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