ENGINEERING JOURNAL

Article

Acoustic-Phonetic Approaches for Improving Segment-Based Speech Recognition for Large Vocabulary Continuous Speech

Krerksak Likitsupin[1,a], **Proadpran Punyabukkana**[1,b], **Chai Wutiwiwatchai**[2,c], **and Atiwong Suchato**[1,d]

1 Spoken Language Systems Research Group, Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand
2 Human Language Technology Laboratory, National Electronics and Computer Technology Center, Pathum Thani 12120, Thailand
E-mail: [a]krerksak@gmail.com, [b]proadpran.p@chula.ac.th, [c]chai.wutiwiwatchai@nectec.or.th, [d]atiwong.s@chula.ac.th (Corresponding author)

Abstract. Segment-based speech recognition has shown to be a competitive alternative to the state-of-the-art HMM-based techniques. Its accuracies rely heavily on the quality of the segment graph from which the recognizer searches for the most likely recognition hypotheses. In order to increase the inclusion rate of actual segments in the graph, it is important to recover possible missing segments generated by segment-based segmentation algorithm. An aspect of this research focuses on determining the missing segments due to missed detection of segment boundaries. The acoustic discontinuities, together with manner-distinctive features are utilized to recover the missing segments. Another aspect of improvement to our segment-based framework tackles the restriction of having limited amount of training speech data which prevents the usage of more complex covariance matrices for the acoustic models. Feature dimensional reduction in the form of the Principal Component Analysis (PCA) is applied to enable the training of full covariance matrices and it results in improved segment-based phoneme recognition. Furthermore, to benefit from the fact that segment-based approach allows the integration of phonetic knowledge, we incorporate the probability of each segment being one type of sound unit of a certain specific common manner of articulation into the scoring of the segment graphs. Our experiment shows that, with the proposed improvements, our segment-based framework approximately increases the phoneme recognition accuracy by approximately 25% of the one obtained from the baseline segment-based speech recognition.

Keywords: Segment-based speech recognition, distinctive features, distinctive features-based speech recognition, speech recognition.

ENGINEERING JOURNAL Volume 20 Issue 2
Received 5 May 2015
Accepted 23 September 2015
Published 18 May 2016
Online at http://www.engj.org/
DOI:10.4186/ej.2016.20.2.179

DOI:10.4186/ej.2016.20.2.179

1. Introduction

Segment-based speech recognition [1] has been proven that it can overcome traditional frame-based segment recognition such as the Hidden Markov Models (HMMs) [2] in both American English language and Thai language. For the American English language, the Massachusetts Institute of Technology's SUMMIT [1], a segment-based speech recognition system, has shown successful results in several recognition tasks. SUMMIT archive appears to have 24.4% phonetic-recognition error rate on the TIMIT (Texas Instruments–Massachusetts Institute of Technology) speech corpus [3], and word-recognition error rate is 6.1% on weather inquiry tasks. For the Thai language, our previous research [4] compared phoneme recognition accuracies between segment-based and frame-based by using a large vocabulary Thai continuous speech corpus. The results showed that segment-based speech recognition yielded slightly better accuracy when segmental models alone were applied, and approximately 5% increase when both segmental and boundary models were used.

Nonetheless, the segment-based speech recognition still has some drawbacks. Our previous research [5] revealed that segment graphs generated by a probabilistic segmentation [6–8] may generate errors due to inserted false boundaries and deleted boundaries. In this work, we defined the segment error as the transcribed segments not appeared in the segment graph. The acceptable tolerance level for a segment boundary is at ±20 milliseconds. This work categorizes the segment error into two types.

The first type is the segment error due to the inserted false boundaries, an example of which is shown in Fig. 1, in which the "kl" segment is missing from the segment graph since the "l-x" boundary is falsely hypothesized.

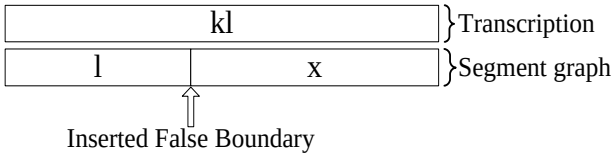

Fig. 1. A segment error due to an inserted false boundary.

The second type of segment error is the segment error due to the deleted boundary (Fig. 2). In this figure, the "a" and "s" segments do not appear in the segment graph since a boundary is missing.

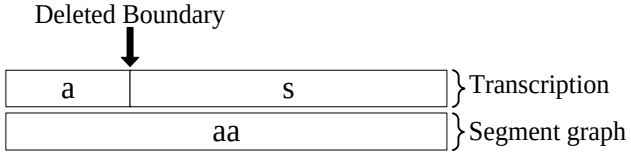

Fig. 2. A segment error due to a deleted boundary.

Table 1 lists the segment error rate of the segment graph, which is the ratio of the number of transcribed segments not appearing in the segment graph to the total number of transcribed segments, obtained via the twenty best hypotheses of the baselined frame-based recognizer on evaluation test set of LOTUS corpus [9, 10]. The errors due to inserted false boundary were handled in previous research [5]. This research will be focusing on the remaining error.

Table 1. Segment errors generated from a probabilistic segmentation.

Error types	Error (%)
Errors due to inserted false boundary	15.80
Errors due to deleted boundary	11.90
Total errors	27.70

The two types of segment errors directly affect the segment-based speech recognition's performance because segment-based speech recognizer will score the segments and paths traversing the segment graph, and searching for optimal hypotheses. To improve the segment graph, we added a segment error recovery step to attempt to recover possible missing segments in a new segment-based speech recognition

framework as shown in Fig. 3. The processes labeled 1 and 2 in the proposed segment-based framework attempt to recover possible missing segments caused by falsely inserted boundaries which done by our previous research [5]. This research expanded upon previous work by recovering segment errors caused by falsely deleted segment boundaries. The changes of distinctive feature values reflecting speech manners (labeled 4), together with some considerations on raw acoustic discontinuities (labeled 3), are used in assessing boundaries in the segment graph in an attempt to improve its quality. Second, we proposed an improved process to segment scoring by evaluating and combining the probabilities of a segment being one type of sound unit of a certain specific common manner of articulation into our probabilistic framework of segment-based speech recognition (labeled 5). Third, due to the limited resources of Thai speech corpus to train well-trained acoustic models for segment-based speech recognition, we tuned the performance of acoustic models by reducing dimensions of feature vector by using a principal component analysis (PCA) (labeled 6).

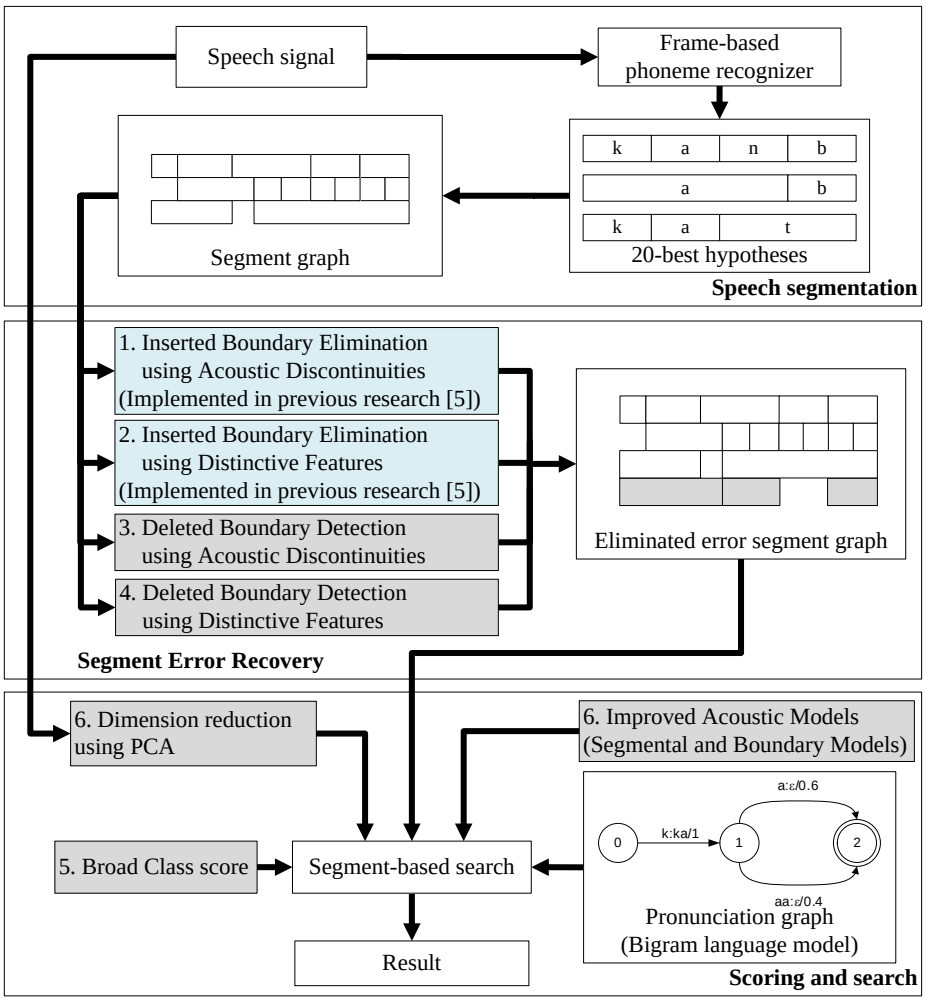

Fig. 3. The segment-based recognition framework with the proposed segment error recovery, broad class score, and dimension reduction.

This paper is structured as follows. Section 2 describes the background knowledge of Acoustic-Phonetics which was incorporated with our segment-based framework. The next section explains in detail about our proposed segment-based speech recognition framework. Section 3.1 and Section 3.2 revisit the methods for eliminating segment errors due to inserted boundaries originally propose from our previous research [5], while the rest of section 3 describe additional proposals originated in this work to the framework. Section 4 defines our experiment setting and experiment details. The experiment results and discussions are shown in the Section 5. In section 6, we conclude the work as well as suggest aspects that should be studied further in the future.

2. Acoustic-Phonetic Approaches

To improve the segment-based speech recognition accuracy, we focused on improving the segment graph quality by increasing the number of correct segments in the segment graph and proposing additional criteria for segment scoring during the decoding process. Acoustic-phonetic information was utilized to achieve these improvements.

2.1. Acoustic Discontinuities (Acoustic Approach)

In general, the degrees of acoustic discontinuities on the boundaries are much higher than the ones within segments. There are many researches proposed segmentation algorithm by locating abrupt acoustic changes such as Glass and Zue [11, 12], Wang et al. [13], and Leelaphattarakij et al. [14]. In this work, we also attempted to capture such acoustic discontinuities through some acoustic measurements in order to verify each and every hypothesized boundary that is highly likely to be one of the actual boundaries (or not). Moreover, the acoustic discontinuities are also used to detect a missing segment. In this work, we applied acoustic discontinuity measuring to the method used in [11, 14], in which Euclidean distance between MFCC vectors were calculated according to the formula listed in Eq. (1).

$$d(a,b) = \sqrt{\sum_{i=0}^{Dim} (a_i - b_i)^2} \tag{1}$$

The $d(a,b)$ is Euclidean distance between frame a and b where a_i refer to MFCC vector at i^{th} dimension of the speech frame a. We creates a vector of Euclidean distance of three frames of the speech signal prior to a hypothesized boundary and ones from three frames after the boundary i.e. $d(p1,a1), d(p2,a2), d(p3,a3)$ as shown in Fig. 4 then models the acoustic discontinuities of segment boundaries and non-boundaries by using Multivariate Gaussian distribution models. The boundary classifier was trained from the training set of LOTUS corpus, in which boundaries were located via some forced alignment.

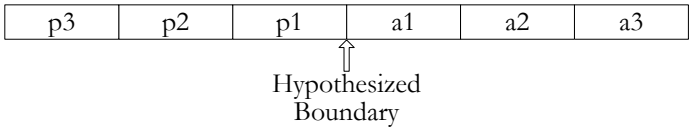

Fig. 4. The discontinuities measurement at hypothesized boundary.

2.2. Distinctive Features (Phonetic Approach)

Sounds in world languages are typically classified based on their corresponding articulatory mechanisms that occur in the human speech production process. Distinctive features are binary-valued (either + or -) features, each of which is associated with a certain major articulatory characteristic. For the set of distinctive features we adopted in this work, distinctive features are categorized into three categories: source characteristics, manners of articulation, and places of articulation. Source characteristics indicate the vibration of the vocal folds, manners of articulation represent phonological structures of speech production, and places of articulation determine major articulators in the vocal tract.

Many research reports have tried to classify speech based on the corresponding articulatory mechanisms occurring in the human speech production process. Most of them focus on distinctive features, which are binary-valued (either + or −) features representing a major articulatory characteristic. Liu [15] and Dareyoah et al. [16] tried to detect vowel landmark. Many works reported cases regarding distinctive feature-based speech recognition framework, ranging from studying and proposing measurable acoustic parameters (APs) to proposing and evaluating complete distinctive feature-based speech recognition tasks in limited domains [17–20], which yielded satisfactory results. Juneja and Espy-Wilson [21–23] also proposed acoustic parameters (APs) for classifying speech signals into defined manner classes. They also proposed a segmentation algorithm, and complete event-based speech recognition on a limited domain, respectively. Tang et al. [24] used manners and places of articulation for speech recognition in order to reduce the classifier's complexity. While Tang used linguistic features, Borys and Hasegawa-Johnson [25] used distinctive feature-based support vector machines (SVMs) to recognize phone sequences. Many

reports on distinctive feature-based speech recognition have claimed that distinctive features have many advantages, such as enabling a linguistic-based hierarchical division of classification problems, being a remedy for data insufficiency due to the usage of lower-dimensional feature vectors, and being more robust. Here, we also utilized distinctive feature information as an additional source in assisting segment-based speech recognition.

In this work, we focused on only four manner features, i.e. "Speech", "Sonorant", "Syllabic" and "Continuant", which were adopted from the researches of Juneja and Espy-Wilson [21, 22]. The speech feature distinguishes between speech ([+Speech]) and silence ([-Speech]). 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 silence, vowels, sonorant consonants, fricatives, and stop consonants, as shown in Fig. 5.

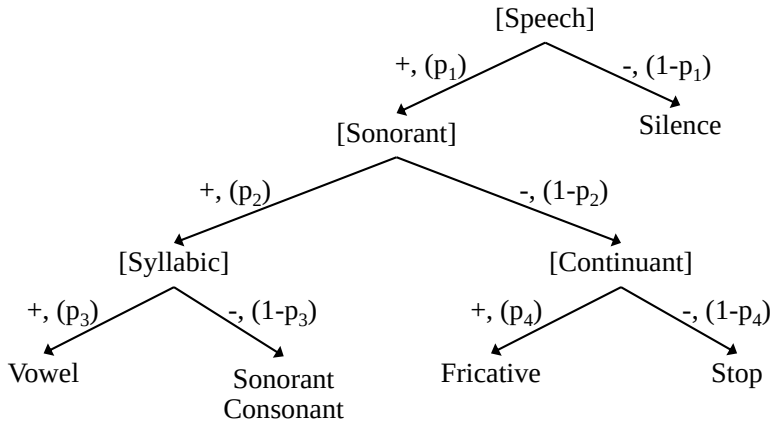

Fig. 5. A hierarchical structure of speech manners [21, 22]. Each label at the branch represents the binary value of the parent manner and its associated probability.

Given an observation vector o_t, we can calculate the probabilities of o_t generated from a type of sound belonging to a broad class, at a leaf of the structure in Fig. 5, by successively conditioning on the distinctive features associated with the sound carrying the values corresponding to all nodes from the root node of the structure traversing to the parents of the leaf nodes of interest. Eq. (2), (3), (4), (5), and (6) demonstrate the calculation of a given observation vector o_t generated from silence, vowel, sonorant, fricative, and stop, respectively:

$$\begin{aligned} P(Silence|o_t) &= P(-Speech|o_t) \\ &= (1-p_1) \end{aligned} \tag{2}$$

$$\begin{aligned} P(Vowel|o_t) &= P(+Speech,+Sonorant,+Syllabic|o_t) \\ &= P(+Speech|o_t)P(+Sonorant|+Speech,o_t)P(+Syllabic|+Speech,+Sonorant,o_t) \\ &= p_1p_2p_3 \end{aligned} \tag{3}$$

$$\begin{aligned} P(Sonorant|o_t) &= P(+Speech,+Sonorant,-Syllabic|o_t) \\ &= P(+Speech|o_t)P(+Sonorant|+Speech,o_t)P(-Syllabic|+Speech,+Sonorant,o_t) \\ &= p_1p_2(1-p_3) \end{aligned} \tag{4}$$

$$\begin{aligned} P(Fricative|o_t) &= P(+Speech,-Sonorant,+Continuant|o_t) \\ &= P(+Speech|o_t)P(-Sonorant|+Speech,o_t)P(+Continuant|+Speech,-Sonorant,o_t) \\ &= p_1(1-p_2)p_4 \end{aligned} \tag{5}$$

$$\begin{aligned} P(Stop|o_t) &= P(+Speech,-Sonorant,-Continuant|o_t) \\ &= P(+Speech|o_t)P(-Sonorant|+Speech,o_t)P(-Continuant|+Speech,-Sonorant,o_t) \\ &= p_1(1-p_2)(1-p_4) \end{aligned} \quad (6)$$

Note that it is very reliable for detecting the binary value of the [Speech] feature. Therefore, we set the value of p_1 to a value very close to unity for [+Speech] and close to zero for [-Speech] in order to compensate for the fact that the probabilities of a sound being a vowel, a sonorant consonant, a fricative, or a stop are computed from a multiplication of three probabilities, while the one for silence is just $1 - p_1$.

Table 2 shows the Thai phonemes which were categorized to phonemes broad class.

Table 2. Thai phonemes in each board class.

Broad Class	Manner Features	Thai Phonemes
Silent	[-Speech]	sil, sp
Vowel	[+Speech][+Sonorant][+Syllabic]	a, aa, i, ii, v, vv, u, uu, e, ee, x, xx, o, oo, @, @@, q, qq, ia, iia, va, vva, ua, uua
Sonorant Consonant	[+Speech][+Sonorant][-Syllabic]	m, m^, n, n^, ng, ng^, w, w^, j, j^, r, l, l^
Fricative	[+Speech][-Sonorant][+Continuant]	c, ch, ch^, f, f^, fr, fl, s, s^, h
Stop	[+Speech][-Sonorant][-Continuant]	p, p^, t, t^, k, k^, z, ph, th, kh, b, d, pr, phr, tr, kr, khr, pl, phl, thr, kl, khl, kw, khw, br, bl, dr

In order to classify the manners from speech signals, we adapted acoustic measurements from the work of Juneja et al. [21], evaluated by ANOVA for their appropriateness to the classification task. Acoustic measurements individually showing good discriminating abilities for each distinctive feature are listed in Table 3.

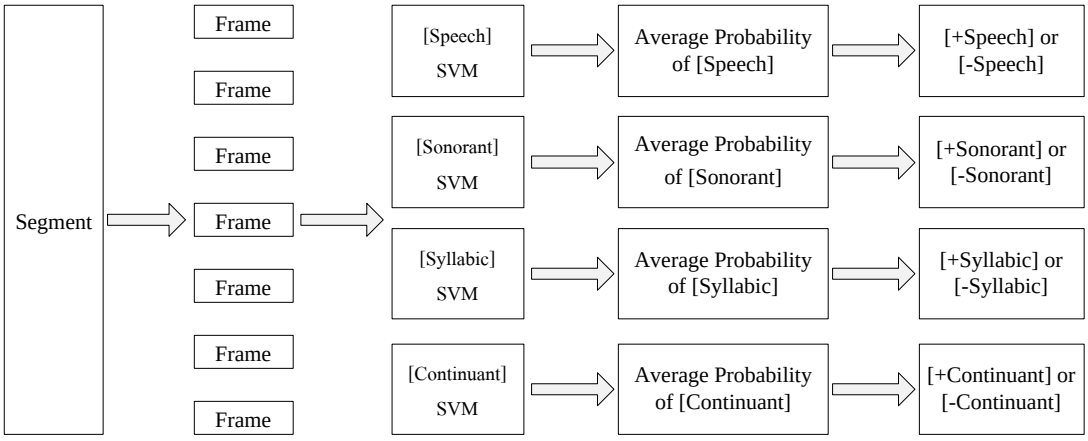

Fig. 6. The segment manners classifications by using support vector machine

Table 3. Acoustic measurements (F_3 is a third formant frequency, f_s is a sampling frequency, and E[a, b] denotes energy in the frequency band [aHz, bHz]).

Manners	Acoustic measurements
[Speech]	E[100, 400] E[0, F_3-1000] E[F_3, f_s/2] E[640, 2800] E[2000, 3000] Degree of voicing Degree of aperiodicity
[Sonorant]	E[100 ,400]

	E[0, F_3-1000]
	E[640, 2800]
	Degree of voicing
	Degree of aperiodicity
	E[0, F_3-1000]
	E[F_3-1000, f_s/2]
[Syllabic]	E[640, 2800]
	E[2000, 3000]
	E[100, 400]
	E[0, F_3-1000]
[Continuant]	E[F_3, f_s/2]
	Degree of voicing
	Degree of aperiodicity

The binary value of each distinctive feature is determined using an SVM classifier. The SVM classifier is trained from 300,000 samples obtained from TR and PD set of the LOTUS corpus by using the acoustic measurements shown in Table 3. However, to identify the manner of a segment, we must average probabilities of every frame in the segment as shown in Fig. 6.

A radial basis function is selected as the SVM kernel based on preliminary results evaluated upon a development test set. The overall conclusion about the binary value of each distinctive feature of an entire segment is then determined by the ET set.

However, in the transcriptions of the LOTUS corpus, the closure region and the release burst region of a stop consonant are included in the same segment. Therefore, to detect a stop consonant ([−Continuant]), we first detect its closure, which is treated as [−Speech], in the front portion of the segment. If the closure is detected, this segment will be classified as [−Continuant]. A toolkit called LIBSVM [26] was used for all SVM implementations.

Phonotactically, we can safely assume that adjacent segments cannot have the same set of binary values of the four manner distinctive features. Therefore, we will verify and detect the segment boundaries based on manner change.

3. Proposed Segment-Based Speech Recognition Framework

To improve the segment graph, we previously proposed in [5] an attempt to recover missing segments caused by boundaries falsely inserted to the segment graphs. In section 3.1 and section 3.2, we revisit the method proposed in the previous work. Another source of errors comes from actual boundaries being omitted in the segment graphs. Consequently, this results in that some correct segments do not exist in the segment graphs at all. In this work, we attempted to recover such segments by analysing acoustic discontinuities in the speech signals as well as changes in manners of articulation portrayed via related distinctive features.

Additionally, we also utilized the probability of a segment being a sound in certain broad classes obtained in the segment error recovery step during the scoring and searching process. Moreover, dimension reduction was also integrated to the process after the feature extraction and used in the scoring process too. Figure 3 illustrates our proposed framework of actions (labeled 1 to 6) in order to improve the segment-based phoneme recognition result. The following sub-section describes the details of every proposed action.

3.1. Inserted Boundary Elimination Based on Measurements of Acoustic Discontinuities at Hypothesized Boundaries (This section was adopted from previous research [5])

The acoustic discontinuities of hypothesized boundaries are measured and classified based on multivariate Gaussian distribution models. If a hypothesized boundary is classified as a falsely inserted 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.

DOI:10.4186/ej.2016.20.2.179

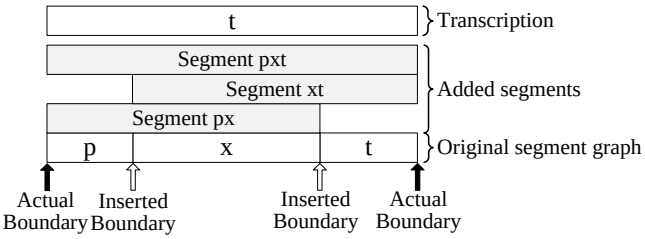

Fig. 7. An example of boundary insertion elimination using acoustic discontinuities. (This figure was adopted from [5]).

Figure 7 shows an example of segment recovery by the inserted boundary elimination algorithm. Let us assume that the boundary between the "p" and "x" segments (boundary "p-x") is classified as a falsely inserted boundary (hollow arrow). "Segment px", 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 xt" is then added. "Segment pxt" is also added due to the hypothesized false boundary "p-Segment xt". This process will continuously verify the boundaries of the newly adding segments too.

3.2. Inserted Boundary Elimination Based on Distinctive Features Determining Manners of Articulation (This section was adopted from previous research [5])

In this step, we attempt to verify the hypothesized boundaries, which were generated from the acoustic segmentation, by detecting manner change. The segments located before and after the hypothesized boundaries are classified by SVM into four manners. If no manner change is detected, the hypothesized boundary is treated as a highly possible falsely inserted boundary. Consequently, a segment will be added to the original segment graph, in the same fashion as the adding mechanism performed in the acoustic discontinuity case. This mechanism of adding segments is performed until no new segments are added.

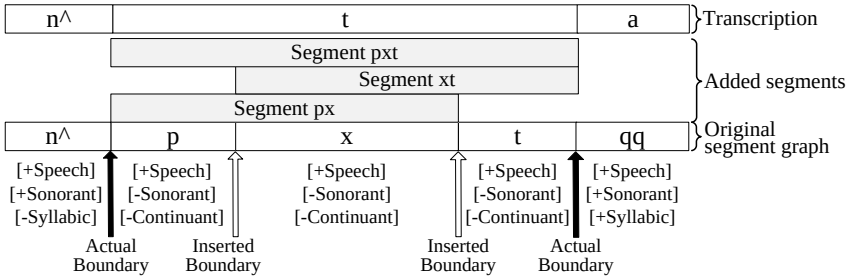

Fig. 8. An example of segment error recovery based on inserted boundary elimination using manner change. (This figure was adopted from [5]).

Figure 8 shows an example of segment error recovery based on inserted boundary elimination by detecting manner changes. In this example, the "n^" segment is classified as [+Speech][+Sonorant][−Syllabic] (i.e. a nasal consonant) while the "p" segment is classified as [+Speech][−Sonorant] [−Continuant] (i.e. a stop consonant). The boundary between both segments shows some changes in manners and is classified as an actual boundary (solid arrow). On the other hand, segments involving the "p-x" and "x-t" boundaries are all classified as [+Speech][-Sonorant][-Continuant]. Therefore, these boundaries do not reflect any changes in manners. Consequently, they are classified as possible inserted false boundaries (hollow arrow). Thus, new segments, i.e. "Segment px", "Segment xt", and Segment pxt", are added into the segment graph accordingly.

3.3. Deleted Boundary Detection Based on Measurements of Acoustic Discontinuities

Besides the falsely inserted boundaries, we also handle the segment errors due to some boundaries being missed by the frame-based recognizer that produces the segment graph. We turn to the assistance of acoustic discontinuities, for which the assumption is that locations with high degrees of acoustic

discontinuity could reasonably qualify as segment boundaries despite being ignored by the frame-based recognizer. An MFCC vector is extracted from each possible set of three adjacent speech frames. The Euclidean distance between two MFCC vectors extracted from any adjacent sets of three frames are measured and monitored. The location where the Euclidean distance is locally maximal is selected as a possible deleted boundary that is missed by the frame-based recognizer. Two segments, corresponding to the result of splitting the segment where each possible deleted boundary resides, are added to the segment graph in an attempt to recover segment errors.

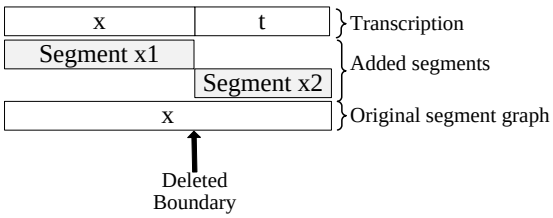

Fig. 9. An example of segment error recovery based on deleted boundary detection using acoustic discontinuities.

Figure 9 shows an example of segment recovery by the deleted boundary detection algorithm. Let us assume that GMM classifier detected the abrupt change at the arrow, we will split the "x" segment to "Segment x1" and "Segment x2", and add them to the segment graph.

3.4. Deleted Boundary Detection Based on Distinctive Features Determining Manners of Articulation

In this section, we detect the deleted boundary by detecting change of manner based on the assumption that the adjacent segments cannot have the same set of binary values of the four manners. First, the acoustic measurements of every frame in the hypothesized segments, which are listed in Table 3, were extracted. Second, all frames in each segment were classified into four manner classes. To avoid manner classification error, we checked the integrity of manner features at the current frame by verifying manner features of the adjacent frames. Third, manner change was detected to identify the deleted boundary. After that, the detected boundary will be verified by the boundary classifier, as for the boundary detection by acoustic discontinuities mentioned in section 2.1. If the detected boundary is verified as the boundary, the segment will be split into two segments based on the newly detected boundary.

Figure 10 shows an example of segment error recovery based on deleted boundary detection by detecting manner changes. We assume that the algorithm detects the manner change, from [+Speech][-Sonorant][-Continuant] to [+Speech][+Sonorant][+Syllabic], at arrow. We defined this position as a deleted boundary, the "t" segment was split to "Segment t1" and "Segment t2". Both new segments will be added to the segment graph.

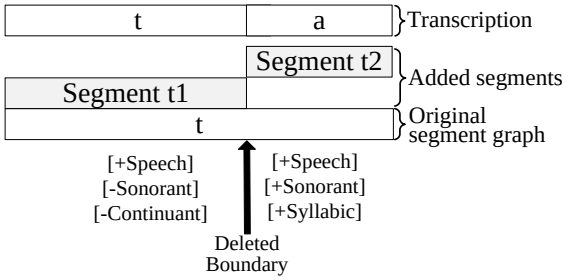

Fig. 10. An example of segment error recovery based on deleted boundary detection using distinctive feature.

3.5. Improving the Scoring Process using Broad Class Scores

One advantage of the segment-based speech recognizer is that it easily integrates the acoustic-phonetic knowledge into a segment-based framework. Therefore, we will integrate the probability of a segment being a sound in certain broad classes into the segment-based scoring and searching process.

Given a set of segment-based observation vectors of all segments in a segment graph A, we solved the phoneme recognition task by finding an optimal phoneme sequence U^* from the equation

$$U^* = \arg\max_{S,U} P(S,U \mid A) \tag{7}$$

where S is a possible segmentation (i.e. a chosen path traversing the segment graph) and U is a sequence of possible phonemes. Let x_i be the observation vector of s_i, the i^{th} segment in S, a segmentation of interest with n segments. Here, we also introduce the multiplicative broad class score $BC(s_i, u_i)$, which should be proportional to how confident we are that s_i belongs to the broad class of u_i, the i^{th} phoneme in U. Deploying the anti-phone modeling technique proposed in [1], with $\bar{\alpha}$ representing the anti-phone model and biasing the probability associated with each segment with $BC(s_i, u_i)$, we can re-arrange Eq. (7) to obtain:

$$U^* = \arg\max_{S,U} \prod_{i=1}^{n} \frac{BC(s_i,u_i)P(x_i \mid u_i)}{P(x_i \mid \bar{\alpha})} P(s_i \mid u_i)P(U) \tag{8}$$

In our experiments, the $BC(s_i, u_i)$ for each s_i is actually $P(b_i \mid s_i, A)$, the probability of s_i being the broad class b_i. Such probabilities were obtained from multiplying the probabilities of associated manners of articulation provided by the SVM distinctive feature classifiers mentioned in the section 4.2. The procedure to obtain broad class scoring by using SVM was explained in Fig. 11.

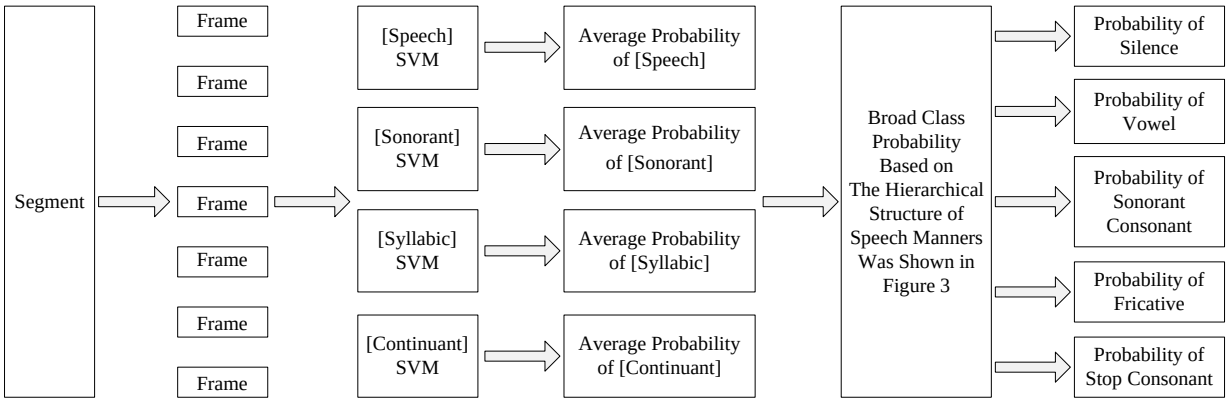

Fig. 11. Broad class scoring by using support vector machine.

3.6. Improving Acoustic Models by Reducing the Dimensions of Feature Vectors using a Principal Component Analysis (PCA)

Another disadvantage of segment-based speech recognition is that it requires more training data to train the acoustic models. In a frame-based speech recognition, one frame can be considered one training sample while one sample in the segment-based are requires an entire sound unit. Although LOTUS is the largest public large-vocabulary Thai continuous speech corpus for research, it still difficult to train well-tuned segment-based acoustic models. GMMBAYES [27], a Gaussian mixture model toolbox we used in this experiment, requires at least $D + D$ samples to train a single mixture diagonal covariance Gaussian model, where D is number of dimension of feature vector. It also requires $D + (D \times \left(\frac{D+1}{2}\right))$ samples to train a single mixture full covariance Gaussian model. For our baseline segment-based speech recognizer, we use the total of 117 dimensions in the feature vectors of the segmental models and seventy-eight dimensions for the boundary models. It appeared that only just sixty of seventy-six phonemes have sufficient amount of training data from both TR and PD set to train the single mixture diagonal covariance Gaussian model while there were only sixteen phonemes that have sufficient training data to train the single mixture full covariance Gaussian model. Therefore, we duplicated some training data to match the minimum required

number of training data. However, the trained acoustic models were not good enough. More training data were required to train the Gaussian mixture models.

Due to this limitation, we solved the problem by reducing the dimension of both segmental and boundary feature vectors. A principal component analysis (PCA) was conducted to analyse factors and generate the new feature vectors.

4. Experimental Settings and Details

4.1. Thai Continuous Speech Corpus (LOTUS Corpus)

This work uses LOTUS, a public large-vocabulary Thai continuous-speech corpus, in all experiments. LOTUS was recorded with forty-eight speakers, both male and female, at 16 kHz in a clean environment via a dynamic close-talk microphone. It contains two main speech data sets: a phonetically distributed sentence set (PD) and another set containing speech utterances that cover the five thousand most frequently used Thai words. This set consists of three subsets: a training set (TR), a development test set (DT), and an evaluation test set (ET). TR, PD, DT, ET contain 3007, 801, 500, and 500 sentences, respectively. All sets were labelled with complete seventy-six phoneme labels. Speech utterances in PD and TR were used for training, while those in DT and ET were used as the development testing set and the performance evaluation set, respectively.

4.2. American English Continuous Speech Corpus (TIMIT Corpus)

To compare the recognition accuracies with context-dependent acoustic model an American English speech corpus (TIMIT) was invoked in our experiment. TIMIT corpus is a well-known American English continuous speech corpus, comprising thirty-nine phoneme units. It contains a total of 6,300 sentences, which were spoken by 630 speakers from eight major dialect regions of United States. In this work, all TIMIT's training set was used to train the context-independent and context-dependent acoustic model. The phoneme-based bigram language model was also trained from labels of the training sets. The TIMIT's testing set was used for evaluating the phoneme recognition results.

4.3. Frame-Based Recognizer

An HMM-based speech recognizer was selected as the frame-based baseline system. We used the hidden Markov toolkit (HTK) [28] as the toolkit for training and testing the recognizer. A thirty-nine dimensional mel-frequency cepstral coefficient (MFCC) feature vector was extracted from each speech frame. Single mixture, eight mixtures, sixteen mixtures, and thirty-two mixtures context-independent Multivariate Gaussian distributions with a diagonal and full covariance matrix were used to model seventy-six Thai phonemes. We used a phoneme bigram language model trained from the transcription of the TR set as an additional constraint. The HMM-based recognizer was tuned by varying the basic HMM topology and the number of re-estimation iterations. Parameters yielding the best phoneme recognition accuracy, based on speech utterances in the development set, were used for the baseline recognizer. We chose the best HMM topology, i.e. with the highest accuracy on the development test set, from three kinds of HMM topologies: 3-state left to right (LTR); 5-state LTR; and a mixture of 3- and 5-state LTR, where 3-state LTR models were used for short vowels, and 5-state ones were used for others. Moreover, we also found the best result by varying the number of re-estimation iterations from one to fifty iterations.

4.4. Segment-Based Recognizer

The segment-based recognizer in this work used segmental and boundary feature vectors, adapted from Halberstadt and Glass's measurements [29]. We used the concatenation of three 39 MFCC feature vectors which were extracted from three parts of the segment: the first 30%, the next 40%, and the last 30%. Anti-phone models [1] were also used to model non-lexicon segmental units. Boundaries between segments were also modeled explicitly using another set of features apart from the segmental representation. Three speech frames, located twenty milliseconds apart from one another, on each side of a boundary were picked for representing the boundary. In this case, 13 MFCCs were extracted from each frame and then concatenated into a seventy-eight dimensional boundary feature vector. Therefore, phoneme boundaries were required

for training the acoustic models. In this work, we used a forced alignment process to obtain the boundaries of the phoneme labels, based upon acoustic evidence of the speech utterances. The segmental, anti-phone and boundary representations were modeled using single mixture Gaussian distributions with diagonal and full covariance matrix. The segment graph was generated from twenty best hypotheses from the baseline HMM-based phoneme recognizer. A phoneme bigram language model was also used for the language constraint, in addition to the language constraint in the HMM-based recognizer. This setting was applied to both the baseline and proposed recognizers.

4.5. Recognition Evaluation (Phoneme Recognition Accuracy)

In this work, we use phoneme recognition accuracy to indicate the performance of all recognizers. In computing the phoneme recognition accuracy, the recognition results are compared with their corresponding actual transcriptions. The HResults program, bundled within the Hidden Markov Model Toolkit (HTK) [28], was used for the calculation. The recognition accuracy is computed by

$$Accuracy = \frac{H - I}{N} \times 100\% \tag{9}$$

Note that H is $N-(D+S)$ where D is the number of deletions, S is the number of substitutions, I is the number of insertions, and N is the total number of labels in the corresponding transcriptions.

4.6. Experimental Details

For a fair comparison, in the following experiments we still use the same speech corpus as described in the segment error analysis section. Also, the parameters and configuration of the first-pass phoneme recognizer and the segment-based acoustic models are set to the values used in the segment error analysis.

The first experiment evaluated the ability of the selected acoustic measurements to determine the binary values of each manner feature. Approximately 300,000 frames of training examples, which were selected from TR and PD of LOTUS corpus, were used to train the SVM classifiers, while the segment classification results were evaluated on the development test set. Probabilities of speech frames in a segment being a certain binary value were averaged in order to obtain the final binary value for that segment. There were 33,323 segment evaluated in the development test set.

In the second experiment, we studied the effect of the proposed inserted boundary elimination and deleted boundary detection methods. The segment error rates and the number of segments added to the graph were compared. This experiment can give some rough ideas about the trade-off between the size of the segment graph and the inclusion of actual segments. The merit of the proposed method cannot be evaluated until the phoneme recognition experiment is conducted.

For the third experiment, phoneme recognition accuracies of the proposed segment-based framework, the baseline segment-based recognizer, and the typical HMM-based phoneme recognizer were measured and compared. Moreover, we also compared the recognition results of segment-based speech recognition, with and without the broad class scores, to observe their contributions. The reason that we conducted the phoneme recognition task is because we do not want a higher-level constraint, such a word-based language model, to affect the recognition accuracies. Also, it is commonly known that if we add a well-trained language model, the recognition accuracy of each recognizer in the paper will be improved.

In the fourth experiment, we tried to improve the quality of acoustic models by reducing the dimensions used for the segment and boundary models via Principal Component Analysis (PCA). Consequently, new acoustic models with Gaussian Mixture Models were trained based on the PCA components. Phoneme recognition accuracies were compared with the original models with full dimensions.

American English phoneme recognition experiments were also conducted using the TIMIT corpus to compare the phoneme recognition accuracies among the ones of a frame-based speech recognizer, our segment-based speech recognizer, and the case when our segment-based speech recognizer was applied with the proposed boundary insertion and elimination procedure. The acoustic models for all American English phonemes were trained based on the training dataset of TIMIT and the evaluation was based on the test dataset of the corpus.

5. Results and Discussions

The manner classification correction percentages are showed in Table 4. The correction percentages for [Speech], [Sonorant], [Syllabic] and [Continuant] are 93.47%, 93.54%, 80.66% and 73.92%, respectively.

Table 4. Manners classification results.

Manners	Correction (%)
[Speech]	93.47
[Sonorant]	93.54
[Syllabic]	80.66
[Continuant]	73.92
All manners	88.55

The SVM manner classifiers in the first experiment yield very good recognition classification results. These numbers indicate that the chosen acoustic measurements can perform the manner classification reasonably well. A significant source of errors comes from stop consonants being classified as silences; this is due to the fact that, in Thai, stop releases are always absent in certain positions and they are normally weak in other positions. Such errors are difficult to cope with using spectral shape-focused measurements without explicit formant tracking. The detection of the [Continuant] is also a major error source in terms of accuracy rate. We could argue that the lack of acoustic measurements directly capturing segment durations contributes to the confusion between stop releases and weak fricatives, leading to such errors. However, we can assume that the overall classification accuracy of 88.55% provides a reasonable basis for continuing to explore the main proposal to improve the overall segment-based phoneme recognition framework by manner classifiers.

The results of the second experiment are shown in Table 5. This table lists the percentage of segment errors using different error compensation methods. The rightmost column shows the ratio of the number of segments contained in the segment graphs belonging to each corresponding method to the number of segments in the original segment graph. For every method evaluated, we can see reductions in segment errors. However, the contribution of each method varies. Regarding cases of falsely inserted boundaries, the results suggest that acoustic discontinuities are better than manner distinctive features; however, the size of the segment graph is bigger than in the case of the manner feature. Still, the two methods complement each other, leading to an improvement when they are combined.

Table 5. Segment errors of different segment graphs.

Segment graph	Inserted boundary elimination		Deleted boundary detection		Ratio of segment graph size
	Segment error (%)	Improvement (%)	Segment error (%)	Improvement (%)	
No elimination	15.80	-	11.90	-	1
1) Inserted boundary with discontinuities	11.71	25.89	11.90	-	1.73
2) Inserted boundary with manners	12.10	23.42	11.90	-	1.60
1) and 2) together	10.34	34.56	11.90	-	1.93
3) Deleted boundary with discontinuities	15.80	-	10.96	7.90	1.53
4) Deleted boundary with manners	15.80	-	11.06	7.06	1.51
3) and 4) together	15.80	-	10.18	14.45	1.78

Unfortunately, the compensation for falsely deleted boundaries from both acoustic discontinuities and manner features provides only slight improvement. With the two combined, the improvement is less than 15%. The resulting segment graph size for every method is less than twice the size of the original graph. However, the phoneme recognition accuracy needs to be looked at before we can determine the worthiness of these larger segment graphs.

DOI:10.4186/ej.2016.20.2.179

Table 6. Resulting phoneme recognition accuracies.

Acoustic models	Methods	% Accuracy		% Accuracy with broad class score	
		Inserted boundary elimination	Deleted boundary detection	Inserted boundary elimination	Deleted boundary detection
Frame-based	-	47.21	47.21	-	-
Frame-based (20-best)	-	49.39	49.39	-	-
Segmental	None	47.70	47.70	47.72	47.72
Segmental	Discontinuities	52.92	41.88	52.92	42.16
Segmental	Manners	52.21	44.79	52.32	45.29
Segmental	Both methods	53.44	38.50	53.54	39.25
Segmental and Boundary	None	51.47	51.47	51.58	51.58
Segmental and Boundary	Discontinuities	57.50	48.74	57.59	48.91
Segmental and Boundary	Manners	56.91	49.69	56.96	49.86
Segmental and Boundary	Both methods	58.18	46.84	58.26	47.20

The third experiment demonstrated the phoneme recognition accuracies of the segment-based systems with and without inserted boundary elimination and deleted boundary detection, as well as that of baseline frame-based recognition. The recognition results in Table 6 show the satisfactory improvement of proposed methods. Since segment graphs used for the segment-based cases were generated from twenty best hypotheses from frame-based recognizers, we also included a case when all 20-best hypotheses from the frame-based recognizer were considered.

For the falsely inserted boundary eliminations case, the resulting phoneme recognition accuracies are improved considerably compared to the cases without the elimination, although the segment graph sizes are almost twice as large compared to the original graphs using both elimination methods. Therefore, we can infer that the more actual segments included in the graph, the more the segment-based recognizer will achieve accurate final phoneme recognition results. Compared to the baseline probabilistic segmentation cases, error eliminations of 12.03% and 13.04% were achieved in the case of segmental models and the combination of both segmental and boundary models, respectively. The performances of the original segment-based recognition and segment-based recognition using the segment graph from inserted boundary elimination are also shown to be higher than the performances of both baseline frame-based recognitions.

Table 7. Detailed phoneme recognition results of proposed deleted boundary detection methods.

Acoustic models	Deleted boundary detection methods	Correct labels	Insertions	Deletions	Substitutions
Segmental	None	20,380	4,572	2,027	11,007
Segmental	Discontinuities	20,314	6,292	1,565	11,605
Segmental	Manners	20,746	5,749	1,684	11,054
Segmental	Both methods	20,357	7,467	1,367	11,760
Segmental and Boundary	None	21,723	4,488	1,834	9,927
Segmental and Boundary	Discontinuities	21,827	5,495	1,508	10,149
Segmental and Boundary	Manners	21,985	5,346	1,627	9,872
Segmental and Boundary	Both methods	22,010	6,326	1,353	10,121

Although deleted boundary detection could reduce segment errors, it does not demonstrate any significant improvement in phoneme recognition accuracies. Furthermore, it impairs the recognition accuracies in many cases. Eq. (9) shows that the insertions directly affect the accuracies. Table 7 shows the detection method seems to introduce too many additional insertions that outweigh the contribution of the recovery of the deletions.

DOI:10.4186/ej.2016.20.2.179

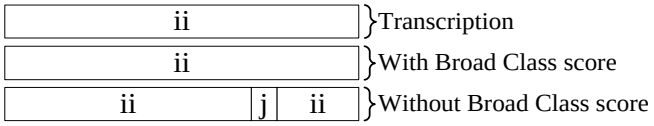

Fig. 12. A comparison of phoneme recognition results with and without the broad class score.

The two rightmost columns of Table 6 show the phoneme recognition accuracies of segment-based approaches with the broad class scores weighted into the probabilistic framework. We can observe that using such scores yields better recognition accuracies in all cases. Figure 12 shows that in some cases of the contribution of broad class scores, a hypothesis with the /j/ segment will not be selected due to a penalty introduced by the broad class scores. Even though it is obvious that the scores will not assist in reducing phoneme recognition errors due to confusion among phonemes in the same classes, we can still obtain some degree of improvement in the overall accuracies.

In the fourth experiment, PCA was used for reducing the dimensions of the feature vectors used for the segment models and the boundary models. We selected the components which have Eigenvalues more than one. After the dimension reduction using PCA, the number of dimension of segment feature vectors was reduced from 117 to 34 while the one of boundary feature vectors was reduced from 78 to 20. Based on PCA-reduced feature vectors, we can finally trained the Gaussian mixture models with full covariance matrices, the task that could be achieved with the original sets of dimensions.

Table 8 shows the phoneme recognition accuracies obtained from the segment-based recognizers using Gaussian models with diagonal covariance matrices and with full covariance matrices in various conditions. The results show that segment-based speech recognizer with full covariance matrices yielded better accuracies in all conditions. The best condition granted 3.16% improvement over its corresponding diagonal covariance matrix case.

In Table 8, we can also compare the phoneme recognition accuracies among ones obtained from frame-based recognizers with different numbers of components (1, 8, 16, and 32) in the Gaussian Mixture Models of the acoustic models and ones from the segment-based cases. Among the frame-based cases, the accuracies increase with the increasing numbers of components as typically expected. Despite being able to utilize only single component Gaussian Mixture Models, we can obtain rather similar phoneme recognition accuracy for the phoneme recognition accuracy of the segment-based case with boundary models as well as boundary elimination methods to the accuracies of the multi-component models of the frame-based approach. Unfortunately, experiments with multi-component models with full covariance matrices could not be conducted on the frame-based cases due to insufficient training speech resources.

Table 8. Comparison of phoneme recognition accuracies.

Acoustic models	**Inserted boundary elimination methods**	**Number of Mixture**	**Diagonal covariance matrix**		**Full covariance matrix**	
			% Accuracy	**% Accuracy with broad class score**	**% Accuracy**	**% Accuracy with broad class score**
Frame-based	-	1	47.21	-	55.90	-
Frame-based	-	8	56.36	-	-	-
Frame-based	-	16	58.19	-	-	-
Frame-based	-	32	58.75	-	-	-
Segmental	None	1	47.70	47.72	48.89	48.93
Segmental	Discontinuities	1	52.92	52.92	53.47	53.49
Segmental	Manners	1	52.21	52.32	53.37	53.43
Segmental	Both methods	1	53.44	53.54	53.58	53.65
Segmental and Boundary	None	1	51.47	51.58	53.02	53.11
Segmental and Boundary	Discontinuities	1	57.50	57.59	59.13	59.23
Segmental and Boundary	Manners	1	56.91	56.96	58.71	58.77
Segmental and Boundary	Both methods	1	58.18	58.26	59.85	59.94

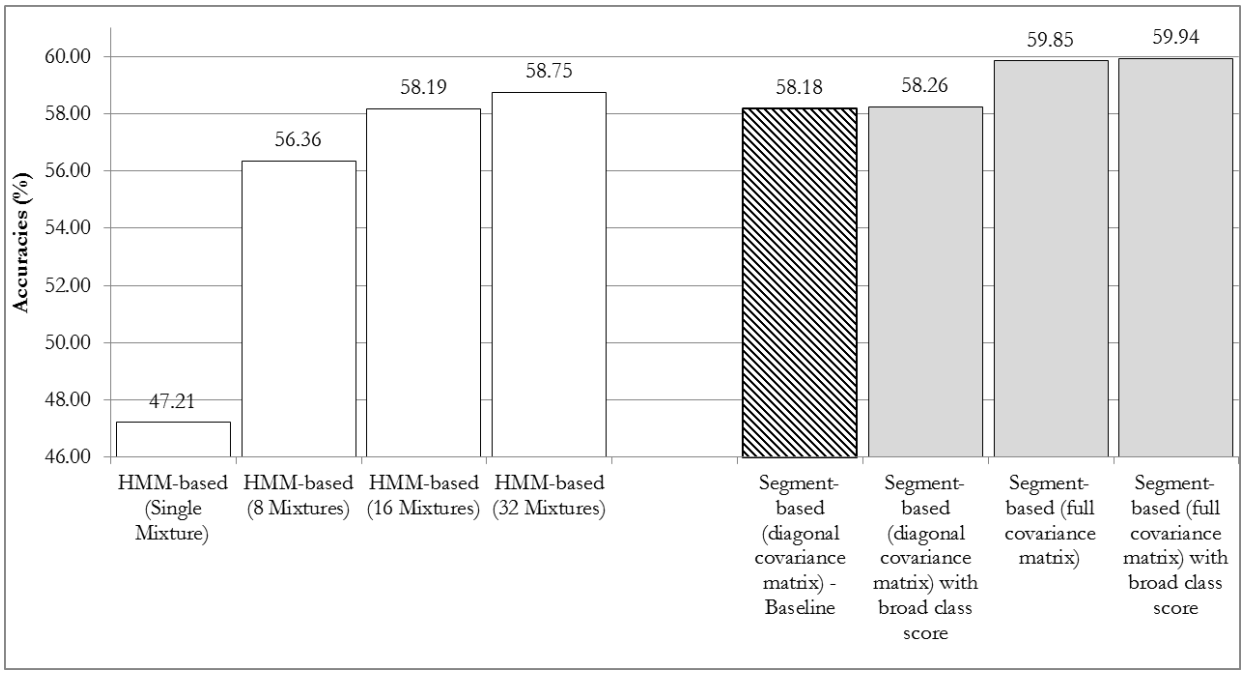

Fig. 13. Phoneme recognition accuracies comparison bar chart.

In this work, we explored the idea of restoring segments missing from the segment graphs due to deleted boundary errors, utilized broad class probabilities in scoring segments, and performed PCA dimensional reduction that enable the training of full covariance matrices despite limited training speech resources. Although the detection of incorrectly deleted boundaries did not introduced significant improvement, Figure 13 summarizes the improvement in phoneme recognition accuracies obtained from the other two aspects. The utilization of the broad class probabilities increases the 58.18% phoneme recognition accuracy of the diagonal covariance matrix segment-based case to 58.26%. Being able to train the full covariance matrices yields a 2.87% improvement (58.18% to 59.85%). Combining the two aspects, we obtained a 3.03% improvement (58.18% to 59.94%) over the baseline segment-based case. The best phoneme recognition accuracy is 1.19% higher than (or 2.02% improvement over)the 32-component GMM baseline frame-based HMM case.

Table 9. Resulting phoneme recognition accuracies on TIMIT.

Acoustic models	**% Accuracy**
Context-independent Frame-based (Monophone)	52.85
Context-dependent Frame-based (Triphone)	54.53
Segmental and Boundary	56.18
Segmental and Boundary (with inserted boundary elimination by both methods)	57.33

The results in Table 9 shows that the segment-based speech recognition with and without inserted boundary elimination outperform both of the frame-based context-independent (Monophone), and the context-dependent (Triphone) cases on the TIMIT dataset. It is important to point out that our phoneme recognition experiments are focused on studying the improvement of acoustic models. There were no higher-level constraints, such as lexical constraints and language models, deployed to help improving the overall resulting recognition accuracies. Only phoneme bigrams were straight-forwardly utilized in the decoding. Therefore, it is not entirely relevant to compare the recognition accuracies from the settings of our experiments with the recognition accuracy of TIMIT dataset using MIT's SUMMIT recognizer reported in [1].

The improvement when using inserted boundary elimination was 2.05% over the case when it was not used. This improvement percentage is not as large as the one evaluated on the LOTUS corpus because segment errors of the corresponding segment graphs in the TIMIT case were not as significant as the errors found in the segment graphs of the LOTUS case.

6. Conclusions and Future Works

This paper reported improvements in acoustic modeling of the first Thai segment-based speech recognition system on Thai phoneme recognition tasks, which usually yield phoneme recognition accuracies of less than 50% due the limited resource nature of Thai. In addition to the segment graph improvement proposed in our previous work, this paper contributes in three main aspects. Firstly, attempts to detect falsely deleted boundaries were implemented and tested in order to improve the quality of segment graphs so that the correct segment inclusion rate of the graph increased. However, deleted boundary detection did not yield significant improvement as in the case of the deletion of falsely inserted boundaries. The second aspect that was proposed in this work was the utilization of broad Acoustic-Phonetic class scoring, a procedure which could be added in a straightforward manner to the scoring of the decoding of the segment graphs. Such scoring methods could not be achieved in frame-based speech recognition approaches. The Acoustic-Phonetic scores were demonstrated to contribute to the overall recognition improvement. The third aspect was a demonstration of mitigating the insufficiency in speech resources of Thai, which has always been the most significant obstacles of achieving high pure phoneme recognition accuracies. In our case, segment and boundary models benefited from the feature vector dimension reduction provided by PCA. The best phoneme recognition accuracies of our segment-based framework yielded an improvement of more than 25% compared to the original segment-based system. Still, there were still rooms for improvement as pointed out in the earlier discussion. The fact that deletion of true boundaries accounted for 11.90% of the errors in the segment graphs indicated a possible future work regarding such the issue. More explicit extraction of Acoustic-Phonetic constraints such as formant tracks, segment duration modeling, and places of articulation could be incorporated to the scoring of the segment graphs. Furthermore, other modeling techniques with higher levels of model complexities such as Conditional Random Fields [30, 31] and Deep Neural Networks [32] could be experimented to replace the modeling with Gaussian Mixtures. Evaluating phoneme recognition performances based purely on the acoustic of the speech signals should be a good indicator to the performance of phoneme recognition tasks in which higher-level constraints cannot be used. Word recognition tasks utilizing this segment-based framework were left for our future works.

Acknowledgements

This research was supported by the Thailand Graduate Institute of Science and Technology (grant no. TG-44-09-088D).

References

[1] J. R. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, pp. 137–152, Apr. 2003.

[2] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, Feb. 1989, vol. 72, no. 2, pp. 257–286.

[3] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallet, and N. Dahlgren. (1990). *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus* [CDROM].

[4] K. Likitsupin, S. Tangruamsub, P. Punyabukkana, and A. Suchato, "Phoneme recognition from Thai continuous speech using a segment-based approach," in *Proc. The 11th National Computer Science and Engineering Conference*, Bangkok, 2007, pp. 218–222.

[5] K. Likitsupin, A. Suchato, P. Punyabukkana, and C. Wutiwiwatchai, "Improving segment-based speech recognition by recovering missing segments in segment graphs—A Thai case study," in *Proc. International Symposium on Communications and Information Technologies*, Vientiane, 2008, pp. 268–273.

[6] J. W. Chang and J. R. Glass, "Segmentation and modeling in segment-based recognition," in *Proc. The European Conference on Speech Communication and Technology*, Rhodes, 1997, pp. 1199–1202.

[7] S. C. Lee and J. R. Glass, "Real-time probabilistic segmentation for segment-based speech recognition," in *Proc. International Conference on Spoken Language Processing*, Sydney, 1998, pp. 1803–1806.

[8] S. C. Lee, "Probabilistic segmentation for segment-based speech recognition," M.S. thesis, EECS, MIT, Cambridge, MA, 1998.

[9] S. Kasuriya, V. Sornlertlamvanich, P. Cotsomrong, S. Kanokphara, and N. Thatphithakkul, "Thai speech corpus for Thai speech recognition," in *Proc. The Oriental Chapter of International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques*, Singapore, 2003, pp. 54–61.
[10] C. Wutiwiwatchai, P. Cotsomrong, S. Suebvisai, and S. Kanokphara, "Phonetically distributed continuous speech corpus for Thai language," in *Proc. The Third International Conference on Language Resources and Evaluation*, Las Palmas, 2002, pp. 869–872.
[11] J. R. Glass and V. Zue, "Multi-level acoustic segmentation of continuous speech," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, New York, 1988, pp. 429–432.
[12] V. Zue, J. R. Glass, M. Philips, and S. Seneff, "Acoustic segmentation and phonetic classification in the SUMMIT System," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, 1989, pp. 389–392.
[13] D. Wang, L. Lu, and H.–J. Zhang, "Speech Segmentation without Speech Recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, 2002, pp. 468–471.
[14] P. Leelaphattarakij, P. Punyabukkana, and A. Suchato, "Locating phone boundaries from acoustic discontinuities using a two-staged approach," in *Proc. International Conference on Spoken Language Processing*, Pittsburgh, 2006, pp. 673–676.
[15] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," Ph.D. dissertation, EECS, MIT, Cambridge, MA, 1995.
[16] P. Dareyoah, A. Suchato, and P. Punyabukkana, "A study of acoustic measurements for voicing detection in speech with room-level SNR," in *Proc. The Sixth Symposium of Natural Language Processing (SNLP 2005)*, 2005, pp. 109–114.
[17] A. Salomon, C. Y. Espy-Wilson, and O. Deshmukh, "Detection of speech landmarks: Use of temporal information," *JASA*, vol. 115, no. 3, pp. 1296–1305, Mar. 2004.
[18] A. Salomon and C. Y. Espy-Wilson, "Automatic detection of manner events based on temporal parameters," in *Proc. The Sixth European Conference on Speech Communication and Technology*, Budapest, 1999, pp. 2797–2800.
[19] T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for automatic detection of nasal manner," *Speech Communication*, vol. 43, no. 3, pp. 225–239, Aug. 2004.
[20] T. Pruthi and C. Y. Espy-Wilson, "Automatic classification of nasals and semivowels," in *Proc. The Fifteenth International Congress of Phonetic Sciences*, Barcelona, 2003, pp. 3061–3064.
[21] A. Juneja and C. Y. Espy-Wilson, "A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition," *JASA*, vol. 123, no. 2, pp. 1154–1168, Feb. 2008.
[22] A. Juneja and C. Y. Espy-Wilson, "Speech Segmentation using Probabilistic Phonetic Feature Hierarchy and Support Vector Machines," in *Proc. International Joint Conference on Neural Networks*, Portland, OR, 2003, pp. 675–679.
[23] A. Juneja and C. Y. Espy-Wilson, "Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning," in *Proc. the Ninth International Conference on Neural Information Processing*, Singapore, 2002, pp. 726–730.
[24] M. Tang, S. Seneff, and V. Zue, "Two-stage continuous speech recognition using feature-based models: A preliminary study," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, U.S. Virgin Islands, 2003, pp. 49–54.
[25] S. Borys and M. Hasegawa-Johnson, "Distinctive feature based SVM discriminant features for improvements to phone recognition on telephone band speech," in *Proc. The Ninth European Conference on Speech Communication and Technology*, Lisbon, 2005, pp. 697–700.
[26] C. C. Chang and C. J. Lin, (2012). *A Library for Support Vector Machines*, [Website]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm
[27] J. Kamaramen and P. Paalanen. (2003). *GMMBAYES—Bayesian Classifier and Gaussian Mixture Model ToolBox* [Source Code]. Available: http://www2.it.lut.fi/project/gmmbayes/downloads/src/gmmbayestb
[28] S. Young , G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, 3rd ed. Cambridge: Cambridge University, 2006.
[29] A. K. Halberstadt and J. R. Glass, "Heterogeneous measurements and multiple classifiers for speech recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Sydney, 1998, pp. 995–998.

[30] S. Boonsuk, A. Suchato, P. Punyabukkana, C. Wutiwiwatchai, and N. Thatphithakkul, "Language recognition using latent dynamic conditional random field model with phonological features," *Mathematical Problems in Engineering*, vol. 2014, pp. 1–16, Feb. 2014.
[31] N. Kertkeidkachorn, P. Punyabukkana, and A. Suchato, "A hidden conditional random field-based approach for Thai tone classification," *Engineering Journal*, vol. 18, no. 3, pp. 99–122, Jul. 2014.
[32] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.