ABSTRAK

Dalam era pembelajaran daring dan pekerjaan jarak jauh, konferensi video menjadi penting sebagai alat komunikasi. Namun, kualitas audio seringkali terganggu oleh faktor-faktor seperti kebisingan latar belakang dan mikrofon berkualitas rendah.

Penelitian ini berfokus pada teknik *Speech Enhancement* berbasis *Deep Learning* dengan model DeepFilterNet3 dan menganalisis kinerjanya dalam konteks konferensi video pembelajaran daring. Model ini menggunakan pendekatan *Complex Mask* (CM) untuk menyempurnakan suara dengan memfilter derau yang tidak diinginkan, dan dilatih menggunakan dataset Voicebank, Demand, dan MIT IR Survey sebagai dataset *Clean Speech*, *Noise*, dan RIR.

Hasil penelitian menunjukkan bahwa model *Self-Trained* terbaik, dicapai pada epoch 115 dengan pengujian *test loss* sebesar 1,05138, *MultiResSpecLoss* sebesar 1,02696, dan *LocalSnrLoss* sebesar 0,02442. Secara keseluruhan, dibandingkan dengan model *Pre-Trained* dan RNNoise, model *Pre-Trained* berbasis DeepFilterNet3 menunjukkan kinerja unggul dalam metrik akurasi seperti PESQ, CSIG, CBAK, COVL, STOI, SiSDR, dan SegSNR. Namun, model *Self-Trained* juga menunjukkan potensi dalam meningkatkan kualitas suara dalam konferensi video untuk pembelajaran daring. Dalam metrik kecepatan, waktu respon, dan RTF, RNNoise memiliki kecepatan yang lebih tinggi dengan nilai RTFavg sebesar 0,001. Kedua versi DeepFilterNet3 memiliki RTF_{avg} masing-masing sebesar 0,081 dan 0,088. Meskipun kedua versi model DeepFilterNet3 memiliki RTF yang lebih lambat dibandingkan dengan RNNoise, kompleksitasnya masih dapat diterima untuk aplikasi tertentu.

Kata kunci: *Speech Enhancement*, DeepFilterNet3, konferensi video, pembelajaran daring, metrik akurasi, metrik kecepatan.

ABSTRACT

In the era of online learning and remote work, video conferencing has become important as a communication tool. However, audio quality is often compromised by factors such as background noise and low quality microphones.

This study focuses on the Deep Learning-based Speech Enhancement technique with the DeepFilterNet3 model and analyzes its performance in the context of online learning video conferencing. This model uses the Complex Mask (CM) approach to enhance speech by filtering out unwanted noise, and is trained using the Voicebank, Demand, and MIT IR Survey datasets as the Clean Speech, Noise, and RIR datasets.

The results showed that the best Self-Trained model was achieved at epoch 115 with a test loss of 1.05138, MultiResSpecLoss of 1.02696, and LocalSnrLoss of 0.02442. Overall, compared to the Pre-Trained and RNNoise models, the DeepFilterNet3-based Pre-Trained models show superior performance in accuracy metrics such as PESQ, CSIG, CBAK, COVL, STOI, SiSDR, and SegSNR. However, the Self-Trained model also shows potential in improving voice quality in video conferencing for online learning. In speed, response time, and RTF metrics, RNNoise has a higher speed with an RTFavg value of 0.001. Both versions of DeepFilterNet3 have an RTFavg of 0.081 and 0.088. Although both versions of the DeepFilterNet3 model have a slower RTF compared to RNNoise, their complexity is still acceptable for certain applications.

***Keywords**: Speech Enhancement, DeepFilterNet3, video conferencing, online learning, accuracy metrics, speed metrics.*