StutterZero: End-to-End Speech Conversion to Transcribe and Correct Stutters

Xu, Qianheng (School: Millburn High School)

Over 70 million in the world suffer from stuttering, a speech impairment characterized by involuntary disruptions during normal speech. This can create communication barriers and limiting access to voice-enabled technologies, often failing to recognize disfluent speech. Children and teenagers are more likely to suffer from anxiety and depression because of their stutter. This study develops two deep learning solutions for stuttering intervention: (1) a hybrid ASR-based pipeline, a fine-tuned Whisper-Small ASR model to produce accurate transcripts and MeloTTS text-to-speech generation (Word Error Rate (WER) = 0.04, Word Information Preservation (WIP) = 0.95) , and (2) an end-to-end multitask model, which employs a encoder-multitask decoder model on log-Mel spectrograms to directly generate fluent speech (WER=0.116, WIP=0.898). Both models significantly outperform the baseline Whisper-Tiny ASR model (WER = 0.43,WIP = 0.47), with the end-to-end approach offering computational efficiency by bypassing intermediate text representations. Importantly, this research introduces the first ever end-to-end model to correct stutters. Beyond technical advances, this work makes three key contributions: (i) the release of two extended, publicly available stuttered-fluent paired datasets, supporting future speech AI research, (ii) empirical validation that end-to-end, multitask stuttering correction is possible and highly performant, (iii) and a framework for stuttering correction, applicable to equitable voice technology and inclusive clinical tools. These findings provide both immediate tools for assistive technology development and a foundation for future speech pathology research, advancing accessibility and equity in voice AI.

Awards Won: