JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i3.11137

PERBANDINGAN PERFORMA MODEL SISTEM PENGENALAN SUARA TRADISIONAL DAN MODERN PADA BAHASA INDONESIA

Nabila Imesa[1*], Ratna Atika[2], A. Rahman[3]

[1,2,3]Program Studi Sarjana Terapan Teknik Elektro, Politeknik Negeri Sriwijaya; Jl. Srijaya Negara, Bukit Besar, Palembang, Sumatera Selatan 30139, Indonesia; Telp/Fax (0711) 353414

Keywords:
Automatic Speech Recognition, Bahasa Indonesia, Transformer Vanilla, Word Error Rate, Robot

Corespondent Email:
nabilaimesa11@gmail.com

Abstrak. Sistem pengenalan suara merupakan komponen penting pada robot untuk mendukung komunikasi alami antara manusia dan robot. Keterbatasan dataset berbahasa Indonesia dapat memengaruhi kinerja model pengenalan suara, sehingga diperlukan analisis untuk menentukan model yang sesuai. Penelitian ini bertujuan membandingkan performa empat model, yaitu *Hidden Markov Model* (HMM), HMM-*Gaussian Mixture Model* (HMM-GMM), HMM-GMM-*Deep Neural Network* (HMM-GMM-DNN), dan *Transformer Vanilla* menggunakan dataset suara bahasa Indonesia yang dikumpulkan secara mandiri. Dataset terdiri atas 1.250 data audio dari 50 kalimat, yang meliputi 25 kalimat tanya dan 25 kalimat perintah dengan 25 kali pengulangan. Model HMM, HMM-GMM, dan HMM-GMM-DNN dilatih menggunakan *Kaldi Toolkit*, sedangkan *Transformer Vanilla* diimplementasikan menggunakan *Python*. Evaluasi dilakukan menggunakan *Word Error Rate* (WER) dan *Character Error Rate* (CER). Hasil penelitian menunjukkan bahwa *Transformer Vanilla* memiliki performa terbaik dengan WER sebesar 0,0222 dan CER sebesar 0,0164, dibandingkan HMM, HMM-GMM, dan HMM-GMM-DNN. Hasil ini menunjukkan bahwa *Transformer Vanilla* lebih efektif dalam mengenali ucapan bahasa Indonesia pada dataset penelitian.

Abstract. *Speech recognition is an important component of robots to support natural communication between humans and robots. The limited availability of Indonesian-language datasets can affect the performance of speech recognition models; therefore, an analysis is needed to determine the most suitable model. This study aims to compare the performance of four models, namely Hidden Markov Model (HMM), HMM-Gaussian Mixture Model (HMM-GMM), HMM-GMM-Deep Neural Network (HMM-GMM-DNN), and Vanilla Transformer using a self-collected Indonesian speech dataset. The dataset consists of 1,250 audio samples from 50 sentences, including 25 interrogative sentences and 25 command sentences, with 25 repetitions each. The HMM, HMM-GMM, and HMM-GMM-DNN models were trained using the Kaldi Toolkit, while the Vanilla Transformer was implemented using Python. Performance evaluation was conducted using Word Error Rate (WER) and Character Error Rate (CER). The results show that the Vanilla Transformer achieved the best performance, with a WER of 0.0222 and a CER of 0.0164, outperforming HMM, HMM-GMM, and HMM-GMM-DNN. These results indicate that the Vanilla Transformer is more effective in recognizing Indonesian speech in the research dataset.*

1. PENDAHULUAN

Interaksi manusia-robot (*Human-Robot Interaction*/HRI) berbasis suara telah menjadi salah satu modalitas komunikasi paling alami dan banyak diminati dalam pengembangan robot layanan, karena memungkinkan

pengguna berinteraksi tanpa memerlukan antarmuka fisik seperti tombol atau layar sentuh [1]. Efektivitas interaksi tersebut sangat ditentukan oleh persepsi pengguna terhadap kemampuan robot dalam memahami maksud ucapan secara cepat dan akurat, sehingga kegagalan sistem dalam mengenali perintah suara dapat menurunkan kepercayaan pengguna terhadap robot secara keseluruhan [2]. Oleh karena itu, komponen *Automatic Speech Recognition* (ASR) menjadi salah satu subsistem paling krusial pada robot pelayan, terutama untuk aplikasi pada lingkungan layanan publik seperti restoran, perkantoran, rumah sakit, maupun fasilitas pendidikan yang menuntut respons cepat dan akurat terhadap perintah verbal pengguna [3]. Robot pelayan pada umumnya perlu memahami dua jenis ujaran yang paling sering muncul dalam interaksi sehari-hari, yaitu kalimat tanya untuk meminta informasi dan kalimat perintah untuk meminta robot melakukan suatu tindakan, sehingga akurasi ASR pada kedua jenis kalimat tersebut menjadi tolok ukur penting keberhasilan sistem interaksi suara secara keseluruhan.

Pengembangan sistem ASR pada dasarnya bertumpu pada dua pendekatan besar. Pendekatan pertama adalah model statistik konvensional berbasis *Hidden Markov Model* (HMM) yang memodelkan hubungan temporal antara unit fonetik dan fitur akustik, umumnya dikombinasikan dengan *Gaussian Mixture Model* (GMM) sebagai model emisi. Pendekatan ini menjadi fondasi ASR selama beberapa dekade karena kebutuhan komputasi yang relatif rendah, tetapi performanya sangat bergantung pada kualitas dan jumlah data pelatihan yang tersedia [4]. Pendekatan kedua adalah model berbasis deep learning, di mana *Deep Neural Network* (DNN) menggantikan GMM sebagai model akustik untuk mengekstrak fitur yang lebih kaya dan mampu menangani variasi suara lintas bahasa maupun dialek dengan lebih baik [5]. Kombinasi antara HMM, GMM, dan DNN pada arsitektur hybrid HMM-GMM-DNN diharapkan dapat memanfaatkan keunggulan masing-masing komponen, yaitu kemampuan HMM dalam memodelkan struktur sekuensial ujaran serta kemampuan DNN dalam mempelajari representasi fitur akustik yang lebih kompleks dan diskriminatif dibandingkan model statistik murni.

Dalam beberapa tahun terakhir, arsitektur *Transformer* yang memanfaatkan mekanisme *self-attention* telah membawa perubahan besar pada bidang ASR karena kemampuannya menangkap ketergantungan jangka panjang secara paralel tanpa bergantung pada struktur rekuren, sehingga memungkinkan pengembangan sistem end-to-end yang lebih sederhana dan akurat dibandingkan pendekatan konvensional [6], [7]. Namun demikian, performa model-model tersebut sangat bergantung pada ketersediaan dataset berkualitas tinggi, dan bahasa Indonesia masih tergolong sebagai bahasa dengan sumber daya terbatas (*low-resource language*) dalam konteks pengembangan ASR, sehingga penelitian yang secara khusus mengevaluasi berbagai algoritma pada dataset berbahasa Indonesia masih sangat dibutuhkan [8]. Beberapa penelitian terbaru telah menunjukkan potensi besar arsitektur *Transformer* untuk bahasa Indonesia; modifikasi arsitektur *Transformer* dengan penambahan blok konvolusional dan *Vision Transformer* (ViT) misalnya berhasil menurunkan WER menjadi 0,158 dan CER menjadi 0,118 pada dataset NSS-ID [9], sementara penelitian lanjutan turut membandingkan performa Transformer Vanilla dengan model Whisper untuk tugas speech-to-text bahasa Indonesia[10].

Meskipun berbagai algoritma ASR telah banyak diteliti secara terpisah, kajian yang secara khusus membandingkan performa model tradisional (HMM, HMM-GMM), model hybrid (HMM-GMM-DNN), dan model modern (*Transformer Vanilla*) pada satu dataset bahasa Indonesia yang identik masih sangat terbatas. Sebagian besar penelitian yang ada hanya menguji satu algoritma atau satu jenis dataset secara terpisah, sehingga belum memberikan gambaran menyeluruh mengenai bagaimana keempat pendekatan tersebut bekerja pada kondisi data yang sama, khususnya untuk konteks aplikasi robot pelayan yang memerlukan pengenalan kalimat tanya dan kalimat perintah secara akurat.

Berdasarkan uraian tersebut, penelitian ini bertujuan mengisi kesenjangan tersebut melalui studi komparatif yang membandingkan performa empat model ASR, yaitu *Hidden Markov M*odel (HMM), *Hidden Markov Model-*

Gaussian Mixture Model (HMM-GMM), *Hidden Markov Model-Gaussian Mixture Model-Deep Neural Network* (HMM-GMM-DNN), dan *Transformer Vanilla*, menggunakan dataset suara bahasa Indonesia yang dikumpulkan secara khusus untuk merepresentasikan komunikasi antara manusia dan robot pelayan berupa kalimat tanya dan kalimat perintah. Evaluasi performa dilakukan menggunakan parameter *Word Error Rate* (WER) dan *Character Error Rate* (CER) pada setiap model, sehingga dapat diketahui algoritma yang paling sesuai untuk diterapkan pada subsistem komunikasi suara robot.

2. TINJAUAN PUSTAKA

2.1 *Automatic Speech Recognition* (ASR)

Automatic Speech Recognition (ASR) adalah teknologi yang memungkinkan sistem komputer secara otomatis mengubah sinyal ucapan manusia menjadi teks atau perintah yang dapat diproses mesin. Sistem ASR modern umumnya terdiri atas beberapa komponen utama yang saling terintegrasi, yaitu modul ekstraksi fitur akustik, model akustik, model bahasa, dan dekoder [11]. Proses kerja ASR diawali dari akuisisi sinyal suara menggunakan mikrofon, yang kemudian didigitasi dan diproses melalui rangkaian tahapan pemrosesan sinyal; pada tahap ekstraksi fitur, sinyal suara direpresentasikan dalam bentuk vektor fitur yang menonjolkan informasi fonetik relevan, seperti *Mel-Frequency Cepstral Coefficients* (MFCC). Model akustik memetakan vektor fitur tersebut ke unit-unit fonetik, sedangkan model bahasa memilih urutan kata yang paling mungkin berdasarkan konteks linguistik, dan proses dekoding menggabungkan keduanya untuk menghasilkan transkripsi teks terbaik. Perkembangan ASR telah berevolusi dari pendekatan berbasis model statistik seperti HMM-GMM menuju arsitektur deep learning yang sepenuhnya terintegrasi (*end-to-end*), di mana komponen akustik dan linguistik dilatih bersama dalam satu kerangka jaringan saraf [11].

2.2 *Hidden Markov Model* (HMM)

Hidden Markov Model (HMM) adalah model statistik yang memodelkan sistem dengan state tersembunyi yang tidak dapat diamati langsung, namun dapat diestimasi berdasarkan sekuens observasi yang dihasilkan. Parameter HMM didefinisikan sebagai $\lambda = (A, B, \pi)$, dengan A adalah matriks probabilitas transisi antar-state, B adalah probabilitas emisi setiap observasi dari setiap state, dan π adalah distribusi probabilitas state awal [12]. Dalam sistem ASR, HMM memodelkan hubungan antara unit fonetik sebagai state tersembunyi dan fitur akustik MFCC sebagai observasi, dengan proses optimasi parameter menggunakan algoritma Baum-Welch berbasis Expectation-Maximization (EM) secara iteratif.

2.3 *Gaussian Mixture Model* (GMM)

Gaussian Mixture Model (GMM) merupakan model probabilistik yang merepresentasikan distribusi data sebagai superposisi beberapa distribusi *Gaussian,* digunakan sebagai model emisi pada HMM untuk memodelkan distribusi statistik fitur MFCC yang kompleks pada setiap state fonem, menggantikan distribusi Gaussian tunggal yang kurang fleksibel [13]. Setiap komponen *Gaussian* pada GMM ditentukan oleh parameter rata-rata (μ), kovarians (Σ), dan bobot campuran (π), yang diestimasi melalui algoritma EM yang mengalternasikan *E-Step* untuk menghitung probabilitas keanggotaan setiap titik data, dan M-Step untuk memperbarui parameter hingga konvergen. Kombinasi HMM-GMM memungkinkan pemodelan distribusi fitur akustik yang lebih representatif dan fleksibel dibandingkan HMM dengan emisi tunggal.

2.4 *Deep Neural Network* (DNN)

Deep Neural Network (DNN) merupakan pengembangan dari *Artificial Neural Network* yang terdiri atas banyak lapisan tersembunyi (multi-layer) yang mampu mempelajari representasi fitur secara hierarkis dari data mentah, dengan struktur berupa lapisan input, beberapa hidden layer, dan lapisan output, di mana setiap neuron melakukan operasi perkalian bobot, penambahan bias, dan penerapan fungsi aktivasi non-linear [14]. Dalam konteks ASR, DNN menggantikan komponen GMM pada model HMM-GMM untuk memperkirakan probabilitas posterior state fonem berdasarkan fitur akustik. Model HMM-GMM-DNN memanfaatkan keunggulan HMM dalam pemodelan sekuensial, GMM sebagai penyelaras awal yang menghasilkan label per frame, dan DNN sebagai pengekstrak

fitur yang lebih kuat melalui proses forward propagation dan backpropagation.

2.5 *Transformer Vanilla*

Arsitektur *Transformer* memanfaatkan mekanisme *self-attention* untuk memodelkan ketergantungan jangka panjang dalam data sekuensial tanpa bergantung pada struktur rekuren maupun konvolusional [15]. Dalam sistem ASR, *Transformer Vanilla* bekerja secara *end-to-end* dengan memetakan representasi akustik langsung menjadi teks transkripsi. Arsitektur ini terdiri atas *encoder* dan *decoder* yang masing-masing dilengkapi mekanisme *multi-head self-attention*, *feed-forward network,* dan lapisan normalisasi (*Add & Norm*). Input berupa *log-Mel spectrogram* diproses melalui linear projection dan positional *encoding* sebelum masuk ke tumpukan lapisan *encoder*, sedangkan *decoder* menambahkan masked *multi-head attention* serta *encoder-decoder attention* untuk menghasilkan prediksi token transkrip secara bertahap [7]. Beberapa penelitian menunjukkan bahwa arsitektur *Transformer* yang dirancang secara efisien mampu mempertahankan akurasi pengenalan suara yang tinggi dengan kompleksitas komputasi yang lebih rendah dibandingkan model *Transformer* berskala besar [16].

2.6 *Word Error Rate* dan *Character Error Rate*

Evaluasi sistem ASR dilakukan menggunakan dua metrik utama. *Word Error Rate* (WER) mengukur proporsi kesalahan pada tingkat kata, sedangkan *Character Error Rate* (CER) menggunakan formulasi yang sama namun dihitung pada tingkat karakter [13], [17]. Kedua metrik dihitung menggunakan persamaan berikut:

$$\text{WER} = = \frac{(\text{S} + \text{D} + \text{I})}{\text{N}} \qquad (1)$$

$$\text{CER} = \frac{(\text{S} + \text{D} + \text{I})}{\text{N}} \qquad (2)$$

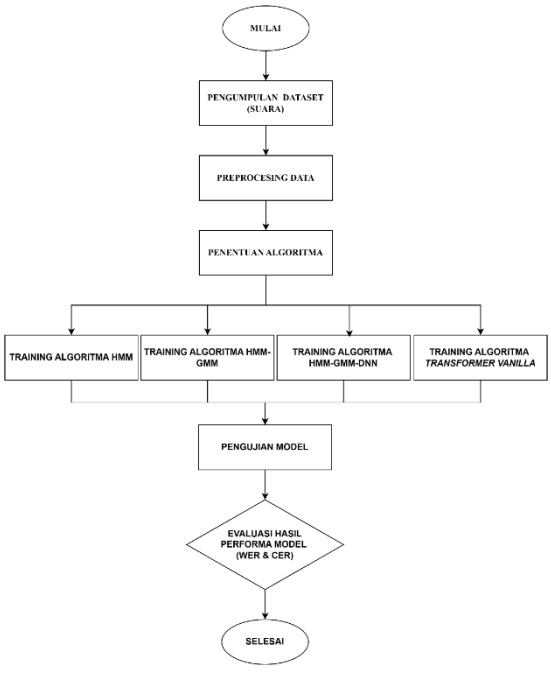
dengan S adalah jumlah kata/karakter yang disubstitusi (*substitution*), D adalah jumlah kata/karakter yang hilang (*deletion*), I adalah jumlah kata/karakter tambahan (*insertion*), dan N adalah jumlah total kata/karakter pada teks referensi. Nilai WER dan CER berkisar antara 0 dan 1, dengan nilai yang lebih kecil menunjukkan performa pengenalan suara yang lebih baik. Penggunaan kedua metrik secara bersamaan memberikan gambaran komprehensif mengenai akurasi sistem, di mana WER lebih sensitif terhadap kesalahan kata secara keseluruhan sedangkan CER memberikan analisis kesalahan yang lebih granular pada tingkat karakter [13], [17].

2.7 *Kaldi Toolkit* dan *Python*

Kaldi merupakan toolkit *open-source* berbasis *machine learning* dan *deep learning* yang banyak digunakan untuk mengembangkan sistem ASR karena menyediakan modul ekstraksi fitur, pelatihan model akustik, integrasi model bahasa, dan decoding secara menyeluruh [18]. Adapun Python digunakan sebagai bahasa pemrograman untuk mengimplementasikan model *Transformer Vanilla* karena sintaksnya yang sederhana serta dukungan library seperti PyTorch dan NumPy yang mempermudah pengembangan sistem berbasis deep learning [19].

3. METODE PENELITIAN

Penelitian ini diawali dengan tahap pengumpulan data, dilanjutkan dengan *preprocessing* data, penentuan algoritma, pelatihan (*training*) data, pengujian model, dan evaluasi performa menggunakan metrik WER dan CER. Tahapan penelitian secara lengkap ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

3.1 Dataset Penelitian

Dataset suara bahasa Indonesia disusun secara mandiri untuk merepresentasikan komunikasi antara manusia dan robot humanoid pelayan. Kalimat yang digunakan dibagi menjadi dua kategori, yaitu 25 kalimat tanya dan 25 kalimat perintah, sehingga total terdapat 50 kalimat penelitian. Seluruh kalimat direkam oleh satu penutur perempuan dengan 25 kali pengulangan pada setiap kalimat, sehingga diperoleh 1.250 berkas audio. Perekaman dilakukan di Laboratorium Politeknik Negeri Sriwijaya pada ruang yang dilengkapi busa peredam suara menggunakan mikrofon yang terhubung ke laptop dengan perangkat lunak Audacity, format audio WAV, sampling rate 16 kHz, kanal mono, dan resolusi 16-bit.

3.2 *Preprocessing* Data

Tahap preprocessing bertujuan meningkatkan kualitas data, menyeragamkan format audio, serta mempersiapkan representasi fitur sesuai karakteristik masing-masing model. Proses ini meliputi pengolahan sinyal audio, normalisasi teks, tokenisasi, dan ekstraksi fitur. Model HMM, HMM-GMM, dan HMM-GMM-DNN menggunakan fitur *Mel-Frequency Cepstral Coefficients* (MFCC), sedangkan model Transformer Vanilla menggunakan *log-Mel spectrogram* sebagai representasi masukan.

3.3 Penentuan *Algoritma dan Training* Data

Algoritma yang dibandingkan pada penelitian ini mencakup *Hidden Markov Model* (HMM), *Hidden Markov Model-Gaussian Mixture Model* (HMM-GMM), *Hidden Markov Model-Gaussian Mixture Model-Deep Neural Network* (HMM-GMM-DNN), dan *Transformer Vanilla*. Pemilihan keempat algoritma tersebut bertujuan membandingkan cara kerja model statistik tradisional, model hybrid berbasis *deep learning*, dan model *modern* berbasis *attention* dalam mengenali ucapan bahasa Indonesia. Setiap model dilatih secara terpisah menggunakan dataset yang sama agar perbedaan performa yang diperoleh lebih dipengaruhi oleh karakteristik algoritma dibandingkan oleh perbedaan data. Proses pelatihan model HMM, HMM-GMM, dan HMM-GMM-DNN dilakukan menggunakan *Kaldi Toolkit* melalui skrip *train_mono.sh, train_deltas.sh*, dan *train_pnorm_fast.sh* secara berurutan, sedangkan model *Transformer Vanilla* diimplementasikan menggunakan *framework PyTorch* pada bahasa pemrograman *Python.*

Pelatihan model HMM diawali dengan ekstraksi fitur MFCC dari data audio serta normalisasi transkrip untuk membangun *lexicon* dan *language* model. Parameter model dioptimasi menggunakan algoritma *Baum-Welch* (*Expectation-Maximization*) secara iteratif hingga nilai *log-likelihood konvergen*. Pada model HMM-GMM, proses diawali dengan alignment menggunakan model HMM monophone, dilanjutkan dengan inisialisasi GMM untuk memodelkan distribusi fitur akustik secara lebih fleksibel. Pada model HMM-GMM-DNN, hasil *forced alignment* dari model HMM-GMM digunakan sebagai target pelatihan DNN, dengan proses *forward propagation*, perhitungan *Cross-Entropy Loss,* dan *backpropagation* menggunakan optimizer Adam. Adapun model *Transformer Vanilla* dilatih secara *end-to-end*: fitur *log-Mel spectrogram* dan token transkrip diproses melalui input *embedding* dan *positional encoding*, kemudian direpresentasikan oleh *encoder* dan *decoder* berbasis *multi-head self-attention,* dengan *Cross-Entropy Loss* yang dikombinasikan dengan label smoothing serta optimizer Adam untuk memperbarui parameter model pada setiap epoch.

3.4 Evaluasi Performa Model

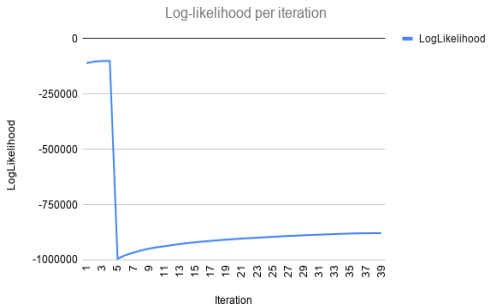
Setelah proses pelatihan selesai, setiap model diuji menggunakan data testing yang sama dan tidak digunakan selama pelatihan maupun validasi, sehingga hasil keempat model dapat dibandingkan secara objektif. Transkripsi hasil prediksi setiap model kemudian dibandingkan dengan transkrip referensi (*ground truth*) untuk memperoleh nilai *Word Error Rate* (WER) dan *Character Error Rate* (CER) menggunakan *skrip score.sh* dan *score_cer.sh* pada *Kaldi Toolkit* untuk model berbasis HMM, serta perhitungan WER/CER secara terprogram untuk model *Transformer Vanilla*. Nilai WER dan CER dari keempat model selanjutnya dibandingkan untuk menentukan algoritma dengan performa pengenalan suara terbaik pada dataset penelitian.

4. HASIL DAN PEMBAHASAN

Pelatihan model dilakukan menggunakan empat model pengenalan suara, yaitu *Hidden Markov Model* (HMM), *HMM-Gaussian Mixture Model* (HMM-GMM), *HMM-GMM-Deep Neural Network* (HMM-GMM-DNN), dan *Transformer Vanilla*. Model HMM, HMM-GMM, dan HMM-GMM-DNN dilatih menggunakan *Kaldi Toolkit,* sedangkan model *Transformer Vanilla* diimplementasikan menggunakan *Python*. Evaluasi kinerja masing-masing model dilakukan menggunakan *Word Error Rate* (WER) dan *Character Error Rate* (CER) sebagai dasar perbandingan performa antar model. Pada tahap pengumpulan data, diperoleh dataset mentah sebanyak 1.250 data audio yang terdiri atas 50 kalimat, yaitu 25 kalimat tanya dan 25 kalimat perintah dengan masing-masing kalimat direkam sebanyak 25 kali. Dataset tersebut kemudian digunakan sebagai data pelatihan dan pengujian untuk membandingkan kemampuan masing-masing model dalam mengenali ucapan bahasa Indonesia.

4.1 *Hidden Markov Model* (HMM)

Pelatihan model HMM dilakukan menggunakan fitur MFCC sebagai representasi sinyal ujaran. Selama proses pelatihan, parameter model dioptimalkan menggunakan algoritma *Baum–Welch* dengan tujuan memaksimalkan nilai *log-likelihood*. Perkembangan nilai *log-likelihood* pada setiap iterasi digunakan sebagai indikator untuk mengetahui proses konvergensi model. Hasil pelatihan model HMM ditunjukkan pada Gambar 2

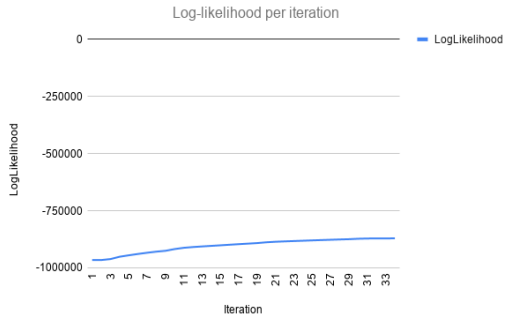


Gambar 2 Grafik *Log-likelihood* Model HMM

Berdasarkan Gambar 2, pada iterasi awal terjadi penurunan nilai *log-likelihood* yang cukup signifikan sebagai bagian dari proses inisialisasi parameter model menggunakan algoritma *Baum–Welch.* Setelah proses inisialisasi selesai, nilai *log-likelihood* meningkat secara bertahap hingga mencapai kondisi yang relatif stabil pada iterasi ke-39. Peningkatan tersebut menunjukkan bahwa parameter model semakin mampu merepresentasikan hubungan antara fitur akustik dan fonem pada data pelatihan.

4.2 HMM-GMM.

Proses pelatihan model HMM-GMM dikembangkan dengan menambahkan *Gaussian Mixture Model* (GMM) pada setiap state HMM untuk memperoleh representasi distribusi fitur akustik yang lebih baik. Proses pelatihan dilakukan menggunakan algoritma *Expectation Maximization* (EM), sedangkan perkembangan nilai *log-likelihood* digunakan sebagai indikator konvergensi model. Hasil pelatihan ditunjukkan pada Gambar 3.



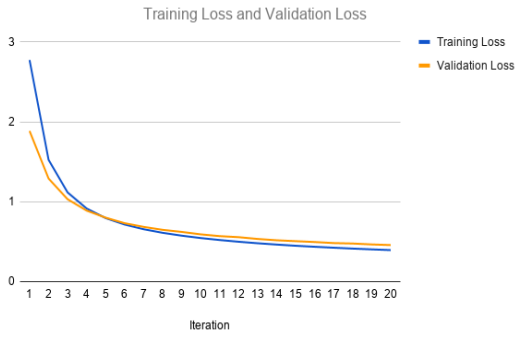
Gambar 3 Grafik *Log-likelihood* Model HMM-GMM

Berdasarkan Gambar 3, nilai *log-likelihood* meningkat secara konsisten sejak iterasi pertama hingga mencapai kondisi konvergen pada iterasi ke-33. Dibandingkan model HMM, kurva *log-likelihood* HMM-GMM menunjukkan pola yang lebih stabil tanpa penurunan yang signifikan pada tahap awal pelatihan. Hal ini menunjukkan bahwa penambahan komponen GMM mampu memberikan representasi distribusi fitur akustik yang lebih baik sehingga proses optimasi parameter berlangsung lebih stabil.

4.3 HMM-GMM-DNN

Model ini memanfaatkan hasil *alignment* dari model HMM-GMM sebagai target supervisi dalam proses pelatihan jaringan saraf. Performa pelatihan diamati melalui

perkembangan nilai *training loss* dan *validation loss* pada setiap epoch. Hasil pelatihan model ditunjukkan pada Gambar 4.

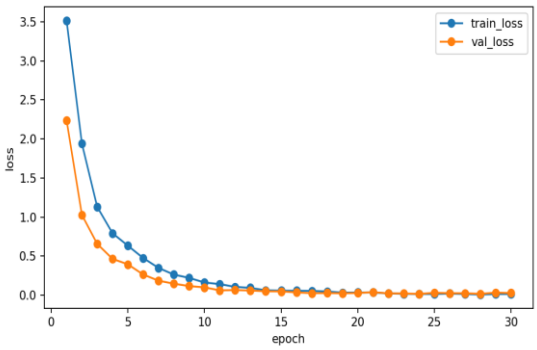


Gambar 4 Grafik Training dan Validation Loss Model

Berdasarkan Gambar 4, nilai *training loss* menurun secara bertahap dari sekitar 2,8 pada awal pelatihan menjadi sekitar 0,4 pada epoch ke-20. Nilai *validation loss* juga menunjukkan tren penurunan yang serupa dengan selisih yang kecil terhadap *training loss*. Pola tersebut menunjukkan bahwa model dapat mempelajari karakteristik data pelatihan tanpa menunjukkan indikasi *overfitting* yang signifikan. Penurunan *loss* yang baik belum diikuti oleh peningkatan performa pengenalan suara secara keseluruhan. Hal ini terlihat dari nilai WER yang masih lebih tinggi dibandingkan model HMM-GMM, yang mengindikasikan bahwa ukuran dan variasi dataset belum cukup untuk mengoptimalkan kemampuan jaringan saraf secara maksimal.

4.4 Model *Transformer Vanilla*

Pelatihan model *Transformer Vanilla* dilakukan menggunakan fitur *log-Mel spectrogram* sebagai masukan utama. Selama proses pelatihan, perkembangan *training loss* dan *validation loss* digunakan untuk mengevaluasi proses optimasi model. Hasil pelatihan model *Transformer Vanilla* ditunjukkan pada Gambar 5.



Gambar 5 Grafik *Training* dan *Validation Loss* Model *Transformer Vanilla*

Berdasarkan Gambar 5, nilai *training loss* mengalami penurunan yang sangat cepat dari sekitar 3,5 pada epoch pertama menjadi kurang dari 0,1 pada epoch ke-15, kemudian terus menurun hingga mendekati 0 pada epoch ke-30. Nilai *validation loss* juga menunjukkan pola penurunan yang konsisten, yaitu dari sekitar 2,2 pada epoch pertama hingga mendekati 0 pada epoch ke-30, dengan selisih yang sangat kecil terhadap *training loss*. Penurunan yang signifikan pada beberapa epoch awal menunjukkan kemampuan mekanisme *multi-head self-attention* dalam menangkap pola kompleks data suara bahasa Indonesia. Konvergensi dalam mendekati nilai 0 mengindikasikan bahwa model dapat mempelajari karakteristik data dengan cepat, sedangkan kestabilan kedua kurva pada epoch-epoch akhir mengindikasikan bahwa proses pelatihan telah mencapai konvergensi tanpa menunjukkan *overfitting* yang signifikan. Dibandingkan ketiga model lainnya, *Transformer Vanilla* menunjukkan proses optimasi yang paling stabil, sehingga mampu menghasilkan representasi fitur yang lebih baik dan berkontribusi terhadap nilai WER dan CER yang paling rendah pada tahap evaluasi

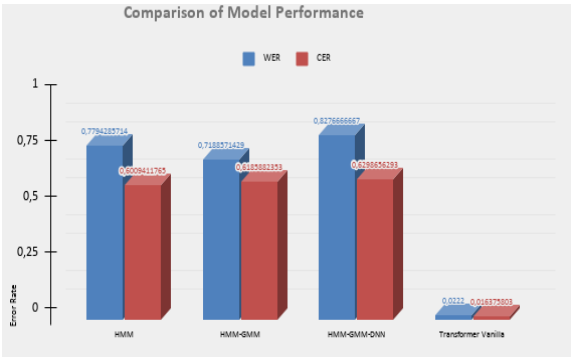
4.5 Evaluasi dan Perbandingan Performa Model

Evaluasi dilakukan untuk mengukur performa dari masing-masing model dalam mengenali *speech* Bahasa Indonesia menggunakan data pengujian (*testing)*. Performa model dievaluasi dengan membandingkan hasil pengenalan model terhadap transkrip referensi menggunakan metrik *Word Error Rate* (WER) dan *Character Error Rate* (CER). Nilai WER digunakan untuk

mengukur tingkat kesalahan pada tingkat kata, sedangkan CER digunakan untuk mengukur tingkat kesalahan pada tingkat karakter. Semakin kecil nilai WER dan CER yang diperoleh, semakin baik kemampuan model dalam mengenali suara. Hasil evaluasi keempat model ditunjukkan pada Tabel 1 dan Gambar 6.

Tabel 1 Hasil Perbandingan Performa Keempat Model

Model	WER	CER
HMM	0,7794	0,6009
HMM-GMM	0,7189	0,6126
HMM-GMM-DNN	0,8277	0,6299
Transformer Vanilla	0,0222	0,0164



Gambar 6 Perbandingan Performa setiap Model

Berdasarkan Tabel 1 dan Gambar 6, hasil penelitian menunjukkan bahwa *Transformer Vanilla* memperoleh nilai WER dan CER terendah dibandingkan ketiga model lainnya, yaitu WER sebesar 0,0222 dan CER sebesar 0,0164. Hasil ini mengindikasikan bahwa mekanisme *self-attention* pada *Transformer Vanilla* lebih efektif dalam mengenali *speech* pada dataset sehingga menghasilkan akurasi pengenalan yang lebih tinggi.

Di antara algoritma model tradisional, HMM-GMM memperoleh nilai WER yang lebih rendah dibandingkan HMM, yaitu 0,7189 dan 0,7794. Hal ini menunjukkan bahwa penambahan *Gaussian Mixture Model* (GMM) mampu meningkatkan representasi distribusi fitur akustik sehingga proses pengenalan suara menjadi lebih baik terbukti memberikan representasi distribusi fitur akustik yang lebih baik.

Sebaliknya, hasil HMM-GMM-DNN nilai WER 0,8277 dan CER 0,6299 tertinggi. Meskipun proses pelatihannya menunjukkan penurunan *loss* yang stabil, performa model belum optimal. Kondisi ini diduga dipengaruhi oleh ukuran dan variasi dataset yang terbatas, sehingga pembelajaran jaringan saraf belum mampu melakukan generalisasi secara efektif.

5. KESIMPULAN

a. Model modern *Transformer Vanilla* terbukti lebih unggul dibandingkan model tradisional HMM, HMM-GMM, dan model hybrid HMM-GMM-DNN pada dataset speech bahasa Indonesia yang digunakan, dengan WER sebesar 0,0222 dan CER sebesar 0,0164, dibandingkan HMM (WER 0,7794; CER 0,6009), HMM-GMM (WER 0,7189; CER 0,6186), dan HMM-GMM-DNN (WER 0,8277; CER 0,6299).
b. Jumlah dan keragaman dataset berpengaruh terhadap performa model selama pelatihan maupun pengujian; dataset yang lebih besar dan beragam berpotensi meningkatkan kemampuan model dalam mempelajari karakteristik ucapan bahasa Indonesia.
c. Untuk pengembangan selanjutnya, disarankan menggunakan dataset dengan jumlah data dan variasi penutur (jenis kelamin, usia, karakteristik suara) yang lebih besar, membandingkan arsitektur ASR yang lebih mutakhir seperti Conformer, Whisper, atau wav2vec 2.0, serta mengimplementasikan langsung model terbaik pada perangkat keras robot pelayan untuk pengujian pada kondisi interaksi nyata

UCAPAN TERIMA KASIH

Penulis memanjatkan puji dan syukur kepada Tuhan Yang Maha Esa atas rahmat, kekuatan, dan kemudahan yang diberikan selama proses penelitian. Penulis juga menyampaikan terima kasih kepada orang tua serta seluruh pihak yang telah memberikan doa, dukungan, bantuan, dan motivasi sehingga penelitian ini dapat terlaksana dan diselesaikan dengan baik.

DAFTAR PUSTAKA

[1] M. M. Reimann, F. A. Kunneman, C. Oertel, and K. V. Hindriks, "A Survey on Dialogue Management in Human-robot Interaction," *ACM Trans. Hum. Robot. Interact.*, vol. 13, no. 2, Jun. 2024, doi: 10.1145/3648605.

[2] Z. Janeczko and M. E. Foster, "A Study on Human Interactions with Robots Based on Their Appearance and Behaviour," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jul. 2022. doi: 10.1145/3543829.3544523.

[3] N. Mawaddah, C. R. Tamba, D. Hutagalung, F. Nainggolan, E. Lubis, and E. Jamzuri, "Automatic Speech Recognition for Human-Robot Interaction on The Humanoid Robot Barelang 7," European Alliance for Innovation n.o., Feb. 2024. doi: 10.4108/eai.7-11-2023.2342940.

[4] T. F. Abidin, A. Misbullah, R. Ferdhiana, L. Farsiah, M. Z. Aksana, and H. Riza, "Acoustic Model with Multiple Lexicon Types for Indonesian Speech Recognition," *Applied Computational Intelligence and Soft Computing*, vol. 2022, 2022, doi: 10.1155/2022/3227828.

[5] K. Nugroho, E. Noersasongko, Purwanto, Muljono, and D. R. I. M. Setiadi, "Enhanced Indonesian Ethnic Speaker Recognition using Data Augmentation Deep Neural Network," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4375–4384, Jul. 2022, doi: 10.1016/j.jksuci.2021.04.002.

[6] A. Gulati *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," May 2020, [Online]. Available: http://2005.08100

[7] A. Loubser, P. De Villiers, and A. De Freitas, "End-to-end automated speech recognition using a character based small scale transformer architecture," *Expert Syst. Appl.*, vol. 252, Oct. 2024, doi: 10.1016/j.eswa.2024.124119.

[8] A. Adila, D. Lestari, A. Purwarianti, D. Tanaya, K. Azizah, and S. Sakti, "Enhancing Indonesian Automatic Speech Recognition: Evaluating Multilingual Models with Diverse Speech Variabilities," Oct. 2024, [Online]: http://2410.08828

[9] R. Atika, S. Dwijayanti, and B. Y. Suprapto, "Improving Speech-To-Text For The Indonesian Language Using A Modified Transformer," *Eastern-European Journal of Enterprise Technologies*, vol. 1, no. 9 (139), pp. 78–90, 2026, doi: 10.15587/1729-4061.2026.350949.

[10] R. Atika, S. Dwijayanti, and B. Y. Suprapto, "Comparing Transformer Models for Indonesian Speech-to-Text: Vanilla and Whisper," in *2025 International Conference on Artificial Intelligence and Technological Solutions: For Good Health, Well-Being, and Sustainable Water Management in Support of SDGs 3, 6, and 9, ICAITech 2025 - Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 328–334. doi: 10.1109/ICAITech66481.2025.11387506.

[11] H. Ahlawat, N. Aggarwal, and D. Gupta, "Automatic Speech Recognition: A survey of deep learning techniques and approaches," Dec. 01, 2025, *KeAi Communications Co.* doi: 10.1016/j.ijcce.2024.12.007.

[12] T. F. Abidin, A. Misbullah, R. Ferdhiana, L. Farsiah, M. Z. Aksana, and H. Riza, "Acoustic Model with Multiple Lexicon Types for Indonesian Speech Recognition," *Applied Computational Intelligence and Soft Computing*, vol. 2022, 2022, doi: 10.1155/2022/3227828.

[13] C. Park, M. Chen, and T. Hain, "Automatic Speech Recognition System-Independent Word Error Rate Estimation," May 2024.

[14] P. Chhabra and S. Goyal, "A Thorough Review on Deep Learning Neural Network," Jan. 2023.

[15] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 2023, [Online]. Available: http://1706.03762

[16] A. R. Sajun, I. Zualkernan, and D. Sankalpa, "A Historical Survey of Advances in Transformer Architectures," May 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app14104316.

[17] S. Zhang, J. Hale, M. Renwick, Z. Vrzić, and K. Langston, "An Evaluation of Croatian ASR Models for Čakavian Transcription," May 2024.

[18] J. Linke, S. Wepner, G. Kubin, and B. Schuppler, "Using Kaldi for Automatic Speech Recognition of Conversational Austrian German," Jan. 2023, [Online]. Available: http://2301.06475

[19] M. Müller, "Leveraging Python in AI and Machine Learning: A Survey of Techniques and Educational Approaches in Software Engineering," Jan. 2025. [Online]. Available: www.ijeais.org/ijeais