บทคัดย่อ

รหัสโครงการ: TRG5780202

ชื่อโครงการ: การรู้จำเสียงร้องที่ไม่ขึ้นกับเสียงรบกวนพื้นหลังโดยใช้แบบรูปสเปกโทรแกรมเชิงภาพ

ชื่อนักวิจัย: ดร.พีระพล ขุนอาสา

E-mail Address: peerapol_utt@hotmail.com, peerapol@uru.ac.th

ระยะเวลาโครงการ : 2 ปี

การรู้จำเสียงพูด (Speech Recognition) คือ การที่คอมพิวเตอร์สามารถรับรู้ เสียงของมนุษย์ได้โดยอัตโนมัติ โดยทั่วไปแล้วจะอาศัยระบบโปรแกรมคอมพิวเตอร์ที่สามารถแปลงเสียงพูด (Audio File) เป็นข้อความตัวอักษร (Text) โดยสามารถแจกแจงคำพูดต่างๆ ที่มนุษย์สามารถพูดใส่ไมโครโฟน โทรศัพท์หรืออุปกรณ์อื่นๆ และเข้าใจคำศัพท์ทุกคำอย่างถูกต้องเกือบ 100% โดยเป็นอิสระจากขนาดของกลุ่มคำศัพท์ ความดังของเสียงและลักษณะการออกเสียงของผู้พูด โดยระบบจะรับฟังเสียงพูดและตัดสินใจว่าเสียงที่ได้ยินนั้นเป็นคำๆใด โดยทั่วไปนั้นการรับรู้ของมนุษย์มีประสิทธิภาพสูงสามารถที่จะจำแนกประเภทเสียงต่างๆได้ แต่ปัญหาการรู้จำเสียง(Audio recognition) โดยใช้คอมพิวเตอร์นั้น คือการให้คอมพิวเตอร์สามารถจำแนกเสียงประเภทต่างๆ ไม่ว่าจะเป็น เสียงเพลง เสียงพูด หรือเสียงร้องได้เหมือนโสตประสาทของมนุษย์นั้นมีความซับซ้อนสูงและยากที่จะเลียนแบบ

ในงานวิจัยชิ้นนี้ เราสนใจในกลุ่มหัวข้อการสืบค้นข้อมูลจากเพลง(Music Information Retrieval) โดยเฉพาะปัญหาในการรู้จำคำร้อง(Singing voice recognition) จากเพลงนั้นมีความยุ่งยากและซับซ้อนมาก เนื่องจากคำที่ออกเสียงจากการร้องเพลงนั้นมีคุณสมบัติที่แตกต่างจากการพูดโดยทั่วๆ ไปหลายประการ เช่น รูปแบบการออกเสียงเฉพาะตัวบุคคลที่ขึ้นกับประเภทของดนตรี ระยะเวลาในการออกเสียงคำที่แตกต่างกันจากปัจจัยการเอื้อนหรือการลากเสียง ลูกคอในการร้องเพลง และจังหวะของเพลงต่างๆ จึงทำให้อัลกอลิทึมเดิมที่ใช้กับปัญหาการรู้จำเสียงพูดนั้นไม่ประสบผลสำเร็จเท่าที่ควร

อีกปัจจัยหนึ่งที่ลดทอนคุณภาพของการรู้จำคือเพลงจะมีเสียงดนตรีประกอบจะมีผลเทียบได้กับเสียงรบกวน(Noise) ซึ่งจะทำให้ประสิทธิภาพในการรู้ลดลง โดยทั่วไปแล้วการแก้ปัญหาคือการใช้ตัวรองเสียงรบกวน (Noise filter) ชนิดต่างๆ เข้ามากรองเสียงดนตรีออกไป แต่การใช้ตัวกรองเสียงรบกวนนั้นจะเป็นการเพิ่มขั้นตอนเข้าไปทำให้เวลาในการประมวลผลเพิ่มขึ้นตาม อีกทั้งการใช้ตัวรองเสียงรบกวนอาจไปทำลายเสียงที่ต้องการใช้จริงไปด้วย

ดังนั้นในงานวิจัยชิ้นนี้จึงมีจุดประสงค์หลักคือ รู้จำคำร้องต่างๆ ในเพลงโดยที่ไม่ใช้ตัวรองเสียงรบกวน (Noise filter) มากรองหรือขจัดเสียงดนตรี เสียงรบกวนออกไป งานวิจัยนี้อาศัยหลักการของการรู้จำรูปภาพเข้ามาประยุกต์ใช้ในการแก้ปัญหา อย่างแรกนำเอาสัญาณเสียงที่ได้ไปแปลงให้อยู่ในรูปแบบของภาพที่เราเรียกว่า สเปกโตรแกรม(Spectrogram) ที่มีลักษณะข้อมูลเป็นแบบ MxN มิติ แต่เนื่องจาก สเปกโตรแกรม(Spectrogram) มีขนาดของมิติข้อมูลที่สูงเกินไปการนำเอามาใช้งานโดยตรงนั้นเป็นไปได้ยากและต้องใช้เครื่องประมวลผลที่ประสิทธิภาพสูงมาก อีกทั้งคำร้องแต่ละคำเมื่อนำมาแปลงเป็น สเปกโตรแกรม(Spectrogram) ขนาดจะไม่เท่ากันไม่สามารถนำไปทำการรู้จำได้ ดังนั้นงานวิจัยนี้จึงนำเอา สเปกโตรแกรม(Spectrogram) มาลดขนาดของมิติข้อมูลลงก่อนนำไปทำการรู้จำด้วยเทคนิคการปรับขนาดภาพ(Image resizing algorithm) สำหรับขั้นตอนในการรู้จำนั้นงานวิจัยนี้เลือกใช้ Feed-Forward neural Network

จากการทดลองพบว่างงานวิจัยนี้สามารถแก้ปัญหาการรู้จำเสียงร้องหรือคำร้องในเพลงที่มีเสียงดนตรีพื้นหลังได้เป็นอย่างดี โดยประสิทธิภาพในการรู้จำอยู่ที่ 90.0% ขึ้นไปในทุก ๆ ชุดทดลอง อีกทั้งยังสามารถรู้จำได้หลายๆ ภาษาพร้อมๆ กันในชุดข้อมูลเดียวกันอีกด้วย

คำหลัก: การรู้จำแบบ, นิวรอนเน็ตเวิร์ค, ประมวลผลสัญญาณดิจิตอล, การสกัดคุณลักษณะเด่น

Abstract

Project Code: TRG5780202

Project Title: Single Signal Entity Approach for Thai Singing Word Recognition Using Images of Power Spectrogram and Image Processing Techniques

Investigator: Dr. Peerapol Khunarsa

E-mail Address: peerapol_utt@hotmail.com, peerapol@uru.ac.th

Project Period: 2 Years

Singing word recognition is one of the interesting research topics in the area of Music Information Retrieval (MIR). The first approach to solve this problem used successful techniques in Automatic Speech Recognition (ASR). Moving from monophonic to polyphonic audio signal, the problem has become more complex. The background instrumental accompaniment is regarded as the noise source degrading the performance of the recognition system. The papers proposed a statistical learning method for recognition of the word in a singing signal with background music and for classification of singing voice region in a polyphonic audio signal.

The goal of this paper is to solve singing word recognition without using any method to separated instrumental from background music. The papers also applied the concept of image recognition by using a สเปกโตรแกรม(Spectrogram) as an image to solve the problem. An audio signal that accompanies music was analyzed and transformed into a สเปกโตรแกรม(Spectrogram). A dimension of สเปกโตรแกรม (Spectrogram) is very high and time interval of each singing word is not equal. Then we apply image resizing algorithm to solve both problem. To recognize it, the whole สเปกโตรแกรม(Spectrogram) was sliced and formed as a feature vector for a neural classifier. Several classification functions are compared, such as Fisher classifier, K-nearest

neighbor and Feed-Forward can effectively recognize the word in music with the accuracy rate more than 90.0% Especially, we can recognize Cross-Language Music Data.

Keyword Spectrogram, Singing voice recognition, Automatic speech recognition (ASR), Feed-Forward Neural Network.