ENGINEERING JOURNAL

Article

An Automatic Real-time Synchronization of Live Speech with Its Transcription Approach

Nat Lertwongkhanakool[a]**, Natthawut Kertkeidkachorn**[b]**, Proadpran Punyabukkana**[c,*]**, and Atiwong Suchato**[d]

Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand
E-mail: [a]Nat.L@Student.chula.ac.th, [b]Natthawut.K@student.chula.ac.th, [c]Proadpran.P@Chula.ac.th (Corresponding author), and [d]Atiwong.S@Chula.ac.th

Abstract. Most studies in automatic synchronization of speech and transcription focus on the synchronization at the sentence level or the phrase level. Nevertheless, in some languages, like Thai, boundaries of such levels are difficult to linguistically define, especially in case of the synchronization of speech and its transcription. Consequently, the synchronization at a finer level like the syllabic level is promising. In this article, an approach to synchronize live speech with its corresponding transcription in real time at the syllabic level is proposed. Our approach employs the modified real-time syllable detection procedure from our previous work and the transcription verification procedure then adopts to verify correctness and to recover errors caused by the real-time syllable detection procedure. In our experiment, we empirically customized the acoustic features and the parameters to ensure that there is no inconsistency between the speech and its transcription. Results are compared with two baselines which have been applied to the Thai scenario. Experimental results indicate that, our approach outperforms two baselines with error rate reduction of 75.9% and 41.9% respectively and also can provide results in the real-time situation. Besides, our approach is applied to the practical application, namely ChulaDAISY. Practical experiments show that ChulaDAISY applied with our approach could reduce time consumption for producing audio books.

Keywords: Automatic speech-text synchronization, syllable detection, real-time alignment, live speech and transcription alignment, transcription verification.

ENGINEERING JOURNAL Volume 19 Issue 5
Received 24 August 2014
Accepted 5 February 2015
Published 31 October 2015
Online at http://www.engj.org/
DOI:10.4186/ej.2015.19.5.81

1. Introduction

An automatic synchronization of audio and text plays a key role as a core technology for developing many kinds of applications, in which text and audio usually are presented together, such as subtitling of TV news [1, 2], alignment of lyrics and acoustic music signal [3], e-learning lesson [4] and Digital Accessible Information System (DAISY) audio books [5]. Such applications could utilize the automatic synchronization of text and audio in order to align between transcription and speech. Currently, the trend of the automatic synchronization of audio and text merely focuses on a specific method of the synchronization at an utterance level and also determine only the condition, in which audio and text utterances have been already prepared availably [6–9].

Although many techniques are introduced for the automatic synchronization of text and audio at utterance levels, some applications require the automatic synchronization of text and audio in a finer level than an utterance level. This requirement has been discovered, while working on a project, namely ChulaDAISY [10]. ChulaDAISY is an application tool which aims to automatically generate audio books in DAISY 3 format [11] by gathering audio speech of volunteers who read transcriptions or contents in a book through the ChulaDAISY application. After the construction process of DAISY audio books has been finished, a user then can listen to audios and also will be able to navigate through an audio book reader, e.g. AMIS [12]. The audio book reader would play utterances together with highlight texts corresponding to those utterances because the binding between audios and texts are done at utterance level. In most languages, in particular English, an utterance is generally a sentence or a part of a sentence which is separated by spaces or punctuation marks. Therefore, the binding between audio and text at utterance does not cause any problem. However, the synchronization of text and audio in the utterance level is not suitable for Thai since sentences in Thai are constructed by consecutive words that are written continuously without any spaces or punctuation marks that could identify the end of sentences [12]. Even though white spaces could appear in some cases, they generally depend on decisions of writers to put them in order to emphasize or separate phases to match their desirable meaning. Consequently, in the audio book construction process, it is difficult for a user to read utterances without considering what they actually should read for each utterance. The workaround is to define the process for constructing DAISY audio books in Thai by giving users to prepare the transcriptions which they will read manually. The preparation of the transcription could be done by placing or removing white spaces and punctuation marks in suitable positions in order to separate or combine utterances before each utterance is read. According to our practical experience, the process for preparing transcriptions is time-consuming and methods of preparing transcription are varied by one user to another. Due to the nature of sentences in Thai that lack clear separation marks between sentences, the most appropriate approach to cope with the Thai nature problem is to take words or syllables level into account instead of considering utterances.

Furthermore, ChulaDAISY application aims to assist users to read books as natural as they can. Hence, the automatic synchronization of given transcription and live speech in a real-time situation is promising, since users do not have to worry whether they separated utterances suitably. In the same way, it also means that the manual text preparation process can be eliminated from the audio book construction process.

Several studies for automatic synchronization of transcription and speech at the utterance level mostly concern in the non-real-time situation. In our previous work [14], an algorithm for the automatic synchronization of live speech and its transcriptions in the real-time situation at the syllabic level is introduced. Although our previous work provides reasonable results, the results still are some drawbacks of our previous approach because it only uses audio data to calculate. Hence, this approach could not apply for some practical applications. In our previous approach, when an error of the synchronization between text and audio occurred, the error is accumulated. As a result, it would cause errors to later positions even though the algorithm could align later text and later audio correctly. To make it more robust, we therefore improve our previous work in order to handle cumulative errors by revising correctness of the alignment between audio and text using a more robust algorithm, a transcription verification approach. We also show a case study of ChulaDAISY that applies this algorithm in practice.

The rest of this article is organized as follows. In section 2, the background knowledge on Thai sound system is given. In the following section, related works to our problem in the automatic synchronization of transcription and speech are reviewed and discussed. Section 4 and 5 respectively describe the baselines and the proposed method of synchronization. Details of experiments are presented in the following section. A case study of ChulaDAISY is explained in section 7. Finally, the results and conclusions are shown in the last section.

2. Background Knowledge

Thai syllable structure comprises of three units: 1) Consonantal unit, 2) Vowel unit and 3) Tonal unit as described in Fig. 1.

Syllable		
Onset	Rhyme	
Onset	nucleus	Coda
initial consonant	vowel	final consonant

Fig. 1. Thai syllable structure diagram.

According to the syllable structure in Thai, the consonantal unit can appear two positions in a syllable. One is at beginning of syllable, called an initial consonant (C_i) at the onset period, and another one appears at the ending of syllable, referred as a final consonant (C_f) at coda period. In Thai there are 33 consonants which could appear at the initial position. In 33 consonants, 21 consonants are non-cluster consonants (C_i), while 12 consonants are cluster consonants (C_i(C_i)). For final consonants, there are 9 possible consonants which could manifest at the coda period.

For vowel in Thai, there are two categories of vowel units (V) including monophthongs and diphthongs. Monopthongs have 9 differences in qualities and each of them includes two differences in quantities relating to their duration: lax and tense. In case of diphthongs, there are three sounds, each of which also has two quantitative differences similar to monopthongs.

For tone unit (T), there are five tones in Thai including the mid tone, the low tone, the falling tone, the high tone and the rising tone. Thai syllable structure diagram in Fig. 1 does not show the position of tone unit in the syllable, since tone unit could span along the syllable duration.

Generally, a syllable is a sequence of speech having a maximum or peak of inherent sonority between two minima of sonority [15]. This maximum or peak of inherent sonority usually find in a nucleus portion which is typically a vowel unit or a sonorant consonant like the sound /l/, /m/, /n/and /r/ [16].

In writing, syllables can be extracted through their graphical unit using predefined structures or grammars. In Thai, the structure of syllable is formed as in Eq. (1) [17],

S = (C_i (C_i) V(V)(C_f))^T (1)

where S is syllable, C_i is initial consonant, C_f is final consonant, V is vowel and T is tone.

3. Related Work

Related works are roughly divided into three research fields: 1) Speech-Text Alignment, 2) Syllable Detection and 3) Transcription Verification. The details of research works in each field are described as follows.

3.1. Speech-Text Alignment

Most of speech-text alignment techniques are invented to apply for an application which presents text and speech simultaneously, such as a subtitling system and a lyrics synchronization system. The first purpose for developing speech-text alignment is to assist disabilities; in particular hearing impaired person, in order to make them can access broadcast information easily. In order to perform speech-text alignment, the conventional method intends to use transcribers. It clearly leads to a human error problem, since amount of speech and text, which needs to be aligned, is increased swiftly and continuously. Ando [2] therefore introduced an Automatic Speech Recognition (ASR) system for transcribed the audio. Later, there were studies in [6, 7, 18, 19] that investigated and introduced the automatic speech-text alignment. However, these studies still suffered from errors of transcription, since transcriptions of these texts were produced by ASR systems. Thus, to avoid such problem in our task, we presume that the transcription of text is already correctly available. The study that assumed the similar assumption was proposed by Gao [20]. Gao introduced the automatic real-time alignment between live speech and text by detecting the end time of speech utterances. The end time detection task is the problem to identify positions where the end of speech utterances happened. If the end time was detected, the system would align the segment of transcription

DOI:10.4186/ej.2015.19.5.81

with corresponding speech segment. The frame-synchronous likelihood ration test technique was applied to verify the end of speech segments. Gao's approach obtained accuracy of 85.6% at error range 0.5 second. Although the Gao's approach gave reasonable results, the alignment was done at the utterance level. In our work, we aim to focus on the alignment at syllable level because it is more suitable for Thai.

3.2. Syllable Detection

Several studies in the syllable detection mostly concerned about energy values of speech signals in order to indicate positions of syllables and separate them into segments. The first algorithm for detecting syllable was proposed by Mermelstein's study [15]. Since a syllable position is usually in ranges' of energy values of speech signals are maximized, the convex hull algorithm could be applied to detect that part in the power spectrum intensity representing energy values of speech signal. The highest value in the power spectrum intensity is regarded as the landmark for the nucleus of a syllable, while minimum values before and after the maximum value in the power spectrum intensity are expected to be boundaries of the syllable. In Mermelstein's study, the power spectrum intensity band between 500 Hz to 4000 Hz was considered. Then, the convex hull algorithm was computed on such frequency band in order to find boundaries between all syllables. Later, Pfitzinger [21] presented a new frequency band instead of the frequency band between 500 Hz to 4000 Hz introduced by Mermelstein. In their analysis, the frequency band between 500 Hz to 4000 Hz was mostly affected by voiced consonants or noise in environment due to characteristic of their energy. As a result, Pfitzinger presented two new frequency bands for the syllable detection. The first frequency band is the range between 250 Hz to 2500 Hz, whereas the second frequency band is the band between 7 Hz to 13 Hz. Their approach yielded an error rate of 12.87% for the reading speech condition, while provided error rate of 21.03% in the normal speech case. Juneja [22] also introduced distinctive feature sets including the frequency band for the syllable detection. The frequency band proposed by Juneja consisted of two frequency bands: between 640 Hz to 2800 Hz and between 2000 Hz to 3000 Hz. Although their study showed good results, they still suffered with phonetic ambiguity as reported in [21]. For example, the word "support" has two syllables while the word "sport" has only one syllable, even though their pronunciations are similar. In contrast with Thai, Thai syllable is a well-defined unit since most syllables in Thai are composed by one vowel. Consequently, vowels could be assigned as the nucleus of syllable so the syllable detection task is similar to the vowel detection task in case of Thai.

In Thai, Dareeyoah [23] used an autocorrelation method for measuring periodicity of speech segments together with the conventional convex hull algorithm in the frequency range in which frequencies less than 300 Hz were filtered. They gained the correctness of vowel detection at 84% accuracy based on the large vocabulary Thai continuous speech recognition corpus (LOTUS). When comparing Dareeyoah's study with Howitt's study [16] which assigned the frequency band between 300 Hz to 900 Hz only, the results indicated that Dareeyoah's results outperformed Howitt's results. Howitt's approach yielded the correctness of 75% on the same corpus. Recently, Rochkittichareon [24] introduced distinctive features for identifying an acoustic landmark of Thai using the same method as Juneja's study but appropriately adjusted some parameters for Thai. They not only chose two frequency bands of energy between 640 Hz to 2800 and between 2000 Hz to 3000 Hz but also selected the frequency band of sound intensity between 0 Hz to 900 Hz. They used a ratio of frequency bands of energy between 0 Hz to 400 Hz and between 400 Hz to 6000 Hz to represent acoustic features. The results of this approach gained correctness of 84.47% on the LOTUS speech corpus and also had been reporter that the performance of Rochkittichareon approach provided better results than Dareeyoah's study. Even though the solution for Thai vowel detection acquired the results up to 85%, they still cannot perform the vowel detection in real-time situation.

3.3. Transcription Verification

As reported [23, 24], the vowel detection algorithm for Thai yielded the average of correctness more than 80%. However if there were errors occurred, it would be cumulative since the alignment between speech and text is done continuously. As a result, the error rate will become more and more increased. To avoid this situation, the transcription error detection algorithm could verify the correctness of transcription by not only checking correctness of the syllable level but considering also upper levels of syllable such as words, sentences and utterances in order to guarantee correctness and adjust the alignment position more correctly. There are many studies involved in the transcription verification technique [25, 26]. The study in [25] utilized the transcription verification for developing a speech corpus and verifying correctness of the speech corpus. Jiang [26] introduced three applications based on the confidence measure technique. The first application, namely Combination of Predictor Features, was to use confidence measure score to determine whether the result was correct. Secondly, the confidence measure as posterior probability typically used to examine reliabilities of speech recognition results acquired from Hidden Markov Model (HMM)-based techniques during the decoding period of N-best choice. The third application was to use confidence measure score for the utterance verification task. In this approach two hypotheses were given. One was null hypothesis (H_0) meaning that the interested utterances were corresponded to the considered transcription, whereas the alternative hypothesis (H_1) assumed that the utterances and the transcription were not consistency. Then H_0 or H_1 was chosen by considering the confidence measure score which hypothesis should be accepted.

The reviews above had been showed that there were still lacks of a syllable in the real-time situation in Thai which provided high results. We therefore aimed to propose the algorithm for syllable detection in the real-time situation by considering the vowel detection algorithm in Thai together with the transcription verification, in which all correct transcription of all text is known, using the confidence measure score in order to reduce cumulative error occurred from the error of syllable detection algorithm.

4. Baseline

Two systems are chosen as baselines for automatic synchronization of speech and transcription. The details of each baseline are as follows.

4.1. Baseline 1

For Baseline 1, we considered the work from [14] and found its weakness in landmark detection. In that work, the audio stream was continuously passed into the Real-Time Syllable Endpoint Detection module that extract the feature representation of each speech frame to determine whether the current speech frame was the endpoint, which means that the current syllable transcription has already been spoken. Unless the endpoint was successfully found, the segment would be sent to the Landmark Detection module to locate the detected candidate peak of the syllabic nuclei to identify number of spoken syllables from such segment.

Since the candidate peak was immediately detected as the location of the syllabic nuclei, it could have introduced some discrepancies. Therefore, in this work we adopt the SVM-based approach [23, 24] to classify vowels in speech signals, by scoring peak module to detect the actual syllable nuclei.

In the SVM-based approach, two acoustic feature sets were investigated. The first acoustic feature set introduced by Dareeyoah [23] consists of the periodicity value computed by the maximum value of autocorrelation between 60 Hz to 320 Hz and the intensity value of speech signals, that high-pass filtered at 300 Hz. The second set was presented in Rochkittichareon's study [24], where the intensity value of speech signals with the frequency band between 640-2,800 Hz, the intensity value of speech signals with the frequency band between 2,000 to 3,000 Hz, the maximum intensity value of the frequency below 900 Hz and the ratio of the intensity value with the frequency below 400 Hz to the frequency band between 400 to 6,000 Hz were used as acoustic features.

The parameters in this preliminary experiment consist of four parameters, the window length, the interval, the energy threshold and the dip threshold. The acoustic features as mentioned above are extracted with the window length at 25 msec and the interval at 15 msec. The energy threshold and the dip threshold of the convex hull algorithm are set differently for each approach according to the training data. The previous work [14] set the parameters to 60 dB and 10 dB respectively. Our proposed SVM-based approach with the first acoustic feature set the parameters to 56 dB and 6 dB respectively, whereas the SVM-based approach with the second acoustic feature set the parameters to 60 dB and 8 dB respectively. There are 440 speech utterances, randomly selected from the PD set as the training data, while the rest of the data in the PD set is the test data. Experimental results are shown in Table 1.

Table 1. The results of the real-time syllable detection procedure.

Approaches	Acoustic Feature Sets	Accuracy
Our Previous work [14]	-	82.02%
The SVM-based Approach	Set 1	84.70%
	Set 2	82.65%

The results in Table 1 showed that the SVM-based approach with the first acoustic feature set outperformed other approaches. In this article, the SVM-based approach with the first acoustic feature set, therefore, is chosen as baseline 1. The major difference between the baseline 1 and our approach in this article is that the transcription verification procedure is augmented in order to reduce the cumulative errors occurred by false alignment so that the errors would not be dramatically increased and also could remedy the alignment problem.

4.2. Baseline 2

For Baseline 2, Gao's approach [20] is considered. In Gao's method, they verified the transcription continuously with incoming speech utterances until the end of speech utterances. In their verification process, the end time detection technique was employed. The end time detection task aimed at identifying positions where the end of speech utterances occurred. If the end time was detected, the system would align the segment of transcription with corresponding speech segment. Gao employed frame-synchronous likelihood ratio test technique to verify whether the current speech segment was the endpoint.

However, Gao's work was conducted at utterance level, while ours focuses on syllabic level. Therefore, in our experiment, we apply Gao's approach and adjust the synchronization of live speech and transcription to the syllabic level based on speaker rate. Our data set suggests that a normal speaker could speak three syllables per second. The speaker rate assumption is utilized to generate possible hypotheses for the transcription verification task. The verification is performed every second in order to select the best hypotheses so that the transcription and speech would be aligned. This approach is similar to the transcription verification in our approach. Nevertheless, the number of syllables for creating the hypotheses is assumed based on the speaker rate because we need to investigate only the ability of transcription verification procedure without the assumption of the boundary of spoken syllable. Consequently, one second is set to be the trigger to start the process of this approach.

5. The Proposed Method

The synchronization of live speech and transcription consists of two mainly procedures: the Syllable Detection procedure and the Transcription Verification procedure, as shown in Fig. 2. According to the process in Fig. 2, audio stream is segmented into small speech frames and then passed to the syllable detection procedure in order to detect a number of syllables that correspond to speech frames on the real-time situation. After that, the alignment between the number of detected syllables and the text transcription is updated by changing the position of the pointer that indicates the position of corresponding speech and text to the correct position and repeats the procedure continuously until the end of the synchronization process. When the syllable detection procedure is done on one period, the transcription verification procedure is applied in order to verify the correctness of the alignment. The transcription verification procedure is necessary to reduce cumulative errors of the syllable detection procedure as discussed in Section 3. The details of each procedure are described in the following subsections.

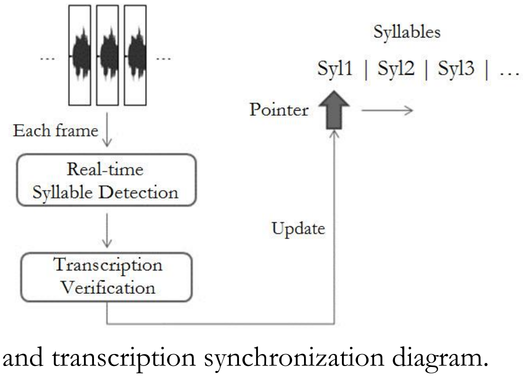

Fig. 2. The process of speech and transcription synchronization diagram.

5.1. Real-time Syllable Detection

There are four necessary modules in the real-time syllable detection procedure: 1) Feature Extraction, 2) Real-Time Syllable Endpoint Detection, 3) Landmark Detection and 4) Scoring Peak, as illustrated in Fig. 3. Firstly, the audio stream is segmented into small speech frames, and then the speech frames are passed to the feature extraction module continuously in order to extract acoustic feature vectors from a speech frame. After that, the acoustic feature vectors are conveyed into the real-time syllable endpoint detection module to determine whether the corresponding speech frame of this acoustic feature vector is an endpoint of syllable. In case that the endpoint of the syllable is detected, it could be concluded that a syllable had been spoken. Then, the pointer indicating the position of the alignment between speech and text should be updated to the next syllable. Although the syllable detection using the endpoint detection algorithm yields reasonable results, it still could miss to detect some syllables because some endpoints might not be detected. To reduce such the error, the acoustic feature vectors extracted from the speech frames starting from the last endpoint to the latest endpoint are kept and conveyed into the landmark detection module.. This module is designed to locate the other landmarks of the syllables in the segment. Before the location is accepted to be the landmark of the syllable, the scoring peak module is applied to compute the landmark score in order to verify the correctness of the detected landmark. After all landmark positions are confirmed, the pointer indicating the alignment position is adjusted to the new position according to the number of landmarks or detected spoken syllables.

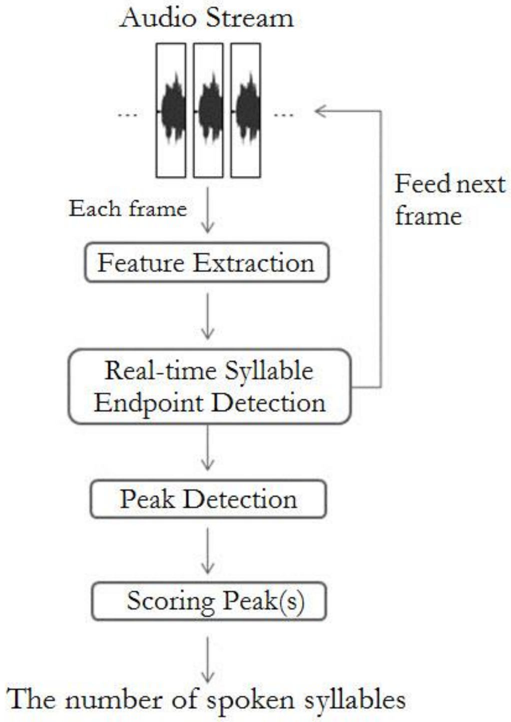

Fig. 3. The real-time syllable detection procedure diagram.

5.1.1. Feature extraction

In the feature extraction module, an acoustic feature vector is extracted from each speech frame. The location of the syllable could be detected by loudness due to the aspect of syllables or vowels that could be found at the loudest point of a syllable. The sound intensity, which is the loudness function of speech, is therefore chosen to represent an acoustic feature in our task. To capture the sound intensity for each speech frame, the band-pass filter range between 640 Hz – 2800 Hz is set up because the intensity of this band is not affected by noise in the high frequency band caused by environments or fricative voiced. Consequently, this intensity can represent the characteristic of syllables clearly as reported by [27]. After speech frames are passed into that band pass filter, the short-time energy algorithm then applies to compute the intensity of each speech frame. Besides for each speech frame, which is delivered to the feature extraction module, the window size is set at 25 msec. while the overlap period between each speech frame is assigned at 10 msec.

5.1.2. Real-time syllable endpoint detection

Each of speech frames extracted by the feature extraction module is tracked by the real-time syllable endpoint detection module in order to detect the endpoint of the syllable. The real-time syllable endpoint detection module determines whether the intensity of a speech frame is greater than the setup threshold. If the intensity is more than the threshold, still, it will be considered to be syllable duration. Otherwise, it is assumed that it is an end point or a starting point of a syllable. According to the example shown in Fig. 4, the starting point of syllable is at the point A, at which the intensity of a speech frame is greater than the threshold. Based upon our preliminary experiment on the training dataset, we define the threshold at 56 dB, since it provides the best results as reported in our previous work [14]. As we mentioned, the endpoint should be at the point B, at which the intensity of a speech frame is less than the threshold. Although the assumption of threshold for defining the endpoint is reasonable, it is still inaccurate in some practical cases.

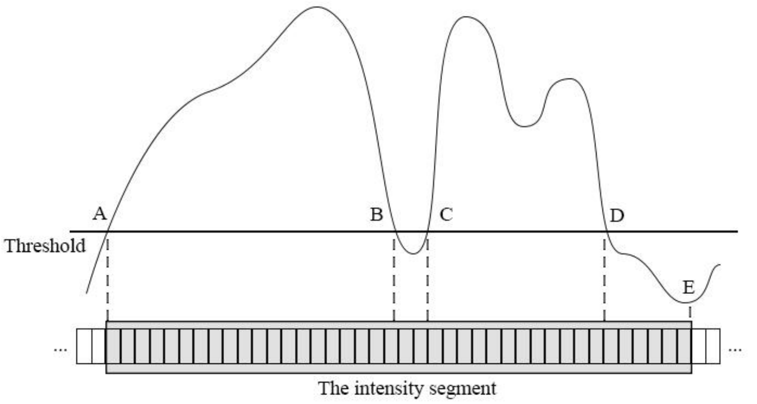

Fig. 4. The example of endpoint detection.

Considering the example in Fig. 4, we can notice that the middle of the syllable between the point B and the point C is less than the threshold in only a few frames. Consequently, it leads to detect the incorrect endpoint of syllable because the intensity of speech frames drops dramatically due to some factors, such as noises and the speech production condition, and then the intensity of speech frames raises above the threshold again in a few speech frames later. To avoid the situation that the intensity of a few speech frames significantly decreased for a while, we therefore take neighbor frames of such speech frames into account by considering the next six consecutive frames in order to determine whether it is the real endpoint. As in an example with this strategy, the point E is considered as the endpoint instead, whereas the point B and D are not. Additionally, not only the endpoint is determined in this module, but also the intensity segment of each speech frame is gathered from the starting point to the endpoint. As in Fig. 4, the intensity segment will be conveyed to the next module from the point A to E.

5.1.3. Landmark detection

To reduce the miss detection of the syllable endpoint detection, the intensity segment from the last module is forwarded into the landmark detection module for locating the landmark of syllable, after the endpoint is detected. The landmark detection is necessary because all of syllable landmarks or syllabic nuclei can be verified in the segment. In the same way, the number of syllables in the segment can be computed. To locate the landmark of syllable, Mermelstein's convex hull algorithm is chosen [15]. The convex hull algorithm is proposed to find the maximum and the minimum of an interested period. In our case, the interested period is the intensity segment. The maximum sometimes refers to the peak. In syllable detection task, the peak of intensity is considered as the candidate syllabic nucleus. The process of convex hull algorithm is performed by the following steps. Firstly, Dips of the convex hull are computed by the difference between the maximum and the minimum of the intensity of the intensity segment. Then if the dip is greater than the threshold's dip, the convex hull algorithm would be divided the considered segment into sub segments and recursively calculate dip on every new generated sub segments until it could not be segmented. Eventually, peaks of sub segments, which could not be divided, are locations of syllable nuclei.

5.1.4. Scoring peaks

After obtaining the locations of the syllable landmarks in the intensity segment, each candidate landmark is then verified whether it is the actual syllable nucleus of syllables by using a statistical machine learning approach. The support vector machine (SVM)-based classifier is applied to classify landmarks relied on two acoustic features. One is the intensity of the landmark location that is passed into the high-pass filter frequency band at 300 Hz. Another acoustic feature is obtained from the maximum of the autocorrelation value between 60 Hz to 320 Hz in the frequency band below 900 Hz as Dareeyoah's work [23] suggested.

After determining all actual syllabic nuclei from all candidates, the pointer indicating the alignment position is adjusted its position to the new position according to the number of the actual landmarks or detected spoken syllables. This result will be the input information for the next approach, transcription verification.

5.2. Transcription Verification

The syllable detection procedure has a chance to miss detecting some syllables in the real-time situation due to the limitations mentioned in [14]. The transcription verification procedure therefore is necessary to verify and correct such errors. In the transcription verification procedure, a speech segment gathered by the syllable detection procedure is considered simultaneously together with hypothetic transcriptions in order to improve performances of the synchronization. The transcription verification procedure consists of the following steps as presented in Fig. 5.

The segment of speech frames from the previous procedure is passed into the MFCCs extraction module to extract acoustic feature vectors, while the number of syllables, which are computed by the syllable detection procedure, is adopted in the hypothesis generation module to generate possible hypotheses. After extracting the acoustic feature vectors from the speech segment and generating possible hypotheses, the force alignment module is performed in order to compute a confident score of each hypothesis. Eventually, the scoring module is to rank the hypotheses according to their confident score and then the pointer of the alignment between speech and text is adjusted from the last considered position of the syllable detection procedure as the next new considered position according to the highest score hypothesis.

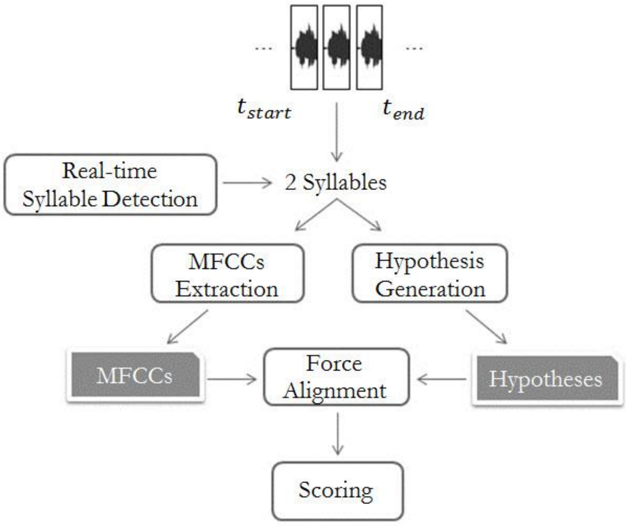

Fig. 5. The transcription verification procedure diagram.

5.2.1. MFCCs extraction

In the MFCCs extraction module, acoustic features are extracted from speech signals in order to represent the characteristic of speech signals. The standard Mel-frequency cepstral coefficients (MFCCs) are employed. The MFCCs with their derivative consisting of the delta of MFCCs and the acceleration of MFCCs are computed as well. The deltas of MFCCs are differences between the MFCCs in each time position, while the accelerations are differences between the deltas of MFCCs. The degree of MFCCs is set at 13; in consequence, the acoustic feature vectors are represented by MFCCs with their derivative having 39 dimensions. Those acoustic feature vectors are computed from the speech signal every 10 msec with the window size at 25 msec. After calculating the 39 MFCCs, the acoustic feature vectors are put into the force alignment module.

5.2.2. Hypothesis generation

The hypothesis generation module generates possible transcriptions of text, which are likely to match with the speech segment. To build possible hypotheses, the hypothesis generation module takes the number of syllables predicted by the real-time syllable detection procedure together with the entire transcription and then expands boundaries of hypotheses. Given the number of expanding hypotheses (*n*), the hypothesis generation module would create *2n+1* hypotheses. For example, given a number of expanding boundary hypotheses at 1, three hypotheses therefore are created as shown in Fig. 6.

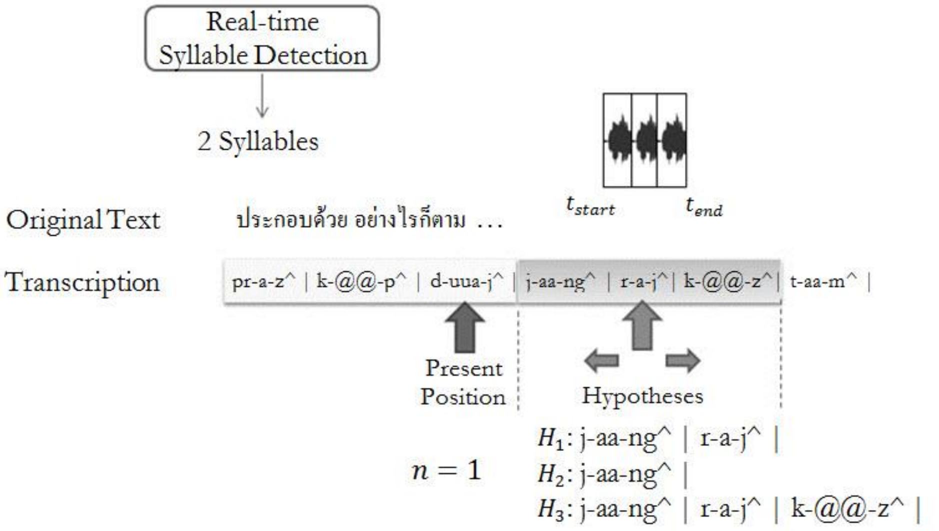
Real-time Syllable Detection
2 Syllables
t_{start}
t_{end}
Original Text
ประกอบด้วย อย่างไรก็ตาม ...
Transcription
pr-a-z^ | k-@@-p^ | d-uua-j^ | j-aa-ng^ | r-a-j^ | k-@@-z^ | t-aa-m^ |
Present Position
Hypotheses
$n = 1$
H_1: j-aa-ng^ | r-a-j^ |
H_2: j-aa-ng^ |
H_3: j-aa-ng^ | r-a-j^ | k-@@-z^ |

Fig. 6. The example of generating possible hypotheses for alignment when setting *n* at 1.

As illustrated in Fig. 6, the first hypothesis is the number of syllables that is similar to the number of syllables computed from the real-time syllable detection procedure. In the second hypothesis, the number of syllables of the hypothesis is less than the first the hypothesis by one syllable. The number of syllables of the third hypothesis is more than the number of syllables of the first hypothesis by one syllable. After knowing the number of syllables for each hypothesis, the hypothesis generation module matches amount of syllables to the corresponding transcription and then generates the transcription for each hypotheses.

Based upon this idea and preliminary experiment, the correctness of the synchronization between speech and text could be verified more firmly if longer speech segments are processed. The longer segment obtained, the more information acquired. Therefore, the number of previous segments of speech and transcription could be combined with the present segment to make the considered segment longer for *n* previous segments as well. The result of each hypotheses will be more precisely but more time-consuming.

5.2.3. Force alignment

After obtaining the acoustic feature vectors and the possible hypotheses, the force alignment is performed. The acoustic feature vectors and the possible hypotheses are input into the force alignment module in order to compute log-likelihood scores for each hypothesis so that those scores will be ranked in the scoring module. The force alignment of an ASR system is to restrict and force the possible output of recognition results to the given hypothesis in order to acquire likelihood scores of matching between an acoustic input and its corresponding transcription. The force alignment also provides boundaries of transcription. Each acoustic model is computed how likely between the acoustic model and the speech signal are and then each log-likelihood score is combined in the decoder step to find the exact score that transcription and speech are matched. For constructing the force alignment module, the Hidden Markov Model (HMM)-based approach is chosen.

5.2.4. Scoring

The log-likelihood scores obtained from the force alignment module are used to select the best hypothesis in the scoring module. The log-likelihood score of each hypothesis is used to order the rank of each hypothesis. The higher rank it is, the more likely it will be the correct hypothesis. After the ranking step, the best hypothesis among other hypotheses is selected. Then the pointer indicating position of alignment is updated by using the number of syllables of the best hypothesis in order to adjust the position of alignment toward the correct position, which could reduce the cumulative error from the real-time syllable detection procedure alone.

6. The Details of Study

In this section, details' study of our synchronization method of speech and text is reported. Beginning with the setting of experiment, details of the experiments and results are presented respectively. Eventually, the discussion about our method is discussed. Each detail is described as follows.

6.1. Experimental Setup

In the experimental setup, there are three parts: the data set, the baseline and the evaluation method. The data set describes the data set using in the experiment. Then, baseline introduces baseline systems in the experiment in order to compare their results with our approach. After that, the evaluation method is presented to be a standard for assessing the performance of the systems.

6.1.1. Data set

The Large Vocabulary Thai Continuous Speech Recognition Corpus (LOTUS) [28] is selected as the data set in the experiment. The LOTUS corpus consists of four major data set including Phonetically Distribution (PD) set, Training (TR) set, Development Test (DT) set and Evaluation Test (ET) set. In our experiment we only chose two sets of the LOTUS, the PD set and the TR set, for conducting experiments. Therefore, we have 1260 utterances on the PD set, while acquired 3007 utterances on the TR set.

6.1.2. Evaluation method

To measure the error of the synchronization of transcription and speech, the measurement, temporal aberration by [6], is applied and adapted to describe how the synchronization is at a moment. Nonetheless, in our approach which is performed in the syllabic level, the temporal aberration is changed to be the error aberration. The result will be tested with the reference time-aligned transcription on every endpoint of each segment. Similarly, when the pointer adjusts to the next syllable position, called phrase (P_i), the error aberration at P_i is computed by the difference of a number of syllables between the reference transcription and the result from the beginning to the ending time of the current phrase of transcriptions as shown in Eq. (2). In Eq. (2), S_{Pi} denoting cumulative syllables are counted from the start of the spoken utterance until the current phrase (Pi). S_{Ri} is a number of exact syllables counted from the reference transcription from the beginning to the current ending time as same as the S_{Pi}. The error aberration at phrase Pi (E_{Pi}) can calculate from the difference between S_{Pi} and S_{Ri} for each phrase.

$$E_{Pi} = S_{Pi} - S_{Ri} \tag{2}$$

The E_{Pi} value shows that at the current phrase where the speech utterance is spoken, the alignment between speech utterance and transcription is in which states. Three states: the deletion state, the insertion state and the correct state, could be determined by the characteristic of the E_{Pi} value. The deletion state happens when the E_{Pi} value is less than zero. It could infer that the alignment position is slower than the reference position. In case that the E_{Pi} value equals zero, it could refer to the correct state, which means the synchronization is correctly aligned. If the E_{Pi} value is greater than zero, it is in the insertion state, in which the alignment position is faster than the reference position. The example of the error aberration results illustrates in Fig. 7.

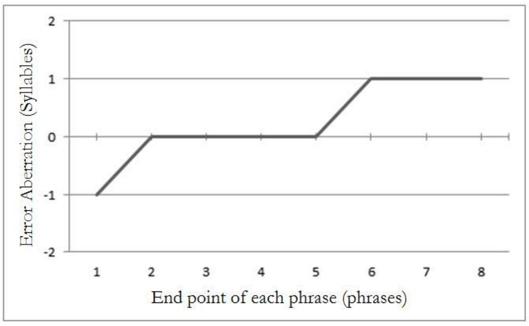

Fig. 7. The example of the error aberration (E_{Pi}) value at every phrase compared between the reference and the result.

After computing the E_{Pi} value for each phrase of the sentence (*S*), the total error aberration ($E_{err}(S)$) of the reference time-aligned transcription of *S* is calculated in order to count the absolute total number of E_{Pi} via Eq. (3), where *n* is a number of phrases in the *S*. For example, the total error aberration value is at 4 for Fig. 7. Furthermore, the reference deviation ($R.D.$) is also computed by Eq. (4) to measure the deviation from the reference line. The reference line is a case where E_{Pi} value is zero. In the Eq. (4), the E_{Pi} value is the error aberration of each phrase of sentence *S* consisting of *n* phrases.

$$E_{err}(S) = \sum_{i=1}^{n} |E_{Pi}| \tag{3}$$

$$R.D. = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_{Pi})^2} \tag{4}$$

Since every syllables consists vowels as the nucleus of the syllable structure in Thai language, thus, to count syllables for our method and the baseline methods, we define vowels as a representative of each syllable.

In the experiment, time processing using for the synchronization is also considered because the condition of the synchronization in our scenario is in the real-time situation. The time processing is measured by taking the elapsed time into consideration for every phrase and then estimating the average processing time relied on every processing phrase.

6.2. Experiments and Results

In the experiment, our approach is evaluated and is compared with the baseline 1 and the baseline 2 on the PD set of LOTUS. First The TR data set is used for training the acoustic models for the transcription verification procedure. The HMM-based model with five states and the transition from left to right is selected to model acoustic parameters. The model is constructed by not considering neighbor context, referred to mono-phone, and is estimated probability by one Gaussian mixture model with the 39-dementional MFCC vectors. The PD data set in the experiment is divided into two data sets. One is the data set for development set consisting of 30% in the PD data set or 504 utterances both male and female. The development set purposed for training two parameters in the transcription verification module which are the numbers of previous segments and the number of expanding hypotheses as describe in the hypotheses generation module. Another set consisting of the rest of utterances in PD data set (1176 utterances or 70% of the data set) is used to test our approach and the baselines. All Thai phonemes of transcriptions using in the experiment is converted into C form and V form, where C stands for consonants and V denotes vowels, because we intend to determine whether the position of the synchronization is in the correct reference portion without considering their lexical meaning.

To conduct the experiment, parameters are set for our approach and baselines as follows. In baseline 1, the threshold of intensity for determining the end point of syllable is set at 56 dB and neighbor speech frames, of which the intensity unexpectedly drops, are configured at six consecutive frames according to the preliminary experiment in our previous work [14]. Also the score peak module's parameters are set as described in section 5.1.2.

For baseline 2, the HMM-based model was trained as mentioned above on the TR data set, in which the data do not get involved in the PD set. Other configurations are the number of expanding hypotheses and the number of previous segments for reconsidering the alignment. Based on our preliminary experiment in the development set, the number of expanding hypotheses is set at 3 whereas the number of previous segments is configured at 5.

For our approach, the configuration of the real-time syllable detection procedure is similar to baseline 1. However, some setup for the transcription verification procedure is different from baseline 2. In our approach, the expanding hypotheses parameter is set at 1 which had been reported as the best result for our preliminary experiment in the development set.

Since the experiment is conducted on the existing speech data, a speech audio is split into small segments and those segments are streamed continuously into the synchronization process of approaches on every 125 msec. until the end of audio in order to imitate the live-speech situation. This scenario is similar to gather the speech data from the real microphone with 2000 frame buffer size.

Results of the experiment are reported in Table 2. The results indicate that our approach provides total error aberration less than other baselines with the reasonable time processing.

Table 2. The result of experiment on automatic real-time alignment of speech and text.

	Baseline 1	Baseline2	Our Approach
Total Deletion	4031	1216	1363
Total Insertion	5895	1229	1068
Total Error Aberration	9926	2445	2431
Total Error Aberration per number of phrases	686.20	284.23	165.10
Average R.D.	0.7294	0.4386	0.2904
Maximum Error Aberration	11	20	19
Average Runtime per Phrase (seconds)	0.020	0.265	0.168

6.3. Discussion

As shown in Table 2, our approach gives the total error aberration less than baseline 1 and baseline 2 because there are some drawbacks in the baselines. For baseline 1, the significant drawback comes from the false synchronization and causes cumulative errors continuously. When the false synchronization continuously happens, the total error aberration is increasing dramatically if the error occurs in the same states. As in Fig. 8, the error is happened due to the deletion state at phrase 2 and phrase 4. After the deletion state appears at phrase 2, it is cumulative to phrase 3 because there is no method for recovering such error. Consequently, the total error aberration dramatically increases due to cumulative errors from phrase 4 and 5. As in Fig. 9, we can notice that the total error aberration of baseline 1 is greater than our approaches; however, the baseline 1's processing time is less than others.

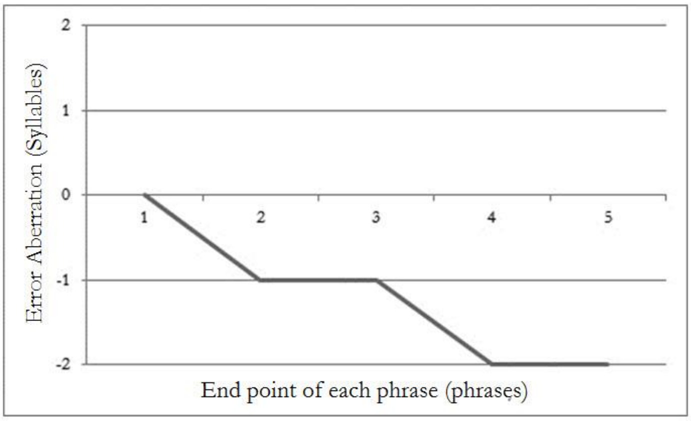

Fig. 8. The example of false alignment that happened in the same state.

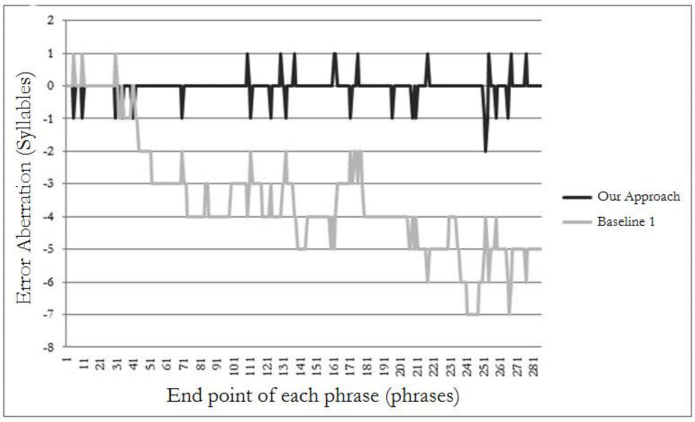

Fig. 9. The error aberration comparison of our approach and the baseline 1.

Considering results of baseline 2, we found that although the total error aberration between our approach and baseline 2 are slightly different. On the other hand, the total error aberration per number of phrases, which is the normalized result between our approach and baseline 2, is greatly different. Since baseline 2 processes the synchronization every 1 second, it has less total number of phrases than our approach but yields more error in each phrase. It also directly affects time processing since the expanding hypotheses parameter is set to 3 in order to cover all possible hypotheses while our approach is adequately set to 1. In our approach, the transcription verification procedure is relied on the number of syllables detected by the real-time syllable detection procedure so it does not need to assume any time constraints like baseline 2. Consequently, our approach yields the total error aberration per number of phrases less than baseline 2.

The experiment results show that the R.D. value of our approach outperforms the others. Although the R.D. value of our approach is the smallest, the maximum error aberration is not minimal. It leads us to think that why the maximum error aberration is not consistence with the R.D. value. When investigating the results, we discovered that there are only 41 sentences from 1176 sentences (3.49%) having the R.D. value more than 1 and those sentences mainly cause the total error aberration of these sentences tending to increase continuously for our approach. This error situation could be inferred as miss synchronization. The rest of sentences do not have the error aberrations more than 3 syllables since they are recovered by the transcription verification procedure in the next phrases. Consequently, our approach gives the better performance for the overall result.

In addition, we also investigate in those 41 sentences which caused the miss synchronization error to explore what factors affect our approach. As a result, we found most common two situations that cause the error in each phrase. The first situation is occurred commonly when there is a period that speakers stop to breathe. The duration of some phonemes might be expanded longer than usual since the system could not know whenever speakers would stop to breathe or we called silence or short pause. For example, considering the consonant 'r' in Fig. 10, the length of duration is longer than usual, whereas the consonant 'r' should be small according to its linguistic characteristic. Due to the characteristic of the force alignment algorithm which tries to force the label 'r' of transcription to the speech segment, the log-likelihood score of the correct hypothesis might be lower than the false one as shown in Fig. 11. The false hypothesis will be determined as the best hypothesis instead. Typically, the insertion error and the deletion error are mostly caused by this situation. However, this situation could mostly be recovered in the next few phrases.

Another situation occurs when some words in transcription are spoken faster or slower than usual. For example, when speakers utter the word "สหรัฐอเมริกา" (\s-a h-a r-a-t^ z-a m-e r-i k-a\) consisting of seven syllables whose most of their vowels are lax vowels and the speakers tend to speak pretty fast, the real-time syllable detection procedure could not detect all syllables in this word and could miss detecting three or four syllables. In case that the number of expanding hypotheses are less than a number of miss detection, the errors could not be recovered because the transcription verification procedure could not expand possible hypotheses to cover the correct hypothesis. According to the results, we found that the synchronization will keep suffering from the same state of errors continuously from this phrase and the total error aberration of this utterance will increase dramatically. Consequently, it causes cumulative errors and then makes our approach give the maximum error aberration value than baseline 1 in this case.

According to the second situation that causes an error, one factor that affects our approach is the speaking rate. Our approach correctness becomes worse if speakers utter too fast than usual. In the future work, if our approach could predict the speaking rate of each speaker individually, then the correctness of the synchronization would be improved. Another factor that partially affects our approach is the loudness. Due to the syllable detection procedure, the threshold of the real-time syllable endpoint detection module and the landmark detection module is a statistical number obtaining from the training data. It is possible that sometimes in practice users will speak overwhelmingly louder or softer than the average loudness in the training data. This situation would cause the failure on our approach in the real-time syllable detection procedure. To avoid such errors, our approach might need to take more preliminary data to predict these two parameters, the speaking rate and the loudness, before applying to practical applications.

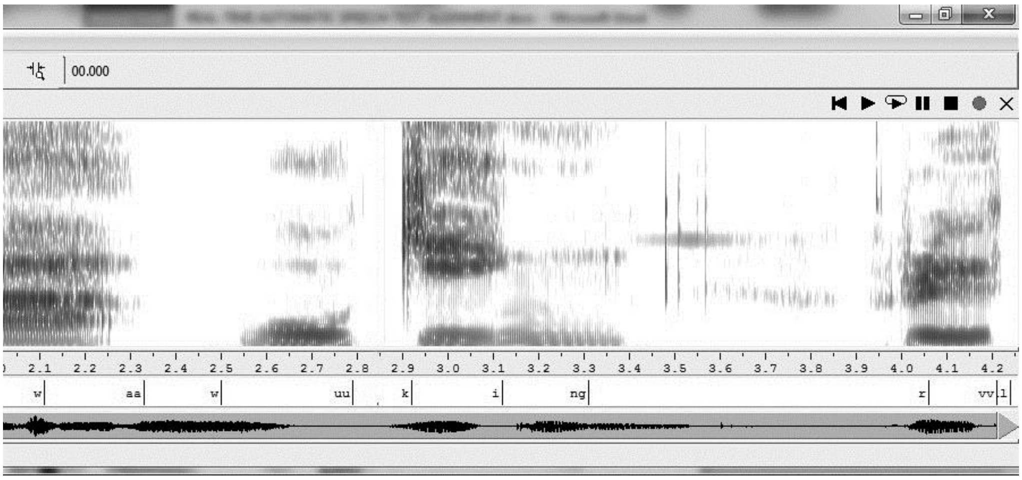

Fig. 10. The alignment example of the correct hypothesis.

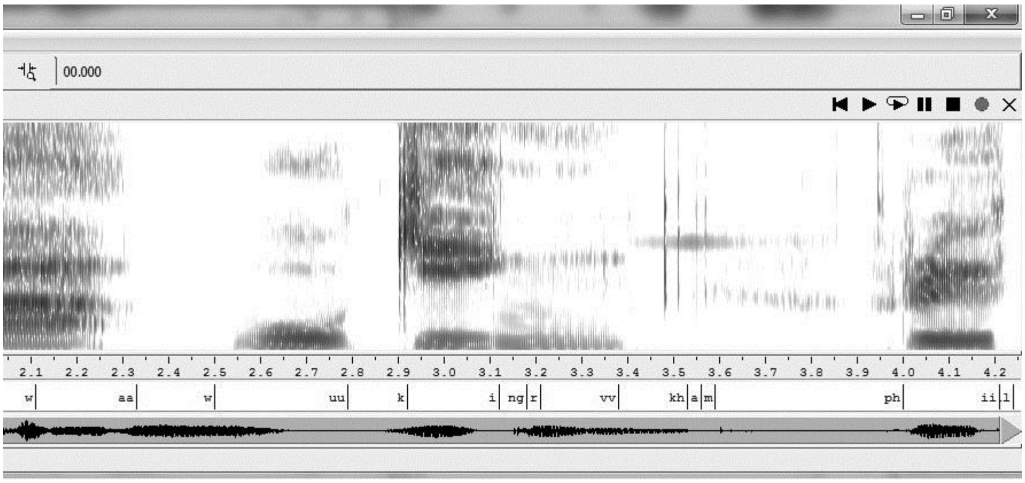

Fig. 11. The alignment example of the false hypothesis.

In terms of processing time, the experiment results indicate that our approach could be used in the real-time situation since our approach consumed the average runtime per phrases at 168.00 msec., whereas according to LOTUS [28] speakers require averagely 404.20 msec. to utter one syllable. Consequently, our approach can perform synchronizing one syllable before speakers could speak the next syllables. Therefore, it can be concluded that our approach can also support the synchronization of live speech and transcription in the real-time situation.

7. The Application

In this section, the application using the real-time synchronization of live speech and its transcription is introduced. ChulaDAISY [10], which is an automated audio book generator in DAISY format, is chosen as a case study.

Generally, when constructing a full text and audio digital taking book, we have to manually synchronize between text in a book and speech audio of speakers who read the book. In ChulaDAISY, to make the synchronization task become easier, it allows a user to prepare and manipulate text in the application in order that a user could read text of the book straight-forwardly without worrying about the synchronization of transcription and speech. Still, the preparation and the manipulation of text are time-consuming, especially in case of Thai in which boundaries of sentences are difficult to determine due to a characteristic of the language. Our approach is therefore applied to ChulaDAISY in order to eliminate time for the preparation and the manipulation of text so that the productivity of producing audio books could be improved. Details of ChulaDAISY experiments are as follows.

7.1. Experiments and Results

Conventional ChulaDAISY and ChulaDAISY with our approach are used to create full text and audio digital taking books in DAISY 3 format. Their time processing for generating the audio books in each approach is measured in order to investigate whether ChulaDAISY with our approach could produce the audio books faster than ChulaDAISY alone.

In the experiment, a participant who is familiar with conventional ChulaDAISY is invited to produce audio books in three trials. In three trials, we mock up scenarios, in which the participant asks to produce one-page audio book, two-pages audio book and three-pages audio book respectively. Each trial imitates the process for producing audio books in a practical situation. For the conventional ChulaDAISY, the participant has to manipulate and prepare text, whereas for ChulaDAISY applied with our approach, the participant could start reading text of the audio book suddenly. All experiments are conducted in a silent chamber and a unidirectional microphone is used to record speech audios. The processing time in the experiment for each approach measures from the starting the process of creating the audio book until the audio book is generated.

Results of the processing time in each trial are listed in Table 3. The experimental results indicate that ChulaDAISY with our approach requires less processing time for producing audio books.

Table 3. The results of processing time using for producing audio books.

Text Trail	Number of words	Processing Time (min.)	
		ChulaDAISY	ChulaDAISY with our approach
Trial 1	328	11.46	2.42
Trial 2	595	15.90	4.99
Trial 3	1085	24.07	8.92

7.2. Discussion

Considering results of the processing time in each trial, they shows that ChulaDAISY with our approach could reduce processing time in every trial more than two times of the conventional ChulaDAISY. This conforms to our assumption in which our approach could provide the method to produce the audio books in a faster way.

Since our experiment attempted to synchronize speech and its transcription based on the assumption that both speech audio and its transcription are consistent, we found that when our approach is applied to generate audio books practically, the participant could make three types of speaking errors that affected our approach. Three types of speaking error are an insertion error, a deletion error and a substitution error. The insertion error is the error when the participant utters some additional syllables or words that are not in the transcription. On the other hand, the deletion error is the error when the participant skips uttering some syllables or words in the transcription. Lastly, the substitution error is the error when the participant utters some words to be other words that are not in the transcription. It is obvious that those three types of errors break our assumption. As a result, our approach could not perform well in such scenarios. According to the experiments, we discovered that in our approach would definitely suffer from the insertion error and the deletion error. Although the insertion error and the deletion error could cause negative effects on our approach, the substitution error might not give much trouble because our approach only considers a number of syllables, not a lexical meaning, as a key of synchronization in the syllabic level. Thus, as long as the error utterance has the same length of syllables equally to its corresponding transcription, our approach would be robust from the substitution error.

When these types of error occur, the system would probably choose the wrong hypothesis in that phrase and become cumulative errors of the next phrases. When this comes to practical uses, users have to re-utter the utterance that they do not utter correctly.

8. Conclusion

This article introduces the automatic synchronization of live speech and its transcription in the real-time situation at the syllabic level. In our approach, the real-time syllable detection procedure and the transcription verification procedure are proposed. In the real-time syllable detection procedure, the syllable detection technique is utilized to locate syllable positions and calculate total spoken syllables. In the transcription verification procedure, the HMM-based approach is selected to compute log-likelihood scores of hypotheses generating from the output of the real-time syllable detection procedure in order to verify and determine the best hypothesis for the synchronization. Experimental results show that our approach outperforms two baselines which have been applied for Thai. Considering the processing time for the synchronization, it also indicates that our approach could perform in the real-time situation because our approach utilizes the processing time less than the time at which speakers produce a syllable.

Besides, our approach is applied to ChulaDAISY. In the experiment, the results indicate that ChulaDAISY with our approach could reduce time consumption for producing audio books due to the text preparation process. As a result, the productivity of producing audio books could be improved. Nevertheless, there are the speaking errors from users which could cause problem to our approach.

In the future work, we plan to handle such errors by considering the speaking rate as the parameter to customize the number of expanding hypotheses during the automatic synchronization between live speech and its transcription so that our approach could generate the correct hypothesis in the transcription verification procedure. Furthermore, the silence detection technique could help our approach to perform the transcription verification procedure more accurately.

Acknowledgement

This research was funded by the Chulalongkorn University Graduate Scholarship to commemorate the 72nd Anniversary of His Majesty King Bhumibol Adulyadej given by Graduate School Chulalongkorn University.

References

[1] J. E. Garcia, A. Ortega, E. Lleida, T. Lozano, E. Bernues, and D. Sanchez, "Audio and text synchronization for TV news subtitling based on automatic speech recognition," in *Broadband Multimedia Systems and Broadcasting, 2009. BMSB '09.* IEEE International Symposium, May 13-15, 2009, pp. 1–6.

[2] A. Ando, T. Imai, A. Kobayashi, H. Isono, and K. Nakabayashi, "Real-time transcription system for simultaneous subtitling of Japanese broadcast news programs," *Broadcasting, IEEE Transactions*, vol. 46, pp. 189–196, Sept. 2000.

[3] Y. Wang, M.-Y. Kan, T. L. Nwe, A. Shenoy and J. Yin, "LyricAlly: Automaitc Synchronization of Acoustic Musical signals and textual lyrics," in *Proceedings of the 12th ACM International Conference on Multimedia*, Oct. 10–16, 2004, pp. 212–219.

[4] S.-W. Li, H.-T. Lin and H.-Y. Chen, "How Speech/Text Alignment Benefits Web-based Learning," In *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 259-260.

[5] Hindenburg Systems ApS, Hindenburg ABC. [Online]. Available: http://hindenburg.com/products/hindenburg-abc/. [Accessed: 23 July 2014].

[6] D. Damm, H. G. Grohganz, F. Kurth, S. Ewert, and M. Clausen. "SyncTS: Automatic synchronization of speech and text documents," in *Proceedings of the AES 42nd International Conference*, 2011, pp. 98–107.

[7] Y. Wang, M.-Y. Kan, T. L. Nwe, A. Shenoy, and J. Yin, "LyricAlly: Automatic synchronization of acoustic musical signals and textual lyrics," in *Proceedings of the 12th Annual ACM International Conference on Multimedia, New York*, NY, USA, 2004.

[8] X. Anguera, N. Perez, A. Urruela, and N. Oliver, "Automatic synchronization of electronic and audio books via TTS alignment and silence filtering," in *Proceedings of ICME 2011*, 2011, pp. 1–6.

[9] D. Iskandar, Y. Wang, M.-Y. Kan, and H. Li, "Syllabic level automatic synchronization of music signals and text lyrics," in *Proceedings of ACM Multimedia*, New York, USA, 2006, pp. 659–662.

[10] P. Punyabukkana, N. Lertwongkhanakool, N. Kertkeidkachorn, S. Vorapatratorn, P. Hirankan, and A. Suchato, "ChulaDAISY: an automated DAISY audio book generation," in *Proceedings of the Sixth International Convention for Rehabilitation Engineering and Assistive Technology (i-CREATe 2012)*, Singapore, July 24–26, 2012.

[11] The DAISY Consortium, the ANSI/NISO Z39.86 Specification for the Digital Talking Book. [Online]. Available: http://www.daisy.org/z3986/2005/Z3986-2005.html, [Accessed: 23 July 2014].

[12] The DAISY Consortium, AMIS: DAISY 2.02 & DAISY 3 Playback Software. [Online]. Available: http://www.daisy.org/amis/amis-daisy-2.02-daisy-3-playback-software. [Accessed: 23 July 2014].

[13] P. Charoenpornsawat and V. Sornlertlamvanich, "Automatic sentence break disambiguation for Thai," In *Proceedings of ICCPOL2001*, Korea, 2001, pp. 231–235.

[14] N. Lertwongkhanakool, P. Punyabukkana, and A. Suchato, "Real-time synchronization of live speech with its transcription," in *Proceeding of Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Krabi, Thailand, May 15–17, 2013, pp. 1–5.

[15] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *Journal of The Acoustical Society of America*, vol. 58, no. 4, pp. 880–883, 1975.

[16] A. W. Howitt, "Automatic syllable detection for vowel landmarks," Ph.D. thesis, Massachusetts Institution of Technology, Jul. 2000.

[17] S. Luksaneeyanawin, "Three-dimensional phonology: A historical implication," in *The Third International Symposium on Language and Linguistics*, Bangkok, Thailand, 1992, pp. 75–90.

[18] A. Katsamanis, M. P. Black, P. G. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Proceedings of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[19] A. Haubold and J. R. Kender, "Alignment of speech to highly imperfect text transcriptions," in *Proceedings of the IEEE International Conference on Multimedia and Expo 2007*, pp. 224–227.

[20] J. Gao, Q. Zhao, and Y. Yan, "Automatic synchronization of live speech and its transcripts based on a frame-synchronous likelihood ratio test," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP'2010)*, 2010, pp. 1622–1625.

[21] H. R. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *Proceedings of Fourth International Conference on Spoken Language*, 1996, pp. 1261–1264.

[22] A. Juneja and C. Espy-Wilson, "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines," in *Proceedings of the International Joint Conference on Neural Networks*, 2003, pp. 675–679.

[23] P. Dareeyoah, "Vowel landmark detection in Thai continuous speech," M.S. thesis, Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, 2006.

[24] W. Rochkittichareon, A. Suchato, and P. Punyabukkana, "Broad phonetic class segmentation study for Thai automatic speech recognition," in *Proceedings of Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2012, pp. 1–4.

[25] M. Pitz, S. Molau, R. Schlüter, and H. Ney, "Automatic transcription verification of broadcast news and similar speech corpora," in *Proceedings of DARPA Broadcast News Workshop*, 1999.

[26] H. Jiang "Confidence measures for speech recognition: A survey," *Journal of Speech Communication*, vol. 45, no. 4, pp. 455–470, Apr. 2005.

[27] N. N. Bitar, "Acoustic analysis and modeling of speech based on phonetic features," Ph.D. thesis, Boston University, 1998.

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