R&D on Spoken Language Technologies at The Chinese University of Hong Kong
香港中文大學

P. C. Ching（程伯中）

Spoken Language Processing Group

- Two Laboratories :

  Speech Processing Lab (EE Dept)

  Human-Computer Communications Lab (SEEM Dept)

- The Team :

  3 Faculties - Helen Meng (蒙美玲), Tan Lee (李丹), and P.C. Ching (程伯中)

  6 Research Staff

  13 Graduate Students

- Research Directions :

  Highly usable voice-enabled interface technologies

  Ease of access : *anyone, anywhere, anytime, any device*

- Funding and Sponsors :

  HK Government - *RGC, ITF*

  Local Industry - *Group Sense, SmarTone, Reuters HK, TVB*

  Collaborators/Sponsors - *US NSF, JHU, PKU, CAS, Intel, Microsoft China, SpeechWorks, IVRS*

Three Focal Areas

- Resources and Infrastructure :

  Chinese Text Corpora, Lexicon, Pronunciation Dictionary, Speech Databases **(Cantonese)**

- Spoken Dialog Systems and Component Technologies :

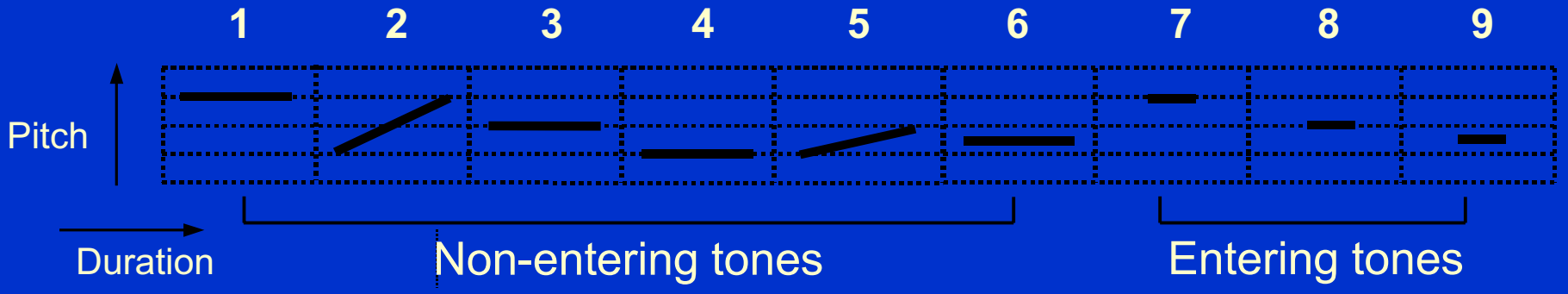
  LVCSR, NL Understanding, Dialog Modeling Speech Generation **(CUFOREX, ISIS)**

- Translingual Speech Retrieval :

  Audio indexing, IR on Spoken Contents, Embedded Machine Translation **(MEI)**

Characteristics of Cantonese

- a monosyllabic and tonal language :



- homophone : (jyu4 予、如、魚、愚 ....)

  homograph : (行徑 hang4 ging3,行路 haang4 lau6, 銀行 ngan4 hong4,操行 cou1 hang6 . .)

- ambiguous word-boundary, gender, tense logic
- very colloquial:

  佢好『腌尖』, 鍾意著『單吊西』、飲齋啡、食『多士』

Cantonese Spoken Language Databases

* CUCorpora (1999) - ***the first publicly available Cantonese microphone speech database***

- 150 speakers, 70 hr. studio recording
- 16 KHz sampling, 16-bit resolution
- a wide phonetic coverage
- phonemic and orthographic transcription

CUSYL
1800 Cantonese tonal syllables
2M & 2F + pitch marks
1 CD

CUWORD
2527 polysyllabic words
13M & 15 F speakers
5 CD

CUSENT
5719 sentences
40 M & 40 F
3 CD

CUDIGIT
Digit strings - length 1 to 14
50 M & 50 F speakers
2 CD

CUCMD
100 navigation commands
50 M & 50 F speakers
1 CD

http://dsp.ee.cuhk.edu.hk/speech/corpus.html

* CUCall (2002) - collection of telephone speech data over fixed-line and mobile networks
  - over 1000 speakers
  - 8 KHz sampling, 8-bit resolution (μ-law)
  - 200 hours of data, phonemically transcribed

Sentences

continuous Cantonese sentences based on CUSENT for phonetic coverage

Passages

Passages extracted from newspapers - capture long speech characteristics

Digit strings

randomly generated digit strings of length 1 to 16

Spontaneous conversation

spontaneous answers to prompted questions - capture speaking styles

Application specific words

local place names, command words, stock names, foreign currencies,

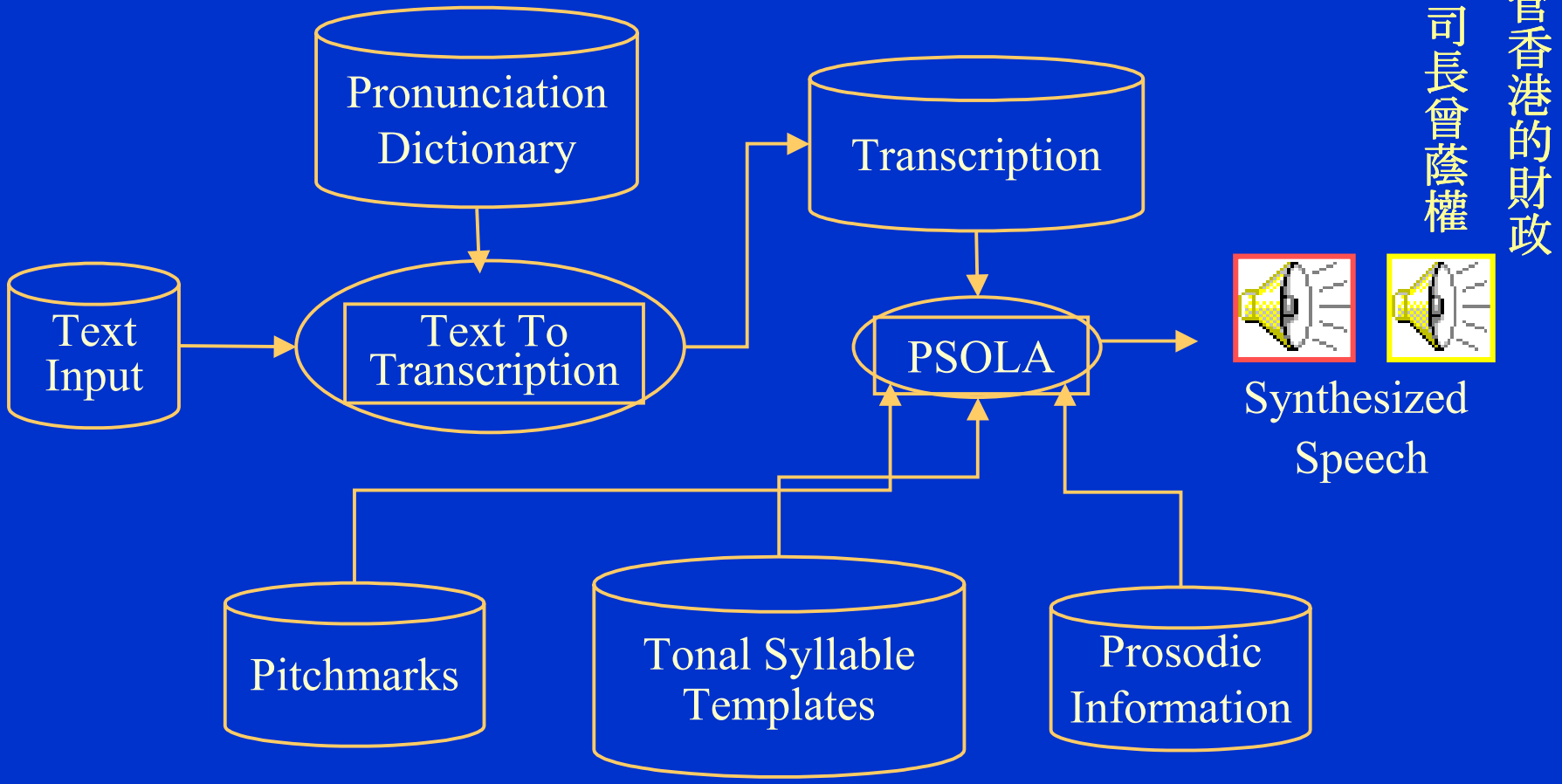
http://dsp.ee.cuhk.edu.hk/speech/cucall.html

Cantonese Text-to-Speech

* CUTalk (2000) ***- the first PSOLA-based Cantonese synthesizer***

  - 1857 tonal syllable speech templates
  - cross-syllable co-articulation
  - manually labeled pitchmarks
  - over 90,000 words pronunciation dictionary
  - text segmentation by maximum matching
  - word-level duration and energy normalization
  - prosody modeling : duration, F0, intensity (focus, phrasing and final lengthening)

CUTalk Architecture



曾經主管香港的財政

政務司司長曾蔭權

API using DLL
Window and Linus

supports: big5 and GB
multi-threading
English spelling

* CUVocal *- the first corpus-based concatenative TTS for Cantonese*

- handles mixed language with English text
- portable across Chinese dialects
- amendable to domain-optimization for enhanced naturalness in specific applications
- unit selection based on phonologically-motivated distinctive features

CU Vocal : Air Travel Domain-Optimized

你於七月二十五日訂購了一張成人及三張兒童的英國航空公司 單程商務客位機票航班編號係 BA375 將於當地時間十三時二分由紐約出發 預計於當地時間十七時二十四分抵達大阪

ASR for Cantonese

* Acoustic Model *(HMM)*
  - feature parameters : ***MFCC + energy, Δ, ΔΔ***
  - training data : **CUSENT, CUCALL**
  - acoustic units : ***base syllable(BS), tonal syllable(TS), initial-final (IF), initial-tonal-final (ITF)***
  - context independent/dependent models

syllable recognition accuracy

	BS	TS	IF	ITF
CI	*56.13 (638)*	*52.25 (1615)*	*62.05 (81)*	*64.99 (301)*
BI			*78.37 (1031)*	*77.68 (1993)*
TRI			*79.47 (8458)*	*78.66 (45556)*

  - decision tree based sharing

* Language Model

- LVCSR
  - word bigram, ~6,500 words lexicon
  - training text : 5 local newspapers over 1 year perplexity of the bigram is ~160
  - character accuracy ~83%
  - ***expanded lexicon of 20K words***

- Domain-specific ***(stock/currency trading, air travel)***
  - class-based bigram **(~1,600 wd ; 270 cl)**
  - training data : from newspapers/magazines and collected conversational queries
  - character accuracy ~95%

* Search Algorithm

- time synchronous Viterbi search
- multiple layers of beam pruning : *acoustic level, language level, cross-word context*
- LM Lookahead : *static and compressed*
- exact heuristic for A* stack search to generate the true N-best list

Robustness : *parallel model combination to remove additive noise, blind adaptive FRESH filtering to separate different speech sources*

Speaker Adaptation : *MLLR with confidence measure*

Mixed Language Recognition

Spoken Language Understanding

- human-computer dialog on limited domain
- 3 core components : semantic parser, discourse analysis, dialog manager
- handcrafted grammar by experts (rule-based) on annotated corpus (data-driven)

Objectives of our recent work :

- reduce handcrafting grammar rules
- enhance portability across domains and languages
- improve efficiency in parsing

I. Semi-Automatic Grammar Induction

- agglomerative word clustering technique
- optional prior knowledge as catalyst
- manual refinement of induced grammar
- expedite grammar development
- reduce reliance on - handcrafting grammar
  - annotating corpora
- semi-automatic approach enhances portability across domains / languages

Word Clustering

- Inspired by language model [McCandless 93]
- Unannotated training corpus
- Spatial clusters (*SC*) ⇒ semantic categories

$$Div(p_1||p_2) = \sum_{i=1}^{V} p_1(i)\frac{p_1(i)}{p_2(i)} + \sum_{i=1}^{V} p_2(i)\frac{p_2(i)}{p_1(i)}$$

$$Dist(e_1,e_2) = Div(p_1^{left}||p_2^{left}) + Div(p_1^{right}||p_2^{right})$$

- Temporal clusters (*TC*) ⇒ phrasal structure

$$MI(e_1,e_2) = P(e_1,e_2)\log\frac{P(e_1|e_2)}{P(e_2)}$$

- Alternates between *SC* and *TC* formation

Induced ATIS Grammar

SC4 --> december | february…
SC7 --> nashville | toronto | tampa … (places)
SC24 --> serve | serves
SC28 --> monday | wednesday | thursday
TC39 --> first class
TC44 --> one way
TC145 --> flights from SC7 to SC7 (a phrase)

With prior knowledge injection
(seed categories)
SC2 --> atlanta | baltimore | boston… (i.e. city names)
SC3 --> monday | tuesday | wednesday… (i.e. days of the week)
SC15 --> from | departing from | leave from...
SC16 --> to | arriving to | arrive to...
TC38 --> flights SC15 SC2 SC16 SC2 on SC3 (i.e. a phrase)

Post-processing

- Insert meaningful nonterminal tags
  e.g., city_name --> nashville | toronto | tampa..
- Completing terminal set
  e.g., months of the year --> Jan | Feb .. | Dec.
- Merging

TC211 --> flights from SC7 to SC12
TC274 --> flights from SC7 to SC29
TC292 --> flights from SC12 to SC7
SC7, SC12, SC29 are all city names

- Pruning

SC31 --> today | uses | TC81 (in the morning)...
SC32 --> either | or | travel | know | take...
SC45 --> airport | florida
SC69 --> TC81 (city_name to) | TC166 (denver to)...

Portability to Chinese

- Parallel Chinese corpus from ATIS-3
  - Cantonese colloquialisms

English: *show me all united airlines first class flights*
Chinese: 話我知所有聯合航空既頭等航班

English: *show me all the northwest flights from new york to milwaukee that leave at seven twenty a m*
Chinese:
話俾我知所有係上晝七點廿分
由紐約飛去密耳瓦基既航機

English: *How many first class flights does united have today*
Chinese: 今日有幾多班聯合航空既頭等航機起飛

- Tokenization and clustering

SC_1 → 波班克|蒙特利爾|鹽湖城
(translation: burbank | montreal | salt lake city)

TC_{50} → 米契爾 國際 機場
(translation: general mitchell international)

TC_{68} → 單程航班
(translation: one way flight)

- Word order differences

flight_number
e.g. “flight four one seven” 四 一 七 航機

arrival_time
e.g. “arrive before five pm” 下晝 五 點 前 到 達

Evaluation Results

- Error rate in semantic concepts

ATIS-3 Class A sentences	Error Rate from Induced Grammar G_{SA}	Error Rate from Handcrafted Grammar G_H
Training Set	6.9%	6.3%
1993 Test Set	16.1%	8.3%
1994 Test Set	17.1%	13.0%

- Understanding using induced grammar

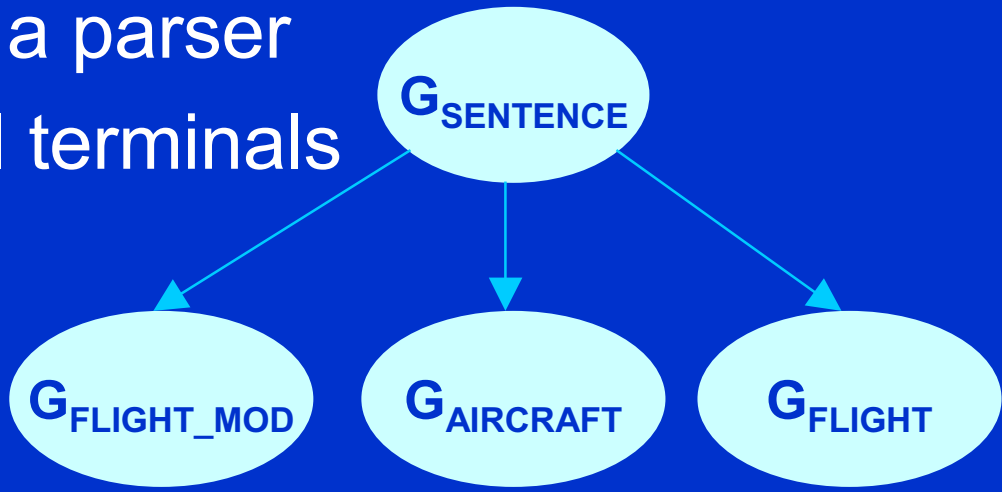
	Test 1993		Test 1994	
Understanding	G_{CSA}	G_{SA}	G_{CSA}	G_{SA}
Full	*77.7%*	*80.4%*	*74.1%*	*76.8%*
No	*6.0%*	*3.1%*	*3.9%*	*1.4%*

II. Lattice Parsing with Multiple Grammars

- Concerns in parsing natural language
  - efficiency (scalability)
  - ambiguity
  - robustness
- Grammar partitioning
  - reduce exponential growth of states in LR parsing
  - save time to generate parsing table
- Parser composition
  - integrate parsers with specialized grammars

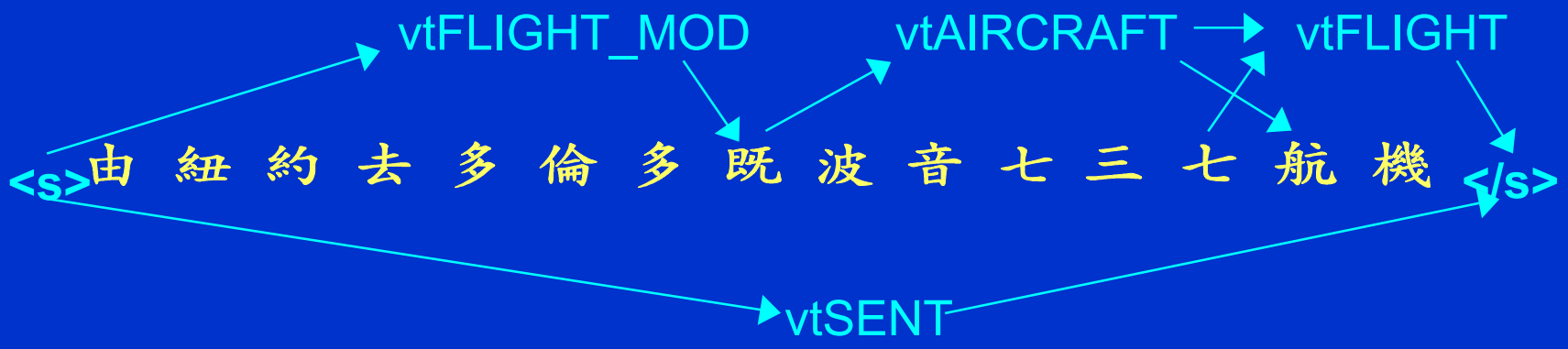
Grammar Partitioning

- Properties of sub-grammars
  - modular, no overlap
  - union gives original rule set
  - each operates with a parser
  - interfaced by virtual terminals
  - ordered in levels
- Virtual terminal
  - non-terminal
  - output of a sub-grammar/parser
  - input of another sub-grammar/parser



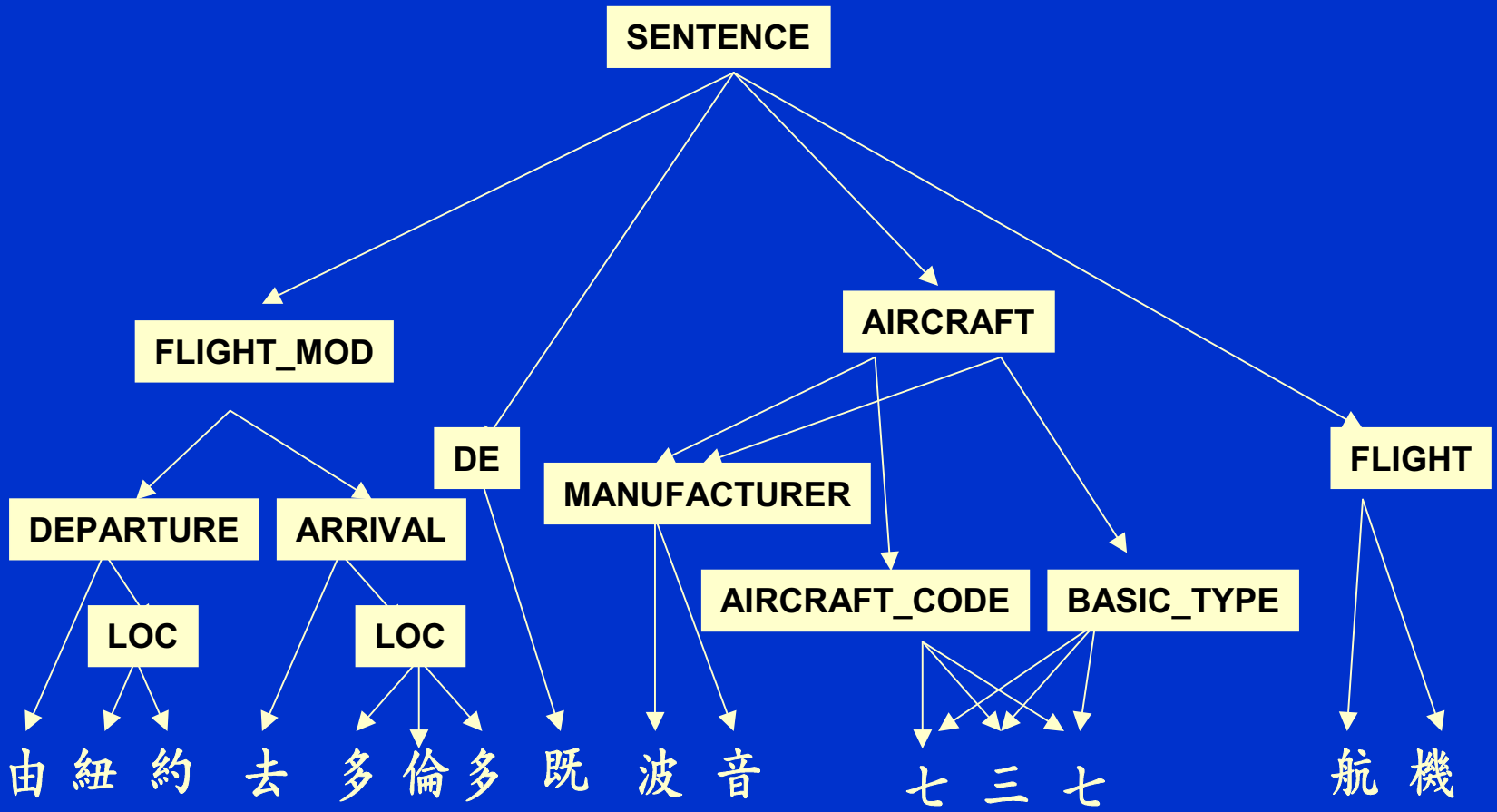
Parser Composition

- Cascading
  - bottom-up robust parsing
  - lattice with multiple granularity LMG



  - adapt GLR parsers to process lattice
  - virtual terminals added to LMG
  - output parse forest

Parse Forest Output



Parser Statistics

- ATIS-3 (English)

Grammar Statistics	No Partitioning	Partitioned Grammar
# of rules	1,650	1,818
# of virtual terminals	N/A	65
Total # states in parser	72,869	3,350

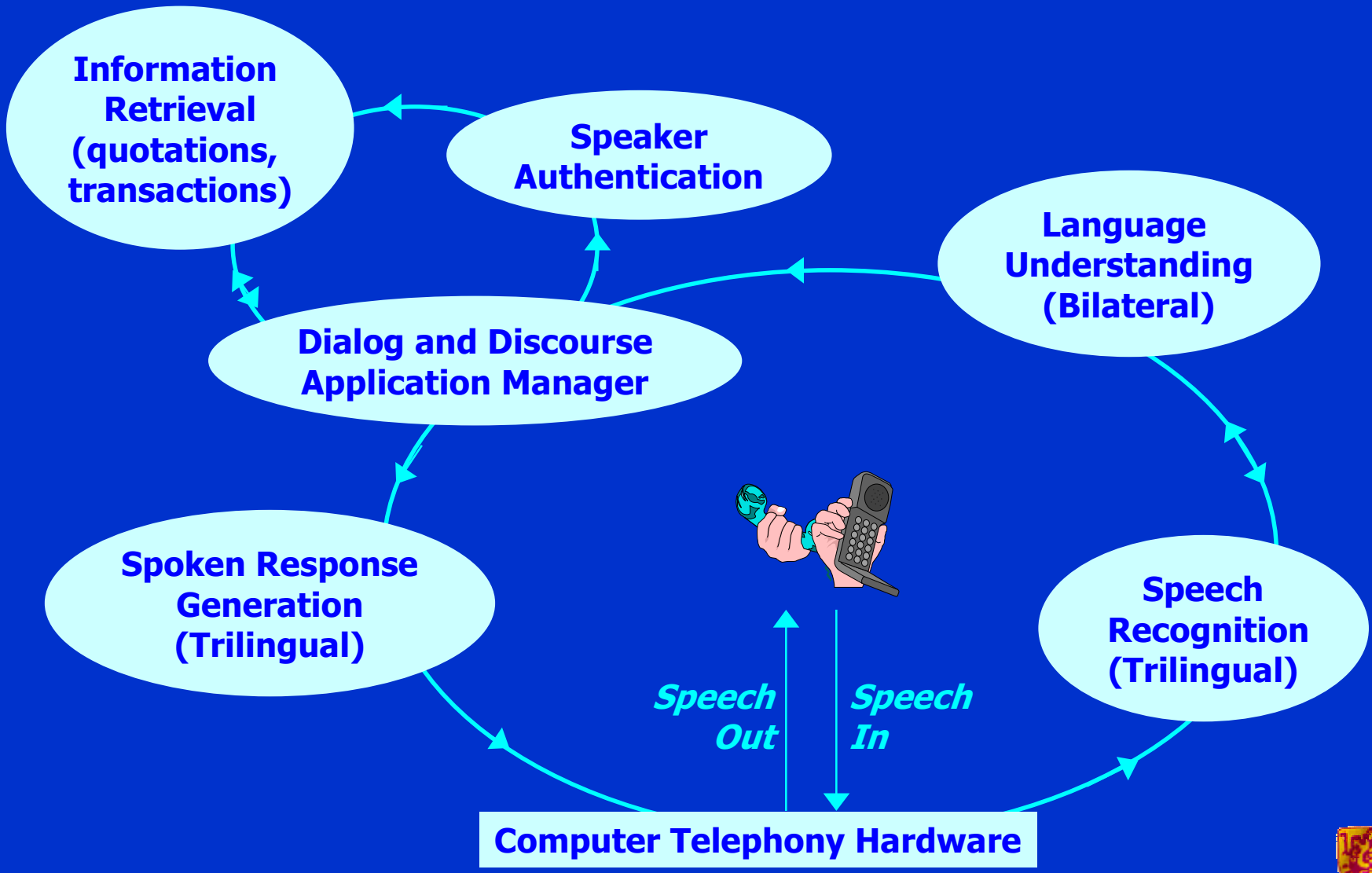
- Translated ATIS-3 (Chinese)

Grammar Statistics	No Partitioning	Partitioned Grammar
# of rules	1,538	1,639
# of virtual terminals	N/A	63
Total # states in parser	29,734	3,896

Integration into a Dialog System

- ISIS (Intelligent Speech for Information Systems)
  - collaboration with Peking Uiversity
  - trilingual **(Cantonese, Putonghua, English)**
  - stocks domain
- NLU component
  - disambiguates numeric expressions in stocks domain
  - handles out-of-vocabulary words
- Adaptive learning, on line interaction interleaves with off-line delegation

ISIS: System Overview



CU FOREX

中文大學 寰宇之聲

Principal Investigator

Professor Helen Meng

HCCL Human-Computer Communications Laboratory

Department of Systems Engineering and Engineering Management

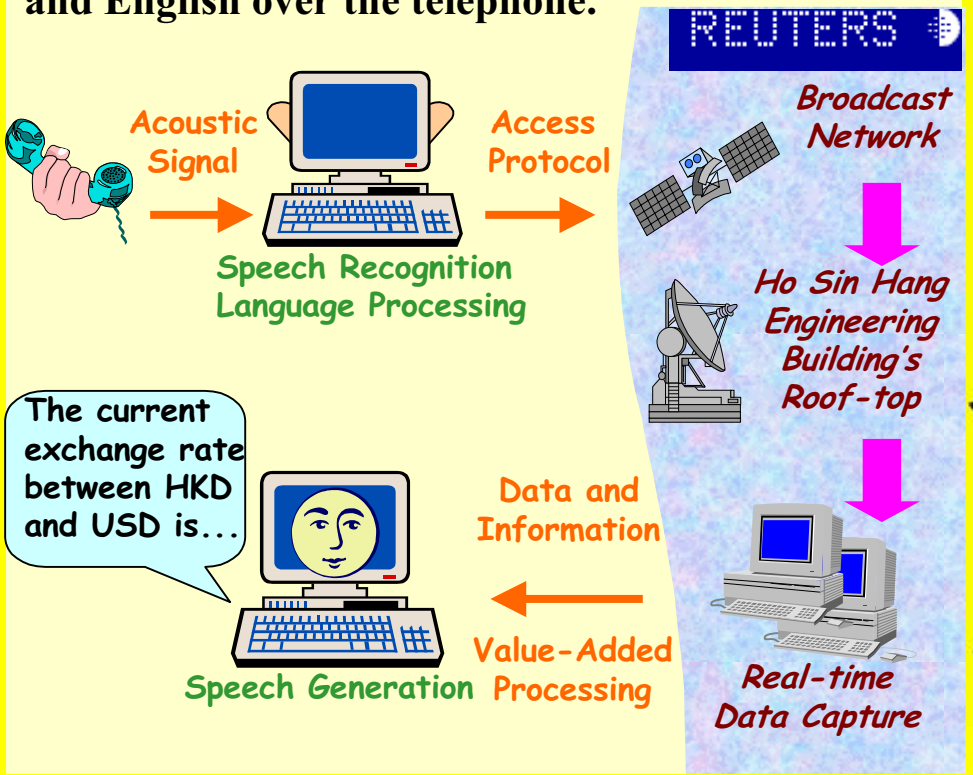
Industrial Sponsors

SpeechWorks

IVRS (INTERNATIONAL) LTD.

REUTERS

CU FOREX is a bilingual hotline for real-time foreign exchange enquiries. It integrates a *plethora of speech recognition and language processing technologies* to handle both Cantonese and English over the telephone.

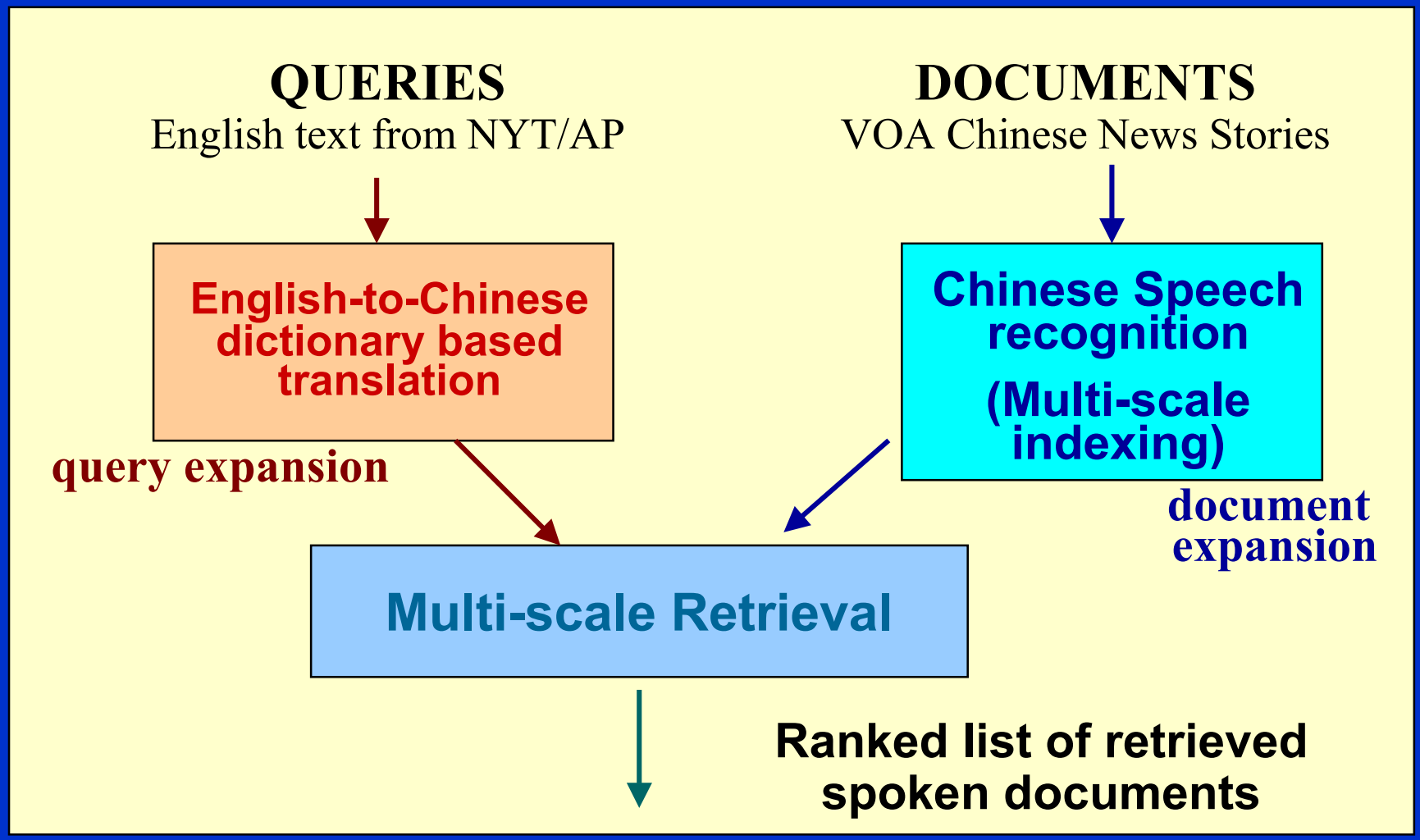


Telephone : (852) 2603-7830 or (852) 2603-7884

http://www.se.cuhk.edu.hk/hccl/

Spoken Document Retrieval

- Cross-lingual *(English-Chinese)*



- queries are translated from English to Chinese using phrase based approach
- Chinese queries with multiple translations are discounted in weighting
- the processed queries are used to retrieve the Mandarin spoken documents
- the retrieval process is susceptible to
  - speech recognition error
  - out-of-vocabulary error due to unmatched translated terms
- queries are expanded both before and after translation

- ***Pre-translation Query Expansion***
  - based on a side corpus
  - include co-occurrence terms
  - select expansion terms according to TF/IDF
  - control the number of expanded terms
- ***Post-translation Query Expansion on target language***
- ***Multi-scale audio indexing/Retrieval*** - best word sequence + character and syllable bigram
- ***Document Expansion***
  - based on an expansion set of documents
  - augment relevant terms to the document
  - cover synonyms by co-occurrence terms

Evaluation Results

- experiements based on Topic Detection and Tracking Corpora (TDT-2, TDT-3)
- Mean Average Precision of Document Retrieval

	No Query Expansion	With Query Expansion
Word	*0.4160*	*0.4637*
Character Bigrams	*0.4917*	*0.5255*
Syllable Bigrams	*0.4215*	*0.4477*

	Word	Character	Word+Character
TDT-2	*0.4641*	*0.5163*	*0.5182*
TDT-3	*0.462*	*0.475*	*0.4815*

Man-machine Communication using Natural Languages

possible but plenty of roadblocks and challenges