ABSTRAK

PENGEMBANGAN TEKNOLOGI TEXT-TO-SPEECH (TTS) BAHASA INDONESIA MENGGUNAKAN VOICE CLONING BERBASIS KONVERSI GRAPHEME-TO-PHONEME (G2P)

OLEH

MUHAMMAD AL FAHREZI

Berbagai kondisi medis seperti *stroke*, *cerebral palsy*, cedera otak, penyakit neurodegeneratif, dan kanker kepala–leher dapat menyebabkan kehilangan kemampuan bicara sehingga memerlukan teknologi bantu komunikasi. Text-to-Speech (TTS) menjadi solusi yang memungkinkan teks diubah menjadi suara sintesis untuk mendukung komunikasi pengguna. Penelitian ini mengembangkan sistem TTS Bahasa Indonesia dengan menerapkan voice cloning berbasis Grapheme-to-Phoneme (G2P) untuk meningkatkan ketepatan pelafalan. Sistem dibangun menggunakan arsitektur Variational Inference with Adversarial Learning for End-to-End Text-to-Speech (VITS) dengan modul G2P berbasis BERT untuk menghasilkan representasi fonem kontekstual. Evaluasi menunjukkan model G2P mencapai akurasi 0,9907 dan Phoneme Error Rate (PER) 2,15%. Namun, integrasi G2P pada sistem TTS belum meningkatkan kualitas suara, di mana sistem tanpa G2P memperoleh Mean Opinion Score (MOS) 3,80 dan dengan G2P sebesar 2,89. Hasil ini menunjukkan perlunya optimasi lebih lanjut pada integrasi G2P dalam model VITS untuk meningkatkan kualitas suara sintetis Bahasa Indonesia.

Kata kunci: *Text-to-Speech*, *Grapheme-to-Phoneme*, *Voice Cloning*, BERT, VITS, Bahasa Indonesia

ABSTRACT

DEVELOPMENT OF AN INDONESIAN TEXT-TO-SPEECH (TTS) SYSTEM USING VOICE CLONING BASED ON GRAPHEME-TO-PHONEME (G2P) CONVERSION

by

MUHAMMAD AL FAHREZI

Various medical conditions such as stroke, cerebral palsy, brain injury, neurodegenerative diseases, and head–neck cancer can lead to speech impairment, thereby requiring assistive communication technologies. Text-to-Speech (TTS) *systems provide a solution by converting text into synthesized speech to support user communication. This study develops an Indonesian Text-to-Speech system by implementing voice cloning with a Grapheme-to-Phoneme* (G2P) *approach to improve pronunciation accuracy. The system is built using the Variational Inference with Adversarial Learning for End-to-End Text-to-Speech* (VITS) *architecture, integrated with a* BERT *based G2P module to generate contextual phoneme representations. Evaluation results show that the* G2P *model achieves an accuracy of 0.9907 and a Phoneme Error Rate* (PER) *of 2.15%. However, the integration of* G2P *into the* TTS *system does not improve speech quality, as the system without* G2P *achieves a Mean Opinion Score* (MOS) *of 3.80, while the system with* G2P *obtains a* MOS *of 2.89. These findings indicate that further optimization is required in integrating the* G2P *module into the* VITS *model to enhance the quality of synthesized Indonesian speech.*

Keywords: *Text-to-Speech*, *Grapheme-to-Phoneme*, *Voice Cloning*, BERT, VITS, Indonesian *Language*