

การสร้างเสียงสังเคราะห์ด้วยการโคลนเสียงส่วนบุคคล

โดย นางสาวชนนิก สีหาบุตร

นายณัฐธรรศ ชูรัตน์

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาและประเมินประสิทธิภาพของระบบโคลนเสียงส่วนบุคคลสำหรับการสร้างเสียงสังเคราะห์ โดยบูรณาการเทคนิค Retrieval-Based Voice Conversion (RVC) ร่วมกับ Text-to-Speech (TTS) กระบวนการวิจัยเริ่มจากการเก็บรวบรวมข้อมูลเสียงต้นฉบับจากบุคคลเป้าหมายเพื่อนำมาฝึกฝนโมเดล RVC โดยใช้ ContentVec (HuBERT-based) สำหรับการสกัดคุณลักษณะเสียง และประยุกต์ใช้ TTS เพื่อสังเคราะห์เสียงพูดจากข้อความที่กำหนด ระบบสามารถทำงานได้ใน 2 โหมด คือ การแปลงเสียงจากเสียงต้นฉบับ และการสร้างเสียงใหม่จากข้อความ

การประเมินประสิทธิภาพของระบบดำเนินการผ่านการทดสอบกับกลุ่มตัวอย่างจำนวน 40 คน โดยใช้แบบสอบถามออนไลน์ผ่าน Google Forms ซึ่งประเมินใน 5 มิติ ได้แก่ ความชัดเจนของเสียง ความเป็นธรรมชาติของเสียง ความถูกต้องของอารมณ์ คุณภาพเสียงโดยรวม และเสียงรบกวน ผลการวิจัยพบว่าระบบที่พัฒนาขึ้นสามารถสร้างเสียงสังเคราะห์ที่มีคุณภาพสูง โดยได้คะแนนเฉลี่ยมากกว่า 4.40 จาก 5.00 ในทุกด้าน ทั้งในภาษาไทยและภาษาอังกฤษ

นอกจากนี้ งานวิจัยยังได้พัฒนาระบบฝึกรับรู้เสียงภาษาอังกฤษโดยใช้เทคโนโลยีการโคลนเสียงร่วมกับการแปลงเสียงเป็นข้อความ (Speech-to-Text) ซึ่งช่วยให้ผู้เรียนสามารถฟังเสียงของตนเองที่ออกเสียงได้อย่างถูกต้องและได้รับข้อมูลย้อนกลับทันที ผลลัพธ์จากงานวิจัยนี้มีศักยภาพในการประยุกต์ใช้หลากหลายด้าน ทั้งการสร้างเสียงทดแทนสำหรับผู้มีความบกพร่องทางการพูด การพัฒนาสื่อคอนเทนต์และวิดีโอ การสร้างเสียงตัวละครในเกมและแอนิเมชัน และการผลิตสื่อการเรียนการสอนที่มีประสิทธิภาพ

Development of Synthetic Speech through Personal Voice Cloning

By Miss. Chonchanok Seehabut

Mr. Nutthuch Choorat

ABSTRACT

This research aims to develop and evaluate the effectiveness of a personalized voice cloning system for synthetic speech generation by integrating Retrieval-Based Voice Conversion (RVC) with Text-to-Speech (TTS) technologies. The methodology involves collecting high-fidelity voice samples from target individuals to train an RVC model, using ContentVec (HuBERT-based) for voice feature extraction, while leveraging TTS capabilities to generate synthetic speech from specified text inputs. The system operates in two modes: converting voice from original audio and generating new speech from text.

System performance was evaluated through a user study involving 40 participants who assessed synthesized speech samples via Google Forms questionnaires. The evaluation employed five metrics: voice clarity, naturalness, emotion accuracy, overall quality, and noise artifacts. The experimental results demonstrate that the developed system effectively produces high-quality synthetic voices, achieving average scores above 4.40 out of 5.00 across all dimensions for both Thai and English languages.

Additionally, the research developed an English pronunciation training system utilizing voice cloning technology combined with Speech-to-Text (STT), enabling learners to hear their own voices with correct pronunciation and receive immediate feedback. The findings from this research have potential applications across multiple domains, including assistive speech technology for individuals with speech impairments, content and video production, character voice creation for games and animation, and effective educational media development.