Laporan Speech Recognition

Speech To Text

Kelompok 8

Disusun oleh:

Calvin Lim	Ketua	2502015762	Binusian 2025
Michael Vic Chow	Anggota	2502037454	Binusian 2025
Kenichi Halim	Anggota	2502039163	Binusian 2025

UNIVERSITAS BINA NUSANTARA KEMANGGISAN
2023

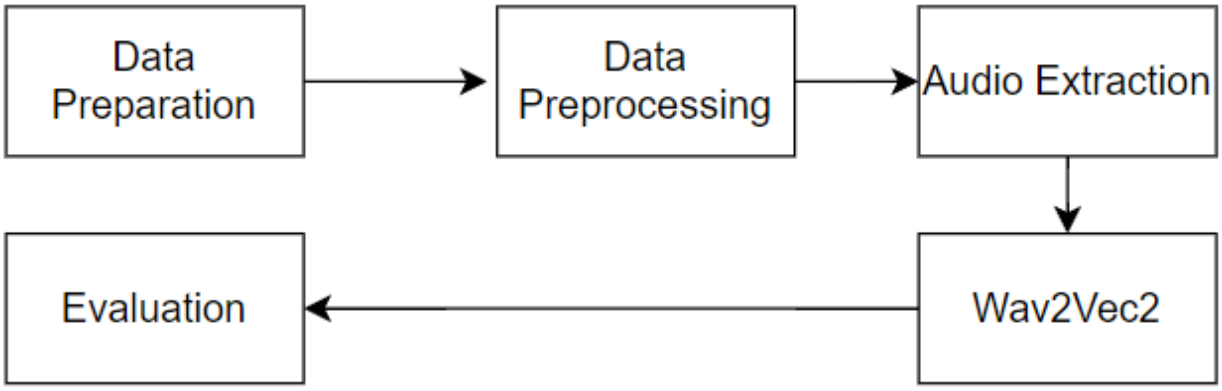
Table of Content

Bab 1 Introduction

Speech recognition, juga dikenal sebagai Speech to Text (STT), adalah teknologi yang memungkinkan komputer atau sistem untuk mengubah ucapan manusia menjadi teks tertulis. Dengan bantuan teknologi ini, pengguna dapat berkomunikasi dengan perangkat elektronik, seperti komputer, smartphone, atau perangkat lainnya, menggunakan suara mereka sebagai input. Speech recognition telah menjadi bagian integral dari berbagai aplikasi dan layanan, termasuk sistem pengenalan suara, asisten virtual, penerjemah bahasa, dan banyak lagi.

Speech recognition telah mengalami perkembangan pesat sejak diperkenalkan pertama kali. Awalnya, teknologi ini terbatas dalam hal akurasi dan keterbatasan model bahasa yang digunakan. Namun, dengan kemajuan teknologi dan metode pemodelan bahasa yang lebih canggih, speech recognition telah mencapai tingkat akurasi yang signifikan dalam beberapa tahun terakhir. Salah satu metode yang digunakan dalam implementasi Speech to Text adalah wav2vec2. Metode ini menggunakan pendekatan deep learning dan telah menunjukkan hasil yang mengesankan dalam pengenalan ucapan. Pendekatan wav2vec2 menggabungkan pembelajaran pemrograman ulang (self-supervised learning) dengan model bahasa yang canggih untuk meningkatkan akurasi dan kinerja pengenalan ucapan. Pada laporan ini kami akan menganalisa performa metode Wav2Vec mengenai speech recognition, khususnya dalam konteks speech to text. Kami akan menjelaskan konsep dasar, metode, serta evaluasi yang mengenai tingkat akurasi yang akan kami jelaskan.

Bab 2. Methodology



Gambar 1. Tahap-tahap Experiment

Proses Experiment yang dilakukan sesuai pada Gambar 1, dimulai dari pengumpulan dataset, Data Preprocessing untuk meningkatkan tingkat akurasi untuk evaluation dengan cara memperbaiki dataset, Audio Extraction supaya dapat digunakan oleh model, model Wav2Vec2

untuk konversi audio menjadi text (Speech to Text) dengan nama transcription, terakhir melakukan evaluasi model menggunakan Word Error Rate (WER) untuk mengukur akurasi transcription text

2.1 Dataset

Pada Experiment ini, kami menggunakan dataset “The LJ Speech Dataset” yang akan dijadikan untuk menganalisa hasil evaluasi Speech To Text model Wav2Vec2. Dataset mengandung sebesar 13,100 short audio clip dengan lama waktu setiap clip berjangka 1 hingga 10 detik dan total lama waktunya sekitar 24 jam sebagai sumber dataset hasil eksperimen dengan total size sebesar 3 GB, beserta dataset dalam bentuk format csv bernama metadata yang memiliki 3 variabel. Dataset audio akan diread dan membuat column baru bernama file_path audio ke file metadata untuk mempermudah mengakses audio. Dataset file metadata memiliki 4 variable dan dapat diperhatikan pada tabel 1 di bawah ini.

Variable	**Description**	**Value**
File	Nama File	Name
Text	Text audio versi 1	String
Text2	Text audio versi 2	String
Path	Lokasi file audio	String

Table 1**.** Variabel yang terdapat pada Dataset

2.2 Data Preprocessing

Data mentah dapat mengurangi hasil akhir kualitas prediksi, baik karena data yang kosong dan data rusak. Maka dari itu, agar data yang digunakan tidak menyebabkan kecacatan dalam model diperlukan melakukan data preprocessing. Tahap-tahap yang dilakukan dibagi ke dalam beberapa segmen, yaitu:

2.2.1 Feature Selection

Feature Selection merupakan tahap untuk mengurangi parameters yang tidak digunakan pada pembelajaran model. Parameters yang dikurangi adalah Variable File dan Text2.

2.2.2 Removing Special Characters

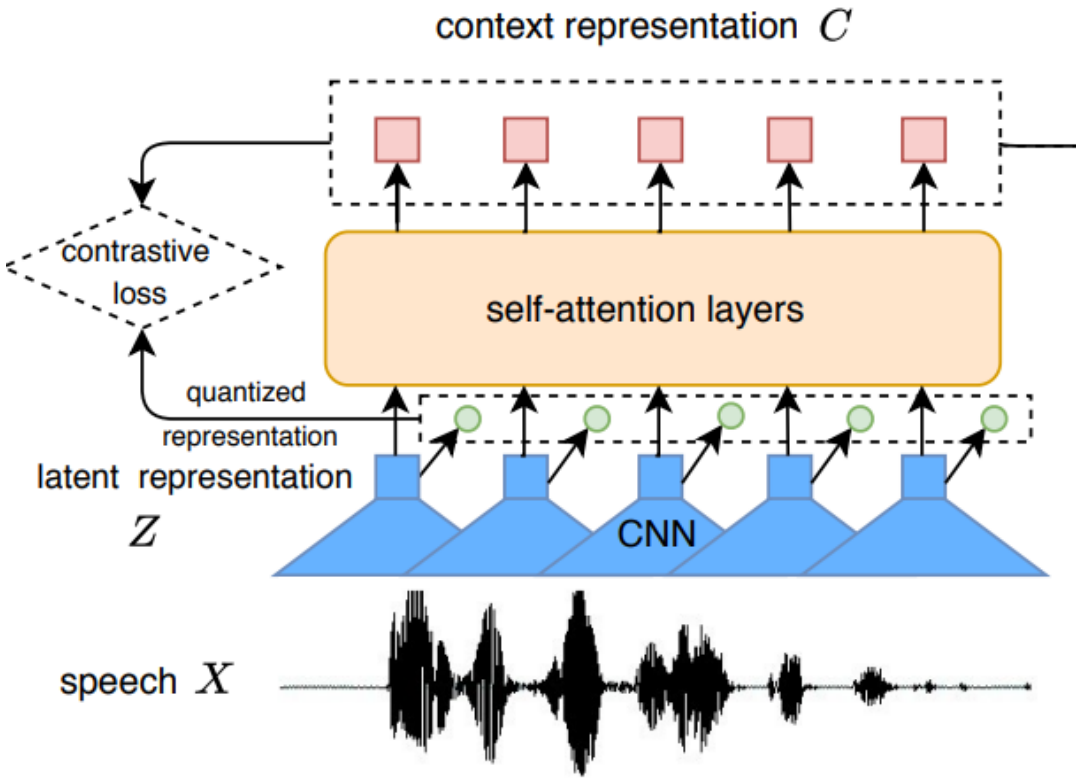
Merupakan tahap untuk menghilangkan huruf spesial seperti tanda tanya ‘?’, tanda seru ‘!’, tanda kurung’()’ dan yang lainnya. Karena transcription yang diberikan tidak bisa mengeluarkan huruf spesial sehingga perlu di hilang huruf spesial pada dataset upaya hasil evaluasi tidak diinterupsi oleh data sampah.

2.3 Audio Extraction

Untuk menggunakan Model, file audio diperlukan mengekstrak audio tersebut menggunakan library Librosa yang merupakan salah satu fitur ekstraksi yaitu MFCC (Mel Frequency Cepstral Coefficients). Fitur ekstraksi mengembalikan dua nilai, nilai pertama audio berisi data audio dalam bentuk array numpy berisi representasi numerik dimana setiap elemen array mewakili amplitudo suara pada waktu tertentu, nilai kedua sample_rate berisi kecepatan sample audio yaitu jumlah sample yang diambil dalam satu detik audio, nilai ini digunakan untuk menginterpretasikan waktu dengan benar saat bekerja dengan audio. Sample_rate yang kita gunakan sebesar 16000 Hz karena model Wav2Vec model telah di pre-train pada 16000 Hz frequency.

2.4 Wav2Vec2

ASR Model wav2cec 2.0 telah membuktikan hasil sukses di test Standar Librispeech benchmark dengan membatasi label data dan menggunakan unlabeled data yang sangat banyak untuk di pre-trainkan



Gambar 2. Struktur Wav2vec2.0

Seperti yang ditunjukkan dalam Gambar 2, wav2vec2.0 terdiri dari beberapa convolution layers dan self-attention layers. Struktur ini secara luas digunakan dalam model ASR end-to-end terkini. Convolution layer melakukan down-sample pada speech X dan menghasilkan representasi laten yang lebih terkompresi Z. Secara khusus, Z mewakili sinyal raw audio X yang diambil sampelnya dengan kecepatan 16k dengan jarak sekitar 20ms dan jangkauan penerimaan sebesar 25ms. Self-attention layers membangun representasi kontekstual C dan menangkap konten tingkat tinggi dari input Z. Kemampuan pemodelan dependensi konteks yang kuat mereka memberdayakan model untuk membuat pilihan yang tepat selama pelatihan kontrastif berikutnya dengan menggunakan Z yang tersembunyi.

2.5 Evaluation

Ketika mengevaluasi model Wav2Vec, word error rate (WER) akan digunakan. WER menghitung dengan cara melihat seberapa banyak kata yang salah di transcript oleh model dibandingkan dengan data text sebenarnya. WER dihitung seperti yang ditunjukan di Equation 1

$$WER = \frac{I + D + S}{N}$$

Equation 1. WER

Dimana semua kata akan ditambahkan bersama dalam pembilang yang ditranskripsi dengan tidak benar (Insertion I), sepenuhnya dihapus (Deletion D), atau digantikan dengan kata yang salah (Substitution S). angka ini akan dibagi dengan total keseluruhan kata sebenarnya (N). WER akan dihitung menggunakan function yang tersedia di library torchmetrics

```
average = result['wer'].mean()
print(f'WORD ERROR RATE = {average}%')
```

```
WORD ERROR RATE = 9.255954198473283%
```

Gambar 3. Hasil WER

Seperti yang ditunjukan pada gambar 3. Nilai rata-rata WER sebesar 9.25% sangatlah bagus karena memberitahukan kita bahwa kata error muncul 9 kali setiap 100 kata, sehingga dapat disimpulkan semakin kecil nilai WER mendekati 0 maka semakin bagus tingkat akurasi Speech to Text.

3. Conclusion

Pada hasil penelitian Dataset “The LJ Speech Dataset”, penggunaan metode Wav2Vec2 dapat menghasilkan hasil transcription Speech to Text sangat bagus. Wav2Vec menyediakan pendekatan inovatif dan efektif untuk ASR Speech to Text dengan memanfaatkan representasi audio yang ditingkatkan dan pelatihan dari pre-trained kontrastif. Hal ini telah membantu dalam meningkatkan kualitas transkripsi dan keakuratan sistem ASR pada berbagai aplikas iyang memerlukan konversi audio menjadi teks, namun model Wav2Vec memiliki arsitektur yang kompleks dan membutuhkan konsumsi memori dan komputasi yang tinggi, Wav2Vec juga sensitif terhadap kualitas audio yang rendah sehingga kebisingan latar belakang dapat memberikan kualitas atau performa model yang menurun dan menghasilkan transkripsi yang tidak akurat. Dalam mempertimbangkan penggunaan Wav2Vec, penting untuk memahami kekurangan-kekurangan ini dan menyesuaikan penggunaan sesuai dengan kebutuhan dan kendala yang ada.

Link Kodingan : https://www.kaggle.com/code/mifuniree/stt-wav2vec2/edit