Rancang Bangun Sistem Automatic Speech Recognition untuk Bahasa Indonesia Berbasis Wav2Letter dengan Loss Function CTC dan ASG = Development of Automatic Speech Recognition System for Indonesian Language Based on Wav2Letter with Loss Function CTC and ASG

Arief Saferman, author

Abstrak

Selama masa pandemi COVID-19, teknologi Automatic Speech Recognition (ASR) menjadi salah satu fitur yang sering digunakan pada komputer untuk mencatat di kelas online secara realtime. Teknologi ini akan bekerja dimana setiap suara yang muncul akan langsung dikenali dan dicatat pada halaman terminal. Dalam penelitian ini, model ASR Wav2Letter akan digunakan menggunakan CNN (Convolution Neural Network) dengan loss function CTC (Connectionist Temporal Classification) dan ASG (Auto Segmentation Criterion). Selama proses pembuatannya, berbagai hyperparameter acoustic model dan language model dari model ASR Wav2Letter terkait dengan implementasi batch normalization¸ learning-rate, window type, window size, n-gram language model, dan konten language model diuji pengaruh variasinya terhadap performa model Wav2Letter. Dari pengujian tersebut, ditemukan bahwa model ASR Wav2Letter menunjukkan performa paling baik ketika acoustic model menggunakan metode ASG dengan learning-rate 9 × 105 , window size 0.1, window type Blackman, serta 6-gram language model. Berdasarkan hasil akurasi WER CTC unggul 1,2% dengan 40,36% berbanding 42,11% dibandingkan ASG, namun jika dilihat lamanya epoch dan ukuran file model, loss function ASG memiliki keunggulan hampir dua kalinya CTC, dimana ASG hanya membutuhkan setengah dari jumlah epoch yang dibutuhkan oleh CTC yakni 24 epoch berbanding dengan 12 epoch dan ukuran file model ASG setengah lebih kecil dibandingkan CTC yakni 855,2 MB berbanding dengan 427,8 MB. Pada pengujian terakhir, model ASR Wav2Letter dengan loss function ASG mendapatkan hasil terbaik dengan nilai WER 29,30%. Berdasarkan hasil tersebut, model ASR Wav2Letter dengan loss function ASG menunjukkan perfoma yang lebih baik dibandingkan dengan CTC. <hr /> During the COVID-19 pandemic, Automatic Speech Recognition technology (ASR) became one of features that most widely used in computer to note down online class in real-time. This technology works by writing down every word in terminal from voice that is recognized by the system. ASR Wav2Letter model will use CNN (Convolutional Neural Network) with loss function CTC (Connectionist Temporal Classification) and ASG (Auto Segmentation Criterion). While developing Wav2Letter, various hyperparameter from acoustic model and language model is implemented such as batch normalization, learning rate, window type, window size, n-gram language model, and the content of language model are examined against the performance of Wav2Letter model. Based on those examination, Wav2Letter shows best performance when it uses ASG loss function learning rate 9 × 105 , window size 0.1, window type Blackman, and 6-gram language model. With that configuration, WER of CTC outplay ASG around 1.2% with 40.36% compare to 42,11%, but another parameter shows ASG are way more superior than CTC with less time epoch training which are 24 epoch for CTC against 12 epoch for ASG and the size of memory model shows CTC has bigger size than ASG with 855.2 MB against 427.8 MB. In the last test, ASR Wav2Letter model with ASG loss function get the best WER value around 29.3%. Based on those results, ASR Wav2Letter Model shows its best performance with ASG loss function than CTC.