CSE6328.3
Speech & Language Processing

YORK UNIVERSITÉ UNIVERSITY redefine THE POSSIBLE.

No.1

Introduction

Prof. Hui Jiang
Department of Computer Science and Engineering
York University

CSE6328 Course Outline: *"Speech & Language Processing"*

- Part I: Introduction (2 weeks)
  - Overview of speech and language technologies
  - Basic Knowledge of speech and spoken language
  - Math foundation: review
- Part II: Basic theory of pattern classification/verification (4—5 weeks)
  - Bayesian decision rule; Model estimation methods
  - Generative models: Gaussian, GMM, Markov Chain, HMM, Graphical models
  - Discriminative models: SVM, Neural networks (NN) and beyond
- Part III: Case studies (3—4 weeks)
  - Automatic speech recognition
  - Spoken language processing
- Part IV: Advanced topics – YOUR PARTICIPATION !! (1—2 weeks)
  - Choose an advanced topic in machine learning
  - Choose a journal article in speech/language or other applications
  - Self-study and oral presentation in class

Course Info

- **Course Web site:** ***http://www.cs.yorku.ca/course/6328/***
- **Course Format:**
  - **Lectures (10 weeks):**
    - **Covers basic data modeling, pattern classification theory;**
    - **Introduces some selected applications in speech recognition and spoken language processing.**
  - **Students' in-class presentations (2 weeks): based on basic theories in class, self-study recently published technical article and orally present it in class.**
    - **Choose an advanced topic in machine learning**
    - **Choose an application in speech/language or other domains**
- **Evaluation:**
  - **One assignment (10%) (roughly first 1/3 of the course)**
  - **Two lab projects (50%): report + oral presentation(?)**
  - **Advanced topic self-study and in-class presentation (30%)**
  - **Class Participation (10%)**

Reference Materials

- **Lecture notes**
- **Assigned reading materials through the course**
- **Reference books:**

[1] Pattern Recognition and Machine Learning by C. M. Bishop. (Springer, ISBN 0-387-31073-8)

[2] *Pattern Classification* (2nd Edition) by R. O. Duda, P. Hart and D. Stork. (John Wiley & Sons, Inc., ISBN 0-471-05669-3)

[3] *Spoken Language Processing: a guide to theory, algorithm, and system development* by X.D. Huang, A. Acero, H.W. Hon. (Prentice Hall PTR, ISBN 0-13-022616-5)

[4] *Foundations of Statistical Natural Language Processing* by C. D. Manning and H. Schutze. (The MIT Press, ISBN 0-262-13360-1

- **Prerequisite:**
  - First course in probability or statistics
  - First course in linear algebra or matrix theory
  - C/C++/Java and perl/shell programming skill (for project)

Speech Research and Technology

- Speech Communication
- Speech Production and Perception
- Speech Analysis and Synthesis
- Speech and Audio Coding & Compression
- Speech Recognition and Understanding
- Speaker Identification and Verification
- Speech Enhancement
- Language Identification
- Dialogue Processing

Language Research and Technology

- Written vs. Spoken Languages
- Computational Linguistics
- Corpus-Based Language Technologies
- Statistical Language Modeling
- Language Analysis and Generation
- Statistical Part-of-Speech Tagging
- Modeling Syntax and Semantics
- Statistical Text Understanding / Text Mining
- Probabilistic Parsing
- Text Categorization
- Statistical Machine Translation
- Information Retrieval

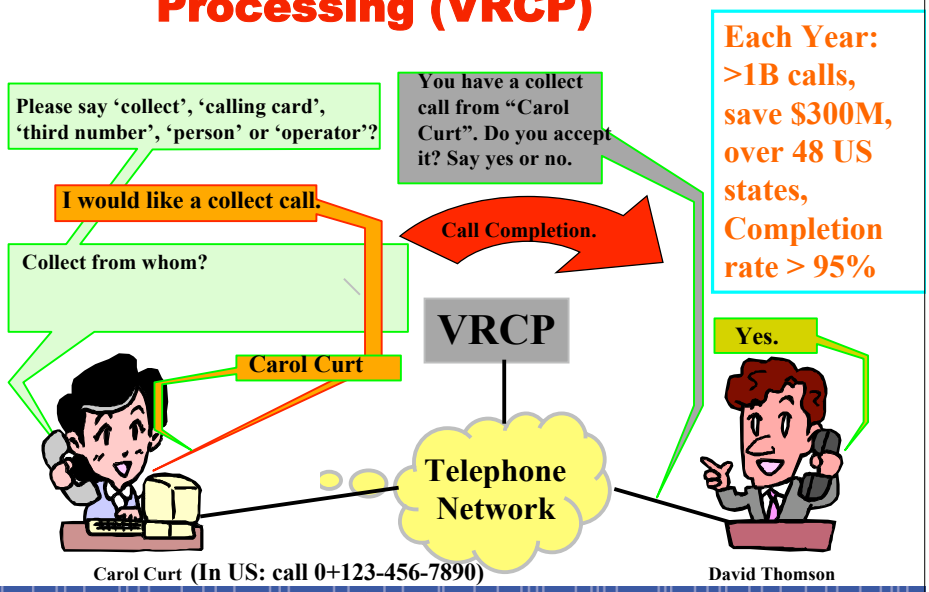
Applications of Speech and Language Technologies

- Voice typewriter – dictation systems:
  - IBM, Microsoft, Nuance, etc.
- Applications in telecommunications:
  - AT&T, Lucent Bell Labs, Nuance, Philips, Motorola, etc.
  - Automatic Call Centers.
  - Google 411, Microsoft Tellme, Free 411.
- Applications related to the Internet:
  - Google's voice search
  - Apple's Siri (voice assistant)
  - More and more to emerge

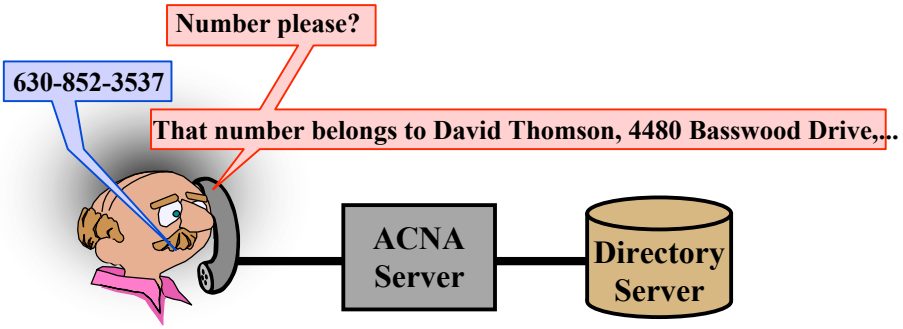
Voice Typewriter (Dictation system)

- IBM ViaVoice® System
  http://www.ibm.com/software/voice/viavoice/
- Microsoft Speech SDK and Whisper® System
  http://www.microsoft.com/speech/
- Nuance's Naturally Speaking System
  http://www.dragontalk.com/NATURAL.htm
  http://www.scansoft.com/naturallyspeaking/

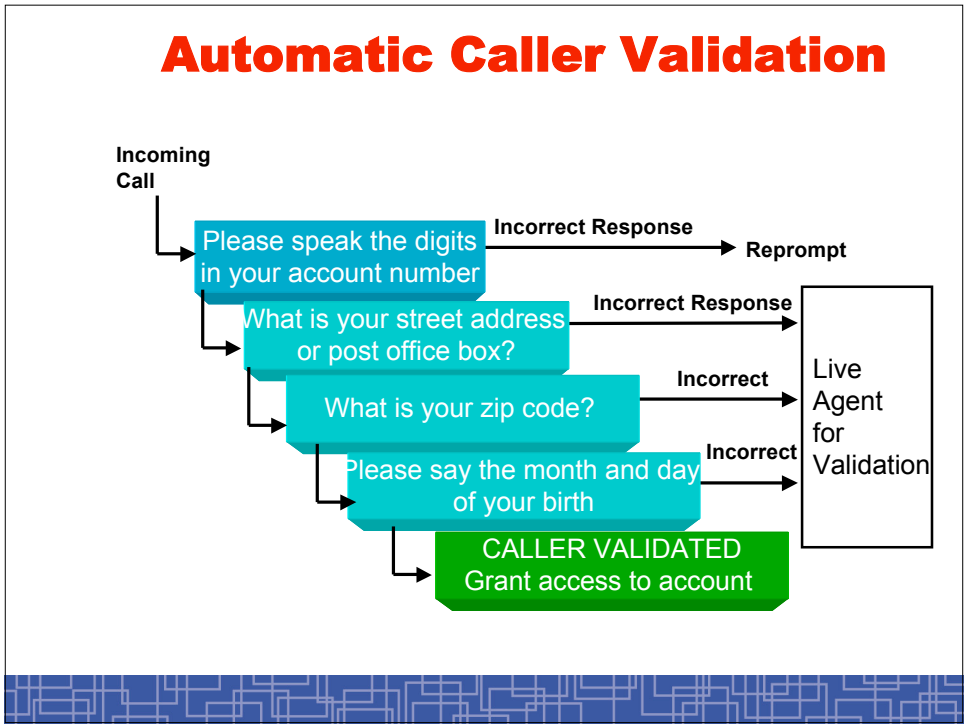
AT&T Voice Recognition Call Processing (VRCP)

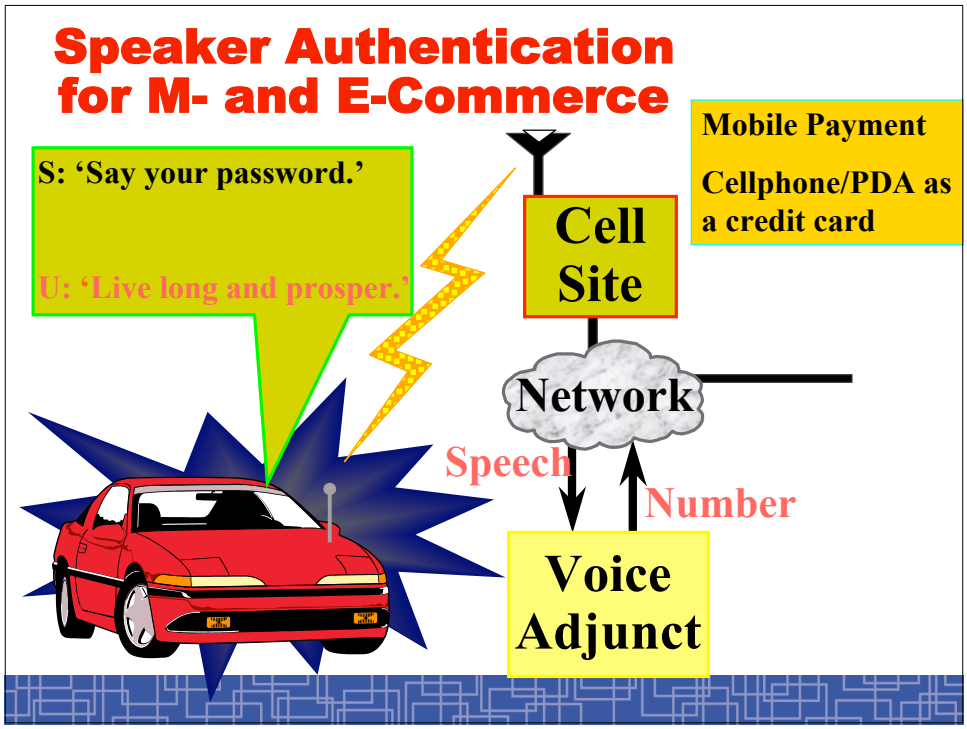


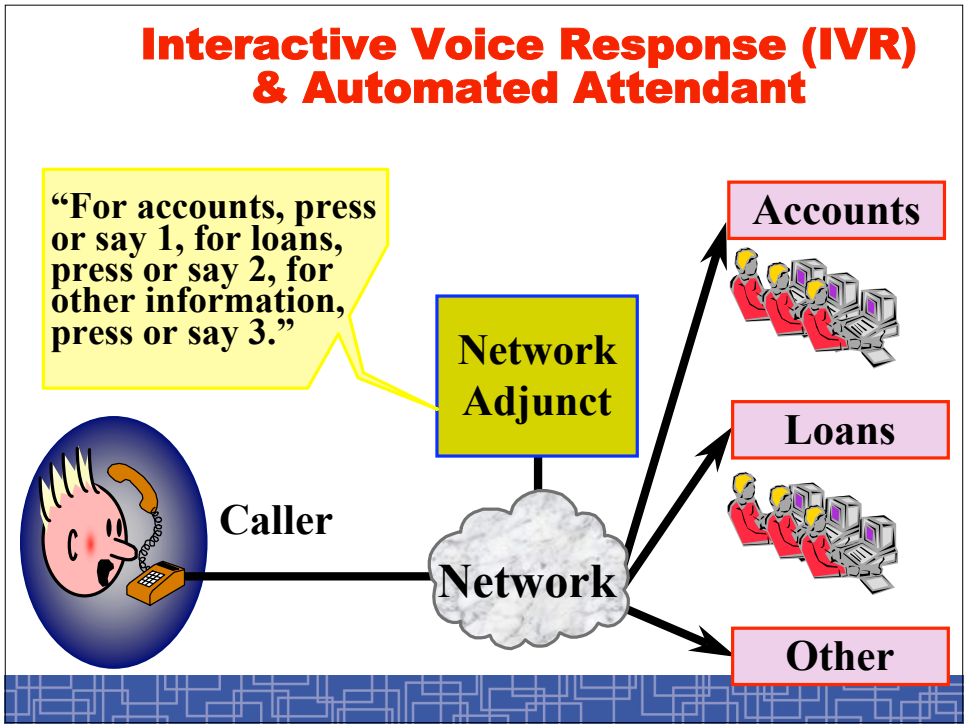
Automated Caller Name & Address (ACNA)

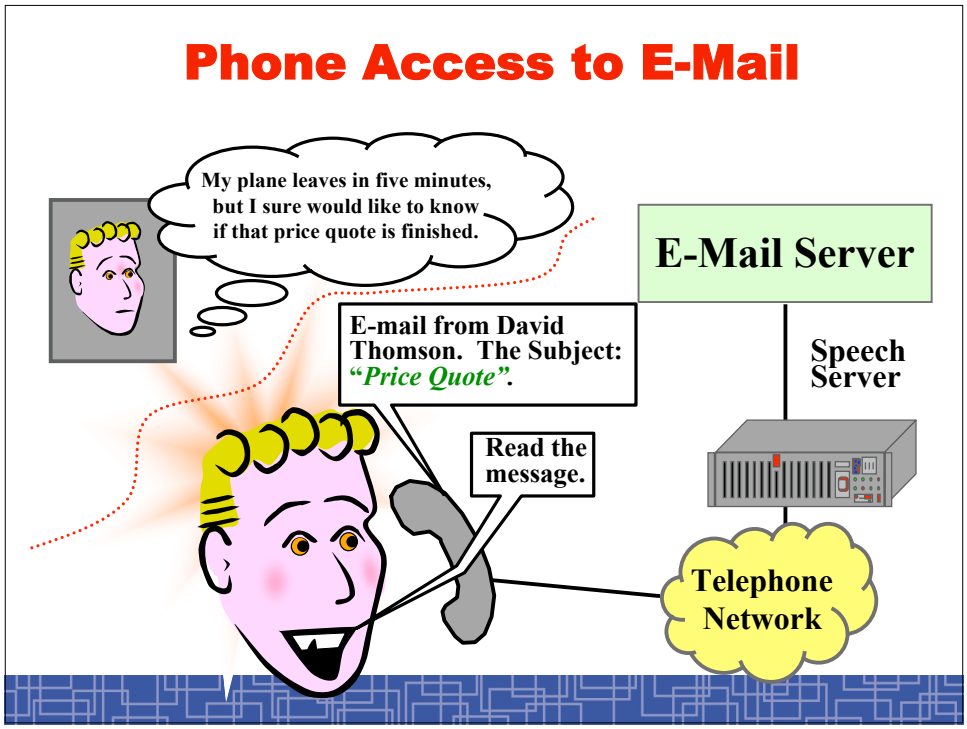


- Input via DTMF or high-accuracy connected digit recognition
- Information provided via high-quality text-to-speech
- Reverse directory service

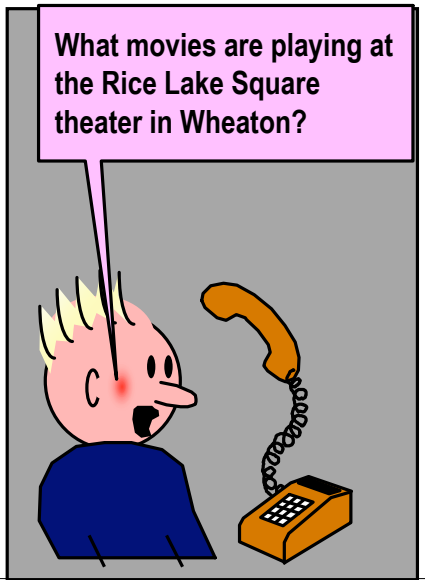

Automatic Caller Validation
Incoming Call
Please speak the digits in your account number
Incorrect Response
Reprompt
What is your street address or post office box?
Incorrect Response
What is your zip code?
Incorrect
Please say the month and day of your birth
Incorrect
Live Agent for Validation
CALLER VALIDATED
Grant access to account


Speaker Authentication for M- and E-Commerce
S: 'Say your password.'
U: 'Live long and prosper.'
Mobile Payment
Cellphone/PDA as a credit card
Cell Site
Network
Speech
Number
Voice Adjunct


Interactive Voice Response (IVR)
& Automated Attendant
"For accounts, press or say 1, for loans, press or say 2, for other information, press or say 3."
Network Adjunct
Accounts
Loans
Other
Caller
Network


Phone Access to E-Mail
My plane leaves in five minutes, but I sure would like to know if that price quote is finished.
E-Mail Server
E-mail from David Thomson. The Subject: "Price Quote".
Speech Server
Read the message.
Telephone Network

Natural Language Example: Movie Locator



Moviefone (777-film)

Other applications:

- "This is the operator, how can I help you?"
- "Hi, I'd like a large pizza with pepperoni with mushrooms toppings."
- "Play all messages from Tom Smith."
- Business (restaurant) locator, yellow pages
- Travel information systems (train/flight)
- TellMe, UA FlightInfo, Google-411
- L&H, Nuance, SpeechWorks, Philips, etc.

The Bell Labs "Natural Language Call Router"

- Input: user request (in speech or text)
- Output: desired destination related to the request (in a call center)
- Data Preparation: user (request, destination) pairs are grouped to train routing matrix using a data-driven approach
- Technologies: speech recognition, language modeling, call routing, dialogue generation

The United Airlines Flight Information System

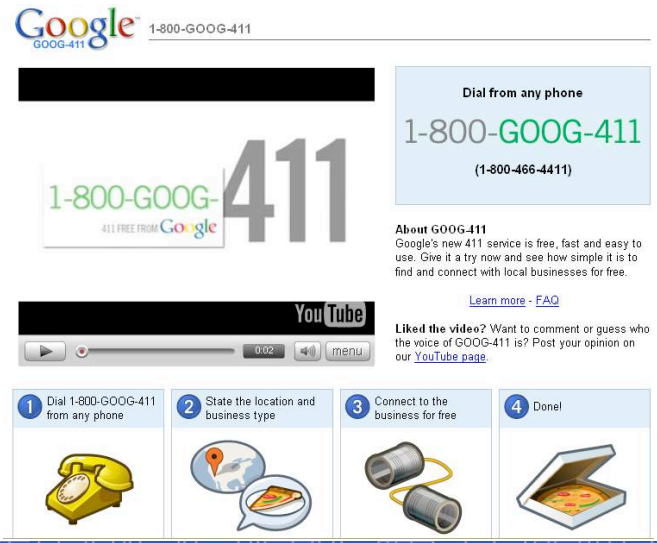
- Input: user request (in speech)
- Output: desired flight information
- Data Preparation: flight schedule info is converted into groups of finite state grammars, prompts are used to guide users
- Technologies: speech recognition, language modeling, user modeling

The Stock Information System

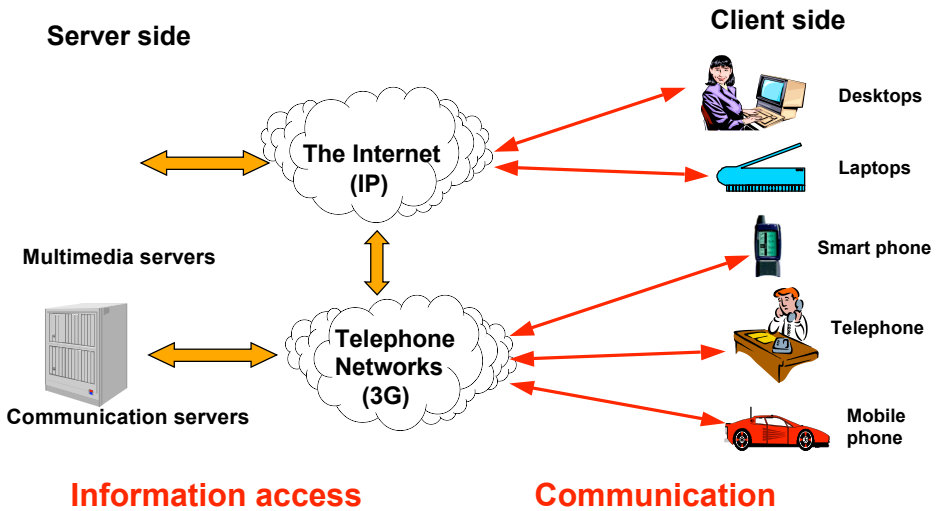
- Input: user request (in speech)
- Output: desired update stock info
- Data Preparation: convert stock names into pronunciation entries
- Technologies: speech recognition, pronunciation modeling, database, text-to-speech synthesis

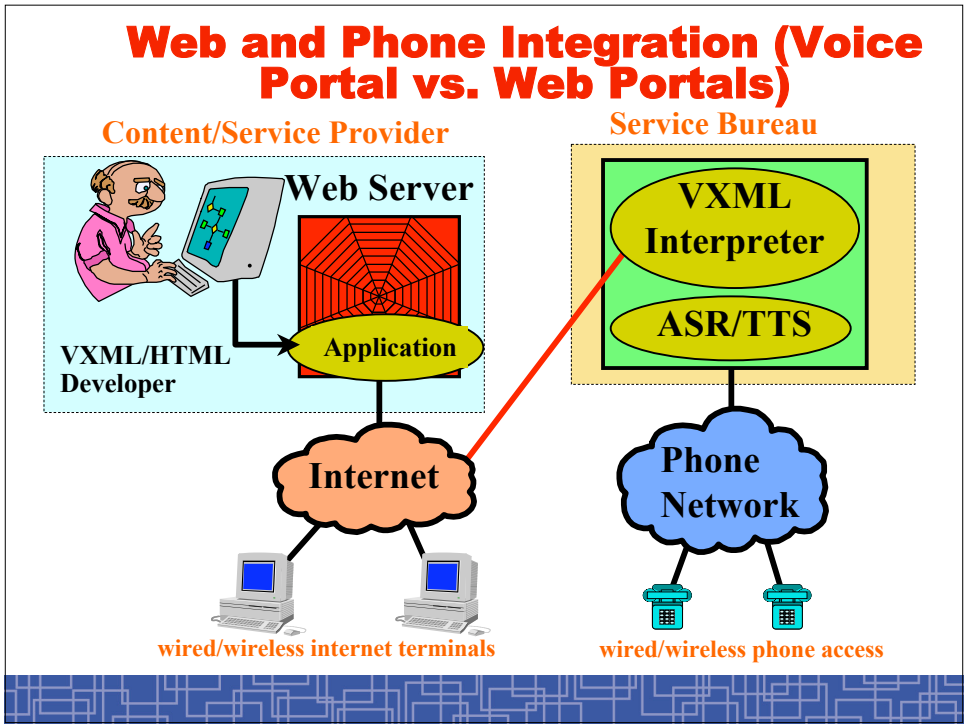
The Google 411 Service

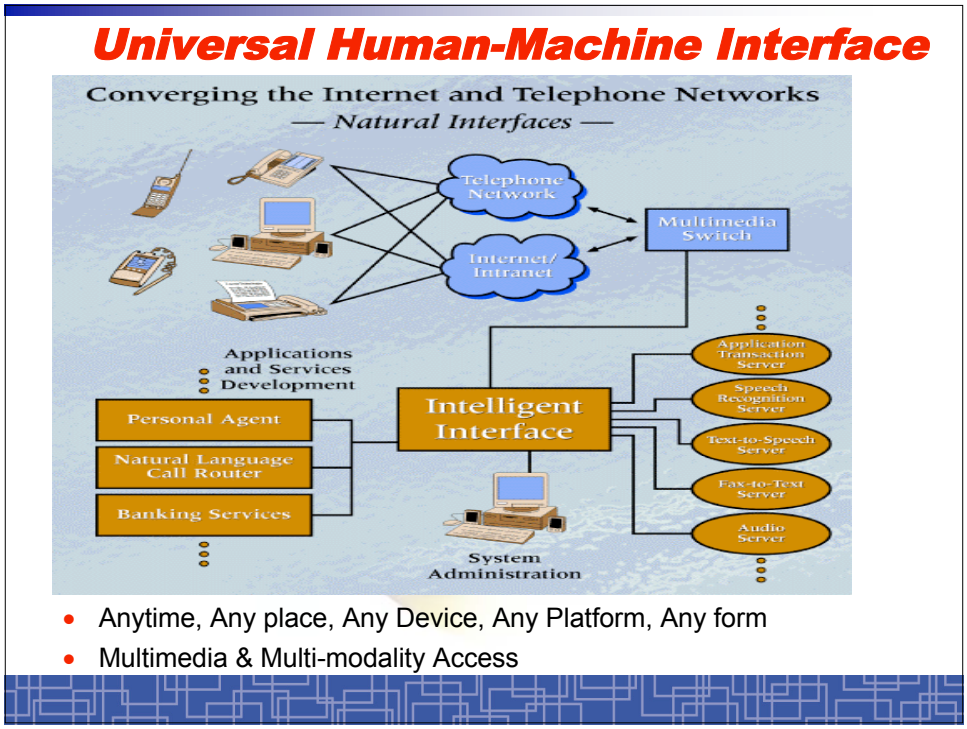
- http://www.google.com/goog411/



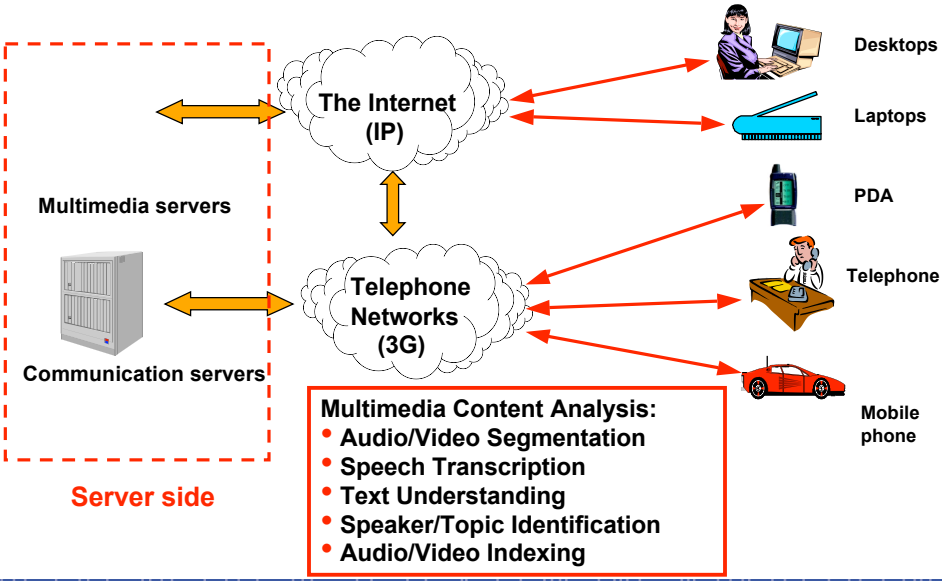
The Internet Era



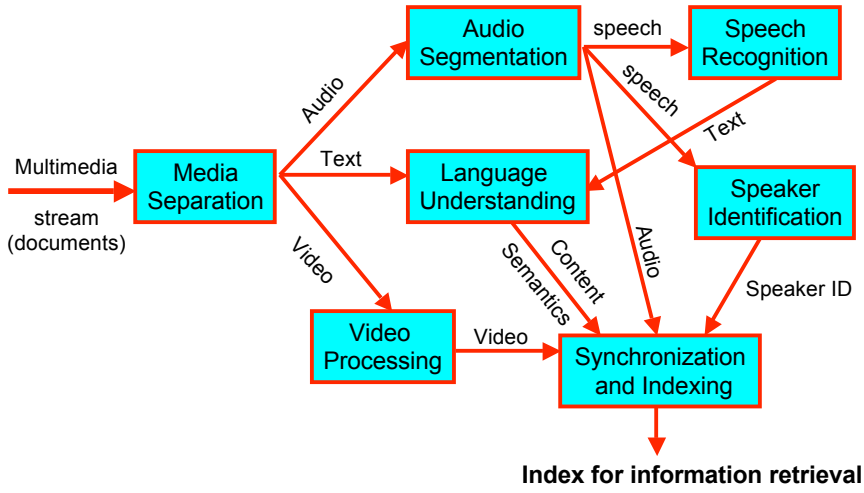

Web and Phone Integration (Voice Portal vs. Web Portals)
Content/Service Provider
Service Bureau
Web Server
Application
VXML/HTML Developer
VXML Interpreter
ASR/TTS
Internet
Phone Network
wired/wireless internet terminals
wired/wireless phone access


Universal Human-Machine Interface
Converging the Internet and Telephone Networks
— Natural Interfaces —
Telephone Network
Internet/ Intranet
Multimedia Switch
Applications and Services Development
Personal Agent
Natural Language Call Router
Banking Services
Intelligent Interface
System Administration
Application Transaction Server
Speech Recognition Server
Text-to-Speech Server
Fax-to-Text Server
Audio Server
• Anytime, Any place, Any Device, Any Platform, Any form
• Multimedia & Multi-modality Access

Research Challenges: Server side



Multimedia Content Analysis for Information retrieval



Video and Audio Segmentation

Archiving & Browsing Multimedia Data

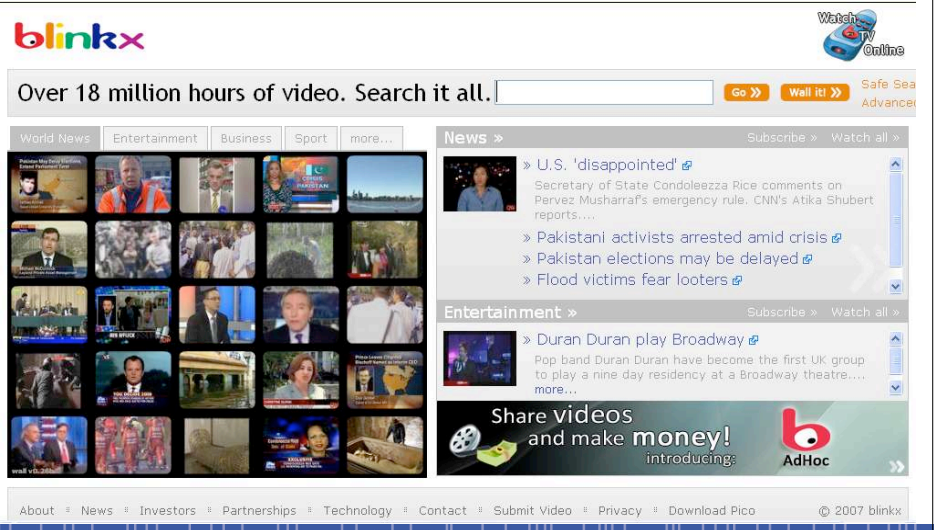
- Input: user request (in speech or text)
- Output: desired audio/video segments

- Data Preparation: video/audio segments with semantic description and (recognized) text for easy browsing, like MPEG7 descriptions (creation of indexing info for access is key)

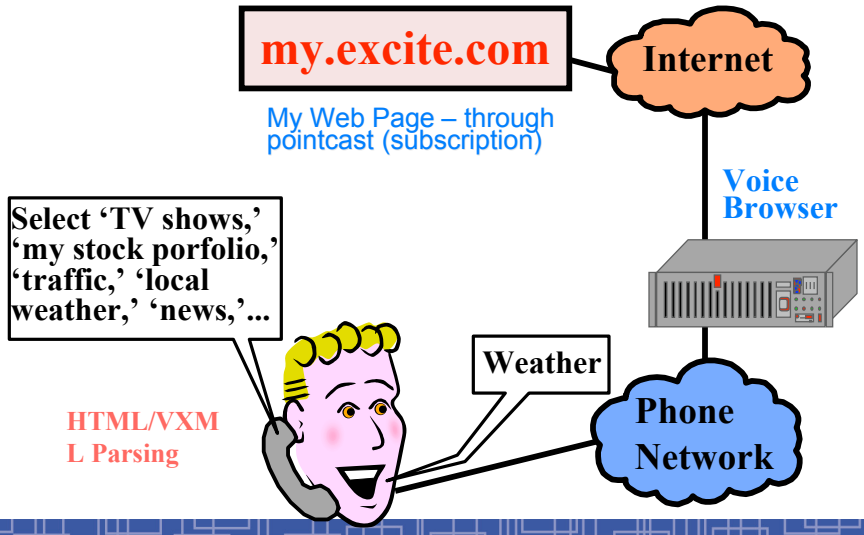
- Technologies: speech recognition, video processing, multimedia segmentation and data mining, fusion of audio/video/caption information and presentation, etc.

Video Search -- Blinkx

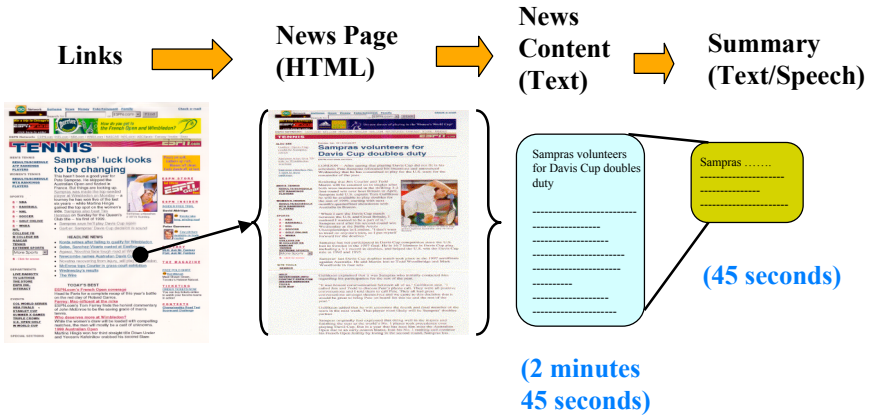
- WWW: http://www.blinkx.com/



Voice Browsing Applications



Web Document Processing, Navigation & Presentation



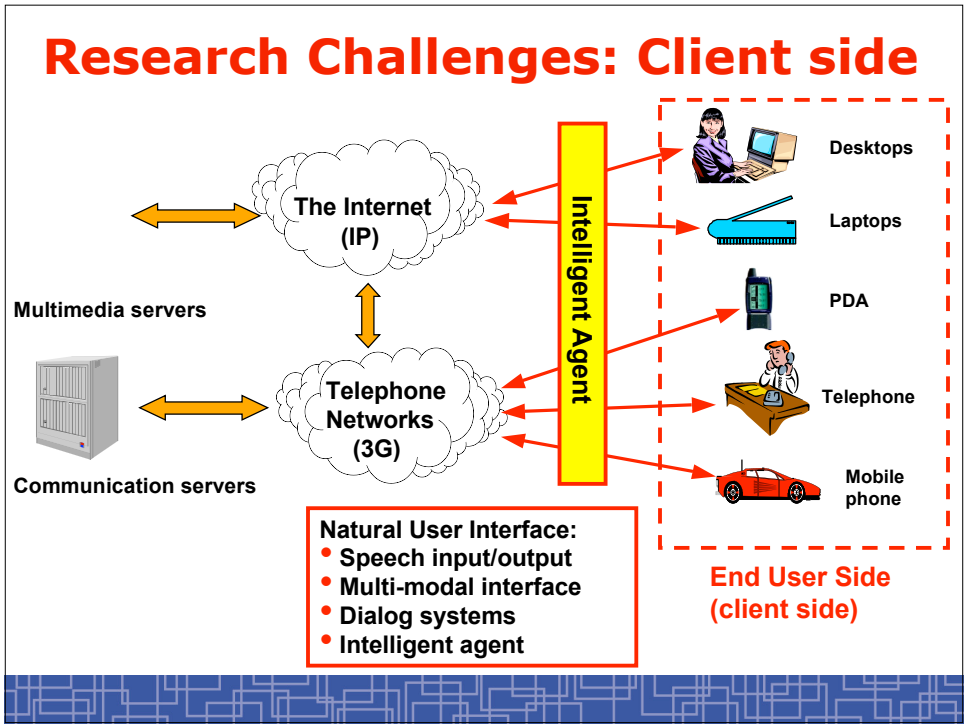
Google's Voice Search

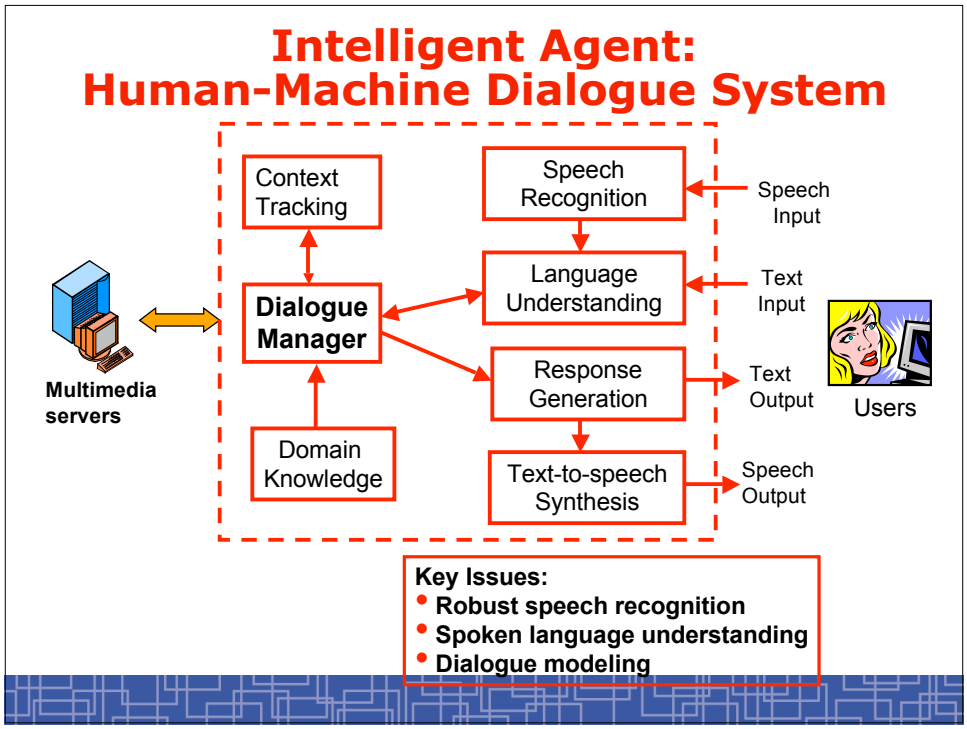
- Google's voice search

Search by voice

Speak your queries instead of typing.

Watch video Learn more

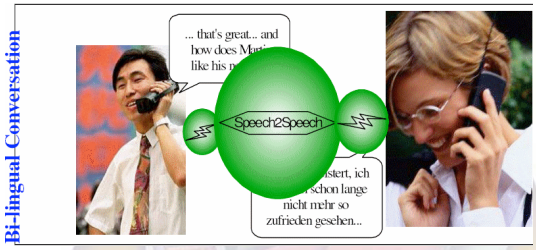

Research Challenges: Client side
The Internet
(IP)
Multimedia servers
Telephone
Networks
(3G)
Communication servers
Intelligent Agent
Desktops
Laptops
PDA
Telephone
Mobile
phone
Natural User Interface:
• Speech input/output
• Multi-modal interface
• Dialog systems
• Intelligent agent
End User Side
(client side)

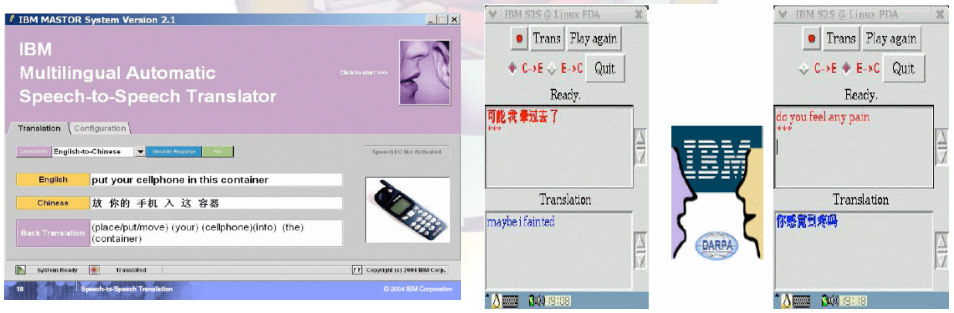

Intelligent Agent:
Human-Machine Dialogue System
Context
Tracking
Speech
Recognition
Speech
Input
Language
Understanding
Text
Input
Dialogue
Manager
Response
Generation
Text
Output
Users
Multimedia
servers
Domain
Knowledge
Text-to-speech
Synthesis
Speech
Output
Key Issues:
• Robust speech recognition
• Spoken language understanding
• Dialogue modeling

Apple's Siri

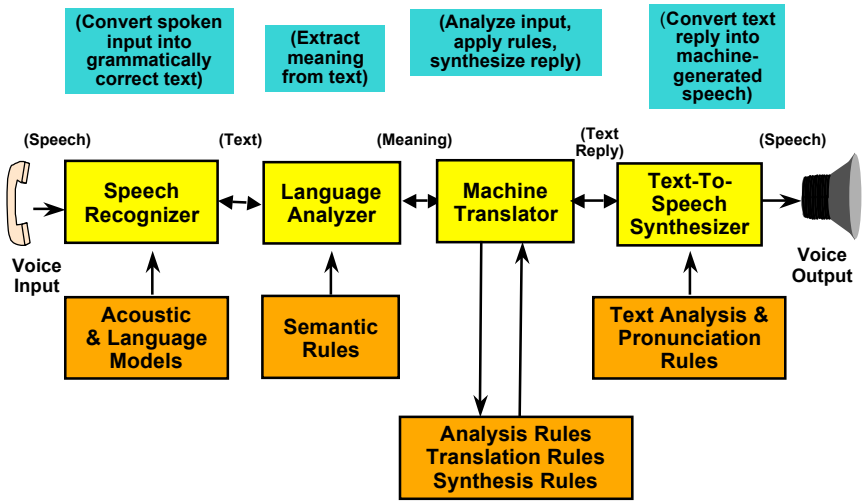
- Voice Assistant: *Siri*

Speech-To-Speech Translation





Multilingual Speech-to-Speech Translation

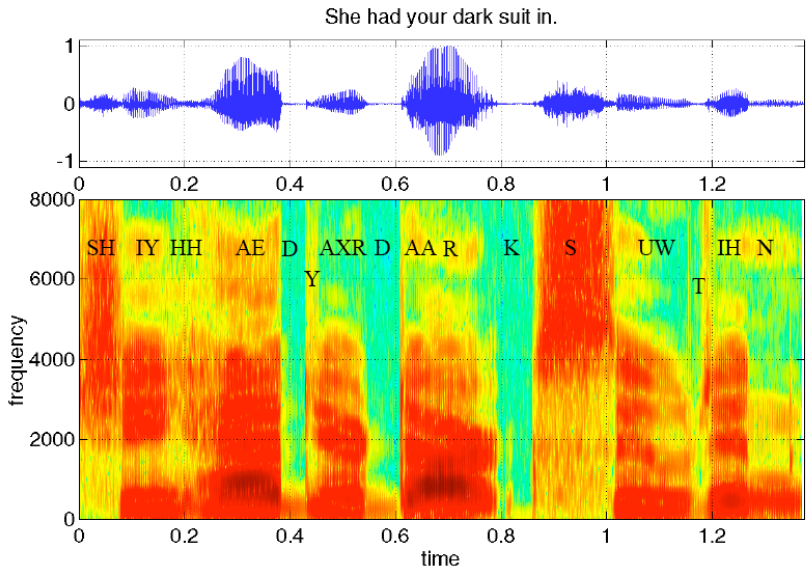


Statistical Pattern Classification

- Feature extraction:
  - Need to know objects to extract good features
  - Varies a lot among different applications (speech, audio, text, image, video, biological sequences, etc)
- Statistical model training/learning
- Inference, matching, decision

The basic theories common to various applications

Speech Signals



Fundamental Speech Units

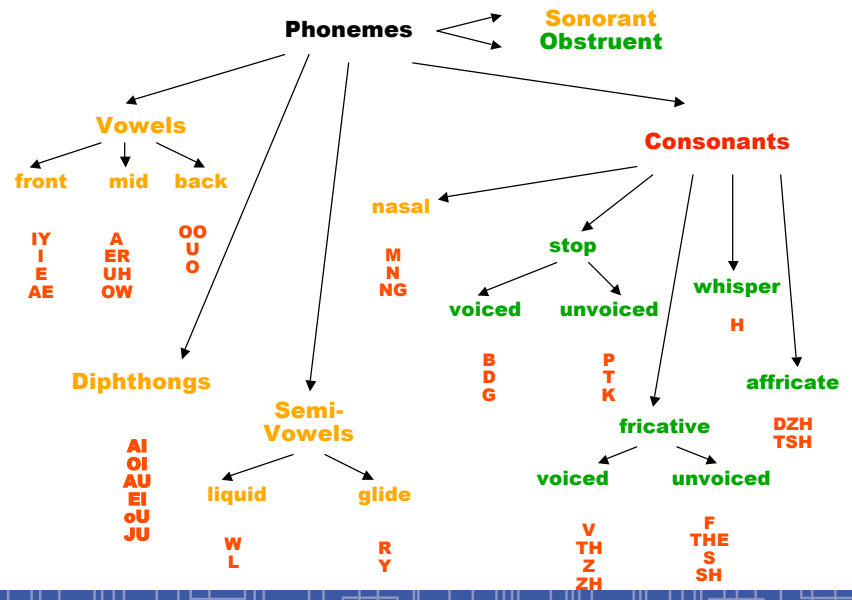
- Sentence/utterance ➔ Phrase ➔ Word ➔ Syllable ➔ Phone
- Phone
  - abstract name is called "phoneme"
  - infinite number of acoustic realization
  - monophone: context-independent phone
  - allophone: context-dependent phone
- Other considerations:
  - Language dependency
  - Task or vocabulary dependency
    - digit (small size but critically important)
- Example:

  Sentence: How do they turn out later ?

  Syllables: How do they turn out la-ter?

  Phones: h aw d uh dh eh t er n aw t l ai t er

Phonemes in American English



Coarticulation

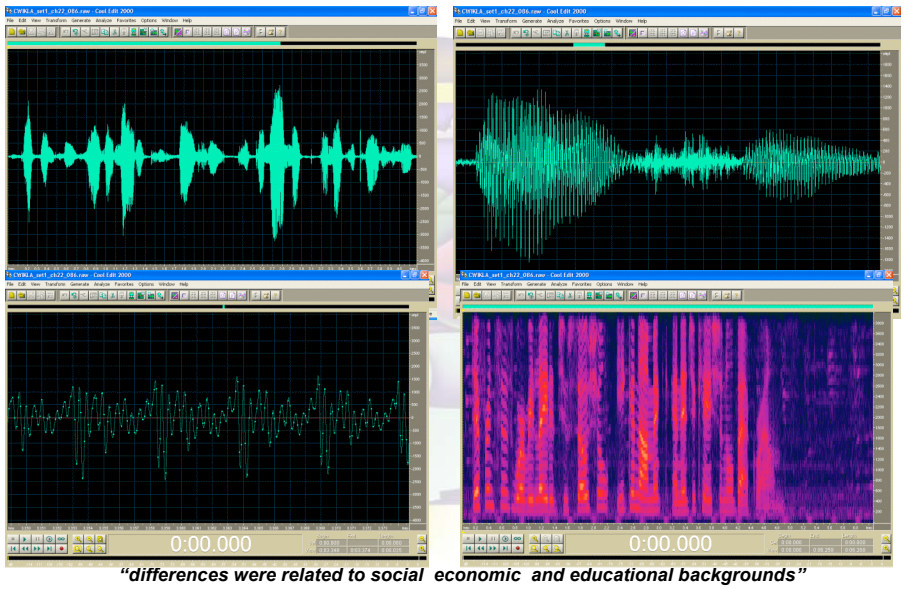
- Phones exhibit consistent acoustic characteristics if pronounced in isolated; but large acoustic variations may appear if uttered in different contexts.
- Coarticulation: acoustic realization of a phone is largely affected by its neighboring contexts.
- Reason: in speech production, articulatory gestures follow dynamics constrained by mechanical time constants associated with the articulator to keep the effort of muscles to a minimum.
- In speech recognition, how to model a phone:
  - Context-independent phone modeling – monophone: treat each phoneme equally no matter where it appears.

    (in American English ➔ 42 distinct phone units to model)
  - Context-dependent phone modeling:

    Left (or right) biphone: a phone unit varies based on left(right) neighboring phone. (American English → 42X42 distinct units)

    Triphone: a phone varies based on both left and right adjacent phones. (American English → 42X42X42 distinct units)

Acoustic Realization: speech waveform

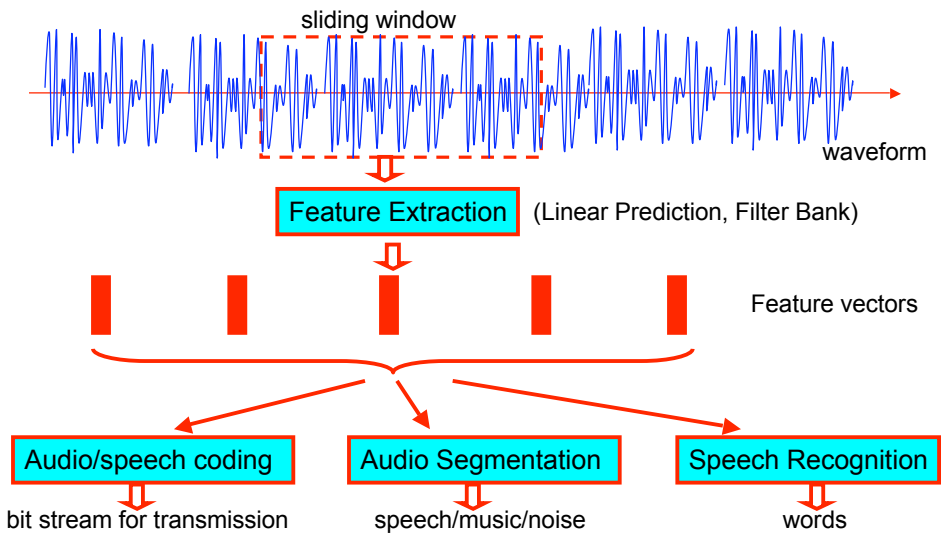


"differences were related to social economic and educational backgrounds"

Speech Waveform: Digital form

```
 ........
1400:   -529   -405   -601  -1038     51    323   -324    115    698    465
1410:    485    251    166    433   -346   -908   -303   -414   -773   -475
1420:     65    406    672    566   1160   1000   -354    519    417   -702
1430:   -728   -487   -769   -511   -719   -811    227    149   -130    476
1440:    726    439    556    273    175     49   -718   -733   -363   -661
1450:   -754    -11    318    684    782   1088    999   -108    559    409
1460:   -704   -789   -509   -833   -735   -762   -712    205     80    -88
1470:    576    847    390    552    369    170   -193   -833   -719   -481
1480:   -739   -707    143    408    811    888   1321    685   -101    815
1490:     33   -963   -795   -498   -966   -741   -809   -456    399     66
1500:     -5    817    892    294    496    279     -9   -696   -820   -698
1510:   -534   -753   -254    392    757    985   1265   1187   -266    657
1520:    517   -887  -1134   -406   -830   -987   -568   -691    239    424
1530:     15    507   1212    474    325    435    -24   -784   -741   -812
1540:   -653   -532   -278    240    982    999   1221   1196   -463    630
1550:    500  -1023  -1331   -298   -819  -1110   -597   -520    344    443
1560:     49    526   1297    406    184    367   -438   -883   -589   -949
1570:   -704    -90    -74    261   1413   1188   1332    292   -234    895
1580:   -213  -1468  -1065   -191  -1017   -838   -640     20    688    379
1590:    157    941   1170    194     88   -313   -689   -674   -952   -938
1600:   -124    257    -30   1089   1539   1506    545   -636    687    269
1610:  -1439  -1751   -253   -534  -1033   -691    -74    862    709    156
1620:    555   1408    382   -249   -600   -476   -632  -1063   -938    -96
1630:    548     57    902   1527   1922    309   -874    618    315  -1606
1640:  -1961   -326   -416   -789   -673    118    970    917    256    494
1650:   1231    439   -591   -940   -278   -724  -1031   -728    223    613
1660:    420   1039   1578   2126   -260  -1047    674     16  -2048  -1771
......
```

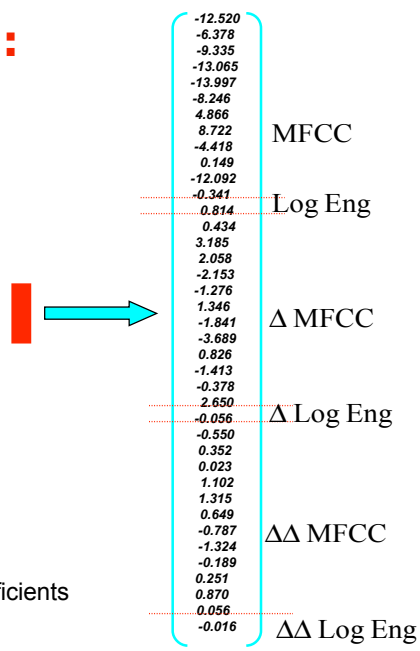
Feature Extraction: speech analysis



Feature Extraction: feature vector

- **Each feature vector is a 12—39 dimension real vector.**
- **A typical setting in most speech recognition system (typically a 39-D vector)**
  - **static MFCC's (12)**
  - **log-energy (1)**
  - **delta MFCC (12)**
  - **delta eng (1)**
  - **delta-delta MFCC (12)**
  - **delta-delta eng (1)**

MFCC: Mel-Frequency Cepstral Coefficients



What's MFCC?

Step 1:

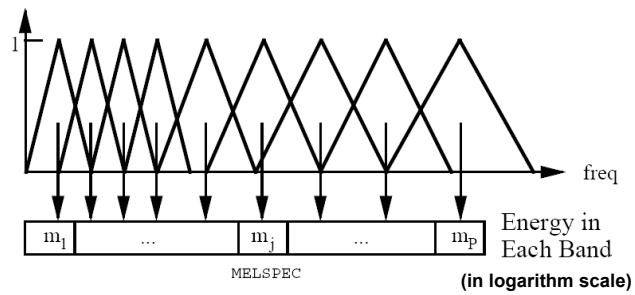


Fig. 5.3 Mel-Scale Filter Bank

Step 2: DCT (Discrete Cosine Transform) to de-correlate

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

MFCC: Mel-Frequency Cepstral Coefficients

Energy Measure

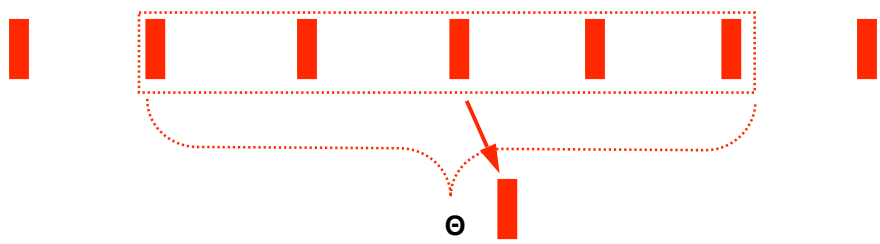
- **For each frame, log-energy is calculated as:**

$$E = log \sum_{n=1}^{N} s_n^2$$

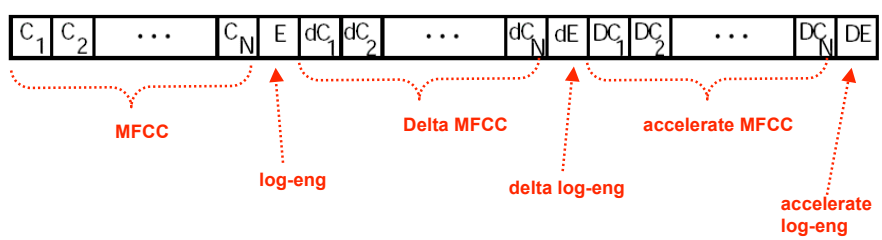
What Delta (Δ) and Acceleration(ΔΔ) Coefficients?

1. Delta coefficients: difference of MFCC among consecutive frames.
2. Acceleration coefficients: difference of delta among consecutive frames

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2}$$



Feature Vector Layout for each frame of speech



In most cases, N=12 ➔ 39-dimension feature vector for each frame

Spoken Language Processing

- Style: written vs. spoken language
  - Written → formal; spoken → casual
- Disfluencies in spoken language:
  - Filled pauses: *um*
  - Repetitions: *… the—the …*
  - Repairs: *… on Thursday – on Friday …*
  - False starts: *I like … – what I always get is …*
- Lots of ungrammatical sentences exist in spoken language.
- In spoken language system:
  - speech recognition errors
- Obviously, spoken language processing is much harder.
- Our goal: build spoken language systems in some very constrained domains to perform <u>shallow understanding</u>:
  - Topic identification
  - Key-word spotting ➔ to obtain gist and/or key message.
  - etc.

Feature Extraction to represent Text Document

- Text document ==> bag of words
- Key words vs. stop words
- "Russia cancels an <u>exhibition</u> of Russian <u>art</u> in London over fears the <u>art</u> could be seized to settle legal claims."
- => "exhibition(1) art(2)" <=> "art(2) exhibition(1)"
- Raw Feature vector:
- May need some normalization:
  - e.g. *TF-IDF,…*

$$\begin{bmatrix} 0 \\ \vdots \\ 2 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \end{bmatrix}$$

Statistical Machine Translation

- **Score and rank all possible combinations…**