Cluster Analysis of Prominent Features for Determining Stress Levels in Thai Speech

Patavee Charnvivit,
Nuttakorn Thubthong, and **Sudaporn Luksaneeyanawin**, Non-members

ABSTRACT

Software testing is an important activity in software development process. Testers have to generate test cases to test a system. At least, test cases consist of test input values and expected results. In functional testing or black-box testing, test designers can generate test cases from a requirements specification document which includes diagrams such as UML Diagrams. In this research, we consider UML use case diagrams and propose an approach for generating test cases from use cases based on a limited entries decision table. These test cases cover all success and alternative scenarios in a use case as well as all events that contain include and extend relationship.

Keywords: Software Testing, Use cases, Test Cases, Decision table

1. INTRODUCTION

Stress is one of prosody features, which play an important role in many areas of speech technology. Applying the knowledge of the prominent features can improve naturalness of text-to-speech synthesis (TTS) system [1, 2], as well as recognition rate of automatic speech recognition (ASR) system [3–6].

Stress is referred as the relative perceptual prominence of a syllable in a word [7]. Listeners can perceive the different level of stresses in an utterance. The number of levels of stress appear to vary from language to language. Most studies classified the degree of stress into three levels [8–11], while some studies (including Thai studies) classified it into two levels [12–14].

The degree of stress is continuous. It can be represented by prominent features. A typical way of stress annotation is to digitized the degree of stress to several discrete levels such as heavy stress, normal stress and weak stress. It is difficult to find an optimized number of stress level in order to have a clear definition for each level. It is also a very difficult task to most labelers to identify all stressed syllables in an utterance directory [15].

Manuscript received on June 15, 2006

The authors are with the Department of Computer Engineering, Faculty of Engineering Chulalongkorn University, Bangkok, 10330, Thailand; E-mail: hommekid@yahoo.com, Taratip.S@chula.ac.th

To this end, supervised learning techniques might be used to quantize the degree of stress to discrete levels. This paper proposes a method based on clustering techniques for categorizing degree of stress into several stress levels. By considering the acoustic correlation of stress in literatures, the number of prominent features were chosen based on duration and pitch to represent the degree of stress. These features were extracted from each syllable in a Thai speech dataset. Clustering techniques, i.e. EM algorithm and the model explorer algorithm [34], were employed to classify all syllables into several reasonable groups according to stress levels. The cluster analysis results revealed the correlation between the prominent features and the level of stress, which can be utilized as a guideline to label a stress by hand.

In this paper, we first describe the speech dataset. Then, the prominent features based on duration and pitch contour are described in Section 3. In Section 4, two cluster analyses of their prominent features are discussed. Finally, conclusions are drawn in Section 5.

2. THAI SPEECH DATASET

Thai speech dataset used in this study was produced by Centre for Research in Speech and Language Processing (CRSLP), Chulalongkorn University, Thailand. The content includes approximately 5.4 hours of formal reading style speech and 1.8 hours of casual reading style speech. The former was collected from two male and two female speakers, while the latter was collected from one male and one female speakers. The dataset was manually labeled with onset-rhyme units.

3. PROMINENT FEATURES

Many researchers have studied the acoustic correlation of stress in several languages [7, 16, 17]. The correlation appears to vary from language to language. Most researches have confirmed that duration is the most important acoustic correlate of stress. Lea [18] found that, beside duration and energy, F_0 were also correlated with lexical stress in English. Some studies have also used spectral features, such as spectral change, measured as the average change of spectral energy over the middle part of a syllable [10, 19]; and spectral tilt, measured as spectral energy in various frequency sub-bands [17, 20].

In Thai, some studies [13, 14, 21, 22] indicated that duration is the predominant cue in signaling the distinction between stressed and unstressed syllables. Potisuk et al. [13] also found that the intrinsic pitch contour for each tone still preserved its shape across stress categories. Therefore, we employed duration and pitch as features in this study.

3.1 Duration Feature

The duration can be calculated in a number of ways. For example, one could use the duration of a syllable, the duration of the vowel in the syllable or the duration of the rhyme in the syllable. Stress is assumed to be a feature of a syllable, but for practical purposes, stress can be attributed to the vowel [23]. Many researches used vowel duration for representing stressed/unstressed syllables [7, 17, 23, 24] but some researchers proposed to use the rhyme duration for representing them [13, 16, 25].

From [21], the rhyme portion has been shown to be a better part for stress recognition in Thai when compared to the whole syllable unit. Therefore, only the rhyme portion of each syllable was considered in our experiments. Each rhyme duration was converted into log ms. The log transformation was used to create more normal probability distributions for duration [26] and more conducive to modeling with a Gaussian mixture distribution [4].

Since variation of syllable structure and speaking rates correlate to rhyme duration of syllables, the log duration was normalized by the z-score technique using log duration mean and standard deviation of each speaker and each syllable structure. In this study, syllables were classified into four categories: CV:, CV, CV(:)S and CV(:)O, where C, V, V:, V(:), S and O are initial consonant, short vowel, long vowel, short or long vowel, sonorant ending, and obstruent ending, respectively. The normalized duration feature is referred to as z_d.

3.2 Pitch Features

Pitch was extracted and manually corrected using PitchEditor module of PRAAT program [27]. The pitch value of unvoiced portion was set by linear interpolation. Normally, the shape of pitch contour of a syllable mainly depends on syllabic tone. However, there are other interacting factors affecting the shape of pitch contour, e.g., intonation, coarticulation, stress, and speaker's gender [14]. This study aimed to cluster syllables to different groups of stress level based on the prominent features. Thus the other factors on the prominent features, except stress, should be removed.

Intonation is defined as a combination of tonal features into larger structural units associated with the acoustic parameter of pitch and its distinctive variations in speech process [28]. To eliminate this effect,

Table 1: *The number of syllables in each speech data group.*

Tone	Male	Female
mid	11,378	11,785
low	7,282	7,602
fall	6,697	6,959
high	4,496	4,726
rise	3,167	3,371

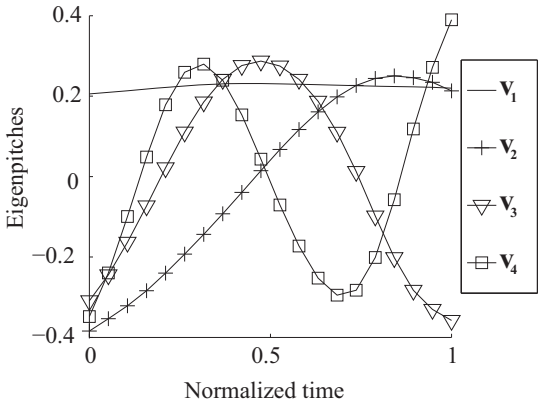


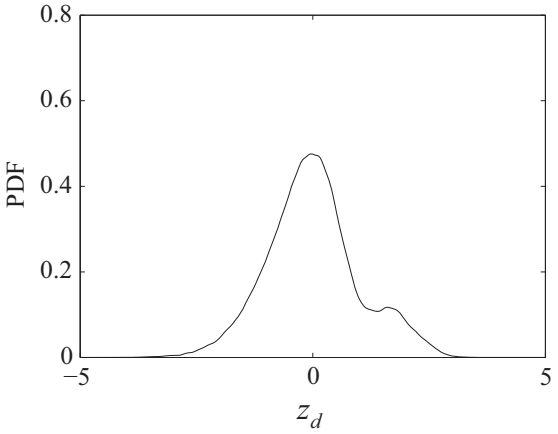
Fig.1: *The four most significant eigenpitches of Thai syllables.*

the pitch contour of each utterance was adjusted by center-point intonation normalization [14].

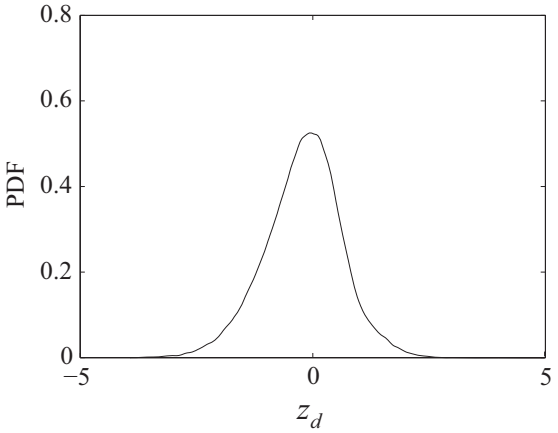
Due to variation of pitch contour is mainly depended on syllabic tone and speaker's gender factors. The speech dataset was divided into ten groups and each group was analyzed separably. These groups are referred to as male-mid, male-low, male-fall, etc. The number of syllables in each group is shown in Table 1.

Coarticulation is the effect of neighboring syllables on the pitch shape of the considering syllable. Potisuk et al. [29] used three-tone sequences to measure this effect. There are 175 possible three-tone sequences, i.e., 5^3 (in the middle of a sentence) + 5^2 (at the beginning of a sentence) + 5^2 (at the end of a sentence) [30]. Unfortunately, the grouping technique cannot be applied in this case since the number of data in each group is too small to analyze. Therefore, the pitch feature was performed without a consideration of the effect from coarticulation.

The principal components analysis (PCA) technique [31] was used to describe the shape of pitch contours. Tian and Nurminen [32] showed that PCA is useful for extracting feature vectors from the pitch contours of Mandarin syllables. They also found that the tonal patterns are preserved in the eigenpitch representation. To determine the eigenpitches, N sampling points of pitch of all syllables in the speech dataset were used to calculate the covariance matrix $N \times N$. The eigenvectors $\{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_N}\}$ of the covariance matrix are the principal compo-



(a)



(b)

Fig.2: *Estimated PDF of z_d of (a) all syllables in the speech data and (b) only non-tonic syllables in the speech data.*

nents or eigenpitches. Their corresponding eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ are numerically related to the variance of the data of that component; the higher the eigenvalues the more significant of that component.

In this study, the number of sampling points (N) was set to 20. By analyzing the speech dataset, the four most significant eigenpitches of Thai syllables were used as shown in Figure 1. The first eigenvector describes the pitch level. The rest of the eigenvectors are used to model the pitch variation. The pitch feature vector of each syllable was simply the dot product of the sampled points of the pitch contour and the four eigenpitches.

4. CLUSTER ANALYSIS

This study was attempted to find out the natural clusters in the data (prominent feature vectors of syllables) and estimating the correct number of clusters (representing stress level). In this section, we first examined the cluster analysis of duration feature, and then explored the cluster analysis of the combination of duration feature and pitch features.

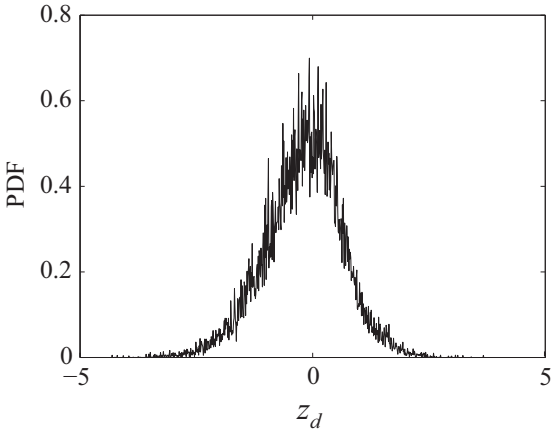


Fig.3: *Estimated PDF of zd of only non-tonic syllables in the speech data with σ = 0.001.*

4.1 Cluster Analysis of Duration Feature

Since the normalized duration feature z_d has only one dimension, the cluster of the speech dataset could be analyzed by investigating its probability density function (PDF). We estimated the PDF of z_d of all syllables in speech dataset by using Parzen window method with Gaussian kernel [33]. In order to inspect the clusters, we first used Gaussian kernel with a large standard deviation (σ) and then we gradually decreased σ until the PDF was split into multiple clusters or σ was reduced to 0.001. When σ was reduced to 0.07, the PDF separated into two dominant clusters, as shown in Figure 2 (a). We suspected that the cluster separated from the main cluster is the cluster of the last syllables of utterances. Generally, the most prominent stress, called tonic stress, almost always found in a syllable in utterance final position.

Then we removed the last syllable of each utterance from the speech dataset and reestimated the PDF by using the same σ (0.07). We found that the second cluster was removed as shown in Figure 2 (b). We continued the analysis to find further dominant clusters by examining the PDF until σ was reduced to 0.001. As a result shown in Figure 3, no obvious cluster was found. This indicates that duration feature can be used to classify the speech dataset into two main clusters; the right one is the cluster of the last syllables of utterances (tonic syllables) and the left one is the cluster of syllables in other positions (non-tonic syllables)

4.2 Cluster Analysis of Combination of Duration Feature and Pitch Features

In this section, the feature vectors to be analyzed were composed of z_d and 4D pitch features. Unlike Section 4.1, the 5D feature vectors could not be visu-

Input: X {a dataset}, k_{max} {maximum number of clusters}, *num_subsamples* {number of subsamples}

Output: $S(i,k)$ {list of similarities for each k and each pair of sub-samples}

Require: A clustering algorithm: cluster(X,k); a similarity measure between labels: $s(L_1, L_2)$

1. $f = 0.8$
2. **for** $k = 2$ **to** k_{max} **do**
3. **for** $i = 1$ **to** *num_subsamples* **do**
4. $sub_1 =$ subsamp(X, f) {a sub-sample with a fraction f of the data}
5. $sub_2 =$ subsamp(X, f)
6. $L_1 =$cluster(sub_1, k)
7. $L_2 =$cluster(sub_2, k)
8. $Intersect = sub_1 \cap sub_2$
9. $S(i,k) = s(L_1(Intersect), L_2(Intersect))$ {Compute the similarity on the points common to both subsamples}
10. **end for**
11. **end for**

Fig.4: *The model explorer algorithm. [34]*

ally examined the cluster structure from the PDF. Thus, EM algorithm was applied to automatically cluster the data. In order to determine the number of clusters to be close to the natural structure of the data, a stability based method proposed by [34] was employed. The method can be used with any clustering algorithm; it provides the means of defining an optimum number of clusters, and can also detect the lack of structure in the data.

To explain the method, we start with the definition of notation. Let $X = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$, and $\mathbf{x_i} \in \Re^d$ be the dataset to be clustered. A labeling L is a partition of X into k subsets $S_1, \ldots, S_k$. We use the following representation of a labeling by a matrix C, with components:

$$C_{ij} = \begin{cases} 1 \text{ if } \mathbf{x_i} \text{ and } \mathbf{x_j} \text{ belong to the same cluster and } i \neq j, \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

Let labeling L_1 and L_2 have matrix representations $C^{(1)}$ and $C^{(2)}$, respectively. The dot product of the labelings is defined as:

$$\langle L_1, L_2 \rangle = \langle C^{(1)}, C^{(2)} \rangle = \sum_{i,j} C_{ij}^{(1)} C_{ij}^{(2)} \quad (2)$$

To measure the similarity between two labelings,

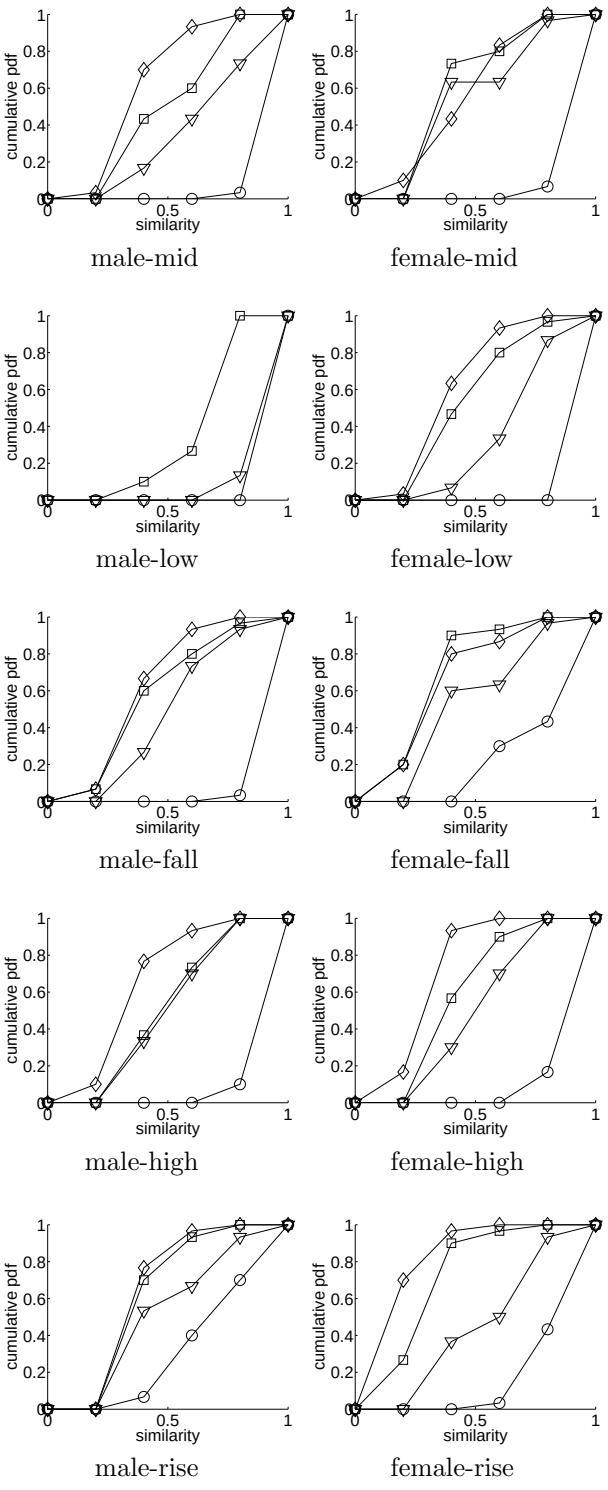


Fig.5: *Cumulative distributions of the similarity score for ten data groups.*

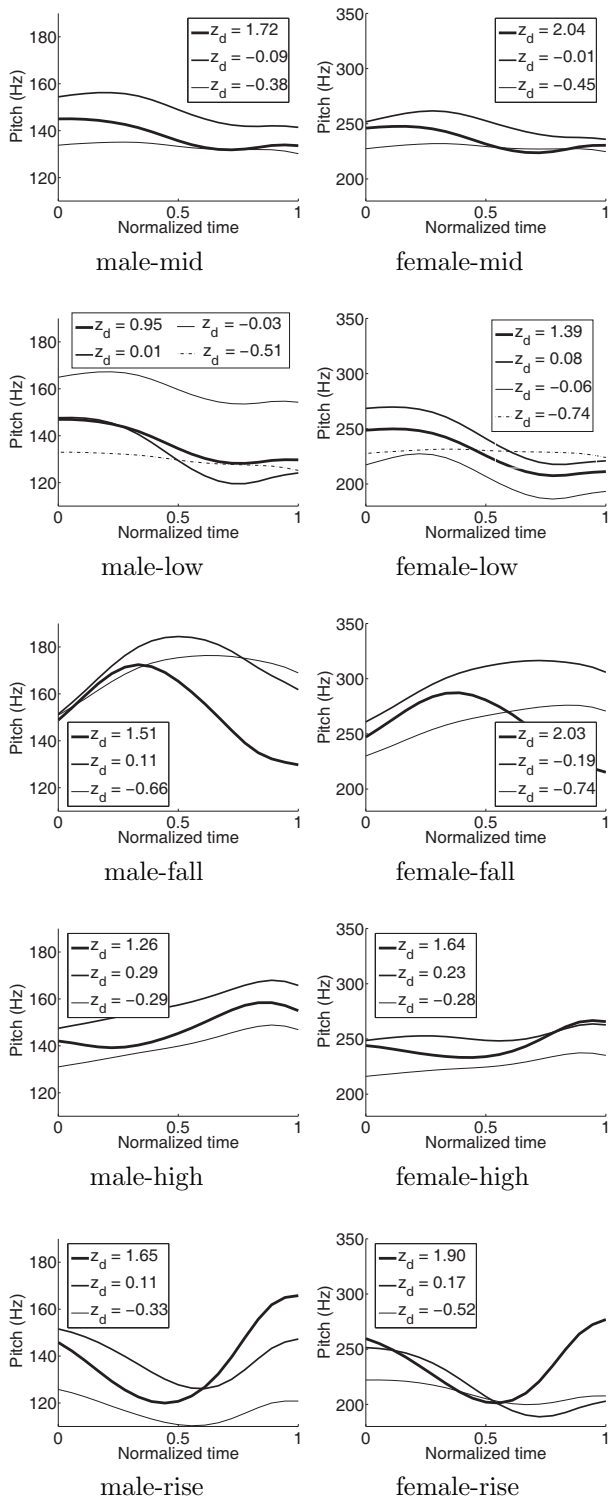


Fig.6: *Mean vectors of prominent clusters for ten data groups represented as the reconstructed pitch contours and their z_d (the thicker contour indicates the stronger stress).*

Jaccard coefficient is used:

$$J(L_1, L_2) = \frac{\langle C^{(1)}, C^{(2)} \rangle}{\langle C^{(1)}, C^{(1)} \rangle + \langle C^{(2)}, C^{(2)} \rangle - \langle C^{(1)}, C^{(2)} \rangle} \quad (3)$$

The idea of this method is that when one looks at two sub-samples of a cloud of data points, with a sampling ratio f (fraction of points sampled) not much smaller than 1 ($f > 0.5$), one usually observes the same general structure. Thus it is reasonable to postulate that a partition into clusters has captured the "inherent" structure in a dataset if partitions into k clusters obtained from running the clustering algorithm with different subsamples are similar. This algorithm is called the model explored algorithm presented in Figure 4.

In this analysis, clustering of only non-tonic syllables was focussed. The k_{max} and $num\_subsamples$ were set to 5 and 30, respectively. To determine the optimum k, Ben-Hur et al. [34] suggested to choose the value where there was a transition from a similarity score distribution that was concentrated near one to a wider distribution. This could be quantified by a jumping in the area under the cumulative distribution function.

The cumulative distributions of the similarity for each speech data group (separated according to syllabic tone and speaker's gender) are shown in Figure 5.

It is noticeable that clustering of male-low group into five clusters is impossible because the covariance matrix always has zero determinant. This usually occurs when applying EM algorithm with too many expected clusters.

We make several observations regarding the cumulative distributions. For $k = 2$, the similarity scores of all groups are concentrated near 1.0, since all data groups can be classified into two clusters. However, the distributions of female-fall, male-rise, and female-rise groups are weaker concentrations than the others.

For $k = 3$, the scores of most groups except the male-low and female-low groups are widely distributed. Only low tone syllables (especially for male speakers) can be reasonably categorized into three clusters.

For $k > 3$, all data groups have widely distributed similarity scores. There is no longer one preferred a cluster.

We then visually determined the best k for each data group. With the best k, we run EM algorithm for all data in that group to get the mean vector of each cluster. The mean vectors of prominent clusters for each data group are represented as the reconstructed pitch contours and their z_d compared with the mean vectors of the tonic syllables as shown in Figure 6.

The level of stress for each cluster was determined by measuring Mahalanobis distance from the mean vector of that cluster to the center of the Gaussian model. The closest cluster to the tonic cluster is defined as the strongest stress. The farthest cluster is considered as the weakest stress. The level of stress for each cluster is represented as the thickness of the contour. The thicker contour represents the stronger stress level.

By considering the contours of each cluster, we found that, in syllables with the strongest stress level, the pitch contours of each tone are quite different from each other. For syllables with the weakest stress level, the shapes of contours among the five Thai tones are rather flat. Moreover, the pitch contours of mid tone and low tone of the syllable with weakest stress level are very confuse. This is a problem of neutral tone. The neutral tone always occurs in unstressed syllables. It have no pitch value of its own, but acquires its pitch value according to context. This makes the tone recognition of syllables with the weakest stress level to be a hard problem.

5. CONCLUSION

This study analyzed the characteristics of the prominent features of Thai syllables to determine the appropriate number of stress levels in Thai speech. Two clustering analyses based on duration and pitch features were explored. For the first one, the duration feature was analyzed by investigating its probability density function using Parzer window method with Gaussian kernel. We found that the duration features extracted from all syllables in the speech dataset provide two clusters representing tonic and non-tonic syllables. To discover further clusters, we continued the analysis of the remaining non-tonic syllables. No obvious cluster was found. This confirms that duration feature can be used to classify the speech dataset into tonic syllables and non-tonic syllables.

For the second analysis, the duration feature was incorporated with the pitch features to discover further clusters within non-tonic syllables. EM algorithm and the model explorer algorithm were combined to determine clustering structure of these non-tonic syllables by separably analyzing them according to syllabic tones and speaker's genders. The results show that there are a two clusters for most groups of non-tonic syllables. Only groups of low tone for both genders provide three clusters. According to empirical results, both analyses reveals that, in most cases, the degree of stress in Thai should be digitized to three levels.

References

[1] H. Mixdorff, "Speech Technology, ToBI and Making Sense of Prosody," *International Conference on Speech Prosody*, 2002.

[2] P. Mittrapiyanuruk, C. Hansakunbuntheung, V. Tesprasit, and V. Sornlertlumvanich, "Improving Naturalness of Thai Text-to-Speech Synthesis by Prosodic Rule," *Proceeding of International Conference on Spoken Language Processing*, 2000.

[3] C. Wang and S. Seneff, "Lexical Stress Modeling for Improved Speech Recognition of Spontaneous Telephone Speech in the JUPITER Domain," *Proceeding of the European Conference on Speech Communication and Technology*, pp. 2761–2765, 2001.

[4] K. Livescu and J. Glass, "Segment-Based Recognition on the PhoneBook Task: Initial Results and Observations on Duration Modeling," *Proceeding of the European Conference on Speech Communication and Technology*, 1437–1440, 2001.

[5] Y. R. Wang and S. H. Chen, "Tone Recognition of Continuous Mandarin Speech Assisted with Prosodic Information," *Journal of the Acoustical Society of America*, Vol. 96, No. 5, pp. 1738–1752, 1994.

[6] N. Thubthong and B. Kijsirikul, "A Syllable-based Connected Thai Digit Speech Recognition Using Neural Network and Duration Modeling," *Proceeding of IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 785–788, 1999.

[7] G.S. Ying L.H. Jamieson R. Chen and C.D. Michell, "Lexicl Stress Detection on Stress-minimal Word Pairs," *Proceeding of International Conference on Spoken Language Processing*, pp. 1612–1615, 1996.

[8] J. Högberg and K. Sjölander1, "Cross Phone State Clustering Using Lexical Stress and Conext," *Proceeding of International Conference on Spoken Language Processing*, pp. 474–477, 1996

[9] L. Hitchcock and S. Greenberg, "Vowel Height is Intimately Associated with Stress Accent in Spontaneous American English Discourse," *Proceeding of the European Conference oon Speech Communication and Technology*, pp. 79–82, 2001.

[10] A. Aull and V. Zue, "Lexical Stress Determination and Its Application to Large Vocabulary Speech Recognition," *Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1549–1552, 1985.

[11] P. Jande, "Stress Patterns in Swedish Lexicalised Phrases," *Proceedings of Fonetik*, pp. 70–73, 2001.

[12] M. Lai, Y. Chen, M. Chu1, Y. Zhao and F. Hu, "A Hierarchical Approach to Automatic Stress Detection in English Sengences," *Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 753–756, 2006.

[13] S. Potisuk, J. Gandour and M.P. Harper, "Acoustic Correlates of Stress in Thai," *Phonet-*

ica, volume 53, pp. 200–220, 1996.

[14] N. Thubthong, B. Kijsirikul, and S. Luksaneeyanawin, "Tone Recognition in Thai Continuous Speech based on Coarticulation, Intonation and Stress Effects," In *Proceeding of International Conference on Spoken Language Processing*, pp. 1169–1172, 2002.

[15] M. Chu, Y. Wang and L. He, "Labeling Stress in ContinuousMandarin Speech Perceptually," *Proceeding of International Congress of Phonetic Sciences*, 2003

[16] S. Potisuk, M. P. Harper, and J. Gandour, "Using Stress to Disambiguate Spoken Thai Sentences Containing Syntactic Ambiguity," *Proceeding of International Conference on Spoken Language Processing*, pp. 805–808, 1996

[17] D. van Kuijk and L. Boves, "Acoustic Characteristics of Lexical Stress in Continuous Telephone Speech," *Speech Communication*, Vol. 27, pp. 95–111, 1999.

[18] W. A. Lea, "Prosodic aids to speech recognition," In W. A. Lea (Ed.), *Trends in Speech Recognition*, pp. 166–205, Englewood Cliffs, New Jersey: Prentice-hall, Inc., 1980.

[19] A. Waibel, *Prosody and Speech Recognition*, London: Pitman, 1988.

[20] A. Sluijter and V. van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *Journal of the Acoustical Society of America*, Vol. 100, No. 4, pp. 2471–2485, 1996.

[21] N. Thubthong and B. Kijsirikul, "Stress and Tone Recognition of Polysyllabic Words in Thai Speech," *Proceeding of International Conference on Intelligent Technologies*, pp. 356–364, 2001.

[22] R. Nitisaroj, "Perception of stress in Thai," *Journal of the Acoustical Society of America*, Vol. 116, No. 4, pp. 2645, 2004.

[23] D. van Kuijk and L. Boves, "Acoustic characteristics of lexical stress in continuous speech," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1655–1658, 1997.

[24] D. van Kuijk and L. Boves "Using lexical stress in continuous speech recognition for Dutch," *Proc. Int. Conf. Spoken Language Processing*, pp. 1736–1739, 1996.

[25] A. Sluijter and V. van Heuven, "Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch," *Phonetica*, Vol. 52 pp. 71–89, 1995.

[26] H. Chung, "Duration Models and the Perceptual Evaluation of Spoken Korean," *Proceeding of the International Conference on Speech Prosody*, 2002.

[27] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer," *Institute of phonetic science, University of Amsterdam, Netherlands*, 2005.

[28] A. Botinis and B. Granström and B. Möbious, "Developments and Paradigms in Intonation Research," *International Journal on Speech Communication*, Vol. 33, No. 4, pp. 263–296, 1994.

[29] S. Potisuk, M. P. Harper, and J. Gandour, "Classification of Thai Tone Sequences in Syllable-Segmented Speech Using the Analysis-by-Synthesis Method," *IEEE Transactions on Speech Audio Processing*, Vol. 7, No. 1, pp. 95–102, 1999.

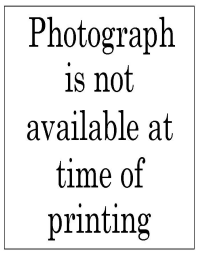
[30] N. Thubthong and B. Kijsirikul, "Tone Recognition of Continuous Thai Speech under Tonal Assimilation and Declination Effects using Half-Tone Model," *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, Vol. 9, No. 6 pp. 815–825, 2001.

[31] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Anacademic Press, Dordrecht, 2001.

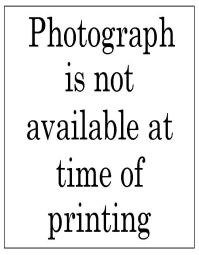
[32] J. Tian and J. Nurminen, "On Analysis of Eigenpitch in Mandarin Chinese," In *Proceeding of the* 4^{th} *International Symposium on Chinese Spoken Language Processing*, 2004.

[33] E. Parzen, "On Estimation of a Probability Density Function and Mode," In *Annals of Mathematical Statistics*, Vol. 33, pp. 1065–1076. 1962.

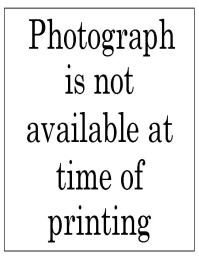
[34] A. Ben-Hur, A. Elisseeff, and I. Guyon, "A Stability Based Method for Discovering Structure in Clustered Data," *Pacific Symposium on Biocomputing*, Vol. 7, pp. 6–17, 2002.



Patavee Charnvivit



Nuttakorn Thubthong



Sudaporn Luksaneeyanawin