BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian ini mengacu pada beberap studi terdahulu yang relevan dengan penerjemahan bahasa isyarat. Sebagian besar penelitian fokus pada recognition meski penelitian ini audio/teks ke gestur, meski begitu studi tersebut bisa memberikan fondasi utama untuk sistem yang akan dibangun.

2.1.1 *Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks [8]*

Penelitian dengan judul "*Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks*" yang dilakukan oleh Stoll, Stephanie pada tahun 2020 adalah penelitian yang mengusulkan sistem penerjemahan teks ke bahasa isyarat dengan menggunakan pendekatan NMT (*Neural Machine Translation*) dan GAN (*Generative Adversarial Networks*). Prosesnya melewati dua tahap. Pertama, sistem akan men-terjemahkan teks ke dalam bahasa alami menjadi urutan gloss (representasi kata dalam bahasa isyarat), kemudian gloss tersebut dikonversi menjadi urutan pose tubuh dalam bentuk skeleton menggunakan jaringan generator berbasis GAN. Hasil pose skeleton selanjutnya dapat diubah menjadi video avatar atau animasi 3D. Penelitian ini diuji pada dataset German Sign Language dan menunjukkan peningkatan signifikan dalam naturalitas gerakan disbanding baseline berbasis RNN. Ketergantungan pada dataset yang sangat besar dan teranotasi lengkap menjadi tantangan besar untuk implementasi pada bahasa isyarat low-resource seperti BISINDO.

Beberapa poin penting yang bisa diambil:

1. Pendekatan dua tahap (text→gloss→pose) efektif dalam meningkatkan struktur sintaksis gerakan.
2. GAN membantu menghasilkan gerakan lebih realistis dan halus.
3. Membutuhkan dataset besar dengan pasangan teks-gloss dan skeleton.

2.1.2 *Text2Sign Diffusion: A Generative Approach for Gloss Free Sign Language Production [9]*

Penelitian berjudul "*Text2Sign Diffusion: A Generative Approach for Floss Free Sign Language Production*" yang dilakukan oleh Feng L, dkk. Penelitian ini memperkenalkan pendekatan generative berbasis *diffusion* model untuk menghasilkan pose bahasa isyarat langsung dari teks tanpa memerlukan anotasi *gloss*. Model ini menggunakan text encoder seperti T5 atau BERT untuk memetakan makna kalimat, lalu menghasilkan urutan pose manusia secara bertahap menggunakan denoising diffusion process. Kelebihan utama metode ini adalah kemampuannya untuk bekerja secara gloss-free yang membuatnya menjadi cocok untuk dataset yang belum memiliki anotasi lengkap. Evaluasi dilakukan dengan menggunakan metrik pose error (MPJPE) dan WER (*Word Error Rate*), ini menunjukkan hasil ebih baik dibandingkan model GAN atau *Transformer*.

Beberapa poin penting yang bisa diambil:

1. Tidak memerlukan anotasi gloss karena lebih efisien untuk dataset kecil.
2. Diffusion model menghasilkan gerakan yang lebih natural dan kontinu.
3. Meskipun gloss-free, metode ini memerlukan sumber data komputasi dan waktu training yang substansial yang membuat kurang praktis untuk pengembangan prototipe cepat.

2.1.3 *Signs as Tokens: A Retrieval-Enhanced Multilingual Sign Language Generator (SOKE) [10]*

Penelitian yang dilakukan oleh Zuo R, dkk ini mengkonversi gestur menjadi token diskrit layaknya kata pada teks. Model ini memanfaatkan autoregressive transformer yang memprediksi token gestur secara berurutan berdasarkan konteks teks input. Dengan pendekatan tokenisasi ini, model mampu menangani berbagai bahasa isyarat dan meningkatkan efisiensi pelatihan karena setiap gestur di-representasikan dalam format yang dapat dipelajari Bersama. Penelitian ini dievaluasi menggunakan dataset RWTH-PHOENIX-Weather 2014T(DGS) dan How2Sign (ASL), menunjukkan peningkatan pada skor BLEU dan FGD (Frechet Gesture Distance)

Beberapa poin penting yang bisa diambil:

1. Representasi token membuat gestur dapat diproses seperti bahasa alami.

2. Mampu menangani banyak bahasa isyarat sekaligus (multilingual).
3. Efisien untuk pelatihan dan transfer antar bahasa.

2.1.4 *SignAligner: Harmonizing Complementary Pose Modalities for Coherent Sign Language Generation* *[11]*

Penelitian yang dilakukan oleh Wang Xu, dkk ini berfokus Pada peningkatan kualitas gerakan dalam video hasil bahasa isyarat. Metode ini memperkenalkan *pose harmonization module* untuk menggabungkan informasi dari berbagai modalitas pose seperti tangan, wajah dan tubuh utama agar lebih sinkron secara temporal. Model dilatih dengan loss gabungan antara *temporal smoothness, inter-joint consistency dan spatial alignment*. Eksperimen pada dataset CSL-Daily menunjukkan peningkatan besarpada naturalitas visual dan konsistensi antar frame dibanding baseline berbasis transformer murni.

Beberapa poin penting yang bisa diambil:

1. Menggabungkan beberapa modalitas pose
2. Meningkatkan naturalitas dan kontinuitas gerakan.
3. Bisa diadaptasi untuk memperhalus hasil generasi pose BISINDO.

2.1.5 *Empowering Sign Language Communication: Intergrating Sentiment and Semantics for Facial Expression Synthesis* *[12]*

Penelitian yang dilakukan oleh Azefedo, Rafael et.all ini menyoroti pentingnya ekspresi wajah dalam sistem bahasa isyarat. Model yang dikembangkan mengintegrasikan analisis semantic dan sentiment dari teks untuk menentukan ekspresi wajah yang sesuai selama proses penerjemahan gestur. Modul *Facial Expression Generator* dilatih bersamaan dengan pose body generator menggunakan jaringan *multimodal attention.* Evaluasi menunjukkan bahwa ekspresi wajah yang dihasilkan meningkatkan keterpahaman tanda hingga 15% dalam uji persepsi pengguna.

Beberapa poin penting yang bisa diambil:

1. Ekspresi wajah berpengaruh besar pada keterpahaman.
2. Integrasi semantic dan emosi dapat meningkatkan hasil translasi.
3. Relevan untuk pengembangan BISINDO yang juga menggunakan ekspresi wajah sebagai komponen utama.

2.1.6 Pengembangan Website Speech to Video Bahasa Isyarat Indonesia (BISINDO) Berbasis Algoritma Long Shot Term Memory [13]

Penelitian yang dilakukan oleh Intan Octaviani et.all ini bertujuan untuk mengonversi input suara menjadi video BISINDO. Sistem ini memanfaatkan proses *speech-to-text* menggunakan API Google Speech, kemudian teks tersebut dipetakan ke urutan video gesture BISINDO yang sudah direkam sebelumnya. Model LSTM digunakan untuk menentukan urutan kata dan sinkronisasi antar video gestur agar hasilnya lebih natural. Penelitian ini menghasilkan prototipe berbasis web yang mampu men-terjemahkan kalimat sederhana ke urutan gestur BISINDO.

Beberapa poin penting yang bisa diambil:

1. Sudah menerapkan BISINDO dalam konteks edukasi.
2. Fokus pada integrasi speech-to-video, bukan generasi otomatis dari pose.
3. Masih terbatas pada kosakata statis dan belum mendukung generalisasi getur dinamis.

2.1.7 Audio to Sign Language Translation for Deaf People [14]

Penelitian dengan judul "*Audio to Sign Language Translation for Deaf People*" yang dilakukan olek Patel et.all mengusulkn sistem penerjemahan audio ke bahasa isyarat untuk membantu komunitas tuli. Sistem ini memiliki alur kerja berupa input audio yang diproses menggunakan Google Speech API untuk menghasilkan teks, kemudian teks tersebut diproses dengan Natural Language Processing (NLP) agar sesuai dengan tata bahasa Indian Sign Language (ISL). Setelah itu kata-kata diterjemahkan dengan pendekatan *dictionary-based* yang dipetakan ke gambar atau GIF statis yang me-representasikan gestur ISL.

Beberapa poin penting yang bisa diambil:

1. Pendekatan berbasis dictionary-based memungkinkan penerjemahan cepat namun sangat terbatas pada kosakata yang sudah di-definisikan.
2. Sistem menghasilkan gestur berupa GIF statis sehingga kurang natural dan tidak fleksibel untuk konteks edukasi yang kompleks.

3. Penelitian ini menekankan pentingnya pipeline otomatis dari audio→teks→gestur, namun belum memanfaatkan deep learning maupun avatar 3D

Tinjauan terhadap studi-studi generative seperti *Text2Sign* dan *Text2Sign Diffusion* menegaskan tantangan utamanya yaitu kebutuhan akan dataset yang besar dan sumber daya komputasi tinggi yang tidak realistis untuk BISINDO saat ini. Sementara itu studi tentang *SignAligner* dan *Facial Expression Synthesis* menggaris-bawahi betapa penting kualitas gerakan dan komponen non-manual. Di sisi lain, studi pada sub-bab 2.1.6 dan 2.1.7 menunjukkan pentingnya efisiensi *dictionary-based* namun masih terbatas pada playback video dan GIF statis. Oleh karena itu, penelitian ini berupaya mengisi celah tersebut dengan meng-implementasikan pendekatan *template concatenation* yang hemat sumber daya dan sangat efisien waktu sebagai solusi yang praktis dan *low resource* untuk BISINDO dengan memanfaatkan representasi *keypoint* dinamis yang fleksibel.

2.2 Tinjauan Teori

Bab ini menguraikan teori utama yang menjadi fondasi dalam penelitian, mulai dari konsep dasar BISINDO sebagai bahasa target & teknik pemrosesan bahasa alami (NLP). Semua teori ini secara langsung mendukung pengembangan sistem sintesis gestur yang diusulkan dalam penelitian ini.

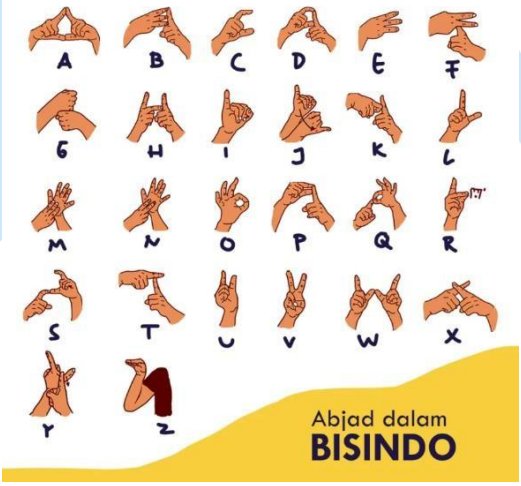
2.2.1. *Deep Learning*

Deep learning adalah cabang dari machine learning yang menggunakan arsitektur *artificial neural network* dengan layer 3 atau lebih yang mampu belajar dari representasi data dalam bentuk yang kompleks dibanding dengan *shallow neural network*. Dalam penelitian ini, deep learning tidak digunakan untuk melatih model terjemahan gestur, peran utamanya sebagai *feature extractor* melalui *mediapipe holistic*. Kemampuan *deep learning* ini digunakan untuk menghasilkan koordinat tubuh yang akurat, yang kemudian menjadi template masukan untuk sistem *template concatenation*.

2.2.2. BISINDO (Bahasa Isyarat Indonesia)

Bisindo (Bahasa Isyarat Indonesia) merupakan bahasa yang digunakan oleh teman tuli di Indonesia. BISINDO mempunyai struktur yang linguistik, tata bahasa dan kosakata yang khas, hal ini yang membuat BISINDO berkembang secara alami. Sifat utama BISINDO adalah bahasa alami yang artinya berkembang secara organik tanpa adanya campur tangan formal dari pihak luar. Hal ini membuat strukturnya berbeda dari SIBI (Sistem Isyarat Bahasa Indonesia) yang merupakan sistem terstruktur berdasarkan tata Bahasa Indonesia yang formal dan lengkap. SIBI sendiri diciptakan oleh pihak pemerintah dengan maksud untuk kemajuan pendidikan formal, tapi sampai sekarang masih sering terasa kaku bagi penutur BISINDO karena memaksa pola kalimat yang tidak sesuai dengan tata bahasa alami teman tuli. Teman tuli pada umumnya lebih nyaman menggunakan BISINDO karena:

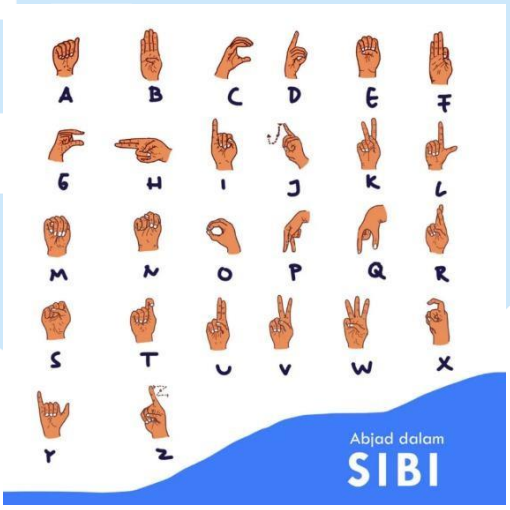
1. Lebih natural karena tidak mengikuti pola bahasa Indonesia formal tapi mengikuti urutan tata bahasa yang berkembang di komunitas Tuli.
2. Lebih ekspresif karena mengandalkan kombinasi gerakan tangan, ekspresi wajah dan bahasa tubuh untuk mengambil makna secara utuh.
3. Punya identitas budaya karena dianggap sebagai bagian dari identitas dan warisan budaya teman tuli di Indonesia.
4. Lebih mudah dipahami sesama pengguna karena BISINDO lebih mudah dipahami tanpa proses penerjemahan mentah seperti SIBI karena digunakan dalam interaksi sehari-hari.



Gambar 2. 1 Bentuk Gestur BISINDO

Karakter penting lain ada penggunaan *classifier*, terutama dalam komunikasi formal atau edukasi. *Classifier* adalah gestur yang tidak menerjemahkan kata, melainkan menggambarkan bentuk, ukuran, atau cara bergerak suatu objek/konsep dalam ruang. Seperti untuk menjelaskan konsep ilmiah, juru bahasa BISINDO akan menggunakan classifier untuk mendemonstrasikan proses daripada meng-isyaratkan kata per kata, yang menjadikannya lebih mudah dipahami secara visual.

Dalam penelitian ini, BISINDO dipilih sebagai target penerjemahan karena lebih representatif terhadap kebutuhan komunikasi teman tuli di Indonesia dibandingkan SIBI. Dengan berfokus pada BISINDO, sistem yang dikembangkan diharapkan dapat lebih natural, dapat dipahami pengguna dan sesuai dengan konteks komunikasi sehari-hari.



Gambar 2. 2 Bentuk Gestur SIBI

2.2.3. NLP (Natural Language Processing)

NLP adalah cabang dari AI yang berfokus pada interaksi antara komputer dan bahasa manusia yang tujuan utamanya adalah membuat sitm yang bisa menghasilkan, memahami dan meng-interepresentasikan bahasa alami, baik dalam bentuk audio/teks sehingga bisa digunakan untuk berbagi aplikasi seperti translator bahasa dan chatbot. NLP sendiri merupakan paduan dari konsep lingusitik dan ilmu komputer jadi pemrosesan bahasa yang tidak hanya memperhatikan arti kata secara individual saja tetapi juga hubungan antar kata, struktur kalimat dan konteks. Peran NLP dalam sistem ini bukan

untuk melakukan parsing sintaksis kompleks seperti *Natural Language Understanding* (NLU) atau *Natural Language Generation* (NLG), melainkan fokus pada dua fungsi untuk menjamin full coverage pada kosakata terbatas:

1. *Synonym Mapping* dari NLP menggunakan kamus *dictionary-based* untuk memetakan kata-kata yang tidak ada di dataset inti ke sinonim yang tersedia templatenya. Mekanisme ini secara efisien meningkatkan cakupan kosakata fungsional.
2. Tokenisasi dan Fallback merupakan proses memecah kalimat menjadi token kata. Kalau suatu kata masih te-identifikasi OOV (*out of vocabulary*) setelah *synonym mapping*, NLP akan memicu *fingerspelling fallback* dimana kata tersebut akan dieja menggunakan urutan gestur huruf.

Dengan demikian, NLP dalam penelitian ini berfungsi sebagai modul ringan yang sangat efisien waktu dan memastikan 100% translatability dari input teks bebas dengan mengandalkan *lexicon* dan *rule-based.*

2.2.4. Pendekatan Template Concatenation

Model *Template Concatenation* merupakan pendekatan non-generatif dalam penerjemahan *audio/text-to-gesture*. Sistem bekerja dengan mengandalkan *library template dictionary* based daripada melatih model *deep learning* untuk memprediksi gerakan baru. Keunggulan dari metode ini ada pada kecepatan produksi dan akurasi gerakan karena data diambil dari data nyata, ini menjadi solusi praktis untuk bahasa dengan keterbatasan data training seperti BISINDO.

2.2.5. Evaluasi Kinerja Sistem

Metrik utama untuk sistem sintesis non-real-time adalah efisiensi waktu produksi, diukur dari input teks hingga hasil file output video .mp4 yang siap publikasi. Pengukuran ini krusial untuk membandingkan kinerja sistem otomatis dengan waktu pasca produksi manual yang memakan waktu. Kinerja waktu total sistem dapat di-definisikan sebagai akumulasi dari tiga komponen utama: $T_{Total} = T_{NLP} + T_{Sintesis} + T_{Rendering}$, dimana:

- T_{NLP} adalah waktu pemrosesan teks, termasuk pencocokan kosakata dan *fallback.*
- $T_{Sintesis}$ adalah waktu komputasi inti yang mencakup concatenation keypoint dan smoothing gerakan.
- $T_{Rendering}$ adalah waktu input/output, termasuk proses encoding akhir oleh moviepy untuk menghasilkan file .mp4.

Dalam sistem sintesis video otomatis, $T_{Rendering}$ seringkali menjadi bottleneck utama karena memerlukan operasi encoding yang intensif. Oleh karena itu metrik T_{Total} yang rendah akan menjadi bukti kuat bahwa sistem ini unggul dalam memangkas waktu produksi bagi stakeholder media.

2.2.6. Aspek Keterpahaman dan Visualisasi

Kualitas sebuah sistem penerjemah bahasa isyarat dinilai dari keterpahaman pesan oleh komunitas tuli. Karena penelitian ini menggunakan representasi 2D, desain visualisasi menjadi sangat penting untuk memfokuskan perhatian penerima pesan. Penggunaan warna kontras yang strategis pada keypoint dan penyesuaian ketebalan garis dapat membantu menonjolkan bentuk tangan dan lengan yang krusial untuk BISINDO, sehingga meningkatkan daya tarik dan kejelasan visual.

2.2.7. Pre-processing Data Gestur

Tahapan pre-processing ini krusial agar template gestur ter-standarisasi dan siap untuk proses *concatenation*. Dengan mediapipe holistic, normalisasi dan augmentasi, data menjadi lebih representative serta memastikan keandalan template dalam me-representasikan gerakan secara universal tanpa terpengaruh variasi perekaman. Tahap ini bertujuan untuk mengubah data mentah hasil ekstraksi keypoint menjadi format yang terstruktur, bersih dan konsisten sehingga template gerakan dapat diakses dan digunakan secara optimal oleh algoritma *template concatenation*.

MediaPipe Holistic

MediaPipe Holistic adalah framework open-source yang dikembangkan oleh Google untuk mendeteksi landmark manusia secara real time. Framework ini menggabungkan 3 model utama, yaitu pose (33 titik), hand (21 titik untuk masing-masing tangan) dan face mesh (468 titik). Dengan menggabungkan ketiga mdel ini, *MediaPipe Holistic* bisa menangkap informasi gerakan tubuh secara menyeluruh seperti posisi tangan yang detil dan ekspresi wajah. Dalam penelitian ini tidak semua titik akan digunakan karena titik wajah sendiri berjumlah sangat banyak dan sebagian besar tidak relevan untuk pengenalan BISINDO. Pendekatan ini tidak hanya mengurangi beban pada sistem komputasi tapi meminimalisir *noise* dari titik yang tidak berkontribusi pada pengenalan gerakan.

Struktur Keypoints

Struktur *keypoints* mengacu pada format dan urutan koordinat yang dihasilkan oleh *mediapipe holistic*. Setiap titik (*landmark*) memiliki parameter x, y, z yang mewakili koordinat spasial 3D yang dinormalisasi.

Normalisasi Skala Global

Normalisasi dilakukan untuk memsatikan bahwa semua nilai koordinat berada dalam rentang yang seragam, sehingga memastikan konsistensi dimensi fisik antar template yang krusial untuk proses blending. Normalisasi ini membantu mengatasi variasi jarak subjek terhadap kamera dan memastikan bahwa template dari video yang berbeda memiliki representasi numerik yang sebanding

2.2.8. *Smoothing dan Blending*

Dalam pendekatan *template concatenation*, tantangan utama adalah memastikan transisi yang mulus antara dua template gestur yang berurutan. Diskontinuitas posisi keypoint pada titik sambungan sering terjadi karena perbedaan posisi akhir gestur pertama dan posisi awal gestur kedua, yang mengakibatkan gerakan tampak kaku atau tidak alami. Untuk mengatasi diskontinuitas ini, akan diterapkan proses gerakan transisi atau blending menggunakan teknik interpolasi. Teknik ini bertujuan untuk menciptakan zona

tumpang tindih yang secara bertahap memindahkan keypoint dari posisi akhir template A ke posisi awal template B selama periode waktu tertentu.

Secara matematis posisi P_t pada frame transisi dihitung berdasarkan interpolasi linier antara posisi P_{akhir} gestur sebelumnya dan P_{awal} gestur berikutnya menggunakan factor bobot yang bergerak secara linier seiring waktu.

2.2.9. *Fingerspelling* dan Visualisasi Gestur

Fingerspelling adalah teknik dalam bahasa isyarat dimana tiap huruf alfabet akan di-representasikan dalam bentuk ejaan. Dalam BISINDO, biasanya *fingerspelling* hanya untuk mengeja kata yang tidak punya gestur spesifik seperti nama orang atau istilah asing. Tantangan utama dalam *fingerspelling* ini adalah bentuk tangan yang sangat halus antar huruf dan kecepatan gerakan yang bisa bervariasi antar pengguna.