PROPOSAL TUGAS AKHIR

Perbandingan Wav2Vec2 dan Data2Vec dalam Penelusuran Ayat Al-Qur'an Berdasarkan Fitur Laten Audio

Disusun Oleh:

Mujahid Ansori Majid 1197050093

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SUNAN GUNUNG DJATI

BANDUNG

2025

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

PENDAHULUAN

1. Latar Belakang

Perkembangan teknologi modern telah memengaruhi metode tradisional dalam penghafalan Al-Qur'an, mendorong munculnya pendekatan berbasis digital yang mengintegrasikan kemajuan dalam audio dan kecerdasan buatan. Keterbatasan waktu dan tantangan pengalih perhatian dalam era kontemporer menjadikan proses hafalan mandiri memerlukan dukungan sistem cerdas [1] [2].

Penerapan teknologi seperti *Automatic Speech Recognition* (ASR) telah terbuktu efektif. Penelitian sebelumnya, seperti yang ditunjukan oleh Shaklawoon et al [3] [4], berhasil mengembangkan sistem yang mendeteksi kesalahan dalam urutan ayat dan pengucapan dengan mengukur *Word Error Rate* (WER) dan *Character Error Rate* (CER) dan hasilnya dengan kedua metriks tersebut telah dicapai angka yang kecil. Inovasi yang telah dilakukan, banyak yang mengadopsi metodologi *deep learning end-to-end* [5] [6], menjadi dasar penting dalam pengembangan alat bantu hafalan Al-Qur'an. Dalam spektrum yang lebih luas, Wav2vec2 dan model sejenis telah mencapai progres signifikan dalam ASR bahasa Arab, semakin memperkuat kemampuan sistem untuk mengenali rekaman tilawah yang kompleks.

Meskipun ASR telah berhasil memberikan solusi transkripsi, studi lanjutan mengungkapkan bahwa model modern seperti Wav2vec2 dan Data2Vec memiliki potensi yang jauh melampaui sekadar konversi suara-ke-teks. Kedua model ini dikembangkan menggunakan paradigma *Self-Supervised Learning* (SSL), yang memungkinkan mereka mempelajari representasi audio mendalam (*latent embeddings*) dari sejumlah besar data audio tanpa memerlukan label transkripsi yang mahal.

Representasi laten yang dihasilkan inilah yang menjadi kunci untuk tugas-tugas non-transkripsi, seperti penelusuran kesamaan audio (*audio similarity search)* atau penelurusan informasi audio (AIR). Tujuan dari tugas *retrieval* adalah mencari ayat yang paling mirip di *database* berdasarkan kemiripan vektor suara (*embeddings*), bukan teks hasil ASR. Hal ini sangat relevan untuk konteks hafalan Al-Qur'an, d mana pengguna mungkin ingin mencocokkan potongan bacaan dengan ayat aslinya tanpa mengandalkan transkripsi yang rentan terhadap *error* tajwid. Konsep *similarity search*

ini didukung oleh bergai riset yang menunjukkan efektivitas penggunaan vektor *embeddings* dan *cosine similarity* untuk percocokan konten audio secara langsung [7]

Wav2Vec2, yang menggunakan *contrastive learning*, sangat efektif untuk ASR. Sementara itu, Data2Vec diperkenalkan sebagai model yang *modality-agnostic*, yang dinilai lebih fleksibel dan mampu menghasilkan representasi yang lebih kontekstual dengan arsitektur *teacher-student* [8]. Kedua model ini telah terbukti sukses digunakan di domain Arab untuk tugas non-ASR seperti *speaker identification* [9], [10].

Namun, meskipun potensi ini sudah terlihat, terdapat kesenjangan penelitian yang signifikan. Sebagian besar studi komparatif yang melibatkan Wav2Vec2 dan Data2Vec, termasuk yang berfokus pada tilawah Al-Qur'an, masih terfokus pada evaluasi kinerja transkripsi (*WER/CER*). Perbandingan langsung yang menguji kemampuan inheren kedua model dalam menghasilkan representasi audio yang berkualitas tinggi untuk tujuan Retrieval Ayat Al-Qur'an—yang diukur melalui metrik pencarian kesamaan (*similarity search*) seperti Mean Average Precision (mAP)—masih terbatas.

Oleh karena itu, riset ini diarahkan untuk melakukan perbandingan komprehensif antara Wav2Vec2 dan Data2Vec. Perbandingan ini akan secara khusus mengukur sejauh mana kualitas *latent embeddings* dari kedua model dapat mendukung tugas penelusuran ayat Al-Qur'an berbasis audio, guna menyusun rekomendasi implementasi teknologi yang paling produktif dan adaptif dalam menunjang aktivitas penghafalan di tengah arus digitalisasi.

2. Rumusan Masalah

Berdasarkan latar belakang di atas, maka rumusan masalah pada penelitian ini adalah:

1. Bagaimana perbandingan kualitas retrieval ayar qur'an berbasis audio menggunakan model wav2vec2 dan data2vec?

3. Batasan Masalah

Batasan masalah dari penelitian ini adalah sebagai berikut:

1. Penelitian dibatasi pada penggunaan model pretrained Data2Vec dan Wav2Vec2, tanpa pelatihan ulang (fine-tuning) pada data khusus Al-Qur'an

2. Eksperimen hanya dilakukan pada teks Al-Qur'an berbahasa Arab, terutama Surah-surah pendek seperti Al-Fatihah dan Juz Amma
3. Fokus sistem adalah pada pencarian kemiripan ayat, bukan pada koreksi tajwid, deteksi kesalahan pelafalan, atau penilaian kualitas tilawah

4. Tujuan Penelitian

Tujuan dari penelitian ini sebagai berikut:

1. Mengetahui Mengevaluasi kinerja model Wav2Vec2 dalam retrieval ayat Al-Qur'an berbasis audio.
2. Mengevaluasi kinerja model Data2Vec dalam retrieval ayat Al-Qur'an berbasis audio.
3. Membandingkan kedua model untuk menentukan model yang lebih sesuai digunakan pada sistem pencarian ayat Al-Qur'an.

5. Manfaat Penelitian

Manfaat yang diharapkan dari hasil penelitian ini adalah sebagai berikut:

1. Memberikan pemahaman ilmiah yang lebih dalam tentang efektivitas model representasi audio seperti Data2Vec dan Wav2Vec2 dalam domain bahasa Arab, khususnya untuk tugas pencocokan kemiripan teks berbasis suara.
2. Memberikan acuan bagi peneliti dan pengembang aplikasi keislaman dalam memilih model *embedding* audio terbaik untuk diterapkan pada sistem pembelajaran Al-Qur'an berbasis suara

6. *The State of The Art*

Beberapa Penelitian telah dilakukan untuk membuat software maintainability yang baik dari masa ke masa. Dalam upaya mengembangkan sebuah aplikasi dengan metode yang telah dikembangkan maka dibutuhkannya proses studi literatur. Berikut merupakan penelitian yang sebelumnya telah dilakukan dengan metode yang serupa:

a. Alexei Baevski, dkk (2020). *Wav2vec 2.0* memperkenalkan kerangkan *self-supervised learning* untuk mempelajari representasi ucapan langsung dari *raw audio.* Arsitektur terdiri dari *convolutional feature encoder* untuk mengekstrasi fitur awal, *Transformer network* untuk memodelkan dependensi jangka panjang, dan *quantizatino module* untuk mengubah

representasi menjadi kode diskrit yang digunakan dalam *contrastive learning*. Pendekatan ini secara signifikasn mengurangi kebutuhan data berlabel, menutup kesenjangan performa antara sistem ASR *fully-supervised* dan *low-resource*, serta mencapai *state-of-the-art* pada *benchmark* Librispeech 100h. Metode ini membuka arah baru pengenalan suara di bahasa dengan sumber daya terbatas melalui representasi ucapan yang lebih efisien secara data. [11]

b. Alexei Baevski, dkk (2022). Data2vec mengusulkan kerangka self-supervised learning multimodal yang seragam, mampu mempelajari representasi laten dari ucapan, teks dan citra menggunakan metode tunggal. Berbeda dengan pendekatan terdahulu yang memprediksi target spesifik modalitas, metode ini mempresdiksi representasi kontekstual laten dari input lengkap berdasarkan versi yang telah dilakukan *masking*. Proses pembelajaran dilakukan melalui mekanisme *self-distillation* dengan arsitekur Transformer, di mana model pelajar (*student*) mengestimasi keluaran model guru (*teacher*) yang dibekukan parameternya. Evaluasi empiris menunjukkan bahwa data2vec mencapai kinerja *state-of-the-art*. [8]

c. Omar Mohamed & Salah A. Aly (2021). *Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset*. Makalah ini memperkenalkan model deep learning untuk pengenalan emosi dalam ucapan bahasa Arab menggunakan representasi audio canggih seperti wav2vec 2.0 dan HuBERT. Model-prinsip self-supervised dijalankan pada dataset tanpa label besar dan kemudian di-fine-tune pada dataset kecil (BAVED). Representasi fitur ini digunakan pada classifier berupa MLP dan Bi-LSTM. Hasil eksperimen menunjukkan bahwa wav2vec 2.0 memberikan performa terbaik dalam akurasi pengenalan emosi, konvergensi lebih cepat, dan stabilitas pelatihan dibanding HuBERT. Model ini mencapai akurasi hingga 89% pada dataset BAVED. Pendekatan ini menunjukkan efektivitas representasi self-supervised untuk tugas SER dalam bahasa Arab, terutama dengan data terbatas [9].

d. Wilhelm Öberg (2025). Paragraf ini sangat relevan dengan penelitian Anda karena secara eksplisit berfokus pada Pencarian Audio *Query-by-Example*, yang memiliki kesamaan tugas dengan pencarian kemiripan ayat yang Anda

lakukan, yaitu mencari konten berdasarkan *query* audio. Penelitian ini menggunakan Wav2Vec 2.0 sebagai basis ekstraksi *embedding*, sama persis dengan salah satu model utama yang Anda uji. Secara metodologis, studi ini mengimplementasikan Contrastive Learning (melalui *Triplet Loss*) untuk menyelaraskan *embedding* dan menciptakan ruang vektor yang lebih terstruktur. Pendekatan ini adalah justifikasi penting yang menjelaskan mengapa Wav2Vec2 dapat menghasilkan representasi laten yang efektif untuk tugas *retrieval*. Lebih lanjut, konsep pencarian yang sepenuhnya bebas transkripsi (*transcription-free*) yang diusung oleh penelitian ini secara kuat mendukung argumen inti Anda, yaitu bahwa pencarian ucapan dapat dilakukan secara langsung menggunakan fitur laten audio tanpa perlu bergantung pada hasil transkripsi ASR yang mahal dan rawan kesalahan [12].

e. Patrick Cormac, dkk (2024). Dalam upaya memahami mekanisme internal model *self-supervised learning* untuk pemrosesan ucapan, sebuah penelitian mendalam dilakukan untuk menilai organisasi fitur ucapan dalam *embedding* Wav2Vec 2.0. Penelitian ini membuktikan bahwa representasi laten tersebut tidak hanya mengandung informasi fonetik secara mentah, tetapi juga mengorganisasikannya secara struktural dengan cara yang konsisten dan selaras dengan teori fonologi linguistik. Organisasi fitur yang terstruktur ini—mencakup elemen-elemen seperti fitur *voicing*, *plosive*, dan *bilabial*—menjadi dasar krusial yang memungkinkan model membedakan nuansa halus dalam ucapan berdasarkan kemiripan fonetik. Untuk mencapai kesimpulan ini, penelitian tersebut menggunakan metode *probing* yang canggih, di mana *Multilayer Perceptron (MLP) probes* dilatih untuk memprediksi kehadiran 30 fitur fonetik, dan hasilnya dianalisis menggunakan Association Rule Mining. Penggunaan teknik *probing* dan analisis asosiasi ini memberikan bukti empiris yang kuat mengenai penjelasan (*explainability*) terhadap kapabilitas fitur laten yang diekstraksi oleh model *self-supervised* berbasis *Transformer* [13]

f. Qijie Shao, dkk (2025). Memperkenalkan metode decoupling quantization melalui dua K-means quantizer untuk memisahkan informasi bahasan dan fonem dalam proses *self-supervised learning*. Berbeda dengan Data2Vec yang melakukan rata-rata lintas sehingga fitur tercampur (bahasa, fonem,

pembicara). DQ-Data2vec menggunakan jumlah klaster yang sesuai dengan jumlah bahasa dan fonem untuk menghasilkan representasi yang lebih terpisah dan bersih. Eksperimen pada dataset CommonVoice menunjukkan peningkatan signifikan, dengan pengurangan relatif 9,51% pada **phoneme** error rate (PER) dan 11,58% pada word error rate (WER) dibandingkan Data2vec dan UniData2vec. Pendekatan ini memperlihatkan potensi besar untuk pengenalan suara multibahasa, baik dalam skenario self-supervised maupun weakly-supervised [14].

g. Yang Xie, dkk (2021). Dalam konteks pemanfaatan fitur laten untuk tugas non-pengenalan ucapan (non-ASR), penelitian ini memberikan bukti penting mengenai kemampuan model yang dilatih sendiri (*self-supervised*). Secara spesifik, penelitian tersebut memvalidasi penggunaan fitur Wav2Vec yang telah dilatih secara *self-supervised* untuk tugas klasifikasi yang kompleks, seperti deteksi *spoofing* (suara palsu), yang memerlukan diskriminasi fitur yang sangat halus. Keberhasilan model dalam tugas ini, bahkan terhadap masalah *out-of-distribution*, mengindikasikan kapabilitas generalisasi yang kuat dari representasi Wav2Vec. Secara metodologis, penelitian ini mengadopsi struktur Siamese Network yang dikombinasikan dengan fungsi kerugian Contrastive Loss. Pendekatan ini secara inheren bertujuan untuk mempelajari metrik kemiripan yang optimal dalam ruang representasi, yang krusial untuk mengelompokkan sampel serupa dan memisahkan sampel yang berbeda. Oleh karena itu, arsitektur ini secara efektif memperkuat prinsip-prinsip yang mendasari sistem *retrieval* yang didasarkan pada perbandingan kemiripan vektor [15].

h. Nik Vaessen dan David A. (2022). Penelitian ini berfokus pada pemanfaatan wav2vec2 untuk tugas speaker recognition melalui pendekatan fine-tuning. Penulis mengusulkan dua varian metode: (1) *single-utterance classification*, yang memandang pengenalan speaker sebagai masalah klasifikasi dan menggunakan *cross-entropy loss* maupun *additive angular softmax loss* untuk memperbesar margin antar speaker; serta (2) *utterance-pair classification*, yang memandang pengenalan pembicara sebagai masalah verifikasi pasangan ucapan, dengan *binary cross-entropy loss* untuk memprediksi apakah dua ucapan berasal dari pembicara yang sama. Hasil eksperimen menunjukkan bahwa kedua varian tersebut secara efektif

meningkatkan performa wav2vec2 pada tugas pengenalan pembicara, dan membuktikan fleksibilitas model dalam menyesuaikan diri dengan formulasi klasifikasi maupun verifikasi [10].

i. Ahmed Adel, dkk (2024). Mengenai aspek adaptasi model *self-supervised* ke domain data yang spesifik, penelitian ini memberikan wawasan tentang efektivitas *Continued Pre-training* (CPT). Metode CPT terbukti menjadi teknik yang paling efektif untuk mengadaptasi model dasar Wav2Vec 2.0 yang umum ke domain audio spesifik yang dicirikan oleh kondisi akustik yang menantang (seperti lingkungan kelas yang bising). Adaptasi ini secara eksplisit berkontribusi pada peningkatan *robustness* model, yang berarti representasi *embedding* yang diekstraksi menjadi lebih stabil dan konsisten meskipun terdapat variasi dalam kualitas rekaman, karakteristik *speaker*, atau kondisi lingkungan. Oleh karena itu, CPT merupakan metode adaptasi domain yang signifikan yang dapat dipertimbangkan untuk mengoptimalkan kinerja model *self-supervised* seperti Wav2Vec 2.0 atau Data2Vec, terutama ketika diterapkan pada domain ucapan yang sangat spesifik seperti rekaman Al-Qur'an [16]

j. Alexis Conneau, dkk (2020). memperkenalkan pendekatan *unsupervised speech recognition* yang menggabungkan *wav2vec 2.0* dengan *k-means clustering* dan *sequence-to-sequence language modeling*. Alih-alih bergantung pada transkrip berlabel, representasi akustik dari wav2vec 2.0 di-*quantize* menggunakan k-means, menghasilkan token diskrit. Token ini kemudian dilatih dengan model bahasa untuk menghasilkan transkripsi teks. Hasil eksperimen pada dataset *Librispeech* menunjukkan bahwa sistem ini dapat mencapai performa kompetitif bahkan tanpa supervisi, membuka arah baru untuk pengenalan suara pada bahasa dengan sumber daya terbatas.

Tabel 1 State of the art

No	Judul Jurnal dan Peneliti	Metode	Tujuan
1	wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations Peneliti: R Prajana, Prof.Kavitha S N (2021)	Wav2vec2, *contrastive masked prediction of quantized units*	Tujuan utama dari Wav2Vec 2.0 adalah memungkinkan pembelajaran representasi suara secara efektif tanpa membutuhkan dataset berlabel dalam jumlah besar. Penulis berupaya mempelajari representasi langsung dari sinyal audio mentah, mengurangi ketergantungan pada fitur buatan, serta menunjukkan bahwa pre-training self-supervised dapat secara signifikan meningkatkan kinerja automatic speech recognition (ASR) meskipun hanya menggunakan data berlabel yang terbatas. [11]
2	data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language	Data2vec, *self-distillation of contextual latent representations*	Data2vec bertujuan mengembangkan sebuah framework self-supervised yang seragam penerapannya untuk berbagai modalitas (gambar, suara, teks).

No	Judul Jurnal dan Peneliti	Metode	Tujuan
	Peneliti: Alexei Baevski, 2022		Dalam pendekatannya, data2vec mengkombinasikan *masked prediction* dan *self-distillation* dengan rata-rata lapisan sebagai target untuk memprediksi representasi laten yang bersifat kontekstual dari seluruh input. Pendekatan ini telah terbukti memberikan performa state-of-the-art atau setara di benchmark-benchmark utama speech recognition, image classification, dan natural language understanding [8]
3.	Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset Penulis: Omar Mohamed, dkk (2021)	Menggunakan representasi kontekstual dari Wav2vec2 dan HuBERT. Setelah itu diklasifikasikan menggunakan MLP dan Bi-LSTM	Mengembangkan model pengenalan emosi dalam ucapan bahasa Arab dengan memanfaatkan representasi self-supervised. Validasi pada dataset BAVED menunjukkan model berbasis wav2vec 2.0 mencapai akurasi hingga 89%, mengungguli model HuBERT base (87%) dan HuBERT large (84%)

No	Judul Jurnal dan Peneliti	Metode	Tujuan
4.	Query-by-Example Audio Search using Acoustic Word Embeddings: Transforming wav2vec 2.0 Embeddings using Contrastive Learning Peneliti: Wilhelm Öberg (2025)	*Embedding Transformation* menggunakan *Projection Network* dengan Contrastive Learning (Triplet Loss) pada *Embedding* Wav2Vec 2.0.	Penelitian ini menginvestigasi peningkatan kinerja Pencarian Audio *Query-by-Example* (QbE) melalui transformasi *embedding* Wav2Vec 2.0 menggunakan Contrastive Learning (khususnya *Triplet Loss*). Tujuannya adalah melatih jaringan proyeksi untuk menyelaraskan *embedding* kata yang serupa dan memfilter karakteristik *speaker* dan kebisingan, sehingga menghasilkan ruang *embedding* yang lebih terstruktur. Metode ini berkontribusi pada pencarian audio yang sepenuhnya bebas transkripsi dan sangat relevan untuk tugas *retrieval* berbasis kemiripan fitur laten.
5.	Searching for Structure: Appraising the Organisation of Speech Features in	Probing Methods (MLP Probes) dilatih untuk 30 fitur fonetik, diikuti oleh Association Rule	Bertujuan mengungkap organisasi fitur fonetik dalam representasi laten Wav2Vec 2.0 dan keselarasan mereka dengan

No	Judul Jurnal dan Peneliti	Metode	Tujuan
	wav2vec 2.0 Embeddings Peneliti: Patrick Cormac, dkk (2024)	Mining (Apriori) pada aktivasi *probe* dan matriks fitur teoritis. Hasil dianalisis melalui Graph Visualization	teori linguistik. Penelitian ini menemukan bahwa *embedding* Wav2Vec 2.0 menangkap struktur asosiasi fitur yang sangat selaras dengan aturan fonologis teoretis, menunjukkan representasi yang robust dan koheren
6.	DQ-Data2vec: Decoupling Quantization for Multilingual Speech Recognition Peneliti: Qijie Shao (2025)	Data2vec backbone, dua K-means quantizer untuk memisahkan fitur bahasa dan fonem	Bertujuan mengungkap **organisasi fitur fonetik** dalam representasi laten Wav2Vec 2.0 dan keselarasan mereka dengan teori linguistik. Penelitian ini menemukan bahwa *embedding* Wav2Vec 2.0 menangkap struktur asosiasi fitur yang **sangat selaras** dengan aturan fonologis teoretis, menunjukkan representasi yang **robust dan koheren**
7.	Siamese Network with Wav2vec Feature for Spoofing Speech Detection Peneliti: Yang Xie, dkk (2022)	Two-Phase Learning: 1. Siamese Network + Contrastive Loss pada fitur Wav2vec. 2. MLP Classifier pada *embedding* hasil Phase 1.	Mengusulkan sistem dua fase untuk mengatasi masalah generalisasi terhadap serangan *spoofing* ucapan (*out-of-distribution*) yang tidak dikenal. Sistem ini memanfaatkan fitur Wav2vec yang sudah

No	Judul Jurnal dan Peneliti	Metode	Tujuan
			dilatih (*pretrained*) sebagai fitur input diskriminatif. Pendekatan ini secara signifikan meningkatkan kinerja, mengurangi *Equal Error Rate* (EER) dari 4.07% (SOTA sebelumnya) menjadi 1.15% pada *benchmark* ASVspoof 2019.
8.	FINE-TUNING WAV2VEC2 FOR SPEAKER RECOGNITION Peneliti:	*Single-utterance classification* dengan cross-entropy loss atau additive angular softmax loss dan *utterance-pair classification* dengan binary cross-entropy loss	Mengoptimalkan embedding Wav2Vec2 agar lebih diskriminatif untuk identifikasi dan verifikasi speaker, sekaligus menunjukkan bahwa pendekatan self-supervised dapat diaplikasikan secara efektif di luar ASR, khususnya pada pengenalan pembicara.
9.	CPT-Boosted Wav2vec2.0: Towards Noise Robust Speech Recognition for Classroom Environments Peneliti:	Continued Pre-training (CPT) dengan *unlabeled* dan *labeled* data domain target pada model Wav2vec 2.0 yang telah dilatih awal (*pre-trained*).	Penelitian ini menguji efektivitas CPT dalam mengadaptasi Wav2Vec 2.0 ke domain bising seperti lingkungan kelas. CPT terbukti menjadi alat yang kuat, mengurangi *Word Error Rate* (WER) hingga lebih dari 10% dan

No	Judul Jurnal dan Peneliti	Metode	Tujuan
	Ahmed Adel Attia, dkk (2024)		meningkatkan *robustness* model terhadap berbagai jenis kebisingan dan kondisi mikrofon, menjadikannya lebih unggul dari metode adaptasi domain lainnya.
10.	Unsepervised Cross-Lingual Representation Learning for Speech recognition Peneliti: Alexis Conneau, dkk (2020)	Menggabungkan *wav2vec 2.0* untuk representasi akustik, *k-means clustering* untuk kuantisasi token, serta *sequence-to-sequence language model* untuk menghasilkan transkripsi.	Mengembangkan representasi ucapan lintas bahasa dengan melatih satu model pada banyak bahasa sekaligus. Tujuannya adalah meningkatkan performa ASR, khususnya di bahasa *low-resource*, dengan membagi representasi fonetik antarbahasa. XLSR menunjukkan reduksi error signifikan (72% pada PER di CommonVoice, 16% pada WER di BABEL) dan menghasilkan model multilingual yang kompetitif terhadap model individual terbaik.

Penelitian mengenai representasi ucapan yang dilatih sendiri (*self-supervised*) telah mengalami kemajuan pesat, dimulai dengan fondasi seperti Wav2vec 2.0 dan Data2vec [8], [11] yang mampu mempelajari representasi laten koheren dari *raw audio* dan berbagai modalitas. Namun, mayoritas penelitian awal dan penelitian lanjutan berkonsentrasi pada peningkatan kinerja ASR (*automatic speech recognition*), seperti yang terlihat pada studi *cross-lingual* [17] dan peningkatan *robustness* terhadap kebisingan melalui CPT-Boosted Wav2vec2.0 [16]. Penelitian lain juga berfokus pada tugas *non-retrival*, termasuk pengenalan pembicara dan deteksi *spoofing* atau pemalsuan suara yang memanfaatkan *siamase network* pada fitur wav2vec2 [15], Meskipun demikian, beberapa studi telah menggeser fokus ke kapabilitas *retriavel* dari *embedding* ini. Misalnya penelitian yang dilakukan oleh Wilhelm Öberg mengenai audio search menggunakan *embedding* dari wav2vec2 [12]

Penelitian wav2vec 2.0 [11] menunjukkan bahwa representasi *self-supervised* dari *raw* audio dapat secara signifikan meningkatkan kinerja ASR (*automatic speec recognition*), bahkan dengan data berlabel terbatas, mencapai *state-of-the-art* di LibriSpeech [18]. Adapun penelitian lanjutannya berupa XLSR (*cross-lingual speech recognition*) [17], memperluas model ke 53 bahasa dan berhasil menurunkan *error rate* pada benchmark lintas bahasa secara drastis, sehingga mendukung skenario *low-resource*. Data2vec [8] memperkenalkan paradigma baru berbasis *self-distillation*, yang mampu menggeneralisasi ke berbagai modalitas, dan terbukti model tersebut kompetitif di bergai tugas dengan data suara, citra dan NLP. Penelitian lain seperti AV-data2vec maupun AV2vec mengembangkan representasi multimodal audio-visual, memperlihatkan keunggulan pada *downstream tasks* yang membutuhkan integrasi suara dan citra.

Selain itu, terdapat pula penelitian mengenai *fine-tuning* Wav2Vec2 untuk pengenalan pembicara [17] menunjukkan bahwa model ini mampu mengenali identitas suara dengan baik dan konsisten. Dalam konteks bahasa Arab, studi lain mengenai pengenalan emosi dari suara dengan Wav2Vec2 [9] berhasil mencapai akurasi tinggi hingga 89% meskipun menggunakan *dataset* yang relatif kecil.

Berdasarkan kajian terhadap penelitian-penelitian sebelumnya, terlihat bahwa kebanyakan studi mengenai Wav2Vec2 maupun Data2Vec berfokus pada *automatic speech recognition* (ASR), *cross-lingual representation*, *speaker recognition*, maupun

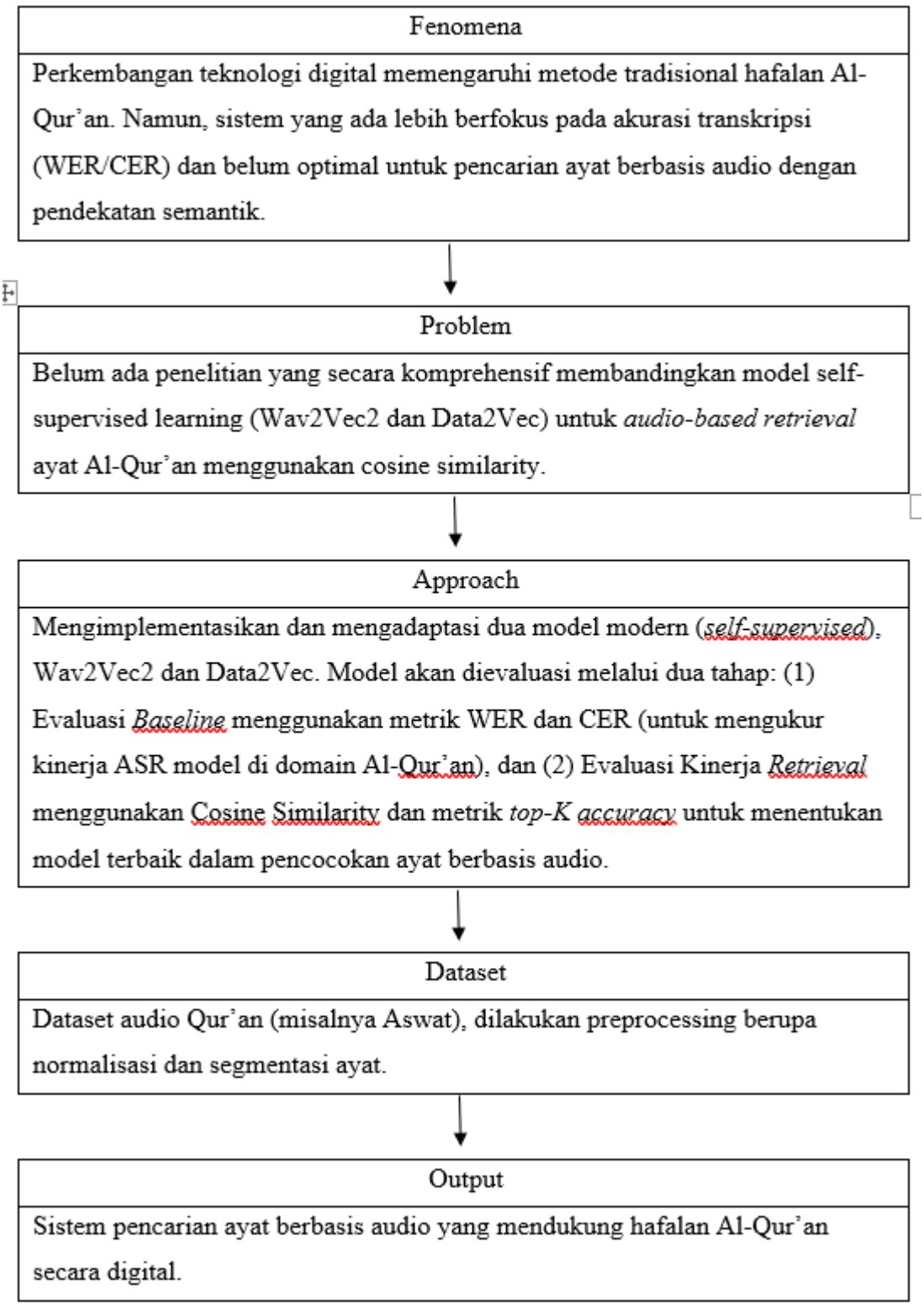
pengembangan multimodal audio-visual. Pendekatan tersebut memang menunjukkan kemajuan signifikan dalam menurunkan Word Error Rate (WER), meningkatkan akurasi pengenalan suara, dan memperluas cakupan ke berbagai bahasa serta modalitas.

Penelitian ini menghadirkan perbedaan penting dengan menempatkan kedua model tersebut dalam konteks retrieval audio ayat Al-Qur'an. retrieval audio ayat Al-Qur'an. Kajian literatur menunjukkan bahwa perbandingan langsung antara Wav2Vec2 dan Data2Vec, khususnya untuk evaluasi kualitas *embedding* dalam tugas *similarity search* atau *audio retrieval*, masih sangat jarang atau belum pernah dilakukan, terutama di domain spesifik ini**.** Alih-alih menitikberatkan pada performa transkripsi, penelitian ini mengevaluasi kemampuan Wav2Vec2 dan Data2Vec untuk mencocokkan cuplikan bacaan dengan ayat terkait. Dengan menggunakan pendekatan komparatif berbasis kesamaan *embedding*, penelitian ini menawarkan kontribusi baru yang belum banyak dieksplorasi sebelumnya.

Keunggulan penelitian ini terletak pada fokus khusus pada domain bacaan Al-Qur'an yang memiliki karakteristik unik, baik dari sisi fonetik maupun aturan tajwid yang membedakannya dari wacana lisan pada umumnya. Selain itu, penelitian ini mengusulkan penerapan *retrieval* audio sebagai sarana praktis yang dapat mendukung kegiatan hafalan dan pembelajaran Al-Qur'an, sehingga nilai manfaatnya tidak hanya terbatas pada aspek teknis, tetapi juga pada aspek pendidikan. Keunggulan lain terletak pada pendekatan evaluasi yang digunakan, yakni dengan membandingkan dua model *self-supervised* modern melalui metrik berbasis kemiripan *embedding*, sehingga performa sistem lebih terukur dalam konteks pencarian bacaan yang mirip dengan input audio, bukan sekadar akurasi transkripsi teks.

Dengan demikian, penelitian ini diharapkan dapat memperkaya pemahaman mengenai pemanfaatan model representasi audio mutakhir untuk mendukung sistem pembelajaran Al-Qur'an yang lebih adaptif dan aplikatif.

Kerangka pemikiran dalam penelitian tugas akhir ini akan dipaparkan pada Gambar 2 sebagai berikut:



Gambar 1 Kerangka Pemikiran

7. Studi Literatur

7.1 Model Representasi Laten *Self-Supervised* (SSL)

Self-Supervised Learning (SSL) adalah paradigma pembelajaran mesin di mana model dilatih untuk mempelajari representasi yang berguna dari data yang tidak memiliki *label* dengan menghasilkan labelnya sendiri. SSL terletak di antara *Supervised Learning* (yang membutuhkan label manual ekstensif) dan *Unsupervised Learning* (yang mencari pola tanpa label).

Inti dari SSL terletak pada penciptaan "Tugas Preteks" (*Pretext Task*). Tugas ini memaksa model untuk memahami dan memprediksi bagian tersembunyi (*masked*) atau bagian yang hilang (*corrupted*) dari data input itu sendiri.

7.1.1 Wav2vec 2.0: Representasi Kontekstual Melalui Contrastive Learning

Wav2vec 2.0 merupakan kerangka self-supervised yang menetapkan fondasi kuat dalam pembelajaran representasi ucapan pembelajaran representasi ucapan. Arsitektur ini menggunakan *Feature Encoder* konvolusional untuk mengekstrak representasi skala waktu rendah dari sinyal audio mentah, yang kemudian diumpankan ke jaringan Transformer untuk menghasilkan rerpesentasi konteksual. Mekanisme pelatihannya berpusat pada Contrastive Loss, yang berfungsi memaksa model untuk membedakan segmen ucapan yang di-*masking* dari sejumlah kandidat negatif. Proses ini menghasilkan *embedding* sangat efektif dalam mengambil detail fonetik dan akustik yang merupakan kunci dalam tugas retrieval berbasis *similarity* [11].

7.1.2 Data2Vec: *Embedding* generalis melalui Self-Distillation

Sebagai model pembanding, Data2Vec memperkenalkan pendekatan *self-supervised* yang bersifat *modality-agnostic*, memungkinkan pembelajaran representasi yang dapat digunakan pada ucapan, teks maupun citra. Perbedaan utama Data2Vec terletak pada mekanisme *Self-Distillation*, di mana *Student Model* dilatih untuk memprediksi representasi laten kontekstual yang dihasilkan oleh *Teacher Model*. Pendekatan ini bertujuan menciptakan *embedding* yang lebih umum dan kurang terikat pada suatu modalitas. Karakteristik Data2Vec yang

generalis ini menjadikannya model yang ideal untuk dikomparasi dengan Wav2vec 2.0 yang spesifik ucapan, guna mengevaluasi *trade-off* antara generalisi model dan kualitas *embedding* untuk tugas *retrieval* fonetik [8].

7.2 Validasi Kualitas Fitur Laten untuk Pencarian Fonetik

Sub-bab ini berfokus pada pembuktian bahwa representasi laten yang diekstrak oleh model SSL tidak hanya bersifat abstrak, tetapi terstruktur dan cukup diskriminatif untuk digunakan dalam tugas *Similarity Search* berbasis fonetik.

7.2.1 Bukti Struktur Fitur Fontek dalam *Embedding*

Representasi laten yang dihasilkan oleh Wav2Vec 2.0 terbukti memiliki kualitas dan organisasi yang melampaui sekadar representasi akustik sederhana. Penelitian *Searching for Structure*. Memberikan justifikasi teoretis yang kuat dengan menerapkan *Probing Methods*, melatih pengklasifikasi sederhana pada *embedding* yang dibekukan—untuk memprediksi fitur-fitur fonetik spesifik (seperti *voicing, plosive, dan bilabial*). Hasil analisis asosiasi lebih lanjut membuktikan bahwa embedding secara struktural mengorganisasikan fitur fonetik dengan cara yang konsisten dan selaras dengan prinsip fonologi linguistik teoritis. Pembuktian bahwa *embedding* memiliki makna fonetik yang terstruktur ini merupakan landasan ilmiah utama untuk hipotesis penelitian ini: bahwa representasi tersebut mampu membedakan dan mengukur kemiripan antar bacaan Al-Quran berdasarkan kesamaan fonetik dan aturan tajwid yang terinternalisasi.

7.2.2 Generalisasi dan Adaptasi *Embedding* ke Tugas Kemiripan

Validitas *embedding* Wav2Vec 2.0 untuk tugas *Similarity Search* diperkuat oleh kemampuan generalisasinya di luar *Aumatic Speech Recognition* (ASR). Fitur laten ini terbukti efektif dalam memetakan kemiripan di berbagai tugas non-ASR lain. Contohnya studi Siamese Network menunjukkan bagaimana fitur Wav2Vec, ketika dipasangkan dengan Contrastive Loss, dapat secara optimal mempelajari metrik jarak untuk tugas diskriminasi. Seperti *anti-spoofing.* Keberhasilan ini menegaskan bahwa embedding tersebut secara inheren cocok utnuk pemetaan kemiripan.

7.2.3 Peningkatan *Robustness* Model Terhadap Domain Baru

Kualitas *embedding* untuk retieval juga bergantung pada stabilitasnya di tengah variasi data. Dalam konteks adaptasi model ke domain baru seperti audio Al-Quran yang mungkin memiliki keragaman *speaker* dan kualitas rekaman, *Continued Pre-training* (CPT) merupakan teknik yang relevan menunjukkan bahwa CPT efektif dalam meningkatkan robustness model dengan mengadaptasi *embedding* ke kondisi akustik yang berbeda (misalnya, kebisingan lingkungan). Pemarapan ini membenarkan bahwa *embedding* dapat dipertahankan stabilitasnya melalui teknik adaptasi domain, memastikan konsistensi fitur yang diekstrak dari data Al-Quran yang bervariasi.

7.3.1 Prinsip *Similarity Search* Berbasis Vektor

Metodologi penelitian ini berpusat pada pencarian kemiripan (similarity search) di ruang vektor laten, yang bertujuan menemukan data dalam database yang memiliki representasi vektor terdekat dengan representasi vektor *query* yang diberikan. Dalam konteks penelusuran ayat Al-Qur'an, tugas ini berarti mencari indeks ayat yang paling mirip (closest match) berdasarkan embedding audio. Pendekatan ini secara fundamental berbeda dari tujuan Automatic Speech Recognition (ASR) konvensional karena model tidak dilatih untuk menghasilkan teks, melainkan untuk mempertahankan kedekatan fonetik dan kontekstual dalam ruang vektor. Keberhasilan dalam tugas ini secara langsung mengukur kualitas embedding untuk memetakan kemiripan audio.

7.3.2 Metrik Kemiripan Vektor: *Cosine Similarity*

Cosine Similarity diimplementasikan sebagai metrik utama untuk mengukur kedekatan antara embedding yang diekstrak dari kedua model SSL (Wav2Vec 2.0 dan Data2Vec). Cosine Similarity mengukur sudut antar dua vektor, sehingga mengukur kesamaan arah dan orientasi fitur sambil mengabaikan perbedaan magnitudo vektor. Metrik ini sangat efektif dalam konteks *similarity* search karena ia fokus pada kesamaan konten fitur yang diwakili oleh arah vektor. Kinerja retrieval kemudian dievaluasi menggunakan metrik berbasis pemeringkatan

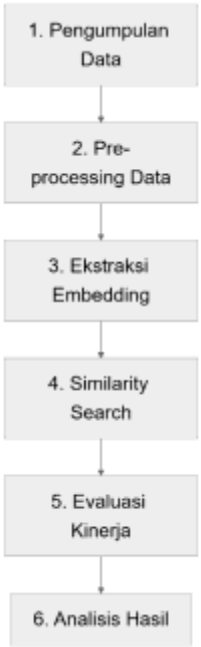
(ranking), seperti Top-K Accuracy, untuk menilai seberapa efektif setiap model menempatkan ayat target yang benar dalam *K* hasil pencarian teratas.

7.3.3 Metrik *Baseline* Komparatif ASR (WER/CER)

Meskipun fokus penelitian adalah *Similarity Search*, metrik ASR tradisional, seperti *Word Error Rate* (WER) dan *Character Error Rate* (CER), akan diukur sebagai metrik *baseline* komparatif. Pengukuran WER/CER bertujuan untuk menilai kapabilitas transkripsi awal model dalam domain audio Al-Qur'an. Data ini akan digunakan sebagai konteks untuk menganalisis korelasi antara kinerja transkripsi dan kinerja *retrieval* model. Dengan demikian, penekanan utama dan klaim kontribusi penelitian diletakkan secara eksklusif pada metrik *Similarity Search* (Cosine Similarity dan Top-K Accuracy).

8. Metode Penelitian

Penelitian ini akan dilakukan melalui enam tahapan utama yang berurutan, dimulai dari pengumpulan data hingga analisis perbandingan hasil *retrieval* kedua model seperti yang dipaparkan dalam Gambar 2.



Gambar 2 Metode Penelitian

1. **Pengumpulan data**

Tahap awal penelitian berfokus pada persiapan dan pengumpulan sumber data audio dan teks Al-Quran. Untuk sumber audio, penelitian ini akan menggunakan dataset audio Al-Quran yang terstruktur dan tersedia untuk umum, seperti koleksi Aswat atau dataset serupa yang relevan. Dataset audio ini harus dilengkapi dengan data teks Al-Quran yang berpasangan secara akurat dengan audio, serta memiliki penandaan (anotasi) ayat yang jelas untuk membangun database *retrieval*. Setelah data terkumpul, akan dilakukan pembagian data menjadi dua set utama: Database (D) dan Query Set (Q). Database (D) akan mencakup keseluruhan audio ayat yang berfungsi sebagai target pencarian (*search space*). Sementara itu, Query Set (Q) akan terdiri dari kumpulan audio, baik dalam bentuk potongan ayat maupun ayat penuh, yang akan digunakan sebagai input query untuk menguji dan mengevaluasi kinerja sistem *retrieval*.

2. **Tahap *Pre-processing* Data**

Tahap kedua adalah *pre-processing* data, yang bertujuan untuk menstandarisasi seluruh data audio agar siap digunakan untuk ekstraksi fitur (*feature extraction*). Proses ini dimulai dengan Normalisasi Audio, di mana semua berkas audio dalam *Database* (D) dan *Query Set* (Q) akan diseragamkan ke *sampling rate* dan format yang konsisten, misalnya 16kHz, mono. Langkah krusial berikutnya adalah Segmentasi Ayat, yang memastikan bahwa setiap berkas audio telah tersegmentasi secara akurat sesuai dengan batasan ayat Al-Qur'an yang benar. Selain itu, sebagai langkah opsional untuk mendukung perhitungan *baseline* WER/CER, akan dilakukan Pembersihan Teks Arab, seperti penghilangan tanda diakritik (*harakat*) minor untuk menjaga konsistensi data transkripsi.

3. **Tahap Ekstraksi Representasi Laten (Vector *Embedding*)**

Tahap ketiga merupakan inti dari penelitian ini, yaitu menghasilkan vector *embedding* dari kedua model *Self-Supervised* (SSL) yang penghilangan tanda diakritik (*harakat*) minor untuk menjaga konsistensi data transkripsi. Langkah awal adalah pemilihan Model, di mana penelitian ini akan menggunakan model Wav2vec2 dan Data2Vec yang telah dilatih (*pre-trained*), idealnya menggunakan versi *large* atau versi yang telah diadaptasi secara spesifik ke domain bahasa Arab, jika tersedia. Proses ekstraksi fitur kemudian dilakukan secara pararel untuk kedua model. Untuk Wav2vec2, setiap berkas audio yang telah di *pre-processing* dari Database (D) dan

Query Set (Q) diumpankan ke *feature encoder* model. Representasi laten (*hidden states*) yang dihasilkan dari lapisan *Transformer* terakhir kemudian diekstrak. Setelah itu, dilakukan *Pooling Temporal* (misalnya, *Mean Pooling*) pada *hidden states* tersebut untuk mengagregasi seluruh urutan fitur menjadi atau vektor tunggal berdimensi tetap per ayat (atau per *query*), yang kemudian disimpan sebagai *Embedding* Wav2Vec2. Proses yang identik diulangi untuk Model 2 (Data2Vec) guna menghasilkan *Embedding* Data2Vec. Hasil akhir tahap ini adalah dua database vektor yang komprehensif dan siap dibandingkan.

4. Tahap Implementasi *Similarity Search*

Tahap keempat berfokus pada implementasi mekanisme pencarian kemiripan (*Similarity search*) menggunakan *embedding* vektor yang telah diekstrak, guna menguji perbandingan kedua model. Untuk setiap audio query q_i di *Query Set*. Proses pencarian dilakukan secara terpisah untuk setiap model. Pertama, vektor *query* $V_{q,i}$ dari model terkait diekstrak. Kemudian, dilakukan perhitungan cosine similarity antara $V_{q,i}$ dan seluruh vektor yang tersimpan di dalam *Database (D)* yang bersesuaian ($D_{Wav2vec2}$ atau $D_{Data2vec}$). Hasil perhitungan ini menghasilkan daftar peringkat (*ranked list*) ayat-ayat di database, yang diturunkan dari skor kemiripan cosine dari tertinggi hingga terendah. Penentuan *Retrieval* dilakukan dengan menganggap ayat yang memiliki skor kemiripan Cosine tertinggi sebagai hasil prediksi model yang paling mendekati *query* audio.

5. Tahap Evaluasi Kinerja (Perbandingan)

Tahap kelima bertujuan untuk membandingkan kedua model SSL secara kuantitatif. Evaluasi Kinerja *Retrieval* akan menjadi metrik utama. Metrik yang digunakan meliputi *Top-K Accuracy* untuk mengukur persentase *query* yang berhasil menempatkan ayat target yang benar di dalam *K* peringkat atas (misalnya, Top-1, Top-5, dan Top-10), di mana *Top-1 Accuracy* berfungsi sebagai metrik keberhasilan pencarian yang paling ketat. Selain itu, Mean Average Precision (MAP) akan dihutung sebagai metrik yang lebih komprehensif untuk menilai kualitas urutan peringkat dari seluruh hasil *retrieval*.

6. Tahap Analisis Hasil

Tahap akhir penelitian adalah merumuskan temuan. Perbandingan Kuantitatif akan dilaukan dengan membandingkan secara langsung metrik *Top-K Accuracy* antara Wav2vec2 dan Data2Vec untuk menentukan model SSL mana yang unggul dalam tugas *Audio Retrieval* ayat Al-Quran. Selain itu, akan dilakukan Analisis Kualitatif melalui studi kasus mendalam terhadap *query* yang mengalami kegagalan dan keberhasilan. Analisis ini meliputi pengidentifikasian kasus di mana satu model unggul (misalnya, Data2Vec mencari ayat dengan konteks yang berbeda namun kemiripan fonetik yang kuat), serta menginvestigasi kasus kesalah *retrieval* (misalnya, *query* suatu ayat mencari ayat lain yang memiliki kesamaan *matras* suara yang sangat tinggi) untuk menilai sensitivitas fonetik setiap model. Kesimpulan penelitian yang kemudian dirumuskan utuk menentukan model terbaik dan mengaitkan temuan empiris dengan justrifikasi teoritis struktur *embedding* yang telah dibahas dalam tinjauan pustaka.

Lokasi Penelitian

Lokasi penelitian dapat dilakukan dimana saja dikarenakan penelitian tidak membutuhkan tempat khusus dalam pengambilan data maupun metode yang digunakan.

Jadwal Penelitian

NO	KEGIATAN	MINGGU												HASIL KESELURUHAN
		1	2	3	4	5	6	7	8	9	10	11	12	
1	Studi Literatur													Landasan teori yang kuat mengenai mekanisme *Self-Supervised Learning* (Wav2vec2 dan Data2Vec), validasi kualitas *vector embedding* untuk pencarian fonetik, dan penentuan metrik evaluasi *Similarity Search* (*Top-K Accuracy* dan MAP)
2	Pengumpulan dataset													Dataset siap digunakan (pembersihan, normalisasi, tokenisasi)
3	Pembuatan Embedding (Wav2vec2 & Data2Vec)													Vektor embedding untuk setiap ayat/bacaan Qur'an
4	Implementasi Sistem Pencarian Semantik (Cosine Similarity + FAISS)													Prototipe sistem pencarian berbasis embedding
5	Evaluasi Model													Hasil evaluasi performa model Wav2Vec2 vs Data2Vec

6	Penarikan Kesimpulan & penyusunan laporan													Kesimpulan akhir dan naskah laporan penelitian

DAFTAR PUSTAKA

[1] M. H. Jarrahi, D. L. Blyth, and C. Goray, “Mindful work and mindful technology: Redressing digital distraction in knowledge work,” *Digit. Bus.*, vol. 3, no. 1, p. 100051, June 2023, doi: 10.1016/j.digbus.2022.100051.

[2] Moh. A. Imam Sofii, “Menghafal Al Qur`an Di Era Digital: Problematis Dan Metodologis.,” *Al Furqan J. Ilmu Al Quran Dan Tafsir*, vol. 7, no. 1, pp. 1–17, June 2024, doi: 10.58518/alfurqon.v7i1.2436.

[3] O. Shaklawoon, A. Shafter, M. Abuzaraida, A. Zeki, and Z. Mahmod, *Monitoring the memorization of the Holy Qur’an based on Speech Recognition and NLP Techniques*. 2023.

[4] M. Mutathahirin, A. Jaafar, and N. R. Kamaruzaman, “A Systematic Literature Review (SLR) on Quranic Memorization: Benefits, Methods, and Innovations,” vol. 6, no. 2.

[5] A. A. Harere and K. A. Jallad, “Quran Recitation Recognition using End-to-End Deep Learning,” May 10, 2023, *arXiv*: arXiv:2305.07034. doi: 10.48550/arXiv.2305.07034.

[6] S. Al-Fadhli, H. Al-Harbi, and A. Cherif, “Speech Recognition Models for Holy Quran Recitation Based on Modern Approaches and Tajweed Rules: A Comprehensive Overview,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 12, 2023, doi: 10.14569/IJACSA.2023.0141297.

[7] Y. Shohoud, M. Shoman, and S. Abdelazim, “Quranic Conversations: Developing a Semantic Search tool for the Quran using Arabic NLP Techniques,” Nov. 09, 2023, *arXiv*: arXiv:2311.05120. doi: 10.48550/arXiv.2311.05120.

[8] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, “data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language,” Oct. 25, 2022, *arXiv*: arXiv:2202.03555. doi: 10.48550/arXiv.2202.03555.

[9] O. Mohamed and S. A. Aly, “Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset,” Oct. 09, 2021, *arXiv*: arXiv:2110.04425. doi: 10.48550/arXiv.2110.04425.

[10] N. Vaessen and D. A. van Leeuwen, “Fine-tuning wav2vec2 for speaker recognition,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 7967–7971. doi: 10.1109/ICASSP43922.2022.9746952.

[11] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” Oct. 22, 2020, *arXiv*: arXiv:2006.11477. doi: 10.48550/arXiv.2006.11477.

[12] W. Öberg, “Query-by-Example Audio Search using Acoustic Word Embeddings: Transforming wav2vec 2.0 Embeddings using Contrastive Learning,” 2025.

[13] P. C. English, J. D. Kelleher, and J. Carson-Berndsen, “Searching for Structure: Appraising the Organisation of Speech Features in wav2vec 2.0 Embeddings,” in *Interspeech 2024*, ISCA, Sept. 2024, pp. 4613–4617. doi: 10.21437/Interspeech.2024-2047.

[14] Q. Shao, L. Dong, K. Wei, S. Sun, and L. Xie, “DQ-Data2vec: Decoupling Quantization for Multilingual Speech Recognition,” Jan. 23, 2025, *arXiv*: arXiv:2501.13497. doi: 10.48550/arXiv.2501.13497.

[15] Y. Xie, Z. Zhang, and Y. Yang, “Siamese Network with wav2vec Feature for Spoofing Speech Detection,” in *Interspeech 2021*, ISCA, Aug. 2021, pp. 4269–4273. doi: 10.21437/Interspeech.2021-847.

[16] A. A. Attia, D. Demszky, T. Ogunremi, J. Liu, and C. Espy-Wilson, “CPT-Boosted Wav2vec2.0: Towards Noise Robust Speech Recognition for Classroom Environments,” in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10890830.

[17] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised Cross-lingual Representation Learning for Speech Recognition,” Dec. 15, 2020, *arXiv*: arXiv:2006.13979. doi: 10.48550/arXiv.2006.13979.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5206–5210. doi: 10.1109/ICASSP.2015.7178964.