ABSTRAK

Automatic Speech Recognition (ASR) atau pengenalan wicara otomatis merupakan sistem yang dirancang untuk mengidentifikasi dan menerjemahkan ucapan manusia menjadi teks. Perkembangan machine learning telah membawa algoritma pengenalan wicara berkembang dari arsitektur tradisional menuju deep learning. Penelitian-penelitian terdahulu banyak membahas perbandingan kinerja model akustik pada bahasa-bahasa utama, namun penelitian serupa untuk bahasa Jawa masih sangat terbatas, padahal bahasa Jawa memiliki struktur fonologis dan kosakata yang kompleks dibandingkan bahasa Indonesia.

Penelitian ini bertujuan untuk membandingkan kinerja model akustik Time Delay Neural Network (TDNN) dan Time Delay Neural Network Factorized (TDNN-F) dalam pengenalan wicara bahasa Jawa. Proses penelitian menggunakan Montreal Forced Aligner (MFA) untuk menghasilkan alignment map dan berkas TextGrid sebagai acuan pelabelan fonem per frame. Fitur akustik diekstraksi menggunakan Mel-Frequency Cepstral Coefficients (MFCC) dan direduksi dimensinya menggunakan Linear Discriminant Analysis (LDA). Pencarian kombinasi hyperparameter optimal untuk masing-masing model dilakukan menggunakan metode random search.

Evaluasi kinerja kedua model dilakukan menggunakan metrik Character Error Rate (CER) pada korpus bahasa Jawa dengan durasi tuturan aktif 17,51 jam. Hasil penelitian menunjukkan bahwa model TDNN-F memperoleh CER terbaik sebesar 33,80%, sedangkan model TDNN memperoleh CER terbaik sebesar 39,50%. Hasil uji Paired T-Test menunjukkan bahwa perbedaan kinerja antara kedua model bersifat signifikan secara statistik ($p = 0{,}000149$), dengan TDNN-F secara konsisten unggul pada seluruh sepuluh pasangan eksperimen. Hasil ini menunjukkan bahwa arsitektur TDNN-F lebih sesuai untuk pengenalan wicara bahasa Jawa pada kondisi low-resource.

Kata Kunci : *Automatic Speech Recognition,* Bahasa Jawa, *Montreal Forced Aligner, Time Delay Neural Network , Time Delay Neural Network Factorized*

ABSTRACT

Automatic Speech Recognition (ASR) is a system designed to identify and convert human speech into text. Advances in machine learning have driven speech recognition algorithms to evolve from traditional architectures toward deep learning. Prior studies have largely focused on comparing acoustic model performance for major languages, yet research on Javanese remains very limited, despite Javanese having a more complex phonological structure and vocabulary than Indonesian.

This study aims to compare the performance of the Time Delay Neural Network (TDNN) and Time Delay Neural Network Factorized (TDNN-F) acoustic models for Javanese speech recognition. The research process employed the Montreal Forced Aligner (MFA) to generate an alignment map and TextGrid files as references for frame-level phoneme labeling. Acoustic features were extracted using Mel-Frequency Cepstral Coefficients (MFCC) and dimensionally reduced using Linear Discriminant Analysis (LDA). The optimal hyperparameter combination for each model was determined using a random search method.

The performance of both models was evaluated using the Character Error Rate (CER) metric on a Javanese corpus with 17.51 hours of active speech duration. The results show that the TDNN-F model achieved the best CER of 33.80%, while the TDNN model achieved the best CER of 39.50%. A paired t-test indicated that the difference in performance between the two models was statistically significant (p = 0,000149), with TDNN-F consistently outperforming TDNN across all ten Eksperiment pairs. These findings suggest that the TDNN-F architecture is better suited for Javanese speech recognition under low-resource conditions.

Keywords: *Automatic Speech Recognition, Javanese*, *Montreal Forced Aligner, Time Delay Neural Network* , *Time Delay Neural Network Factorized*