INTISARI

DIARISASI PEMBICARA RINGAN BERBASIS *END-TO-END* DENGAN *GRAPH EMBEDDING* DAN *ENCODER-DECODER ATTRACTOR* BERBASIS *LEGENDRE MEMORY UNIT* PADA AUDIO BERBAHASA INDONESIA

Oleh

Waffiq Maaroja

23/530204/PPA/06718

Mayoritas model diarisasi pembicara dilatih pada bahasa dengan sumber data yang melimpah, sehingga mengakibatkan kinerja yang kurang baik pada bahasa dengan sumber data rendah seperti bahasa Indonesia. Selain itu, banyak sistem bergantung pada RNN sekuensial (seperti LSTM) pada modul *backend*, yang menimbulkan hambatan komputasi untuk aplikasi *real-time*.

Penelitian ini mengusulkan modul diarisasi pembicara *end-to-end* (backend) yang efisien untuk mengatasi tantangan tersebut. Modul *Embedding Refiner* berbasis *Graph Attention Network* (GAT) digunakan untuk meningkatkan diskriminasi pembicara. Selain itu, penelitian ini menggantikan *attractor* berbasis LSTM tradisional dengan *Legendre Memory Unit* (LMU) yang dapat diparalelkan untuk rangkuman konteks yang efisien. Penelitian ini juga mengembangkan dataset diarisasi simulasi bahasa Indonesia.

Studi ablasi menunjukkan bahwa *refiner* GAT sangat penting untuk kinerja dan *attractor* berbasis LMU secara signifikan lebih akurat serta lebih dari 10x lebih cepat dibandingkan dengan LSTM/GRU pada level komponen. Meskipun sistem masih menggunakan model *pretrained* untuk ekstraksi *embedding* awal, komponen *backend* yang diusulkan (GAT-EEND dengan EDA berbasis LMU) terbukti 41,4% lebih ringan, 38,5% lebih cepat, dengan pengurangan 26 poin dalam DER pada audio 10 pembicara dibandingkan dengan model *benchmark* SA-EEND.

Kata Kunci: Diarisasi pembicara, *end-to-end, deep learning,* GNN, LMU, *low-resource language,* Bahasa Indonesia

ABSTRACT

LIGHTWEIGHT END-TO-END SPEAKER DIARIZATION WITH GRAPH EMBEDDING AND ENCODER-DECODER ATTRACTOR BASED ON LEGENDRE MEMORY UNIT FOR INDONESIAN-LANGUAGE AUDIO

By

Waffiq Maaroja

23/530204/PPA/06718

The majority of speaker diarization models are trained on high-resource languages, leading to suboptimal performance on low-resource languages such as Indonesian. Furthermore, many systems rely on sequential RNNs (such as LSTM) in the backend module, creating computational bottlenecks for real-time applications.

This research proposes an efficient end-to-end speaker diarization (backend) module to address these challenges. A Graph Attention Network (GAT)-based Embedding Refiner module is employed to enhance speaker discriminability. Additionally, this study replaces the traditional LSTM-based attractor with a parallelizable Legendre Memory Unit (LMU) for efficient context summarization. This research also develops a simulated Indonesian diarization dataset.

Ablation studies demonstrate that the GAT refiner is critical for performance, while the LMU-based attractor is significantly more accurate and over 10x faster than LSTM/GRU at the component level. Although the system still utilizes a pretrained model for initial embedding extraction, the proposed backend component (GAT-EEND with LMU-based EDA) proved to be 41.4% lighter and 38.5% faster, with a 26-point reduction in DER on 10-speaker audio compared to the SA-EEND benchmark model.

Keywords: speaker diarization, end-to-end, deep learning, GNN, LMU, low-resource language, Bahasa Indonesia