EmoAvatar: Talk-Driven Avatar Emotional Expression for Virtual Streamers - Avatar's Lively Emotion Controling Solution

by

Aiju Kinoshita

A thesis submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Science (Hons.)

in

Computer Science

Ontario Tech University

Supervisor: Dr. Steven Livingstone, Dr. Ali Neshati

April 2025

Abstract

The rise of Virtual YouTubers (VTubers) has transformed digital entertainment, yet the manual control of avatar expressions poses challenges such as performer fatigue and reduced expressiveness. This study addresses these issues by developing a voice-driven automatic avatar emotion animation selector, leveraging Speech Emotion Recognition (SER) models to automate avatar expressions based on vocal cues. Our approach utilizes the Emotion2Vec model to predict and animate emotional states in real-time. Results indicate that while Emotion2Vec performs robustly on English datasets, it achieves moderate success on Japanese data, highlighting the need for further adaptation and data augmentation. The application demonstrates minimal lag, ensuring seamless integration with VTuber platforms and enhancing the viewer experience. These findings underscore the potential of SER models to revolutionize VTuber performances, reducing physical demands on performers and broadening accessibility. Future work will explore addaptation on Japanese language dataset, integrating visual emotion recognition, and expanding SER model capabilities across diverse languages.

Acknowledgements

I would like to express my deepest gratitude to **Dr. Steven Livingstone** from the Affective Data Science Lab (ADSL), Ontario Tech University, and **Dr. Ali Neshati** from the User Experience and Interaction Design (UXID) Lab, Ontario Tech University, for their invaluable guidance and supervision throughout the course of this project.

Special thanks to the CTO team members at **Hololive Production, Cover Corp**—for offering special insights into their system development and the challenges of implementing speech emotion recognition in real-world applications. I am especially grateful to VTuber, **Usada Pekora**, whose activity has been a personal inspiration and a driving motivation behind this project.

I also acknowledge the creators of the Hatsune Miku 3D models used in this work, particularly the Appearance Miku model by **ままま, アラン・スミシー** and Hatsune Miku NT model by **止丸**, both under ©Crypton Future Media, Inc., for enabling expressive avatar representations during prototype development.

Finally, I extend my heartfelt thanks to my friends and family for their unwavering support and encouragement throughout this journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, the digital entertainment landscape has witnessed a significant transformation with the rise of Virtual YouTubers, or VTubers. These virtual personas, characterized by their computer-generated avatars and anime-inspired aesthetics, have captivated audiences across platforms such as YouTube and Twitch. Originating in Japan in the early 2010s, the VTubing phenomenon has expanded worldwide, with more than 10,000 active VTubers by January 2020 [?]. The influence of VTubers is substantial with nearly 433 million hours watched in the third quarter of 2024 alone []. The market is projected to grow from USD 2.86 billion in 2025 to USD 4.50 billion by 2030, highlighting the enduring appeal and potential of this digital medium [27]. At the heart of a VTuber's charm is their hyper-expressive performance, which is often achieved through manual input to change facial expressions or through the use of sophisticated facial tracking technology. However, this reliance on manual control can lead to fatigue and a reduction in expressiveness over time, particularly during long streaming sessions. As an avid consumer of VTuber content, I observed that while VTubers initially utilize their animated avatars to their fullest, maintaining such expressiveness becomes challenging and often leads to a reliance on simpler expressions or camera facial-tracked systems.

This observation sparked my interest in developing a solution that could automate the expression selection process, enhancing the overall performance and engagement of VTubers. My research focuses on EmoAvatar, a voice-driven automatic avatar emotion animation selector system that utilizes Speech Emotion Recognition (SER) to predict and trigger appropriate avatar expressions based solely on the user's voice. By leveraging the capabilities of AI models, this system aims to seamlessly integrate with various VTuber applications, providing a scalable and versatile tool for content creators. The challenge lies not only in accurately predicting emotions from voice data, but also in adapting existing models to languages with fewer resources, such as Japanese. VTubers are particularly popular in Japan, yet the availability of Japanese speech emotion datasets is limited.

This research investigates the use of existing datasets and the fine-tuning of English-based SER models to cater to Japanese speech, addressing a significant gap in the field. Through this project, we aim to alleviate the physical demands on VTubers, enhance the expressiveness of streamers, and contribute to the growing field of AI-driven digital expression. By automating avatar animations in response to vocal emotions, we hope to enrich the VTuber experience for both creators and audiences, paving the way for future innovations in the realm of virtual entertainment.

1.0.1 Related Work

Virtual Avatars

Virtual avatars have become a central tool in online content creation, particularly for VTubers who engage audiences using expressive digital personas. These avatars serve not just as digital stand-ins for real humans, but as expressive, stylized characters capable of delivering a wide emotional range through visual motion. Their success relies heavily on real-time motion capture, facial animation, and gesture control, often mediated through lightweight or studio-grade tracking systems. At-home 3D streaming setups, such as those described in COVER Corp.'s "Ouchi-3D" system, combine multiple input modalities — including smartphone-based facial tracking, motion capture suits, and predefined animation triggers — to create a compelling virtual performance [11]. The goal in these systems is to offer high expressiveness with minimal friction, enabling creators to operate independently outside studio environments. These setups often support real-time body tracking (e.g., with devices like Sony's mocopi or Rokoko suits) alongside gesture presets (e.g., "wave," "sit," "jump") for simplified control.

Existing Avatar Animation Systems

Software platforms like 3tene, Animaze (FaceRig), and VRChat provide VTubers and virtual avatars with accessible tools for real-time avatar control. However, each has its own set of limitations and use cases. 3tene, a lightweight VTuber software, primarily uses webcam input to track facial features including head motion, eye movement, and mouth shape. [1] Head tracking captures tilting, turning, or nodding movements, which is then mirrored by the avatar's head in real time. Eye tracking detects the user's gaze direction and blinking, allowing the avatar to make eye contact and react naturally. Mouth tracking analyzes lip movements to enable real-time lip-syncing with spoken dialogue. These tracking features rely on standard camera input and computer vision algorithms to replicate the user's facial behavior on the avatar. Animaze, an earlier and widely used system, maps facial expressions using webcam-based facial tracking or Leap Motion devices. [2] Although Animaze supports some automatic expression mapping like Expression Targeting, expression switching in Animaze often requires keyboard shortcuts, limiting natural emotional flow during live performances. VRChat, one of the most popular platforms for social VR interaction, supports avatar expression switching via a radial menu, which users manually activate using a controller. [40] While functional, this method requires conscious input that can interrupt live interactions or performances. These systems, while capable of reflecting a range of physical expressions, often lack autonomous emotional expression, especially when it comes to speech-driven emotional output. This gap opens an opportunity for voice-driven emotional inference, bridging vocal tone with expressive animation.

Limitations of Current Emotion-Animation Systems

Despite advances in avatar tracking, current avatar animation platforms have several limitations:

- Manual Input Dependency: 3tene, Animaze, VRChat, and many expressive avatar software rely on explicit user input (keyboard or menu) to express emotions, which can disrupt flow and reduce immersion.

- Limited Emotional Responsiveness: Camera-based tracking systems offer facial fidelity (blinks, head tilt) but lack semantic emotional awareness derived from voice or context.

- Disconnection Between Modalities: There's a gap between vocal emotional cues and visual avatar expression, leading to unnatural character behavior, particularly when emotional tone in voice is not reflected visually such as crying or getting mad.

The proposed EmoAvatar system aims to bridge this gap by enhancing existing avatar animation platforms with real-time speech emotion recognition, enabling virtual avatars to automatically reflect nuanced vocal emotions through facial expressions.

Speech Emotion Recognition

Speech Emotion Recognition (SER) has evolved from traditional feature-engineered models to deep learning–based architectures capable of interpreting complex vocal patterns. [4, 7, 17], Early SER models depended on hand-crafted features such as MFCCs (Mel-Frequency Cepstral Coefficients), pitch, and prosody. Recent advancements utilize deep CNNs, transformers, and self-supervised models (SSL) such as wav2vec2.0, which learn rich audio representations from raw waveform input. [10, 38, 41] The Emotion2Vec model, based on data2vec family of SSL models and is explicitly optimized for emotion representation by fine-tuning on emotional speech datasets. Studies using deep-learning models have reported accuracy rates above 80% on datasets like RAVDESS and IEMOCAP, with multi-language models showing increasing generalization across speakers and languages [24]. However, while SER systems can now reliably classify emotions in offline tasks, their real-time integration with avatar systems remains underexplored. Additionally, most existing classifiers are trained primarily on English-language data, which limits their performance and generalizability on applying multilingual settings such as Japanese VTuber.

Emotion2Vec

[24] Emotion2Vec is a self-supervised learning (SSL) model specifically designed to extract emotion-aware speech representations across a wide range of emotional tasks. Unlike generic SSL models that focus broadly on speech content (e.g., phonemes or speaker identity), Emotion2Vec is tailored for emotion-centric representations, offering both task and language generalizability.

The model leverages a teacher-student architecture within an online distillation framework, combining both utterance-level loss and frame-level loss during pretraining. This dual-objective strategy ensures the model captures both global emotional context and fine-grained temporal emotional cues. The utterance-level loss learns holistic emotional tones across an audio clip, while the frame-level loss is applied to masked segments to preserve local emotion dynamics.

Emotion2Vec is initialized from powerful self-supervised models; data2vec and data2vec2.0, and pre-trained on 262 hours of emotional speech collected from diverse English-language datasets, including IEMOCAP, MELD, MEAD, CMU-MOSEI, and MSP-Podcast. The training process maintains a frozen encoder during downstream tasks, requiring only lightweight linear layers for emotion classification — a design that supports real-time or resource-constrained deployment.

In empirical evaluations, Emotion2Vec consistently outperforms other state-of-the-art models including WavLM, HuBERT, and Vesper, across benchmarks like IEMOCAP, RAVDESS, SAVEE, and MELD. Notably, it also demonstrates superior language generalization, achieving strong performance on 9 out-of-domain emotional speech datasets in languages such as Japanese, German, Urdu, and Persian — a critical advantage for global VTuber applications.

Additionally, Emotion2Vec exhibits strong task transferability, achieving state-of-the-art results not only in speech emotion recognition (SER) but also in song emotion recognition, emotion prediction in dialogue, and speech-based sentiment analysis. Visualization through UMAP further supports its superior emotional class separability and representational compactness compared to baseline SSL models.

Data2Vec and Data2Vec2.0

[5, 6] Data2Vec is a general-purpose self-supervised learning (SSL) framework introduced by Meta AI that unifies the training objective across speech, vision, and language modalities. Rather than predicting modality-specific tokens (such as words, visual patches, or speech units), Data2Vec instead predicts contextualized latent representations of the full input sequence. This enables the model to capture global and semantically rich features. It adopts a teacher-student architecture where the student model learns to predict the internal representations of the teacher model—an exponential moving average of the student's own parameters—based on a masked view of the input. This approach eliminates the need for discrete quantization or reconstruction of raw inputs, which were central to earlier SSL models like wav2vec2.0 and HuBERT. In speech tasks, Data2Vec showed competitive or superior performance compared to prior SSL methods, particularly in speech recognition and representation learning.

Data2Vec2.0 builds upon the original framework with a focus on efficiency improvements and faster convergence. It introduces an asymmetric encoder-decoder architecture that avoids encoding masked tokens altogether, significantly reducing computation, and a lightweight convolutional decoder speeding up training across all modalities. The model also reuses the same teacher-generated contextualized targets across multiple masked versions of a sample (multi-mask training), which amortizes the cost of computing teacher representations. Additionally, inverse block masking is employed to ensure local continuity in the unmasked regions, allowing for better context retention during learning. These changes allow Data2Vec2.0 to match or exceed the performance of state-of-the-art models like Masked Autoencoders (MAE), wav2vec2.0, and RoBERTa, while reducing training time by 2–16× depending on the task and modality.

Together, these two versions of Data2Vec demonstrate that contextualized target prediction is an effective and scalable paradigm for self-supervised learning across modalities. Their use as initialization backbones for Emotion2Vec ensures that the emotion-specific fine-tuning benefits from strong, general-purpose representations while maintaining efficiency and versatility for downstream emotion-related tasks.

1.0.2 Japanese and English Datasets

The datasets used from multiple sources, with a focus on those that included emotional labels for individual speech segments. Given the popularity of VTubers in Japan, datasets such as JVNV, STUDIES, OGVC, and HCUDB1 were selected as they are well-known resources for Japanese speech emotion recognition tasks. [8, 12, 25, 30, 42] However, additional evaluation was recommended due to the limited number of annotators involved in these datasets (shown in Chapter 3). For model performance evaluation, Japanese dataset evaluation results were compared against accuracy on the RAVDESS English speech dataset, as it is a widely used benchmark in the field of speech emotion recognition. [24]

JVNV

[42] JVNV (Japanese emotional speech corpus with Verbal content and Nonverbal Vocalizations) is a Japanese-language emotional speech corpus developed by Shinnosuke Takamichi in 2023. It comprises 1,617 utterances across six emotion categories: anger, disgust, fear, happy, sad, and surprise. The dataset features speech from four speakers (two female and two male), with utterances recorded in a soundproof environment at a 48 kHz sampling rate. Although the dataset does not provide intensity labels, it offers a detailed distribution of samples by gender and emotion. Each emotion category contains between 249 and 306 utterances, demonstrating relatively

balanced representation across classes. The dataset is annotated with 30 ratings on a subset of 60 balanced utterances, with each rater labeling 30 utterances. JVNV is also multimodal, providing both speech and transcriptions, and is freely available for non-commercial use upon request.

STUDIES (ITA)

[30] The STUDIES corpus is a Japanese emotional speech dataset designed for developing AI agents capable of empathetic dialogue, created by Yuki Saito and colleagues in 2022. The STUDIES corpus comprises multiple components, including a broader dialogue dataset featuring multi-turn empathetic conversations which requires significant pre-processing, and a processed and validated labeled subset known as STUDIES (ITA).

To focus development effort on EmoAvatar rather than extenive audio data engineering, the ITA subset was selected. The STUDIES (ITA) dataset is constructed around the scenario of a female cram school teacher casually chatting with students, with emotion labels and dialogue lines obtained through crowd-sourcing and later refined by the developers. It includes a total of 724 utterances, covering five emotional categories: angry, happy, neutral, sad, and recitation. The dataset was recorded by a single female speaker in a professional studio setting. Although it does not provide intensity annotations, it offers a structured distribution of 100 utterances per emotion category (except recitation, which contains 324).

STUDIES is multimodal, including transcriptions and dialog context metadata, making it valuable for emotion recognition in conversational settings. It is freely available for non-commercial research use upon request.

OGVC (Vol2)

[8] is a Japanese emotional speech corpus developed by Yoshiko Arimoto and colleagues, consisting of two distinct volumes: naturalistic emotional speech (Vol. 1) and acted emotional speech (Vol. 2). Volume 1 contains over 9,000 spontaneous utterances recorded during multiplayer online game sessions, designed to elicit natural emotional reactions. However, the raw audio is provided as mono recordings per speaker session and requires considerable pre-processing to align conversations and extract usable segments, which poses a significant overhead for direct model training. In contrast, Volume 2 is an acted and clearly labeled short short-audio dataset which fits our needs more effectively, so it was selected in this project. Volume 2 consists of 2,656 acted utterances in Japanese. It spans eight emotion categories: joy, accept, fear, surprise, sad, disgust, anger, and anticipate. The speech data was recorded by four speakers (2 male, 2 female) with a balanced gender distribution. Importantly, the dataset includes 4-level discrete emotion intensity labels, offering richer expressivity for modeling affective speech. It also provides transcriptions for each sample, making it suitable for both acoustic and multimodal emotion analysis. OGVC is publicly available for non-commercial research purposes.

HCUDB1

[25] HCUDB (Hiroshima City University Emotional Speech Corpus) is a Japanese emotional speech corpus developed to support emotion recognition research in spoken language. The corpus consists of two volumes: HCUDB1 and HCUDB2. Both volumes contain speech recordings in WAV format (48 kHz, 16-bit, mono), but differ in structure and annotation detail.

Since HCUDB2 has not yet been validated, HCUDB1 was used in this project. It includes 11 emotional categories, such as excited, surprise, angry, fear, happy, neutral, dislike, disgust, relaxed, sleepy/tired, and sad. The dataset features recordings from 14 speakers (8 female, 6 male) and incorporates 3-level discrete emotion intensity labels. Utterances were recorded in a soundproof professional studio, and the dataset includes 10 human raters who evaluated the emotional quality of the samples. HCUDB1 is also multimodal, containing speech data, speaker instructions, and evaluation results. It is publicly accessible for non-commercial use upon request and serves as a rich resource for Japanese-language SER research.

RAVDESS

[22] The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) is a well-known benchmark dataset for speech emotion recognition in English. Released in 2018 by Livingstone and Russo, it comprises 1,440 audio-only utterances performed by 24 professional actors (12 male, 12 female). It features eight emotion categories—neutral, calm, happy, sad, angry, fearful, disgust, and surprised—with each emotion (except neutral) recorded at both normal and strong intensity levels. All utterances were produced in a controlled studio environment. Each file was rated 10 times by 247 raters to assess emotional validity and intensity. The dataset is multimodal (audio, video-only, audio-visual) and includes transcriptions, making it a robust resource for both unimodal and multimodal SER studies. RAVDESS is freely available under a CC BY-NC-SA 4.0 license.

Chapter 2

Method

In this chapter, we describe the full architecture, components, and implementation details of the **EmoAvatar** system.

Beginning with an overview of the complete architecture and then dives into the individual components—from audio processing and emotion prediction to the integration with VTuber animation software. Each section is crafted to be understandable with general knowledge of AI and computer vision, even if they are not specialists in VTuber platforms or SER systems.

Through this breakdown, we aim to show how each module contributes to the seamless and intelligent animation of a virtual avatar, and how these modules interact in real time. Key emphasis is placed on practical challenges, such as latency and emotion-to-expression mapping, and how these were addressed to maintain usability in live streaming environments.

2.0.1 EmoAvatar Architecture

EmoAvatar integrates real-time Speech Emotion Recognition (SER) with virtual avatar animation platforms to enable seamless and intelligent expression control based

on vocal input. The architecture was designed with modularity and real-time interaction in mind, allowing for flexible system development and easy integration with VTuber software such as 3tene.
Our system enables VTubers to control facial expressions using their voice alone, reducing the physical and cognitive burden of managing emotions manually during live performances. This approach not only improves the user experience for streamers but also enhances the perceived realism and emotional coherence of the virtual avatar for viewers. By automating expression control, the system allows for smoother and more natural transitions between emotional states, making performances feel more lifelike. Ultimately, this reduces the acting burden on the VTuber—freeing them to focus more on content creation and audience engagement, rather than constantly managing their avatar's expressions. The architecture is composed of four main components:

Audio Input and Preprocessing: A microphone is setup to continuously capture speech voice input. The captured audio is then processed to detect silence, to apply basic audio engineering steps for best result, and to match the model's input requirements. These tasks are handled by the Audio Class, which produces standardized audio files that match the requirements of the prediction model.

EmoAvatar Frontend: A EmoAvatar processing pipeline and a graphical user interface (GUI) that allows user to start or stop the EmoAvatar system and configure parameters such as audio threshold for silence detection and recording duration. While the current prototype includes logging features for debugging, these are intended for development use, and could be removed on a production setting.

Prediction Class: Handled in backend to predict emotion using the Emotion2Vec model. It receives preprocessed audio, performs inference using FunASR's API, extracts the predicted emotion label, and sends the result via a socket connection to the animation system. The backend is optimized for utterance-level prediction and ensures minimal delay to support real-time expression control.

Avatar Animation Software: The software supports real-time facial and motion tracking as well as external expression input, enabling VTubers to animate their avatars live. In this project, we used 3tene, a symple and lightweight yet include all the feature we need to drive VTuber animation that allows integration with camera-based tracking and keyboard-triggered expressions.

- Head, eye, and mouth tracking - Camera input (3tene): 3tene uses a standard webcam to capture head movement, eye direction, and lip sync. This enables the avatar to mimic the user's natural facial motions during live performance.

- Avatar expression control: High-level facial expressions (e.g., angry, happy, sad) are triggered using keyboard shortcuts in 3tene. In our system, these shortcuts are automatically activated by the SER model's output, allowing the avatar's expressions to change in sync with the speaker's vocal emotion. While physical tracking handles continuous motion (eye blink, mouth, head tilt), SER predictions control discrete expressions such as joy, surprise, or fear.

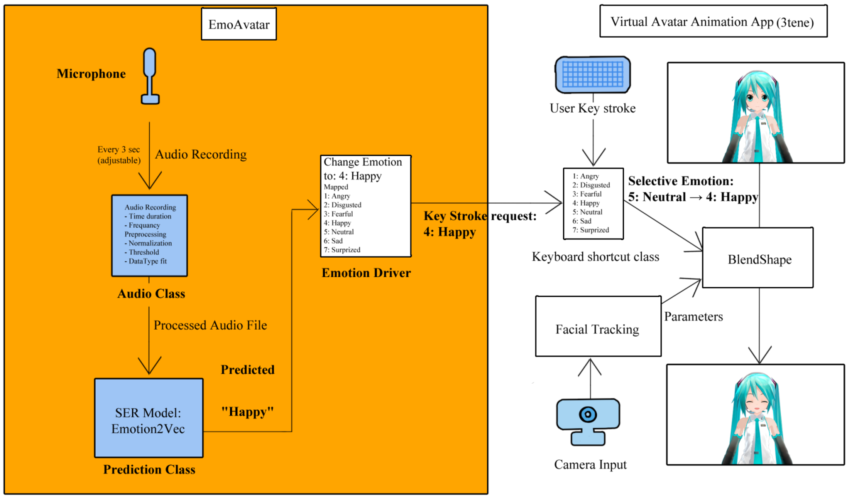


Figure 2.1: EmoAvatar Architecture

2.0.2 Audio Class

The Audio Class serves as the initial stage in the EmoAvatar pipeline, capturing and preparing raw voice input from for emotion inference. It is responsible for microphone recording, signal preprocessing, and standardization to meet the format and quality requirements of the Speech Emotion Recognition (SER) model.

Audio input is handled using the `sounddevice` and `soundfile` libraries in Python. The recording function is encapsulated in the `record_audio()` method, which accepts three parameters: output file path, recording duration, and sampling rate. In this project, the default configuration uses a 3-second window and a 16 kHz mono-channel format to ensure consistency with the pretrained Emotion2Vec model. This 3-second setting was chosen based on empirical observations during testing with the 3tene VTuber platform, where it provided a practical trade-off between latency and emotional prediction performance in live scenarios. These parameters are configurable and can be adjusted by developers to meet the demands of different applications. For example, increasing the recording duration may improve the model's confidence by supplying more speech data, though it introduces additional delay. Conversely, a shorter window can minimize lag—ideal in cases where the model is capable of accurate predictions from brief utterances. This flexibility allows EmoAvatar to adapt to diverse avatar control software.

The key steps in the recording and preprocessing pipeline are as follows:

- **Microphone Capture**: Voice input is recorded using the system's default microphone. The recorded signal is stored in 32-bit float format to retain high precision for downstream tasks.

- **Amplitude Normalization**: The waveform is normalized to a fixed amplitude scale, ensuring consistent volume across recordings. This is crucial for reducing sensitivity to variations in speaker loudness or environmental acoustics

  $$\hat{x} = \frac{x}{\max(|x|)}$$

  Where:

  - x is the original audio signal (vector of sample values)
  - $\hat{x}$ is the normalized signal
  - $\max(|x|)$ is the maximum absolute value in the audio signal

- **Channel Handling**: If stereo input is detected, the system retains only the first channel to enforce a consistent single-channel format.

- **Data Typing and Output**: The preprocessed waveform is cast to 32-bit float and saved to disk using the `soundfile` library. This output is used as the input to the SER model. [24]

In addition to recording, the module also contains a utility function, `is_audio_silent()`, which performs a basic energy-based silence check by comparing the maximum signal amplitude against a fixed decibel threshold. This allows the system to skip processing when the captured audio is considered silent or below meaningful vocal thresholds.

2.0.3 Prediction Class

The backend module of the EmoAvatar system is responsible for processing incoming audio and predicting the speaker's emotional state using a pre-trained deep learning model.

Speech Emotion Recognition Models and Setup

The core of this study is the implementation of Speech Emotion Recognition (SER) to automate avatar expressions. We employed the **Emotion2Vec** model, an extension of Meta's Data2Vec series. [24] The model is loaded using the `FunASR` framework, which provides a simplified interface for executing inference on utterance-level audio files.

Emotion Prediction Process

The backend accepts a standardized `.wav` file (`recorded_audio.wav`) as input. Using the `generate()` method, the model performs inference with utterance-level granularity, returning both label names and corresponding confidence scores. The result is parsed to identify the emotion with the highest probability score.

To ensure clean communication, the predicted emotion label is filtered to remove any non-alphabetic characters using a regular expression. This step guards against artifacts in the label strings that may arise from formatting or encoding inconsistencies.

Real-Time Output Transmission

Once the emotion is identified, it is transmitted over a local TCP socket to the Emotion Driver via the `send_emotion_to_display()` function. This function connects to a predefined port (`localhost:5000`) and streams the emotion as a UTF-8 string. This real-time pipeline allows the Emotion Driver to receive emotional states as they are detected and apply them immediately.

2.0.4 Emotion Driver

¡¡ The **Emotion Driver** acts as a middleware that converts emotion labels detected by the SER backend into expression commands for the VTuber avatar animation platform. In this project, we utilize simulated keyboard inputs to trigger facial expressions in the **3tene** software, which maps specific keys to avatar expressions. This approach provides a simple yet effective method for real-time expression control without requiring direct modification of the VTuber software.

Emotion-to-Key Mapping and Customization

The driver uses a predefined mapping between predicted emotions and keybindings supported by the 3tene application and the target avatar model. For example:

Emotion	Key
happy	j
sad	s
angry	a
surprised	x
disgusted	s
fearful	x
neutral	n

These mappings reflect the configuration of the current avatar model and 3tene environment, but they are fully customizable. Developers can adapt the key mappings to fit different VTuber applications or avatar rigs, or bypass keyboard input entirely by interfacing directly with platforms that support expression control through API calls.

Emotion Compatibility and Fallbacks

Not all avatars or VTuber platforms support the full range of emotions that may be predicted by the SER model. To address this, the Emotion Driver includes a mechanism to handle unsupported emotions.

In our prototype, the model supports *angry*, *joy*, *sad*, *surprise*, and *neutral.* Both *fearful* and *disgusted* are not supported in the avatar model, so we re-mapped to *surprise* and *sad* accordingly.

Developers are encouraged to define fallback rules—for example, mapping unsupported emotions such as *disgust* or *fear* to *surprise* expression—to ensure consistent avatar behaviour without visual artifacts or interruptions in expression flow. This ensures that even when the SER model produces a label not supported by the avatar configuration, the system continues to operate seamlessly by dropping or remapping that emotion in a controlled and visually coherent way.

Efficiency and Filtering

To reduce redundant triggers and maintain smooth avatar behavior, the driver tracks the current active emotion and only sends input when a new emotion differs from the previous one. A cooldown interval (e.g., 1 second) is introduced between expression switches to avoid flooding the animation system with rapid transitions, improving both responsiveness and viewer experience.

Integration Flexibility

This module is intended to be extensible. While this implementation uses key simulation via `pynput` for 3tene, it can be adapted to other applications, such as VRChat or Unity-based platforms, by modifying the communication protocol or expression command format. The separation between the SER backend and the animation frontend ensures that the EmoAvatar system can be ported to diverse environments with minimal effort.

2.0.5 EmoAvatar Frontend

The frontend module drives the high-level pipeline logic and provides a simple graphical user interface (GUI) through which the user can interact with the EmoAvatar system. Built using `Tkinter`, it allows users to start or stop the system and configure runtime parameters such as the recording window duration and silence detection threshold. These settings directly affect how the system captures and filters incoming audio for emotion prediction.

When activated, the frontend initiates a background loop that periodically records short audio segments using our Audio Class and forwards them to the Prediction Class to predict the emotion. This loop continues until the user stops it, allowing for continuous real-time inference.

Notably, this frontend does not support manual overriding of predicted emotions, nor does it support live editing of expression mappings. Instead, it is designed to act as a lightweight controller, making the system accessible and configurable for end-users with minimal interaction.

2.0.6 Combining with VTuber App

3tene

3tene is a real-time 3D and 2D avatar animation software developed by plusplus Inc., widely adopted by VTubers and online creators for its intuitive interface and high compatibility with webcams and motion tracking devices. The software enables facial and head tracking using a standard webcam, translating the user's expressions and movements into real-time animations. It supports multiple formats of avatar models, including VRM and Live2D, making it accessible for a broad range of users from hobbyists to professionals.

Facial expressions in 3tene are controlled via keyboard shortcuts, which are mapped to predefined avatar expressions such as happiness, sadness, anger, and surprise. These mappings can be customized by the user based on their avatar configuration. 3tene also supports body motion presets and environmental settings, though, facial expression switching remains on our most prominent features for enhancing expressiveness during live VTuber performances.

Since 3tene does not currently offer an external API for real-time expression control, our automation is achieved by simulating these keyboard inputs programmatically. This makes it a practical integration point for external systems such as EmoAvatar, which use SER outputs to drive avatar expressions.

User Input and Motion Tracking

¡talk detail¿ VTuber applications like 3tene utilize camera-based facial tracking to map a user's facial behavior onto a virtual avatar in real time. This tracking encompasses several key components:

- **Head Tracking**: Captures movements such as tilting, nodding, and turning of the user's head, allowing the avatar to reflect subtle changes in posture and attention direction.
- **Eye Tracking**: Detects gaze direction and blinking patterns, enabling the avatar to make eye contact and display eye movements that align with natural conversation flow.
- **Mouth Tracking**: Synchronizes the avatar's lip movements with the user's speech by visually tracking mouth shape and movement, supporting dynamic lip-syncing during speech.

These features contribute significantly to the avatar's realism and help maintain user immersion. However, they operate independently of emotional tone. EmoAvatar addresses this gap by adding voice-based emotional context, enabling the avatar to convey affective reactions in sync with vocal delivery, providing a more holistic and expressive virtual representation of the user, enhancing both the performer's expressiveness and the viewer's engagement.

Integrate EmoAvatar to 3tene

Once the Emotion Driver receives an emotion prediction from the SER model (e.g., “Happy”), it determines the corresponding key mapping that is associated with that emotion in the target VTuber software—in this case, 3tene. For example, “Happy” might be mapped to key 4, and key 4 virtually pressed by the Emotion Driver. This mimics a user pressing the corresponding expression-switch key within the 3tene interface.

These key inputs are passed into 3tene’s Keyboard Shortcut Class, which is responsible for managing the internal switching of expression states. For instance, if the avatar is currently in a neutral state and receives a key stroke request to express happiness, the system will transition from “5: Neutral” to “4: Happy.” The selected emotion is then sent to the BlendShape Class, which controls the facial mesh deformation and triggers the corresponding expression animation. This class blends the SER-based expression change with existing real-time tracking data including eye tracking, eyebrow movement, head rotation and tilt, and lip synce based on webcam input. This blending ensures that while facial expressions reflect the emotional prediction from voice, the avatar’s gaze direction, blinking, and head motion remain reactive to the user’s camera-based tracking — maintaining natural realism and interactivity. As a result, SER-driven transitions appear smooth and organically synchronized with the physical motion captured in real-time.

This integrated approach allows EmoAvatar to extend 3tene’s expressive capability with minimal intrusion, enabling VTubers to automate emotion-driven animations while preserving camera-based motion fidelity.

Chapter 3

Experiments

This chapter presents the experiments conducted to evaluate the performance of the Emotion2Vec model on Japanese emotional speech datasets compared with English dataset, with the primary goal of assessing the model's applicability to real-time avatar expression control in VTuber applications.

3.0.1 Datasets

To evaluate the model's performance in multilingual contexts, we selected several speech emotion corpora in both Japanese and English. The Japanese datasets—JVNV, STUDIES (ITA), OGVC (Vol2), and HCUDB1—were chosen for their relevance to VTuber use cases and their range of emotional categories. RAVDESS was used as a strong baseline for English from its structured and balanced emotional speech clips make it a reliable benchmark for SER models. [24]

JVNV [42]

- **Dataset Name:** JVNV (Japanese emotional speech corpus with Verbal content and Nonverbal Vocalizations)
- **Author:** Shinnosuke Takamichi (2023)

- **Language:** Japanese
- **Emotion Categories (6):**
  - Anger, Disgust, Fear, Happy, Sad, Surprise
- **Speakers:** 4 (2 female, 2 male)
- **Intensity Labels:** No
- **Emotion Distribution:**

Emotion	Female	Male	Total
Anger	130	119	249
Disgust	133	125	258
Fear	137	128	265
Happy	148	132	280
Sad	132	125	257
Surprise	153	133	306

Table 3.1: JVNV Emotion Distribution

- **Total Utterances:** 1617
- **Recording Condition:** Acted in a soundproof room, sampling frequency: 48 kHz
- **Rater:**
  - 30 ratings on 60 emotion-balanced samples
  - Each rater labeled 30 utterances
- **Multimodal:** Yes (speech + transcriptions)
- **Publicly Available:** Yes (Freely available for non-commercial use upon request)

STUDIES (ITA) [30]

- **Dataset Name:** STUDIES (ITA)
- **Author:** Yuuki Saitou (2022)
- **Language:** Japanese

- **Emotion Categories (5):** Angry, Happy, Neutral, Sad, Recitation
- **Speakers:** 1 (Female)
- **Intensity Labels:** No
- **Emotion Distribution:**

Emotion	**Total**
Angry	100
Happy	100
Neutral	100
Sad	100
Recitation	324

Table 3.2: STUDIES (ITA) Emotion Distribution

- **Total Utterances:** 724
- **Recording Condition:** Acted by professional speaker, studio-recorded
- **Rater:** N/A
- **Multimodal:** Yes (speech + transcriptions + dialog context)
- **Publicly Available:** Yes (Freely available for non-commercial use upon request)

OGVC (Vol2) [8]

- **Dataset Name:** OGVC
- **Author:** Yoshiko Arimoto (2012)
- **Language:** Japanese
- **Emotion Categories (8):** Joy, Accept, Fear, Surprise, Sad, Disgust, Anger, Anticipate
- **Speakers:** 4 (2 Male, 2 Female)
- **Intensity:** Yes – 4 level discrete emotion categories

- **Emotion Distribution:**

Emotion	**Female**	**Male**	**Total**
Joy	168	168	336
Accept	160	160	320
Fear	160	160	320
Surprise	192	192	384
Sad	168	168	336
Disgust	160	160	320
Anger	160	160	320
Anticipate	160	160	320

Table 3.3: OGVC (Vol 2) Emotion Distribution

- **Total Utterances:** 2656
- **Recording Condition:** Acted (balanced speaker distribution)
- **Rater:** N/A
- **Multimodal:** Yes (speech + transcriptions)
- **Publicity:** Yes (Freely available for non-commercial use upon request)

HCUDB1 [25]

- **Dataset Name:** HCUDB1
- **Author:** Hiroshima City University (2024)
- **Language:** Japanese
- **Emotion Categories (11):** Excited, Surprise, Angry, Fear, Happy, Neutral, Dislike, Disgust, Relaxed, Sleepy/Tired, Sad
- **Speakers:** 14 (8 female, 6 male)
- **Intensity:** Yes – 3 level discrete emotion categories
- **Emotion Distribution:**

Emotion	**Female**	**Male**	**Total**
Excited	210	210	420
Surprise	210	210	420
Angry	210	210	420
Fear	210	210	420
Happy	210	210	420
Neutral	210	210	420
Dislike	210	210	420
Disgust	210	210	420
Relaxed	210	210	420
Sleepy/Tired	210	210	420
Sad	210	210	420

Table 3.4: HCUDB1 Emotion Distribution

- **Total Utterances:** 4620
- **Recording Condition:** Acted by professionals, recorded in a soundproof studio
- **Rater:** Yes – 10 raters been asked for
- **Multimodal:** Yes (speech + speech instruction + evaluations)
- **Publicity:** Yes (Freely available for non-commercial use upon request)

RAVDESS [22]

- **Dataset Name:** RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)
- **Author:** Steven R. Livingstone and Frank A. Russo (2018)
- **Language:** English (North American accent)
- **Emotion Categories (8):** Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised
- **Speakers:** 24 (12 Male, 12 Female)
- **Intensity:** Yes – Each emotion (except Neutral) expressed at two levels: normal and strong

- **Emotion Distribution:**

Emotion	**Female**	**Male**	**Total**
Neutral	48	48	96
Calm	96	96	192
Happy	96	96	192
Sad	96	96	192
Angry	96	96	192
Fearful	96	96	192
Disgust	96	96	192
Surprised	96	96	192

Table 3.5: RAVDESS Emotion Distribution

- **Total Utterances:** 1440 (Audio-only speech files)
- **Recording Condition:** Acted by professional actors; recorded in a controlled studio environment
- **Rater:** Yes – Each file rated 10 times on emotional validity, intensity, and genuineness by 247 individuals
- **Multimodal:** Yes (speech + transcriptions + video-only + audio-visual)
- **Publicity:** Yes (Freely available for non-commercial use under CC BY-NC-SA 4.0 license)

3.0.2 Dataset Preprocessing and Label Alignment

To ensure a consistent evaluation across all datasets, we adopted a standardized audio preprocessing and prediction pipeline.

Each dataset was first scanned recursively to extract all audio files. A predefined emotion mapping dictionary was used to assign high-level emotion categories (e.g., *angry*, *happy*, *sad*) to each audio sample based on file path or filename metadata. Files not matching any known emotion code were categorized under *other*. For each audio file:

- The audio was loaded using the `soundfile` library.

- Audio was resampled to 16kHz to match the input requirement of the Emotion2Vec model.

- Stereo channels were converted to mono by selecting the first channel.

- Waveforms were normalized to [-1, 1] range using peak normalization.

- Audio was cast to 32-bit floating point to match model expectations.

Emotion Remapping

In the benchmarking of Emotion2Vec on Japanese datasets, a key preprocessing step involved mapping raw file-based emotion labels to a consistent, high-level emotional taxonomy. The purpose of this was to align labels from multiple sources with the output classes supported by the pretrained Emotion2Vec model.

Different emotional speech corpora use varying labels and granularity for emotion annotations. For instance, while one dataset may label an utterance as "anger," another might use "furious" or "annoyance" to denote a similar affective state. This label inconsistency poses a challenge when comparing model performance across datasets or integrating them into a unified system like EmoAvatar.

The implementation assigns unmatched emotions to the "other" class. This ensures the model does not misclassify emotions it wasn't trained on or that fall outside its classification scope. This fallback strategy aligns with best practices in recent SER research, which recommend excluding or consolidating low-frequency or unsupported emotions into a general category to avoid misinterpretation. [4, 13]

3.0.3 Emotion2Vec Statistical and Benchmarking Approaches

Benchmarking the SER models involved evaluating their performance against same metrics with Emotion2Vec were been measured such as Weighted Accuracy (WA), Unweighthed Accuracy (UA), and Weighted F1-score(WF1) to compare the accuracy on Japanese Dataset.

- **Weighted Accuracy (WA)**

  The overall accuracy considering class imbalance — i.e., the proportion of correctly predicted samples over the total samples.

  $$\text{WA} = \frac{\text{Number of correct predictions}}{\text{Total number of samples}}$$

- **Unweighted Accuracy (UA)**

  The average of the recall (accuracy) for each emotion class, giving equal weight to each class regardless of its frequency.

  $$\text{UA} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{True Positives in class } c}{\text{Total samples in class } c}$$

  where C is the total number of emotion classes.

- **Weighted F1 Score (WF1)**

  The harmonic mean of precision and recall, weighted by the number of true instances for each class. This is especially useful when dealing with class imbalance.

  $$\text{WF1} = \sum_{c=1}^{C}\frac{N_c}{N}\cdot F1_c$$

  where N_c is the number of samples in class c, N is the total number of samples, and $F1_c$ is the F1 score of class c.

Each processed waveform was passed to the `predict_emotion()` function, which returned a predicted label. To handle robustness issues, prediction failures (e.g., due to corrupted files or unexpected formats) were logged and excluded from the final accuracy computation. Post-processing involved:

- Filtering out samples with failed predictions (`None`).
- Aligning prediction outputs with ground-truth labels.
- Computing accuracy metrics only on successfully processed samples.

The benchmarking procedure allowed us to directly measure the model's classification capabilities in both matched (English) and mismatched (Japanese) language conditions, revealing cross-linguistic performance differences that inform the generalizability and limitations of the model. This approach enabled a fair and replicable comparison of Emotion2Vec's effectiveness across diverse linguistic and emotional data distributions Statistical analysis of the model's predictions helped benchmarking its model performance on Japanese speech, particularly for nuanced emotional expressions that are critical in VTuber performances.

3.0.4 EmoAvatar Performance

Unlike traditional classification tasks, evaluating the performance of real-time avatar animation systems like EmoAvatar poses a unique challenge. There is currently no standardized quantitative metric to directly assess the quality of avatar expressiveness, responsiveness, or realism as perceived by viewers during live interaction.

Instead, performance was qualitatively assessed through observation in a simulated streaming environment.

While no formal matrix is applied in this evaluation, we observe the system's real-time responsiveness, low latency, and consistency between vocal tone and avatar behavior which are key factors in evaluating the success of the prototype. Further usability studies and user feedback collection would be beneficial in future work to more systematically evaluate emotional fidelity and perceived expressiveness.

Chapter 4

Result

4.0.1 Emotion2Vec Benchmarknig

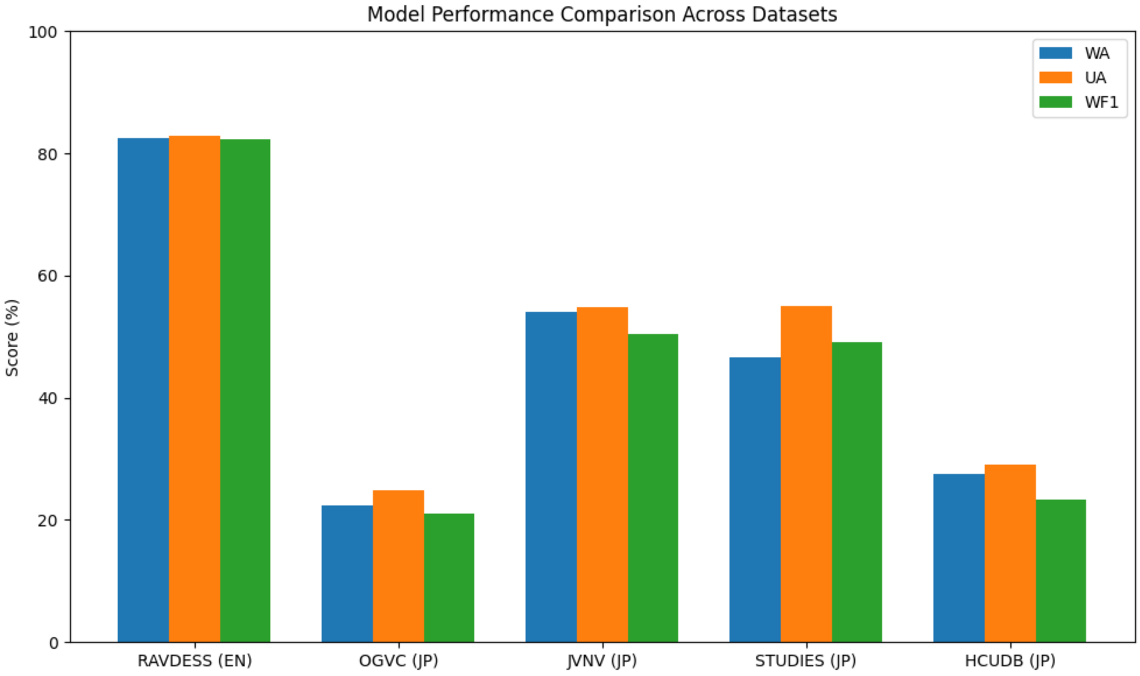


Figure 4.1: Model Performance Comparison across Datasets

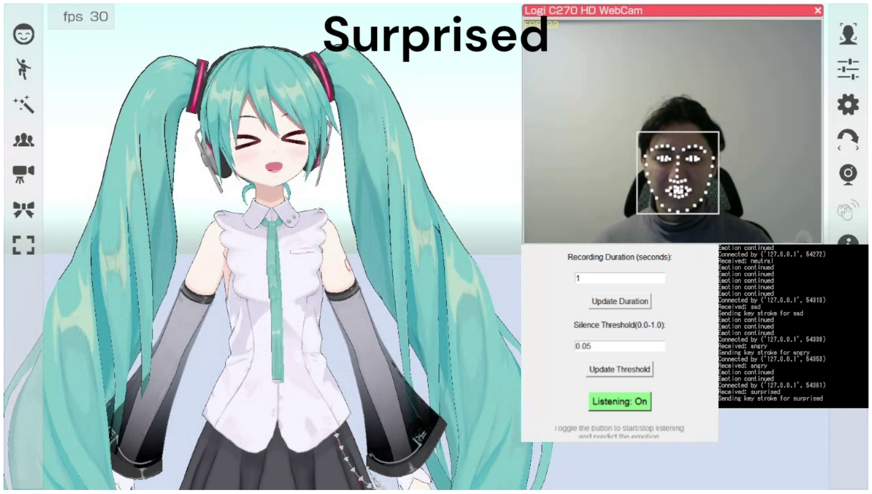
Dataset	**WA (%)**	**UA (%)**	**WF1 (%)**
RAVDESS (EN)	82.43	82.86	82.39
OGVC (JP)	22.44	24.91	20.95
JVNV (JP)	54.06	54.90	50.51
STUDIES (JP)	46.69	55.08	49.08
HCUDB (JP)	27.51	29.04	23.41

Table 4.1: Performance of the Emotion2Vec model on English (RAVDESS) and Japanese datasets

To evaluate the performance of the Emotion2Vec model across languages, we conducted benchmark testing using five emotional speech datasets — one in English (**RAVDESS**) and four in Japanese (**OGVC, JVNV, STUDIES, HCUDB**). The evaluation used three metrics: **Weighted Accuracy (WA)**, **Unweighted Accuracy (UA)**, and **Weighted F1-score (WF1)**. The results are summarized in the Table 4.1 and visualized in Figure 4.1.

While Emotion2Vec model shows strong performance on English data, achieving over 82% across all three metrics on the RAVDESS dataset, performance on Japanese datasets is notably lower, with WA scores ranging from **22.4% (OGVC)** to **54.1% (JVNV)**. The STUDIES dataset showed particularly high UA (55.08%), indicating relatively balanced performance across emotional classes despite overall lower WA.

4.0.2 EmoAvatar



Recording Duration (seconds):
1

Silence Threshold(0.0-1.0):
0.05

Toggle the button to start/stop listening and predict the emotion

```
Emotion continued
Connected by ('127.0.0.1', 54272)
Received: neutral
Emotion continued
Emotion continued
Emotion continued
Emotion continued
Emotion continued
Connected by ('127.0.0.1', 54313)
Received: sad
Sending key stroke for sad
Emotion continued
Emotion continued
Emotion continued
Connected by ('127.0.0.1', 54339)
Received: angry
Sending key stroke for angry
Connected by ('127.0.0.1', 54353)
Received: angry
Emotion continued
Emotion continued
Connected by ('127.0.0.1', 54361)
Received: surprised
Sending key stroke for surprised
```

Figure 4.2: EmoAvatar Demonstration

This video highlights the **EmoAvatar** system to trigger facial expressions based on vocal emotion in real time, demonstrating the end-to-end integration between speech emotion recognition and avatar animation using 3tene.

A public demonstration video showcasing the system in action is available at:
https://youtu.be/xmYV-EMWq7I

And the demonstrated EmoAvatar scripts can found under the project repository:
https://github.com/AKinoshita0915/EmoAvatar

Chapter 5

Discussion

In this project, we explored the classification of vocal emotions to automate avatar expressions for VTubers, using the vocal emotion classification, with Emotion2Vec serving as the core speech emotion recognition (SER) model. Our approach aiming to alleviate the physical and cognitive demands VTubers face when managing expressions manually. By utilizing the user's voice to drive avatar expressions, we sought to enhance both performer expressiveness and audience immersion and engagement. The Emotion2Vec model demonstrated strong performance in classifying key emotional states on English datasets, however, its accuracy dropped substantially when tested on Japanese emotional speech, highlighting a significant language-related performance gap. This underscores the need for further model adaptation, such as fine-tuning on Japanese datasets and expanding the range of training data available in low-resource languages. These findings also point toward broader challenges in multilingual SER. Future work may involve the development of Japanese-specific emotion recognition architectures, augmentation of underrepresented emotional categories, and integration of phonetic and prosodic features unique to Japanese speech. Alternative architectures such as convolutional neural networks (CNNs) or hybrid models could also be explored to improve generalizability.

From a system perspective, a practical limitation encountered was a slight delay between the user's vocal input and the avatar's response. Although this latency remained within tolerable limits, it affected the overall sense of real-time interactivity. We also observed that environmental noise negatively impact SER prediction accuracy—suggesting that incorporating real-time noise suppression or speech enhancement modules could further increase system robustness.

Despite theses challenges, the integration of voice-driven animations with real-time SER processing capabilities marks a significant advancement in VTuber technology. **EmoAvatar** shown reduce the need for manual control enables more fluid, emotionally resonant interactions between streamers and their audiences.

5.0.1 Future work

Looking ahead, several key areas for future work include:

- **Optimization for Japanese Speech**: Fine-tuning models on Japanese data or building language-specific SER modules will be essential for broader applicability within the VTuber community, where Japanese is dominant.

- **Real-time Performance Improvements**: Reducing audio-processing latency and noise-cancelling will further enhance interactivity and durability, especially in live settings.

- **Context-Aware Emotion Recognition**: Incorporating language models (LLMs) to interpret not only tone but also spoken content could significantly improve emotion prediction fidelity and contextual alignment.

- **Platform Adaptability**: Extending support beyond 3tene to platforms such as VRChat, Unity-based 3D environments, or AGI-driven avatars could unlock new use cases across virtual entertainment, education, and social interaction.
- **Multimodal Emotion Fusion**: Combining SER with facial expression recognition, gesture analysis, and dialogue sentiment will allow for more holistic avatar animation.
- **Expression Mapping Frameworks**: Defining configurable fallback rules and dynamic expression blending will enable the system to adapt more flexibly to different avatar models and emotional taxonomies.

Chapter 6

Conclusions

This study has addressed the significant challenge of manual avatar expression control in VTuber performances, which often leads to fatigue and limits expressiveness. By developing a voice-driven automatic avatar emotion animation selector, we utilized Speech Emotion Recognition (SER) models to automate avatar expressions based on vocal inputs. The Emotion2Vec model was implemented to predict and animate emotional states in real-time.

The results demonstrated that while the Emotion2Vec model performs effectively on English datasets, its performance on Japanese data dropped significantly , indicating need for further fine-tuning and expansion of existing datasets. Despite these challenges, the application achieved smooth integration with VTuber platforms and enhancing the interactive experience for viewers. This advancement suggests a promising direction for reducing the physical demands on VTubers and increasing accessibility for content creators across diverse linguistic backgrounds.

As we look to the future, there is significant potential to integrate visual emotion recognition with voice-driven predictions, providing a more comprehensive approach to avatar animation. Although this study did not include full fine-tuning of the model to adapt specifically to Japanese speech, this remains a significant area for future research. Additionally, it was observed that Japanese SER datasets themselves offer substantial room for further investigation and development. Expanding SER model capabilities to accommodate a broader range of languages and emotional expressions will further transform the VTuber landscape, offering innovative tools for more dynamic and engaging performances. Further more, there is a big room to expand the system beyond Vtber apps, into platforms like VRChat, Unity-based 3D animation, or even AGI-driven avatars. Continued research and development in this area will pave the way for even greater advancements in digital entertainment and virtual interactions.

Appendix: Accessing the Japanese Speech Emotion Dataset

For researchers interested in accessing the Japanese Speech Emotion datasets used in this study, such as STUDIES and OGVC, the following outlines the general procedure and my personal experience with the registration and access process.

1. General Access Process

Most Japanese academic speech corpora are distributed via the National Institute of Informatics (NII) or partner institutions. These datasets typically require researchers to:

- Visit the dataset's official application page.
- Agree to the terms of use, including non-commercial restrictions and annual reporting.
- Submit an online or PDF-based application form.

Use translation features such like built-in translator for Google Chrome for Japanese-language pages, as many of the sites and forms are not in English.

2. Registeration

STUDIES Corpus

To access the STUDIES corpus, visit the NII Speech Resources Consortium application page: `https://www.nii.ac.jp/dsc/idr/speech/submit/STUDIES.html`

Registration process:

1. Visit the website above and find the STUDIES dataset.
2. Fill out the online application form. It is available in Japanese, though, you can fill in English.
3. If you're an international applicant, the administrator may contact you for an English version form.
4. Submit the application form and await approval.
5. Once approved, the download link will be provided via email.

OGVC Corpus

The steps to access the OGVC corpus are similar to the STUDIES corpus.

Access: `http://doi.org/10.32130/src.OGVC`

UUDB Corpus

The steps to access the UUDB corpus are similar to the STUDIES corpus.

Access: `https://research.nii.ac.jp/src/en/UUDB.html`

HCUDB Corpus

The steps to access the HCUDB corpus are similar to the STUDIES corpus.

Access: `https://www.nii.ac.jp/dsc/idr/speech/submit/HCUDB.html`

JVNV Corpus

JVNV Corpus is available on the research website:
`https://sites.google.com/site/shinnosuketakamichi/research-topics/jvnv_corpus`

3. Follow the Registration Steps

Complete the registration form as instructed on the website. Ensure that all required fields are filled out accurately to facilitate the approval process.

4. Annual Reporting Requirement

If you use the dataset in a publication, you are expected to notify the consortium by submitting a report. If you have not used the data or published during the year, you may skip this step. A reminder is typically sent via email at year-end.

For further inquiries, consult the Speech Resources Consortium portal:
`https://research.nii.ac.jp/src/en/`

Bibliography

[1] 3tene - vtuber real-time facial capture software. `https://3tene.com/`. Accessed: 2025-04-15.

[2] Animaze 2d avatar special actions guide. `https://www.animaze.us/manual/gettingstarted2d/specialactions?utm_source=chatgpt.com`, 2025. Accessed: 2024-12-20.

[3] Aaron van den Oord, Oriol Vinyals, K. K. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937v2* (2018).

[4] Akçay, M. B., and Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication 116* (2020), 56–76.

[5] Alexei Baevski, Wei-Ning Hsu, Q. X. A. B. J. G. M. A. data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555* (2022).

[6] Alexei Baevski, Wei-Ning Hsu, Q. X. A. B. J. G. M. A. data2vec 2.0: General-purpose self-supervised learning with fast convergence. *arXiv preprint arXiv:2301.10936* (2023).

[7] Ando, A. 音声感情認識の技術動向——深層学習に基づく手法とその最新研究——. 日本音響学会誌 *79*, 1 (2023), 72–79.

[8] Arimoto, Y., and Kawatsu, H. Online gaming voice chat corpus with emotional label (ogvc). *unpublished. https://mac-lab. org/research-projects/ogvc.*

[9] Arimoto, Y., Kawatsu, H., Ohno, S., and Iida, H. Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment. *Acoustical Science and Technology 33*, 6 (2012), 359–369.

[10] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems 33* (2020), 12449–12460.

[11] カバー株式会社. おうち3d配信を支えるトラッキングシステムについて. `https://note.cover-corp.com/n/n4869b6758c67`, 2024. Accessed: 2025-04-12.

[12] 山本悠, 鈴木裕太, 土屋奎太, 陳キュウ, et al. Gan による音声感情を反映させたフォント自動生成システムの改良. vol. 2024, pp. 555–556.

[13] 有本泰子. コーパス使いますか? 作りますか?——感情音声分析のためのコーパス構築——. 日本音響学会誌 *79*, 1 (2022), 64–71.

[14] 森 建人, and 矢野 良和. 短時間発話からの音声感情認識のための音声データ選別法に関する検討. 日本知能情報ファジィ学会 ファジィ システム シンポジウム 講演論文集 *25* (2009), 68–68.

[15] 生形優也, 田村仁, et al. 深層学習を用いた音声感情認識. 第 *86* 回全国大会講演論文集 *2024*, 1 (2024), 409–410.

[16] 秋山 大知, 石川 智希, 井本 桂右, 新妻 雅弘, 山西 良典, AND 山下 洋一. 音声を用いた感情認識のための学習話者の選択. vol. 76, pp. 554–561.

[17] DE LOPE, J., AND GRAÑA, M. An ongoing review of speech emotion recognition. *Neurocomputing 528* (2023), 1–11.

[18] DERINGTON, A., WIERSTORF, H., ÖZKIL, A., EYBEN, F., BURKHARDT, F., AND SCHULLER, B. W. Testing correctness, fairness, and robustness of speech emotion recognition models. *IEEE Transactions on Affective Computing* (2025), 1–14.

[19] INOUE, T., AND SATO, S. Mind of machine: Toward understanding and measuring emotion in computing systems. Tech. rep., IPSJ SIG Technical Reports, 2005.

[20] ISSA, D., FATIH DEMIRCI, M., AND YAZICI, A. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control 59* (2020), 101894.

[21] KOSAKA, T., AIZAWA, Y., KATO, M., AND NOSE, T. Acoustic model adaptation for emotional speech recognition using twitter-based emotional speech corpus. 1747–1751.

[22] LIVINGSTONE, S. R., AND RUSSO, F. A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE 13*, 5 (2018), e0196391.

[23] LUBIS, N., GOMEZ, R., SAKTI, S., NAKAMURA, K., YOSHINO, K., NAKAMURA, S., AND NAKADAI, K. Construction of Japanese audio-visual emotion database and its application in emotion recognition. In *Proceedings of the Tenth*

International Conference on Language Resources and Evaluation (LREC'16) (Portorož, Slovenia, May 2016), N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., European Language Resources Association (ELRA), pp. 2180–2184.

[24] Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., and Chen, X. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185* (2023).

[25] Mera, K. Hiroshima city university japanese emotional speech corpus (hcudb). `https://research.nii.ac.jp/src/en/HCUDB.html`. Accessed: 2025-04-19.

[26] Mitsuyoshi, S. Research on the phonetic recognition of feeling and a system for emotional physiological brain signal analysis. University of Tokushima.

[27] Mordor Intelligence Source: https://www.mordorintelligence.com/industry-reports/vtuber-market. Virtual youtuber market size share analysis - growth trends forecasts (2025 - 2030). `https://www.mordorintelligence.com/industry-reports/vtuber-market`, 2024.

[28] Mori, H. Utsunomiya university spoken dialogue database for paralinguistic information studies (uudb). Speech Resources Consortium, National Institute of Informatics, 2008. Dataset.

[29] Nose, T., et al. A synthesis method of emotional speech using subspace constraints in prosody. *Acoustical Science and Technology 30*, 4 (2009), 279–287.

[30] Saito, Y., Nishimura, Y., Takamichi, S., Tachibana, K., and Saruwatari, H. Studies: Corpus of japanese empathetic dialogue speech towards friendly voice agent. *arXiv preprint arXiv:2203.14757* (2022).

[31] Santoso, J., Yamada, T., Ishizuka, K., Hashimoto, T., and Makino, S. Speech emotion recognition based on self-attention weight correction for acoustic and text features. *IEEE Access 10* (2022), 115732–115743.

[32] Shinya Mori, Tsuyoshi Moriyama, S. O. Emotional speech synthesis using subspace constraints in prosody.

[33] Shunji Mitsuyoshi, F. R. *人間の感情を測定する*. *IEEJ Journal 125*, 10 (2005). (In Japanese).

[34] Shunji Mitsuyoshi, F. R. Mindo of machine. In マルチメディア通信と分散処理ワークショップ (2005), vol. 2005. (In Japanese).

[35] Song, M., Triantafyllopoulos, A., Yang, Z., Takeuchi, H., Nakamura, T., Kishi, A., Ishizawa, T., Yoshiuchi, K., Jing, X., Karas, V., Zhao, Z., Qian, K., Hu, B., Schuller, B. W., and Yamamoto, Y. Daily mental health monitoring from speech: A real-world japanese dataset and multitask learning analysis. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023), pp. 1–5.

[36] Takamichi, S., Kürzinger, L., Saeki, T., Shiota, S., and Watanabe, S. Jtubespeech: corpus of japanese speech collected from youtube for speech recognition and speaker verification. *arXiv preprint arXiv:2112.09323* (2021).

[37] Takeishi, E., Nose, T., Chiba, Y., and Ito, A. Construction and analysis of phonetically and prosodically balanced emotional speech database. 16–21.

[38] Ullah, R., Asif, M., Shah, W. A., Anjam, F., Ullah, I., Khurshaid, T., Wuttisittikulkij, L., Shah, S., Ali, S. M., and Alibakhshikenari, M. Speech emotion recognition using convolution neural networks and multi-head convolutional transformer. *Sensors 23*, 13 (2023), 6212.

[39] UserLocal Inc. Vtuber live viewership data. https://virtual-youtuber.userlocal.jp/document/ranking. Accessed Jan. 2025.

[40] VRChat Inc. Expression menu and controls - vrchat creator companion. https://creators.vrchat.com/avatars/expression-menu-and-controls/, 2024. Accessed: 2025-04-15.

[41] Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., and Schuller, B. W. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence 45*, 9 (2023), 10745–10759.

[42] Xin, D., Jiang, J., Takamichi, S., Saito, Y., Aizawa, A., and Saruwatari, H. Jvnv: A corpus of japanese emotional speech with verbal content and nonverbal expressions. *IEEE Access 12* (2024), 19752–19764.

[43] Xin, D., Takamichi, S., and Saruwatari, H. Jnv corpus: A corpus of japanese nonverbal vocalizations with diverse phrases and emotions. *Speech Communication 156* (2024), 103004.