MODELING PREDICTIVE PERCEPTUAL REPRESENTATION OF THAI INITIAL CONSONANTS

P. Phienphanich¹, C. Onsuwan²,³, C. Tantibundhit¹,³, N. Saimai¹, T. Saimai¹

¹ Department of Electrical and Computer Engineering, Thammasat University, Thailand
² Department of Linguistics, Thammasat University, Thailand
³ Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), Thammasat University, Thailand

tchartur@engr.tu.ac.th, consuwan@tu.ac.th

ABSTRACT

This work is an extension of our previous attempt to construct a spatial representation of 21 initial consonants in Thai by partitioning them into homogeneous clusters based on empirical measures of confusability and distance among phonemes. The measures were taken from perceptual identification performance of 28 listeners (seven full subjects) when stimuli were presented in noise. In present study, two methods of clustering, namely Multidimensional scaling analysis and k-means clustering were employed, yielding six different classifications and four perceptually relevant categories: intra-cluster short distance, intra-cluster long distance, inter-cluster short distance, and inter-cluster long distance. Another set of perceptual experiment (eight listeners; two full subjects) was carried out to verify the predictions. The findings reveal that the derived perceptual clusters and defined categories fit relatively well with the listeners’ performance. Distinctive feature systems in phonological theory appear to provide some basis for the clustering of phonemes.

Index Terms—Thai, initial consonant, confusion matrix, distance/similarity, perceptual representation/space

1. INTRODUCTION

Insights into perceptual classifications of phonemes (vowels and consonants) provide a possible way to understand how speech sounds are actually represented and heard by a listener (rather than being explained in terms of how they are articulated) [1]. The informative way of discovering underlying perceptual structure and relationship among phonemes is crucial for a number of areas in speech and hearing research, including speech recognition [1]. Perceptual spaces or representations could be constructed from empirical measures of perceptual similarity or distance deriving from perceptual tests such as classification tests [2] and identification tests [3].

A number of well-known studies examined confusion patterns of English consonants [4][5] and vowels [3][5]. In a classic work by Shepard (1972), based on the data presented in [3][4], he methodically proposed and derived perceptual representations of English consonants and vowels [6].

To the best of our knowledge, apart from our previous work [7], no attempt has been done to construct any kind of perceptual representation of Thai phonemes. The work presented here is an extension of our previous attempt [7] [8] to construct a spatial representation of Thai initial consonants. Our goal is to apply the findings to our ongoing research in development of a testing method for assessing hearing sensitivity and deficits. Summary of our previous work and the findings are given below.

1.1. Perceptual representation of consonant sounds in Thai [7]

In Tantibundhit et al. (2011), five phoneme clusters were proposed. It was a two-dimensional approximation derived from listeners’ responses (28 listeners; seven full subjects) of initial phoneme identification at the SNR level of −18dB, using a method of Thai Diagnostic Rhyme Test (TDRT), which we developed in [9]. In order to obtain similarity scores, and derive a distance matrix, a confusion matrix was first constructed. In fact, similarity score between each pair of phonemes was calculated from confusion scores based on Shepard’s method [8], i.e., \( S_y = (P_{iy} + P_{jy}) / (P_{ii} + P_{jj}) \), where \( S_y \) 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 \( (d_y) \) was derived from the similarity score, i.e., \( d_y = -\ln(S_y) \) [8]. Finally, a distance matrix was constructed from the calculated distances.

From the calculated distances, a perceptual space for each Thai initial phoneme was sketched in Fig. 1 [7]. Each axis represents perceptual similarity on some undefined, but relevant dimension. Five possible clustering of phonemes,
dashed-line circles, were later drawn on a basis of their common phonological features [10] [11], i.e., glide (/u/ and /l/), glottal constriction (/p/ and /h/), nasality (/n/ and /n/), aspirated obstruent (/pʰ/), /tʰ/, /çʰ/, and /kʰ/), and a combination of liquid and unaspirated obstruent (/p/, /h/, /t/, /l/, /d/, /l/, /ß/, and /r/). Interestingly, /r/, the most confusable phoneme, is located in the middle of the perceptual space and towards the center of its own group.

1.2. Methodological issues in assessing perceptual representation of consonant sounds in Thai [8]

In Tantibundhit et al. (2012), two experimental methods, ABX vs. AXB were evaluated to assess perceptual sensitivity to phonemic similarity based on perceptual representation of Thai initial consonants [7]. Percent correct responses and reaction times (RT) were collected from the ABX and AXB tasks presented in noise [8].

Thirty phoneme pairs were selected to represent three degrees of similarity [7]: (1) highly similar (consonants within the same cluster, with relatively small perceptual distance), e.g. /l/-l, n-ŋ, d-b/ (2) moderately similar (consonants within the same cluster that are separated by large distance, e.g. /k-/b, n-/ñ; or consonants from different clusters but are separated by small distance), e.g. /j/-l, ŋ-k/l and (3) clearly distinct (consonants from different clusters that are separated by large distance), e.g. /½-/m, h-d, J-9/ [8].

From those groupings, it was predicted that consonants from the highly similar group should be the hardest to discriminate, followed by the moderately similar and clearly distinct groups. Interestingly, the findings agreed with the earlier predictions. In both ABX and AXB, listeners’ percent correct responses across three levels of similarity varied significantly with the highest score in the clearly distinct group, and lowest score in the highly similar group.

However, RTs among the moderately similar group showed some irregularities, suggesting that different kinds of perceptual sensitivity or process were involved for across-cluster differences and distance differences.

1.3. Present study

This work presents details of how we employed two clustering methods: Multidimensional scaling analysis and k-means clustering to derive Thai initial phoneme classifications and perceptually relevant predictions. The organization of the paper is as follows: Sections 2 and 3 provide experimental setup, including a construction of model and predictions, and experimental results. Section 4 discusses the results and mentions future work.

2. EXPERIMENTAL SETUP

In this study, two sets of perceptual experiments using Thai Diagnostic Rhyme Test (TDRT) [7] are presented and referred to as Tests 1 and 2. Test 1 was taken from [7] as a baseline for model construction (28 participants equivalent to seven full subjects). Using a smaller set of listeners (eight participants equivalent to two full subjects), Test 2 was carried out to verify model predictions.

2.1. Test 1: test stimuli and procedure (Tantibundhit et al., 2011 [7])

2.1.1. Test stimuli

The 21 rhyming words differing only in initial consonants along with filler words were read 5 times in a carrier sentence 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 [7]. Then,
one of the 5 tokens of rhyming word was selected based on impressionistic hearing evaluation and spectrographic inspection [7].

Then, the selected tokens were corrupted by additive white Gaussian (AWG) noise with 4 SNR levels, i.e., −6, −12, −18, and −24dB [7]. In each trial, listeners hear a target stimulus and are asked to choose what they just hear between two 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 [7].

2.1.2. Test procedure

TDRT for initial phonemes consists of 210 rhyming pairs \(\binom{21}{2}C\) 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 rhyming stimuli and 80 trials for filler words. Rather than giving a very long test of 500 trials × 4 SNR levels which would create a test of 2,000 trials, we increased a number of subjects four times and divided the test equally by 4 SNR levels. Then, the 500 trials are corrupted by one of 4 SNR levels of AWG noise. 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.

2.2. Test 2: test stimuli and procedure

Test stimuli and procedure for Test 2 were conducted in the same fashion as Test 1 using eight Thai participants equivalent to two full subjects.

2.3. Derived clusters and predicted perceptual categories

Clusters and perceptual categories are generated as follows:

**Step 1:** Distance matrix calculation

Confusion matrices of Test 1 (Table 3 of [7]) and Test 2 (shown in Table 3) were constructed from the data at SNR level of −18dB. Similarity score between each pair of phonemes is calculated from confusion score. Then, perceptual distance \(d_{ij} = -\ln(S_{ij})\) is derived from the similarity score \(S_{ij} = \frac{(P_u + P_d)}{(P_u + P_d)} \). Finally, a distance matrix of Test 1 (Table 3 of [7]) and Test 2 (shown in Table 3) are constructed from the calculated distances.

**Step 2:** Phoneme clustering

A multidimensional scaling algorithm (MDS) was used to represent the information contained in a distance matrix, i.e., place each phoneme in \(N\)-dimensional space from experimental data, \(N = 11\) is the dimension of the smallest space in which all phoneme can be embedded) such that the between-object distances are preserved as much as possible [13]. Specifically, each phoneme is represented by a point in a multidimensional space and they are arranged in this space so that the distances between pairs of phonemes reflect the strongest possible relation to the similarities among the pairs of phonemes. Specifically, two similar phonemes (short distance) are represented by two points that are close together, while two different phonemes (long distance) are represented by two points that are far apart. The space is \(N\)-dimensional Euclidean space [13].

Then, the \(k\)-means clustering was used to partition 21 phonemes into six clusters in which each observation belongs to the cluster with the nearest mean [14]. Finally, the derived perceptual representation of 21 Thai initial consonants is shown in Fig. 2.

\[
d = |d_{ij}|_{i \neq n} \rightarrow \text{MDS} \rightarrow \text{Set of phoneme vectors in space} \rightarrow \text{k-means} \rightarrow \{X_i\}, 1 \leq i \leq n, k \leq n
\]

where \(\{X_i\}\) is a set of phoneme clusters.

**Table 1: Intra-, Inter-cluster thresholds.**

\[
\begin{array}{cccccc}
\text{Cluster} & 1^\text{st} & 2^\text{nd} & 3^\text{rd} & 4^\text{th} & 5^\text{th} & 6^\text{th} \\
1^\text{st} & 3.6 & 6.6 & 6.2 & 6.6 & 6.8 & 7.4 \\
2^\text{nd} & 6.6 & 4.6 & 7.2 & 7.1 & 6.9 & 7.6 \\
3^\text{rd} & 6.2 & 7.2 & 3.8 & 5.1 & 6.8 & 7.0 \\
4^\text{th} & 6.6 & 7.1 & 5.1 & 4.3 & 5.1 & 6.8 \\
5^\text{th} & 6.8 & 6.9 & 6.8 & 5.1 & 4.4 & 6.5 \\
6^\text{th} & 7.4 & 7.6 & 7.0 & 6.8 & 6.5 & 5.6 \\
\end{array}
\]

**Step 3:** Categorizing \(T = \{i, j | 1 \leq i, j \leq n, i \neq j\}\)

Intra-cluster and inter-cluster thresholds of each cluster are calculated by the \(k\)-means clustering as shown in Table 1.

**Intra-cluster threshold**

\[
\{d_{ij} | \exists i, j \in X_l, l \neq j\} \rightarrow \text{k-means} \rightarrow \theta_{ij} = \mu, 1 \leq l \leq k \rightarrow t_{ij} = \begin{cases} 1, & d_{ij} \leq \theta_{ij} \\ 2, & d_{ij} > \theta_{ij} \end{cases}
\]

**Inter-cluster threshold**

\[
\{d_{ij} | \exists i, j \in X_l, m \neq n\} \rightarrow \text{k-means} \rightarrow \theta_{jm} = \mu, 1 \leq l \leq m \rightarrow t_{ij} = \begin{cases} 3, & d_{ij} \leq \theta_{jm} \\ 4, & d_{ij} > \theta_{jm} \end{cases}
\]
Table 2: Phoneme pairs in four categories.

<table>
<thead>
<tr>
<th>Category (Total number of phonemes)</th>
<th>Intra SD (19)</th>
<th>Intra LD (10)</th>
<th>Inter SD (85)</th>
<th>Inter LD (96)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p/ - /t/</td>
<td>/k/ - /k/</td>
<td>/d/ - /d/</td>
<td>/p/ - /b/</td>
<td>/p/ - /j/</td>
</tr>
<tr>
<td>/t/ - /k/</td>
<td>/d/ - /d/</td>
<td>/p/ - /b/</td>
<td>/p/ - /j/</td>
<td>/n/ - /n/</td>
</tr>
<tr>
<td>/d/ - /d/</td>
<td>/p/ - /b/</td>
<td>/p/ - /j/</td>
<td>/n/ - /n/</td>
<td>/n/ - /n/</td>
</tr>
<tr>
<td>/p/ - /j/</td>
<td>/n/ - /n/</td>
<td>/n/ - /n/</td>
<td>/n/ - /n/</td>
<td>/n/ - /n/</td>
</tr>
</tbody>
</table>

Table 3: Confusion matrix

3.2 Percent correct responses

Figure 3: Example of how to deal with ‘infinite’ distance value.

Following from Tantibundhit, et al. (2012) [8], and taking into account cluster relationships (intra- vs. inter-cluster) and distances (long vs. short), four perceptual relevant categories could be proposed: intra-cluster short distance (Intra SD), intra-cluster long distance (Intra LD), inter-cluster short distance (Inter SD), and inter-cluster long distance (Inter LD).

Step 4: Category defining

Finally, four categories are predefined as follows:
- \( t_{ij} = 1 \) is an intra-cluster short distance (Intra SD).
- \( t_{ij} = 2 \) is an intra-cluster long distance (Intra LD).
- \( t_{ij} = 3 \) is an inter-cluster short distance (Inter SD).
- \( t_{ij} = 4 \) is an inter-cluster long distance (Inter LD).

Perceptually, Intra SD phonemes should be highly similar and the easiest to confuse with one another, followed by Intra LD phoneme and Inter SD phonemes. Finally, Inter LD phonemes should be highly distinct from one another and the least confusable. Examples of phoneme pairs in the four perceptual categories are given in Table 2.

2.4. Treatment of ‘infinite’ distance value

When a distance between a pair of phonemes is approaching infinity (\( d_{ij} \rightarrow \infty \)), the phonemes \( i \) and \( j \) cannot be accurately illustrated in the space. Therefore, we propose to substitute an infinite distance value with a finite one such that all phonemes can be more effectively illustrated in the space. Specifically, a triangle inequality, which states that for any triangle, summation of the lengths of any two sides must be greater than or equal to the length of the remaining side [12], is manipulated. Then, any three phonemes (\( i, j, \) and \( m \)) in the space are considered as vertices of a triangle. Therefore, distance of each phoneme pair can be expressed as:

\[
\min(d_{im} + d_{mj})
\]

Finally, \( d_{ij} \) with infinite distance value is replaced with \( d'_{ij} = \min(d_{im} + d_{mj}) \), where \( m \neq i \neq j \).

3 EXPERIMENTAL RESULTS

3.1 Confusion matrix

To be in line with Test 1 [7], data from SNR level of −18dB was used to construct the confusion matrix, as shown in Table 3. From Table 3, /t/ and /l/ are the most confusable phonemes, while /j/ is the least confusable phoneme followed by /w/, /bl/, and /pl/. This overall pattern agrees with what was found for Test 1 [7].

3.2 Percent correct responses

Figure 4 (a) illustrates box plots of average percent correct responses of Test 1 vs. Test 2. From the plots, the two tests have comparable means (75.88% vs. 74.10%) and comparable medians (76.04% vs. 75.00%) with larger standard variation in Test 2 (2.76% vs. 4.47%). Two-sample t-test reveals that there is a significant difference between Test 1 and Test 2 \( t(32)=3.6038, p < 0.01 \).

It should be pointed out that within each test, two data sets were divided, one from all phonemes (Full Set) and the other from selected phonemes (Partial Set). Specifically, 40 pairs of phonemes were randomly chosen from 210 pairs of phonemes. It should be noted that the cross validation algorithm was applied to all possible cases resulting in 74 +
2^4 = 2,417 cross validations. From the plots shown in Fig. 4(b), Full Set and Partial Set have comparable means (75.89% vs. 75.83%), comparable medians (75.79% vs. 76.25%), and comparable standard deviations (3.95% vs. 4.92%). In all, there is no significant difference between Full Set and Partial Set [t(2416) = 1.0899, p = 0.2759].

Finally, average percent correct responses of four perceptual categories (shown in Fig. 4c): Intra SD, Intra LD, Inter SD, and Inter LD were compared, i.e., with means of 62.74%, 72.79%, 84.09%, and 99.47%, medians of 63.16%, 75.00%, 84.62%, and 100%, and standard deviations of 7.58%, 10.73%, 5.60%, and 1.78%. The results reveal a gradual increase of correct responses, moving from Intra SD to Inter LD. ANOVA shows that percent correct responses of the four categories are significantly different \( F(1,19332) = 663.94, p < 0.01 \). It should be noted that there are \( 7^4 \times 4 + 2^4 \times 4 = 19,336 \) possible cases.

4 DISCUSSIONS AND FUTURE WORK

Even though the five clusters from the two-dimensional representation of Tantibundhit, et al. (2011) [7] were classified on the basis of already-existing phonological features (i.e., glide, glottal constriction, nasality, aspirated obstruent, and a combination of liquid and unaspirated obstruent), they do not differ drastically from the six clusters derived from the present approach, which are data-driven. Specifically, the two approaches share a relatively tight cluster of /pʰ tʰ tʰ kʰ/, which seems to be very specific to the Thai sound system. It should be noted that /s/ may rather be considered as no initial consonant (or vocalic onset) following the findings of Harris (2001) which reported that there is no real glottal stop production in word-initial position in Thai [15].

There is a strong assumption that distinctive feature systems in phonological theory have definable articulatory and acoustic correlates [10] [11]. It appears that they could also provide some basis for the understanding of the six perceptual clusters. Adopting the distinctive feature systems used in Halle and Clements (1983) [11], one could propose some generalizations along this line:

- the 1st cluster could be called ‘spread glottis’.
- the 2nd cluster could be called ‘nasal sonorant’, excluding /h/ and /r/.
- the 3rd cluster could be called ‘glottal stop’.
-the 3rd cluster could be called ‘anterior’.
- the 4th cluster could be called ‘high’, excluding /al/.
- the 5th cluster could be called ‘(non-nasal) sonorant’.
- the 6th cluster could be called ‘labial’.

Test 2 was carried out in order to verify the generalizations made in Test 1 with regard to the six clusters and four perceptual categories. Overall, the results from Test 2 agree with those from Test 1, showing comparable means and medians in percent correct responses (Fig. 4 a)). However, there is a significant difference in percent correct responses between Test 1 and Test 2 which possibly due to smaller numbers of subjects with larger variance of percent correct responses.

Figure 4 (c) nicely shows a gradual increase of correct responses as we move from one category to the next, in the ascending order of Intra SD, Intra LD, Inter SD, and Inter LD. Moreover, percent correct responses among the four categories are significantly different, confirming our predictions that Intra SD phonemes should be the most difficult to differentiate and Inter LD the easiest. The fact that the Intra phonemes (Intra SD and Intra LD) are harder to differentiate (lower percent correct responses) than the Inter phonemes (Inter SD and Inter LD) clearly suggests that cluster relationship is possibly more important than distances when it comes to constructing a perceptual representation. This brings out the importance of discovering inherent/hidden structure among the phonemes.

The finding that no significant difference was found in the use of Full Sets versus Partial Sets could be very beneficial for the development of testing. It shows that within each well-defined perceptual category, it is not necessary to use every pair of phonemes to achieve desirable outcome.

We plan to continue with this current approach and construct perceptual representations of Thai vowels and lexical tones. The findings will be crucial to our ongoing development of a testing method for assessing hearing sensitivity and deficits.

6. REFERENCES