User-defined gestures for audio manipulation and signal processing

Julia Kadie
jkadie@stanford.edu
Stanford University
Palo Alto, California, USA

Rhett Owen
rcowen@stanford.edu
Stanford University
Palo Alto, California, USA

Amber Patrick
amberpat@stanford.edu
Stanford University
Palo Alto, California, USA

Daphne Skiff
bskiff@stanford.edu
Stanford University
Palo Alto, California, USA

ABSTRACT

The current suite of tools allowing the manipulation of audio fall into a limited set of basic controls for users. It would be game-changing to define touch-based gestures for audio manipulation that appeal to a broad set of people. In this project, we elicited gestures for 22 audio effects from 10 participants of varying familiarity with amateur and professional audio tools. The 220 gestures were then documented and analyzed to develop a set of user-defined gestures for audio manipulation including play, pause, echo, and audio source switching. We saw that while many user gestures were influenced by existing audio tools, there was enough confliction among others to identify the emergence of more than one novel gesture.

CCS CONCEPTS

• **Human-centered computing → Gestural input**.

KEYWORDS

Surface, tabletop, gestures, gesture recognition, guessability, signs, referents, audio, signal-processing.

ACM Reference Format:
Julia Kadie, Rhett Owen, Amber Patrick, and Daphne Skiff. 2022. User-defined gestures for audio manipulation and signal processing. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (CS 347).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recently, the standard for interaction-based input into a system has changed from external input devices like computer keyboards and mouses to direct user interaction with a surface. Moreover, researchers in Human Computer Interaction have been exploring gestural input more often due to a rise in interactive tabletops and tablets [6].

In the field of audio-manipulation and signal processing, advanced audio-editing systems and professional tools are only just moving from traditional input like drum pads, mixing boards, and keyboard/synthesizers with knobs or sliders to software like Pro Tools, Ableton, and Logic. Additionally, music-listening is an interesting subspace within gestural vs traditional input because many listening methodologies combine input from hardware, like headphone-buttons, with control input from a computer or smartphone.

Gestural input allows users to synthesize their mental models from across hardware and software into one gesture that combines these types of input. To explore gestural input in the context of audio manipulation, we used a guessability study methodology that presents the effects of gestures to our study subjects and elicits the causes meant to invoke them. We used video analysis, robust notetaking, and think-aloud protocol during this study to acquire our data. Our data showcased user's mental models for the effects we were interested in understanding. Our paper's methodology is inspired by the work of Wobbrock et al, which was the first to employ users in the development of a gesture set.

This work contributes the following to surface computing and audio research: (1) a quantitative and qualitative characterization and taxonomy of user-defined surface gestures for audio manipulation (2) a user-defined gesture set for audio manipulation (3) insight into users' mental models when manipulating audio and (4) insight for the development of global audio gestures without specificity to a particular app. Designers making audio manipulation interfaces can draw upon our research to create better gestures.

2 RELATED WORK

Related prior work includes the systems defining surface-based touch gestures and gesture elicitation to facilitate audio output.

2.1 Systems Defining Surface-based Touch Gestures

The prevalence of gestures reliant on physical interaction between a user and their device can be attributed to the popularity of hardware like tablets, track pads, and smartphones, among others. Users were included in our approach of sourcing gestures through the

CS 347, 2022, Stanford

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

process outlined in Good et al. [3], where we first provided the treated audio clip to exemplify the action and then requested an accompanying sign that would cause that action.

Wobbrock et al. utilized a congruous approach for creating a set of gestures by first showing the visual effect of a gesture and then documenting the causal gesture performed by non-technical users on a Microsoft Surface prototype [6]. This resulted in a taxonomy of gestures to encompass various mental models motivating the gestures. Additional analysis on enjoyment and ease of completion for each gesture helped with the feasibility assessment when applied to a touch interface.

Engeln et al. [2] created a system for direct manipulation of audio parameters using a touch-based gestural system dependent on the gesture set produced by Wobbrock et al. This solidifies the feasibility of touch gestures for audio manipulation, however the use of gestures intended for visual effects makes it unclear if those are the optimal gestures for the task.

2.2 Gesture Elicitation to Facilitate Audio Output

A subset of prior work has directly translated utilized user input into the creation of a gesture set to facilitate audio outputs. Leng et al. established viable freeform hand gestures to navigate musical tasks in the VR space through the use of hand tracking technology [4]. The outcome was a set of gestures that included repetition for different tasks, resulting in a smaller result than expected. Additionally, the set included gestures directly translated from controls of conventional models.

Drossos et al. incorporated gestural components into a professional micro-controller board to enable sound mixing [1]. Rather than ignoring recognition to allow for any gesture like Wobbrock et al. and Leng et al., emphasis was placed on each gesture's spatial positioning and proximity to conventional orchestral conducting to produce the desired mixing effect.

Rahman et al. demonstrated the efficacy of a gesture-controlled audio system within modern vehicles using pre-established gesture controls [5]. Similar to Drossos et al., a specific mental model inherently influenced gestures through the inclusion of live audio and haptic feedback.

While the aforementioned systems collected user generated gestures within a set with the intended application to technological controls, the unique combination of touch-based surface gestures as applied to audio manipulation was left unaddressed.

3 DEVELOPING A USER-DEFINED GESTURE SET

This section describes our approach to developing a user-defined gesture set. This holds basis in prior work [6].

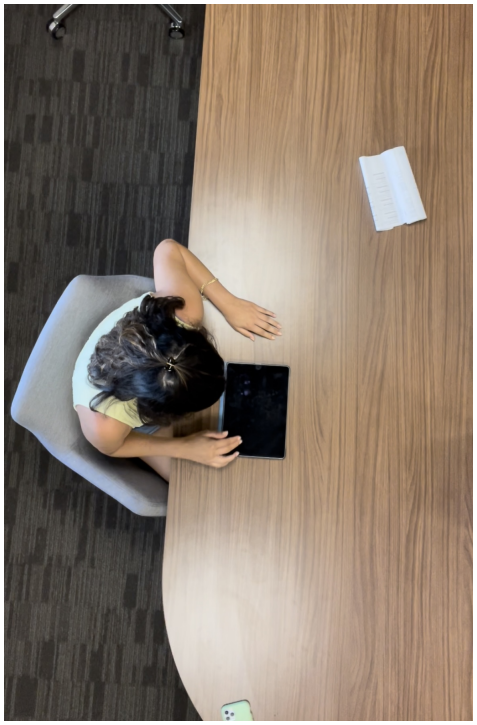

Figure 1: Our experiment setup.

3.1 Overview and Rationale

Each participant first heard an untreated control audio clip and then heard a second clip, which was the audio after the desired audio effect was applied.

3.2 Referents and Signs

Each participant saw the effect of a gesture and was asked to perform the gesture they thought would cause that effect. In linguistic terms, the effect of a gesture is the *referent* to which the gestural *sign* refers. We developed a user-defined gesture set based on the *agreement* that participants exhibited in their gestures. *Agreement* increases as more participants use the gesture for creating a certain audio effect.

3.3 Participants

3.3.1 Demographics. 10 participants volunteered for the study. 7 were women. All participants were Stanford students. All participants were right handed.

3.3.2 Prior audio experience. 9 participants were amateur music editors and 1 participant was an expert music editor who was exposed to common and professional audio-emitting platforms and devices. All participants used Spotify as their primary music-listening platform. All but one participant listened to music using wireless headphones. All participants typically controlled song choice, volume, etc with a smartphone as opposed to a computer.

3.4 Apparatus

The study was conducted on an Apple iPad tablet measuring 9.74" x 7.02" (although 3 of the participants completed the study on a table without a tablet due to Covid-19). We made a 22 different audio clips to present the before and after of an audio effect. The main song used for the display of audio effects was "Fly Me to the Moon" by Frank Sinatra. For example, for the *reverb* referent, the audio recording played for the participant played 5 seconds of "fly me to the moon" unedited, a 1-second pause, and then the same 5 seconds of "Fly Me to the Moon" with a reverb effect applied.
A video camera/iPhone positioned above the user recorded hand gestures, shown in Figure 1. In addition, 1-4 authors observed each session and took detailed notes.

3.5 Procedure

We presented all participants with referents in the same pre-defined order. We informed the participants that they could interact with the tablet in any way with however many hands and fingers they wanted to. For each referent, participants performed a gesture. After each gesture, participants rated their gesture on three 7-point Likert scales concerning quality of fit to the effect, ease of performance and satisfaction with touch interface. An exit survey was given to the participants asking them to discuss their prior experience with music-listening and music-editing. This survey asked about streaming platforms used, devices used for audio, software used for editing audio, and devices used for controlling audio.

4 RESULTS

Our results include a gesture taxonomy, a user-defined set of gestures for audio, and observations about participant mental models.

4.1 Classification of Gestures for Audio

Based on the physical features of each performed gesture and each participant's underlying mental model, we have constructed a four-category taxonomy of gestures for audio (Table 1). Our taxonomy builds off of the classification proposed by Wobbrock [6], but with a few key differences:

- Wobbrock's taxonomy includes a *Physical* category in the *Nature* dimension to describe gestures that would interact with physical objects on the table-top surface the same as they would virtually on the touch screen. Our experimental design did not include visually-displayed objects on the touch-sensing tablet, and by nature, our study aimed to identify gestures for audiological interactions with no implied visual analogue. Thus, a physical category is not present.
- None of the 220 observed gestures could be described as *Symbolic*, meaning no gestures resembled a symbol relevant to the intended audio effect. Thus, it was not included.
- A very large proportion of performed gestures were identified to be motivated by a similar category of mental model: *Spatially-Mapped*. This refers to gestures that map a feature of audio to a particular dimension of physical space. Examples are dragging vertically to increase or decrease volume, or horizontally to scrub to different points in a song or audio clip. A category was appended to the *Nature* dimension to account for this observation.
- No observed gestures had *Mixed-dependencies* in *Binding*, and thus no category was specified.

Taxonomy of Gestures for Audio

Form	Static pose	Hand pose is held in one location
	Dynamic pose	Hand pose changes in one location
	Static pose and path	Hand pose is held as hand moves
	Dynamic pose and path	Hand pose changes as hand moves
	One-point touch	Static pose with one finger
	One-point path	Static pose & path with one finger
Nature	Spatially-mapped	Gesture assumes a spatial mapping of an audio feature
	Metaphorical	Gesture indicates a metaphor
	Abstract	Gesture-referent mapping is arbitrary
Binding	Object-centric	Location defined w.r.t. imagined object features
	World-dependent	Location defined w.r.t. world features
	World-independent	Location can ignore world features
Flow	Discrete	Response occurs after the user acts
	Continuous	Response occurs while the user acts

Table 1: Taxonomy of all observed gestures for audio (220 total).

With our taxonomy of gestures for audio as reference, we then took all 220 participant-produced gestures and classified each in all four dimensions based on salient physical features and participant-shared mental models. Figure 3 shows a complete breakdown of each dimension by category, which reveals important trends in gestures for audio. We can see that one-point gestures (performed with only one finger) comprised the majority of performed gestures, though 42% of gestures were performed with two or more fingers—an increase from the 37% observed by Wobbrock [6]. The *Spatially-mapped* category dwarfed the other subcategories within the *Nature* category, which reveals a strong trend in participant mental models away from arbitrary mappings and gestures meant to evoke other physically-intractable systems or phenomena, and towards gestures that imagine audio features in spatial dimensions.

4.2 A User-defined Gesture Set

The central motivation of our research is the proposal of a clearly-defined set of gestures for audio manipulation and signal processing.

4.2.1 Agreement. To gain a sense of the degree of consensus among participants based on each proposed audio effect, we calculated agreement scores to understand where participants converged in mental models and where they diverged. For each audio effect, the *agreement score A* was calculated as:

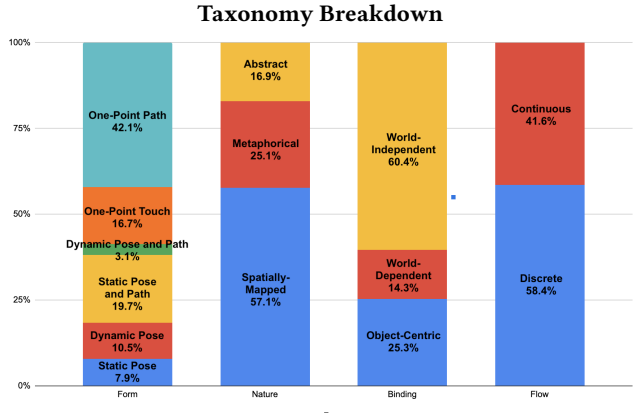

Figure 2: Percentage of observed gestures (220 total) in each taxonomy category.

$$A = \sum_{P_i \subseteq P_r} \left(\frac{P_i}{P_r} \right)^2$$

where P_r is the set of proposed gestures for referent r, and P_i is the subset of identical gestures within P_r. These scores helped us understand which audio effects should be assigned corresponding gestures first. We prioritized the most ubiquitous assignments over the high-conflict ones since effects with higher agreement tended to be simpler effects more commonly used by amateur users (e.g. *Increase/Decrease Volume*), rather than in more complicated audio editing programs (e.g. *Echo*).

4.2.2 Properties. All 22 audio effects were assigned at least one gesture. Eighteen of these referents are either *dichotomous* or *symmetric* gestures. The two *dichotomous*, *Play* and *Pause*, are given the same gesture, though they have directly opposite effects depending on the current context. The sixteen *symmetric* referents are coupled with one another such that reversing the gesture in direction or location performs the opposite effect. In our study, all effects were presented separately to ensure that participants would have consistent symmetric mental models for these referents. This meant that for the large majority of participants, if they chose to drag upwards to increase volume, they correspondingly chose to drag downwards to decrease volume rather than picking a new mental model to represent volume altogether. Thus, many of the gestures are presented together to make their symmetric relationship more clear.

4.3 Mental Model Observations

In compiling our set of gestures, we came to more intimately understand certain trends in mental models underlying each produced gesture.

4.3.1 Choice of Axis. Certain *Symmetric* pairs of referents were overwhelmingly assigned *Spatially-Mapped* gestures as expected, but we were interested to see emerging trends in which axis certain audio effects were mapped. For example, a clear distinction between *Speed Up/Down* and *Pitch Up/Down* quickly became clear when we observed a majority participants using the horizontal axis to represent speed, and the vertical axis to represent pitch. Though either assignment of axis to referent effect would produce gestures of equal ease of performance, the clear preference of mapping perhaps points to subconscious reinforcement of representing time (and consequently, speed) horizontally, and pitch vertically (e.g. in Western music notation). The fact that both pairs of referents use "up" and "down" as inverse abstract directions points to the fact that this spatial mapping isn't linguistically influenced, but rather symbolically in our socially-distributed tools and notation for audio.

4.3.2 Non-Touch Gestures. Though participants were encouraged to perform gestures that could be detected and understood by the touch-sensing apparatus, a few strong mental models for certain referents elicited gestures that could only be detected by a device's accelerometer—in particular, two gestures defied our constraints. Participant 4, when prompted to produce a gesture for the *Switch Audio Source* effect, chose to pick up the touch device and physically tilt the display in the direction of the desired playback device. While this gesture wasn't common enough to be included in this set, it did reinforce the strong agreement in the dominant mental model of making directional reference to a location in the participant's physical context. The chosen gesture for this referent uses the same mental model. The second gesture that defied our constraints was produced for the *Shuffle* effect—a surprising number of participants took the touch device and physically shook it in their hands to match the referent. Such a large number of participants shared the same idea and mental model that we chose to include it in the final gesture set, alongside a the second-most agreed gesture which was touch-detectable. These observations help us understand that participants often conceptualized audio in three dimensions, beyond the two dimensions accommodated by the provided touch apparatus, as well as in a more tactile way than simple axes and visual metaphors allow.

4.3.3 Audio and Continuity. Among referents that weren't inherently discrete effects (e.g. *Play/Pause, Looping*), participants overwhelmingly produced gestures belonging to the *Continuous* category of the *Binding* dimension. This means that these gestures produced the desired effect in real-time, updating in proportion to the extent of the gesture rather than increasing/decreasing a property in discrete chunks. This trend amongst continuous referents helps us understand how participants may have a latent preference for finer-resolution controls that update in real-time for audio in particular. We can hypothesize that this trend is due to the essentially time-dependent nature of audio that correspondingly makes us prefer time-dependent effect rather than discrete ones. For strict visual referents, we've seen how discrete mental models are more prevalent [6], so we kept this context-specific trend of continuity in mind when constructing our final set of gestures for audio.

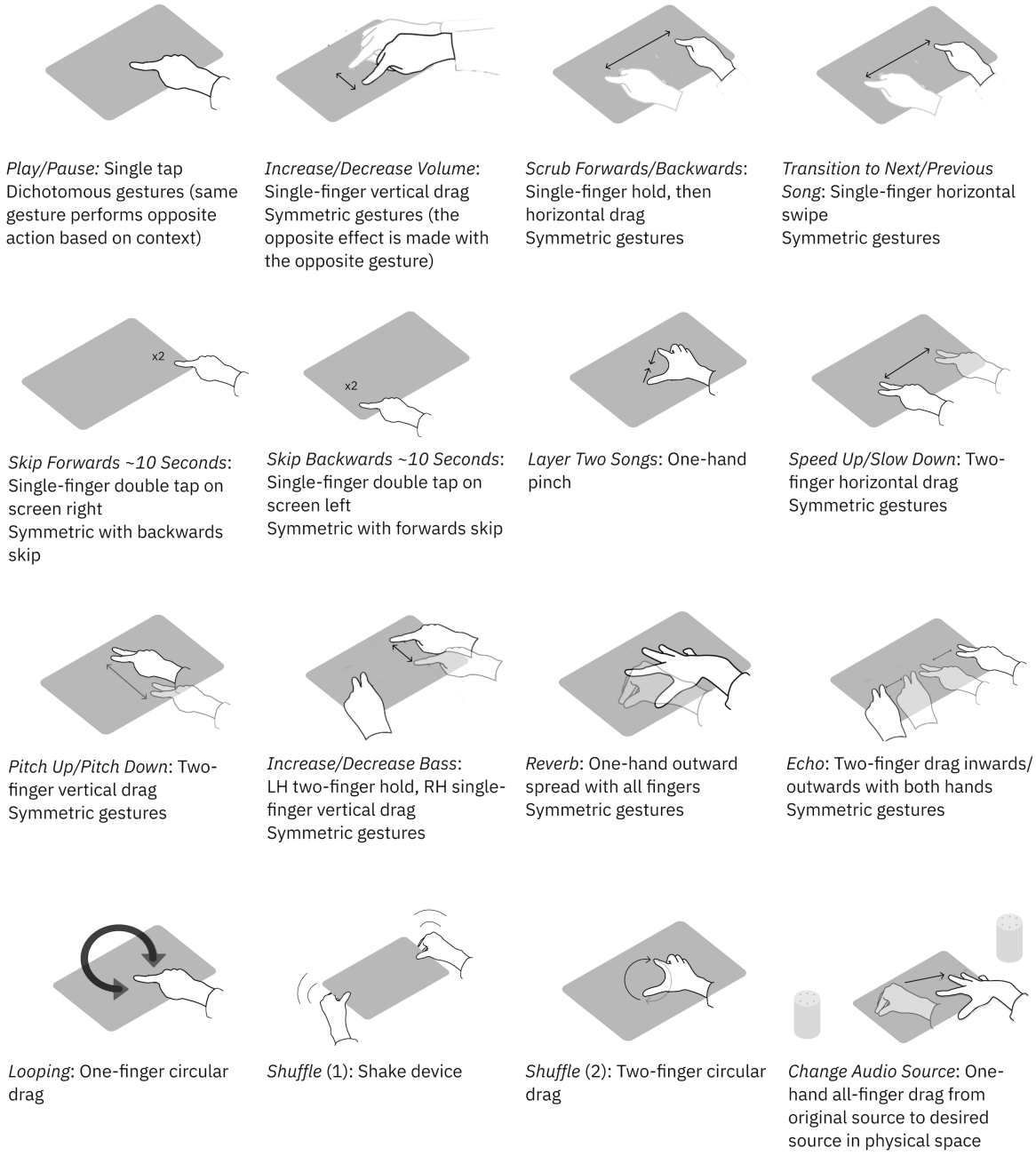

Figure 3: The user-defined gesture set. Hand figure of lower opacity shows previous placement of hand in a single gesture, rather than a second hand needed to perform the gesture.

5 DISCUSSION

5.1 Survey Responses

While agreement scores were the primary measure in determining a final gesture set, the aggregated survey responses gave us insight into the decision making process behind users' gesture choices. Figure 4 shows their average responses to each survey question, as well as a scaled version of the agreement scores.

Many of the expected trends emerged in our analysis of these results. There was a general downward trend in all three survey categories as the audio tasks increased in complexity. In most cases,

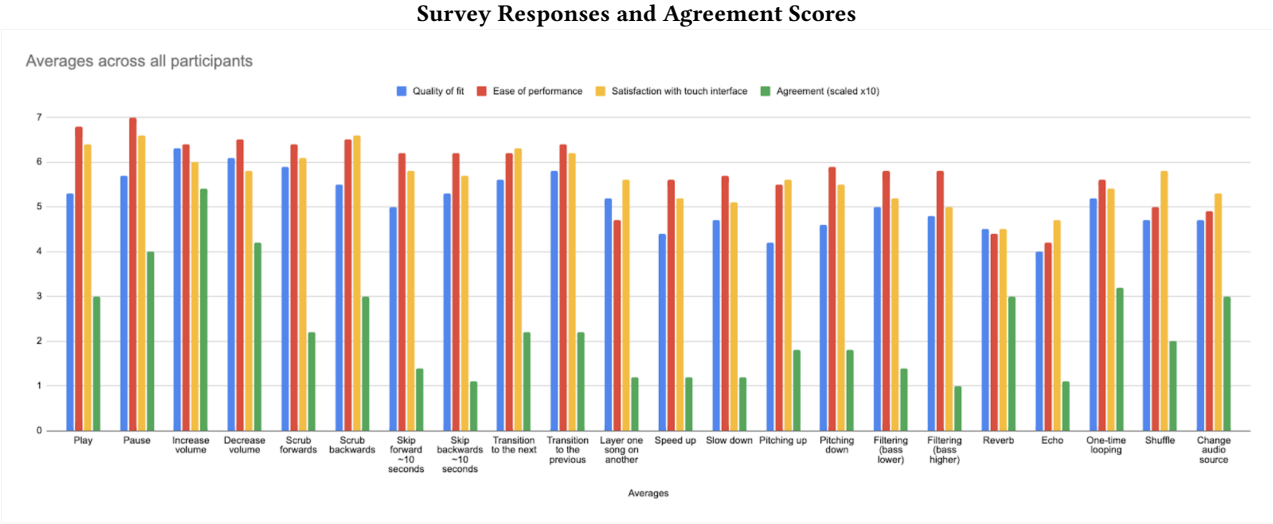

Figure 4: Quality of fit, ease of performance, and satisfaction with touch interface plotted alongside scaled agreement scores.

there was also a decrease in agreement scores as the audio tasks became more complex. However, there were some notable cases in which the agreement scores and survey responses did not correlate with one another. For example, *Skip Forwards/Backwards 10 Seconds* had high survey responses, but had low agreement scores. This likely indicates that the audio task slots into varying mental models in different ways, and different users were able to produce gestures that were easy to perform and fit the task well, but differed widely from one another. We also observed the opposite pattern with effects like *Reverb*. With this task, users had low survey responses, but had high agreement scores. We took this to mean that users have produced a novel gesture, and while it fits well with the given task, users remain unfamiliar with it due to the lack of widespread usage in existing platforms.

5.2 Unexpected Trends in User Gestures

As discussed in our results section, there were many peculiarities in the gestures proposed by the participants in this study. While some of this variation can be explained in the context of varying mental models, it is worth addressing some of the themes that came up in our participants' gestures and the possible reasoning behind them.

5.2.1 High Self-Confliction. As per the authors' expectations, the majority of participants performed gestures where each audio task was represented by a distinct gesture. For application in real-world technology, this would be the intended result. However, many of the participants provided a set of gestures where one or more of their gestures were overloaded. For example, participant 1 performed a particularly narrow range of gestures for the presented audio effects, reusing gestures such as a single finger drag for multiple different effects. While there was no formal follow-up interview discussing why the participant chose to reuse certain gestures, we can use context from the recordings as well as their survey responses in order to hypothesize why these gestures were repeated. One explanation could be that the user decided that the gesture would be more fitting to later effects, and decided to reuse gestures in order to reflect this belief. However, participant 1 actually gave lower quality of fit scores to the tasks for which the one finger drag was reused, which would counter this argument. Another explanation could be confusion about the audio effect being performed. If users inquired about the particular effect being performed, they would be told exactly what the effect was. However, this information was not provided beforehand, and this could have led to confusion that resulted in users reusing gestures.

5.2.2 Low Quality of Fit. Some of the trends in reported quality of fit scores fell in line with the predictions of the authors – in general, more complicated tasks resulted in gestures with lower quality of fit scores. However, some of the quality of fit scores were much lower than predicted, with participant 7 giving quality of fit scores as low as 1. Given that the participants had no constraints on their gestures, it is a surprising discovery that users would perform gestures with such a bad quality of fit. One possible explanation would be that the task being performed is extremely complicated and difficult to fit to a gesture. However, the two tasks that participant 7 ranked as a 1 for quality of fit were actually part of the less technical group of tasks. From analyzing the footage of these gestures, these low quality of fit scores are likely due to a factor that we did not measure in this experiment – the fun of the gestures. Given that participant 7's gestures were quite complicated, it is likely that they valued novelty and uniqueness in a gesture over quality of fit to the effect.

5.3 Limitations and Next Steps

Taking the aforementioned anomalies in user behavior into consideration, we can address many limitations and future steps for this research project. To avoid confusion among users about what is occurring in a particular audio clip, future implementations of this kind of research could explicitly inform users about the task being performed before they are asked to produce a gesture. In addition, it may be worthwhile to walk them through the entire set of tasks before they are asked to produce gestures. By doing this, we avoid the possibility of a user finding a 'better fit' for one of

their previously performed gestures at a later point in the study.

Another future direction of this study would be to test users' reception of this gesture set when applied in an audio application. In order to do this kind of research, a touch-based audio manipulation app would need to be created that implements the gesture set proposed by this paper. Users could be prompted to perform different tasks with the app and then report their satisfaction with the interface. As a control group, a version of the app could also be tested that implements a standard set of gestures such as those provided by Spotify or Apple Music.

5.4 Applications

The central aim of the construction of this gesture set is its general applicability to a diverse range of audio systems. Nine of ten participants noted that streaming services were the primary audio manipulation platforms they used on a regular basis, with only one expert participant that worked regularly with a digital audio workstation (DAW). Despite the skew of our participants towards amateur experience in audio manipulation, the gestures comprising this set have very realistic potential for use in more advanced signal processing and audio editing. Our proposed gesture for *Increase/Decrease Bass*, for instance, can see extended use as a filtering gesture to modify a larger range of frequencies depending on the global location of anchored hand. *Pitch Up/Down, Reverb*, and *Echo* are all referents that only see use most commonly in DAWs, and thus are well-suited for that environment. There is also a strong potential application for the use of these gestures in a live music performance setting, especially among *continuous* gestures that update the referent audio effect in time and proportion to the performed gesture, playing to the strengths of this set's expressiveness and intuitiveness.

6 CONCLUSION

In this paper, we developed a set of intuitive surface-based touch gestures for audio manipulation non-specific to a particular app. We explored over 150 different unique gestures performed in response to 22 different audio prompts which ranged from simple audio tasks to complex signal processing. In addition to presenting a gesture set, we have also developed a taxonomy similar to that of Wobbrock et al. to help capture the general categories and varying mental models behind gestures for audio-based tasks. Our proposed gesture set is a strong candidate for usage in any kind of surface-based environment such as on tablets, smartphones, and track-pads. This paper not only identifies commonly used surface-based gestures that could be applied to audio manipulation, but also proposes novel gestures which have strong agreement among participants and show promise for real-world use.

7 REFERENCES

[1] Drossos, K., Floros, A., Koukoudis, K. (2013). Gestural user interface for audio multitrack real-time stereo mixing. *Proceedings of the 8th Audio Mostly Conference on - AM '13.* https://doi.org/10.1145/2544114.2544123.

[2] Engeln, L., Kammer, D., Brandt, L., Groh, R. (2018). Multitouch enhanced visual audio-morphing. *NIME*, 152–155.

[3] Good, M. D., Whiteside, J. A., Wixon, D. R., Jones, S. J. (1984). Building a user-derived interface. *Communications of the ACM*, 27(10), 1032–1043. https://doi.org/10.1145/358274.358284.

[4] Leng, H. Y., Norowi, N. M., Jantan, A. H. (2017). A user-defined gesture set for music interaction in immersive virtual environment. *Proceedings of the 3rd International Conference on Human-Computer Interaction and User Experience in Indonesia.* https://doi.org/10.1145/3077343.3077348.

[5] Rahman, A. S. M. M., Saboune, J., El Saddik, A. (2011). Motion-path based in car gesture control of the multimedia devices. *Proceedings of the First ACM International Symposium on Design and Analysis of Intelligent Vehicular Networks and Applications - DIVANet '11.* https://doi.org/10.1145/2069000.2069013.

[6] Wobbrock, J. O., Morris, M. R., Wilson, A. D. (2009). User-defined gestures for surface computing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/1518701.1518866.

8 APPENDIX

A EXIT SURVEY GIVEN TO PARTICIPANTS

Questions:

(1) What software do you mostly use for streaming music? [Select one]
  (a) Youtube
  (b) Spotify
  (c) Apple Music
  (d) Other

(2) When listening to music, do you typically listen... [Select one]
  (a) Out loud/on a speaker
  (b) On airpods or other wireless headphones
  (c) On wired headphones

(3) When listening to music, do you typically control song choice, volume, etc with... [Select one]
  (a) A computer
  (b) A smartphone

(4) Do you have experience with any of the following music editing hardware? [Select all that apply]
  (a) Keyboard/Synthesizer with knobs or sliders
  (b) Drum Pad
  (c) Mixing Board
  (d) None of these

(5) Do you have experience with any of the following music editing software? [Select all that apply]
  (a) GarageBand
  (b) Ableton
  (c) Pro Tools
  (d) Logic
  (e) None of these