International Journal on Advanced Science Engineering Information Technology

Vol.15 (2025) No. 2
ISSN: 2088-5334

Generative AI-Driven Multimodal Interaction System Integrating Voice and Motion Recognition

DaeSung Jang [a], JongChan Kim [a,*]

[a] *Department of Computer Engineering, Sunchon National University, Jungang-ro, Suncheon-si, Republic of Korea*
*Corresponding author: *seaghost@scnu.ac.kr*

***Abstract*—This research proposes a two-way interactive algorithm based on voice and motion recognition using generative AI technology to overcome the limitations of existing systems limited to simple command recognition. Current voice and motion recognition technologies are essential in enabling interaction between smart devices and users to enhance user experience. Still, they are mainly limited to recognizing and executing prescribed commands, which do not meet the diverse and complex needs of users. To solve these problems, this research aims to develop a technology that fuses and integrates voice and motion data based on advanced learning and prediction capabilities of generative AI, provides customized data optimized for each user's personality and situation in real-time, and enables more natural and efficient interactions. The main research content includes developing data analysis and processing algorithms that can integrally process multiple input channels, designing generative AI-based models for providing customized data to users, and implementing a two-way interactive system that maintains a natural conversation flow. In particular, the research is intended to combine generative AI language models with computer vision technology to comprehensively analyze user voice and motion data, enabling smart devices to understand and respond to user intent accurately. These technologies can potentially revolutionize the user experience in various areas, including smart homes, healthcare, education, and more. This study's results are expected to significantly contribute to the development of next-generation smart device interaction systems that could improve both efficiency and engagement of interactions.**

***Keywords*— Generative AI; bidirectional; interaction; speech; motion; recognition; personalized data.**

Manuscript received 18 Dec. 2024; revised 12 Feb. 2025; accepted 3 Mar. 2025. Date of publication 30 Apr. 2025.

I. Introduction

With the advent of the Fourth Industrial Revolution, artificial intelligence (AI) is becoming a key driver in modern society and across industries, with speech recognition and motion recognition technologies dramatically enhancing the user experience and revolutionizing interactions with smart devices. Speech recognition technologies use automatic speech recognition (ASR) and natural language processing (NLP) to transcribe a user's voice into text to perform commands and provide information. In contrast, motion recognition technologies use computer vision and sensors to precisely recognize user actions and gestures and enable intuitive input methods. These technologies are widely utilized in a variety of industries, including smart homes, healthcare, automotive, education, and entertainment, to simplify user-device interactions and provide intuitive and efficient experiences [1]-[6].

Among speech recognition technologies, the most popular models utilized in research and technology development are Google Speech-to-Text API and OpenAI's Whisper. Google Speech-to-Text API is widely used in research and development environments due to its high accuracy, multi-language support, and real-time processing capabilities. It is an automated speech recognition technology based on Deep Neural Networks (DNNs) that accurately extracts text from speech data. It can flexibly respond to speech intonation, dialect, and background noise, allowing it to effectively process speech data in a variety of environments [1]. Researchers are utilizing this model to develop various applications, including smart home device control, voice assistants, and multilingual translation systems. The technology's API provides a developer-friendly interface and is customizable based on pre-trained models, making it ideal for research. OpenAI's Whisper represents state-of-the-art speech recognition with models that offer high accuracy and versatility. Trained in large amounts of multilingual speech data, Whisper excels at language translation, understands the context of conversations, and performs well in environments with high background noise. The model considers not only the pronunciation and intonation of speech but also its contextual meaning and is particularly well suited to handle complex voice commands. Whisper is available as open source, giving researchers the flexibility to experiment and customize it, and can be used in a variety of applications [7]-[14].

Regarding motion recognition technology, Intel RealSense is the most widely used in research and technology development. Intel RealSense combines depth sensing technology with infrared cameras to provide a platform for precise recognition of user actions and gestures. Technology is capable of real-time motion tracking and generating sophisticated spatial data, which is utilized in various research and application areas, including virtual reality (VR), augmented reality (AR), rehabilitation in healthcare, and robotics. Intel RealSense is easy to integrate with AI frameworks such as OpenCV and TensorFlow and provides a robust SDK to help researchers easily process data and develop models. In particular, the technology enables detailed tracking of hand gestures, facial expressions, and body movements, making it ideal for designing advanced systems that analyze multi-motion data. While we intended to utilize OpenPose in our initial research, the technology has stagnated since its last update in 2020. Therefore, we adopted Google's MediaPipe and AlphaPose as alternatives to OpenPose. MediaPipe provides lightweight technology that can operate in real-time, even in mobile environments, and AlphaPose is strong in complex pose estimation and multi-person tracking. These state-of-the-art technologies complement the limitations of OpenPose and lay the foundation for effectively achieving our research goals [15]-[22].

In this study, we adopt the Llama3 model as the main component of generative AI. Llama3 is a state-of-the-art language model with the ability to learn from large-scale data and excellent natural language generation capabilities. The model is optimized to understand context and interact with users naturally based on speech recognition and motion data. Llama3 processes voice and motion data as a single, unified input to facilitate interaction between multiple data types, maintain natural conversations based on a deep understanding of user questions and commands, and provide rapid responses based on data processed in real-time via Whisper and MediaPipe. This can enable the implementation of a two-way interactive system, which is the main goal of this research, and contribute to the realization of user-centered smart device interaction [23]-[30].

II. Materials and Method

A. Data Preprocessing: Preprocessing Voice and Motion Data

In this study, the data collection and preprocessing process is mainly divided into speech and motion data. The data preprocessing step is essential for implementing an interactive system based on speech and motion recognition. Data preprocessing is the process of transforming raw data to make it suitable for model training, and it plays a vital role in improving data quality and optimizing the system's performance.

Speech data is collected using public speech datasets (LibriSpeech, Common Voice) and samples collected from real-world environments. The collected data is subjected to denoising algorithms (Spectral Subtraction, Wiener Filter) to remove background noise and clutter, and to enhance the generality of the model by including different accents, dialects, and languages. The sampling rate is then unified to 16 kHz or 44.1 kHz and the amplitude of the speech signal is standardized to reduce variation between data. This process converts speech data into a consistent format and ensures training reliability.

OpenAI's Whisper is the core technology utilized in this study for preprocessing speech data and developing models. Whisper is trained on large-scale multilingual speech data and provides high accuracy in the presence of background noise, accents, and dialects. Whisper is designed with multi-task learning to handle speech recognition, translation, and contextual understanding of conversations and effectively reflects various changes in speech data. In particular, Whisper supports advanced feature extraction using Mel-Frequency Cepstral Coefficients (MFCC) and spectrograms, which are ideal for identifying emotion and intent in speech. The model is open-sourced to researchers and can be customized for various applications.

It extracts features such as Mel-Frequency Cepstral Coefficients (MFCC), spectrogram, pitch, and energy from speech data to provide information about the emotion and intent of a conversation. Furthermore, data augmentation techniques such as time-shifting, frequency distortion, and noise addition are applied to increase the diversity of the speech data. This process helps the model to perform reliably under different speech conditions.

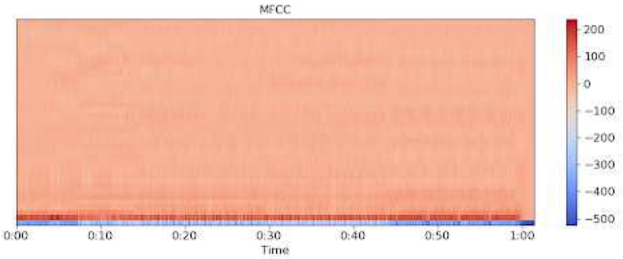


Fig. 1 MFCC(Mel-Frequency Cepstral Coefficient) Algorithm graphs

Motion data is collected using depth sensors such as Intel RealSense and Microsoft Kinect, which consider different lighting conditions, user body types, and background environments. Any possible gaps in the collected data are compensated for using linear interpolation or a Karman filter. The 3D coordinate data extracted from the sensors is converted to a reference coordinate system, and the data is normalized by scaling the user's body shape and range of motion to a standard size. During this process, key features such as joint position, angle, velocity, and acceleration are extracted, and Principal Component Analysis (PCA) is used to reduce data size and speed up learning.

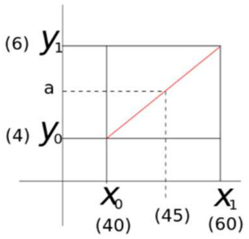


Fig. 2 Linear Interpolation Graphs

Motion data is augmented with augmentation techniques to prepare it for different environments. Rotation and translation transformations, sensor noise, and motion speed transformations are applied to increase the model's versatility. Time axes are synced, and data is sorted based on timestamps for integrated processing of speech and motion data. Speech and motion data are organized into a unified dataset using the same labeling scheme, and their sizes and units are normalized to ensure model training stability.

The preprocessed data is quality-assessed to validate its suitability. The data is checked for consistency, diversity, denoising, and feature extraction accuracy, and outliers are removed using visualizations such as spectrograms, joint trajectories, and 3D coordinate graphs. These preprocessing steps provide high-quality integration of speech and motion data to provide optimized data for model training and increase the proposed system's accuracy and reliability. In turn, this forms an important basis for optimizing the interaction between the user and the device and maximizing the effectiveness of the interactive system.

B. Whisper: A Versatile Automated Speech Recognition Model

Whisper is an automatic speech recognition (ASR) model developed by OpenAI. It is a versatile model that can recognize a wide range of languages and dialects. The model is trained on a large and diverse audio dataset and can perform tasks such as multilingual speech recognition, speech translation, and language identification.

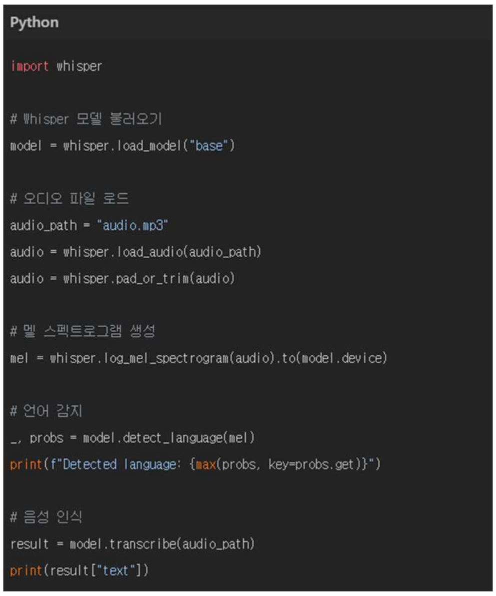
The main feature of Whisper is that it uses a transformer-based sequence-to-sequence model for high accuracy. It can recognize multiple languages and dialects, and it can also perform speech translation and language identification. It can perform various speech-processing tasks, including speech recognition, speech translation, spoken language identification, and speech activity detection. There are five model sizes for speech recognition: tiny, base, small, medium, and large, each with different speeds and accuracy.

Whisper has achieved human-level accuracy and robustness in English speech recognition and can also perform speech translation into multiple languages and multilingual speech recognition. We also measured its zero-shot performance on various datasets and found that it made 50% fewer errors than other models.

TABLE I
COMPARE AUTOMATIC SPEECH RECOGNITION MODELS

ASR System	Key features and performance
Whisper	- High accuracy across languages and tasks - Tolerance for accents, background noise, and technical terms - 50% fewer errors in zero-shot performance
Kaldi	- Open source ASR toolkit - Loved by many researchers and developers over the years - Lower accuracy and robustness than Whisper
wav2vec 2.0	- Facebook Development - Superior performance - Lower performance than Whisper in many languages and tasks
Google ASR	- Commercial ASR systems - Low recognition rates for different accents and non-native speech - Lower performance than Whisper
IBM ASR	- Commercial ASR systems - Low recognition rates for different accents and non-native speech - Lower performance than Whisper
Microsoft ASR	- Commercial ASR systems - Low recognition rates for different accents and non-native speech - Lower performance than Whisper

These comparisons show that the Whisper model is robust in a variety of environments and conditions. To use Whisper, you need to configure your environment by installing the Whisper model using the Python package manager pip『pip install -U Openai-whisper』 to install the Whisper model and configure your environment. The ffmpeg command-line tool must be installed on your system.


```
Python

import whisper

# Whisper 모델 불러오기
model = whisper.load_model("base")

# 오디오 파일 로드
audio_path = "audio.mp3"
audio = whisper.load_audio(audio_path)
audio = whisper.pad_or_trim(audio)

# 멜 스펙트로그램 생성
mel = whisper.log_mel_spectrogram(audio).to(model.device)

# 언어 감지
_, probs = model.detect_language(mel)
print(f"Detected language: {max(probs, key=probs.get)}")

# 음성 인식
result = model.transcribe(audio_path)
print(result["text"])
```

Fig. 3 Transcribe Whisper Model Audio Files to Text

Fig. 3. The code shows converting an audio file to text using the Whisper model. The transcribe method breaks the audio file into smaller parts, each analyzed sequentially to generate text. The Whisper model is utilized in a variety of real-world applications. The main applications include content creation and management (transcribing podcasts and interviews, transcribing meeting minutes), accessibility (captioning videos, real-time captioning), research and data analysis (transcribing interviews and focus groups, transcribing survey responses), legal services (transcribing court proceedings, transcribing meetings and consultations),

healthcare (transcribing patient notes, transcribing medical meetings), education (transcribing lectures, online training content, finance and business, transcribing earnings calls, transcribing customer calls), and multilingual applications (multilingual voice search, voice translation). In addition, the Whisper model can be leveraged in various language-based applications. Through these applications, Whisper can increase efficiency and improve accessibility in various industries.

C. MediaPipe, AlphaPose: Real-time pose estimation

Several technologies can replace OpenPose or be applied to your research, depending on your research objectives and the features you need (e.g., 2D/3D pose estimation, real-time, accuracy, etc. As newer technologies and platforms emerge, the possibilities for replacing or complementing OpenPose are growing. Google's MediaPipe is a lightweight library that provides a real-time solution for mobile devices and web environments. MediaPipe offers a variety of modules, including face tracking (Face Mesh), hand tracking (Hands), and pose estimation (Pose), and can extract 2D joint data from a single RGB camera, with some modules supporting 3D information as well. The technology is lighter and faster to run than OpenPose, and is suitable for various applications, including mobile and web-based applications, AR/VR environments, and sports motion analysis. However, it has limitations because it may not be as sophisticated as OpenPose for complex pose estimation.

BlazePose is a MediaPipe-based technology designed to accurately analyze specific body poses (e.g., yoga, fitness). It supports 3D pose estimation and is optimized to work efficiently in mobile environments. It can be used in applications such as sports and fitness apps and posture analysis and correction tools and is particularly suited to environments where real-time processing is critical. However, it can be less accurate in complex environments with many occlusions.

AlphaPose is a PyTorch-based pose estimation technology developed by Shanghai Jiao Tong University that boasts higher accuracy than OpenPose. AlphaPose performs well in complex postures and crowded environments and has excellent multi-person tracking capabilities. It can be widely used in sports analysis, motion recognition research, and crowd behavior analysis. However, its processing speed can be slow and needs further optimization for real-time applications.

DeepLabCut is a highly customizable open-source platform that supports posture tracking for animals and humans. Researchers can train the model to track specific postures or joints, and it also supports 3D pose estimation using multiple cameras. It can be used in various fields, including behavioral analysis research, biomechanics, and neuroscience research, but it can be complicated to set up and learn.

Detectron2, developed by Meta, performs human joint estimation based on Mask R-CNN and integrates object detection, segmentation, and pose estimation. Technology offers many features and strong customization options, making it ideal for research and development projects. However, its real-time capabilities can be limited, making it better suited for deep analytics research than real-time applications.

PoseNet is a lightweight pose estimation technology based on TensorFlow.js that runs in a web browser and is optimized for mobile. It is ideal for AR/VR applications and rapid prototyping, but it has limitations with lower accuracy than OpenPose.

OpenMMLab's MMPose is a modern framework integrating various deep learning techniques to provide high accuracy. It provides pre-trained models for multiple pose estimation datasets. It has a modular structure designed to be researcher-friendly, making it suitable for pose estimation research and complex motion analysis. However, some drawbacks can make the initial setup complex.

Each of these technologies has its strengths and limitations, and the right choice depends on your research goals. If real-time and lightweight are essential, MediaPipe, BlazePose, and PoseNet are good choices, while AlphaPose and DeepLabCut are good choices if you need accuracy and complex pose analysis. If you need versatility and scalability, you can consider Detectron2 and MMPose. Depending on the research environment and requirements, these technologies can be used alone or in combination to overcome the limitations of OpenPose and enable efficient research and development.

TABLE III
COMPARE MOTION RECOGNITION MODELS

MediaPipe	Mobile/web real-time, lightweight	Limitations of complex pose estimation	AR/VR, Sports, Fitness
AlphaPose	Highly accurate, multi-person tracking	Processing can be slow	Crowd analytics, sports, behavioral research
DeepLabCut	Animal/human trackable, 3D support	Initial setup complexity	Biomechanical and neuroscience research
Detectron2	Versatile, deep learning customizable	Real-time limitations	Research and development, object detection
PoseNet	Lightweight, web/mobile friendly	Low accuracy	AR/VR, Rapid Prototyping

OpenPose is the leading deep learning-based technology for 2D and 3D pose estimation and has been used in many studies. However, it has limited recent updates and performance limitations in some real-time and lightweight environments. As a result, MediaPipe and AlphaPose are two modern alternatives to OpenPose that can be used to replace or complement it. Each of these two technologies has its unique strengths and characteristics, and depending on your research goals and needs, they offer effective alternatives to OpenPose.

MediaPipe is a multi-platform computer vision library developed by Google that is ideal for applications where lightweight and real-time processing are essential. In addition to human pose estimation, MediaPipe offers solutions for hand tracking, face tracking, object recognition, and more, and is designed to run in real time on mobile devices or web environments. It runs smoothly in resource-constrained environments and is faster and more efficient than OpenPose.

MediaPipe's main module, Pose, can estimate 2D and 3D joint data, as well as joint tracking for hands (Hands) and landmark tracking for faces (Face Mesh), and can be combined with various computer vision applications. These characteristics make it an adequate replacement for OpenPose in mobile and web-based applications, real-time fitness coaching, AR/VR interfaces, sports motion analysis, and other environments that require real-time and lightweight performance. However, it has the limitation that it may not be as sophisticated as OpenPose for complex posture or motion analysis.

AlphaPose is a PyTorch-based deep learning framework developed by Shanghai Jiao Tong University that provides high accuracy in pose estimation. It performs better than OpenPose in complex poses and crowded environments and is particularly strong at tracking multiple people. AlphaPose utilizes the latest CNN technology to produce reliable results in complex environments and can replace OpenPose in research and applications that require accurate pose estimation. It is suitable for crowd behavior analysis, motion analysis of sports events, motion-based healthcare solutions, research, and AI model development, etc., that require multi-person tracking and supports precise motion analysis compared to OpenPose. However, processing speeds can be relatively slow, requiring additional optimization for real-time applications.

MediaPipe and AlphaPose each have unique strengths that compensate for OpenPose's limitations and can be chosen according to your research goals. Media Pipe is an excellent alternative to OpenPose for mobile and web-based applications where lightweight and real-time are essential, while AlphaPose provides better results than OpenPose in environments where complex behavior analysis or multi-person tracking requires accuracy and precision.

With OpenPose's limited recent updates and inability to provide real-time processing and lightweight execution in some situations, MediaPipe and AlphaPose are essential alternatives. MediaPipe is ideal for simple motion tracking, such as in smart homes, while AlphaPose is utilized in applications where precision is critical, such as sports analytics, rehabilitation, and crowd tracking. By choosing or combining these two technologies for your research and applications, you can overcome the limitations of OpenPose and increase the versatility and efficiency of study. This makes MediaPipe and AlphaPose the optimal choice to replace or complement OpenPose.

With OpenPose's limited recent updates and inability to provide real-time processing and lightweight execution in some situations, MediaPipe and AlphaPose are essential alternatives. MediaPipe is ideal for simple motion tracking, such as in smart homes, while AlphaPose is utilized in applications where precision is critical, such as sports analytics, rehabilitation, and crowd tracking. By choosing or combining these two technologies for your research and applications, you can overcome the limitations of OpenPose and increase the versatility and efficiency of your study. This makes MediaPipe and AlphaPose the optimal choice to replace or complement OpenPose.

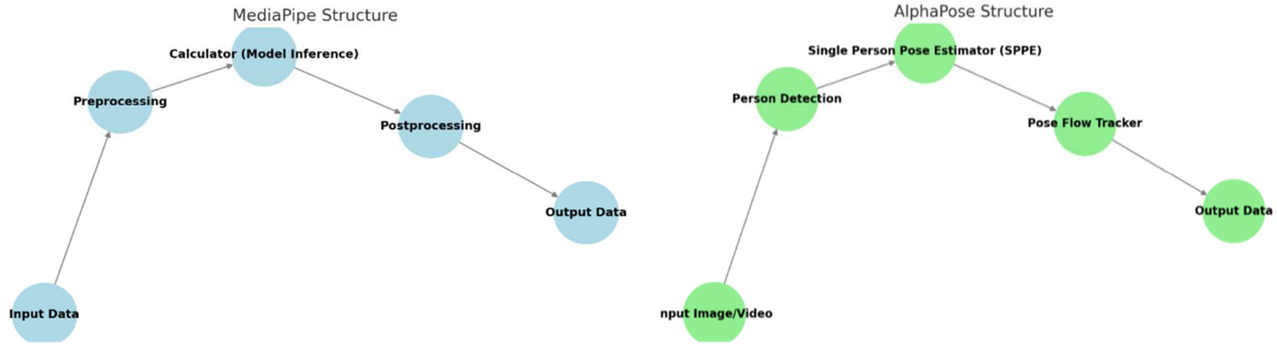


Fig. 4 Structure of MediaPipe and AlphaPose

III. Results and Discussion

A. Experimental Environment

Processing voice and motion data using Whisper, MediaPipe, and AlphaPose technologies and implementing a two-way interactive system based on the Llama3 model is systematic, including data collection, preprocessing, learning, integration, and real-time processing. By integrating voice and motion data, the system aims to revolutionize interaction with smart devices.

Voice data is collected and preprocessed through Whisper. Whispers convert speech data to text in various environments and provide high accuracy to handle background noise, accents, dialects, etc. The speech data uses Spectral Subtraction and Wiener Filter for noise removal and standardizes the sampling rate to 16 kHz or 44.1 kHz. Whisper extracts key features such as MFCC, spectrogram, pitch, and energy from speech to generate data suitable for sentiment and intent analysis. The converted text and acoustic features are stored with a timestamp and provided in a consistent format for subsequent processing and model training.

Motion data is collected utilizing MediaPipe and AlphaPose. MediaPipe is lightweight technology for real-time data collection in mobile and web environments, ideal for handling simple motions and gestures. AlphaPose provides high-precision pose estimation and multi-person tracking and collects data in complex motion and crowd environments. The collected motion data is subjected to 3D coordinate normalization, missing value imputation (using linear interpolation and Karman filter), and dimensionality reduction via PCA. The preprocessed data is stored in a structured format of coordinates, angles, and velocities for model training and real-time processing.

The speech and motion data are organized into a unified dataset and synchronized around a time axis. All data is aligned with unique timestamps, and standard labeling is applied to strengthen the association between speech and motion. Data is stored in JSON or NoSQL formats for scalability and accessibility. The cleaned data is then used as input to train the Llama3 model.

Llama3 models integrate speech and motion data to understand the contextual relationships of user input and generate appropriate responses. It handles multi-input data based on data processed by Whisper, MediaPipe, and AlphaPose, and uses a transformer-based structure to learn the interaction between speech and motion. Combining text data converted from Whisper with pose data from AlphaPose and MediaPipe analyzes contextual relationships and implements interactive interfaces with users through generative AI. The system is optimized by leveraging GPU acceleration and transformer structures for real-time user input processing and rapid response generation.

B. *Dataset Type and Environment Variables*

Environmental data directly affecting the system are external factors that occur during the collection and processing of voice and motion data and play an essential role in system performance and data quality. The main environmental variables for speech data include background noise, acoustic conditions, and speaker characteristics. Background noise is divided into different noise levels and types, such as quiet room, public, street, and traffic noise, and also includes white noise, conversation, music, and machine sounds. Acoustic conditions affect the quality of speech data by factors such as reverberation (echo) caused by the size of the room and the reflectivity of the walls, microphone quality, and the distance between the microphone and the speaker. Speaker characteristics include different accents and dialects, speech rate, and pitch (high/low), which significantly impact model training and recognition accuracy.

TABLE IIIII
UNIFIED DATA SCHEMA

time stamp	Datetime	Time information for synchronizing voice and motion data	2024-12-29T12:34:56.789Z	Common
audio_raw	Binary	Raw voice data	Binary Audio Data	Whisper
audio_text	String	The result of converting speech to text	"Turn on the living room lights"	Whisper
audio_features	JSON	Features extracted from speech data (MFCC, spectrogram, pitch, energy, etc.)	{ "mfcc": [...], "pitch": 220.5 }	Whisper
Motion_pose	JSON	Motion data (integrating 2D and 3D pose data)	{ "keypoints_2d": [...], "joints_3d": [...] }	MediaPipe, AlphaPose
Motion_features	JSON	Features extracted from motion data (velocity, acceleration, joint angles, etc.)	{ "velocity": 1.5, "angle": 45.0 }	MediaPipe, AlphaPose
device_id	String	Unique ID of the device that collected the data	device12345	Common
user_id	String	Unique IDs to associate data with specific users	user56789	Common
label	String	Labels that indicate the intent or purpose of the data	Turn on light	Common
environment	JSON	Data collection environmental variables (lighting, noise level, user location, etc.)	{ "lighting": "low", "noise": "high"}	Common
data_ source	String	Where the data comes from and how it was created	Whisper, MediaPipe, AlphaPose	Common

Table 3 shows a scheme for the data collected by the sensors. The main variables in motion data are lighting conditions, camera conditions, user behavior, and the physical environment. Lighting conditions are factors such as brightness level (day, night, cloudy weather), type of lighting (natural, LED, fluorescent), and shadow and reflection effects that affect the quality of motion data. Camera conditions include camera position and angle (single or multiple cameras, frontal and side view), resolution (low and high), and distance from the user. User behavior includes static (sitting, standing) and dynamic (walking, running, gesturing, etc.), as well as how fast the motion is. In the physical environment, ground conditions (flat floor, tilted ground, slippery surface) and complexity (simple vs. complex background) affect motion recognition accuracy.

When processing speech and motion data together, timestamp synchronization, multi-user environments, connectivity between devices, and a combination of environmental variables are essential. Timestamp synchronization is necessary to reduce the time difference between voice and motion data and minimize input delays or processing speed differences. In multi-user environments, it is essential to separate and label voice and motion data collected simultaneously from multiple users. In addition, the network connectivity of data collection devices (microphones, cameras) and signal delays and synchronization errors between devices can also affect data quality. Environmental factors such as lighting, noise, and motion speed must also be addressed to maintain data quality.

This environmental data must be thoroughly considered in the research design phase to ensure the system's reliability and effectiveness. The system is designed to maintain stable performance in realistic usage scenarios through experiments under various conditions.

C. *Reference Data to Ensure Data Reliability*

To ensure the reliability of the data, various external reference data were used in this system's research. We utilized standard speech datasets such as LibriSpeech, Common Voice, and TED-LIUM for speech data. These datasets cover various accents, dialects, and noise conditions and are suitable for lengthy contextual processing and learning complex sentence structures. In addition, background noise datasets such as UrbanSound8K and CHiME Speech Separation and Recognition were used to evaluate performance in real-world noise conditions. Datasets such as TIMIT were also referenced for specific accent or pronunciation recognition.

For motion data, standard datasets such as the COCO Keypoints Dataset, MPII Human Pose Dataset, and AMASS Dataset were utilized to train and evaluate 2D and 3D pose data. In particular, specialized datasets such as Human 3.6M (H3.6M) and KITTI Vision Benchmark Suite were used to enable precise pose estimation for complex actions and in various environments. These datasets include various behavioral sequences and postures to enhance the precision of motion recognition.

Multimodal datasets such as AVSpeech and Ego4D are utilized to analyze speech and motion data. These datasets provide data with synchronized speech and video and are useful for performance validation in real-time interaction scenarios. Labeling reference datasets such as the OpenPose Keypoints Dataset and Labelled Speech in Noise (LSN) provide the standards needed to label pose and speech data accurately.

To evaluate the reliability of the data, we used the Noisy Audio and Motion Dataset and the Simulated Multi-User Dataset to verify the model's robustness under various noise conditions and multi-user environments. In addition, cross-domain datasets such as the Multi-Domain Audio Dataset and the Diverse Pose Dataset were used to validate model performance in various environments, including indoor, outdoor, sports, yoga, and daily behavior.

These datasets provide a standard for training and testing speech and motion data and are used to evaluate the system's accuracy and generality under different conditions. We merged publicly available datasets with experimentally collected data to maximize the system's reliability and effectiveness. This could ensure the model's robustness and adaptability in real-world applications.

D. Expected Experimental Results

The expected outcome of this project is focused on optimizing the performance of interactive systems by integrating speech and motion data to optimize the performance of interactive systems. By leveraging the OpenAI Whisper model, we expect to achieve high speech recognition accuracy in various noise environments, dialects, and intonation conditions, as well as the ability to understand complex contexts and accurately handle multiple commands. MediaPipe and AlphaPose could enable high precision in 2D and 3D pose estimation, and the ability to identify and track motion in multi-user environments. These technologies are expected to maintain stable performance even in poor lighting conditions and complex backgrounds.

Voice and motion data synchronized in real-time to enable intuitive interaction with users. After data integration, Llama3-based models utilized to improve dialog generation and response quality. This enables real-time, two-way conversations that combine voice and motion data and can combine users' voice commands and gestures to perform contextually appropriate actions. In addition, by learning from data in different user environments, the system adapts to user behavior and language patterns to deliver personalized interactions.

Fig. 5 shows this system's data processing and output process. The system receives voice and motion data as input and goes through denoising, feature extraction, data normalization, and time synchronization to create an integrated dataset. It then utilizes the Llama3 model to process the data and outputs the system's response or behavior. This diagram visualizes the entire data flow and processing steps.

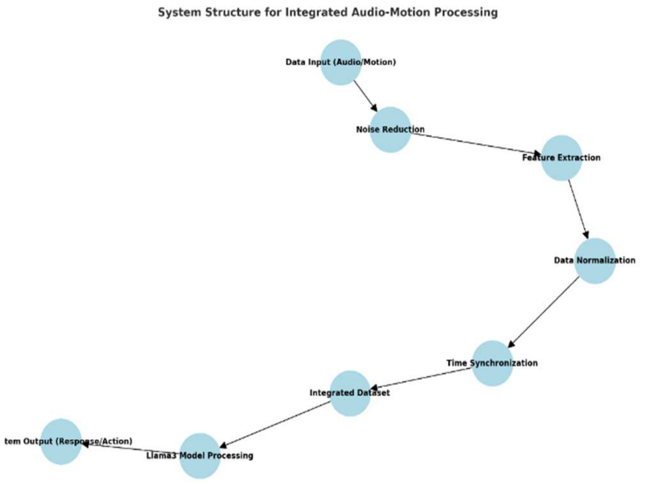


Fig. 5 System Structure for Integrated Audio-Motion Processing

In terms of performance evaluation metrics, for speech recognition, we aim for an accuracy rate of 95% or higher for simple commands, 85% or higher for complex commands, and 90% or higher for recognition in noisy environments; for motion recognition, we expect an accuracy rate of 98% or higher for single motion recognition, 90% or higher for complex motion sequences, and 85% or higher for motion separation and recognition success in multi-user environments. In addition, voice and motion data synchronization time is designed to be 200ms or less, and the user response time is designed to be 500ms or less to ensure real-time.

In conclusion, this project aims to create a user-centered, intuitive, and efficient two-way interactive system by integrally processing voice and motion data. By achieving high recognition accuracy and real-time response performance, it is expected to overcome the limitations of existing single-input-based systems and ensure reliability and stability in various environments.

IV. CONCLUSION

This paper proposes a two-way interaction system that integrally processes speech and motion data by leveraging state-of-the-art technologies such as OpenAI's Whisper, Google MediaPipe, and AlphaPose. The system addresses issues such as the lack of contextual understanding in existing speech and motion recognition systems, difficulties in analyzing complex gestures, and real-time processing limitations.

The research methodology starts with raw data collection and includes preprocessing steps such as denoising, feature extraction, data normalization, and synchronization. The processed data is then integrated and fed into the Llama3 model to enable contextually accurate response generation. This end-to-end approach could increase speech and motion data recognition and response reliability and accuracy.

Whisper technology recognizes speech with high accuracy, even in noisy environments, while MediaPipe and AlphaPose are expected to improve the precision of motion recognition in various user environments and behaviors. The combination of these technologies creates a seamless experience that allows users to interact with their devices intuitively through voice and gesture. This represents a significant step forward in multimodal interface research. In future research, we can compare the performance of the actual implemented system with the expected experimental results.

References

[1] J.-S. Han, C.-I. Lee, Y.-H. Youn and S.-J. Kim, "A study on realtime hand gesture recognition technology by machine learning-based mediapipe", *Journal of System and Management Sciences*, vol. 12, no. 2, pp. 462-476, 2022.

[2] C. Graham and N. Roll, “Evaluating OpenAI’s Whisper ASR: Performance analysis across diverse accents and speaker traits,” *JASA Express Letters*, vol. 4, no. 2, Feb. 2024, doi: 10.1121/10.0024876.

[3] G. Park and J. Kim, “Multivariate Variable-Based LSTM-AE Model for Solar Power Prediction,” *International Journal on Advanced Science, Engineering and Information Technology*, vol. 15, no. 1, pp. 293–299, Feb. 2025, doi: 10.18517/ijaseit.15.1.20944.

[4] A. Badiola-Bengoa and A. Mendez-Zorrilla, “A Systematic Review of the Application of Camera-Based Human Pose Estimation in the Field of Sport and Physical Exercise,” *Sensors*, vol. 21, no. 18, p. 5996, Sep. 2021, doi: 10.3390/s21185996.

[5] Y. Kim and H.-J. So, "Unpacking multimodal representation strategies in explainer videos from TPACK perspectives," *J. Educ. Inf. Media* (Korean Assoc. Educ. Inf. Media), vol. 30, no. 3, pp. 933-954, Jun. 2024, doi: 10.15833/kafeiam.30.3.933.

[6] Bae, H. J., Jang, G. J., Kim, Y. H., and Kim, J. P. “LSTM (Long Short-Term Memory)-Based Abnormal Behavior Recognition Using AlphaPose,” *KIPS Transactions on Software and Data Engineering*, vol. 10, no. 5, pp. 187-194, May. 2021.

[7] A. Lee, H. Ryu, H. Choi, and Y. Koo, "The DX museum strategy based on multimodal blueprint: Generation Z," *Arch. Des. Res.*, vol. 37, no. 4, pp. 149-178, 2024.

[8] Y. Lee and H. Kwon, "Research on improving unknown track reading by applying multimodal model," *J. Internet Comput. Serv.*, vol. 25, no. 6, pp. 155-162, Dec. 2024.

[9] J. Lee, "An efficient XR content authoring method based on collaborative service using multi-modal objects," *KIISE Trans. Comput. Pract.*, vol. 29, no. 4, pp. 190-195, Apr. 2023, doi:10.5626/ktcp.2023.29.4.190.

[10] D.-S. Jang and J.-C. Kim, "Two-way interactive algorithms based on speech and motion recognition with generative AI technology," *J. Korea Inst. Electron. Commun. Sci.*, vol. 19, no. 2, pp. 397-402, Apr. 2024.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16×16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2021. [Online]. Available: https://arxiv.org/abs/2010.11929.

[12] J. Chang and H. Nam, "Exploring the feasibility of fine-tuning large-scale speech recognition models for domain-specific applications: A case study on Whisper model and KsponSpeech dataset," *Phonetics Speech Sci.*, vol. 15, no. 3, pp. 83-88, Sep. 2023, doi:10.13064/ksss.2023.15.3.083.

[13] C. Oh, C. Kim, and K. Park, "Building robust Korean speech recognition model by fine-tuning large pretrained model," *Phonetics Speech Sci.*, vol. 15, no. 3, pp. 75-82, Sep. 2023, doi:10.13064/ksss.2023.15.3.075.

[14] H.-W. Cha and J.-M. Ma, "Proposal and evaluation of military automatic speech recognition model based on transformer," *J. Korea Acad.-Ind. Cooper. Soc.*, vol. 25, no. 3, pp. 102-107, Mar. 2024, doi:10.5762/kais.2024.25.3.102.

[15] K. J. Son and S. H. Kim, "A study on the evaluation methods for assessing the understanding of Korean culture by generative AI models," *Trans. Korea Inf. Process. Soc.*, vol. 13, no. 9, pp. 421-428, Sep. 2024.

[16] H.-C. Jung, K.-S. Shin, H.-D. Kim, and S.-B. Park, "Clinical trials utilizing LLM-based generative AI," *J. Korea Soc. Comput. Inf.*, vol. 29, no. 12, pp. 169-180, Dec. 2024, doi: 10.9708/jksci.2024.29.12.169.

[17] P. T. J. Miguel and K. Nah, "Integrating multimodal and generative AI in design research: Enhancing ethnographic methods with data-driven analysis," *J. Des. Res.* (Korea Inst. Des. Res. Soc.), vol. 8, no. 4, pp. 27-38, Dec. 2023, doi: 10.46248/kidrs.2023.4.27.

[18] D. Saisanthiya and P. Supraja, "Neuro-facial fusion for emotion AI: Improved federated learning GAN for collaborative multimodal emotion recognition," *IEIE Trans. Smart Process. Comput.*, vol. 13, no. 1, pp. 61-68, Feb. 2024, doi: 10.5573/ieiespc.2024.13.1.61.

[19] D. Min, S. Nam, and D. Choi, "A study on improving the accuracy of Korean speech recognition texts using KcBERT," *J. KIISE*, vol. 51, no. 12, pp. 1115-1124, Dec. 2024, doi: 10.5626/jok.2024.51.12.1115.

[20] G. Bae, C. Kim, S. Hwang, Y. Lee, and J. Kong, "Development of digital exhibition contents using generative AI and prompt engineering," *J. Korea Multimed. Soc.*, vol. 27, no. 8, pp. 959-968, Aug. 2024, doi: 10.9717/kmms.2024.27.8.959.

[21] S. Park and K. Kim, "AI image generation study utilizing ChatGPT and Midjourney," *J. Digit. Art Eng. Multimed.*, vol. 10, no. 4, pp. 501-510, Dec. 2023, doi: 10.29056/jdaem.2023.12.06.

[22] J. M. Lee, Y. K. Choi, and H. S. Kang, "A study on the motion and voice recognition smart mirror using Grove gesture sensor," *J. Korea Inst. Electron. Commun. Sci.*, vol. 18, no. 6, pp. 1313-1320, 2023.

[23] H.-J. Bae, G.-J. Jang, Y.-H. Kim, and J.-P. Kim, "LSTM (long short-term memory)-based abnormal behavior recognition using AlphaPose," *KIPS Trans. Softw. Data Eng.*, vol. 10, no. 5, pp. 187-194, 2021.

[24] J.-M. Lee, H.-J. Bae, G.-J. Jang, and J.-P. Kim, "A study on the estimation of multi-object social distancing using stereo vision and AlphaPose," *KIPS Trans. Softw. Data Eng.*, vol. 10, no. 7, pp. 279-286, Jul. 2021.

[25] N. Kwak, "A study on hand-face hybrid gesture interface using MediaPipe models," *J. Internet Things Converg.*, vol. 10, no. 5, pp. 1-11, Oct. 2024.

[26] H.-S. Kim, J.-Y. Jeong, B.-J. Choi, and M.-K. Moon, "Visualization system for dance movement feedback using MediaPipe," *J. Korea Inst. Electron. Commun. Sci.*, vol. 19, no. 1, pp. 217-224, Feb. 2024.

[27] R. Song, Y. Hong, and N. Kwak, "User interface using hand gesture recognition based on MediaPipe hands model," *J. Korea Multimed. Soc.*, vol. 26, no. 2, pp. 103-115, Feb. 2023, doi:10.9717/kmms.2023.26.2.103.

[28] M. Udurume et al., "Real-time multimodal emotion recognition based on multithreaded weighted average fusion," *J. Ergon. Soc. Korea*, vol. 42, no. 5, pp. 417-433, Oct. 2023, doi: 10.5143/jesk.2023.42.5.417.

[29] M. Udurume, A. Caliwag, W. Lim, and G. Kim, "Emotion recognition implementation with multimodalities of face, voice and EEG," *J. Inf. Commun. Converg. Eng.*, vol. 20, no. 3, pp. 174-180, Sep. 2022, doi:10.56977/jicce.2022.20.3.174.

[30] J.-H. Lee, "The expanded user interfaces and immersion by the multisensory stimulation in peripheral environment," *J. Digit. Contents Soc.*, vol. 21, no. 5, pp. 987-996, May 2020, doi:10.9728/dcs.2020.21.5.987.