

# Automatic Speech Recognition – Research and Standards

**S. Umesh**

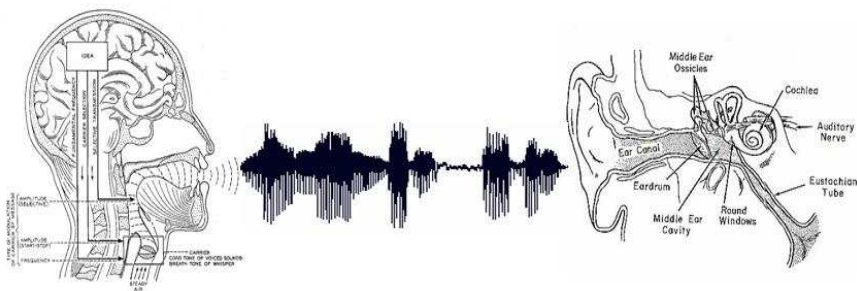
(with Raghavendra, Kishore Prahlad, Hema Murthy)

Department of Electrical Engineering  
Indian Institute of Technology Madras  
May 7<sup>th</sup>, 2010

# Outline

- Automatic Speech Recognition (ASR)
- ASR engines from academia
- ASR engines from industry
- Flexibility & Limitation of academia ASRs
- Existing Standards

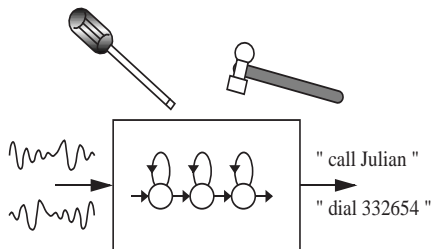
# Automatic Speech Recognition (ASR)



- **ASR technology:** allows a computer to recognize words that a person speaks into a microphone or telephone. Convert the input speech into text.
- Articulators produce sounds which the ear conveys to the brain for processing.

# Automatic Speech Recognition (ASR)

ASR - Convert Speech signal to words



- Most languages: only 50-60 distinct sound units make up the words
- Example :

and - sil /a/ /n/ /d/ sil

yes - sil /y/ /E/ /s/ sil

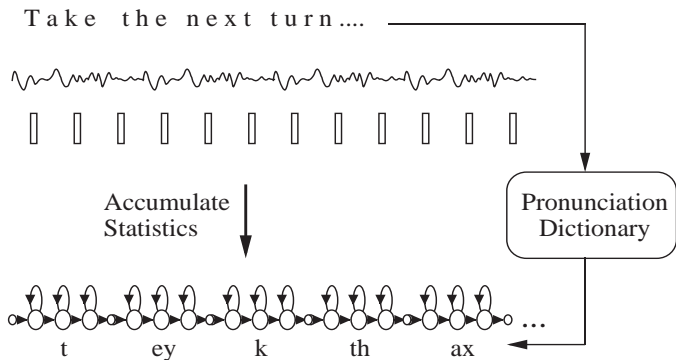
no - sil /n/ /ow/ sil

# Pronunciation Dictionary

Pronunciation Dictionary: Expand words to corresponding sounds

|       |           |
|-------|-----------|
| A     | ah        |
| A     | ax        |
| A     | ey        |
| CALL  | k ao l    |
| DIAL  | d ay ax l |
| EIGHT | ey t      |
| PHONE | f ow n    |
| SEVEN | s eh v n  |
| TO    | t ax      |
| TO    | t uw      |
| ZERO  | z ia r ow |

# Training of a Speech Recognition System



# Language Model (LM)

- Example:

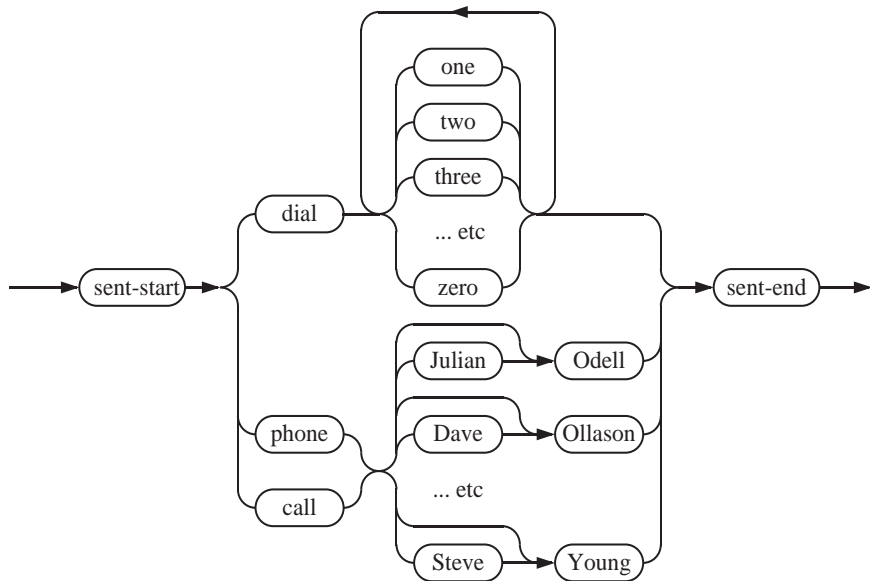
"It's fun to recognise speech?"

"It's fun to wreck a nice beach?"

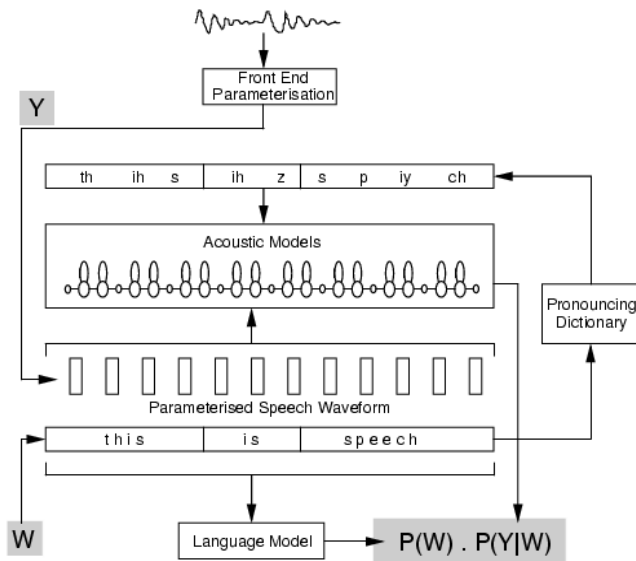
Although the "sound-sequence" may be similar, LM will tell us that the first sentence is more likely

- Language models are used to restrict the combination of words
- Permissible words following each word are given explicitly in LM

# Grammar

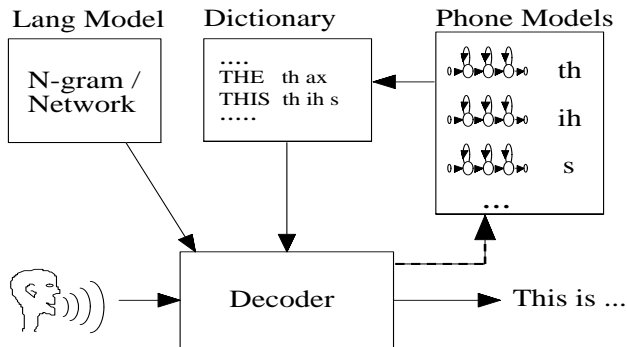


# Statistical View Point of ASR



# Recognition

- Speech Recognition Grammar Specification (SRGS)
- Pronunciation Lexicon Specification (PLS)



# ASR Engines from Academia

- Sphinx (CMU)
- HTK (Cambridge University)
- SUMMIT (MIT)
- SONIC (University of Colorado)
- Julius (CSRC, Japan)
- CSLU (OGI school of Science and Engineering)

## ASR engines from Industry

- Loquendo ASR
- Dragon Naturally Speaking
- TeliSpeech Recognizer
- IBM ViaVoice
- MacSpeech
- Simmortel Voice
- e-Speaking
- VoiceFinger
- LumenVox Speech Engine

## Flexibility & Limitations of Academia ASRs

- Very Flexible: Lot of freedom to make changes in different modules
- Geared towards promoting research in different modules
- Interoperability between ASRs is difficult
  - An module working Sphinx cannot be easily made to work in HTK
  - Input, output specifications differ between ASRs
  - Storage formats are very different
- To use ASR in various applications, system should support standards.
  - Allow easy interoperability
  - To have a plug-n-play ASR
- Industry engines follow standards but still do not allow easy inter-operability between various industry engines.

# Conclusion

- Standards exist for many modules of ASRs but not all
- For some applications, current standards may not provide enough flexibility
- Academic ASRs do not follow standards but provide flexibility
- Implementation aspects of Standards in ASR differ significantly from research aspects of ASR

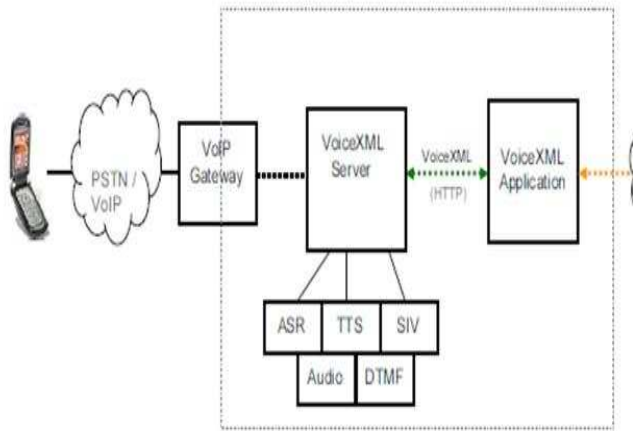
# Standards

- The various kinds of standards over internet, distributed system and desktops are as follows.
  - Internet; Voice Extensible Markup Language (VXML)
  - Distributed environment; Media Resource Control Protocol (MRCP)
  - Desktop; Speech Application Programming Interface (SAPI)
  - ASRs developed by the academia does not follow above standards where as industry developed systems (such as supports most one or two
- Microsofts SAPI
- W3C Standards
  - Voice XML
  - Media Resource Control Protocol (MRCP)
  - Speech Recognition Grammar Specification (SRGS)
  - Pronunciation Lexicon Specification (PLS)

# VXML

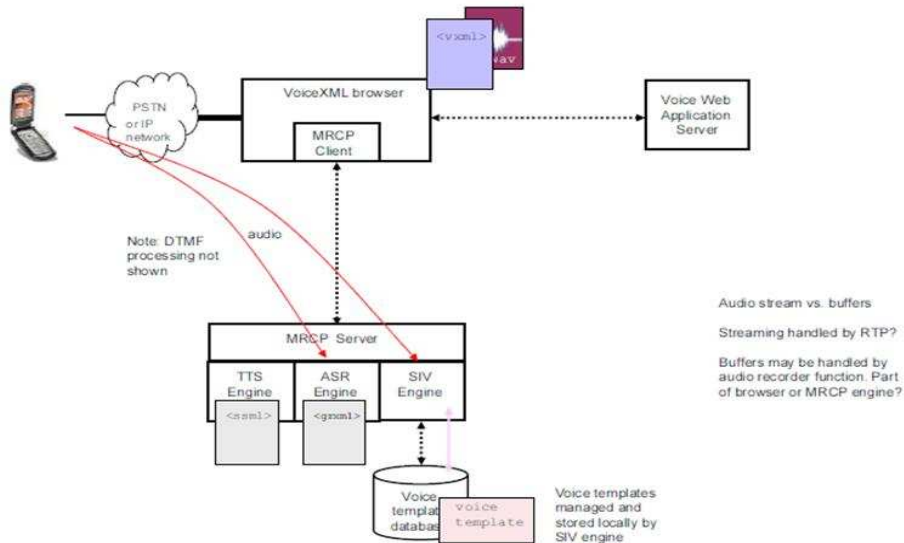
- VoiceXML (VXML) is the W3C's standard XML format
- Specify interactive voice dialogues between a human and a computer.
- A kind of programming language that helps computers and other devices operate through telephone lines.
- Allows voice applications to be developed and deployed in an analogous way to HTML for visual applications.
- As HTML documents are interpreted by a visual web browser, VoiceXML documents are interpreted by a voice browser or IVR.
- VoiceXML has tags that instruct the voice browser to provide speech synthesis, automatic speech recognition, dialog management, and audio playback.

# Voice XML Application Architecture



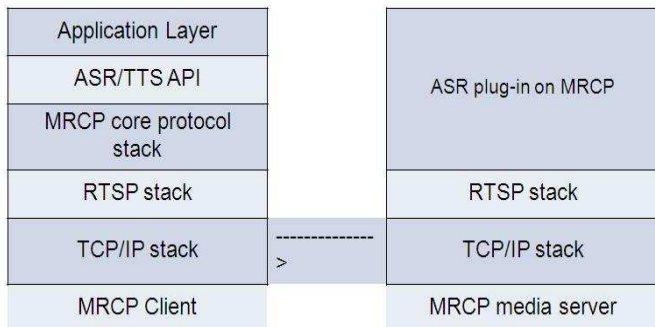
- Media Resource Control Protocol (MRCP) helps to talk to different speech engines (TTS, ASR, Speaker id) in an efficient fashion.

# VXML, MRCP



# MRCP

- MRCP is designed to provide a mechanism for a client device requiring audio/video stream processing to control processing resources on the network.
- The MRCP protocol defines the requests, responses, and events needed to control the media processing resources.
- Architecture



# SAPI

- The SAPI is an API developed by Microsoft to allow the use of speech recognition within Windows applications.
- In general all versions of the API have been designed such that a software developer can write an application to perform speech recognition and synthesis by using a standard set of interfaces, accessible from a variety of programming languages.
- In addition, it is possible for a 3rd-party company to produce their own Speech Recognition engine or adapt existing engines to work with SAPI. In principle, as long as these engines conform to the defined interfaces they can be used instead of the Microsoft-supplied engines.