

IMPLEMENTASI MODEL *VARIATIONAL AUTOENCODER WITH ADVERSARIAL LEARNING FOR END-TO-END TEXT-TO-SPEECH* (VITS) UNTUK BAHASA INDONESIA

Nama Mahasiswa : Yoga Tiara Wiguna
NIM : 11211086
Dosen Pembimbing Utama : Bima Prihasto, S.Si., M.Si., Ph.D.
Pembimbing Pendamping : Bobby Mugi Pratama, S.Si., M.Han.

ABSTRAK

Penelitian ini bertujuan untuk mengimplementasikan dan mengevaluasi model *Variational Autoencoder With Adversarial Learning For End-To-End Text-to-Speech* (VITS) dengan menggunakan audio Bahasa Indonesia. Penelitian pada model VITS sebelumnya menggunakan *dataset* Bahasa Inggris dan audio yang dihasilkan sudah cukup baik. Maka dari itu dilakukan penelitian ini dengan menggunakan Bahasa Indonesia, dikarenakan di Indonesia pengembangan teknologi masih menghadapi kendala, seperti keterbatasan *dataset* berkualitas dan minimnya penelitian terkait. Peneliti menggunakan transkrip dari *dataset* TITML-IDN, ASR-SindoDuSC, ASR-IndoCSC, dan *audiobook* novel novel The Art of War sebagai bahan untuk membuat *dataset* yang baru, lalu membandingkan nilai evaluasi model VITS yang menggunakan Stochastic Duration Predictor (SDP) dengan VITS yang menggunakan Deterministic Duration Predictor (DDP). Peneliti juga melakukan penerapan pelatihan *adversarial* pada *duration predictor* untuk memprediksi durasi pengucapan. Evaluasi dilakukan tidak hanya dengan pendekatan subjektif menggunakan *Mean Opinion Score* (MOS) tetapi juga dengan pendekatan objektif menggunakan *Resemblyzer cosine similarity*. *Dataset* yang digunakan sebanyak 343 dengan audio cenderung formal, namun pada 1250 data audio yang digunakan lebih bervariasi baik formal maupun informal. Dari kedua *dataset*, 1250 *dataset* menunjukkan performa lebih baik dalam menghasilkan audio karena dapat menghasilkan nilai rata-rata *cosine similarity* sebesar 0,91124, *Mean Opinion Score* (MOS) didapatkan sebesar 4,54 pada salah satu jenis transkripnya. Selanjutnya dilakukannya perbandingan SDP dan DDP, dari hasil penelitian ditemukan bahwa SDP lebih natural daripada DDP dari segi durasi audio. Pada penelitian ini juga melakukan penambahan *adversarial learning* pada kedua jenis *duration predictor* dan hasilnya juga dapat lebih meningkatkan lagi kualitas audio dari keragaman durasi pengucapannya.

Kata kunci : TTS, VITS, SDP, DDP, Audio Bahasa Indonesia