


โครงการ "การเรียนรู้จำเสียงพูดภาษาไทยโดยใช้นิวโรฟัซซี่"

Thai Speech Recognition using NeuroFuzzy

โดย ผศ.ดร. ประทิต สันติประภพ และคณะ

กุมภาพันธ์ 2549

โครงการ "การเรียนรู้จำเสียงพูดภาษาไทยโดยใช้นิวโรฟัซซี่"

Thai Speech Recognition using NeuroFuzzy

โดย ผศ.ดร. ประทิต สันติประภพ และคณะ

กุมภาพันธ์ 2549

สัญญาเลขที่ BRG/4/2542

โครงการ "การเรียนรู้จำเสียงพูดภาษาไทยโดยใช้นิวโรฟัซซี่"
Thai Speech Recognition using NeuroFuzzy

คณะผู้วิจัย	สังกัด
ผศ. ดร. ประทิต สันติประภพ	คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยอัสสัมชัญ
ผศ. ดร. ธิติพงศ์ ตันประเสริฐ	คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยอัสสัมชัญ
นายตรัยพงษ์ จันทร์จรุง	คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยอัสสัมชัญ

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัย

กิตติกรรมประกาศ

ในลำดับแรกคณะผู้วิจัยใคร่ขอขอบพระคุณสำนักงานกองทุนสนับสนุนงานวิจัย และมหาวิทยาลัยอัสสัมชัญ ที่ได้สนับสนุนงานวิจัยชิ้นนี้ด้วยดีมาโดยตลอด ถึงแม้ว่างานวิจัยนี้จะใช้เวลามากกว่าที่คาดการณ์ไว้พอสมควร เนื่องจากต้องขยายขอบเขตของงานวิจัยจากที่ได้นำเสนอไว้เมื่อตอนเริ่มต้นโครงการ ทั้งนี้ เพื่อให้ผลที่ได้รับมีความสมบูรณ์ครบตามเป้าประสงค์ของการพัฒนากรอบการทำงานสำหรับรู้จำเสียงเสียงพูดภาษาไทย นอกจากนี้คณะผู้วิจัยใคร่ขอขอบพระคุณคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยอัสสัมชัญ สำหรับความอนุเคราะห์ในส่วนของอุปกรณ์และเครื่องมือที่ใช้ในงานวิจัย สุดท้ายนี้คณะผู้วิจัยใคร่ขอขอบพระคุณอาจารย์และนักศึกษาของคณะวิทยาศาสตร์ฯ ที่ได้กรุณาพูดคำและประโยคต่างๆ เป็นจำนวนมาก เพื่อบันทึกไว้เป็นตัวอย่างเสียงสำหรับงานวิจัย ซึ่งงานวิจัยนี้จะไม่สามารถสำเร็จลงได้เลยถ้าไม่มีตัวอย่างเสียงดังกล่าว

บทคัดย่อ

ภาษาไทย

การวิจัยเกี่ยวกับการรู้จำเสียงพูดของมนุษย์ (Speech Recognition) นั้น ในระยะหลังจนถึงปัจจุบันมีความสำคัญมากขึ้นเป็นลำดับ รวมทั้งปริมาณงานวิจัยที่เกี่ยวข้องก็มีมากขึ้นด้วย งานวิจัยที่นำเสนอในรายงานฉบับนี้เป็นความพยายามที่จะพัฒนากระบวนการรู้จำเสียงพูดภาษาไทย กรอบการทำงานที่ได้พัฒนาขึ้นสามารถแบ่งออกได้เป็น 3 ส่วน ซึ่งส่วนแรกคือการตัดแบ่งพยางค์ในสัญญาณเสียงพูดที่นำเข้ามาสู่ระบบ ในส่วนแรกนี้ได้มีการพัฒนาอัลกอริทึมที่ใช้เทคโนโลยี Fuzzy Inference System สำหรับการคำนวณค้นหาขอบเขตของแต่ละพยางค์ในสัญญาณเสียงพูด ในส่วนที่สองนั้น แต่ละพยางค์ที่ได้ถูกตัดแบ่งไว้จะถูกนำมาประมวลผลเพิ่มเติม เพื่อที่จะทำการรู้จำหน่วยเสียง (Phoneme) ของพยางค์นั้นๆ กล่าวคือ เสียงพยัญชนะต้น เสียงสระ เสียงพยัญชนะปลาย และเสียงวรรณยุกต์ โดยอาศัยเทคโนโลยี Hidden Markov Model และ Artificial Neural Network ณ จุดนี้สัญญาณเสียงพูดที่นำเข้ามาสู่ระบบ ได้ถูกประมวลขึ้นมาเป็นพยางค์ที่ผ่านการรู้จำที่ถูกจัดเรียงกันเป็นลำดับ ในส่วนที่สามจะนำพยางค์ที่ผ่านการรู้จำเหล่านั้นมาจัดกลุ่มเป็นคำ ซึ่งผลลัพธ์ที่ได้จะเป็นคำที่ผ่านการรู้จำที่ถูกจัดเรียงกันเป็นลำดับ ในงานวิจัยนี้ได้นำเสนอแนวทางการประมวลผลในส่วนที่สามเป็น 2 แนวทางคือการใช้เทคโนโลยี Genetic Algorithm และการใช้ Ambiguous Probability ทั้งนี้ทั้งสองแนวทางดังกล่าวจะต้องมีการกำหนด Word Domain ของคำศัพท์ที่จะทำการรู้จำ ซึ่งจะใช้เป็นพื้นฐานในการสร้าง Word Model ขึ้นมาสำหรับใช้ในการรู้จำคำ นอกจากนั้นแล้วในการประมวลผลของส่วนที่สามนี้ พยางค์ที่อาจจะมีความผิดพลาดจากการรู้จำจะได้รับการปรับปรุงให้ถูกต้องตาม Word Model ที่ได้สร้างขึ้นมา จะเห็นได้ว่ากรอบการทำงานสำหรับรู้จำเสียงพูดภาษาไทยที่ได้พัฒนาขึ้นมา มีความแตกต่างเป็นอย่างมากจากระบบรู้จำเสียงพูดที่ทำงานในลักษณะของ Template Matching ที่มีใช้กันอยู่ในสินค้าทางเทคโนโลยีทั่วไป ซึ่งระบบในลักษณะดังกล่าวสามารถรู้จำได้เฉพาะคำศัพท์ที่ได้รับการฝึกฝนหรือจดจำไว้ก่อน จึงทำให้สามารถใช้งานได้โดยจำกัด ไม่สามารถใช้ในการรู้จำคำพูดที่ไม่ได้ทำการฝึกฝนหรือจดจำไว้ก่อนโดยทั่วไปได้ ในทางกลับกันกรอบการทำงานที่พัฒนาขึ้นในงานวิจัยนี้ มีเป้าหมายที่รู้จำเสียงพูดภาษาไทยโดยทั่วไปในระดับพยางค์ ไม่จำกัดอยู่เฉพาะคำศัพท์ที่ได้จดจำไว้ก่อนเท่านั้น ผลลัพธ์สุดท้ายที่ได้จากกรอบการทำงานสำหรับรู้จำเสียงพูดภาษาไทยนี้จะอยู่ในรูปของคำอ่านมาตรฐาน ที่สามารถนำไปใช้ในการวิจัยด้านการทำความเข้าใจเสียงพูดภาษามนุษย์ (Natural Language Understanding of Spoken Speeches) ต่อเนื่องในอนาคตได้โดยสะดวก ทั้งนี้ ณ ช่วงเวลาที่เริ่มต้นดำเนินงานวิจัย คณะผู้วิจัยได้เข้าใจว่า 2 ส่วนแรกของกรอบการทำงาน น่าจะเพียงพอต่อการรู้จำเสียงพูดภาษาไทย และคาดว่าจะสามารถพัฒนาอัลกอริทึมเพื่อดำเนินงานใน 2 ส่วนดังกล่าวขึ้นมาได้โดยอาศัยเทคโนโลยี NeuroFuzzy เท่านั้น แต่หลังจากที่ได้ดำเนินการวิจัยและพัฒนาไปเป็นเวลามากกว่า 1 ปี ก็ได้พบว่าเทคโนโลยี NeuroFuzzy เพียงอย่างเดียวไม่เพียงพอต่อการทำงานตามที่คาดหวังไว้ จึงได้มีการ

นำเทคโนโลยี Hidden Markov Model เข้ามาเสริม นอกจากนั้นยังพบว่าการประมวลผลออกมาเป็นพยางค์ที่ผ่านการรู้จำที่ถูกจัดเรียงกันเป็นลำดับนั้น ก็ยังไม่เป็นคำตอบที่น่าพอใจในการที่จะนำไปพัฒนาระบบงานเพิ่มเติมในการทำความเข้าใจเสียงพูดภาษามนุษย์ เพราะว่าในพยางค์ที่ผ่านการรู้จำที่ถูกจัดเรียงกันเป็นลำดับนั้น จะมีข้อผิดพลาดจากการรู้จำเล็กๆ น้อยๆ ที่ไม่สามารถหลีกเลี่ยงได้ปรากฏอยู่เสมอ ซึ่งทำให้มีความจำเป็นที่จะต้องพัฒนาส่วนที่สามของกรอบการทำงานขึ้นมา ซึ่งทั้งหมดนี้ ได้ส่งผลให้งานวิจัยมีขอบเขตที่ขยายมากขึ้นจากเดิม และใช้เวลาในการดำเนินการวิจัยมากกว่าที่คาดไว้เดิมพอสมควร อนึ่ง จะเห็นได้ว่าปัญหาการรู้จำเสียงพูดภาษาไทยนี้ เป็นปัญหาที่มีความยากและซับซ้อนอยู่ในตัวเองมาก เสียงพูดของมนุษย์โดยทั่วไปจะมีความไม่แน่นอนอยู่เสมอ แม้แต่คำพูดคำเดียวกันที่พูดโดยคนคนเดียวกันสองครั้งก็ยังมีความแตกต่างกันในรายละเอียด ทั้งนี้นอกจากความยุ่งยากและซับซ้อนของปัญหาการรู้จำเสียงพูดของมนุษย์โดยทั่วไปแล้ว กรอบการทำงานสำหรับการรู้จำคำเสียงพูดภาษาไทยที่พัฒนาขึ้นมา ยังต้องรองรับลักษณะพิเศษต่างๆ ของภาษาไทย เช่น เสียงวรรณยุกต์ และเสียงประสมด้วย

ภาษาอังกฤษ

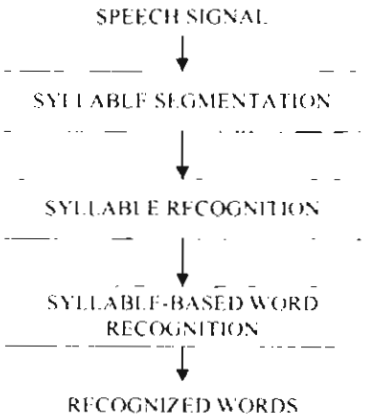
Speech recognition has been a growing filed of research for quite some time in terms of its importance as well as the number of active researches. The present research looks into a particular problem of recognizing Thai connected speech. The framework developed in this research consists of three parts. The first part called syllable segmentation starts with the segmentation of an input speech signal into a sequence of syllables at the syllables' boundaries with algorithms based on Fuzzy Inference System (FIS). Then, for each segmented syllable signal, its phonemes, namely leading consonant, vowel, ending consonant and tone, are recognized in the second part called syllable recognition using Hidden Markov Model (HMM) and Artificial Neural Network (ANN). At this point, the input speech signal has been processed into a sequence of recognized syllables. Subsequently, in the third and last part called syllable-based word recognition, the sequence of recognized syllables is segmented into a series of words with respect to a given word domain which is in turn used as a basis for the development of word models. This is accomplished by means of either a Genetic Algorithm (GA) based approach or an Ambiguous Probability based approach. In addition to the segmentation, this third part also attempts to correct any misrecognized syllables according to the word models developed. The three-part framework described here is in stark contrast to a mere template matching scheme employed in many commercially available products. With such a scheme, speech recognition is only limited to a certain number of vocabularies that have been trained or memorized. Even though practical, an application domain of the template matching scheme is rather restricted since speeches cannot be recognized in general: only pre-memorized vocabularies can subsequently be recognized. On the other hand, the framework developed in this research is meant for the recognition of any spoken Thai speeches. Here, all syllables of Thai language can be recognized and represented in a standard phonetic representation. This, therefore, forms a basis for a future research into natural language understanding of spoken Thai speeches. Originally, the research started off attempting to solve only the first two parts of the framework by means of NeuroFuzzy technology. After a couple of years of research and development effort, it has been proven that the NeuroFuzzy alone is inadequate in solving such complex problems; and the Hidden Markov Model had to be included. In addition, it has also been shown that merely producing a sequence of recognized syllables is not totally useful for speech understanding applications to be further developed since there are always some, albeit small, inherent errors in the recognized syllables. Hence, the third part of syllable-based word recognition had to be developed. As the result, the research with its widened scope took much longer than originally anticipated. It can be observed that the problem being tackled here is intrinsically difficult. Spoken speeches do contain uncertainty. Even when the same word or phrase is spoken by the same person twice, there are always some subtle differences. Moreover, in addition to typical challenges encountered in a speech recognition problem, the research also needs to address the peculiarities of Thai speeches ranging from tone to diphthong.

Key words: Speech Recognition, Signal Processing, Fuzzy Inference System, Artificial Neural Network, Hidden Markov Model, Genetic-Algorithm

สารบัญ

สรุปเนื้อหาของงานวิจัย

งานวิจัยนี้ได้ทำการศึกษาค้นคว้าองค์ความรู้เกี่ยวกับเสียงพูดภาษาไทยและงานวิจัยที่เกี่ยวข้องกับการรู้จำเสียงพูดของมนุษย์ และได้ศึกษาถึงเทคโนโลยีทาง Soft Computing ต่างๆ ที่รวมถึง เทคโนโลยี Fuzzy Inference System (FIS) เทคโนโลยี Artificial Neural Network (ANN) เทคโนโลยี Hidden Markov Model (HMM) และเทคโนโลยี Genetic Algorithm (GA) จากนั้นจึงได้ทำการทดลองค้นคว้ากับตัวอย่างเสียงที่ได้บันทึกไว้จากผู้พูดทั้งหญิงและชาย เพื่อพัฒนากรอบการทำงาน (Framework) สำหรับรู้จำเสียงพูดภาษาไทย ซึ่งกรอบการทำงานที่พัฒนาขึ้นประกอบด้วยส่วนประกอบหลัก 3 ส่วนคือ ส่วนการตัดแบ่งพยางค์ในสัญญาณเสียงพูดที่นำเข้ามาสู่ระบบ (Syllable Segmentation) ส่วนการรู้จำพยางค์โดยการประมวลผลเพื่อรู้จำหน่วยเสียงของพยางค์ (Syllable Recognition) และส่วนการรู้จำคำจากพยางค์ผ่านการรู้จำที่ถูกจัดเรียงเป็นลำดับ (Syllable-based Word Recognition) ดังที่แสดงอยู่ในรูปด้านล่างนี้



รูปแสดงถึงส่วนประกอบของกรอบการทำงานสำหรับรู้จำเสียงพูดภาษาไทย

รายละเอียดของงานวิจัยที่ประกอบด้วยการองค์ความรู้เบื้องต้นต่างๆ ที่ใช้ในการพัฒนากรอบการทำงานดังกล่าว รายละเอียดการออกแบบระบบในแต่ละส่วนประกอบของกรอบการทำงาน การทดลองและผลการทดลองที่ได้ บทวิเคราะห์ และบทสรุปรวมถึงแนวทางสำหรับดำเนินการวิจัยต่อเนื่อง ได้ทำการเรียบเรียงไว้เป็นหนังสือตามรายละเอียดด้านล่างเพื่อการเผยแพร่ ซึ่งหนังสือดังกล่าวได้นำส่งมอบมาประกอบกับรายงานฉบับนี้แล้วจำนวน 108 เล่ม (สำหรับประกอบรายงานจำนวน 8 เล่ม และสำหรับเผยแพร่จำนวน 100 เล่ม)

Pratit Santiprabhob, *Thai Speech Recognition Framework: A Syllable-based Approach*, AU Press, Assumption Univeristy, Thailand 2005, ISBN: 974-615-255-6

ผลที่ได้รับ

ผลที่ได้รับจากงานวิจัยนี้ทั้งทางตรงและทางอ้อมประกอบด้วย วิทยานิพนธ์ของนักศึกษาระดับปริญญาโท สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยอัสสัมชัญ จำนวน 5 ฉบับ เอกสารงานวิจัยที่ได้รับการนำเสนอในการประชุมวิชาการระดับนานาชาติ จำนวน 9 ฉบับ และหนังสือ จำนวน 1 เล่ม ตามรายละเอียดดังต่อไปนี้

วิทยานิพนธ์ของนักศึกษาระดับปริญญาโท สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยอัสสัมชัญ จำนวน 5 ฉบับ ประกอบด้วย

Jirawat Chaiareerat. Thai Syllable Segmentation for Connected Speech with Fuzzy System. Master Thesis. Assumption University. December 2000

Ronnarit Cheirsilp. Phoneme-Based Thai Syllable Recognition by Means of Soft Computing. Master Thesis. Assumption University. December 2000

Traipong Chancharung. Hybrid Neural Network and Fuzzy System for Phonetic Classification of Thai Speech. Master Thesis. Assumption University. December 2001

Nunmanus Dachapratumvan. Thai Syllabic Correction for Connected Thai Speech Recognition by Means of Statistic Model. Master Thesis. Assumption University. December 2002

Wanna Supasirirojana. Thai Word Decoder Based on Genetic Algorithm. Master Thesis. Assumption University. December 2002

เอกสารงานงานวิจัยที่ได้รับการนำเสนอในการประชุมวิชาการระดับนานาชาติ จำนวน 9 ฉบับ ประกอบด้วย

Traipong Chancharung and Pratit Santiprabhob. Hybrid Neural Network and Fuzzy System for Phonetic Classification of Thai Speech. In Proc.18th IASTED International Conference Applied Informatics. pp. 801-804. Innsbruck. Austria. February 2000.

Jirawat Chaiareerat and Pratit Santiprabhob. Thai Syllables Segmentation for Connected Speech with Fuzzy System. In Proc. International Conference on Artificial Intelligence (IC-AI'2000). Las Vagus. Navado. USA. July 2000

Ronnarit Cheirsilp and Pratit Santiprabhob. Phoneme-based Thai Speech Recognition Using Fuzzy System and Neural Network. In Proc. International Conference on Artificial Intelligence (IC-AI'2000). Las Vagus. Navado. USA. July 2000

Pratit Santiprabhob. Traipong Chancharung. Ronnarit Cheirsilp and Jirawat Chaiareerat. A Neurofuzzy Framework for Thai Speech Recognition. In Proc.World Multiconference on Systemics. Cybernetics and Informatics (SCI 2000). Volume III:

Virtual Engineering and Emergent Computing, pp. 602-607, Orlando, Florida, USA, July 23-26 2000

Jirawat Chaiareerat and Pratit Santiprabhob. Fuzzy-Based Thai Syllable Segmentation for Connected Speech using Energy and Different Cepstral. In Proc. 3rd International Conference on Intelligent Technologies and Third Vietnam-Japan Symposium on Fuzzy Systems and Applications (Intech/VJFuzzy'2002), pp. 334 – 343, Hanoi, Vietnam, December 3-5, 2002

Ronnarit Cheirsilp and Pratit Santiprabhob. Phoneme-Based Thai Syllable Recognition by Means of Soft Computing. In Proc. 3rd International Conference on Intelligent Technologies and Third Vietnam-Japan Symposium on Fuzzy Systems and Applications (Intech/VJFuzzy'2002), pp. 325 - 333, Hanoi, Vietnam, December 3-5, 2002

Nunmanus Dachapratumvan and Pratit Santiprabhob. Thai Syllabic Correction in Connected Thai Speech Recognition. In Proc. 3rd International Conference on Intelligent Technologies and Third Vietnam-Japan Symposium on Fuzzy Systems and Applications (Intech/VJFuzzy'2002), pp. 314- 319, Hanoi, Vietnam, December 3-5, 2002

Wanna Supasirirojana and Pratit Santiprabhob. Thai Word Decoder Based on Genetic Algorithm. In Proc. 3rd International Conference on Intelligent Technologies and Third Vietnam-Japan Symposium on Fuzzy Systems and Applications (Intech/VJFuzzy'2002), pp. 320- 324, Hanoi, Vietnam, December 3-5, 2002

Pratit Santiprabhob, Jirawat Chaiareerat, Ronnarit Cheirsilp, Nunmanus Dachapratumvan, and Wanna Supasirirojana, A Framework for Connected Speech Recognition for Thai Language. In Proc. International Conference on Intelligent Technologies 2003 (InTech' 2003), Chiang Mai, Thailand, December 17-19, 2003.

หนังสือ จำนวน 1 เล่ม คือ

Pratit Santiprabhob, Thai Speech Recognition Framework: A Syllable-based Approach, AU Press, Assumption University, Thailand 2005, ISBN: 974-615-255-6

ภาคผนวก

HYBRID NEURAL NETWORK AND FUZZY SYSTEM FOR PHONETIC CLASSIFICATION OF THAI SPEECH

TRAIPONG CHANCHARUNG
traipong@loxinfo.co.th
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok , Thailand

PRATIT SANTIPRABHOB
pratit@s-t.au.ac.th
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok, Thailand

ABSTRACT

Nowadays, the speech recognition is widely used to interface with computer programs or electronic devices by human voice. Most of the speech recognition systems recognize the human speech by detecting words or syllables previously learned. Existing methodologies limit the number of words or syllables to be stored in the system. Therefore, such methodologies do not provide the ability to learn words or syllables beyond their limits. On the other hand, learning phonemes leads to the ability to learn all syllables. Some speech recognition systems use several mathematical functions to calculate and find out the result. This approach may take several minutes. In stead of using mathematical functions, the proposed real-time system using Fuzzy System and Neural Network which can provide faster output. Some existing systems use Neural Network to recognize phonemes but there is no existing methodology that can be applied to recognize all the phonemes of the Thai language. This framework provides a way to analyze Thai speech by recognizing the syllable sounds in real-time using Fuzzy System and Neural Network. One syllable sound is divided into four phonemes. Five modules of the proposed framework altogether categorize and recognize the phonetic constituents of Thai syllable.

Keywords: Thai Speech Recognition, Intelligent System, Fuzzy system, Neural Network

1. INTRODUCTION

At present, we always use computers in our daily activities because they can manipulate any data quickly and correctly. We normally interface with a computer via a keyboard or a mouse, but both of their usage always causes some errors and inefficiency. Many times, mistyping and slow working are common. To decrease such errors, interfacing with computers by sound is preferred.

In [1], a review of the use of Neural Network in speech recognition is thoroughly documented. Most of all the existing systems for the speech recognition recognizes by word or syllable using Neural Network [2,3,4]. According to the recognition of the existing systems, their limitations are up to the number of word and syllable. Such limitations cause less than optimal learning. On the other hand, learning phonemes as discussed in [5,6,7] is different yet a better approach because the number of phonemes is fixed. Thus phonemes can be exhaustively learned.

This framework proposes a way to recognize Thai sound by firstly parsing it into syllables and then classifying them into phonemes. Because phoneme is the smallest unit of a language, it is easy to be learned, remembered and applied with any applications. In the framework, five modules are used to classify four phonemes of the Thai syllable.

2. THE STRUCTURE OF THAI SOUND

The structure of Thai sound can be divided into 4 types of phoneme: leading consonant, vowel, tone and ending consonant. There are 32 leading consonant, 24 vowels, 5 tones and 9 ending consonant. For vowel, it can be separated into 2 sub groups: 12 short-vowels and 12 long-vowels. However, there are only 9 main vowel per group because other 3 vowels of each group are the mix of main vowel.

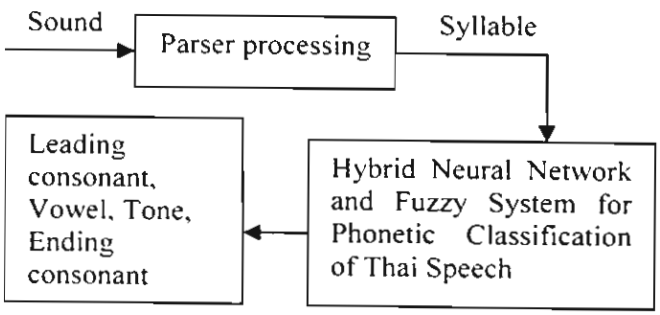


Figure 1: Over all process.

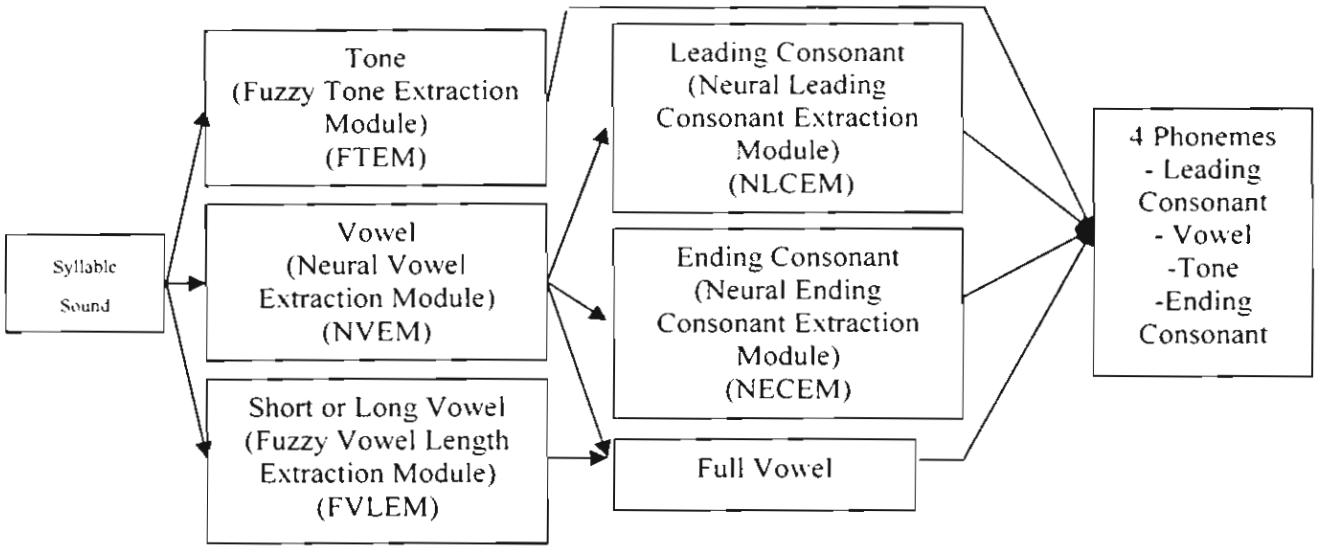


Figure 2: Overall framework

3. FEATURE OF THAI PHONETIC CLASSIFICATION

The framework is started from parsing human speech with sampling rate 11025 Hz. on mono channel (8 bits) into syllable sound with parser processing. The amplitude of syllable will be passed through the system of 'Hybrid Neural Network and Fuzzy System for Phonetic Classification of Thai Speech' in order to be analyzed and clustered into 4 kinds of phoneme (Leading consonant-Vowel-Tone-Ending consonant) as shown in Figure 1. For 'Hybrid Neural Network and Fuzzy System for Phonetic Classification of Thai Speech', it composes of 5 modules as show in Figure 2.

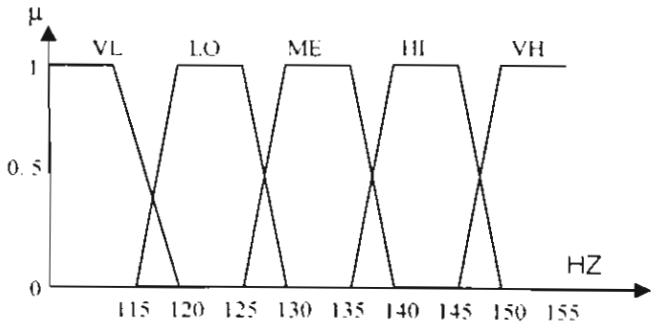


Figure 3: Fuzzy number of FTEM

3.1 FUZZY TONE EXTRACTION MODULE (FETM)

There are 5 tones of Thai sound (Low Tone, Mid Tone, High Tone, Rising Tone and Falling Tone). To recognize tone, this module has to convert all sampling into frequency and then apply the Fuzzy Logic to recognize it. The frequency will calculate from the sampling interval. All of frequency will be clustering into 5 groups. The frequency can be very low (VL), low (LO), medium (ME), high (HI) and very high (VH). The Fuzzy number representation is shown in Figure 3. There are 24 Fuzzy Rules Base for recognize tone and some of them are show below:

Rule 1:
If (group1 = ME) and (group2 = ME) and (group3 = ME) and (group4 = LO) and (group5 = VL) then Tone = Mid Tone

Rule 2:
If (group1 = VH) and (group2 = VH) and (group3 = VH) and (group4 = HI) and (group5 = ME) then Tone = Falling Tone

Rule 3:
If (group1 = VL) and (group2 = VL) and (group3 = VL) and (group4 = LO) and (group5 = VH) then Tone = Rising Tone

Rule 4:
If (group1 = ME) and (group2 = HI) and (group3 = HI) and (group4 = HI) and (group5 = VH) then Tone = High Tone

Rule 5:
If (group1 = LO) and (group2 = VL) and (group3 = VL) and (group4 = VL) and (group5 = VL) then Tone = Low Tone

The defuzzification method of FTEM is MOM (Mean of Maximum).

3.2 NEURAL VOWEL EXTRACTION MODULE (NVEM)

This module extract essential information for further processing. So this module will recognize only 12 out of 24 vowels of Thai sound. "Back Propagation" algorithm of Neural Network is applied here to analyze the amplitude. There is only one hidden layer. There are 100 input nodes, 15 hidden nodes and 9 output nodes. The learning rate is 0.7 and the max acceptable error is 0.005. Next, the processes of NVEM are setup as following:

- Selecting one sampling interval from the sound sample at any point after the 40th percentile where the shape of the sound wave becomes steady. Then normalize sample interval to 100 samples (normalize time) and normalize the amplitude to maximum value (127) and use Neural Network to recognize 9 major sounds.

- If the major sound is "i" or "u" or "ω" then check for minor sound.

- For 3 minor sounds, select one sampling interval from the sound sample at any point after the 70th percentile where the shape of the sound wave becomes steady. Then normalize sample interval to 100 samples (normalize time) and normalize the amplitude to maximum value (127) and then use Neural Network to recognize. If minor sound is "a" then it show that this sound is minor sound. Otherwise, this sound is major sound.

- The result of NVEM is sent to FVLEM.

3.3 FUZZY VOWEL LENGTH EXTRACTION MODULE (FVLEM)

FVLEM uses fuzzy rules to classify the length of vowel (Short, Long) according to the number of sampling interval. All of sampling interval must also have the maximum amplitude value more than 30. The fuzzy number is shown in Figure 4.

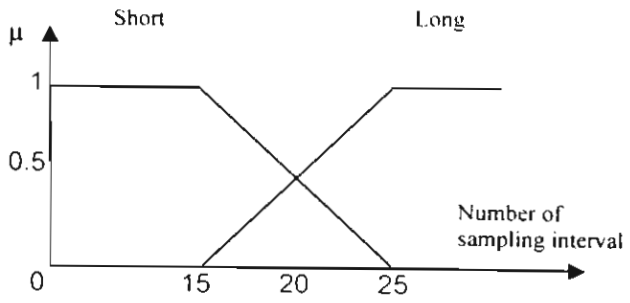


Figure 4: Fuzzy number of FVLEM

3.4 NEURAL LEADING CONSONANT EXTRACTION MODULE (NLCEM)

There are 32 leading consonants of Thai sound. NLCEM selects one sampling interval from the beginning of the sampled sound for recognizing of 32 leading consonant. It normalizes sample interval to 100 samples (normalize time) and normalizes the amplitude to maximum value (127). Simple "Back Propagation" algorithm of Neural Network is used here to recognize them. Each leading consonant is up to vowel, which is meant that it uses 9 Neural Networks to find out the value of consonant in each vowel. Hence only one hidden layer is used. A Neural Network contains 100 input nodes, 20 hidden nodes, and 32 output nodes. The learning rate is 0.7 and the max acceptable error is 0.005.

Because of the length of leading consonant sample is vary, it is not necessary to be compared with all Neural Network that gives the fast result.

3.5 NEURAL ENDING CONSONANT EXTRACTION MODULE (NECEM)

There are 9 ending consonants of Thai sound. NECEM selects one sampling interval from the ending of the sampled sound to recognize 9 ending consonants. Then, it normalizes sample interval to 100 samples (normalize time) and normalizes the amplitude to maximum value (127). Since each of ending sound is based on vowel as consonant, it uses 9 Neural Networks apply Back Propagation algorithm. A Neural Network contains 100 input nodes, 15 hidden nodes and 9 output nodes. The learning rate is 0.7 and the max acceptable error is 0.005.

4. EXPERIMENTAL RESULT

There are 10 people speakers in this experiment. Each of them speaks 5 times of all of possible data: 5 tones, 9 vowels, 32 leading consonants and 9 ending consonants. After training all of data, there are other 50 people tests this application. They speak two times for all of data. The results are shown in Figure 5-9.

Tone	Mid	Low	Falling	High	Rising
Correct	98%	85%	90%	98%	95%
Incorrect	2%	15%	10%	2%	5%

Figure 5: The result of FTEM

Vowel	Short	Long
Correct	95%	85%
Wrong	5%	15%

Figure 6: The result of FVLEM

Vowel	Correct	Incorrect
A	99%	1%
I	80%	20%
ω	82%	18%
U	90%	10%
e	98%	2%
ɣ	99%	1%
o	95%	5%
ɛ	95%	5%
ɔ	93%	7%
ia	80%	20%
ωa	82%	18%
ua	90%	10%

Figure 7: The result of NVEM

LC	Correct	Incorrect
k	92%	8%
kh	95%	5%
ng	92%	18%
c	85%	15%
ch	86%	14%
j	91%	9%
d	91%	9%
t	80%	20%
th	90%	10%
n	85%	15%
b	83%	17%
p	94%	6%
ph	95%	5%
f	95%	5%
m	88%	12%
r	90%	10%

LC	Correct	Incorrect
l	88%	12%
w	96%	4%
s	92%	8%
?	80%	20%
h	82%	16%
kw	76%	24%
kl	78%	22%
kr	78%	22%
khw	82%	18%
khl	83%	17%
khr	83%	17%
phl	86%	14%
phr	86%	14%
pl	88%	12%
pr	82%	18%
tr	96%	4%

Figure 8: The result of NLCEM
(LC = Leading Consonant)

EC	Correct	Incorrect
n	85%	15%
ng	85%	15%
k	92%	8%
t	90%	10%
p	95%	5%
m	85%	15%
j	88%	12%
w	90%	10%
§	95%	5%

Figure 9: The result of NECEM
(EC = Ending Consonant)

5. CONCLUSION

This paper presents a hybrid Neural Network and Fuzzy System for Thai phonetic classification. Each Thai syllable is consisted of a leading consonant, a vowel, a tone and an ending consonant. Each of them requires different classification techniques for which different modules, namely FTEM, NVEM, FVLEM, NLCEM and NECEM have been developed. The results of all modules are combined in order to represent the Thai syllable from the input speech. As the result, this proposed framework can recognize 34,560 syllable sounds (32 x 24 x 5 x 9) that are the basis of Thai sound. This system can be embedded in any applications such as word processing, games or electronic machine likes robot that require voice interface in Thai language.

This framework is limited on gender because the sounds of man and woman have different base frequencies. Basically, the frequency of women's sound is higher than that of men's about two times. Hence, the training data must be prepared separately for each gender. However, for the same gender, the system is speaker independent. It can be used with any people after it successfully passes the learning process.

6. REFERENCES

[1] Richard P. Lippmann, Review of Neural Networks for Speech Recognition, *Neural Computation*, 1, 1989, 1-38.

[2] L. Fissore, P. Laface, and F. Ravera, Using Word Temporal Structure in HMM Speech Recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2, 1997, 975-978.

[3] Cheng-Yuan Liou and Chwan-Yi Shiah, Perception of Speech Signals Using Self-Organization on Linear Neuron Array, *International Joint Conference on Neural Networks*, 1, 1993, 251-253.

[4] L. Bottou, F.Fogelman Soulié, P.Blanchet, and J.S.Liénard, Speaker-Independent Isolated Digit Recognition: Multilayer Perceptron vs. Dynamic Time Warping, *Neural Networks*, 3, 1990, 453-465.

[5] Åge J. Eide and Terje Lindén, Recognizing Norwegian Vowels using Neural Network, *International Neural Network Society*, 4, 1994, 524-537.

[6] David J. Pepper and Mark A. Clement, Phonetic Recognition Using a Large Hidden Markov Model,

IEEE Trans. Signal Processing, 40(6), 1992, 1590-1595.

[7] X.D.Huang, Phoneme Classification Using Semicontinuous Hidden Makov Models, *IEEE Trans. Signal Processing*, 40(5), 1992, 1062-1067

THAI SYLLABLES SEGMENTATION FOR CONNECTED SPEECH WITH FUZZY SYSTEM

JIRAWAT CHAIAREERAT
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok, Thailand
Fax: (662) 3198753
Tel: (662) 3004543

PRATIT SANTIPRABHOB
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok, Thailand
Fax: (662) 3198753
Tel: (662) 30045453

ABSTRACT

In speech recognition system, The fundamental task is to do segmentation the syllable of speech signals. The efficiency of syllable segmentation effects the performance of the entire recognition system. This process should be fast and use less time of computation to make good response time of the system.

This paper proposes a real-time framework and algorithm to segment Thai syllables by Fuzzy System. Four Energy functions will use as parameter of the speech and two Fuzzy Inference Systems (FIS) are used to determine the starting and ending point of the syllable.

Keywords: Thai Speech Recognition, Syllable Segmentation, Endpoint Detection, Signal Processing, and Fuzzy System.

1. INTRODUCTION

Most of the errors in the automatic speech recognition system are the inaccurate detection of the beginning and ending boundaries of test and reference patterns [3].

The major usage of the automatic speech recognition system is in real-time system, so the speed and response time of the system is important.

From those reasons, the good syllable segmentation algorithm should be fast and has a good accuracy. Nowadays there are several algorithms proposed to syllable segmentation. The following are a typical word boundary detection algorithm [3].

1) Energy based algorithms with automatic threshold adjustment. They are intuitive approaches based on energy levels and duration of silence and speech. Sometimes several pairs of boundaries are yielded in order of their rank of being correct.
2) Noise adaptive algorithms: use the log of the RMS signal energy, the zero-crossing rate, duration information, and a set of heuristics. The thresholds used for the energy and the zero-crossings are adapted automatically from a few frames provided by the signal environment.
3) Algorithms using frequency-based features. This algorithm performs FFT transformations and computes the energy in the frequency-band 250-3500Hz, and logarithm of the RMS energy.

The Energy and Noise adaptive algorithm need good decision system to determine the start and endpoint of syllable. The Frequency-based algorithm uses much time consuming and computation overhead. This is not good for real-time system

This paper proposes the algorithm to segment the Thai syllables with fast speed and acceptable accuracy, using energy as the parameter of the speech and use Fuzzy Inference System to determine the starting and ending point of the syllables.

2. DESCRIPTION OF THE ALGORITHM

The proposed algorithm consists of five steps: Pre-Processing, En-framing, Energy computation, Energy Graph Smoothing and Starting-Ending point detection step. First the speech signal is Pre-Processing to reduce noise and normalize the maximum amplitude to 1 , then En-framing these signal and compute the energy from each frame. The energy graph have got from the last step must be smoothed the last step, use FIS to make decision for ending and starting point of each syllable in the sentence.

2.1 PRE-PROCESSING

The speech signal will be pre-emphasized with first order low-pass filter as show in equation 1. The factor of pre-emphasize is 0.95

$$\widetilde{S}_n = S[i] - aS[i-1] \quad (1)$$

[1].

Where s[I] is a speech signal at position I
A is the factors of pre-emphasize

Then the maximum amplitude of speech signal will be normalized to 1 by the following

$$S[I]' = S[I] / MAX(S) \quad (2)$$

equation

Where S[I] is a speech signal at position I
MAX(S) is the maximum amplitude of the speech signal

2.2 EN-FRAMING

The speech signal will be en-frame with, 256 samples per frame and the overlapping 50% of each frame were assumed. This overlapping is essential for a smooth transition feature from one frame to another frame.

2.3 ENERGY COMPUTATION

For computing the energy of each frame, four following energy functions [1] are used:

1) Absolute energy

$$E_n = \sum_{i=1}^{w} |S_n[i]| \quad (3)$$

2) Root mean square energy

$$E_n = \left[\frac{1}{w}\sum_{i=1}^{w} S_n^2[i]\right]^{1/2} \quad (4)$$

3) Square energy

$$E_n = \sum_{i=1}^{w} S_n^2[i] \quad (5)$$

4) Teager energy

$$E[i] = S^2[i] - S[i+1]S[i-1] \quad (6)$$

$$E_n = \sum_{i=1}^{w} E[i] \quad (7)$$

Where
w is given as the window of frame
$S_n[i]$ *is speech sample in frame number n*
$E[i]$ *is the energy at position i*
E_n *is the energy of frame n*

2.5 SMOOTHING ENERGY GRAPH

Without smoothing energy contour, boundary detection is very difficult because the contour has a lot of local maximum and minimum energy. This process will construct the smoothing energy curve.

$$S[I]' = \frac{1}{2W+1}\sum_{i=I-W}^{I+W} S[I]$$

Where
S[I] is speech signal at position i
W is size of smoothing window

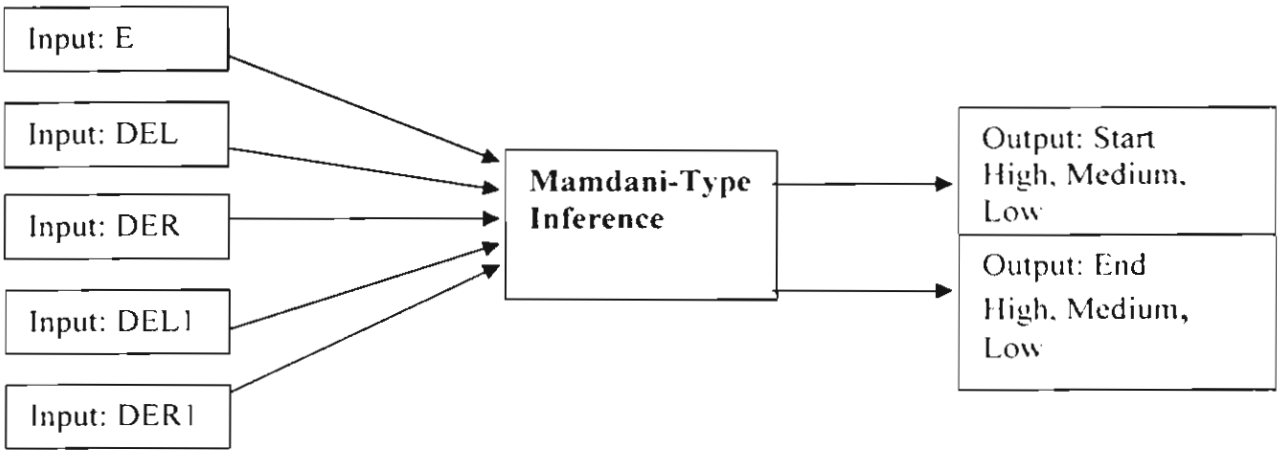


Figure 1. FIS Model

2.5 STARTING AND ENDING POINT DETECTION

The Fuzzy Inference System (FIS) is used to make a decision at each point whether: it is starting point or ending point or neither.
This paper proposes two FIS models for detection the starting and ending point of syllables:

2.5.1 FIS1 MODEL

This FIS model is Mamdni-type Inference. The defuzzification method of FIS1 is centroid. FIS1 has 21 rules, 5 inputs and 2 outputs as show in the Figure 1, Table 1 and 2 respectively.

INPUT	*FORMULAR*
E : Energy	E[I]
DEL : Delta E Left	Max(I-5,I) – E[I]
DER: Delta E Right	Max(I,I+5) – E[I]
DEL1 : Delta E Left 1 Point	E[I-1] – E[I]
DER1 : Delta E Right 1 Point	E[I+1] -E[I]

Table 1.

Where E[I] is the energy at position I
Max(I,I+5) is the maximum of energy from position I to I+5

OUTTPUT	*DESCRIPTION*
START	The values from range 0 to 1,for indicating the starting point. If 1 this point is the strongest starting point , if 0 this point is the weakest starting point
END	The value from range 0 to 1, for indicate the ending point. If 1 this point is the strongest ending point , if 0 this point is the weakest ending point

Table 2.

First, input parameter of every point in energy graph to FIS1 and keep out put from FIS1 for every point, then do the following steps from first point to the last point:

1. *Find the first Point that has Starting Output from FIS1 is High. That point is the first starting point.*
2. *Go to Next Point*
3. *If Ending Output from FIS1 is*
   - *Medium ,*
     If State is ending

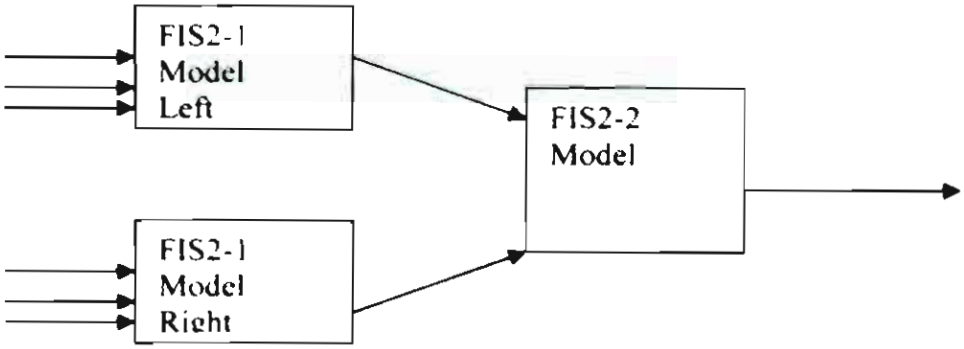


Figure 2. FIS2 Model

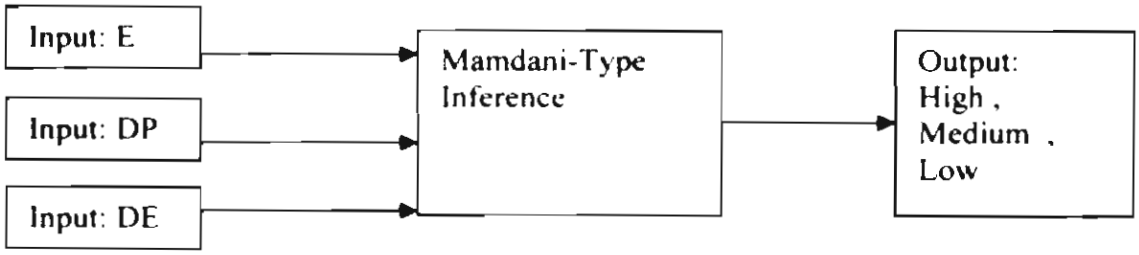


Figure 3. FIS2-1 Model

If Length from last Starting Point > 25. Add this point as New Ending Point. State is Starting
Else If Length from last Starting Point >25. Expand last Ending Point to this Point

- *High ,*
  If Length from last Starting Point <=1,
  If State is ending,
  If Last Starting Point is Medium Output, Remove Last Starting Point, Replace last Ending Point with this point. State is starting
  Else add this point as New Ending Point
  Else Expand Last ending to this Point
  Else
  If State is ending then Add this point as new Ending point. State is starting
  Else Expand Last Ending Point to this Point

4. *If Starting Output from FIS1 is*
   - *Medium*
     If State is Ending and Length from last Starting Point > 25 , Add this point as new Starting Point , State is Starting
   - *High*
     If State is Ending, Add this point as new Starting Point, State is starting
5. *Go to Step 2*

INPUT	FORMULA For LEFT	FORMULA For RIGHT
E : Energy	E[I]	E[I]
DP : Delta to Peak	Max(I-S1,I) – E[I]	Max(I-S2,I) – E[I]
DE : Delta E	E[I] - E[S1]	E[I] - E[S2]

Where each point represents by [S1, I, S2]

S1 is the previous possible point
I is current possible point
S2 is the next possible point

Table 3. Input of FIS2-2

2.5.2 FIS2 MODEL

This FIS model has 2 levels, which are Mamdni-type Inference FIS2-1 and FIS2-2. The FIS2 Model has shown in Figure 2. The Defuzzification method of both FIS2-1 and FIS2-2 is centroid. FIS2-1 has 25 rules, 3 inputs and 1 output as shown in the Figure 3 and Table 3. And FIS2-1 has 9 rules, 2 inputs and 1 output as show in Figure 4 and Table 4.

The following algorithm was used to find all of the possible point to be starting and ending. Then compute the parameter of each possible point for input to FIS2, get the output and make decision that is it should be boundary point of syllables

Algorithm to find possible point:

1. *I = first point, First Possible Point is I.*
2. *DE = E[I+1] – E[I]*
3. *If DE > 0, Go to 4*
   If DE < 0, Go to 5
   If DE = 0, Go to 6
4. *Go to Next Point and Compute DE until :*
   DE = 0 then Add I to Possible Point List, Go to next I, Go to Step 2.
   DE < 0 then Go to next I, Go to Step 5
5. *Go to Next Point and Compute DE until DE >= 0 then Add I to Possible Point List, Go to next I, Go to Step 2.*
6. *Go to Next Point and Compute DE until DE! = 0 then Add I to Possible Point List, Go to next I, Go to Step 2.*

3 EXPERIMENTAL RESULTS

There are 5 speakers in this experiment. Each of them speaks 36 ambiguous Thai sentences [5], each sentences contains 5-10 words. Use FIS1 and FIS2 to detect the starting and ending point of syllables from the four energy graphs (ABS Energy, Square Energy, RMS Energy and Teager Energy) with smoothing and not smoothing. The ABS Energy graph produces the best accuracy as show in Table 9 for accuracy by sentence, in Table 10 for accuracy by syllables.

4 CONCLUSION

Four energy algorithms have been tested on the proposed detection syllable method. From the experiment, the results show that ABS Energy produce the best accuracy. The RMS Energy produces near accuracy with ABS Energy but take more time for computing. Square Energy and Teager Energy produce low accuracy because of losing syllable problem.

For the Boundary Decision Model, FIS1 with smoothing window = 2 produce the best accuracy. For FIS2, it will not work without smoothing, however FIS2 with smoothing window = 2, produce the lower accuracy than FIS1.

Errors of the syllable detection are due to "Low Volume Syllable", "Too Long Syllable" and "Too Short Syllable"

5 REFERENCES

[1] Nuthacha Jittiwarangkul, Somchai Jitapunkul, Sudaporn Luksaneeyanawin, Visarut Ahkuputra,Chai Wutiwiwatchai. "Thai Syllable Segmentation for Connected Speech Based on Energy" 0-7803-5146-0/98 IEEE 1998

[2] He Qiang, Zhang Youwei "On Prefiltering and Endpoint Detection of Speech Signal" 0-7803-4325-6/98 Procedding of ICSP '98

[3] [Viying Zhang, Xiaoyan Zhu, Yu Hao "A Robust and fast Endpoint detection algorithm for isolated word recognition" 0-7803-4253-4/97 1997 IEEE International Conference on Intelligent Processing Systems , October 28-31, Beijing, Chaina

[4] John R. Deller. JR., John G. Proakis, John H,L. Hansen. "Discrete-Time Processing of Speech Signal." Prentice-Hall , 1993

[5] Nutthacha Jittiwarangkul, "Syllable segmentation algorithm for Thai connected speech" Master's Thesis , Chulalongkorn University, 1998

[6] Ahkuputra, V. "Speaker Independent. Thai Polysyllabic Word Recognition System Using Hidden Makov Model". Master's Thesis , Chulalongkorn University, 1996

[7] Prathumthan, T. "Thai Speech Recognition using Syllable Unit". Master's Thesis, Chulalongkorn University, 1987

[8] Thanwa Sripramong,, "Thai Speech Analysis in Harmonic-Frequency Domain". Master's Thesis, King Mongkut's Institue of Technology Ladkrabang, 1994

[9] Kongsupanich S. "The Transformation of Thai Morphemes to Phonetic Symbols For Thai Speech Synthesis System". Master's Thesis, King 1997

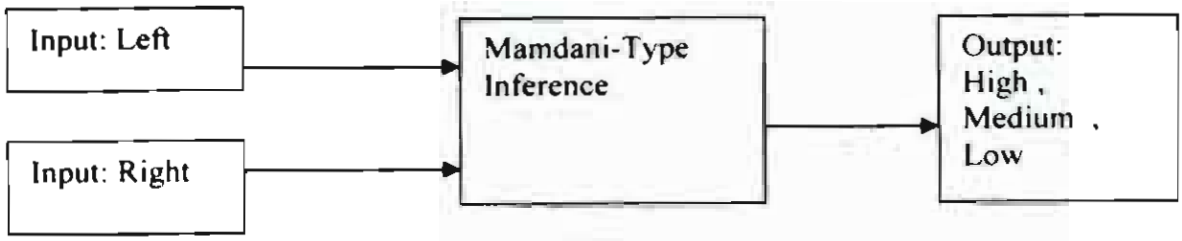


Figure 4. FIS2-2 Model

INPUT	Description
Left	The value received from the first FIS2-1
Right	The value received from the second FIS2-2

Table 4. Input of FIS2-2

Model	All Sentences	Wrong	Percent	Correct	Percent
FIS1 with No Smoothing	180	29	16.11%	151	83.39%
FIS1 with Smoothing Window size = 2	180	31	17.22%	149	82.78%
FIS2 with Smoothing Window size = 2	180	35	19.44%	145	80.56%

Table 5. Accuracy by sentence

Model	All Words	Wrong	Percent	Correct	Percent
FIS1 with No Smoothing	1355	32	2.36%	1323	97.64%
FIS1 with Smoothing Window size = 2	1355	31	2.29%	1324	97.71%
FIS2 with Smoothing Window size = 2	1355	36	2.66%	1319	97.34%

Table 6. Accuracy by word

PHONEME-BASED THAI SPEECH RECOGNITION USING FUZZY SYSTEM AND NEURAL NETWORK

RONNARIT CHEIRSILP
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok, Thailand
Fax: (662) 3198753
Tel: (662) 3004543

PRATIT SANTIPRABHOB
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok, Thailand
Fax: (662) 3198753
Tel: (662) 3004543

ABSTRACT

The Thai speech can be recognized by word or syllable. Most Thai speech recognition systems are word-based and syllable-based system. And most systems use several signal-processing functions in recognition process. These systems have a limit number of word and syllable that they can recognize.

This research paper proposed the Thai speech recognition system that recognizes syllable by its phoneme representation using Fuzzy and Neural Network (NN) technique. The proposed system is designed to have desirable properties:

1. *The proposed system is not limited to the number of word or syllable.*
2. *The proposed system is general enough to be speaker-independent speech recognition system*
3. *The proposed system has small computation time to be able to be real-time system.*
4. *The proposed system has acceptable recognition error rate.*

Keywords: Thai Speech Recognition, Fuzzy System, Neural Network and Signal Processing

1. INTRODUCTION

Speech recognition is an exciting and challenging technology that changes the way we interact with computers in the future. This technology has been under development for more three decades at university, corporate and government research labs.

From the review of Thai speech recognition systems, most systems recognize Thai speech by words [1,2,3,4]. Most of them [1,2,4] use several signal-processing functions that require a lot of computation time and power.

In this research, the proposed system presents a way to recognize Thai syllable sound by their phonetic representation using Fuzzy and NN system. This paper is organized as follows. In Section 2, the structure of Thai syllable included Thai sound and phonetic representation of Thai syllable is introduced. In Section 3, the details of this proposed system are discussed. In Section 4, The initial experimental results are discussed. In Section 5, the conclusion is discussed

2. THE STRUCTURE OF THAI SYLLABLE AND PHONEME

A Thai syllable can be divided into 4 phonemes: leading consonant, vowel, tone, and ending consonant. There are 32 leading consonants, 24 vowels, 5 tones and 9 ending consonants in Thai language. Vowels can also be separated into 2 groups: 12 short-vowels and 12 long-vowels. Each group has only 9 major vowels (pure vowel) and 3 minor (diphthong or the combination of 2 pure vowels) vowels. Both groups of vowel are identical regardless to length of vowel.

A Thai syllable sound can be divided into three main parts regarding to redundant pattern in its amplitudes. Three main parts of a Thai syllable sound: leading consonant part, vowel part, and ending consonant part.

3. PROPOSED SYSTEM

The proposed system recognizes the Thai syllables by classify them to phonemes. The system consists of 4 processes. A Thai syllable signal sample is passed through all 4 processes the proposed system in parallel and then phonetic representation of Thai syllable is returned as the result. The following are all 4 processes in the proposed system:

3.1 VOWEL RECOGNITON PROCESS (VRP)

Each Thai syllable contains at least one major vowel. In Thai syllable signal sample, Each 9 major vowels has a different period sample at the middle (about 25-60%) of the signal. The main point of this process is to recognize these 9 period samples. And the rest of Thai vowel can be derived from these 9 major vowels. VRP is divided into 2 parts. First part is to acquire a normalized vowel period sample from the syllable signal sample. The following steps are applied to acquire a normalized vowel period:

1. The middle part of the Syllable signal sample is extracted to be used as the input of step 2
2. A period sample is extracted from step 1 output signal sample using Average Magnitude Difference Function (AMDF) method [5].
3. The period sample is normalized to have time length equal to 20
4. All amplitudes of the period sample is normalized using the following formula:

$$s(i)' = s(i)/a$$

where s(i) is period sample at time i.
a is the absolute maximum amplitude of period sample.

The second part is to recognize the vowel using the result from the first part as the input. In this research, One-Hidden layer feedforward neural network (OFNN) is applied to classify 9 vowels. OFNN consists of 20 input nodes, 15 hidden nodes and 9 output nodes. Back-propagation (gradient descent with variable learning rate) algorithm is used for OFNN training. If the output of OFNN is the major vowel that can be combined with another major vowel in order to construct diphthong vowel. The processes in the first and second part of VRP are repeated again to find out the second major vowel. But in step 1 of the first part the different middle part of syllable signal sample is extracted at the different position. If second vowel is the same as first vowel, the vowel of the input syllable sound is major vowel. Otherwise, the vowel of the input syllable sound is minor (diphthong) vowel.

3.2 TONE RECOGNITION PROCESS (TRP)

In Thai language, each tone has a different frequency feature. Fuzzy inference system (FIS) is used in TRP. The FIS model is Mamdani-type FIS. FIS has 63 rules, 4 inputs and 5 outputs. The defuzzification method is centroid.

All 4 inputs (d1, d2, d3, and d4) of FIS are gathered using the following steps:

1. The input Syllable signal sample is divided into 30 windows equally.
2. Apply AMDF algorithm to each window in order to find 1 period sample.
3. Find the frequency of each window by using the following formula:

$$f = s / p$$

where f is frequency (Hz)
s is Sampling rate (Hz)
p is a period sample time (sec) from the step 2

As the results, we got 30 values of frequency.

4. The frequency values for step 3 are sequentially grouped into 5 groups
5. The median frequency is used to represent each group.
6. The input of FIS is finally calculated using the following formula:

$$d1 = (100 / M1) * M2 - 100$$
$$d2 = (100 / M1) * M3 - 100$$
$$d3 = (100 / M1) * M4 - 100$$
$$d4 = (100 / M1) * M5 - 100$$

where dN is the n-th input of the FIS.
MN is the median of n-th group.

The output of FIS is the tone of the input syllable signal sample

3.3 LEADING CONSONANT RECOGNITION PROCESS (LRP)

The leading consonants of Thai syllable sounds are found out in this process. LRP is also divided into 2 parts as VRP. The first part is similar to the first part of VRP but the difference is the beginning part of syllable signal sample is used instead of the middle part. The second part is to recognize the leading consonant using the result from the first part as the input. In this part, One-Hidden layer feedforward neural network (OFNN) can be applied to classify 32 leading consonants. OFNN consists of 20 input nodes, 15 hidden nodes and 32 output nodes. Back-propagation (gradient descent with variable learning rate) algorithm is used for OFNN training. Fuzzy Associative Memory, constructed using the weight of OFNN, can also be used to classify 32 leading consonant to increase the process speed.

3.4 ENDING CONSONANT RECOGNITION PROCESS (ERP)

The ending consonants of Thai syllable sounds are found out in the process. ERP is also divided into 2 parts as VRP. The first part is similar to the first part of VRP but the difference is the ending part of syllable signal sample is used instead of the middle part. The second part is to recognize the ending consonant using the result from the first part as the input. In this part, One-Hidden layer feedforward neural network (OFNN) can be applied to classify 9 ending consonants. OFNN consists of 20 input nodes, 15 hidden nodes and 9 output nodes. Back-propagation (gradient descent with variable learning rate) algorithm is used for OFNN training. Fuzzy Associative Memory, constructed using the weight of OFNN, can also be used to classify 9 ending consonant to increase the process speed.

4. EXPERIMENTAL RESULTS

Two initial experiments were conducted in VRP and TRP. For both initial experiments, the syllable sounds recorded at 8 bit, mono and 11kHz sampling rate.

In VRP experiment, there are 4 speakers in the experiment. VRP experiment was conducted as the following steps:

1. 154 syllable sound data, which can be classified into 9 groups by vowel, is gathered for 2 speakers.
2. An OFNN was setup and configured as OFNN in VRP.
3. The OFNN from step 2 is trained with 154 syllable sound data from step 1.
4. 9 syllable sound data, which can be classified into 9 groups by vowel, is gathered from each speaker.
5. Then the trained OFNN from step 3 is tested with altogether 36 syllable sound data from step 4. And the results are shown in Table 1.

In TRP, there are 2 speakers in the experiment.

1. 53 syllable sound data, which can be classified into 5 groups by tone, is gathered from the first speaker.
2. A FIS was setup and configured as the FIS in TRP.
3. The FIS from step 2 was tuned with all syllable sound data from step 1.
4. 62 syllable sound data, which can be classified into 5 groups by tone, is gathered from the second speaker.
5. Then the tuned FIS from step 4 is tested with altogether 115 syllable sound data from step 1 and 4. And the results are shown in Table 2.

Vowel	Number	Correct	Correct(%)	Incorrect	Incorrect(%)
1	4	4	100.00	0	0.00
2	4	4	100.00	0	0.00
3	4	4	100.00	0	0.00
4	4	3	75.00	1	25.00
5	4	4	100.00	0	0.00
6	4	3	75.00	1	25.00
7	4	3	75.00	1	25.00
8	4	3	75.00	1	25.00
9	4	2	50.00	2	50.00
Total	36	30	83.33	6	16.67

Table 1: VRP Initial Experimental results

Tone	Number	Correct	Correct(%)	Incorrect	Incorrect(%)
1	24	22	91.67	2	8.33
2	18	14	77.78	4	22.22
3	24	20	83.33	4	16.67
4	25	24	96.00	1	4.00
5	22	22	100.00	0	0.00
Total	113	102	89.76	11	10.24

Table 2: TRP Initial Experimental results

5. CONCLUSION

The combination result, from all 4 processes, is the phonetic representation of input syllable sound. The recognition accuracy is depended on process of gathering a period signal sample. The system performance can be improved in 2 aspects: 1) the recognition accuracy rate and 2) The speed of the system.

The absolute energy function may be using together with AMDF in VRP to improve the process of gathering a period signal sample that represent the vowel in the syllable. And this would yield the result in increasing the recognition accuracy rate of the system. The use of FAM instead of OFNN would increase the speed of the whole process. Further research and experiments should be conducted on LRP and TRP.

6. REFERENCES

[1] Pensiri, R. and Jitapunkul, S. "Speaker-Independent Thai Numerical Voice Recognition by using Dynamic Time Warping", Proceedings of the 18th Electrical Engineering Conference, 977-981, 1995

[2] Ahkuputra, V., Jitapunkul, S., Pornsukchandra, W. and Luksaneeyanawin, S. "A Speaker-Independent Thai Polysyllabic Word Recognition Using Hidden Markov Model", Proceedings of the 1997 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, 593-599, August 1997.

[3] Pornsukchandra, W. and Jitapunkul S. "Speaker-Independent Thai Numeral Speech Recognition using LPC and the Back Propagation Neural Network", Proceedings of the 19th Electrical Engineering Conference, 977-981, 1996

[4] Areepongsa, S. and Jitapunkul, S. "Speaker Independent Thai Numeral Speech Recognition by Hidden Markov Model and Vector Quantization", Proceedings of the 1997 International Symposium on Natural Language Processing, 370-378, 1995

[5] John R. Deller. JR., John G. Proakis and John H.L. Hansen. "Discrete-Time Processing of Speech Signal.", Prentice-Hall, 1993

A NEUROFUZZY FRAMEWORK FOR THAI SPEECH RECOGNITION*

PRATIT SANTIPRABHOB, TRAIPONG CHANCHARUNG,
RONNARIT CHEIRSILP, JIRAWAT CHAIAREERAT
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok, Thailand
pratit@s-t.au.ac.th

ABSTRACT

Speech recognition is a growing field of research in terms of both importance and the number of active researches. There are numerous approaches to tackling the problem. Some opt for a hard and complex computation. On the other hand, some do incline to soft computing. The results of those researches range from very basic template matching systems that can just memorize and recognize a few vocabularies of a single speaker to systems that parse speech into phonemes or some basic sound constituents and try to recognize them. Note that the latter approach is required in order to fully recognize all syllables of a given language. A hard and complex computation answer to the problem normally incurs a lot of overhead and is very difficult to perform in real-time. In addition, natural language speeches do contain uncertainties; even the same word spoken by the very same person twice do contain some subtle differences. Moreover, different languages do have their own unique characteristics. For example, Thai language contains a concept of tone which does not exist in typical western languages. This paper outlines a framework with a soft computing approach to Thai speech recognition. The proposed framework utilizes both fuzzy systems and neural networks with an aim to achieve a system for Thai speech recognition at phoneme-level.

KEYWORDS: Speech recognition, Fuzzy system, Neural network, Soft computing

1 INTRODUCTION

One important aspect in making computer more intelligent and easier to use by general public is to make the user interface friendlier. Speech recognition will undoubtedly provides users with the long-awaited and real-friendly user interface. However, in order to make the speech recognition useful in general, a template matching approach that memorizes and retrieves only limited set of vocabularies is inadequate. There have been a number of researches that try to improve upon the basic template matching approach by recognizing/learning syllables or even words by means of neural networks. Examples may be found in [1, 2, 3]. The latter even though is an improvement upon the former, it is still limited in the domain of application since the number of syllables or words to be learned must be bounded to a certain subset of those exist in a given language. To provide a universally deployable system, a more general approach that attempts to recognize sound constituents or phonemes and assemble them into syllables is required. This is because the number of sound constituents or phonemes in any given language is finite and fixed. Exhaustive learning can be achieved with a right architecture. Examples of such an approach are given in [4, 5, 6].

As for Thai speeches, there have been attempts as reported in [7, 8, 9] to recognize Thai words. However, the approach taken is to utilize a combination fuzzy system and neural network or some other techniques to recognize only a set of pre-trained vocabularies. On the other hand, this paper proposes a more general framework for Thai speech recognition at phoneme-level. The proposed framework takes an approach to recognizing Thai speeches by recognizing sound constituents or phonemes and assembling the recognized sound constituents into syllable representation. Both neural networks and fuzzy systems are used at different parts of the framework. Section 2 starts with the discussion of the scope of the framework followed by a brief explanation of Thai sound structure in Section 3. Then, techniques used in

* This research is supported in part by The Thailand Research Fund

syllable segmentation and phonetic classification of Thai sounds are discussed in Sections 4 and 5, respectively. Early experimental results are reported in Section 6. Finally, a conclusion is then given in Section 7.

2 SCOPE OF THE FRAMEWORK

The proposed framework takes a stream of digitized Thai sound, i.e. a spoken word or phrase. The input sound is sampled at the rate of 11,025 Hz and represented using 8 bits per sample. The input is recorded through a regular microphone coming with a PC-based standard multimedia set. Then, two main processing steps are performed on the sound.

- First, the sound is broken into syllables by means of syllable segmentation. This is quite a crucial step since the effectiveness and the accuracy of the second step depends on the correctness of this syllable segmentation.
- Secondly, each identified syllable is further analyzed and recognized in terms of a group of four sound constituents, namely vowel, leading consonant, ending consonant, and tone. This step may be called phonetic classification.

The recognized sound constituents are then represented for each syllable in a computer interpretable format. This representation can then be used as input into a natural language understanding system which could try to group syllables into words, and eventually determine their meanings. A schematic diagram of the overall framework is shown in Figure 1.

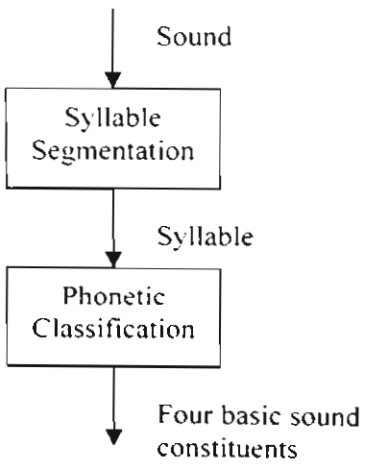


Figure 1: Overview of the framework

3 THAI SOUND STRUCTURE

Each syllable of Thai sound can be divided into four sound constituents: vowels, leading consonants, ending consonants and tones. Note that not every syllable has all the four sound constituents, e.g. ending consonant sound may not present. However, every single syllable must have a vowel sound and a tone.

- There are 24 vowels which can be further divided into 12 short vowels and 12 long vowels. Each short vowel has its long vowel counterpart, and vice versa. Of the 12 vowels in each group, there are only 9 main vowels (pure vowels). Each of the remaining three is a mixture of some of the main vowels. Each has a major vowel and a minor vowel to be detected separately.
- There are 38 different leading consonant sounds. 33 of them are from authentic Thai vocabularies, and the other five are from the vocabularies imported from English language. Among the 33 Thai leading consonants, 21 of them are individual consonants and the remaining 12 are clustered consonants.
- On the other hand, only 9 ending consonant sounds are used in Thai speech.
- A distinctive feature of Thai sound is the present of five different tones. The exactly same basic sound, i.e. same vowel, leading and ending consonants, pronounced with different tones result in different words with different meanings.

Vowels play an important role in the phonetic classification process. Since the patterns of both the leading and ending consonant sounds are embedded in the pattern of the vowel of the syllable, vowels must be identified first.

4 SYLLABLE SEGMENTATION

Syllable segmentation refers to a process of identifying the starting and ending boundaries of syllables in a given sound stream. This is the very crucial beginning of speech recognition. Inaccurate detection of syllable boundaries almost certainly leads to a failure in phonetic classification [10]. For any practical system, not only the accuracy but also the speed of computation is essential. In order to keep the computational overhead to the minimal, the proposed framework only utilizes the absolute energy. The input sound signal must be pre-processed by means of pre-emphasizing, amplitude normalization and en-framing before the absolute energy can be computed.

Pre-emphasizing is performed first to increase signal-to-noise ratio of the input signal. This is accomplished by passing the input signal through a first-order filter specified in (1).

$$\tilde{S}[i] = S[i] - aS[i-1] \quad (1)$$

where

$S[i]$ *is a speech signal at time i*

a is the pre - emphasized factor

Then, the pre-emphasized signal is normalized amplitude-wise according to (2) in order to re-scale the amplitudes to the maximum of 1. This would let us apply the same set of fuzzy rules regardless of how loud a particular speaker speaks.

$$S'[i] = \tilde{S}[i] / \max(\tilde{S}) \quad (2)$$

where

$\max(\tilde{S})$ *is the largest applitude of the input steam*

Then, the speech signal needs to be en-framed with appropriate overlapping between consecutive frames for a smooth transition between frames. The parameters used here are 256 samples per frame and the overlapping of 50%. The absolute energy (of each frame) can then be computed as specified in (3)

$$E_n = \sum_{i=1}^{w} | S'_n[i] | \quad (3)$$

where

$S'_n[i]$ *is a normalized speech sample in frame number n*

w is the number of samples per frame

Other energy computations including root mean square energy, square energy and Teager energy have also been tried out. The outcomes show that the absolute energy produces a very satisfactory result with the least computational overhead. Thus, the absolute energy is adopted in our framework.

The next step is to smooth out the energy graph in order to reduce local bumps and drops (max's and min's). This would help us reduce ambiguity level when we try to identify the starting and ending points of each syllable. In our framework, (4) is used to smooth out the energy graph.

$$E'_n = \frac{1}{2v+1} \sum_{i=n-v}^{n+v} E_i \quad (4)$$

where

v is the size of the smoothing window i.e. the number of the energy measurments to the left and right of this n measurement

Once the smoothed energy graph is obtained, it is subtracted by the silent energy level, i.e. the energy level generated during a silent period by the microphone system used. Finally, this adjusted smoothed energy graph is then used as input into a fuzzy rule-based system that determines the starting and ending points of each syllable in the speech signal. This fuzzy rule-based system basically tries to detect the changes in the energy graph. When a syllable ends, energy level drops. On the other hand, energy level rises when a new syllable begins. There are five input variables for this fuzzy rule-based system.

E: which is the current energy measurement E'_j

DELmax: which is $\max(E'_{j-5}, E'_j) - E'_j$

DERmax: which is $\max(E'_j, E'_{j-5}) - E'_j$

DEL1: which is $E'_{j-1} - E'_j$

DER1: which is $E'_{j-1} - E'_j$

There are two output variables.

START: which is the (truth) degree of the frame j being a starting point of a syllable.

END: which is the (truth) degree of the frame j being an ending point of a syllable.

The input variable E is defined in terms of four fuzzy sets: *Very Low*, *Low*, *Medium*, *High*. Note that here we want to identify the very low or low energy spot. The input variable DELmax is also defined in terms of four fuzzy sets: *Very Low*, *Low*, *Medium*, *High*. High value of this signifying the downward slope of energy graph which leads to an ending point of a syllable. The input variable DERmax is similarly defined in terms of the four fuzzy sets: *Very Low*, *Low*, *Medium*, *High*. However, the membership functions of these fuzzy sets are slightly different from those of the fuzzy sets defined for DELmax since the upward slope is normally steeper than the downward one. High value of DERmax signifies the upward slope leaving from a starting point of a syllable. The DEL1 and DER1 are used only to confirm the sign of the slope into and out of the point of concern. Here we only need to identify whether it is positive value or not, therefore, only one fuzzy set is defined for each input, i.e. *Positive*. In our rules either

Positive or the negation of it is used. The fuzzy sets of E are defined on [0, 256], while the other fuzzy sets are defined on [-256, 256].

On the other hand, each of the two output variables has the fuzzy terms of *Low*, *Medium*, and *High* defined on the truth values between 0 and 1. The basic *max-min* inference method is used together with the *centroid* defuzzification method. The output values for both output variables are then computed for every frame.

Examples of the fuzzy rules used in the syllable segmentation step are shown below.

Rule 1:
If (E = L) and (DELmax = None) and (DERmax = M) and (DEL1 = None) and (DER1 = P) then (START = H) and (END = None)

Rule 2:
If (E = VL) and (DELmax = not VL) and (DERmax = None) and (DEL1 = P) and (DER1 = None) then (START = None) and (END = H)

For the output values at each frame, the post-processing is performed in order to determine the starting and ending points of each syllable. This also involves memorizing the current state of the boundary point detected. The post-processing algorithm is as follows.

1. Set *state* to Ending. Find the first frame that has a resulting high degree of START. This is marked as the first starting frame of the first syllable in the signal steam. Set *state* Starting.
2. Search for a frame that has a medium or high degree of END or a medium or high degree of START.
   - if the END degree is high, then
     if the *state* is Starting, then
     case of the length from the starting frame **is**
     >= the minimum medium-syllable size:
     mark this frame an ending frame, and set *state* Ending
     < the minimum short syllable size:
     if the last START degree is medium, then
     remove the last starting frame and replace the last ending frame with this frame, and set *state* Ending
     else do nothing
     others:
     if the last START degree is medium and the length of previous syllable < minimum long-syllable size, then
     remove the last starting frame and replace the last ending frame with this frame, and set *state* Ending
     else mark this frame an ending frame, and set *state* Ending
     end case
     else replace the last ending frame with this frame
   - ⇒ if the END degree is medium, then
     if the *state* is Starting, then
     if the length from the last starting frame >= the minimum long-syllable size, then
     mark this frame an ending frame, and set *state* Ending
     else do nothing
     else replace the last ending frame with this frame
   - ⇒ if the START degree is high, then
     if the *state* is Ending, then
     mark this frame a starting frame, and set *state* Starting
     else do nothing
   - ⇒ if the START degree is medium, then
     if the *state* is Ending then
     if the length from the last starting frame >= the minimum long-syllable size, then
     mark this frame a starting frame, and set *state* Starting
     else do nothing
     else do nothing
3. If it is not the end of the speech signal stream, go back to 2.

As a result of the post-processing algorithm, starting and ending frames that demarcate syllables are located. The digitized sound sample at the middle of the frame, i.e. sample number 128 of the frame, is used as a demarcating sample. The identified syllables are then used as input into the second step of phonetic classification.

5 PHONETIC CLASSIFICATION

The architecture of the proposed phonetic classification part for this Thai speech recognition framework is shown in Figure 2. This part of the framework is divided into five different modules. It is assumed that the syllable's starting and ending points have been correctly detected by the syllable

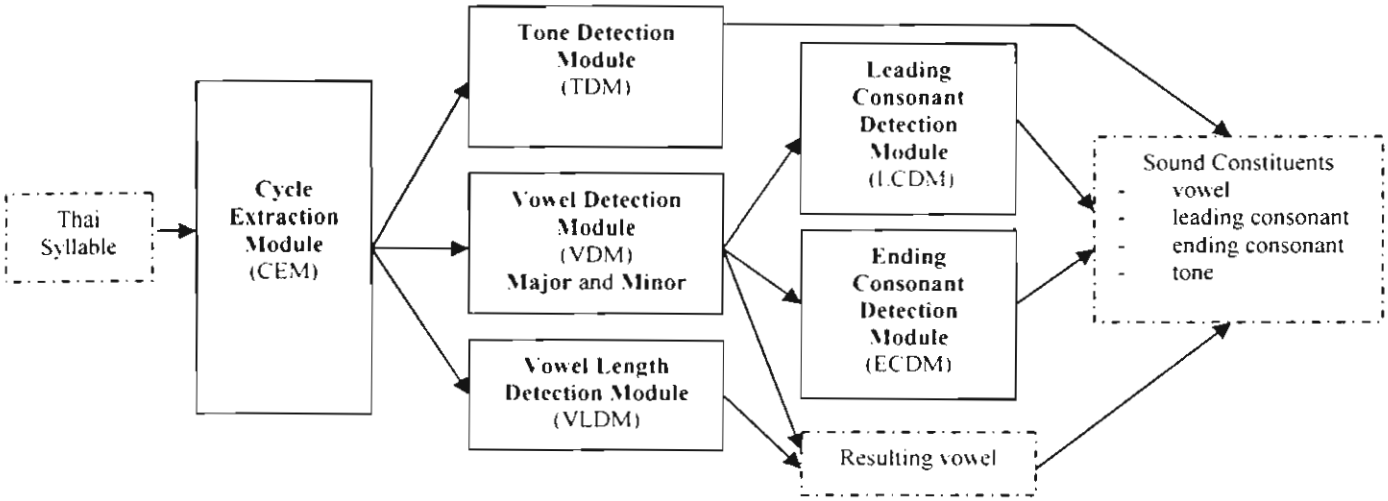


Figure 2: Subsystem of phonetic classification part

segmentation part. The phonetic classification part processes one syllable at a time. One major task in this part is to identify a sample cycle at different intervals of the input signal. This is used to calculate fundamental frequency for the purpose of tone detection, and also used to extract an input signal of one cycle for the vowel and consonant detection. In our framework, the Average Magnitude Difference Function (AMDF) is used to identify sample cycles.

A pre-processing step starts by dividing a given syllable into 30 equal intervals. Then, a sample cycle of each interval is extracted by means of AMDF. Here, we have both the shape and the size (in terms of number of samples) of each cycle extracted. These 30 sample cycles are later used as input to different detection modules. This pre-processing step is performed in the Cycle Extraction Module (CEM)

The Tone Detection Module (TDM) uses the sample cycle of each interval to determine the interval's fundamental frequency by means of (5).

$$f = \frac{s}{p} \qquad (5)$$

where

f is frequecy (Hz)

s is sampling rate (samples / sec)

p is period size (samples / cycle)

Then, the 30 intervals are grouped into five groups of six intervals each. The median frequency among those of the six intervals is used to represent each group.

Before applying fuzzy rules, a pre-processing is needed to normalize the resulting fundamental frequencies. This reduces the frequency discrepancies between male's and female's voice. In our framework, the frequencies of a given syllable are normalized with respect to the representative frequency of the first group of the syllable. Since each tone in Thai language has a different movement pattern of frequencies with respect to time, the differences between the representative frequency of each group and that of the first group are used as input into our fuzzy rule-based system. Hence, four input variables are used in our system, each is calculated by (6).

$$df_i = \frac{100}{f_1} * f_{i+1} - 100 \qquad (6)$$

Each of the four input variables has the same set of seven fuzzy sets namely, *Negative Big*, *Negative Medium*, *Negative Small*, *Zero*, *Positive Small*, *Positive Medium*, *Positive Big* defined on [-50,50]. While, the output comprises five variables – one for each tone. There are three fuzzy sets: *Low*, *Medium*, *High* for each output variable. They are defined on the truth values between 0 and 1. Examples of the fuzzy rules employed in TDM are shown below.

Rule 1:
If (df1 = NS) and (df2 = NS) and (df3 = NS) and (df4 = NB) then (T1 = L) and (T2 = H) and (T3 = M) and (T4 = L) and (T5 = L)

Rule 2:
If (df1 = ZE) and (df2 = ZE) and (df3 = NS) and (df4 = NM) then (T1 = H) and (T2 = L) and (T3 = L) and (T4 = L) and (T5 = L)

As for the different neural network based detection modules, the sample cycles extracted from each syllable need to be normalized by both time and amplitude before being used as their input. Each sample cycle is re-sampled into 100 sample points which correspond to input nodes of the neural networks.

The Vowel Detection Module (VDM) requires the use of back-propagation neural networks for training on sample data. This is since there is no structure knowledge that can easily be captured into fuzzy rules in this case unlike the TDM. The neural network used has 100 input nodes, one hidden layer with 15 nodes, and 9 output nodes corresponding to the number of different major vowel types. Note, however, that the resulting neural network can later be mapped onto fuzzy associative memory (FAM) which can basically be used as a fuzzy rule-based system. This VDM needs to detect both the major and minor parts of a vowel if that vowel is of the mixture type. The very same neural network is used for the detection of both vowel parts. The only difference is on the interval from where the input sample cycle is extracted.

After the vowel is detected, another two separate sets of neural networks can be trained to detect the leading consonant – Leading Consonant Detection Module (LCDM) and the ending consonant – Ending Consonant Detection Module (ECDM). The consonant detection must be done after the vowel is detected since the pattern of the consonants is embedded in and depended upon the vowel of the syllable. The LCDM contains 9 neural networks: one for each main vowel type. Each neural network has 100 input nodes, one hidden layer with 15 nodes, and 38 output nodes corresponding to 38 possible leading consonant sounds. Similarly, the ECDM also has 9 neural networks corresponding to the 9 main vowels. Each neural network has 100 input nodes, one hidden layer with 15 nodes, but only 9 output nodes representing 9 different ending consonant sounds. Likewise, these neural networks can also be mapped onto FAM for a more efficient operation after the training is completed.

Note that in addition to VDM, a vowel length detection module (VLDM) is also required to classify the recognized vowel whether it is a long or a short vowel. This module simply calculates the time span of the syllable's vowel by looking at how many intervals the vowel sound occupies. Then, the fundamental frequency calculated for each of those occupied intervals is used to calculate approximate time period of the interval. The total time span is then compared to a threshold to determine the vowel length.

All these five detection modules together analyze each input syllable and result in the four basic sound constituents of Thai syllable, namely vowel, leading consonant, ending consonant, and tone.

6 EXPERIMENTAL RESULTS

The proposed framework is being constructed piece-wise using MATLAB*. Testing has only been performed on individual parts. Multiple speakers both males and females are enlisted to provide sample data for system training as well as to provide test data for the system. The initial results are quite promising. The syllable segmentation part yields an accuracy of over 95%. The tone detection module yields an accuracy level of about 90%, while the vowel detection module achieves an accuracy level of over 80%. Early trials on both leading and ending consonant detection modules also show quite satisfactory results. The system developed is in the process of integration and tuning. It is expected that the overall accuracy level of the integrated system would reach the level of at least 80% when tested with life data.

7 CONCLUSION

The proposed framework is an attempt to provide a system that could be used to recognize Thai speech in general. It first segments a given sound stream into syllables. Then, each syllable is analyzed in terms of its four basic constituents. In order to reduce computational overhead of traditional hard computing, neural network and fuzzy rule-based system technologies are employed in different parts of the framework. Neural networks are required in the training phase of certain detection modules, and would be mapped onto fuzzy rule-based systems (in terms of fuzzy associative memory) for the operational phase. This would result in a system that can cope with ambiguity inherently presenting in human speech, yet can achieve an efficient computation due to the use of fuzzy rule-based system.

8 REFERENCES

[1] L. Bottou, F.Fogelman Soulié, P.Blanchet, and J.S.Liénard. Speaker-Independent Isolated Digit Recognition: Multilayer Perceptron vs. Dynamic Time Warping. *Neural Networks*, 3, 1990, pp. 453-465.

[2] Cheng-Yuan Liou and Chwan-Yi Shiah. Perception of Speech Signals Using Self-Organization on Linear Neuron Array, *International Joint Conference on Neural Networks*, 1, 1993, pp. 251-253.

[3] L. Fissore, P. Laface, and F. Ravera. Using Word Temporal Structure in HMM Speech Recognition. *Proc. IEEE International*

Conference on Acoustics, Speech, and Signal Processing, 1997, pp. 975-978.

[4] X.D.Huang. Phoneme Classification Using Semicontinuous Hidden Makov Models, *IEEE Trans. Signal Processing*, 40(5), 1992, pp. 1062-1067

[5] David J. Pepper and Mark A. Clement. Phonetic Recognition Using a Large Hidden Markov Model, *IEEE Trans. Signal Processing*, 40(6), 1992, pp. 1590-1595.

[6] Åge J. Eide and Terje Lindén, Recognizing Norwegian Vowels using Neural Network, *International Neural Network Society*, 1994, pp. 524-537.

[7] C. Wutiwiwatchai, S. Jitapankul, V. Ahkuputra, E. Maneenoi, S. Luksaneeyanawin. Thai Polysyllabic Word Recognition using Fuzzy-Neural Network, *Proc. International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[8] C. Wutiwiwatchai, S. Jitapankul, V. Ahkuputra, E. Maneenoi, S. Luksaneeyanawin. A New Strategy of Fuzzy-Neural Network for Thai Numeral Speech Recognition. *Proc. International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[9] V. Ahkuputra, S. Jitapunkul, N. Jittiwarangkul, E. Maneenoi, S. Karuriya. A Comparison of Thai Speech Recognition Systems using Hidden Markov Model, Neural Network, and Fuzzy-Neural Network, *Proc. International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[10] Viying Zhang, Xiaoyan Zhu, Yu Hao. A Robust and Fast Endpoint Detection Algorithm for Isolated Word Recognition, *Proc. IEEE International Conference on Intelligent Processing Systems*, Beijing, China, 1997.

Fuzzy-based Thai Syllable Segmentation for Connected Speech using Energy and Different Ceptral*

Jirawat Chaiareerat[1] and Pratit Santiprabhob[2]
Intelligent System Laboratory
Department of Computer Science
Faculty of Science and Technology
Assumption University
Bangkok, 10240,Thailand
Email: jirawat[1], pratit[2]@s-t.au.ac.th

Abstract: In speech recognition system, the fundamental task is to perform syllable segmentation of speech signals. The efficiency of the syllable segmentation affects the performance of the entire recognition system. This process should be fast and use little computational power to make good response time of the system.

This paper proposes a framework with an algorithm to segment Thai syllables by means of Fuzzy System. Four features of a given speech (High Amplitude Rate, Absolute Energy, Zero Crossing Rate and Different Ceptral) are used as parameters in the proposed framework. A Fuzzy Inference Systems (FIS) is constructed to determine the starting and ending point of the syllables using those features.

Keywords: Thai Speech Recognition, Syllable Segmentation, Endpoint Detection, Signal Processing, and Fuzzy System.

1. Introduction

Most of the errors in the automatic speech recognition system are the inaccurate detection of the beginning and ending boundaries of test and reference patterns [4].

The major usage of the automatic speech recognition system is in real-time system, so the speed and response time of the system is important.

From those reasons, the good syllable segmentation algorithm should be fast and has a good accuracy. Nowadays there are several algorithms proposed for syllable segmentation.

The followings are typical word boundary detection algorithms.

1) Energy based algorithms with automatic threshold adjustment [2][3]: They are intuitive approaches based on energy levels and duration of silence and speech. Sometimes several pairs of boundaries are yielded in the order of their ranks of being correct.

2) Noise adaptive algorithms [3]: They use the logarithm of the RMS signal energy, the zero-crossing rate, duration information, and a set of heuristics. The thresholds used for the energy and the zero-crossings are adapted automatically according to a few sample frames provided by the environment signal.

3) Algorithms using frequency-based features [4]: These algorithms perform FFT transformations, and compute the energy in the frequency-band 250-3500Hz and logarithm of the RMS energy.

4) Algorithm using Energy and Fuzzy System [1]: These algorithms are the same as Energy based algorithm but instead of using threshold, they used Fuzzy System to determine the starting and ending point.

The first three algorithms are based on threshold, which is not flexible enough. The fourth algorithm is better but using only energy cannot archive a good accuracy when working with a spontaneous speech.

This paper proposes an algorithm to segment Thai syllables using High Amplitude Rate (HAR), Absolute Energy (ABS Energy), Zero Crossing Rate (ZCR) and Different Ceptral (DC) of a given speech as parameter to the system. Finally, Fuzzy Inference System is used to determine the starting and ending points of syllables.

Obtaining DC requires a lot of computational overhead but the computed DC can also be used as speech features of syllable recognition system. Therefore, we do not consider about DC computation time as it can make the system slow.

The main framework is designed to be a part of the Connected Speech Recognition System consisting of three sub-systems: Syllable Segmentation, Syllable Recognition and Word Associative system as shown in figure 1.

* This research is supported in part by The Thailand Research Fund.

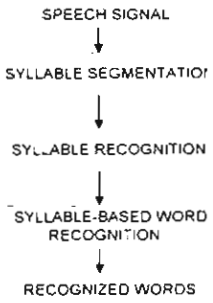


Figure 1: The Connected Speech Recognition System.

2. Description of algorithm

The proposed algorithm consists of three steps: Parameters computation, Threshold based Segmentation and Fuzzy based Segmentation. First, the speech signal is pre-processed. And then the parameters are calculated. Some parameters are used to segment the speech signal. Finally, the FIS is used to identify the ending point and starting point of each syllable in each speech segment.

2.1. Parameters Computation

First, the speech signal is pre-emphasized with first order low-pass filter. The factor of pre-emphasize is 0.95 [3][7].

The pre-emphasized speech signal is then en-framed with, 30 ms long for each frame and the overlapping factor of 20 ms; this overlapping is essential for a smooth transition from one frame to another frame [3].

For each frame, the following four parameters are calculated:

1. High Amplitude Rate (HAR): HAR is the number of sampling that have the amplitude greater than the threshold.
2. Absolute Energy (ABS Energy): ABS Energy is sum of the absolute value of the amplitude in the frame [1][2][3].
3. Zero Crossing Rate (ZCR): ZCR is the rate of the signal crossing at zero level [7].
4. Different Ceptral (DC): DC is the different of ceptral [7] between the frame *i-1* and *i+1* when *i* is the current frame.

Finally, the contours of above four parameters are then smooth by Moving Average Smoothing [2].

2.2. Threshold based Segmentation

In this step, we eliminate the silent speech by using the threshold to segment the speech signal into the group of syllables (speech segment).

The speech signal is searched from the first frame to find the pairs of starting frames and ending frames as shown in the figure 2. We use the following rules to determine whether frame *i* is starting frame or ending frame or neither.

If Energy[i]>E_th1 or HAR[i]>HAR_th1 then frame I is staring frame.

If Energy[i]<E_th2 or ZCR[i]=0 or HAR[i]<HAR_th2 then frame I is ending frame.

Where:

Energy[i] is the ABS Energy at frame i
HAR[i] is the HAR at frame i
ZCR is the ZCR at frame i
E_th1 and E_th2 are Energy thresholds calculated from the background noise at the beginning of the speech signal.
HAR_th1 and HAR_th2 are HAR thresholds calculated from the background noise at the beginning of the speech signal.

The result from this step is the speech segments, which will be segmented again in next step.

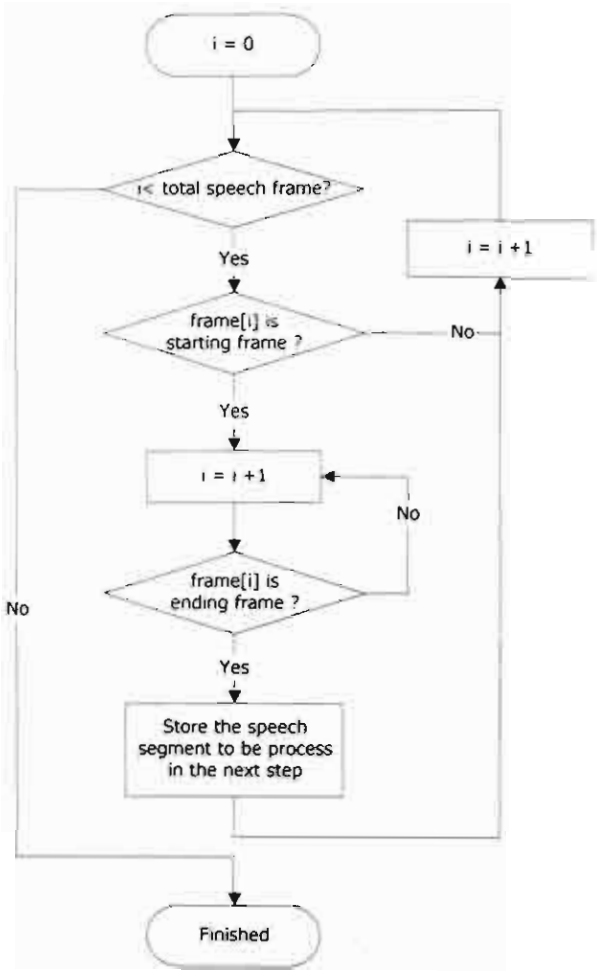


Figure 2: The algorithm to detect the starting and ending frames.

2.3. Fuzzy based Segmentation

Each speech segments from 2.2 is segmented again in this step. There are four steps in segmentation. First, PeakE frames are examined from the speech segment. Then, Emin frames are identified. For each Emin frames, five fuzzy input variables are calculated. Finally, Fuzzy Inference

System is used to determine the frame whether it is boundary frame or not from those four variables.

2.3.1. Identifying PeakE frames

Each frame of the speech segment is determined whether it is PeakE frame or not. The frame will be PeakE frame if it agrees the following condition.

If E[i] > MAX(E[i-3..i+3]) or MAX(E[i-3..i+3])-MAX(MIN(E[i-3:i]),MIN(i..i+3)) then frame is PeakE frame.

Where:

E[i] is the ABS Energy at frame I

E[i:j] is the array of ABS Energy from frame I to frame j

MAX(A) is the maximum value of all elements in array A.

MAX(a,b) is the maximum value from a and b

MIN(A) is the minimum value of all elements in array A.

2.3.2. Identifying Emin frames

Between two PeakE frames, the minimum energy frame will be searched and identified as Emin frame.

2.3.3. Fuzzy Input calculation

For each Emin frame, we calculate the following five Fuzzy input variables.

1. EM: the ABS energy of the current Emin frame.
2. ZM: the minimum ZCR between the previous PeakE frame and the next PeakE frame.
3. DEL: the difference of the EM and the ABS Energy of the previous PeakE frame.
4. DER: the difference of the EM and the ABS Energy of the next PeakE frame.
5. DCMAX: the maximum DC between the previous PeakE frame and the next PeakE frame.

2.3.4. Fuzzy Inference

Five fuzzy input variables from 2.3.3 are sent to the Fuzzy Inference System (FIS) [6][8] to determine whether the current Emin frame is the Boundary frame or not. The Boundary frame is the ending frame of syllable and the beginning of the next syllable. Therefore, each syllable in the speech segment can be identified by the locations of Boundary frames. After the Boundary frames are located, the center speech signal samples of these frames are determined as the Boundary point of the syllables.

The proposed FIS model is Mamdni-type Inference engine using centroid defuzzification method [8]. The model has seven rules, five inputs and one output. The fuzzy sets of each variable are shown in table 1.

Variables	Fuzzy Set
EM	Low, Medium, High
ZM	Low, Medium, High
DEL	Very Low, Low, Medium, High
DER	Low, Medium, High
DCMAX	Low, Medium, High
OUTPUT	Low, Medium, High

Table 1: The Fuzzy Set of each variable.

The proposed FIS basically try to detect the changes in the contours of energy and DC. At the Boundary frame, the ABS Energy is low, the different of the ABS Energy with the previous PeakE frame and the next PeakE frame is high and the DCMAX is also high.

The followings are the rules of the proposed FIS.

1 if (EM is Low) and (DEL is High) and (DER is High) then OUTPUT is High

2 if (EM is Low) and (DCMAX is High) then OUTPUT is High

3 if (ZM is Low) and (DCMAX is High) then OUTPUT is High

4 if (DEL is High) and (DER is High) and (DCMAX is High) then OUTPUT is High

5 if (DEL is Very Low) and (DER is High) and (DCMAX is Low) then OUTPUT is Low

6 if (EM is Medium) and (DEL is Medium) and (DER is not Low) and (DCMAX is High) then OUTPUT is High

7 if (EM is Medium) and (DEL is High) and (DER is not Low) and (DCMAX is High) then OUTPUT is High

3. Experimental Result

There are four speakers in this experiment. Each of them speaks two sets of sentences: 36 ambiguous Thai sentences [3] (Set A) and 53 Thai general spontaneous speaking sentences randomly selected from conversations and books (Set B). Generally, each sentence in Set B is longer and spoken faster than Set B. The average syllables per sentences of set A and set B are 7.5 and 10.3 respectively.

Each sentence is recorded using 16-bit quantization level, and sampling rate at 11.025 KHz. The results are shown in the tables 2 and 3.

From experimental result, there are three kinds of incorrect segmentation:

1. Merge: two syllables are merged into a syllable.
2. Location: wrong location of the syllables or one syllables is spliced into two syllables.
3. Lost: a syllable is lost.

The accuracy of the sentences set A and B is 93.08% and 92.96% respectively.

Speakers	Total Syllables	Correct		Incorrect					
				Merge		Location		Lost	
		Syllables	Percent	Syllables	Percent	Syllables	Percent	Syllables	Percent
A	271	258	95.20	8	2.95	5	1.85	0	0.00
B	271	251	92.62	6	2.21	12	4.43	2	0.74
C	271	249	91.88	9	3.32	13	4.80	0	0.00
D	271	251	92.62	8	2.95	12	4.43	0	0.00
Total	**1084**	**1009**	**93.08**	**31**	**2.86**	**42**	**3.87**	**2**	**0.18**

Table 2: The results of the sentences set A.

Speakers	Total Syllables	Correct		Incorrect					
				Merge		Location		Lost	
		Syllables	Percent	Syllables	Percent	Syllables	Percent	Syllables	Percent
A	537	506	94 23	18	3 35	13	2 42	0	0 00
B	555	522	94 05	18	3.24	11	1 98	4	0 72
C	557	516	92 64	30	5.39	11	1 97	0	0 00
D	552	502	90 94	36	6 52	14	2 54	0	0 00
Total	**2201**	**2046**	**92.96**	**102**	**4.63**	**49**	**2.23**	**4**	**0.18**

Table 3: The results of the sentences set B.

4. Conclusion

The proposed system is an attempt to provide a Thai Syllables segmentation system to be integrated with the Thai Syllables Recognition system and Thai Syllable-based word recognition in order to become a complete Thai Speech Recognition System.

Four features have been used and incorporated with FIS in the proposed syllable segmentation algorithms. From the experimental result, the accuracy of the system is acceptable and the difference of the accuracy between set A and set B is only 0.01%.

In the sentences set A, most of the errors are Location Errors, but in set B they are Merge Errors. The reason is that most of the sentences in set B are longer and spoken faster than the sentences in set A.

From the experiment, the causes of each the error type is identified as the followings.

1. Merge:
   - Too fast speaking.
   - "Open Syllable" which the successive syllable can be spoken continuously without a drop of energy (ex. /ma/lee/).
2. Location:
   - Too slow speaking (Some syllables can be considered as either two syllables or one syllable when they are very slowly spoken. For example, the syllable "/tua/" can be recognized as "/tu/ua/").
3. Lost:
   - Too low volume speaking.

The system is being fine-tuned to reduce such errors to the minimal.

5. Further Research

Using the automatic learning FIS system such as Adaptive-Network-based Fuzzy Inference Systems [6] will be helpful to optimize the FIS in the proposed system.

More research works on this topic can be done on the training-based system with a large Thai speech corpus [5].

6. References

[1] Jirawat C. and Pratit S., *Thai syllables segmentation for connected speech with fuzzy system*, Proceedings of the ICAI, vol. 1, pp.387-392, 2000.

[2] Jittiwarangkul N., Jitapunkul S., Luksaneeyanawin S., Ahkuputra V., Wutiwiwatchai C., *Thai Syllable Segmentation for Connected Speech based on Energy*, Proceedings of the IEEE APCCAS, pp. WP1-8.1, Nov. 1998.

[3] Nutthacha Jittiwarangkul, *Syllable segmentation algorithm for Thai connected speech*, Master's Thesis, Chulalongkorn University, 1998.

[4] J-C junqua, Brain Mak, and Ben Reaves, *A robust algorithm for word boundary detection in the presence of noise*, IEEE Transactions on Speech and Audio Processing, vol. 2, no. 3, Jul 1994.

[5] S. H. Chen, Y. F. Liao, S. M. Chiang, and S. Chang, *An RNN-based Pre-classification Method for Fast Continuous Mandarin Speech Recognition*, IEEE Trans. Speech and Audio Processing, Vol.6, No.1, pp.86-90, Jan. 1998.

[6] Jang, J. -S. R., *ANFIS: Adaptive-Network-based Fuzzy Inference Systems*, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 23, No. 3, pp. 665-685, May 1993.

[7] L. Rabiner, B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall , 1993.

[8] MathWorks, Inc., *Fuzzy Logic Toolbox for use with MATLAB User Guide, Version 2*, 1999.

Phoneme-Based Thai Syllable Recognition by Means of Soft Computing*

Ronnarit Cheirsilp and Pratit Santiprabhob
Intelligent System Laboratory
Department of Computer Science
Faculty of Science and Technology
Assumption University
Bangkok, 10240,Thailand
Email:ronnarit, pratit@s-t.au.ac.th

Abstract: Thai speech can be recognized at syllable-level. Most Thai speech recognition systems use the word-based approach. This approach has a limit number of words and syllables that they can recognize. This paper proposes a phoneme-based recognition subsystem for Thai connected speech recognition system, which recognizes syllable sounds from their phonemes using soft computing: Neural Network (NN) and Hidden Markov Model (HMM). In the proposed system, a Thai syllable sound is recognized as a sequence of four phonemes. Namely, they are leading consonant, vowel, ending consonant and tone. The proposed system is also designed to have a capability to recognize all Thai syllable sounds and take the advantages of NN and HMM technologies.
Keywords: Speech recognition, Hidden Markov Model, and Neural Network.

1. Introduction

Speech is used as a primary media for human beings to communicate to each other. For a word "speech" here, we mean a group of syllables, words, phrases, or sentences. People can understand speech because they can recognize it. In the present, computers have become a part of our day life. Speech recognition is an exciting and challenging technology that changes the way we interact with computers in the future.

Based on our view, the connected speech recognition system consists of three parts as shown in figure 1. In this system, the speech to be recognized is segmented into syllables first. Next, each syllable is recognized. Finally, a sequence of recognized syllables is decoded to a sequence of recognized words. This paper proposes a phoneme-based syllable recognition system, which is a mapping between syllables and their phonetic representations. This system is designed to be a part of our conceptual Thai connected speech recognition system

In our reviews, five successful Thai speech recognition systems have been reviewed. Conclusively, the techniques used in these systems are Dynamic Time Wrapping [6], Conventional Neural Network [7], Modified Back Propagation Neural Network [5], Neural Network with Fuzzy MF Preprocessor [11], and Hidden Markov Models [2]. From [1][3], the conclusion on recognition performance of Thai numerals using all of the above techniques has been made. The HMM technique [2] yields the best recognition rate.

From our review of these techniques, the following disadvantages have been discovered.

1) All of the above systems use the word-based speech recognition approach. They can recognize only the small vocabulary such as numbers, names and commands.
2) For [2], the system cannot be applied with the connected word. Because it has to detect number of syllable before it can perform recognition.
3) All of the above approaches are not suitable for large vocabulary recognition because they require a lot of computation.

To overcome the above approaches, our approach is designed to be general speech recognition system. Instead of recognizing the whole word, our approach is to segment the word or speech into syllable units and perform the recognition on them.

* This research is supported in part by The Thailand Research Fund.

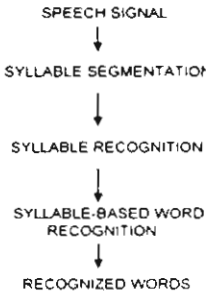


Figure 1: Our Conceptual Connected Speech Recognition System.

2. Thai syllable structure

In Thai language, a syllable consists of a set of phonemes. Each Thai syllable sound is comprised of four phonemes: leading consonant, vowel, ending consonant, and tone.

There are 38 different leading consonants. Actually, 33 of them are from Thai vocabularies, and the other five are borrowed from English vocabularies. These 38 leading consonants can also be classified into two groups: non-cluster and cluster. There are 21 non-clusters and 17 clusters. The cluster is the combination of two different leading consonants.

There are 24 vowels, which can be divided into two groups regarding to the length of vowel sound: 12 short-vowels and 12 long-vowels. Each group can be subdivided into nine major vowels (pure vowel) and three minor vowels (diphthong or the combination of two different major vowels).

There are only nine ending consonants in Thai phonemes. Not all ending consonants can occur with all vowel phonemes. Some syllables may not have ending consonant. Null ending consonant is also counted as one ending consonant class.

Only five tones are present in Thai syllable sounds.

3. Thai syllable analysis

The syllable sounds that consist of different phoneme classes are recorded using 16-bit quantization level, and sampling rate at 11.025 KHz. These syllable signals are then analyzed. By conducting this analysis on these syllable signals, we observed that each of three phoneme classes that constitute the syllable have the distinct signal sound patterns. These phoneme classes are leading consonant, vowel and ending consonant. From our analysis, the sound patterns of leading consonant, vowel and ending consonant can be located at the beginning, middle and ending parts of syllable signal respectively. It is difficult to locate the sound patterns of pure leading consonant and ending consonant in the syllable signal. But it is easy to locate the sound pattern of leading consonant in combination with vowel, and ending consonant in combination with vowel. When each of these sound patterns is played, we heard its sound like a single-syllabled word. And there is limited number of these sound patterns. From these reasons, we decided to treat them, as they are isolated words. Hence, the HMM technique, which gives the best recognition rate for isolated word recognition from our reviews, is used for recognizing these phonemes.

For tone phonemes, we can distinguish them using the fundamental frequency contours of the syllables. The fundamental frequency contours of all five Thai tone phonemes from [4] are given in figure 2. Therefore, we decided to use the Neural Network technique in recognizing tone phonemes.

Figure 2: Fundamental Frequency Contours of All Five Tones.

4. Proposed system

Each Thai syllable sound comprises four different types of phoneme, namely leading consonant, vowel, ending consonant, and tone. Actually, there are approximately several thousands of Thai syllable. It is not practical to recognize them all. In the other hand, there are altogether approximately 76 phonemes in Thai language. Therefore, the components of a syllable should be recognized instead.

The overall system comprises five processes namely leading consonant, vowel, and ending consonant (LVET) feature extraction process, leading consonant recognition process (LRP), vowel recognition process (VRP), ending consonant recognition process (ERP) and tone recognition process (TRP). A block diagram of the overall system is given in figure 3. The details of five processes are described in the following subsections.

The speech feature extraction process has a duty to extract the needed speech features for all four recognition processes. Note that each recognition process requires its own set of speech features in order to function. And each recognition process is responsible for recognizing each phoneme part of the syllable as corresponding to its name. For example, the leading consonant recognition process is responsible for recognizing leading consonant phonemes.

For each unknown syllable signal s, which is to be recognized, the processing in figure 3 must be carried out. The steps in the processing are as follows:

1) The LVET feature extraction process extracts four different sets of speech features from the speech signal for leading consonant, vowel, ending consonant and tone recognition process.
2) The speech features for vowel and tone recognition processes are then processed by vowel recognition and tone recognition processes simultaneously. The indices of both recognized core-vowel *cv* and vowel *v* are generated as the output of vowel recognition. Then the core-vowel index *cv* is passed to leading and ending consonant recognition process. The tone recognition process generates the index of recognized tone *t* as its output.
3) The leading and ending consonant speech features and an index of recognized core-vowel *cv* are then passed to leading consonant and ending consonant recognition process simultaneously. The output of leading consonant and ending consonant recognition are the indices of recognized leading consonant *lc* and ending consonant *ec* respectively.
4) Finally, all indices of recognized leading consonant *lc*, vowel *v*, ending consonant *ec*, and tone *t* altogether present a recognized syllable *rs*.

Note that leading and ending consonant recognition process depends on the core-vowel recognition from the vowel recognition process.

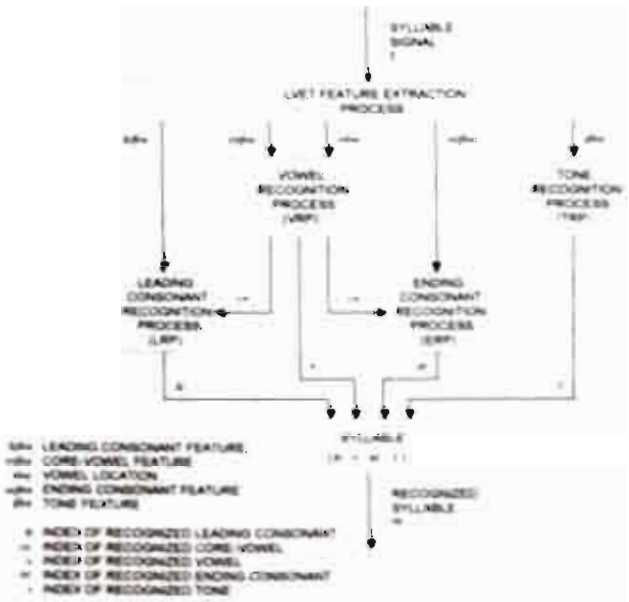


Figure 3: A Block Diagram of the Proposed System.

4.1. LVET feature extraction process

This process is responsible for extracting all features needed for the four recognition processes (LRP, VRP, ERP, and TRP). This process is separated into two parts leading consonant, vowel and ending consonant features extraction and tone feature extraction. LRP, VRP, and ERP all use the same feature type but TRP use the different one.

4.1.1. Leading consonant, vowel, and ending consonant feature extraction

The following steps are carried out in order to extract features for leading consonant, vowel and ending consonant recognition.

LPC analysis and energy measurement. A block diagram of this step is given in figure 4. LPC analysis and energy measurement are performed on the syllable signal. The details of LPC analysis and energy measurement are described in [8][9]. The result from LPC analysis and energy measurement is a sequence of cepstral coefficients and energy vectors representing each signal frame. We call them together as *cep_e*.

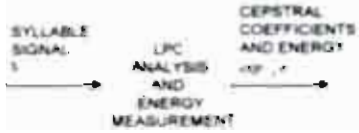


Figure 4: A Block Diagram of LPC Analysis and Energy measurement.

Vowel location detection. The feature vector *cep_e* is used in determining the vowel location of the syllable signal. The vowel location *vloc* contains two values. They are the starting and ending frame numbers of vowel location in a given syllable signal. In order to determine the vowel location, the following steps are performed.

1) First step is to determine the Euclidean distance between cepstral feature vector frames. Then a series of difference cepstral values is generated.
2) The next one is to find the starting frame of vowel. Two thresholds are used, namely cepstral and energy threshold (the average energy of the syllable signal). The starting frame of vowel is the first frame in a sequence that lies between two frames, which have difference cepstral value higher than the cepstral threshold. Moreover, the signal energy of the searched frame must higher than the energy threshold. Searching of this frame number must be done in forward direction starting from the first frame in the sequence to the last one.
3) To determine the ending of the frame, same step in 2) is processed, but, this time, opposite direction starting from the last frame to the first one.

LVE Feature Segmentation. After the vowel location *vloc* has been determined. The *cep_e* feature vector is segmented into three parts based on vowel location. These three parts are *lcfea*, *cvfea*, and *ecfea*. They are used as the feature vectors of LRP, VRP, and ERP respectively. A

block diagram of this segmentation is given in figure 5. The following cases are applied in segmentation.

- For leading consonant feature *lcfea*, all feature frames between its first frame of *cep_e* and the starting frame number of vowel location, are segmented and used as *lcfea* for LRP.
- For core-vowel feature *cvfea*, all feature frames between the starting and the ending frame numbers of vowel location, are segmented and used as *cvfea* for VRP.
- For ending consonant feature *ecfea*, all feature frames between the ending frame number of vowel location and the last frame of *cep_e* feature vector, are segmented and used as *ecfea* for ERP.

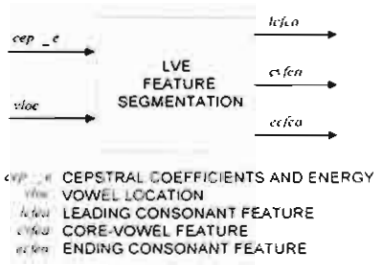


Figure 5: A Block Diagram of LVE Feature Segmentation.

4.1.2. Tone feature extraction

A tone feature vector *tfea* is computed over a syllable signal. The extraction steps are as follows.

1) *Cepstral Pitch Detection:* A fundamental frequency contour f_0 is computed over a syllable signal using the cepstral pitch detection method. The method has already been described in [10].
2) *Tone Feature Normalization:* The fundamental frequency contour f_0 is normalized to have the same fundamental frequency contour level (reference point) and a specific number of elements as required in the TRP. That is because male and female have different fundamental frequency contour level and also the number of element in tone feature vector is depended on the length of the syllable sound. In order to normalize fundamental frequency contour to have N values, the following steps are carried out.
   a) Block the fundamental frequency contour values into N+1 frames with 50% overlap.
   b) Compute the mean value of each frame. These values are the elements of a new fundamental frequency.
   c) Normalize the new fundamental frequency contour values to have the same reference point that is the first value. The following formula is used.

$$f_0(n) = \left(\frac{f_0(n)}{f_0(1)} \times r \right) - r, \quad n = 2,3,\ldots, N+1$$

where $f_0(n)$ is the n^{th} value of new fundamental frequency contour and r is any positive integer number.

   d) Remove the first value of normalized fundamental frequency contour out. The rest values are altogether a tone feature vector *tfea*.

4.2. Tone recognition process (TRP)

In this process, tone phoneme is recognized. A neural network is employed as recognition engine. A block diagram of this process is given in figure 6. A tone feature vector *tfea* from the LVET feature extraction process is processed. Finally, an index of recognized tone t is returned. In order to do tone recognition, the following steps are performed.

1) A neural network is configured and it is trained to classify all five tones.
2) For each unknown tone t, its tone feature vector *tfea* is used as the input of the neural network from a previous step.

We call the neural network used in TRP as Tone Neural Network (TONENN). The TONENN is a three layers feed forward neural network consisting of input, hidden, and output layer. There are requirements that input layer must have the same size as a tone feature vector *tfea*, the hidden layer must have a sufficient number of nodes for the network to perform tone classification and the output layer must have five nodes in the output layer corresponding to five tone phonemes. Each output node produces the value between 0 and 1. In this context, each output node value represents the probability for each tone phoneme class. The TONENN is supposed to be trained with a sufficient number of sample data. For each unknown tone to be recognized, its tone feature vector *tfea* is given to TONENN. After a tone feature vector *tfea* is passed to the TONENN, the network will generate five values as the number of output nodes. Each value implicitly represents a probability for one tone phoneme class. The index of the output node that gives the maximum value will be selected as the index of recognized tone t.

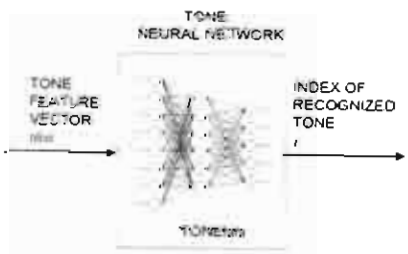


Figure 6: A Block Diagram of TRP.

4.3. Vowel recognition process (VRP)

In this process, vowel phonemes are recognized. 12 different vowel phonemes are called core-vowel in this process. We can recognize vowel by recognizing core-vowel and its length. A block diagram of this process is given in figure 7.

To recognize an unknown vowel, first its core-vowel feature vectors from LVET feature extraction process are processed by core-vowel recognition and an index of recognized core-vowel is returned. Next, its vowel location *vloc* is determined whether it is short or long vowel in vowel length determination step. Finally, these two results are used to determine an index of recognized vowel out of 24 vowel indices. The details of core-vowel recognition and vowel length determination are described as the followings.

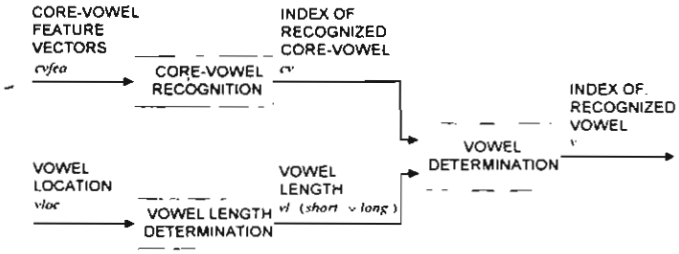


Figure 7: A Block Diagram of VRP.

Core-vowel recognition: A Hidden Markov Model (HMM) is used to represent each core-vowel class. Hence, 12 HMMs are required. A block diagram of core-vowel recognition is given in figure 8. The type of HMM used in this process is Continuous Density Hidden Markov Model (CDHMM). The details of CDHMM are described in [8][9].

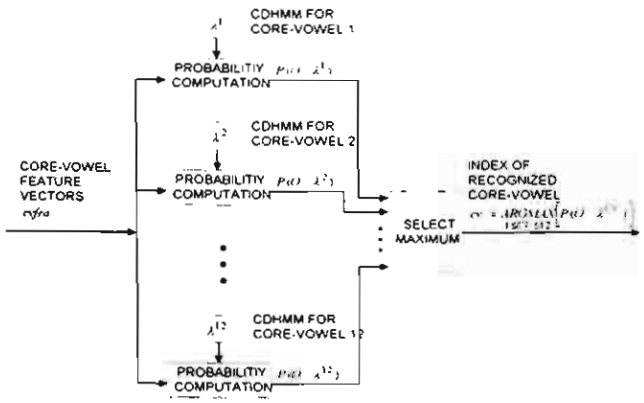


Figure 8: A Block Diagram of Core-Vowel Recognition

In order to do core-vowel recognition, the following steps are performed.

1) For each core-vowel class *cv*, a CDHMM λ^{cv} is built. And then the parameter set of each CDHMM is estimated in order to optimize the likelihood of the training set observation sequences.
2) For each unknown core-vowel, which is to be recognized, its core-vowel feature vectors *cvfea* are used as the observation sequence in the computation of model probabilities for all possible CDHMM models $P(O \mid \lambda^{cv})$. Finally, the index of CDHMM model, which has the highest probability, is selected as the index of recognized core-vowel *cv*.

Vowel length determination: The vowel location *vloc* contains the frame numbers of the starting and the ending points of vowel in a syllable signal. The vowel length is computed by subtracting the ending frame number with the starting frame number and then adding one to the result of subtraction because we want to include the starting and the ending frames. A simple threshold method is then used to determine whether the vowel is short or long. If the vowel length exceeds the threshold, it is long vowel. Otherwise, it is short vowel.

4.4. Leading consonant recognition process (LRP)

In this process, leading consonant feature vectors from LVET feature extraction process and an index of recognized core-vowel for VRP are processed. Finally, the index of recognized leading consonant is returned as the output. This means that the recognition of leading consonant is based on the recognition of core-vowel in VRP. A block diagram of LRP is given in figure 9.

For unknown leading consonant, which is to be recognized, the following steps are performed.

1) The index recognized core-vowel from VRP is used to select the leading consonant HMM bank.
2) Leading consonant feature vectors are used as the input of the selected LCHMM bank from the previous step. LCHMM bank processes the leading consonant feature vectors and returns the index of recognized leading consonant as its output.

The details of the LCHMM bank are described as the following.

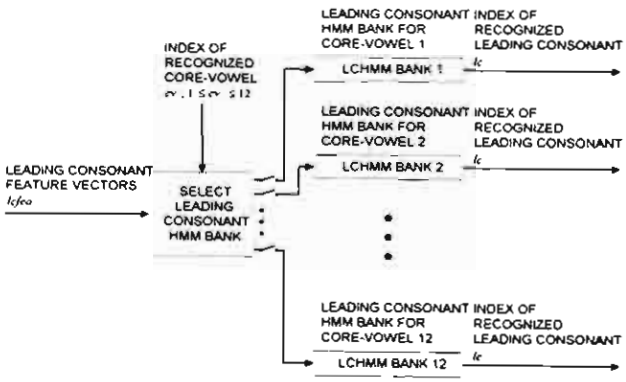


Figure 9: A Block Diagram of LRP.

LCHMM bank: Figure 10 shows a block diagram of a LCHMM bank. Each LCHMM bank is designed to cover all 38 leading consonant classes. Each LCHMM consists of up to 38 HMMs. This means that for each leading consonant class, there is a HMM corresponding to it. Each HMM in LCHMM bank is a CDHMM. In order to do leading consonant recognition, the following steps are performed.

1) For each leading consonant class *lc* of core-vowel *cv* class, a CDHMM λ_{cv}^{lc} is built. And then the parameter set of each HMM is estimated in order to optimize the likelihood of the training set observation sequences.
2) For each unknown leading consonant and known core-vowel *cv*, which is to be recognized, its leading consonant feature vectors *lcfea* are used as the observation sequence in the computation of model probabilities for all possible HMM models $P(O \mid \lambda_{cv}^{lc})$. Finally, the index of HMM model, which has the highest probability, is selected as the index of recognized leading consonant *lc*.

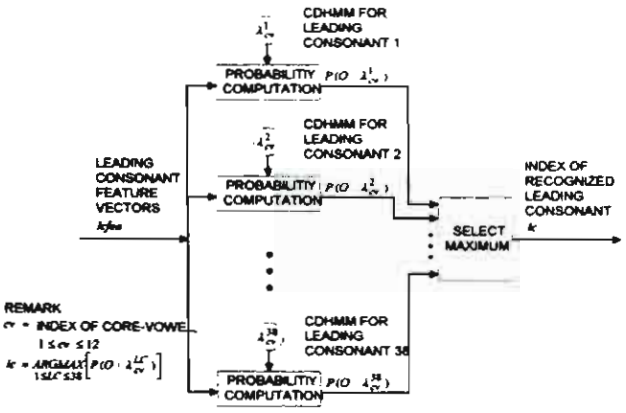


Figure 10: A Block Diagram of a LCHMM Bank.

4.5. Ending consonant recognition process (ERP)

In this process, the ending consonant feature vectors and an index of recognized core-vowel from VRP are processed. Finally, the index of recognized ending consonant is returned as the output. This means the recognition of ending consonant is based on the recognition of core-vowel in VRP. A block diagram of ERP is given in figure 11.

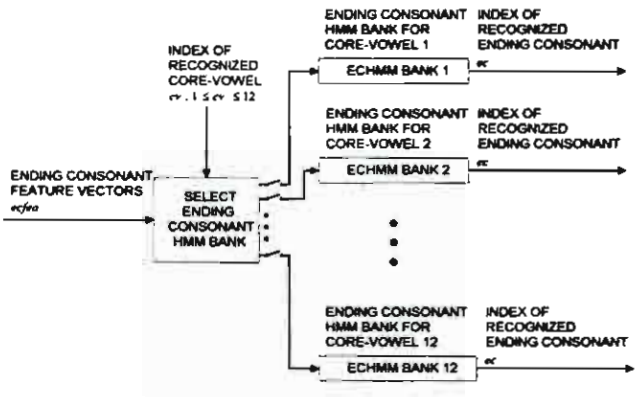


Figure 11: A Block Diagram of ERP.

For unknown ending consonant, which is to be recognized, the following steps are performed.

1) The index recognized core-vowel is used to select the ending consonant HMM (ECHMM) bank. There are 12 ECHMM banks as the number of core-vowels. Each ECHMM bank has the different number of ending consonant classes to be recognized. That is because not all ending consonant can occur with all core-vowels. The details of each ECHMM are described in the next section.
2) Ending consonant feature vectors are used as the input of the selected ECHMM bank from the previous step. The selected ECHMM bank processes the ending consonant feature vectors and returns the index of recognized ending consonant as its output.

ECHMM bank: each ECHMM bank consists of at most 9 HMMs because not all ending consonant can occur with all core-vowels. This means that for each ending consonant class, there is a HMM representing it. Each HMM in ECHMM bank is a CDHMM. A block diagram of an ECHMM bank is given in figure 12. In order to do ending consonant recognition, the following steps are performed.

1) For each ending consonant class *ec* having core-vowel class *cv*, a CDHMM λ_{cv}^{ec} is built. And then the parameter set of each HMM is estimated in order to optimize the likelihood of the training set observation sequences.
2) For each unknown ending consonant having core-vowel *cv*, which is to be recognized, its ending consonant feature vectors *ecfea* are used as the observation sequence in the computation of model probabilities for all possible HMM models $P(O \mid \lambda^{ec})$. Finally, the index of HMM model, which has the highest probability, is selected as the index of recognized ending consonant *ec*.

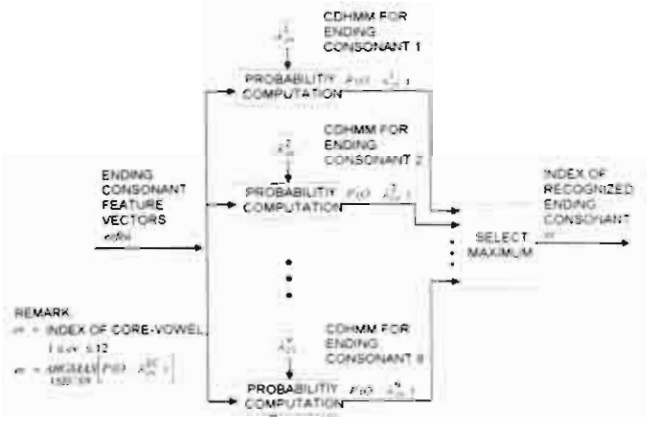


Figure 12: A Block Diagram of an ECHMM Bank.

5. Experimental results

There are three sets of syllables. They are all possible combinations of 38 leading consonants and 24 vowels used in LRP and VRP experiments, all possible combinations of 12 vowels and 9 ending consonants for ERP experiment, and randomly selected syllables comprises all five tone phonemes for TRP and VRP core-length determination experiments. All syllable sounds are gathered for two male and two female speakers. Each speaker spoke each syllable of all syllable sets 10 times. The syllable sounds were recorded using 16-bit quantization level, and sampling rate at 11.025 KHz. Four sets of experiments were conducted for TRP, VRP, LRP, and ERP respectively.

For TRP, there are two experiments conducted. The TONENNs for both experiments are configured to have 9 input nodes, 60 hidden nodes, and 5 output nodes according to 5 tone phonemes. The training algorithm is gradient descent with variable learning rate. Goal and maximum number of training epochs are set to 0.005 and 1000 respectively. The first TONENN is trained using 20% of data and the second one is trained using 50% of data. The recognition result of TRP is shown in table 1.

For VRP, two sets of experiments are conducted. The first set is for core-vowel recognition and the second set is for vowel length determination.

Two experiments are conducted for VRP core-vowel recognition, LRP, and ERP. Both experiment are almost the same except the number of training data. The number of training data for both experiments is 20% and 50% of data respectively.

For core-vowel recognition, LRP, and ERP experiments, the same parameters are used to extract the *cep* c features. They are 30ms frame size, 10ms frame rate, 14 LPC orders, and 12 cepstral coefficients orders. Each CDHMM used in these experiments is configured to have 3 states and 3 mixtures. Their experimental results are shown in table 2, 3, and 4 respectively.

For vowel length determination experiments, the data are divided into two sets: 20 % tuning set and 80% testing set. The vowel length threshold is determined from the tuning set. Finally, the experimental result of vowel length determination is shown in table 5.

Tone No.	Amount	Train 20%		Train 50%	
		Train (%)	Test (%)	Train (%)	Test (%)
1.[Mid]	1920	89.58	88.54	89.48	92.29
2.[Low]	1800	91.39	93.89	91.56	94.67
3.[Falling]	1120	88.84	87.28	85.89	90.00
4.[High]	1560	94.23	91.59	92.82	94.10
5.[Rising]	800	100.00	100.00	100.00	99.75
Total	7200	92.08	91.61	91.33	93.75

Table 1: TRP Experimental Result.

Vowel No.	Amount	Train 20%		Train 50%	
		Train (%)	Test (%)	Train (%)	Test (%)
1.[/i/,(I Ð]	1680	95.83	92.78	97.02	95.12
2.[/w/,(I Ð]	1680	92.56	87.20	83.81	83.33
3.[/u/,(I Ð]	1680	91.37	86.83	90.12	90.95
4.[/e/,(à Ð]	1680	95.83	92.49	95.95	96.31
5.[/3/,(à I Ð]	1680	94.05	88.32	91.31	90.48
6.[/o/,(à Ð]	1680	88.39	86.24	89.17	88.45
7.[/ae/,(à Ð]	1680	97.92	94.20	96.67	95.24
8.[/^/,(à Ð]	1680	95.83	91.89	94.29	94.52
9.[/a/,(I Ð]	1680	95.54	94.57	95.60	94.76
10.[/ia/,(à Ð]	1680	95.83	93.60	95.24	95.95
11.[/wa/,(à Ð]	1680	93.15	94.05	92.86	94.52
12.[/ua/,(I Ð]	1680	96.43	90.55	94.76	96.07
Total	20160	94.39	91.06	93.07	92.98

Table 2: VRP Core-Vowel Experimental Result.

Leading No.	Amount	Train 20%		Train 50%	
		Train (%)	Test (%)	Train (%)	Test (%)
1.[/ph/ ,(¾)]	960	99.48	70.31	91.88	88.96
2.[/th/ ,(·)]	960	98.44	73.57	91.04	93.13
3.[/kh/ ,(¤)]	960	98.96	77.60	95.83	90.63
4.[/p/ ,(»)]	960	98.96	57.68	91.67	86.88
5.[/t/ ,(µ)]	960	97.92	68.23	96.46	91.04
6.[/k/ ,(¡)]	960	98.96	64.06	94.17	90.00
7.[/?/ ,(¿)]	960	98.44	62.50	92.92	87.29
8.[/b/ ,(°)]	960	98.96	71.88	95.42	93.75
9.[/d/ ,(´)]	960	98.44	64.32	93.75	88.75
10.[/f/ ,(½)]	960	98.44	69.27	94.58	91.04
11.[/s/ ,(É)]	960	100.00	78.91	97.71	94.38
12.[/h/ ,(È)]	960	96.35	59.64	90.83	84.38
13.[/ch/ ,(°)]	960	100.00	85.81	97.08	96.67
14.[/c/ ,(´)]	960	98.44	79.30	96.88	95.42
15.[/m/ ,(Á)]	960	99.48	83.59	96.67	94.58
16.[/n/ ,(')]	960	97.40	77.99	96.04	92.71
17.[/nj/ ,(§)]	960	98.44	80.60	97.08	97.92
18.[/l/ ,(Á)]	960	99.48	73.57	94.58	91.04
19.[/j/ ,(Á)]	960	99.48	87.50	98.13	98.13
20.[/w/ ,(Q)]	960	96.35	79.56	95.83	90.42
21.[/r/ ,(Á)]	960	98.96	61.07	85.63	71.25
22.[/phl/,(¾Á)]	960	95.83	65.76	83.75	82.71
23.[/phr/,(¾Á)]	960	96.88	59.51	90.21	85.21
24.[/thr/,(· Á)]	960	97.40	69.92	91.46	87.29
25.[/khl/,(¤Á)]	960	95.31	65.10	90.83	89.17
26.[/khw/,(¤Q)]	960	98.44	77.08	93.33	90.21
27.[/khr/,(¤Á)]	960	97.92	67.58	89.17	86.88
28.[/pl/ ,(»Á)]	960	96.88	53.52	84.58	81.46
29.[/pr/ ,(»Á)]	960	94.79	52.08	84.79	77.33
30.[/tr/ ,(µÁ)]	960	97.92	52.47	85.83	80.00
31.[/kl/ ,(¡ Á)]	960	97.40	47.53	84.79	80.21
32.[/kw/ ,(¡ Q)]	960	97.92	74.48	93.75	92.71
33.[/kr/ ,(¡ Á)]	960	97.92	60.55	91.67	83.13
34.[/dr/ ,(´ Á)]	960	98.44	59.51	87.08	81.67
35.[/fl/ ,(¿ Á)]	960	96.88	66.15	86.25	90.21
36.[/fr/ ,(¿ Á)]	960	100.00	75.00	90.42	88.33
37.[/bl/ ,(° Á)]	960	98.44	64.19	86.25	86.04
38.[/br/ ,(° Á)]	960	94.79	54.56	86.04	77.29
Total	36480	98.01	68.21	91.69	88.00

Table 3: LRP Experimental Result.

Ending No.	Amount	Train 20%		Train 50%	
		Train (%)	Test (%)	Train (%)	Test (%)
1.[/p/,(°)]	480	100.00	79.69	99.58	95.83
2.[/t/,(')]	480	100.00	78.39	100.00	95.83
3.[/k/,(¡)]	480	100.00	69.53	98.75	95.00
4.[/?/,(Ð)]	480	100.00	83.07	100.00	97.08
5.[/m/,(A)]	480	100.00	96.35	99.58	97.92
6.[/n/,(')]	480	100.00	94.01	100.00	98.33
7.[/nj/,(§)]	480	100.00	85.16	100.00	97.92
8.[/j/,(A)]	280	100.00	97.77	100.00	100.00
9.[/w/,(Q)]	200	100.00	95.63	100.00	100.00
Total	3840	100.00	85.38	99.74	97.24

Table 4: ERP Experimental Result.

Vowel Length	Amount	Tuning 20%	
		Tuning(%)	Testing(%)
1.[Short]	2680	96.08	95.01
2.[Long]	4520	86.06	86.56
Total	7200	89.79	89.70

Table 5: Vowel Length Experimental Result.

6. Conclusion

This paper has presented the phoneme-based Thai syllable recognition system using Continuous Density Hidden Markov Model (CDHMM) and Neural Network (NN) techniques. The system consists of five processes. The first process is responsible for doing feature extraction. The rest four processes are responsible for doing phoneme recognition. As a syllable consists of four phonemes: leading consonant, vowel, ending consonant and tone. They are namely as leading consonant recognition process (LRP), vowel recognition process (VRP), ending consonant recognition process (ERP) and tone recognition process (TRP). Cepstral coefficients and signal energy frames extracted from a syllable signal are the base feature for LRP, VRP, and ERP. The vowel location is detected using the differences of cepstral coefficients frame and energy and then used to partition the base feature into three parts for LRP, VRP, and ERP respectively. The fundamental frequency contour extracted using cepstral pitch detection techniques is used as feature for TRP. The CDHMM technique is applied as the recognition engine in LRP, VRP, and ERP. The NN technique is applied as the recognition engine of TRP.

The best recognition rates of the leading consonant, core-vowels, and ending consonants phonemes are 88.00%, 92.98%, and 97.24% respectively. Note that in order to get these recognition rates, 50% of data are required as the training set of each CDHMM. The recognition rate of vowel length is 89.70%. From our experimental results, the correctness of syllable segmentation, the correctness of vowel location detection, the size of training set, and the quality of training set play the important role. For tone phonemes, the best average recognition rate is 93.75%. Note that 50% of data are required as the training set of NN. The NN used in TRP has 9 input nodes, 60 hidden nodes, and 5 output nodes according to 5 tone phonemes. The training algorithm is gradient descent with variable learning rate. Goal and maximum number of training epochs are set to 0.005 and 1000 respectively.

7. Future researches

In the future, there are three challenging works for us. The first one is to find the ways to improve the recognition results of our system. The higher recognition result can be achieved by adding more HMM models to each phoneme class. This causes increasing of the recognition time. Second, this system should be integrated with syllable segmentation system. Finally, the syllable-based word recognition system is the next challenging step to fulfill our conceptual Thai connected speech recognition.

8. References

[1] Ahkuputra, V., Jitapunkul, S., Maneenoi, E., Kasuriya, S., and Amornkul, P., *Comparison of Different Techniques On Thai Speech Recognition*, The 1998 IEEE Asia-Pacific Conference on Circuits and Systems, pp. 177-180 1998.

[2] Ahkuputra, V., Jitapunkul, S., Pornsukchandra, W., and Luksaneeyanawin, S., *A Speaker-Independent Thai Polysyllabic Word Recognition Using Hidden Markov Model*, Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, pp. 593-599, 1997.

[3] Jitapunkul, S., Luksaneeyanawin, S., Ahkuputra, V., Maneenoi, E., Kasuriya, S., and Amornkul, P., *Recent Advances of Thai Speech Recognition in Thailand*, The 1998 IEEE Asia-Pacific Conference on Circuits and Systems, pp. 173-176, 1998.

[4] Luksaneeyanawin, S., *Linguistics Research and Thai Speech Technology*, International Conference on Thai Studies, 1993.

[5] Maneenoi, E., Jitapunkul, S., Wutiwuwatchai, C., and Ahkuputra, V., *Modification of BP Algorithm for Thai Speech Recognition*, Proceeding of the 1997 International Symposium on Natural Language Processing, 1997.

[6] Pensiri, R. and Jitapunkul, S., *Speaker-Independent Thai Numerical Voice Recognition by using Dynamic Time Warping*, Proceedings of the 18th Electrical Engineering Conference, pp. 977-981, 1995.

[7] Pornsukchandra, W. and Jitapunkul, S., *Speaker-Independent Thai Numeral Speech Recognition using LPC and the Back*

Propagation Neural Network, Proceedings of the 19th Electrical Engineering Conference, pp. 977-981, 1996.

[8] Rabiner, L. R., *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceeding of IEEE, Vol. 77, No.2, pp. 257-286, February, 1989.

[9] Rabiner, L. R. and Juang, B. H., *Fundamentals of Speech Recognition*, A. Oppenheim, Series Editor, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[10] Rowden, C., *Speech Processing*, London: McGraw-Hill, 1992.

[11] Wutiwiwatchai, C., *Speaker Independent Thai Numeral Speech Recognition Using Neural Network and Fuzzy Technique*, Master's thesis, Chulalongkorn University, 1997.

Thai Syllabic Correction in Connected Thai Speech Recognition

Nunmanus Dachapratumvan
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok, Thailand
E-mail: g4219712@au.ac.th

Pratit Santiprabhob
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok, Thailand
E-mail: pratit@s-t.au.ac.th

Abstract: The basic syllable rules for Thai language can be defined by use a phoneme encoding method. Thai characters can be classified into the following sets: leading consonant, vowel symbols, Ending consonant and tone makers. Based on this character set and their appearances, the rule of syllabic boundary and ambiguous phonetic sound can be formulated to created Thai Ambiguous Phonetic Dictionary. Word Segmentation algorithm used for separating connected syllables in Thai Speech Recognition to be the possible word segmentation. Some of each word segment will be know or unknown in Thai meaning. Average Likelihood gives the correction unknown Thai words to be Thai word meaning system.
Key words: Thai Phonetic Rule Base, Thai Phoneme Ambiguous Dictionary, Word Segment Algorithm, and Average Likelihood Thai word correction.

1. Introduction

Thai syllabic correction has become one of the most essential things for integrating with the empirical result of Phoneme-base Thai speech recognition system [1]. From the empirical result of Phoneme-base Thai Speech recognition system [1] representation Thai phonetic: Leading Consonant, Vowel, Ending Consonant and Tone. All four parts represent to a syllable.

From the review of Thai Word decoding or Thai Word Segmentation approaches. Thai Word decoding [5,6,9,11] almost approaches use for text retrieved to find the unknown word meaning to word correction. Word Segmentation approaches [7,8,10] are important for separating syllables connected from Thai speech recognition to be each word segments. A detail survey of this technique almost found with Word Segmentation from Thai text sentence. Another approach is important for Thai word properties are Thai Rule-Base System [3,4] use for created Thai Regular Grammar to apply with Thai Syllable Speech Recognition error.

In this research, the proposed system presents a way to correction Thai syllabic connected speech recognition using Word Segmentation algorithm search in Thai Ambiguous Phonetic Dictionary and correction a word with Average Likelihood. This paper is organized as follows: In Section 2, the detail of Thai ambiguous Phonetic Word Model Section 3 and 4, the detail of propose system show the algorithm of Word Segment Algorithm and Average Likelihood Algorithm to find the correction word in Thai word system.

2. Proposed system

The propose system Thai Syllabic Correction by classifying them to correction with Thai Rule base system. The system consists of three processing Thai Ambiguous Phonetic Word Model, Word Segment Algorithm and Average Likelihood Algorithm.

A Thai syllable format can be divided to 4 parts. Those are leading consonant, vowel, ending consonant and Tone. A Thai syllable format using properties of each leading consonant, vowel and ending consonant. And also properties of Thai ambiguous sound modify with Phoneme-Base Thai Speech recognition system using Fuzzy system and Neural Network to create the probability of Thai Syllabic model.

2.1 Thai phoneme word dictionary

The format of Thai syllables properties and the properties of ambiguous phonetic sound can be generated to be Thai Phoneme Word Dictionary. The basic Thai Syllable Format Rules can define to be Thai Rule base Regular Grammar. The syllables building in Thai phonetic system consists of Leading Consonant, Vowel, and Ending Consonant. Thai Phoneme Words have to verify with Thai Rule base Regular Grammar. Thai Phoneme Word Dictionary keeps the words that can possible to occurred error phoneme from Phoneme-base Thai speech recognition system [1].

2.1.1 Modify leading, vowel marker properties

There are many approaches to derive vowel properties. Some approaches consider only the appearance of each vowel symbol. The following rules are used to determine their segmentation.

- The vowels, "-ะ /a/, - /ua/, โ -ะ /o/, เ -ะ /æ/, เ -าะ /ɔ/, เ -ะ /e/" always require at least one leading consonant and no final consonant follows.
- The vowel all form "เ -ะ /e/, เ -ะ /æ/, โ -ะ /o/, เ -าะ /ɔ/" always has a leading consonant.

- The vowels เ -/ee/, แ -/ææ/ always precedes consonants.

2.1.2 Modify Final Consonant Properties

The following letters are never used as final consonants ข /kh/,ต /t/,จ /ch/,ณ/n/,ฝ /f/,ผ /ph/

2.1.3 Rule-based Method

Although Thai grammar has many exceptions, the majority of syllable usage still follows these rules.

Characters	Meaning
|	Matches either the preceding or the following regular expression.
C, V	Consonant, Vowel

Table1: Character Meaning

Thai characters set symbols
C = {C_i:1<= i<= 44 set of Thai Leading Consonants}
C_f = { c_i | c_j = letters which consider as a Final Consonant, except /kh/,ต /t/,จ /ch/,ณ/n/,ฝ /f/,ผ /ph/}
V = { v_i|v_j = vowels which always place after consonant }

Table2: Thai Characters set symbols

In summary the basic regular expressions were compiled and emphasized to the consonant rules.

	Regular Expression	Vowel Patterns	Example
R1	C (V)(T)(Cf)		
R2	C (T)า(Cf)		
R3	C (T)ำ(Cf=ม)		
R4	C (T)ะ C ั (T)(Cf=า) C ะ(T) (Cf=น) C ัวะ	ะ ัว ั ัวะ	
R5	เ C (T)Cf เ C (T)ะ	เ เ-ะ	
R6	เ | C (T) (Cf=ย) เ | C (T) (Cf=ย)	เ เ	
R7	แC (V)(T)(Cf) แC (V)(T)ะ	แ แ-ะ	
R8	เ C (T)(Cf) เ C (T)า(Cf=า) เ C (T)ี(Cf) เ C (T)าะ เ C (T)ะ เC(V)(T)ียะ(Cf|ะ) เC(V)(T)ือะ (Cf|ะ)	เ เ-า เ-ี เ-าะ เ-ะ เ-ียะ เ-ือะ	
R9	C (T)าCf	า	

Table3: Thai Regular Expression

2.1.4 Thai ambiguous phonetic system

From the empirical result of syllable recognition of Phoneme Base speech recognition [1] can classified the groups of leading consonant ambiguous phonetic sound base on vowel phoneme. The results of Phoneme base speech recognition can analyst by percent of sound matching show in Table 4.

/ph/ (พ)	/th/ (ท)	/kh/ (ค)	/h/ (ห)	
	0.8421	0.1052	0.0526	0.0001
/th/ (ท)	/ph/ (พ)	/ch/ (ช)	/h/ (ห)	
	0.6176	0.2941	0.0882	0.0001
/kh/ (ค)	/k/ (ก)	/h/ (ห)	/ph/ (พ)	
	0.8235	0.0882	0.0882	0.0001
/p/ (ป)	/w/ (ว)	/l/ (ล)	/t/ (ต)	
	0.3637	0.3181	0.3181	0.0001
/t/ (ต)	/s/ (ส)	/r/ (ร)	/l/ (ล)	
	0.5	0.2812	0.2187	0.0001
/k/ (ก)	/kh/ (ค)	/th/ (ท)	/h/ (ห)	
	0.8285	0.0857	0.0857	0.0001
/?/ (อ)	/h/ (ห)	/nj/ (ญ)	/k/ (ก)	
	0.4285	0.3809	0.1905	0.0001
/b/ (บ)	/d/ (ด)	/w/ (ว)	/m/ (ม)	
	0.6486	0.2702	0.0811	0.0001
/d/ (ด)	/n/ (น)	/b/ (บ)	19.[/j/ .(ย)]	
	0.5	0.4117	0.0882	0.0001
/f/ (ฝ)	/s/ (ส)	/th/ (ท)	/ch/ (ช)	
	0.8055	0.1389	0.0555	0.0001
/s/ (ส)	/ch/ (ช)	/f/ (ฝ)	/c/ (จ)	
	0.3749	0.3125	0.3125	0.0001
/h/ (ห)	/th/ (ท)	/kh/ (ค)	/k/ (ก)	
	0.3448	0.3448	0.3103	0.0001
/ch/ (ช)	/c/ (จ)	/s/ (ส)	/kh/ (ค)	
	0.6857	0.1714	0.1428	0.0001
/c/ (จ)	/ch/ (ช)	/s/ (ส)	/th/ (ท)	
	0.5588	0.3529	0.0882	0.0001
/m/ (ม)	/nj/ (ญ)	/n/ (น)	/w/ (ว)	
	0.4594	0.4324	0.1081	0.0001
/n/ (น)	/d/ (ด)	/nj/ (ญ)	/m/ (ม)	
	0.4324	0.3243	0.2432	0.0001
/nj/ (ญ)	/n/ (น)	/m/ (ม)	/l/ (ล)	
	0.5526	0.3947	0.0526	0.0001
/l/ (ล)	/r/ (ร)	/s/ (ส)	/d/ (ด)	
	0.7586	0.1379	0.1034	0.0001
/j/ (ย)	/h/ (ห)	/l/ (ล)	/d/ (ด)	
	0.5	0.3461	0.1538	0.0001
/w/ (ว)	/r/ (ร)	/m/ (ม)	/b/ (บ)	
	0.5312	0.3125	0.1562	0.0001
/r/ (ร)	/l/ (ล)	/s/ (ส)	/w/ (ว)	
	0.4166	0.3333	0.25	0.0001

Table4: Thai Ambiguous Phoneme Matching Values with Vowel /i/ อิ and /ii/ อี

From the empirical result of syllable recognition process [1], the system has leading consonants training 100 times.

Let S = Time of second match
T = Time of third match
F = Time of fourth match

Percent of First match (1st) = 1.0

Percent of Second match (2nd) $= \frac{S}{\sum \{S, T, F\}}$

Percent of Third match (3rd) $= \frac{T}{\sum \{S, T, F\}}$

Percent of Forth match (4th) $= \frac{F}{\sum \{S, T, F\}}$

Percent of other match = 1-(%2nd+%3rd+%4th)

2.1.5 Create Thai ambiguous phonetic dictionary (AMPD)

From the result of Thai Ambiguous Phonetic System can generate the words into Thai ambiguous phonetic dictionary. The groups of word model have to separate to 4 groups.

- First Group for monosyllable
- Second group for two syllables
- Third group for three syllables
- Fourth group for four syllables

Example of word model in third group:

Sentence:	สวัสดี		
Phonetic:	สะ	หวัด	ดี

From this word, generate to

สะ =

/s/ (ส)	/c/ (จ)	/f/ (ฝ)	/ch/ (ช)	
	0.5151	0.303	0.1818	0.0001

หวัด =

/w/ (ว)	/h/ (ห)	/b/ (บ)	/k/ (ก)	
	0.4736	0.3684	0.1579	0.0001

ดี =

/d/ (ด)	/th/ (ท)	/b/ (บ)	/j/ (ย)	
	0.5	0.4117	0.0882	0.0001

The word model of สวัสดี is metric of [n, 4], where n is number of syllables.

	1	2	3
1	สะ 1	หวัด 1	ดี 1
2	จะ 0.5151	หัด 0.4736	นี 0.5
3	ฝะ 0.303	บัด 0.3684	บี 0.4117
4	ชะ 0.1818	กัด 0.1579	ยี 0.0882

Table 5: Word Model Matrix

3. Word Segmentation Algorithm

Word segmentation algorithm use for separating words from connected speech in a sentence. Word segmentation algorithm is searching by the group of word model. Let

N = # of word model.
T = a given sentence to be word segmented.
Tij = a word of T starting at first position of word to end position of word.
PD = a word model in Thai Ambiguous Phonetic Dictionary (AMPD)
G = syllable group (length of syllables)

The algorithm consists of three steps as follows:

1. if G not in 1st group ,
   For G = 4 to 2, searching from four syllables to two syllables
   For each Ti, i=1,...,n, find a word, in G group of PD satisfying the following conditions:
   - Si syllables matches found. Let i' = i-1+length of Si. Therefore Wi=Si,i'
   - Wi is not a word in PD. Go to next Ti

   If G equal 1 syllable, separate all syllables segment by don't search in PD.

For example,

1	2	3	4	5	6	7
สะ	บัด	ดี	นัก	เรียน	คอน	เช้า

Phonetic speech recognition error:

1	2	3	4	5	6	7
ชะ	บัด	ดี	นัก	เรียน	คอน	เช้า

Generate Word Segment to:

N	G	Word	Wi = Si,i'	AWM
1	3	สะ-บัด-ดี	S1,3	0.1567
2	2	นัก-เรียน	S4,5	1
3	2	คอน-เช้า	S6,7	0.2948
4	1	สะ	S1,1	0.1818
5	1	บัด	S2,2	0.3684
6	1	ดี	S3,3	1
7	1	นัก	S4,4	1
8	1	เรียน	S5,5	1
9	1	คอน	S6,6	0.1785
10	1	เช้า	S7,7	1

Let

WM = value of each syllable from word model matrixes (table 4)
AWM = Average Word Model

$$AWM = \frac{\sum WM_{i,j}}{G}$$

2. From the example in step 1, we can construct the corresponding overlapping graph as shown in Figure1.

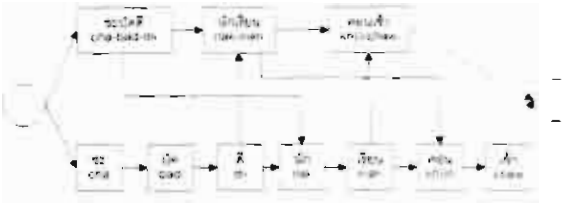


Figure1: Word Segment Overlapping Graph

3. For each component of the graph, find the average value for all paths in graph.

No.	Word Segmented	NWS	AWS
1	จะ|บัด|ดี | นักเรียน |คอน|เช้า	3	0.701983333
2	จะ|บัด|ดี | นักเรียน |คอน | เช้า	4	0.6292
3	จะ|บัด|ดี | นัก | เรียน |คอน|เช้า	4	0.7764875
4	จะ|บัด|ดี | นัก | เรียน |คอน | เช้า	5	0.70336
5	จะ| บัด| ดี | นัก|เรียน |คอน|เช้า	5	0.51789
6	จะ| บัด| ดี | นักเรียน |คอน | เช้า	6	0.50005
7	จะ| บัด| ดี | นัก | เรียน |คอน|เช้า	6	0.598241667
8	จะ| บัด| ดี | นัก | เรียน |คอน |เช้า	7	0.571471429

Table6: Average values of Word Segment

Let NWS = Number of word segment

$$\textit{Average Word Model (AWS)} = \frac{\sum_{1..NWS} AWM}{NWS}$$

4. Averages likelihood

From the word segmentation algorithm can obtained a set of words that know and unknown meaning. So Average Likelihood give the solution for correction unknown word meaning to be the real word in Thai meaning system.

if G not in 1st group
  for each Word Segmented (Table6) in path
    if AWM = 1
      - Give word correction value (WCV) to be 1

    Else
      For i=1 to 4
        - Change wrong syllable to be syllable of [1,i] in word model matrixes
        - Get the correction syllable
        - Give word correction value (WCV) to be 1

      Next

  Next

Else

  If AWM =1,
    - Give word correction value (WCV) to be 1

  Else
    For i=1 to 4
      if [i,1] = syllable value of [1,1] with a syllable word model matrixes
        - Give word correction value (WCV) to be AWM of [i,1]
        - Get the correction syllable

      Else not exist in [i,1]
        - Give the word correction value (WCV) to be 0.0001 for unknown word meaning

    Next

$$\textit{Average Word Correction (AWC)} = \frac{\sum WCV}{NWS}$$

$$\textit{Average Max Likelihood (AML)} = Max(AWC * AWS)$$

So AML gives a Maximum value of Word Segmentation path that the syllables have to correction unknown word to be the know word in Thai meaning system.

5. Experimental Result

The experimental result is mean by using word segmentation accuracy rate (WSAR), Thai word correction accuracy rate (WCAR) and correction rate (CR) is shown the value of performance system.

The initial experiment is base on the following condition:

- 500 Thai word model in Thai word meaning system
- 1,500 ((500*4)-500) for Thai Phoneme word model
- 4 maximum syllables for each word
- 4 groups of word model
- 9 rules of Thai Rule Base System.

The experiment was conducted according to the following steps:

1. Maximum 15 Syllables for each sentence.
2. Thai Ambiguous Phonetic Dictionary was setup the word matrixes model in the discussion of AMPD in section 2.1.5.
3. The Word Segment Algorithm was applied for separating all syllables to be segment of words and created likelihood word segment overlap graph.
4. User Average Likelihood Algorithm to correction unknown Thai words meaning to be know Thai word meaning.

5. The experimental result is shown in Table 7.

#of Syllables	#of Thai Word	# of Word Segment	WSAR(%)	WCAR(%)	CR(%)
15	9	8	88.89	60	100
10	10	10	100	80	90
10	5	4	80	60	100
6	2	2	100	50	100
6	3	2	66.67	66.7	100
3	1	2	50	33.34	100

Table7: The Experimental Result

6. Conclusion

The proposed system is an attempt to provide Thai Syllabic Connected to integrated with the empirical result of syllable recognition process error.

The system was corrected unknown Thai word to be Thai word meaning by Thai Ambiguous Phonetic word model and get the maximum value of Average Likelihood to separated the connected Thai syllables to be segments of a word.

The correction rate is dependent upon the process of determining appropriate sample cycles, which plays the most important role in this system. Applying Thai Rule-Base System with Thai Ambiguous Phonetic System to create the matrixes of word model, can give the knowledge base of unknown word model in each matrixes. Word Segment algorithm and Average Likelihood is also important to cutoff the connected syllable in Thai speech recognition to be segment of Thai word meaning.

All the processes are being tested with data representing subtle features of Thai word. Once the system is accomplished, the overall accuracy of approximately 90% is expected.

7. References

[1] Cheirsilip Ronnarit, Santiprabhob Pratit, *Phoneme-based Thai Speech Recognition System Using Fuzzy System and Neural Network*, IC-AI'2000, July 2000.

[2] Chalireerat Jirawat, Santiprabhob Pratit, *Thai Syllable Segmentation For Connected Speech With Fuzzy System*, IC-AI'2000, July 2000.

[3] Jaruskulchai Chuleerat, *An Automatic Indexing For Thai Text Retrieval*, The School of Engineering and Applied Science, George Washington University, July, 1998.

[4] "อภิษา ศาลา, ลลิตา นฤปิยะกุล บุญเจริญ ศิริเนาวกุล", *การออกเสียงคำหลายพยางค์ในภาษาไทยในคอมพิวเตอร์*, วารสารวิจัยและพัฒนา มจธ ปีที่23 ฉบับที่ 1 มกราคม-เมษายน 2543

[5] Zheng Fang, Wu Jian, Song Zhanjiang, *The Syllable-Synchronous Network Search Algorithm For Word Decoding In Continuous Chinese Speech Recognition*, IEEE ICASSP, Volume(2), 601-604, September, 1999.

[6] Zheng Fang, Wu Jian, Song Zhanjiang, *Improving The Syllable-Synchronous Network Search Algorithm For Word Decoding In Continuous Chinese Speech Recognition*, J. Computer Science & Technology, 15(5), 461-471, September, 2000.

[7] Witoon Kanlayanawat and Somchai Prasitjutrakul, *Automatic indexing for Thai text with unknown words using trie structure*, In Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'97), pages 115--120, Phuket, Thailand, 1997.

[8] S. Charnyapornpong, *A Thai Syllable Separation Algorithm*, M.Eng Thesis, Asian Institute of Technology, Aug. 1983.

[9] A. Kawtrakul, C. thumkonon, and S. Seriburi, *A Statistical Approach to Thai Word Filtering*, Proc. Of the second Symposium on Natural Language Processing, pp. 398-406, 1995.

[10] Y. Poovoranwan and V. Imarom, *Dictionary-base Thai Syllable Segmentaion (in Thai)*, 9th Electrical Engineering Conference, 1986.

[11] Karoonboonyanan, T., Sornlertlamvanich, V. and Meknavin, S., *A Thai Soundex System for Spelling Correction*, Proceeding of the National Language Processing Pacific Rim Symposium 1997, 1997, pp. 633-636.

Thai Word Decoder Based on Genetic Algorithm

Wanna Supasirirojana
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok, Thailand
E-mail: g4219711@au.ac.th

Pratit Santiprabhob
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok, Thailand
E-mail: pratit@s-t.au.ac.th

Abstract: Word decoding plays an important role in connected speech recognition. There are many decoding techniques based on a language model. This paper proposes to use an alternative technique that does not require a language model. The authors employ genetic algorithm as decoding technique to identify the most likely word sequence from the Thai syllable sequence input. The fitness function used in this genetic algorithm is based on the Thai language rule on ambiguous phonetic and empirical result of syllable recognition process.
Key words: Thai Word, Decoding, Genetic Algorithm

1. Introduction

In the field of speech recognition, the decoder is a core of recognition and it is time consuming process. There are many techniques of decoding algorithm develop in the connected speech recognition system.

From the review of decoding technique, many researchers used language model as a knowledge resource input into the decoder. Syllable network and word search tree are used in [2,.3], N-Gram base language model is use in [1,2,3]. The viterbi-decoding algorithm [1,2,3] and network search algorithm [1,2,3] are used as a search algorithm.

This paper proposes a new framework of decoding technique base on genetic algorithm [5] that does not require a language model. The propose decoding technique finds the most likely word sequence from an input syllable sequence that is a result of syllable recognition process. The process generates a set of candidate word sequence to be a solution by picks up a sequence of random word that satisfies the initial population condition. A selection and crossover operation are applied to create a new candidate word sequence in a next generation. The evaluation function for the candidate is fitness function. Fitness function is calculated as by using matrix base on ambiguous degree and Thai language rule.

2. Thai word domain

The proposed algorithm in this research is applied within the boundary of Thai word domain. Word domain composes of words in Thai dictionary and word pronunciation. Each word has 4 maximum syllables. An ID is assigned to each word in order to make it easy to reference form decoding algorithm. An example of Thai words is shown in table 1.

Thai syllable composes of 4 phonemes:

- Leading (L): there are 38 different Thai leading consonants
- Vowel (V): there are 24 different Thai vowel
- Ending (E): there are 9 different Thai ending consonants
- Tone (T): there are 5 different Thai tone

Id	Word	Pronounce (LVET)
7	กำลัง (kam -la ŋ)	6 9 5 1; 18 9 7 1
8	กำลังใจ (kam -la ŋ-cai)	6 9 5 1; 18 9 7 1; 14 9 8 1
11	เก็บ (kep)	6 4 1 2
61	ตลอด (ta-l)t)	5 9 4 2; 18 20 2 2
62	ตลาด (ta-la:t)	5 9 4 2;18 21 2 2
158	เศรษฐกิจ (se:t-tha-kit)	11 16 2 2;2 9 4 2;6 1 2 2
163	สนับสนุน (sa-nap-sa-nun)	11 9 4 2;16 9 1 2;11 9 4 2;16 3 6 5
187	เหตุผล (he:t-phon)	12 16 2 2;1 6 6 5

Table 1: Example of Thai Word Domain.

3. Ambiguous matrix

Thai Ambiguous matrix is base on the Thai language rule on ambiguous phonetic and empirical result of syllable recognition process [6]. There are 4 groups of Thai ambiguous phonetic: ambiguous leading (ALM), Ambiguous vowel (AVM), ambiguous ending (AEM) and ambiguous tone (ATM)

In this research, there are 16 ambiguous leading matrix based on a couple of vowels and the other 3 matrixes for ambiguous vowel, ambiguous ending and ambiguous tone.

	/th/ (ท)	/kh/ (ค)	/k/ (ก)	/h/ (ห)
/ph/ (พ)	32	5	1	2

Table 2: Value of Ambiguous Leading /ph/(พ).

	/ph/ (พ)	/kh/ (ค)	/k/ (ก)	/h/ (ห)	/ch/ (ช)
/th/ (ท)	25	5	2	3	10

Table 3: Value of Ambiguous Leading /th/ (ท).

The partial results of ambiguous leading /ph/ (พ) and /th/ (ท) with vowel /i/ (อี) and /i:/ (อี) from syllable recognition process [6] with 40 number of test set are shown in table 2 and table 3 respectably. The row in the table is value of ambiguous leading corresponding to the leading in the leftmost column. From talble 2, there are 40 number of test set of recognition the leading "/ph/ (พ)" 32 of 40 recognize as /th/ (ท) and 4 of 40 recognize as /p/ (ป).

Base on the empirical result of ambiguous leading consonant in table 2 and 3, the ambiguous degree (AD) of each phoneme is calculated by using equation (1).

$$AD = \frac{Value\ of\ ambiguous\ phoneme}{total\ number\ of\ test\ set} \qquad (1)$$

From the equation (1), AD 1 is set to 1 for an ambiguous degree of phoneme itself.

The ambiguous degree is a value between 0 and 1. The partial ambiguous matrix for leading consonant /ph/ (พ) and /th/ (ท) with vowel /i/ (อี) and /i:/ (อี) is shown in table 4.

	/ph/(พ)	/th/(ท)	/kh/ (ค)	/k/(ก)	/h/(ห)	/ch/(ช)
/ph/(พ)	1	0.8	0.125	0.025	0.05	0
/th/(ท)	0.625	1	0.125	0.05	0.075	0.25

Table 4: The partial Ambiguous Matrix for Leading Consonant.

4. Word decoding

The authors propose the decoding technique based on genetic algorithm (GA) to find the most likely Thai word sequence given the Thai syllable sequence. Thai syllable sequence from the syllable recognition process is a string of syllable consists of phonemes number. There are many characteristics of GA used in this research.

4.1. Decoding process

The decoding process is started with a set of word sequences to be solution by initial population (refer to sub-section 4.3). The number of generation and acceptable fitness value are set as a condition to stop a decoding process. The fitness value (refer to sub-section 4.4) is calculated for each word sequence in the population. If fitness value of the current word sequence is not good to be a solution then the new generations of word sequences is generated from selecting two parents (refer to sub-section 4.5) and apply crossover operation (refer to sub-section 4.6).

There are 2 steps of stopping decoding process in this research.

First step is assumed that the input syllable from the syllable recognition process has a 100% recognition rate. The acceptable fitness value is set to 1. The fitness function is using the normal degree as a degree of fitness. The process is stopped when there is a fitness value of word sequence (result) equal to the acceptable fitness value. If more than 80% of word sequence in the current population has the same fitness value then assume that the input syllable from the syllable recognition process has a recognition rate less than 100% so the decoding process is restart and second step is applied

For second step, the fitness function is using the ambiguous degree as a degree of fitness. Second step is stopped when more than 80% of word sequence in the current generation has the same fitness value. Word sequence that has a maximum fitness value is picking up as a result.

4.2. Encoding of word sequence

The authors used permutation encoding as a encoding method to represent each word sequence by a string of word ID. A variable-length encoding scheme is applied because the number of word in word sequence is varying according to the number of syllable of input syllable. An example of encoding of word sequence is shown in figure 1.

Word sequence A:
(lu:k-ca:ŋ mi: he:t-phon thi: maj hen dua:j kap na:j-ca:ŋ)

144	44	128	187	78	130	188	54	4	90

Word sequence B:
(ro:ŋ-ria:n pha-lit nak-ria:n thi: mi: khun-na-phap)

144	44	128	187	78	130	188	54	4	90

Figure 1: Encoding of Word Sequence.

4.3. Initial population

Initial population generates a set of candidate word sequence. Word sequence is created by pick up a random word that satisfy the initial population condition and add that word to the end of current word sequence then pick up the next word and repeat random process until the length of syllable in word sequence is equal to the length of input syllable.

The initial population conditions are

- The random word has either vowel or tone that equal to vowel or tone on the same position of input syllable.
- Length of the word sequence must be equal to the length of input syllable.

4.4. Fitness function

$$FN = \sum_{i=1}^{w}\left(\sum_{j=1}^{n} \text{aDegree}_j\right) \Big/ NS \qquad (2)$$

where w = # of word in chromosome
n = # of syllable in each word
NS = # of syllable in chromosome

$$\text{aDegree}_j = \begin{cases} \text{ambiguous degree} \\ \text{normal degree} \end{cases}$$

The fitness function is calculated by using equation (2).

From the equation (2), the ambiguous degree of each phoneme can be obtained from ambiguous matrix as describe in section 3. The normal degree is the set to 1 if syllable at j is equal to syllable at the same position of input syllable, otherwise it is set to 0. The ambiguous degree is used to measure the similarity of the input syllable and candidate word sequence.

4.5. Selection and crossover point

Two word sequences (father, mother) are random selected from the population by using roulette wheel as a selection method. The random mother is not accepted as a new mother until it is verified as a good mother.

Mother is verified by choosing the crossover points of mother and father that preserve the syllable length of the new word sequences (offspring) as mention in the initial population condition and apply crossover operation (refer to sub-section 4.6) to produce the new word sequences. The fitness function is applied to new word sequences. The mother is good if any of the new word sequence has a better fitness value compare with the parent. A flowchart of selection is shown in figure 2.

4.6. Crossover

Swapping words between two word sequences at the crossover points performs the crossover operation. New word sequences from crossover operation are applied with fitness function. The new word sequence with the better fitness value is put in a new population. An example of crossover is shown in figure 3.

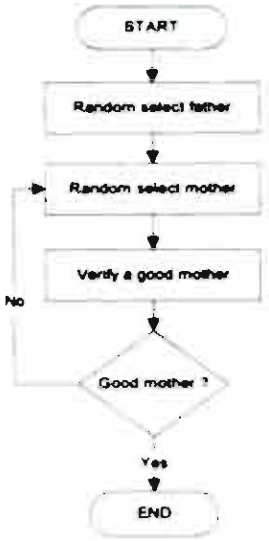


Figure 2: Selection Flowchart.

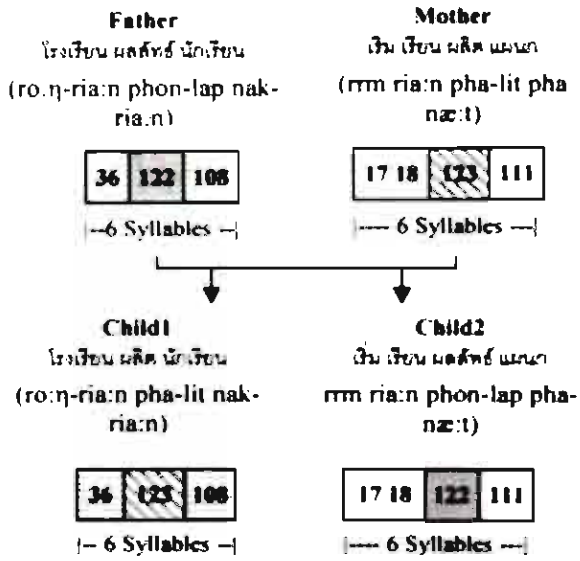


Figure 3: An Example of Crossover Operation.

5. Experimental result

The experimental result is measure by using decoding accuracy rate (DAR) and correction rate (CR). The number of word that the system decodes correctly measures DAR. Correction rate (CR) is number of word that the system makes it correct from the error of input.

The initial experiment is based on the following condition:

- 220 Thai words in Thai word domain.

- 4 maximum syllables for each word.
- 4 group of ambiguous matrix base on [6].

The authors perform 2 experiments from the above conditions.

First experiment is using 200 input syllables with 100% recognition rate from syllable recognition process [6]. The example graph display fitness value for each generation is shown in figure 4. Table 5 is shown the DAR of the experiment.

Initial Condition	DAR
400 initial population size, 12 generations.	100%
200 initial population size, 12 generations.	97%
200 initial population size, 5 generations.	80%

Table 5: Accuracy Rate of Experiment 1.

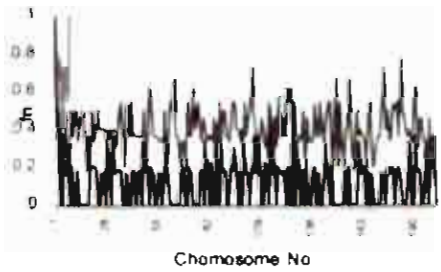


Figure 4: Fitness Value of Experiment 1.

Second experiment is using the different 200 input syllable sequences with a recognition rate from syllable recognition process [6] less than 100%. From the experiment, 85% of word sequence in the 6th-generation has the same fitness value as shown in figure 5. The system recognizes that there is some error with input syllable so the second step of decoding process is used. DAR is 95% and CR is 85%.

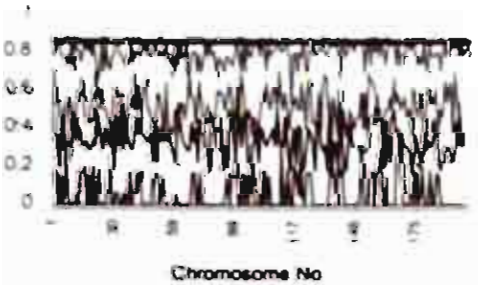


Figure 5: Fitness Value of Experiment 2.

6. Conclusion

This paper proposed the alternative decoding technique that does not require a language model by using empirical result of syllable recognition process.

From the decoding technique being test, the population size, number of generation and recognition rate of the input syllable are factors that effect to the accuracy rate. The population size should be set according to the number of word in word domain. A small number of generations cause the low accuracy rate because in some case the solution is found in the next generation from the maximum generation. The accuracy rate is decrease when an input syllable from a syllable recognition process has recognition rate less than 75% per syllable.

The decoding technique used in this paper is a time-consuming process. Further research should be focus on this problem by applied the mutation operator and considering the condition of selecting a candidate solution in the initial population.

7. References

[1] Jie, Zhao, *Network and N-Gram Decoding in speech recognition*, Mississippi State University, October, 2000.

[2] Zheng Fang, Wu Jian, Song Zhanjiang, *The Syllable-Synchronous Network Search Algorithm For Word Decoding In Continuous Chinese Speech Recognition*, IEEE ICASSP, Volume(2), 601-604, September, 1999.

[3] Zheng Fang, Wu Jian, Song Zhanjiang, *Improving The Syllable-Synchronous Network Search Algorithm For Word Decoding In Continuous Chinese Speech Recognition*, J. Computer Science & Technology, 15(5), 461-471, September, 2000.

[4] Neeraj Deshmukh, *Decoder Strategies*, *Institute for Signal and Information Processing*, Mississippi State University, 1997.

[5] Lawence Davis, *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York , 1991

[6] Cheirsilip Ronnarit, Santiprabhob Pratit, *Phoneme-based Thai Speech Recognition System Using Fuzzy System and Neural Network*, IC-AI'2000, July, 2000.

[7] K T Lua , *Part of Speech Tagging of Chinese Sentences Using Genetic Algorithm*, Chinese Computing Conference, 45-49,June, 1996.

[8] Hugo M. Ayala, *Natural Language Generation Using Genetic Algorithms*, MIT Department of Mechanical Engineering,1995

A Framework for Connected Speech Recognition for Thai Language*

Pratit Santiprabhob[1], Jirawat Chaiareerat[2], Ronnarit Cheirsilp[3],
Nunmanus Dachapratumvan[4], and Wanna Supasirirojana[5]
Intelligent System Laboratory
Department of Computer Science
Faculty of Science and Technology
Assumption University
Bangkok, 10240,Thailand

Email: pratit[1], jirawat[2], ronnarit[3] @s-t.au.ac.th, g4219712[4], g4219711[5] @au.ac.th

Abstract: Connected speech recognition problem for Thai language, like the similar problem in other languages, involves three sub-problems: 1) syllable segmentation, 2) syllable recognition and 3) syllable-based word recognition. This paper presents a framework upon which a speech recognition system can be built. The approach taken in our framework differs from a so-called word-based approach in which whole words are trained to be later recognized. Our approach attempts to recognize syllables based on their constituent phonemes; the recognized syllables are then grouped into words within a given context of discourse. The four constituent phonemes of Thai syllables are leading consonant, vowel, ending consonant and tone. The proposed framework utilizes several soft computing techniques in different parts. As for the signal processing portion of the framework, Fuzzy System (FS) is used in the syllable segmentation part while the Neural Network (NN) and Hidden Markov Model (HMM) are used in the syllable recognition part. On the other hand, Genetic Algorithm (GA) and rule-based system techniques are used to develop alternative methods to recognizing words from given set of syllables

Keywords: Speech Recognition, Hidden Markov Model, Neural Network, Fuzzy System, Genetic Algorithm, Rule-Based System

1. Introduction

Speech is a primary means of human communications. It is the most natural way for humans to convey ideas, to exchange information, to give instruction, etc. A speech is an intelligible group of words. Thus, the foundation for the understanding of human speech is the understanding of spoken words which in turn requires the recognition of spoken words to first be achieved. Our proposed framework outlines methods that can be used to solve this spoken words (or speech) recognition problem. This is indeed an exciting yet challenging research area. Speech is seen as the way humans will interact with computers in the future. In general, humans can speak about two times faster than a proficient typist can type. In addition, this mode of man-machine interaction allows for hand-free operation such as giving on-board computer an instruction while driving a car.

* This research is supported in part by the Thailand Research Fund

Techniques for recognizing words as trained are widely commercially available. These words are not connected, individual words that can be encoded as templates. On the other hand, recognizing connected speech is a totally different problem with a magnitude of difficulty. Our proposed framework is conceptually depicted in Figure 1. In the first step, the given speech is segmented into syllables. Then, in the second step, each syllable is attempted a recognition from its constituent phonemes. Eventually, in the third step, the recognized syllables are decoded into words within a given context of discourse.

Various researchers have developed different alternatives to the problem of Thai speech recognition. Different techniques are used such as Dynamic Time Wrapping [1], Conventional Neural Network [2], Modified Back Propagation Neural Network [3], Neural Network with Fuzzy MF Preprocessor [4] and Hidden Markov Model [5]. From the studies in [6] and [7], the Hidden Markov Model (HMM) as used in [5] is identified as the technique that yields the best recognition rate.

However, there are a number of limitations observed with regard to the research works cited.

1) All of the research works utilizes the word-based speech recognition approach. The whole words are trained/encoded. Hence, the approaches can recognize only a small set of vocabularies such as numbers, names and commands.
2) The approach outlined in [5] is not readily applicable to the connected speech recognition problem in general since the number of syllables has to be determined before the recognition can be undertaken.
3) In all of the approaches, computational requirement grows in proportion to the number of vocabularies they are trained/encoded to recognize.

In order to overcome the limitations discussed above, our proposed framework attempts to recognize connected speech in terms of syllabic units. This requires that words in a given connected speech be segmented into syllables before recognition can be achieved.

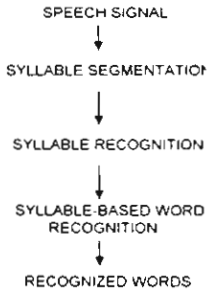


Figure 1: Conceptual Framework for Connected Speech Recognition

2. Framework Architecture

The proposed framework consists of three parts: firstly, syllable segmentation which is described in Section 2.1, secondly, syllable recognition which is outlined in Section 2.2, and thirdly, syllable-based word recognition whose two alternatives are discussed in Section 2.3.

2.1. Syllable Segmentation

The segmentation algorithm used in our framework is based on the concepts of energy and Different Ceptral as explained in [8]. The segmentation algorithm consists of three steps: parameters computation, threshold based segmentation and fuzzy based segmentation. First, the speech signal is pre-processed to enhance the signal quality. Then, necessary parameters are calculated. These parameters are used in segmenting the speech signal. Finally, a fuzzy inference system is used to identify the ending point and starting point of each syllable in each resulting segment.

2.1.1. Parameters Computation

First, the speech signal is pre-processed by means of signal pre-emphasizing technique as described in [9]. The signal is then en-framed into 30 milliseconds long frames with 20 millisecond overlapping factor between frames. For each frame, fours parameters are computed: High Amplitude Rate (HAR), Absolute Energy, Zero Crossing Rate (ZCR), and Different Ceptral (DC). Detailed descriptions of these parameters can be found in [8]. A graph representing each of the four parameters is respectively constructed. Finally, the contours of each graph are then smoothed according to the Moving Average Smoothing algorithm [10].

2.1.2. Threshold based Segmentation

In this step, the threshold-based segmentation algorithm eliminates the silent portions of a given speech using a set of threshold values calculated from the beginning part of the speech. Here, the original speech signal is segmented into groups of syllables called speech segments.

The algorithm works as follows. The speech signal is searched from the first frame to find the pairs of starting frames and ending frames. The following rules are then applied to determine whether a frame *I* is starting frame or ending frame or neither.

If Energy[i]>E_th1 or HAR[i]>HAR_th1 then frame I is staring frame.
If Energy[i]<E_th2 or ZCR[i]=0 or HAR[i]<HAR_th2 then frame I is ending frame.

Where:
Energy[i] is the ABS Energy at frame i
HAR[i] is the HAR at frame i
ZCR is the ZCR at frame i
E_th1 and E_th2 are Energy thresholds calculated from the background noise at the beginning of the speech signal.
HAR_th1 and HAR_th2 are HAR thresholds calculated from the background noise at the beginning of the speech signal.

The algorithm is depicted in Figure 2. The results obtained in this step are the speech segments, which will further be segmented into syllables in next step.

2.1.3. Fuzzy based Segmentation

Each speech segments resulted from the threshold-based segmentation algorithm is once again segmented in this step. The ultimate results are syllables to be recognized. There are four steps in segmentation. First, local peak energy frames, so-called PeakE frames are identified in each speech segment. Then, local minimum energy frames

between two PeakE Frames, so-called Emin frames are also identified. The identification rules for these frames are given in [8].

For each Emin frame, five fuzzy input variables are defined, namely 1) the absolute energy of the current Emin frame – EM, 2) the minimum ZCR between the two surrounding PeakE frames, 3) the difference between the EM and the absolute energy of the preceding PeakE frame – DEL, 4) the difference between the EM and the absolute energy of the following PeakE frame – DER, and .5) the maximum DC between the two surrounding PeakE frames – DCMAX. Finally, a Fuzzy Inference System (FIS) is constructed to determine the frame whether it is a boundary frame or not based on these five input variables. Fuzzy terms for each of the parameters and the fuzzy rules are defined in [8]. Here, Mamdani-type FIS with centroid defuzzification method [11] is employed.

After the boundary frames are located, the center speech signal sample of each frame is used as a boundary point to demarcate the boundary between syllables.

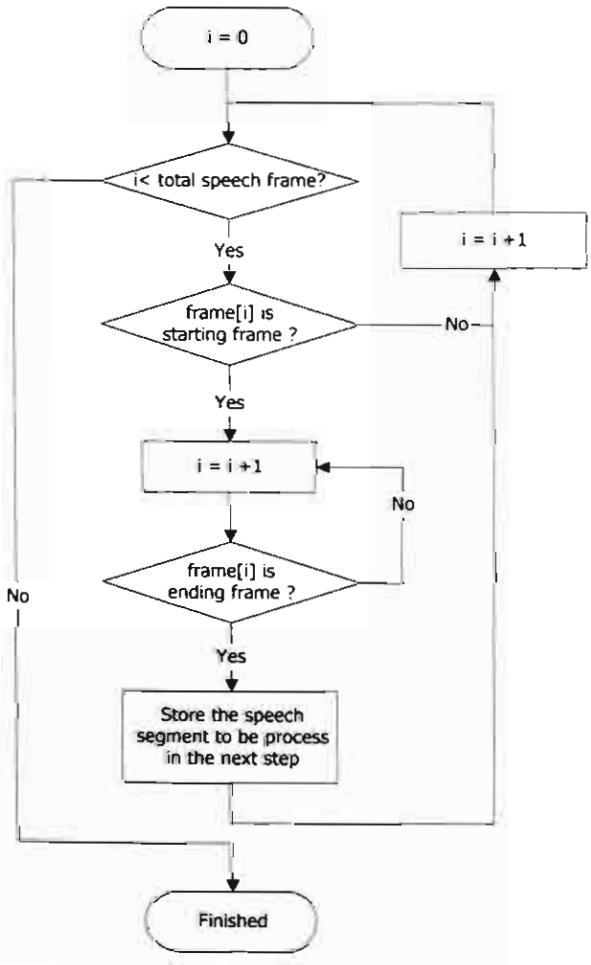


Figure 2: The algorithm to detect the starting and ending frames.

2.2. Syllable Recognition

Each Thai syllable sound comprises four different types of phoneme, namely leading consonant, vowel, ending consonant, and tone. In order to recognize a Thai syllable, all these four constituents of that syllable must be recognized.

The proposed syllable recognition system comprises five processes namely 1) leading consonant, vowel, ending consonant and tone (LVET) feature extraction process, 2) leading consonant recognition process (LRP), 3) vowel recognition process (VRP), 4) ending consonant recognition process (ERP) and 4) tone recognition process (TRP). A block diagram of the overall recognition system which is described in detail in [12] is given in Figure 3.

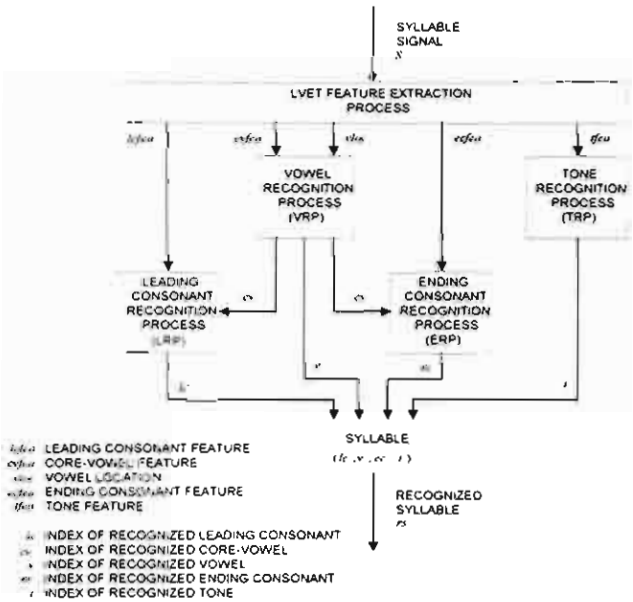


Figure 3: A block diagram of the syllable recognition system.

2.2.1. LVET Feature Extraction Process

This process is responsible for extracting all the features needed from each segmented syllable signal for the four following recognition processes, i.e. LRP, VRP, ERP, and TRP.

The Linear Predictor Coefficient (LPC) analysis as defined in [9] and [13] is conducted to determine the Ceptral Coefficient and energy feature vector, so-called CEP_E feature vector. In addition, fundamental frequency contour [14] is extracted from each segmented syllable signal.

Then, vowel location is detected based on the differences between CEP_E feature vectors of the frames of that syllable. Ceptral and energy thresholds are used to determine the beginning and ending frames of the vowel part. Details of this vowel location detection algorithm is given in [12]

Subsequently, the CEP_E feature vector is segmented into three feature vectors, *lcfea*, *cvfea* and *ecfea*, according to the vowel location. These three feature vectors are to be used as the inputs of LRP, VRP, and ERP, respectively. A Block diagram of this particular process, so-called LVE feature segmentation, is given in Figure 4.

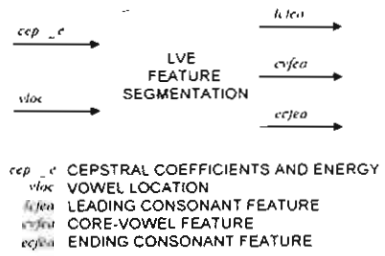


Figure 4: A block diagram of LVE feature segmentation.

On the other hand, the fundamental frequency contour is used to construct a tone feature vector, *tfea* which becomes an input into TRP.

2.2.2. Tone Recognition Process (TRP)

In this process, the tone phoneme is recognized. A neural network is employed as the recognition engine. A block diagram of this process is given in Figure 5. The tone feature vector, *tfea* from the LVET feature extraction process is processed. As a result of the recognition for each syllable, an index of a recognized tone *t* is returned.

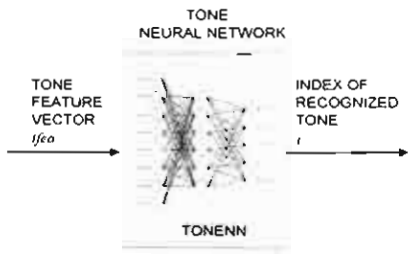


Figure 5: A block diagram of TRP.

2.2.3. Vowel Recognition Process (VRP)

In this process, vowel phonemes are recognized. In order to recognize vowels two elements must be determined, type of vowel and vowel length. There are 12 different types of vowel, so-called core vowels and two vowel lengths, short and long, in Thai language. A block diagram of this VRP is given in Figure 6. The process is divided into the core vowel recognition part and the vowel length determination part, which are further discussed below.

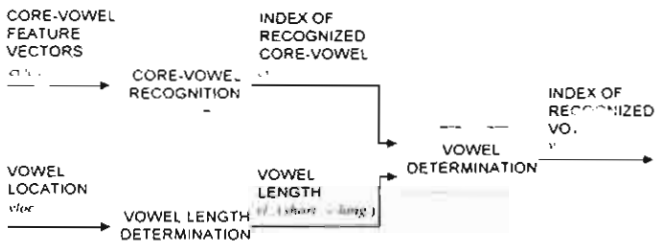


Figure 6: A block diagram of VRP.

Core-vowel recognition: A Hidden Markov Model (HMM) is used to represent each core-vowel class. Hence, 12 HMMs are included. A block diagram of core-vowel recognition is given in Figure 7. The type of HMM used in this process is Continuous Density Hidden Markov Model (CDHMM) whose details are described in [9][13].

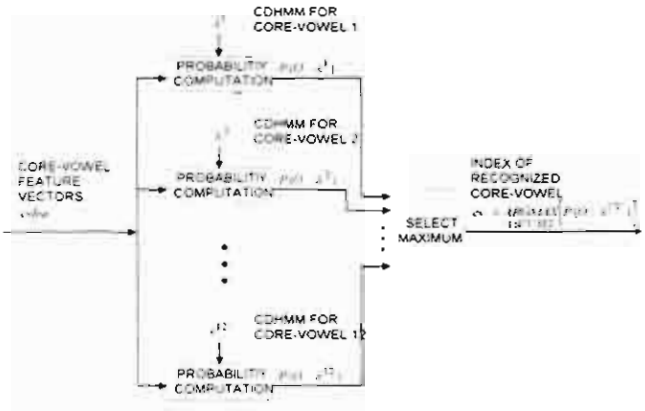


Figure 7: A block diagram of core-vowel recognition

Vowel length determination: The vowel location has been identified with the frame numbers of the starting and the ending points of the vowel with respect to each segmented syllable signal. The vowel length can easily computed in terms of number of frames from these starting and ending points. A simple threshold method is then used to determine whether the vowel is short or long. If the vowel length exceeds the threshold, it is long vowel. Otherwise, it is short vowel.

2.2.4. Leading Consonant Recognition Process (LRP)

Here, leading consonant feature vectors, *lcfea* from LVET feature extraction process and the index of recognized core-vowel from VRP are processed. As a result of this LRP process, an index of recognized leading consonant is returned. This means that the recognition of leading consonant depends on the recognized core-vowel type from the VRP. A block diagram of LRP is given in Figure 8.

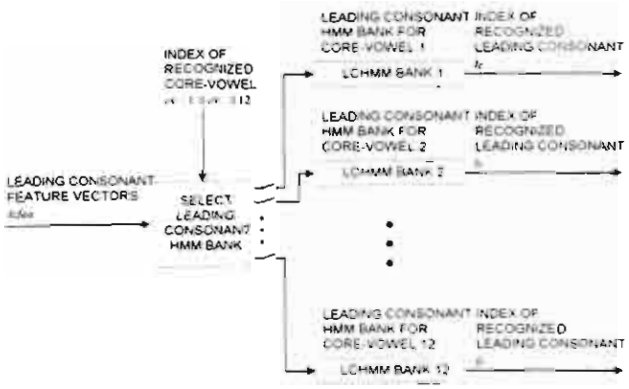


Figure 8: A block diagram of LRP.

For each core-vowel type, there is an LCHMM bank. Each LCHMM bank is designed to cover all possible 38 leading consonant classes of Thai language. Each LCHMM consists of 38 HMMs. This means that for each leading consonant class, there is a HMM corresponding to it. Each HMM in LCHMM bank is also a CDHMM. Figure 9 shows a block diagram of an LCHMM bank.

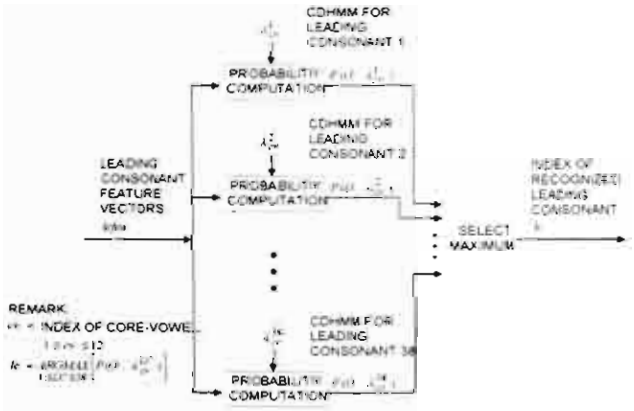


Figure 9: A block diagram of an LCHMM bank.

2.2.5. Ending Consonant Recognition Process (ERP)

In this process, the ending consonant feature vectors, *ecfea* and the index of recognized core-vowel from VRP are similarly processed. An index of recognized ending consonant is returned as the output. Observe that the recognition of ending consonant is also based on the core-vowel recognized in VRP. A block diagram of this ERP is shown in Figure 10.

Like in the case of the leading consonant, there is a corresponding ECHMM bank for each core-vowel. Each ECHMM bank consists of at most 9 HMMs because not all ending consonants can be associated with every core-vowel. An HMM in each ECHMM bank represents an ending consonant class associated with the corresponding core-vowel. Each HMM in ECHMM bank is also a CDHMM. A block diagram of an ECHMM bank is depicted in Figure 11.

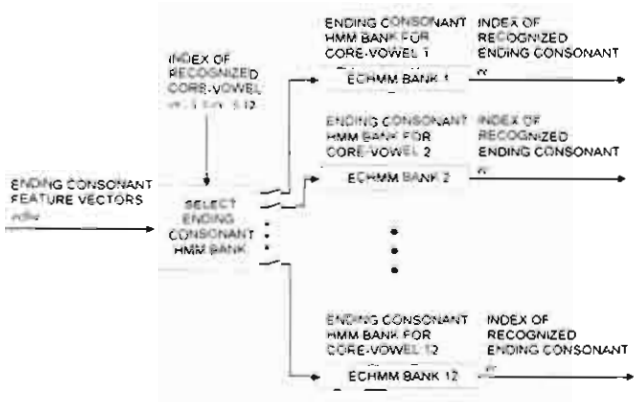


Figure 10: A block diagram of ERP.

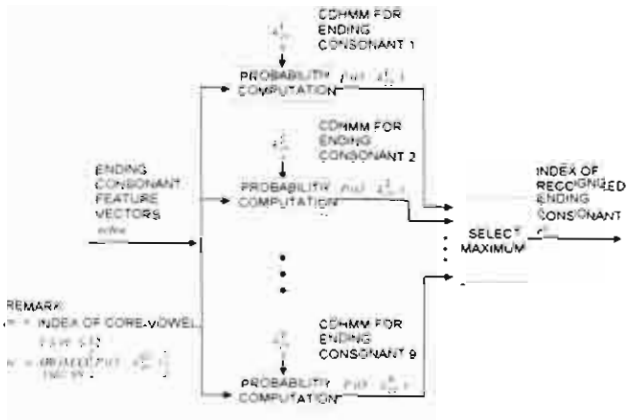


Figure 11: A block diagram of an ECHMM bank.

2.3. Syllable-Based Word Recognition

After the syllables are recognized, one more difficult task awaits us. This is the grouping of the recognized syllables into meaningful words. We propose two approaches to accomplish this particular task. They are Rule-based Word Recognition, and Genetic Decoding Algorithm as outlined in Sections 2.3.1 and 2.3.2, respectively. Both of these approaches require the context of discourse in order to associate syllables into appropriate vocabularies. Note that the syllable recognition as described in Section 2.2 does not always produce perfect results. These two word recognition approaches attempt to make appropriate corrections while trying to recognize the words.

2.3.1. Rule-based Word Recognition

In this first alternative, an ambiguous phonetic dictionary of Thai language is constructed for words within the context of discourse. Words are considered according to the number of syllables contained.

For each word, a word model matrix is constructed. This matrix contains potential variations to the pronunciation of the word as possibly recognized

by the processes described in Section 2.2. Probabilities of the alternative variations to the pronunciation are calculated for each syllable of the correct word. These alternative variations are called second, third, forth and other matches. The recognition probabilities of the variations for each syllable are summed to 1. An example of a table showing potential alternative variations to different leading consonants of a vowel /a/ is shown in Figure 12.

9 /a/ ะ 21 /aa/ า				
1.[/ph/ ,(พ)]	2.[/th/ ,(ท)] 30	3.[/kh/ ,(ค)] 6	12.[/h/ ,(ห)] 2	38
	0.7894	0.1578	0.0527	0.0001
2.[/th/ ,(ท)]	1.[/ph/ ,(พ)] 21	3.[/kh/ ,(ค)] 8	13.[/ch/ ,(ช)] 8	37
	0.5675	0.2162	0.2162	0.0001
3.[/kh/ ,(ค)]	2.[/th/ ,(ท)] 26	6.[/k/ ,(ก)] 5	1.[/ph/ ,(พ)] 2	33
	0.7878	0.1515	0.0606	0.0001
4.[/p/ ,(ป)]	8.[/b/ ,(บ)] 20	10.[/f/ ,(ฝ)] 5	1.[/ph/ ,(พ)] 4	29
	0.6896	0.1724	0.1379	0.0001
5.[/t/ ,(ต)]	13.[/ch/ ,(ช)] 10	18.[/l/ ,(ล)] 7	2.[/th/ ,(ท)] 5	22
	0.4545	0.3181	0.2273	0.0001
6.[/k/ ,(ก)]	3.[/kh/ ,(ค)] 12	17.[/nj/ ,(ง)] 8	21.[/r/ ,(ร)] 6	26
	0.4615	0.3076	0.2308	0.0001
7.[/?/ ,(อ)]	17.[/nj/ ,(ง)] 13	5.[/t/ ,(ต)] 9	3.[/kh/ ,(ค)] 6	28
	0.4643	0.3213	0.2143	0.0001
8.[/b/ ,(บ)]	15.[/m/ ,(ม)] 20	4.[/p/ ,(ป)] 9	9.[/d/ ,(ด)] 5	34
	0.5882	0.2647	0.1470	0.0001
9.[/d/ ,(ด)]	19.[/j/ ,(ย)] 13	14.[/c/ ,(จ)] 5	16.[/n/ ,(น)] 5	23
	0.5652	0.2174	0.2173	0.0001
10.[/f/ ,(ฝ)]	11.[/s/ ,(ส)] 32	20.[/w/ ,(ว)] 4	13.[/ch/ ,(ช)] 1	37
	0.8649	0.1080	0.0270	0.0001
11.[/s/ ,(ส)]	14.[/c/ ,(จ)] 17	10.[/f/ ,(ฝ)] 10	13.[/ch/ ,(ช)] 6	33
	0.5151	0.3030	0.1818	0.0001
12.[/h/ ,(ห)]	3.[/kh/ ,(ค)] 10	2.[/th/ ,(ท)] 9	1.[/ph/ ,(พ)] 9	28
	0.3571	0.3214	0.3214	0.0001
13.[/ch/ ,(ช)]	14.[/c/ ,(จ)] 15	2.[/th/ ,(ท)] 11	11.[/s/ ,(ส)] 6	32
	0.4688	0.3437	0.1875	0.0001
14.[/c/ ,(จ)]	11.[/s/ ,(ส)] 23	13.[/ch/ ,(ช)] 10	19.[/j/ ,(ย)] 4	37
	0.6216	0.2702	0.1081	0.0001
15.[/m/ ,(ม)]	17.[/nj/ ,(ง)] 21	4.[/p/ ,(ป)] 5	8.[/b/ ,(บ)] 4	30
	0.7000	0.1666	0.1333	0.0001
16.[/n/ ,(น)]	17.[/nj/ ,(ง)] 23	9.[/d/ ,(ด)] 11	21.[/r/ ,(ร)] 2	36
	0.6388	0.3056	0.0555	0.0001
17.[/nj/ ,(ง)]	16.[/n/ ,(น)] 18	15.[/m/ ,(ม)] 9	9.[/d/ ,(ด)] 5	32
	0.5625	0.2812	0.1562	0.0001
18.[/l/ ,(ล)]	21.[/r/ ,(ร)] 12	16.[/n/ ,(น)] 6	9.[/d/ ,(ด)] 5	23
	0.5217	0.2608	0.2174	0.0001
19.[/j/ ,(ย)]	9.[/d/ ,(ด)] 37	14.[/c/ ,(จ)] 2	17.[/nj/ ,(ง)] 1	40
	0.9249	0.0500	0.0250	0.0001
20.[/w/ ,(ว)]	4.[/p/ ,(ป)] 10	8.[/b/ ,(บ)] 7	17.[/nj/ ,(ง)] 7	24
	0.4166	0.2917	0.2916	0.0001
21.[/r/ ,(ร)]	18.[/l/ ,(ล)] 16	8.[/b/ ,(บ)] 5	6.[/k/ ,(ก)] 4	25
	0.6399	0.2000	0.1600	0.0001
22.[/ph/ ,(พล)]	1.[/ph/ ,(พ)] 21	18.[/l/ ,(ล)] 5	2.[/th/ ,(ท)] 3	29
	0.7241	0.1723	0.1034	0.0001
23.[/ph/ ,(พร)]	12.[/h/ ,(ห)] 21	3.[/kh/ ,(ค)] 5	6.[/k/ ,(ก)] 3	29
	0.7241	0.1723	0.1034	0.0001
24.[/kh/ ,(คล)]	3.[/kh/ ,(ค)] 26	18.[/l/ ,(ล)] 5	26.[/kh/ ,(คว)] 2	33
	0.7878	0.1515	0.0606	0.0001
25.[/kh/ ,(คล)]	3.[/kh/ ,(ค)] 26	18.[/l/ ,(ล)] 5	26.[/kh/ ,(คว)] 2	33
	0.7878	0.1515	0.0606	0.0001
26.[/kh/ ,(คว)]	3.[/kh/ ,(ค)] 26	20.[/w/ ,(ว)] 5	6.[/r/ ,(ร)] 2	33
	0.7878	0.1515	0.0606	0.0001
27.[/kh/ ,(ครว)]	3.[/kh/ ,(ค)] 26	2.[/th/ ,(ท)] 5	6.[/r/ ,(ร)] 2	33
	0.7878	0.1515	0.0606	0.0001
28.[/p/ ,(ปล)]	4.[/?/ ,(ป)] 9	8.[/b/ ,(บ)] 8	18.[/l/ ,(ล)] 6	23
	0.3913	0.3477	0.2609	0.0001
29.[/p/ ,(ปร)]	4.[/?/ ,(ป)] 9	8.[/b/ ,(บ)] 8	2.[/th/ ,(ท)] 6	23
	0.3913	0.3477	0.2609	0.0001
30.[/t/ ,(ตร)]	5.[/t/ ,(ต)] 16	11.[/s/ ,(ส)] 9	21.[/r/ ,(ร)] 7	32
	0.5000	0.2812	0.2187	0.0001
31.[/k/ ,(กล)]	33.[/k/ ,(กร)] 12	27.[/kh/ ,(ค)] 8	18.[/l/ ,(ล)] 6	26
	0.4615	0.3076	0.2308	0.0001
32.[/k/ ,(กว)]	6.[/k/ ,(ก)] 12	27.[/kh/ ,(ค)] 8	20.[/w/ ,(ว)] 6	26
	0.4615	0.3076	0.2308	0.0001
33.[/k/ ,(กร)]	27.[/kh/ ,(ค)] 12	17.[/nj/ ,(ง)] 8	21.[/r/ ,(ร)] 6	26
	0.4615	0.3076	0.2308	0.0001

Figure 12: An example of potential alternative pronunciation variations

Using data from appropriate tables, a word model matrix for any give word in the context of discourse can be constructed. An example of such a matrix is given in Figure 13.

	1	2	3
1	สะ (sa)	หวัด (wat)	ดี (di:)
	1	1	1
2	จะ (ca)	ปัด (pat)	นี (ni:)
	0.5151	0.4166	0.5
3	ฝะ (fa)	บัด (bat)	บี (bi:)
	0.303	0.2917	0.4117
4	ชะ (cha)	ฉัด (njat)	ยี (ji:)
	0.1818	0.2916	0.0882

Figure 13: An example of word matrix

Utilizing the word matrices, each given sentence is run through an algorithm called Word

Segmentation. This algorithm separates words contained in a sentence of a connected speech. Details of this algorithm are described in [15]. It can be summarized in three steps as follows.

1) Determine all possible word models of different lengths, say one syllable to four syllables, according to the recognized syllables of a sentence. Calculate the recognition probability of each word model based on the probabilities of its syllables.
2) Construct a graph containing all possible combinations of word models in the given sentence. An example of such a graph is shown in Figure 14.
3) For each path (combination) of word models in the graph, calculate the average recognition probability of the path.

Figure 14: An example of a graph containing all possible combinations of word models

Among the resulting paths (combinations) of word models, the one with the highest average recognition probability is chosen as a candidate. This candidate is then subject to another algorithm called Averages Likelihood which attempts to correct errors left over from the syllable recognition process. Details of this algorithm are also given in [15]. In essence, this algorithm basically looks at each word model in that candidate sentence, for any word model with a recognition probability lower than 1, an attempt is made to change its syllable(s) whose recognition probability is lower than 1 to the corresponding syllable(s) of the correct word for the model. Note that this correction can only be done for words defined in the context of discourse, i.e. the words need to be included in the dictionary of the system.

2.3.2. Genetic Decoding Algorithm

Unlike the very structured rule-based algorithm described in the previous section, the decoding process presented in this section is based on the concept of Genetic Algorithms (GA) [16].

First of all, appropriate ambiguous matrices need to be constructed for the four types of phoneme, i.e. leading consonant, vowel, ending consonant and tone. Each of these matrices contains possible variations of incorrect recognition of concerned phonetic value, e.g. a given leading consonant together with corresponding ambiguous degrees. Each ambiguous degree is basically a probability of that particular incorrect recognition among all the incorrect recognitions. A partial ambiguous matrix for two leading consonants is shown in Figure 15.

	/ph/(ผ)	/th/(ท)	/kh/ (ค)	/k/(ก)	/h/(ห)	/ch/(ช)
/ph/(ผ)	1	0.8	0.125	0.025	0.05	0
/th/(ท)	0.625	1	0.125	0.05	0.075	0.25

Figure 15: A partial ambiguous matrix for leading consonants

The decoding process then starts with a set of potential word sequences as the initial population. These word sequences for the initial population are selected with the following two conditions.

- Words are randomly selected in such a way that they have the same vowel or tone to those in the same positions in the input word sequence.
- The number of syllables in a word sequence is equal to the number of syllables in the input word sequence.

The fitness value is calculated for each word sequence according to the fitness function below

$$FN = \sum_{i=1}^{w}\left(\sum_{j=1}^{n} \text{aDegree}_j\right) \Big/ NS$$

where w = # of words in chromosome
n = # of syllables in each word
NS = # of syllables in chromosome

$$\text{aDegree}_j = \begin{cases} \text{ambiguous degree} \\ \text{normal degree} \end{cases}$$

If the fitness value of the current word sequence is not good enough to be a solution then a new generation of word sequences is generated by selecting two parents and applying the crossover operation. An example showing the crossover operation is given in Figure 16.

The number of generations and acceptable fitness value are set as a condition to stop the decoding process. There are two alternatives to stopping the decoding process.

In the first alternative, it is assumed that the input syllable from the syllable recognition process has a 100% recognition rate. The acceptable fitness value is then set to 1. The fitness function uses the normal degree as a degree of fitness. The process is stopped when there is a fitness value of a word sequence equal to the acceptable fitness value. If, however, more than 80% of word sequences in the current population have the same fitness value then

it can be concluded that the input syllable from the syllable recognition process has a recognition rate less than 100%, i.e. there are some errors. In such a case, the decoding process is restarted with the second alternative.

For second alternative, the fitness function uses the ambiguous degree from a corresponding ambiguous matrix as a degree of fitness. This alternative is stopped when more than 80% of word sequences in the current generation have the same fitness value. The word sequence that has a maximum fitness value is picking up as the result of this decoding process. The details of this genetic word decoding algorithm can be found in [17].

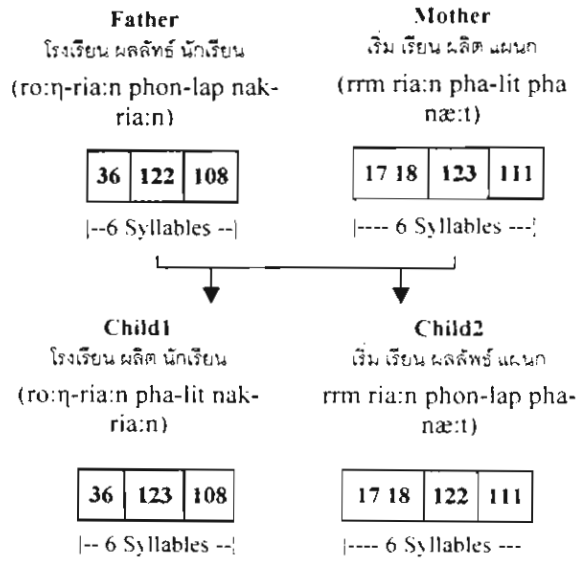


Figure 16: An example of the crossover operation

3. Conclusion

This paper presents an overall framework upon which a connected speech recognition system for Thai language can be built. This speech recognition problem is divided into three research problems, namely syllable segmentation problem, syllable recognition problem and syllable-based word recognition problem.

The first two problems tackle the signal processing portion. As for the syllable segmentation process, it needs to be tuned to fit the speaking style of representative speakers. The one used in our experiments is tuned for moderate speaking speed with typical loudness. It has also been observed that the quality of syllable recognition depends on the quality of the training set. The system tends to perform better when recognizing speeches of speakers whose sample words are included in the training set. In addition, even though we attempt to recognize syllables based on their phonemes, words that are included in the training set tend to be recognized better than those that are not. An important factor here is on the varying pattern of the speech signal when two syllables are connected. The syllables are recognized better when the training set contains their connecting pattern.

It should clearly be seen that the signal processing portion alone cannot achieve a high recognition rate in most cases. Two techniques to improve the recognition rate by means of associating recognized syllables with words from a given context of discourse are proposed. The rule-based word recognition approach is quite traditional and very well-structured, while the genetic word decoding algorithm follows a soft computing paradigm. Both show promising results. However, it should be observed that both techniques rely on the empirical result concerning the probability of incorrect recognition with respect to each phonetic value.

With the current stage of advancement in computing platform, it can be concluded that the connected speech recognition can still only be practically achieved within a given context of discourse. Enough words from the context need to be included in the training set for the syllable recognition process as well as in the dictionary for the syllable-based word recognition process in order to obtain reasonably high recognition rate.

4. References

[1] Pensiri, R. and Jitapunkul, S., *Speaker-Independent Thai Numerical Voice Recognition by using Dynamic Time Warping*, Proceedings of the 18th Electrical Engineering Conference, pp. 977-981, 1995.

[2] Pornsukchandra, W. and Jitapunkul, S., *Speaker-Independent Thai Numeral Speech Recognition using LPC and the Back Propagation Neural Network*, Proceedings of the 19th Electrical Engineering Conference, pp. 977-981, 1996.

[3] Maneenoi, E., Jitapunkul, S., Wutiwuwatchai, C., and Ahkuputra, V., *Modification of BP Algorithm for Thai Speech Recognition*, Proceeding of the 1997 International Symposium on Natural Language Processing, 1997.

[4] Wutiwiwatchai, C., *Speaker Independent Thai Numeral Speech Recognition Using Neural Network and Fuzzy Technique*, Master's thesis, Chulalongkorn University, 1997.

[5] Ahkuputra, V., Jitapunkul, S., Pornsukchandra, W., and Luksaneeyanawin, S., *A Speaker-Independent Thai Polysyllabic Word Recognition Using Hidden Markov Model*, Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, pp. 593-599, 1997.

[6] Ahkuputra, V., Jitapunkul, S., Maneenoi, E., Kasuriya, S., and Amornkul, P., *Comparison of Different Techniques On Thai Speech Recognition*, The 1998 IEEE Asia-Pacific Conference on Circuits and Systems, pp. 177-180 1998.

[7] Jitapunkul, S., Luksaneeyanawin, S., Ahkuputra, V., Maneenoi, E., Kasuriya, S., and Amornkul, P., *Recent Advances of Thai Speech Recognition in Thailand*, The 1998 IEEE Asia-Pacific Conference on Circuits and Systems, pp. 173-176, 1998.

[8] Chaiareerat, J. and Santiprabhob P., *Fuzzy-based Thai Syllable Segmentation for Connected Speech using Energy and Different Ceptral*, Proceedings of InTech/VJFuzzy, pp.334-337, December, 2002.

[9] Rabiner, L. R. and Juang, B. H., *Fundamentals of Speech Recognition*, A. Oppenheim, Series Editor, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[10] Jittiwarangkul N., Jitapunkul S., Luksaneeyanawin S., Ahkuputra V., Wutiwiwatchai C., *Thai Syllable Segmentation for Connected Speech based on Energy*, Proceedings of the IEEE APCCAS, pp. WP1-8.1, Nov. 1998.

[11] MathWorks, Inc., *Fuzzy Logic Toolbox for use with MATLAB User Guide*, Version 2, 1999.

[12] Cheirsilp, R. and Santiprabhob P., *Phoneme-Based Thai Syllable Recognition by Means of Soft Computing*, Proceedings of InTech/VJFuzzy, pp.325-333, December, 2002.

[13] Rabiner, L. R., *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceeding of IEEE, Vol. 77, No.2, pp. 257-286, February, 1989.

[14] Rowden, C., *Speech Processing*, London: McGraw-Hill, 1992.

[15] Dachapratumvan, N. and Santiprabhob P., *Thai Syllablic Correction in Connected Thai Speech Recognition*, Proceedings of InTech/VJFuzzy, pp.314-319, December, 2002.

[16] Lawrence D., *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York, 1991.

[17] Supasirirojana, W. and Santiprabhob P., *Thai Word Decoder Based on Genetic Algorithm*, Proceedings of InTech/VJFuzzy, pp.320-324, December, 2002.

โครงการ "การเรียนรู้จำเสียงพูดภาษาไทยโดยใช้นิวโรฟัซซี่"

Thai Speech Recognition using NeuroFuzzy