Speech Signal Processing - From Lab to Land

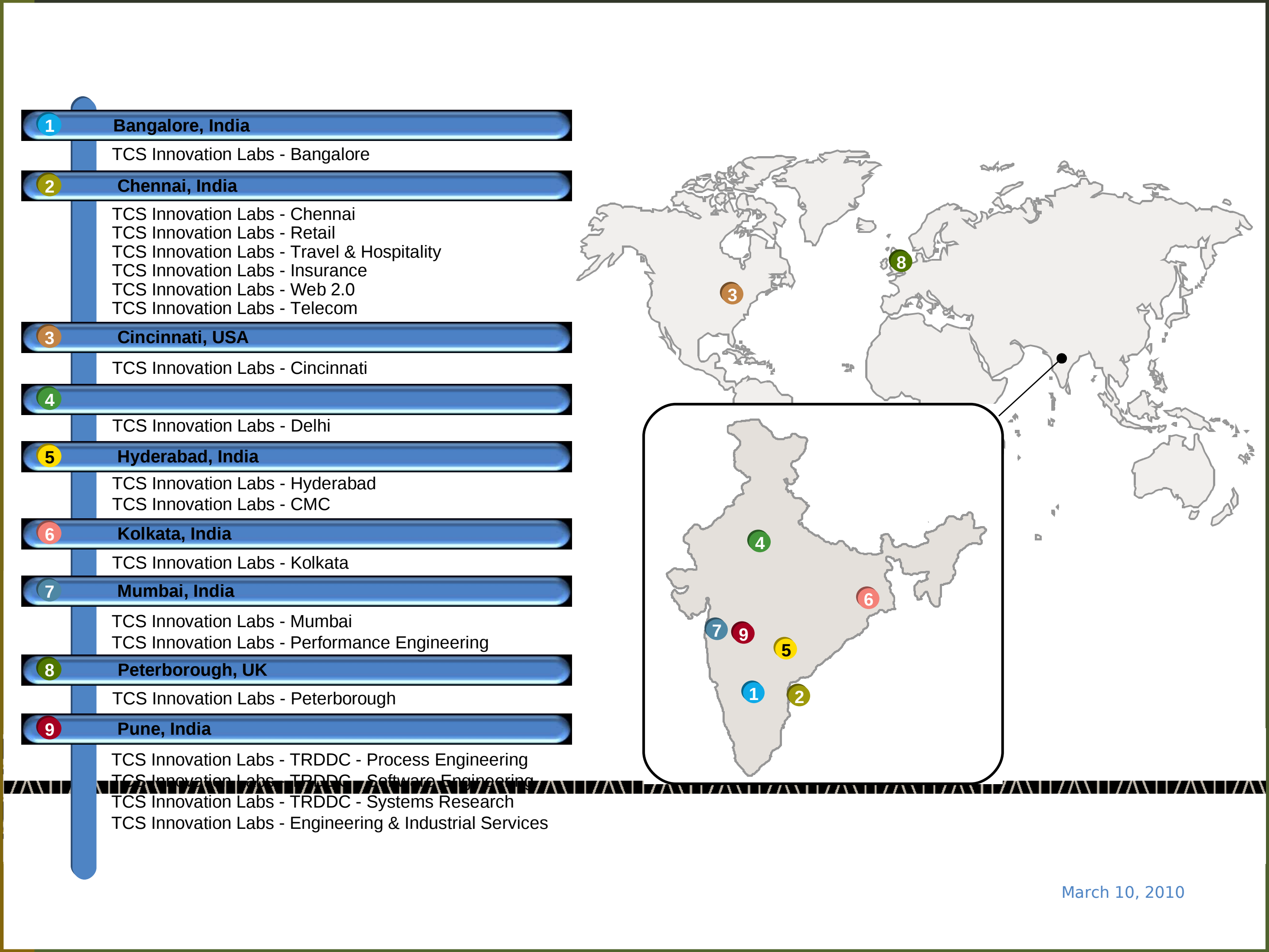
Sunil Kumar Kopparapu

SunilKumar.Kopparapu@TCS.COM

Speech and Natural Language Processing Group
TCS Innovation Labs - Mumbai
Yantra Park, Thane (West), Maharastra

March 2010

TCS Innovations Lab



1 **Bangalore, India**

TCS Innovation Labs - Bangalore

2 **Chennai, India**

TCS Innovation Labs - Chennai
TCS Innovation Labs - Retail
TCS Innovation Labs - Travel & Hospitality
TCS Innovation Labs - Insurance
TCS Innovation Labs - Web 2.0
TCS Innovation Labs - Telecom

3 **Cincinnati, USA**

TCS Innovation Labs - Cincinnati

4

TCS Innovation Labs - Delhi

5 **Hyderabad, India**

TCS Innovation Labs - Hyderabad
TCS Innovation Labs - CMC

6 **Kolkata, India**

TCS Innovation Labs - Kolkata

7 **Mumbai, India**

TCS Innovation Labs - Mumbai
TCS Innovation Labs - Performance Engineering

8 **Peterborough, UK**

TCS Innovation Labs - Peterborough

9 **Pune, India**

TCS Innovation Labs - TRDDC - Process Engineering
TCS Innovation Labs - TRDDC - Software Engineering
TCS Innovation Labs - TRDDC - Systems Research
TCS Innovation Labs - Engineering & Industrial Services

March 10, 2010

TCS Innovations Labs - Mumbai

- Research and Development
  - Speech,
  - Script, Image
  - Natural Language Processing (Information Retrieval)
  - Wireless sensor networks
- Innovative Mobile Applications
  - mKRISHI (agro advisory system for farmers)
  - Chit Chat (audio based social networking)

Will try and stick to speech ... meanwhile a short Video

Speech - One Slide Intro!

- 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

Origins of speech? unknown and subject to much debate

Some Basics

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 Signal Processing

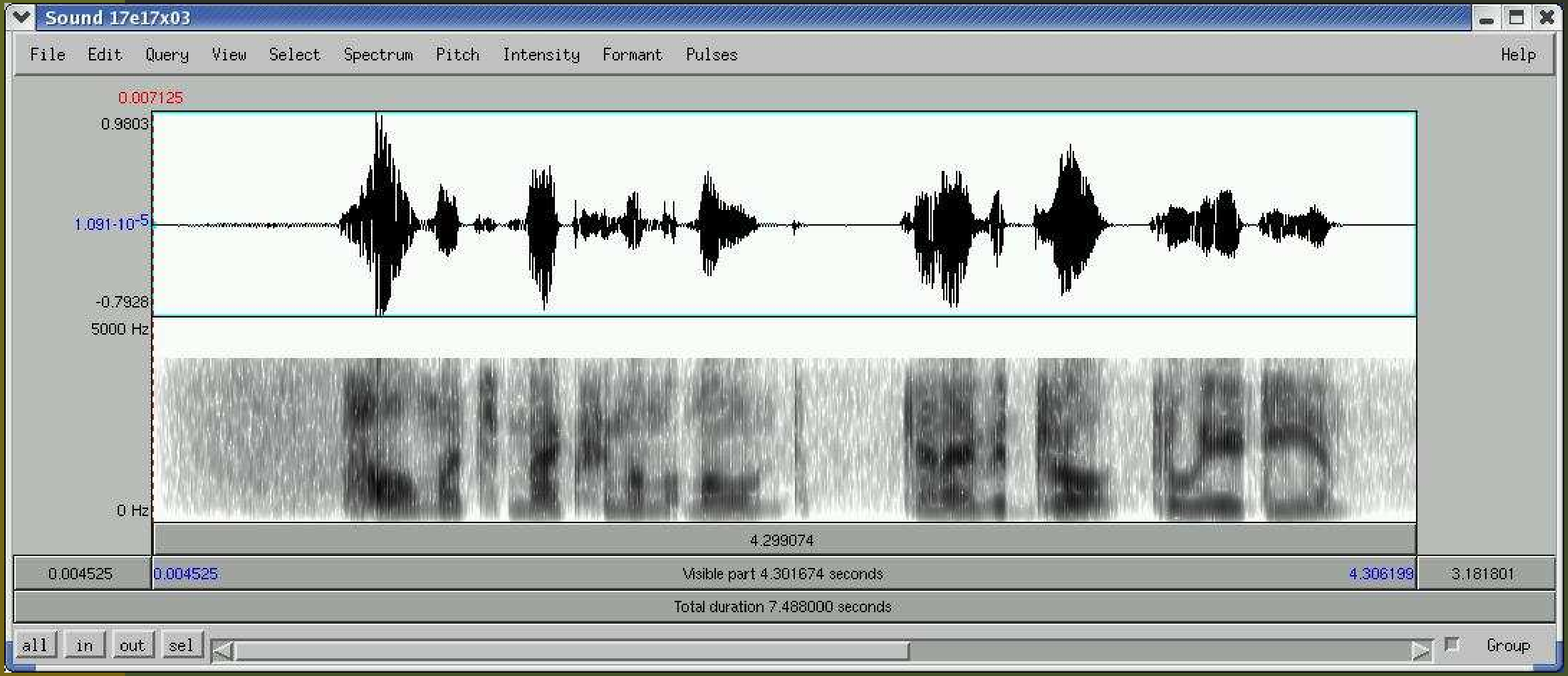
Speech signal processing refers to acquisition, manipulation, storage, transfer and output of human utterances by a computer. The main goals are recognition, synthesis and ~~speech compression~~.

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

Speech Recognition Example

- **Input** Speech, 16kHz, 8 bit
- **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 dictionary — 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*

To a Speech recognition Engine ...



which has to be recognized as ....
How much maximum amount can I withdraw through ATM

Speaker Recognition Example

Recognition

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

Response: *Sunil*

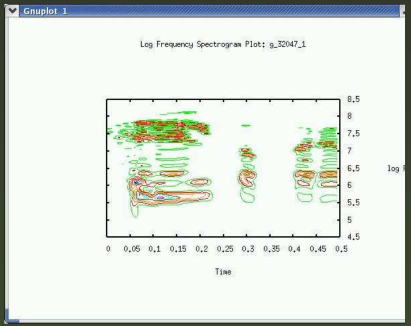
Verification

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

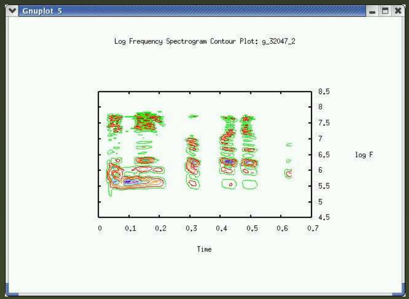
Response: *Yes/No*

Speech Reco - Speaker Verification

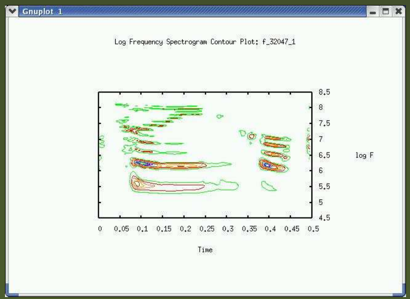
Recognize *Sunil Kopparapu*



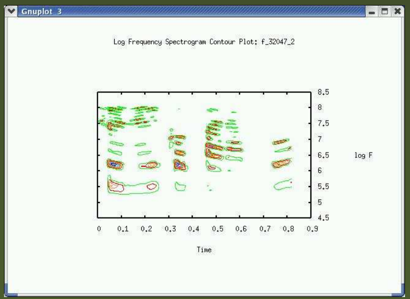
X_1



X_2



Y



Z

Visual: Log-Frequency Contour; Parameters need to be extracted accordingly

Speech Synthesis Example

Speech Synthesis 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.

Digress ...

Lab Objective

- Utilize the speech signal processing knowledge to research and build in-house a state of the art speech recognition, synthesis, and speaker verification and Speaker identification engine
- Understand and explore aspects of speech recognition technology which can be used to enhance the performance in actual use
- Utilize in-house engine to custom build applications, rather than use an off the shelf speech recognition system from a third party vendor as a black-box.
- Speech solution for Indian masses

Indian Challenge

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

But, India needs speech solution most

- Large illiteracy
- Speech is the most natural interaction channel

Where is the use?

Speech Solutions? Towards Land

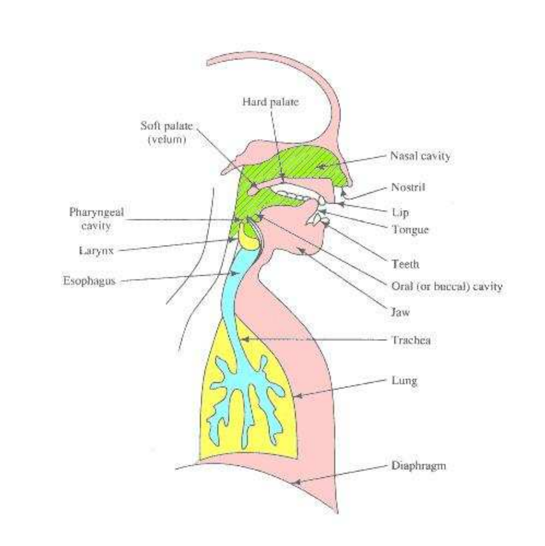
- Self Help Applications
  Banking, Insurance
- Automated Speech based Transactions
  Indian Railway; Banking; Mandi Bhav
- Speech Analytics
  Voice Call center; customer satisfaction index
- Multilingual Video, Audio Annotation
  searchable video
- Speech Training
  Accent training, Music learning
-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

Speech Processing

Two school of thoughts

- Speech Production
  We should model the source of speech production because that is the origin of speech sound.
- Speech Perception
  But 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 cavity** (throat), **oral cavity** (mouth), **nasal cavity** (nose).
- Speech is produced by a cooperation of
  - lungs, glottis (with vocal cords) and
  - articulation tract (mouth and nose)

Speech Production - Common Terms

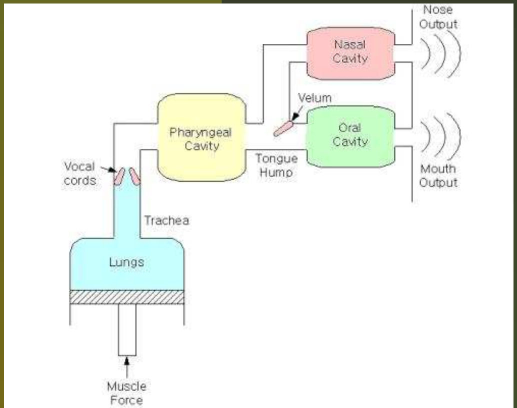
- Pharyngeal and oral cavities are grouped into one unit and called **vocal tract** and nasal cavity is often called the **nasal tract**.
- **vocal tract** begins at the output of the **larynx** (vocal cords, or glottis) and terminates at the input to the **lips**. The **nasal tract** begins at the **velum** (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.

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

Speech Production Schematic



Mechanism for creating speech a simplified representation

Source-Filter Model

- **Source:** The lungs (and the 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 - 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 (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 (-12 dB/octave due to excitation source; +6 dB/octave due to the radiation compensation) 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.

How does one use this?

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 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

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

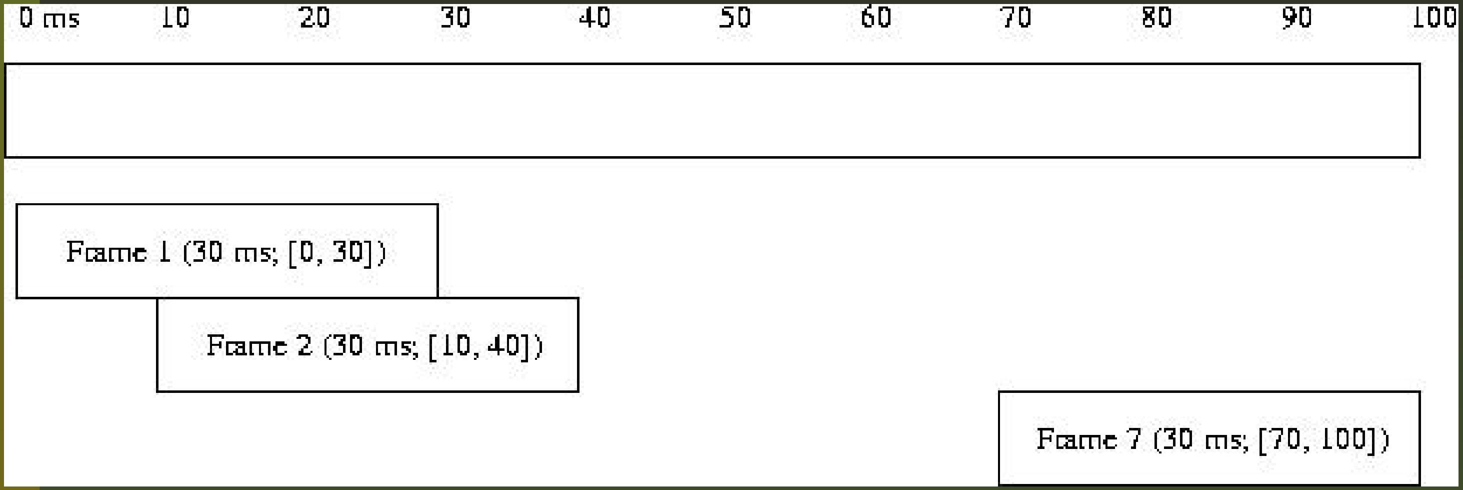
Speech Recognition Preprocessing(1)

- Speech is non-stationary. Meaning, the statistics of the speech signal change with time.
- To develop a statistical model for speech we need to consider smaller *portions* of speech.
- Typically 10-20 ms of speech called speech frames where the signal can be considered to be stationary *a key assumption in all current speech recognition systems*

Speech Recognition Preprocessing(2)

- Dividing the speech signal into frames,
- Removing non-speech signal,
- Pre-emphasizing the signal to spectrally flatten the signal 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 and
- Tapering (Windowing) the frames (Hamming window) to minimize signal discontinuities at the beginning and end of the frame.

Speech Analysis Frames



- If Speech signal duration is T (= 100 ms), then
- total total number of frames is given by (T - F_{size})/F_{shift} = (100 - 30)/10 = 7,
- provided frame size (F_{size}): 30 ms *typical for 8 kHz speech signal* and frame shift (F_{shift}): 10 ms *typically,* $F_{shift} = F_{size}$*/3,* F_{size}*/2*

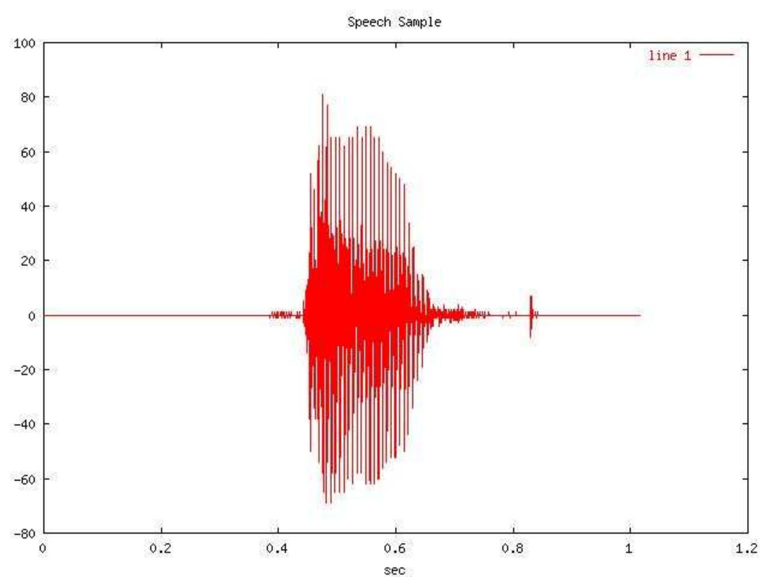
End Silence Detection

- 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 $\mathcal{A}^1_{max}, \mathcal{A}^1_{max}, \cdots \mathcal{A}^N_{max}$ (maximum amplitude in each frame). Compute,

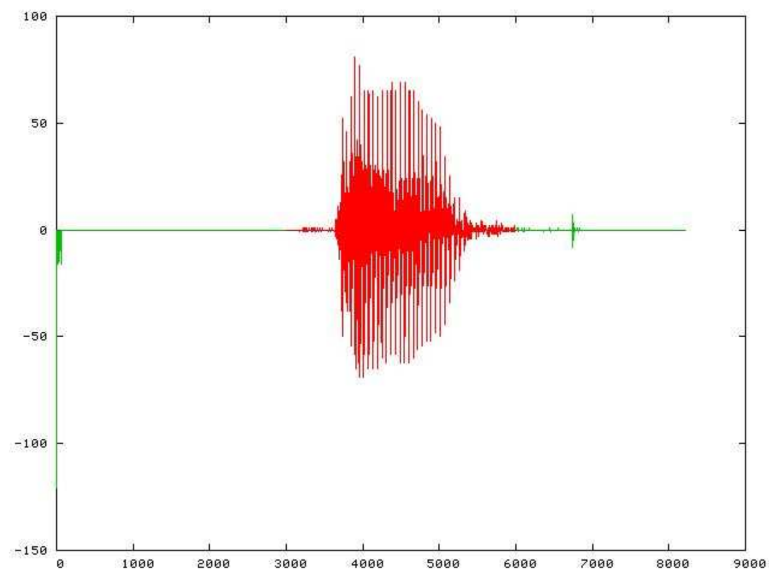
$$\mathcal{T} : \frac{(\mu + 2\sigma^2) + \mathcal{A}_{max}}{15}$$

- Frame i is a speech frame if $(\mathcal{A}^i_{max} > \mathcal{T})$. 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/

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

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

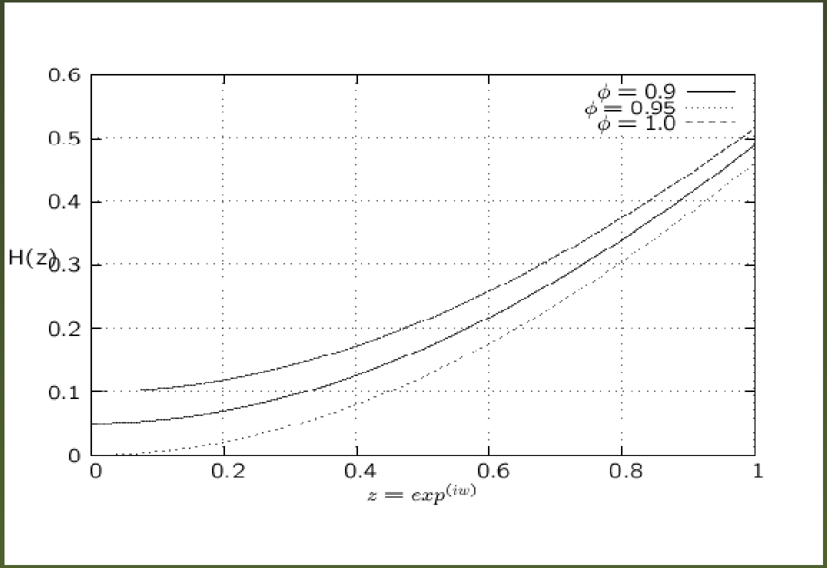
- In the time domain this is equivalent to the difference equation

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

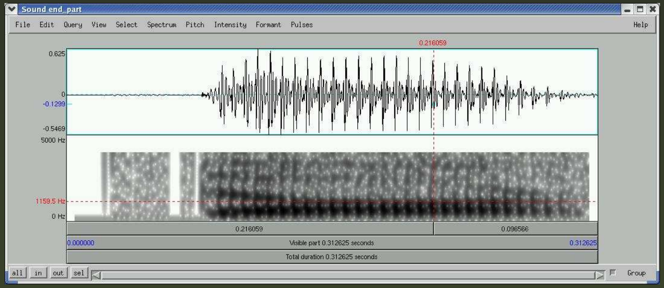
Pre-Emphasis (2)

- 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.

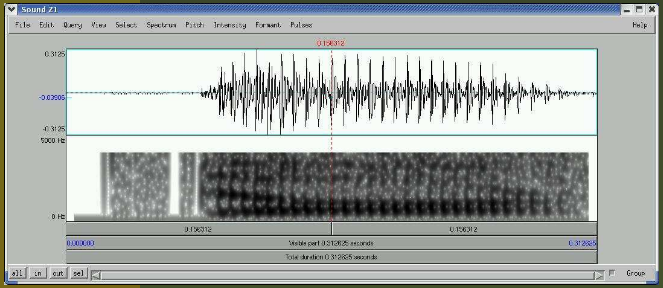
Note: The overall effect is to emphasize the high frequency content and deemphasizing the low frequency content. This is done to compensate for the attenuation caused by the radiation from the lips.



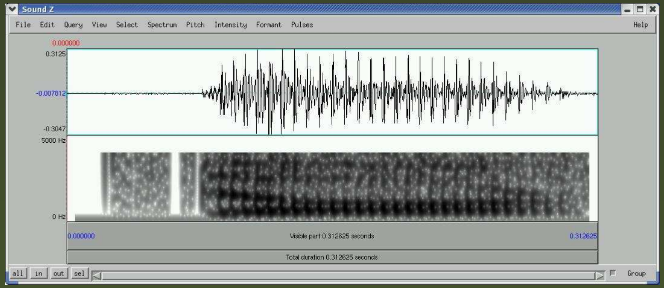
Ex: Pre-Emphasis ($\phi = 1.0, 0.9$)


Sound end_part
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses
Help
0.216059
0.625
0
-0.1299
-0.5469
5000 Hz
1159.5 Hz
0 Hz
0.216059
0.096566
0.000000
Visible part 0.312625 seconds
0.312625
Total duration 0.312625 seconds
all in out sel
Group

/Dark/

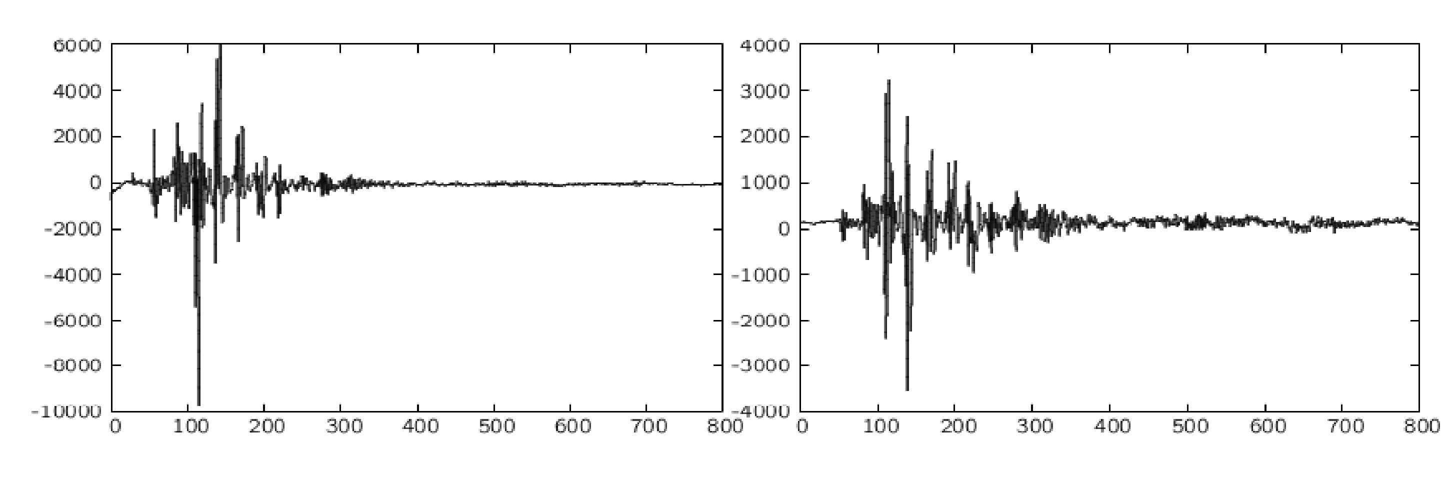

Sound Z1
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses
Help
0.156312
0.3125
-0.03906
-0.3125
5000 Hz
0 Hz
0.156312
0.156312
0.000000
Visible part 0.312625 seconds
0.312625
Total duration 0.312625 seconds
all in out sel
Group

/Dark $\phi = 0.9$/


Sound Z
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses
Help
0.000000
0.3125
-0.007812
-0.3047
5000 Hz
0 Hz
0.000000
Visible part 0.312625 seconds
0.312625
Total duration 0.312625 seconds
all in out sel
Group

/Dark $\phi = 1.0$/

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.

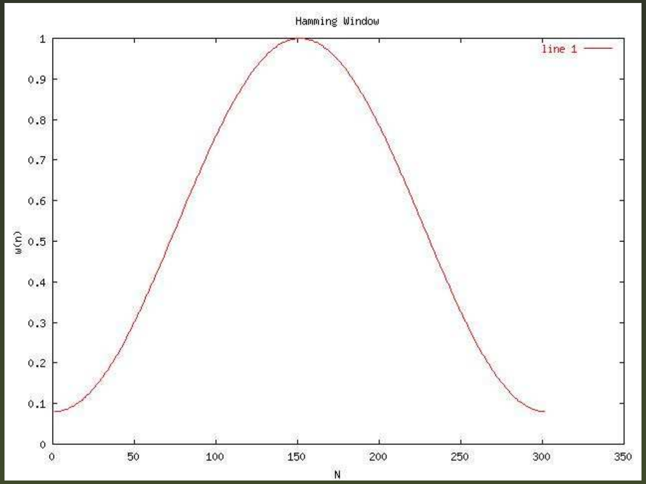
$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \le n \le N-1$$

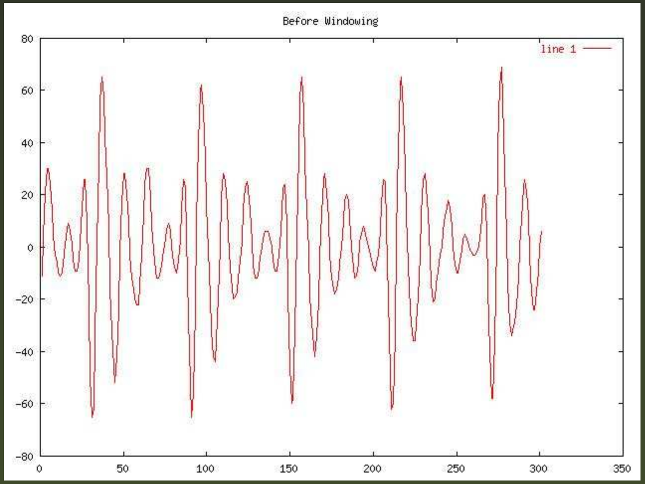
- The windowed signal ($s_w(n)$) is obtained as

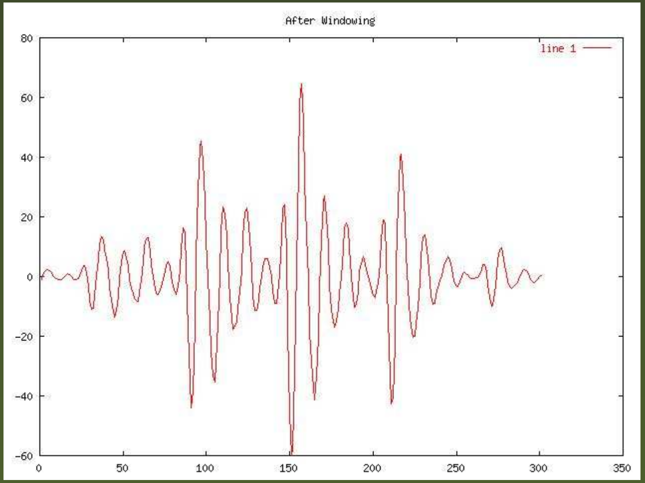
$$s_w(n) = s_f(n)w(n)$$

where $s_f(n)$ is the speech frame

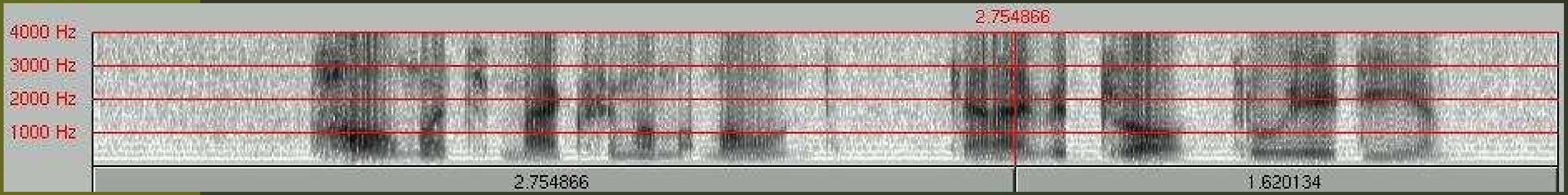
Windowing: Example (Time)



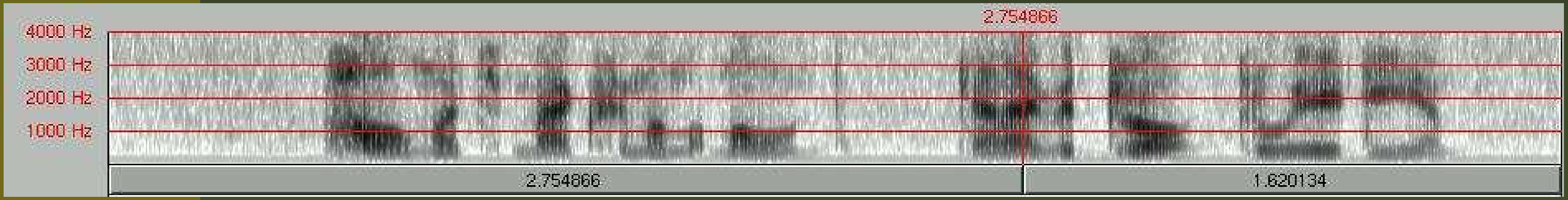




Windowing: Example (Spectrogram)

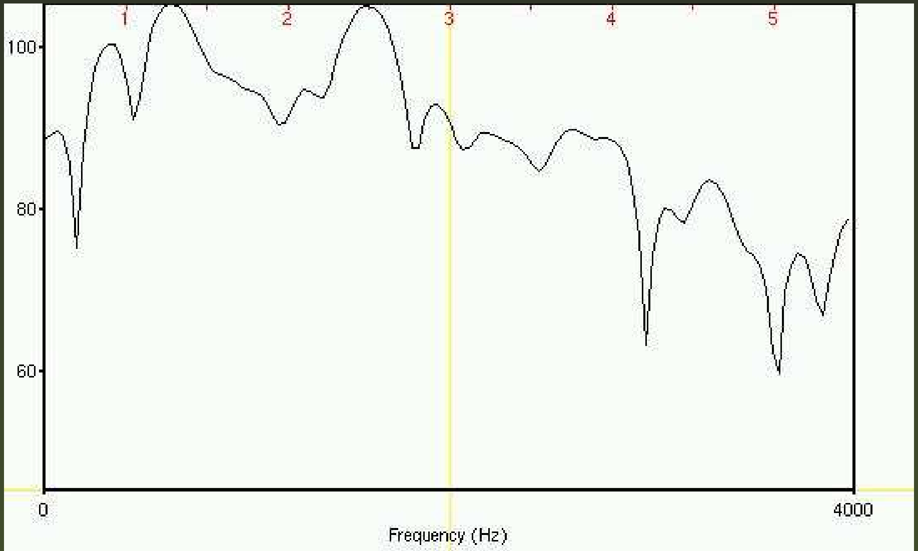


No Windowing

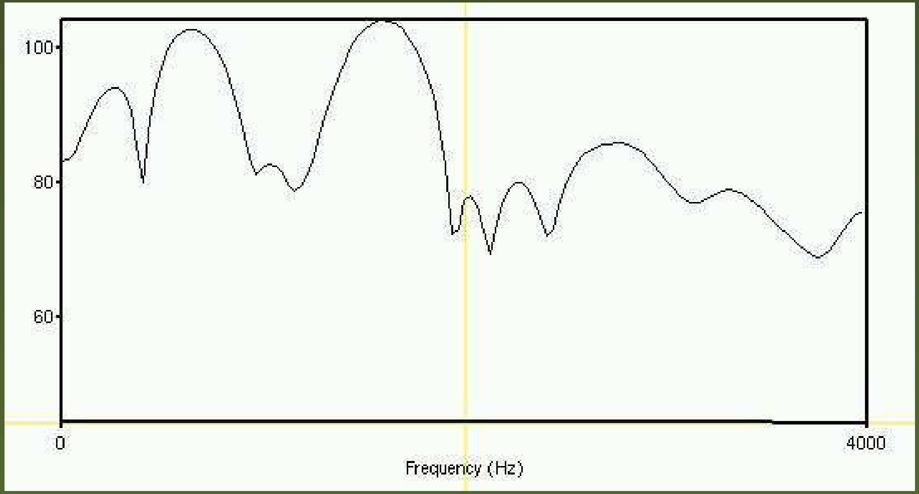


With Windowing

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. These large number of samples by themselves do not *e*xplicitly exhibit information that can be used to build statistical models.
- It is typical to extract a few (typically 30-40) parameters, by suitably processing the signal, from the large number of speech samples. These fewer parameters (related to human speech generation or perception model) are *r*epresentative of all the 80000 speech samples.

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

LPC: Speech Parameters

- Speech can be 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 chosen to satisfy $\frac{f_s}{1000} + 2$ where f_s is the sampling frequency of the signal s_f.
- For a 8 kHz speech signal typically p is chosen as 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

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

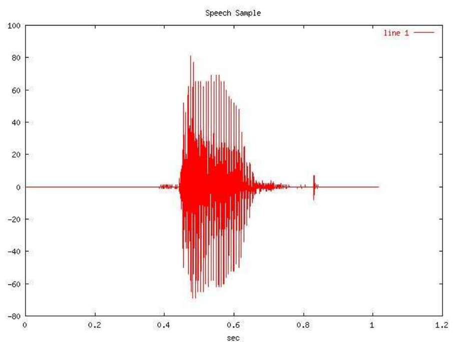
  where $e(n)$ is the excitation source signal and $h(n)$ is the vocal tract system model, then in the cepstral domain we have

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

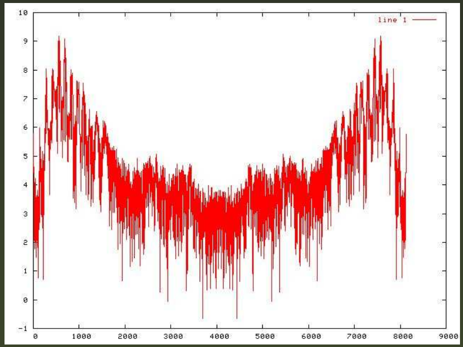
- The ability of the cepstrum of a frame of speech to separate the excitation source from the vocal tract system model is often exploited in speech signal processing.

Source-Signal Separation in Cepstral

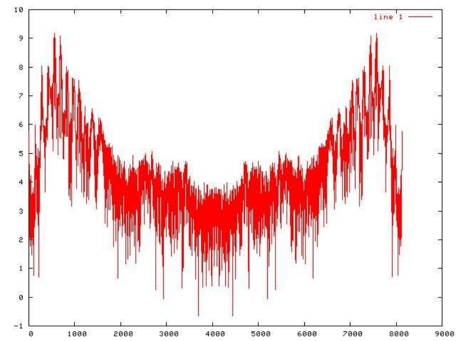
$e(n) * h(n) = s(n)$



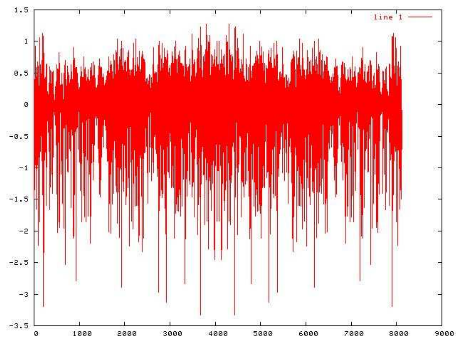
$\leftrightarrow$



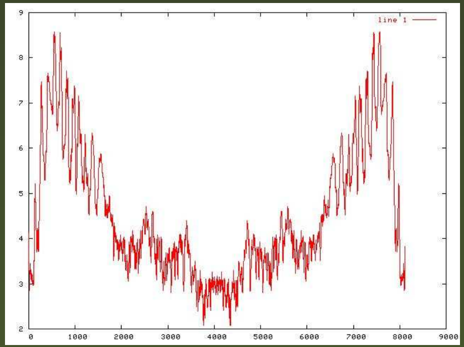
$C_s(\omega)$



$=$



$+$



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

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

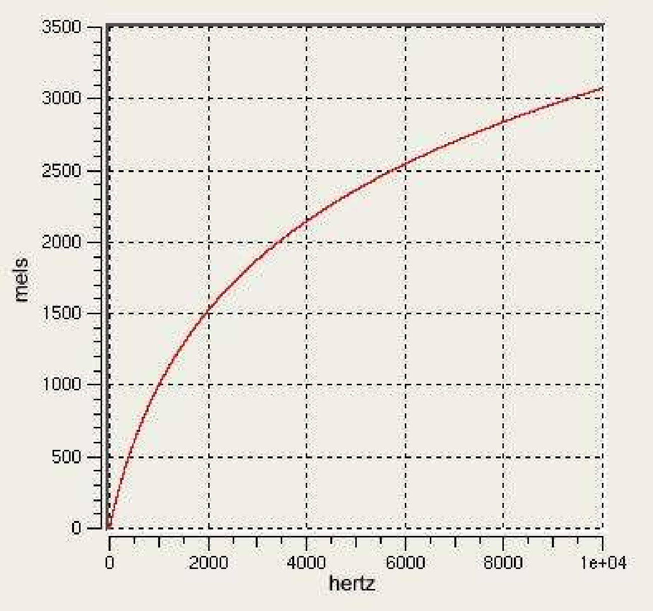
$$\begin{aligned} c_1 &= a_1 \\ c_n &= a_n + \sum_{m=1}^{n-1} \left(\frac{m}{n}\right) a_m c_{n-m} \quad \text{for} \quad 2 \leq n \leq p \\ c_n &= \sum_{m=1}^{p} \left(\frac{n-m}{n}\right) a_m c_{n-m} \quad \text{for} \quad n > p \end{aligned}$$

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

Parameter: MFCC, ΔMFCC (1)

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

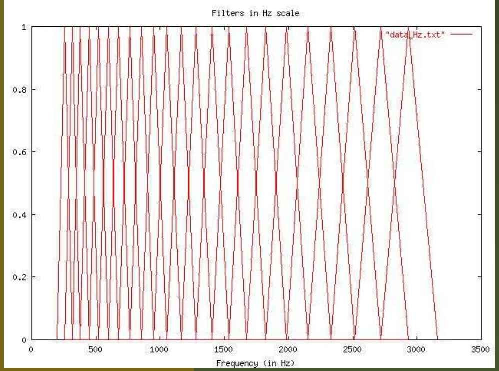
$$Mel(f) = 2595 \log_{10}\left\{1 + \frac{f}{700}\right\}$$



Parameter: MFCC, ΔMFCC (2)

24 mel *(filter)* channels (in f scale *(in mel scale filter bandwidth is equal; range 240 - 3400 Hz*)

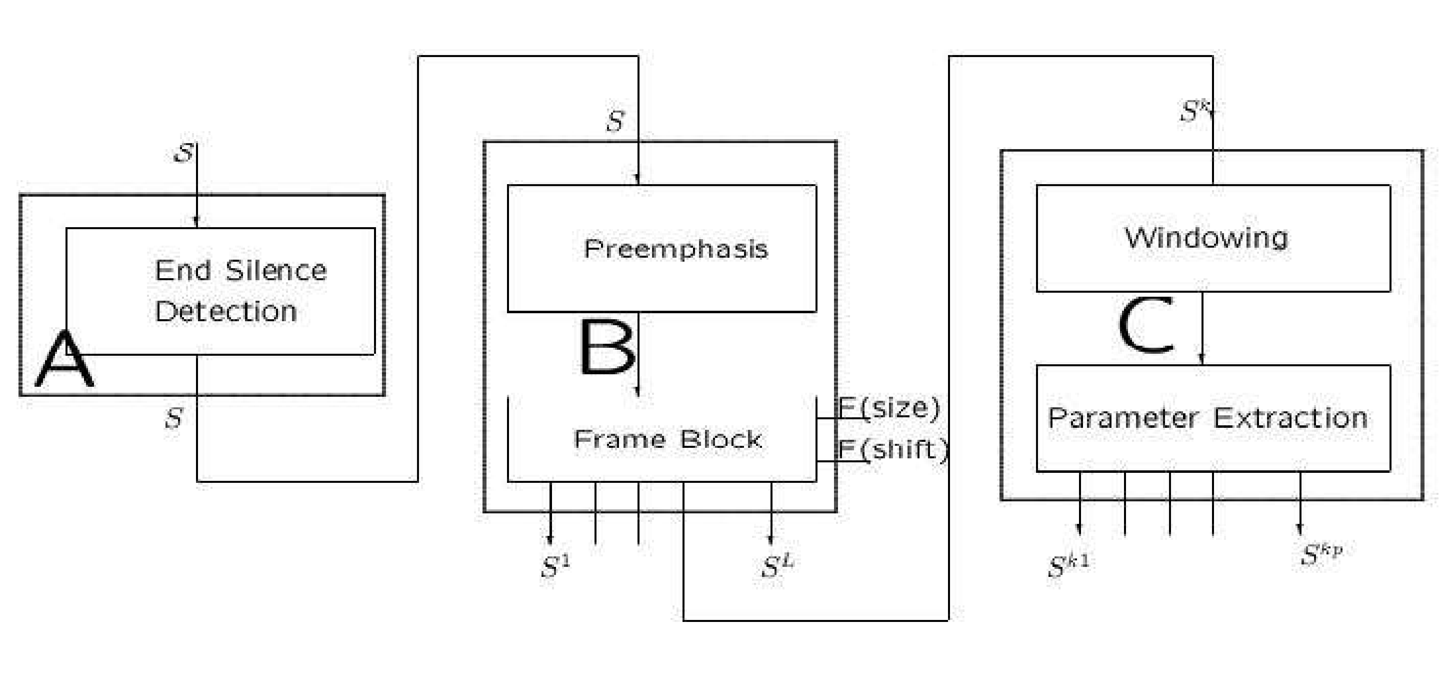
Energy in each filter band (e_1, e_2, $\cdots$ e_{24})



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

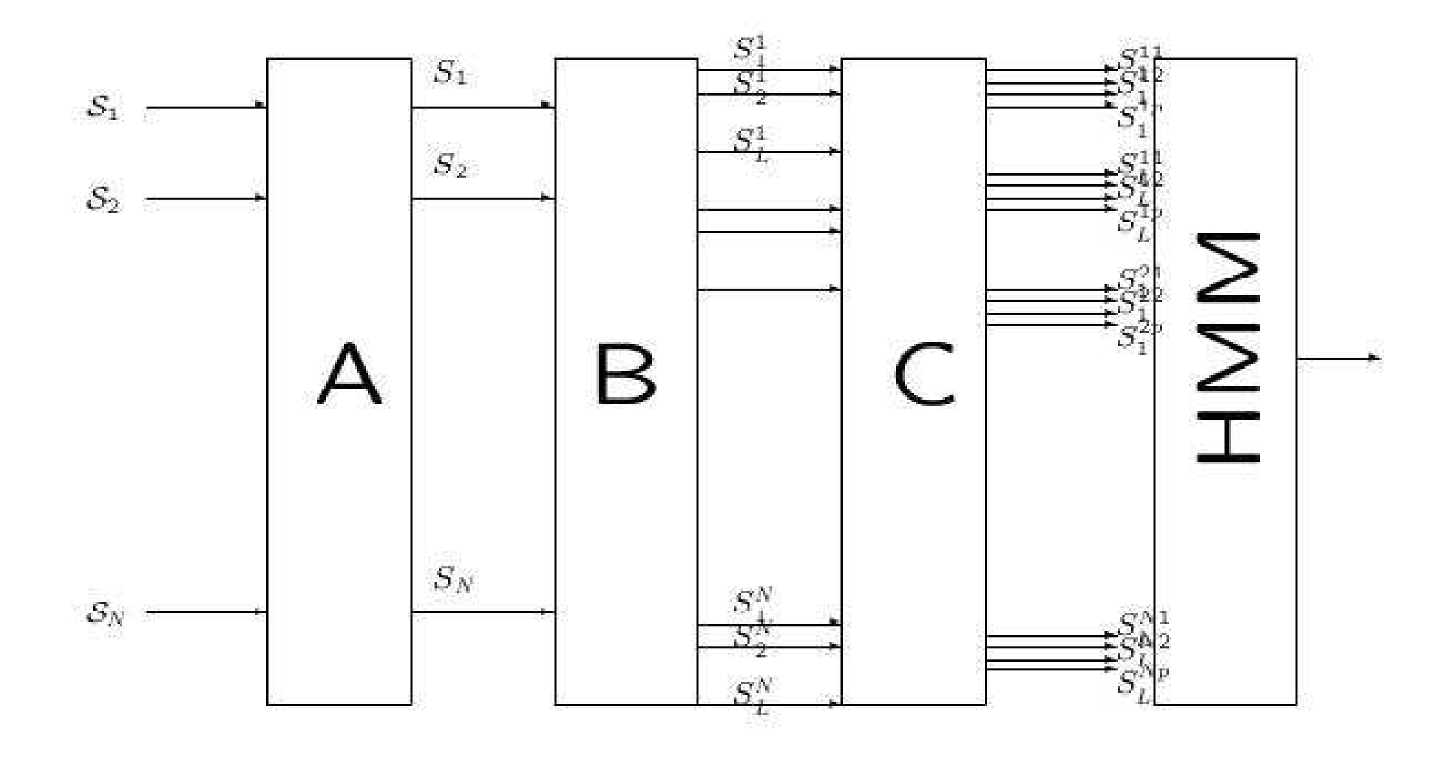
$$\Delta m_k = \frac{\sum_{l=1}^{L} l(m_{k+l} - m_{k-l})}{2\sum_{l=1}^{L} l^2}$$

Speech to Parameter Extraction



Speech sample converted into pre determined parameters. Note the 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.

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

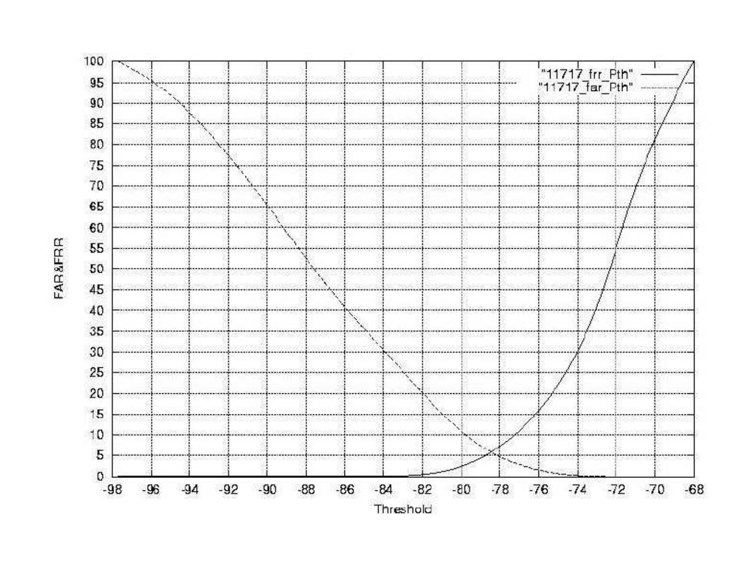
Types of Speaker Verification

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

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 Unvoiced Speech

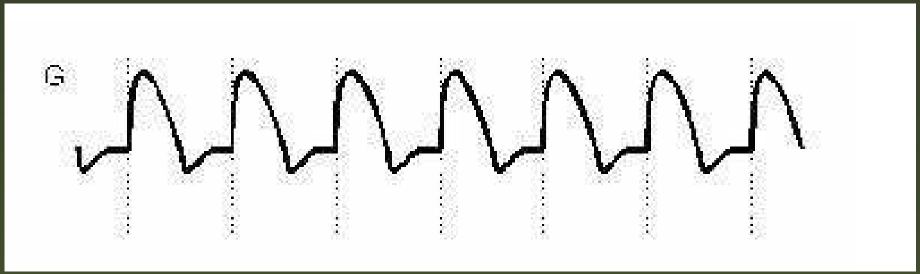
- When the vocal cords are tensed, the air flow causes them to vibrate, producing so-called **voiced speech** sounds.
- When the vocal cords are relaxed, in order to produce a sound, the air flow passes through a constriction in the vocal tract and thereby become turbulent, producing so-called **unvoiced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

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.)*

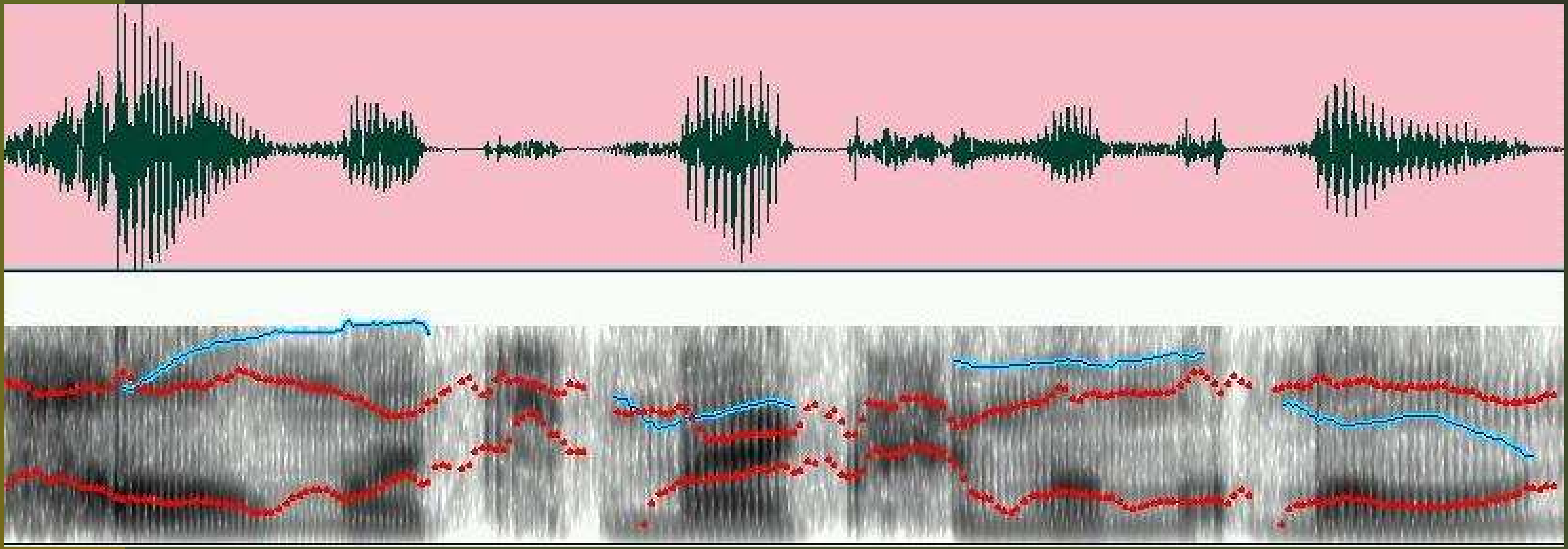
Voiced Speech - Formant?

- 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.

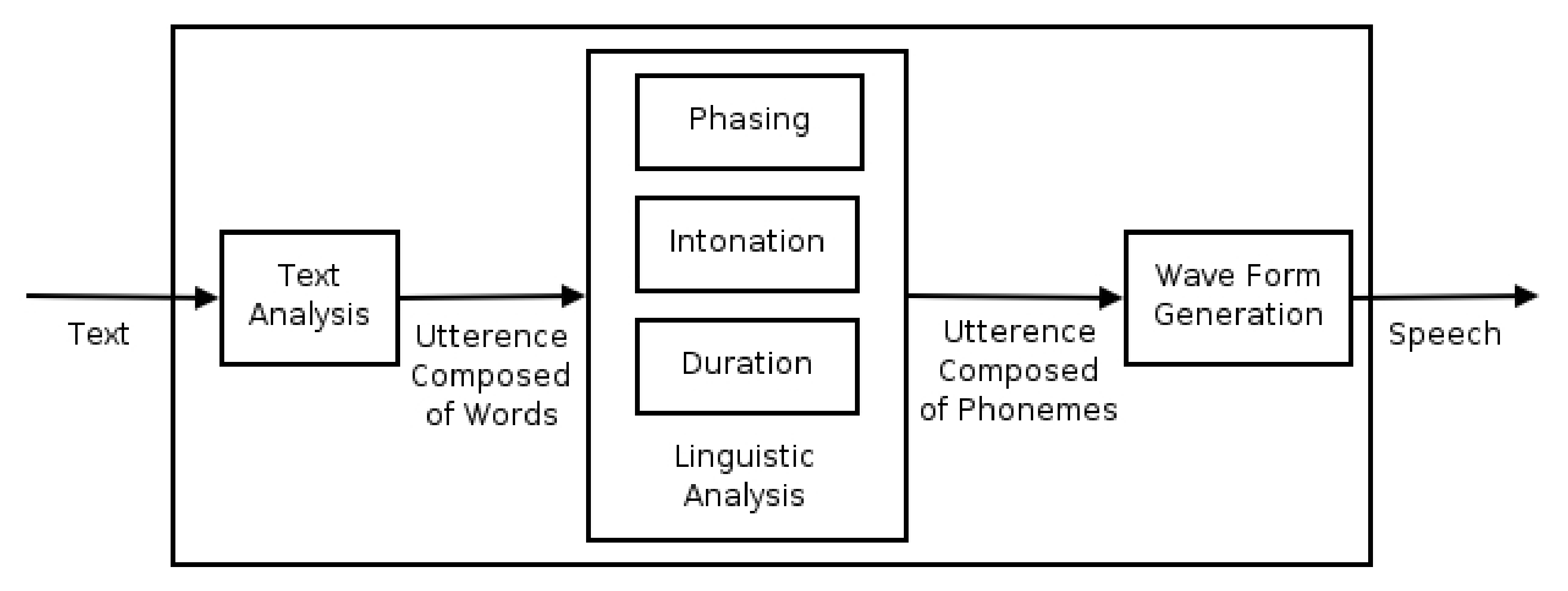
Formant and Pitch in Spectrogram



/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)*

How do we use all this?

From Lab to Land ....
Keyword Based Indexing of Multilingual News Videos

The Scene

- Television Broadcasting in India
  - Main source of news and entertainment
  - Officially 22 languages
  - Television broadcast in 11 major languages (Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu and Urdu)
  - Several 24×7 news channels
  - Most other non-news channels have specific news slots in their broadcast everyday

The Scene

- Television Broadcasting in India
  - Main source of news and entertainment
  - Officially 22 languages
  - Television broadcast in 11 major languages (Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu and Urdu)
  - Several 24 × 7 news channels
  - Most other non-news channels have specific news slots in their broadcast everyday

Number of television channels growing ...
broadcast in more languages ...

The Scene

- Post Broadcast
  - Archived ... hours of news
  - Stored safely .... (in tapes, on hard disks!)
- Usability of the broadcast? at a later time ...
  - Can we reuse it?
  - Can we get to specific news?
  - Can we search?
- Answer
  - No

The Scene

- Post Broadcast
  - Archived ... hours of news
  - Stored safely .... (in tapes, on hard disks!)
- Usability of the broadcast? at a later time ...
  - Can we reuse it?
  - Can we get to specific news?
  - Can we search?
- Answer
  - No

But we can .. most Western news channels ....

The Scene

- Why not for Indian channels?
  - Simple. **No** closed captioned text. Unlike most Western channels!
  - No searchable text associated with news broadcast
- Who would need it anyway?
  - Security Agencies (get cues to monitor happenings)
  - Individuals (Searchable information on media sharing websites)
  - Broadcast channels (reuse the feed as per need)

The Scene

- Why not for Indian channels?
  - Simple. **No** closed captioned text. Unlike most Western channels!
  - No searchable text associated with news broadcast
- Who would need it anyway?
  - Security Agencies (get cues to monitor happenings)
  - Individuals (Searchable information on media sharing websites)
  - Broadcast channels (reuse the feed as per need)

So ... What can be done?

The Scene

- Annotate (or Index) news broadcast
- How can we do this?
  - Full transcription of the news broadcast
    Good to have; difficult to automate
  - Spot keywords only video
    Sufficient; relatively easy to automate
- For Indian channels ...
  - Need for annotating multilingual videos exists
  - Indexing can be based only on audio and visual cues
    absence of closed caption text

The Scene

- Annotate (or Index) news broadcast
- How can we do this?
  - Full transcription of the news broadcast
    Good to have; difficult to automate
  - Spot keywords only video
    Sufficient; relatively easy to automate
- For Indian channels ...
  - Need for annotating multilingual videos exists
  - Indexing can be based only on audio and visual cues
    absence of closed caption text

How can we do this?

The Problem

- In the absence of closed caption text ...
  - Enable indexing of multilingual videos
    - spot **only the keywords** in news broadcast
    - in different languages
    - on different television channels
  - Using
    - The audio track of the news broadcast (audio cue)
    - Ticket text ... (visual cue)
      News Ticker? small screen space dedicated to presenting headlines or some important news.

The Problem

- In the absence of closed caption text ...
  - Enable indexing of multilingual videos
    - spot **only the keywords** in news broadcast
    - in different languages
    - on different television channels
  - Using
    - The audio track of the news broadcast (audio cue)
    - Ticket text ... (visual cue)
      News Ticker? small screen space dedicated to presenting headlines or some important news.

Are there challenges? in keyword spotting

Challenges

- Multiple languages
- No closed captioned text
  - Rely on audio and visual processing
- Audio or Visual to extract keywords? or Both?
  - *Maturity* of audio and visual processing work in *Indian languages* not evolved yet
  - Which keyword list to use?
  - *Construction* of keyword list in different languages?
  - Use of *small and dynamic* or *large and static* keyword list

Challenges

- Multiple languages
- No closed captioned text
  - Rely on audio and visual processing
- Audio or Visual to extract keywords? or Both?
  - *Maturity* of audio and visual processing work in *Indian languages* not evolved yet
  - Which keyword list to use?
  - *Construction* of keyword list in different languages?
  - Use of *small and dynamic* or *large and static* keyword list

How?

Approach

- Assumption
  Keyword based indexing is sufficient
- Observation
  - Keywords in news broadcast are either (a) proper nouns or (b) common nouns
  - Keywords are dynamic; they change with time
- Use RSS news feed to create dynamic keyword list (in English)
- How? Named entity detection, minimal parsing, n-gram analysis, statistical, ... several approaches

Approach

- Assumption
  Keyword based indexing is sufficient
- Observation
  - Keywords in news broadcast are either (a) proper nouns or (b) common nouns
  - Keywords are dynamic; they change with time
- Use RSS news feed to create dynamic keyword list (in English)
- How? Named entity detection, minimal parsing, n-gram analysis, statistical, ... several approaches

Good for English ... for Multilingual ....

Multilingual Keyword list

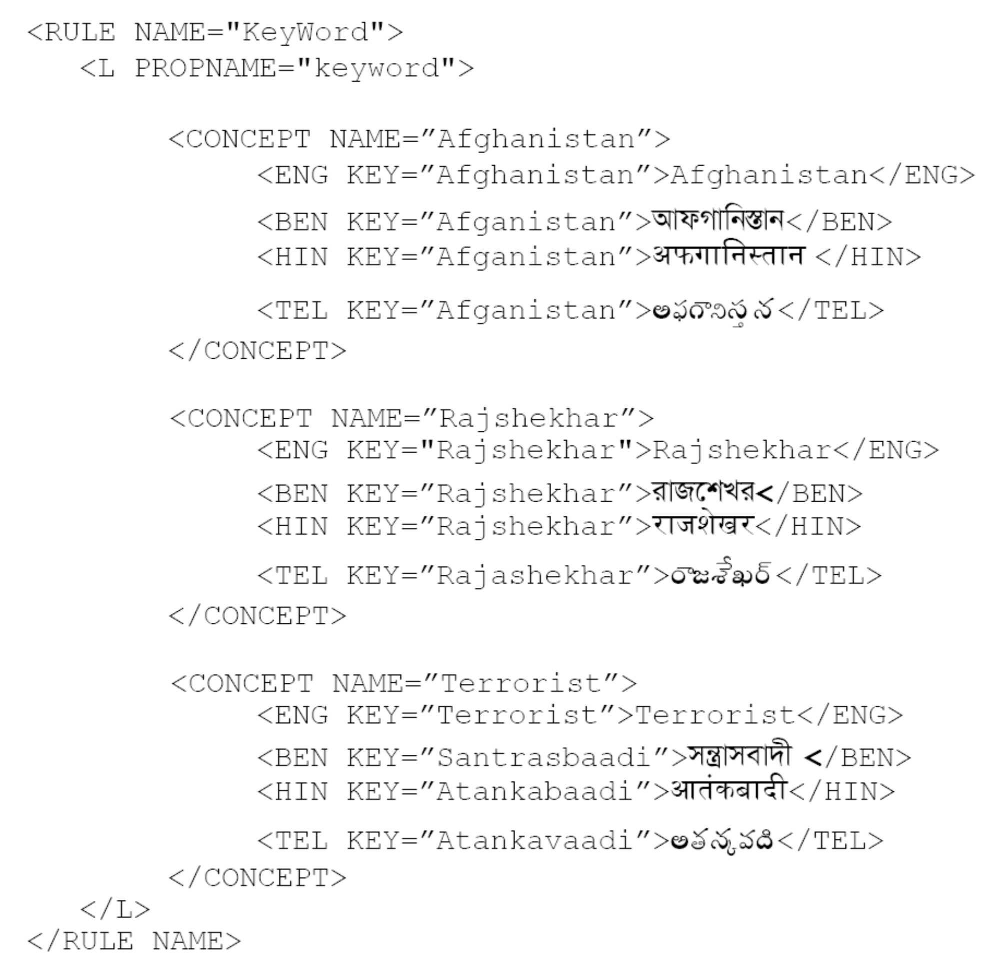
- Identify English keyword equivalents in other Indian languages (How?)
    - Use TDIL Indian language translation tools (some language pairs exist)
    - Proper nouns - majorly independent of language (transliteration)
    - Common nouns (on-line English to X dictionary)
- Create a multilingual keyword list
    - input for keyword spotting in audio and
    - input for keyword spotting in visual

Multilingual Keyword list

- Identify English keyword equivalents in other Indian languages (How?)
  - Use TDIL Indian language translation tools (some language pairs exist)
  - Proper nouns - majorly independent of language (transliteration)
  - Common nouns (on-line English to X dictionary)
- Create a multilingual keyword list
  - input for keyword spotting in audio and
  - input for keyword spotting in visual

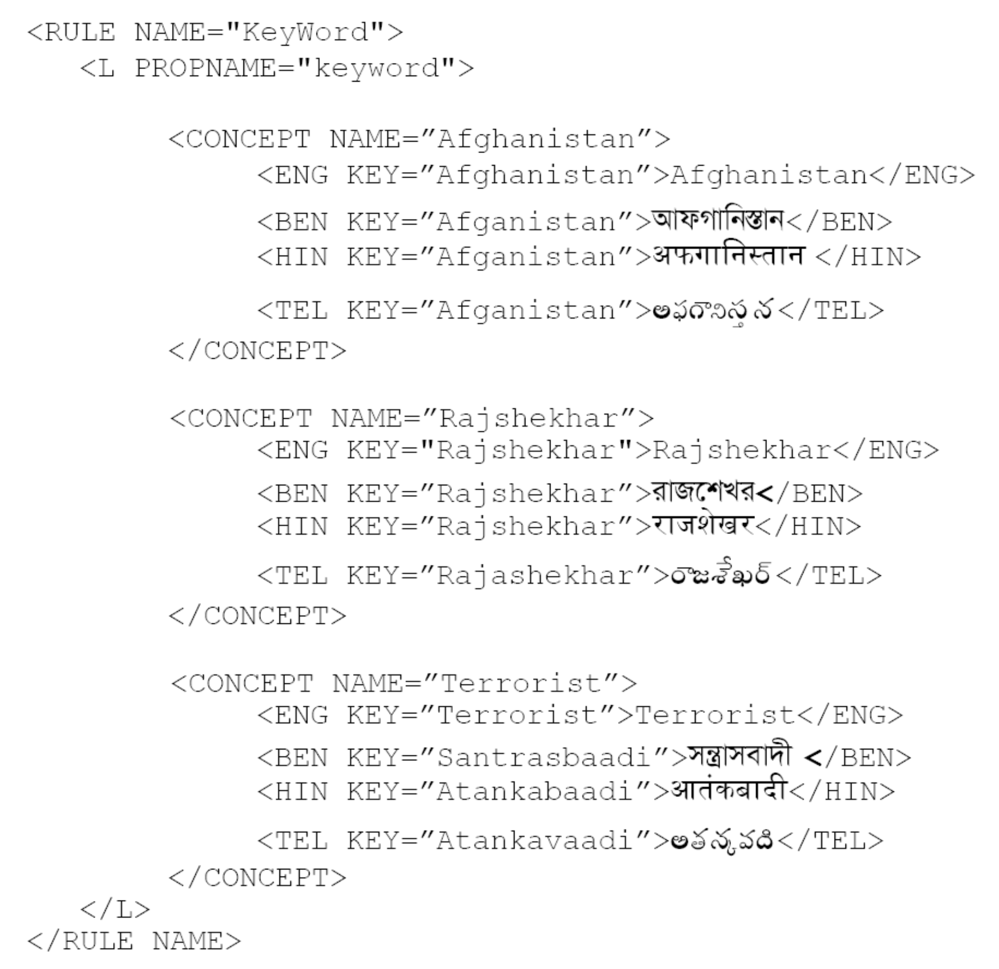
Multilingual keyword list? What is it?

Sample Multilingual Keyword list



```
<RULE NAME="KeyWord">
   <L PROPNAME="keyword">

        <CONCEPT NAME=”Afghanistan”>
             <ENG KEY=”Afghanistan”>Afghanistan</ENG>
             <BEN KEY=”Afganistan”>আফগানিস্তান</BEN>
             <HIN KEY=”Afganistan”>अफगानिस्तान </HIN>
             <TEL KEY=”Afganistan”>అఫగానిస్తన</TEL>
        </CONCEPT>

        <CONCEPT NAME=”Rajshekhar”>
             <ENG KEY="Rajshekhar">Rajshekhar</ENG>
             <BEN KEY=”Rajshekhar”>রাজশেখর</BEN>
             <HIN KEY=”Rajshekhar”>राजशेखर</HIN>
             <TEL KEY=”Rajashekhar”>రాజశేఖర్</TEL>
        </CONCEPT>

        <CONCEPT NAME=”Terrorist”>
             <ENG KEY=”Terrorist”>Terrorist</ENG>
             <BEN KEY=”Santrasbaadi”>সন্ত্রাসবাদী </BEN>
             <HIN KEY=”Atankabaadi”>आतंकबादी</HIN>
             <TEL KEY=”Atankavaadi”>అతన్కవడి</TEL>
        </CONCEPT>
   </L>
</RULE NAME>
```

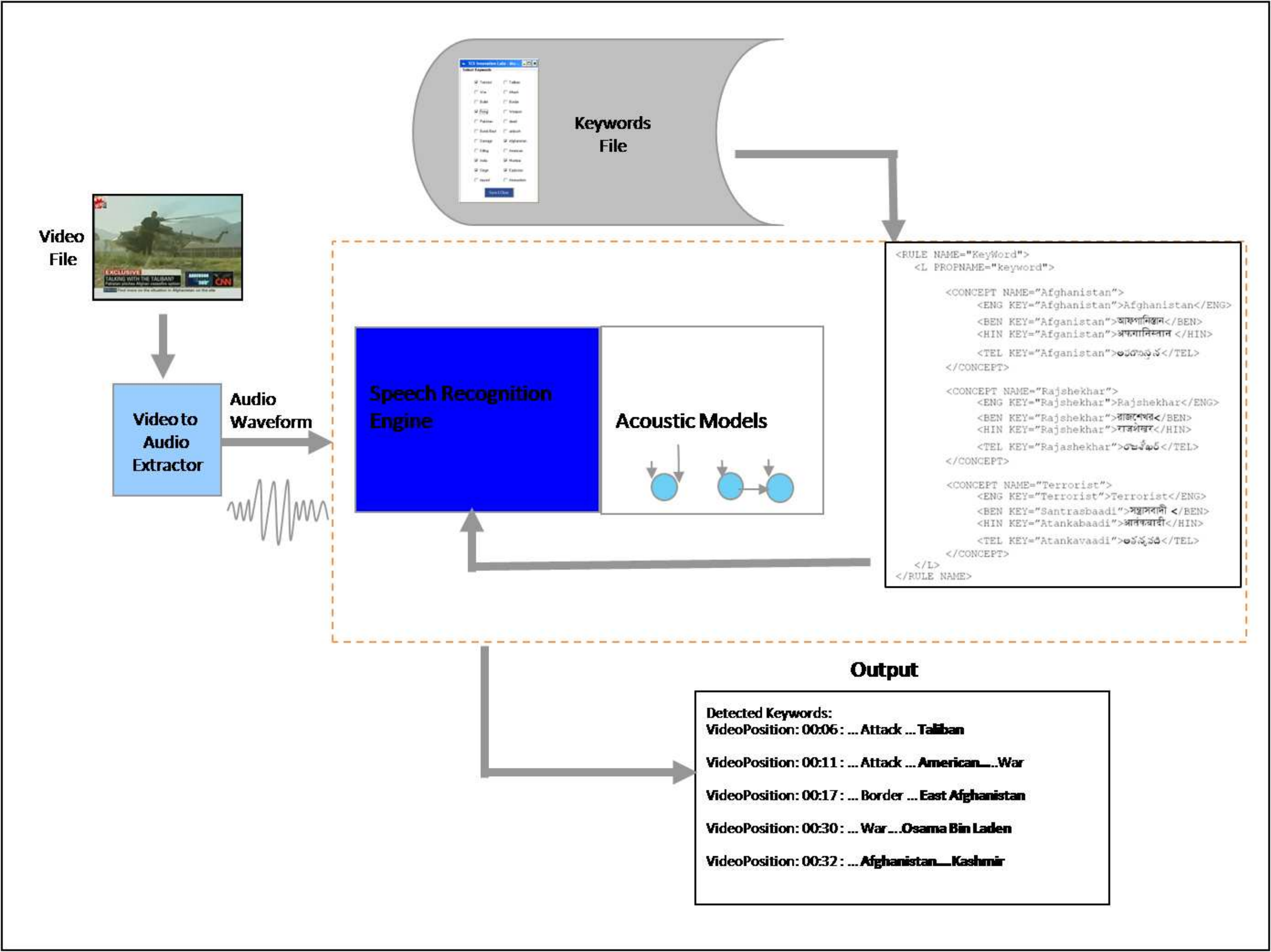
Sample Multilingual Keyword list



```
<RULE NAME="KeyWord">
   <L PROPNAME="keyword">

        <CONCEPT NAME=”Afghanistan”>
             <ENG KEY=”Afghanistan”>Afghanistan</ENG>
             <BEN KEY=”Afganistan”>আফগানিস্তান</BEN>
             <HIN KEY=”Afganistan”>अफगानिस्तान </HIN>
             <TEL KEY=”Afganistan”>అఫగానిస్తన</TEL>
        </CONCEPT>

        <CONCEPT NAME=”Rajshekhar”>
             <ENG KEY="Rajshekhar">Rajshekhar</ENG>
             <BEN KEY=”Rajshekhar”>রাজশেখর</BEN>
             <HIN KEY=”Rajshekhar”>राजशेखर</HIN>
             <TEL KEY=”Rajashekhar”>రాజశేఖర్</TEL>
        </CONCEPT>

        <CONCEPT NAME=”Terrorist”>
             <ENG KEY=”Terrorist”>Terrorist</ENG>
             <BEN KEY=”Santrasbaadi”>সন্ত্রাসবাদী </BEN>
             <HIN KEY=”Atankabaadi”>आतंकबादी</HIN>
             <TEL KEY=”Atankavaadi”>అతన్కవది</TEL>
        </CONCEPT>
   </L>
</RULE NAME>
```

How is it used?

In Audio KW Spotting



```
<RULE NAME="KeyWord">
   <L PROPNAME="keyword">

        <CONCEPT NAME="Afghanistan">
             <ENG KEY="Afghanistan">Afghanistan</ENG>
             <BEN KEY="Afganistan">আফগানিস্তান</BEN>
             <HIN KEY="Afganistan">अफगानिस्तान </HIN>
             <TEL KEY="Afganistan">అఫ్గానిస్తాన్</TEL>
        </CONCEPT>

        <CONCEPT NAME="Rajshekhar">
             <ENG KEY="Rajshekhar">Rajshekhar</ENG>
             <BEN KEY="Rajshekhar">রাজশেখর</BEN>
             <HIN KEY="Rajshekhar">राजशेखर</HIN>
             <TEL KEY="Rajashekhar">రాజశేఖర్</TEL>
        </CONCEPT>

        <CONCEPT NAME="Terrorist">
             <ENG KEY="Terrorist">Terrorist</ENG>
             <BEN KEY="Santrasbaadi">সন্ত্রাসবাদী </BEN>
             <HIN KEY="Atankabaadi">आतंकवादी</HIN>
             <TEL KEY="Atankavaadi">ఆతంకవాది</TEL>
        </CONCEPT>
   </L>
</RULE NAME>
```

Detected Keywords:
VideoPosition: 00:06 : ... Attack ... **Taliban**

VideoPosition: 00:11 : ... Attack ... **American**.....War

VideoPosition: 00:17 : ... Border ... **East Afghanistan**

VideoPosition: 00:30 : ... War.....**Osama Bin Laden**

VideoPosition: 00:32 : ... **Afghanistan**.....**Kashmir**

Keyword spotting in audio

- Using keyword list in different languages
- Use public domain Speech recognition engine
  Sphinx?
- Need creation of pronunciation lexicon
  Indian languages phonetic
- Need language specific acoustics models
  different for different languages? one for all? in between?

Keyword spotting in audio

- Using keyword list in different languages
- Use public domain Speech recognition engine
  Sphinx?
- Need creation of pronunciation lexicon
  Indian languages phonetic
- Need language specific acoustics models
  different for different languages? one for all? in between?

Are there challenges?

Challenges in Audio KWS

- Acoustic models for Indian languages do not exist
- Transcribed speech data for some languages exist (expensive and *toy like* compared to English!)

Challenges in Audio KWS

- Acoustic models for Indian languages do not exist
- Transcribed speech data for some languages exist (expensive and *toy like* compared to English!) *Can we do something?*

Challenges in Audio KWS

- Acoustic models for Indian languages do not exist
- Transcribed speech data for some languages exist (expensive and *toy like* compared to English!) *Can we do something?*
- Build or use acoustics models for one Indian language (simpler than building for all! Can we use existing (English) acoustic models??)
- Use this for keyword spotting (Largely Indian language phonetic and we are doing only keyword spotting anyway!)

Challenges in Audio KWS

- Acoustic models for Indian languages do not exist
- Transcribed speech data for some languages exist (expensive and *toy like* compared to English!)
  Can we do something?
- Build or use acoustics models for one Indian language
  (simpler than building for all! Can we use existing (English) acoustic models??)
- Use this for keyword spotting
  (Largely Indian language phonetic and we are doing only keyword spotting anyway!)

Does it work?

Some Experiments

- Used Microsoft SAPI (ASR Engine)
  (default English (US) acoustic models)
- Developed a self-help application based on spotting keywords
- Works well for Indian English accent
  (if the keywords are words not in the dictionary - we add the pronunciations)
- Works equally well with Hindi!
  (pronunciations lexicon for Hindi keywords words also added)

Some Experiments

- Used Microsoft SAPI (ASR Engine)
(default English (US) acoustic models)
- Developed a self-help application based on spotting keywords
- Works well for Indian English accent
(if the keywords are words not in the dictionary - we add the pronunciations)
- Works equally well with Hindi!
(pronunciations lexicon for Hindi keywords words also added)

Definitely we can do better with acoustic models of one Indian language! Visual KW Spotting?

Keyword spotting in visual

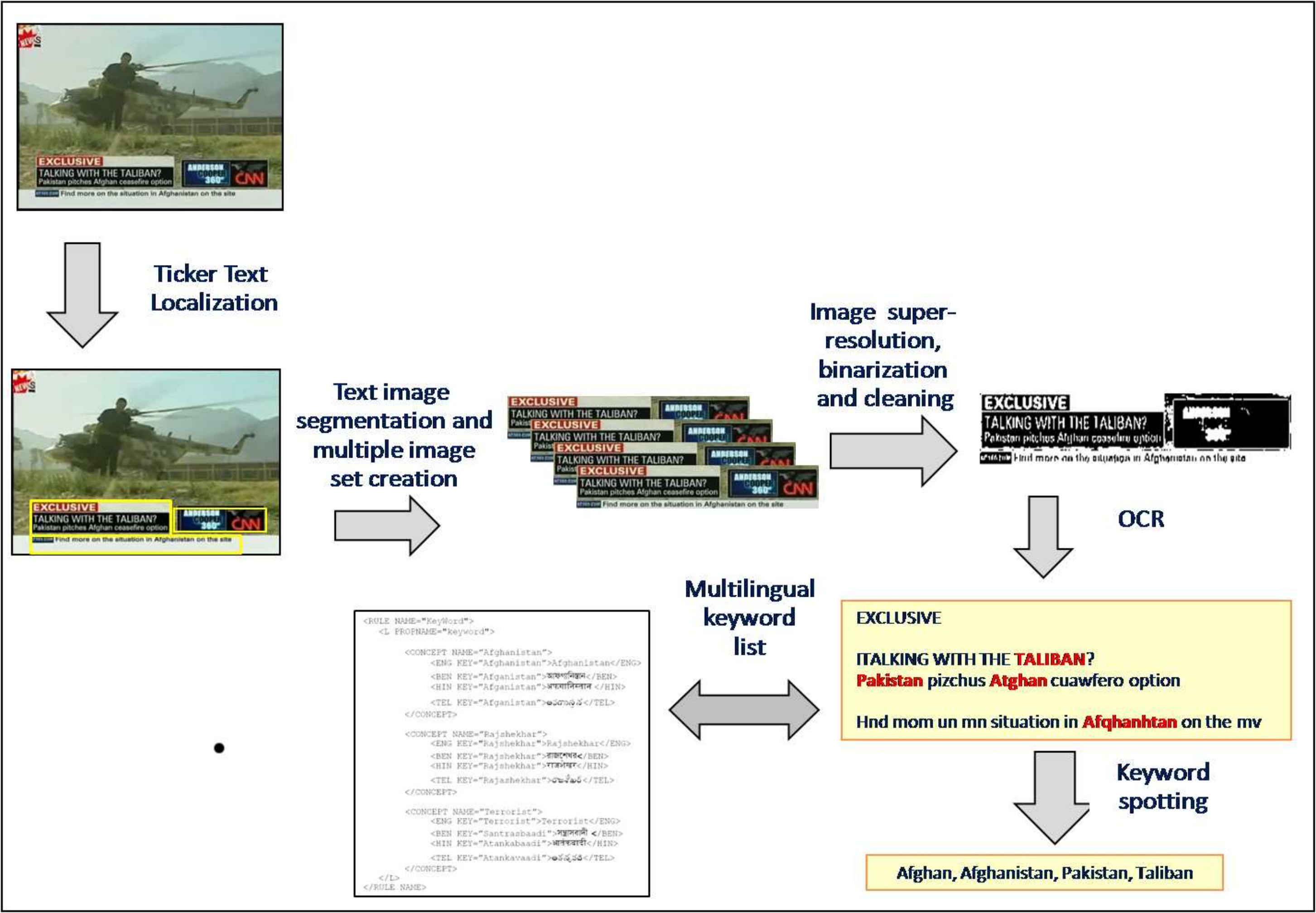
- Using keyword list in different languages
- English
  - Video OCR
  - sufficiently mature
  - increased accuracies with dynamic keyword list

Keyword spotting in visual

- Using keyword list in different languages
- English
  - Video OCR
  - sufficiently mature
  - increased accuracies with dynamic keyword list

How is this done?

Video OCR (English)


EXCLUSIVE
TALKING WITH THE TALIBAN?
Pakistan pitches Afghan ceasefire option
Find more on the situation in Afghanistan on the site
ANDERSON COOPER 360°
CNN
Ticker Text Localization
Text image segmentation and multiple image set creation
Image super-resolution, binarization and cleaning
OCR
Multilingual keyword list
Keyword spotting

```
<RULE NAME="KeyWord">
   <L PROPNAME="keyword">

         <CONCEPT NAME="Afghanistan">
              <ENG KEY="Afghanistan">Afghanistan</ENG>
              <BEN KEY="Afganistan">আফগানিস্তান</BEN>
              <HIN KEY="Afganistan">अफगानिस्तान </HIN>
              <TEL KEY="Afganistan">ఆఫ్గానిస్తాన్</TEL>
         </CONCEPT>

         <CONCEPT NAME="Rajshekhar">
              <ENG KEY="Rajshekhar">Rajshekhar</ENG>
              <BEN KEY="Rajshekhar">রাজশেখর</BEN>
              <HIN KEY="Rajshekhar">राजशेखर</HIN>
              <TEL KEY="Rajashekhar">రాజశేఖర్</TEL>
         </CONCEPT>

         <CONCEPT NAME="Terrorist">
              <ENG KEY="Terrorist">Terrorist</ENG>
              <BEN KEY="Santrasbaadi">সন্ত্রাসবাদী </BEN>
              <HIN KEY="Atankabaadi">आतंकवादी</HIN>
              <TEL KEY="Atankavaadi">ఆతంకవాది</TEL>
         </CONCEPT>
   </L>
</RULE NAME>
```

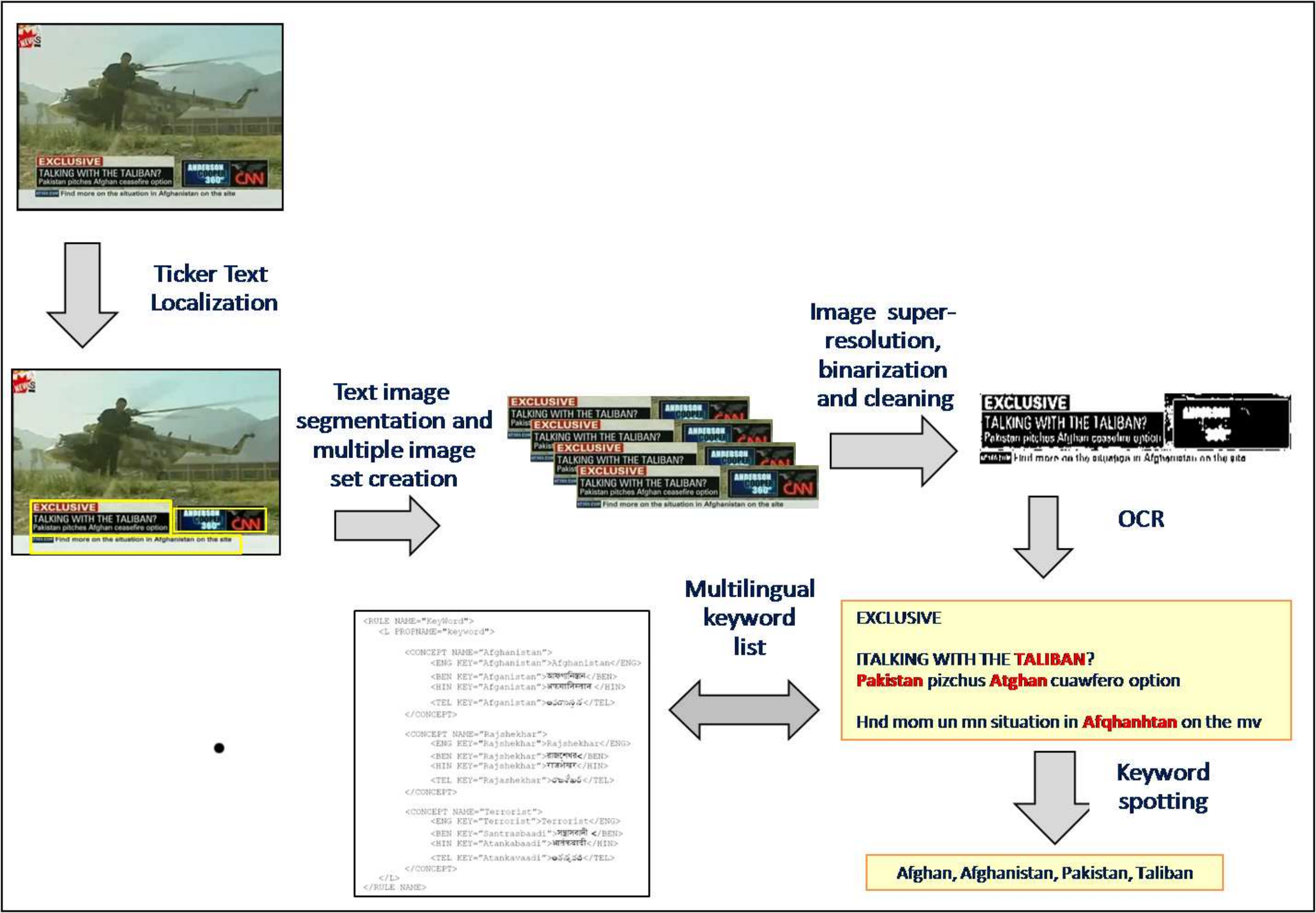
EXCLUSIVE

ITALKING WITH THE **TALIBAN**?
Pakistan pizchus **Atghan** cuawfero option

Hnd mom un mn situation in **Afqhanhtan** on the mv

Afghan, Afghanistan, Pakistan, Taliban

Video OCR (English)



Ticker Text Localization
Text image segmentation and multiple image set creation
Image super-resolution, binarization and cleaning
OCR
Multilingual keyword list
Keyword spotting

EXCLUSIVE

ITALKING WITH THE **TALIBAN**?
Pakistan pizchus **Atghan** cuawfero option

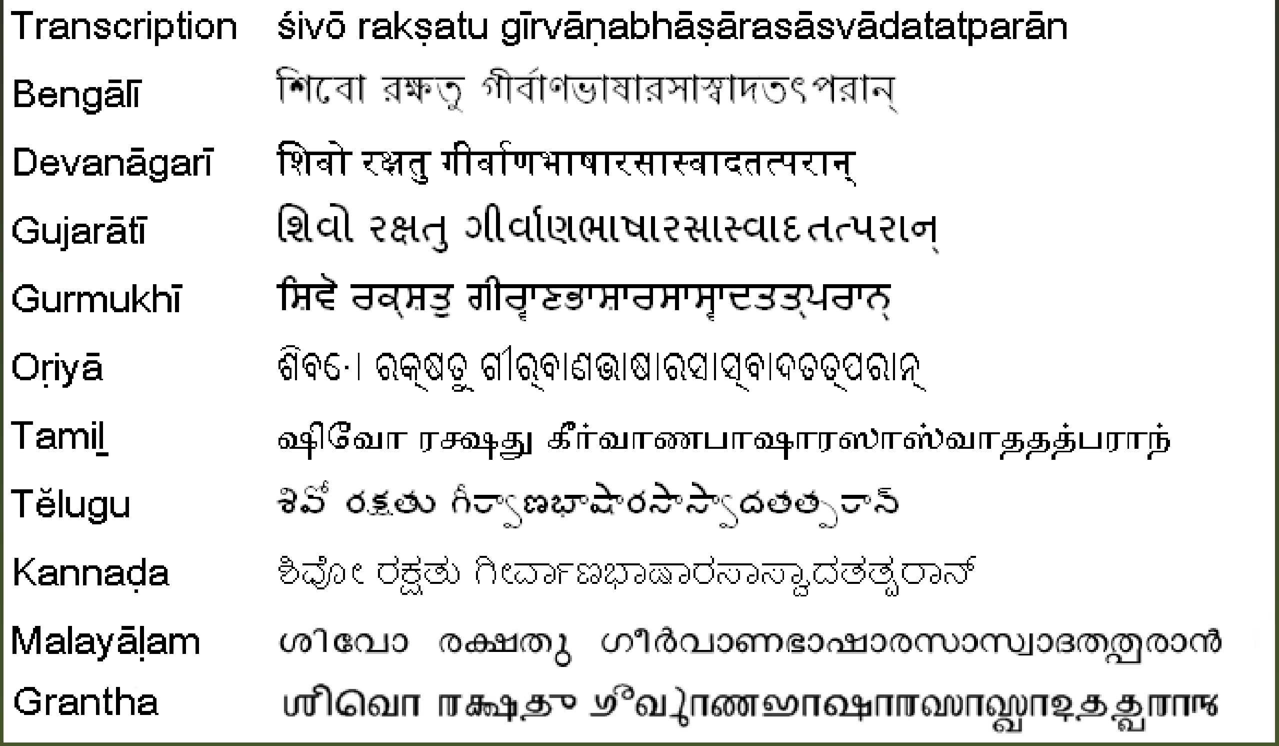
Hnd mom un mn situation in **Afqhanhtan** on the mv

Afghan, Afghanistan, Pakistan, Taliban

```
<RULE NAME="KeyWord">
   <L PROPNAME="keyword">

        <CONCEPT NAME="Afghanistan">
            <ENG KEY="Afghanistan">Afghanistan</ENG>
            <BEN KEY="Afganistan">আফগানিস্তান</BEN>
            <HIN KEY="Afganistan">अफगानिस्तान </HIN>
            <TEL KEY="Afganistan">ఆఫ్ఘనిస్తాన్</TEL>
        </CONCEPT>

        <CONCEPT NAME="Rajshekhar">
            <ENG KEY="Rajshekhar">Rajshekhar</ENG>
            <BEN KEY="Rajshekhar">রাজশেখর</BEN>
            <HIN KEY="Rajshekhar">राजशेखर</HIN>
            <TEL KEY="Rajashekhar">రాజశేఖర్</TEL>
        </CONCEPT>

        <CONCEPT NAME="Terrorist">
            <ENG KEY="Terrorist">Terrorist</ENG>
            <BEN KEY="Santrasbaadi">সন্ত্রাসবাদী </BEN>
            <HIN KEY="Atankabaadi">आतंकवादी</HIN>
            <TEL KEY="Atankavaadi">ఆతంకవాది</TEL>
        </CONCEPT>
   </L>
</RULE NAME>
```

Does it work for Indian Languages?

Indian Language Script OCR

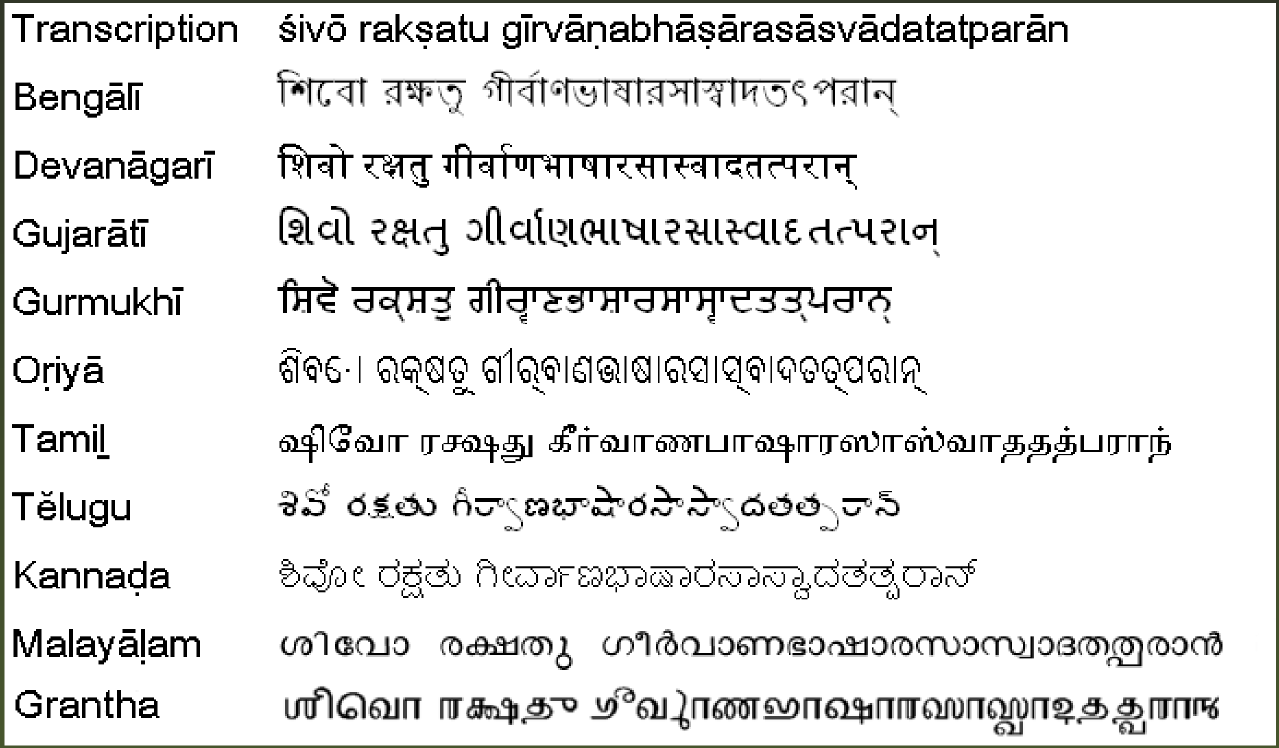


Transcription	śivō rakṣatu gīrvāṇabhāṣārasāsvādatatparān
Bengālī	শিবো রক্ষতু গীর্বাণভাষারসাস্বাদতৎপরান্
Devanāgarī	शिवो रक्षतु गीर्वाणभाषारसास्वादतत्परान्
Gujarātī	શિવો રક્ષતુ ગીર્વાણભાષારસાસ્વાદતત્પરાન્
Gurmukhī	ਸ਼ਿਵੋ ਰਕ੍ਸ਼ਤੁ ਗੀਰ੍ਵਾਣਭਾਸ਼ਾਰਸਾਸ੍ਵਾਦਤਤ੍ਪਰਾਨ੍
Oṛiyā	ଶିବୋ ରକ୍ଷତୁ ଗୀର୍ବାଣଭାଷାରସାସ୍ବାଦତତ୍ପରାନ୍
Tamil̲	ஷிவோ ரக்ஷது கீர்வாணபாஷாரஸாஸ்வாததத்பராந்
Tĕlugu	శివో రక్షతు గీర్వాణభాషారసాస్వాదతత్పరాన్
Kannaḍa	ಶಿವೋ ರಕ್ಷತು ಗೀರ್ವಾಣಭಾಷಾರಸಾಸ್ವಾದತತ್ಪರಾನ್
Malayāḷam	ശിവോ രക്ഷതു ഗീർവാണഭാഷാരസാസ്വാദതത്പരാൻ
Grantha	𑌶𑌿𑌵𑍋 𑌰𑌕𑍍𑌷𑌤𑍁 𑌗𑍀𑌰𑍍𑌵𑌾𑌣𑌭𑌾𑌷𑌾𑌰𑌸𑌾𑌸𑍍𑌵𑌾𑌦𑌤𑌤𑍍𑌪𑌰𑌾𑌨𑍍

Source: http://www.myscribeweb.com/Phrase_sanskrit.png

- Script - Complex; Work being done in few languages
- Poor accuracies ;-(using approaches for English

Indian Language Script OCR



Transcription	śivō rakṣatu gīrvāṇabhāṣārasāsvādatatparān
Bengālī	শিবো রক্ষতু গীর্বাণভাষারসাস্বাদতৎপরান্
Devanāgarī	शिवो रक्षतु गीर्वाणभाषारसास्वादतत्परान्
Gujarātī	શિવો રક્ષતુ ગીર્વાણભાષારસાસ્વાદતત્પરાન્
Gurmukhī	ਸ਼ਿਵੇ ਰਕ੍ਸ਼ਤੁ ਗੀਰ੍ਵਾਣਭਾਸ਼ਾਰਸਾਸ੍ਵਾਦਤਤ੍ਪਰਾਨ੍
Oṛiyā	ଶିବୋ ରକ୍ଷତୁ ଗୀର୍ବାଣଭାଷାରସାସ୍ବାଦତତ୍ପରାନ୍
Tamiḻ	ஷிவோ ரக்ஷது கீர்வாணபாஷாரஸாஸ்வாததத்பராந்
Tĕlugu	శివో రక్షతు గీర్వాణభాషారసాస్వాదతత్పరాన్
Kannaḍa	ಶಿವೋ ರಕ್ಷತು ಗೀರ್ವಾಣಭಾಷಾರಸಾಸ್ವಾದತತ್ಪರಾನ್
Malayāḷam	ശിവോ രക്ഷതു ഗീർവാണഭാഷാരസാസ്വാദതത്പരാൻ
Grantha	𑌶𑌿𑌵𑍋 𑌰𑌕𑍍𑌷𑌤𑍁 𑌗𑍀𑌰𑍍𑌵𑌾𑌣𑌭𑌾𑌷𑌾𑌰𑌸𑌾𑌸𑍍𑌵𑌾𑌦𑌤𑌤𑍍𑌪𑌰𑌾𑌨𑍍

Source: http://www.myscribeweb.com/Phrase_sanskrit.png

- Script - Complex; Work being done in few languages
- Poor accuracies ;-(using approaches for English

New approach?

Recognition Free Approach? Maybe

- We know the keyword list (dynamic and update!)
- Generate images of keywords
  (different fonts and sizes; usually not very different!)
- Match in the images space
  - ticker text can be segmented into word images;
  - compare with generated keyword images;
  - some work done at IIIT Hyderabad
    (http://cvit.iiit.ac.in/projects/videoprocessing/)
- Recognition accuracies still poor

Recognition Free Approach? Maybe

- We know the keyword list (dynamic and update!)
- Generate images of keywords
  (different fonts and sizes; usually not very different!)
- Match in the images space
  - ticker text can be segmented into word images;
  - compare with generated keyword images;
  - some work done at IIIT Hyderabad
    (http://cvit.iiit.ac.in/projects/videoprocessing/)
- Recognition accuracies still poor

Recognition accuracies neither good in audio nor in visual .... can we do something?

Integrate

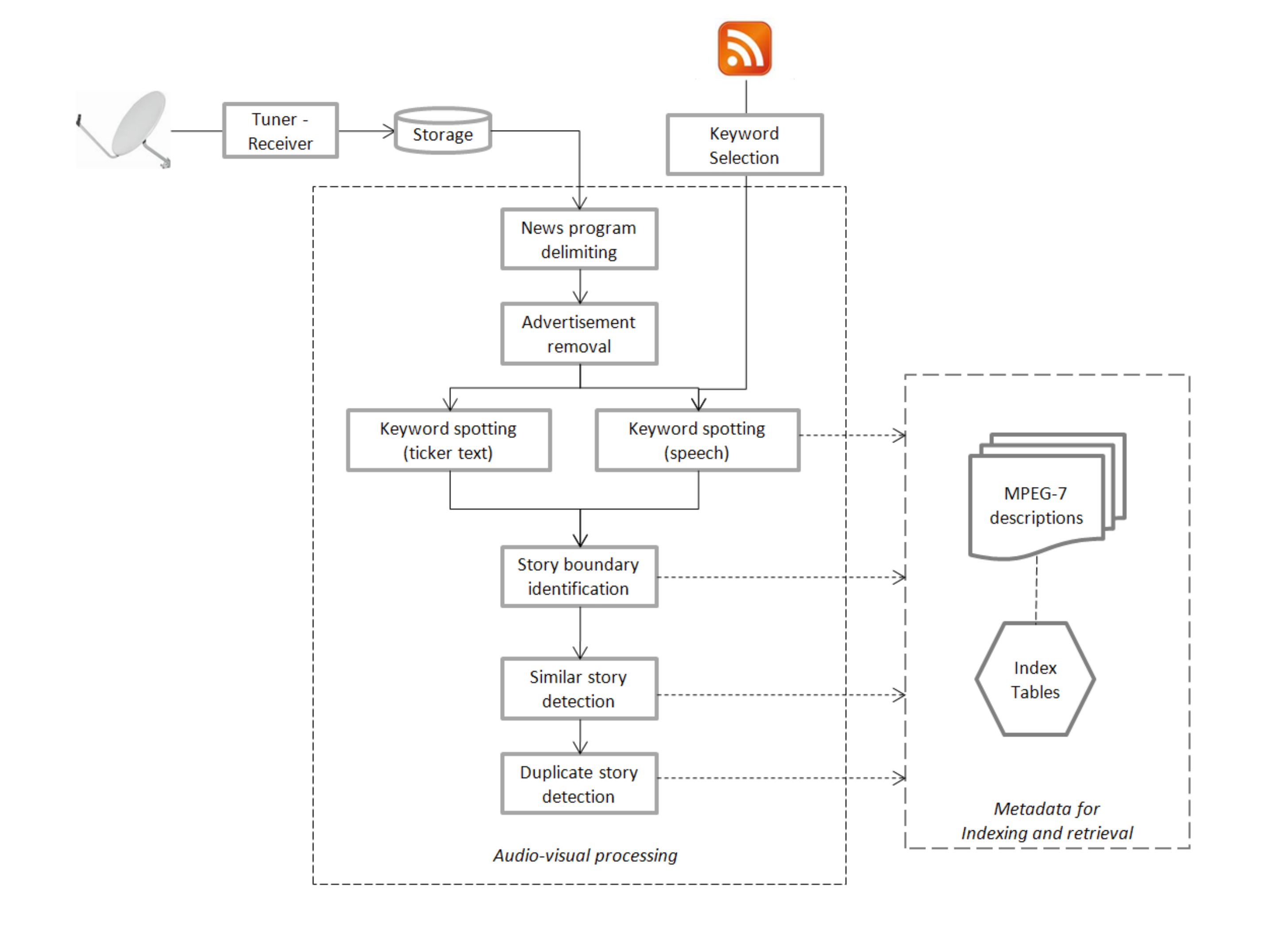
- Fuse the output of
  - audio channel keyword spotting and
  - keyword spotting in visual
- To get better keyword spotting accuracies in the news broadcast
  - in terms of precision
  - in terms of recall
- Validated by some preliminary experiments

Integrate

- Fuse the output of
  - audio channel keyword spotting and
  - keyword spotting in visual
- To get better keyword spotting accuracies in the news broadcast
  - in terms of precision
  - in terms of recall
- Validated by some preliminary experiments

In Summary ...

Summary


Tuner -
Receiver
Storage
Keyword
Selection
News program
delimiting
Advertisement
removal
Keyword spotting
(ticker text)
Keyword spotting
(speech)
Story boundary
identification
Similar story
detection
Duplicate story
detection
Audio-visual processing
MPEG-7
descriptions
Index
Tables
Metadata for
Indexing and retrieval

Approach Summary

- Use RSS feed to create keyword list
  (in English; use NL processing)
- Identify English keyword in other Indian languages
  (common nouns use word dictionary; transliteration for proper nouns)
- Create a multilingual keyword list
  (source for keyword spotting)
- Keyword spotting in audio
  (in different languages; same acoustic models; use Sphinx)
- Keyword spotting in visual
  (ticker text in different languages; Recognition free)

What is Novel?

- Smart Use of RSS feed to construct a keyword list
- Results in dynamic and small KW list
  increased keyword spotting accuracies
- Creation of a multilingual keyword list
- Using on-line resources
  dictionary and transliteration
- Combining audio and visual to improve KW spotting
  when acoustics data is small and the script to be recognized is complex!

What is Novel?

- Smart Use of RSS feed to construct a keyword list
- Results in dynamic and small KW list
  increased keyword spotting accuracies
- Creation of a multilingual keyword list
- Using on-line resources
  dictionary and transliteration
- Combining audio and visual to improve KW spotting
  when acoustics data is small and the script to be recognized is complex!

Special (difficult?) problems need appropriate solutions

Thank You

- Queries?
- Comments
- Suggestions?

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