BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian terdahulu mengenai penggunaan teknologi pengenalan suara otomatis untuk transkripsi telah banyak dilakukan dengan berbagai pendekatan, termasuk penggunaan model pembelajaran mendalam seperti wav2vec 2.0. Studi-studi yang relevan terkait pengembangan sistem transkripsi otomatis ini menunjukkan kemajuan dalam akurasi pengenalan suara lintas bahasa dan dialek. Dalam konteks penggunaan wav2vec 2.0 untuk aplikasi transkripsi khotbah, teknologi ini menunjukkan potensi besar untuk mentranskripsikan audio mentah menjadi teks secara otomatis dengan akurasi tinggi, terutama dalam kondisi audio yang bervariasi dan penggunaan bahasa lokal. Sub bab berikut ini akan membahas beberapa studi terkait yang menggunakan model-model jaringan saraf seperti wav2vec 2.0 dan adaptasinya terhadap kebutuhan bahasa lokal serta dialek, dengan fokus khusus pada penerapan di konteks keagamaan seperti dokumentasi khotbah di GKJW Mundusewu.

2.1.1 wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations

Penelitian Baevski dkk ini bertujuan untuk mengembangkan kerangka kerja pembelajaran mandiri (self-supervised learning) dalam pengenalan ujaran dengan memanfaatkan representasi audio mentah. Melalui pendekatan wav2vec 2.0, penelitian ini berfokus pada pengkodean ujaran dan masking pada ruang laten, yang memungkinkan pembelajaran kontekstual dan penyusunan unit ujaran secara kuantitatif. Dengan demikian, penelitian ini berupaya meningkatkan akurasi pengenalan ujaran, khususnya dalam kondisi data terbatas, dengan memperlihatkan bahwa pelatihan menggunakan data tanpa label dalam jumlah besar dapat secara signifikan mengurangi kebutuhan akan data berlabel.

Model	Unlabeled data	LM	dev		test	
			clean	other	clean	other
10 min labeled						
Discrete BERT [4]	LS-960	4-gram	15.7	24.1	16.3	25.2
BASE	LS-960	4-gram	8.9	15.7	9.1	15.6
		Transf.	6.6	13.2	6.9	12.9
LARGE	LS-960	Transf.	6.6	10.6	6.8	10.8
	LV-60k	Transf.	4.6	7.9	4.8	8.2
1h labeled						
Discrete BERT [4]	LS-960	4-gram	8.5	16.4	9.0	17.6
BASE	LS-960	4-gram	5.0	10.8	5.5	11.3
		Transf.	3.8	9.0	4.0	9.3
LARGE	LS-960	Transf.	3.8	7.1	3.9	7.6
	LV-60k	Transf.	2.9	5.4	2.9	5.8
10h labeled						
Discrete BERT [4]	LS-960	4-gram	5.3	13.2	5.9	14.1
Iter. pseudo-labeling [58]	LS-960	4-gram+Transf.	23.51	25.48	24.37	26.02
	LV-60k	4-gram+Transf.	17.00	19.34	18.03	19.92
BASE	LS-960	4-gram	3.8	9.1	4.3	9.5
		Transf.	2.9	7.4	3.2	7.8
LARGE	LS-960	Transf.	2.9	5.7	3.2	6.1
	LV-60k	Transf.	2.4	4.8	2.6	4.9
100h labeled						
Hybrid DNN/HMM [34]	-	4-gram	5.0	19.5	5.8	18.6
TTS data augm. [30]	-	LSTM			4.3	13.5
Discrete BERT [4]	LS-960	4-gram	4.0	10.9	4.5	12.1
Iter. pseudo-labeling [58]	LS-860	4-gram+Transf.	4.98	7.97	5.59	8.95
	LV-60k	4-gram+Transf.	3.19	6.14	3.72	7.11
Noisy student [42]	LS-860	LSTM	3.9	8.8	4.2	8.6
BASE	LS-960	4-gram	2.7	7.9	3.4	8.0
		Transf.	2.2	6.3	2.6	6.3
LARGE	LS-960	Transf.	2.1	4.8	2.3	5.0
	LV-60k	Transf.	1.9	4.0	2.0	4.0

Gambar 2.1 Hasil WER

2.1.2 Pengolahan Korpus Dataset Audio Bacaan Al-Qur'an Menggunakan Metode Wav2Vec 2.0

Tujuan penelitian ini adalah untuk mengembangkan dan memproses korpus dataset audio bacaan Al-Qur'an menggunakan metode Wav2Vec 2.0. Penelitian ini berfokus pada pembuatan model pengenalan suara otomatis (*Automatic Speech Recognition*/ASR) untuk bacaan Al-Qur'an dengan memanfaatkan pembelajaran mandiri (self-supervised learning) guna meningkatkan akurasi pengenalan teks dari suara. Selain itu, penelitian ini bertujuan mengatasi keterbatasan dataset audio berlabel dan menangani variasi aksen serta pelafalan pembaca dengan optimal. Diharapkan, hasil penelitian ini dapat mendukung pengembangan teknologi pembelajaran Al-Qur'an berbasis deep learning serta meningkatkan minat generasi muda dalam mempelajari Al-Qur'an melalui pemanfaatan teknologi modern.

No.	Teks Target	Teks Prediksi
1	قُلْ أَعُوذُ بِرَبِّ ٱلنَّاسِ	قُلْ أَعُوذُ بِرَبِّنَا
2	لَمْ يَكُنِ ٱلَّذِينَ كَفَرُوا۟ مِنْ أَهْلِ ٱلْكِتَٰبِ وَٱلْمُشْرِكِينَ مُنفَكِّينَ حَتَّىٰ تَأْتِيَهُمُ ٱلْبَيِّنَةُ	لَمْ يَكُنْ الَّذِينَ كَفَرُوا مِنْ أَهْلِ الْكِتَابِ وَالْمُشْرِكِينَ مُ فَكِينَ حَتَّى تَأْتِيَهُمُ الْبَيِنَ
3	لَكُمْ دِينُكُمْ وَلِيَ دِينِ	لَكُمْ دِينكُم وَلِيدِي
4	إِذَا زُلْزِلَتِ ٱلْأَرْضُ زِلْزَالَهَا	إِذَا زُئْزِلَتِ الأَرْضِ لْزَالَهَا
5	وَمِنْ ءَايَٰتِهِۦٓ أَنْ خَلَقَ لَكُم مِّنْ أَنفُسِكُمْ أَزْوَٰجًا لِّتَسْكُنُوٓا۟ إِلَيْهَا وَجَعَلَ بَيْنَكُم مَّوَدَّةً وَرَحْمَةً ۚ إِنَّ فِى ذَٰلِكَ لَءَايَٰتٍ لِّقَوْمٍ يَتَفَكَّرُونَ	وَمِنْ آيَاتِ أَنْ خَلَقَ لَكُم مِنْ أَنْفُسِكُمْ أَزْوَاجًا لِتَسْكُنُو إِلَيْهَا وَجَعَلَ بَيْنَكُمْ وَدَّةً وَرَحْمَ إِنَّ فِي ذَلِكَ لَآيَاتٍ لِقَومِإِيَّةً فَكُرُو

Gambar 2.2 Prediksi wav2vec dan contoh target

Pengujian	Akurasi	Presisi	Recall	F1	WER
1	65.52%	0.83	0.66	0.73	0.5
2	27.97%	0.30	0.28	0.29	0.75
3	41.94%	0.57	0.42	0.48	0.75
4	46.15%	0.49	0.46	0.47	1.0
5	10.94%	0.12	0.11	0.12	0.9

Gambar 2.3 Hasil prediksi wav2vec2 dengan accuracy, precision, recall, F1-score, dan WER

2.1.3 A Method Improves Speech Recognition with Contrastive Learning in Low-Resource Languages

Penelitian Sun dkk ini membahas mengenai metode peningkatan pengenalan suara otomatis, terutama dalam konteks bahasa dengan sumber daya rendah, melalui pembelajaran kontrastif. Fokus utamanya

adalah mengatasi tantangan dalam pengenalan suara otomatis, di mana sering kali diperlukan banyak data yang berlabel untuk melatih model yang efektif. Metode yang diusulkan, False Negatives Impact Elimination (FNIE), bertujuan untuk mengeliminasi dampak sampel negatif palsu dalam pembelajaran kontrastif sehingga dapat meningkatkan kualitas kumpulan sampel negatif dan hasil pengenalan suara. Penelitian ini menunjukkan bahwa metode FNIE berhasil meningkatkan akurasi pengenalan suara pada bahasa Inggris, Mandarin, Uyghur, dan Uzbek di lingkungan dengan sumber daya terbatas.

Method	K	K'	N	Mandarin *CER%*	
wav2vec 2.0	100	100	0	27.853	
FNIE	101	100	1	27.791	
	102	100	2	27.335	
	104	100	4	26.755	
Method	K	K'	N	**English**	
				CER%	*WER%*
wav2vec 2.0	100	100	0	11.753	23.998
FNIE	101	100	1	12.115	24.062
	102	100	2	11.805	23.353
	104	100	4	12.18	24.254

Gambar 2.4 Perbandingan penghapusan jumlah sampel negatif palsu yang berbeda.

2.1.4 Exploring Wav2Vec 2.0 On Speaker Verification and Language Identification

Pemanfaatan model self-supervised wav2vec 2.0 untuk dua tugas utama dalam pemrosesan suara, yaitu speaker verification (SV) dan language identification (LID), dibahas dalam penelitian yang dilakukan oleh Fan dkk. Meskipun awalnya dikembangkan untuk pengenalan suara otomatis, wav2vec 2.0 ternyata mampu menangkap informasi penting terkait identitas pembicara dan bahasa dalam tahap pre-training tanpa supervisi. Melalui visualisasi dengan t-SNE dan serangkaian eksperimen, ditemukan bahwa fitur yang dihasilkan oleh model ini memiliki kemampuan membedakan pembicara dan bahasa, terutama pada lapisan bawah transformer. Dengan melakukan fine-tuning pada dataset VoxCeleb1 untuk SV dan AP17-OLR untuk LID, model ini mencapai hasil yang kompetitif,

bahkan mencetak EER terbaik untuk SV sebesar 3.61%. Selain itu, pendekatan multi-task learning juga dieksplorasi untuk menyederhanakan proses fine-tuning dengan menggabungkan dua tugas dalam satu model, meskipun dengan sedikit penurunan performa. Hasil penelitian ini menunjukkan bahwa pre-training wav2vec 2.0 efektif tidak hanya untuk pengenalan suara, tetapi juga dapat diperluas untuk tugas-tugas lain dalam domain pemrosesan suara.

2.1.5 Methods to Optimize Wav2Vec with Language Model for Automatic Speech Recognition in Resource Constrained Environment

Penelitian Haswani dan Mohankumar membahas metode optimasi model Wav2Vec 2.0 untuk sistem pengenalan suara otomatis (ASR) di lingkungan dengan sumber daya terbatas seperti perangkat dengan CPU 2 inti dan RAM 4GB. Metode WSLR (Wav2Vec with Stride Chunking and Language Model for Resource-constrained devices) diusulkan, yang menggabungkan teknik stride chunking, integrasi language model (LM) berbasis 4-gram, dan post-processing untuk menghasilkan transkripsi yang lebih akurat dan terbaca. Dengan pendekatan ini, mereka berhasil menurunkan Word Error Rate (WER) menjadi 0.85, mengungguli model Wav2Vec standar dan model ASR lain seperti Citrinet dan QuartzNet. Teknik stride chunking yang digunakan memungkinkan pemrosesan audio panjang tanpa membuat sistem crash, dengan cara memecah audio menjadi potongan-potongan kecil yang saling tumpang tindih untuk mempertahankan konteks. Hasil eksperimen menunjukkan bahwa sistem yang diusulkan tidak hanya efisien secara komputasi, tetapi juga tetap memberikan performa tinggi pada perangkat dengan keterbatasan memori, menjadikannya cocok untuk integrasi pada perangkat pintar seperti speaker berbasis voice assistant dan aplikasi pengenalan suara offline di perangkat mobile.

2.2 Teori GKJW

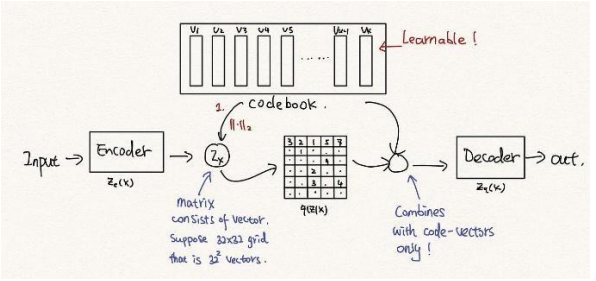
Greja Kristen Jawi Wetan (GKJW) adalah persekutuan gereja-gereja yang berbasis daerah di Jawa Timur yang dideklarasikan kali pertama pada tanggal 11 Desember 1936 di salah satu Jemaat Kristen Jawa terkemuka saat itu, yakni Mojowarno, Kabupaten Jombang. Keberadaan gereja berkenaan dengan adanya kristenisasi penduduk pribumi di Jawa pada awal abad XIX, ketika angin kebebasan beragama diembuskan oleh Revolusi Perancis, yang berimbas pada kebijakan kolonial Hindia Belanda di bawah Gubernur Jenderal H.W.Daendels. Dalam pengkristenan ini tidak lepas dari jasa orang-orang peranakan Belanda yang bukan berasal dari kalangan teolog atau pendeta, serta beberapa orang Jawa yang gemar *ngelmu* (mencari pengetahuan hakiki tentang Tuhan) seperti tokoh - tokoh Coolen, J.E. Jellesma, Kiai Tunggul Wulung, Kiai Sadrach(Ainiyah et al., 2020).

2.3 Teori Wav2vec 2.0

Wav2vec 2.0 adalah kerangka kerja untuk pembelajaran mandiri (*self-supervised learning*) yang dirancang untuk mempelajari representasi suara dari sinyal audio mentah. Model ini memanfaatkan *feature encoder* berbasis konvolusi untuk mengubah sinyal audio menjadi representasi laten. Representasi ini kemudian dimasking secara acak pada ruang laten, memungkinkan model untuk belajar konteks dari data suara yang tidak terlihat. Langkah selanjutnya melibatkan jaringan Transformer yang digunakan untuk menghasilkan representasi kontekstual melalui *self-attention*, yang memungkinkan model menangkap hubungan antar fitur suara sepanjang waktu.

Untuk mengatasi tantangan pada data tanpa label, wav2vec 2.0 menggunakan modul kuantisasi untuk mengubah representasi laten menjadi unit diskret yang berfungsi sebagai target pembelajaran. Model ini dilatih menggunakan tugas kontrasif, di mana ia harus membedakan representasi laten yang benar dari kumpulan *distractors*. Setelah proses pre-training, model di-*fine-tune* menggunakan data berlabel dengan

memanfaatkan *Connectionist Temporal Classification* (CTC) sebagai fungsi loss untuk menghubungkan representasi suara ke teks transkripsi. Kombinasi pembelajaran mandiri, masking, kuantisasi, dan *fine-tuning* ini memungkinkan wav2vec 2.0 bekerja secara efektif bahkan pada dataset yang terbatas, menjadikannya solusi unggul untuk pengenalan suara otomatis di berbagai bahasa dan aksen.



Gambar 2.5 Arsitektur wav2vec 2.0

2.4 Teori Android Studio

Android studio adalah IDE (*Integrated Development Environment*) resmi untuk pengembangan aplikasi Android dan bersifat open source atau gratis. Peluncuran Android Studio ini diumumkan oleh Google pada 16 mei 2013 pada event Google I/O Conference untuk tahun 2013. Sejak saat itu, Android Studio mengantikan Eclipse sebagai IDE resmi untuk mengembangkan aplikasi Android.

Android Studio sendiri dikembangkan berdasarkan Intellij IDEA yang mirip dengan Eclipse disertai dengan ADT plugin (*Android Development Tools*). Android Studio memiliki fitur:

a. Projek berbasis pada Gradle *Build*
b. *Refactory* dan pembenahan bug yang cepat.
c. Tools baru yang Bernama “Lint” diklaim dapat memonitor kecepatan, kegunaan, serta kompetibelitas aplikasi dengan cepat.

d. Mendukung *Proguard* dan *App-signing* untuk keamanan.

Memiliki GUI aplikasi Android lebih mudah. Didukung oleh Google *Cloud Platform* untuk setiap aplikasi yang dikembangkan(Subari & Ramadhan, 2022).

2.5 Teori UML

Unified Modelling Language (UML) merupakan sebuah bahasa yang divisualisasikan dalam bentuk gambar atau grafik yang berfungsi untuk memberikan gambaran dan spesifikasi dalam pembangunan dan dokumentasi dari sebuah pengembangan sistem berorientasi objek (*object oriented*). UML memberikan sebuah standar pembuatan blue print sistem, yang dapat terdiri dari konsep proses bisnis, pembuatan class yang dapat dituangkan pada bahasa pemrograman tertentu, rancangan basis data, serta komponen-komponen yang dibutuhkan dalam pengembangan sistem(Siska Narulita et al., 2024).

2.6 Teori Activity Diagram

Activity diagram merepresentasikan aliran proses atau aktivitas dalam sebuah sistem yang akan dibangun, mulai dari proses awal, keputusan-keputusan yang terjadi di dalam sistem, hingga bagaimana sebuah proses berakhir. Activity diagram juga memvisualisasikan proses-proses paralel yang terjadi ketika sistem dieksekusi. Tahapan atau langkah-langkah yang terjadi di dalam sistem digambarkan dalam diagram ini. Setiap use case minimal terdapat satu activity diagram. Activity diagram dirancang berdasarkan satu atau beberapa use case yang ada pada use case diagram. Activity diagram merepresentasikan proses yang berjalan pada sebuah system(Siska Narulita et al., 2024). Berikut adalah simbol yang ada dalam *activity* diagram:

Simbol	Nama	Keterangan
	Status awal	Sebuah diagram aktivitas memiliki sebuah status awal.
	Aktivitas	Aktivitas yang dilakukan sistem,aktivitas biasanya diawali dengan kata kerja.
	Percabangan / Decision	Percabangan dimana ada pilihan aktivitas yang lebih dari satu.
	Penggabungan / Join	Penggabungan dimana yang mana lebih dari satu aktivitas lalu digabungkan jadi satu.
	Status Akhir	Status akhir yang dilakukan sistem, sebuah diagram aktivitas memiliki sebuah status akhir
	Swimlane	Swimlane memisahkan organisasi bisnis yang bertanggung jawab terhadap aktivitas yang terjadi

Gambar 2.6 Activity Diagram