Forced Alignment and Speech Recognition Systems

Overview

- Uses of automatic speech recognition technology
- Principles of forced alignment and speech recognition systems
- Some practicalities
- Evaluating alignment quality

ASR technology - Existing uses

As a toolbox:

Pre-built generic application used as a tool

> speech recognition

> forced alignment

for lexical transcription or time stamps

As a methodology:

Forms an integral part of the experimental procedure

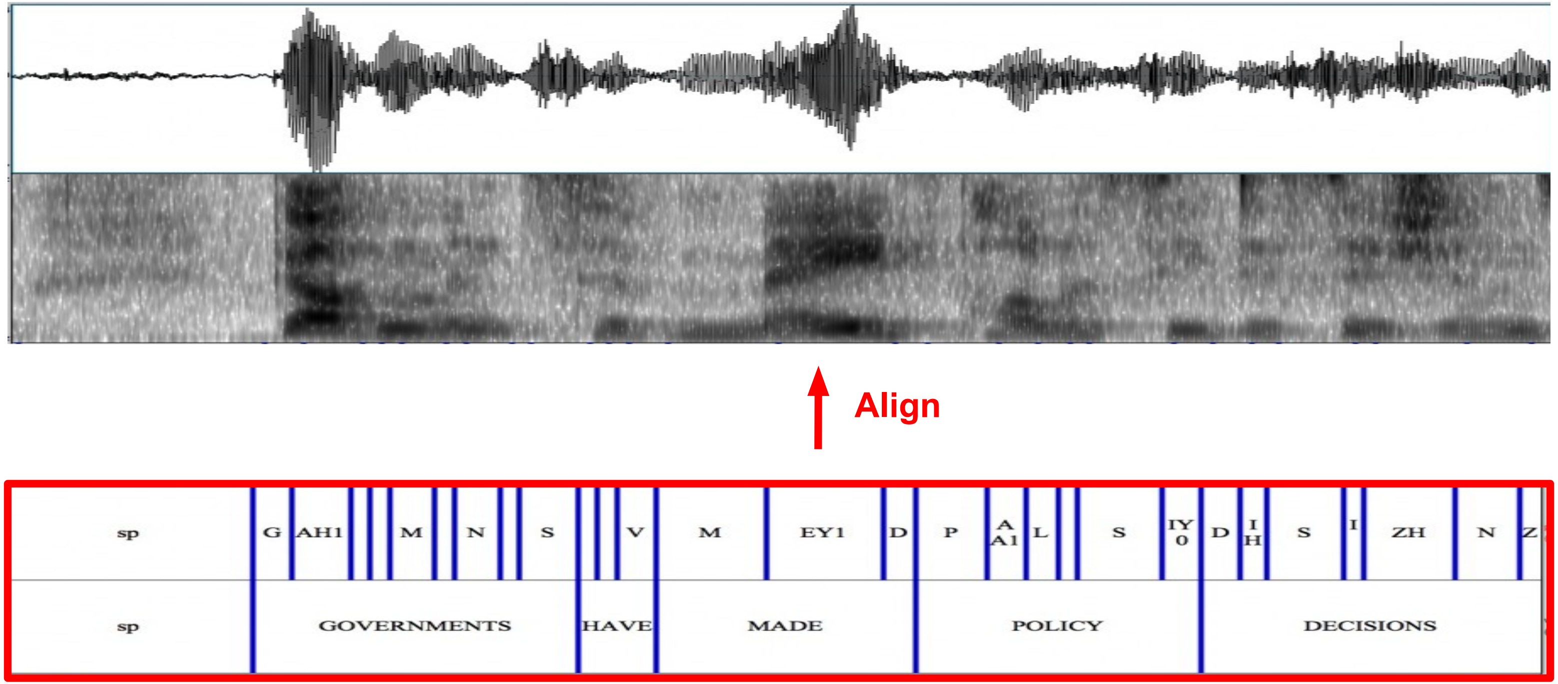
> tune for unusual pronunciations

> extract probability of words/phonemes matching models

> detect assimilation, deletion, insertion

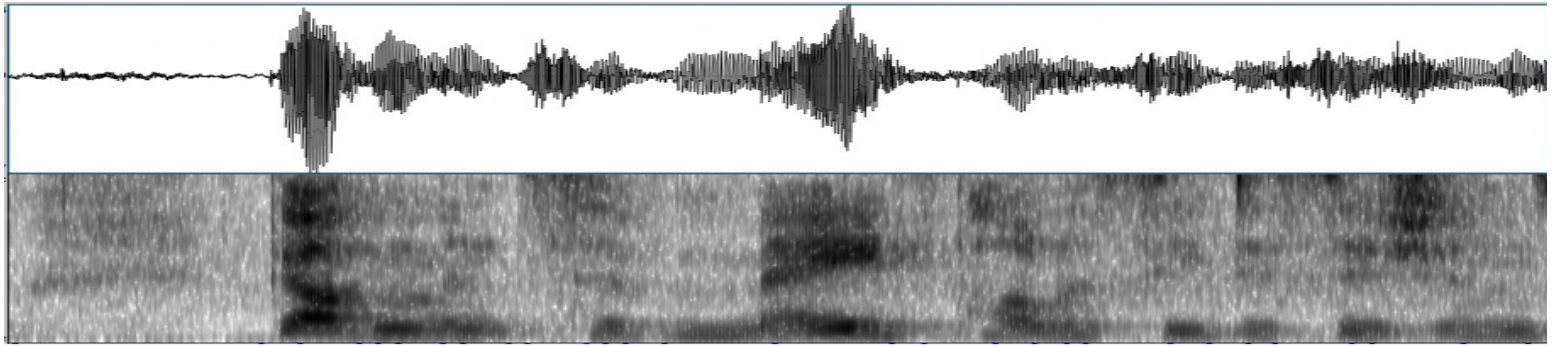
Forced alignment

With transcription:
Already know exactly what is in the audio.

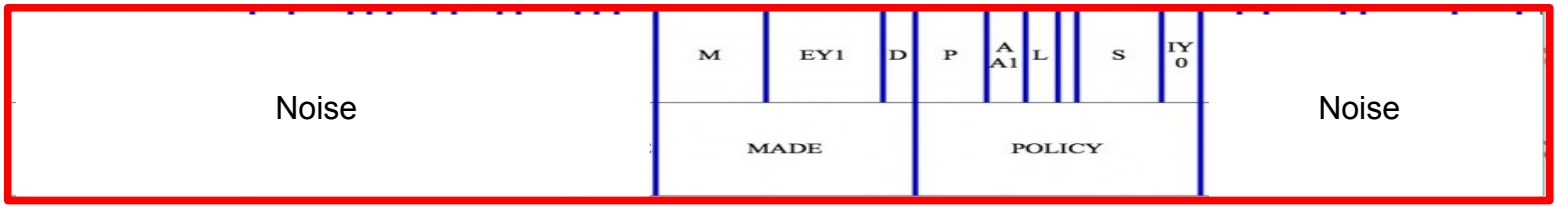


Forced alignment

With some transcription:
Know what is in some of the audio.

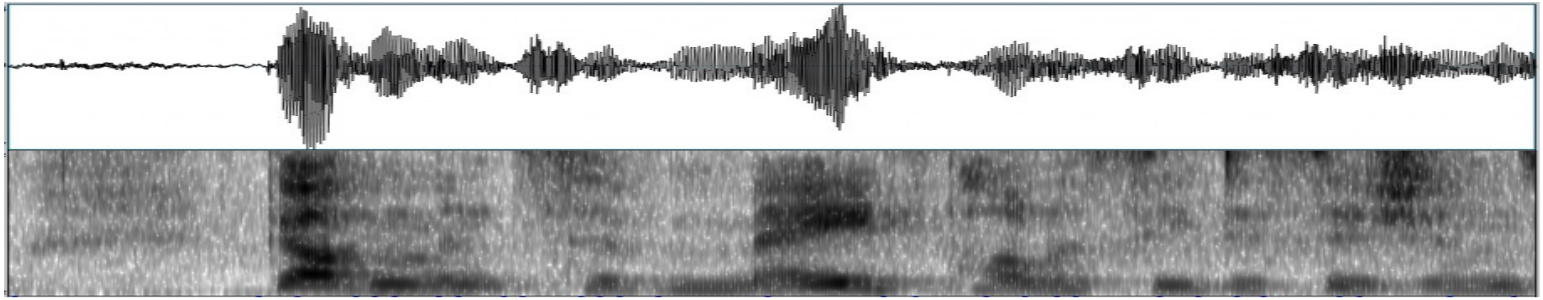


Align

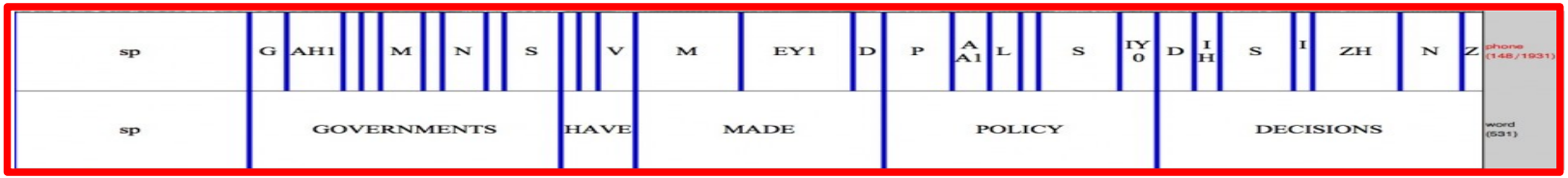


Automatic speech recognition

No transcription:
Don't know what's in the audio

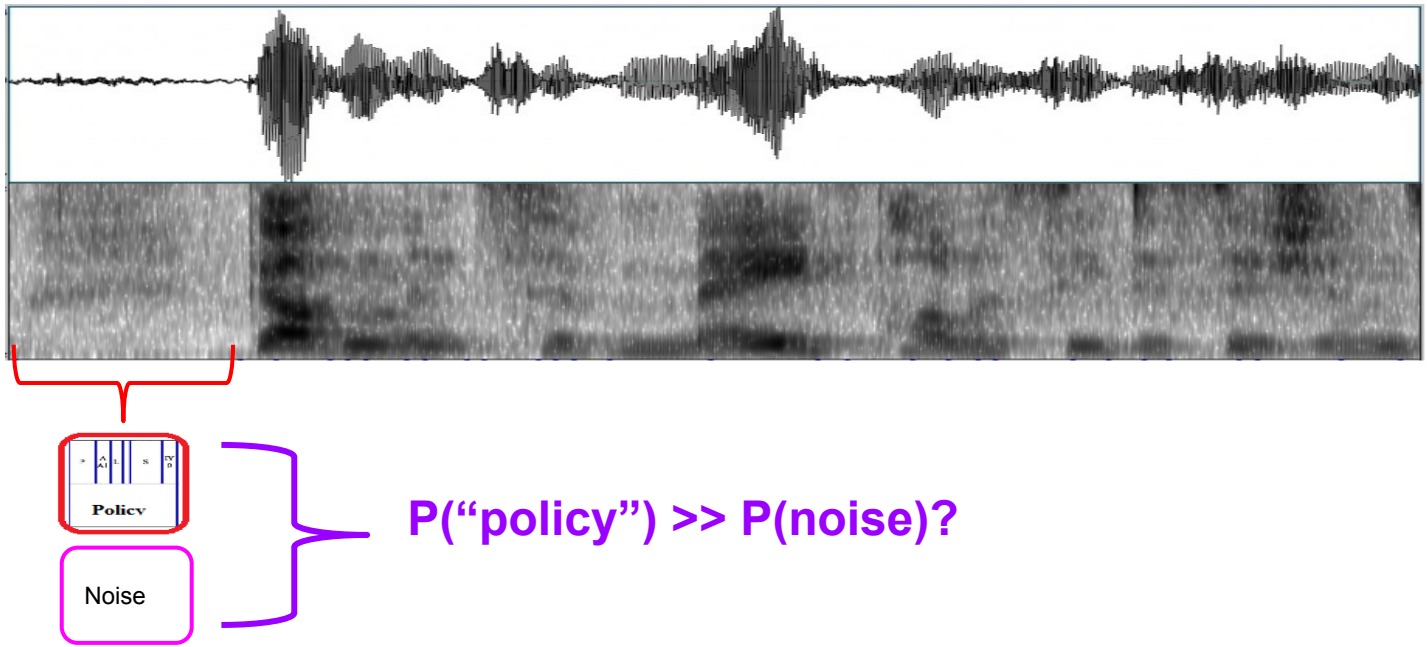


Estimate



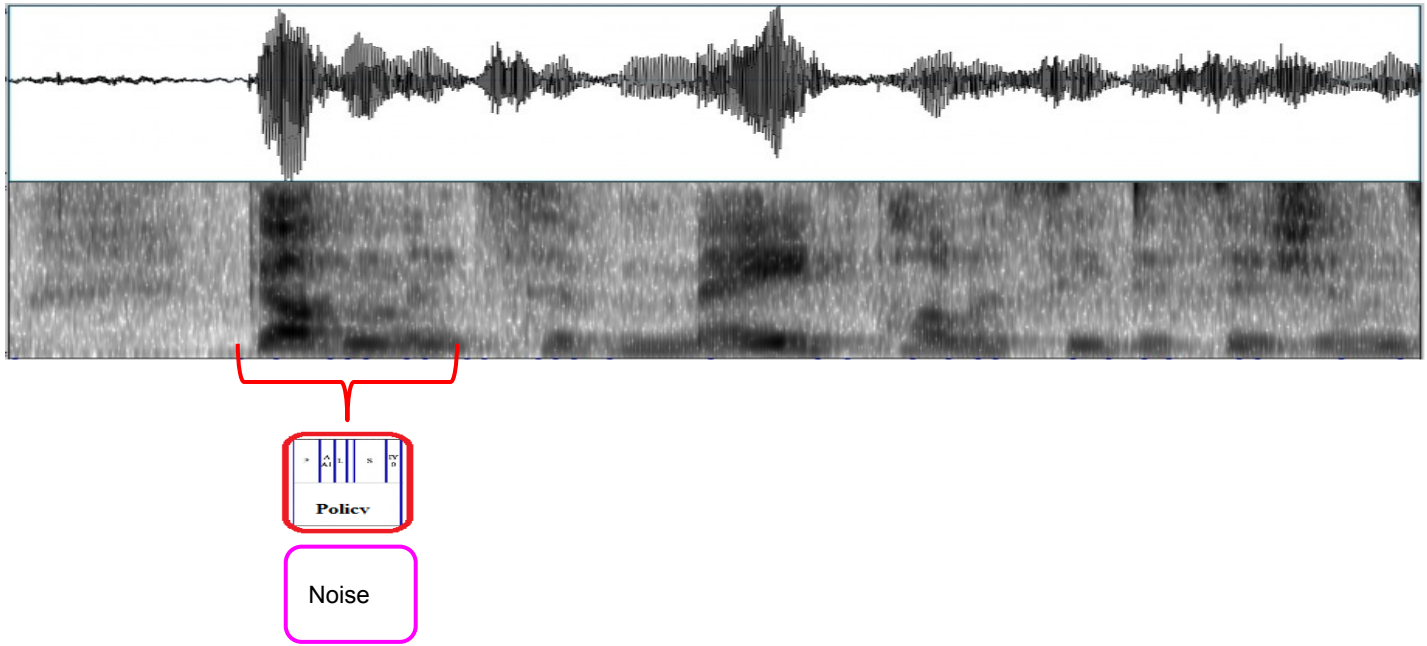
Word spotting

Possible transcription:
Looking for a word/phrase



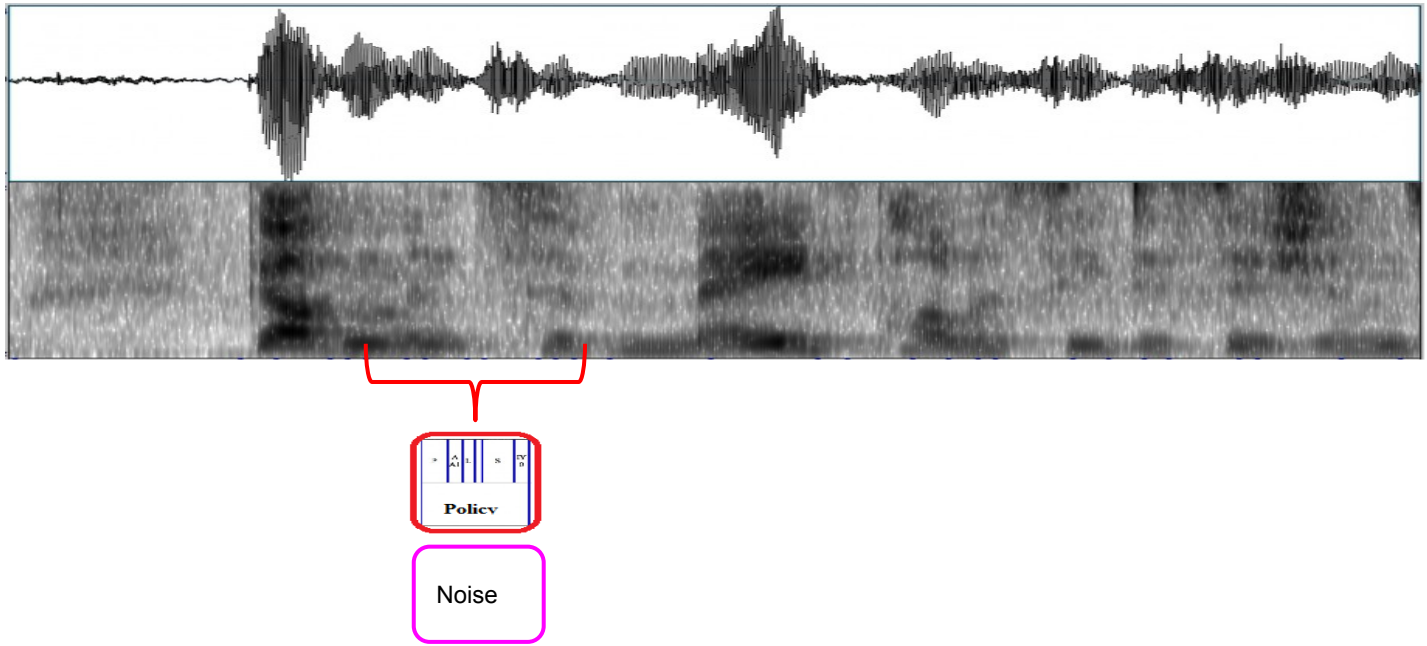
Word spotting

Possible transcription:



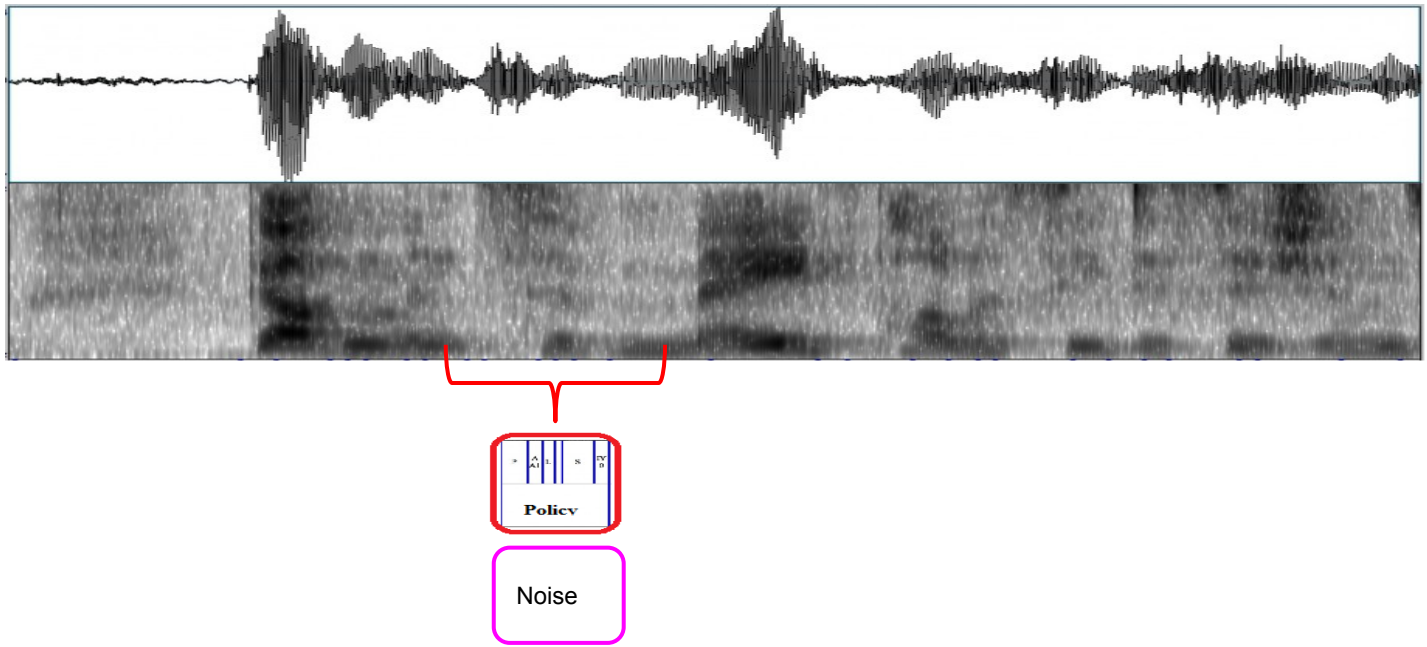
Word spotting

Possible transcription:



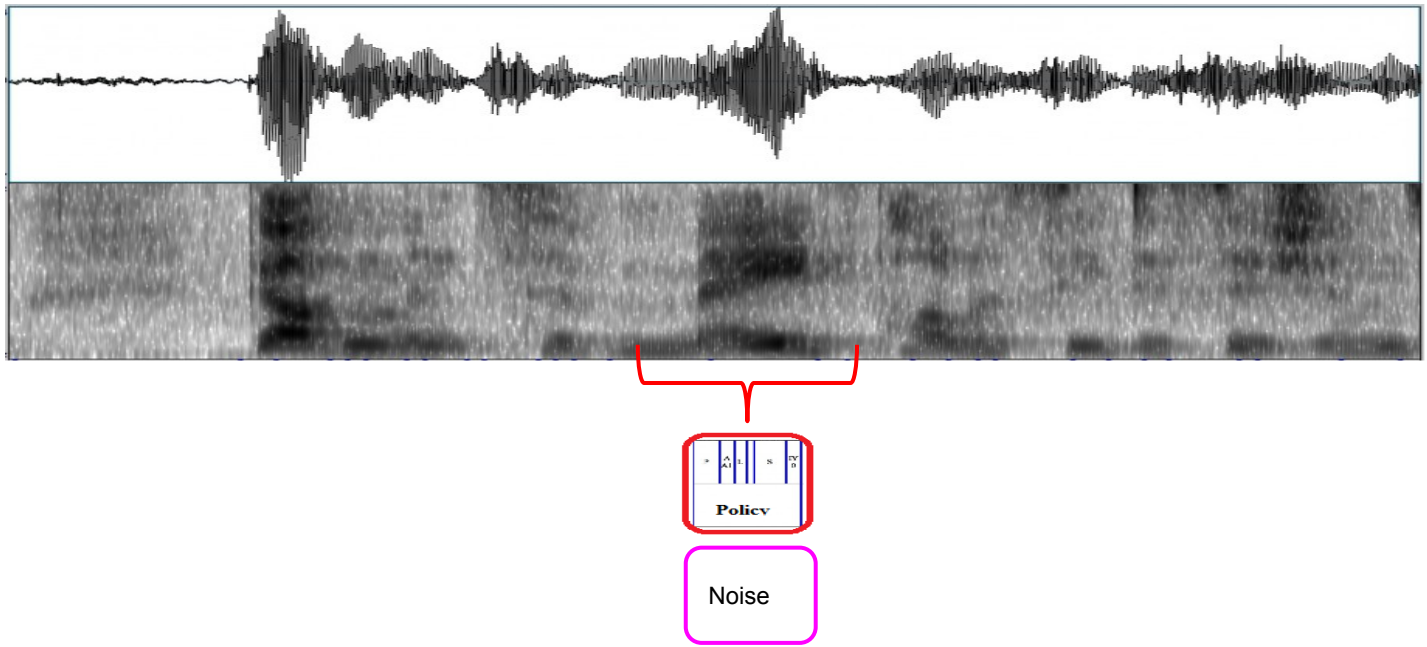
Word spotting

Possible transcription:



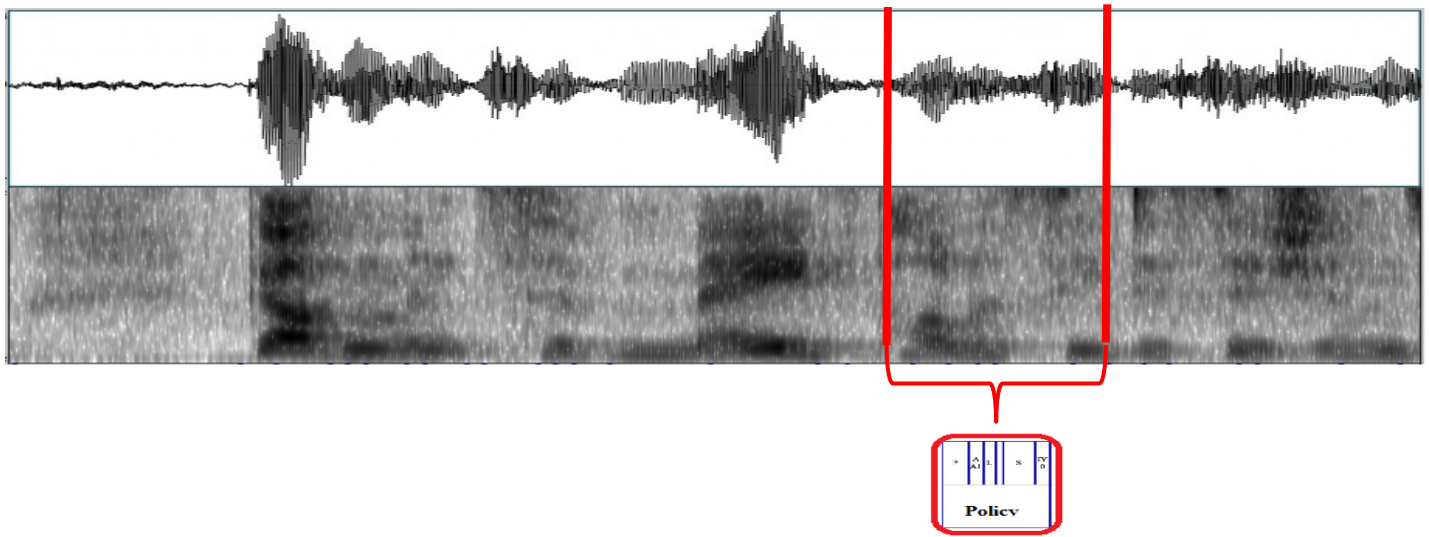
Word spotting

Possible transcription:



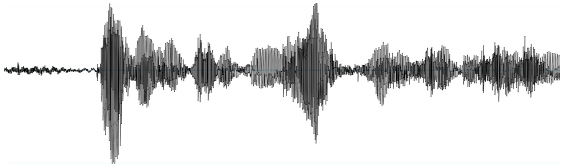
Word spotting

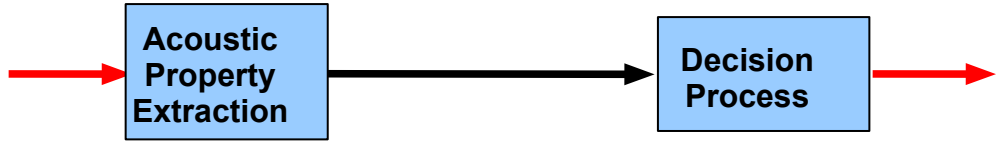
Possible transcription:

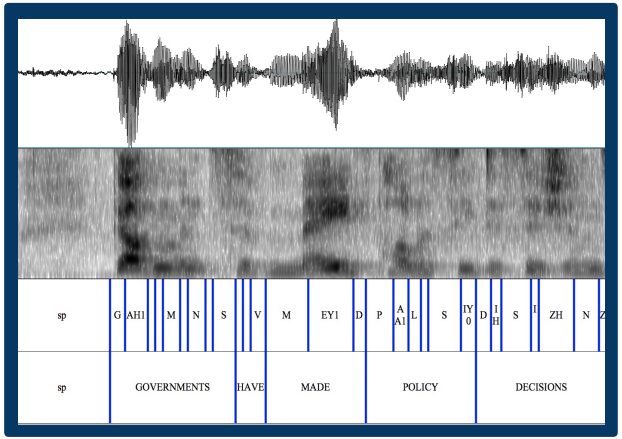


Yes! P(“policy”) >> P(noise)

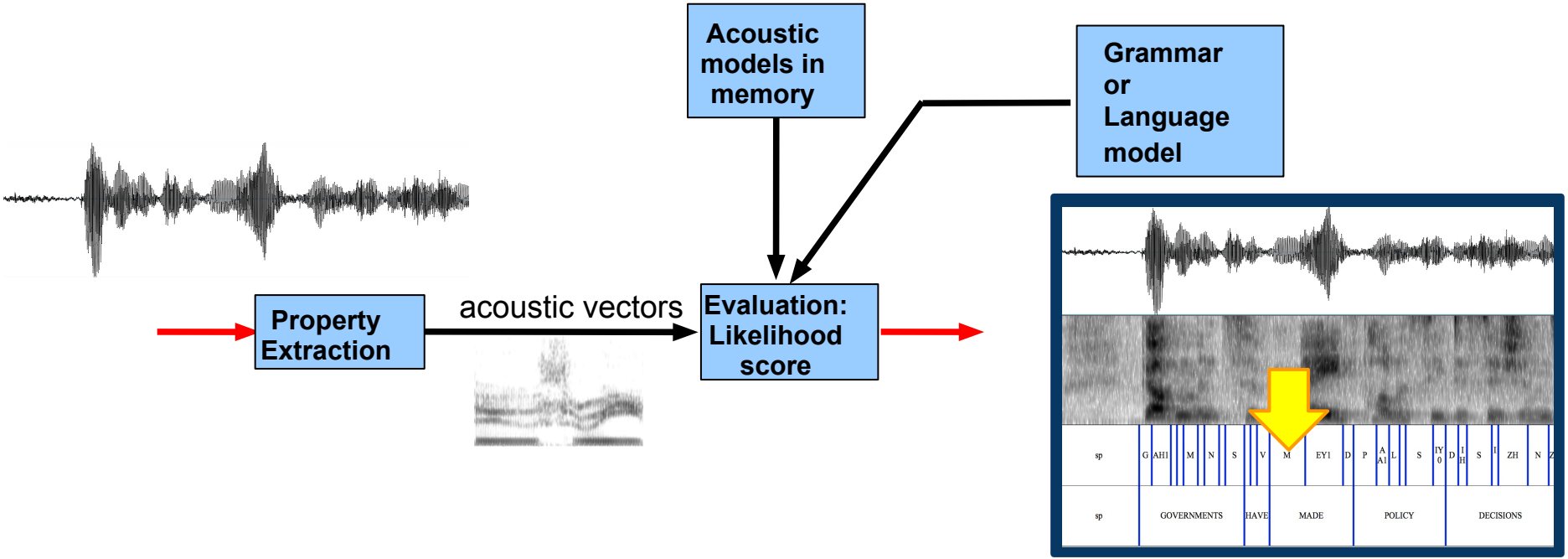
Early Automatic Speech Recognition (ASR) Systems





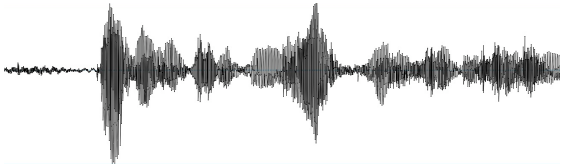


Automatic Speech Recognition System



Forced Alignment system

Words	phonemes
ADVERSITY	AE0 D V ER1 S AH0 T IY0
BLUEST	B L UW1 IH0 S T
BLUESTEIN	B L UH1 S T AY0 N

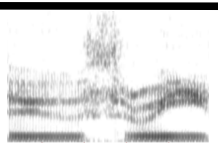


Acoustic models in memory

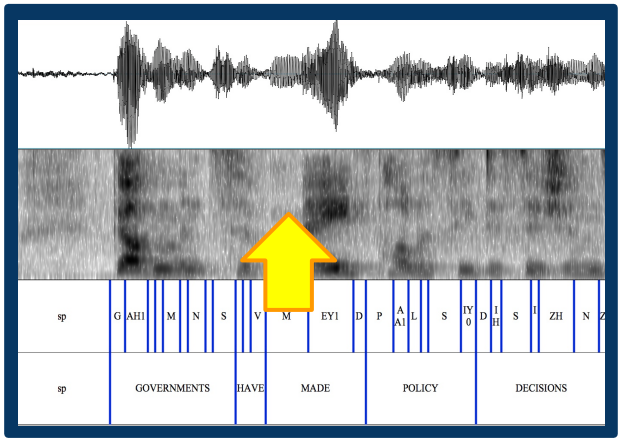
Lexicon & Transcription

Property Extraction

acoustic vectors

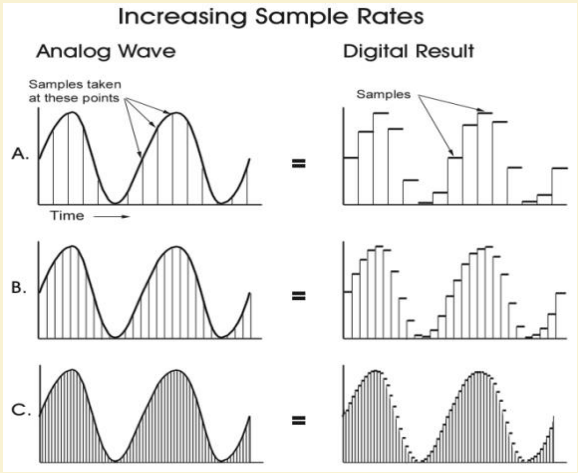


Evaluation: Likelihood score

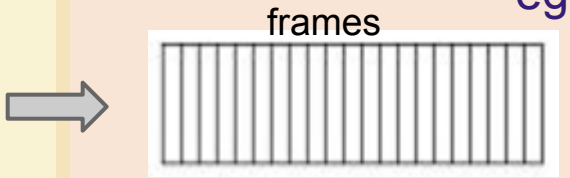


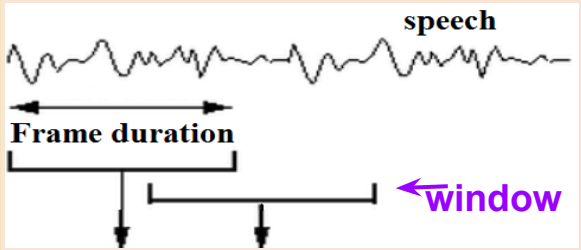
Speech transformation:

analogue conversion to digital

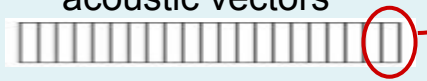


Split digitized audio into overlapping frames(~10ms)

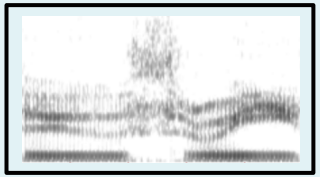




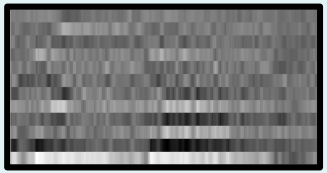
Pre-processor/Front end
eg. FFT, LPC, MFCC



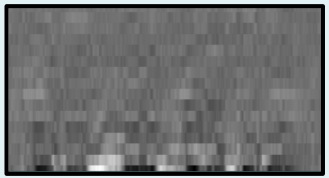
FFT



LPC



MFCC



0.12
-0.32
0.32
0.09
-0.17
0.18
-1.11
-1.18
0.15
0.06
0.39
-1.27

Speech transformation for ASR:

Split digitized audio into overlapping segments

frames

Pre-processor/Front end
eg. FFT, LPC, MFCC

acoustic vectors

Sampling rate

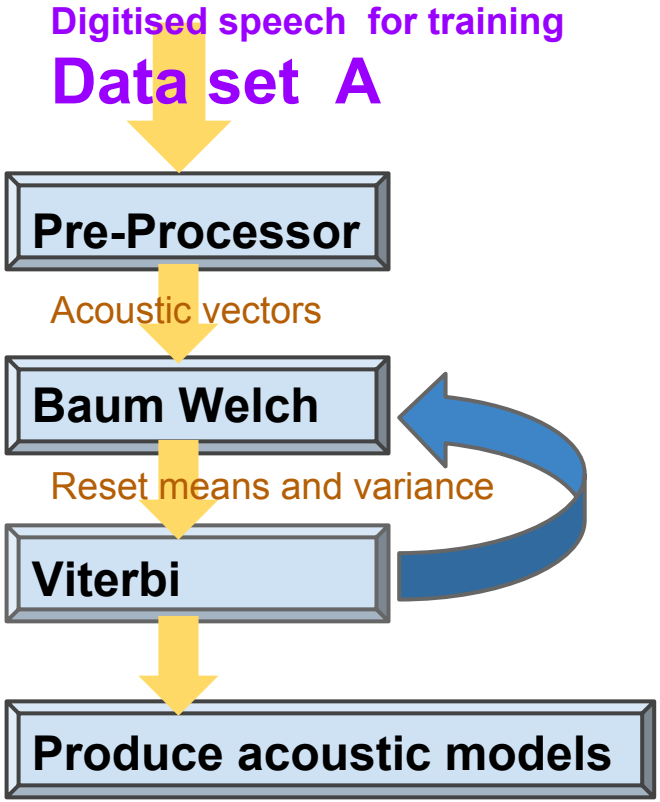
Window duration, shape and overlap

Pre-processor

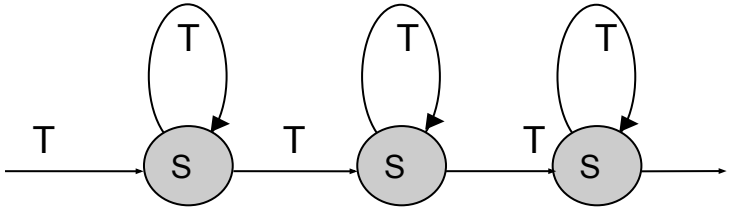
Forced Alignment can be based on:

- **(Mono)phones**
- Diphones
- Triphones
- Words
- Syllables
- Broad phonetic classes (eg. stop, sonorant, vowel)

Training the system Using Hidden Markov Models (HMM)



Calculate the acoustic models

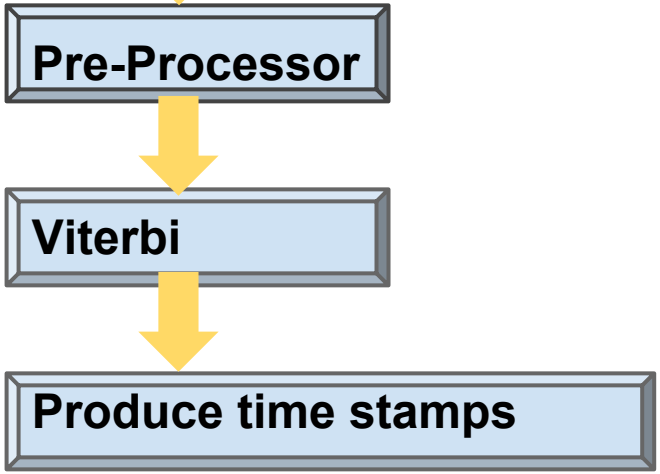


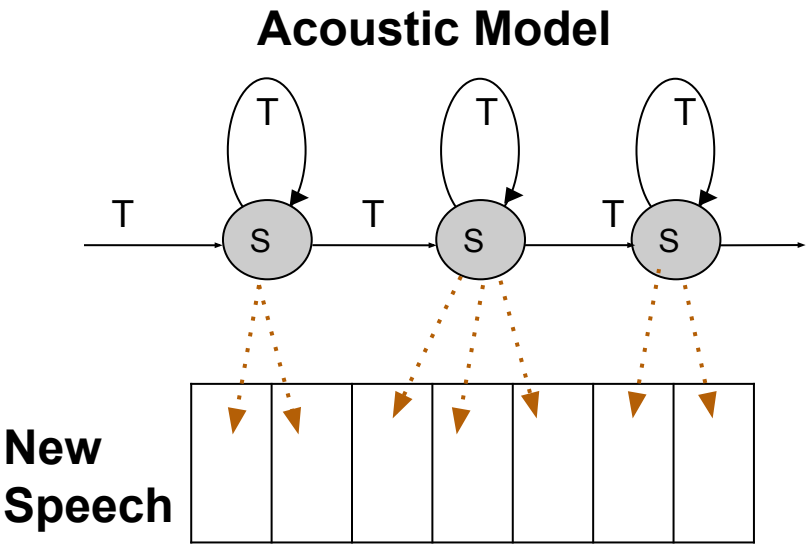
State S = Probability of being in a state
Transition T = Probability of moving to a state

Forced Alignment:

Digitised speech for alignment

Data set B





State S = Probability of being in a state
Transition T = Probability of moving to a state

Open Source Alignment/Recognition Systems:

Toolkit for development:

HTK htk.eng.cam.ac.uk

Kaldi kaldi.sourceforge.net

Ready systems:

P2FA www.ling.upenn.edu/phonetics/p2fa

Julius julius.sourceforge.jp/en_index.php

SPPAS aune.lpl-aix.fr/~bigi/sppas/

Sphinx sourceforge.net/projects/cmusphinx

HTK / P2FA prerequisites

INPUT TO HTK / P2FA

- Audio sampling rate (16kHz best)
- Lists of phone and silence model names
- Dictionary/Lexicon (ARPAbet)

ADVERSE	AH0 D V ER1 S
ADVERSELY	AE0 D V ER1 S L IH0
ADVERSELY	AE0 D V ER1 S L IY0
ADVERSITIES	AH0 D V ER1 S IH0 T IH0 Z
ADVERSITY	AE0 D V ER1 S AH0 T IY0
BLUEST	B L UW1 IH0 S T
BLUESTEIN	B L UH1 S T AY0 N
BLUESTEIN	B L UH1 S T IY0 N
BLUESTINE	B L UW1 S T AY2 N
ZZZ	sp
ZZZ	ns
…	

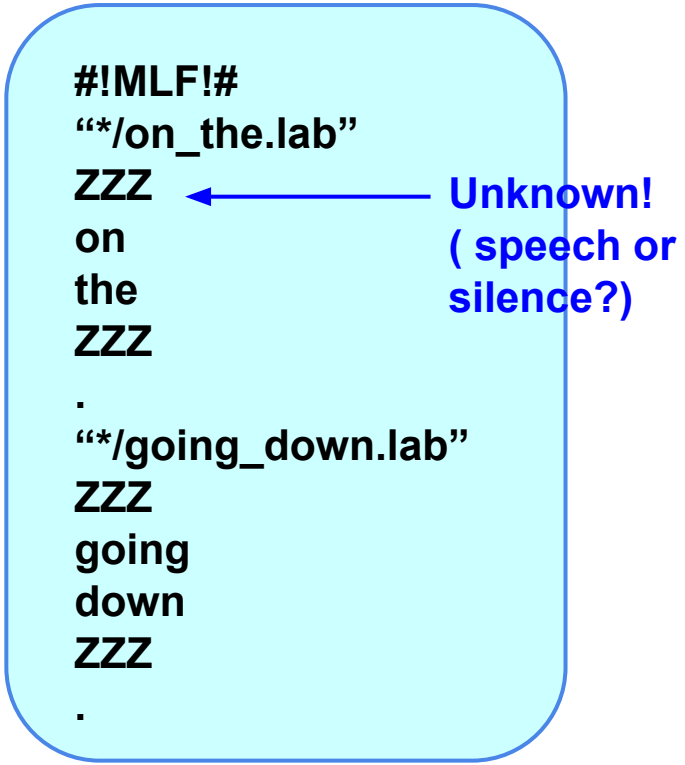
Enter all likely pronunciations (BNC_dict.txt)

- Orthographic transcription

HTK / P2FA prerequisites

Orthographic transcription:

- Only use subset of ASCII - best to keep to letters, numbers, underscore
- “+”, “-”, “\” etc. have special meanings in HTK, and P2FA
- Make one big file with transcriptions for all the wav file names: "Master Label File" (extension .MLF)



HTK result formats (also .mlf file)

#!MLF!#

"*/on_the.rec"

0 200000 sp 62.947620 ZZZ

200000 1100000 AA1 137.138046 ON

1100000 1500000 N 110.194252

1500000 1500000 sp -0.156736 sp

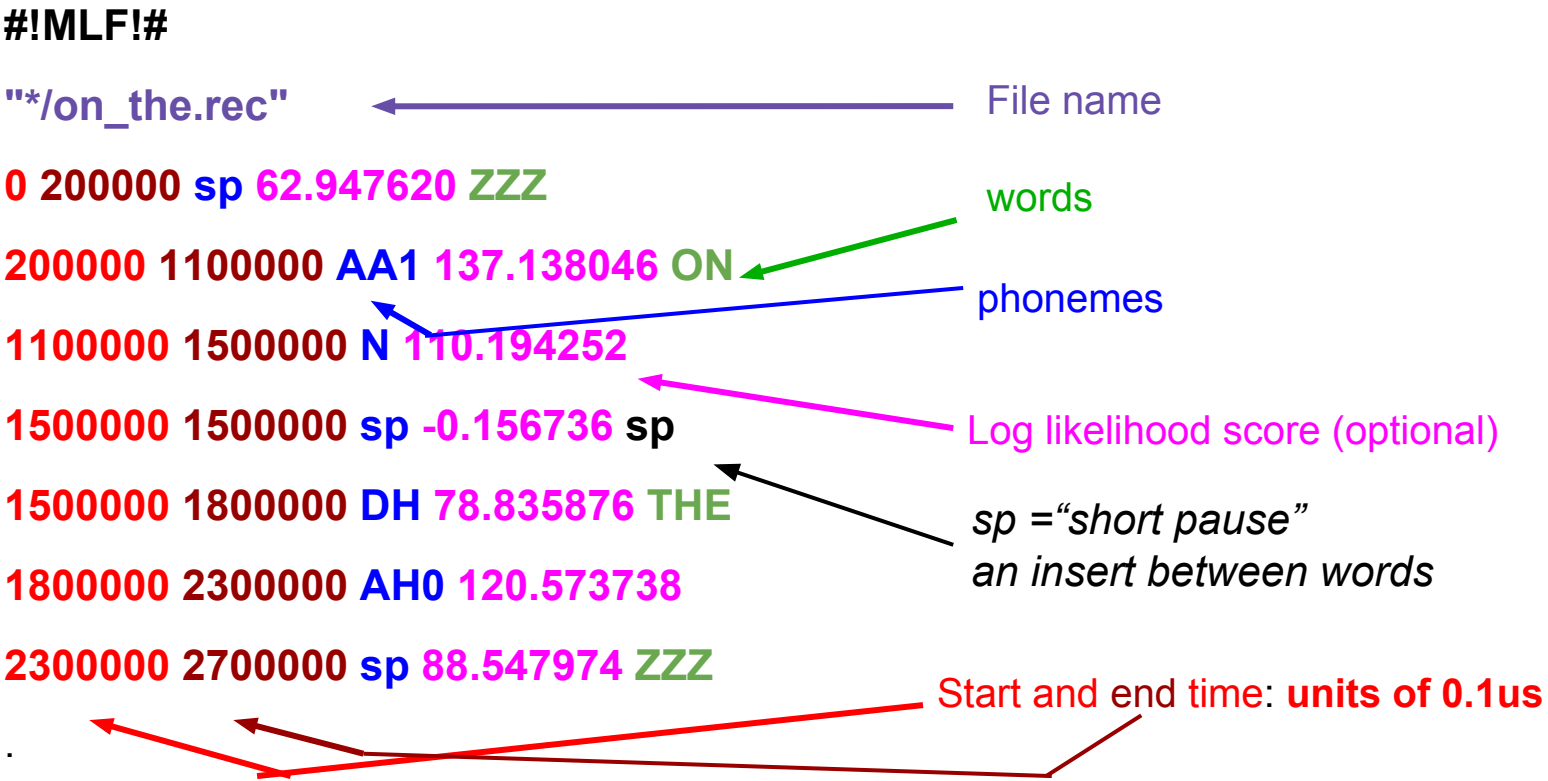
1500000 1800000 DH 78.835876 THE

1800000 2300000 AH0 120.573738

2300000 2700000 sp 88.547974 ZZZ

.

HTK result formats (*.mlf)



Converting .mlf to .TextGrid

```
#!MLF!#
"*/on_the.rec"
0 200000 sp 62.947620 ZZZ
200000 1100000 AA1 137.138046 ON
1100000 1500000 N 110.194252
1500000 1500000 sp -0.156736 sp
1500000 1800000 DH 78.835876 THE
1800000 2300000 AH0 120.573738
2300000 2700000 sp 88.547974 ZZZ
.
```

mlf2textgrid.py

exclude short pause with zero duration

```
File type = "ooTextFile short"
"TextGrid"

0.00
2.70
<exists>
2
"IntervalTier"
"phone"
0.00
2.70
2
0.00
0.20
"sp"
0.20
1.10
"AA1"
1.10
1.50
"N"
1.50
1.80
"DH"
1.80
2.30
"AH0"
```

```
File type = "ooTextFile short"
"TextGrid"

0.01
1.01
<exists>
2
"IntervalTier"
"phone"
0.01
1.01
2
0.01
0.055
"ih1"
0.055
1.01
"n"
"IntervalTier"
"words"
0.01
1.01
1
0.01
1.01
"IN"
```

PRAAT - SHORT FORMAT

PRAAT - SHORT FORMAT

```
File type = "ooTextFile short"
"TextGrid"

0.01
1.01
<exists>
2
"IntervalTier"
"phone"
0.01
1.01
2
0.01
0.055
"ih1"
0.055
1.01
"n"
"IntervalTier"
"words"
0.01
1.01
1
0.01
1.01
"IN"
```

Audio start and end time

Number of tiers

Tier name

Tier start and end time

Number of labels

Label start and end time

Label

PRAAT - SHORT FORMAT

```
File type = "ooTextFile short"
"TextGrid"

0.01
1.01
<exists>
2
"IntervalTier"
"phone"
0.01
1.01
2
0.01
0.055
"ih1"
0.055
1.01
"n"
"IntervalTier"
"words"
0.01
1.01
1
0.01
1.01
"IN"
```

Audio start and end time

Number of tiers

Tier name

Tier start and end time

Number of labels

Label start and end time

Label

PRAAT - SHORT FORMAT

```
File type = "ooTextFile short"
"TextGrid"

0.01
1.01
<exists>
2
"IntervalTier"
"phone"
0.01
1.01
2
0.01
0.055
"ih1"
0.055
1.01
"n"
"IntervalTier"
"words"
0.01
1.01
1
0.01
1.01
"IN"
```

Audio start and end time

Number of tiers

Tier name

Tier start and end time

Number of labels

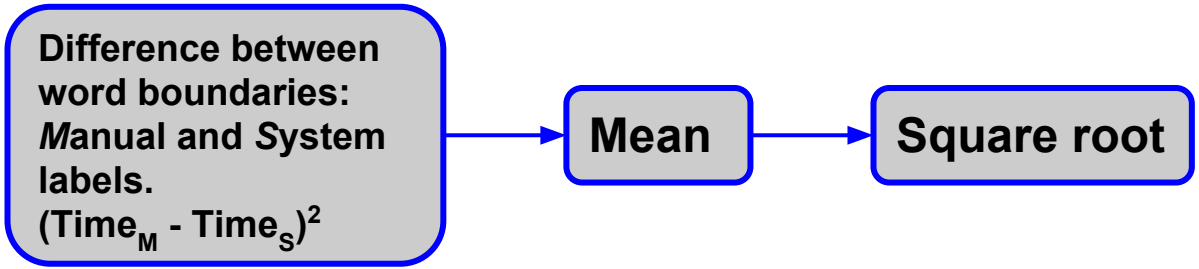
Label start and end time

Label

Evaluating alignment quality

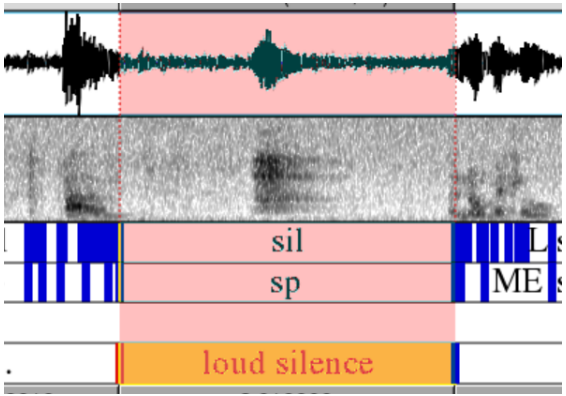
- Listening test (results can be as a percentage)
- Gold standard - manual hand labels

Root mean square (rms) difference



- Speech recognition labels - select correct words and compare boundaries

Evaluating alignment quality



- Gross Error
  - Speech in Silence
  - Silence in Speech
  - phoneme/word durations (too long or too short)