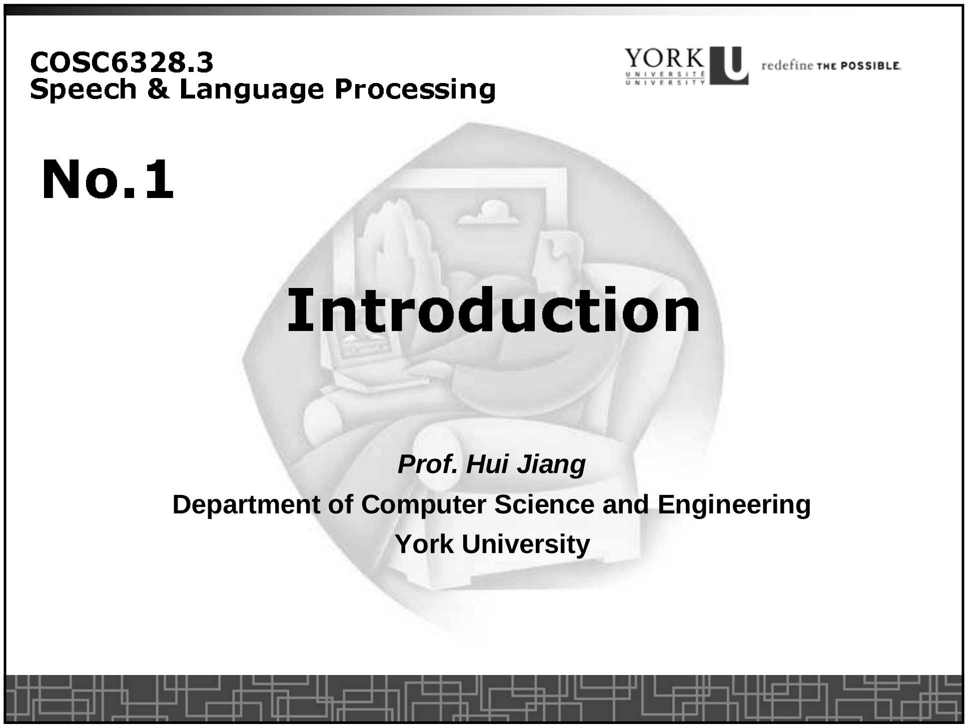


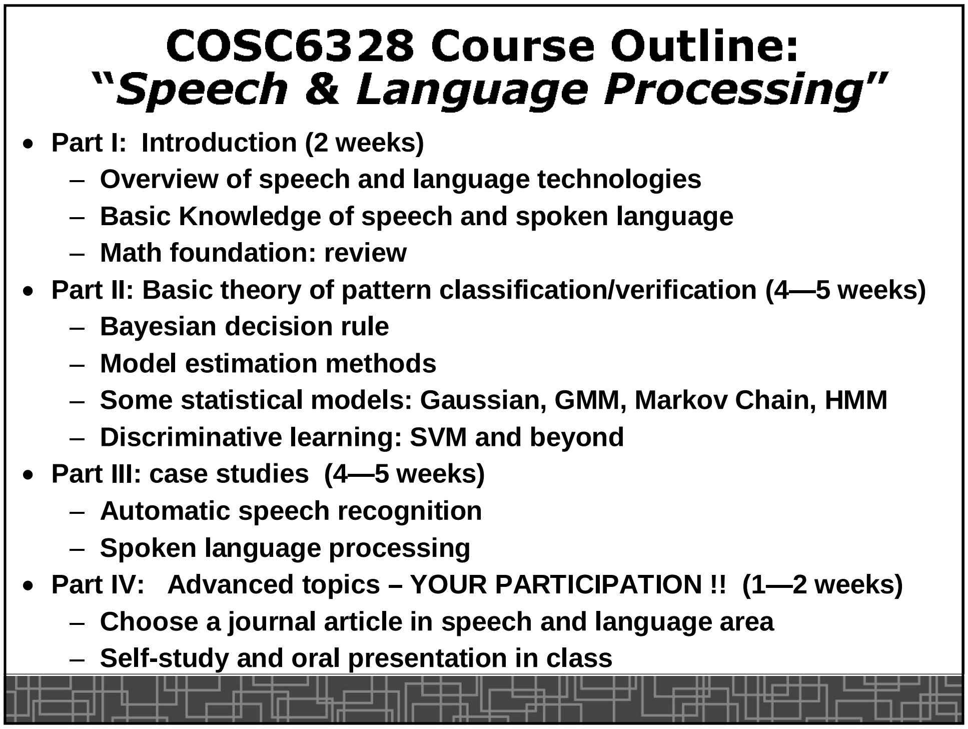
YORK UNIVERSITÉ UNIVERSITY
redefine THE POSSIBLE.

COSC6328.3
Speech & Language Processing

No.1

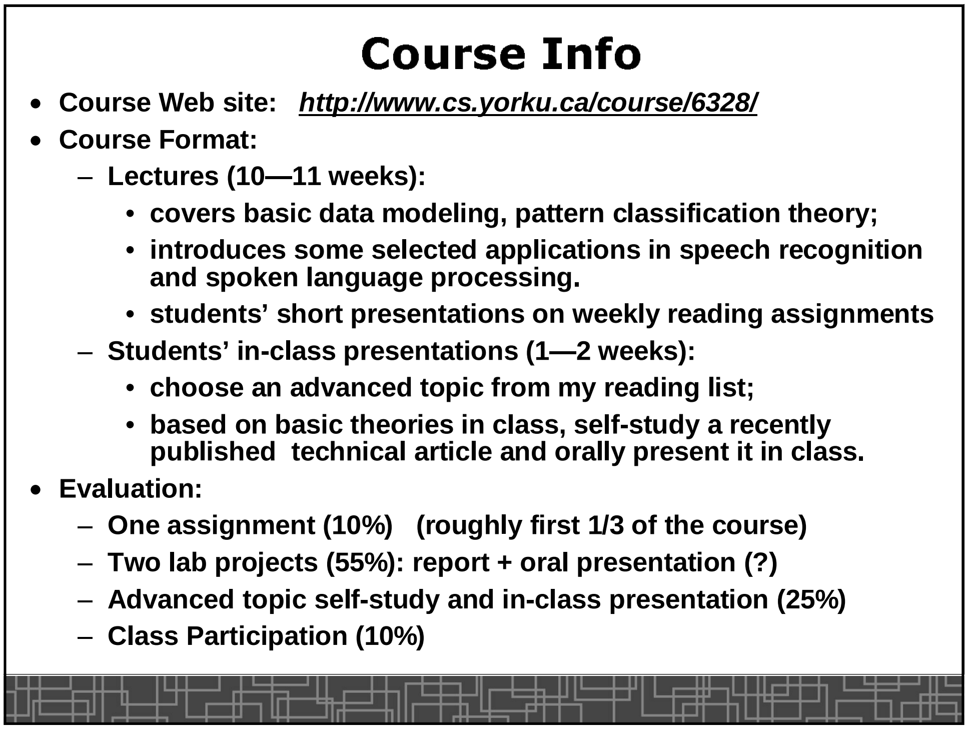
Introduction

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



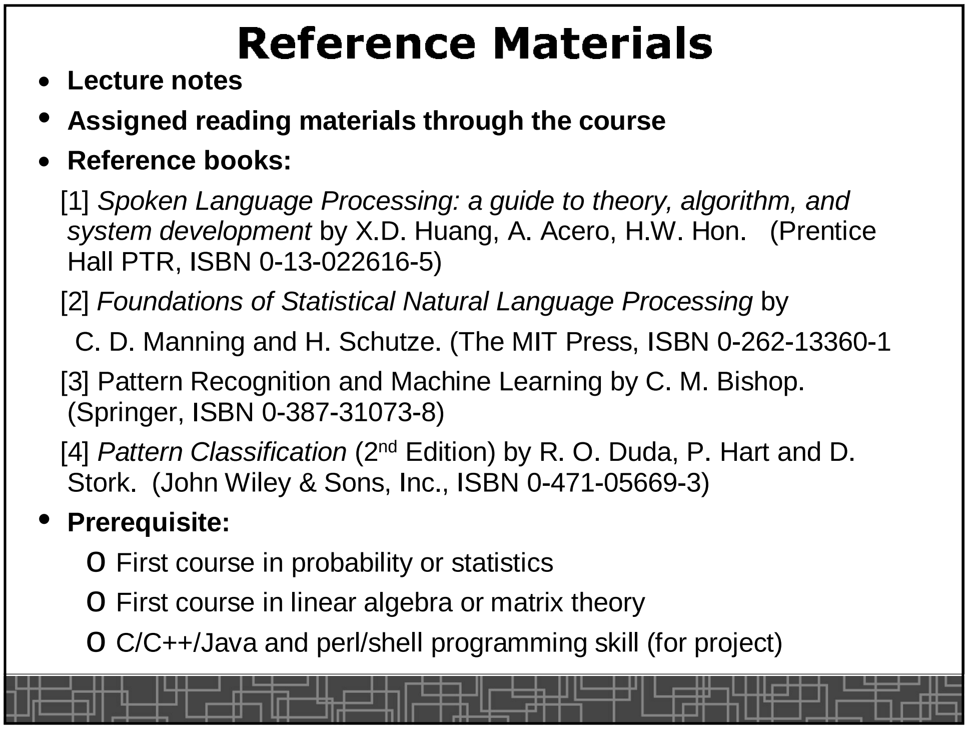
COSC6328 Course Outline: *"Speech & Language Processing"*

- **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**
  - **Some statistical models: Gaussian, GMM, Markov Chain, HMM**
  - **Discriminative learning: SVM and beyond**
- **Part III: case studies (4—5 weeks)**
  - **Automatic speech recognition**
  - **Spoken language processing**
- **Part IV: Advanced topics – YOUR PARTICIPATION !! (1—2 weeks)**
  - **Choose a journal article in speech and language area**
  - **Self-study and oral presentation in class**



Course Info

- Course Web site: ***http://www.cs.yorku.ca/course/6328/***
- Course Format:
  - Lectures (10—11 weeks):
    - covers basic data modeling, pattern classification theory;
    - introduces some selected applications in speech recognition and spoken language processing.
    - students' short presentations on weekly reading assignments
  - Students' in-class presentations (1—2 weeks):
    - choose an advanced topic from my reading list;
    - based on basic theories in class, self-study a recently published technical article and orally present it in class.
- Evaluation:
  - One assignment (10%) (roughly first 1/3 of the course)
  - Two lab projects (55%): report + oral presentation (?)
  - Advanced topic self-study and in-class presentation (25%)
  - Class Participation (10%)



Reference Materials

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

[1] *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)

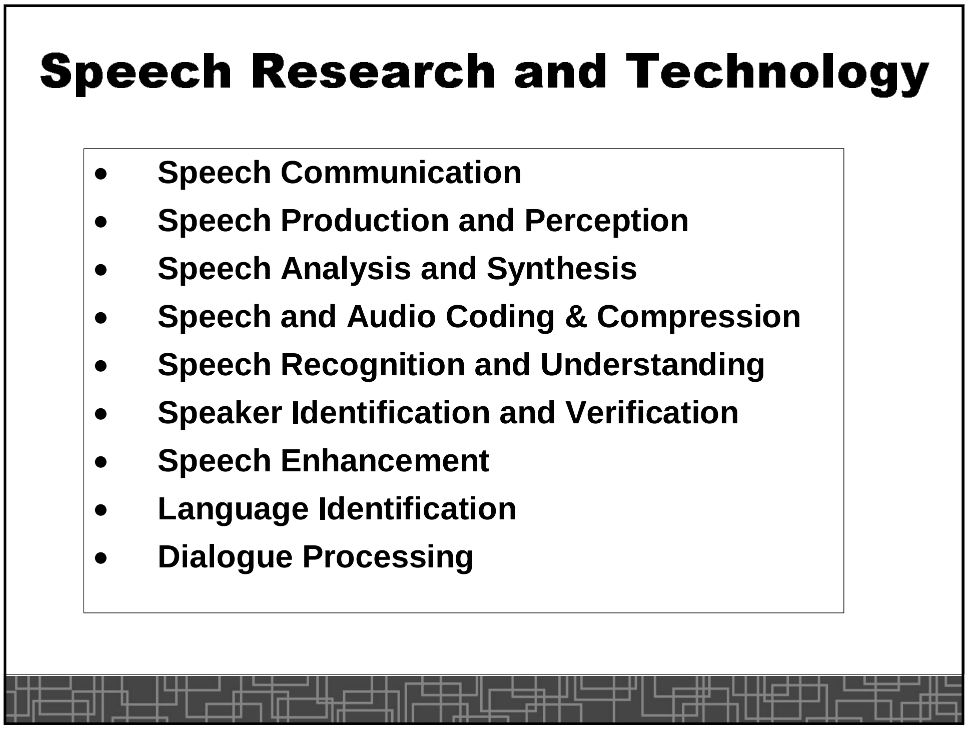
[2] *Foundations of Statistical Natural Language Processing* by

C. D. Manning and H. Schutze. (The MIT Press, ISBN 0-262-13360-1

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

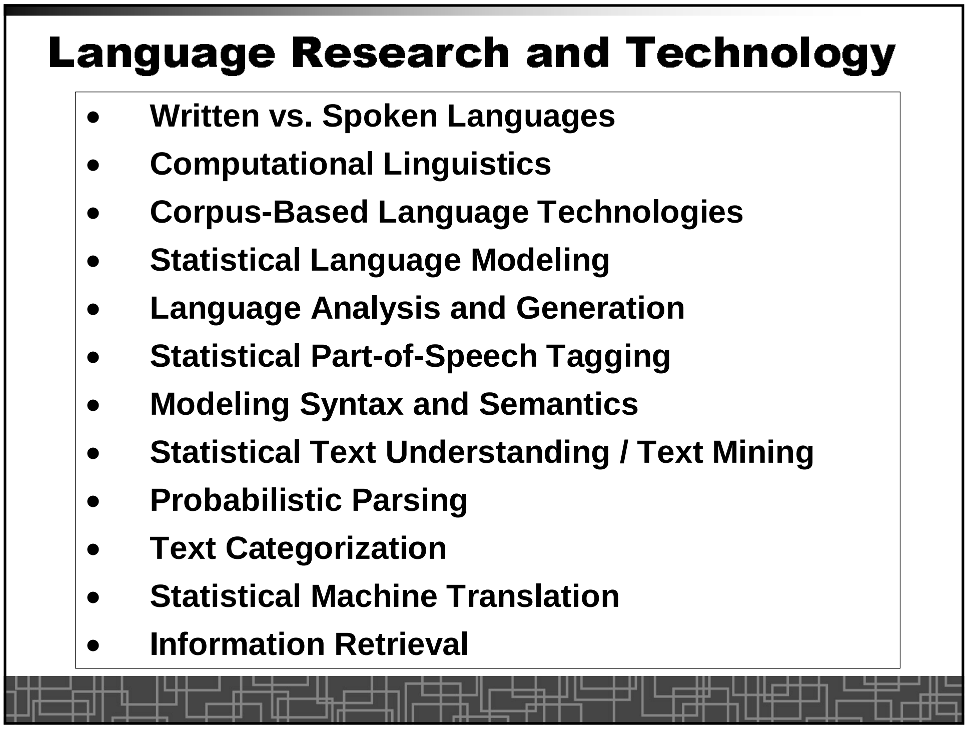
[4] *Pattern Classification* (2nd Edition) by R. O. Duda, P. Hart and D. Stork. (John Wiley & Sons, Inc., ISBN 0-471-05669-3)

- **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)



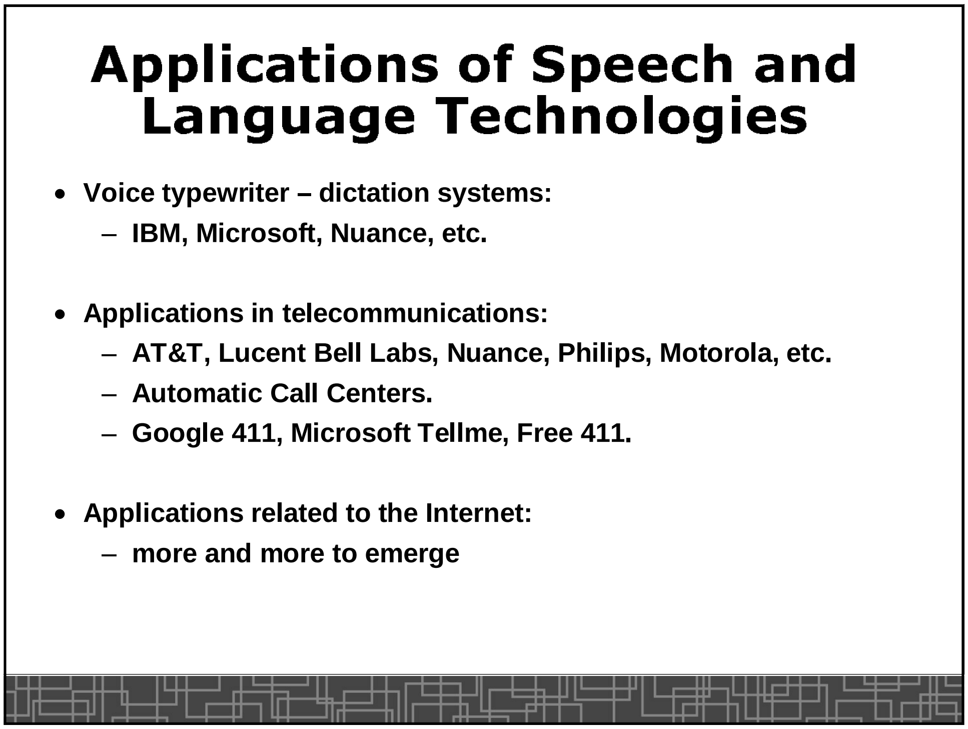
Speech Research and Technology

- 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



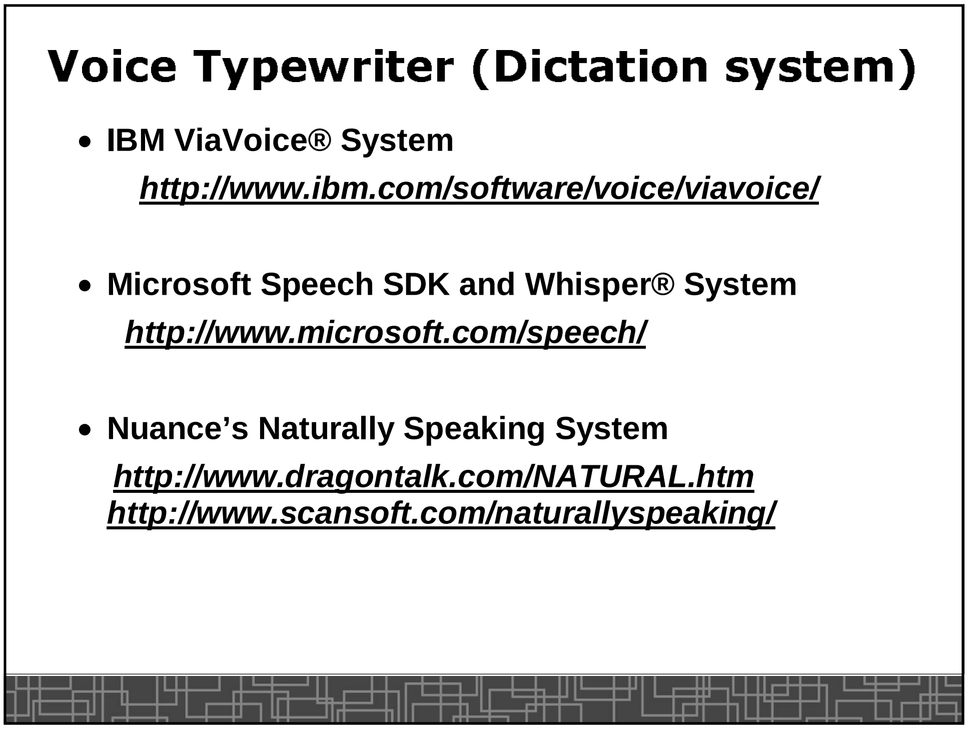
Language Research and Technology

- 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:
  - more and more to emerge



Voice Typewriter (Dictation system)

- IBM ViaVoice® System

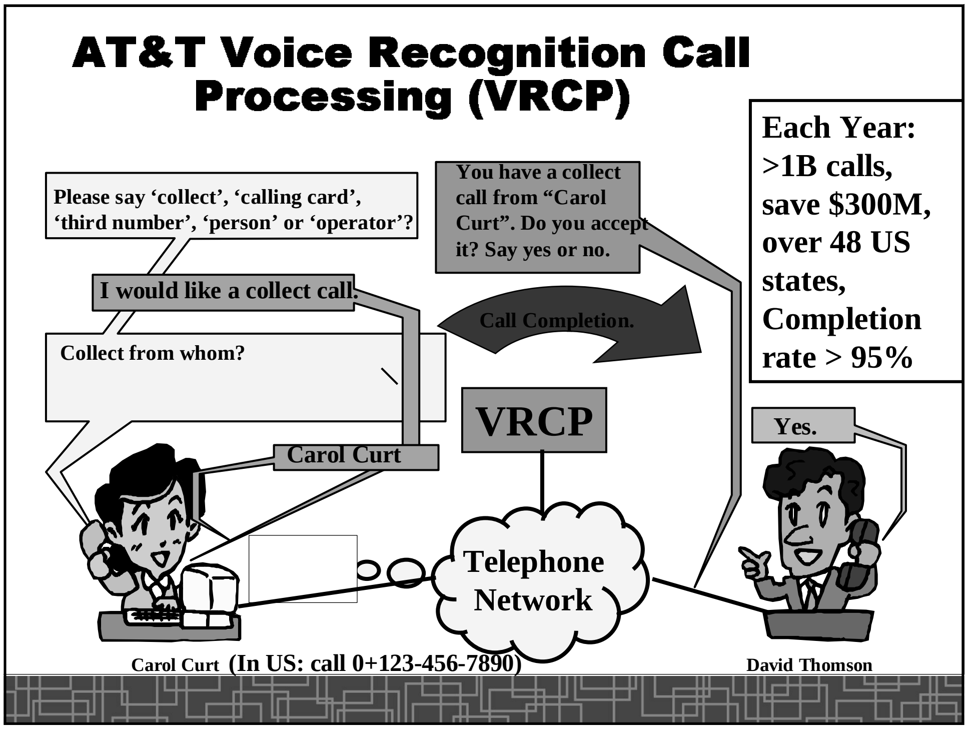
  http://www.ibm.com/software/voice/viavoice/

- 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)

Each Year:
>1B calls,
save $300M,
over 48 US
states,
Completion
rate > 95%

Please say ‘collect’, ‘calling card’, ‘third number’, ‘person’ or ‘operator’?

I would like a collect call.

Collect from whom?

Carol Curt

You have a collect call from “Carol Curt”. Do you accept it? Say yes or no.

Call Completion.

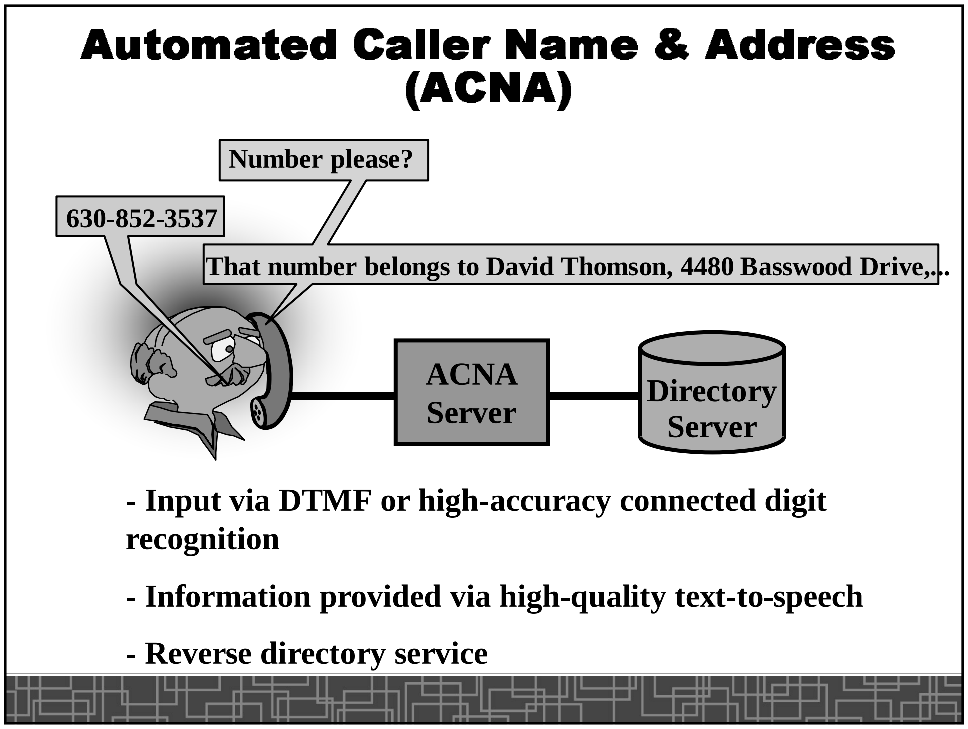
VRCP

Telephone Network

Yes.

Carol Curt (In US: call 0+123-456-7890)

David Thomson



Automated Caller Name & Address (ACNA)

Number please?

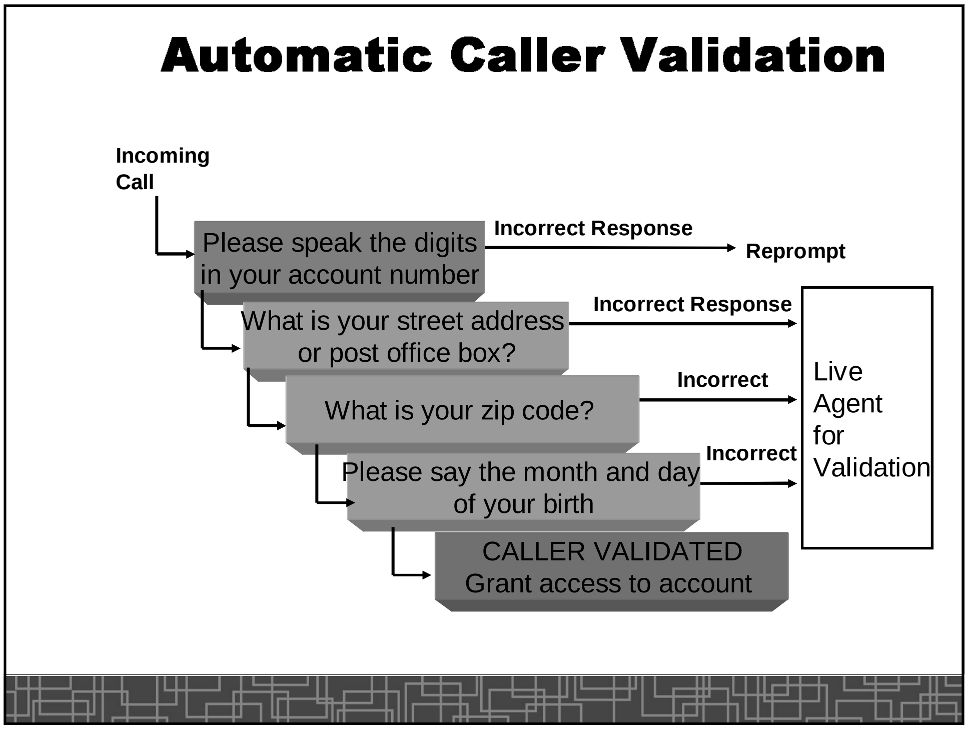
630-852-3537

That number belongs to David Thomson, 4480 Basswood Drive,...

ACNA Server

Directory Server

- 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?

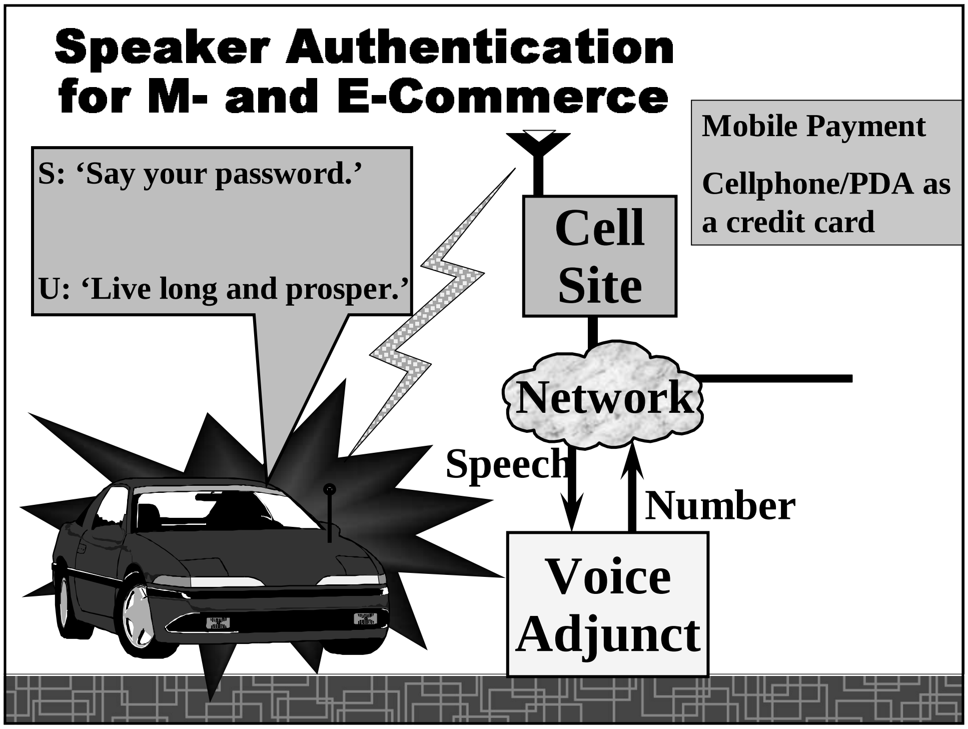
Incorrect

Please say the month and day of your birth

Incorrect

Live Agent for Validation

CALLER VALIDATED
Grant access to account



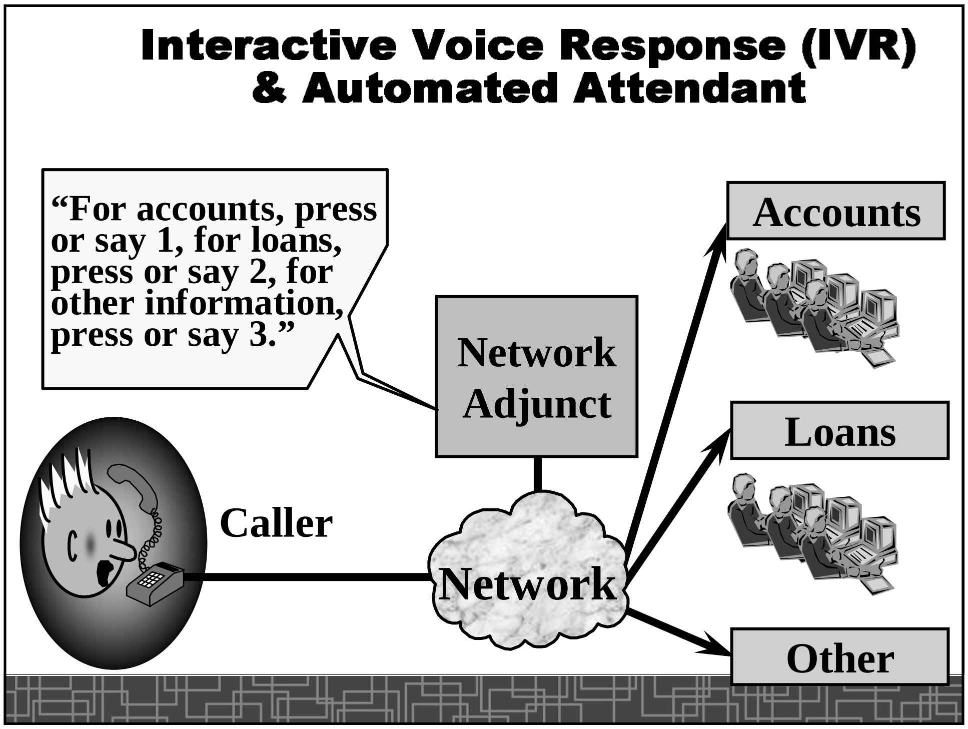
Speaker Authentication for M- and E-Commerce

S: 'Say your password.'

U: 'Live long and prosper.'

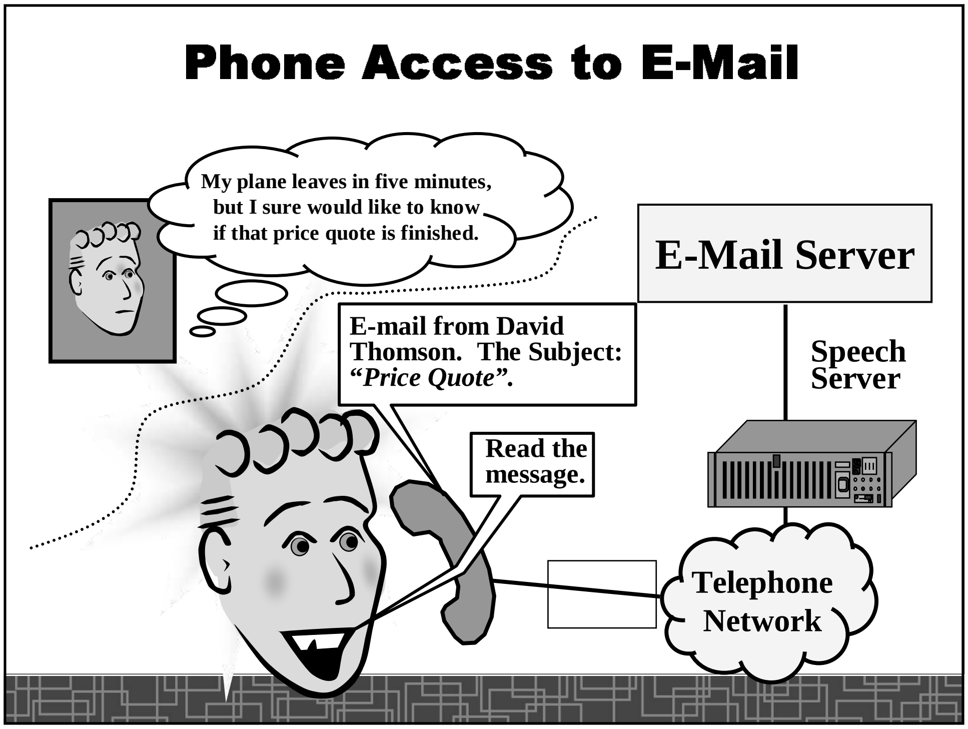
Mobile Payment

Cellphone/PDA as a credit card



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."

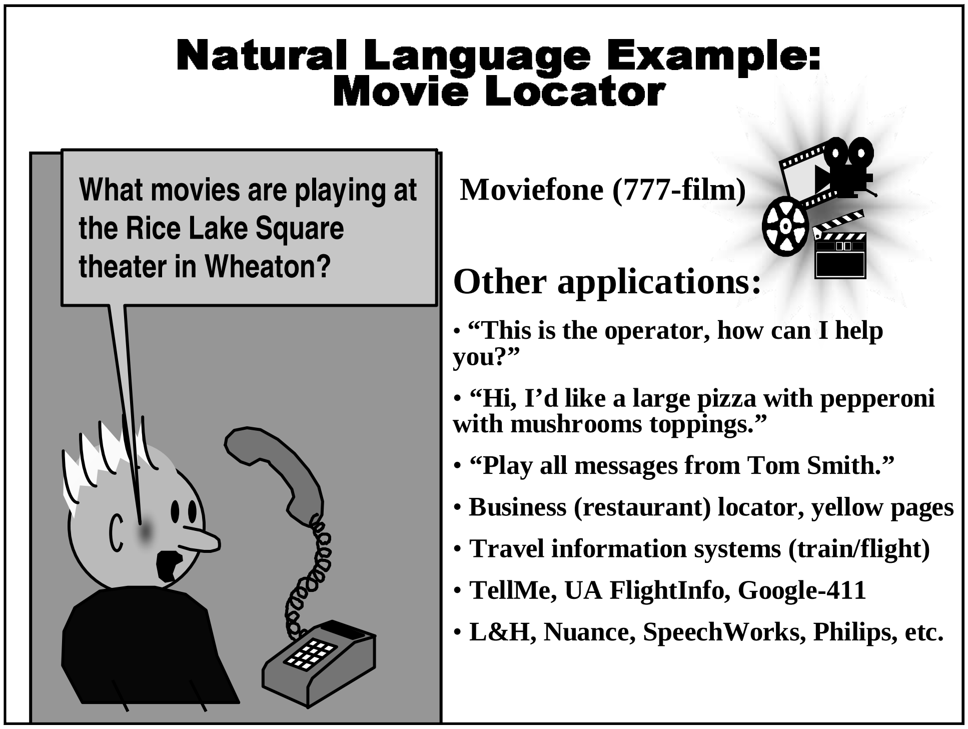


Phone Access to E-Mail

My plane leaves in five minutes, but I sure would like to know if that price quote is finished.

E-mail from David Thomson. The Subject: *"Price Quote"*.

Read the message.



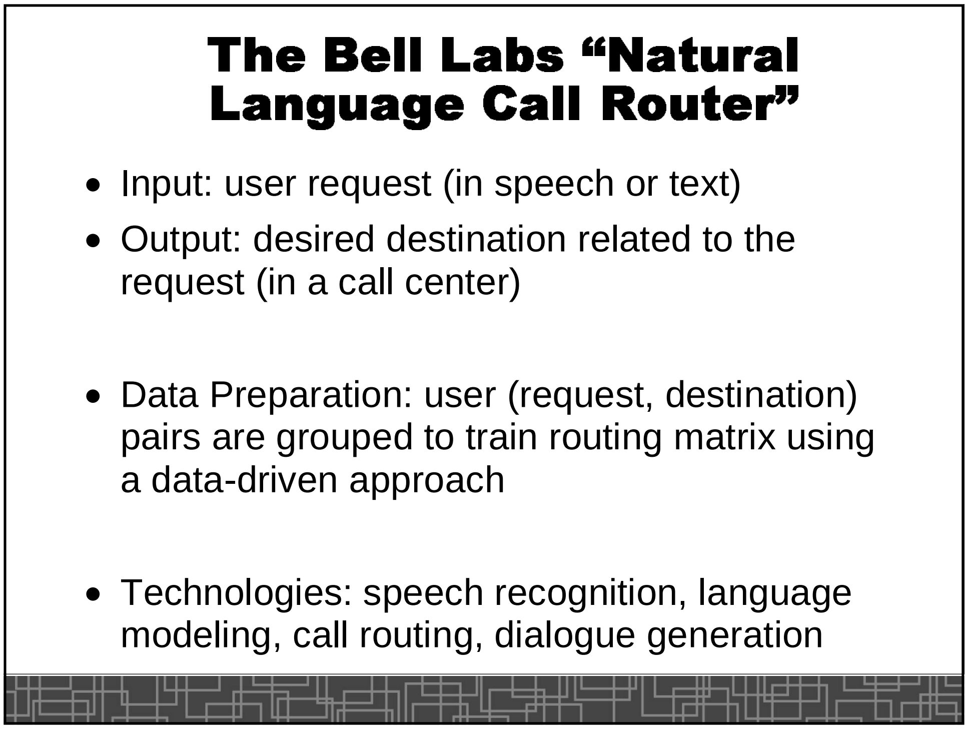
Natural Language Example: Movie Locator

What movies are playing at the Rice Lake Square theater in Wheaton?

Moviefone (777-film)

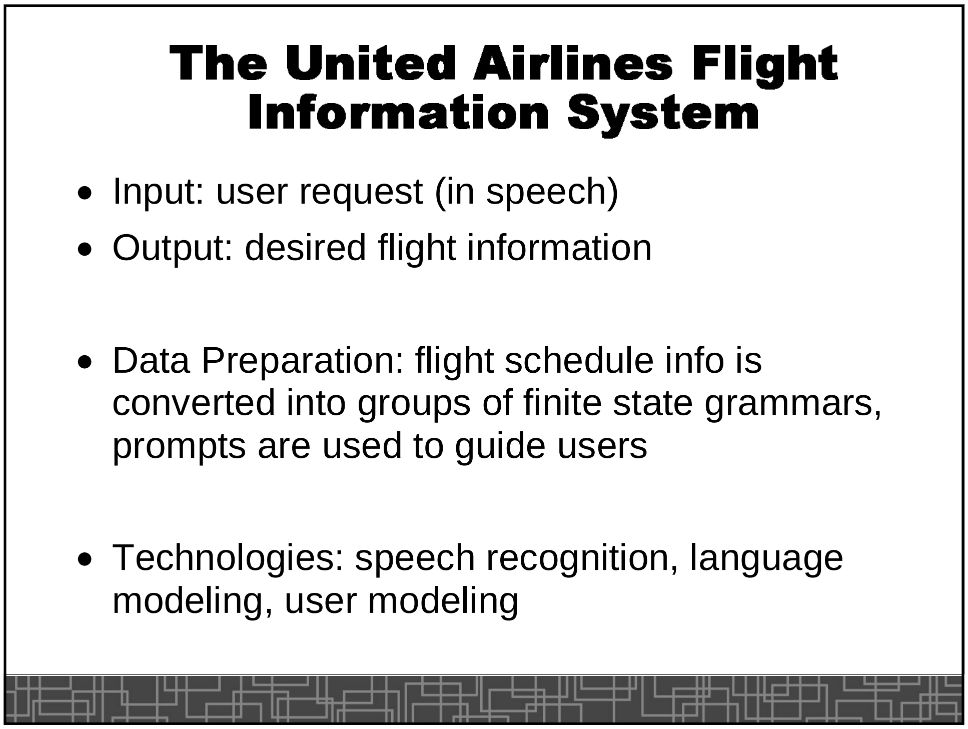
Other applications:

- "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.



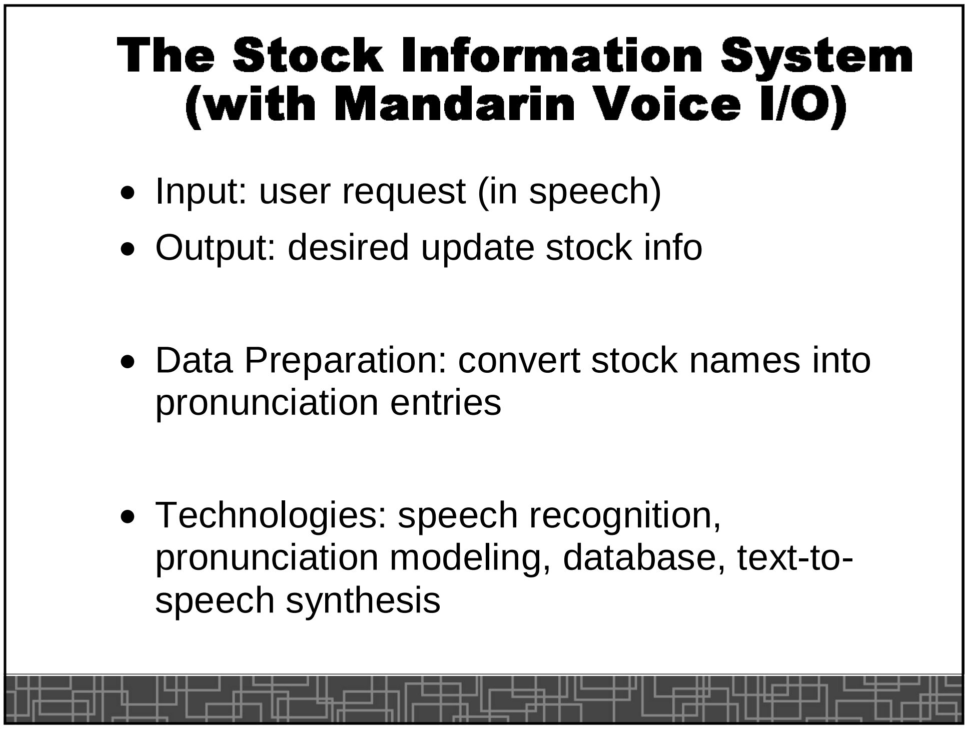
The Bell Labs "Natural Language Call Router"

- 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

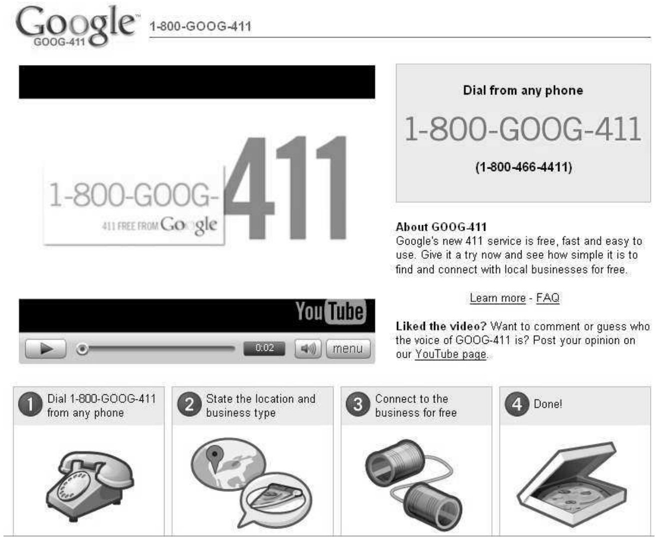


The Stock Information System (with Mandarin Voice I/O)

- 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/



Dial from any phone

1-800-GOOG-411

(1-800-466-4411)

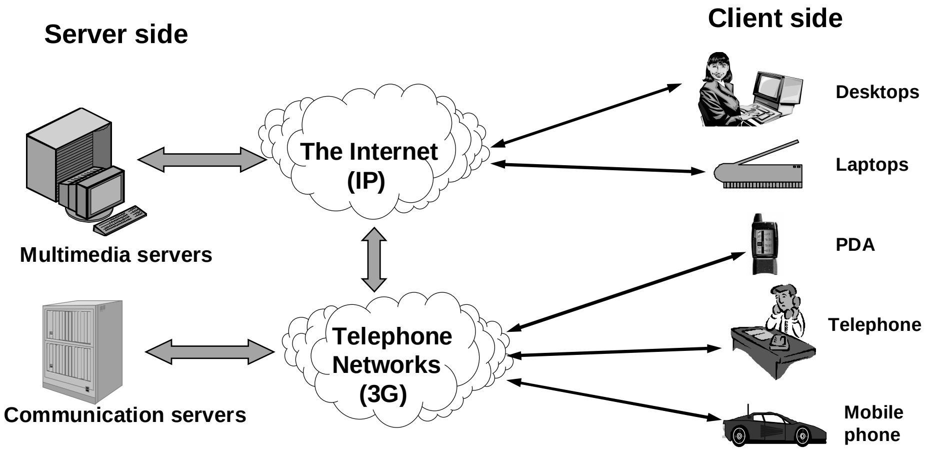
About GOOG-411
Google's new 411 service is free, fast and easy to use. Give it a try now and see how simple it is to find and connect with local businesses for free.

Learn more - FAQ

Liked the video? Want to comment or guess who the voice of GOOG-411 is? Post your opinion on our YouTube page.

1. Dial 1-800-GOOG-411 from any phone
2. State the location and business type
3. Connect to the business for free
4. Done!

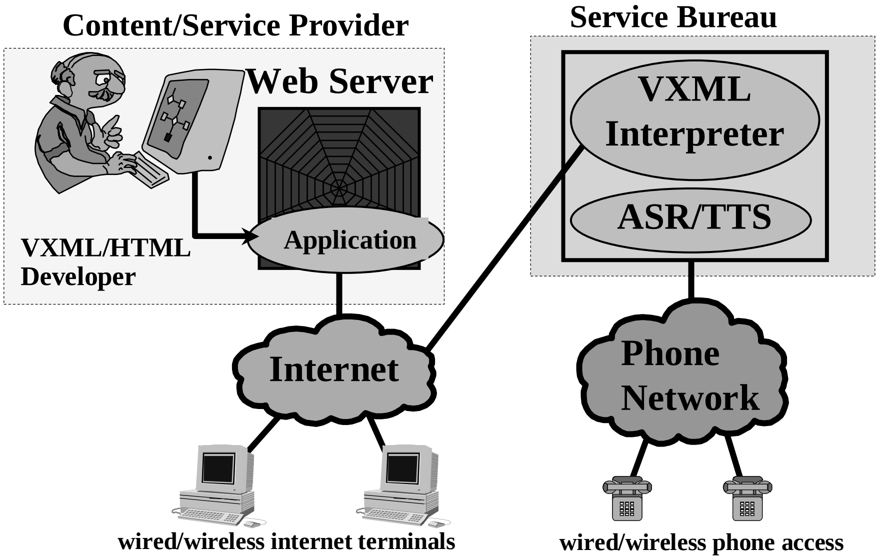
The Internet Era



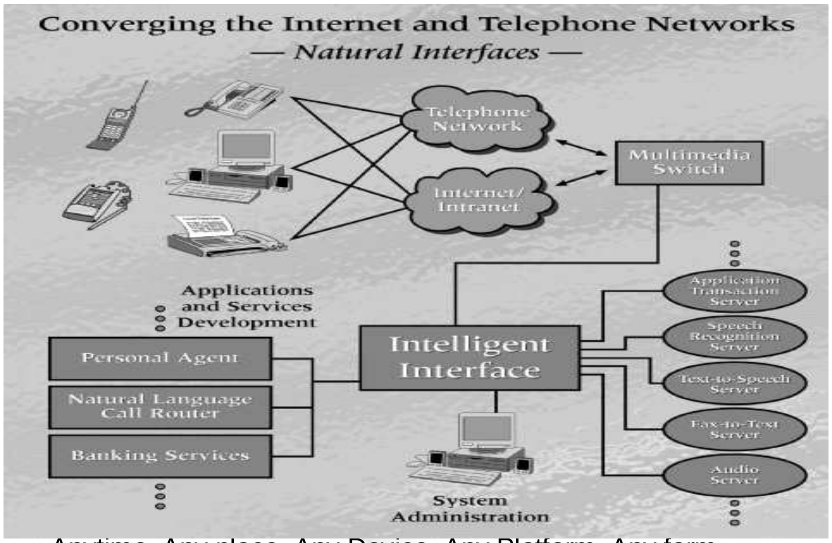
Information access

Communication

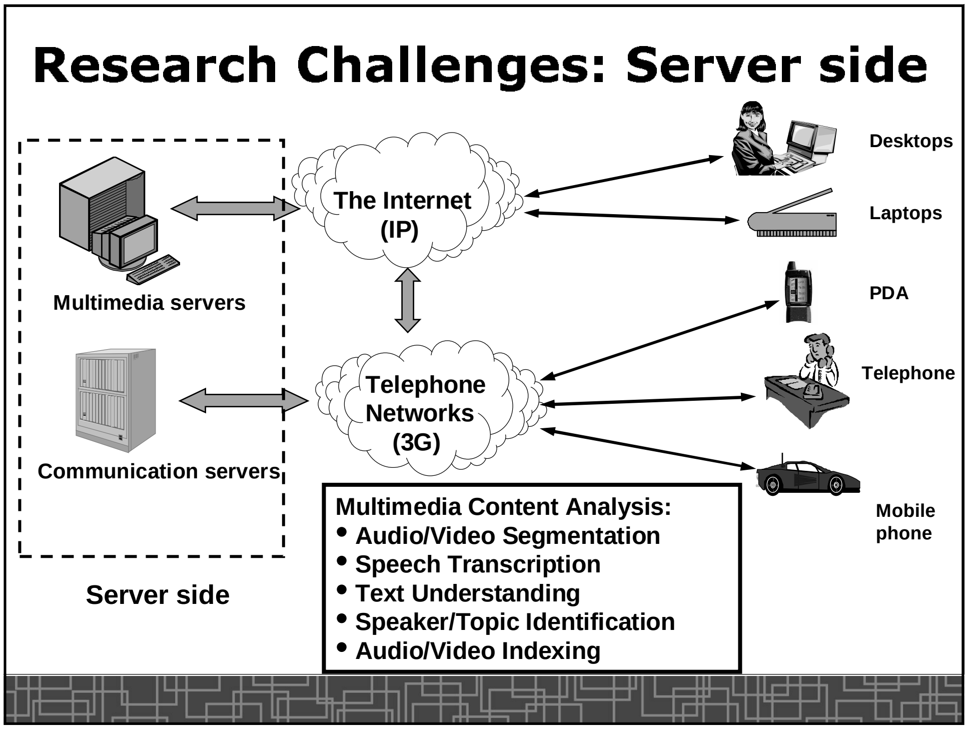
Web and Phone Integration (Voice Portal vs. Web Portals)



Universal Human-Machine Interface



- Anytime, Any place, Any Device, Any Platform, Any form
- Multimedia & Multi-modality Access



Research Challenges: Server side

Multimedia servers

Communication servers

Server side

The Internet
(IP)

Telephone
Networks
(3G)

Desktops

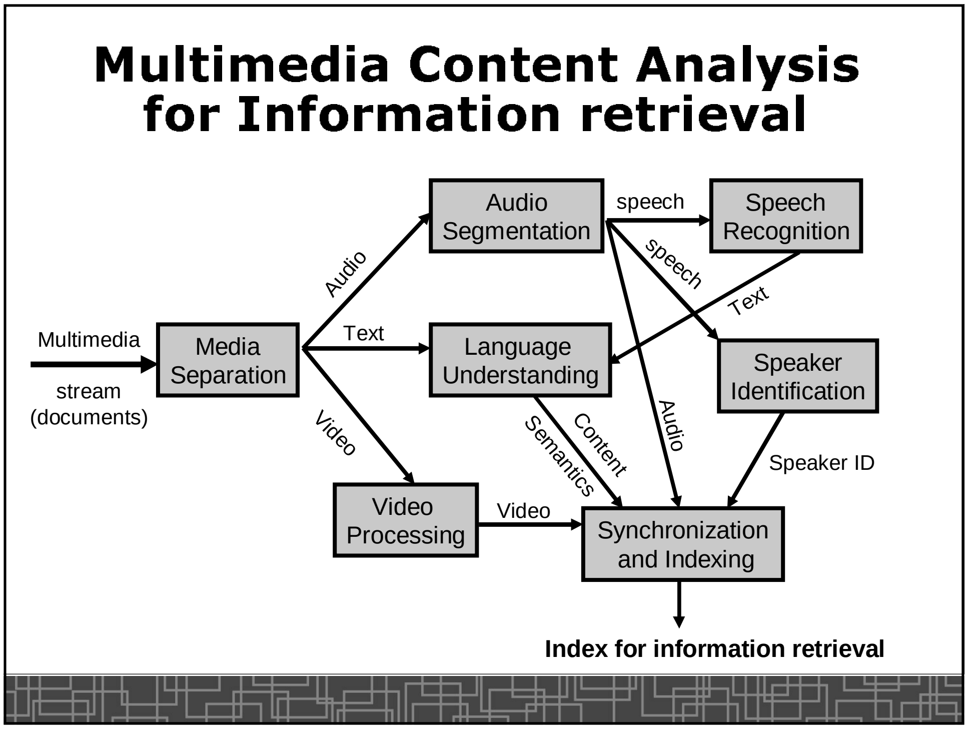
Laptops

PDA

Telephone

Mobile
phone

Multimedia Content Analysis:
- Audio/Video Segmentation
- Speech Transcription
- Text Understanding
- Speaker/Topic Identification
- Audio/Video Indexing



Multimedia Content Analysis for Information retrieval

Multimedia stream (documents)

Media Separation

Audio

Text

Video

Audio Segmentation

speech

Speech Recognition

speech

Text

Language Understanding

Speaker Identification

Content

Semantics

Audio

Speaker ID

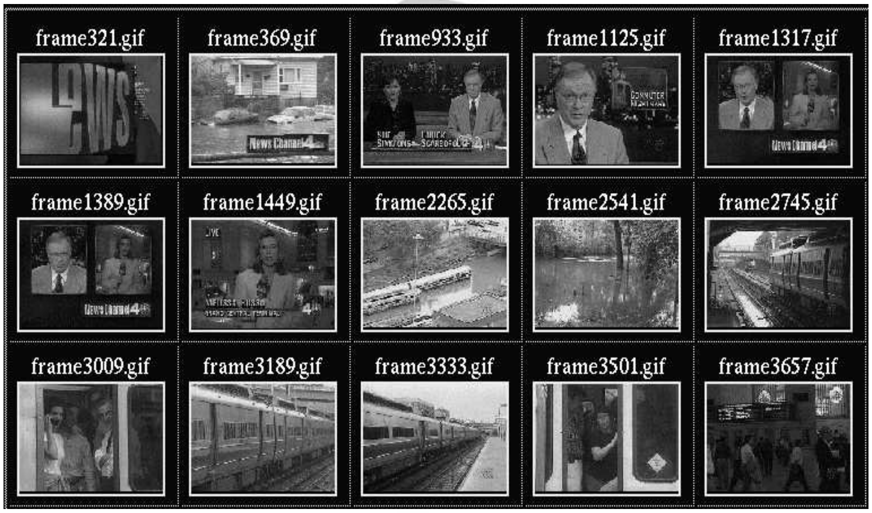
Video Processing

Video

Synchronization and Indexing

Index for information retrieval

Video and Audio Segmentation

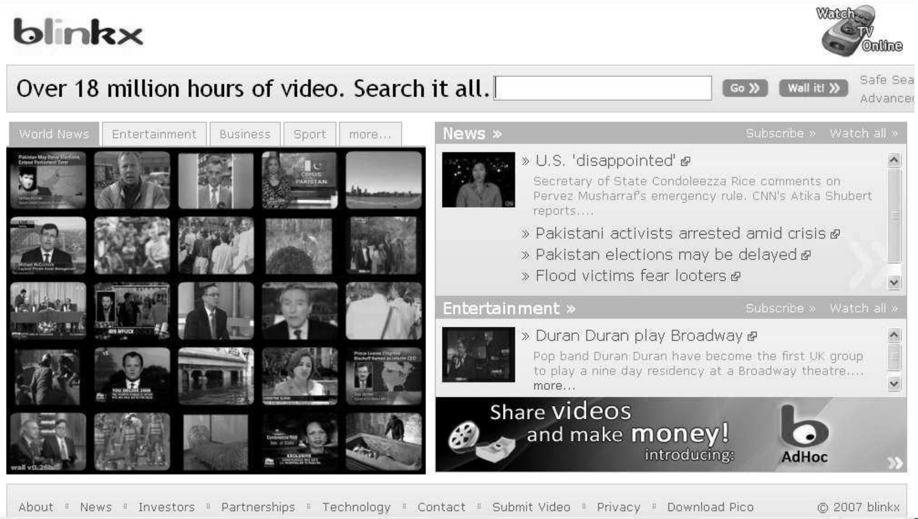


Archiving & Browsing Multimedia Data

- 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/



Over 18 million hours of video. Search it all.

» U.S. 'disappointed'

Secretary of State Condoleezza Rice comments on Pervez Musharraf's emergency rule. CNN's Atika Shubert reports....

» Pakistani activists arrested amid crisis

» Pakistan elections may be delayed

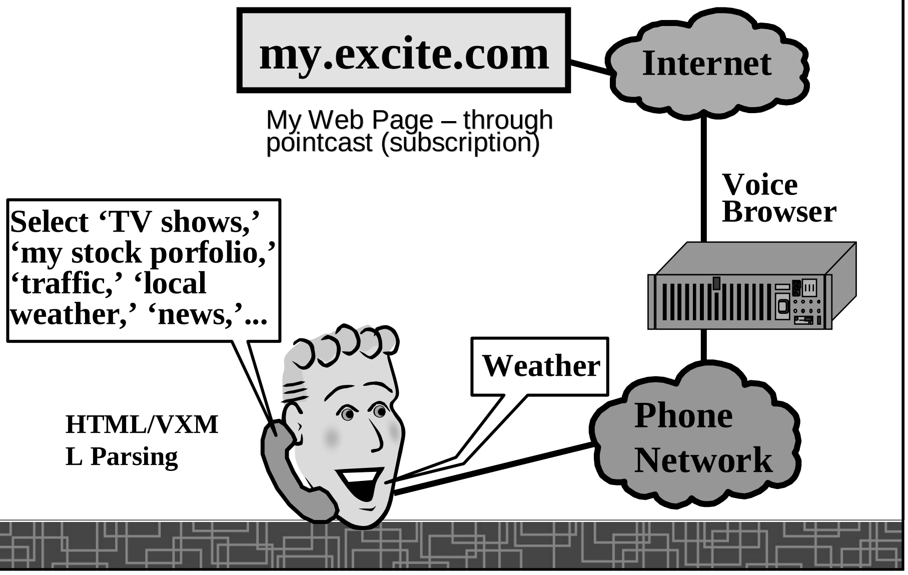
» Flood victims fear looters

» Duran Duran play Broadway

Pop band Duran Duran have become the first UK group to play a nine day residency at a Broadway theatre.... more...

About · News · Investors · Partnerships · Technology · Contact · Submit Video · Privacy · Download Pico   © 2007 blinkx

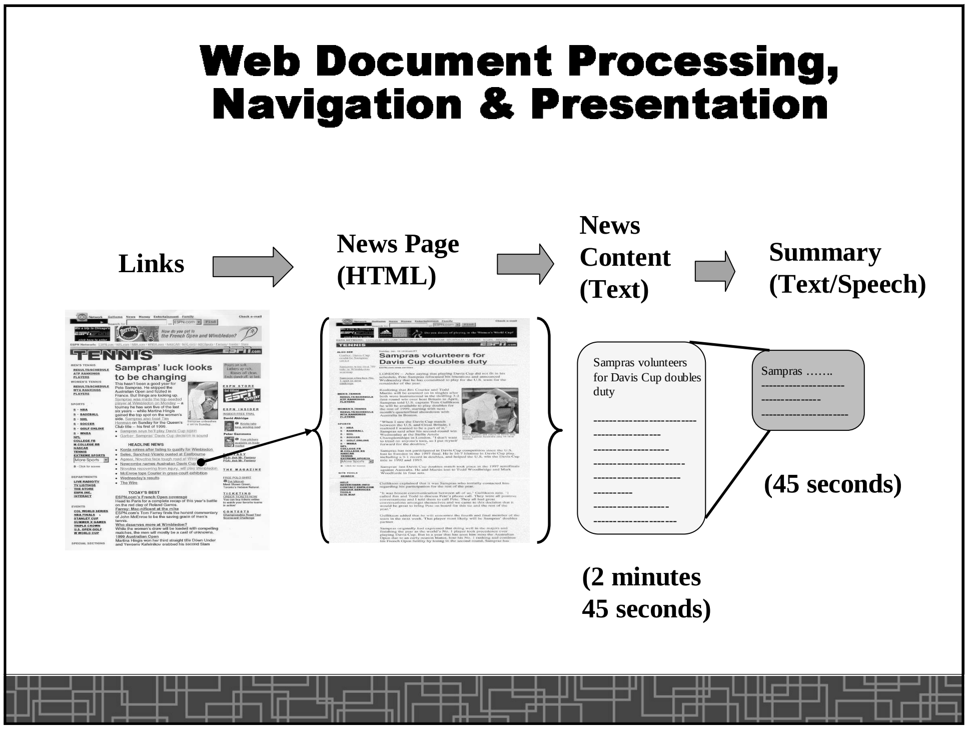
Voice Browsing Applications



My Web Page – through pointcast (subscription)

Select 'TV shows,' 'my stock porfolio,' 'traffic,' 'local weather,' 'news,'...

HTML/VXM L Parsing



Web Document Processing, Navigation & Presentation

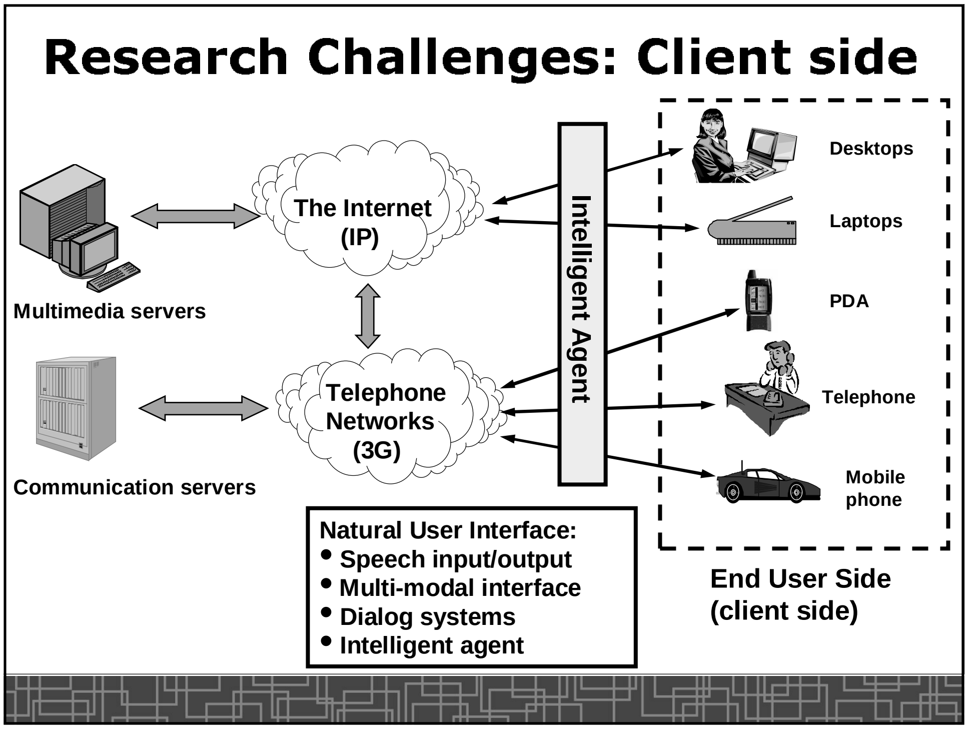
Links → News Page (HTML) → News Content (Text) → Summary (Text/Speech)

Sampras volunteers for Davis Cup doubles duty

(2 minutes 45 seconds)

Sampras …….

(45 seconds)



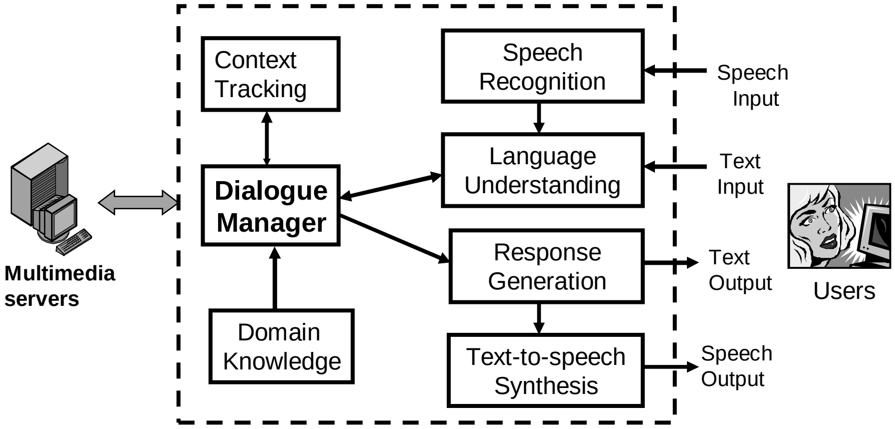
Research Challenges: Client side

Natural User Interface:

- Speech input/output
- Multi-modal interface
- Dialog systems
- Intelligent agent

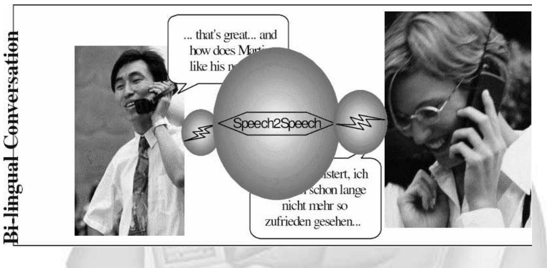
End User Side (client side)

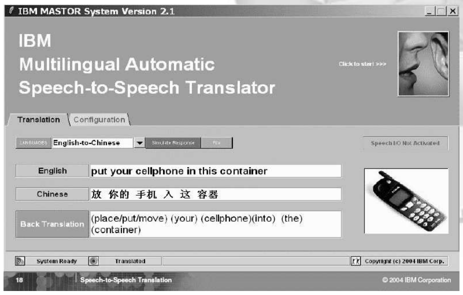
Intelligent Agent: Human-Machine Dialogue System



Key Issues:
- Robust speech recognition
- Spoken language understanding
- Dialogue modeling

Speech-To-Speech Translation


Bi-lingual Conversation
... that's great... and how does Marti... like his n...
Speech2Speech
...istert, ich ...schon lange nicht mehr so zufrieden gesehen...


IBM MASTOR System Version 2.1

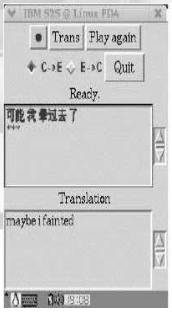
IBM Multilingual Automatic Speech-to-Speech Translator

English-to-Chinese

English	put your cellphone in this container
Chinese	放 你的 手机 入 这 容器
Back Translation	(place/put/move) (your) (cellphone)(into) (the) (container)

System Ready | Translated | Copyright (c) 2004 IBM Corp.

18 Speech-to-Speech Translation © 2004 IBM Corporation


IBM S2S @ Linux PDA

◆ C->E ◇ E->C

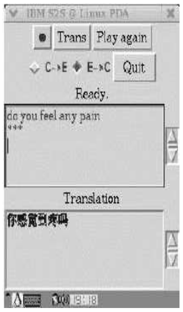
Ready.

可能我晕过去了

Translation

maybe i fainted

19:08


IBM S2S @ Linux PDA

◇ C->E ◆ E->C

Ready.

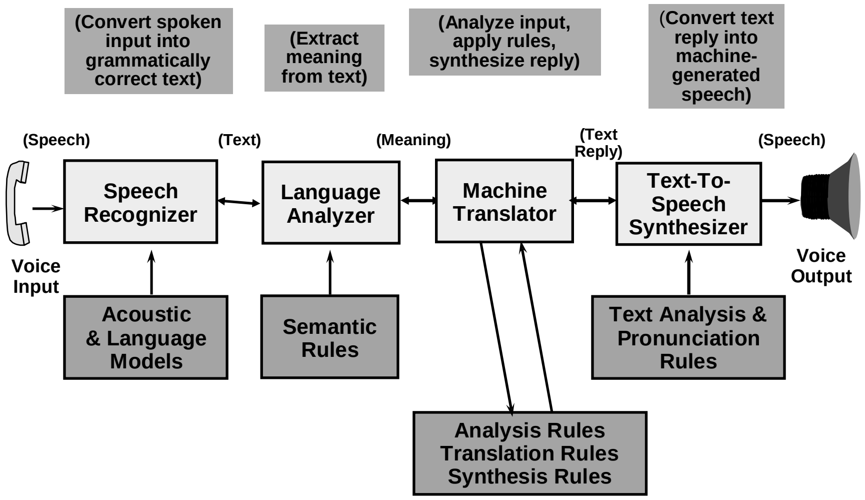
do you feel any pain

Translation

你感觉到疼吗

19:18

Multilingual Speech-to-Speech Translation



(Convert spoken input into grammatically correct text)

(Extract meaning from text)

(Analyze input, apply rules, synthesize reply)

(Convert text reply into machine-generated speech)

(Speech)

(Text)

(Meaning)

(Text Reply)

(Speech)

Voice Input

Speech Recognizer

Language Analyzer

Machine Translator

Text-To-Speech Synthesizer

Voice Output

Acoustic & Language Models

Semantic Rules

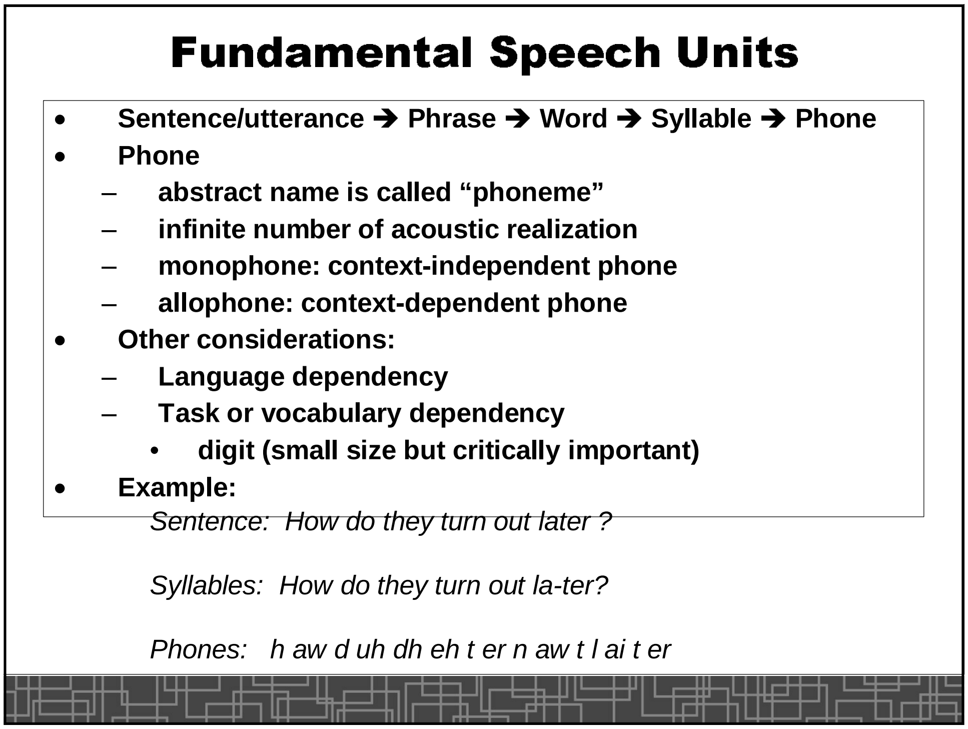
Text Analysis & Pronunciation Rules

Analysis Rules
Translation Rules
Synthesis Rules

Statistical Pattern Classification

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

} The basic theories common to various applications



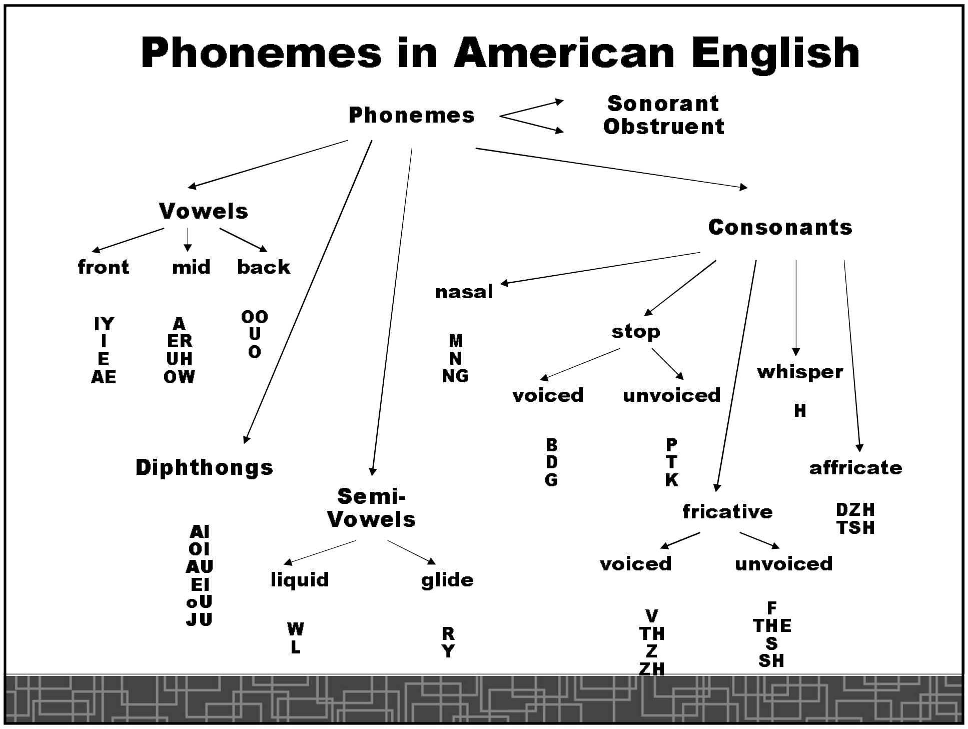
Fundamental Speech Units

- 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

Phonemes

Sonorant
Obstruent

Vowels

front
IY
I
E
AE

mid
A
ER
UH
OW

back
OO
U
O

Diphthongs
AI
OI
AU
EI
oU
JU

Semi-Vowels

liquid
W
L

glide
R
Y

Consonants

nasal
M
N
NG

stop

voiced
B
D
G

unvoiced
P
T
K

whisper
H

fricative

voiced
V
TH
Z
ZH

unvoiced
F
THE
S
SH

affricate
DZH
TSH

Coarticulation

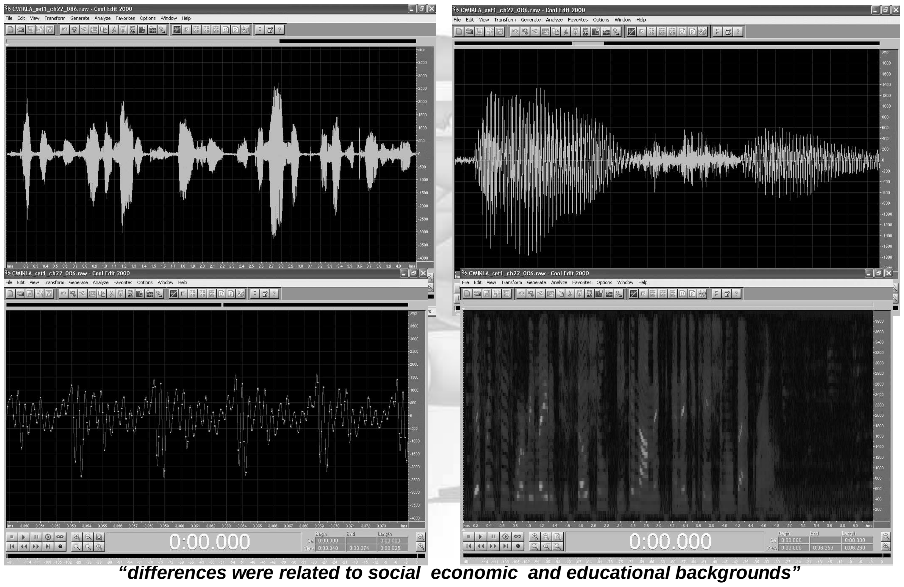
- Phones exhibit consistent acoustic characteristics if pronounced in isolated; but large acoustic variations may appear if uttered in different contexts.
- Coarticulation: acoustic realization of a phone is largely affected by its neighboring contexts.
- Reason: in speech production, articulatory gestures 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:
  - <u>Context-independent phone modeling – monophone</u>: treat each phoneme equally no matter where it appears.

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

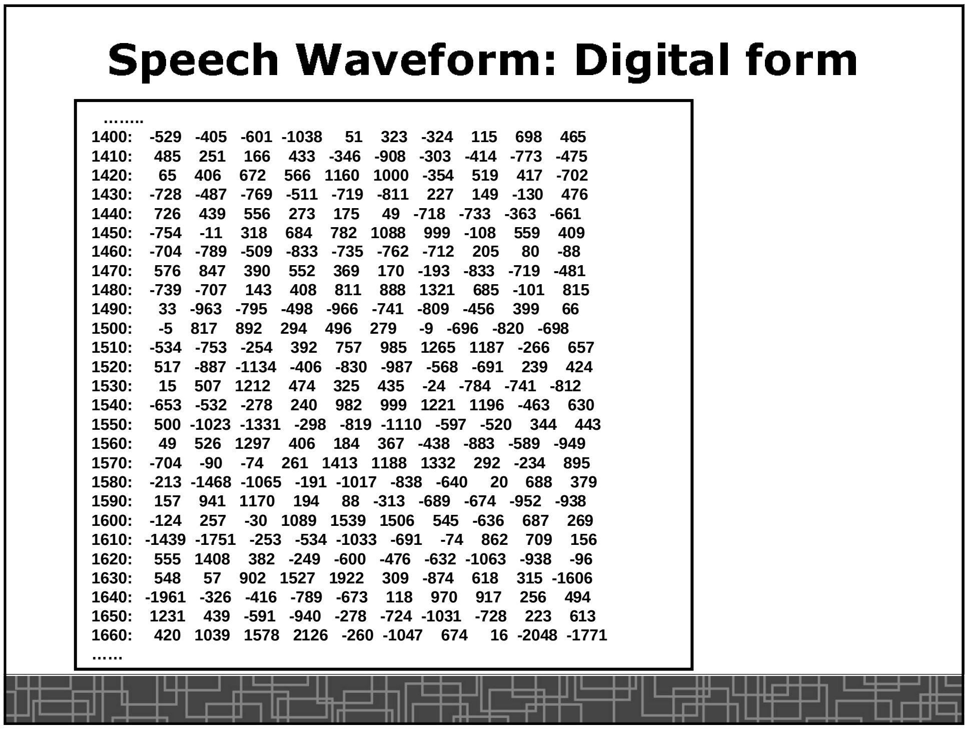
    <u>Left (or right) biphone</u>: a phone unit varies based on left(right) neighboring phone. (American English → 42X42 distinct units)

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

Acoustic Realization: speech waveform

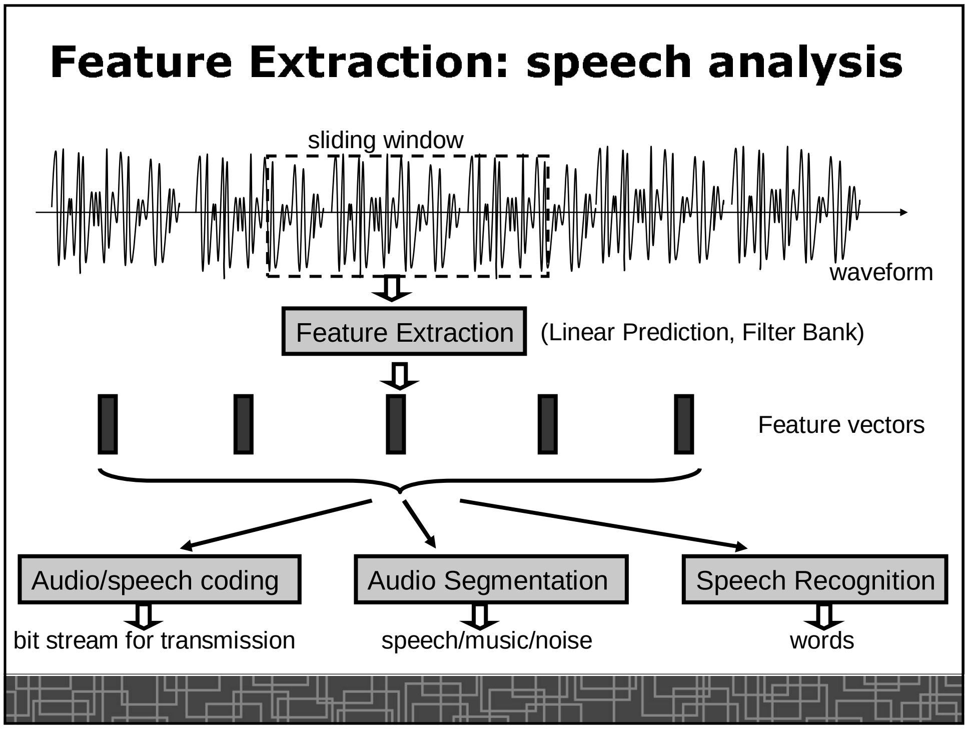

CWIKLA_set1_ch22_086.raw - Cool Edit 2000
File Edit View Transform Generate Analyze Favorites Options Window Help
smpl
3500
3000
2500
2000
1500
1000
500
0:00.000
0:00.000

"differences were related to social economic and educational backgrounds"

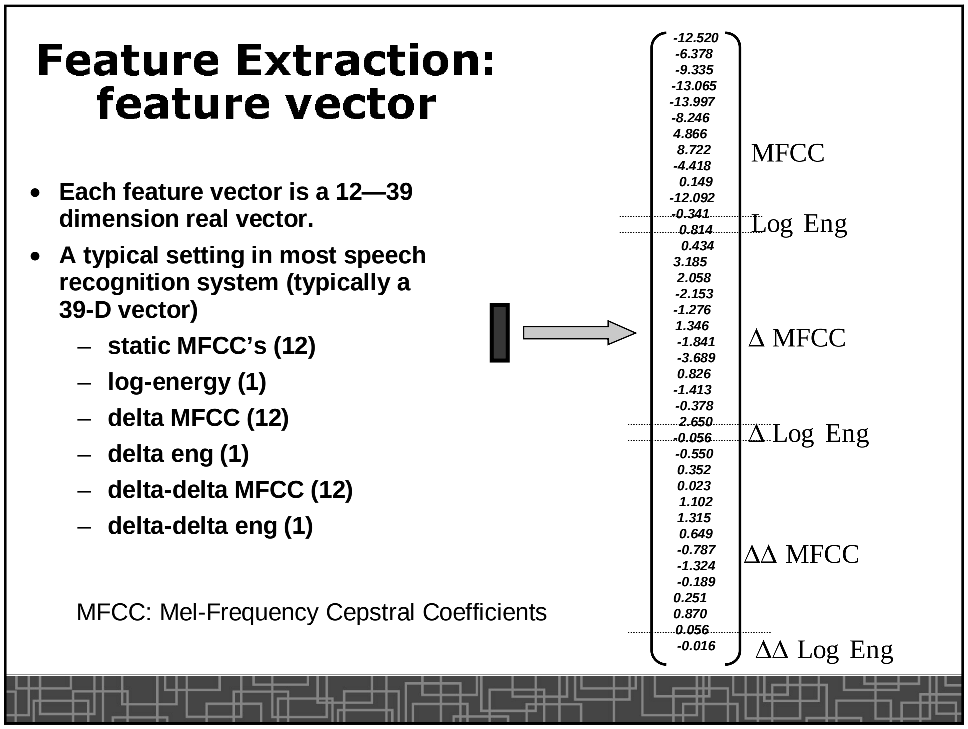


Speech Waveform: Digital form

```
  ........
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

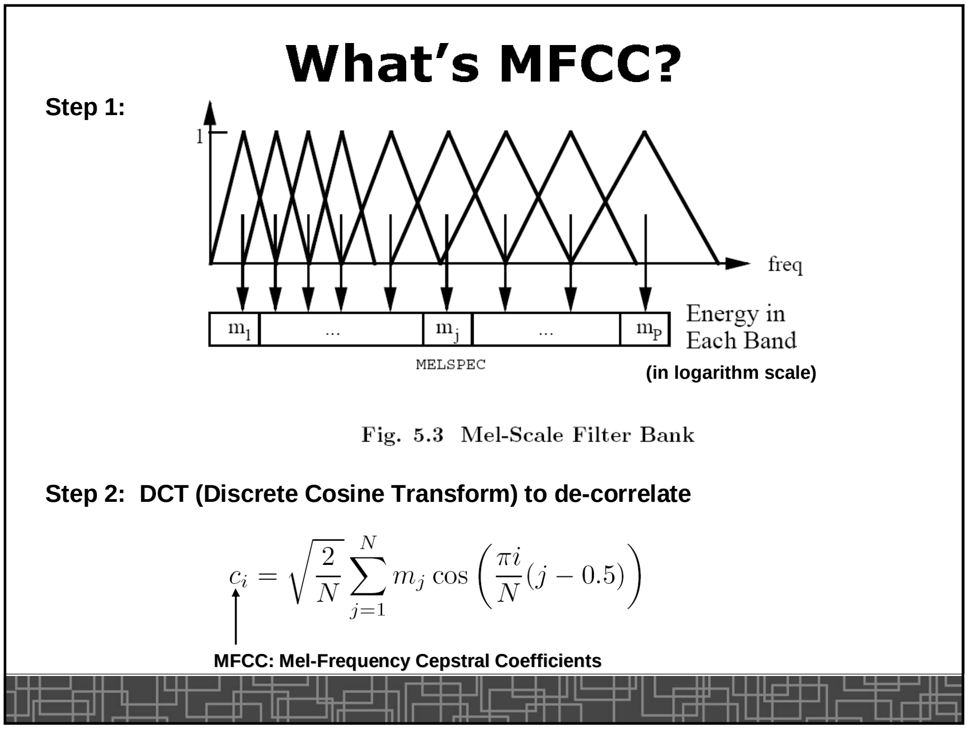


Feature Extraction: feature vector

- 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

$$\begin{pmatrix} -12.520 \\ -6.378 \\ -9.335 \\ -13.065 \\ -13.997 \\ -8.246 \\ 4.866 \\ 8.722 \\ -4.418 \\ 0.149 \\ -12.092 \\ -0.341 \\ 0.814 \\ 0.434 \\ 3.185 \\ 2.058 \\ -2.153 \\ -1.276 \\ 1.346 \\ -1.841 \\ -3.689 \\ 0.826 \\ -1.413 \\ -0.378 \\ 2.650 \\ -0.056 \\ -0.550 \\ 0.352 \\ 0.023 \\ 1.102 \\ 1.315 \\ 0.649 \\ -0.787 \\ -1.324 \\ -0.189 \\ 0.251 \\ 0.870 \\ 0.056 \\ -0.016 \end{pmatrix}$$



What's MFCC?

Step 1:

(in logarithm scale)

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

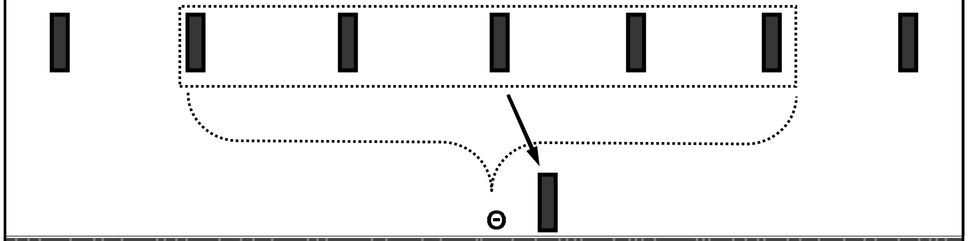
- For each frame, log-energy is calculated as:

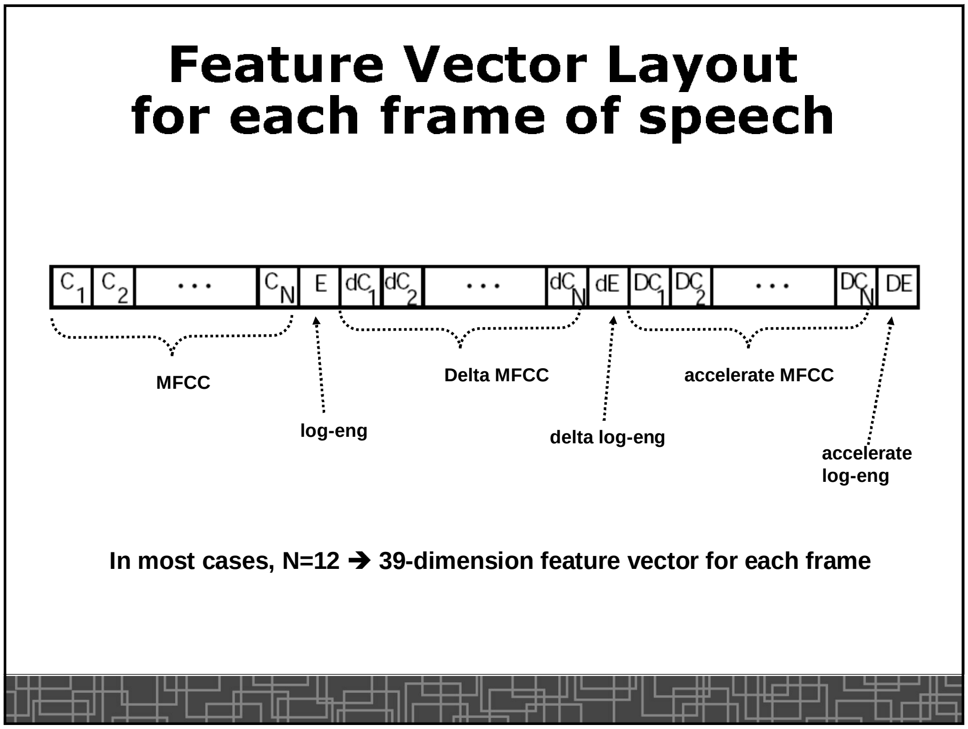
$$E = log \sum_{n=1}^{N} s_n^2$$

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

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

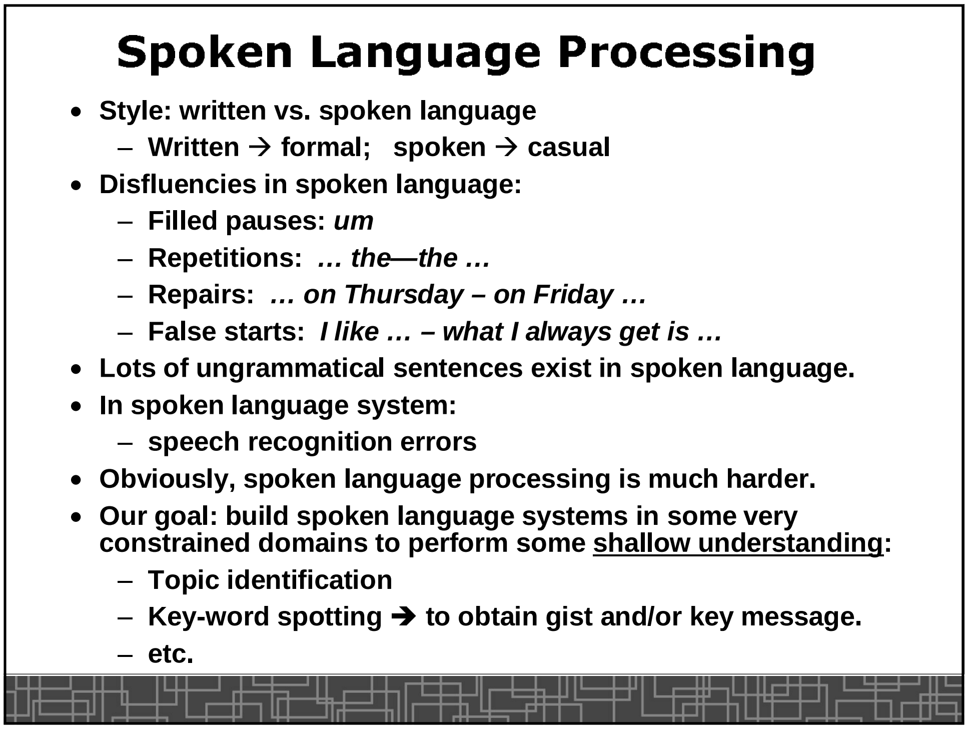




Feature Vector Layout for each frame of speech

C_1	C_2	...	C_N	E	dC_1	dC_2	...	dC_N	dE	DC_1	DC_2	...	DC_N	DE

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 some <u>shallow understanding</u>:
  - Topic identification
  - Key-word spotting ➔ to obtain gist and/or key message.
  - etc.