Recent Activities of Oriental COCOSDA

S. Itahashi, H. Meng, P.Y.Hui, W.K.Lo, H.C.Wang,
Li Haizhou, Virach Sornlertlamvanich ,
Sawit Kasuriya , Chatchawarn Hansakunbuntheung

Contents

1. Introduction of Oriental COCOSDA
2. Speech Corpora and Standardization in Japan
3. Recent Activities in
   - Korea and China
   - Hong Kong
   - Taiwan
   - Singapore
   - Thailand
4. Summary

Asian Activities: Oriental COCOSDA

To exchange ideas, share information, discuss regional matters on creation, utilization, dissemination of spoken language corpora of oriental languages, assessment methods of speech input/output systems, and

To promote speech research on oriental languages.

Proposed in 1994,

Preparatory meeting with 11 people

in Hong Kong in 1997

Annual workshops since 1998

in Japan, Taiwan, China, Korea, Thailand

International Workshop on East-Asian Language Resources and Evaluation
– Oriental COCOSDA WORKSHOP –

1998 1st Meeting, Tsukuba, Japan (30 talks)
1999 2nd Meeting, Taipei, Taiwan (44 talks)
2000 3rd Meeting, Beijing, China (8 talks)
2001 4th Meeting, Taejon, Korea (11 talks)
2002 5th Meeting, Hua Hin, Thailand (24 talks)
2003 6th Meeting, Sentosa, Singapore (Oct. '03)
2004 7th Meeting, Delhi, India (Nov. '04, planned)

Peculiarities in East Asian Languages

1. Variety : Many Languages belong to different language families
2. Different Orthographic Systems
   Chinese characters, Korean syllabic alphabet
   Japanese Kana alphabet, Thai characters, etc.
3. Non-unique romanization
   Chinese pinyin, Japanese romanization system, etc.
   . Needs different processing from European languages

Some Activities in East Asia

Japan

Korea

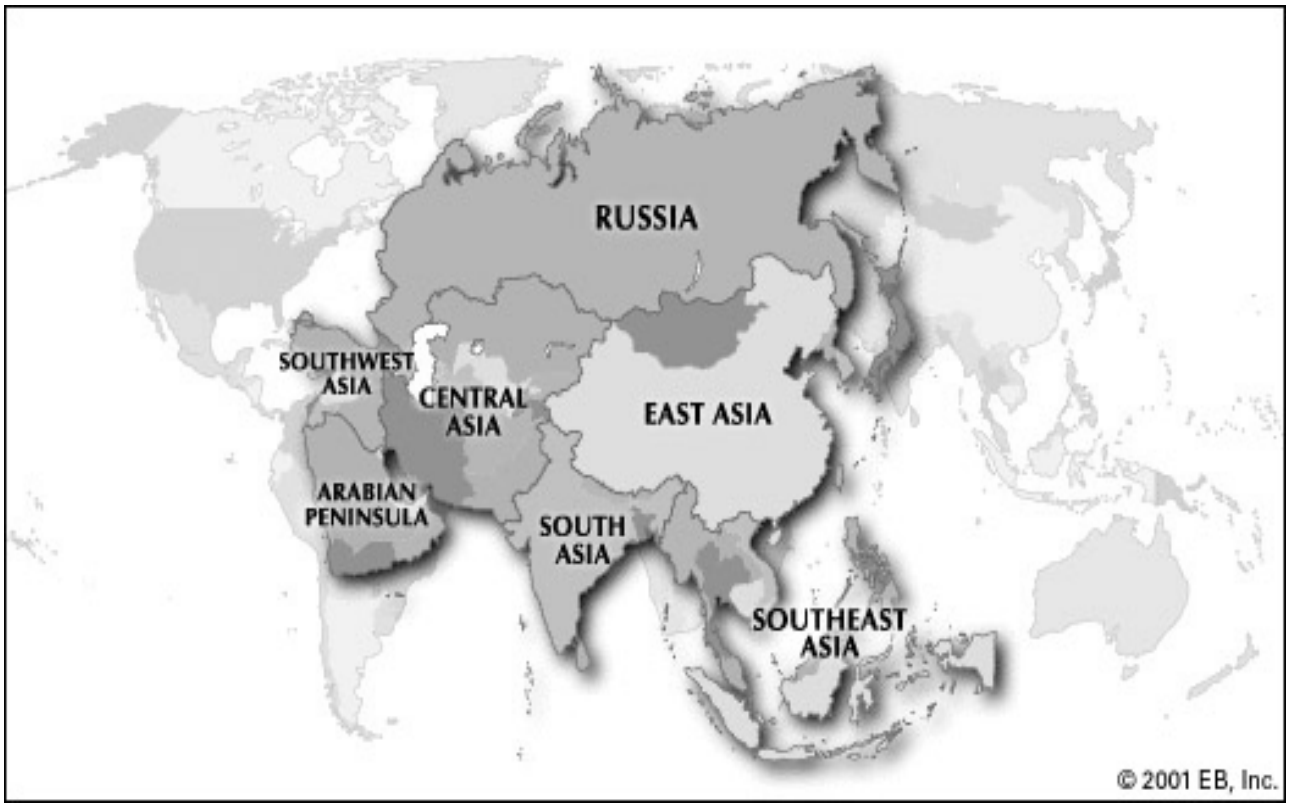
China

Hong Kong

Taiwan

Singapore

Thailand



Japanese Activities

GSK (Language Resource Association)

Launched in 1999

- Home page of available corpora
- Seeking for financial support

Being renovated as an NPO (Oct./Nov. 2003)

President: Prof. H. Tanaka (TIT)

10 Board Members

27 Steering Committee Members

Secretaries: Intergroup Corp.

http://tanaka-www.cs.titech.ac.jp/gsk/, gsk-eng.html

Speech Related Projects in Japan

1. Spontaneous Speech Engineering – Corpus and Processing Technology – (1999–2003, Prof. S. Furui)
2. Integrated Acoustic Information Research (1999–2003, Prof. F. Itakura)
3. Realization of Advanced Spoken Language Processing from Prosodic Features (2000–2003, Prof. K. Hirose)
4. Expressive Speech Processing (2000–2003, Dr. N. Campbell)

Read Speech Corpus for Foreign Language Learning

Research Project on "Advanced Utilization of Multimedia to Promote Higher Education Reform" (2000–2002)

Japanese Speech Database Ready by Non-native Speakers

- A. 100 PB sentences
- B. 115 minimal pair words
- C. 108 sentences including B
- D. some dialogues

Spoken by 140 students

5 CD-ROMs

English Speech Database read by Japanese students has also been created (reported at COCOSDA 2001).

Standardization in Japan

1) Standard of Speech Synthesis System
   Performance Evaluation Methods
   by JEITA (2003)
   Revised version of JEIDA Guidelines (2000)
2) Standard of Symbols for Japanese Text-To-Speech Synthesizer
   by JEIDA (2000)

JEITA: Japan Electronics and Information Technology Industries Association

JEIDA: Japan Electronic Industry Development Association

New Project in Japan

CoE Project (5 years from 2003)

Systematization of Large Scale Intellectual Resources and Construction of their Utilization Infrastructure

Project Leader:

Prof. Sadaoki Furui

(Tokyo Institute of Technology)

Korea

SITEC

(Speech Information Technology & Industry Promotion Center)

founded in 2001 (Korean LDC/ELRA)

Wonkwang University as host organization (7 full-time staffs)

1. Creation & distribution of speech corpora including
   standardization of products, performance assessment methods
2. Co-hosting seminars and training programs, publishing books and newsletters related to speech information technology.
3. Assistance to joint researches, academic activities, etc.

China

The National Program of Key Fundamental Research (973 Program)

- Construction of Chinese Corpora
  - Knowledge bases & Lexicons

Chinese LDC, launched in 2002

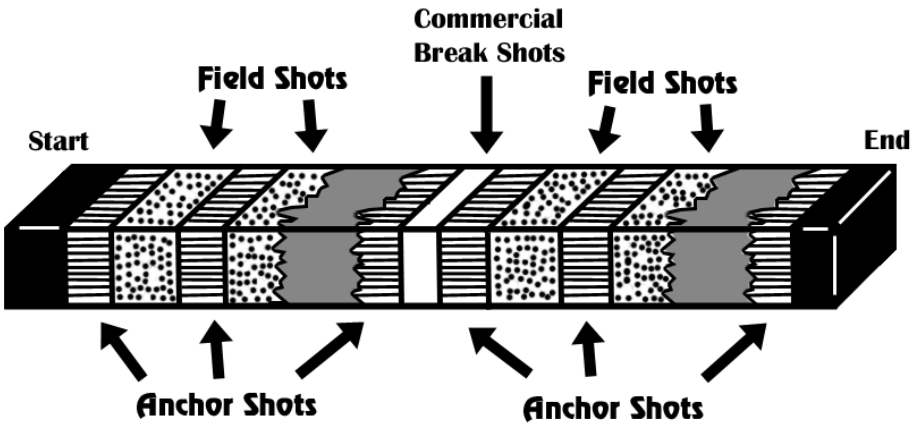
- Creation of linguistic corpora
- Management & distribution of language sources
- Promotion of sharing language resources

AoE-IT Multimedia Repository Hong Kong

- Area of Excellence in Information Technology initiative of the Hong Kong SAR Government
- Repository of Cantonese television news video
  - Content provided by the HK Television Broadcasts Ltd. (TVB)
  - Includes audio, video and text
- Multimedia Markup Language (MmML)
  - Designed for annotation of the AoE-IT Repository
  - Represents multimedia in a structured annotation hierarchy
  - References W3C's SMIL 2.0 specifications
- Related research
  - Repository supports research in Cantonese spoken document retrieval and audio-video fusion for multimedia retrieval

Overview

- Collection of Cantonese television news programs
- Temporal structure of a typical news program



- Anchor shot
  - homogeneous scenes with studio-quality, articulated speech
- Field shots
  - many scene changes with spontaneous speech recorded from variable acoustic conditions, contains language changes

Overview (cont)

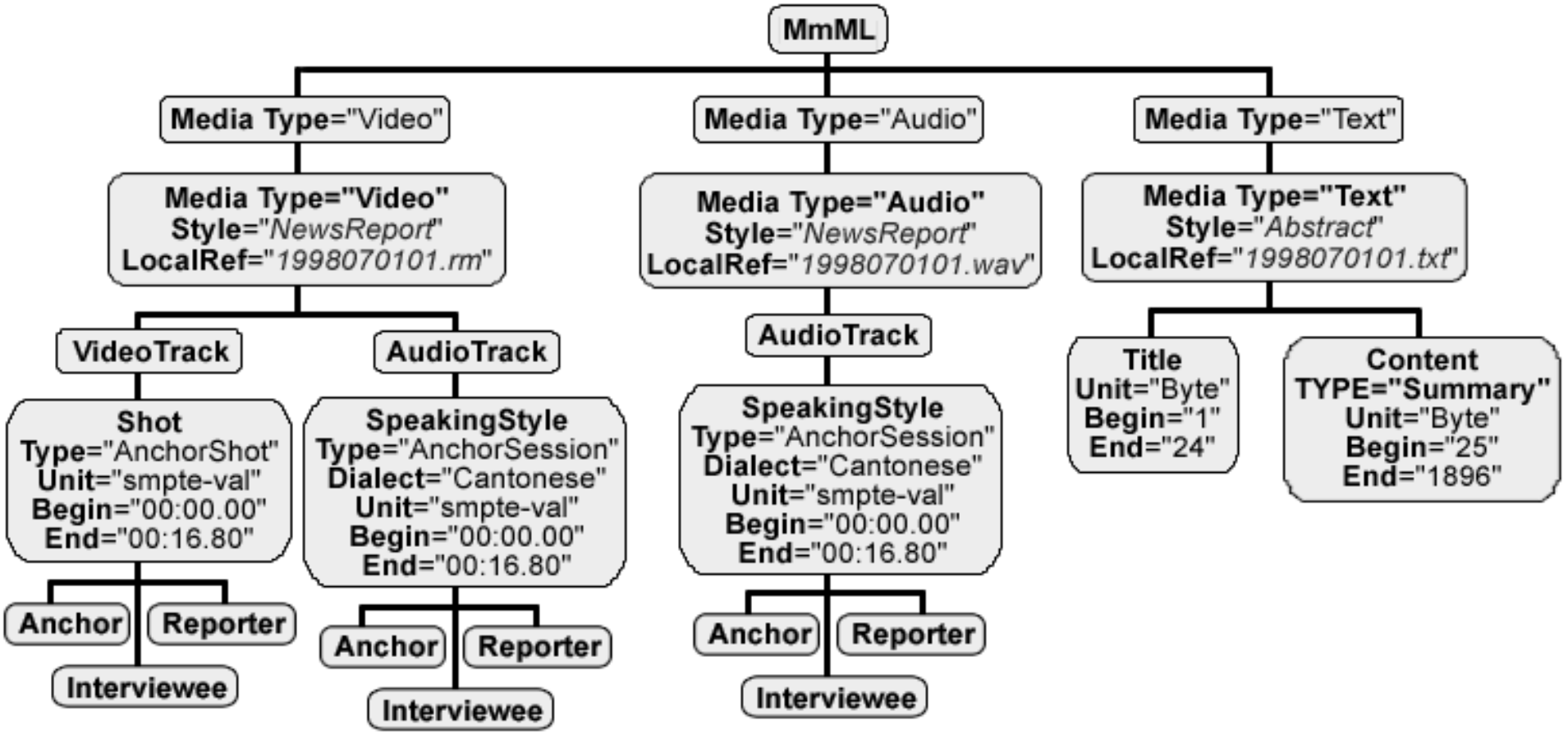
- Manually segmented news stories from news programs
- Each story stored as an individual video file, accompanied by a textual summary with a title
- News stories in two formats:

	Real Media portion	**MPEG-1 portion**
No. of news stories	1722	325
Period	7 Jul. to 31 Dec. 1999	7 Jul. to 31 Jul. 1999
Total # hours of video	39.9	7.4
Average duration per story	1 min 23 sec	1 min 22 sec
Min. story duration	7.9 sec	7 sec
Max. story duration	8 min 14 sec	6 min 50 sec

Multimedia Markup Language (MmML)

- Designed for annotating the multimedia repository
- References W3C's Synchronized Multimedia Integration Language (SMIL) 2.0 specifications
- http://www.w3.org/TR/smil20/
- Describes details / content of non-textual (binary) objects in textual format, e.g.
  - Type of shots (anchor shots from studio, field shots with many scene changes)
  - Type of speech (articulated speech from the anchor, spontaneous speech from the reporters / interviewees)
  - Language and dialect (switching among Cantonese, Putonghua and English)
- Can incorporate user-defined elements / attributes

MmML (cont)


MmML
Media Type="Video"
Media Type="Audio"
Media Type="Text"
Media Type="Video"
Style="NewsReport"
LocalRef="1998070101.rm"
Media Type="Audio"
Style="NewsReport"
LocalRef="1998070101.wav"
Media Type="Text"
Style="Abstract"
LocalRef="1998070101.txt"
VideoTrack
AudioTrack
AudioTrack
Title
Unit="Byte"
Begin="1"
End="24"
Content
TYPE="Summary"
Unit="Byte"
Begin="25"
End="1896"
Shot
Type="AnchorShot"
Unit="smpte-val"
Begin="00:00.00"
End="00:16.80"
SpeakingStyle
Type="AnchorSession"
Dialect="Cantonese"
Unit="smpte-val"
Begin="00:00.00"
End="00:16.80"
SpeakingStyle
Type="AnchorSession"
Dialect="Cantonese"
Unit="smpte-val"
Begin="00:00.00"
End="00:16.80"
Anchor
Reporter
Interviewee
Anchor
Reporter
Interviewee
Anchor
Reporter
Interviewee

Two Mandarin Speech Corpora Developed in Taiwan

Hsiao-Chuan Wang
Department of Electrical Engineering
National Tsing Hua University, Hsinchu, Taiwan, 300-13

Cocosda 2003

Introduction

Two Mandarin speech corpora currently developed in Taiwan;

MATBN is a Mandarin Chinese Broadcast News corpus based on the TV news in Taiwan.

TAICAR is the Mandarin speech collected in car environment.

These corpora are designed for the development and evaluation of speech recognition systems.

MATBN – Mandarin Chinese Broadcast News Corpus

A 3-year project for collecting broadcast news in Taiwan was initiated in August 2002.

This project was sponsored by National Science Council and assisted by Public Television Service Foundation in Taiwan.

Speech Signal

- recorded in TV broadcasting studio,
- converted into a single channel, 16 kHz, 16 bits/sample, linear PCM, WAV format

The corpus has been segmented, labeled, and transcribed manually using a tool developed by DGA and LDC, called “Transcriber”

Two annotation sets designed by Dr. Chiu-Yu Tseng and Dr. Shu-Chuan Tseng of Academia Sinica are applied for annotating process.

The first interim 40-hour Mandarin Chinese broadcast news corpus

779 news stories --

3 anchors (2 males and 1 female) -- 300 minutes,

130 distinct field reporter -- 850 minutes,

1300 distinct interviewee speech -- 650 minutes.

104 headline sections (about 80 minutes)

21 advertising sections (about 13 minutes)

40 weather reporting sections (about 190 minutes)

40 ending sections (about 12 minutes).

TAICAR – In-Car Mandarin Speech

A project for collecting in-car read speech data was coordinated by National Cheng Kung University.

Six microphones – 6-channel speech data

- 1 headphone -- for the driver
- 1 microphone -- behind the steering wheel
- a microphone array of 4 microphones in line with 30 cm space -- in the front window.

The recording is categorized into two cases;

speed below 50 KM/h

speed in 70~100 KM/h

Each speaker provides the speech of two cases.

Recording materials (MAT materials) -- Mandarin syllables, words, and short phrases.

Speakers -- 120 males and 120 females.

Activities on corpora and assessment in Singapore

- Active organizations in Singapore
  - Institute of Infocomm Research http://www.i2r.a-star.edu.sg/
  - National University of Singapore
  - Nanyang Technological University
  - The Chinese and Oriental Language Information Processing Society http://www.colips.org
  - InfoTalk Corp. Ltd http://www.infotalkcorp.com

Activities on corpora and assessment in Singapore

- Active projects in 2002/2003
  - In-car Speech Corpus for Singapore English
    - 175 speakers completed, ~100 utterances each
    - Scripts in general domains
    - 4 channels from a microphone array plus one channel of beam-forming result
    - Manually transcribed
    - Car cruising at 40Kmh to 90Kmh with windows up
    - Carried out by InfoTalk Technology, Knowles Electronics
    - Spoken by Singaporean
    - Benchmarking report available
    - Singapore English linguistic and phonetic research report available

Activities on corpora and assessment in Singapore

- Active projects in 2002/2003
  - Singapore English Telephony Database
    - Targeted 6000 speakers
    - Scripts for stock names, words, numbers, names of people and street
    - Telephony channel
    - Manually transcribed
    - Carried out by the Institute of Infocomm Research
    - Spoken by Singaporean or Singapore Residents
    - Both fixed line and mobile phone, from home, school, offices, in canteens, on the road, etc.
    - Data collection underway

NECTEC-ATR Thai Speech Database

- 40 speakers (20 males and 20 females)
- Three sets (isolated words (DB1), PB sentences (DB2), and dialogues (DB3))
- The qualities of speech are around 20 dB
- The size of this corpus is 42 hours

Summary of NECTEC-ATR Thai Speech Database

Attribute	DB1	DB2	DB3
No. of sentences	None	398	1,637
No. of words	5,771	N/A	18,787
No. of syllables	11,182	5,501	23,308
No. of phones	29,432	14,472	61,489
No. of unique words	5,771	3,377	736
No. of unique syllables	1,668	1,160	586
No. of unique phones	68	72	65
No. of unique biphones	1,296	953	1,619
No. of unique triphones	8,244	6,032	4,437

Thai Large vocabulary continuous spECh (TLEC)

- 248 speakers (124 males and 124 females)
- Four sets (PD set, TR set, DT set and ET set)
- The qualities of speech are 20 dB in office environment and 30 dB in clean speech environment.

Summary of phonetically distributed sentence set

Attribute	*PD set*
No. of sentences	802
No. of vocabularies	2,269
No. of words	7,847
No. of syllables	12,702
No. of phonemes	38,106

Summary of 5,000-words vocabulary sentence set

Attribute	*TR set*	*DT set*	*ET set*
No. of sentences	3,007	500	500
No. of vocabularies	5,000	1,622	1,630
No. of words	55,504	8,076	8,290
Difference from TR	0	3,378	3,370
Difference from DT	0	0	609
Difference from ET	0	617	0

Summary

1. Outline of Oriental COCOSDA
2. Activities in

   Japan, Hong Kong, Taiwan, Singapore, Thailand, China and Korea.

Oriental COCOSDA 2003

October 1–3, Singapore

http://cslp.comp.nus.edu.sg/colips/conference/cocosda2003/

Oriental COCOSDA 2004

Nov. 2004, Delhi, India