Perancangan Model Pengenalan Emosi pada Percakapan Berbahasa Indonesia dengan Ekstraksi Fitur Mel dan Metode Transfer Learning = Design of Speech Emotion Recognition Model for Indonesian Language with Mel Features and Transfer Learning Methods

Martin Hizkia Parasi, author

Abstrak

Perkembangan teknologi pemrosesan ucapan sangat pesat akhir-akhir ini. Namun, fokus penelitian dalam Bahasa Indonesia masih terbilang sedikit, walaupun manfaat dan benefit yang dapat diperoleh sangat banyak dari pengembangan tersebut. Hal tersebut yang melatarbelakangi dilakukan penelitian ini. Pada penelitian ini digunakan model transfer learning (Inception dan ResNet) dan CNN untuk melakukan prediksi emosi terhadap suara manusia berbahasa Indonesia. Kumpulan data yang digunakan dalam penelitian ini, diperoleh dari berbagai film dalam Bahasa Indonesia. Film-film tersebut dipotong menjadi potongan yang lebih kecil dan dilakukan dua metode ekstraksi fitur dari potongan audio tersebut. Ekstraksi fitur yang digunakan adalah Mel-Spectrogram dan Mel Frequency Cepstral Coefficient (MFCC). Data yang diperoleh dari kedua ekstraksi fitur tersebut dilatih pada tiga model yang digunakan (Inception, ResNet, serta CNN). Dari percobaan yang telah dilakukan, didapatkan bahwa model ResNet memiliki performa yang lebih baik dibanding Inception dan CNN, dengan rata-rata akurasi 49%. Pelatihan model menggunakan hyperparameter dengan batch size sebesar 16 dan dropout (0,2 untuk Mel-Spectrogram dan 0,4 untuk MFCC) demi mendapatkan performa terbaik. <hr /> Speech processing technology advancement has been snowballing for these several years. Nevertheless, research in the Indonesian language can be counted to be little compared to other technology research. Because of that, this research was done. In this research, the transfer learning models, focused on Inception and ResNet, were used to do the speech emotion recognition prediction based on human speech in the Indonesian language. The dataset that is used in this research was collected manually from several films and movies in Indonesian. The films were cut into several smaller parts and were extracted using the Mel-Spectrogram and Mel-frequency Cepstrum Coefficient (MFCC) feature extraction. The data, which is consist of the picture of Mel-spectrogram and MFCC, was trained on the models followed by testing. Based on the experiments done, the ResNet model has better accuracy and performance compared to the Inception and simple CNN, with 49% of accuracy. The experiments also showed that the best hyperparameter for this type of training is 16 batch size, 0.2 dropout sizes for Mel-spectrogram feature extraction, and 0.4 dropout sizes for MFCC to get the best performance out of the model used.