ABSTRACT

Emotion is a critical aspect that remains underexplored in facilitating richer and more natural human-computer interactions. The expression of emotion in speech signals presents a complex challenge for analysis, and the application of identical methodologies across different languages often yields varying performance outcomes. In the context of Indonesian Speech Emotion Recognition (SER), there is a clear need for advancements, particularly in recognizing more specific emotions or expanding the range of emotion classes. This study focuses on analyzing the nonverbal components of speech signals to enhance emotion recognition performance in Indonesian speech. The analysis of nonverbal elements is crucial, as emotions are frequently more readily identified through variations in acoustic features, such as pitch and rhythm, rather than through spoken words.

In this research, Mel Frequency Cepstral Coefficients (MFCC) are employed as a feature extraction technique to represent the nonverbal components of speech. By applying adjustments to MFCC coefficients and utilizing five distinct windowing techniques, we derived more specific MFCC features, referred to as modified MFCC. These features were subsequently analyzed for the identification and classification of emotions from speech signals. Emotion identification and classification were conducted using two approaches: Dynamic Time Warping (DTW), a distance-based method, and a one-layer Long Short-Term Memory (LSTM) model, representing a machine learning approach. The classification of six emotion classes using DTW achieved a maximum F measure 0.858, while the LSTM model reached highest F measure 0.773. Both results indicate significant improvements compared to previous studies, which reported F measure 0.582 for six emotion classes using LSTM.

Keywords : Speech Emotion Recognition (SER), nonverbal components, Mel Frequency Cepstral Coefficients (MFCC), Dynamic Time Warping (DTW), Long Sort Term Memory (LSTM), Indonesian speech, accuracy

INTISARI

Emosi adalah aspek yang belum sepenuhnya dieksplorasi untuk memberikan interaksi yang lebih kaya dan lebih alamiah antara manusia dan komputer. Ekspresi emosi dalam sebuah isyarat tutur merupakan hal yang kompleks untuk dianalisis. Penggunaan metode yang sama pada ragam bahasa yang berbeda mungkin menghasilkan performa yang berbeda. Penelitian pengenalan emosi isyarat tutur atau *Speech Emotion Recognition* (SER) bahasa Indonesia, perlu ditingkatkan terlebih untuk emosi yang lebih spesifik atau kelas emosi yang lebih banyak. Analisis komponen nonverbal dari isyarat tutur menjadi fokus penelitian ini untuk meningkatkan kinerja pengenalan emosi isyarat tutur berbahasa Indonesia. Analisis komponen nonverbal menjadi penting karena emosi seringkali lebih mudah dikenali melalui perubahan fitur akustik seperti nada dan ritme daripada dari kata-kata yang diucapkan.

Mel Frequency Cepstral Coefficients (MFCC) sebagai salah satu metode ekstraksi fitur, digunakan sebagai representasi komponen nonverbal dari isyarat tutur. Nilai-nilai koefisien pada MFCC diatur dan lima jenis *windowing* digunakan untuk mendapatkan fitur MFCC yang lebih spesifik (*modified* MFCC). Fitur-fitur *modified* MFCC yang didapatkan selanjutnya dianalisis untuk melakukan identifikasi dan klasifikasi emosi dari isyarat tutur. Identifikasi dan klasifikasi dilakukan menggunakan algoritma *Dynamic Time Warping* (DTW) sebagai metode dengan pendekatan jarak dan model *Long-Sort Term Memory* (LSTM) satu lapisan sebagai metode dengan pendekatan *machine learning*. Klasifikasi untuk enam kelas emosi menggunakan DTW mendapatkan Nilai F tertinggi 0,858 sedangkan menggunakan LSTM mendapat nilai F tertinggi 0,773. Keduanya menunjukkan peningkatan dari penelitian sebelumnya dengan nilai F 0,582 untuk enam kelas emosi dengan LSTM.

Kata kunci– pengenalan emosi isyarat tutur, komponen nonverbal, Mel Frequency Cepstral Coefficients (MFCC), Dynamic Time Warping (DTW), Long Sort Term Memory (LSTM), isyarat tutur bahasa Indonesia