Final Report

Project Title Subjective Intelligibility Testing of Thai Speech for Initial and Final Consonants

By Assoc. Prof. Dr. Charturong Tantibundhit

12/ 2014

Contract No. MRG5480272

Final Report

Project Title Subjective Intelligibility Testing of Thai Speech for Initial and Final Consonants

Researcher	Institute
1 Assoc. Prof. Dr. Charturong Tantibudhit	Thammasat University
2 Prof. Dr. Apirat Siritaratiwat	Khon Kaen University

This project granted by the Thailand Research Fund

Abstract

Project Code : MRG5480272

Project Title : Subjective Intelligibility Testing of Thai Speech

Investigator : Assoc. Prof. Dr. Charturong Tantibundhit

E-mail Address : tchartur@engr.tu.ac.th

Project Period : 3 years

Abstract: We methodically designed and developed a subjective intelligibility testing of Thai speech for initial and final consonants based on the diagnostic rhyme test (DRT). The Thai DRT (TDRT) consists of 2 test sets, one for initials (TDRT-I) and the other final consonants (TDRT-F). The test for initials is designed to equally compare 21 phonemes pairwise, which results in 210 stimulus pairs. The TDRT-F compares 8 final phonemes, yielding 84 stimulus pairs. The tests are well-constructed using real monosyllabic words. TDRT-I and TDRT-F have main ad- vantages in that percent intelligibility scores in each stimulus pair as well as confusion patterns across all phonemes can be evaluated and compared. To confirm its validity, we carried out a series of experiments. The subjective intelligibility tests were conducted on 28 Thai normal hearing listeners in four SNR levels (-6, -12, -18, and -24 dB for TDRT-I and TDRT-F) and subsequently on eight sensorineural hearing loss patients (with and without hearing aids) using clean stimuli (for TDRT-I). Average intelligibility scores, percent correct responses, and confusion matrices were obtained. Comparisons of confusion patterns in both subject groups showed that for both initial and final consonants, voicing was the most robust contrast while place-of-articulation was the least. Specifically, for initials, /r/ is the most confusable phoneme, while /w/, /j/, and /p/ are among the least. Perceptual representation spaces, derived from confusion matrices, yielded five non-overlapping groupings: glide, glottal constriction, nasality, aspirated obstruent, and a combination of liquid and unaspirated obstruent. The results suggested that patients' perceptual difficulty could be attributed to the nasality grouping, normally well separated for normal hearing listeners, shifting close to the glottals and aspirated obstruents. Hearing aids seemed to improve perception of all phonemes by 10%, with /tɕʰ/, /kʰ/, /s/, and /h/ (call unvoiced) showing significant improvement rate. Lastly, the signal detection theory (SDT) bias values of *c* among all possible 108 pairs of unvoiced vs. voiced phonemes revealed that normal hearing subjects are in favour of unvoiced phonemes. The hearing loss patients (with and without hearing aids) showed the same bias pattern. Interestingly, the hearing aids seem to substantially increase more biases for the unvoiced category.

Keywords : Thai diagnostic rhyme test, subjective intelligibility, confusion matrix, similarity score, perceptual representation space, *c*-value

Executive summary

It is important to note that this work is among the very first papers that investigates and provides a detailed analysis of confusion patterns of Thai initial and final phonemes. Specifically, a subjective intelligibility testing of Thai speech for initial and final consonants, based on the diagnostic rhyme test (DRT), was developed. Using real monosyllabic words, the Thai DRT for initials (TDRT-I) and for finals (TDRT-F) were designed to equally compare 21 phonemes and 8 phonemes, respectively. Its strength lies in the fact that the percent intelligibility scores as well as confusion patterns across all 21 phonemes can be evaluated and compared.

To test the validity of our approach, the subjective intelligibility tests were conducted on Thai normal hearing listeners (Group I) in four SNR levels (for TDRT-I and TDRT-F) and on hearing loss patients (with and without hearing aids for TDRT-F) (Group II) using clean stimuli. Comparisons of confusion patterns in both groups showed that for both initial and final consonants, voicing was the most robust contrast while place-of-articulation was the least. Specifically, for initials, /r/ is the most confusable phoneme, while /w/, /j/, and /p/ are among the least. Perceptual representation of Group I, derived from confusion matrices, yielded five nonoverlapping groupings: glide, glottal constriction, nasality, aspirated obstruent, and a combination of liquid and unaspirated obstruent. Interestingly, perceptual representation of Group II suggested that their perceptual difficulty could be attributed to the nasality grouping shifted closer to the glottals and aspirated obstruents. Hearing aids seemed to improve perception of all phonemes by 10%, with only /tɕʰ/, /kʰ/, /s/, and /h/ (all unvoiced) showing significant improvement. The signal detection theory (SDT) bias values of c among all possible 108 pairs of unvoiced vs. voiced phonemes further revealed that Group I were in favor of unvoiced phonemes. Likewise, Group II showed the same bias pattern. Interestingly, the hearing aids seemed to substantially increase more biases for the unvoiced category.

In addition, our paper offers insightful cross-linguistic observations. Perceptual similarity and distance scores were computed to yield perceptual representations of Thai and English phonemes (Miller and Nicely, 1955). Generally, English phonemes can be divided into three clusters while Thai has five clusters. Nasal sounds are grouped together nicely in both languages. Voicing (voiced and unvoiced), one of the most distinct perceptual properties in English, appears to be a less robust feature in Thai, where aspiration plays a more significant role. It is interesting that English obstruents 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.

1. Introduction

This paper describes in details a series of experiments starting with the development of the Thai diagnostic rhyme test for initials (TDRT-I) and finals (TDRT-F) (Tantibundhit et al., 2011c). Two experiments, using TDRT-I and TDRT-F were conducted on normal-hearing listeners (Tantibundhit et al., 2011b) and hearing-loss patients (for TDRT-I) (Tantibundhit et al., 2011a). Experimental results (including confusion matrices) were partially given in Tantibundhit et al. (2011a,b,c) but here are presented in full. Moreover, derived perceptual representations are compared and discussed in detail. Lastly, the analysis of signal detection theory (SDT) values of c (criterion) is further examined for both sets of data. In this section, previous and relevant work, related to subjective intelligibility testing, analyses of perceptual confusions, is reviewed.

1.1 Subjective Intelligibility Testing

Speech intelligibility and speech quality are two distinct properties. Speech quality is subjective in nature and difficult to reliably evaluate. Specifically, it reflects how an utterance is produced and also includes speech attributes such as natural, raspy, hoarse, etc (Loizou, 2013). Speech intelligibility, on the other hand, refers to what is being said, i.e., the meaning or the content of the spoken words (Loizou, 2013). Therefore, speech intelligibility is one of the essential attributes of the speech signal and needs to be preserved by speech enhancement algorithms (Tantibundhit et al., 2007, 2010). Several algorithms have been developed specifically to enhance speech intelligibility in background noise (Tantibundhit et al., 2007, 2010). Evaluating intelligibility of the enhanced compared with the original speech is often conducted using a subjective intelligibility testing (Loizou, 2013). Several intelligibility tests have been proposed for English by using rhyming words presented in six-response (House et al., 1965) or in pair-response (Voiers, 1983).

House et al. (1965) developed a test by restricting response choices to a finite set of six rhyming words called the modified rhyme test (MRT). The test was composed of 50 sets, each of which was composed of six monosyllabic consonant-vowel-consonant (CVC) words. Twenty-five sets differed in their initial consonants, e.g., led, shed, red, bed, fed, wed, while the rest differed in their final consonants, e.g., bat, bad, back, bass, ban, bath (House et al., 1965).

Voiers (1983) refined the MRT and created a diagnostic rhyme test (DRT), which is widely used for a subjective testing for measuring the intelligibility of speech coders (Loizou, 2013). The DRT was an A/B forced comparison test based on word pairs differing in their initial consonants by one of six distinctive features (Voiers, 1983). The DRT test material was composed of a word list of 96 rhyming pairs, e.g., veal - feel. As the DRT was developed specifically for English, it has some limitations when evaluating intelligibility of a tonal language such as Chinese (McLoughlin, 2008).

McLoughlin (2008) developed a New Chinese diagnostic rhyme test (NCDRT). The NCDRT was composed of a test set of phonemes in Chinese, which were classified under six distinctive features similar to the DRT (McLoughlin, 2008). Although the subjective intelligibility testing of a tonal language such as Chinese is well underway (McLoughlin, 2008), subjective intelligibility testing of another tonal language, Thai, with several acoustic and phonemic differences from that of Chinese (Comrie, 1990) has yet to be developed.

In our previous work, we designed and developed an intelligibility testing of Thai speech specifically for its initial (TDRT-I) and final consonants (Tantibundhit et al., 2011c). The test was designed to facilitate an evaluation of percent intelligibility responses in each stimulus pair and to systematically compare confusion responses across all initial and final phonemes (Tantibundhit et al., 2011c). Specifically, several useful frameworks, namely DRT (Voiers, 1983), NCDRT (McLoughlin, 2008), MRT (House et al., 1965), and the analysis method of balanced confusion matrix (Miller and Nicely, 1955) were integrated. Moreover, an A/B forced choice and monosyllabic (CV(V)(C)) rhyming pairs, which differ only in one sound either in an initial or final position (the tone was kept identical) were used. The words were well selected from real and commonly used words in the language (Tantibundhit et al., 2011c).

1.2. Analyses of Perceptual Confusions

Analyses of perceptual confusions among phonemes (speech sounds) provide valuable information in determining and understanding speech perception in general and cross-linguistically (Johnson, 2003). 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 (Stevens, 1981). This insight is crucial for a number of areas in speech research, including speech recognition (Mermelstein, 1976). 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 listener's language experience (Strange, 1995). 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 (Stevens, 1981).

A number of studies have focused on confusion analyses of English consonants, e.g., Miller and Nicely (1955). Among them, a classic report from Miller and Nicely (1955), where perception of English word-initial consonants (16 phonemes) in an open-response task was conducted under

different bandwidths (in between 200-6,500 Hz) and different signal to noise ratios (SNRs) (-18, -12, -6, 0, +6, and +12 dB).

Shepard (1972) proposed a method to assess a psychological representation of speech sounds by computing similarity and distance scores from confusion matrices. He applied his formula and method to the English perceptual data from Miller and Nicely (1955). The analysis showed that the perceptual representation of English consonants could be grouped according to two phonological dimensions (adapted from Jacobson et al. (1952)), that of nasality and a combination of voicing and frication, suggesting that nasality, voicing, and frication are the strongest perceptual features for English consonants (Shepard, 1972).

Benkí (2003) examined 10 English word-initial phonemes using four degrees of SNR (-14, -11, -8, and -5 dB) in an open-response task. His investigation was expanded to include confusion matrices of 10 English final phonemes and nine vowels. His findings confirmed that voicing feature is stronger than feature for place of articulation and that initial consonants are more distinct than finals (Benkí, 2003).

Cross-linguistically, Singh and Black (1966) explored speaker-listener errors of phonemic and non-phonemic intervocalic sounds of four languages, i.e., Arabic, English, Hindi, and Japanese. 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 (Tingsabadh and Abramson, 1993), resembles English, with 24 phonemes, 22 of which (except /ŋ/ and /ʒ/) occur word-initially (Giegerich, 1992)). However, the two languages differ phonologically in many aspects. For instance, Thai has a 3-way stop/affricate distinction (voiced, unvoiced unaspirated, and aspirated), while there exists a 2-way distinction (voiced and unvoiced) in English. Moreover, English has 11 fricatives/affricates, whereas Thai has only four (Comrie, 1990). Therefore, 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, i.e., either for normal hearing listeners or for sensorineural hearing loss patients. As a result, there remains a large gap for the knowledge of perceptual representation in general, and as for whether or to what extent hearing loss affects this perceptual property. Our previous studies (Tantibundhit et al., 2011a,b) are hoped to fill this remaining gap. Our effective method of Thai diagnostic rhyme test for initials (TDRT-I) (Tantibundhit et al., 2011c) enabled us to systematically collect listener's responses of initial phoneme identification in different noise conditions and constructed a valid approximation of the perceptual representation (Tantibundhit et al., 2011b). Along with an investigation in the hearing-loss population (Tantibundhit et al., 2011a), our goal is to provide

some insights into the abstract yet consequential representations in the case of Thai speech sounds both for normal hearing listeners and for sensorineural hearing loss patients.

1.3. Organization of the Report

In this report, we combine and explain in more detail our previous studies, i.e., Tantibundhit et al. (2011a,b,c, 2012). In the following sections, Thai phonology is reviewed in Section 2. Design and development of TDRT-I and TDRT-F, subjective intelligibility tests, experimental results regarding percent intelligibility responses in each stimulus pair, perceptual confusions across all phonemes, and analysis of confusion matrices are presented in Section 3. In Section 4, similarity scores between each pair of phonemes and perceptual distances calculated from confusion scores (Shepard, 1972) are explained. Distances and perceptual spaces of Thai phonemes for normal hearing subjects and for sensorineural hearing loss patients are also given in this section. Section 5 presents the SDT bias values of c, which are used to highlight and quantify confusion asymmetries that exist in certain initial phoneme pairs (Benkí, 2003). After confusion matrices for 21 initial phonemes are constructed, the bias measure c (Macmillan and Creelman, 2005) is calculated. Investigation of the SDT bias values of c in initial phoneme pairs specifically for any combination between an unvoiced and voiced phonemes is carried out for normal hearing listeners and for sensorineural hearing loss patients. Finally, we discuss the results, implications, and future work in Section 6.

2. Thai Phonology Review

Thai is a tonal language with 21 consonantal phonemes in initial position /p/, /ph/, /b/, /t/, /th/, /d/, /tɕ/, /tɕ^{h}/, /k/, /kh/, /ʔ/, /f/, /s/, /h/, /m/, /n/, /ŋ/, /l/, /r/, /w/, and /j/ and 8 consonantal phonemes in final position /k/, /t/, /p/, /ŋ/, /n/, /m/, /j/, and /w/ (Tingsabadh and Abramson, 1993). Each of the nine monophthongs in Thai occurs phonemically short or long (/i/, /ii/, /e/, /ee/, /ɛ/, /ɛɛ/, /ɯ/, /ɯɯ/, /ɤ/, /ɤɤ/, /a/, /aa/, /u/, /uu/, /o/, /oo/, /ɔ/, and /ɔɔ/) (Comrie, 1990; Tingsabadh and Abramson, 1993). 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 ˋ, 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 (Comrie, 1990; Tingsabadh and Abramson, 1993).

3. TDRT-I and TDRT-F Designs, Intelligibility Tests (Experiments 1 and 2), Experimental Results, and Confusion Matrices

3.1. TDRT-I and TDRT-F Design

The goal of TDRT-I and TDRT-F designs were to come up with a subjective intelligibility test set specifically for Thai initial and final consonants, keeping in mind that not only should the tests allow us to evaluate percent intelligibility responses in each stimulus pair, but to systematically compare confusion responses across all phonemes. In addition, the test should not be too long to cause fatigue Loizou (2013). To do so, monosyllabic rhyming word pairs differing only in one sound in an initial or final position are constructed and selected as follows:

3.1.1 TDRT-I

1) Multiple sets of Thai monosyllabic ($C_iV^T(V)(C_f)$) words, each of which differs only in their initial phoneme are pooled.
2) Vowel /aa/ along with mid tone are chosen because it is one of the most frequently used vowels (Kosawat et al., 2009) and when combined with mid tone yields the most possible number of rhyming words, i.e., 21 rhyming words for 21 phonemes: /pāa/ ปา, /p^hāa / พา, /bāa / บา, /tāa / ตา, /t^hāa / ทา, /dāa / ดา, /tɕāa/ จา, /tɕ^{h}āa/ ชา, /kāa/ กา, /k^hāa/ คา, /pāa/ อา, /fāa/ ฟา, /sāa/ ซา, /hāa/ ฮา, /māa/ มา, /nāa/ นา, /ŋāa/ งา, /lāa/ ลา, /rāa/ รา, /wāa/ วา, and /jāa/ ยา.
3) Each rhyming word is paired with 20 others of different initial phoneme. This results in a total combination of 210 stimulus pairs of rhyming words, which can be expressed mathematically as a combination of 21 choose 2 ($^{21}C_2$).

Complete list of rhyming words for initial consonants with their translation is shown in Table 1.

Table 1: *A set of 21 rhyming words differing in their initial consonants across 21 phonemes (Tantibundhit et al., 2011c).*

no.	transcription	Thai script	translation
1.	/pāa/	ปา	throw
2.	/p^{h}āa/	พา	bring
3.	/bāa/	บา	teacher
4.	/tāa/	ตา	eye
5.	/t^{h}āa/	ทา	paint
6.	/dāa/	ดา	advance along a wide front
7.	/t͡ɕāa/	จา	talk
8.	/t͡ɕ^{h}āa/	ชา	tea
9.	/kāa/	กา	crow
10.	/k^{h}āa/	คา	stick
11.	/ʔāa/	อา	uncle
12.	/fāa/	ฟา	F musical note
13.	/sāa/	ซา	lessen
14.	/hāa/	ฮา	laugh
15.	/māa/	มา	arrive
16.	/nāa/	นา	field
17.	/ŋāa/	งา	ivory
18.	/lāa/	ลา	donkey
19.	/rāa/	รา	fungus
20.	/wāa/	วา	2 meters (Thai unit)
21.	/jāa/	ยา	medicine

3.1.2 TDRT-F

1) Pairs of monosyllabic ($C_iV^T(V)C_f$) words, each of which differs only in their final consonant phoneme (the tone in each pair remains identical) are garnered.
2) Two types of initial consonants C_i are chosen to create the rhyming words, namely voiceless unaspirated plosives (/p/, /t/, and /k/) and voiceless aspirated plosives (/p^{h}/, /t^{h}/, and /k^{h}/). The initial plosives are chosen over other types of initial consonant as they can be combined with the most possible types of rime unit (the sequence of vowel and final consonant).
3) Six initial plosives are subsequently combined with all 18 vowels: 9 short and 9 long vowels and with all 5 tones (6×18×5=540). For example, initial consonant /t/ when combined with a vowel /a/, a low tone ◌̀, and 8 different final phonemes will produce /tàk/ ตัก, /tàt/ ตัด, /tàp/ ตับ, /tàŋ/ ตั้ง, /tàn/ ตั่น, /tàm/ ต่ำ, /tàj/ ไต่, and /tàw/ เต่า. Altogether, 540 possible words are created.

4) Out of the 540 words, only 84 pairs of real words (84 stimulus pairs) that are commonly used are selected. These stimulus pairs comprise 3 instances of each rhyming word paired with 7 others of different final phonemes, which can be expressed mathematically as *three times a combination of 8 choose 2 ($3\times{}^{8}C_{2}$)*.

Complete list of rhyming words for final consonants with their translation is shown in Table A-I and A-II.

Table A-I: A set of 84 rhyming words differing in their final consonants across 8 final phonemes (Tantibundhit et al., 2011c).

pair no.	transcription	Thai script	translation	-	transcription	Thai script	translation
1.	/tàp/	ตับ	liver	-	/tàt/	ตัด	cut
2.	/kòp/	กบ	frog	-	/kòt/	กด	press
3.	/pɔ̀ɔp/	ปอบ	ogre	-	/pɔ̀ɔt/	ปอด	lung
4.	/tòp/	ตบ	slap	-	/tòk/	ตก	fall
5.	/kàp/	กับ	and	-	/kàk/	กัก	confine
6.	/tɔ̀ɔp/	ตอบ	answer	-	/tɔ̀ɔk/	ตอก	hammer
7.	/t^{h}úp/	ทุบ	pond	-	/t^{h}úm/	ทุ้ม	bass
8.	/k^{h}âap/	คาบ	hold in the mouth	-	/k^{h}âam/	ข้าม	skip
9.	/k^{h}áp/	คับ	tight	-	/k^{h}ám/	ค้ำ	prop up
10.	/p^{h}óp/	พบ	meet	-	/p^{h}ón/	พ้น	pass
11.	/k^{h}óp/	คบ	associate	-	/k^{h}ón/	ค้น	seek
12.	/k^{h}áp/	คับ	tight	-	/k^{h}án/	คั้น	squeeze
13.	/kèp/	เก็บ	keep	-	/kèŋ/	เก่ง	excellently
14.	/k^{h}âap/	คาบ	hold in the mouth	-	/k^{h}âaŋ/	ข้าง	side
15.	/tàp/	ตับ	liver	-	/tàŋ/	ตั่ง	stool
16.	/tàp/	ตับ	liver	-	/tàw/	เต่า	turtle
17.	/t^{h}áp/	ทับ	overlay	-	/t^{h}áw/	เท้า	foot
18.	/k^{h}áp/	คับ	tight	-	/k^{h}áw/	เค้า	outline
19.	/p^{h}âap/	ภาพ	picture	-	/p^{h}âaj/	พ่าย	lose
20.	/kàp/	กับ	and	-	/kàj/	ไก่	chicken
21.	/k^{h}âap/	คาบ	hold in the mouth	-	/k^{h}âaj/	ค่าย	camp
22.	/pàt/	ปัด	sweep	-	/pàk/	ปัก	stab down
23.	/pàat/	ปาด	slice off	-	/pàak/	ปาก	mouth
24.	/tàt/	ตัด	cut	-	/tàk/	ตัก	scoop
25.	/k^{h}òt/	ขด	coil	-	/k^{h}òm/	ข่ม	oppress
26.	/k^{h}át/	คัด	select	-	/k^{h}ám/	ค้ำ	prop up
27.	/k^{h}út/	คุด	curl	-	/k^{h}úm/	คุ้ม	protect
28.	/kɔ̀ɔt/	กอด	hug	-	/kɔ̀ɔn/	ก่อน	before
29.	/k^{h}ùt/	ขุด	dig	-	/k^{h}ùn/	ขุ่น	be turbid
30.	/t^{h}àat/	ถาด	tray	-	/t^{h}àan/	ถ่าน	charcoal
31.	/pèt/	เป็ด	duck	-	/pèŋ/	เป่ง	be ripe
32.	/tìt/	ติด	close	-	/tìŋ/	ติ่ง	protrusion
33.	/t^{h}ít/	ทิด	man who resumes secular life	-	/t^{h}íŋ/	ทิ้ง	discard
34.	/pàt/	ปัด	sweep	-	/pàw/	เป่า	blow
35.	/k^{h}àt/	ขัด	rub	-	/k^{h}àw/	เข่า	knee
36.	/k^{h}ít/	คิด	think	-	/k^{h}íw/	คิ้ว	eyebrow
37.	/k^{h}âat/	คาด	anticipate	-	/k^{h}âaj/	ค่าย	camp
38.	/kàt/	กัด	bite	-	/kàj/	ไก่	chicken
39.	/t^{h}àat/	ถาด	tray	-	/t^{h}àaj/	ถ่าย	take a picture
40.	/t^{h}úk/	ทุกข์	suffering	-	/t^{h}úm/	ทุ้ม	bass
41.	/k^{h}úk/	คุก	prison	-	/k^{h}úm/	คุ้ม	protect
42.	/t^{h}òk/	ถก	discuss	-	/t^{h}òm/	ถ่ม	spit

Table A-II: A set of 84 rhyming words differing in their final consonants across 8 final phonemes (Tantibundhit et al., 2011c).

pair no.	transcription	Thai script	translation	-	transcription	Thai script	translation
43.	/pʰók/	พก	carry	-	/pʰón/	พ้น	pass
44.	/kʰôok/	โคก	mound	-	/kʰôon/	โค่น	fall
45.	/pòk/	ปก	cover	-	/pòn/	ป่น	powdered
46.	/tàak/	ตาก	air	-	/tàaŋ/	ต่าง	differ
47.	/kʰòok/	โขก	knock	-	/kʰòoŋ/	โข่ง	Pila (gastropod)
48.	/kʰàak/	ขาก	spit	-	/kʰàaŋ/	ข่าง	spinning top
49.	/pʰàk/	ผัก	vegetable	-	/pʰàw/	เผ่า	tribe
50.	/pʰàak/	ผาก	parched	-	/pʰàaw/	ผ่าว	scorching
51.	/pàk/	ป๊ก	stick	-	/pàw/	เป่า	blow
52.	/kúk/	กุ๊ก	cook	-	/kúj/	กุ๊ย	thug
53.	/tàak/	ตาก	air	-	/tàaj/	ต่าย	rabbit
54.	/kàak/	กาก	garbage	-	/kàaj/	ก่าย	rest on
55.	/tùm/	ตุ่ม	pimple	-	/tùn/	ตุ่น	mole
56.	/kâam/	ก้าม	claw	-	/kâan/	ก้าน	stem
57.	/tām/	ตำ	pound	-	/tān/	ตัน	clog
58.	/tɔ̄ɔm/	ตอม	swarm	-	/tɔ̄ɔŋ/	ตอง	banana leaf
59.	/pôm/	ป้อม	fortress	-	/pôŋ/	ป้อง	cover up
60.	/tēm/	เต็ม	full	-	/tēŋ/	เต็ง	favorite
61.	/kām/	กำ	grasp	-	/kāw/	เกา	scratch
62.	/tām/	ตำ	pound	-	/tāw/	เตา	stove
63.	/kâam/	ก้าม	claw	-	/kâaw/	ก้าว	step
64.	/tɤ̄ɤm/	เติม	add	-	/tɤ̄ɤj/	เตย	screw pine
65.	/tāam/	ตาม	follow	-	/tāaj/	ตาย	die
66.	/pāam/	ปาล์ม	palm	-	/pāaj/	ปาย	Pai district
67.	/kɛ̀n/	แก่น	core	-	/kɛ̀ŋ/	แก่ง	islet
68.	/kōon/	โกน	shave	-	/kōoŋ/	โกง	cheat
69.	/tɛ̄ɛn/	แตน	wasp	-	/tɛ̄ɛŋ/	แตง	melon
70.	/kān/	กัน	keep out	-	/kāw/	เกา	scratch
71.	/pân/	ปั้น	mold	-	/pâw/	เป้า	target
72.	/tān/	ตัน	clog	-	/tāw/	เตา	stove
73.	/pɔ̄ɔn/	ปอน	sloppy	-	/pɔ̄ɔj/	ปอย	tuft
74.	/tùn/	ตุ่น	mole	-	/tùj/	ตุ่ย	puffy
75.	/pâan/	ป้าน	obtuse	-	/pâaj/	ป้าย	plate
76.	/pʰêŋ/	แพ่ง	civil	-	/pʰêw/	แผ้ว	clear
77.	/tī̂ŋ/	ติ๋ง	admonish	-	/tī̂w/	ติ๋ว	cram for an examination
78.	/tâŋ/	ตั้ง	establish	-	/tâw/	เต้า	breast
79.	/kôŋ/	ก้อง	echo	-	/kôj/	ก้อย	little finger
80.	/tàaŋ/	ต่าง	differ	-	/tàaj/	ต่าย	rabbit
81.	/kōoŋ/	โกง	cheat	-	/kōoj/	โกย	shovel
82.	/kàw/	เก่า	old	-	/kàj/	ไก่	chicken
83.	/kʰâaw/	ข้าว	rice	-	/kʰâaj/	ค่าย	camp
84.	/kʰāaw/	คาว	fishy	-	/kʰāaj/	คาย	spit out

3.2. Intelligibility Tests for Normal Hearing (Experiment 1) and Sensorineural Hearing Loss Subjects (Experiment 2)

In Experiment 1, the subjective intelligibility tests for initial consonants were conducted in four conditions of additive white Gaussian noise (AWG) on normal hearing listeners (Tantibundhit et al., 2011c). In Experiment 2, the tests were carried out in clean condition only on sensorineural hearing loss subjects (with and without hearing aids (Tantibundhit et al., 2011a)). To create test stimuli, all 21 initial rhyming words along with filler words were read five times in a carrier sentence (ฉันชอบ … อีกแล้ว /tɕʰǎn tɕʰɔ̂ɔp … ʔiik lɛ̀ɛw/) and recorded at a sampling rate of 44.1 kHz in a sound-attenuated chamber by a 36-year-old Thai male speaker who was born and grew up in Bangkok. Then, each target word stimulus was excised from the carrier sentence. To avoid audible discontinuity problems at the splice points, the starting point of each stimulus began approximately 10 msec prior to the onset of initial consonant and its end point included some durational adjustments to the last sound segment at a precise location. Every splice was done at a zero crossing.

Then, one of the five tokens of each target word that was the clearest, most typical, and most natural sounding was selected based on impressionistic hearing evaluation and spectrographic inspection. Average duration of the stimuli was 324.4 msec.

For normal hearing listeners, four signal-to-noise ratios (SNR) of -6, -12, -18, and -24 dB were chosen based on our preliminary findings such that intelligibility scores are in a range to avoid floor and ceiling effects, i.e., much higher than 50% but not approaching 100% (subjects so well perceived stimuli) or close to 50% (the scores are indistinguishable from guesswork) (Loizou, 2013). Experiment 1 (presented in four SNR conditions) was performed individually on untrained 28 normal hearing subjects over headphones in a quiet room, while Experiment 2 (presented in clean condition) was performed individually on untrained eight sensorineural hearing loss patients with and without hearing aids over speakers in a sound booth at Thammasat University hospital. The patients were recruited by an otolaryngologist at Thammasat University hospital. Table 3 shows background information of these eight participants. In each trial, listeners heard a target stimulus and were asked to choose what they just heard between two rhyming words, appearing on the computer screen. If they did not recognize the stimulus, they were instructed to guess before moving on to the next trial. Sequence of individual trials as well as sequence of word in each A/B pair for intelligibility tests for initial consonants were randomized in real tests.

For normal hearing subjects (Experiment 1), 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 (Loizou, 2013). Alternatively, by increasing a number of subjects four times, we could stay with the 500 trials and divide the test equally by four SNR levels. Consequently, the 500 trials

Table 2: *Distributions of rhyming word groupings for initial consonants for normal hearing listeners (Experiment 1) (modified from Table 1 of Tantibundhit et al. (2011c)).*

	Rhyming and Filler Word			
Subject	Group A	Group B	Group C	Group D
I	−6 dB	−12 dB	−18 dB	−24 dB
II	−24 dB	−6 dB	−12 dB	−18 dB
III	−18 dB	−24 dB	−6 dB	−12 dB
IV	−12 dB	−18 dB	−24 dB	−6 dB

Table 3: *Background information of eight hearing impaired adults (with moderate to moderately severe degrees of hearing loss).*

Subject	Age (years)	Sex	Average hearing loss		Air		SL/PB	
			R	L	R	L	R	L
1	62	F	44.17	55	47	55	30/96%	30/92%
2	73	F	80	66.67	72	55	26/84%	35/88%
3	19	M	108.33	48.57	110	35	N.A.	30/92%
4	52	F	90	68.57	97	65	5/84%	35/92%
5	66	F	71.43	81.43	63	67	25/92%	25/88%
6	79	F	63.33	62.50	60	58	65/88%	65/92%
7	62	F	88.33	63.33	88	68	10/88%	25/100%
8	61	F	70	55	35	32	20/92%	30/96%

are corrupted by one of four SNR levels of AWG noise stated earlier, i.e., Groups A, B, C, and D, each of which has an SNR level of -6 dB, -12 dB, -18 dB, and -24 dB, respectively as summarized in Table 2. With regard to distributions of the rhyming words, subjects' performance per SNR level was equally distributed yielding 105 trials/SNR level (420 trials/4 SNR levels). Each of the 105 trials was 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 were randomized in the test. It should be noted that 28 listeners are equivalent to seven complete subjects.

For sensorineural hearing loss adults (Experiment 2), the test 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 was presented once as a stimulus in a trial, resulting in a total of 420 trials for initial consonants and 80 trials for filler words. The test was presented twice to the patients, first where they removed the hearing aids and later where the hearing aids were kept on.

Table 4: *Average percent intelligibility for normal hearing subjects.*

Consonant	SNR (dB)			
	−6 dB	−12 dB	−18 dB	−24 dB
Initial	93.1%	87.1%	77.4%	24.1%

Table 5: *Average percent intelligibility for sensorineural hearing loss subjects.*

Consonant	without hearing aids	with hearing aids
Initial	49.9%	70.4%

Table B: Average percent intelligibility for final consonants for normal hearing subjects.

Consonant	SNR (dB)			
	-6dB	-12dB	-18dB	-24dB
Final	91.67%	84.01%	67.35%	27.21%

3.3. Experimental Results and Confusion Matrices

Table 4 shows percent intelligibility scores for initial consonants for normal hearing listeners across 4 SNR levels. It should be pointed out that the average percent correct response, which does not necessarily match the intelligibility score, is calculated from total number of correct responses divided by total number of stimuli. Percent intelligibility scores are calculated from

$$P_e = \frac{N_r - N_w}{T} \times 100\%, \tag{1}$$

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 (Voiers, 1983).

It is clear that percent intelligibility scores were decreasing as increasing level of noise. Furthermore, the results showed that subject's performance at SNR level of at -24 dB was far below a guesswork (50%) and could be excluded from analysis of confusions.

Balanced confusion matrices at all SNR levels for normal hearing listeners were obtained from the test responses and shown in Tables 6—9. It should be noted that confusion matrices are constructed from correct responses, not intelligibility scores. The results showed that across SNR levels of -6, -12, and -18 dB, /r/ was the most confusable initial consonant and it was mostly misperceived as /d/, /t/, /tɕ/, /b/, and /k/. On the other hand, /w/, /j/, and /p/ were among the least confusable initial consonants.

A separate analysis across the three levels for listeners' misidentified responses showed that /t/ and /t^h/ were the most favored while /w/ and /r/ the least. Focusing at the -18 dB level in which

the intelligibility score is neither too high nor too low, voicing was the most robust contrast while place-of-articulation was the least. In addition, at this SNR level, /r/ was the most confusable phoneme followed by /t^h/, while /j/ was the least confusable phoneme, followed by /w/. At -18 dB, a separate analysis for listeners' misidentified responses revealed that the listeners favored /t/, /p/, /t^h/, /$tɕ^h$/, and /tɕ/ and disfavored /w/ and /r/ over other phonemes.

Table 5 shows percent intelligibility scores for sensorineural hearing loss subjects, i.e., 49.9% without hearing aids and 70.4% with hearing aids. Overall, in terms of intelligibility, hearing aids seemed to improve hearing performance by 20.5% (10.3% in terms of percent correct response). In addition, balanced confusion matrices for sensorineural hearing loss subjects without hearing aids and with hearing aids are shown in Tables 10 and 11, respectively. Interestingly, for both cases, /r/ was the most confusable phoneme (in line with the normal hearing subjects), followed by /k^h/ for the subjects without hearing aids and followed by /ŋ/ with hearing aids. In addition, /r/ was mostly misperceived as /b/, /tɕ/, and /$tɕ^h$/ without hearing aids while mostly misperceived as /p/, /d/, and /k/ for the subjects with hearing aids. Interestingly, /p/ was the least confusable phoneme for both cases. A separate analysis for listeners' misidentified responses revealed that the subjects with hearing aids favored /p/ and /tɕ/ and disfavored /w/ and /r/, while the subjects without hearing aids favored /t/, /p/, and /s/ and disfavored /j/ and /h/.

Paired t-test difference between subject's performance with and without hearing aids of 21 initial phonemes showed that hearing aids can significantly improve speech perception of four phonemes, i.e., /$tɕ^h$/ [$t(7) = 2.8924$, $p = 0.023$], /k^h/ [$t(7) = 4.0566$, $p = 0.005$], /s/ [$t(7) = 2.5252$, $p = 0.040$], and /h/ [$t(7) = 2.4279$, $p = 0.046$].

Table B shows percent intelligibility scores for final consonants for normal hearing subjects. Comparing with Table 4, the results show that the initial consonants were perceived better than the final consonants except at the SNR level of -24dB. Confusion matrices for final consonants across SNR levels of -6, -12, -18, and -24 dB are shown in Table C—F, respectively. For the results, /k/ is the most confusable consonant and it was mostly misperceived as /t/, which is also a voiceless non-continuant. Interestingly, at the -18dB level, for both initial and final consonants, voicing was the most robust contrast while place-of-articulation was the least.

Table 6: *Confusion matrix for normal hearing listeners at SNR = −6 dB (Experiment 1).*

Stimulus	Response																				Total
	/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/	140	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140
/p^h/	0	137	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	140
/b/	1	0	139	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140
/t/	0	0	0	137	0	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	140
/t^h/	0	3	0	0	136	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	140
/d/	0	0	1	0	0	139	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140
/tɕ/	0	0	0	0	0	0	139	1	0	0	0	0	0	0	0	0	0	0	0	0	0	140
/tɕ^h/	0	0	0	0	0	0	0	139	0	0	0	0	1	0	0	0	0	0	0	0	0	140
/k/	0	0	0	0	0	0	0	0	140	0	0	0	0	0	0	0	0	0	0	0	0	140
/k^h/	0	3	0	0	6	0	0	0	0	130	0	0	0	1	0	0	0	0	0	0	0	140
/ʔ/	0	0	0	0	0	0	0	1	0	0	137	0	0	0	0	1	1	0	0	0	0	140
/f/	1	0	0	0	0	0	0	0	0	0	0	139	0	0	0	0	0	0	0	0	0	140
/s/	3	0	0	4	0	2	1	1	0	1	0	1	126	0	0	0	0	0	0	0	1	140
/h/	0	1	0	0	2	0	0	0	0	0	0	0	0	137	0	0	0	0	0	0	0	140
/m/	0	0	0	1	0	0	0	0	0	0	0	0	0	0	139	0	0	0	0	0	0	140
/n/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	135	5	0	0	0	0	140
/ŋ/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	139	0	0	0	0	140
/l/	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	135	3	0	0	140
/r/	3	1	2	4	1	5	3	2	3	2	1	1	3	1	2	3	3	1	99	0	0	140
/w/	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	139	0	140
/j/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	137	140
Misidentified response	9	8	3	9	11	7	4	7	4	5	1	4	4	2	2	5	9	3	3	1	1	

Table 7: *Confusion matrix for normal hearing listeners at SNR = −12 dB (Experiment 1).*

Stimulus	Response																					Total
	/p/	/p^h/	/b/	/t/	/t^h/	/d/	/ʨ/	/ʨ^h/	/k/	/k^h/	/ʔ/	/f/	/s/	/h/	/m/	/n/	/ŋ/	/l/	/r/	/w/	/j/	
/p/	138	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	140
/p^h/	0	132	0	0	1	0	0	1	0	2	0	0	0	2	0	0	0	0	0	2	0	140
/b/	0	0	131	0	2	1	0	0	0	0	0	1	0	0	0	0	0	1	2	1	1	140
/t/	1	0	0	126	3	0	1	2	0	0	0	3	4	0	0	0	0	0	0	0	0	140
/t^h/	0	4	0	0	125	0	1	1	0	3	2	0	1	3	0	0	0	0	0	0	0	140
/d/	0	0	0	0	0	140	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140
/ʨ/	0	0	0	0	1	0	137	1	0	0	0	0	1	0	0	0	0	0	0	0	0	140
/ʨ^h/	1	3	0	3	2	0	3	115	3	5	0	2	3	0	0	0	0	0	0	0	0	140
/k/	0	0	0	0	0	0	0	0	139	0	0	0	1	0	0	0	0	0	0	0	0	140
/k^h/	0	1	0	2	4	0	0	3	0	123	2	0	0	5	0	0	0	0	0	0	0	140
/ʔ/	0	1	0	0	0	0	0	0	0	0	130	0	0	3	1	3	2	0	0	0	0	140
/f/	0	0	1	0	0	0	0	0	0	0	0	138	1	0	0	0	0	0	0	0	0	140
/s/	3	0	0	5	0	2	2	0	1	0	0	3	124	0	0	0	0	0	0	0	0	140
/h/	0	0	0	0	0	0	0	0	0	0	2	0	0	134	1	1	1	1	0	0	0	140
/m/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140	0	0	0	0	0	0	140
/n/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140	0	0	0	0	0	140
/ŋ/	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	137	0	0	0	0	140
/l/	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	136	2	0	0	140
/r/	4	3	5	6	2	7	7	4	4	1	0	0	3	0	0	1	0	1	87	1	4	140
/w/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140	0	140
/j/	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	139	140
Misidentified response	11	13	6	17	15	10	14	12	8	11	6	10	14	13	3	7	3	3	4	4	5	

Table 8: *Confusion matrix for normal hearing listeners at SNR = −18 dB (Experiment 1).*

Stimulus	Response																					Total
	/p/	/p^h/	/b/	/t/	/t^h/	/d/	/ʧ/	/ʧ^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	140
/p^h/	1	121	0	1	6	0	1	4	0	2	1	0	1	2	0	0	0	0	0	0	0	140
/b/	0	0	137	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	140
/t/	0	3	2	116	2	0	4	1	3	0	0	5	3	0	0	0	0	0	1	0	0	140
/t^h/	4	3	1	2	112	1	3	4	2	2	0	2	3	1	0	0	0	0	0	0	0	140
/d/	1	0	0	0	0	132	3	0	4	0	0	0	0	0	0	0	0	0	0	0	0	140
/ʧ/	1	0	0	4	1	1	115	5	1	0	0	3	5	0	0	1	0	1	0	1	1	140
/ʧ^h/	2	2	0	1	2	1	1	126	0	2	1	0	1	0	0	0	0	1	0	0	0	140
/k/	1	0	0	3	0	0	2	0	130	0	0	0	1	0	0	0	1	1	0	0	1	140
/k^h/	0	3	1	2	6	0	1	2	1	117	5	0	0	2	0	0	0	0	0	0	0	140
/ʔ/	0	1	0	0	1	0	0	0	1	1	128	0	0	4	1	1	2	0	0	0	0	140
/f/	5	0	1	2	1	2	0	0	0	0	0	126	1	0	0	0	0	1	0	1	0	140
/s/	5	0	1	4	1	1	1	1	1	1	1	5	116	0	0	0	0	1	1	0	0	140
/h/	1	4	0	0	2	0	0	0	0	2	4	0	0	123	1	0	2	1	0	0	0	140
/m/	0	0	0	0	0	0	0	1	1	0	0	0	0	1	130	2	3	0	1	1	0	140
/n/	1	0	0	1	0	0	0	0	0	1	3	1	1	4	2	122	2	0	2	0	0	140
/ŋ/	0	0	0	0	0	1	0	1	0	0	3	0	0	1	1	3	129	1	0	0	0	140
/l/	1	0	0	0	0	1	2	2	1	1	0	0	0	0	1	1	1	124	1	2	2	140
/r/	1	1	4	3	1	5	5	3	6	2	0	1	2	1	1	1	4	4	91	0	4	140
/w/	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	139	0	140
/j/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140	140
Misidentified response	24	17	10	27	24	14	24	24	21	14	18	19	18	16	7	9	16	11	6	6	8	

Table 9: *Confusion matrix for normal hearing listeners at SNR = −24 dB (Experiment 1).*

Stimulus	Response																					Total
	/p/	/pʰ/	/b/	/t/	/tʰ/	/d/	/ʨ/	/ʨʰ/	/k/	/kʰ/	/ʔ/	/f/	/s/	/h/	/m/	/n/	/ŋ/	/l/	/r/	/w/	/j/	
/p/	94	2	2	3	1	2	1	1	4	6	1	4	5	2	1	2	3	1	2	2	1	140
/pʰ/	0	95	3	1	4	2	1	4	2	4	1	2	3	2	4	2	5	1	3	1	0	140
/b/	1	1	107	1	3	2	1	1	1	1	2	4	3	3	3	3	0	2	0	1	0	140
/t/	3	1	1	124	2	0	0	1	2	0	0	3	2	0	0	0	0	0	1	0	0	140
/tʰ/	3	4	1	4	90	1	1	4	1	2	4	3	0	2	1	3	3	5	2	4	2	140
/d/	6	3	5	4	6	58	5	5	5	6	4	6	2	1	5	2	3	3	3	4	4	140
/ʨ/	5	2	1	4	2	1	94	3	5	4	1	5	5	1	1	1	0	0	1	2	2	140
/ʨʰ/	5	3	5	5	2	5	2	57	6	6	4	4	2	5	5	4	5	4	3	4	4	140
/k/	1	0	3	3	1	0	7	2	106	4	0	2	2	2	0	0	0	0	4	3	0	140
/kʰ/	5	5	4	4	3	4	0	1	1	70	4	1	4	4	3	6	4	5	5	2	5	140
/ʔ/	2	1	0	1	1	2	1	0	0	1	108	2	1	5	3	4	3	0	2	2	1	140
/f/	4	1	3	3	4	3	4	2	4	3	4	88	2	2	3	3	1	2	1	1	2	140
/s/	5	4	4	4	4	4	5	2	6	4	1	6	51	3	5	4	6	4	5	7	6	140
/h/	3	7	1	5	2	1	5	2	2	1	5	0	2	85	4	5	3	3	1	1	2	140
/m/	2	0	4	1	1	1	1	1	2	3	2	1	0	0	113	1	3	0	0	1	3	140
/n/	1	0	2	1	1	2	0	0	1	1	5	0	0	0	1	113	5	1	3	2	1	140
/ŋ/	7	4	5	7	5	4	6	6	5	4	7	4	3	4	4	3	52	3	0	3	4	140
/l/	7	4	5	4	5	3	3	3	4	6	6	5	2	4	4	6	4	57	3	3	2	140
/r/	5	4	4	5	1	4	6	4	7	7	3	3	4	2	2	5	6	4	53	4	7	140
/w/	0	0	1	0	0	1	0	0	0	2	1	1	1	1	0	0	0	0	2	128	2	140
/j/	4	4	3	3	4	3	2	5	3	5	2	2	2	4	4	0	3	2	3	1	81	
Misidentified response	69	50	57	63	52	45	51	47	61	70	57	58	45	47	53	54	57	40	44	48	48	

Table 10: *Confusion matrix for sensorineural hearing loss listeners without hearing aids (Experiment 2).*

Stimulus	Response																					Total
	/p/	/pʰ/	/b/	/t/	/tʰ/	/d/	/tɕ/	/tɕʰ/	/k/	/kʰ/	/ʔ/	/f/	/s/	/h/	/m/	/n/	/ŋ/	/l/	/r/	/w/	/j/	
/p/	144	1	1	2	1	2	0	0	2	0	0	2	0	0	1	0	1	1	1	0	1	160
/pʰ/	3	117	3	2	3	2	3	2	1	1	4	2	3	2	2	2	2	3	1	1	1	160
/b/	3	2	119	2	1	2	2	0	1	2	3	2	4	1	2	2	2	2	3	5	0	160
/t/	3	0	2	135	0	1	2	2	3	0	0	1	2	2	2	1	1	2	0	0	1	160
/tʰ/	3	5	1	2	116	1	1	2	4	6	3	3	2	0	2	2	2	1	1	2	1	160
/d/	2	0	0	3	2	129	3	1	2	1	1	2	3	0	2	1	1	3	1	1	2	160
/tɕ/	3	0	1	2	0	1	125	2	3	2	2	3	4	1	1	1	1	2	2	4	0	160
/tɕʰ/	5	3	2	4	2	2	3	109	2	2	2	3	4	2	3	2	3	3	2	1	1	160
/k/	2	1	0	2	4	2	2	1	131	1	0	1	2	1	1	2	1	2	1	1	2	160
/kʰ/	2	5	3	4	2	3	4	5	4	102	4	1	4	1	1	1	5	5	0	2	2	160
/ʔ/	1	1	2	2	3	1	0	1	3	2	128	4	0	4	0	3	1	0	1	1	2	160
/f/	6	2	1	0	0	1	1	1	1	1	0	141	0	0	2	1	0	0	1	0	1	160
/s/	3	1	3	5	3	2	1	1	4	1	2	3	116	2	0	1	2	3	3	2	2	160
/h/	2	1	0	4	2	1	1	3	2	2	5	3	4	113	2	5	5	1	1	1	2	160
/m/	3	2	1	3	3	2	2	1	2	2	4	3	3	1	118	1	3	0	3	1	2	160
/n/	0	4	1	3	5	2	3	3	3	4	3	1	3	3	2	108	4	4	2	2	0	160
/ŋ/	2	4	3	4	4	4	1	3	1	2	4	2	3	2	2	3	108	5	1	1	1	160
/l/	2	1	4	2	1	2	1	1	1	3	2	3	3	1	2	3	3	118	4	1	2	160
/r/	6	4	7	6	3	6	7	7	6	4	3	5	4	3	3	2	1	4	74	3	2	160
/w/	2	0	2	2	0	1	0	1	2	2	0	3	2	1	2	0	0	1	2	137	0	160
/j/	0	3	0	2	1	1	3	0	1	2	0	2	3	1	0	3	1	3	2	2	130	160
Misidentified response	53	40	37	56	40	39	40	37	48	40	42	49	53	28	32	36	39	45	32	31	25	

Table 11: *Confusion matrix for sensorineural hearing loss listeners with hearing aids (Experiment 2).*

Stimulus	Response																					Total
	/p/	/pʰ/	/b/	/t/	/tʰ/	/d/	/tɕ/	/tɕʰ/	/k/	/kʰ/	/ʔ/	/f/	/s/	/h/	/m/	/n/	/ŋ/	/l/	/r/	/w/	/j/	
/p/	153	0	1	3	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	160
/pʰ/	0	141	2	1	3	0	2	0	1	1	1	1	3	0	1	1	1	0	1	0	0	160
/b/	5	1	128	1	1	1	1	1	1	1	1	4	4	1	2	2	2	2	0	0	1	160
/t/	2	1	0	145	0	2	2	2	0	1	0	1	1	1	0	0	1	0	0	0	1	160
/tʰ/	1	3	1	1	137	1	0	3	2	2	0	2	3	1	1	1	1	0	0	0	0	160
/d/	4	1	3	2	1	128	5	1	2	1	0	2	4	1	0	0	1	0	2	1	1	160
/tɕ/	2	0	1	1	1	0	141	3	1	2	2	3	2	0	0	0	0	0	1	0	0	160
/tɕʰ/	4	3	1	3	0	2	1	137	1	2	0	1	2	0	1	2	0	0	0	0	0	160
/k/	2	1	1	1	3	1	4	0	139	0	1	1	2	0	0	1	0	2	0	0	1	160
/kʰ/	1	3	1	1	1	1	2	2	1	141	1	0	2	0	1	0	1	1	0	0	0	160
/ʔ/	1	3	2	4	1	0	0	0	0	2	134	1	2	3	2	3	1	1	0	0	0	160
/f/	4	0	3	2	1	2	1	1	0	0	1	143	0	0	1	0	0	0	0	1	0	160
/s/	2	2	3	4	1	1	4	1	2	0	2	3	131	1	0	1	0	0	2	0	0	160
/h/	1	1	1	1	1	0	1	2	0	0	0	2	1	145	0	0	2	0	1	1	0	160
/m/	2	3	1	2	2	0	2	0	1	1	1	0	1	0	139	2	1	1	0	0	1	160
/n/	3	1	0	2	0	2	3	0	0	2	1	2	0	2	2	132	2	4	1	0	1	160
/ŋ/	1	2	0	2	2	3	4	2	1	0	2	1	3	3	2	2	127	0	1	1	1	160
/l/	2	1	1	2	0	1	2	1	2	1	1	1	1	2	2	0	2	133	3	1	1	160
/r/	6	1	4	4	2	5	4	4	5	3	1	4	4	1	0	3	2	3	96	4	4	160
/w/	0	1	1	1	1	0	2	1	0	1	0	1	2	0	1	1	1	1	1	143	1	160
/j/	0	1	0	0	0	1	1	1	0	0	0	1	1	0	1	1	0	3	0	0	149	160
Misidentified	43	29	27	38	21	23	42	25	20	21	15	32	38	16	17	20	18	18	13	9	13	

Table C: Confusion matrix for final consonants for normal hearing listeners at SNR = -6 dB.

Stimulus	Response							
	/p/	/t/	/k/	/m/	/n/	/ŋ/	/w/	/j/
/p/	143	1	2	0	0	0	0	1
/t/	3	141	2	0	1	0	0	0
/k/	1	8	133	0	3	1	0	1
/m/	0	0	1	143	1	1	1	0
/n/	0	2	0	0	136	6	0	3
/ŋ/	0	1	0	3	2	140	0	1
/w/	0	0	0	1	0	0	146	0
/j/	1	0	0	0	1	0	0	145

Table D: Confusion matrix for final consonants for normal hearing listeners at SNR = -12 dB.

Stimulus	Response							
	/p/	/t/	/k/	/m/	/n/	/ŋ/	/w/	/j/
/p/	139	3	4	0	0	0	0	1
/t/	7	126	8	0	1	1	0	4
/k/	6	12	117	2	1	5	0	4
/m/	0	0	0	141	1	1	0	4
/n/	0	5	0	2	139	1	0	0
/ŋ/	0	0	3	5	5	131	0	3
/w/	0	1	0	0	0	1	145	0
/j/	0	1	0	0	0	2	0	144

Table E: Confusion matrix for final consonants for normal hearing listeners at SNR = -18 dB.

Stimulus	Response							
	/p/	/t/	/k/	/m/	/n/	/ŋ/	/w/	/j/
/p/	119	4	14	4	0	2	0	4
/t/	11	118	4	0	7	0	1	6
/k/	6	12	117	0	3	6	0	3
/m/	1	0	1	119	6	11	3	6
/n/	0	9	0	10	117	3	0	8
/ŋ/	0	1	7	4	3	132	0	0
/w/	0	0	0	3	2	0	142	0
/j/	0	0	1	6	11	9	0	120

Table F: Confusion matrix for final consonants for normal hearing listeners at SNR = -24 dB.

Stimulus	Response							
	/p/	/t/	/k/	/m/	/n/	/ŋ/	/w/	/j/
/p/	108	9	12	4	0	5	0	9
/t/	18	95	11	4	6	4	2	7
/k/	12	12	92	4	7	10	4	6
/m/	8	4	9	94	7	10	4	11
/n/	3	14	5	11	89	15	2	8
/ŋ/	6	5	11	11	9	90	5	10
/w/	7	8	11	9	14	7	79	12
/j/	5	8	5	9	8	11	0	101

4. Distance Matrix and Perceptual Representation Space

To further our analysis of confusion patterns, a number of perceptual representations was constructed based on the data. To do so, similarity scores and distance matrices were first calculated and constructed, followed by construction of perceptual representations, resulting in three separate representations: one for normal listeners at -18 dB and the others for hearing loss patients (without and with hearing aids).

4.1. Similarity Scores

The experimental setup in Section 3 was designed to equally make pairwise comparisons among 21 word-initial phonemes resulting in 210 real-word stimulus pairs. Then, percent correct responses and confusion matrices (Tables6—11) were obtained. Similarity score and perceptual distance for each phoneme pair were systematically derived from the confusion scores based on a method proposed by Shepard (1972). Specifically, the similarity score between each pair of phonemes is calculated from confusion scores by

$$S_{ij} = \frac{P_{ij} + P_{ji}}{P_{ii} + P_{jj}}, \tag{2}$$

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 based on Shepard (1972) by

$$d_{ij} = -\ln S_{ij}. \tag{3}$$

4.2. Distance Matrices and Perceptual Representations

4.2.1. Experiment 1

From the intelligibility test results across 4 SNR levels for 28 normal hearing listeners in Table 4, it is clear that percent intelligibility scores were decreasing as increasing level of noise. While subjects' performance at SNR level of -6 dB and -12 dB was near-perfect, at -24 dB it was far below a guesswork (50%). Therefore, the remaining SNR level of -18 dB was the most interpretable to investigate the perceptual distance. Therefore, the confusion matrix at this SNR level was used to compute similarity scores, and to derive a distance matrix. Table 12 shows the distance scores for Thai initial phonemes at the SNR level of -18 dB.

Table 12: *Distance scores for 21 Thai initial phonemes in Thai at SNR = −18 dB.*

Stimulus	Response																				
	/p/	/pʰ/	/b/	/t/	/tʰ/	/d/	/tɕ/	/tɕʰ/	/k/	/kʰ/	/ʔ/	/f/	/s/	/h/	/m/	/n/	/ŋ/	/l/	/r/	/w/	/j/
/p/		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	∞	∞
/pʰ/			∞	4.1	3.1	∞	5.5	3.7	∞	3.9	4.8	∞	5.5	3.7	∞	∞	∞	∞	5.4	∞	∞
/b/				4.8	5.5	5.6	∞	∞	∞	5.5	∞	5.6	5.5	∞	∞	∞	5.6	∞	4.0	5.6	∞
/t/					4.0	∞	3.4	4.8	3.7	4.8	∞	3.5	3.5	∞	∞	5.5	∞	∞	4.0	∞	∞
/tʰ/						5.5	4.0	3.7	4.8	3.4	5.5	4.4	4.0	4.4	∞	∞	∞	∞	5.3	∞	∞
/d/							4.1	5.6	4.2	∞	∞	4.9	5.5	∞	∞	∞	5.6	5.6	3.8	∞	∞
/tɕ/								3.7	4.4	5.5	∞	4.4	3.7	∞	∞	5.5	∞	4.4	3.7	4.8	5.5
/tɕʰ/									∞	4.1	5.5	∞	4.8	∞	5.6	∞	5.5	4.4	4.3	∞	∞
/k/										5.5	5.6	∞	4.8	∞	5.6	∞	5.6	4.8	3.6	∞	5.6
/kʰ/											3.7	∞	5.5	4.1	∞	5.5	∞	5.5	4.6	∞	∞
/ʔ/												∞	5.5	3.5	5.6	4.1	3.9	∞	∞	∞	∞
/f/													3.7	∞	∞	5.5	∞	5.5	5.4	5.6	∞
/s/														∞	∞	5.5	∞	5.5	4.2	∞	∞
/h/															4.8	4.1	4.4	5.5	5.4	∞	∞
/m/																4.1	4.2	5.5	4.7	5.6	∞
/n/																	3.9	5.5	4.3	∞	∞
/ŋ/																		4.8	4.0	∞	∞
/l/																			3.8	4.9	4.9
/r/																				∞	4.1
/w/																					∞
/j/																					

A perceptual space representation, constructed based on the perceptual distances in Table 12 and sketched in Fig. 1, shows relative locations for each Thai initial phoneme according to Johnson (2003) and Shepard (1972). It should be noted that the perceptual representation graphically represents the confusion patterns in Table 8 as well. 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 appear to be five clustering of phonemes (shown in dashed-line circles) based on their common phonological features adapted from Jacobson et al. (1952), 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.

4.3. Experiment 2

Tables 13 and 14 show the distance scores derived from Tables 10 and 11 from eight hearing loss patients. Figures 2 and 3 illustrate perceptual representations of Thai initial consonants for eight sensorineural hearing loss patients without and with hearing aids. Five clusters could be grouped similar to Fig. 1. Comparison across Figs. 1, 2, and 3 suggests that patients' perceptual difficulty could be attributed to the nasality grouping, which is well separated for normal hearing listeners as well as patients with hearing aids, shifting closer to the glottal constrictions and aspirated obstruents. From Fig. 3, it seems that the hearing aids are beneficial in moving the nasality cluster further away from the nearby groupings.

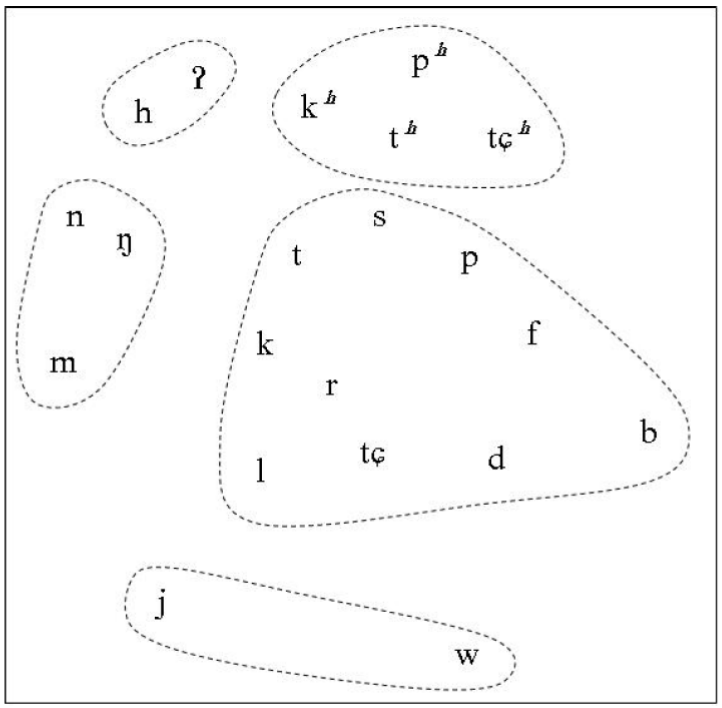

Figure 1: Perceptual representation of 21 Thai initial phonemes in Thai (from Fig. 2 of Tantibundhit et al. (2011b)).

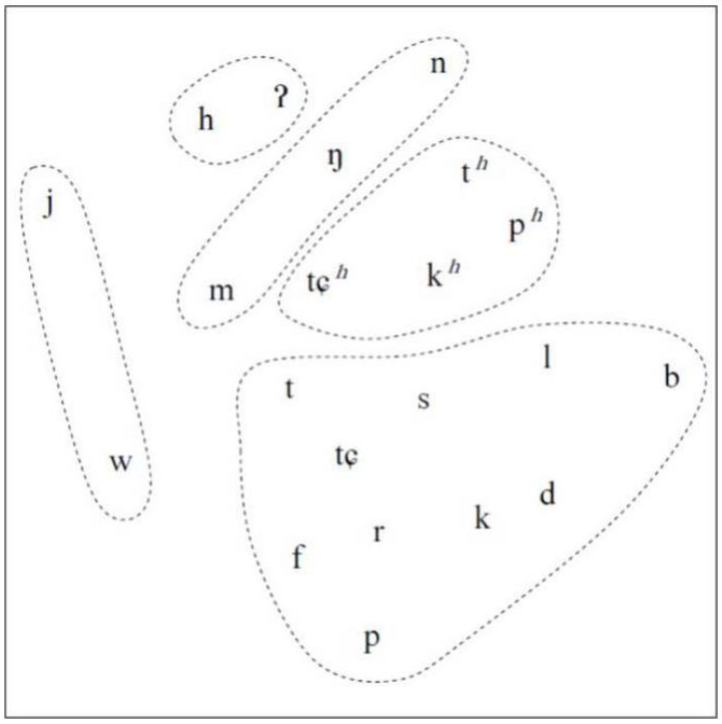

Figure 2: Perceptual representation of 21 Thai initial phonemes for sensori-neural hearing loss patients without hearing aids.

Table 13: *Distance scores for sensorineural hearing loss subjects without hearing aids.*

Stimulus	Response																				
	/p/	/p^h/	/b/	/t/	/t^h/	/d/	/ʧ/	/ʧ^h/	/k/	/k^h/	/ʔ/	/f/	/s/	/h/	/m/	/n/	/ŋ/	/l/	/r/	/w/	/j/
/p/		4.2	4.2	4.0	4.2	4.2	4.5	3.9	4.2	4.8	5.6	3.6	4.5	4.9	4.2	∞	4.4	4.5	3.4	4.9	5.6
/p^h/			3.9	4.8	3.4	4.8	4.4	3.8	4.8	3.6	3.9	4.2	4.1	4.3	4.1	3.6	3.6	4.1	3.6	5.5	4.1
/b/				4.2	4.8	4.8	4.4	4.7	5.5	3.8	3.9	4.5	3.5	5.4	4.4	4.3	3.8	3.7	3.0	3.6	∞
/t/					4.8	4.2	4.2	3.7	4.0	4.1	4.9	5.6	3.6	3.7	3.9	4.1	3.9	4.1	3.6	4.9	4.5
/t^h/						4.4	5.5	4.0	3.4	3.3	3.7	4.5	3.8	4.7	3.8	3.5	3.6	4.8	3.9	4.8	4.8
/d/							4.2	4.4	4.2	4.1	4.9	4.5	3.9	5.5	4.1	4.4	3.9	3.9	3.4	4.9	4.5
/ʧ/								3.8	3.9	3.6	4.8	4.2	3.9	4.8	4.4	4.1	4.8	4.4	3.1	4.2	4.4
/ʧ^h/									4.4	3.4	4.4	4.1	3.8	3.8	4.0	3.8	3.6	4.0	3.0	4.8	5.5
/k/										3.8	4.5	4.9	3.7	4.4	4.4	3.9	4.8	4.4	3.4	4.5	4.5
/k^h/											3.6	4.8	3.8	4.3	4.3	3.7	3.4	3.3	3.8	4.1	4.1
/ʔ/												4.2	4.8	3.3	4.1	3.7	3.9	4.8	3.9	5.6	4.9
/f/													4.5	4.4	3.9	4.8	4.8	4.5	3.6	4.5	4.5
/s/														3.6	4.4	4.0	3.8	3.7	3.3	4.1	3.9
/h/															4.3	3.3	3.5	4.7	3.8	4.8	4.4
/m/																4.3	3.8	4.8	3.5	4.4	4.8
/n/																	3.4	3.5	3.8	4.8	4.4
/ŋ/																		3.3	4.5	5.5	4.8
/l/																			3.2	4.8	3.9
/r/																				3.7	3.9
/w/																					4.9
/j/																					

Table 14: *Distance scores for sensorineural hearing loss subjects with hearing aids.*

Stimulus	Response																				
	/p/	/p^h/	/b/	/t/	/t^h/	/d/	/ʨ/	/ʨ^h/	/k/	/k^h/	/ʔ/	/f/	/s/	/h/	/m/	/n/	/ŋ/	/l/	/r/	/w/	/j/
/p/		∞	3.8	4.1	5.7	4.3	4.6	4.3	5.0	5.0	5.7	4.1	5.0	5.7	5.0	4.6	5.6	5.0	3.7	∞	∞
/p^h/			4.5	5	3.8	5.6	4.9	4.5	4.9	4.3	4.2	5.6	4.0	5.7	4.2	4.9	4.5	5.6	4.8	5.6	5.7
/b/				5.6	4.9	4.2	4.9	4.9	4.9	4.9	4.5	3.7	3.6	4.9	4.5	4.9	4.8	4.5	4.0	5.6	5.6
/t/					5.6	4.2	4.6	4.0	5.6	5.0	4.2	4.6	4.0	5.0	5.0	4.9	4.5	4.9	4.1	5.7	5.7
/t^h/						4.9	5.6	4.5	4.0	4.5	5.6	4.5	4.2	4.9	4.5	5.6	4.5	∞	4.8	5.6	∞
/d/							4.0	4.5	4.5	4.9	∞	4.2	3.9	5.6	∞	4.9	4.2	5.6	3.5	5.6	4.9
/ʨ/								4.2	4.0	4.3	4.9	4.3	3.8	5.7	4.9	4.5	4.2	4.9	3.9	5.0	5.7
/ʨ^h/									5.6	4.2	∞	4.9	4.5	4.9	5.6	4.9	4.9	5.6	4.1	5.6	5.7
/k/										5.6	5.6	5.6	4.2	∞	5.6	5.6	5.6	4.2	3.9	∞	5.7
/k^h/											4.5	∞	4.9	∞	4.9	4.9	5.6	4.9	4.4	5.6	∞
/ʔ/												4.9	4.2	4.5	4.5	4.2	4.5	4.9	5.4	∞	∞
/f/													4.5	5.0	5.6	4.9	5.6	5.6	4.1	5.0	5.7
/s/														4.9	5.6	5.6	4.5	5.6	3.6	4.9	5.6
/h/															∞	4.9	4.0	4.9	4.8	5.7	∞
/m/																4.2	4.5	4.5	∞	5.6	5.0
/n/																	4.2	4.2	4.0	5.6	4.9
/ŋ/																		4.9	4.3	4.9	5.6
/l/																			3.6	4.9	4.3
/r/																				3.9	4.1
/w/																					5.7
/j/																					

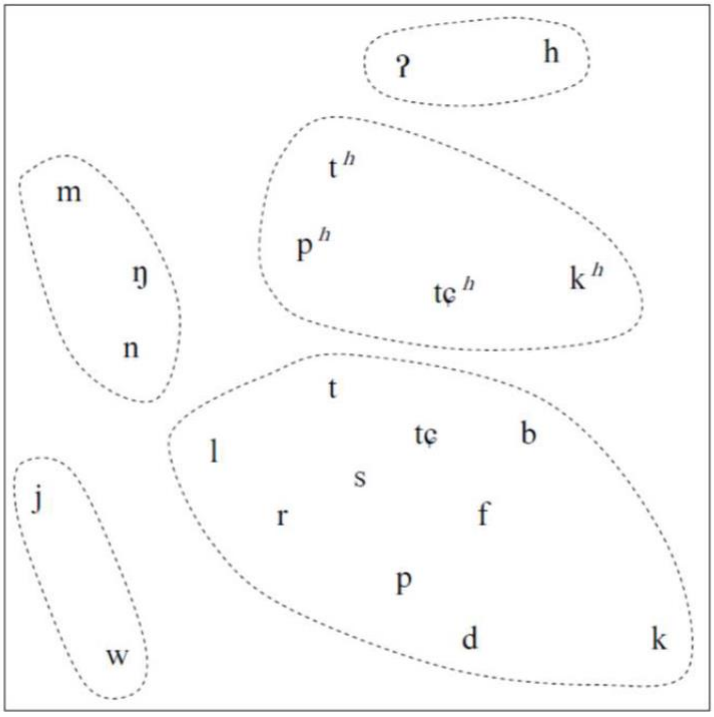

Figure 3: Perceptual representation of 21 Thai initial phonemes for sensori-neural hearing loss patients with hearing aids.

5. Response Bias Measurement

5.1. SDT Bias Values of c

From the confusion matrix at -18 dB (Table 8), it is clear that voicing was the most robust feature as opposed to place and manner of articulation. Therefore, in this section, voicing is our central focus. With its robustness (cross-linguistically), we believe any response between voiced and unvoiced might give us interesting insights into an understanding of the perception of voicing. The SDT bias values of *c*, are used to highlight and quantify confusion asymmetries that exist in certain initial phoneme pairs (Benkí, 2003). After confusion matrices for 21 initial phonemes are constructed, the bias measure *c* (criterion) (Macmillan and Creelman, 2005) is then calculated by

$$c = -0.5[z(H) + z(F)], \tag{4}$$

where *z*(*H*) and *z*(*F*) are *z* scores of hit and false alarm rates extracted from the main confusion matrix. Specifically, for a 2×2 confusion matrix of two initials *x* and *y*, the diagonal entries are the response frequencies for *x* and *y* from the diagonal of the original confusion matrix, which are referred to the hits and correct rejections, respectively. In addition, the response frequency for *y* given stimulus *x* from the original matrix is referred to the miss frequency, while the response frequency for *x* given stimulus *y* is referred to the false alarms. The marginal totals from the 2×2

confusion matrix of two initials *x* and *y* are used to compute the hit rate and false alarm rate (Benkí, 2003; Macmillan and Creelman, 2005).

As mentioned, the focus is to investigate SDT bias values of *c* in initial phoneme pairs specifically for either one of unvoiced phonemes (/p/, /p^h/, /t/, /t^h/, /tɕ/, /tɕ^h/, /k/, /k^h/, /ʔ/, /f/, /s/, and /h/) to either one of voiced phonemes (/b/, /d/, /m/, /n/, /ŋ/, /l/, /r/, /w/, and /j/). Moreover, we compare differences between the SDT bias values of *c* across two groups of subjects, i.e., normal hearing listeners and sensorineural hearing loss listeners. Finally, we examine the extent to which hearing aids have on bias asymmetries in the initial phoneme pairs.

5.2. Experimental Results

Tables 15, 16, and 17 show the SDT bias values of *c* for all combinations of phoneme pairs of any unvoiced and voiced phonemes, e.g., /p/-/b/ and /h/-/j/. It should be important to note that for any phoneme pair, a negative *c*-value means that in identifying the two phonemes, the subjects favored the row (unvoiced) phoneme over the column (voiced) phoneme, and vice versa for a positive *c*-value. For example, in Table 15, for the pair /p/-/b/, the corresponding *c*-value is 0.005, meaning that the subjects “very slightly” favored /b/ over /p/.

Moreover, a negative/positive weighted normalized sum of *c*-value (weighted *c*-value) in the last two columns (for unvoiced phonemes) of each table is calculated by summation of all negative/positive *c*-values in each row of the table divided by summation of absolute *c*-values in that row. Weighted *c*-value in the last two rows (for voiced phonemes) is calculated in the same fashion. Therefore, sum of the absolute of weighted *c*-values will be unity. For example, in Table 15, identification response of phoneme /p/ (row) has weighted *c*-values of -0.977 (for unvoiced) and 0.023 (for voiced) which means that overall for this phoneme, listeners favored unvoiced to voiced phonemes. Table 18 gives a summary of the weighted *c*-value patterns across three groups of listeners.

For normal hearing listeners, overall the subjects show biases in favor of unvoiced phonemes when identifying /p/, /p^h/, /tɕ/, /tɕ^h/, /k/, /k^h/, /ʔ/, /h/, /m/, /n/, /l/, and /r/ rather than voiced phonemes. In fact, among the unvoiced phonemes, 8 out of 12 phonemes (33.33/50%) show biases towards unvoiced and among the voiced ones, 4 out of 9 (22.22/50%) biases towards unvoiced. Likewise, sensorineural hearing loss subjects show more biases towards unvoiced phonemes when identifying /p/, /p^h/, /t/, /t^h/, /tɕ/, /k/, /ʔ/, /f/, /s/, /b/, /n/, /ŋ/, and /r/. In Table 18, 9 out of 12 phonemes (37.5/50%) show biases towards unvoiced and 4 out of 9 voiced phonemes (22.22/50%) do. Finally, hearing aids seem to increase substantial unvoiced biases. In Table 18, 11 out of 12 unvoiced phonemes (45.9/50%) were bias towards unvoiced and all of the voiced phonemes do.

Table 15: *SDT bias values of c: unvoiced vs. voiced phonemes for normal hearing listeners.*

unvoiced	voiced									weighted	
	/b/	/d/	/m/	/n/	/ŋ/	/l/	/r/	/w/	/j/	normalized sum	
/p/	0.005	−0.195	−0.004	−0.211	−0.005	−0.208	−0.267	0.007	0.009	**−0.977**	**0.023**
/p^h/	0.021	0.015	0.012	0.001	0.011	0.004	−0.251	0.023	0.024	**−0.694**	**0.306**
/t/	0.326	0.022	0.019	−0.188	0.018	0.011	−0.266	0.030	0.031	**−0.498**	**0.502**
/t^h/	0.232	0.226	0.025	0.014	0.024	0.017	−0.238	0.036	0.037	**−0.281**	**0.720**
/t͡ɕ/	0.029	−0.186	0.022	0.208	0.019	−0.117	−0.522	0.034	0.230	**−0.604**	**0.396**
/t͡ɕ^h/	0.014	0.203	−0.190	−0.005	−0.191	−0.134	−0.434	0.016	0.018	**−0.792**	**0.208**
/k/	0.009	−0.414	−0.195	−0.011	0.194	−0.009	−0.582	0.011	0.207	**−0.742**	**0.258**
/k^h/	0.224	0.020	0.018	−0.189	0.016	−0.186	−0.350	0.029	0.030	**−0.683**	**0.317**
/ʔ/	0.011	0.005	0.198	−0.222	−0.080	−0.005	−0.058	0.014	0.015	**−0.600**	**0.400**
/f/	0.210	0.303	0.005	−0.202	0.004	0.193	−0.258	0.212	0.018	**−0.328**	**0.673**
/s/	0.225	0.219	0.019	−0.188	0.018	0.209	−0.181	0.030	0.031	**−0.329**	**0.671**
/h/	0.018	0.012	0.010	−0.422	0.139	0.198	−0.254	0.020	0.022	**−0.618**	**0.383**
weighted	**1.000**	**0.563**	**0.457**	**0.120**	**0.616**	**0.490**	**0.000**	**1.000**	**1.000**		
normalized sum	**−0.000**	**−0.437**	**−0.543**	**−0.880**	**−0.384**	**−0.510**	**−1.000**	**−0.000**	**−0.000**		

Table 16: *SDT bias values of c: unvoiced vs. voiced phonemes for sensorineural hearing loss listeners without hearing aids.*

unvoiced	voiced									weighted	
	/b/	/d/	/m/	/n/	/ŋ/	/l/	/r/	/w/	/j/	normalized sum	
/p/	−0.248	−0.021	−0.250	−0.050	−0.185	−0.167	−0.512	−0.244	0.130	**−0.928**	**0.072**
/p^h/	0.086	0.257	0.002	−0.161	−0.161	0.216	−0.378	0.179	−0.192	**−0.547**	**0.453**
/t/	−0.025	−0.220	−0.109	−0.257	−0.318	−0.026	−0.569	−0.233	−0.137	**−1.000**	**0.000**
/t^h/	0.005	−0.111	−0.079	−0.209	−0.159	0.003	−0.311	0.269	0.021	**−0.745**	**0.255**
/tɕ/	−0.140	−0.206	−0.142	−0.243	−0.027	0.120	−0.394	0.368	−0.293	**−0.747**	**0.253**
/tɕ^h/	0.257	0.163	0.230	−0.085	−0.002	0.230	−0.367	0.042	0.183	**−0.291**	**0.709**
/k/	−0.167	−0.003	−0.150	−0.121	−0.035	0.111	−0.495	−0.121	0.128	**−0.820**	**0.180**
/k^h/	0.115	0.261	−0.105	−0.267	0.208	0.143	−0.435	0.058	0.048	**−0.492**	**0.508**
/ʔ/	−0.097	0.001	−0.369	−0.035	−0.309	−0.254	−0.329	0.161	0.240	**−0.776**	**0.224**
/f/	−0.162	−0.146	−0.117	−0.048	−0.289	−0.335	−0.464	−0.303	−0.145	**−1.000**	**0.000**
/s/	−0.056	−0.060	−0.301	−0.229	−0.097	0.004	−0.162	0.033	−0.059	**−0.964**	**0.036**
/h/	−0.142	0.175	0.140	0.101	0.185	0.008	−0.306	0.035	0.158	**−0.359**	**0.642**
weighted	**0.309**	**0.528**	**0.813**	**0.056**	**0.199**	**0.516**	**0.000**	**0.560**	**0.524**		
normalized sum	**−0.692**	**−0.472**	**−0.187**	**−0.944**	**−0.801**	**−0.484**	**−1.000**	**−0.441**	**−0.476**		

Table 17: *SDT bias values of c: unvoiced vs. voiced phonemes for sensorineural hearing loss listeners with hearing aids.*

unvoiced	voiced									weighted	
	/b/	/d/	/m/	/n/	/ŋ/	/l/	/r/	/w/	/j/	normalized sum	
/p/	−0.3524	−0.3820	−0.2513	−0.3250	−0.1818	−0.2602	−0.5274	−0.0116	−0.0046	**−1.0000**	**0.0000**
/p^h/	0.1114	−0.1662	−0.2120	−0.0118	−0.1491	−0.1588	−0.0701	−0.1450	−0.1373	**−0.9041**	**0.0959**
/t/	−0.1710	−0.0242	−0.2420	−0.2525	−0.1540	−0.2509	−0.4366	−0.1499	0.1519	**−0.9171**	**0.0829**
/t^h/	−0.0122	−0.0122	−0.1263	0.1417	−0.1439	−0.0052	−0.3082	−0.1401	0.0145	**−0.8272**	**0.1728**
/ʨ/	−0.0174	−0.4099	−0.2372	−0.3109	−0.3696	−0.2461	−0.3522	−0.2315	−0.1373	**−1.0000**	**0.0000**
/ʨ^h/	−0.0122	0.1169	0.1507	0.2287	−0.2504	−0.1538	−0.4268	−0.1401	−0.1323	**−0.6921**	**0.3079**
/k/	−0.0148	−0.1450	−0.1480	0.1390	−0.1652	−0.0086	−0.4731	0.0049	0.1600	**−0.7586**	**0.2414**
/k^h/	−0.0174	−0.0174	−0.0026	−0.2476	0.1295	−0.0105	−0.3793	−0.1450	0.0095	**−0.8550**	**0.1450**
/ʔ/	0.1212	−0.0080	0.1360	0.2077	−0.1400	−0.0013	−0.2134	0.0113	0.0184	**−0.4231**	**0.5769**
/f/	−0.0831	−0.0215	0.1426	−0.2501	−0.1701	−0.1613	−0.4342	−0.0000	−0.1397	**−0.8984**	**0.1016**
/s/	−0.0652	−0.2762	−0.1376	0.1503	−0.3063	−0.1461	−0.2092	−0.2187	−0.1245	**−0.9080**	**0.0920**
/h/	−0.0223	−0.1710	−0.0073	−0.2525	−0.1072	−0.2509	−0.0751	0.1449	0.0047	**−0.8556**	**0.1444**
weighted	**0.2325**	**0.0668**	**0.2393**	**0.3445**	**0.0571**	**0.0000**	**0.0000**	**0.1199**	**0.3470**		
normalized sum	**−0.7675**	**−0.9332**	**−0.7607**	**−0.6555**	**−0.9429**	**−1.0000**	**−1.0000**	**−0.8801**	**−0.6530**		

Table 18: *Summary of weighted normalized sum bias across normal hearing subjects and sensorineural hearing loss subjects without/with hearing aids. Note that minus sign (−) indicates that on average listeners favored unvoiced to voiced phonemes and vice versa for plus sign (+).*

	normal hearing subjects	sensorineural hearing loss subjects without hearing aids	sensorineural hearing loss subjects with hearing aids
/p/	−	−	−
/p^h/	−	−	−
/t/	+	−	−
/t^h/	+	−	−
/tɕ/	−	−	−
/tɕ^h/	−	+	−
/k/	−	−	−
/k^h/	−	+	−
/ʔ/	−	−	+
/f/	+	−	−
/s/	+	−	−
/h/	−	+	−
/b/	+	−	−
/d/	+	+	−
/m/	−	+	−
/n/	−	−	−
/ŋ/	+	−	−
/l/	−	+	−
/r/	−	−	−
/w/	+	+	−
/j/	+	+	−

6. Discussion and Future Work

6.1. TDRT-I and TDRT-F as An Important Tool

We have developed the subjective intelligibility testing of Thai speech (TDRT-I and TDRT-F) and systematically compared confusion responses across all phonemes for initial and final consonants. Not only has TDRT-I and TDRT-F proved to be a very useful for intelligibility assessment, but it has offered valuable insights into phoneme confusion patterns (Section 3). Moreover, specifically for initial consonants, perceptual representation (Section 4), and response biases (Section 5) between normal hearing listeners and those with sensorineural hearing loss were analyzed.

It is interesting to find that some overall confusion patterns for initials were shared between the two groups of listeners, i.e., /r/ was the most confusable phoneme and /w/, /j/, and /p/ were among the least confusable. However, their perceptual patterns were quite different in details as reflected in their perceptual representations in Figs. 1—3. In view of these representations, it is suggested that some of their hearing deficits (at least for this group of patients) could be attributed to the shifting of some of the phoneme groupings, specifically the nasality which was moved closer to the glottals and aspirated obstruents. Hearing aids seemed to help in separating out the nasality from other groupings. However, the device improved the correct responses by 10% on average with only /tɕ^{h}/, /k*h*/, /s/, and /h/ (all unvoiced) showing significant improvement. The hearing deficits could be further examined when we look at the SDT bias values of *c* (Table 18) among all possible 108 pairs of unvoiced vs. voiced phonemes. The findings show that normal hearing subjects were bias in favour of unvoiced phonemes. More importantly, the unvoiced biases appeared to increase in hearing loss patients and substantially more with the hearing aids.

For final consonants, /k/ is the most confusable consonant and it was mostly misperceived as /t/, which is also a voiceless non-continuant. Interestingly, at the —18dB level, for both initial and final consonants, voicing was the most robust contrast while place-of-articulation was the least.

6.2. Phoneme Occurrence Frequency

The confusion matrices not only show a correct response sensitivity but also patterns of misperceptions. Investigation of normal hearing listeners' misidentified responses reveals that in initial position across the -6, -12, and -18 dB levels, the listeners appeared to highly favor /t/ and /t^{h}/ and disfavored /w/ over other consonants. One interpretation is to connect these biases to the frequency of phoneme occurrences found in a Thai BEST corpus (Kosawat et al., 2009), constructed from various types of written materials. From the data of more than 9 million words, among all initial consonants including clusters, /t^{h}/ occurs at the highest rate (followed by /n/ and /s/), whereas /w/ is among consonants of lowest occurrence, which include /tɕ^{h}/, /h/, /ʔ/, /b/,/ŋ/, and /f/ (Munthuli et al., 2013).

6.3. Comparison with The English Data from Miller and Nicely (1955)

Current speech perception models provide different accounts for a basic unit of analysis such as context dependent allophones (Ingram and Park, 1998), individual exemplar (Johnson, 1997). In this study, we have taken a more traditional approach and consider phoneme as the unit of analysis. TDRT-I 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

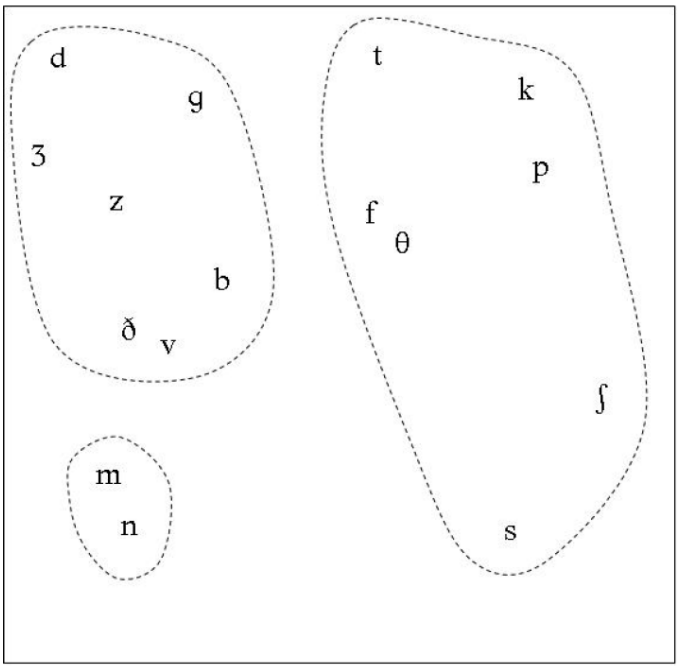

Figure 4: Perceptual representation of 15 initial phonemes (plus ʒ) in English adapted from a confusion matrix of Miller and Nicely (1955) (from Fig. 3 of Tantibundhit et al. (2011b)).

could be computed. This allows us to make some cross-linguistic observations concerning perceptual representations of Thai (Fig. 1) and English (Fig. 4) phonemes.

Figure 4 illustrates a perceptual representation of 16 (out of 22) English initial phonemes at SNR level of 12 dB with a bandwidth of 200—6,500 Hz adapted from a confusion matrix of Miller and Nicely (1955). This condition was chosen for the comparison because it yielded average percent correct response (88.67%), which is relatively close to that of -18 dB of the Thai data (90.85%). It is noteworthy that 8 phonemes in English (/tʃ/, /dʒ/, /ŋ/, /l/, /ɹ/, /w/, /j/, and /h/) were not included in the study of Miller and Nicely (1955). Roughly, English phonemes can be divided into 3 clusters, while Thai can be divided into 5 clusters. Separately, nasal sounds are grouped together in both languages. Voicing (voiced and unvoiced), one of the most distinct perceptual properties in English, appears to be a less 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 (Shepard, 1972; Mermelstein, 1976). On the other hand, in the case of Thai unaspirated obstruents, the separation among fricatives, affricates, and plosives seems less clear. Moreover, the two liquids in Thai seem to belong to the same cluster as the 8 unaspirated obstruents, the finding which calls for further investigation and explanation.

6.4. Status of /r/ and /l/

It is a well-known fact of Thai that the /r/ phoneme has undergone remarkable changes in the past decades. As a result, the /r/ has a variety of allophones including [r], [R], and [l] (the same sound as the existing /l/ phoneme). Moreover, Thai initial clusters containing /r/ have followed similar changes including the r-deletion. The instability and variations of /r/ might be linked to the finding that /r/ was clearly the most confusable phoneme of all. However, it should be noted that /r/ was misperceived as many other phonemes such as /t/, /tɕ/, /k/, /b/, and /l/.

6.5. Future work

Along with TDRT-I, we have developed subjective intelligibility testing of Thai final consonants (Tantibundhit et al., 2011c), vowels (Onsuwan et al., 2013) and tones (Onsuwan et al., 2012). Specifically, equipped with the experimental method and findings of the initial consonants, an adaptive testing is now being developed to detect hearing deficits and to come up with a way at identifying his/her degrees of hearing difficulty for certain groups of speech sound.

7. Acknowledgments

The author would like to thank Prof. Dr. Apirat Siritaratiwat, a mentor of the grant MRG5480272, for his guidance. The authors would also like to thank Sumonmas Thatphithakkul, Patcharika Chootrakool, Dr. Krit Kosawat, and Dr. Nattanun Thatphithakkul (all from National Electronics and Computer Technology Center (NECTEC), Thailand) for their support. Moreover, thank Dr. Nida Rueangwit at Department of Otorhinolarygology, Thammasat University Hospital for sensorineural hearing loss subject recruitment. Finally, thank Tanawan Saimai, Nuntaporn Saimai, and Phongphan Pienpanich, PhD students of the first author, for their assistance with the experiment tests.

References

Benki, J., 2003. Analysis of English nonsense syllable recognition in noise. Phonetica 60 (2), 129–157.

Comrie, B., 1990. The World's Major Languages. Oxford University Press, Oxford.

Giegerich, H., 1992. English Phonology: An Introduction. Cambridge University Press, Cambridge.

House, A., Williams, C., Hecker, H., Kryter, K., 1965. Articulation-testing methods: Consonantal differentiation with a closed-response set. J. Acoust. Soc. Am. 37 (1), 158–166.

Ingram, J., Park, S., 1998. 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.

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

Johnson, K., 1997. Speech perception without speaker normalization: An exemplar model. In: Johnson, K., Mullennix, J. (Eds.), Talker Variability in Speech Processing. Academic Press, CA, pp. 145–165.

Johnson, K., 2003. Acoustic and Auditory Phonetics, 2nd Edition. Wiley-Blackwell, Malden, MA.

Kosawat, K., Boriboon, M., Chootrakool, P., Chotimongkol, A., Klaithin, S., Kongyoung, S., Kriengket, K., Phaholphinyo, S., Purodakananda, S., Thanakulwarapas, T., Wutiwiwatchai, C., 2009. Best 2009: Thai word segmentation software contest. In: Proc. of SNLP. pp. 83–88.

Loizou, P., 2013. Speech Enhancement: Theory and Practice, 2nd Edition. CRC Press, Boca Raton, FL.

Macmillan, N., Creelman, C., 2005. Detection Theory: A User's Guide, 2nd Edition. Lawrence Erlbaum Associates, Mahwah, NJ.

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

Mermelstein, P., 1976. Distance Measures for Speech Recognition-Psychological and Instrumental, Haskins Laboratories Status Report on Speech Research SR-47.

Miller, G., Nicely, P., 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am. 27 (338), 338–352.

Munthuli, A., Sirimujalin, P., Tantibundhit, C., Kosawat, K., Onsuwan, C., 2013. A corpus-based study of phoneme distribution in thai. In: Proc. of SNLP. pp. 114–121.

Onsuwan, C., Tantibundhit, C., Saimai, N., Saimai, T., Chootrakool, P., Thatphithakkul?, S., 2013. Perception of Thai distinctive vowel length in noise. POMA 19 (060115), 1–7.

Onsuwan, C., Tantibundhit, C., Saimai, N., Saimai, T., Thatphithakkul, S., Chootrakool, P., 2012. Analysis of Thai tonal identification in noise. In: Proc. of SST. pp. 173–176.

Shepard, R., 1972. Psychological representation of speech sounds. In: David, E., Denes, P. (Eds.), Human Communication: A Unified View. McGraw- Hill, NY, pp. 67–113.

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

Stevens, K., 1981. Contraints imposed by the auditory system on the properties used to classify speech sounds: Data from phonology, acoustics, and psychoacoustics. In: Myers, T., Laver, J., Anderson, J. (Eds.), The Cognitive Representation of Speech. North-Holland, Oxford, pp. 61–74.

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

Tantibundhit, C., Boston, J., Li, C., Durrant, J., Shaiman, S., Kovacyk, K., El-Jaroudi, A., 2007. New signal decomposition method based speech enhancement. Signal Process 87 (11), 2607–2628.

Tantibundhit, C., Onsuwan, C., Phienphanich, P., Wutiwiwatchai, C., 2012. Methodological issues in assessing perceptual representation of consonant sounds in Thai. In: Proc. of Interspeech.

Tantibundhit, C., Onsuwan, C., Rueangwit, N., Saimai, N., Saimai, T., Thatphithakkul, N., Kosawat, K., Thatphithakkul, S., Chootrakool, P., 2011a. Perceptual deficits in Thai sensorineural hearing loss patients. J. Acoust. Soc. Am. 130 (4), 244.

Tantibundhit, C., Onsuwan, C., Saimai, T., Saimai, N., Thatphithakkul, S., Chootrakool, P., Kosawat, K., Thatphithakkul, N., 2011b. Perceptual representation of consonant sounds in thai. In: Proc. of Interspeech. pp. 3193–3196.

Tantibundhit, C., Onsuwan, C., Thatphithakkul, S., Chootrakool, P., Kosawat, K., Thatphithakkul, N., Saimai, T., Saimai, N., 2011c. Subjective intelligibility testing and perceptual study of thai initial and final consonants. In: Proc. of ICPhS. pp. 1970–1973.

Tantibundhit, C., Pernkopf, F., Kubin, G., 2010. Joint time-frequency segmentation algorithm for transient speech decomposition and speech enhancement. IEEE Trans. Audio Speech Lang. Process. 18 (6), 1417–1428.

Tingsabadh, K., Abramson, A., 1993. Thai. J. Int. Phon. Assoc. 23 (1), 24–28.

Voiers, W., 1983. Evaluating processed speech using the diagnostic rhyme test. Speech Technol. 1 (4), 30–39.

8. Output

8.1 International Journal Publication

Tantibundhit, C. and Onsuwan, C., Applying subjective intelligibility testing to analysis of confusions and perceptual representations of Thai initial consonants. Speech Communication (submitted).

8.2 International Conference

Tantibundhit, C., Onsuwan, C., Phienphanich, P., Wutiwiwatchai, C., 2012. Methodological issues in assessing perceptual representation of consonant sounds in Thai. In: Proc. of Interspeech.

Tantibundhit, C., Onsuwan, C., Rueangwit, N., Saimai, N., Saimai, T., Thatphithakkul, N., Kosawat, K., Thatphithakkul, S., Chootrakool, P., 2011a. Perceptual deficits in Thai sensorineural hearing loss patients. J. Acoust. Soc. Am. 130 (4), 244.

Tantibundhit, C., Onsuwan, C., Saimai, T., Saimai, N., Thatphithakkul, S., Chootrakool, P., Kosawat, K., Thatphithakkul, N., 2011b. Perceptual representation of consonant sounds in thai. In: Proc. of Interspeech. pp. 3193–3196.

Tantibundhit, C., Onsuwan, C., Thatphithakkul, S., Chootrakool, P., Kosawat, K., Thatphithakkul, N., Saimai, T., Saimai, N., 2011c. Subjective intelligibility testing and perceptual study of thai initial and final consonants. In: Proc. of ICPhS. pp. 1970–1973.

9. Appendix

SUBJECTIVE INTELLIGIBILITY TESTING AND PERCEPTUAL STUDY OF THAI INITIAL AND FINAL CONSONANTS

C. Tantibundhit[a], *C. Onsuwan*[b], *S. Thatphithakkul*[c], *P. Chootrakool*[c], *K. Kosawat*[c], *N. Thatphithakkul*[c], *T. Saimai*[a] *& N. Saimai*[a]

[a]Department of Electrical and Computer Engineering, Thammasat University, Thailand;
[b]Department of Linguistics, Thammasat University, Thailand;
[c]National Electronics and Computer Technology Center (NECTEC), Thailand
tchartur@engr.tu.ac.th; consuwan@tu.ac.th

ABSTRACT

We methodically design and develop a subjective intelligibility testing of Thai speech based on the diagnostic rhyme test (DRT). The Thai DRT (TDRT) consists of 2 test sets, one for initials and the other final consonants. The test for initials is designed to equally compare 21 phonemes pairwise, which results in 210 stimulus pairs. The TDRT for finals compares 8 final phonemes, yielding 84 stimulus pairs. These tests are well-constructed using real words. TDRT have two main advantages. It allows us to evaluate percent intelligibility responses in each stimulus pair and to systematically compare confusion responses across all phonemes. To test the validity of our method and to further our investigation, we carry out the subjective intelligibility test on twenty eight Thai listeners using TDRT, which varies in 4 SNR levels (–6, –12, –18, and –24dB). Average intelligibility scores and confusion matrices for initial and final consonants are analyzed.

Keywords: Thai, diagnostic rhyme test, subjective intelligibility, initial/final consonants, confusion matrix

1. INTRODUCTION

Speech intelligibility and speech quality are two distinct properties. Speech quality reflects how an utterance is produced and also includes speech attributes such as natural, raspy, hoarse, etc. Speech intelligibility, on the other hand, refers to what is being said, i.e., the meaning or the content of the spoken words [5]. Therefore, speech intelligibility is one of the essential attributes of the speech signal and needs to be preserved by speech enhancement algorithms [5].

Several algorithms have been developed specifically to enhance speech intelligibility in background noise [5]. Evaluating intelligibility of the enhanced compared with the original speech is often conducted using a subjective intelligibility testing [5]. Several intelligibility tests have been proposed for English by using rhyming words presented in six-response [2] or in pair-response [8]. House *et al*. developed a test by restricting response choices to a finite set of six rhyming words called the modified rhyme test (MRT) [2]. The test was composed of 50 sets, each of which was composed of six monosyllabic consonant vowel-consonant (CVC) words. Twenty-five sets differed in their initial consonants, while the rest differed in their final consonants.

Voiers refined the MRT and created a diagnostic rhyme test (DRT) [8], which is widely used for a subjective testing for measuring the intelligibility of speech coders [5]. The DRT was an A/B forced comparison test based on word pairs differing in their initial consonants by one of six distinctive features [8]. The DRT test material was composed of a word list of 96 rhyming pairs, e.g., *veal - feel*. As the DRT was developed specifically for English, it has some limitations when evaluating intelligibility of a tonal language such as Chinese [6]. McLoughlin devloped a New Chinese diagnostic rhyme test (NCDRT) [6]. The NCDRT was composed of a test set of phonemes in Chinese, which were classified under six distinctive features similar to the DRT [6].

Although the subjective intelligibility testing of a tonal language such as Chinese is well underway [6], subjective intelligibility testing of another tonal language, Thai, with several acoustic and phonemic differences from that of Chinese has yet to be developed. Therefore, this paper proposes an intelligibility testing of Thai speech specifically for its initial and final consonants. The tests are designed to facilitate an evaluation of percent intelligibility responses in each stimulus pair and to systematically compare confusion responses across all initial and final phonemes.

To do so, we have integrated several useful frameworks, namely DRT [8], NCDRT [6], MRT [2], and the analysis method of balanced confusion matrix [7]. Specifically, we use an A/B forced choice and monosyllabic (CV(V)(C)) rhyming pairs, which differ only in one sound either in an initial or final position (the tone is kept identical). These words are well-selected from real and commonly used words in the language. In this paper, a review of Thai Phonology is provided in Section 2, design and development of the TDRT for initial and final consonants in Section 3, experimental setup for the subjective intelligibility tests in Section 4, and experimental results in Section 5. Section 6 discusses the paper and mentions future work.

2. THAI PHONOLOGY REVIEW

Thai is a tonal language with 21 consonantal phonemes in initial position /p/, /p^h/, /b/, /t/, /t^h/, /d/, /tɕ/, /tɕ^h/, /k/, /k^h/, /ʔ/, /f/, /s/, /h/, /m/, /n/, /ŋ/, /l/, /r/, /w/, and /j/ and 9 consonantal phonemes in final position /p/, /t/, /k/, /ʔ/, /m/, /n/, /ŋ/, /j/, and /w/. Final /p/, /t/, /k/ in Thai are unreleased and often glottalized. 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).

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 ˋ, High ´ (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 [1].

3. TDRT DESIGN AND DEVELOPMENT

The goal of this section is to come up with two separate subjective intelligibility test sets specifically for Thai, each for initial and final consonants. In addition, the test should not be too long to cause fatigue [5]. To do so, a number of monosyllabic rhyming word pairs differing only in one sound either in an initial or final position is constructed step by step as follows:

3.1. TDRT for initial consonants

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 [3] and when combined with mid tone yields the most possible number of rhyming words, i.e., 21 rhyming words for 21 phonemes: /pāa/ ปา, /p^{h}āa/ พา, /bāa/ บา, /tāa/ ตา, /t^{h}āa/ ทา, /dāa/ ดา, /tɕāa/ จา, /tɕ^{h}āa/ ชา, /kāa/ กา, /k^{h}āa/ คา, /ʔāa/ อา, /fāa/ ฟา, /sāa/ ซา, /hāa/ ฮา, /māa/ มา, /nāa/ นา, /ŋāa/ งา, /lāa/ ลา, /rāa/ รา, /wāa/ วา, and /jāa/ ยา.

3) Each rhyming word is paired with 20 others of different initial phonemes. This results in a total combination of 210 stimulus pairs of rhyming words[1], which can be expressed mathematically as *a combination of 21 choose 2 ($^{21}C_2$).*

3.2. TDRT for final consonants

1) Pairs of monosyllabic ($C_iV^T(V)C_f$) words, each of which differs only in their final consonant phoneme (the tone in each pair remains identical) are garnered.

2) Two types of initial consonants C_i are chosen to create the rhyming words, namely voiceless unaspirated plosives (/p/, /t/, and /k/) and voiceless aspirated plosives (/p^h/, /t^h/, and /k^h/). The initial plosives are chosen over other types of initial consonant as they can be combined with the most possible types of rime unit (the sequence of vowel and final consonant).

3) Six initial plosives are subsequently combined with all 18 vowels: 9 short and 9 long vowels and with all 5 tones (6×18×5=540). For example, initial consonant /t/ when combined with a vowel /a/, a low tone ˋ, and 8 different final phonemes will produce /tàk/ ตัก, /tàt/ ตัด, /tàp/ ตับ, /tàŋ/ ตั้ง, /tàn/ ตั้น, /tàm/ ตำ่, /tàj/ ไต่, and /tàw/ เต่า. Altogether, 540 possible words are created.

4) Out of the 540 words, only 84 pairs of real words (84 stimulus pairs) that are commonly used are selected[2]. These stimulus pairs comprise 3 instances of each rhyming word paired with 7 others of different final phonemes, which can be expressed mathematically as *three times a combination of 8 choose 2 ($3\times{}^8C_2$).*

4. EXPERIMENTAL SETUP

The goal of this experiment encompasses two aspects. Firstly, to conduct the subjective intelligibility tests for initial and final consonants with 4 conditions of additive white Gaussian noise (AWG) using the developed rhyming words from the previous section. Percent intelligibility scores are calculated from, where P_s, N_r, N_w, and T are percent intelligibility score, numbers of correct responses, numbers of wrong responses, and total numbers to stimuli, respectively [8]. Four signal-to-noise ratios (SNR) of –6, –12, –18, and –24dB were chosen based on our preliminary findings such that intelligibility scores are in a range to avoid floor and ceiling effects, i.e., much higher than 50% (the scores are indistinguishable from guesswork) but not approaching 100% (subjects so well perceived stimuli) [5]. It should be pointed out that the average percent correct response, which does not necessarily match the intelligibility score, is calculated from total number of correct responses divided by total number of stimuli. Secondly, to gain insights into confusion patterns among phonetic categories for initial and final consonants.

To create stimulus materials, all 21 initial rhyming words and 84 pairs of final rhyming words along with filler words were read 5 times in a carrier sentence (ฉันชอบ ... อีกแล้ว /tɕʰǎn tɕʰɔ̂ɔp … ʔìik lɛ́ɛw/) 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, each target word stimulus was excised from the carrier sentence. To avoid audible discontinuity problems at the splice points, the starting point of each stimulus began approximately 10 ms prior to the onset of initial consonant. Moreover, its end point included some durational adjustments to the last sound segment at a precise location. Every splice was done at a zero crossing.

One of the 5 tokens of each target word that was the clearest, most typical, and most natural sounding was selected based on impressionistic hearing evaluation and spectrographic inspection. Average durations of stimuli of initial (C_iV^TV) and final ($C_iV^T(V)C_f$) rhyming words were 324.4ms and 309.1ms, respectively.

The intelligibility tests were performed individually 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. Sequence of individual trials as well as sequence of word in each A/B pair for intelligibility tests for initial and final consonants are randomized in real tests and explained in full details below.

4.1. Test setup for initial consonants

The test 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 consonants 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 [5]. Alternatively, by increasing a number of subjects 4 times, we could stay with the 500 trials and distribute the trials equally across 4 SNR levels, i.e., Groups A, B, C, and D, each of which contains SNR levels of –6dB, –12dB, –18dB, and –24dB as summarized in Table 1.

Table 1: Distributions of rhyming word groupings for initial and final consonants (referred from top header) and the remaining of final phonemes (referred from bottom header).

Subject	Rhyming and Filler Word (Initial and Final)			
	Group A	Group B	Group C	Group D
	Remaining Phoneme (Final)			
	/p/, /t/	/k/, /m/	/n/, /ŋ/	/j/, /w/
I	–6dB	–12dB	–18dB	–24dB
II	–24dB	–6dB	–12dB	–18dB
III	–18dB	–24dB	–6dB	–12dB
IV	–12dB	–18dB	–24dB	–6dB

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.

4.2. Test setup for final consonants

The final consonant test comprises 84 rhyming pairs across 8 final phonemes and 16 pairs of filler

words. To be in line with the initial consonant test, the 200 trials (84×2+16×2) are divided equally into groups of 4 SNR levels, i.e., corrupted by the 4 SNR levels of AWG noise in the same fashion as the initial consonants. With regard to distributions of the rhyming words, subjects' performance per SNR level is equally distributed producing 42 trials/SNR level referred to as Groups A, B, C, and D, respectively as shown in Table 1. Each group of 42 trials is equally distributed across 8 phonemes resulting in 5 trials/SNR level/phoneme plus a remainder of 2 trials. In total, there are 8 remaining trials (2 remaining trials/SNR level×4 SNR levels), each of which corresponding to one of the 8 phonemes. Finally, the remaining 8 phonemes are distributed across 4 SNR levels as shown in Table 1 (referred from bottom header of the table).

5. EXPERIMENTAL RESULTS

Percent intelligibility scores for initial and final consonants across 4 SNR levels shown in Table 2 are calculated by P_e stated earlier in Section 4. In agreement with findings of Miller and Nicely [7], the outcome from Table 2 suggests that the initial consonants were better perceived than the final consonants except at the SNR level of –24dB, where P_e is well below 50% and the score could be indistinguishable from guesswork [6]. Additionally, balanced confusion matrices at all SNR levels are obtained from the test responses of initial and final consonants[3]. Preliminary analysis across 3 SNR levels (–6, –12, and –18dB) according to segment type and phonological feature [4] shows that on average /r/ is the most confusable initial consonant and it was mostly misperceived as /d/, which shares voicing and coronal features. On the other hand, /w/ is the least confusable consonant in both initial and final positions. For final consonants, /k/ is the most confusable consonant and it was mostly misperceived as /t/, which is also a voiceless non-continuant. Interestingly, at the –18dB level, for both initial and final consonants, voicing was the most robust contrast while place-of-articulation was the least.

Table 2: Average percent intelligibility for initial and final consonants.

Consonant	SNR (dB)			
	–6dB	–12dB	–18dB	–24dB
Initial	93.06%	87.14%	77.35%	24.08%
Final	91.67%	84.01%	67.35%	27.21%

6. DISCUSSION AND FUTURE WORK

We have developed the subjective intelligibility testing of Thai speech and systematically compared confusion responses across all phonemes both for initial and final consonants. The confusion matrices not only show a pattern of correct responses but also that of misperceptions. Investigation of listeners' misidentified responses reveals that in initial position across the –6, –12, and –18dB levels, the listeners favored /t/ and /t^h/ and disfavored /w/ over other consonants. One interpretation is to connect these biases to the frequency of phoneme occurrences found in a Thai BEST corpus [3], constructed from various types of written materials. From the data of approximately 9 million words, among all initial consonants including clusters, /t^h/ occurs at the highest rate whereas /w/ is among consonants of lowest occurrence, which include /tɕ^h/, /h/, /ʔ/, /b/,/ŋ/, and /f/ [3]. We are working on the full analysis of confusions and developing subjective intelligibility tests of Thai vowels and tones.

7. REFERENCES

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

[2] House, A.S., Williams, C.E., Hecker, H.M.L., Kryter, K.D. 1965. Articulation-testing methods: Consonantal differentiation with a closed-response set. *J. Acoust. Soc. Am.* 37, 158-166.

[3] Human Language Technology Laboratory, BEST. *http://www.hlt.nectec.or.th/best/*

[4] de Lacy, P. 2007. *Segmental Features.* Cambridge: Cambridge University Press, 311-334.

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

[6] McLoughlin, I. 2008. Subjective intelligibility testing of Chinese speech. *IEEE Trans. Audio Speech Lang. Process.* 16, 23-33.

[7] Miller, G.A., Nicely, P.E. 1955. An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.* 27, 338-352.

[8] Voiers, W.D. 1983. Evaluating processed speech using the diagnostic rhyme test. *Speech Technol.* 1, 30-39.

[1]Complete list at *http://charturong.ece.engr.tu.ac.th/ICPhS2011/Initials.pdf*.
[2]Complete list at *http://charturong.ece.engr.tu.ac.th/ICPhS2011/Finals.pdf*.
[3]Available at *http://charturong.ece.engr.tu.ac.th/ICPhS2011/Confusions.pdf.*

Acoustical Society of America
The Journal of the Acoustical Society of America

Perceptual deficits in Thai sensorineural hearing loss patients

C. Tantibundhit[1], C. Onsuwan[1], N. Rueangwit[1], N. Saimai[1], T. Saimai[1], N. Thatphithakkul[2], K. Kosawat[2], S. Thatphithakkul[2] and P. Chootrakool[2]

J. Acoust. Soc. Am. 130, 2449 (2011); http://dx.doi.org/10.1121/1.3654835

Abstract

This study explored differences in CVV perception in two groups of Thai listeners: with normal hearing and with sensorineural hearing loss (with/without hearing aids). All participants chose one response in each of 210 Thai stimulus rhyming pairs, e.g., /taa/-/naa/. The rhyming monosyllabic words share an /aa/ vowel and mid tone, but differ in their initial phonemes (symmetrically distributed across 21 phonemes). While all stimuli for the normal hearing group were embedded in 4 signal-to-noise ratio levels, clean stimuli were presented to the patients. Comparisons of confusion patterns and perceptual distance were made. In both groups, /r/ is the most confusable phoneme, while /w/ is among the least. Perceptual representations of initial phonemes show five individual clusters: glide, glottal constriction, nasality, aspirated obstruent, and a combination of liquid and unaspirated obstruent. Patients' perceptual difficulty could be attributed to the nasality grouping, which is normally well separated, shifting closer to the glottal constrictions and aspirated obstruents. Hearing aids seem to improve perception of all phonemes by 10%, with /kh/ and /h/ showing the highest improvement rate, and /d/ the lowest. The instruments are beneficial in moving the nasality cluster further away from the nearby groupings.

DOI: http://dx.doi.org/10.1121/1.3654835

Key Topics

Phonetic segments
Deafness
Hearing
Hearing aids

Most read this month

Evaluation of smartphone sound measurement applications[a)]

Chucri A. Kardous and Peter B. Shaw

Coffee roasting acoustics

Preston S. Wilson

Application of subharmonics for active sound design of electric vehicles

Doo Young Gwak, Kiseop Yoon, Yeolwan Seong and Soogab Lee

Most cited this month

Transformed Up-Down Methods in Psychoacoustics

H. Levitt

Theory of Propagation of Elastic Waves in a Fluid-Saturated Porous Solid. II. Higher Frequency Range

M. A. Biot

Stimulated acoustic emissions from within the human auditory system

D. T. Kemp

Methodological Issues in Assessing Perceptual Representation of Consonant Sounds in Thai

C. Tantibundhit[1], C. Onsuwan[2], P. Phienphanich[1], C. Wutiwiwatchai[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 evaluate different experimental methods, ABX vs. AXB, and the use of reaction time (RT) measurement in assessing perceptual sensitivity to phonemic similarity based on perceptual representation of Thai initial consonants [1]. Thirty phoneme pairs are selected to represent varying degrees of similarity: highly similar, moderately similar, and clearly distinct. All the phoneme pairs are presented in noise in ABX and AXB tasks to twenty-two normal hearing Thai listeners. Order of the two tasks is counter-balanced across listener groups. Percent correct responses *(p*(C)), RTs, and preference rating are collected. The findings show that, *p*(C) is significantly higher in AXB than ABX despite no significant difference in RT values. In both ABX and AXB, listeners' *p*(C) across 3 levels of similarity varies significantly with the highest score in the clearly distinct group, and lowest score in the highly similar group. RT values across the 3 levels follow similar patterns but are not always statistically significant. ABX and AXB tasks could systematically be used to assess perceptual representation of speech sounds, with AXB eliciting higher *p*(C) and preference rating. It is suggested that some irregular patterns found in one part of the RT data may reflect some perceptual sensitivity pertaining to perceptual phoneme-cluster boundary.

Index Terms: Thai, initial consonant, perceptual similarity/distance, AXB, ABX, reaction time

1. Introduction

Experimental designs for speech perception research usually involve various forms of identification and/or discrimination tasks [2]. Identification tasks allow experimenters to explore the ways in which listeners assign and categorize sound stimuli. On the other hand, in discrimination tasks, degrees of perceptual difficulty and decision process can be addressed and compared. It is generally assumed that listerners' discrimination ability often exceeds their ability to identify [3], [4].

In the commonly known ABX discrimination, a series of three stimuli is presented, where listeners choose which stimulus, A or B, is most similar to or 'match' the stimulus X [2]. One of many advantages of this task is that it requires a straightforward and simple explanation to participants. However, it has been discussed that a long separation of A from X in ABX may have caused more demands on working memory [2], [5]. Different versions of this task, one of which is AXB [2], [6], have been proposed to avoid this problem. In AXB, the stimulus X is presented between the comparison stimuli, A and B. Similar to ABX, measures of accuracy and reaction could be collected in AXB.

Reaction time (RT) measurement has been used in speech perception research, but perhaps not as widely as it could have been. Reaction time can serve to assess listeners' sensitivity to within-category and across-category differences [2], [3], [7]. As Schneider *et al*. [7] have simply put it, a simple task such as discrimination of stimuli from different categories generally yields short RTs, while a more complex task, such as discrimination of stimuli lying between two categories requires more RTs [7].

Of interest here, we are in the process of developing a perceptual method to assess and evaluate listeners' discrimination ability in varying conditions, including those with hearing deficits. In fact, certain predictions can be made with regards to the previously proposed perceptual representation of Thai initial phonemes [1]. Our focus, at this moment, is to verify them using the well-known ABX and AXB paradigms along with reaction time measures. To the best of our knowledge, it seems that this issue has never been directly addressed.

The organization of the paper is as follows: Section 2 reviews the Thai phonology and the previously proposed perceptual representation of Thai initial phonemes. Section 3 provides details of the experimental design and setup. Section 4 presents experimental results. Section 5 discusses the findings and future work.

2. Thai

2.1. Thai Phonology

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/. Each of the nine monophthongs in Thai occurs phonemically short or long: /i/, /iː/, /e/, /eː/, /ɛ/, /ɛː/, /ɯ/, /ɯː/, /ɤ/, /ɤː/, /a/, /aː/, /u/, /uː/, /o/, /oː/, /ɔ/, and /ɔː/ [8].

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 ˉ, Low ˋ, 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 $C_i(C)V{:}^T(C_f)$, where C_i stands for an initial consonant, C_iC a consonantal cluster, C_f a final consonant, V a short vowel, V: a long vowel, and T a tone [9].

2.2. Perceptual Representation of Thai initial Phonemes

In our previous perceptual investigation [10], we developed and proposed a method for subjective intelligibility testing (identification) of Thai initial and final consonants, Thai diagnostic rhyme test (TDRT). Later, in [1], we provided an approximation of perceptual representation of Thai initial phonemes from listeners' responses (confusion matrix) of initial phoneme identification in noise.

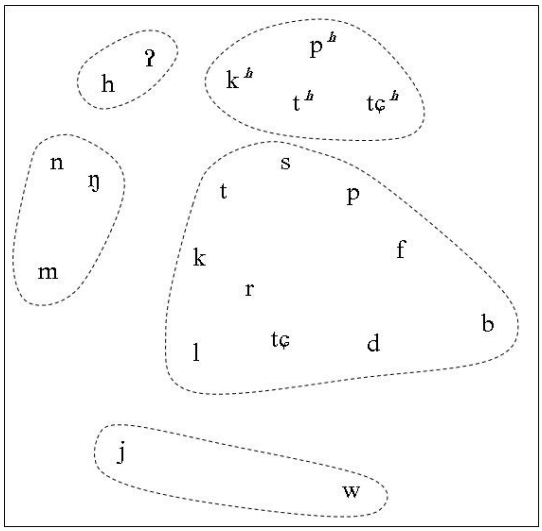

Figure 1: Perceptual space of 21 initial phonemes in Thai from Fig. 1 of [1].

More specifically, based on the perceptual space shown in Fig. 1, it is concluded that there are 5 clusterings of Thai initial consonant sounds, 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/) [1]. From those groupings and relative perceptual space/distance that separate the phonemes, we predict that consonant sounds within the same cluster, with small perceptual distance between them should be the ones that are hardest to discriminate, followed by consonants from the same cluster that are separated by relatively large distance, or consonants in different clusters that are separated by small distance. Lastly, consonants from 2 different clusters that are separated by relatively large distance should be easily discriminated.

To verify our predictions of decision difficulty, we carry out two psychophysical tasks, ABX and AXB, respectively.

3. Experimental Design and Setup

This section describes the experimental design and procedure of two psychophysical tasks, namely ABX and AXB for Thai initial consonant sounds. Details are given as follows:

3.1. Test Stimuli

Table 1: Three types of Thai initial consonant pairs (30 pairs).

I **Highly similar**	II **Moderately similar**	III **Clearly distinct**
l-r, m-ŋ, d-b, n-ŋ, k^h-p^h, s-p, r-tɛ, t-k, t^h-$tɕ^h$, h-ʔ	f-l, j-l, k-b, ŋ-k, p-l, s-t^h, $tɕ^h$-p, h-n, n-m, k^h-$tɕ^h$	f-m, h-d, j-p^h, j-w, k^h-w, n-b, p-p^h, s-ŋ, t-h, ʔ-tɛ

From one of the experiments in [1], [10], where there were two hundred and ten word pairs differing in initial consonant sounds, thirty pairs are chosen for this study. The aim is to make the test considerably short and does not cause subject's fatigue. These 30 pairs represent 3 levels of perceptual similarity: Group I-Highly similar, Group II-Moderately similar, and Group III-Clearly distinct, respectively, as shown in Table 1.

Based on Fig. 1, each of the 10 pairs in Group I is chosen from consonant sounds in the same cluster, with relatively small perceptual distance (most confusable), while each of the 10 pairs in Group II is chosen from 1) either consonants within the same cluster that are separated by relatively large distance, i.e., [k-b], [k^h-$tɕ^h$], [n-m], [p-l], [f-l], or 2) from consonants in different clusters but are separated by small distance, i.e., [j-l], [ŋ-k], [s-t^h], [$tɕ^h$-p], [h-n]. Finally, each of the 10 pairs in Group III is chosen from consonants belonging to different clusters that are separated by relatively large distance. The exception is for the consonant pair [j-w], which came from the same cluster, i.e., its perceptual distance is equal to infinity [1]. Table 1 summarizes the three types of initial consonant pairs.

The 30 rhyming word pairs, containing the target consonants, of the form [Cā:] (with identical tone), 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. During the test trials, the selected tokens are corrupted by additive white Gaussian (AWG) noise –12dB. This SNR level is chosen based on a preliminary experiment tested at various SNR levels, i.e., –6dB, –12dB, –18dB, and –24dB, respectively [1], [10]. The subjects' correct responses were near ceiling at –6dB but indistinguishable from guesswork at –18dB and –24dB.

The test consists of 2 tasks, ABX and AXB. Each task has 30 target trials (from the 30 rhyming word pairs) plus 10 trials of filler word pairs. In the trials of ABX task, three stimuli are presented in order, i.e., A, B, and X, respectively, where A and B are referred to as the standards, while X is the focus. On the other hand, in the AXB task, three stimuli A, X, and B are presented in order. In both tasks, the listeners have to identify whether the focus X they just hear is identical or most similar to A or B by pressing the button A or B appearing on the computer screen. If they do not recognize the stimulus, they are instructed to guess before moving on to the next trial.

To avoid B recency effect, among the 10 pairs in each group, a number of focuses are equally divided between the standards, i.e., five focuses (X) are identical to A and five to B. For example, /lā:/ ลา - /rā:/ รา - /lā:/ ลา is AB(X=A) and /nā:/ นา - /ŋā:/ งา - /ŋā:/ งา is AB(X=B), while /k^hā:/ คา - /k^hā:/ คา - /$tɕ^h$ā:/ ชา is A(X=A)B and /t^hā:/ ทา - /sā:/ ซา - /sā:/ ซา is A(X=B)B, respectively.

Figure 2 illustrates the stimulus condition for ABX and AXB. As shown, in each stimulus, the inner square refers to speech segment (A, B, or X); the outer part includes noise and tapering noise portions. Duration of the speech segment ranges from 265 to 430 msec. After a response is logged, there is a 3-second gap before the new trial begins.

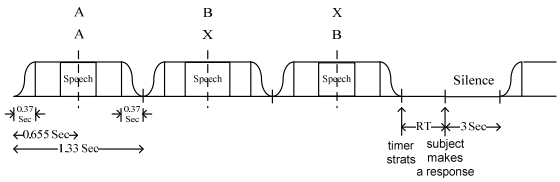

Figure 2: Description of the stimulus condition for ABX and AXB.

3.2. Procedure

Two psychophysical tasks, namely ABX and AXB, are conducted on 22 untrained Thai volunteer subjects (6 female and 16 male) with normal hearing over headphones in a quiet room. The subjects are equally divided into two groups, i.e., Groups A and B. The subjects in Group A perform ABX task first and then AXB task and vice versa for the subjects in Group B. Ordering of individual trials is randomized every time for each subject. After the instruction is given, each task begins with a 5-trial practice session without feedback. There is a short 5-minute break after the first task is completed. All tasks are completed within 30 minutes.

Percent correct responses *(p*(C)), reaction time (RT), and preference rating are collected. Paired difference t-test with 95% CI is performed on the results.

3.3. Reaction Time Measurement

RTs are collected for all responses made during the tasks. RTs record the timing to the initiation of the response from the ending of stimulus presentation (see Fig. 2). Only RT values to the correct responses are analyzed [3]. As some RT values are longer than 10 seconds, we decide to follow an approach taken by [7] and eliminate RT values that are more than two standard deviations (S.D.) away from the mean. This step eliminates approximately 3% of the correct response data.

4. Experimental Results

4.1. Percent Correct Responses and Reaction Time

Table 2: *Average percent correct responses and reaction time.*

	***p*(C)**		**RT (msec)**	
	Average	**S.D.**	**Average**	**S.D.**
ABX	88.2	9.1	691..4	173.5
AXB	94.6	4.8	759.8	251.0

Table 2 presents average percent correct responses (*p*(C)) and reaction time (RT) of the correct responses. The results show that listeners performed on average significantly better in AXB task than ABX task (94.6% vs. 88.2%) [t(21) = -3.1122, $p = 0.0053$]. Moreover, AXB task gives lower S.D. than ABX (4.8 vs. 9.1). Although average RT values for AXB is higher than those for ABX, the difference is not statistically significant [t(21) = -1.0521, $p = 0.2996$].

Out of 22 participants, six are female and they respond slower (lower RTs) than males (818.7 msec vs. 684.9 msec).

4.2. Preference Rating

After the test is completed, all participants are asked which of the two tasks they prefer. Sixteen persons (73%) prefer AXB over ABX.

4.3. Pairwise Discrimination

Table 3 presents percent correct responses (*p*(C)) and reaction time (RT) from correct responses for each phoneme pair in 3 groups. For ABX, listeners' *p*(C) across 3 levels of similarity varies significantly with the highest score in Group III (96.8%), followed by Group II (89.1%), and lowest score in Group I (78.6%). Paired t-tests confirm that (*p*(C)) differences across Groups I-II, II-III, and I-III (each group is composed of 10 pairs of phoneme) are statistically significant; for Groups I-II [t(21) = -3.6965, $p = 0.0013$]; Groups II-III [t(21) = -3.5521, $p = 0.0019$], and Groups I-III [t(21) = -7.2231, $p = 0.0000$]. RT differences across Groups I-II, II-III, and I-III vary with the highest score in Group I (824.3 msec), followed by Group II (647.8 msec), and lowest score in Group III (618.7 msec). Paired t-tests confirm significant differences for Groups I-II [t(9) = 2.8342, $p = 0.0196$] and Groups I-III [t(9) = 3.3069, $p = 0.0091$], but not for Groups II-III.

Table 3: *Percent correct responses (p(C)) and reaction time (RT) from correct responses for each initial consonant pair.*

Type	**Pair**	**ABX**		**AXB**	
		***p*(C)**	**RT (msec)**	***p*(C)**	**RT (msec)**
I Highly similar	l-r	86.4	726.6	100.0	517.8
	m-ŋ	72.7	586.9	100.0	510.6
	d-b	90.9	697.8	95.5	854.3
	n-ŋ	72.7	699.1	95.5	992.5
	k^h-p^h	72.7	895.1	90.9	1154.7
	s-p	95.5	759.3	90.9	1036.2
	r-tɕ	95.5	879.8	86.4	1368.5
	t-k	54.6	976.8	77.3	973.2
	t^h-$tɕ^h$	95.5	738.9	77.3	1164.6
	h-ʔ	50.0	1282.7	63.6	775.7
Average		**78.6**	**824.3**	**87.7**	**934.8**
II Moderately similar	j-l	100.0	518.6	100.0	576.0
	ŋ-k	95.5	710.5	100.0	859.5
	s-t^h	90.9	750.0	100.0	529.7
	$tɕ^h$-p	63.6	675.2	90.9	1030.4
	h-n	63.6	690.7	90.9	842.3
	k-b	95.5	581.3	100.0	668.4
	k^h-$tɕ^h$	100.0	630.0	86.4	611.0
	n-m	95.5	619.1	90.9	798.3
	p-l	100.0	619.2	100.0	805.0
	f-l	86.4	683.3	100.0	575.8
Average		**89.1**	**647.8**	**95.9**	**729.6**
III Clearly distinct	f-m	90.9	609.4	100.0	598.7
	h-d	100.0	633.1	100.0	817.9
	j-p^h	95.5	585.6	100.0	554.2
	j-w	100.0	496.4	100.0	597.4
	k^h-w	100.0	780.5	100.0	544.1
	n-b	100.0	524.5	100.0	911.3
	p-p^h	86.4	567.4	100.0	793.5
	s-ŋ	95.5	661.3	100.0	590.3
	t-h	100.0	705.5	100.0	580.2
	ʔ-tɕ	100.0	623.6	100.0	538.4
Average		**96.8**	**618.7**	**100.0**	**652.6**

Likewise, for AXB, listeners' *p*(C) across 3 levels of similarity varies significantly with the highest score in Group III (100%), followed by Group II (95.9%), and lowest score in Group I (87.7%). Paired t-tests confirm that (*p*(C)) differences across Groups I-II, II-III, and I-III are statistically significant; for Groups I-II [t(21) = -4.5, $p = 0.0002$]; Groups II-III [t(21) = -2.8801, $p = 0.0086$], and Groups I-III [t(21) = -5.9188, $p = 0.0000$]. RT differences across Groups I-II, II-III, and I-III vary with the highest score in Group I (934.8 msec), followed by Group II (729.6 msec), and lowest score in Group III (652.6 msec). Paired t-tests confirm significant differences for Groups I-II [t(9) = 3.4931, $p = 0.0068$] and Groups I-III [t(9) = 2.9671, $p = 0.0158$], but not for Groups II-III.

4.4. Analysis of Group II-Moderately Similar

As stated earlier in Section 3, ten pairs of consonants in Group II can be divided further in two subgroups:

Type 1: consonants from different clusters that are separated by relatively small distance, i.e., [j-l], [ŋ-k], [s-t^h], [$tɕ^h$-p], [h-n] (shown as light grey in Table 3).

Type 2: consonants from the same cluster that are separated by relatively large distance, i.e., [k-b], [k^h-$tɕ^h$], [n-m], [p-l], [f-l] (shown as dark grey in Table 3).

Figure 3 illustrates percent correct responses from Type 1 (left series) and Type 2 (right series), and Fig. 4 the

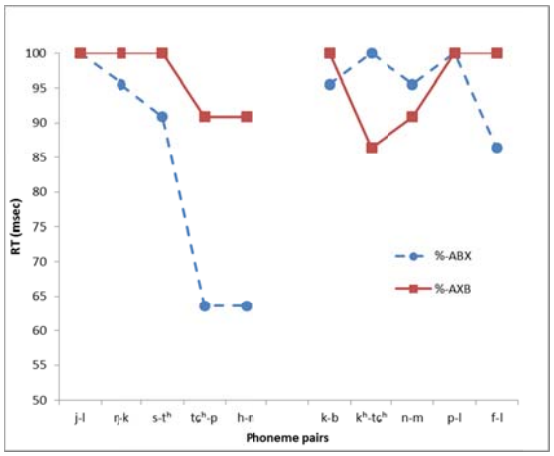

Figure 3: Percent correct responses from Group II consonant pairs; Type 1: consonants from different clusters but separated by small distance (left series) and Type 2: consonants from the same cluster but separated by relatively large distance (right series).

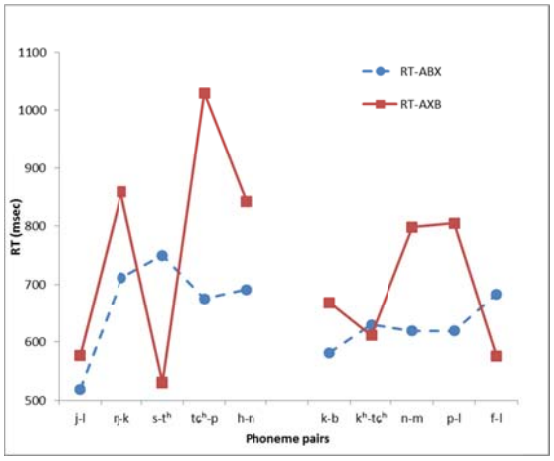

Figure 4: Reaction time values from Group II consonant pairs; Type 1: consonants from different clusters but separated by small distance (left series) and Type 2: consonants from the same cluster but separated by relatively large distance (right series).

corresponding RT values. It can be seen from Fig. 3 and Fig. 4 that percent correct responses as well as RTs among the Type-1 pairs show more fluctuations (wider range) than the Type-2 pairs. This pattern is particularly clear for RTs of Type-1 pairs (Fig. 4, left series). Across ABX and AXB, average *p*(C)) for Type 1 and Type 2 are 89.54% and 95.47%, and average RTs are 718.29 msec and 659.14 msec, respectively.

5. Discussion and Future Work

Overall, based on the perceptual representation, our predictions for decision difficulty are borne out. Consonant sounds within the same cluster, with small perceptual distance between them (Group I) are the hardest to discriminate, with the lowest *p*(C) and the longest RTs in the ABX and AXB tasks. The intermediate level refers to consonants from the same cluster that are separated by relatively large distance, or consonants in different clusters but are separated by small distance (Group II). Lastly, the last group (Group III), consonants from different clusters (except for [j-w]) that are separated by the relatively large distance, is the easiest to discriminate, with the highest *p*(C) and the shortest RTs in the ABX and AXB tasks.

The findings suggest that both ABX and AXB tasks could systematically be used to assess perceptual representation of speech sounds, with AXB eliciting higher *p*(C) and preference rating. The higher *p*(C) in the AXB task may be attributed to the relatively longer exposure to the focus (X) stimulus and fewer demands on working memory [2], [6]. With additional steps to exclude outliers from RT values, RT measurement is also a reliable tool.

Interestingly, RTs of Group II-Type 1 pairs (Fig. 4, left series) show more irregularities than those of Group II-Type 2 pairs (Fig. 4, right series). This suggests that between these two types, a different kind of perceptual process/sensitivity is taking place. The across-cluster difference may exert more fluctuation patterns in RTs than the difference in terms of perceptual distance.

Currently, we are developing an adaptive testing method (using real speech) to detect hearing deficits. More specifically, we plan to incorporate our developed Thai rhyming words and AXB task and to come up with the method that not only could suggest whether a person has hearing problems but also identify his/her degrees of hearing difficulty for certain groups of speech sound.

6. Acknowledgements

The authors would like to thank NECTEC for the support of this project.

7. References

[1] Tantibundhit, C., Onsuwan, C., Saimai, T., Saimai, N., Thatphithakkul, S., Chootrakool, P., Kosawat, K., and Thatphithakkul, N., "Perceptual representation of consonant sounds in Thai," in Proc. Interspeech, 2011, pp. 3193–3196.

[2] Beddor, P. S. and Gottfried, T. L., "Methodological issues in cross-language speech perception research with adults," in W. Strange [Ed], Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research, 207–232, Timonium, MD: York Press, 1995.

[3] Pisoni, D. B. and Tash, J., "Reaction times to comparisons within and across phonetic categories," Perception and Psychophysics, 15: 285–290, 1974.

[4] Macmillan, N. A., Kaplan, H. L., and Creelman, C. D., "The psychophysics of categorical perception," Psychological Review, 84(3): 452-471, 1977.

[5] Wayland, R. and Guion, S., "Perceptual discrimination of Thai tones by naive and experienced learners of Thai," Applied Psycholinguistics, 24: 113–129, 2003.

[6] Harnsberger, J. D., "On the relationship between identification and discrimination of non-native nasal consonants," J. Acoust. Soc. Am., 110(1): 489–503, 2001.

[7] Schneider, K., Dogil, G., and Möbius, B., "Reaction time and decision difficulty in the perception of intonation," in Proc. Interspeech, 2011, pp. 2221–2224.

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

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

[10] Tantibundhit, C., Onsuwan, C., Thatphithakkul, S., Chootrakool, P., Kosawat, K., Thatphithakkul, N., Saimai, T., and Saimai, N., "Subjective intelligibility testing and perceptual study of Thai initial and final consonants," in Proc. ICPhS, 2011, pp. 1970–1973.