

INTISARI

Analisis Perbandingan Model *Automatic Speech Recognition* (ASR) Whisper, Vosk, dan Wav2Vec2 pada Aplikasi Transkripsi Audio Berbasis Web

Muhammad Muflih Raihan

22/503413/SV/21502

Meningkatnya aktivitas daring seperti rapat kerja, perkuliahan, dan seminar mendorong kebutuhan teknologi *Automatic Speech Recognition* (ASR) yang dapat membantu pengguna, khususnya mahasiswa, mendokumentasikan materi secara otomatis tanpa perlu mencatat manual, sehingga dapat lebih fokus mengikuti jalannya diskusi. Penelitian ini membandingkan performa tiga model ASR, yaitu Whisper, Wav2Vec2, dan Vosk, berdasarkan akurasi dan efisiensi pemrosesan. Pengujian dilakukan menggunakan 2.620 sampel *dataset* LibriSpeech, yaitu *dataset* acuan standar untuk menguji sistem pengenalan ucapan, pada kondisi audio jernih dan audio yang diberi gangguan. Hasil menunjukkan bahwa pada kondisi audio jernih, Wav2Vec2 memperoleh akurasi tertinggi sekaligus waktu pemrosesan tercepat dengan nilai WER sebesar 3,748%. Sementara itu, pada kondisi audio yang mengandung gangguan, Whisper menghasilkan akurasi terbaik dengan nilai WER sebesar 13,889% serta menunjukkan ketahanan terhadap gangguan yang paling baik. Berdasarkan keseimbangan antara akurasi dan efisiensi, Whisper dipilih sebagai model terbaik dan diintegrasikan ke dalam aplikasi transkripsi audio berbasis web, yang diharapkan dapat membantu pengguna, khususnya mahasiswa, mendokumentasikan kegiatan berbasis audio secara efisien dan akurat.

Kata kunci: *Automatic Speech Recognition*, Whisper, Vosk, Wav2Vec2, *Word Error Rate*, *Character Error Rate*, *Real-Time Factor*, *Latency*, transkripsi audio.

A Comparative Analysis of Whisper, Vosk, and Wav2Vec2 Automatic Speech Recognition (ASR) Models for Web-Based Audio Transcription Application

Muhammad Muflih Raihan

22/503413/SV/21502

The increasing volume of online activities such as work meetings, lectures, and seminars has driven the need for Automatic Speech Recognition (ASR) technology that can help users, particularly students, automatically document material without the need for manual note-taking, allowing them to better focus on following the discussion. This study compares the performance of three ASR models, namely Whisper, Wav2Vec2, and Vosk, in terms of transcription accuracy and processing efficiency. The evaluation was conducted using 2,620 samples from the LibriSpeech dataset, a standard benchmark dataset for speech recognition systems, under both clean and noisy audio conditions. The results show that, under clean audio conditions, Wav2Vec2 achieved the highest accuracy and the fastest processing time, with a Word Error Rate (WER) of 3.748%. In contrast, under noisy audio conditions, Whisper achieved the highest accuracy with a WER of 13.889% and demonstrated the greatest robustness to noise. Based on the balance between accuracy and efficiency, Whisper was selected as the best-performing model and integrated into a web-based audio transcription application. The application is expected to help users, particularly students, efficiently and accurately document audio-based activities.

Keywords: Automatic Speech Recognition, Whisper, Vosk, Wav2Vec2, Word Error Rate, Character Error Rate, Real-Time Factor, Latency, audio transcription