Speech Research and Corpora in Thailand

Virach Sornlertlamvanich

Information Research and Development Division
National Electronics and Computer Technology Center
(NECTEC), THAILAND
virach@nectec.or.th

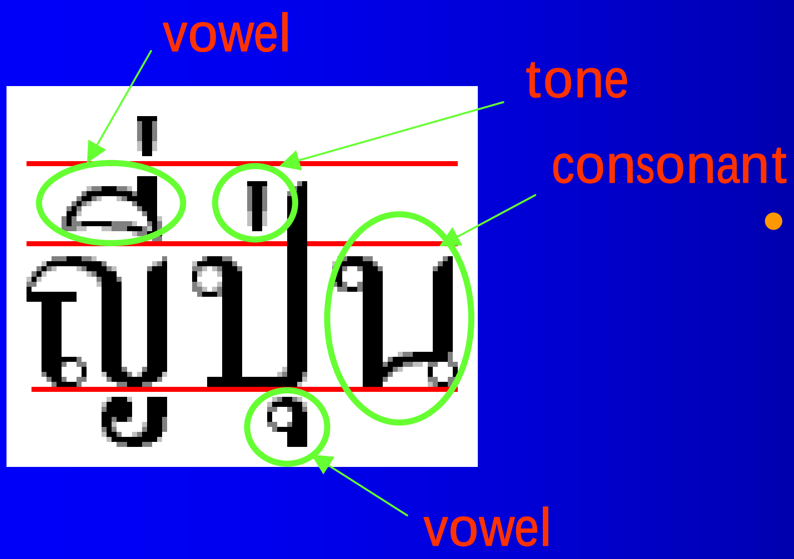
Oriental COCOSDA Workshop 2000, Oct. 16, 2000, Beijing, China.

Introduction to Thai (1): Morphology

- Running text (a paragraph):

สวัสดีครับ ผมชื่อวิรัช ศรเลิศล้ำวาณิช ปัจจุบันเป็นผู้อำนวยการฝ่ายวิจัยและพัฒนาสาขาสารสนเทศ
ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ผมเริ่มสนใจงานวิจัยในสาขาการประมวล
ผลภาษาธรรมชาติตั้งแต่ที่ได้มีโอกาสเข้าร่วมโครงการวิจัยและพัฒนาระบบแปลภาษาในปี 1989

- Writing in 4 levels



- No. of characters (signs)
46 consonants; 18 vowels;
4 tones; 9 symbols; 10 digits
- No word boundary
Ex: “GODISNOWHERE”
1) God is nowhere
2) God is now here
3) God is no where

Introduction to Thai (2): Syntax

- No explicit sentence marker
  - *space character for pausing*
- Sentence pattern
  - (S) (V) (O)

    Ex: ฉัน เห็น เขา

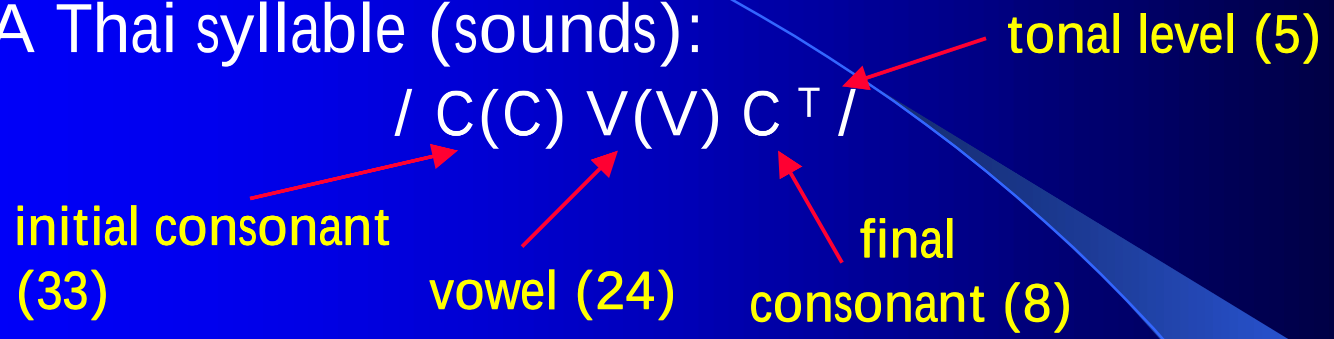
    (I) (saw) (him)
- No inflection forms
  - tenses

    use adverbs and auxiliary verbs
  - plural or singular nouns

    use quantifiers, classifiers or determiners
  - subject-verb agreements
- No syntactic marker
  - *word position*

Introduction to Thai (3): Phonology

- A Thai syllable (sounds):



/ C(C) V(V) C^{T} /

tonal level (5)

initial consonant (33)

vowel (24)

final consonant (8)

- Different tones convey different meanings

  /su:aj4/ = beautiful /su:aj0/ = terrible

- No liaison:

  A word has the same pronunciation, no matter where it is.

- Linking syllable pronunciation:

  ตุ๊กแก (gecko) = tuk4 - kae -> ตุ๊ก = tuk4

  ตุ๊กตา (doll) = tuk4 - ka1 - ta0 -> ตุ๊ก = tuk4 - ka1

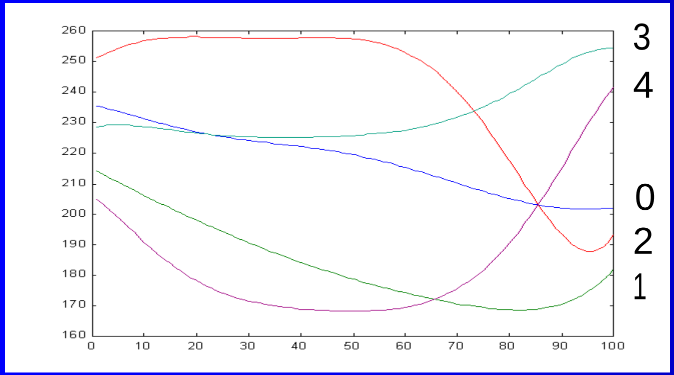
  (grapheme to phoneme conversion)

Introduction to Thai (4): Summary

- Simple grammar
  - easy for generation
  - hard for analysis and recognition
- Sharable problems among Asian languages
  - word segmentation
  - indexing for IR
  - lexical acquisition
  - tone recognition and generation

Research on Speech (1): Recognition

- Tone recognition



Thai Tones

Current state

- Object: Syllable-segmented speech
- Feature: Energy, Zero-crossing, F0
- Method: Neural net,
  Analysis-by-synthesis

Ongoing

- Continuous speech

- Syllable detection

Current state

- Object: Connected speech
- Feature: Energy, Zero-crossing, Duration

Ongoing

- Continuous speech

Research on Speech (2): Recognition

- Isolated word-based recognition

  Current state - Mel-frequency cepstrum (MFC)

  - Neural net, Fuzzy, HMM

  Ongoing - Applications (digits, commands)

- Large vocabulary continuous speech recognition (LVCSR)

  Current state - Isolated phoneme recognition

  - Preparing basic tools for CSR

  Ongoing - Creating LVCSR corpus

Research on Speech (3): Synthesis

- Text analysis

  Current state

  - Word / Phrase / Sentence segmentation by POS tagging model, Rule, Machine learning
  - Letter-to-sound: Rules and Pronunciation dictionary

  Ongoing - Letter-to-sound: PGLR parser (87-94%)

- Speech synthesis

  Current state - Demisyllable-concatenation based

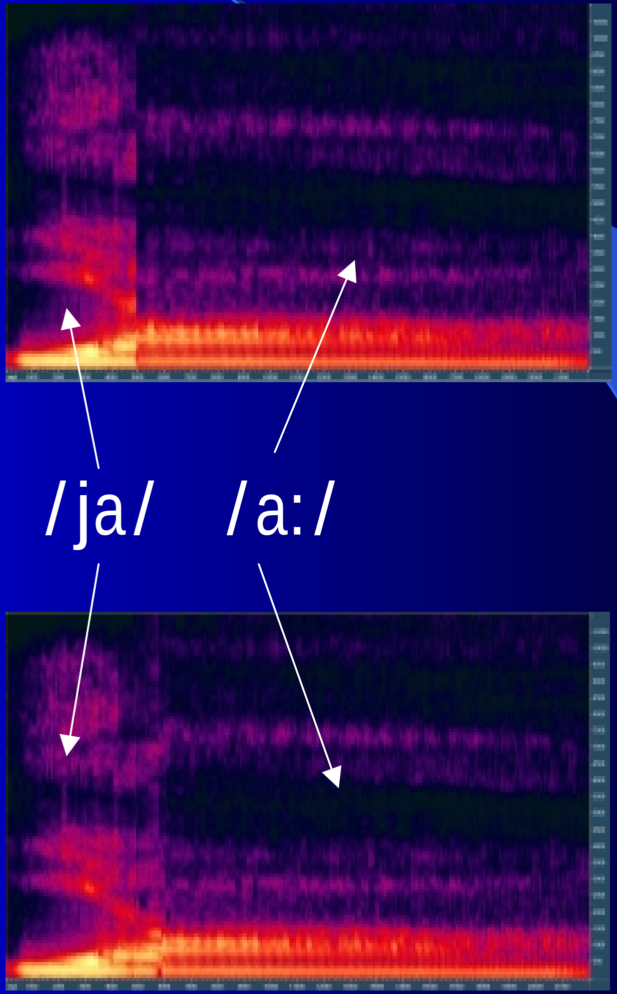
  - LSP-based spectral smoothing
  - Duration adjustment
  - F0 contour smoothing

  Ongoing - Smoothing, Statistical prosody analysis

Research on Speech (4): Synthesis

- LSP parameter smoothing

ยา /ja:/



/ja/ /a:/

Research on Speech (5): Speaker Recognition

- Speaker identification (SID)

Current state - Text-dependent, Closed speaker set, Office environment speech

- Dynamic time warping (DTW; 90-97%), Gaussian mixture model (GMM; 92-98%)

Ongoing - Telephony environment speech

- Speaker verification (SV) - Ongoing

Thai Speech Corpora (1)

- Current state:
  - A number of separated speech corpora
    e.g. Speech database of Thai digits 0-9 for SID
    Speech database of Thai polysyllabic words
- Ongoing:
  - LVCSR corpus for Speech dictation system
    up to 5,000 vocabulary size
    with Phonetically-balanced set
  - Prosody tagging speech corpus
    for statistical prosody analysis
    in improving synthesis system

Thai Speech Corpora (2)

- Basic tools required:

Dictionary	- Manually coding	
	- Corpus-based extraction	
Word segmentation	- Longest matching	(92%)
	- Maximal matching	(93%)
	- POS N-gram	(96%)
	- Machine learning	(97%)
Sentence extraction	- POS N-gram	(85%)
	- Machine learning	(89%)

Thai Speech Corpora (3)

- Basic tools required:

Letter-to-sound - Rule-based and dictionary
- PGLR parser (87%-94%)

Basic tagged corpus - ORDHID: POS tagging corpus
160 documents;
5.75 MB; 311,426 words

Other tools - Automatic sentence selection for phonetically balanced set
- Automatic phoneme labeling

Thai Text to Speech: Demo

สวัสดีครับ ผมชื่อวิรัช ศรเลิศล้ำวาณิช ปัจจุบันเป็นผู้อำนวยการฝ่ายวิจัยและพัฒนาสาขาสารสนเทศ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ผมเริ่มสนใจงานวิจัยในสาขาการประมวลผลภาษาธรรมชาติตั้งแต่ที่ได้มีโอกาสเข้าร่วมโครงการวิจัยและพัฒนาระบบแปลภาษาในปี 1989

Hello, I am Virach Sornlertlamvanich, the director of Information Research and Development Division, National Electronics and Computer Technology Center. I began to interest myself in the research of Natural Language Processing since having a chance in participating in the Machine Translation Research and Development project in 1989.