Automatic Speech Recognition

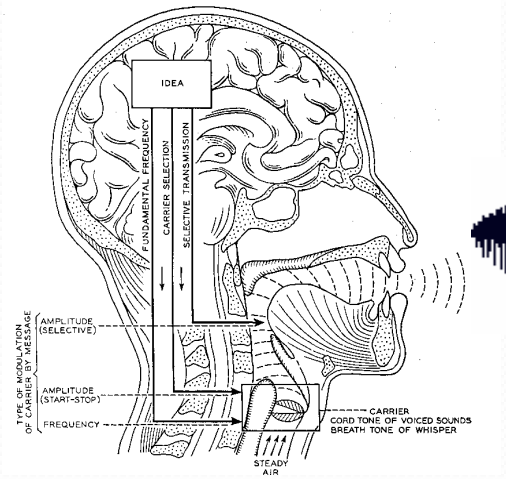
Automatic speech recognition

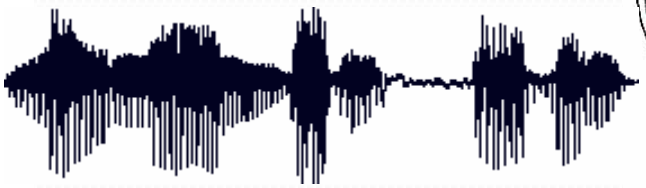
- What is the task?
- What are the main difficulties?
- How is it approached?
- How good is it?
- How much better could it be?

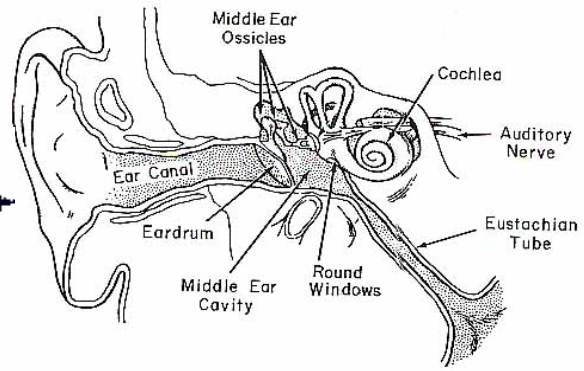
What is the task?

- Getting a computer to understand spoken language
- By “understand” we might mean
  - React appropriately
  - Convert the input speech into another medium, e.g. text
- Several variables impinge on this

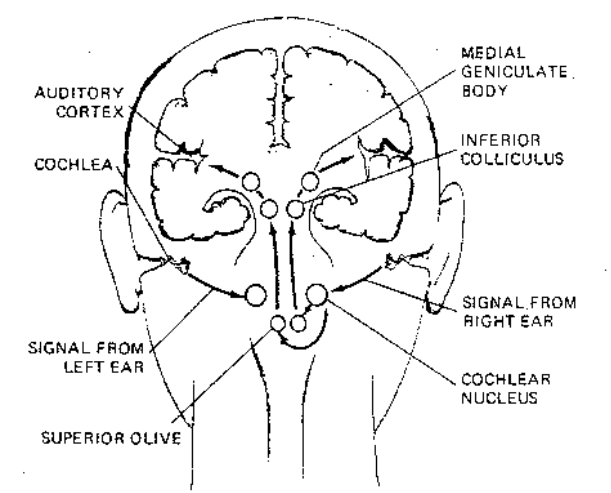
How do humans do it?



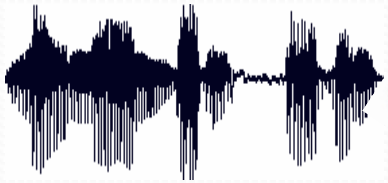




- Articulation produces sound waves which the ear conveys to the brain for processing

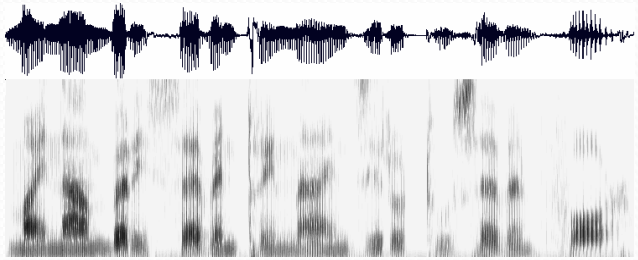


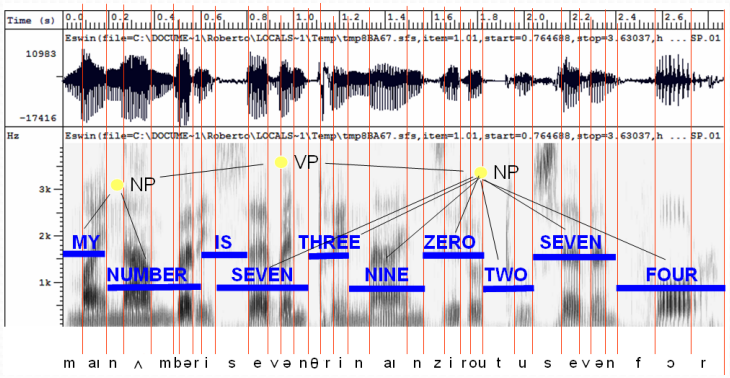
How might computers do it?



Acoustic waveform

Acoustic signal

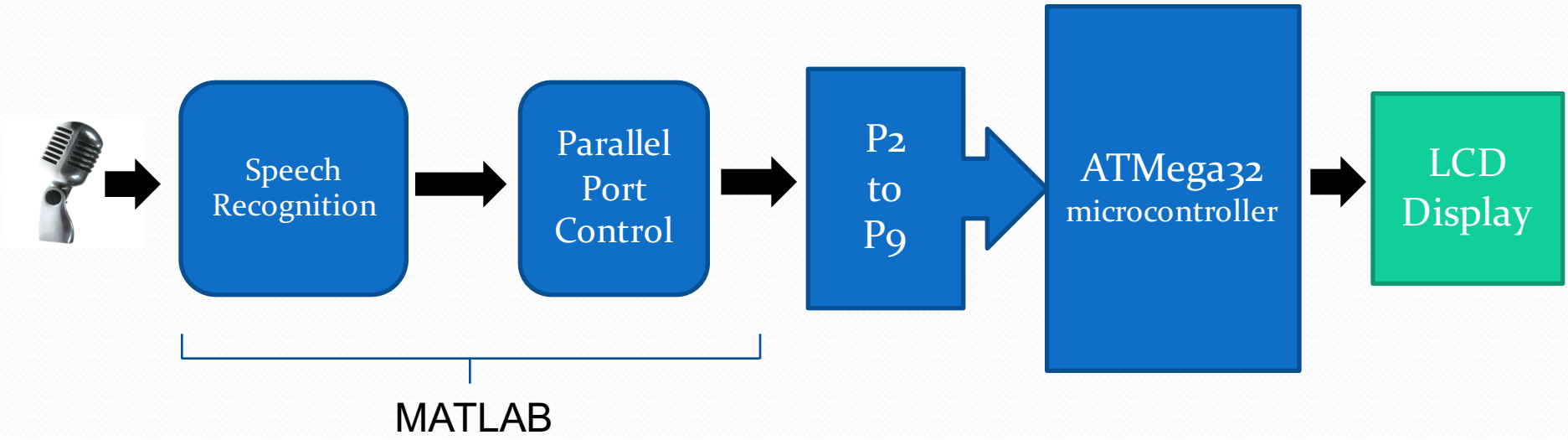




Speech recognition

- Digitization
- Acoustic analysis of the speech signal
- Linguistic interpretation

Basic Block Diagram



What's hard about that?

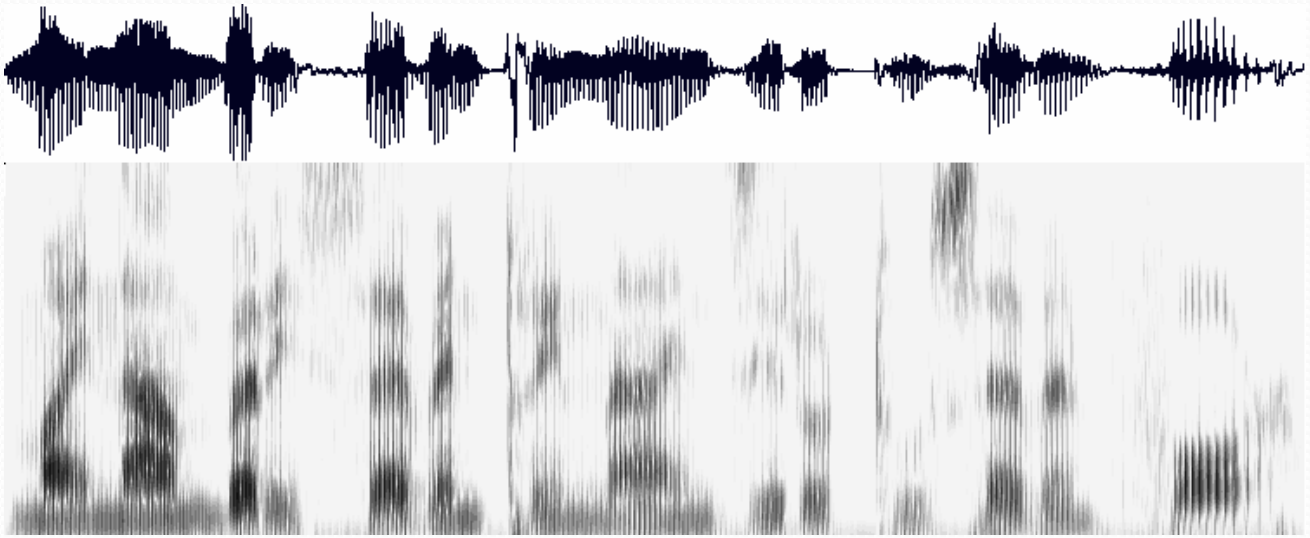
- Digitization
  - Converting analogue signal into digital representation
- Signal processing
  - Separating speech from background noise
- Phonetics
  - Variability in human speech
- Phonology
  - Recognizing individual sound distinctions (similar phonemes)
- Lexicology and syntax
  - Disambiguating homophones
  - Features of continuous speech
- Syntax and pragmatics
  - Interpreting prosodic features
- Pragmatics
  - Filtering of performance errors (disfluencies)

Digitization

- Analogue to digital conversion
- Sampling and quantizing
- Use filters to measure energy levels for various points on the frequency spectrum
- Knowing the relative importance of different frequency bands (for speech) makes this process more efficient
- E.g. high frequency sounds are less informative, so can be sampled using a broader bandwidth (log scale)

Separating speech from background noise

- Noise cancelling microphones
  - Two mics, one facing speaker, the other facing away
  - Ambient noise is roughly same for both mics
- Knowing which bits of the signal relate to speech
  - Spectrograph analysis



Variability in individuals' speech

- Variation among speakers due to
  - Vocal range (fo, and pitch range – see later)
  - Voice quality (growl, whisper, physiological elements such as nasality, adenoidality, etc)
  - ACCENT !!! (especially vowel systems, but also consonants, allophones, etc.)
- Variation within speakers due to
  - Health, emotional state
  - Ambient conditions
- Speech style: formal read vs spontaneous

Speaker-(in)dependent systems

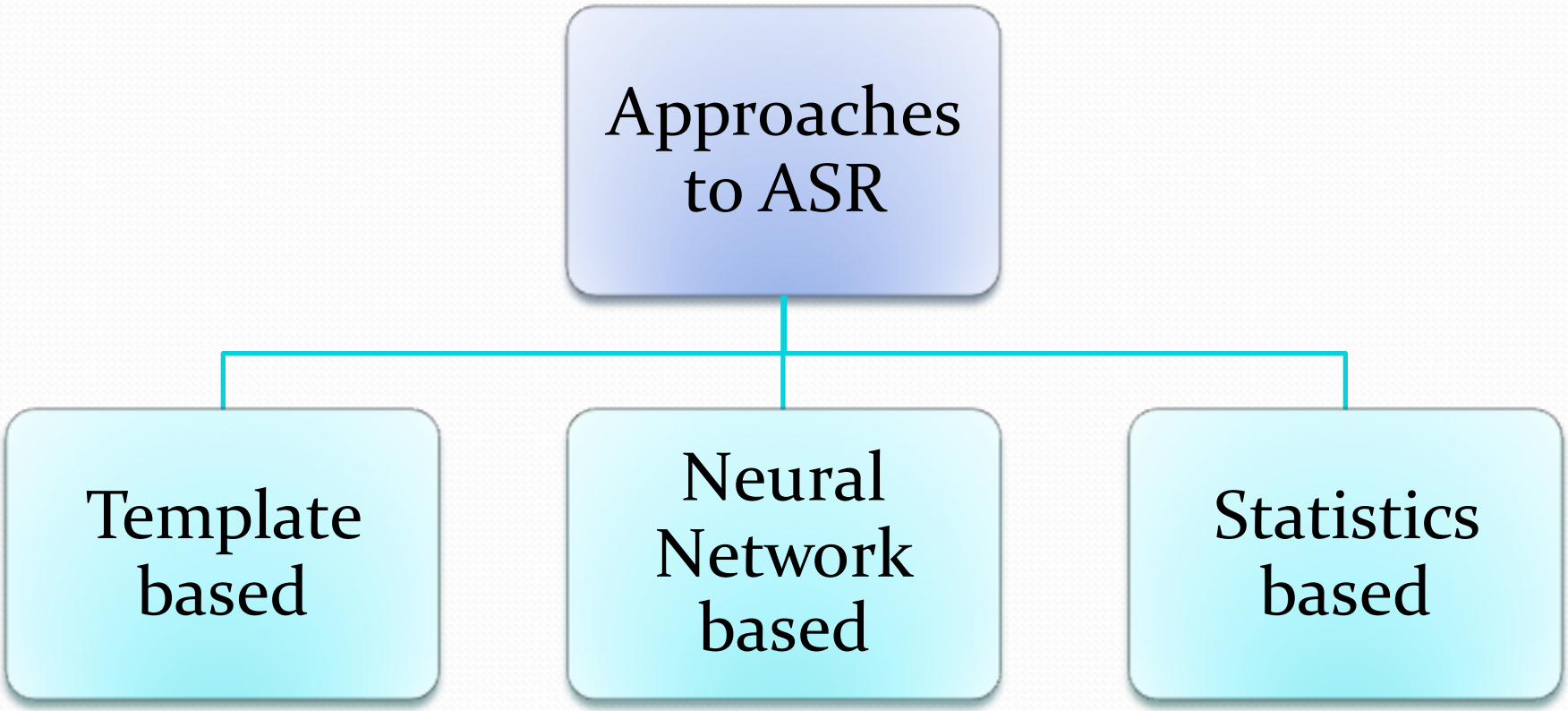
- Speaker-dependent systems
  - Require "training" to "teach" the system your individual idiosyncracies
    - The more the merrier, but typically nowadays 5 or 10 minutes is enough
    - User asked to pronounce some key words which allow computer to infer details of the user's accent and voice
    - Fortunately, languages are generally systematic
  - More robust
  - But less convenient
  - And obviously less portable
- Speaker-independent systems
  - Language coverage is reduced to compensate need to be flexible in phoneme identification
  - Clever compromise is to learn on the fly

(Dis)continuous speech

- Discontinuous speech much easier to recognize
  - Single words tend to be pronounced more clearly
- Continuous speech involves contextual coarticulation effects
  - Weak forms
  - Assimilation
  - Contractions

Performance errors

- Performance “errors” include
  - Non-speech sounds
  - Hesitations
  - False starts, repetitions
- Filtering implies handling at syntactic level or above
- Some disfluencies are deliberate and have pragmatic effect – this is not something we can handle in the near future


Approaches to ASR
Template based
Neural Network based
Statistics based

Template-based approach

- Store examples of units (words, phonemes), then find the example that most closely fits the input
- Extract features from speech signal, then it's "just" a complex similarity matching problem, using solutions developed for all sorts of applications
- OK for discrete utterances, and a single user

Template-based approach

- Hard to distinguish very similar templates
- And quickly degrades when input differs from templates
- Therefore needs techniques to mitigate this degradation:
  - More subtle matching techniques
  - Multiple templates which are aggregated
- Taken together, these suggested …

Neural Network based approach

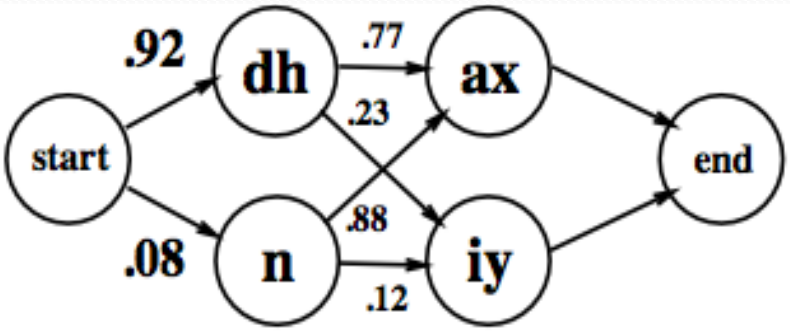
Statistics-based approach

- Collect a large corpus of transcribed speech recordings
- Train the computer to learn the correspondences ("machine learning")
- At run time, apply statistical processes to search through the space of all possible solutions, and pick the statistically most likely one

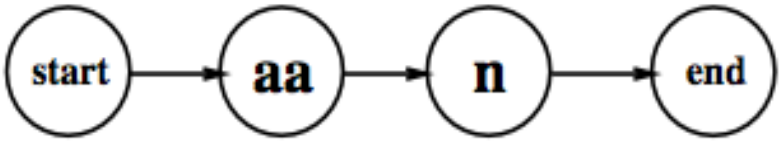
Statistics based approach

- Acoustic and Lexical Models
  - Analyse training data in terms of relevant features
  - Learn from large amount of data different possibilities
    - different phone sequences for a given word
    - different combinations of elements of the speech signal for a given phone/phoneme
  - Combine these into a Hidden Markov Model expressing the probabilities

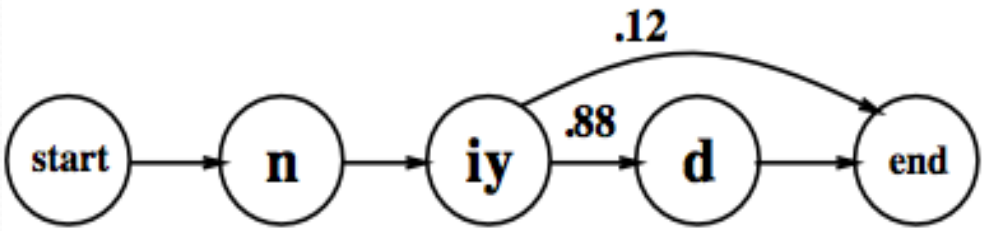
HMMs for some words



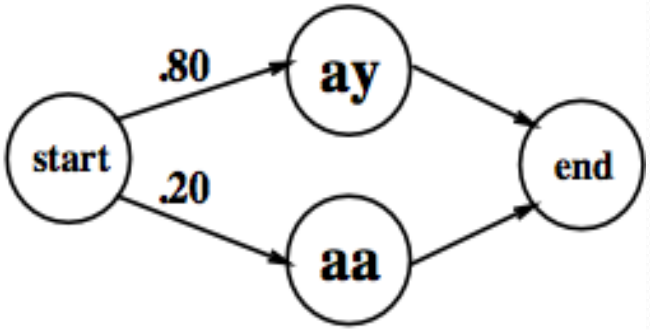
Word model for "the"



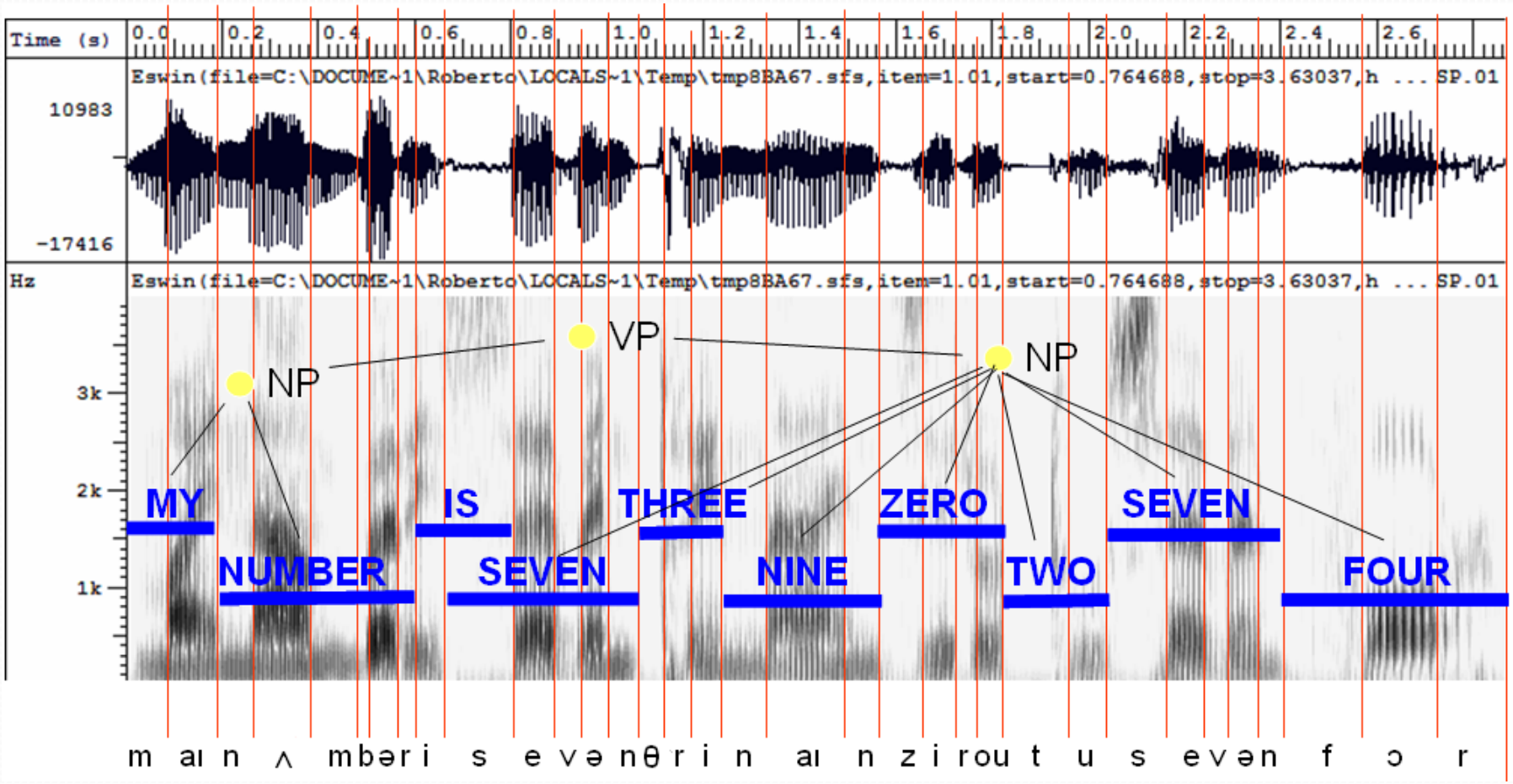
Word model for "on"



Word model for "need"

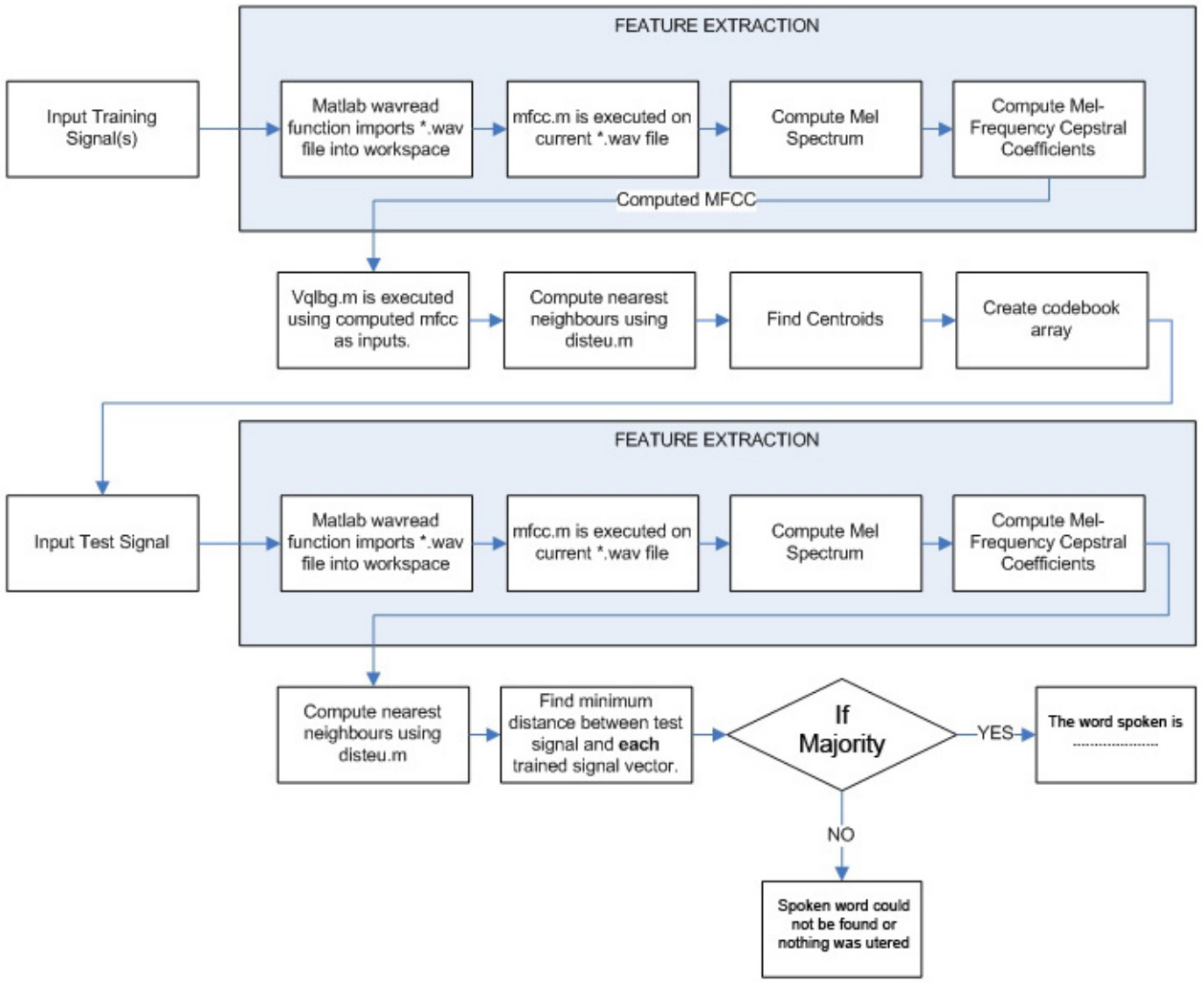


Word model for "I"



- Identify individual phonemes
- Identify words
- Identify sentence structure and/or meaning

SPEECH RECOGNITION BLOCK DIAGRAM



BLOCK DIAGRAM DESCRIPTION

Speech Acquisition Unit

- It consists of a microphone to obtain the analog speech signal
- The acquisition unit also consists of an analog to digital converter

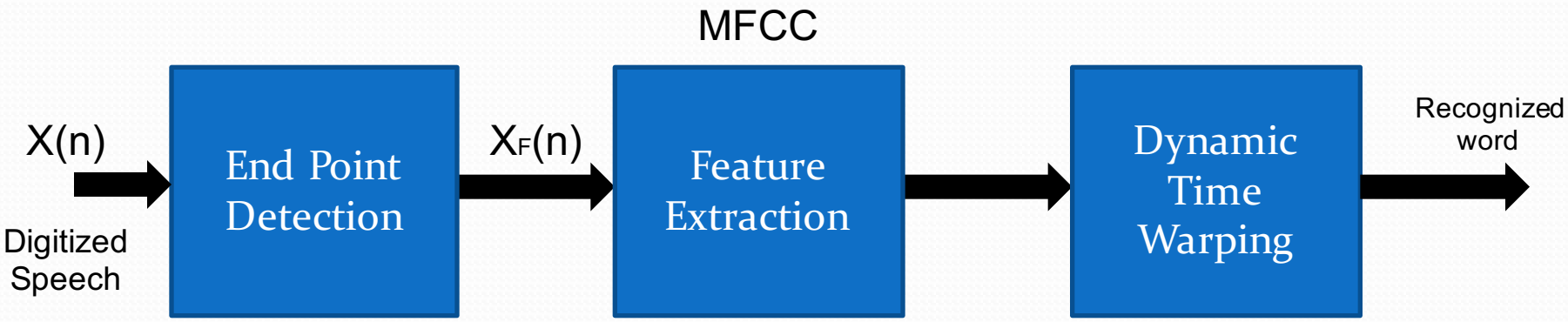
Speech Recognition Unit

- This unit is used to recognize the words contained in the input speech signal.
- The speech recognition is implemented in MATLAB with the help of
- template matching algorithm

Device Control Unit

- This unit consists of a microcontroller, the ATmega32, to control the various appliances
- The microcontroller is connected to the PC via the PC parallel port
- The microcontroller then reads the input word and controls the device connected to it accordingly.

SPEECH RECOGNITION



END-POINT DETECTION

• The accurate detection of a word's start and end points means that subsequent processing of the data can be kept to a minimum by processing only the parts of the input corresponding to speech.

•We will use the endpoint detection algorithm proposed by Rabiner and Sambur. This algorithm is based on two simple time-domain measurements of the signal - the energy and the zero crossing rate.

The algorithm should tackle the following cases:-

1. Words which begin with or end with a low energy phoneme
2. Words which end with a nasal
3. Speakers ending words with a trailing off in intensity or short breath

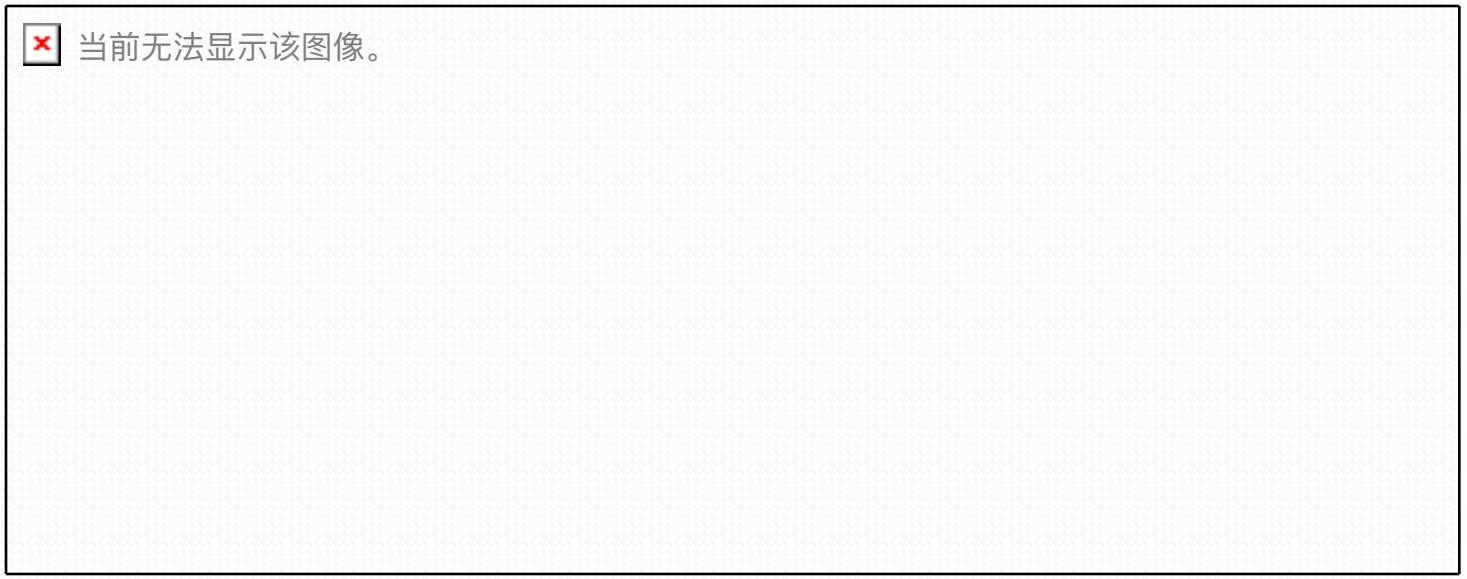
Steps for EPD

•Removal of noise by subtracting the signal values with that of noise

• Word extraction

steps –

1. ITU [Upper energy threshold]
2. ITL [Lower energy threshold]
3. IZCT [Zero crossing rate threshold]



Feature Extraction

- Input data to the algorithm is usually too large to be processed
- Input data is highly redundant
- Raw analysis requires high computational powers and large amounts of memory
- Thus, removing the redundancies and transforming the data into a set of features
- DCT based Mel Cepstrum

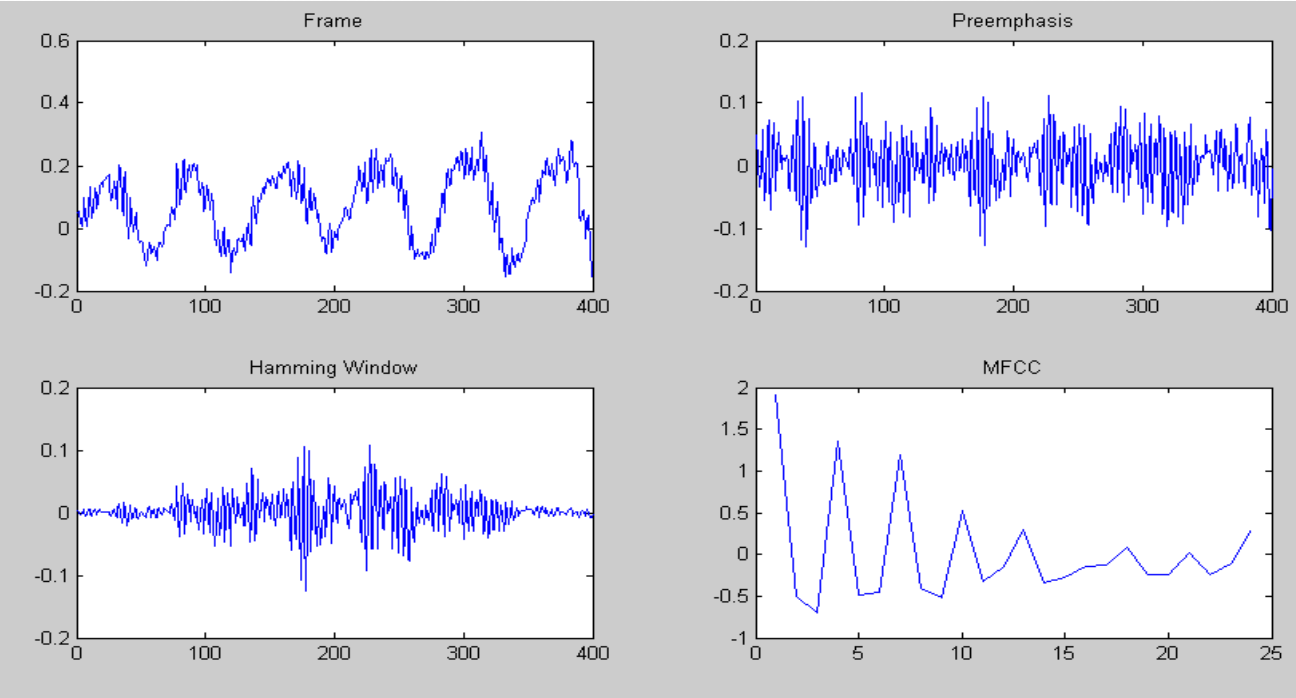
DCT Based MFCC

- Take the Fourier transform of a signal.
- Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
- Take the logs of the powers at each of the mel frequencies.
- Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
- The MFCCs are the amplitudes of the resulting spectrum.

MFCC Computation

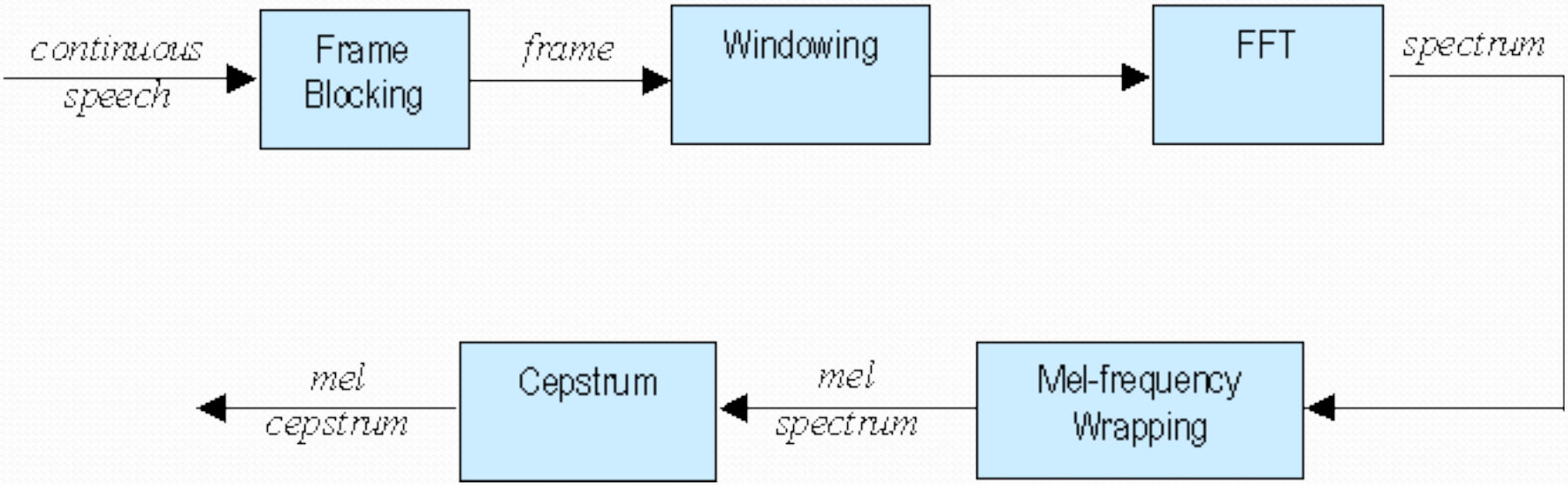
$$y_t^{(m)}(k) = \sum_{m=1}^{M} \log\{|y_t(m)|^2\} \cos\left(k\left(m - \frac{1}{2}\right)\frac{\pi}{M}\right)$$

- As Log Magnitude is real and symmetric IDFT reduces to DCT. The DCT produces highly un-correlated feature $\mathbf{y_t^{(m)}(k)}$. The Zero Order MFCC coefficient $\mathbf{y_t^{(o)}(k)}$ is approximately equal to the Log Energy of the frame.



The number of MFCC co-effecients chosen were 13

Feature extraction by MFCC Processing



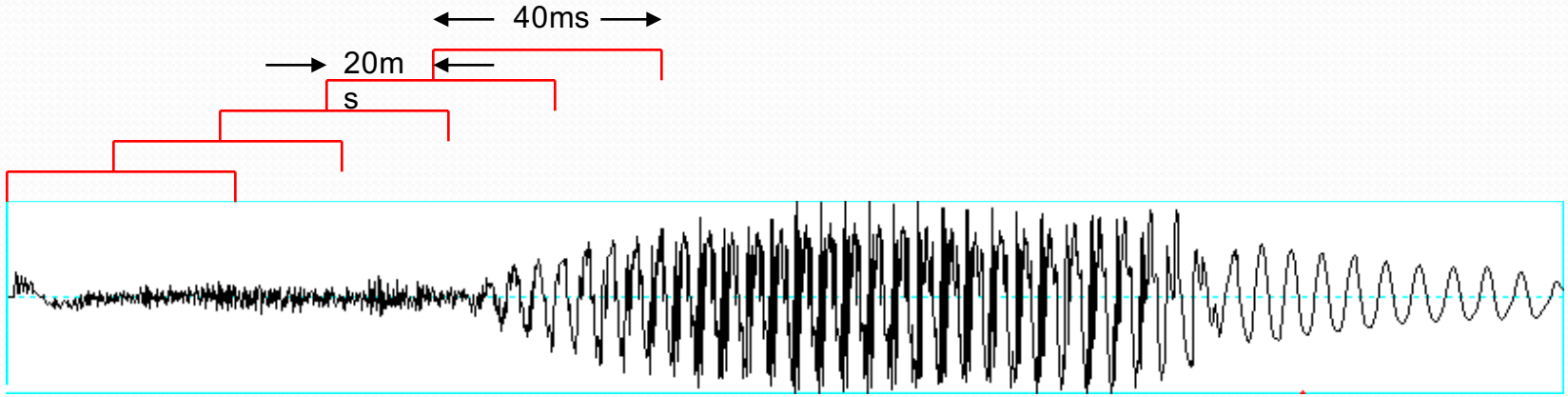
Dynamic Time Warping and Minimum Distance Paths measurement

- Isolated word recognition:
- Task :
  - Want to build an isolated word recogniser
- Method:
  1. Record, parameterise and store vocabulary of reference words.
  2. Record test word to be recognised and parameterize.
  3. Measure distance between test word and each reference word.
  4. Choose reference word ‘closest’ to test word.

Words are parameterised on a frame-by-frame basis

Choose frame length, over which speech remains reasonably stationary

Overlap frames e.g. 40ms frames, 10ms frame shift

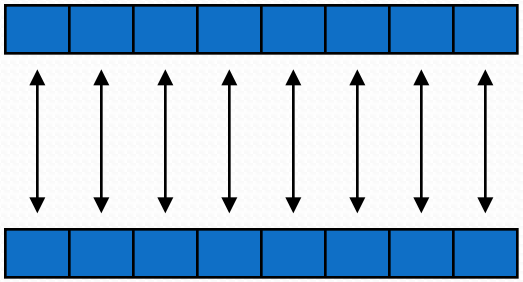


We want to compare frames of test and reference words i.e. calculate distances between them

Calculating Distances

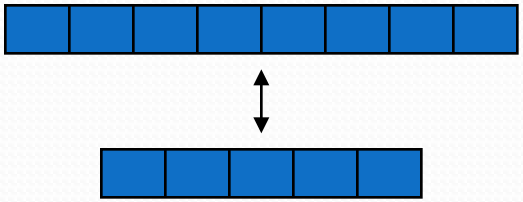
- Easy:

  Sum differences between corresponding frames



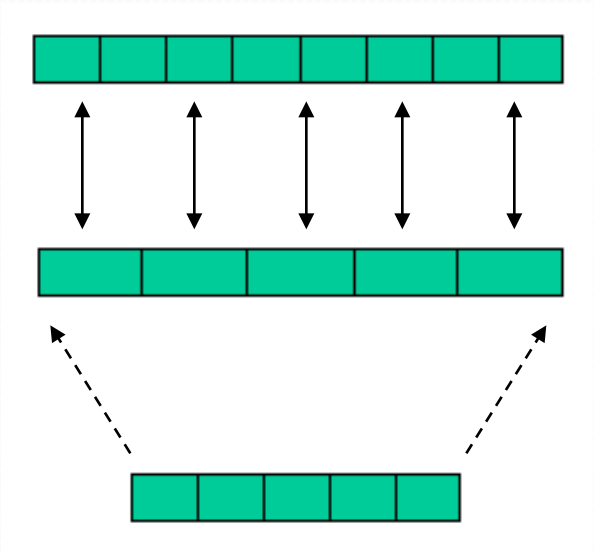
- Hard:

  Number of frames won't always correspond



- Solution 1: Linear Time Warping

  Stretch shorter sound

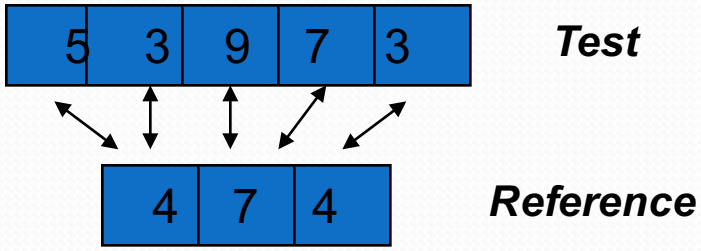


- Problem?

  Some sounds stretch more than others

- Solution 2:

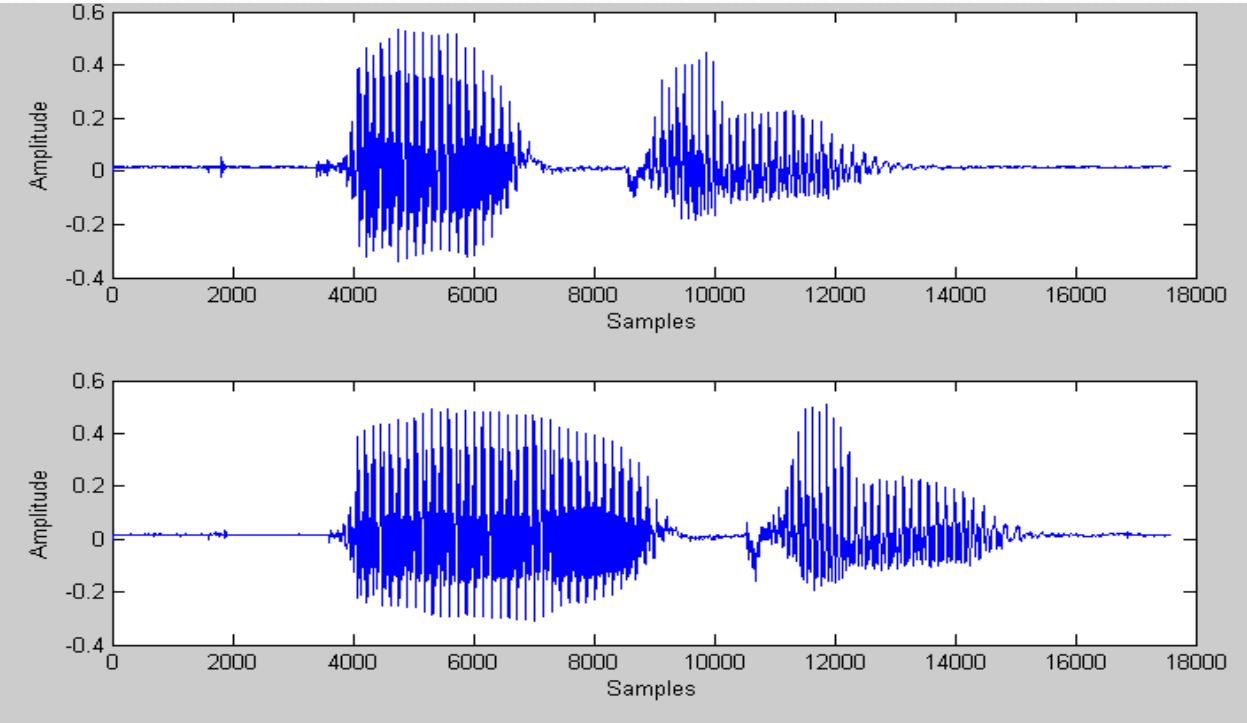
Dynamic Time Warping (DTW)



Using a dynamic alignment, make most similar frames correspond

Find distances between two utterances using these corresponding frames

Dynamic Programming



Waveforms showing the utterance of the word "Open" at two different instants. The signals are not time aligned.

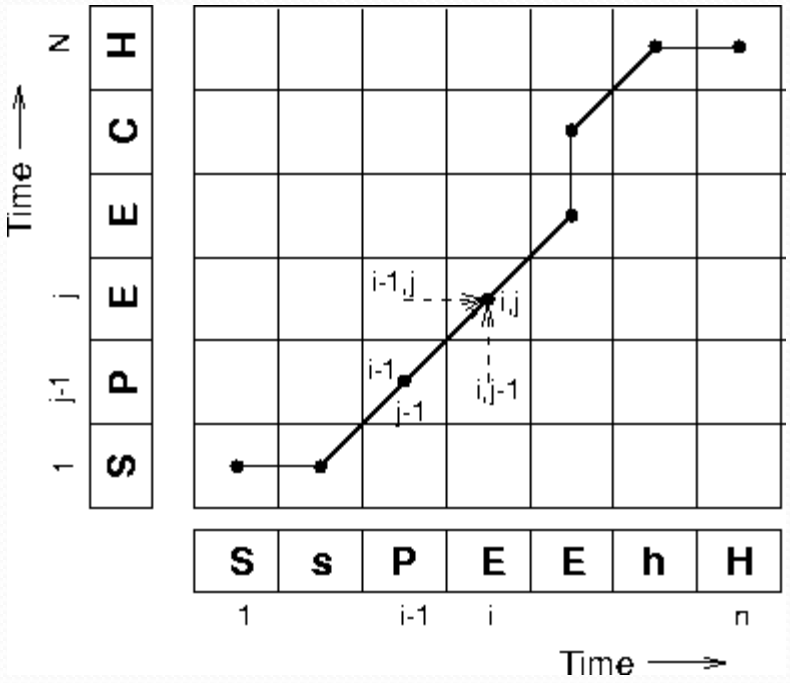
Place distance between frame *r* of *Test* and frame *c* of *Reference* in *cell*(*r*,*c*) of *distance* matrix

Test		1	2	3
3	5	1 x	4 x	1 x
7	4	3 x	0 x	3 x
9	3	5 x	2 x	5 x
3	2	1 x	4 x	1 x
5	1	1 x	2 x	1 x

4	7	4

Reference

DTW Process



Constraints

- Global
  - Endpoint detection
  - Path should be close to diagonal
- Local
  - Must always travel upwards or eastwards
  - No jumps
  - Slope weighting
  - Consecutive moves upwards/eastwards

Empirical Results : Known Speaker

	SONY	SUVARNA	GEMINI	HBO	CNN	NDTV	IMAGINE	ZEE CINEMA
SONY	9	0	1	0	0	0	0	0
SUVARNA	0	10	0	0	0	0	0	0
GEMINI	0	0	8	0	0	0	2	0
HBO	0	0	0	10	0	0	0	0
CNN	0	0	0	0	8	0	2	0
NDTV	0	0	0	0	0	10	0	0
IMAGINE	0	0	0	0	0	0	10	0
ZEE CINEMA	0	0	0	0	0	0	1	9

Empirical Results : Unknown Speaker

	SONY	SUVARNA	GEMINI	HBO	CNN	NDTV	IMAGINE	ZEE CINEMA
SONY	8	0	1	0	0	0	1	0
SUVARNA	0	8	0	0	0	0	0	2
GEMINI	1	0	8	0	0	0	1	0
HBO	0	0	0	10	0	0	0	0
CNN	1	0	0	0	8	0	2	0
NDTV	0	0	0	0	0	10	0	0
IMAGINE	0	0	0	0	0	0	10	0
ZEE CINEMA	0	2	0	0	0	0	0	8

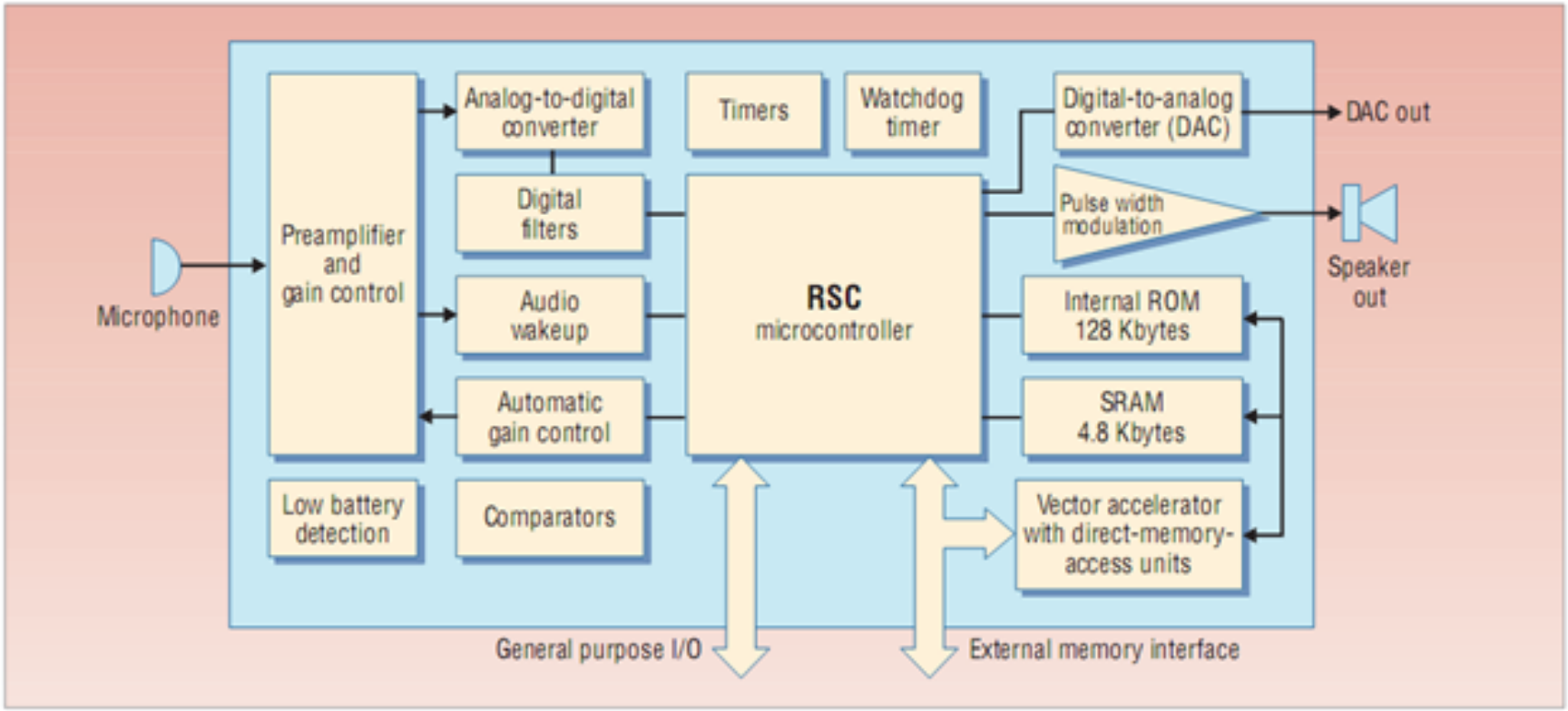
Applications

- Medical Transcription
- Military
- Telephony and other domains
- Serving the disabled

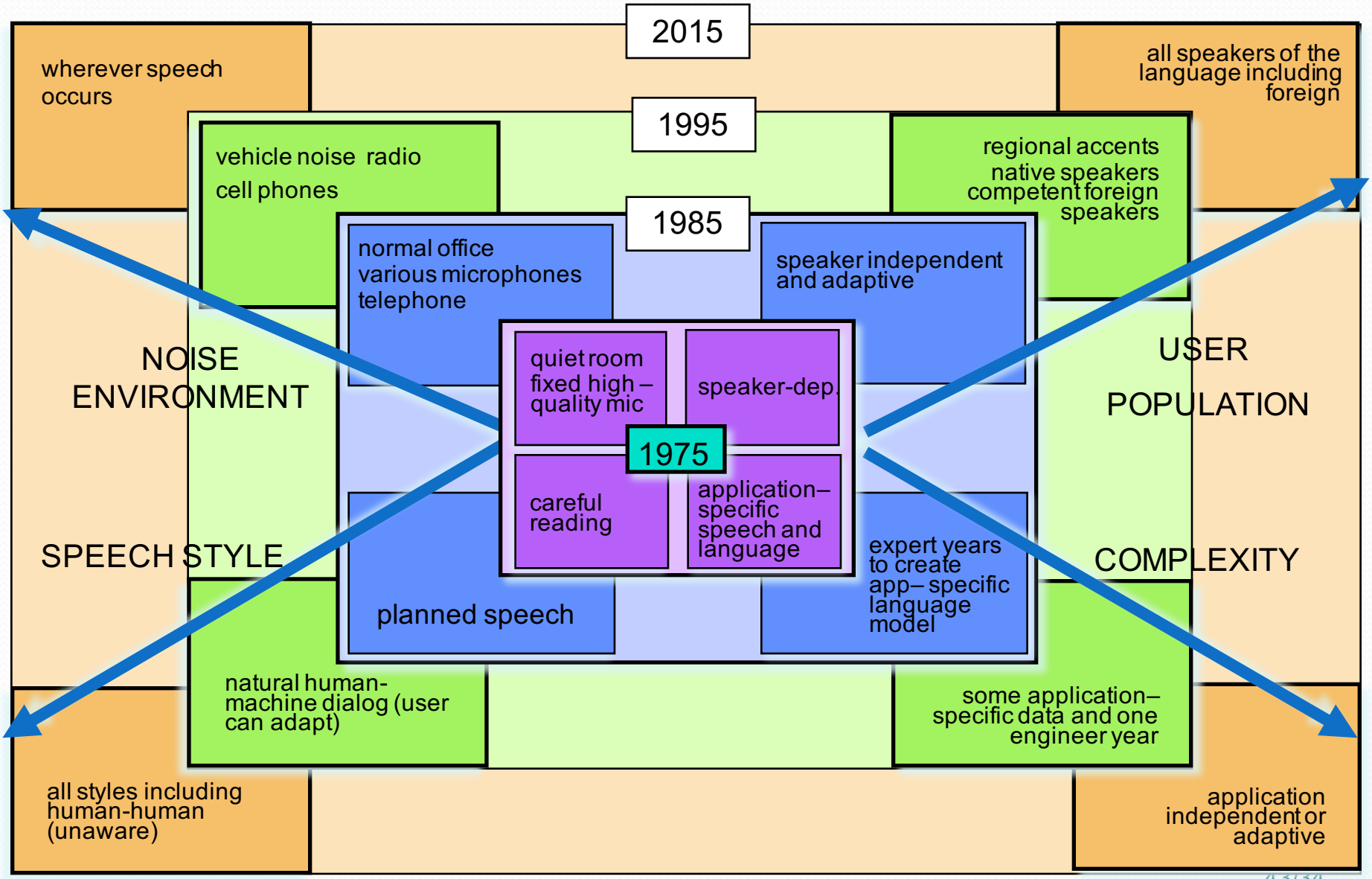
Further Applications

- Home automation
- Automobile audio systems
- Telematics

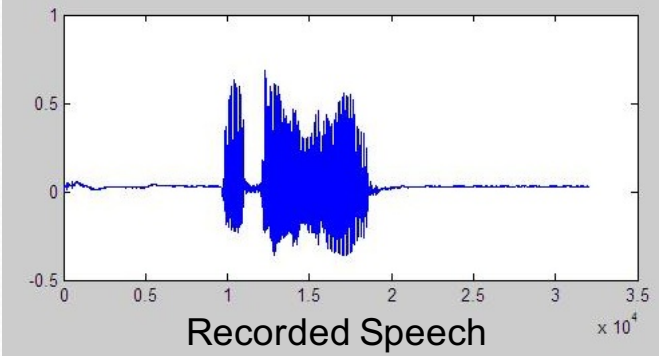
Where, From here?



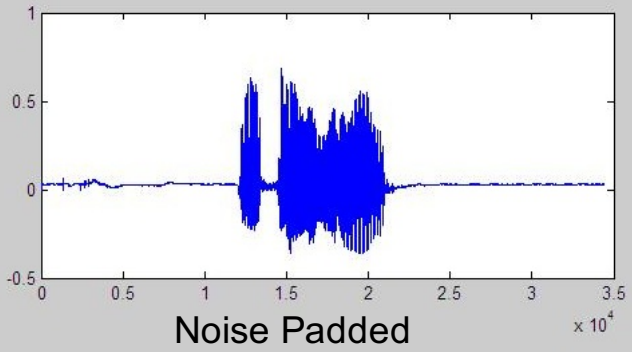
Evolution of ASR



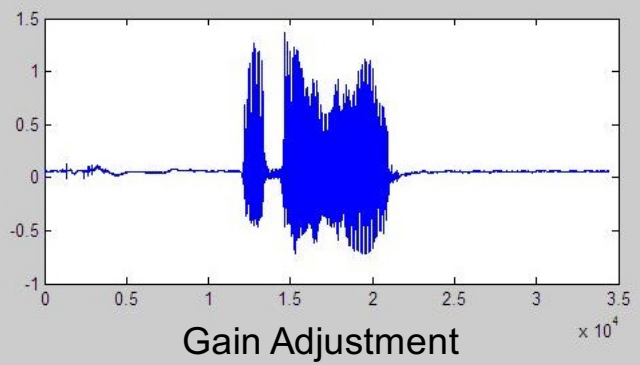
Conclusive remarks



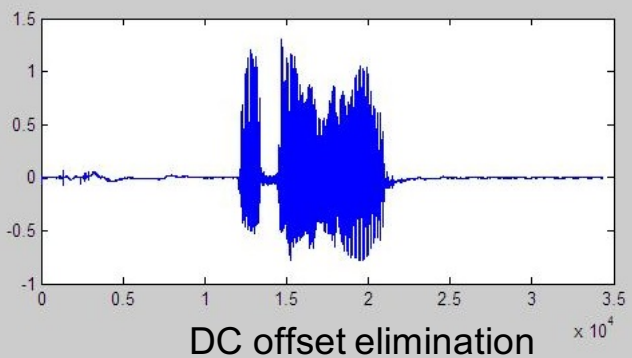
Recorded Speech



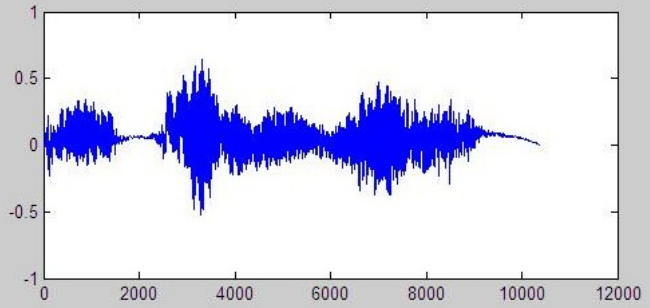
Noise Padded



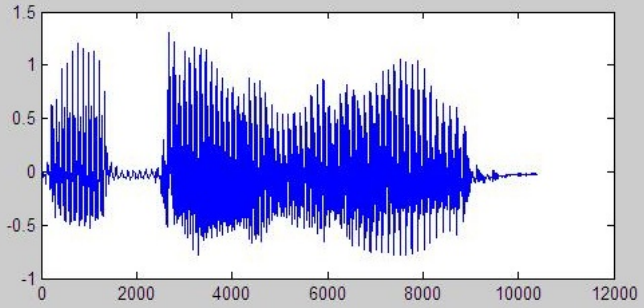
Gain Adjustment



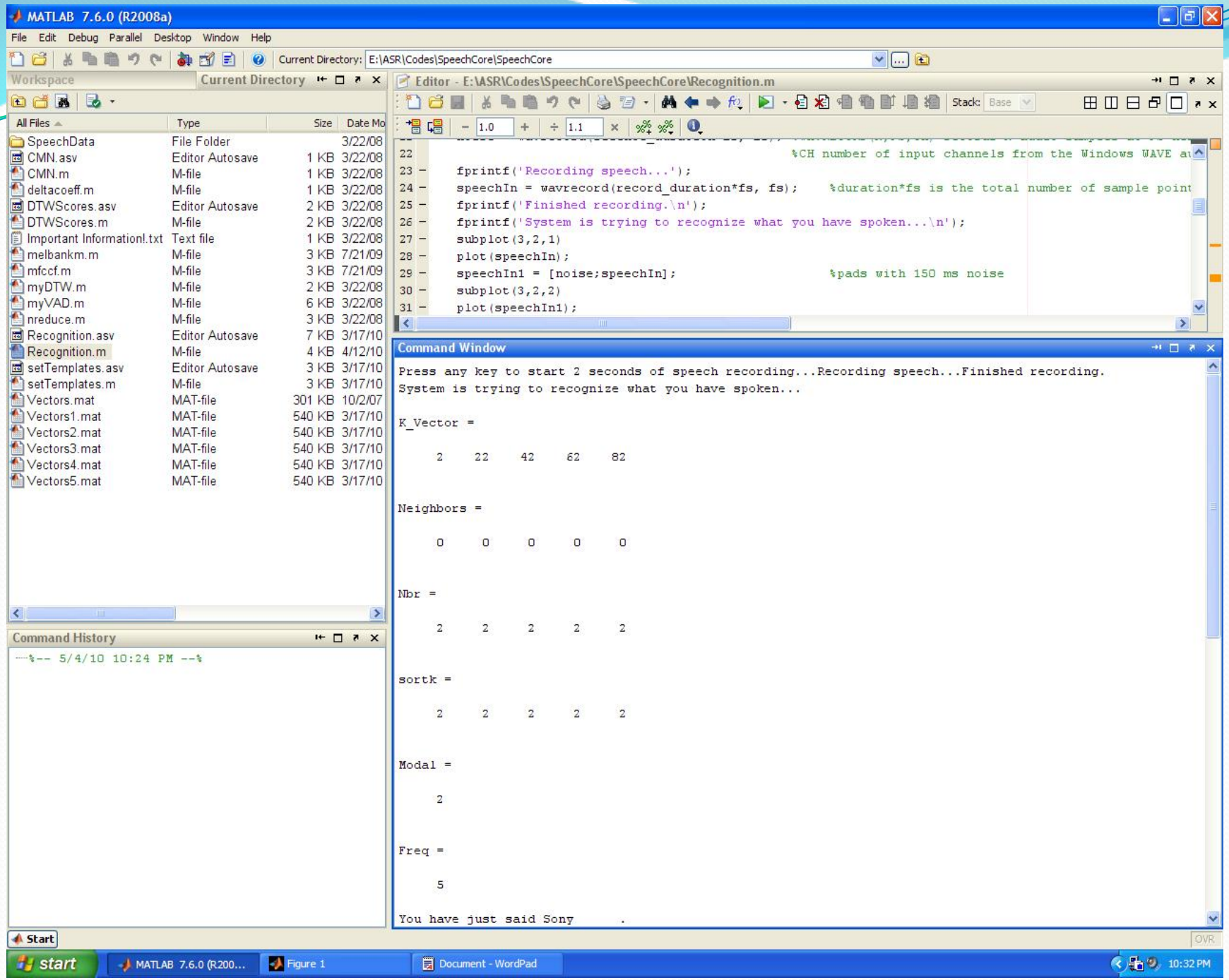
DC offset elimination

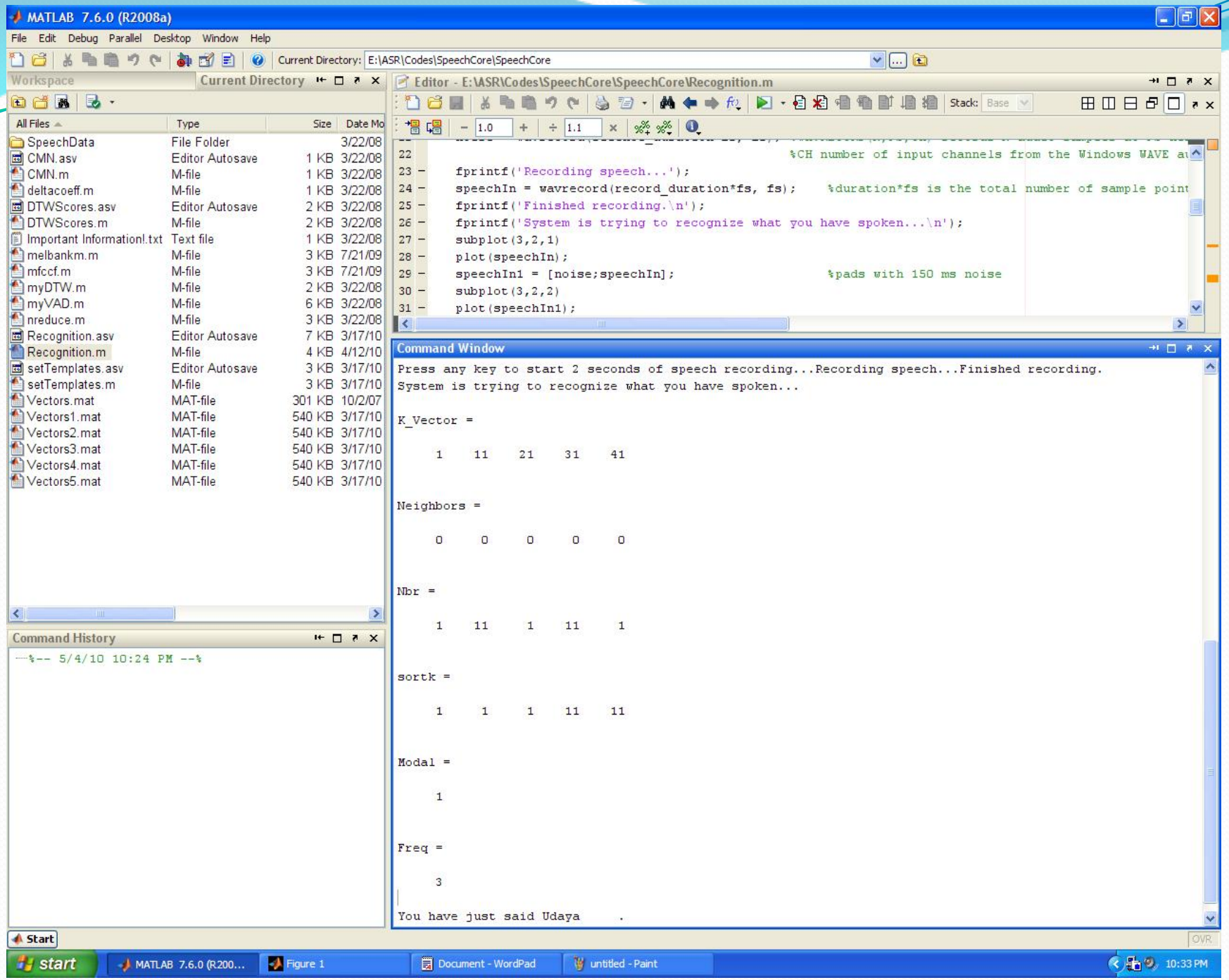


Spectral Subtraction



End Point Detection


MATLAB 7.6.0 (R2008a)
File Edit Debug Parallel Desktop Window Help
Current Directory: E:\ASR\Codes\SpeechCore\SpeechCore
Workspace
Current Directory
All Files Type Size Date Mo
SpeechData File Folder 3/22/08
CMN.asv Editor Autosave 1 KB 3/22/08
CMN.m M-file 1 KB 3/22/08
deltacoeff.m M-file 1 KB 3/22/08
DTWScores.asv Editor Autosave 2 KB 3/22/08
DTWScores.m M-file 2 KB 3/22/08
Important Information!.txt Text file 1 KB 3/22/08
melbankm.m M-file 3 KB 7/21/09
mfccf.m M-file 3 KB 7/21/09
myDTW.m M-file 2 KB 3/22/08
myVAD.m M-file 6 KB 3/22/08
nreduce.m M-file 3 KB 3/22/08
Recognition.asv Editor Autosave 7 KB 3/17/10
Recognition.m M-file 4 KB 4/12/10
setTemplates.asv Editor Autosave 3 KB 3/17/10
setTemplates.m M-file 3 KB 3/17/10
Vectors.mat MAT-file 301 KB 10/2/07
Vectors1.mat MAT-file 540 KB 3/17/10
Vectors2.mat MAT-file 540 KB 3/17/10
Vectors3.mat MAT-file 540 KB 3/17/10
Vectors4.mat MAT-file 540 KB 3/17/10
Vectors5.mat MAT-file 540 KB 3/17/10
Command History
%-- 5/4/10 10:24 PM --%
Editor - E:\ASR\Codes\SpeechCore\SpeechCore\Recognition.m
Stack: Base
%CH number of input channels from the Windows WAVE au
fprintf('Recording speech...');
speechIn = wavrecord(record_duration*fs, fs); %duration*fs is the total number of sample point
fprintf('Finished recording.\n');
fprintf('System is trying to recognize what you have spoken...\n');
subplot(3,2,1)
plot(speechIn);
speechIn1 = [noise;speechIn]; %pads with 150 ms noise
subplot(3,2,2)
plot(speechIn1);
Command Window
Press any key to start 2 seconds of speech recording...Recording speech...Finished recording.
System is trying to recognize what you have spoken...
K_Vector =
2 22 42 62 82
Neighbors =
0 0 0 0 0
Nbr =
2 2 2 2 2
sortk =
2 2 2 2 2
Modal =
2
Freq =
5
You have just said Sony .
Start
start
MATLAB 7.6.0 (R200...
Figure 1
Document - WordPad
10:32 PM


MATLAB 7.6.0 (R2008a)
File Edit Debug Parallel Desktop Window Help
Current Directory: E:\ASR\Codes\SpeechCore\SpeechCore
Workspace
Current Directory
All Files Type Size Date Mo
SpeechData File Folder 3/22/08
CMN.asv Editor Autosave 1 KB 3/22/08
CMN.m M-file 1 KB 3/22/08
deltacoeff.m M-file 1 KB 3/22/08
DTWScores.asv Editor Autosave 2 KB 3/22/08
DTWScores.m M-file 2 KB 3/22/08
Important Information!.txt Text file 1 KB 3/22/08
melbankm.m M-file 3 KB 7/21/09
mfccf.m M-file 3 KB 7/21/09
myDTW.m M-file 2 KB 3/22/08
myVAD.m M-file 6 KB 3/22/08
nreduce.m M-file 3 KB 3/22/08
Recognition.asv Editor Autosave 7 KB 3/17/10
Recognition.m M-file 4 KB 4/12/10
setTemplates.asv Editor Autosave 3 KB 3/17/10
setTemplates.m M-file 3 KB 3/17/10
Vectors.mat MAT-file 301 KB 10/2/07
Vectors1.mat MAT-file 540 KB 3/17/10
Vectors2.mat MAT-file 540 KB 3/17/10
Vectors3.mat MAT-file 540 KB 3/17/10
Vectors4.mat MAT-file 540 KB 3/17/10
Vectors5.mat MAT-file 540 KB 3/17/10
Command History
%-- 5/4/10 10:24 PM --%
Editor - E:\ASR\Codes\SpeechCore\SpeechCore\Recognition.m
Stack: Base
22 %CH number of input channels from the Windows WAVE a
23 - fprintf('Recording speech...');
24 - speechIn = wavrecord(record_duration*fs, fs); %duration*fs is the total number of sample point
25 - fprintf('Finished recording.\n');
26 - fprintf('System is trying to recognize what you have spoken...\n');
27 - subplot(3,2,1)
28 - plot(speechIn);
29 - speechIn1 = [noise;speechIn]; %pads with 150 ms noise
30 - subplot(3,2,2)
31 - plot(speechIn1);
Command Window
Press any key to start 2 seconds of speech recording...Recording speech...Finished recording.
System is trying to recognize what you have spoken...
K_Vector =
1 11 21 31 41
Neighbors =
0 0 0 0 0
Nbr =
1 11 1 11 1
sortk =
1 1 1 11 11
Modal =
1
Freq =
3
You have just said Udaya .
Start
start
MATLAB 7.6.0 (R200...
Figure 1
Document - WordPad
untitled - Paint
10:33 PM