/Speech/ Processing

NWASPA - 2011, Rajiv Gandhi Institute of Technology, Mumbai

Dr. Sunil Kumar Kopparapu

SunilKumar.Kopparapu@TCS.COM

TCS Innovation Labs - Mumbai
Yantra Park, Thane (West), Maharastra

January 2011

Speech

- Unique mode of communication
  - *(probably)* only in the human species
- Speech accounts for nearly 70% of the information and knowledge
  - *a movie minus the spoken dialogs*
- Intelligence in humans
  - *(is)* due to language and speech

Speech

- Unique mode of communication
  - *(probably)* only in the human species
- Speech accounts for nearly 70% of the information and knowledge
  - *a movie minus the spoken dialogs*
- Intelligence in humans
  - *(is)* due to language and speech

How did speech evolve?

Evolution of Speech

Communication necessitated *by need* for survival

- Most species
  - Finding food
  - Avoiding predators
  - Finding a mate (reproduction, perpetuation of species)
- Higher species
  - Emotions (Anger, love, concern etc.)
  - Combination of sounds and body gestures

Evolution of Speech

Communication necessitated *by need* for survival

- Most species
  - Finding food
  - Avoiding predators
  - Finding a mate (reproduction, perpetuation of species)
- Higher species
  - Emotions (Anger, love, concern etc.)
  - Combination of sounds and body gestures

What about humans?

Evolution of Speech in Human

- Sound and gesture based, but with a shift
  - Conveyance of emotions ⟶ communication of concepts
  - Sound and gesture ⟶ predominantly sound

    The effect?
    - Freed the limbs and body
    - Line of sight communication not crucial
    - Facilitated organization of group activity - hunting, fighting

Evolution of Speech in Human

- Sound and gesture based, but with a shift
  - Conveyance of emotions ⟶ communication of concepts
  - Sound and gesture ⟶ predominantly sound
    The effect?
    - Freed the limbs and body
    - Line of sight communication not crucial
    - Facilitated organization of group activity - hunting, fighting

Communicating Concepts needs Language

Natural Language

- Language
  - a systematic creation & usage of a set of symbols
  - a symbol is paired with an intended meaning
  - meaning established through social conventions
  - Language is a system of symbols for encoding and decoding information
- Natural Language
  - languages used by human
  - it is full of intended meaning and
  - the language rules are generally tricky and complex

Natural Language

- Language
  - a systematic creation & usage of a set of symbols
  - a symbol is paired with an intended meaning
  - meaning established through social conventions
  - Language is a system of symbols for encoding and decoding information
- Natural Language
  - languages used by human
  - it is full of intended meaning and
  - the language rules are generally tricky and complex

Speech $\equiv$ *Natural Spoken Language*

From Thought ⟶ Speech

A Top Down Look

- Intended message formed in brain
- Pragmatics, Semantics, Syntax
- Symbol String
- Inertia anticipatory & co-articulation
- Articulatory choreography
- **Speech Signal**

From Thought ⟶ Speech

A Top Down Look

- Intended message formed in brain
- Pragmatics, Semantics, Syntax
- Symbol String
- Inertia anticipatory & co-articulation
- Articulatory choreography
- **Speech Signal**

Content of Speech?

Information in Speech

- Non-linguistic, (*who said it*)
  gender, emotional states, speaker name
- Linguistic (*what (s)he said*)
  Language name and what was said (written language)
- Paralinguistic (*how well said* – manner, clarity or accent, aspects related to quality)
  deliberately added by the speaker, and not inferable from the written text.

Information in Speech

- Non-linguistic, (*who said it*)
  gender, emotional states, speaker name
- Linguistic (*what (s)he said*)
  Language name and what was said (written language)
- Paralinguistic (*how well said* – manner, clarity or accent, aspects related to quality)
  deliberately added by the speaker, and not inferable from the written text.

Goal: Automatically extract information in speech signal

Speech Processing

Speech signal processing refers to acquisition, manipulation, storage, transfer and output of human utterances by a computer. The main goals are recognition, synthesis and ~~speech compression~~.

- *Speech recognition* focuses on capturing human voice as a digital sound wave and converting it into a computer-readable format *(speech to text)*.
- *Speaker Recognition* focuses on verifying or determine the identity of the speaker.
- *Speech synthesis* is the reverse process of speech recognition. A TTS system converts normal language *text into speech*.

Speech - Processing Perspective

- Speech is a *(vocalized)* form of human communication
- *(which is)* based on syntactic combination of lexical and names from large set of words *(vocabularies)*
- Spoken word is created as a combination of speech units
- Smallest unit of speech is called a phone while a phoneme is a mental image of a phone
    - **t**ea and **t**rip have the same phoneme but different phones

Speech - Processing Perspective

- Speech is a *(vocalized)* form of human communication
- *(which is)* based on syntactic combination of lexical and names from large set of words *(vocabularies)*
- Spoken word is created as a combination of speech units
- Smallest unit of speech is called a phone while a phoneme is a mental image of a phone
  - **t**ea and **t**rip have the same phoneme but different phones

X of speech units $\longrightarrow$ *X of Speech*; $X \in \{$ *Recognition, Synthesis* $\}$

Speech Recognition – Example

- **Input** /Speech, 16 kHz, 8 bits per sample/
- **Output**
  1. a phoneme string — <sil> h au m a ch m ae k s i m a m a m A u n T sil k ae n ai w i D r ao th r U E T I e m </sil>
  2. find word boundaries using lexicon — hau mach maeksimam amAunT kaen ai wiDrao thrUE TIem
  3. converting the phoneme strings into text *How much maximum amount can I withdraw through ATM*

Speech Recognition – Example

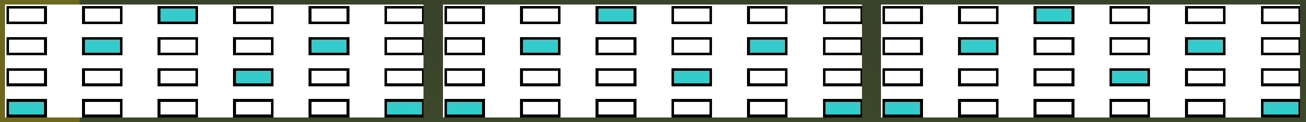
- **Input** /Speech, 16 kHz, 8 bits per sample/
- **Output**

1. a phoneme string — <sil> h au m a ch m ae k s i m a m a m A u n T sil k ae n ai w i D r ao th r U E T I e m </sil>
2. find word boundaries using lexicon — hau mach maeksimam amAunT kaen ai wiDrao thrUE TIem
3. converting the phoneme strings into text *How much maximum amount can I withdraw through ATM*

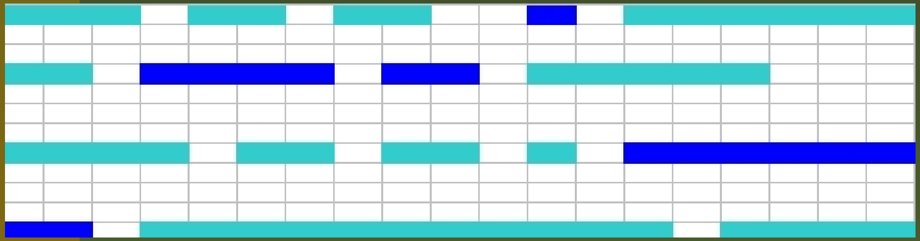
Simple? ...

More Details

- Speech Signal
- Choose several alternatives for each phoneme (*acoustic models*)



- Choose appropriate phoneme to form words (*lexicon*)



- Choose word strings to form sentences (*grammar/language model*)

Speaker Recognition – Example

Recognition

Gallery	Who spoke?
/Edna/	
/Sunil/	
/Akhilesh/	/????/
/Dipti/	
/Devanuj/	

Response: *Sunil*

1:N match

Verification

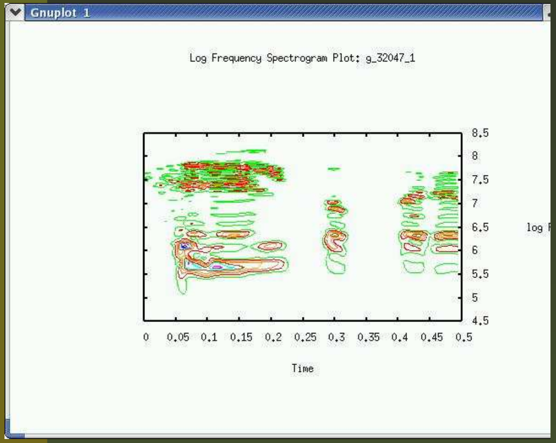
/????/	Is this Sunil?
/????/	Is this Sunil?

Response: *Yes/No*

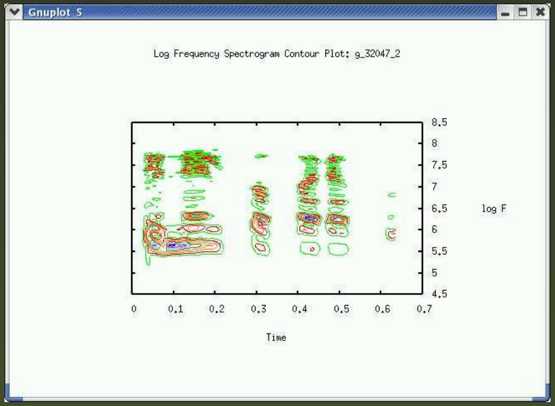
1:1 match

Diff: Spe(aker)ech Recognition

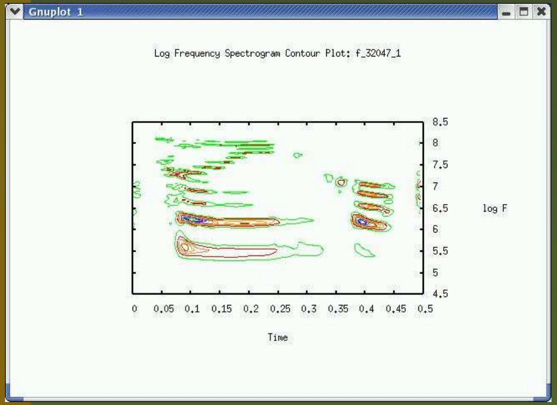
Recognize *Sunil Kopparapu*



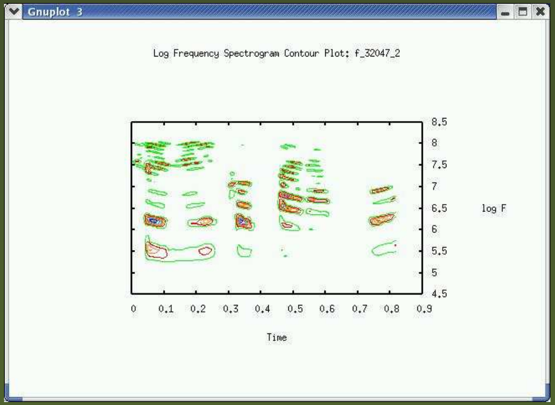
X_1



X_2



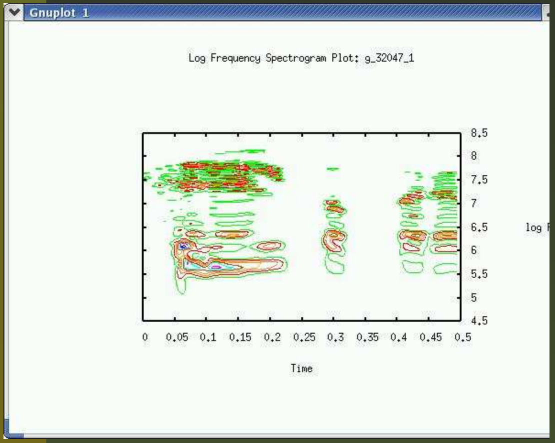
Y



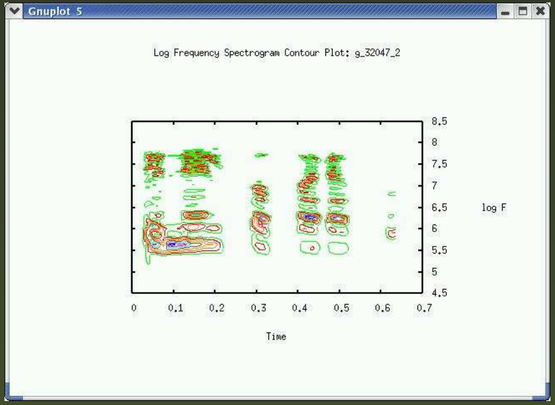
Z

Diff: Spe(aker)ech Recognition

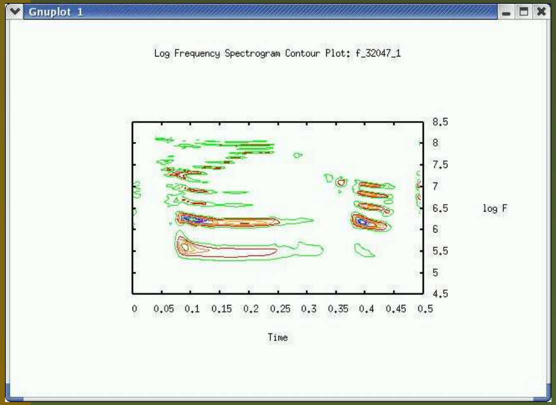
Recognize *Sunil Kopparapu*



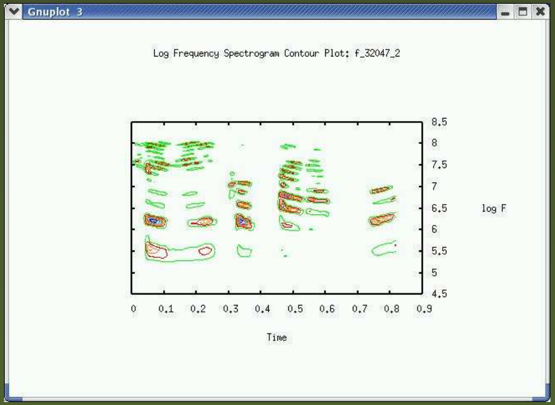
X_1



X_2



Y



Z

How? Feature Extraction, Modeling .... Later

Speech Synthesis – Example

Speech Synthesis (or TTS) is the art of making a machine speak as well as an average literate human is capable of.

- **Input** – *How much maximum amount can I withdraw through ATM*
- **Internal** – hau mach maeksimam amAunT kaen ai wiDrao thrUE TIem
- **Output** – Male,TTS Female,TTS Ideal

Speech Synthesis – Example

Speech Synthesis (or TTS) is the art of making a machine speak as well as an average literate human is capable of.

- **Input** – *How much maximum amount can I withdraw through ATM*
- **Internal** – hau mach maeksimam amAunT kaen ai wiDrao thrUE TIem
- **Output** – Male,TTS Female,TTS Ideal

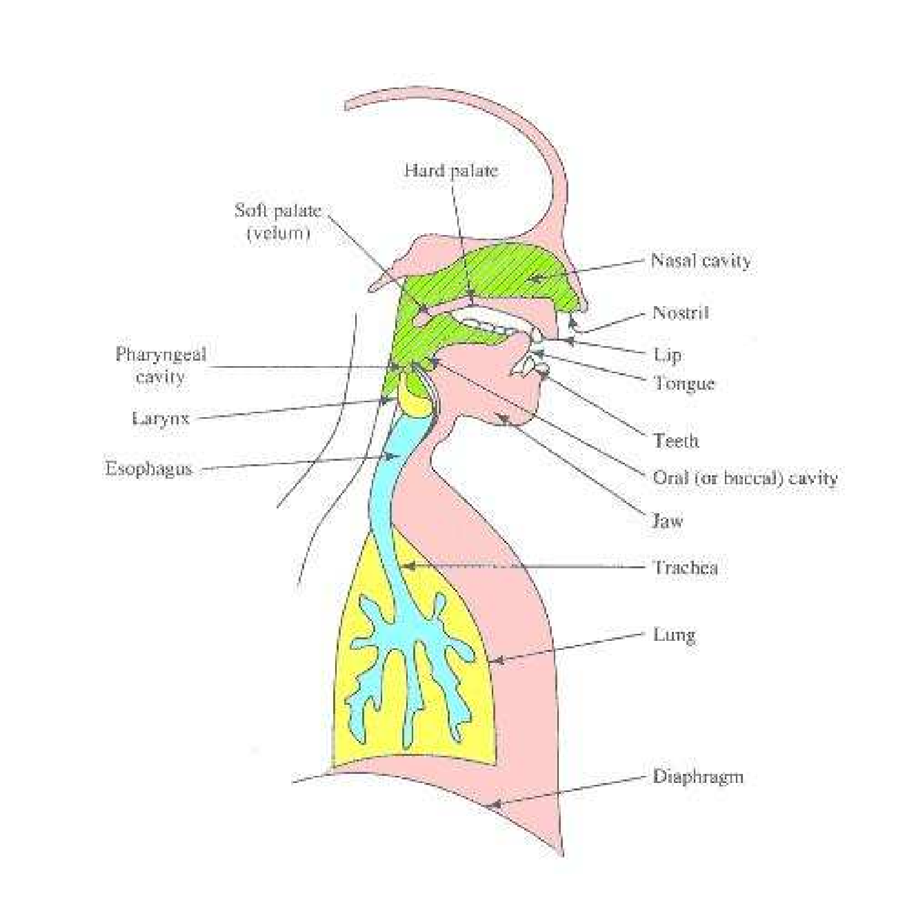
The objective of speech synthesis is *deemed* complete when a human can not distinguish between a human spoken and a machine spoken speech.

How to Process Speech?

Speech Processing primarily driven by has two schools of thought

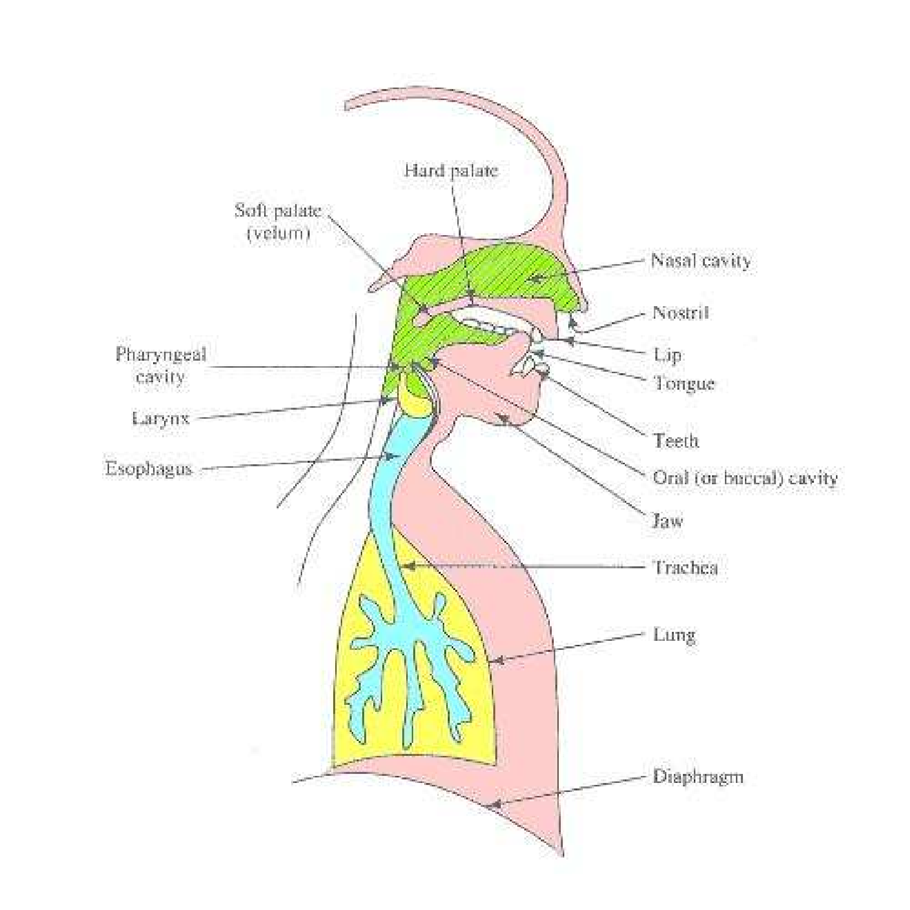
- Speech Production
    - *We should model the source of speech production because that is the origin of speech sound.*
- Speech Perception
    - *It is the ear that perceives the sound so we should process speech based on how human hear it! It doesn't really matter how it was produced.*

Speech Production



Main components **lungs**, **trachea** (wind pipe), **glottis / larynx** (organ of speech production), **pharyngeal** (throat), **oral** (mouth), **nasal** (nose) cavities.

Speech Production



Main components **lungs**, **trachea** (wind pipe), **glottis / larynx** (organ of speech production), **pharyngeal** (throat), **oral** (mouth), **nasal** (nose) cavities.

Speech is produced by a cooperation of lungs, glottis (with vocal cords) and articulation tract (mouth, nose)

Speech Production - Common Terms

- Throat and oral cavities are grouped into one unit and called **vocal tract**
  - **vocal tract** begins at the output of the **larynx** (vocal cords, or glottis) and terminates at the input to the **lips**.
- nasal cavity is often called the **nasal tract**.
  - The **nasal tract** begins at the **velum** (soft palate) and ends at the **nostrils**.
- When the velum is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech (example /n/ in **n**et, /m/ in **m**et)

Speech Production Process

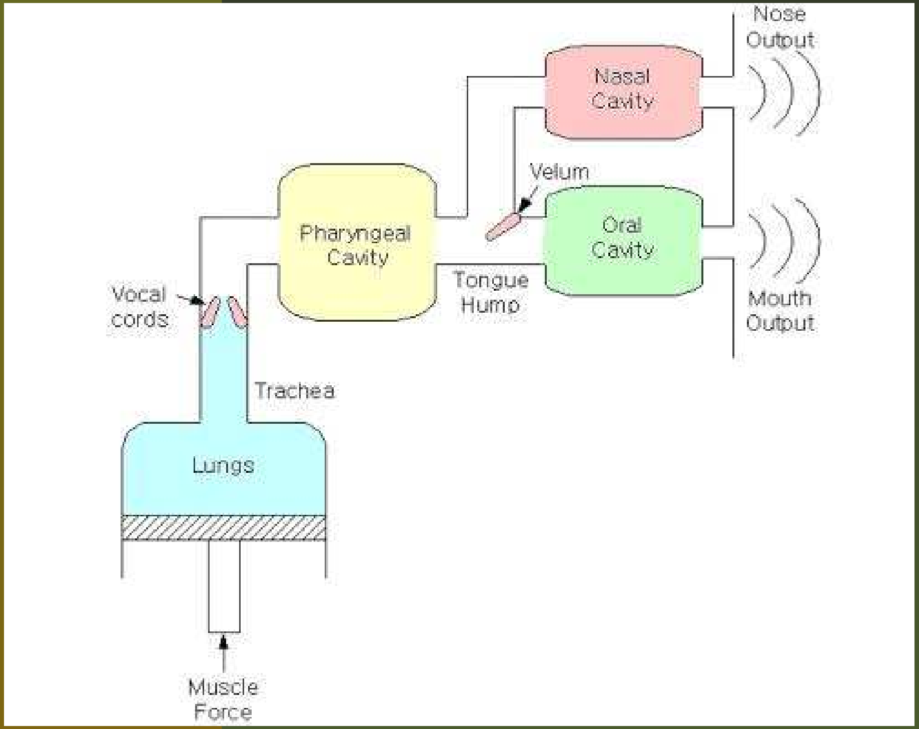
- Air enters the lungs (normal breathing); air is expelled from the lungs through the trachea (wind pipe)
- Tensed vocal cords within the larynx are caused to vibrate by the air flow.
- The air flow is chopped into quasi-periodic pulses which are then
- Modulated in frequency in passing through the throat, the oral cavity, and possibly nasal cavity.

Speech Production Process

- Air enters the lungs (normal breathing); air is expelled from the lungs through the trachea (wind pipe)
- Tensed vocal cords within the larynx are caused to vibrate by the air flow.
- The air flow is chopped into quasi-periodic pulses which are then
- Modulated in frequency in passing through the throat, the oral cavity, and possibly nasal cavity.

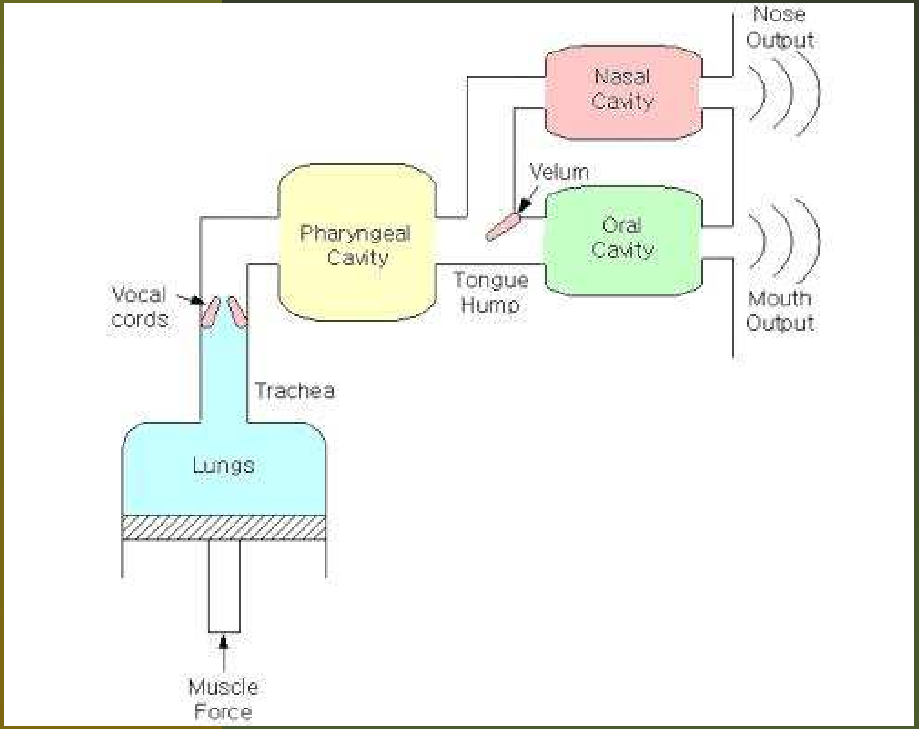
Depending on the positions of the articulators (jaw, tongue, velum, lips, mouth), different sounds are produced.

Speech Generation Simplified



- The lungs (and associated muscles) act as the **source** of air for exciting the vocal mechanism
- **Filter:** Vocal tract (begins at the output of the vocal cords and terminates at the input to the lips)

Speech Generation Simplified



Source-Filter Model

- The lungs (and associated muscles) act as the **source** of air for exciting the vocal mechanism
- **Filter:** Vocal tract (begins at the output of the vocal cords and terminates at the input to the lips)

Voiced – Unvoiced Speech

- When the vocal cords are tensed, the air flow causes them to vibrate, producing so-called **voiced speech** sounds.
- When the vocal cords are relaxed, in order to produce a sound, the air flow passes through a constriction in the vocal tract and thereby become turbulent, producing so-called **unvoiced speech** sounds

Consonants	
Unvoi	Voi
/p/	/b/
/t/	/d/
/k/	/g/
/f/	/v/
/s/	/z/

Place fingers on the voice box (Adam's apple). Pronounces zzzz (vibration?). Pronounces ssss (no vibration?).

Speech - Voiced and Unvoiced

- **Voiced speech**
  generated by the modulation of the air-stream of the lungs by periodic opening and closing of the vocal folds in the glottis or larynx. For vowels and nasal consonants like /m/, /n/.
- **Unvoiced speech**
  generated by a constriction of the vocal tract narrow enough to cause turbulent airflow, which results in noise (in fricatives like /f/, /s/), or breathy voice (where the constriction is in the glottis) and unvoiced plosives like /p/, /t/, /k/

Speech - Voiced and Unvoiced

- **Voiced speech**
  generated by the modulation of the air-stream of the lungs by periodic opening and closing of the vocal folds in the glottis or larynx. For vowels and nasal consonants like /m/, /n/.
- **Unvoiced speech**
  generated by a constriction of the vocal tract narrow enough to cause turbulent airflow, which results in noise (in fricatives like /f/, /s/), or breathy voice (where the constriction is in the glottis) and unvoiced plosives like /p/, /t/, /k/

Signal Processing View

Speech Prod - Signal Proc View (1)

- The source $S(z)$ is modeled by either an impulse train *for voiced speech*, or a random signal *for unvoiced component*
- Effect of the shape of the vocal tract is modeled by $V(z)$ and the radiation characteristics of the lips are taken into account by $L(z)$.
- These three filters can be combined to one single filter $H(z)$

$$H(z) = S(z)V(z)L(z)$$

Speech Prod - Signal Proc View (1)

- The source $S(z)$ is modeled by either an impulse train *for voiced speech*, or a random signal *for unvoiced component*
- Effect of the shape of the vocal tract is modeled by $V(z)$ and the radiation characteristics of the lips are taken into account by $L(z)$.
- These three filters can be combined to one single filter $H(z)$

$$H(z) = S(z)V(z)L(z)$$

How does one use this?

Speech Prod - Signal Proc View (2)

- The source $S(z)$ and lip radiation $L(z)$ are mostly constant and well known a priori,
- there is an overall of -6 dB/octave decay in speech radiated from lips, as frequency increases
- Compensation for this is taken care while speech signal processing is done through **pre-emphasis**.
- the vocal tract transfer function $V(z)$ is the characteristic part to determine the **content** of the speech being uttered.

Speech Prod - Signal Proc View (2)

- The source $S(z)$ and lip radiation $L(z)$ are mostly constant and well known a priori,
- there is an overall of -6 dB/octave decay in speech radiated from lips, as frequency increases
- Compensation for this is taken care while speech signal processing is done through **pre-emphasis**.
- the vocal tract transfer function $V(z)$ is the characteristic part to determine the **content** of the speech being uttered.

Used in Speech Processing

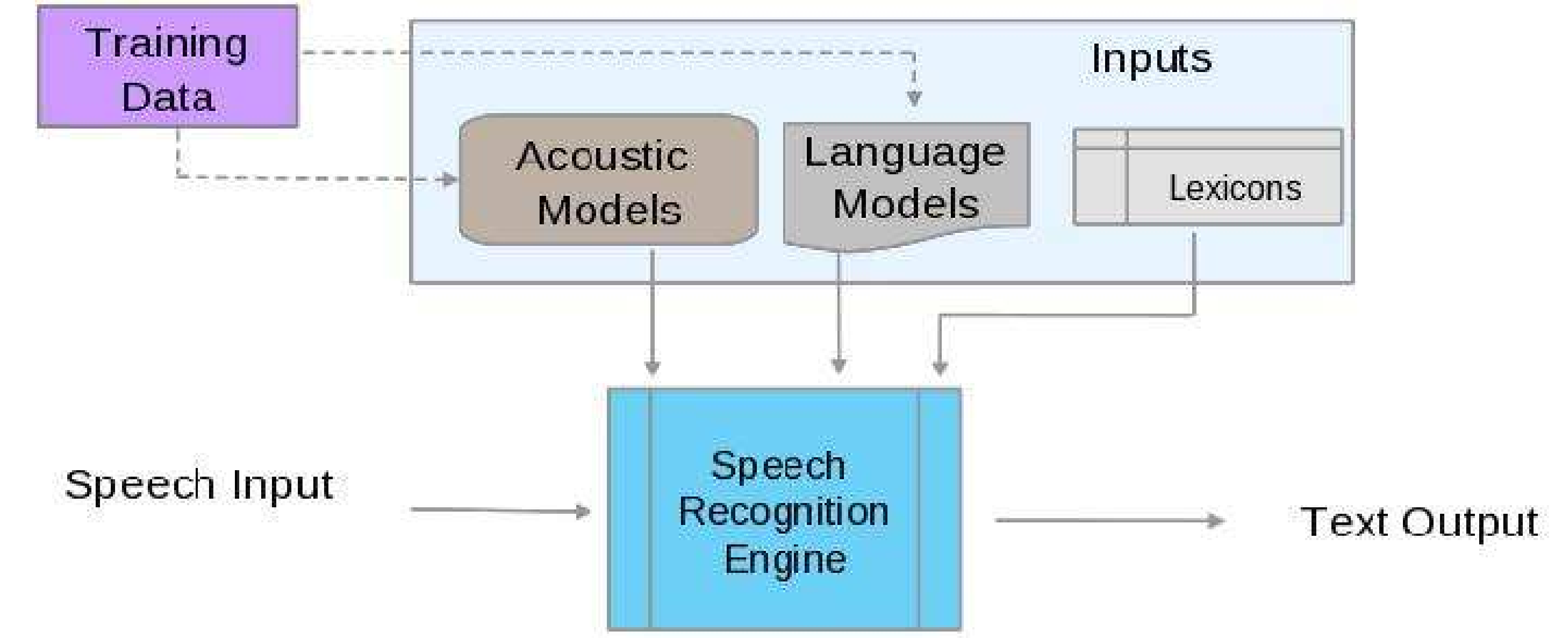
Speech (Speaker) Recognition

- Speech recognition can be loosely termed as the ability of a machine to recognize (content not the intent) what is being spoken.
- The system is called speaker dependent (speaker independent) if it can recognize speech spoken by only a particular person (any person).
- A system that can recognize a limited set of predefined words (a large vocabulary of words) is called an isolated word recognition (continuous speech recognition).

Speech (Speaker) Recognition

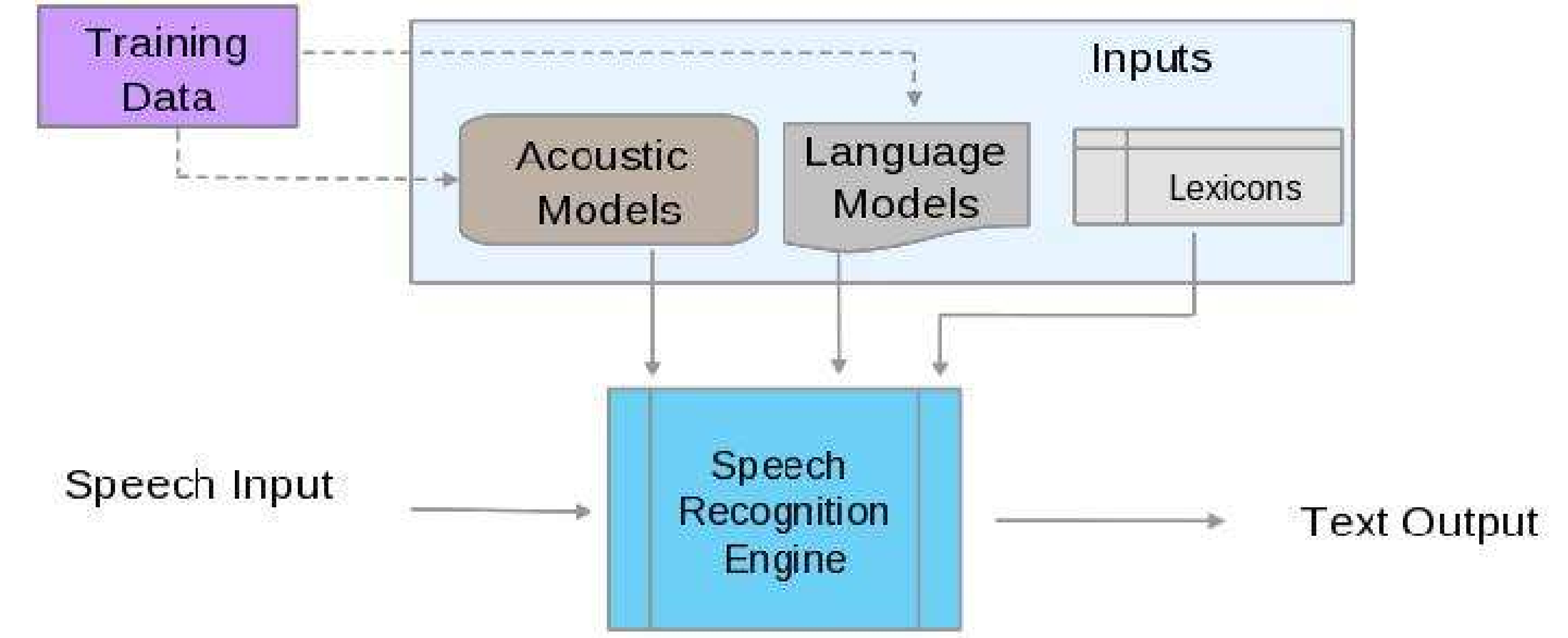
- Speech recognition can be loosely termed as the ability of a machine to recognize (content not the intent) what is being spoken.
- The system is called speaker dependent (speaker independent) if it can recognize speech spoken by only a particular person (any person).
- A system that can recognize a limited set of predefined words (a large vocabulary of words) is called an isolated word recognition (continuous speech recognition).
- Speech Recognition involves (i) Training and (ii) Decoding or recognition

Speech Recognition – Complete View



- Acoustic Model (HMM/Gaussian Mixture Model)
- Lexicon (or Dictionary)
- Language Model (Speech Grammar)

Speech Recognition – Complete View



- Acoustic Model (HMM/Gaussian Mixture Model)
- Lexicon (or Dictionary)
- Language Model (Speech Grammar)

Some Details ...

Speech Recognition - Details

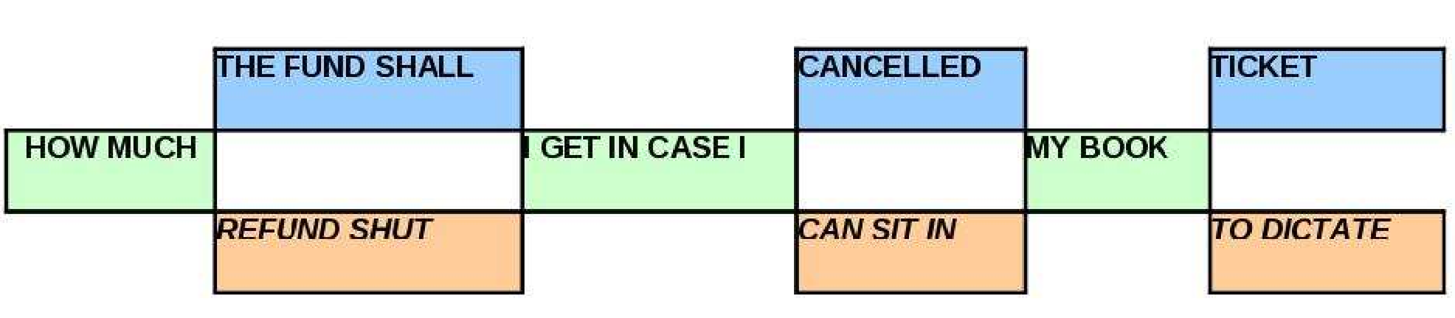
- Acoustic Models
  - models sounds (phonemes) that make up a word
  - modeled using (manually) transcribed audio data
- Lexicon (or Dictionary)
  - list of words and its corresponding phonetic transcription pronunciation (manually constructed)
- Language Model (LM)
  - models the syntax and grammar of the sentences to be recognized
  - constructed using large amount of text corpus (domain)

Sample Lexicon; Grammar

- Lexicon

Word	Phoneme String
INSURANCE	IH N SH UH R AH N S
POLICY	P AA L AH S IY
SURRENDER	S ER EH N D ER
SUPERVISOR	S UW P ER V AY Z ER

- Grammar (n-gram)



THE FUND SHALL

HOW MUCH

REFUND SHUT

I GET IN CASE I

CANCELLED

CAN SIT IN

MY BOOK

TICKET

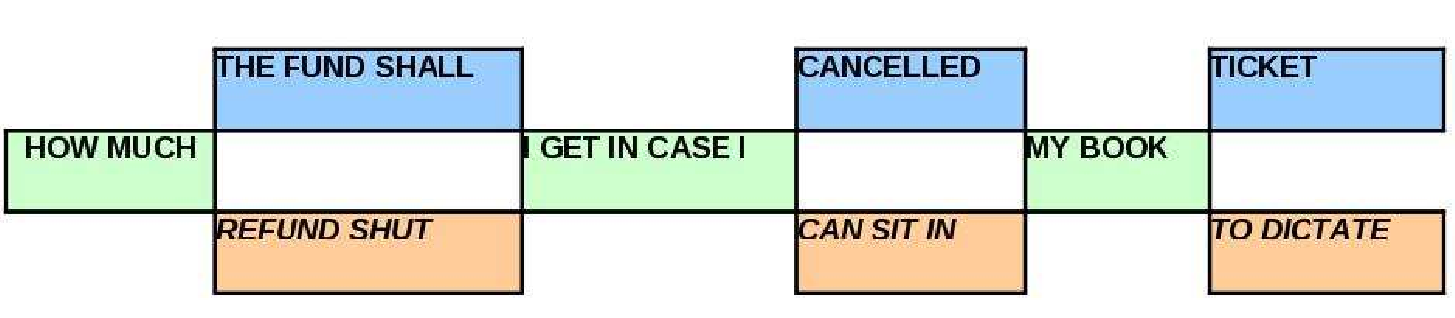
TO DICTATE

Sample Lexicon; Grammar

- Lexicon

Word	Phoneme String
INSURANCE	IH N SH UH R AH N S
POLICY	P AA L AH S IY
SURRENDER	S ER EH N D ER
SUPERVISOR	S UW P ER V AY Z ER

- Grammar (n-gram)



HOW MUCH

THE FUND SHALL

REFUND SHUT

I GET IN CASE I

CANCELLED

CAN SIT IN

MY BOOK

TICKET

TO DICTATE

Use of HMMs in Speech Recognition

Hidden Markov models (HMMs) are best suited for modeling speech

- Statistical models (able to capture large variations which are possible in speech)
- Able to preserve temporal information (important in speech)
- Have been in use for several decades (with no visible replacements spare Artificial Neural Networks)
- Their use has been successfully demonstrated (time and again)

Use of HMMs in Speech Recognition

Hidden Markov models (HMMs) are best suited for modeling speech

- Statistical models (able to capture large variations which are possible in speech)
- Able to preserve temporal information (important in speech)
- Have been in use for several decades (with no visible replacements spare Artificial Neural Networks)
- Their use has been successfully demonstrated (time and again)

When CPU's were slow one used Dynamic Time Warping also called edit distance in CS

Speech Recognition Preprocessing(1)

- Speech is non-stationary
  - statistics of the speech signal change with time
- Make it stationary
- Why?
  - to use rich signal processing literature
  - extract features and build statistical model
- How?
  - process smaller *portions* of speech *(frames)*
  - Typically 10-20 ms of speech – the signal can be considered to be stationary

Speech Recognition Preprocessing(1)

- Speech is non-stationary
  - statistics of the speech signal change with time
- Make it stationary
- Why?
  - to use rich signal processing literature
  - extract features and build statistical model
- How?
  - process smaller *portions* of speech *(frames)*
  - Typically 10-20 ms of speech – the signal can be considered to be stationary

a key assumption in speech processing systems

Speech Recognition Preprocessing(2)

Word Recognition

- Removing non-speech signal,
- Dividing the speech signal into frames,
- Pre-emphasizing the signal
  - spectral flattening to make it less susceptible to finite precision effects in signal processing
  - *to offset 3 dB per octave fall due to the effect of radiation from the lips*
- Tapering (Windowing) the frames (Hamming window) to minimize signal discontinuities at the beginning and end of the frame.

Speech Recognition Preprocessing(2)

Word Recognition

- Removing non-speech signal,
- Dividing the speech signal into frames,
- Pre-emphasizing the signal
  - spectral flattening to make it less susceptible to finite precision effects in signal processing
  - *to offset 3 dB per octave fall due to the effect of radiation from the lips*
- Tapering (Windowing) the frames (Hamming window) to minimize signal discontinuities at the beginning and end of the frame.

In some detail ..

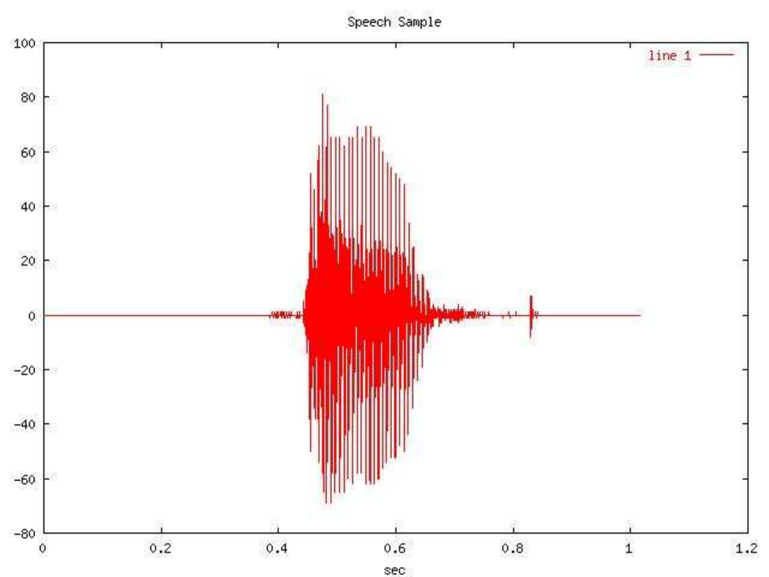
End Silence Detection

- An energy based thresholding ($\mathcal{T}$) is used to determine if each frame window of the speech signal is a speech frame or a non-speech frame.
- For N frames compute. Compute $\mathcal{A}_{max}^{1}, \mathcal{A}_{max}^{1}, \cdots \mathcal{A}_{max}^{N}$ (maximum amplitude in each frame). Compute,

$$\mathcal{T} : \frac{(\mu + 2\sigma^{2}) + \mathcal{A}_{max}}{15}$$

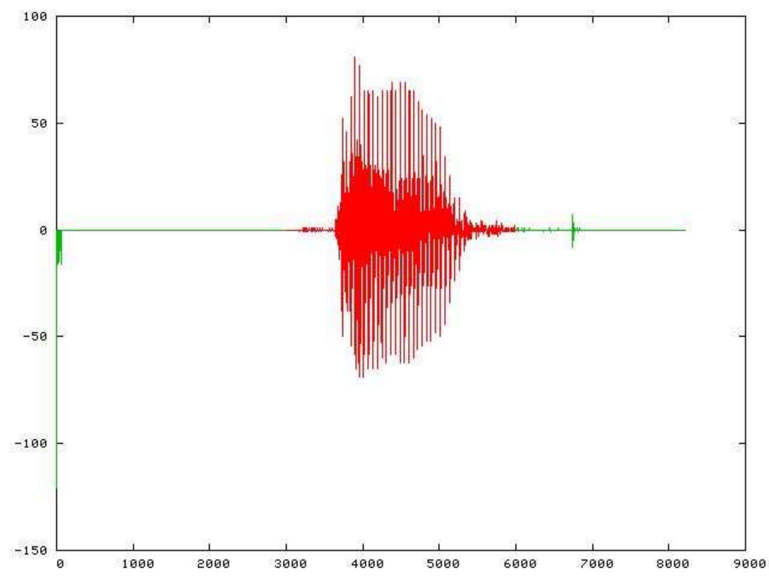
- Frame i is a speech frame if $(\mathcal{A}_{max}^{i} > \mathcal{T})$. The speech frames between the first identified speech frame from the start of the speech file and the last speech frame is the end silence detected speech.

Example: End Silence Detection

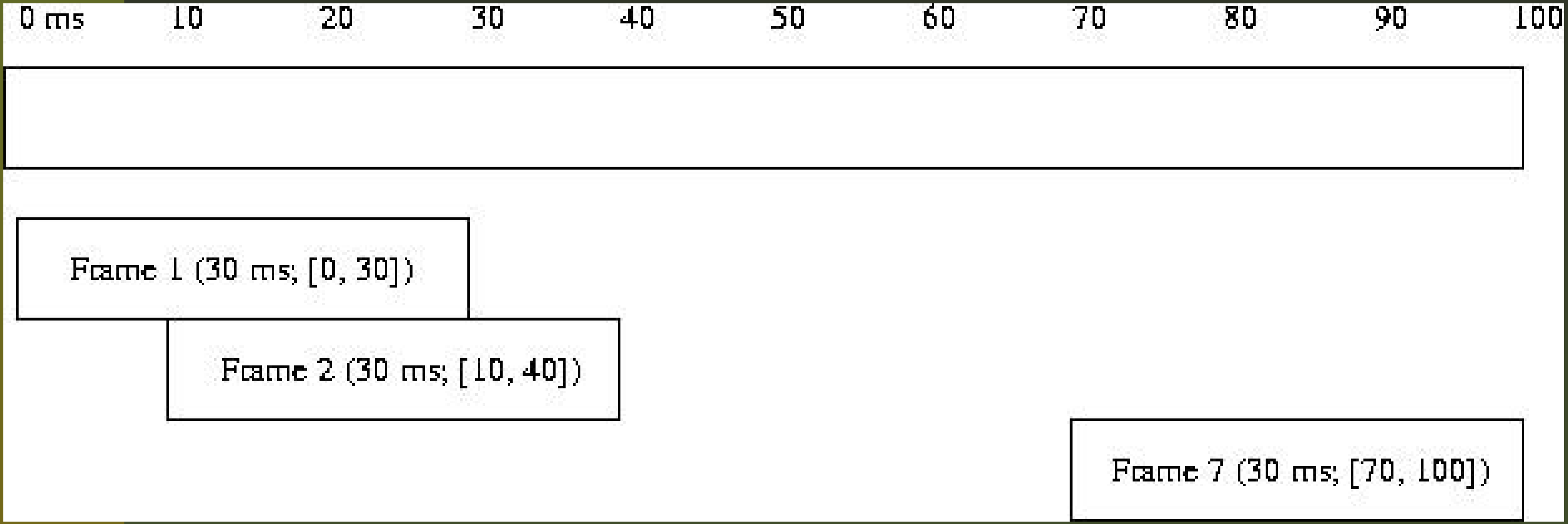


/sil Dark sil/

/Dark/

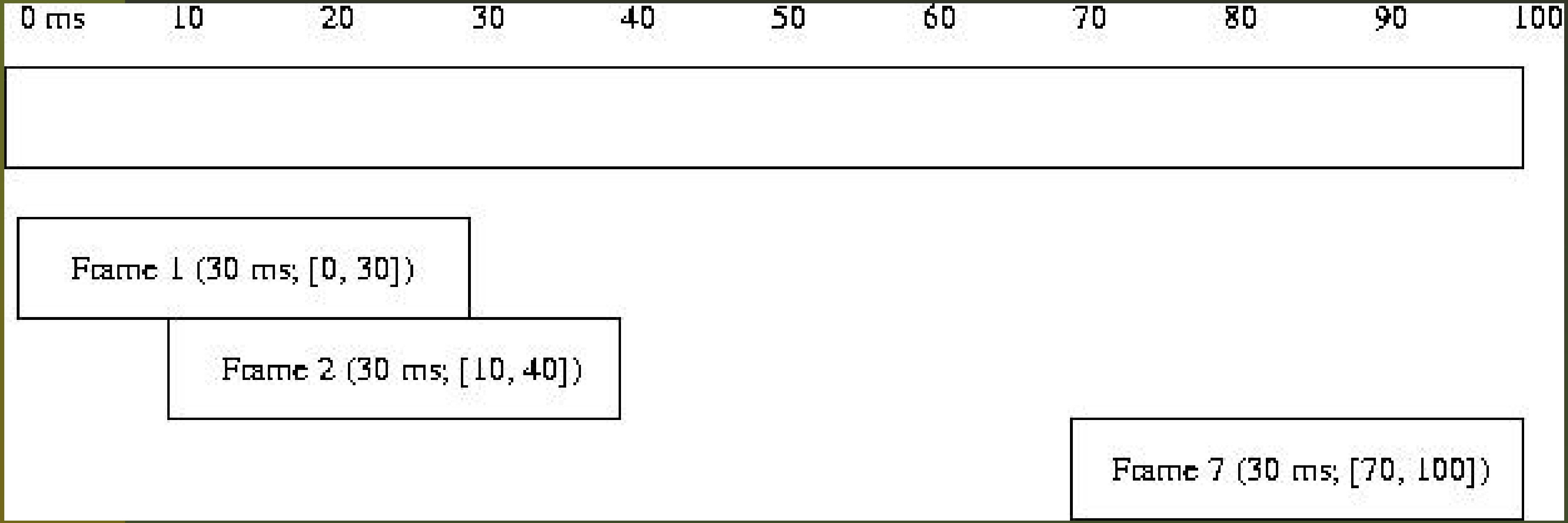


Speech Analysis Frames



- Speech duration T (= 100 ms)
- total number of frames is given by (T - F_{size})/F_{shift} = (100 - 30)/10 = 7,
- typical frame size (F_{size}): 30 ms and frame shift (F_{shift}): 10 ms (*typically*, $F_{shift} = F_{size}/3, F_{size}/2$)

Speech Analysis Frames



- Speech duration T (= 100 ms)
- total number of frames is given by (T - F_{size})/F_{shift} = (100 - 30)/10 = 7,
- typical frame size (F_{size}): 30 ms and frame shift (F_{shift}): 10 ms (*typically*, $F_{shift} = F_{size}/3$, $F_{size}/2$)

Process each frame separately

Pre-Emphasis (1)

- Pre-emphasizing the signal is necessary to spectrally flatten the signal to make it less susceptible to finite precision effects in signal processing. *This is done by a 1^{st} order Finite Impulse Response (FIR) filter.*
- The impulse response $H(z)$ of a pre-emphasis filter is

$$H(z) = 1 - \phi z^{-1} \quad \text{where} \quad \phi \in [0.9, 1.0]$$

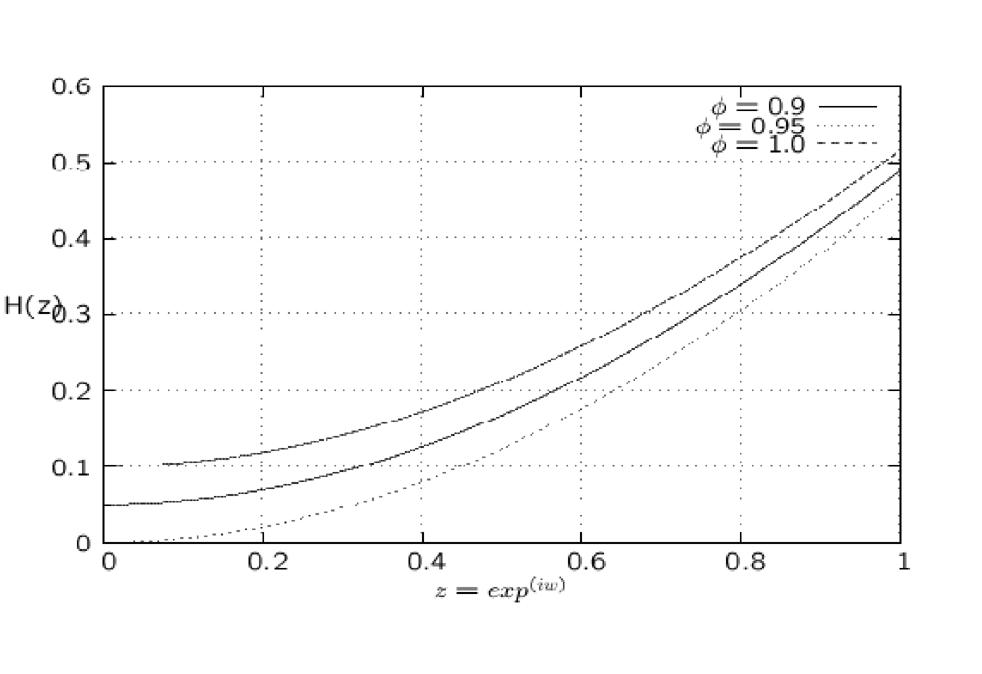
- In the time domain this is equivalent to the difference equation

$$s_p(n) = s(n) - \phi s(n-1)$$

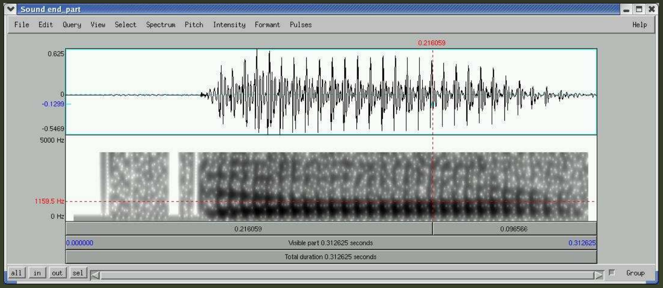
Pre-Emphasis (2)

- where, $s_p(n)$ is the n^{th} sample of the pre-emphasized signal, $s(n)$ is the n^{th} sample of the original signal and ϕ is the pre-emphasis factor.

Note: The effect *emphasize high freq content and deemphasizing low freq content*. Why? to compensate for the attenuation due to lip radiation.



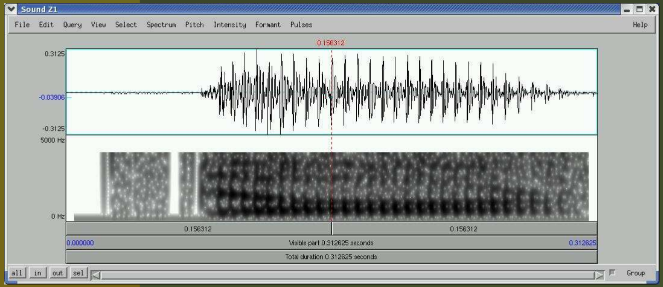
Ex: Pre-Emphasis ($\phi = 1.0, 0.9$)



Visible part 0.312625 seconds

Total duration 0.312625 seconds

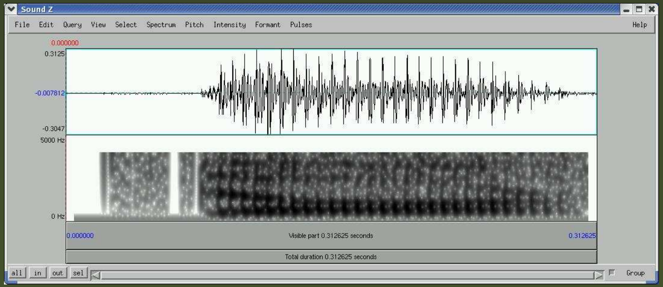
/Dark/



Visible part 0.312625 seconds

Total duration 0.312625 seconds

/Dark $\phi = 0.9$/

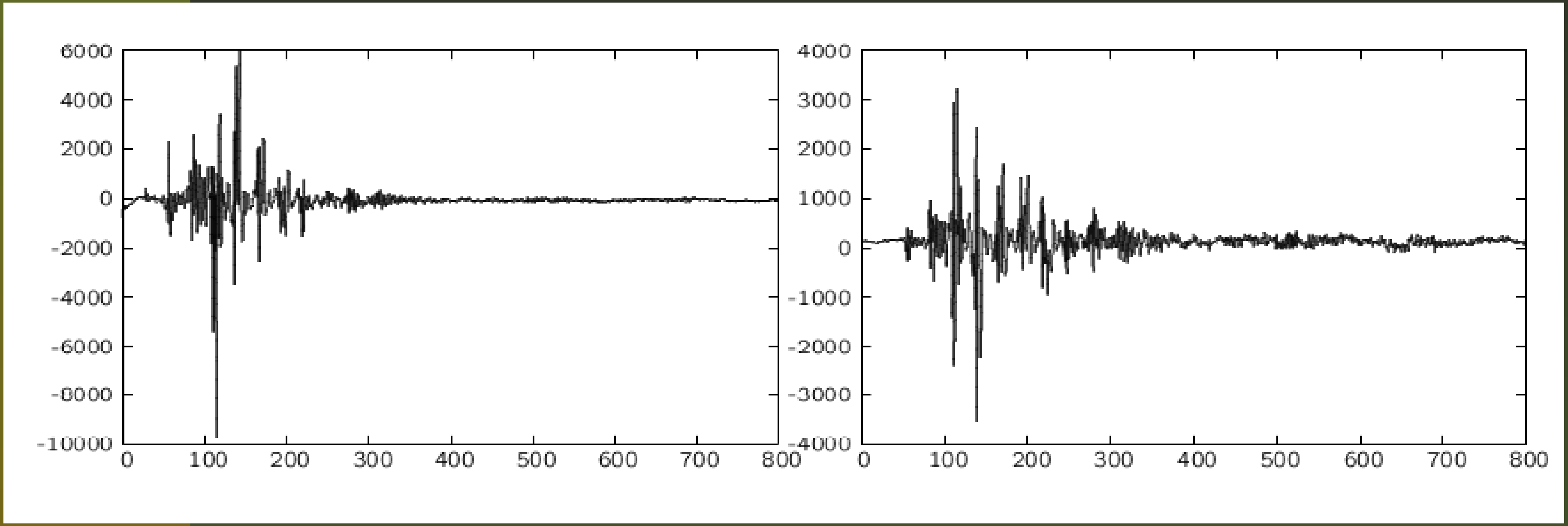


Visible part 0.312625 seconds

Total duration 0.312625 seconds

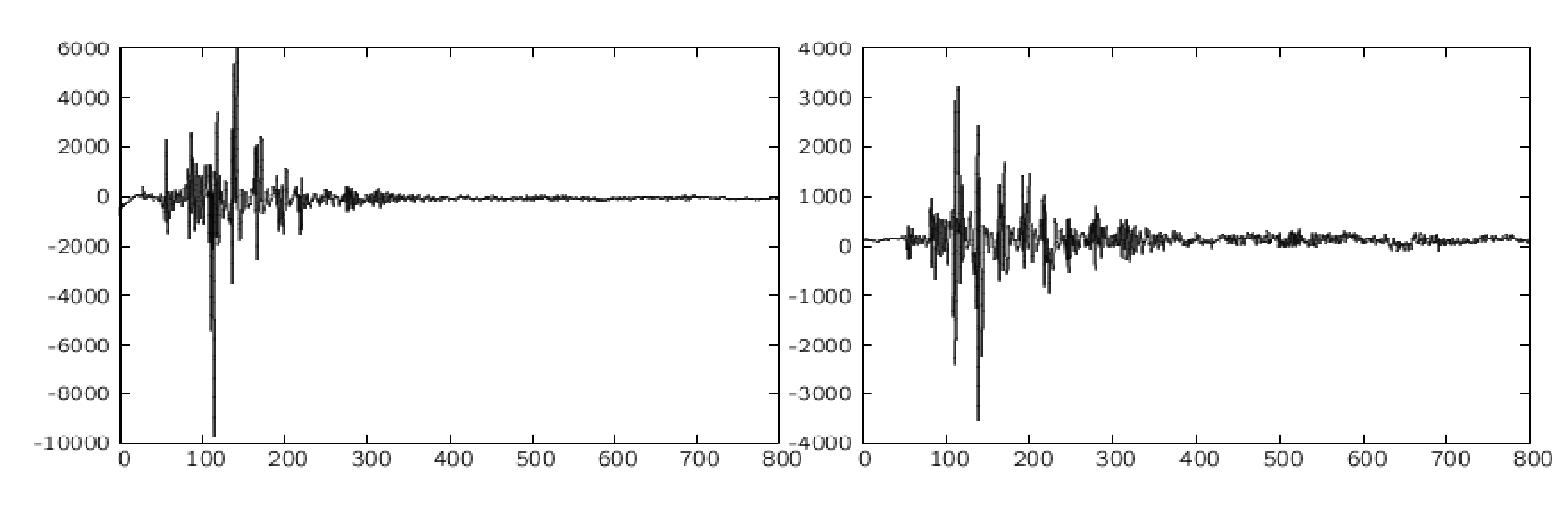
/Dark $\phi = 1.0$/

Pre-Emp ($\phi = 1$) in Freq domain



Before Pre-emphasis After Pre-emphasis

Pre-Emp ($\phi = 1$) in Freq domain



Before Pre-emphasis After Pre-emphasis

Observe: The low-frequency content is attenuated while the high frequency content is enhanced.

Windowing: Hamming

- Windowing is done on each frame of the speech signal to minimize signal discontinuities at the beginning and the end of each frame.
- The signal s_f (N - speech samples in a frame) is multiplied by a Hamming window $w(n)$ length N.

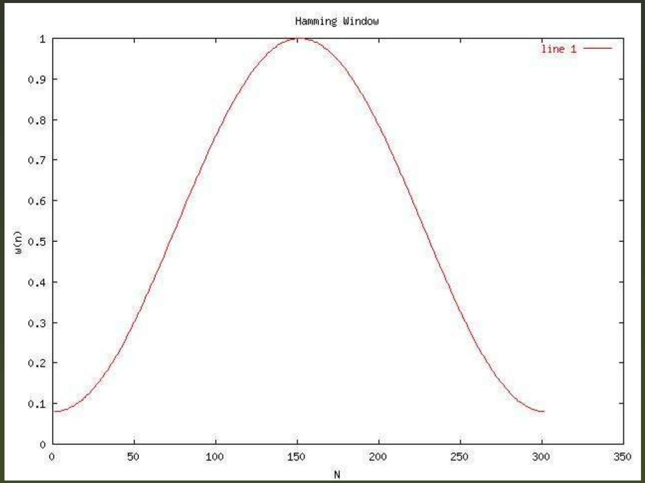
$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \leq n \leq N-1$$

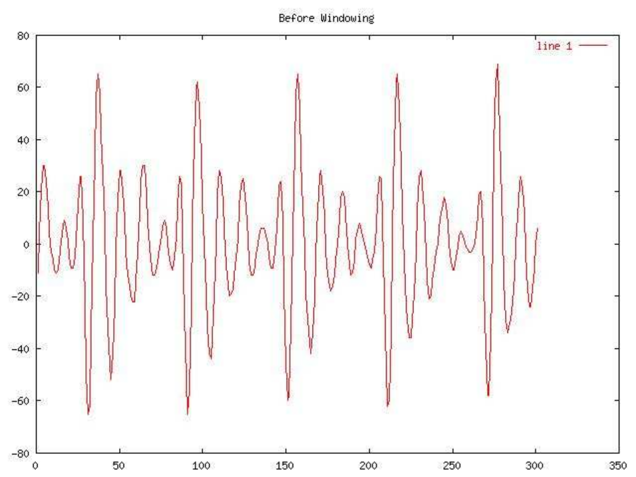
- The windowed signal ($s_w(n)$) is obtained as

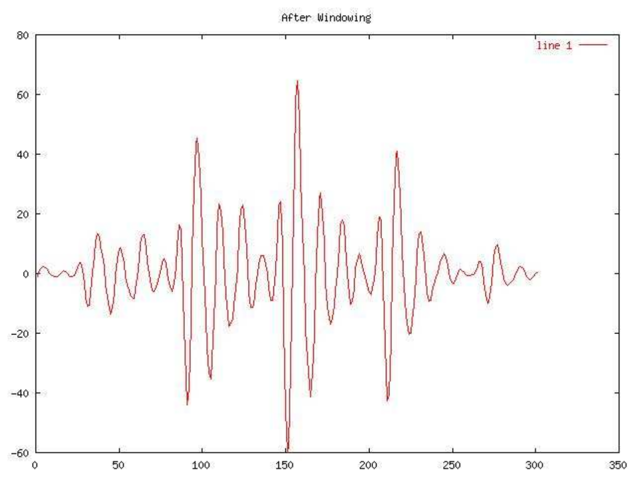
$$s_w(n) = s_f(n) w(n)$$

where $s_f(n)$ is the speech frame

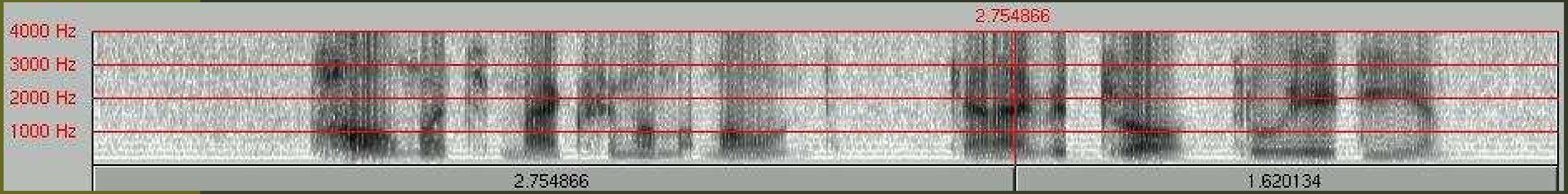
Windowing: Example (Time)



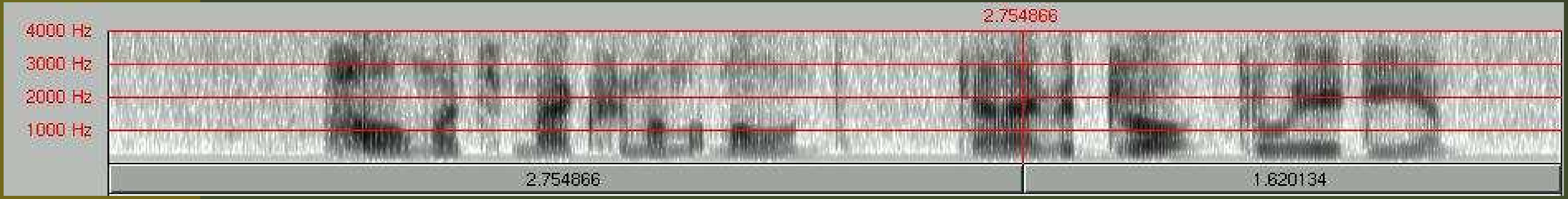




Windowing: Example (Spectrogram)

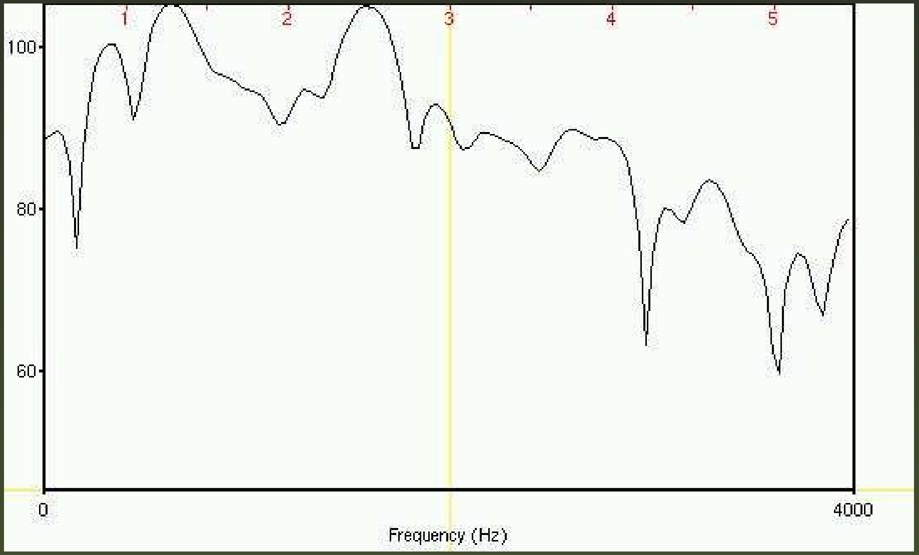


No Windowing

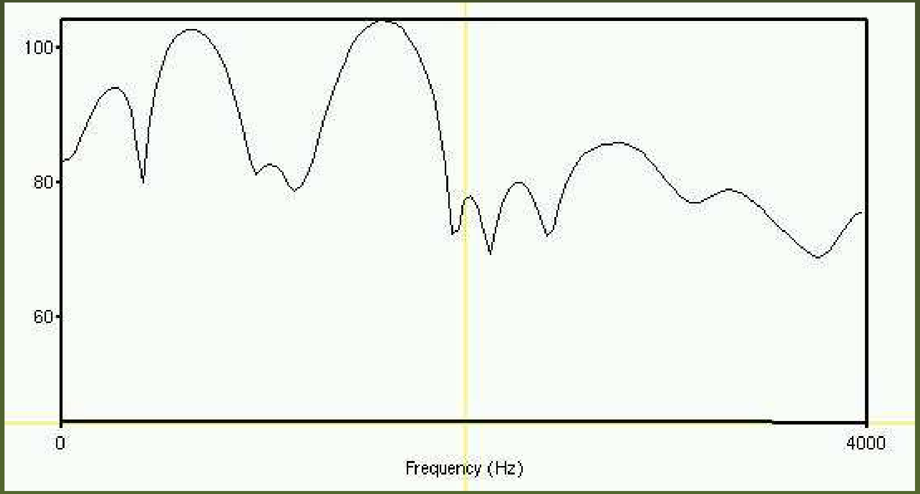


With Windowing

Windowing: (Spectrum Slice 2.75 s)



No Windowing



Windowing

Need for Parameter Extraction

- A speech signal of 5 seconds duration (sampled at 16 kHz) is made up of 80000 speech samples.
  - large number of samples – do not *e*xplicitly exhibit information – to build statistical models.
- So – extract a few (typically 30-40) parameters,
- How? by suitably processing the signal are *representative* of all the 80000 speech samples.
- parameters based on human speech generation or perception model

Example Parameters: LPC, CC-LPC, MFCC, ΔMFCC, Δ^2MFCC, Energy, Pitch, Amplitude, Formants etc.

LPC: Speech Parameters

- Speech modeled as a p^{th} order autoregressive (AR) model,

$$s_f(n) \approx \sum_{i=1}^{p} a_i s_f(n-i)$$

  $s_f(n)$ is n^{th} speech sample in the f^{th} frame and $\{a_1, a_2, \cdots a_p\}$ are the LPC parameters. *(current sample is a weighted sum of previous p samples.)*

- Typically, p is $\frac{f_s}{1000} + 2$ where f_s is the sampling frequency of the signal s_f.
- For a 8 kHz speech p is 10

Computing LPC

$$r(m) = \sum_{n=0}^{N-1-m} s(n)s(n+m) \quad \text{for} \quad m = 1, \cdots p$$

where N is the frame size. For $1 \leq i \leq p$ compute

$$\begin{aligned}
E^{o} &= r(0) \\
k_i &= \frac{r(i) - \sum_{j=1}^{i-1} \alpha_j^{i-1} r(|i-j|)}{E^{(i-1)}} \\
\alpha_i^{(i)} &= k_i \\
\alpha_j^{(i)} &= \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \\
E^{i} &= (1 - k_i^2) E^{i-1}
\end{aligned}$$

The LPC parameter is obtained as $\boxed{a_i = \alpha_i^{(i)}}$

Computing Cepstral Coefficients

- The Cepstral coefficients are the coefficients of the Fourier transform representation of the logarithm magnitude spectrum.
- Consider a speech signal $s(n)$. Define the Fourier transform pair as

$$s(n) \leftrightarrow S(\omega)$$

- The Cepstral, $c_s(n)$ is defined as the inverse Fourier transform $(C_s(\omega) \leftrightarrow c_s(n))$ of $C_s(\omega)$, where

$$C_s(\omega) = \log_e |S(\omega)|$$

Ceptsral Computing in Speech

- If a frame of speech samples is represented by

$$s(n) = e(n) * h(n)$$

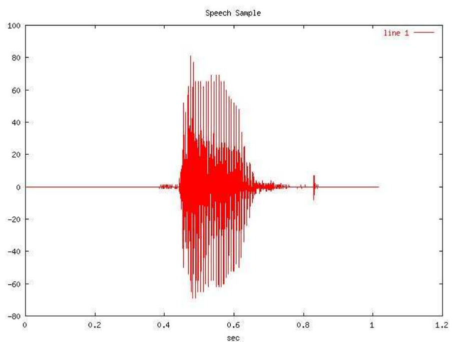
where $e(n)$ is the excitation source signal and $h(n)$ is the vocal tract system model, then in the cepstral domain we have

$$C_s(\omega) = C_e(\omega) + C_h(\omega)$$

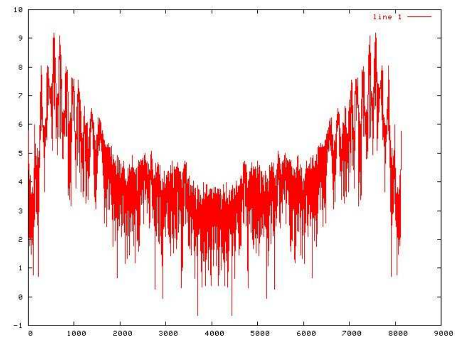
- The ability of the cepstrum of a frame of speech to separate the excitation source from the vocal tract system model is often exploited in speech signal processing.

Source-Signal Separation in Cepstral

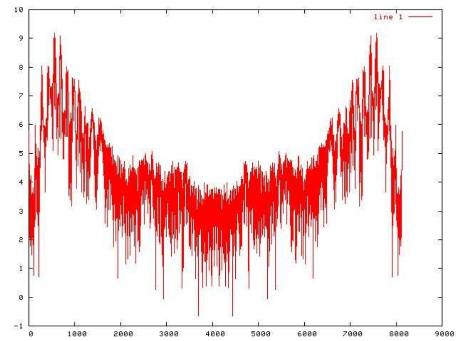
$e(n) * h(n) = s(n)$



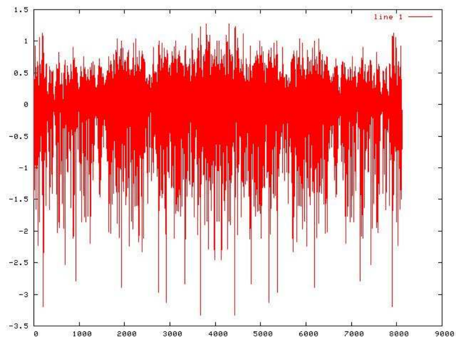
$\longleftrightarrow$



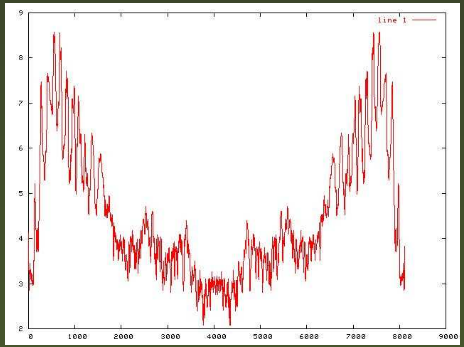
$C_s(\omega)$



$=$



$+$



$$C_s(\omega) = C_e(\omega) + C_h(\omega)$$

Importance of Ceptsral Computing

- Speech can be considered to be produced as a convolution of an excitation source and vocal tract.
- The excitation source; a periodic pulse source (voiced speech) or noise (unvoiced) while the vocal tract model has a slowly varying spectral envelope.
- The Cepstral of a frame of speech enables separation of the excitation source from the vocal tract system model. This results in vocal tract model to lower indices in the Cepstral (time, or quefrency) domain and the excitation to higher Cepstral indices.
- This is often exploited in speech signal processing (lower Cepstral indices for speech recognition, higher Cepstral for speaker recognition).

Par: LPC Ceptsral Coefficients

Given the LPC parameters $a_1, a_2, \cdots, a_p$. The Cepstral parameters $c_1, c_2, \cdots, c_q$ are calculated recursively as

$$\begin{aligned} c_1 &= a_1 \\ c_n &= a_n + \sum_{m=1}^{n-1} \left(\frac{m}{n}\right) a_m c_{n-m} \quad \text{for} \quad 2 \leq n \leq p \\ c_n &= \sum_{m=1}^{p} \left(\frac{n-m}{n}\right) a_m c_{n-m} \quad \text{for} \quad n > p \end{aligned}$$

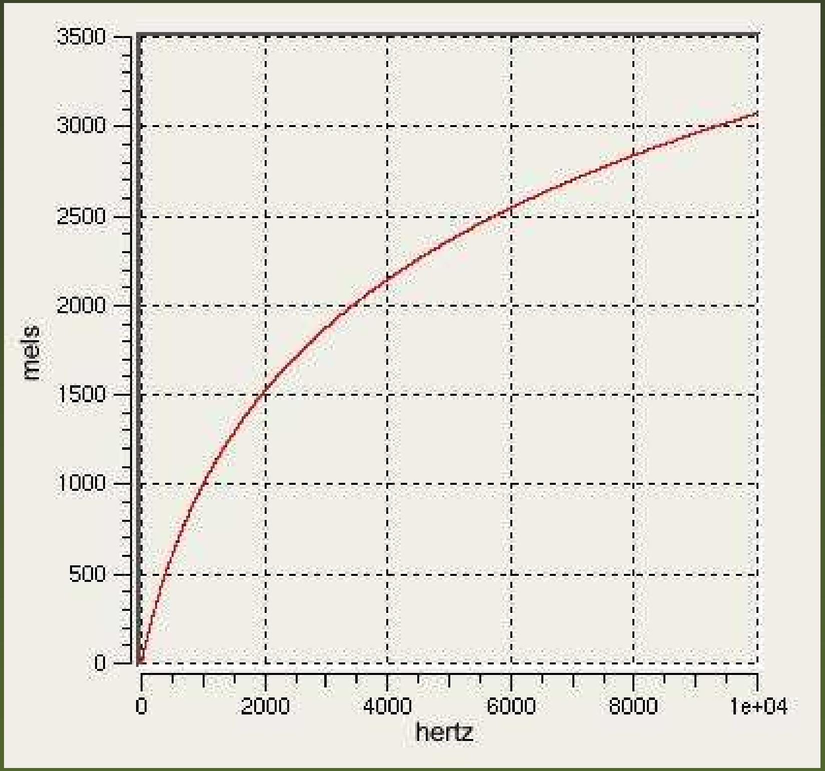
Typically one chooses $q \approx \left(\frac{3}{2}\right) p$

Parameter: MFCC, ΔMFCC (1)

Mel Frequency Cepstral Coefficients are based on the perception model.
Mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another.

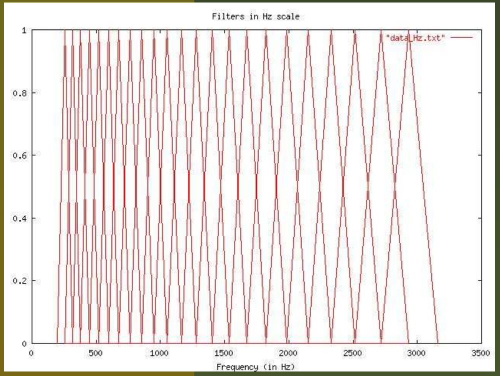
$$Mel(f) =$$

$$2595 \log_{10} \left\{ 1 + \frac{f}{700} \right\}$$



Parameter: MFCC, ΔMFCC (2)

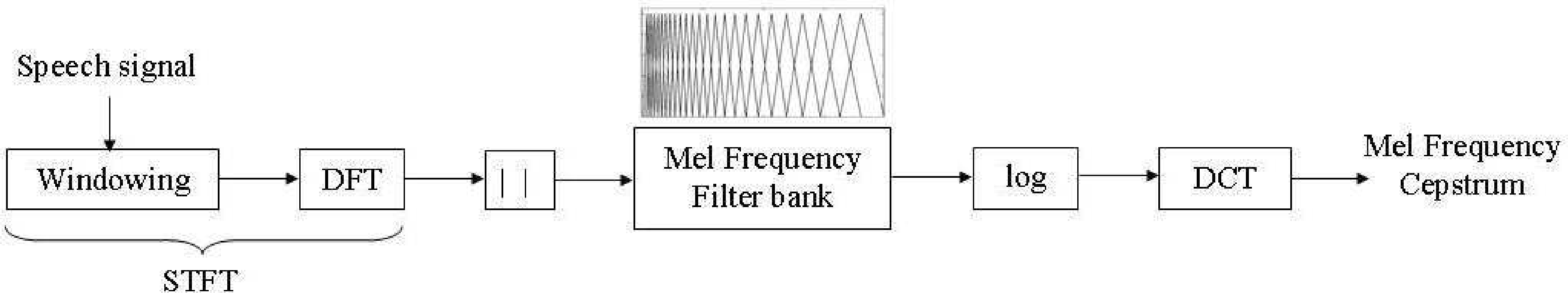
24 mel *(filter)* channels (in f scale *(in mel scale filter bandwidth is equal; range 240 - 3400 Hz*) Energy in each filter band (e_1, e_2, $\cdots$ e_{24})



$$m_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N=24} e_j \cos\left\{\frac{\pi i}{N}(j - 0.5)\right\}$$

$$\Delta m_k = \frac{\sum_{l=1}^{L} l(m_{k+l} - m_{k-l})}{2\sum_{l=1}^{L} l^2}$$

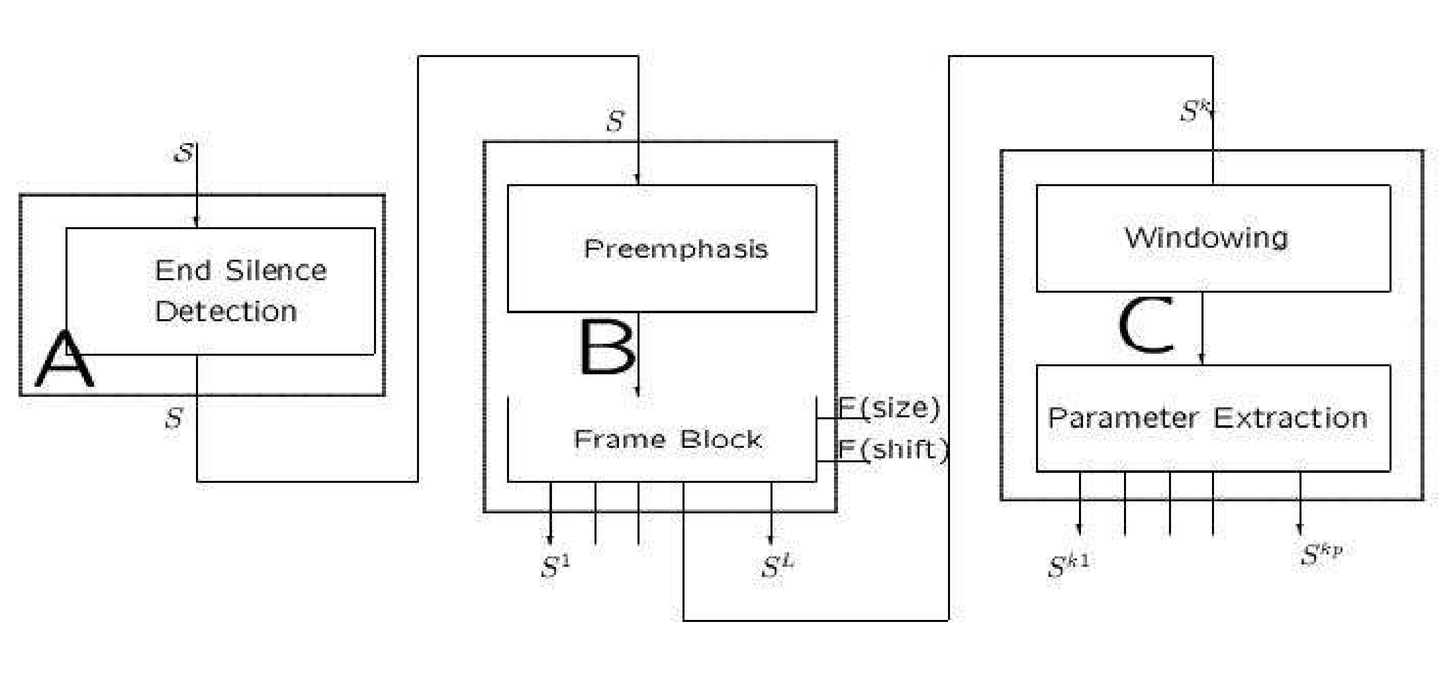
Computing MFCC - Full View



Intersting things to do

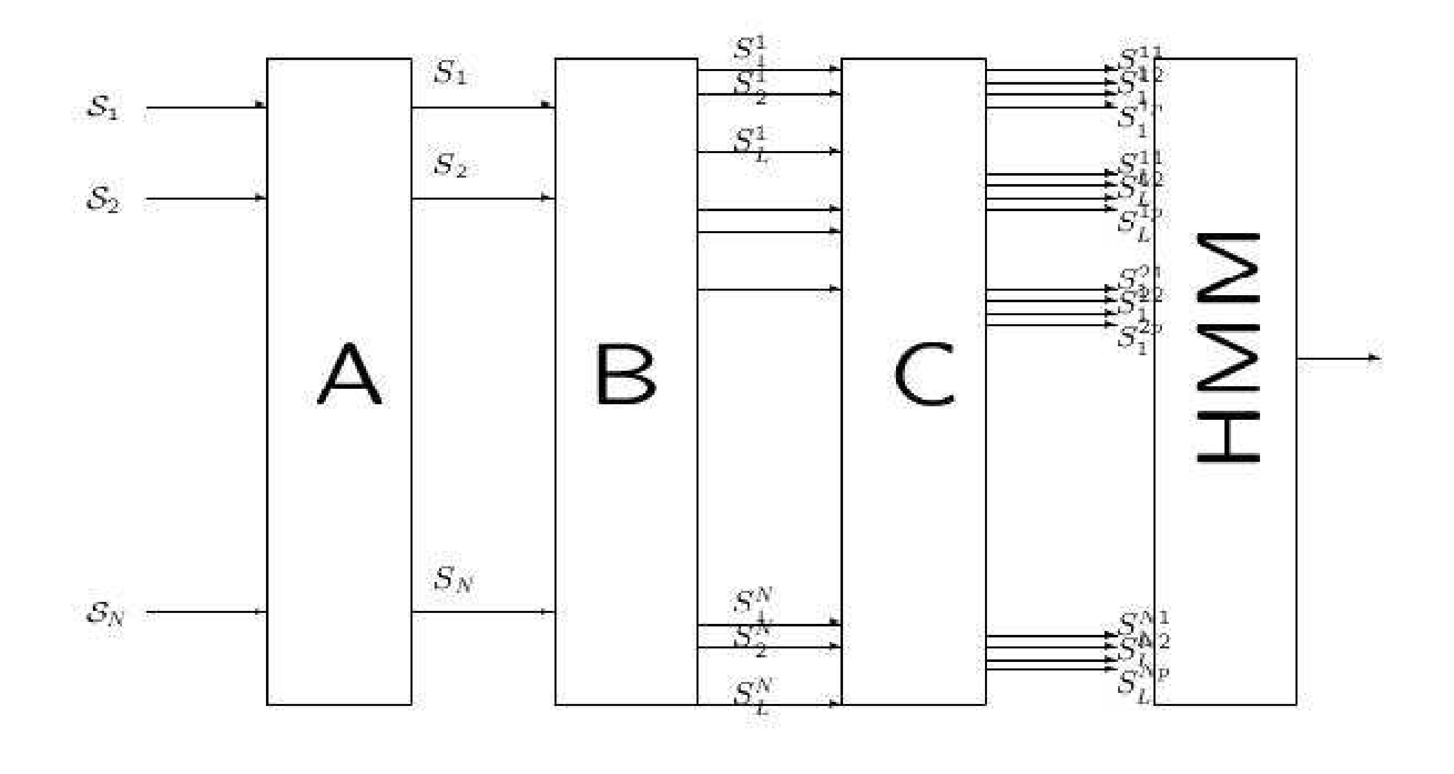
- What should be the choice of mel filter bank if we have to compute MFCC of a resampled speech?
- What is the effect of Noise-in-Speech on MFCC Parameters?

Speech to Parameter Extraction



- Speech converted into pre determined parameters
- preprocessing happening before parameter extraction.

Speech Modeling using HMMs



The parameters extracted (for all the speech files corresponding to the same category – word) are modeled using HMMs.

Hidden Markov Model - Theory

HMMs are doubly stochastic models and is completely defined by a three-tuple $(A, B, \pi) = \lambda$.

- A is the transition probability
- B is the observation probability and
- π is the initial state distribution.

Let

- $\{x_1, \cdots x_N\}$ represents set of states
- $\{y_1, \cdots, y_M\}$ represents set of observations
- X_t (state) and Y_t (observation) denote the random variable at time t.

Hidden Markov Model - Theory(2)

Transition probability $A = \{a_{ij}\}$;

$$a_{ij} = P(X_{t+1} = x_j | X_t = x_i)$$

- describes the probability of the moving from state i (x_i) to j (x_j).
- a_{ii} denotes the probability of staying in the same state i.
- In speech, $a_{ij} \neq 0$ for $j = i$ and $j = i + 1$.

Hidden Markov Model - Theory(3)

Observation probability $B = \{b_i\};$

$$b_i(y_k) = P(Y_t = y_k | X_t = x_i)$$

- describes the probability of the k^{th} observation (y_k) being in state i (x_i)
- B is considered to be a Gaussian with a certain μ and σ^2

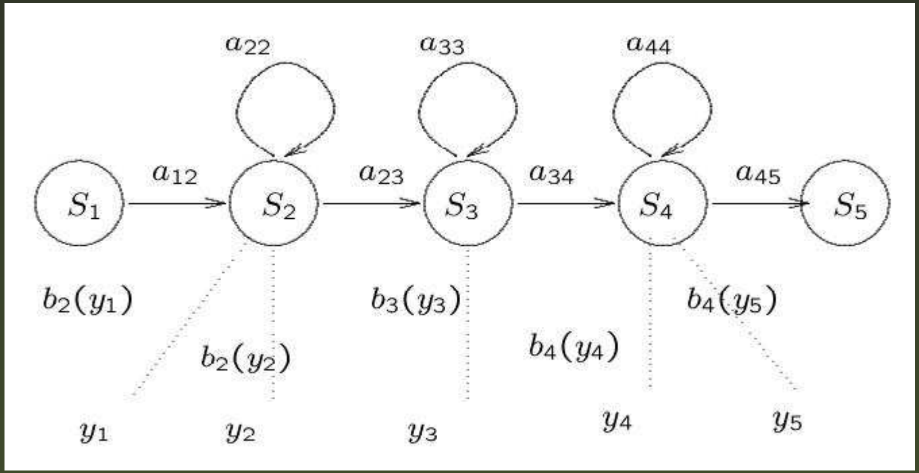
Hidden Markov Model - Theory(4)

Initial state distribution $\pi = \{\pi_i\};$

$$\pi_i = P(X_0 = x_i)$$

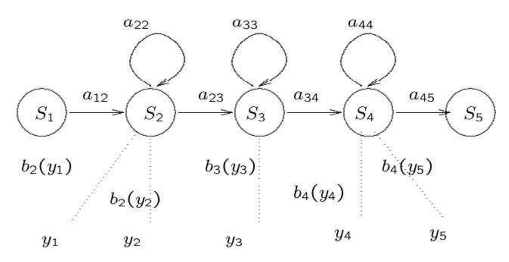
- describes the probability of the initial observation being in the i^{th} state.
- In speech, $\pi_1 = 1$.

HMM - Simplified



- As a generator of vector sequences (y)
- a finite state machine – changes state – each time t – state j is entered, an acoustic speech vector y_t is generated with probability density $b_j(y_t)$
- transition – state i to j is probabilistic governed by discrete probability a_{ij}

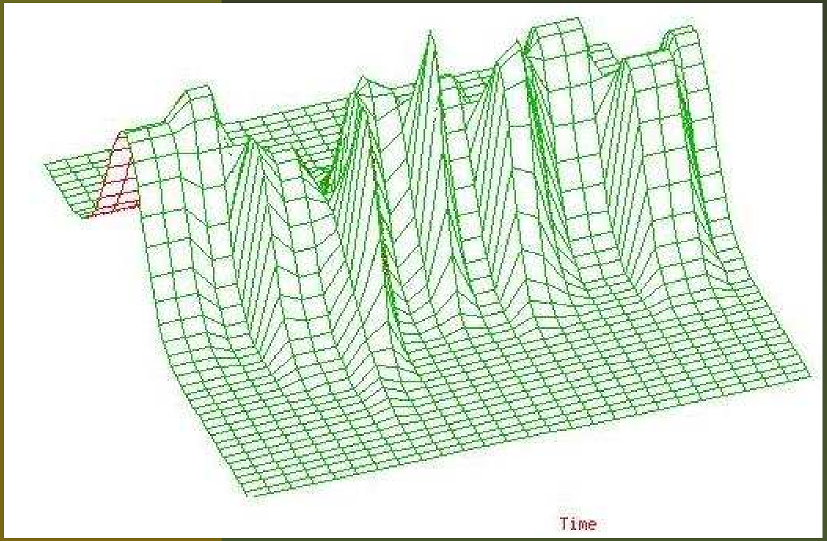
HMM - Simplified(2)



model moves through the states X = 1, 2, 2, 3, 4, 4, 5 to generate the sequence $y_1 \cdots y_5$.

- The joint probability of a vector sequence Y and state sequence X given a model λ $P(Y; X|\lambda) = a_{12}b_2(y_1)\ a_{22}b_2(y_2)\ a_{23}b_3(y_3)\ a_{34}b_4(y_4)\ a_{44}b_4(y_5)$
- Observation sequence is known, the state sequence and the emission probabilities are hidden, that is why it is called doubly hidden Markov Model.

HMM - Visualization of a parameter



1. thickness of Gaussian (along time axis) captures transition probability a_{ii}
2. shape of Gaussian in each state, μ and σ^2, captures observation probability (b_i) of each state
3. The figure as a whole captures the HMM

Three Basic Problems for HMMs

Given a sequence of T observations $Y = y_1, \cdots, y_T$. We can primarily think of solving three kinds of problems.

1. Find $P(Y|\lambda)$: the probability of the observations given the model. *Forward-Backward* procedure used. (Recognition / Decoding)
2. Find the most likely (hidden) state sequence (X) given the model (λ) and observations (Y). *Viterbi Algorithm* used.
3. Given observations (Y), adjust $\lambda = (A, B, \pi)$ to maximize $P(Y|\lambda)$. *Forward-Backward* or *Baum Welch algorithm* used. (Training)

Isolated Word Training

- Predefined set of W words, number of states S
- For each $w \in W$
  - Collect representative speech samples (say $N = 100$ utterances to cover variability in acoustic, people)
  - For each utterance $u \in U$ $(u_1, \cdots, u_N)$
    1. Pre-process (silence, frames, pre-emphasis)
    2. Extract parameters (eg MFCC) for each frame (Y; observations)
  - Initialize HMM λ_w^{init} (for each $s \in S$, for each parameter)

Isolated Word Training(2)

- For each $w \in W$
  - For each utterance $u \in U$ $(u_1, \cdots, u_N)$
    1. Find best state sequence X (using *Viterbi*, λ_{init} and Y)
  - Update HMM λ_w
  - Find the optimal λ_w such that $P(Y|\lambda)$ is maximized. (Using *Baum Welch algorithm*)

Isolated Word Recognition

- Read λ_w, $w \in W$
- Given a speech sample (say T); Pre-process (End silence, frames); Extract parameters (Y_T; observations)
- For each HMM λ_w, $w \in W$
  1. Find best state sequence (using *Viterbi*, λ_w and Y_T)
  2. Calculate $P_w = P(Y_T|\lambda_w)$ for the best state sequence
- For each $w \in W$
  1. Sort P_w in non-increasing order

Isolated Word Recognition(2)

- Select word α as the recognized word if P_α is the first entry in the sorted list, namely,

$$P_\alpha > P_w \ \forall \ \ w \in W \ \ \text{and} \ \ w \neq \alpha$$

How is this used: Speech Solutions?

- Self Help Applications
  Banking, Insurance
- Automated Speech based Transactions
  Indian Railway; Banking; Mandi Bhav
- Speech Analytics
  Voice Call center; customer satisfaction index
- Multilingual Video, Audio Annotation
  searchable video
- Speech Training
  Accent training, Music learning
- Speech Biometrics
  Most speech applications

Speech Solutions

- Some Examples
- How about India? Do we need Speech Solutions?

Indian Challenge

- Many languages (22 official)
- Very many dialects
- Noisy telephone channels
- Non-availability of speech corpus (to train - acoustic models, language model)
- Use of more than one language in the same sentence

But, India needs speech solution most

- Large illiteracy (only spoken Indian language)
- Speech is the most natural interaction channel

Several Examples ... Later ... Synthesis

Speech Synthesis

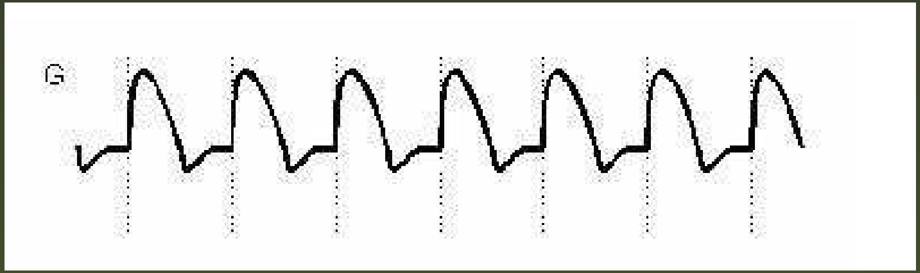
Speech Synthesis - Trade-off

Trade-offs in development of speech synthesizers are based on conflicting demands of

- maximize
  - quality of speech,
- minimize
  - memory space,
  - algorithmic complexity, and computation speed.

Voiced Speech - Pitch?

- For voiced sounds the vocal cords vibrate *(they interrupt the air stream)* and produce a quasi-periodic pressure wave called **pitch impulses**.



- The frequency of the pressure signal is the **pitch frequency**.
- Pitch is the part of the voice signal that defines the speech melody *(When we speak with a constant pitch frequency, the speech sounds monotonous but in normal cases a permanent change of the frequency ensues.)*

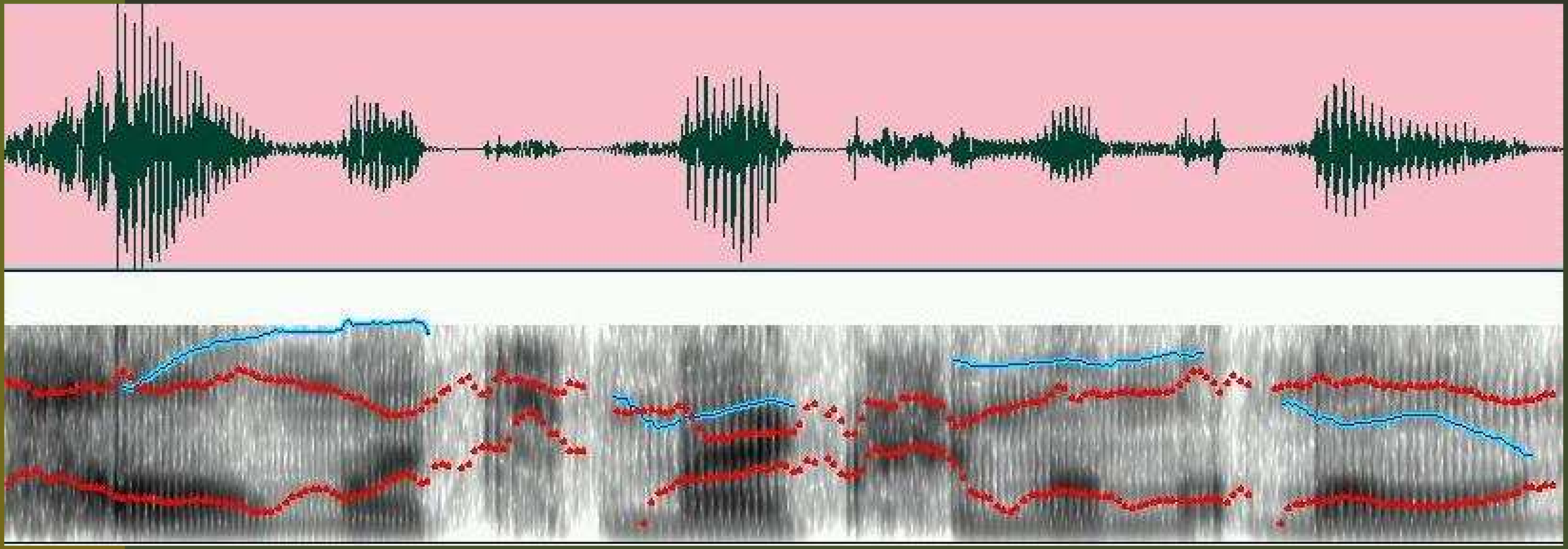
Voiced Speech - Formant?

- The vocal tract can be viewed as an acoustic tube of varying diameter. We can abstract from its curvature and divide it into cylindrical sections of equal width.
- Depending on the shape of the acoustic tube (mainly influenced by tongue position), a sound wave traveling through it will be reflected in a certain way so that interferences will generate resonances at certain frequencies. These resonances are called formants.
- The location of formants largely determine the speech sound that is heard.

Pitch and Formant in Spectrogram

- A wideband spectrogram of the speech signal (spectral analysis on a 15 ms section of the waveform using a 125 Hz bandwidth analysis filter) shows up formants.
- A narrow band spectrogram (spectral analysis on 50 ms section of speech waveform using a 40 Hz bandwidth narrow analysis filter) of speech signal shows up pitch.
- The narrow bandwidth of the analysis filter picks up the individual spectral harmonics corresponding to pitch. These are seen a horizontal lines in the spectrogram.

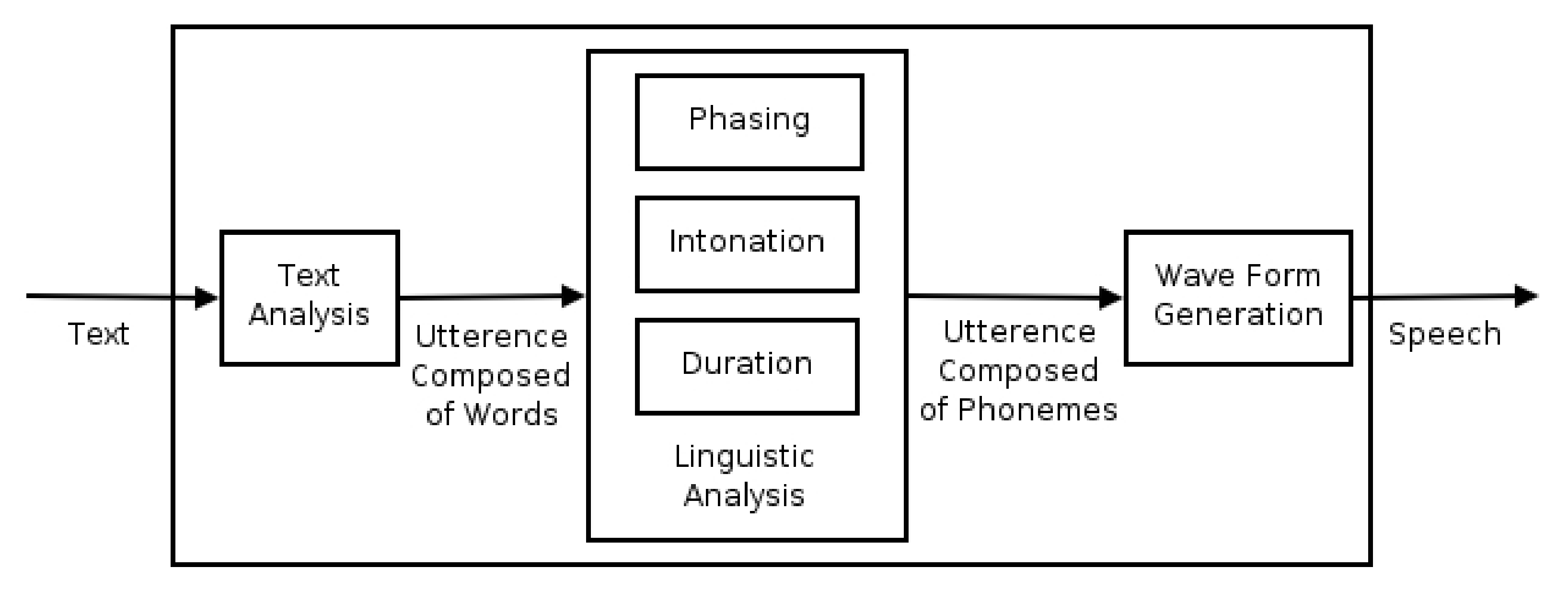
Formant and Pitch in Spectrogram



/How much maximum amount/

- Formant (in Red) - high energy in certain frequencies corresponding to a sound
- Pitch (in Blue) - repetitive pulse frequency

Speech Synthesis Flow



Text to Speech Synthesizer

- Text normalization (1234; "one two three four"; "one thousand two hundred and thirty four"; "twelve thirty four")
- Grapheme to Phoneme Conversion (usually through a look up dictionary and through a set of rules especially for out of vocabulary words)
- Synthesizing Phonemes; either using predefined speech files or generating them on the fly using some apriori information *(gives intelligibility)*
- Introducing prosody or pitch variations on the synthesized speech *(gives naturalness)*

Types of Synthesizer

- Formant synthesis (Rule based synthesis)
- Articulatory synthesis
- HMM-based synthesis (frequency spectrum, pitch, prosody of speech modeled simultaneously by HMM)
- Concatenative synthesis
  - Unit selection synthesis
  - Diphone synthesis
  - Domain-specific synthesis

Formant based Synthesizer

- Uses information *(acoustic model)* about the resonances of the human vocal tract - formants - as the primary source material and requires **no speech database** at runtime.
- Synthesized voice output is derived by varying the *(fundamental frequency)* pitch, spectral components, voicing and noise levels over time to create a waveform that follows the formants of natural speech.
- The output of such systems is *generally* robotic-sounding and would not be mistaken for a real human voice.

Concatenative Synthesizer

- Concatenative synthesizers concatenate *(join)* speech units using stored waveform.
- This requires (a) large memories and (b) good algorithm to smooth the transitions; but it can yield good quality speech.
- Systems differ in the size of the stored speech units.
- A system using phones or diphones provides the largest output range, but may lack clarity; on the other hand, a system using entire words or sentences allows for high-quality output but the range is limited *(useful for domain specific applications)*

Digress: Voice Grafting

Grafting refers to implanting some characteristics of a speech signal of one voice onto another.

- Mahatma Gandhi spoken speech ($S_1(z)V_1(z)$)
- Amateur same sentence ($S_2(z)V_2(z)$)
- Crafted ($S_1(z)V_2(z)$)

Speech Biometrics

Speaker Verification: Overview

- Is the process of verifying the claimed identity of a registered speaker using his voice characteristics.
- The speaker needs to enroll before using the system.
- During enrollment, the speaker speaks a given set of utterances, using which the systems builds statistical models representing the speaker's voice.
- A user claims he is X. Speaks a pass-phrase. The system gives a binary output YES (accept claimed identity) | NO.

Need for threshold to be able to say Yes or No. AttMon

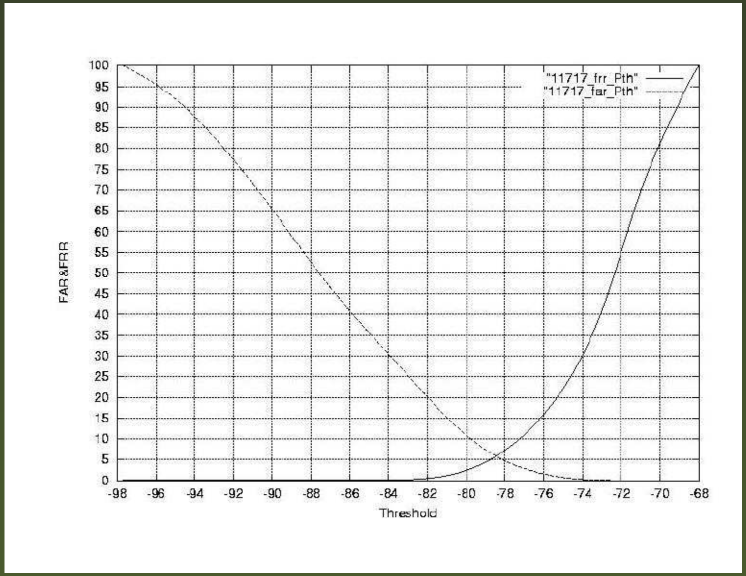
Types of Speaker Verification

1. **Fixed Phrase** – pre-determined phrase used for verification
2. **Fixed Vocabulary** – verification more flexible and practical; training and testing materials for a speaker are generated based on words of a fixed vocabulary
3. **Flexible Vocabulary** – a general set of subword phone models is created during speaker model training
4. **Text-Independent** – user is not constrained to say fixed or prompted phrases

Clearly, both complexity and security increases as we go from fixed phrase to text-independent.

Speaker Verification Terms

- FAR - False Acceptance Ratio
  The percentage of **incorrect successful** verifications.
- FRR - False Rejection Ratio
  The percentage of **incorrect failed** verifications.
- EER - Equal Error Rate
  The value at which FAR equals FRR



Our Speaker Verification System [1]

- Uses the state-of-the-art speaker verification engine tuned for telephone speech

Our Speaker Verification System [1]

- Uses the state-of-the-art speaker verification engine tuned for telephone speech
- Easily customizable and configurable and can work even for a large company setup with thousands of employees. It can easily be integrated into any existing employee database of the company.

Our Speaker Verification System [1]

- Uses the state-of-the-art speaker verification engine tuned for telephone speech
- Easily customizable and configurable and can work even for a large company setup with thousands of employees. It can easily be integrated into any existing employee database of the company.
- Functional for attendance monitoring at several locations of TCS

Our Speaker Verification System [2]

- System accessible over the telephone line (EPBAX) using telephony card interface.

Our Speaker Verification System [2]

- System accessible over the telephone line (EPBAX) using telephony card interface.
- Speaker verification engine, conceived and engineered in-house.

Our Speaker Verification System [2]

- System accessible over the telephone line (EPBAX) using telephony card interface.
- Speaker verification engine, conceived and engineered in-house.
- Select speech feature set which captures the identity of the speaker makes it very robust with very low FRR and FAR

Performance

	T_1	T_2
FAR	6.47%	1.75%
FRR	0.95%	10.90%

- Threshold (T_1) was chosen to be such that the FRR was close to 0% (pass all) and T_2 was chosen so that FRR was approximately 10%.
- Experiments carried out on a set of 15 speakers. Imposter's aware of the pass phrase (i.e. skilled forgery)
- 25 Parameters; Continuous HMM models used
- Ported and tested on BREW SDK2.0 simulator

Thank You

- Queries?
- Comments
- Suggestions?

SunilKumar.Kopparapu@TCS.Com
TCS Innovation Lab - Mumbai
Tata Consultancy Services Limited
Yantra Park, Thane (West), India.