Instant Musical Superpowers by augmenting intrinsic movements, expressions and emotions

Zenon Olenski
designerzen Limited, Research Institute, London / Manchester, UK
https://designerzen.com
zenon@designerzen.com

ABSTRACT

In this paper, we introduce and describe *intuitive ways to control machines using parts of the body*, predominantly through gauging intent from facial expressions, reflexive locomotion, and the orientation of the head and body — without requiring any physical touch or interaction. We then present a fully working implementation in the form of a face-controllable musical instrument that demonstrates how this simplified form of input can be highly expressive yet still retain fine control, subtlety, and finesse, whilst being learned more quickly and easily.. We also suggest future methods for more inclusive human-machine interactions, where accessibility, inclusive expression and fidelity are the primary motivators.

1 Introduction

Through repetition our brains associate actions with outcomes, allowing us to learn new skills. This paper and related projects go on to show that if the desired skill can be mapped on natural movements already employed by the body in daily life, then the time to learn this skill is vastly reduced.

To research this theory we have created a complex and expressive musical instrument completely playable with the face and requiring no specialised hardware or equipment, training or skill—inclusively aimed to work for as many people as possible—opening up new musical avenues and empowering everyone with powerful musical tools.

This accessible instrument aims to grant new musical abilities to those unable or unwilling to use existing instruments, by trying to improve human-machine relations to find simpler and more frictionless ways of controlling technology through augmented integration. By piggybacking onto a person's own natural movements using motion-tracking in real-time, we can reduce the time required to learn a complex skill to mere moments whilst still retaining full expression and precise control.

We demonstrate how to operate an interface by mapping simple, obvious, movements onto a complex control system in an intuitive and ergonomic way, and how, by observing subtle emotional cues and facial expressions, we can comprehend intention and mood using a traditional understanding of human psychology, and modify the data and musical output accordingly. Expressions, movements and emotions can now control and manipulate data, software and even hardware.

This virtual interface has no moving parts, requires no specialist hardware and can run on most modern computers and mobile devices with dedicated GPUs.

Using Machine Learning we can create models that observe the human form allowing them to understand basic intentions in near real-time, where the software improves over time as it gets more familiar with and better at understanding your intentions and how you choose to naturally express them.

Using the face as a control mechanism is intrinsic and familiar—so we convert smiles instantly into musical notes, movements into melodies and expressions into audio controls.

As a result of feedback from younger users, the demonstration software (see sec. 3) is often referred to as the, *"Smile Powered Synthesizer"*, due to the way most people discover how it works, however it's full operation mode is explained in sec. 5.

2 Requirements

- From 1 to 4 people
- Video Camera (webcam)
- *Screen / Projector*
- Modern Computer with GPU
- *Input device / touch screen* (optional)

Apart from a modern machine with a powerful GPU, a screen or projector and a webcam are the only hardware requirements; any laptop or modern phone should work fine. *A mouse, keyboard or touch screen interface is required to initiate play but only a face and mouth is necessary once the application has begun.*

3 Demonstration

As a good representation of the flexibility and fidelity of this approach in commanding machines, an "extended-reality" accessible musical synthesizer—completely controllable by the face—has been made publicly available and can be accessed online at https://interface.place. Other public demonstrations are available in the references [4].

Web Audio Conference WAC-2024, March 15–17, 2024, West Lafayette, IN, USA.

4 Technical Implementation

4.1 Capturing face position

This project streams live video from the web camera onto a digital Canvas where algorithms work in real-time to discern landmarks on any detected faces. When a face is identifiable to the machine it will try its best to follow the contours and facial features through time-making assumptions and predictions that connect together positions over frames—providing an understanding of where the user is in relation to the machine at any point in time and what orientation their body is in.

4.2 Converting face position to music

This positional data is analysed and facial markers are located, mapping out the entire face and its features. Triangulation allows us to calculate the shape of key facial characteristics and determine both their proportional sizes and relative distances in order to extract useful information such as mouth shape, head angle, eye positions and whether the eyelids are open or closed. From this data, basic emotions are inferred, and winks, smiles and frowns are calculated in realtime to give us a series of useful metrics, gestures and events.

4.3 Generating Audio

WebAudio provides both sound synthesis and sample playback. This allows for considerable musical range by allowing previously recorded instruments to form the basis of the new sounds—an oboe therefore sounds like an actual oboe would. We also generate sounds using procedural code for the beats, synths and accompaniment; providing more range, configuration and potential expression. The percussion is created using a range of *WebAudio Nodes*, predominantly *Oscillators*, *Buffers filled with noise* and Filters such as *BiQuadFilter* to shape it and a Compressor to give it more depth.

To allow new instruments to be added and swapped out dynamically, as well as being able to add extra effects (similar to a VST chain), the **Web Audio Module 2** format was implemented as a common interface between all internal instruments to allow interoperability and chaining as well as the option to upload new instruments. All sounds are also passed through some extra dynamics to breathe life into the audio and bring unity between the different parts— *Convolvers*, *Compressors* and *ImpulseFilters* work together to provide reverb, an environment and to dampen the silence.[1]

4.4 Recording Audio / Video / MIDI

Various approaches are taken to record the performance resulting in different media types : audio files, videos, pictures and MIDI performances. Audio is pre-mixed through GainNodes then recorded in a background Worker using MediaRecorder and also passed through an AnalyzerNode in order to read the FFT data, resulting in both an audio recording and its associated visual waveform. Video is recorded using captureStream from the Canvas whilst photographs are individual Canvas frames re-encoded via webWorker. Performances can be saved as well as loaded as MIDI files. Once recorded, sonic media can be looped and redubbed forming a basic DAW.

[1]A common audio trick to smooth chaos and soften noises

4.5 Connecting to other Equipment

WebMIDI is used to communicate to external music equipment via the MIDI 1.0 protocol which, once a compatible device has been connected, will sync tempo and send all notes and modifiers as they are being played by the user. If multiple people are playing, each person can control their own MIDI instrument which remains in sync with all others. When run on a server with NodeJS available it is possible to enable a virtual MIDI port so that it can communicate directly with other MIDI software.

5 Operation

This software was designed to be face controllable but as a web security requirement, one physical touch is necessary before the camera is granted permission and for this reason the player selection screen requires mandatory user interaction with a mouse or touch screen. A more in-depth technical implementation is available [7] and the source code is open-sourced [5].

5.1 You are the controller

Once the software has been started, the face is the primary control mechanism for each user. By altering the angle of the head and the openness of the mouth and eyes (see Fig. 1) the player has full control of any digital musical instrument that communicates via MIDI as well as the internal digital synthesizer that has many natural sounding built-in instruments [6].

5.2 The Head

In order to control the key and octave of the sound, the angle of the head is analysed, turning the head into a rudimentary joystick. Tilting and rotating the head left and right (yaw and roll) will select which key on the keyboard will be played while rocking it up and down (pitching) will select which octave those keys will be played from. Each of these mechanisms are interchangeable based on the ability and desires of the performer.

5.3 Mouth

The mouth acts as it would in real life — an amplitude gate where the more open the mouth is, the louder the sound coming out of it — though in this case the sound outputted is replaced with computer generated sound. Closing the mouth stops the sound entirely. This motion is based on the innate gesture of singing and as such tries to mimic it's dynamics including the shape of the mouth.

5.4 Eyes

As a way of adding extra control to the sound once it has been played, the eyes change the stereo panning to the direction where the user is looking; so if the user looks left, the stereo pan plays more on the left channel than the right channel. Looking up and down can be tied into pitch bending but it is quite hard to see this movement without a good camera and lighting. Closed eyes also offer a useful control mechanism and by closing both eyes for one bar, a person can change which instrument they are playing, for example. Closing one eye or another in sequence can trigger special features similar to gestures.

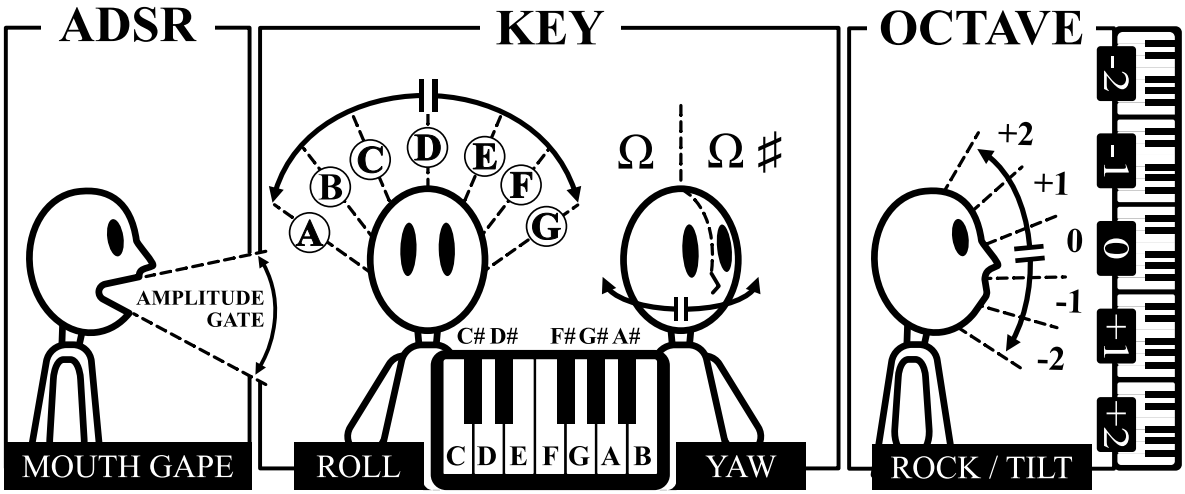

Figure 1: Interface Input Control Mechanism options

5.5 Eyebrows

Both analysed for their orientation and position, the eyebrows tell a story independent of the eyes and add complexity to the sound depending on the instrument. Funnelling your brow can modulate the sound of sadness, while raising your eyebrows can tune the sound to be more joyous.

5.6 Expressions and Emotions

By observing facial landmarks it is possible to recognise certain obvious physical emotions such as happiness, anger, frustration, surprise and depression. This can be used to mould the sound appropriately by setting the scale, choosing the chords, and by adding effects.

5.7 MIDI

Once connected, any MIDI instrument will stay in key and sync with the user's actions and all notes and velocities will get relayed, so that one face can control an entire orchestra or studio. If there are two players and two MIDI instruments, each gets a dedicated channel, otherwise player one broadcasts to all MIDI channels.

6 Conclusions

One of the holy grails in computing is to significantly reduce the complexity of operation, whilst improving the intelligence of a system so that it can guide, rather than be navigated [2].

Already we are seeing software that automatically creates new books, art and music—yet this omits the joy of creation and lacks the hallmarks of intentional expression, emotion and fidelity. That is not to say that computer generated art lacks emotion, but that the process is so automated that it is hard to form a bond over it's creation and lacks significant enough process that it disconnects from that feeling of ownership and originality.

Clearly the next step is feeding machine learned art generators with an expressive, intuitive and enjoyable input mechanism—herein suggested as real-time performance capture. In the same way as the brain adapts to a prosthetic limb which then functions harmoniously with the body, and demonstrated through the "Rubber Hand Illusion" [1], the brain can be quickly convinced to expect any outcome when tied to a repeatable intrinsic movement, as movement is the glue that binds the body with intention [3]. Presumably this is some evolutionary life trait that allows organisms to operate whatever limbs they have, and any limbs they can grow in the future, or in this case, augment.

So in no time at all a new second-nature skill can be learned or an old one regained and this experience is acquired mostly through the act of discovery, experimentation and association—a form of accelerated learning.

By analysing emotion and deducing intention we can offer attenuation and control in a totally automatic, authentic and intuitive way, simplifying operation and harmonising with the user's expectation and desire; in essence allowing the computer to play you as you become the instrument itself.

Accessible technology has to be be easy to access as well as use, and web technologies are the fastest and simplest route—they work on all devices almost instantly—and with computer power increasing every year the requirements will seem humble soon, potentially enabling the addition of more live participants, extra body parts as controls and better facial recognition to maintain tracking in larger groups.

The *"Phantom limb"* is a phenomenon where there is a disconnection of an input mechanism, such as an arm, without the removal of the associated mental expectation, resulting in illusionary sensations and a dissociation with reality. Therefore there are potential remnant side-effects to this approach when used for any long period. In the same way as *'Sea Legs'*, can affect your balance when returning to dry land after being on a boat for a long time, piggybacking any intrinsic movement for a long duration can trick the brain into expectations that can linger and range outside the experience — so *it is important to be mindful of remnants when designing these types of applications.*

6.1 Acknowledgments

Thanks to the **TensorFlow** team at Google for their Machine Learning models, to **Drake Music** Charity for their funding, testing and feedback, the **MIDI Association** and the WAC

for their support. Special acknowledgement to **Audience of the Future**, **Creative Council** and **CreaTech** for promoting the concept in the UK and **SXSW**, **Music Maker Festival**, **DMLab** and *all who have documented the project around the world as it has developed over the years.*

7 References

[1] F. Della Gatta, F. Garbarini, G. Puglisi, A. Leonetti, A. Berti, and P. Borroni. *Decreased motor cortex excitability mirrors own hand disembodiment during the rubber hand illusion*, volume 5. eLife Sciences Publications, Ltd, New York, 2016.

[2] D. C. Engelbart and M. Friedewald. *Augmenting human intellect: A conceptual framework*. Stanford Research Institute, 1962.

[3] M. Jeannerod. *Motor Cognition: What Actions Tell the Self*, volume 42. OUP Oxford, 2006.

[4] Z. Olenski. Give a man a record, they dance for the day, give that man a synthesizer and they dance for a lifetime. https://youtu.be/gTwf6ii6Lak, 2020.

[5] Z. Olenski. Photosynth3 source code. https://github.com/designerzen/InterFACE, 2020.

[6] Z. Olenski. Photosynth3#interface in 90 second. https://youtu.be/-DSDIET5MJo, 2021.

[7] Z. Olenski. Photosynth3#interface. https://zenodo.org/record/6794293, 2022.

APPENDIX

A Installation

A.1 Equipment requirements

- Digital Video Camera or Webcam - high speed preferred over quality
- Screen or a Projector and a wall / screen
- Modern Computer with 3GB HD space and a GPU
- Input devices for initial setup : touch screen, keyboard or mouse

A.2 Installation with Screen

- Ensure the room is well lit
- Mount the screen on the wall in the position that a mirror would take
- Mount the camera as close to the centre top of the screen as possible
- Place stereo speakers either side of the screen or on the floor
- Connect Input devices for initial setup : touch screen, keyboard or mouse
- Connect all cables : PSU, video signal, audio signal and camera
- Goto the homepage and install the PWA to the installation machine
- Be sure to select **Automation** / **Demonstration**, 2-4 player mode and to allow permission for the camera and MIDI port when requested.

A.3 Installation with Projector

As above but replace the screen for a projector. One benefit with this approach is that the camera can be mounted through the screen

- Place short throw projector (preferred) close to the wall / screen or with a long throw projector mount it on the ceiling with enough vertical headroom so as not to hit anybody.