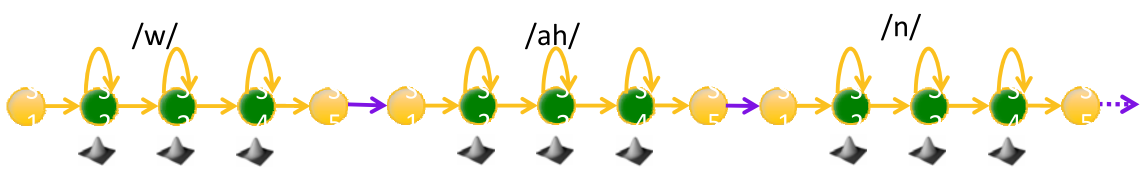
  1. Concatenate the corresponding monophone HMMs

  2. Use the Baum-Welch Algorithm to train the

     HMMs given the MFC features

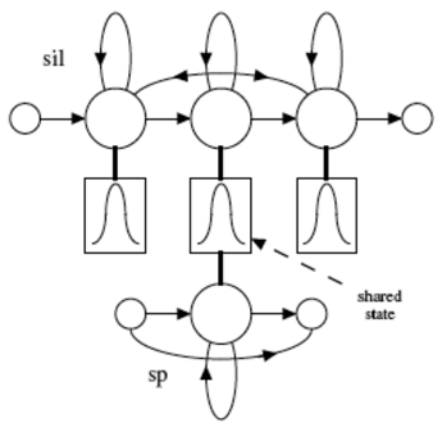
```
#!MLF!#
"*/S0001.lab"
sil
w
ah
n
v
ae
l
ih
d
.. etc
```

One validated acts of school districts…



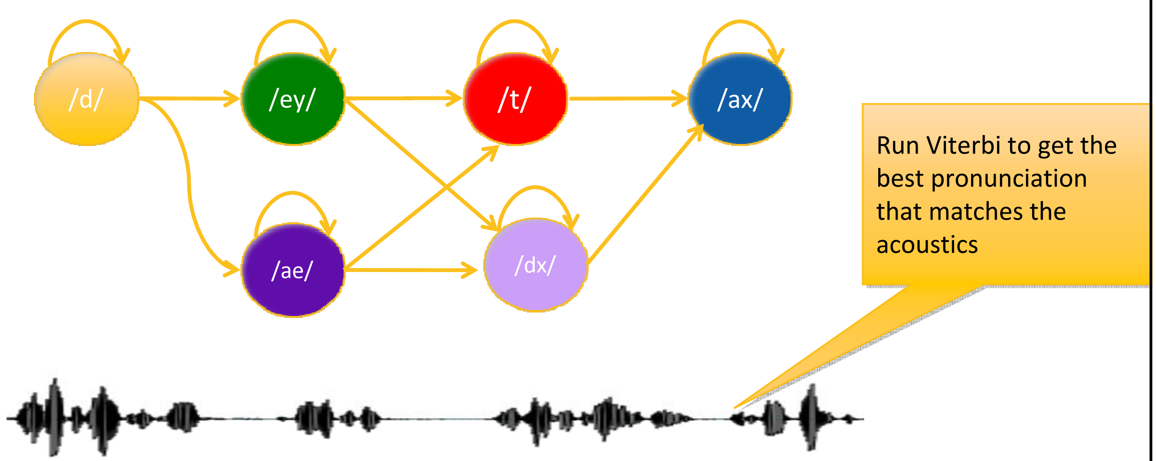
Training

- So far, we have all monophone models trained
- Train the short pause (*sp)* model



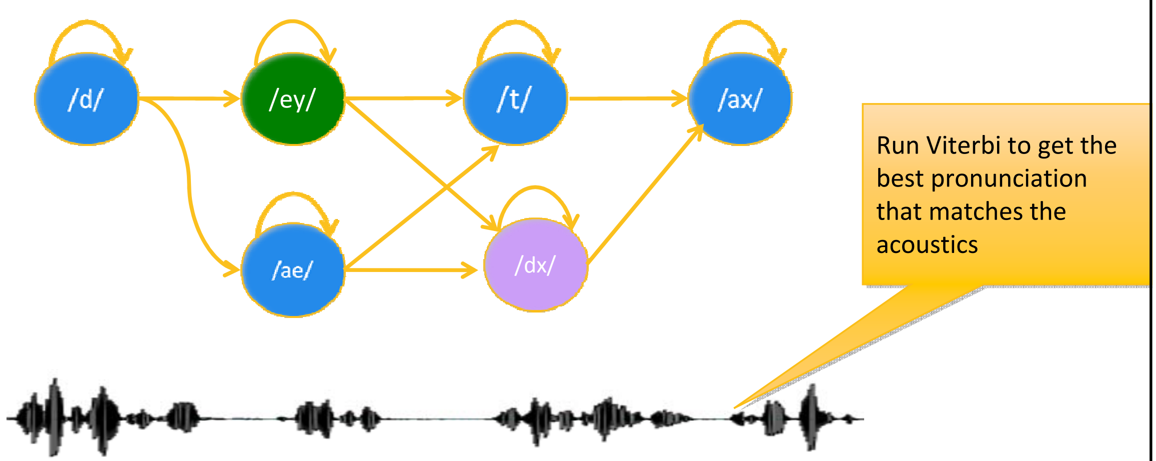
Handling Multiple Pronunciations (HVite)

- The dictionary contains multiple pronunciations for some words.
- Forced alignment



Handling Multiple Pronunciations

- The dictionary contains multiple pronunciations for some words.
- Forced alignment



Retrain

- After getting the best pronunciation
  - => Train again using Baum-Welch algorithm using the best pronunciations

Creating Triphone Models (using HLEd)

- Phones may be realized differently in some contexts

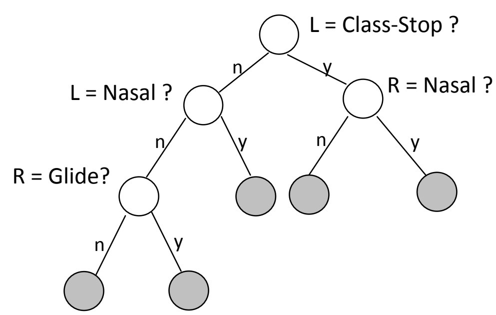
➔ Build context-dependent acoustic models (HMMs)

- Triphones: One preceding and succeeding phone
- Make triphones from monophones
  - Generate a list of all the triphones for which there is at least one example in the training data
  - s-l+ow
  - b-l+aa
  - p-l+aa
  - jh-oy+s
  - f-iy+t
  - …

Tie Triphone (HDMan)

- Clustering by growing decision trees
- All states in the same leaf will be tied

t+ih	t+ae
t+iy	t+ae
ao-r+ax	r
t+oh	t+ae
ao-r+iy	
t+uh	t+ae
t+uw	t+ae
sh-n+t	
sh-n+z	sh-n+t
ch-ih+l	
ay-oh+l	
ay-oh+r	ay-oh+l



After Tying

- Train the acoustic models again using Baum-Welch algorithm (HERest)
- Increase the number of Gaussians for each state
  - HHEd followed by HERest

Decoding (HVite)

- Using the compiled grammar network (WNET)
- Given a new speech file:
  - Extract the mfcc features (.mfc file)
  - Run Viterbi on the WNET given the .(mfc file) to get the most likely word sequence

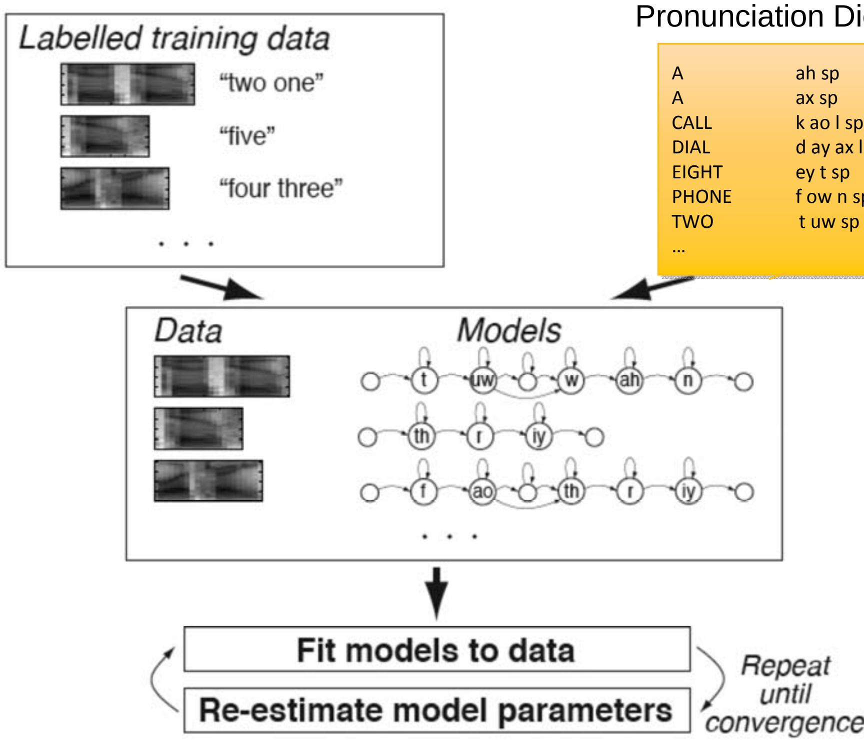
Summary

- MFCC Features
- HMM 3 basic problems
- Steps for Building an ASR using using HTK:
  - Features and data preparation
  - Monophone topology
  - Flat Start
  - Training monophones
  - Handling multiple pronunciations
  - Context-dependent acoustic models (triphones) + Tying
  - Final Training
  - Decoding

Thanks!

HMM – III. Training

•



Pronunciation Dictionary

A	ah sp
A	ax sp
CALL	k ao l sp
DIAL	d ay ax l sp
EIGHT	ey t sp
PHONE	f ow n sp
TWO	t uw sp
…	