ABSTRAK

Nama : Kurniawati Azizah, S.T., M.Phil
Program Studi : Doktor Ilmu Komputer
Judul : Pembelajaran Transfer Hierarkis pada *Text-To-Speech* Berbasis *Deep Learning* untuk Domain *Low-Resource*
Pembimbing : Prof. Dr. Eng. Wisnu Jatmiko S.T., M.Kom

Sistem berbasis *deep neural network* (DNN) pada umumnya membutuhkan data dalam jumlah besar untuk proses pelatihannya. *Text-to-speech* (TTS) berbasis DNN telah memberikan hasil baik pada domain *high-resource language* (HRL), namun masih mengalami masalah kelangkaan data pada domain *low-resource language* (LRL). Penelitian ini mengusulkan strategi untuk mengembangkan tiga kelompok model TTS berbasis DNN pada domain LRL, yaitu: 1) *monolingual single-speaker* (MoSS) TTS, 2) *multilingual multi-speaker* (MLMS) TTS, dan 3) *zero-shot* MLMS TTS yang dapat menangani *zero-shot speaker adaptation*. Dalam mengembangkan ketiga kelompok model TTS tersebut, terdapat dua permasalahan, yaitu: 1) masalah kelangkaan data pada LRL dan 2) masalah performansi *zero-shot speaker adaptation* untuk model *zero-shot* MLMS TTS. Untuk untuk mengatasi masalah *low-resource*, penelitian ini menerapkan skema *hierarchical transfer learning* (HTL) dengan *partial network-based deep transfer learning* (DTL) untuk proses pelatihan model TTS pada domain LRL dengan memanfaatkan ketersediaan data yang melimpah pada domain HRL. Skema HTL merupakan pembelajaran transfer multi-tahap dan diterapkan pada model TTS dengan memanfaatkan *pre-trained* MoSS TTS yang dilatih pada HRL sebagai model sumber pertama. *Pre-trained* MoSS TTS pada HRL dilatih kembali dengan *fine-tuning* pada domain LRL menggunakan arsitektur model yang sama. Kemudian, *partial network-based* DTL diterapkan untuk mentransfer parameter *pre-trained* MoSS TTS ke MLMS TTS dan *zero-shot* MLMS TTS. Khusus pada model *zero-shot* MLMS TTS digunakan *pre-trained* d-*vector speaker encoder* yang dilatih pada HRL sebagai model sumber kedua. Untuk mengatasi permasalahan performansi *zero-shot speaker adaptation*, penelitian mengusulkan penambahan kontrol eksplisit prosodi dari contoh suara penutur target menggunakan *style encoder* untuk pengondisian TTS dan penggunaan *utterance-level speaker reconstruction loss* bersama dengan *frame-level acoustic reconstruction loss* pada saat pelatihan model TTS. Eksperimen menggunakan korpus audio yang tersedia untuk umum menunjukkan bahwa skema HTL menggunakan *partial network-based* DTL yang diusulkan mampu melatih model TTS secara efektif pada domain LRL. Model yang dilatih menggunakan skema ini berhasil menyintesis ucapan yang jelas, alami, dan tidak jauh berbeda dengan suara asli manusia; sedangkan model yang dilatih menggunakan skema biasa gagal menyintesis ucapan yang dapat dipahami. Eksperimen pada *zero-shot* MLMS TTS juga menunjukkan bahwa usulan penambahan *style encoder* dan *loss function* secara signifikan dapat meningkatkan *speaker similarity* dalam tugas *zero-shot speaker adaptation* dibandingkan dengan model *baseline*.

Kata Kunci:
hierarchical transfer learning, low-resource, multi-speaker, multilingual, partial network-based deep transfer learning, text-to-speech, zero-shot speaker adaptation.