Perceptual Representation of Consonant Sounds in Thai

C. Tantibundhit[1], C. Onsuwan[2], T. Saimai[1], N. Saimai[1],
S. Thatphithakkul[3], P. Chootrakool[3], K. Kosawat[3], N. Thatphithakkul[3]
[1] Department of Electrical and Computer Engineering, Thammasat University, Thailand
[2] Department of Linguistics, Thammasat University, Thailand
[3] National Electronics and Computer Technology Center (NECTEC), Thailand
tchartur@engr.tu.ac.th, consuwan@tu.ac.th

Abstract

This work is an attempt to construct a perceptual representation of Thai consonants based on perceptual identification results (from 28 Thais) of 21 phonemes presented in noise. The experiment is designed to equally make pairwise comparisons among 21 word-initial phonemes, which results in 210 real-word stimulus pairs. Percent correct responses and confusion matrices are obtained. Similarity score and perceptual distance for each phoneme pair are systematically derived from confusion scores based on a method proposed by Shepard (Psychological Representation of Speech Sounds, 1972). Piecing this together pair by pair, a perceptual space or representation of Thai consonants takes shape. The perceptual space could roughly be divided into 5 non-overlapping groupings: glide, glottal constriction, nasality, aspirated obstruent, and a combination of liquid and unaspirated obstruent. It is suggested that these phonological classes reflect the most distinct and relevant perceptual properties of Thai consonants. Preliminary cross-linguistic observation is addressed in light of the data of English consonants from Miller and Nicely (J. Acoust. Soc. Am., 1955).
Index Terms: Thai, initial consonant, perceptual similarity/distance, intelligibility, confusion matrix, cross-language study

1. Introduction

Analyses of perceptual confusions among speech sounds provide valuable information in determining and understanding speech perception in general and cross-linguistically [1]. By and large, there are two main motivations behind these types of analysis. First of all, confusion patterns provide essential clues for the understanding of how speech signals are auditorily processed and transformed as some parts of the signals will become more distinct while others suppressed [2]. This insight is crucial for a number of areas in speech research, including speech recognition [3]. Secondly, a number of cross-linguistic perception experiments have shown that perception of speech sounds is not only limited to the input from the auditory system, but also the result of perceptual representations, which are largely shaped by listeners' language experience [4]. Importantly, perceptual confusion patterns, which generally reflect phonological predisposition of speech sounds, will provide a more reasonable explanation for a connection between language, i.e., its sound inventory and (human) auditory constraints [2].

A number of studies have focused on confusion analyses of English consonants. Among them, a classic report from Miller and Nicely [5], where perception of English word-initial consonants (16 phonemes) in an open-response task was conducted under different bandwidths (in between 200–6500Hz) and different signal to noise ratios (SNRs) (−18, −12, −6, 0, +6, and +12dB). Shepard proposed a method to assess a psychological representation of speech sounds by computing similarity and distance scores from confusion matrices [6]. He applied his formula and method to the English perceptual data from [5]. The analysis showed that the perceptual representation of English consonants could be grouped according to 2 phonological dimensions (adapted from [7]), that of nasality and a combination of voicing and frication, suggesting that nasality, voicing, and frication are the strongest perceptual features for English consonants. Benkí examined 10 English word-initial phonemes using 4 degrees of SNR (−14, −11, −8, and −5dB) in an open-response task [8]. His investigation was expanded to include confusion matrices of 10 English final phonemes and 9 vowels. His findings confirmed that voicing feature is stronger than feature for place of articulation and that initial consonants are more distinct than finals [8].

Cross-linguistically, Singh and Black explored speaker-listener errors of phonemic and non-phonemic intervocalic sounds of 4 languages: Arabic, English, Hindi, and Japanese [9]. The findings revealed perceptual similarities and differences across languages. It would be of special interest to compare the English perceptual representation with that of a language, which has a comparable phoneme inventory size. In this respect, a language such as Thai, with 21 phonemes and all appear in word-initial position [10], resembles English, with 24 phonemes, 22 of which (except /ŋ/ and /ʒ/) occur word-initially [11]). However, the two languages differ phonologically in many aspects. For instance, Thai has a 3-way stop/affricate distinction (voiced, voiceless unaspirated, and aspirated), while there exists a 2-way distinction (voiced and voiceless) in English. English has 11 fricatives/affricates, whereas Thai has only four. It will be interesting to see how these differences play out in the phonological representations between the two languages.

To date, a very small number of studies on Thai have investigated confusion of Thai speech sounds, including tones (e.g., [12], [13]). As a result, there remains a large gap for the knowledge of perceptual representation. In part, the work presented here is hoped to fill this remaining gap. Our perceptual representation of Thai phonemes is a valid approximation derived from listeners' responses of initial phoneme identification in noise following a method of Thai diagnostic rhyme test (TDRT), which we developed. Undertaking an investigation from the complete set of 21 initial phonemes, this current study will provide some insight into the abstract yet consequential representation in the case of Thai speech sounds.

The TDRT is an adaptation of several useful frameworks, namely diagnostic rhyme test (DRT) [14], new Chinese diagnostic rhyme test (NCDRT) [15], and the analysis method of balanced confusion matrix [5]. More generally, we manipulated DRT, which is widely used for a subjective testing for measuring the intelligibility of speech coders [14]. The DRT is an A/B forced comparison test based on 96 rhyming pairs, e.g., *veal – feel* [14]. Because the DRT was

developed specifically for English, it has some limitations when evaluating intelligibility of a tonal language such as Chinese [15] and Thai. Therefore, we reviewed NCDRT, which is a subjective intelligibility test for Chinese phoneme and tone [15]. In the end, we developed the A/B forced choice and monosyllabic (CV(V)(C)) rhyming pairs, which differ only in word-initial sounds [14] with identical tone [15]. The designed rhyming words are real and commonly used words in the language [14], [15]. TDRT allows us to systematically compare confusion responses across all phonemes in Thai.

The organization of the paper is as follows: Section 2 reviews the Thai Phonology. Sections 3 and 4 provide details of the design of TDRT for initial phonemes and experimental setup. Section 5 presents experimental results (confusion matrix and distance matrix). Section 6 discusses the paper and mentions future work.

2. Thai Phonology Review

Thai is a tonal language composed of 21 phonemes in initial position /p/, /pʰ/, /b/, /t/, /tʰ/, /d/, /tɕ/, /tɕʰ/, /k/, /kʰ/, /ʔ/, /f/, /s/, /h/, /m/, /n/, /ŋ/, /l/, /r/, /w/, and /j/ and 9 phonemes in final position /p/, /t/, /k/, /ʔ/, /m/, /n/, /ŋ/, /w/, and /j/ [10]. Each of the nine monophthongs in Thai occurs phonemically short or long (/i/ อิ, /ii/ อี, /e/ เอะ, /ee/ เอ, /ɛ/ แอะ, /ɛɛ/ แอ, /ɯ/ อึ, /ɯɯ/ อือ, /ɤ/ เออะ, /ɤɤ/ เออ, /a/ อะ, /aa/ อา, /u/ อุ, /uu/ อู, /o/ โอะ, /oo/ โอ, /ɔ/ เอาะ, and /ɔɔ/ ออ) [10].

Thai syllables consist of a tone and up to two initial consonants followed by a short vowel and a final consonant or by a long vowel and an optional final consonant. There are five tones: Mid /¯/, Low /ˆ/, High /ˊ/ (with a level pitch contour), Falling /ˆ/, and Rising /ˇ/ (with a non-level pitch contour). Thus, Thai syllables may be represented as $C_i(C)V^TC_f$ or $C_i(C)V^TV(C_f)$, where C_i stands for an initial consonant, C_iC a consonantal cluster, C_f a final consonant, V a short vowel, VV a long vowel, and T a tone [16].

3. TDRT Design and Development

Thai diagnostic rhyme test (TDRT) for initial phonemes is well designed such that percent intelligibility can be evaluated and confusion responses across all phonemes can be systematically compared. Moreover, TDRT test is considerable short to avoid subject's fatigue [17]. The following steps present the design and development of TDRT's wordlist.

1) Multiple sets of monosyllabic ($C_iV^T(V)(C_f)$) words, each of which differs only in their initial phoneme are gathered.
2) Vowel /aa/ along with mid tone are chosen because it is one of the most frequently used vowels [18] and when combined with mid tone yields the most possible number of rhyming words, i.e., 21 rhyming words for 21 phonemes: /pāa/ ปา, /pʰāa/ พา, /bāa/ บา, /tāa/ ตา, /tʰāa/ ทา, /dāa/ ดา, /tɕāa/ จา, /tɕʰāa/ ชา, /kāa/ กา, /kʰāa/ คา, /ʔāa/ อา, /fāa/ ฟา, /sāa/ ซา, /hāa/ ฮา, /māa/ มา, /nāa/ นา, /ŋāa/ งา, /lāa/ ลา, /rāa/ รา, /wāa/ วา, and /jāa/ ยา. All 21 words are real words in the language.
3) Each rhyming word is paired with 20 others of different initial phonemes. This results in a total combination of 210 rhyming words[i], which can be expressed mathematically as *a combination of 21 choose 2* ($^{21}C_2$).

4. TDRT Setup

This section describes the experimental setup of TDRT for initial phonemes to obtain confusion responses among phonetic categories. As the next step involves converting confusion scores into similarity scores and distances [6], a detailed explanation of calculation method is given in 4.3.

4.1. Intelligibility Test Setup

The 21 rhyming words (in Section 3) along with filler words were read 5 times in a carrier sentence and recorded at a sampling rate of 44.1kHz in a sound-attenuated chamber by a 36-year-old Thai male speaker who was born and grew up in Bangkok. Then, one of the 5 tokens of each target word was selected based on impressionistic hearing evaluation and spectrographic inspection.

Different types of masking noise have been used to obtain confusion matrices [17]. In this study, to be in line with [5], white Gaussian noise is employed. The selected tokens are corrupted by additive white Gaussian (AWG) noise with 4 SNR levels, i.e., –6, –12, –18, and –24dB, respectively. These levels are chosen based on a preliminary experiment such that percent intelligibility scores, calculated from $P_e = \frac{N_r - N_w}{T} \times 100\%$, where P_e, N_r, N_w, and T are percent intelligibility score, numbers of correct responses, numbers of wrong responses, and total numbers to stimuli, respectively [14], are not either close to 100% (so well perceived stimuli) or close to 50% (indistinguishable from guesswork) [15].

The intelligibility tests were conducted on untrained 28 volunteer subjects with normal hearing over headphones in a quiet room. In each trial, listeners hear a target stimulus and are asked to choose what they just hear between 2 rhyming words, appearing on the computer screen. If they do not recognize the stimulus, they are instructed to guess before moving on to the next trial.

4.2. Balanced Confusion Matrix

Table 1: *Distributions of rhyming word groupings.*

Subject	Rhyming and filler word			
	Group A	Group B	Group C	Group D
I	–6dB	–12dB	–18dB	–24dB
II	–24dB	–6dB	–12dB	–18dB
III	–18dB	–24dB	–6dB	–12dB
IV	–12dB	–18dB	–24dB	–6dB

TDRT for initial phonemes consists of 210 rhyming pairs across 21 initial phonemes and 40 pairs of filler words. To bring out a balanced confusion matrix, the rhyming word in each pair is presented once as a stimulus in a trial, resulting in a total of 420 trials for initial phonemes and 80 trials for filler words.

A straightforward test of 500 trials × 4 SNR levels would create a test of 2,000 trials, which is considerably long and could cause subject's fatigue and learning effect [17]. Alternatively, by increasing a number of subjects 4 times, we could stay with the 500 trials and divide the test equally by 4 SNR levels. Consequently, the 500 trials are corrupted by one of 4 SNR levels of AWG noise stated earlier, i.e., Groups A, B, C, and D, each of which has an SNR level of –6dB, –12dB, –18dB, and –24dB, respectively as summarized in Table 1. With regard to distributions of the rhyming words, subjects' performance per SNR level is equally distributed yielding 105 trials/SNR level (420 trials/4 SNR levels). Each of the 105 trials is equally distributed across 21 phonemes resulting in 5 trials/SNR level/phoneme (420 trials/4 SNR levels/21

phonemes). Finally, ordering of individual trials as well as sequence of words in each A/B pair are randomized in the test.

4.3. Measure of Distance Matrix

Similarity score between each pair of phonemes is calculated from confusion scores based on Shepard's method [6], i.e., $S_{ij} = \frac{P_{ij}+P_{ji}}{P_{ii}+P_{jj}}$, where S_{ij} is the similarity between phoneme i and phoneme j, P_{ij} is an element of confusion matrix when stimulate with phoneme i (row) and perceive as phoneme j (column) and so forth. Then, perceptual distance (d_{ij}) is derived from the similarity score, i.e., $d_{ij} = -\ln(S_{ij})$ [6]. Finally, a distance matrix is constructed from the calculated distances.

5. Experimental Results

Table 2: *Average percent intelligibility for all SNR levels.*

Average percent intelligibity	SNR (dB)			
	–6dB	–12dB	–18dB	–24dB
	93.06%	87.14%	77.35%	24.08%

Percent intelligibility scores (P_e) for initial phonemes across 4 SNR levels are calculated and shown in Table 2. It is clear that percent intelligibility scores are decreasing as increasing level of noise. While subject's performance at SNR level of –6dB and –12dB is near-perfect, at –24dB it is far below a guesswork (50%). Therefore, the remaining SNR level of –18dB is the most interpretable and is used to construct the confusion matrix, to compute similarity scores, and to derive a distance matrix.

The confusion matrix of 21 initial phonemes at SNR level of –18dB is constructed along with the distance scores (underlined figures) and shown in Table 3. From the matrix, /r/ is the most confusable phoneme followed by /t^h/, while /j/ is the least confusable phoneme followed by /w/. A separate analysis for listeners' misidentified responses reveals that the listener favored /t/, /p/, /t^h/, /tɕ^h/, and /tɕ/ and disfavored /w/ and /r/ over other phonemes.

A perceptual space showing relative locations for each Thai initial phoneme is sketched in Fig. 1 according to [1], [6], which graphically represents the confusion patterns in Table 3. Between each pair of phonemes, less distance in space means more confusions. It is worth noting that Fig. 1 is an approximation of 2 dimensional perceptual representation (with limited scaling). Consequently, the value of infinity covers a relatively large distance, if not exceptionally large. There are five clustering of phonemes (shown in dashed-line circles) based on their common phonological features adapted from [7], i.e., glide (/w/ and /j/), glottal constriction (/ʔ/ and /h/), nasality (/m/, /n/, and /ŋ/), aspirated obstruent (/p^h/, /t^h/, /tɕ^h/, and /k^h/), and a combination of liquid and unaspirated obstruent (/p/, /b/, /t/, /d/, /tɕ/, /k/, /f/, /s/, /l/, and /r/). Interestingly, /r/, the most confusable phoneme, is nicely located in the middle of the perceptual space and towards the center of its own group.

Table 3: *Confusion matrix and distance scores (underlined figures) for 21 initial phonemes in Thai at SNR level of –18dB .*

	p	*p^h*	*b*	*t*	*t^h*	*d*	*tɕ*	*tɕ^h*	*k*	*k^h*	*ʔ*	*f*	*s*	*h*	*m*	*n*	*ŋ*	*l*	*r*	*w*	*j*
p	**133**	0	0	4	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
		5.5	∞	4.1	3.9	5.6	5.5	4.9	5.6	∞	∞	3.6	3.9	5.6	∞	5.5	∞	5.6	5.4	∞	∞
p^h	1	**121**	0	1	6	0	1	4	0	2	1	0	1	2	0	0	0	0	0	0	0
			∞	4.1	3.1	∞	5.5	3.7	∞	3.9	4.8	∞	5.5	3.7	∞	∞	∞	∞	5.4	∞	∞
b	0	0	**137**	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
				4.8	5.5	5.6	∞	∞	∞	5.5	∞	5.6	5.5	∞	∞	∞	5.6	∞	4.0	5.6	∞
t	0	3	2	**116**	2	0	4	1	3	0	0	5	3	0	0	0	0	0	1	0	0
					4.0	∞	3.4	4.8	3.7	4.8	∞	3.5	3.5	∞	∞	5.5	∞	∞	4.0	∞	∞
t^h	4	3	1	2	**112**	1	3	4	2	2	0	2	3	1	0	0	0	0	0	0	0
						5.5	4.0	3.7	4.8	3.4	5.5	4.4	4.0	4.4	∞	∞	∞	∞	5.3	∞	∞
d	1	0	0	0	0	**132**	3	0	4	0	0	0	0	0	0	0	0	0	0	0	0
							4.1	5.6	4.2	∞	∞	4.9	5.5	∞	∞	∞	5.6	5.6	3.8	∞	∞
tɕ	1	0	0	4	1	1	**115**	5	1	0	0	3	5	0	0	1	0	1	0	1	1
								3.7	4.4	5.5	∞	4.4	3.7	∞	∞	5.5	∞	4.4	3.7	4.8	5.5
tɕ^h	2	2	0	1	2	1	1	**126**	0	2	1	0	1	0	0	0	0	1	0	0	0
									∞	4.1	5.5	∞	4.8	∞	5.6	∞	5.5	4.4	4.3	∞	∞
k	1	0	0	3	0	0	2	0	**130**	0	0	0	1	0	0	0	1	1	0	0	1
										5.5	5.6	∞	4.8	∞	5.6	∞	5.6	4.8	3.6	∞	5.6
k^h	0	3	1	2	6	0	1	2	1	**117**	5	0	0	2	0	0	0	0	0	0	0
											3.7	∞	5.5	4.1	∞	5.5	∞	5.5	4.6	∞	∞
ʔ	0	1	0	0	1	0	0	0	1	1	**128**	0	0	4	1	1	2	0	0	0	0
												∞	5.5	3.5	5.6	4.1	3.9	∞	∞	∞	∞
f	5	0	1	2	1	2	0	0	0	0	0	**126**	1	0	0	0	0	1	0	1	0
													3.7	∞	∞	5.5	∞	5.5	5.4	5.6	∞
s	5	0	1	4	1	1	1	1	1	1	1	5	**116**	0	0	0	0	1	1	0	0
														∞	∞	5.5	∞	5.5	4.2	∞	∞
h	1	4	0	0	2	0	0	0	0	2	4	0	0	**123**	1	0	2	1	0	0	0
															4.8	4.1	4.4	5.5	5.4	∞	∞
m	0	0	0	0	0	0	0	1	1	0	0	0	0	1	**130**	2	3	0	1	1	0
																4.1	4.2	5.5	4.7	5.6	∞
n	1	0	0	1	0	0	0	0	0	1	3	1	1	4	2	**122**	2	0	2	0	0
																	3.9	5.5	4.3	∞	∞
ŋ	0	0	0	0	0	1	0	1	0	0	3	0	0	1	1	3	**129**	1	0	0	0
																		4.8	4.0	∞	∞
l	1	0	0	0	0	1	2	2	1	1	0	0	0	0	1	1	1	**124**	1	2	2
																			3.8	4.9	4.9
r	1	1	4	3	1	5	5	3	6	2	0	1	2	1	1	1	4	4	**91**	0	4
																				∞	4.1
w	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	**139**	0
																					∞
j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**140**

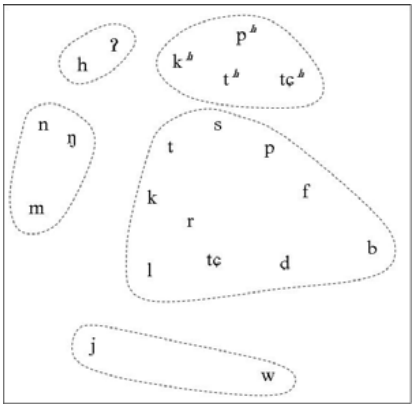

Figure 1: Perceptual space of 21 initial phonemes in Thai.

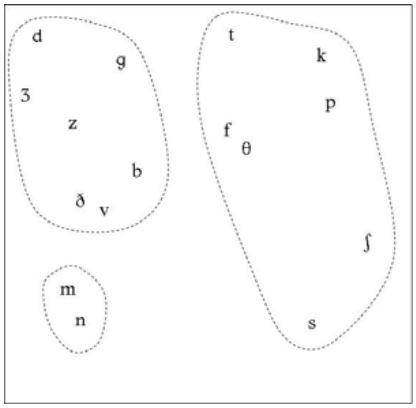

Figure 2: Perceptual space of 15 initial phonemes (plus /ʒ/) in English adapted from a confusion matrix of [5].

6. Discussion and Future Work

Current speech perception models provide different accounts for a basic unit of analysis such as context dependent allophones [19], individual exemplar [20]. In this study, we have taken a more traditional approach and consider phoneme as the unit of analysis. TDRT for initial phoneme is well designed such that subjective intelligibility score can be easily obtained and a balanced confusion matrix efficiently constructed. Then, perceptual similarity and distance scores could be computed.

Figure 2 illustrates a perceptual representation of 16 (out of 22) English initial phonemes at SNR level of +12dB with a bandwidth of 200–6500Hz adapted from a confusion matrix of [5]. This condition is chosen for the comparison because it yielded average percent correct response (88.67%), which is relatively close to that of –18dB of the Thai data (90.85%). It should be pointed out that the average percent correct response, which does not necessarily match the intelligibility score (Table 2), is calculated from total number of correct responses divided by total number of stimuli.

Some cross-linguistic observations could be made concerning perceptual representations of Thai (Fig. 1) and English (Fig. 2) phonemes. (It is noteworthy that 8 phonemes in English (/ʧ/, /ʤ/, /ŋ/, /l/, /ɹ/, /w/, /j/, and /h/) were not included in the study of [5]). Separately, nasal sounds are grouped together in both languages. Voicing, one of the most distinct perceptual properties in English, appears to be a non-robust feature in Thai, where aspiration plays a more significant role. It is interesting that obstruents in English form a cluster which can be further divided into fricatives and plosives. On the other hand, in the case of Thai unaspirated obstruents, the separation among fricatives, affricates, and plosives seems less clear. Moreover, liquids in Thai seem to belong to the same cluster as the 8 unaspirated obstruents, the finding which calls for further investigation and explanation.

We are in the process of developing a method to test subjective intelligibility of final phonemes, vowels, and tones in Thai. Analyses of confusion matrices similar to what is presented will be carried out to broaden an understanding of Thai speech sounds.

7. References

[1] Johnson, K., Acoustic and Auditory Phonetics 2nd edition, Wiley-Blackwell, Malden, MA, 2003.

[2] Stevens, K. N., "Contraints imposed by the auditory system on the properties used to classify speech sounds: data from phonology, acoustics, and psychoacoustics," in T. Myers, J. Laver, J. Anderson [Eds], The Cognitive Representation of Speech, 61–74, North-Holland, 1981.

[3] Mermelstein, P., Distance Measures for Speech Recognition-Psychological and Instrumental. *Haskins Laboratories Status Report on Speech Research* SR-47, 91–103, 1976.

[4] Strange, W., Speech Perception and Linguistic Experience, York Press, MD, 1995.

[5] Miller, G. A., Nicely, P. E., "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am., 27(2): 338–352, 1955.

[6] Shepard, R. N., "Psychological representation of speech sounds," in E.E. David and P.B. Denes [Eds], Human Communication, A Unified View, 67–113, McGraw-Hill, NY, 1972.

[7] Jacobson, R., Fant, G., Halle, M., Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates, MIT Press, Cambridge, MA, 1952.

[8] Benkí, J. R., "Analysis of English nonsense syllable recognition in noise," Phonetica, 60(2): 129–157, 2003.

[9] Singh, S., Black, J. W., "Study of twenty-six intervocalic consonants as spoken and recognized by four language groups," J. Acoust. Soc. Am., 39(2): 372–387, 1966.

[10] Tingsabadh, K., Abramson, A. S., "Thai," Journal of the International Phonetic Association, 23(1): 24–28, 1993.

[11] Giegerich, H. J., English Phonology: An Introduction, Cambridge University Press, Cambridge, 1992.

[12] Gandour, J., Dardarananda, R., "Identification of tonal contrasts in Thai aphasic patients," Brain Lang., 18(1): 98–114, 1983.

[13] Thubthong, N., A study of various linguistic effects on tone recognition in Thai continuous speech. Doctoral dissertation, Chulalongkorn University, Bangkok, 2001.

[14] Voiers, W.D., "Evaluating processed speech using the diagnostic rhyme test," Speech Technol., 1(4): 30–39, 1983.

[15] McLoughlin, I., "Subjective intelligibility testing of Chinese speech," IEEE Trans. Audio, Speech, Lang. Process., 16(1): 23–33, 2008.

[16] Comrie, B., The World's Major Languages, Oxford University Press: Oxford, 1990.

[17] Loizou, P.C., Speech Enhancement: Theory and Practice, CRC Press: Boca Raton, FL, 2007.

[18] Kosawat, K., "BEST," Human Language Technology Laboratory, Online: http://www.hlt.nectec.or.th/best/, accessed on 1 Mar 2011.

[19] Ingram, J. C., Park, S. G., "Language, context, and speaker effects in the identification and discrimination of English /r/ and /l/ by Japanese and Korean listeners," J. Acoust. Soc. Am., 103(2): 1161–1174, 1998.

[20] Johnson, K., "Speech perception without speaker normalization: an exemplar model," in K. Johnson and J. W. Mullennix [Eds], Talker Variability in Speech Processing, 145–165, Academic Press, San Diego, CA, 1997.

[i] Complete wordlist with translation and transcription can be accessed at http://charturong.ece.engr.tu.ac.th/Interspeech2011/Initials.pdf.