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.*