Improving Segment-based Speech Recognition by Recovering Missing Segments in Segment Graphs – A Thai Case Study

Krerksak Likitsupin, Atiwong Suchato, Proadpran Punyabukkana [*] and Chai Wutiwiwatchai [†]

[*] Spoken Language Systems Research Group, Department of Computer Engineering,
Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand
E-mail: g48klk@cp.eng.chula.ac.th, {Atiwong.S, Proadpran.P}@chula.ac.th

[†] Human Language Technology Laboratory
National Electronics and Computer Technology Center, Pathumthani, Thailand
E-mail: chai@nectec.or.th

***Abstract*— In segment-based speech recognition systems, the quality of the segmentation step is a major factor highly affecting their accuracies. This paper proposes methods to reduce missing segments caused by boundary insertion errors in segment graphs, which, in the case of Thai, could be generated from a probabilistic segmentation with limited speech resources. Acoustic discontinuities and manners of articulation are used to verify boundaries of the segment graph. Segments are added to the graph in the case of possible falsely detected boundaries. With the proposed insertion error eliminations, the best phonetic recognition accuracy achieved shows a 13.66% error reduction.**

I. INTRODUCTION

Segment-based speech recognition [1] is an approach to the automatic speech recognition problem where each acoustic feature vector is extracted from variable-length portion of speech signal according to a hypothesized underlying speech unit, called "Segment" rather than from a fixed-length frame as in a more widely-adopted frame-based approach, such as the Hidden Markov Model (HMM) -based speech recognition. This technique has many advantages over the frame-based approach. For example, the segment-based approach makes fewer conditional independent assumptions between observations, it can be easily designed to support the use of heterogeneous feature vectors and classifiers [2], and it is easier to be integrated with speech-specific knowledge such as phonetic boundaries – one of important cues for phonetic contrasts. In English, MIT's SUMMIT [1], a segment-based speech recognition system has shown to be successful in various recognition tasks. The system achieve 24.4% phonetic recognition error rate, while the word recognition error rate is at 6.1% on the TIMIT database [3].

Typically, the segment-based framework requires two main steps. The first step is the segmentation step in which some segmentation algorithms are used to construct a graph listing possible interconnections among hypothesized segments which are simply signal portions corresponding to underlying sound units. Next is the recognition step in which the paths in the graph are scored against the acoustic feature vectors. This could be implemented using various methods including using dynamic programming techniques to search the composed weighted finite state transducer between the segment graph and a pronunciation graph derived from the grammar of the recognition task of interest.

It is obvious that the quality of the segment graph, which could be judged based upon how many correctly hypothesized segments residing in the graph, is a major factor that highly affects the recognition accuracy since segmentation errors are propagated to the recognition process. In many languages, probabilistic segmentations that construct segment graphs from the result of first-pass frame-based phonetic recognition results have been proven to yield good performances. For Thai, segment graphs for such highly accurate segmentation algorithms are still prone to errors. This is partially due to the lack of speech resources that can be utilized to train acoustic models of the qualities of ones in languages with longer history in speech recognition researches. With well-tuned HMMs, a phonetic recognition accuracy of approximately only 50% was achieved when clean speech utterances in the training set of LOTUS corpus [4], the only publicly available large-vocabulary Thai speech corpus, were used to train the acoustic models and a bigram language model was also used to constrain the search.

This paper aims at improving the quality of the segment graph obtained from a typical HMM-based phonetic recognition by adjusting segment availability in the graphs so that possible insertion errors, showed in this paper to be the dominating contributor to the overall errors, are compromised.

II. BACKGROUND KNOWLEDGE

This section describes some background knowledge about the distinctive features which we used as additional constraints for improving the segment graphs in the speech recognition technique experimented in this work.

Distinctive features are parameters of the human speech production mechanism that can be applied as a phonetic classifier. Each distinctive feature has a binary value: either

positive (+) or negative (-). There are three categories of distinctive features typically used to describe sound units in world languages: source characteristics, manners of articulation, and places of articulation. Source characteristic distinctive features indicate the vibration of the vocal folds. Manners of articulation represent phonological structures of speech production, where Places of articulation determines major articulators in the vocal tract. In this work, three manner features are used including “sonorant”, “syllabic” and “continuant”. The sonorant feature determines the resonance of phones. The [+syllabic] value of the syllabic feature indicates that such a phone can be the nucleus of a syllable, e.g. a vowel sound, otherwise its value is [–syllabic]. The [+continuant] value describes the occurrence of a free airflow through an oral cavity while [–continuant] indicates that there is a narrow constriction blocking the air stream in the oral cavity while uttering the sound. We can combine manners of articulation into a hierarchical structure to classify phones into broad classes such as vowels, semi-vowels, nasal consonants, fricatives, and stop consonants as shown in Fig. 1.

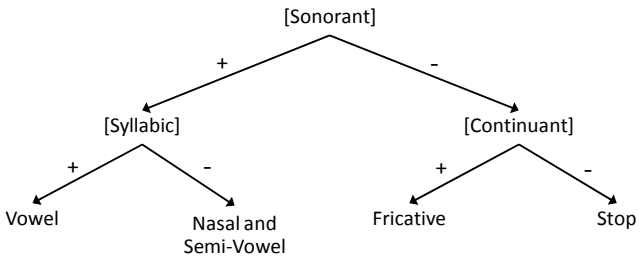

Fig. 1. A Hierarchical structure of speech manners

III. LITERATURE REVIEWS

A. *Speech segmentation for segment-based speech recognition*

To perform acoustic scoring of the incoming speech observations, the signals have to be segmented. This means boundaries between speech units, typically phonemes, have to be hypothesized. There have been a few works dedicated to the segmentation techniques for using with a segment-based recognition approach. Lee [5, 6] developed a probabilistic segmentation algorithm which segments the input speech signal into a segment graph, a directed graph describing how speech units can be interconnected together with their related boundary temporal information, by using a frame-based phonetic recognizer to generate N-best results and then combine these N results into the segment graph. The advantage of this method is that high level constraints such as context-dependent models and language models can be used right in the segmentation step. Clearly, this algorithm is effective when the recognizer used for the segmentation is highly accurate. In contrary to Lee’s approach which relies on a phonetic recognizer, Leelaphattarakij et al. [7] focused on locating phone boundaries directly by detecting evidences for acoustic discontinuities in the signal. Such discontinuities are reflected by differences in Euclidean distances between the spectral feature vectors of the current speech frame and the one following that frame. This method can recall 86.9% of the actual boundaries at 20 milliseconds tolerance level. However, precisions are sacrificed. The method generates twice as many boundaries as the actual number.

B. *Distinctive features-based speech recognition*

To improve the quality of the segment graph obtained probabilistically based on some spectral feature vectors, acoustic-phonetic knowledge should be used more directly. While Leelaphattarakij et al. turned to raw acoustic cues, distinctive features, which belong to a more abstract level than the cues, could also be detected and, in turn, used for assisting in locating or verifying phonetic boundaries. There have been many works showing some successful detection of various distinctive features. For example, Liu [8] was successful in detecting vowel landmarks which are closely related to the [vocalic] distinctive feature in English, while Dareyoah et al. [9] have achieved a good result for this feature in Thai. Identifying distinctive features for consonant places of articulation [10, 11] as well as their [voiced] feature [12] was also shown to yield good results. Pattern classification approaches were also used for detecting distinctive features. Consequently, these features were used in many tasks including speech recognition. Kirchhoff et al. [13] decomposed a complex task of hypothesizing subword unit sequences directly from speech signals into smaller and easier tasks of classifying each speech frame into various binary pseudo-articulatory features, which, in turn, were integrated and utilized as feature vectors for further subword unit classification. She showed that the articulatory feature system achieved superior performances at high noise levels. Borys and Johnson [14] used distinctive feature-based Support Vector Machines (SVMs) to recognize phone sequences. They trained the SVMs to classify manner transitions between phones. Juneja and Espy-Wilson [15] used phonetic features for speech recognition. The phonetic features are extracted from acoustic landmarks located by binary classifiers of speech manners.

In this work, changes in distinctive features reflecting speech manners together with raw acoustic discontinuities are used in assessing boundaries in the segment graph as the attempt to improve the graph quality.

IV. SPEECH CORPUS

A large vocabulary Thai continuous speech corpus called LOTUS [4] was used throughout this work. The corpus contains two speech data sets: phonetically distributed sentence set (PD) and another set containing speech utterances that cover 5,000 most frequently used Thai words. This set consists of another 3 subsets: training set (TR), development test set (DT) and evaluation test set (ET). Each subset of the speech corpus is shipped with complete phonetic label files. Speech utterances in PD and TR were used for training, while ones in DT and ET were used as the development testing set and the performance evaluation set,

respectively. The total number of speaker involved in this corpus is 248 Utterances in the corpus used in this work were recorded at 16 kHz in a clean environment via a dynamic close-talk microphone.

V. PRELIMINARY ANALYSIS

Some preliminary analyses were conducted in order to evaluate the quality of the segment graph obtained from constructing phoneme lattices based on the result of a first-pass Thai phonetic recognizer. The following subsections describe briefly about the preliminary study.

A. System structure

Fig. 2 shows the schematic structure of the segment-based speech recognizer in the preliminary study. Segment graph is constructed from cross fertilizing phonemes in the 20-best list hypothesized by the phonetic recognizer. Segments in the segment graph is then scored using segment-based acoustic models and bigram language model, both trained from the utterances in the training set and their transcriptions.

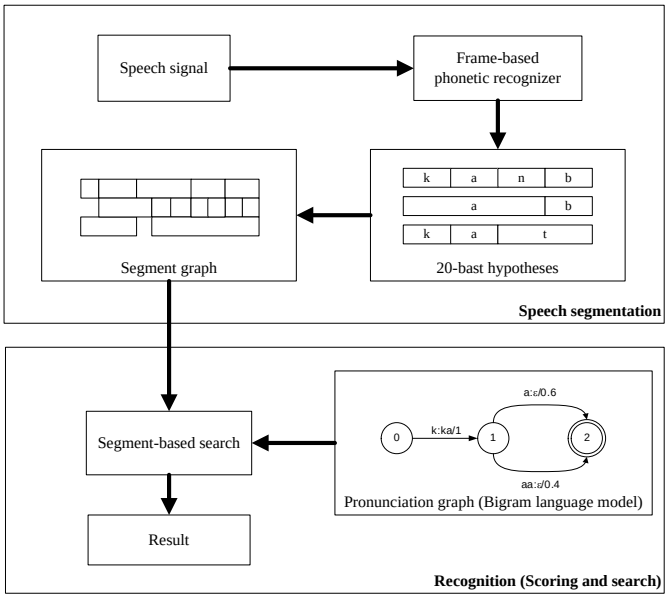

Fig. 2. The system in the preliminary study.

B. Phonetic recognizer

An HMM-based speech recognizer was used as the phonetic recognizer. 39 dimensional MFCC feature vectors were used to represent each speech frame. Multivariate Gaussian distributions were used to models 75 Thai phonemes. A bigram language model trained from the transcription of the TR set was also used as additional constraints.

C. Segment graph evaluation

The segment error which was used for segment graph evaluation is defined as the ratio of the number of transcribed segment not appeared in the segment graph to the total number of transcribed segments. The acceptable tolerance level was ±20 milliseconds. Transcribed segments not appearing in the segment graph could be due to falsely proposed boundaries, i.e. boundary insertion errors, and missed detection of boundaries, i.e. boundary deletion errors. Table I lists the amount of segment errors of the segment graph in this study.

TABLE I
SEGMENT ERRORS IN THE PRELIMINARY STUDY

Segment graph	Error (%)
Total errors	27.7
Errors from boundary insertion errors	15.8
Errors from boundary deletion errors	11.9

D. Segment-based acoustic models

The acoustic representation of each segment is the concatenation of three MFCC feature vectors whose setting is similar to the one used for the frame-based phonetic recognizer. The three feature vectors are extracted from the first 30% in duration of the segment, the next 40%, and the last 30%. Boundaries between segments are also modeled explicitly using another set of features apart from the segmental representation. Three frames, located 20 ms apart from one another, on each side of a boundary are picked for representing the boundary. In this case, 13 MFCCs are extracted from each frame and then concatenated into a 78 dimensional boundary feature vector. Both segmental and boundary representations are modeled using Gaussian distributions. Phonetic boundaries are needed for training the acoustic models. In this work, they are obtained by performing a forced alignment process that aligns phonetic boundaries of the phonetic labels based upon acoustic evidences of the speech utterances.

E. Recognition result

Table II shows the phonetic recognition accuracies of the segment-based system in the preliminary study when the segment graph is obtained from: 1) the probabilistic segmentation, 2) actual boundary information from transcription, and 3) the combination of the segment graph in 1) with the actual boundary information in 2).

TABLE II
PHONETIC RECOGNITION ACCURACIES USING DIFFERENT SEGMENT GRAPHS AND ACOUSTIC MODELS

Segment graph	Accuracy (%)	
	Segmental model only	Segmental + boundary
1) Probabilistic segments	47.70	51.47
2) Actual segments	69.46	76.91
1) and 2)	59.25	65.46

This analysis supports our assumption that a high quality segment graph yields a highly accurate recognition result. We can observe the result from Table II that the phonetic recognition accuracy of the segment-based system depends on the quality of the segment graph. The overall accuracy could be up to almost 80% if segment boundaries are known, while, with probabilistic segmentation, the performance is only at

around 50%. This indicates rooms for improvement in this aspect and that efforts should be spent in improving the quality of the segment graph in order to achieve better phonetic recognition results from the segment-based approach.

VI. PROPOSED METHOD

In this work, we aim at reducing boundary insertion errors which was shown in Table I to contribute to more than one half of the total segment errors. The binary values (+/-) of some manner distinctive features are detected and used, together with discontinuities in the signal, in an attempt to reduce falsely-proposed boundaries. We can implement simple, low-dimensional, and well-trained binary classifiers where good classification results for these features can be expected. Although this makes the segment graph bigger, carefully adding meaningful and high probable segments should lead to a recognition performance that is worthy of the increasing in the size of the segment graph. The method mentioned can be illustrated in Fig. 3.

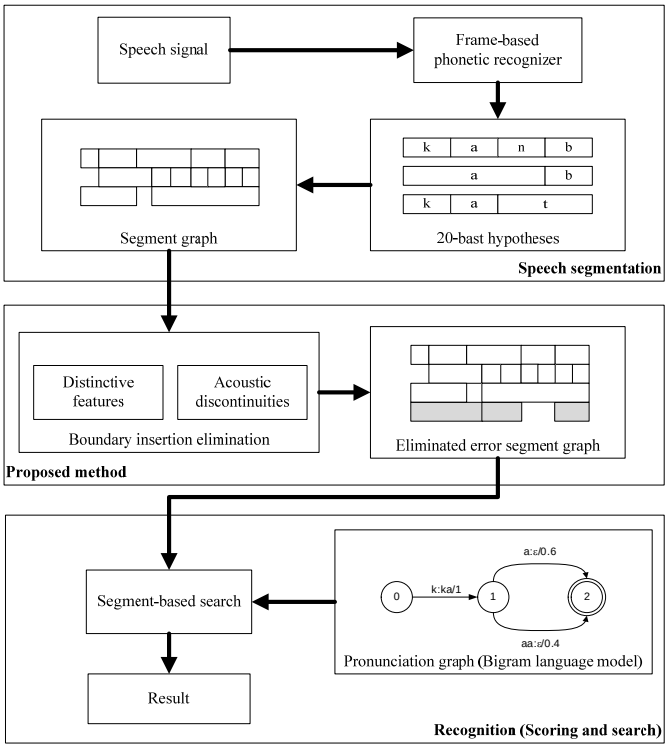

Fig. 3. The Proposed segment-based recognition framework with the proposed insertion error elimination method.

A. *Improving segment graphs based on measurements of acoustic discontinuities at hypothesized boundaries*

Acoustic discontinuities represent the spectral discontinuities of the utterances where their degrees of discontinuity are usually high right at their (acoustic) boundaries. In this work, we measured the acoustic discontinuities from average Euclidean distances (AVD) between MFCC vectors of the 3 adjacent frames on each side of the boundaries. A Gaussian distribution model is used to model measurements from such discontinuities. Our algorithm can be described as the following. Firstly, the AVD value of each boundary in the segment graph to be improved is measured. Secondly, the measurement of each boundary is classified statistically into either an "actual boundary" or an "inserted false boundary". If a boundary is classified as an inserted false boundary, a new segment spanning the original segments on both sides of the boundary is added to the original segment graph. Such a mechanism of adding segments is performed until no new segments are added.

Fig. 4 shows an example of the boundary insertion elimination algorithm. Let's assume that the boundary between the "p" and "x" segments (Boundary "p-x") is classified as an inserted false boundary. "Segment 1" which is a merged segment between the "p" segment and the "x' segment is then added to the segment graph due to a possible insertion error. If the "x-t" boundary is also classified as an inserted false boundary, "Segment 2" is then added. The "Segment 3" is also added due to the hypothesized false boundary "Segment 1-t".

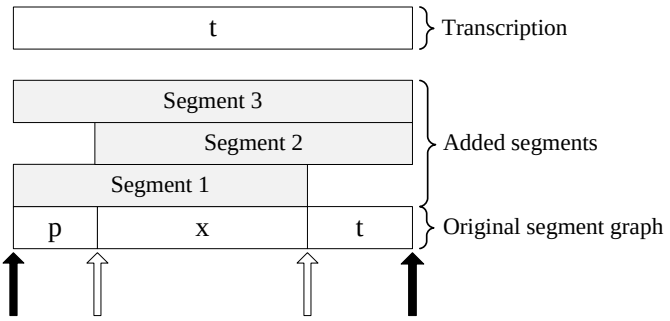

Fig. 4. An example of boundary insertion elimination using acoustic discontinuities. The solid arrows point to the boundaries classified as actual boundaries, while the hollow arrows point to the boundaries classified as inserted false boundary.

B. *Improving segment graph based on distinctive features determining manners of articulation*

Distinctive features used in this work, including [sonorant], [syllabic], and [continuant], are related to manners of speech articulation which are articulator-free features in the human speech production mechanism. To evaluate whether each boundary in the segment graph has a high confidence level of being an actual boundary or not, we conduct the following three steps.

1) Acoustic measurements are measured from each segment in the segment graph. The Acoustic measurements associated with each manner distinctive feature are listed in Table III. These measurements are extracted from the middle of the segment.
2) SVMs are used for determining the binary values of the three manner distinctive features for each segment. This leads to three SVM classifiers.
3) Due to a phonotactical rule of Thai language, we can safely assume that adjacent segments cannot have the

same set of binary values of the three manner distinctive features. Therefore, if no manner changes are detected across a boundary, that boundary will be treated as a highly possible inserted false boundary. Consequently, a segment will be added to the original segment graph correspondingly in the same fashion as the adding mechanism performed in the acoustic discontinuity case. And, such a mechanism of adding segments is performed until no new segments are added.

TABLE III
ACOUSTIC MEASUREMENTS

Manners	Acoustic measurements
Sonorant	Energy between 100-400Hz. Ratio of energy below 2000Hz. to energy between 2000-8000Hz.
Syllabic	Energy between 640-2800Hz. Energy between 2000-3000Hz. Degree of voicing
Continuant	Energy between 2000-8000Hz. Degree of aperiodicity

Fig. 5 shows an example of boundary insertion elimination by detecting manner changes. In this example, the "n^" segment is classified as [+Sonorant][-Syllabic] (i.e. a nasal consonant) while the "p" segment is classified as [-Sonorant][-Continuant] (i.e. a stop consonant). The boundary between both segments shows some changes in manners and is classified as an actual boundary. On the other hand, segments involving the "p-x" and "x-t" boundaries are all classified as [-Sonorant][-Continuant]. Therefore, these boundaries do not reflect any changes in manners. Consequently, they are classified as possible inserted false boundaries. Thus, new segments are added into the segment graph accordingly.

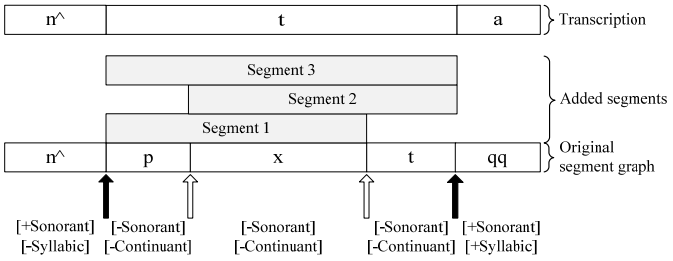

Fig. 5. An example of boundary insertion elimination using distinctive features. The solid arrows point to the boundaries where manner changes are detected and therefore classified as an actual boundary, while the hollow arrows point to the boundaries that do not present any manner changes.

VII. EXPERIMENT DETAILS

The speech corpus described in the preliminary analysis with the same arrangement of the training set and the test set was also used to evaluate our proposed system. The first-pass phonetic recognizer and the segment-based acoustic model settings were also fixed as the one in the preliminary analysis.

First, we evaluated the ability of the selected acoustic measurements in distinguishing between the binary values of each manner features. The training set of the LOTUS corpus was used to train the SVM classifiers, while the classification results were evaluated on the test set. The total number of segments is 33484.

In the next experiment, we studied the effect of the proposed insertion error elimination methods in terms of the segment errors and the number of segments added to the graph. Still, the observation in this experiment just gives some rough idea about the trade-off between the size of the segment graph and the inclusion of actual segments. It does not conclude the merit of the proposed method until the last experiment is conducted.

The last experiment passed the segment graphs obtained from the one before to the segment-based scoring. Phonetic recognition accuracies were measured. A typical HMM-based phonetic recognizer was also included in the comparison.

VIII. RESULTS & DISCUSSION

Table IV shows the manner classification results. The [sonorant] feature is classified with 88.47% accuracy, while the accuracies for [syllabic] and [continuant] are 80.75% and 79.31%, respectively. All manners are classified with high accuracies of around 80% and above using the selected acoustic measurements. Here, we imply from the result that these acoustic measurements lead to the manners that should be accurate enough to benefit the boundary insertion error elimination step.

TABLE IV
MANNER CLASSIFICATION

Manners of articulation	Correction (%)
Sonorant	88.47
Syllabic	80.75
Continuant	79.31

Table V shows the percentage of segment errors for of segment graphs with different insertion error elimination methods. The right-most column shows the ratio of the number of segment contained in the segment graphs belonging to each corresponding method to the number of segments in the original segment graph. When both acoustic discontinuities and manner features are combined in order to determine possible boundary insertion errors, the number of missing actual segments in the segment graph is reduced by 41.58%. However, the size of the segment graph is increased by 2.59 times of the original size. While the elimination method that relies on detecting changes in the manner features did not yield an improvement percentage as much as the ones from the other two methods, the size of the resulting segment graph is considerably smaller. It is left to be seen in the next experiment whether this size increase the phonetic recognition accuracy of the overall segment-based recognizer.

TABLE V
THE SEGMENT ERROR OF EACH SEGMENT GRAPHS

Segment graph	Segment error (%)	Improvement (%)	Ratio of segment graph size
No Elimination	15.80		
1) Discontinuities	10.38	34.30	2.43
2) Manners	11.17	29.30	1.63
1) and 2) together	9.23	41.58	2.59

TABLE VI
RESULTING PHONETIC RECOGNITION ACCURACIES

Acoustic models	Elimination	% Accuracy
Frame-based	-	47.21
Segmental	None	47.70
Segmental	Discontinuities	53.56
Segmental	Manners	52.42
Segmental	Both methods	53.70
Segmental+Boundary	None	51.47
Segmental+Boundary	Discontinuities	58.27
Segmental+Boundary	Manners	57.18
Segmental+Boundary	Both methods	58.50

Table VI shows the phonetic recognition accuracies of the segment-based systems with and without insertion error eliminations as well as the one of the frame-based recognition. With the eliminations, although the segment graphs are more than twice as large as the original graphs in the cases using both elimination methods, the resulting phonetic accuracies are improved considerably compared to the cases with no elimination. Also, regardless of the segment graph sizes, the more actual segmented included in the graph, the more accurate the final phonetic recognition results. Compared to the baseline probabilistic segmentation cases, we have achieved 12.58% and 13.66% error eliminations in the case of segmental models and the combination of both segmental and boundary models, respectively. The performances of all segment-based recognition are also shown in the table to be higher than the one of the typical frame-based recognition.

IX. CONCLUSIONS

We have proposed a method to improve the quality of the segment graph for segment-based speech recognition by attempting to reconstruct segments that are missing due to possible insertion errors. Acoustic discontinuities and manners of articulation are used in evaluating each boundary in the segment graph. The results show satisfactory phonetic recognition accuracy in Thai continuous speech despite the increase in segment graph size in an intermediate step of the system. However, the algorithm reported here in this work only adds new segments into the segment graph. Segment errors due to boundary deletion errors are still remained to be handled in our ongoing future works.

ACKNOWLEDGEMENT

Financial supports from Thailand Graduate Institute of Science and Technology (Grant no. TG-44-09-088D), are gratefully acknowledged.

REFERENCES

[1] J. R. Glass, "A Probabilistic Framework for Segment-Based Speech Recognition," *Computer Speech and Language*, vol. 17, pp. 137–152, 2003.

[2] A. K. Halberstadt and J. R. Glass, "Heterogeneous Measurements and Multiple Classifiers for Speech Recognition," in *ICSLP*, Sydney, Australia, 1998, pp. 995-998.

[3] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallet, and N. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM. NTIS order number PB91-505065," 1990.

[4] S. Kasuriya, V. Sornlertlamvanich, P. Cotsomrong, S. Kanokphara, and N. Thatphithakkul, "Thai Speech Corpus for Thai Speech Recognition," in *The Oriental COCOSDA 2003*, Singapore, 2003, pp. 54-61.

[5] J. W. Chang and J. R. Glass, "Segmentation and Modeling in Segment-Based Recognition," in *Eurospeech*, Rhodes, Greece, 1997, pp. 1199-1202.

[6] S. C. Lee, "Probabilistic Segmentation for Segment-Based Speech Recognition," in *Electrical Engineering and Computer Science*. Doctor of Philosophy Massachusetts: Massachusetts Institute of Technology, 1998, p. 66.

[7] P. Leelaphattarakij, P. Punyabukkana, and A. Suchato, "Locating Phone Boundaries from Acoustic Discontinuities using a Two-staged Approach," in *The Ninth International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh, Pennsylvania, USA, 2006, pp. 673-676.

[8] S. A. Liu, "Landmark Detection for Distinctive Feature-Based Speech Recognition," in *Electrical Engineering and Computer Science*. Doctor of Philosophy Massachusetts, USA: Massachusetts Institute of Technology, 1995, p. 191.

[9] P. Dareyoah, A. Suchato, and P. Punyabukkana, "A Study of Acoustic Measurements for Voicing Detection in Speech with Room-level SNR," in *The Sixth Symposium of Natural Language Processing (SNLP 2005)*, Chiang Rai, Thailand, 2005.

[10] A. Suchato, "Classification of Stop Consonant Place of Articulation," in *Electrical Engineering and Computer Science*. Doctor of Philosophy Massachusetts: Massachusetts Institute of Technology, 2004, p. 181.

[11] A. Suchato and P. Punyabukkana, "Factors in Classification of Stop Consonant Place of Articulation," in *The Ninth European Conference on Speech Communication and Technology (Interspeech 2005)*, Lisbon, Portugal, 2005, pp. 2969-2972.

[12] B. Pholkul, A. Suchato, and P. Punyabukkana, "Stop Consonant Voicing Classification for Computer-Assisted Speech Training of Patients with Cleft Lips and Palates," in *The 1st International Convention on Rehabilitation Engineering & Assistive Technology*, Singapore, 2007, pp. 142-147.

[13] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining Acoustic and Articulatory Feature Information for Robust Speech Recognition," *Speech Communication*, vol. 37, pp. 303-319, 2002.

[14] S. Borys and M. Hasegawa-Johnson, "Distinctive Feature Based SVM Discriminant Features for Improvements to Phone Recognition on Telephone Band Speech," in *The Ninth European Conference on Speech Communication and Technology (Interspeech 2005)*, Lisbon, Portugal, 2005, pp. 697-700.

[15] A. Juneja and C. Espy-Wilson, "A Probabilistic Framework for Landmark Detection Based on Phonetic Features for Automatic Speech Recognition," *The Journal of the Acoustical Society of America*, vol. 123, pp. 1154-1168, 2008.