Using voice to animate avatar expressions

Aiju Kinoshita,
Ali Neshati, and Steven R. Livingstone
Ontario Tech

VTuber

- Online entertainers who stream videos using expressive virtual avatar.
- Originated in Japan, 433 million hours watched in Q3 2024
- Animated in real time via motion tracking and voice input

Manually controlled expressions

- Virtual avatar’s head and eye motion driven by motion tracking,
- However, facial expressions are manually activated-deactivated via keyboard commands.
- Disconnect between Vtuber and its avatar’s expressiveness.

Automating expressions with SER

- Speech Emotion Recognition (SER) can be used to classify a speaker's emotion.
- Output emotion (e.g., happy, sad, angry) can be used to automatically enable matching avatar expressions.
- Automatic transitions make for smoother, less error-prone experience. Lower acting burdens from VTubers.

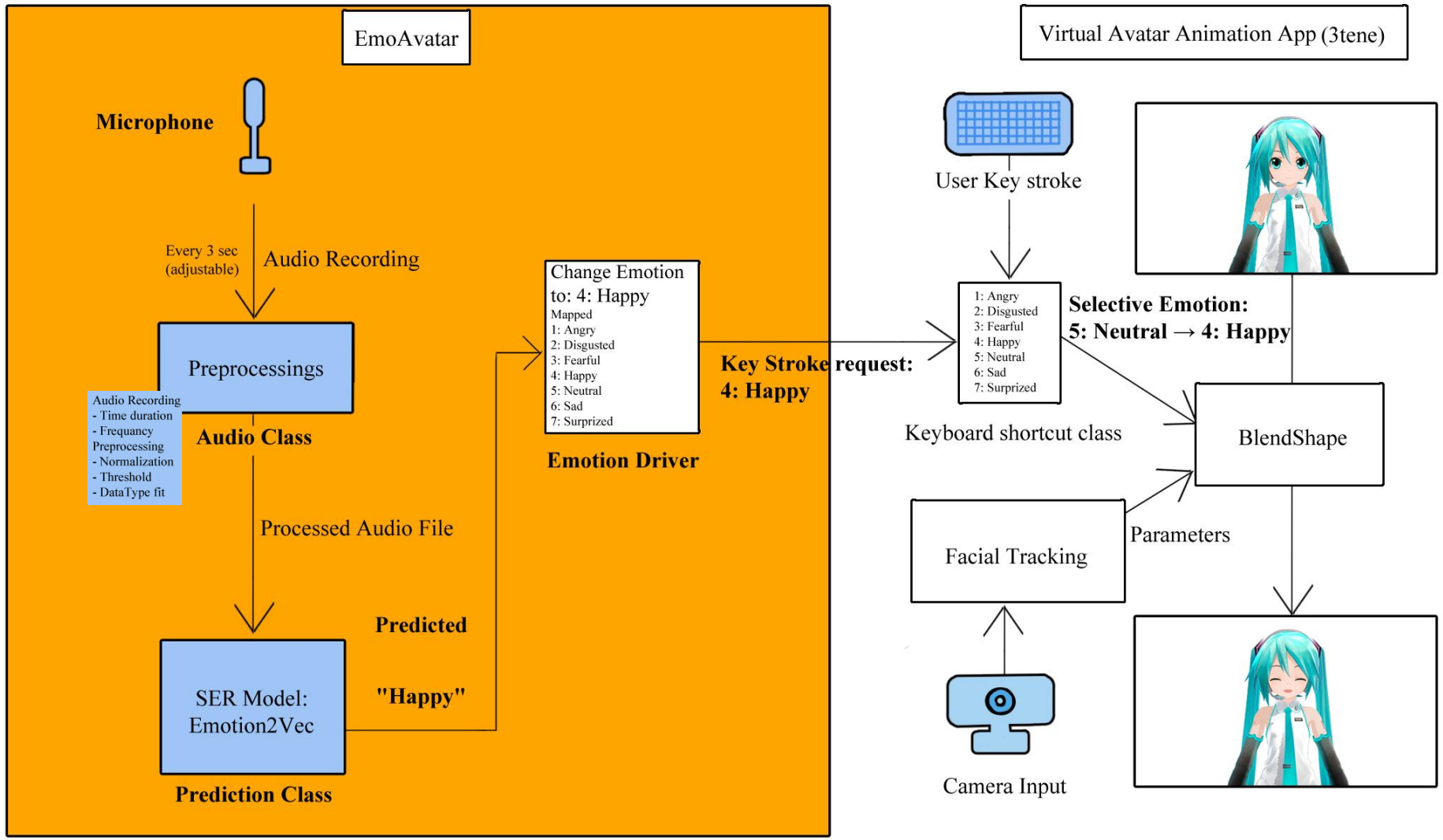
Emotion2Vec and training datasets

Emotion2Vec was chosen for its real-time accuracy, multilingual generalization, and ability to detect nuanced emotions from raw audio without relying on handcrafted features.

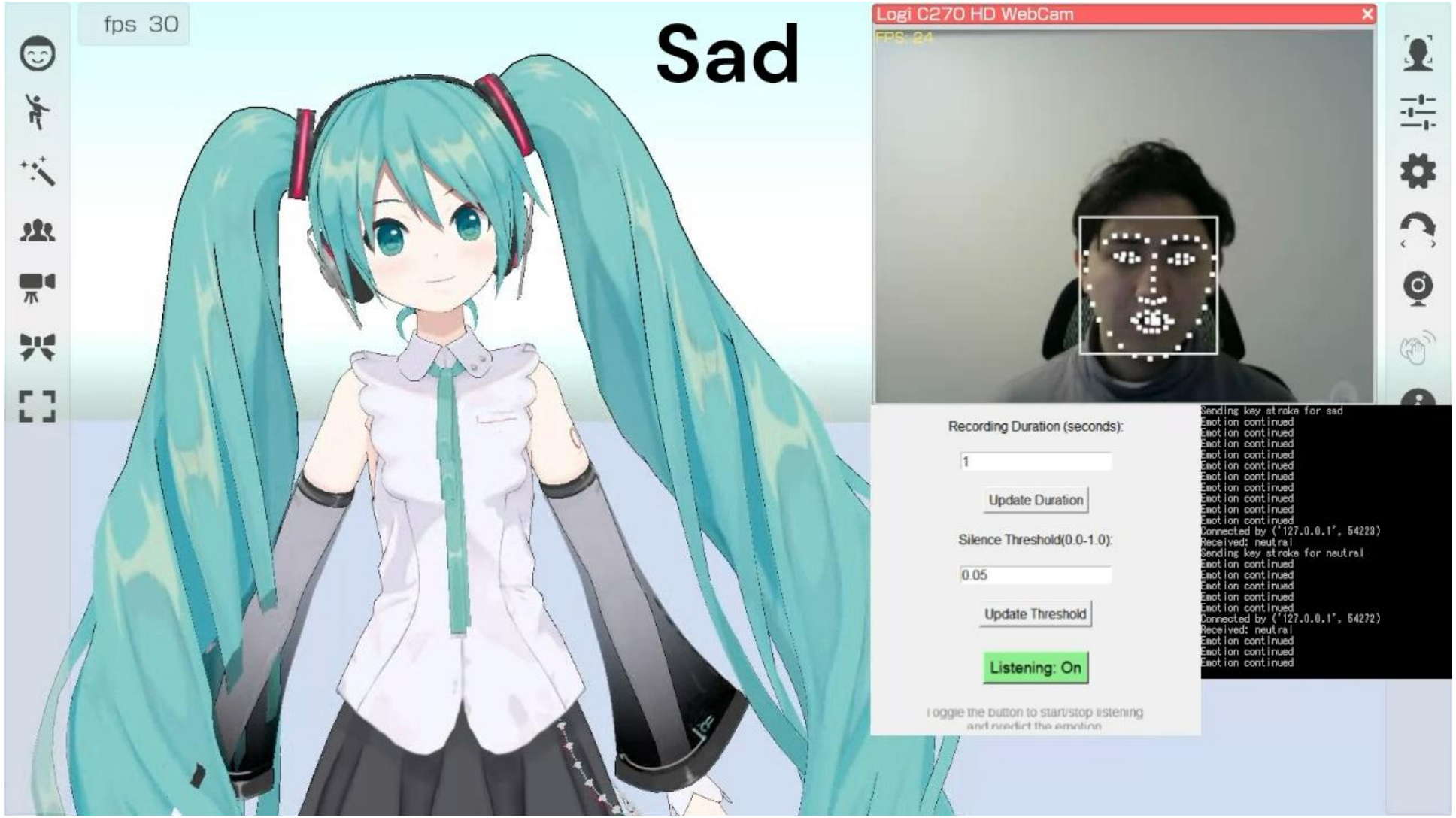
Speech datasets: IEMOCAP, MELD, MEAD, CMU-MOSEI, MSP-Podcast

- Performance on **Japanese speech** remains uncertain — a key focus for our target audience

Method – Model architecture



Results

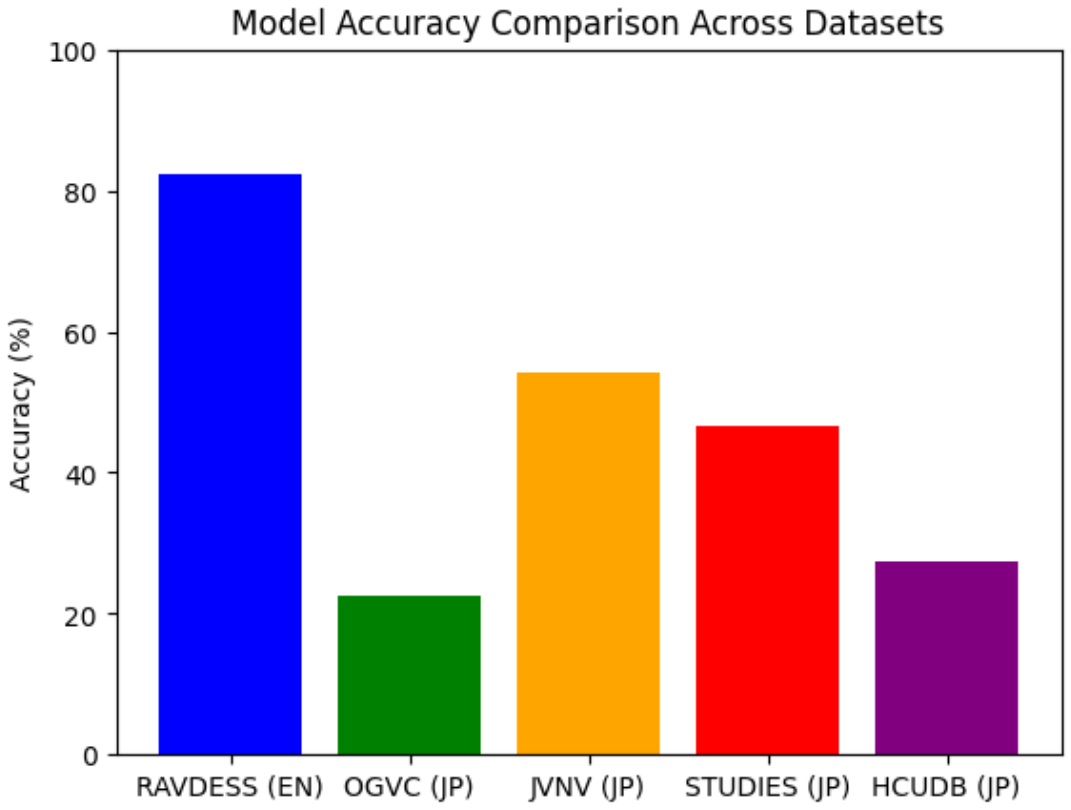

fps 30
Sad
Logi C270 HD WebCam
FPS: 24
Recording Duration (seconds):
1
Update Duration
Silence Threshold(0.0-1.0):
0.05
Update Threshold
Listening: On
Sending key stroke for sad
Emotion continued
Emotion continued
Emotion continued
Emotion continued
Emotion continued
Emotion continued
Emotion continued
Emotion continued
Emotion continued
Emotion continued
Connected by ('127.0.0.1', 54228)
Received: neutral
Sending key stroke for neutral
Emotion continued
Emotion continued
Emotion continued
Emotion continued
Emotion continued
Connected by ('127.0.0.1', 54272)
Received: neutral
Emotion continued
Emotion continued
Emotion continued

Results - benchmarking

Emotion2Vec achieved 82% on English dataset (RAVDESS).

Accuracy dropped to **22~54%** on Japanese speech sets.

Highlights a performance gap due to language mismatch and its phonetic differences.



Discussion

Here we built a Vtuber framework for automating animation of virtual avatars from steamer voice cues.

Good accuracy in English, but not for Japanese speech.

Lag on recording and actual expression.

Future Work:

- Optimization for Japanese speech
- Real-time performance
- Take in consider word context via LLM
- Adaption on other platform (VRChat, 3D animation, AGI, etc…)

Thank you

Special Thanks:

Ali Neshati

Steven Livingstone

Email**: aiju.kinoshita@ontariotechu.net**