

O-COCOSDA 2010 **Thailand Report** November 2010

- **LOTUS**: Read-speech LVCSR evaluation
- **LOTUS-BN**: Broadcast-news speech corpus
- **LOTUS-CELL**: Telephone conversation corpus
- **TSYNC/BSYNC**: Speech synthesis corpora
- **NEWS**: Name transliteration corpus



LOTUS Evaluation

- Kasuriya et al., “Thai speech corpus for speech recognition”, O-COCOSDA 2003.
- Chotimongkol et al., “Toward benchmarking a general-domain Thai LVCSR system”, ECTI-CON 2010, Chiangmai, Thailand.

- **Large vOcabulary Thai continUous Speech (LOTUS) corpus**

- Read speech on newspaper/magazine text
- Clean/office environment, close-talk/unidirectional mic

Attribute	PD set	TR set	DT set	ET set
No. of utterances	802	3,007	500	500
Vocabulary size	2,269	5,000	1,622	1,630
No. of words	7,847	55,504	8,076	8,290
No. of speakers	48	24	12	12

LM	Perplex	WER
TR	121.8	25.7
BEST	157.7	26.6
TR+BEST	81.3	24.4

* Tested on ET
BEST: 5M-word newspaper text



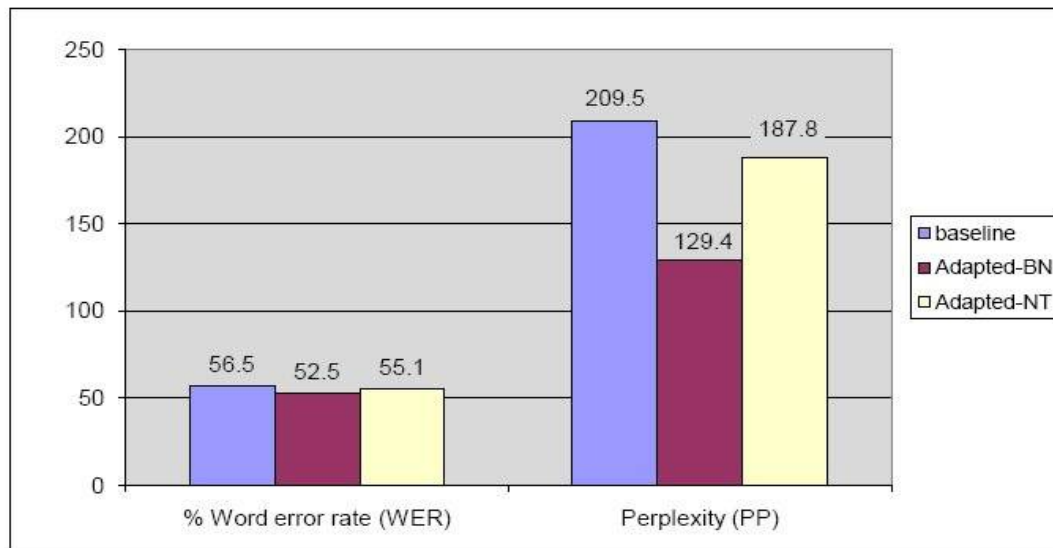
LOTUS-BN and BN Transcription

• Saykham et al., “Online temporal language model adaptation for a Thai broadcast news transcription system”, LREC 2010.

• LOTUS-Broadcast news (LOTUS-BN)

- Collaboration with Tokyo Institute of Technology

Data set	Period	Amount	Words	Vocab
TIT-BN	01/2007 - 03/2007	17 hrs.	223,993	10,303
NECTEC-BN	10/2007 - 03/2008	60 hrs.	529,136	23,993
NewsText	10/2007 - 03/2008	6,074 articles	1,357,702	38,485
Test-BN	04/2008	0.5 hrs.	4,150	1,300



LOTUS-CELL 2.0

• Chotimongkol et al., “The development of a large Thai telephone speech corpus”, O-COCOSDA 2010.

• Structure and conditions

- Three major mobile channels in Thailand
- **Questions:** Name, Place, Spelling, Date, Time, Address, Digit sequence, Amount, Discussion on given topics
- **Tags:** adopted from CMU Communicator Transcription Conventions (including pronunciation variations, incorrect pronunciations, abbreviations, numbers, false-starts, fillers, tone marks)

Part	Description	#Hrs	#Spk	Transcribed currently				
				#Hrs	#Word (x1000)	#Vocab	#False start	#Filler
1	Closed-ended questions	16.5	84	5.8	12.3	1,453	14	86
2	Open-ended questions	14.4	33	1.3	6.2	1,047	41	93
3	Dialogues	59.2	21	5.2	14.5	1,625	177	1,264
Total		90.1		12.3	33.0	3,193	232	1,443

TSYNC 2.0 & BSYNC

• Wutiwiwatchai et al., “An intensive design of a Thai speech corpus”, SNLP 2007, Pattaya, Thailand.

- **Thai speech Synthesis Corpus (TSynC) 2.0**
 - Reducing the size of corpus by 25% from TSynC 1.0 (14 to 10 hrs.) by excluding unnecessary tonal diphones with no effect to the synthesized speech quality
 - Increasing the number of speakers from 1 to 4 (female/male, adult/child)
- **Bilingual speech Synthesis Corpus (BSynC)**
 - Only Thai prompts in TSynC, allowing synthesis of English words from only Thai phonemes
 - Better creating a speech synthesis corpus containing both Thai and English speech from bilingual speakers
 - **Prompts:** TSynC 2.0 for Thai, CMU ARCTICS for English
 - 1 female, 1 male



NEWS 2010

• Wutiwiwatchai and Thangthai, “Syllable-based Thai-English machine transliteration”, NEWS 2010 Workshop in ACL 2010, Uppsala, Sweden.

- **Thai-English name transliteration corpus**
 - A resource available in Named Entity Transliteration Workshop (NEWS) 2010
 - **Entries:** English names and common words mostly transliterated into Thai
 - **Rules:** Following Thailand Royal Institute guideline with only one most frequently-used transliterated form

Total # syllables	39,537
Avg. # syllables/word	2.4
# distinct syllables	4,367 (Thai) 6,307 (English)
# distinct syllable sounds	1,869
Avg. # homophones/syllable sound	2.3 (Thai) 3.4 (English)
Max. # homophones/syllable sound	16 (Thai) 38 (English)

