BAB II
TINJAUAN PUSTAKA

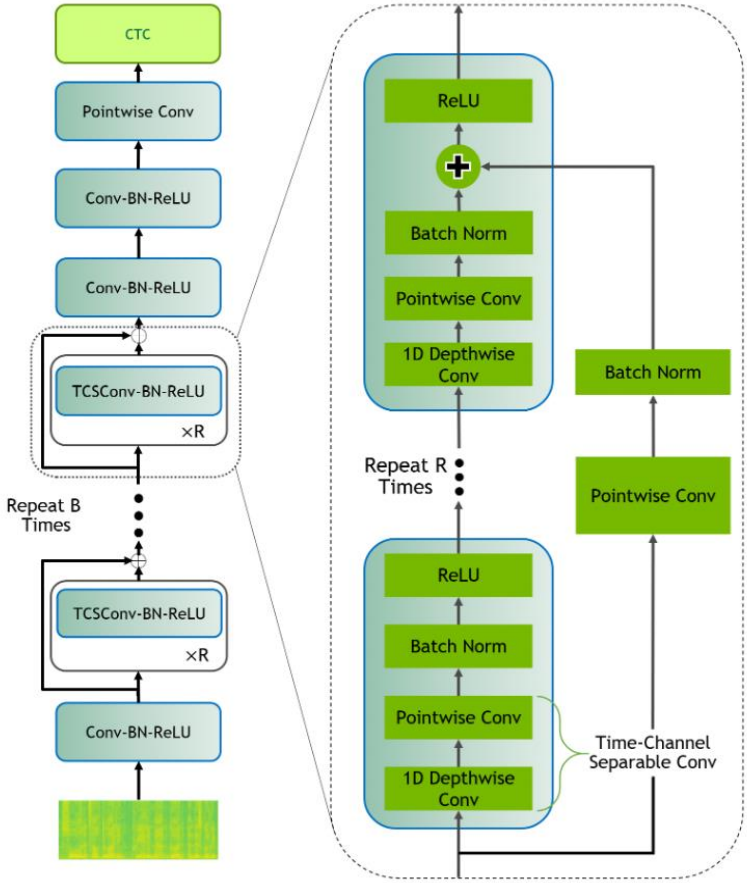
2.1. Penelitian Terdahulu

Terdapat beberapa penelitian terkait mengenai implementasi *speech recognition* yang menjadi referensi dalam pemilihan arsitektur model dan metode pemrosesan data untuk penelitian yang diselenggarakan.

2.1.1. Penelitian berjudul *"Performance vs. Hardware Requirements in state-of-the-art Automatic Speech Recognition"* [8] oleh Alexandru-Lucian Georgescu, Alessandro Pappalardo, Horia Cucu dan Michaela Blott melakukan perbandingan kinerja dan kebutuhan perangkat keras antara 8 (delapan) arsitektur model *speech recognition.* Untuk melakukan perbandingan, masing-masing model dilatih menggunakan *dataset LibriSpeech corpus* [9] dan kinerja model diukur dengan menghitung persentase *word error rate* (WER) model pada saat pengujian. Kebutuhan perangkat keras diukur dari kebutuhan memori untuk mengolah setiap detik data. Dari arsitektur model yang dibandingkan, *QuartzNet* merupakan model yang memiliki kinerja terbaik dengan nilai WER terendah, sedangkan *Kaldi CNN-TDNN* merupakan model dengan kebutuhan memori terendah. Kedua model memiliki kompleksitas yang setara, namun *Kaldi CNN-TDNN* merupakan model *hybrid* yang terdiri dari 3 (tiga) komponen terpisah, sedangkan *QuartzNet* merupakan model *end-to-end* dan dapat digunakan tanpa komponen tambahan.

2.1.2. Penelitian berjudul *"QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions"* [10] oleh S. Kriman *et al.* merupakan sebuah perincian arsitektur model *speech recognition QuartzNet,* yang merupakan model akustik dengan

implementasi *time-channel separable convolution.* Arsitektur model diajukan sebagai solusi penyederhanaan jumlah parameter pada model berbasis *neural network* sehingga membutuhkan daya komputasi yang rendah. *Time-channel separable convolution* sendiri merupakan implementasi dari *depthwise separable convolution* yang membagi proses *convolution* dalam tahap *depthwise convolution* pada setiap *channel* dan *point-wise convolution* pada setiap *frame*. Arsitektur model terdiri dari blok *convolutional layer* C_1, dilanjutkan dengan blok B_i yang merupakan blok *time-channel separable convolutional layer* yang berulang, dan tiga blok *convolutional layer* C_2, C_3 dan C_4. Diagram arsitektur dari model *QuartzNet* ditampilkan pada gambar 2.1.



Gambar 2.1 Arsitektur Model QuartzNet[10]

Penelitian membandingkan *word error rate* pada beberapa model *speech recognition* berbasis *convolutional neural network* yang dilatih dengan *dataset LibriSpeech* dan menunjukkan bahwa model dengan arsitektur *QuartzNet* dapat mencapai nilai *word error rate* terendah walaupun memiliki jumlah parameter terendah diantara model yang diuji.

2.1.3. Penelitian berjudul *"Acoustic Analysis of Dysarthric Speech and Some Implications for Automatic Speech Recognition"* [11] oleh Tina Magnuson dan Mats Blomberg menyelidiki fitur akustik pada pengidap gangguan bicara *dysarthria* pada individu dengan gangguan motorik dengan melakukan analisis pada rekaman pembacaan teks. Fitur akustik yang dimaksud dalam penelitian merupakan kecepatan serta tingkat kejelasan artikulasi fonem dalam skala kata dan kalimat. Hasil yang didapat dari penelitian menunjukkan bahwa kecepatan rata-rata pembicara dengan gangguan bicara lebih lambat dibanding kecepatan rata-rata pada pembicara tanpa gangguan bicara. Penelitian juga menunjukkan adanya jeda antar kata yang tidak teratur dalam pembacaan kalimat oleh individu dengan gangguan bicara.

2.1.4. Penelitian berjudul *"Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition"* [7] oleh Bhavik Vachhani, Chitralekha Bhat dan Sunil Kumar Kopparapu melakukan perbandingan antara jenis augmentasi yang dapat dilakukan untuk mengubah data *normal speech* menjadi *disordered speech*. Metode augmentasi dalam penelitian berdasar analisis yang dilakukan terhadap durasi fonem pada dua *dataset disordered speech*, yaitu *UASpeech corpus* [12] dan TORGO. Terdapat dua metode augmentasi yang diajukan untuk mengubah durasi pada data *normal speech*, yaitu

speed perturbation dan *tempo perturbation.* Metode *speed perturbation* merupakan proses *re-sampling* data menggunakan sebuah nilai faktor, sedangkan metode *tempo perturbation* merupakan proses transformasi pada data sehingga memiliki durasi yang berbeda. Penelitian menggunakan nilai faktor yang berbeda untuk masing-masing metode untuk mendapatkan nilai faktor optimal untuk merepresentasikan data *disordered speech* melalui proses augmentasi dari *normal speech*. Evaluasi dilakukan dengan melatih model *speech recognition* berbasis DNN-HMM menggunakan data hasil augmentasi. Model diuji menggunakan data *disordered speech* dan ditemukan bahwa secara keseluruhan, metode *speed perturbation* merupakan teknik augmentasi yang memberikan *word error rate* terendah.

2.1.5. Seperti penelitian yang disebutkan pada 2.1.4, penelitian berjudul *"Investigation of Data Augmentation Techniques for Disordered Speech Recognition"* [13] oleh M. Geng *et al.* melakukan penyelidikan teknik augmentasi untuk mendapatkan data *disordered speech*. Penelitian menambahkan satu metode augmentasi tambahan, yaitu *vocal tract length perturbation* (VTLP). Metode VTLP merupakan proses transformasi pada data di domain frekuensi, berbeda dengan metode *speed perturbation* dan *tempo perturbation* yang dilakukan di domain waktu. Evaluasi dilakukan dengan melatih model *speech recognition* berbasis *hybrid* DNN menggunakan campuran antara data yang sudah ada dan data hasil augmentasi. Model diuji menggunakan data *disordered speech* dan ditemukan bahwa *word error rate* terendah didapatkan saat metode augmentasi *speed perturbation* diterapkan pada data *normal speech* dan *disordered speech.*

Ringkasan dari penelitian-penelitian yang digunakan sebagai dasar dari penelitian yang dilakukan disajikan dalam tabel 2.1 sebagai berikut.

Sub Bab	Judul Penelitian	Ringkasan
2.1.1.	*Performance vs. Hardware Requirements in state-of-the-art Automatic Speech Recognition* [8]	- Model berbasis CNN memiliki kinerja yang setara dengan model berbasis HMM sehingga dibutuhkan pertimbangan pada faktor-faktor seperti spesifikasi perangkat keras dan arsitektur dalam pemilihan model untuk *speech recognition.* - Model berbasis HMM menggunakan spesifikasi perangkat keras yang lebih rendah namun membutuhkan komponen *acoustic model*, *phonetic model* dan *language model* yang harus dilatih secara terpisah sehingga proses pelatihan model membutuhkan waktu yang lebih lama dibandingkan model berbasis CNN yang dapat berfungsi dengan hanya *acoustic model*. - Ukuran dan jumlah parameter pada model berbasis CNN *QuartzNet* dapat dikurangi tanpa ada peningkatan persentase *word error rate* yang signifikan dengan cara mengurangi jumlah layer yang diulang sehingga cocok digunakan untuk model yang dilatih dengan data yang terbatas.
2.1.2.	*QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions* [10]	
2.1.3.	*Acoustic Analysis of Dysarthric Speech and Some Implications for Automatic Speech Recognition* [11]	- *Normal speech* dan *disordered speech* memiliki perbedaan dari sisi kecepatan dan jeda antara masing-masing suku kata yang

2.1.4.	*Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition* [7]	diucapkan, dimana secara rata-rata, *disordered speech* memiliki kecepatan yang lebih lambat dari *normal speech* dan jeda yang lebih panjang.
2.1.5.	*Investigation of Data Augmentation Techniques for Disordered Speech Recognition* [13]	- Dari tiga metode pemrosesan sinyal audio yang diajukan sebagai cara untuk mengkonversi data *normal speech* menjadi data *disordered speech*, model yang dilatih menggunakan data yang diproses dengan *speed perturbation* menghasilkan persentase *word error rate* terendah saat diuji menggunakan data *disordered speech* yang sebenarnya sehingga *speed perturbation* dapat digunakan sebagai metode untuk meniru data *disordered speech*.

Tabel 2.1 Ringkasan Penelitian Terdahulu

2.2. Tinjauan Teori

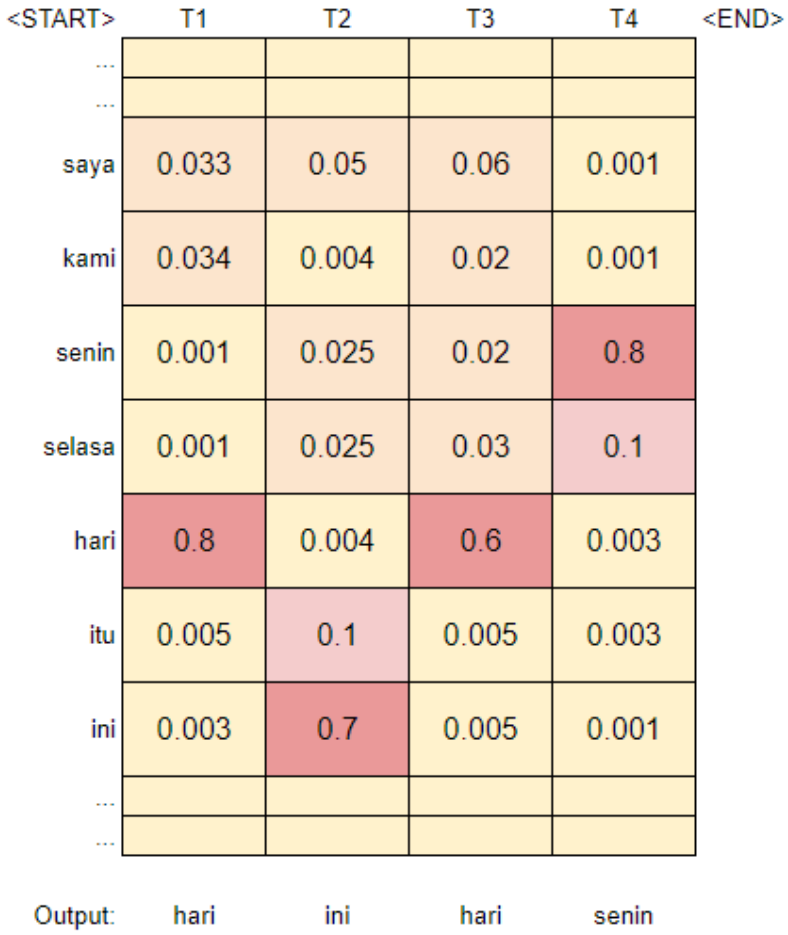
2.2.1. End-to-End Automatic Speech Recognition

End-to-end automatic speech recognition adalah salah satu jenis pendekatan dalam sistem *speech recognition* sebagai alternatif pada sistem *speech recognition* tradisional yang berbasis *Hidden Markov Model* (HMM). Secara garis besar, sistem *end-to-end* terdiri dari komponen *encoder* yang menerima input suara dan komponen *decoder* yang mengubah input suara menjadi output transkrip berupa teks. *Acoustic model* merupakan bagian utama dalam komponen *encoder* pada sistem *speech recognition* yang berfungsi untuk membedakan satuan linguistik dari sinyal audio yang ada pada input. *Acoustic model* dalam sistem *end-to-end* umumnya menggunakan model berbasis *neural network* untuk mengidentifikasi masing-masing satuan linguistik yang terdapat pada sinyal. Komponen *decoder* pada sistem *end-to-end* menerjemahkan input

dengan menggunakan sebuah algoritma *decoding* yang dapat diintegrasi dengan sebuah *language model* untuk meningkatkan akurasi pada *output* teks yang dihasilkan.

2.2.2. *Greedy Search*

Algoritma *greedy search* adalah salah satu algoritma *decoding* yang dapat digunakan untuk menerjemahkan input suara menjadi token berupa karakter maupun kata dengan cara memperhitungkan probabilitas kemunculan token dari distribusi probabilitas yang tersedia di setiap titik waktu dalam satu rangkaian. Token yang memiliki probabilitas tertinggi dipilih sebagai output dari proses penerjemahan. Algoritma *greedy search* memiliki kecepatan *decoding* yang tinggi namun memiliki kekurangan dalam tingkat akurasi akibat tidak mempertimbangkan adanya hubungan antara masing-masing token secara berurut. Algoritma *greedy search* memandang setiap token sebagai entitas yang independen sehingga berkemungkinan untuk melewatkan kombinasi token yang dapat menghasilkan probabilitas lebih tinggi pada urutan token secara keseluruhan. Ilustrasi cara kerja *greedy search* dengan token berupa kata untuk sebuah input audio sepanjang 4 satuan waktu dapat dilihat pada gambar 2.2.

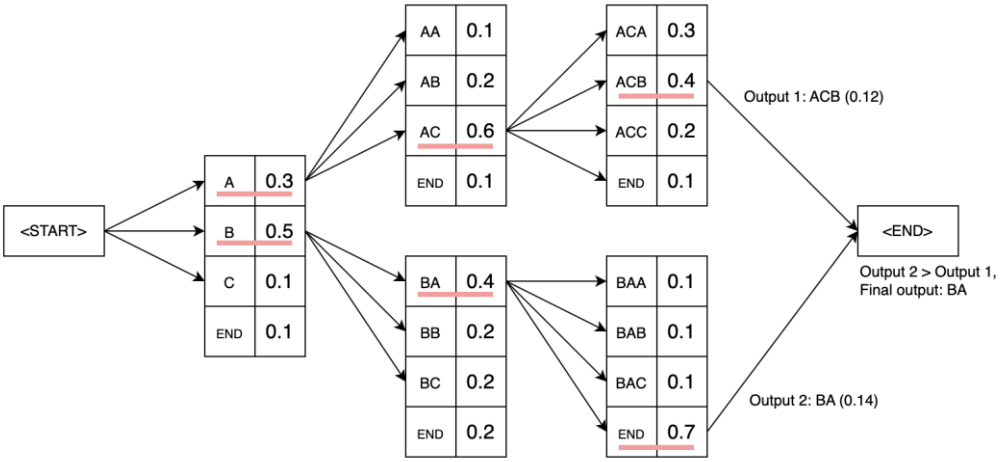


Gambar 2.2 Algoritma Greedy Search

2.2.3. *Beam Search*

Algoritma *beam search* merupakan alternatif pada algoritma *greedy search* untuk melakukan penerjemahan dengan menambahkan parameter *beam size*. *Beam size* didefinisikan sebagai jumlah token yang dipertimbangkan, contohnya untuk *beam size* bernilai 2, dua token dengan probabilitas tertinggi pada setiap titik waktu dipertimbangkan dan urutan token dengan probabilitas tertinggi dipilih sebagai hasil transkripsi. Nilai *beam size* yang lebih tinggi mampu meningkatkan tingkat akurasi dengan cara menyimpan lebih banyak probabilitas untuk dipertimbangkan namun memperlambat proses transkripsi akibat bertambahnya jumlah proses

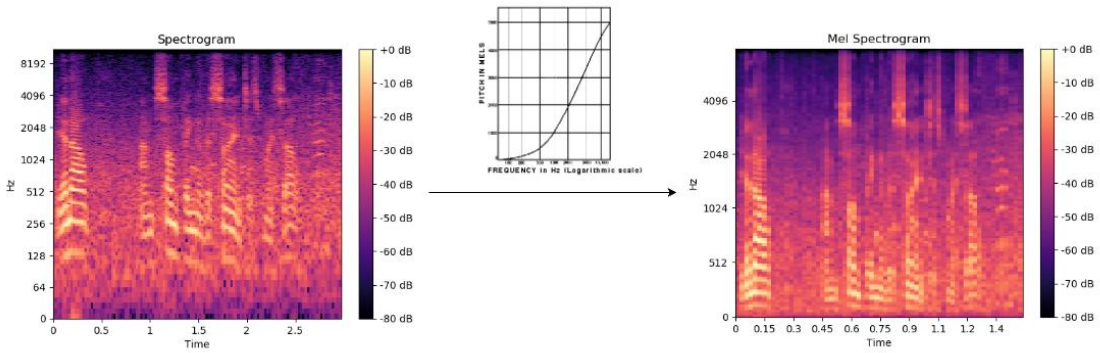
perhitungan yang dilakukan untuk mendapatkan probabilitas masing-masing urutan token. Gambar 2.3 menunjukkan cara kerja algoritma *beam search* menggunakan token berupa karakter dengan nilai *beam size* 2.



Gambar 2.3 Algoritma Beam Search

2.2.4. Mel-spectrogram

Sinyal audio dapat direpresentasikan dalam tiga dimensi dengan sebuah *spectrogram* yang menunjukkan intensitas dan frekuensi gelombang terhadap waktu. *Spectrogram* dari sebuah gelombang suara didapatkan dengan cara melakukan operasi *short-time Fourier transform* terhadap sinyal. *Mel-spectrogram* merupakan *spectrogram* dengan nilai frekuensi yang dipetakan terhadap *mel scale* sehingga jarak antar frekuensi pada gelombang suara ternormalisasi berdasarkan yang didengar telinga manusia. Tahap-tahap konversi sinyal audio menjadi *mel-spectrogram* ditampilkan pada gambar 2.4.



Gambar 2.4 Konversi Sinyal Audio Menjadi Mel-spectrogram

Mel-spectrogram mampu digunakan untuk mengidentifikasi letak dan urutan fonem pada sinyal audio akibat artikulasi masing-masing fonem yang memiliki frekuensi dan pola spesifik dalam representasi menggunakan *spectrogram.*

2.2.5. Convolutional Neural Network

Convolutional neural network (CNN) merupakan sebuah arsitektur *neural network* yang biasa digunakan dalam *computer vision.* Dalam konteks *speech recognition*, model dengan arsitektur CNN dapat digunakan untuk *acoustic modelling* dengan menerima input berupa *spectrogram* dari sinyal gelombang suara. Seperti yang dijelaskan pada bagian 2.2.4, setiap fonem yang direpresentasikan pada *spectrogram* memiliki pola spesifik sehingga dapat diklasifikasi dengan menggunakan model berbasis CNN.