ABSTRACT

Speech Emotion Recognition (SER) is a technology that enables systems to recognize a speaker's emotional state based on speech signals. The development of SER for the Indonesian language still faces several challenges, including the limited availability of local datasets, the use of single acoustic features that are unable to comprehensively represent emotional characteristics, and evaluation methods that do not adequately measure the model's generalization capability to unseen speakers. This study aims to implement the early fusion of Mel-Frequency Cepstral Coefficients (MFCC), Hilbert Multispectrum, and Cochleagram acoustic features within a CNN-BiLSTM architecture enhanced with a Temporal Attention mechanism to improve feature representation and emotion classification performance for Indonesian speech. The study employs the IndoWaveSentiment dataset, which consists of five emotion categories: neutral, happy, surprised, disgusted, and disappointed. The research methodology includes audio preprocessing, data augmentation using additive Gaussian noise, time stretching, pitch shifting, time shifting, and speed perturbation, acoustic feature extraction, early feature fusion through feature concatenation, CNN-BiLSTM model training with Temporal Attention, and model evaluation using the Leave-One-Speaker-Out (LOSO) Cross-Validation method. The experimental results demonstrate that the proposed model achieved an accuracy of 73.33%, a precision of 76.90%, a recall of 73.33%, a macro F1-score of 71.42%, and a macro AUC of 94.40%. Compared with the baseline MFCC-BiLSTM model, which achieved an accuracy of 65.67%, the proposed approach improved classification performance by 7.66%. These findings indicate that combining complementary acoustic features through early fusion and incorporating the Temporal Attention mechanism enhances emotional feature representation and the model's generalization capability in speaker-independent scenarios, making it a promising approach for developing Indonesian Speech Emotion Recognition systems.

Keywords: CNN-BiLSTM, Cochleagram, Early Fusion, Hilbert Multispectrum, Leave-One-Speaker-Out, MFCC, Speech Emotion Recognition, Temporal Attention.

ABSTRAK

Speech Emotion Recognition (SER) merupakan teknologi yang memungkinkan sistem mengenali emosi pembicara berdasarkan sinyal suara. Pengembangan SER untuk bahasa Indonesia masih menghadapi tantangan akibat keterbatasan dataset lokal, penggunaan fitur akustik tunggal yang belum mampu merepresentasikan karakteristik emosi secara komprehensif, serta evaluasi yang belum sepenuhnya mengukur kemampuan generalisasi model terhadap pembicara baru. Penelitian ini bertujuan menerapkan *early fusion* fitur akustik *Mel-Frequency Cepstral Coefficients* (MFCC), *Hilbert Multispectrum*, dan *Cochleagram* pada arsitektur CNN-BiLSTM dengan *Temporal Attention* untuk meningkatkan representasi fitur dan performa klasifikasi emosi ujaran berbahasa Indonesia. Penelitian menggunakan dataset IndoWaveSentiment yang terdiri atas lima kategori emosi, yaitu netral, senang, terkejut, jijik, dan kecewa. Tahapan penelitian meliputi *preprocessing* audio, augmentasi data menggunakan *additive Gaussian noise*, *time stretching*, *pitch shifting*, *time shifting*, dan *speed perturbation*, ekstraksi fitur akustik, *early fusion* melalui konkatenasi fitur, pelatihan model CNN-BiLSTM dengan *Temporal Attention*, serta evaluasi menggunakan metode *Leave-One-Speaker-Out* (LOSO) *Cross Validation*. Hasil penelitian menunjukkan bahwa model yang diusulkan memperoleh *accuracy* sebesar 73,33%, *precision* sebesar 76,90%, *recall* sebesar 73,33%, *macro F1-score* sebesar 71,42%, dan *macro AUC* sebesar 94,40%. Dibandingkan model dasar MFCC-BiLSTM dengan *accuracy* sebesar 65,67%, pendekatan yang diusulkan mampu meningkatkan performa klasifikasi sebesar 7,66%. Hasil tersebut menunjukkan bahwa kombinasi fitur akustik melalui *early fusion* serta penerapan *Temporal Attention* mampu meningkatkan representasi emosi dan kemampuan generalisasi model pada skenario *speaker-independent*, sehingga berpotensi menjadi pendekatan yang efektif dalam pengembangan sistem *Speech Emotion Recognition* berbahasa Indonesia.

Keywords: CNN-BiLSTM, *Cochleagram*, *Early Fusion*, *Hilbert Multispectrum*, *Leave-One-Speaker-Out,* MFCC, *Speech Emotion Recognition*, *Temporal Attention*.