InChorus: Designing Consistent Multimodal Interactions for Data Visualization on Tablet Devices

Arjun Srinivasan[1,2] **Bongshin Lee**[1] **Nathalie Henry Riche**[1] **Steven M. Drucker**[1] **Ken Hinckley**[1]

[1]Microsoft Research
Redmond, WA
{bongshin, nath, sdrucker, kenh}@microsoft.com

[2]Georgia Institute of Technology
Atlanta, GA
arjun010@gatech.edu

ABSTRACT

While tablet devices are a promising platform for data visualization, supporting consistent interactions across different types of visualizations on tablets remains an open challenge. In this paper, we present multimodal interactions that function consistently across different visualizations, supporting common operations during visual data analysis. By considering standard interface elements (e.g., axes, marks) and grounding our design in a set of core concepts including operations, parameters, targets, and instruments, we systematically develop interactions applicable to different visualization types. To exemplify how the proposed interactions collectively facilitate data exploration, we employ them in a tablet-based system, *InChorus* that supports pen, touch, and speech input. Based on a study with 12 participants performing replication and fact-checking tasks with InChorus, we discuss how participants adapted to using multimodal input and highlight considerations for future multimodal visualization systems.

Author Keywords

Multimodal interaction; data visualization; tablet devices; pen; touch; speech.

CCS Concepts

•Human-centered computing → Visualization; Human computer interaction (HCI);

INTRODUCTION

Recent advancements in screen resolution and computing capabilities have made tablet devices a promising platform for data visualization. With this potential in mind, numerous research projects have been investigating visualization tools on tablets (e.g., [2, 9, 19, 20, 35–38]), facilitating interaction through touch and/or pen input. Although these carefully designed systems highlight the potential of tablets along with key design considerations, two fundamental issues persist.

First, the majority of prior research about data visualization on tablets [2, 9, 19, 35, 36] have focused on a single visualization type, optimizing interactions for that chart type. This local optimization could result in a *globally inconsistent interaction experience* when multiple types of visualizations are included as part of one system. For example, prior systems have used the gesture of dragging a finger along the axis to sort a bar chart [9] and select points in a scatterplot [36]. However, when both bar charts and scatterplots are supported by the same system, the gesture of dragging along an axis causes a conflict, resulting in inconsistent functionality across visualizations [37]. Resolving such inconsistencies often requires system designers to introduce specialized gestures such as holding on the axis of a bar chart to enter a transient "sort mode" in which one can swipe to sort [37]. Such subtle differences in gestures can be difficult to remember and may lead to errors while performing tasks, however.

Second, when depending only on pen and/or touch, systems face *increased reliance on menus and widgets* as the number and complexity of operations grow. For example, to filter, users have to select visual marks and tap a delete/keep-only button or adjust sliders and dropdown menus in control panels [20, 36, 37]. Such indirect interactions with interface elements external to the objects of interest (e.g., marks) can divert the users' attention which may prove disruptive to their workflow [9]. Additionally, given the space constraints of tablets, control panels can occlude the visualization and limit the screen space available for the visualization itself.

To address these issues, we propose using multimodal interactions where the directness and precision of pen and touch is complemented by the freedom of expression afforded by speech. For example, we let users perform zoom & pan through familiar pinch and drag touch gestures. On the other hand, given its affordance for drawing free-form strokes, we offload actions like drawing a selection lasso to the pen. Furthermore, considering its expressiveness, we use speech to support operations like filtering where one can filter selected points (e.g., "*Remove these*" or "*Exclude others*") or filter points satisfying specific criteria (e.g., "*Remove all movies except action, adventure, and comedy*").

To design multimodal interactions that function consistently across different types of visualizations, we first surveyed 18 visualization systems to identify common operations and interactions supported in current systems. Then, by considering standard elements in visualization tools (e.g., axes, marks) and grounding our design in a set of core concepts including operations, parameters, targets, and instruments, we systematically develop interactions applicable to different visualization types. To illustrate how the proposed interactions collectively facili-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '20, April 25–30, 2020, Honolulu, HI, USA.
© 2020 Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376782

tate visual data exploration, we employ them in a tablet-based system, *InChorus*. We also leverage InChorus to conduct a user study to assess the practical viability of the proposed interactions and observe if participants adapt to using multimodal input during common visual analysis tasks. Based on a study with 12 participants, we found that participants successfully adapted to using the proposed interactions to complete a series of replication and fact-checking tasks, commenting favorably on the freedom of expression provided by multiple modalities. Reflecting on our experience of designing interactions in InChorus and observations from the user study, we highlight promising research directions for future work on multimodal visualization systems.

In summary, we make the following contributions:

- We present systematically designed multimodal interactions that function consistently across different types of visualizations, supporting core visual data analysis operations.
- Through a prototype system, InChorus, we exemplify how pen, touch, and speech-based multimodal interactions can collectively facilitate visual data exploration on tablets.
- We report findings from a user study, highlighting how multimodal input accommodates varying user interaction patterns and preferences during visual analysis.

RELATED WORK

Pen and Touch-based Visualization Systems

A plethora of systems have investigated the use of touch and/or pen input for interacting with data visualization systems (e.g., [2, 6, 9, 19, 20, 25, 26, 34–36, 39, 46, 54]), examining different devices and form-factors including tablets (e.g., [2, 9, 20, 36]), tabletops (e.g., [11, 18]), and large vertical displays (e.g., [25, 26, 54]), among others.

Although we consider several of these systems when designing multimodal interactions, most relevant to our work are prior systems that are designed for tablets [2, 9, 20, 35–37]. For instance, with TouchWave [2], Baur et al. presented a set of multi-touch gestures to interact with hierarchical stacked graphs on tablets, specifically noting that designing a consistent interaction set was one of the primary challenges they faced. Drucker et al. [9] compared a gesture-based interface to a WIMP-based interface in the context of bar charts. Their results showed that not only were people faster and more accurate with gestures, but also that people subjectively preferred direct, gestural interactions over interacting with controls in the WIMP interface. Sadana and Stasko [36] presented multi-touch interactions for common operations including selections, zooming, and filtering, among others in the context of scatterplots. Following up their work on scatterplots, Sadana and Stasko expanded their system to include other types of visualizations and support multiple coordinated views [37]. However, upon including additional visualizations, they encountered challenges due to inconsistencies in interactions across visualizations, calling for future systems to leverage standard gestures and support consistent interactions and feedback [37]. While these systems focus on touch-only input, with TouchPivot [20], Jo et al. illustrated how pen and touch input can complement WIMP-style interface elements to help novices conduct visual data exploration on tablets.

Our work is motivated by these examples from prior work and addresses a common challenge faced by these systems: inconsistency in interactions for different operations and visualization types. Specifically, we systematically analyze the operations and interactions supported by these systems to identify our target operations and initial set of interactions. Furthermore, while most current systems are optimized for a single type of visualization, to ensure consistency and general applicability of the developed interactions, we design and test our interactions in the context of five popular visualization types as part of the same system.

Speech-based Multimodal Visualization Systems

Recently, there has been an influx of natural language interfaces (NLIs) for data visualizations [1, 8, 13, 17, 21, 40, 43, 45, 53]. Although these systems focus on NL as their primary mode of interaction, they acknowledge the need for multimodal interaction to support limitations of NL such as ambiguity [13, 40]. Furthermore, a majority of current NLIs explore the use of typed NL input in a desktop-setting. However, typing is not an efficient input technique on interactive displays such as tablets where speech becomes a more natural form of input. While the interpretation strategies may be comparable to typed NLIs, speech-based systems require different interface and interaction design considerations due to added complexity with potential speech-to-text recognition errors and the lack of assistive features such as auto-complete.

With this distinction between typed and spoken input in mind, two systems that are most related to our work are Orko [43] and Valetto [21]. Orko [43] supports multimodal touch- and speech-based interaction with node-link diagrams, facilitating common network visualization tasks including finding connections, computing paths, and attribute-based filtering of nodes, among others. Valetto [21] presents a conversational interface for people to query for visualizations using voice on tablets. Once a visualization is created, Valetto allows performing a rotate gesture to flip X/Y axes and swipe to change the visualization type. Although these systems support touch and speech input, they either focus on a specific visualization type [43] or conversational interaction [21], providing little insight into how the multimodal interactions were designed or how future systems can build upon the presented interactions.

In our work, we design pen, touch, and speech-based multimodal interactions for frequent visual analysis operations [15] enabling visual data exploration on tablets. Furthermore, by leveraging common interface elements (e.g., axes, marks) and grounding our design in a set of core concepts applicable to most visualization systems, we also exemplify how future systems can build upon our work and systematically design and describe multimodal interactions with visualizations.

SYSTEMATIC DESIGN OF MULTIMODAL INTERACTIONS FOR DATA VISUALIZATION TOOLS

We surveyed 18 visualization systems [2, 6, 9, 11, 19–21, 26, 35–39, 43, 44, 46, 49, 54] to identify tasks and visualizations to consider as part of our design. To explore a broader design space

	Operation	Parameters (P)	Targets (T)	Instruments (DI or II)	Sample Keywords (K)	Interaction Patterns	Examples	Modalities T	P	S
O1	Bind attribute to an encoding	Attribute	X/Y axes, Color legend	Axis title region, Color legend title, Attribute pills II		*I1:* Drag II to DI	Drag worldwide gross to X-axis title	✓		
						I2: Point on DI and tap II	Point on X-axis title and tap major genre	✓		
						I3: Point on DI and write P	Point on Y-axis title and write running time	✓	✓	
						I4: Point on DI and speak P	Point on color legend title + "*Content Rating*"	✓		✓
						I5: Speak < T , P >	"*Show creative type on the x-axis*"			✓
	Bind multiple attributes to an encoding	Attributes	X/Y axes	Axis title region, Attribute pills II	*Add, include, group by, split by*	*I6:* Point on DI and speak < K , P >	Point on Y-axis title + "*Add budget and gross*"	✓		✓
						I7: Speak < T , K , P >	"*Group x-axis by content rating*"			✓
						w/ modifier active:	w/ modifier active:			
						I8: Drag II to DI	Drag major genre to X-axis title	✓		
						I9: Point on DI and tap II	Point on Y-axis title and tap IMDB Rating	✓		
						I10: Point on DI and write P	Point on X-axis title and write Running Time	✓	✓	
	Remove attributes from an encoding	Attributes	X/Y axes, Color legend	Axis title region, Color legend title	*Remove, unbind, clear, delete*	*I11:* Erase P* from DI	Erase X-axis title to remove attributes mapped to X-axis		✓	
						I12: Point on DI and speak < K , P* >	Point on color legend title + "*Clear*"	✓		✓
						I13: Speak < T , K , P* >	"*Remove budget from the y-axis*"			✓
O2	Change data aggregation level	Aggregation level	X/Y axes	Axis title region		*I14:* Point on DI and write P	Point on Y-axis title and write *sum*	✓	✓	
						I15: Point on DI and speak P	Point on Y-axis title + "*max*"	✓		✓
						I16: Speak < T , P >	"*Show the average values on y*"			✓
	Sort	Sort order, Attribute	X/Y axes	Axis title region	*Order, sort, arrange, reorder*	*I17:* Swipe on DI, direction determines P	Swipe downwards on Y-axis title to sort in descending order by Y-axis attribute values	✓		
						I18: Point on DI and speak < K , P* >	Point on Y-axis title + "*Sort*" to sort in ascending order (default system value) by Y-axis attribute values	✓		✓
						I19: Speak < T , K , P* >	"*Arrange x-axis in decreasing order*"			✓
O3	Filter marks		Marks	Marks	*Filter, exclude others, remove, keep only*	*I20:* (Select +) Erase DI	Erase bar in a bar chart to remove a data category		✓	
						I21: w/ DI selected, speak < K >	Select points in a scatterplot + "*remove others*" to filter unselected points		✓	✓
	Filter by criteria	Attribute values	Marks	Color legend rows	*Filter, exclude others, remove, keep only*	*I22:* Erase P from DI	Erase *Action* row from color legend to remove action movies		✓	
						I23: Speak < K , P >	"*Exclude horror movies*"			✓
O4	Get mark details		Marks	Marks		*I24:* Long press DI	Long press on a bar in a bar chart to see its value	✓	✓	
	Get mark details by value	Attribute values	Marks	Axis scales		*I25:* Drag along DI	Drag along X-axis scale in a horizontal line chart to see values for a specific timestamp	✓		
O5	Change chart type	Char type	Canvas			*I26:* Speak < P >	"*Switch to a line chart*"			✓

Table 1. Proposed multimodal interactions for low-level operations during visual analysis. Operations categories are O1: Bind/unbind visual encodings, O2: Modify axes, O3: Filter, O4: Get details, and O5: Change chart type. Unless explicitly specified as an indirect instrument (II), all instruments are direct instruments (DI). An asterisk (*) indicates a parameter is optional. The rightmost column displays modalities (T: Touch, P: Pen, S: Speech) used in an interaction pattern.

of multimodal interaction for data visualization, we started by considering pen, touch, and speech-based interactions with visualization tools regardless of the target device. Specifically, our selection criteria were that systems (1) involved interactions using one or more of pen, touch, or speech input with a single device and user and (2) focused on general visual data exploration and analysis, excluding systems that placed higher emphasis on externalizing users' thoughts (e.g., [23, 34]) or authoring expressive visualizations (e.g., [22, 25, 51]).

Through the survey, we identified five core categories of operations (Table 1) that were supported by most systems. Furthermore, given their frequent occurrence in the surveyed systems and prevalence in common visualization tools, we decided to focus on histograms, bar charts (including grouped and stacked bar charts), line charts, scatterplots, and parallel coordinates plots as our initial visualization types.

Conceptualizing Multimodal Interaction Design

To design consistent interactions, we needed a standardized nomenclature to describe and compare alternative interactions. Correspondingly, we reviewed the terminology and description of interactions in the surveyed visualization systems' papers. However, since most current systems were optimized for specific visualizations (e.g., [2, 9, 19, 43, 46]) or form-factors (e.g., [20, 21, 26, 37]), there was no common language that let us consistently design and discuss possible interactions. Thus, based on our survey and a review of prior work in the more general space of post-WIMP (e.g., [3, 47]) and multimodal interfaces (e.g., [7, 28, 31, 42]), we identified a set of core concepts that could help us (and future system designers) systematically design and reason about interactions in the context of multimodal visualization systems.

We propose thinking of interactions with multimodal visualization systems in terms of the following four concepts. People interact with visualization systems through a set of one or more low-level ***operations*** (e.g., binding an attribute to an encoding, sorting) to accomplish their high-level tasks (e.g., answering data-driven questions, creating specific visualizations). These operations typically require ***parameters*** (e.g., sorting order, attributes, encodings) and operate on one or more ***targets*** (e.g., selected marks, axis, canvas). Finally, operations are mediated through ***instruments*** in the interface (e.g., marks, axes scales). These instruments can be *direct* (i.e., when the target itself mediates an operation) or *indirect* (i.e., when an operation is performed on a target through a separate instrument).

With these general user interface concepts in mind, we investigated possible interactions for the operations identified through the survey. To ensure the resulting interactions were generalizable, we only considered basic elements (e.g., axes,

marks, attribute pills) present in most visualization tools as our instruments. We then examined both interactions demonstrated in previous systems (e.g., writing an aggregation function name to change the data aggregation level [20], dragging along an axis to sort [9]) as well as novel interactions that were potentially more fluid and consistent (e.g., pointing on an axis with a finger and speaking an attribute name to specify mappings, using the pen's eraser to filter).

Table 1 lists the ten low-level operations and the corresponding set of interactions (*I1-I26*) derived after a series of iterations, along with examples and input modalities. We initially considered *selection* and *zoom/pan* as two additional core operations. However, since these are low-level interface actions and sometimes a precursor to other operations (e.g., filtering a set of marks may require selecting them first), we decided not to include them as standalone operations. In our interaction set, selections can be performed in four ways: 1) tapping with pen/finger directly on a mark, 2) tapping with pen/finger on a legend item to select marks by categories [14], 3) dragging the pen on an axis scale to select marks based on data values, and 4) drawing a free-form lasso with the pen on the chart area. Additionally, zoom and pan are supported through the standard two-finger pinch and single finger drag gestures on the chart area, respectively.

Design Principles

In this section, we describe five underlying principles we had when designing our multimodal interactions. We compiled these principles based on the surveyed papers as well as design guidelines from prior work advocating for post-WIMP visualization interfaces [10, 24, 33].

DP1. *Maintain interaction consistency across visualizations*
To support consistency, we prioritize globally functional interactions (i.e., ones that work across visualization types) over locally optimal interactions (i.e., ones that are specific to a type of visualization). A pattern from Table 1 exemplifying this principle is *I15: Drag along axis scales* to see mark details. Previous systems have inconsistently used this interaction to sort bars in a bar chart [9] and select points in a scatterplot [36]. However, since these are both locally optimal interactions, we use dragging along an axis scale to display mark details which is a common operation across visualizations.

We note that some operations are specific to certain visualization types. For example, sorting an axis is meaningful to bar charts and parallel coordinate plots but not to scatterplots and line charts. Thus, *I17:swiping on the axes* only works for the appropriate visualizations and has no effect in others. We also reserve the swipe interaction for sorting and do not employ it for a different operation elsewhere.

DP2. *Minimize indirection in interactions*
Aligned with the guidelines for fluid interaction [10], we try to enable interactions with direct instruments (e.g., marks, axes), avoiding external controls and indirect instruments that are separated from the view. For instance, to filter marks, people can use *I20: erase marks* directly instead of adjusting external widgets like sliders or dropdown menus. Or to see the details of a mark, one can use *I24: long press on marks* instead of indirectly requesting for details through voice.

Another implication of this design principle is that if an operation is inherently indirect, we offload it to speech since it is also, by nature, indirect. Examples of such indirect operations include changing the visualization type and filtering based on attributes that are not encoded in the current view (e.g., filtering points by *imdb rating* in a scatterplot of *production budget* by *worldwide gross*).

DP3. *Leverage simple and familiar gestures*
Simple gestures that are familiar to users are easier to learn and subjectively preferred for interacting with pen and touch-based visualization systems [9, 20, 37, 49]. To maintain simplicity and promote familiarity, we avoid devising specialized gestures for individual operations. In fact, as illustrated by the patterns in Table 1, all our pen and touch interactions only involve common gestures including tap, point (hold), swipe, and drag. Particularly for cases where one gesture could be mapped to multiple operations, instead of introducing an alternative gesture for one of the operations, we apply a division of labor tactic [16] and offload the interaction to a different modality. For example, due to their ubiquity across devices and applications, we reserve touch-based pinch and drag gestures for zoom and pan, respectively. However, dragging on the chart area is also an intuitive way to perform selection (e.g., by drawing lassos [36]), which is another important action during visual analysis [52]. To resolve this conflict, we leverage a second modality and allow people to draw selection lassos by dragging on the chart area using the pen.

DP4. *Avoid explicit interaction modes*
Interaction modes enable an interface to support a wider range of operations. However, constantly switching between modes (e.g., inking vs. gesture) for the same type of input (e.g., pen) can be disruptive to the users' workflow and are known to be a common source of errors [27,32]. To avoid explicit interaction modes, we assign semantically meaningful actions to different modalities (e.g., touch for pan/zoom, pen for selection) and leverage a combination of modalities to support advanced variations of simpler actions (e.g., using bimanual pen and touch input to compound selections).

DP5. *Strive for synergy not equivalence*
A common myth about multimodal interaction is that all modes of input can support all operations [30]. Instead of designing specialized interactions (e.g., highly customized and complex gestures or widgets) to ensure equivalence, we support equivalence between modalities only if the equivalence is inherently meaningful. For instance, we allow binding attributes to encodings using all three modalities (*I1-I10* in Table 1). On the other hand, because there is no direct interaction (**DP2**) to filter marks based on an attribute that is not encoded in the view using pen or touch, we only allow this via speech (e.g., saying "*Remove movies with an imdb rating under 8*" when the system shows a scatterplot of *budget* and *gross*).

Furthermore, we also leverage *complementarity-based interactions* [28], where different chunks of information are provided by different modalities and subsequently merged together to

perform an operation. In addition to help accomplish **DP2** and **DP3**, complementarity can also facilitate faster interactions and reduce the complexity of speech commands, ultimately improving both the user and system performance [4, 28, 48]. For instance, with touch alone, binding multiple attributes to an axis requires multiple interactions with control panel widgets such as dropdown menus (e.g., [9]). Alternatively, during *I6:pointing on the axis and speaking*, the axis (target) is implicitly determined by touch whereas speech allows specifying multiple attributes (parameters) as part of the same action. Similarly, to support negative filters, instead of providing additional keep-only button or menu item in touch-only systems (e.g., [9, 36]), a system with *I21:select-and-speak* can let people select points by drawing a lasso with a pen and saying "*exclude others*." In this case, the target (marks) is specified through the pen while speech provides the operation (via "*exclude*") and further modifies the target (via "*others*").

Note that the concepts and interactions in Table 1 are by no means an exhaustive or definitive set. They are only one sample set of interactions we designed with **DP1-5** and basic elements of visualization systems in mind. In fact, depending on a system's interface, some of these interactions may not even be applicable. For instance, if a system does not explicitly list attributes as interactive widgets, the *I1:drag-and-drop* and *I2:point-and-tap* interactions involving attribute pills to bind attributes to encodings cannot be used. However, the *I3:point-and-write* and *I4:point-and-speak* interactions for the same operation remain valid since they rely on the X/Y axes of the chart itself.

INCHORUS

To demonstrate how the proposed interactions collectively support visual data exploration on tablets, we employ them in a prototype system, InChorus (Figure 1). In addition to the basic elements such as axes, legend, attribute pills, etc., similar to previous pen and touch systems (e.g., [38, 51]), we also added a modifier button (Figure 1C) that serves two purposes: 1) it allows utilizing bimanual input, which can help avoid explicit mode switches during pen- or touch-only interactions (**DP4**), and 2) it serves as a "record" button to input voice commands. InChorus uses a "push-to-talk" technique: it records speech while a finger is on the modifier button, the X/Y axis title regions, or the color legend title, and executes the recognized command once the finger is lifted.

Interacting with InChorus

We now illustrate key interactions in InChorus through a usage scenario (also depicted in the supplementary video). Imagine that Joe, an analyst at a movie production house, wants to identify movie characteristics his company should focus on for their next investment. To investigate previously released movies, Joe loads a dataset of 709 movies into InChorus. The dataset contains nine attributes for each movie, including *Release Year* (📅 temporal), *Worldwide Gross* (**#** quantitative), and *Major Genre* (**A** categorical), among others; all nine attributes are shown on the left panel (Figure 1A).

Identifying key genres. To get an overview of values for each attribute, Joe *taps* on individual attributes in the side panel

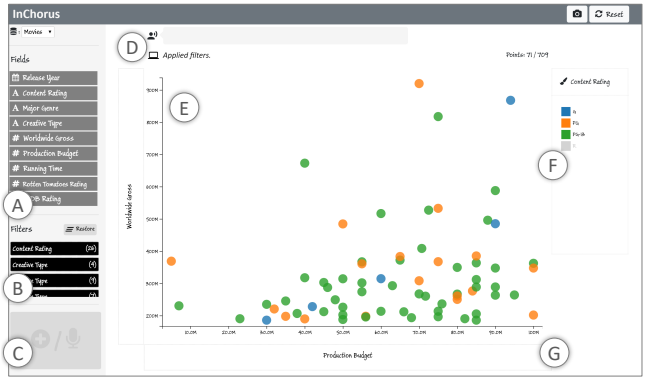


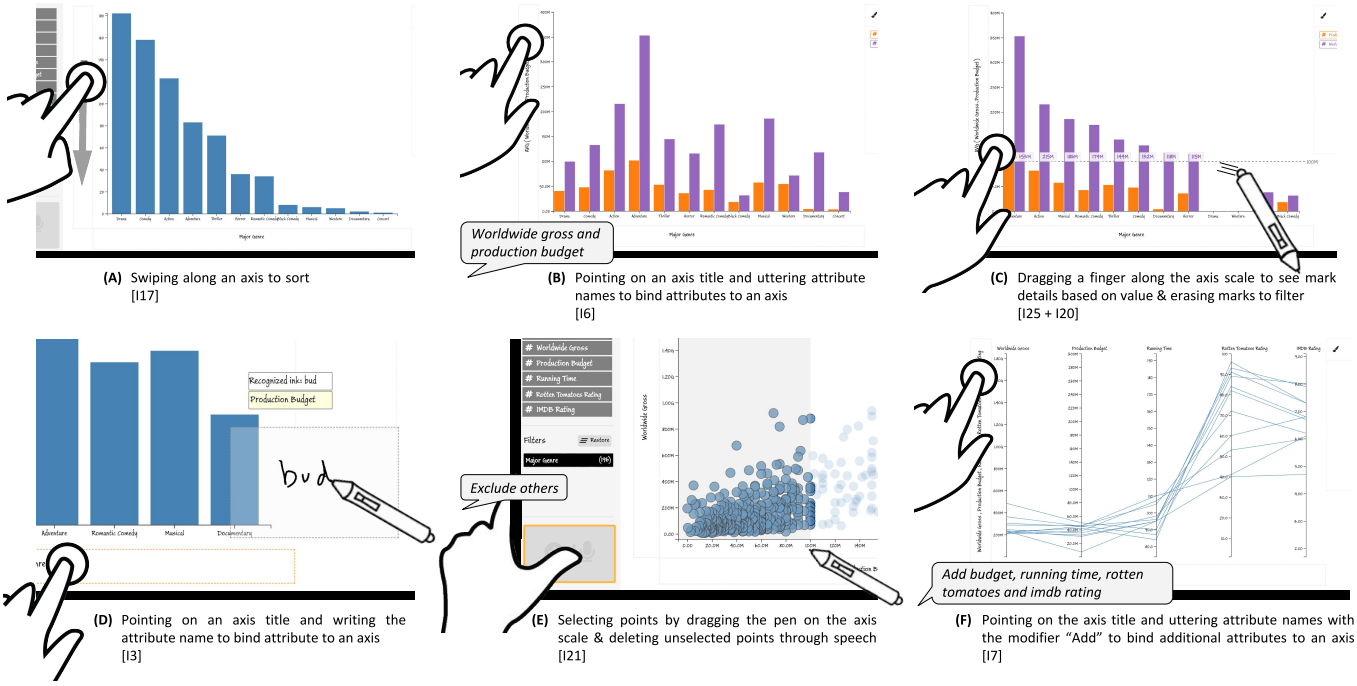
Figure 1. InChorus' interface components. (A) Attribute pills, (B) Active filters, (C) Modifier button, (D) Speech command display and system feedback row, (E) Chart canvas with marks (in this case, circles), (F) Color legend area, and (G) Axis scale and title area.

while pointing (i.e., holding down a finger) on the X-axis title region (Figure 2). As he taps through the attributes, InChorus displays univariate summary visualizations (histograms, bar charts, and line charts) based on the attribute's data type.

To see the popularity of different genres, Joe binds the *Major Genre* attribute to the X-axis, creating a bar chart. Joe decides to sort the bars by the number of movies so he can identify more popular genres faster. As he *swipes downwards* on the Y-axis, InChorus sorts the genres in descending order by count (Figure 3A). To get a sense of the typical return on investment for different genres, Joe adds the gross values to the view by *pointing* on the Y-axis and *saying* "*Worldwide gross and production budget*." This updates the view to a grouped bar chart displaying the average gross and budget for different genres (Figure 3B). Now, to look at the highest grossing genres instead of the most popular ones, Joe again wants to sort the view. However, because two attributes are mapped to the Y-axis, instead of swiping, Joe now *points* on the Y-axis and *says* "*Sort by worldwide gross in descending order*" to clearly express his intent.

To see values corresponding to the bars, Joe *drags* his finger along the Y-axis scale. As he *drags* his finger along the Y-axis scale, InChorus displays a horizontal ruler highlighting the value corresponding to his finger's position and shows

Figure 2. Tapping an attribute while pointing on the x-axis title region binds the data attribute to the x-axis.



(A) Swiping along an axis to sort [I17]

(B) Pointing on an axis title and uttering attribute names to bind attributes to an axis [I6]

(C) Dragging a finger along the axis scale to see mark details based on value & erasing marks to filter [I25 + I20]

(D) Pointing on an axis title and writing the attribute name to bind attribute to an axis [I3]

(E) Selecting points by dragging the pen on the axis scale & deleting unselected points through speech [I21]

(F) Pointing on the axis title and uttering attribute names with the modifier "Add" to bind additional attributes to an axis [I7]

Figure 3. Using InChorus to explore a movies dataset. Sub-figure captions describe the interactions being performed [along with the corresponding interaction pattern labels from Table 1].

details of bars that intersect with the ruler. Inspecting the chart, Joe decides to only focus on high grossing genres. He uses the axis value-based ruler as a cut-off point and *erases* bars corresponding to genres with an average *Worldwide Gross* under 100M, filtering them from the view (Figure 3C).

Shortlisting profitable movies. With genres shortlisted, Joe now wants to compare the budget and gross for individual movies using a scatterplot. To do this, he first *erases* the *Production Budget* from the Y-axis title to remove it. He then *points* on the X-axis title region and starts writing "*budget*" in the ink pad, selecting *Production Budget* from the recommended list of attributes (Figure 3D). This replaces the *Major Genre* attribute on the X-axis with the *Production Budget*, creating a scatterplot. Since he works for a relatively small production house, he decides to focus on lower budget movies. He *drags the pen* along the X-axis scale to select movies with a budget under 100M and *says* "*exclude others*" (Figure 3E) to remove the unselected movies. To further focus on movies with high a return on investment, he also removes movies with a gross of under 200M.

With the filtered scatterplot, Joe starts examining other attributes. To understand what types the movies were, Joe maps the *Creative Type* attribute to the color of the points by *saying* "*Color by creative type*." Noticing that *Contemporary Fiction*, *Kids Fiction*, and *Science Fiction* are most popular, he filters out the other movie types by *erasing* them from the legend. Similarly, mapping the *content rating* to the color of points, Joe removes *R*-rated movies since his company is more interested in movies catering to a universal audience. This filtering results in the view shown in Figure 1.

Inspecting the scatterplot, Joe notices a set of low budget movies that have made a profit of over 5x. He selects these movies at the bottom left corner of the view as well as the three highest grossing movies at the top right corner of the view by *holding the modifier button* and *drawing* two free-form lassos around the points using the pen. Joe then filters out other movies from the chart by *saying* "*remove others*."

Comparing shortlisted movies. Finally, to analyze what characteristics made the shortlisted movies so successful at the box office, Joe wants to inspect and compare the shortlisted movies with respect to the relevant attributes. Joe first removes the *Production Budget* from the X-axis by *erasing* it. He then *points* on the Y-axis (currently showing only the *Worldwide Gross*) and *says* "*Add budget, running time, rotten tomatoes and imdb rating*." InChorus, in response, adds the additional attributes to the Y-axis, creating a parallel coordinates plot. Joe further investigates the shortlisted movies by selecting different value ranges on the parallel axes and identifies a final list of five movies to present as examples of profitable and reliable investments to the management team.

Affordances and Feedback

To make users aware of possible actions, InChorus presents different types of affordances. For instance, when the user points on an axis, InChorus highlights attributes that can be mapped to that axis, renders an ink pad to indicate that written input can be provided, and flashes the command display box and microphone red to indicate that speech recognition is active (Figure 2). Note that the attribute pills are contextually highlighted based on the active view and the attribute type. For example, in Figure 2, *Major Genre* stays dark gray because it is already mapped to the X-axis. Alternatively, if one pointed on the Y-axis with a categorical attribute shown on the X-axis, the system would only highlight numerical attributes

in the panel since we currently do not support visualizations simultaneously showing categorical attributes on both axes.

InChorus also provides constant feedback based on user actions. In addition to visual feedback for direct pen/touch actions (e.g., fading unselected points in Figure 3E, adding an orange stroke when pointing on an axis title as shown in Figure 2), the system also displays three types of textual feedback messages above the chart area (Figure 1D): 1) Success: when the system successfully executes operations in response to user actions (example messages include *Coloring by Major Genre*; *Sorted bars by Worldwide Gross in descending order*), 2) Void action: when users performs a valid operation but the operation has no effect on the view (example messages include *No points meet that filtering criteria*; *Bars are already sorted in descending order by IMDB Rating*), and 3) Error: when users perform invalid actions or the system is unable to interpret a speech command (example messages include *The pen cannot be used in the panel area. Please use touch.*; *Unable to process that command. Please try a different one*).

By providing contextually-relevant affordances before an action and complementing them with feedback after actions, InChorus helps users both know what actions are available as well as interpret the system's reactions to their actions.

Implementation

InChorus is implemented in JavaScript as a web-based application. All visualizations are rendered using D3.js [5]. Pen and touch inputs are collected as standard JavaScript events and processed by custom event handlers. InChorus uses the HTML5 speech recognition API [50] for translating speech-to-text. To improve recognition accuracy, the speech recognizer is trained with the operation-specific keywords (Table 1) and attributes and values in the loaded dataset.

We implemented a custom JavaScript-based lexical parser to interpret the recorded NL commands. The lexicon consists of the attributes and values in the dataset as well as manually defined operation-specific keywords. To identify operations, targets, and parameters, the system compares the tokenized input string to the lexicon. If it is unable to detect a target using the NL command alone, the system employs multimodal fusion and infers the target through the invoking instrument (e.g., axes) and the active view (e.g., selected points).

USER STUDY

We used InChorus as a test bed to assess the general usability of the proposed interactions and gather subjective feedback. In particular, since multimodal interaction with visualizations through pen, touch, and speech is a novel concept, we wanted to assess its practical viability and see whether people actually adapt to using this style of interaction and are able to perform common visual analysis tasks.

Participants and Setup

We recruited 12 participants (P1-P12; six females, five males, and one "undisclosed"), ages 27-55, via email through mailing lists of a large technology company. All participants rated themselves as being fluent English speakers and had a working knowledge of data visualizations (i.e., understood basic visualization types and elements such as axes, legends) but not necessarily specific visualization systems (e.g., Tableau, Microsoft Power BI). In terms of prior experience with input modalities, all participants said they use touch-based systems on a daily basis. Two participants said they use a pen once in a few weeks, five said they had used it in the past but do not use it regularly, and five said they had no experience working with a pen. Eight participants said they use voice-based systems on a daily basis, three said they use it on a weekly basis, and one said she only occasionally uses voice-based systems.

Sessions were conducted in-person in a quiet conference room. Participants interacted with InChorus on Google's Chrome browser on a 12.3" Microsoft Surface Pro set to a resolution of 2736 x 1824. Participants were encouraged to position the device in a way that was most comfortable for them. A 24" monitor was used to display the study instructions and tasks as a slide show. Participants received a $50 gift card as a compensation for their time. All sessions were audio and video recorded.

Procedure and Tasks

Each study session had four phases including a training phase, two task phases, and debriefing. The study protocol and tasks were iteratively refined through 12 pilot sessions. We included two types of tasks to emulate two significantly different visual analysis scenarios. Specifically, the first task phase emulates scenarios where users know the operations they want to perform and have to communicate that intent to the system through interactions. On the other hand, the second task phase emulates scenarios where users first need to think about the task they want to accomplish (e.g., think about the attributes they want to use and the type of chart they want to create), translate that task into system operations (e.g., binding attributes to specific encodings), and finally perform those operations through the supported interactions. Each session lasted between 71-124 minutes (average: 86 min).

Introduction & Training. After they provided consent and filled out a background questionnaire, participants were introduced to the various system operations along with possible interactions for each operation using a dataset of 303 cars with eight attributes (e.g., *Horsepower*, *Acceleration*, *Origin*) for each car. During this phase, as training, participants were free to practice the interactions until they felt confident performing them. This phase lasted approximately between 37-60 minutes (average: 46 min).

Task Phase 1: Replication and Value Identification. In this phase, participants were given four tasks using the IMDB movies dataset introduced as part of the usage scenario earlier. Each task consisted of a set of one to four sub-tasks that displayed a visual state for participants to replicate or values they had to identify. For example, one of the tasks had four sub-tasks requiring participants to 1) recreate a given grouped bar chart, 2) filter out two categories of values, 3) switch to a multi-series line chart, and 4) identify series values for a specific year. This first task phase took approximately 8-26 minutes (average: 13 min) to complete.

Task Phase 2: Fact Verification. Participants were given a dataset of 500 US colleges with nine attributes for each college including a number of numerical (e.g., *Cost*, *Admission Rate*) and categorical attributes (e.g., *Control Type*, *Region*). Following the jeopardy-style evaluation [13] for visualization NLIs, we gave participants five statements (e.g., *There are more public schools in Southwest than the Great Lakes*) that they had to mark as true or false based on their exploration of the data. As with other systems evaluated using this methodology [13, 43], parroting the given statement to the system would not result in the answer. In addition to stating the answer, participants also had to verbally justify their responses and take screenshots of visualizations they used. This phase lasted approximately between 7-28 minutes (average: 14 min).

Debrief. At the end of the session, we had a debriefing phase that included a post-session questionnaire consisting of likert-scale questions regarding their interaction experience and an interview. During the interview, we asked participants general questions about their overall experience, asking them to list interactions they particularly liked/disliked, as well as targeted questions based on our observations during the session.

Results

All participants successfully replicated charts or identified values for the four tasks in the first phase. For the second task phase, nine participants correctly verified all five statements whereas three (P1, P5, P10) correctly verified three out of the five statements. The average completion time for individual tasks was 3 minutes for the first phase and 2:30 minutes for the second.

Interaction Summary

To track participants' interactions, we reviewed the session videos to count the number of times participants attempted operations in Table 1 along with which interaction patterns they used. We identified a total of 1197 attempts, out of which 1000 (83%) executed successfully, 67 (6%) were invalid operations (e.g., mapping an attribute to color with no attribute mapped to the X/Y axes), 15 (1%) used unsupported speech commands (e.g., invoking an axis-value based ruler and saying "*Remove points under this*") or pen/touch gestures (e.g., dragging an attribute from the color legend to the X axis), and 115 (10%) involved erroneous interactions (e.g., speech recognition errors, conflicting pen and touch).

For invalid operations, unsupported and erroneous interactions, participants often reattempted their initial interaction one or more times before switching to a different interaction pattern or operation. Hence, to avoid double counting interaction patterns by including these reattempts, we only summarize the 1000 valid and successful interactions in Figure 4A.

Subjective Feedback

Figure 4B summarizes participants' responses to the post-session questionnaire. In general, participants were positive about the overall experience stating they found the system functionalities well integrated (M = 4.08 out of 5) and that they would want to use the system frequently (M = 4.08). However, all participants noted that there was a learning curve especially during the training phase. That said, participants felt the interactions were intuitive and just needed some getting used to as they had no prior experience with multimodal interfaces. For instance, P6 compared her initial reaction to using a new type of keyboard and said "*...it was a cohesive system. It's just kind of getting the muscle memory down. It's like when you switch to someone else's keyboard you know where everything is but you just can't type.*" However, after the training, not only did she successfully complete all tasks but as highlighted by the counts in Figure 4A, did so using a variety of patterns including unimodal interactions, bimanual interactions, and multimodal interactions combining pen/touch with speech.

While seven participants said they did not find it difficult to switch between or combine modalities, five said it was confusing, in particular to differentiate between pen and touch. For instance, P9 said "*Distinguishing and switching between voice versus not-voice was not difficult but pen versus finger was a tougher one. I don't think I make as much of a distinction between pen and finger and only use the pen to be more specific, it's a finer point as opposed to being a different tool.*" We believe this confusion may have been exacerbated by the fact that the three participants giving this feedback had never used a pen before and the remaining two had used it minimally in the past. Nonetheless, this confusion fuels the open challenge with designing bimanual pen and touch interaction [11, 16], deeming further investigation.

Lastly, on average, participants were neutral about how "natural" it felt to interact with the system (M = 3.33), with four participants stating they felt the system did not support their natural workflow. During the interviews, we noted that more so than the interaction patterns, this stems from the fact that some participants preferred not to manually specify mappings to explore the data. For instance, P5 said "*I'm an excel person. If I had different templates I would have selected different types of graphs and seen what the data looks like and then chose the scatter graph.*"

DISCUSSION AND FUTURE WORK

Towards a Grammar for Multimodal Interaction

Although we focus on tablets as our primary use case in this paper, the underlying conceptual framework (designing and describing multimodal interactions with visualizations in terms of operations, targets, parameters, and instruments) is not limited to tablets or the listed operations. For instance, the *I3:point-and-write* or *I4:point-and-speak* interactions with X/Y axes or the color legend to bind attributes to encodings can be employed as-is in a system like SketchInsight [26] that supports visual data exploration on interactive whiteboards.

In addition to different form-factors, the presented concepts can also be used to design consistent interactions for more advanced visual analysis operations. One such operation might be generalized selection [14] that allows people to specify selection criteria relative to a subset of the data they are interested in. Previous touch-only systems have supported generalized selection through explicit interaction modes and WIMP-style widgets [38]. However, multimodal systems could support this in a more fluid manner by allowing users to specify

A

Operation	Interaction Patterns	Total (with %)	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
Bind attribute to an encoding	Drag Il to Dl	150 (40%)	24	3	26	33			29	1	10		12	12
	Point on Dl and tap Il	177 (47%)	2	27		1	46	29	3	21	9	20	13	6
	Point on Dl and write P	6 (2%)	2							3	1			
	Point on Dl and speak P	42 (11%)	4	1	8	3				3	5	10		8
	Speak < T , P >	1 (0%)							1					
Bind multiple attributes to an encoding	Point on Dl and speak < K , P >	21 (40%)	3			3		3		5	2	4		1
	Speak < T , K , P >													
	Drag Il to Dl	11 (21%)			3	2			5				1	
	Point on Dl and tap Il	20 (39%)		5			2			2	2	5	2	2
	Point on Dl and write P													
Remove attributes from an encoding	Erase (P from) Dl	71 (93%)	2	1	3	10	6	9	7	9	4	14	4	2
	Point on Dl and speak <(K) , P >	5 (7%)	1	1	2									1
	Speak < T , K ,(P) >													
Change data aggregation level	Point on Dl and write P	13 (50%)	1		1	1	2	2	1	2	1		2	
	Point on Dl and speak P	13 (50%)	1	2	1	2			1	2		2		2
	Speak < T , P >													
Sort	Swipe on Dl	119 (94%)	8	6	19	9	4	19	15	6	10	15	8	
	Point on Dl and speak < K , P >	8 (6%)			1									7
	Speak < T , K , P >													
Filter marks	(Select +) Erase Dl	20 (17%)		4		2		2	8		3			1
	w/ Dl selected, speak < K >	100 (83%)	15	10	11	7	10	6	2	7	10	7	7	8
Filter by criteria	Erase P from Dl	29 (73%)			3	1	1	6	4	5		4	2	3
	Speak < K , P >	11 (27%)	1		1	6	1		1	1				
Get mark details	Long press Dl	114 (100%)	2	11	12	12	12	17		17	1	13	7	10
Get mark details by value	Drag along Dl	49 (100%)	2	1	4	7	4	1	9	1	12	1	5	2
Change chart type	Speak < P >	20 (100%)	1	1	1	1	1	1	1	8	1	2	1	1

B

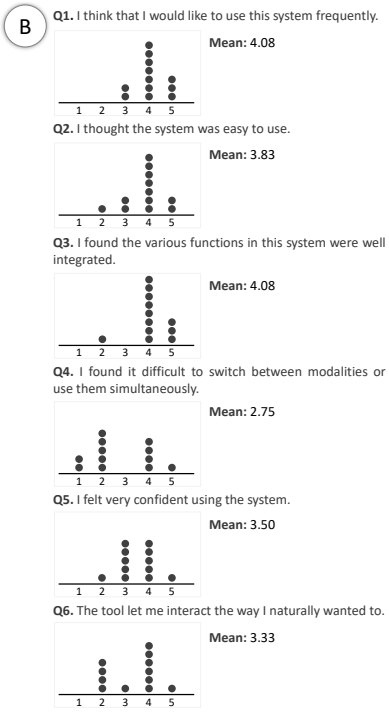


Figure 4. Summary of study results. (A) Operations executed during the study along with the frequency of interaction patterns which were used to perform those operations. Bar widths are normalized for individual operations to facilitate comparison of alternative interaction patterns and (B) Responses to post-session likert-scale questions about the interaction experience.

their selections via pen/touch and generalization criteria via speech. For instance, we can define a new interaction pattern: *Select* Dl *and speak* < K , P > where Dl: marks on the view, P: selection criteria in terms of attributes and values, and K: ["select", "highlight", "lower", ...]. Thus, based on the user selecting a point in a scatterplot and saying "*Select Action movies with a lower budget than this*," the system can infer the operation and leverage the active selection to find and select desired points, all while preserving the current interactions in Table 1.

While these are just examples, in general, standardizing the design of interactions with respect to logical concepts such as operations, targets, parameters, and instruments presents a compelling opportunity to create a unifying grammar to describe multimodal interactions with visualization systems. In addition to streamlining multimodal interaction design, such a grammar could also help develop extensible toolkits and interpreters that let designers create, share, and modify interaction patterns across visualization types and interfaces.

Potential of "Restricted" Natural Language Interfaces

A majority of the work on visualization NLIs has focused on developing interpretation techniques for complex and underspecified commands [13, 17, 40, 41]. While this is the holy grail of NLIs in general, responding to high-level questions and underspecified commands is challenging and highly error prone due to issues such as ambiguity and preserving context across commands. Our work sheds light on a more modest but less error prone class of NLIs for visualization. Specifically, with InChorus, we allow users to issue short keyword-based commands that can be used in conjunction with another modality (**DP5**) or individually to perform low-level operations. We found that these simple commands allowed participants to perform required operations while improving the overall speech recognition and interpretation accuracy: out of a total of 274 utterances, we only had 32 (11%) recognition errors and 7 interpretation errors (2%). This speech recognition error rate of an average of 3% per session is noticeably lower than previous speech based visualization systems like Orko [43] (avg. of 16% per session). With such results in mind, in line with recent work advocating for "restricted" NLIs in complex domains [29], perhaps an opportunity lies in exploring "restricted" NLIs for visualization that build upon simple lexicons and smoothen the transition from current direct manipulation and WIMP-based systems to NLIs. Although they start simple, over time, such systems could evolve to support more complex commands, incrementally exposing users to more advanced system functionality [12, 42].

Synergy between Input Modalities

Our goal was to enable an overall consistent and fluid interaction experience during visual analysis, accommodating users' individual preferences. To this end, instead of aiming to achieve equivalence between modalities, we synergistically combined three input modalities (i.e., pen, touch, and speech), leveraging their unique advantages (**DP5**). The distribution of interaction frequencies in Figure 4A illustrates that the proposed multimodal interactions accommodated varied user preferences while allowing all participants to perform common visual analysis tasks. While some participants (e.g., P1, P8, P9) were open to trying a wider range of interactions and using all modalities and others (e.g., P5, P6, P11) preferred resorting to touch as much as possible, all participants adapted

themselves to using multimodal input and, when needed, successfully used different modalities (individually and in combination) to complete tasks. This was particularly encouraging given that some participants had no experience with some modalities (e.g., five out of 12 participants had never used a pen). P7 aptly summarized his overall experience of interacting with InChorus stating "*It [InChorus] feels completely integrated like this is one thing it's not like this is the pen stuff that I'm doing and now I got to sort through the finger things I can do or the things with voice commands it was like I'm going to interact with this system however best suits me and I'm able to do that nine times out of ten*."

Improving Affordances and Feedback

From the 115 erroneous interactions, we identified seven types of errors that participants encountered—*E1:* dragging an attribute pill outside the screen or dropping it outside the axes/legend (36), *E2:* speech recognition errors (32), *E3:* forgetting to trigger recording of voice commands before speaking (16), *E4:* conflicting pen and touch (e.g., trying to drag an attribute pill with the pen) (15), *E5:* dragging an attribute while pointing on an axis (6), *E6:* forgetting to select from the list of recognized items while writing (5), and *E7:* issuing incomplete speech-only commands (e.g., saying "*Remove under 1200*" without specifying an attribute name) (5). While a majority of these errors were primarily due to inexperience with the interface & input modality (E1, E5, E6) or technological errors (E2), others (E3, E4, E7) were more specific to multimodal input. In addition to the errors, there were also 67/1197 (6%) instances where participants performed a valid interaction but the intended operation was invalid given the state of the view (e.g., sorting a numerical axis in a scatterplot).

As described earlier, we had designed affordances and feedback mechanisms to prevent such errors and invalid operations and help users recover from them. However, during the study, participants rarely noticed the affordances and feedback, often repeating the same interactions multiple times before realizing it was an error or an invalid operation. Going forward, it would be interesting to investigate alternative mechanisms such as multimodal feedback (e.g., auditory and haptic feedback), automated voice recording (e.g., automatically triggering recording after selections are made), and visual aids to clarify support for pen versus touch input (e.g., different background colors or animation overlays for the first time interactions).

Balancing Novelty and Familiarity

The high frequency of drag-and-drop interactions in both Figure 4A and the list of erroneous interactions raises an interesting question about why did participants not switch to alternative patterns upon encountering errors. On further inspection, we found that while most participants switched to the *I2:point-and-tap* interaction, there were some (P3, P4, P7) who persisted with *I1:drag-and-drop* largely due to its familiarity. For instance, P4 had 14 drag-and-drop errors but still preferred it over other patterns for binding attributes to encodings. When asked about why he did not switch to a different pattern, he initially said "*drag-and-drop feels quicker even if I fail than stop and try and articulate the word to match the AI to get what I want out of it.*" He later acknowledged that point-and-tap would have been more accurate and possibly faster, ultimately attributing his preference for drag-and-drop to familiarity. Such observations further motivate the need to explore interfaces that give users the freedom to choose their preferred style of interaction and balance between familiarity and novelty depending on the task at hand (e.g., if speed and efficiency were paramount for the task, P14 may have preferred *I2:point-and-tap* over *I1:drag-and-drop*).

Study and Prototype Limitations

The user study helped us verify that participants were able to successfully adapt and use the proposed multimodal interactions consistently across different visualization types. However, as with most lab studies, our study has some practical limitations and was scoped to a predefined set of tasks, datasets, and visualizations. Changing either of these could affect how users behaved with the system and the inferences made. For instance, in his interview, P9 said "*When there's only a handful of fields here it's easy to drag and select but if there are like hundreds of fields, that's when you would see me writing the name of the field I'm looking for.*" suggesting that his interaction behavior would be different for datasets with a larger number of attributes. Thus, an important next step is to leverage the feedback from this initial study and test the interactions in the context of additional types of datasets and visualizations (e.g., heatmaps, node-link diagrams), perhaps also investigating more open-ended exploratory tasks. Furthermore, because our study focused on adoption and usability of interactions, follow-up experiments are required to more deeply understand the use of multiple modalities along with the associated cognitive challenges and analytical benefits.

Although InChorus was primarily designed to illustrate and test the proposed interactions, the system can be improved and expanded to enhance the overall usability and user experience of interacting multimodally. Some of these improvements include adding an undo/redo feature, providing more details about filtered points, and adding visualization types that support additional configurations of the view (e.g., heatmaps to allow categorical attributes on both axes).

CONCLUSION

We present multimodal interactions for performing visual data analysis operations on tablet devices, designed with the high-level goal of maintaining interaction consistency across different types of visualizations. We describe the design process we followed to systematically develop multimodal interactions by modeling interactions with respect to core concepts including *operations*, *targets*, *parameters*, and *instruments*. Through a user study with 12 participants performing visual analysis tasks with a prototype system implementing the proposed interactions, we discuss how participants adapted to using multimodal interaction and how the freedom of expression afforded by multiple modalities accommodated their interaction preferences. Ultimately, by highlighting the potential benefits of multimodal input and promoting the systematic development of multimodal interactions, we hope to encourage the design of a new generation of visualization systems that can facilitate more natural and fluid human-data interaction, accommodating varying user preferences.

REFERENCES

[1] Jillian Aurisano, Abhinav Kumar, Alberto Gonzalez, Jason Leigh, Barbara DiEugenio, and Andrew Johnson. 2016. Articulate2: Toward a conversational interface for visual data exploration. In *Posters of IEEE VIS '16*.

[2] Dominikus Baur, Bongshin Lee, and Sheelagh Carpendale. 2012. TouchWave: kinetic multi-touch manipulation for hierarchical stacked graphs. In *Proceedings of ACM ITS '12*. 255–264.

[3] Michel Beaudouin-Lafon. 2000. Instrumental interaction: an interaction model for designing post-WIMP user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 446–453.

[4] Richard A Bolt. 1980. "Put-that-there": Voice and gesture at the graphics interface. *SIGGRAPH Comput. Graph.* 14, 3 (July 1980), 262–270. DOI: http://dx.doi.org/10.1145/965105.807503

[5] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2301–2309.

[6] Jeffrey Browne, Bongshin Lee, Sheelagh Carpendale, Nathalie Riche, and Timothy Sherwood. 2011. Data analysis on interactive whiteboards through sketch-based interaction. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*. ACM, 154–157.

[7] Philip R Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen, and Josh Clow. 1997. QuickSet: Multimodal interaction for simulation set-up and control. In *Proceedings of the ACL Conference on Applied Natural Language Processing '97*. 20–24.

[8] Kenneth Cox, Rebecca E Grinter, Stacie L Hibino, Lalita Jategaonkar Jagadeesan, and David Mantilla. 2001. A multi-modal natural language interface to an information visualization environment. *International Journal of Speech Technology* 4, 3-4 (2001), 297–314.

[9] Steven M Drucker, Danyel Fisher, Ramik Sadana, Jessica Herron, and m.c. schraefel. 2013. TouchViz: a case study comparing two interfaces for data analytics on tablets. In *Proceedings of ACM CHI '13*. 2301–2310.

[10] Niklas Elmqvist, Andrew Vande Moere, Hans-Christian Jetter, Daniel Cernea, Harald Reiterer, and TJ Jankun-Kelly. 2011. Fluid interaction for information visualization. *Information Visualization* 10, 4 (2011), 327–340.

[11] Mathias Frisch, Jens Heydekorn, and Raimund Dachselt. 2009. Investigating multi-touch and pen gestures for diagram editing on interactive surfaces. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*. ACM, 149–156.

[12] Anushay Furqan, Chelsea Myers, and Jichen Zhu. 2017. Learnability through adaptive discovery tools in voice user interfaces. In *ACM CHI '17 Extended Abstracts*. 1617–1623.

[13] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G Karahalios. 2015. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of ACM UIST '15*. 489–500.

[14] Jeffrey Heer, Maneesh Agrawala, and Wesley Willett. 2008. Generalized selection via interactive query relaxation. In *Proceedings of ACM CHI '08*. 959–968.

[15] Jeffrey Heer and Ben Shneiderman. 2012. Interactive dynamics for visual analysis. *Queue* 10, 2 (2012), 30.

[16] Ken Hinckley, Koji Yatani, Michel Pahud, Nicole Coddington, Jenny Rodenhouse, Andy Wilson, Hrvoje Benko, and Bill Buxton. 2010. Pen+ touch= new tools. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 27–36.

[17] Enamul Hoque, Vidya Setlur, Melanie Tory, and Isaac Dykeman. 2018. Applying pragmatics principles for interaction with visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 309–318.

[18] Petra Isenberg and Danyel Fisher. 2009. Collaborative Brushing and Linking for Co-located Visual Analytics of Document Collections. In *Computer Graphics Forum*, Vol. 28. Wiley Online Library, 1031–1038.

[19] Jaemin Jo, Bongshin Lee, and Jinwook Seo. 2015. WordlePlus: expanding wordle's use through natural interaction and animation. *IEEE computer graphics and applications* 35, 6 (2015), 20–28.

[20] Jaemin Jo, Sehi L'Yi, Bongshin Lee, and Jinwook Seo. 2017. TouchPivot: blending WIMP & post-WIMP interfaces for data exploration on tablet devices. In *Proceedings of ACM CHI '17*. 2660–2671.

[21] Jan-Frederik Kassel and Michael Rohs. 2018. Valletto: A Multimodal Interface for Ubiquitous Visual Analytics. In *ACM CHI '18 Extended Abstracts*.

[22] Nam Wook Kim, Nathalie Henry Riche, Benjamin Bach, Guanpeng Xu, Matthew Brehmer, Ken Hinckley, Michel Pahud, Haijun Xia, Michael J McGuffin, and Hanspeter Pfister. 2019a. DataToon: Drawing Dynamic Network Comics With Pen+ Touch Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 105.

[23] Yea-Seul Kim, Nathalie Henry Riche, Bongshin Lee, Matthew Brehmer, Michel Pahud, Ken Hinckley, and Jessica Hullman. 2019b. Inking Your Insights: Investigating Digital Externalization Behaviors During Data Analysis. In *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces*. ACM, 255–267.

[24] Bongshin Lee, Petra Isenberg, Nathalie Henry Riche, and Sheelagh Carpendale. 2012. Beyond mouse and keyboard: Expanding design considerations for information visualization interactions. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2689–2698.

[25] Bongshin Lee, Rubaiat Habib Kazi, and Greg Smith. 2013. SketchStory: Telling more engaging stories with data through freeform sketching. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2416–2425.

[26] Bongshin Lee, Greg Smith, Nathalie Henry Riche, Amy Karlson, and Sheelagh Carpendale. 2015. SketchInsight: Natural data exploration on interactive whiteboards leveraging pen and touch interaction. In *2015 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 199–206.

[27] Yang Li, Ken Hinckley, Ken Hinckley, Zhiwei Guan, and James A Landay. 2005. Experimental analysis of mode switching techniques in pen-based user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 461–470.

[28] Jean-Claude Martin. 1998. TYCOON: Theoretical framework and software tools for multimodal interfaces. *Intelligence and Multimodality in Multimedia interfaces* (1998), 1–25.

[29] Jesse Mu and Advait Sarkar. 2019. Do We Need Natural Language?: Exploring Restricted Language Interfaces for Complex Domains.. In *CHI Extended Abstracts*.

[30] Sharon Oviatt. 1999. Ten myths of multimodal interaction. *Commun. ACM* 42, 11 (1999), 74–74.

[31] Sharon Oviatt, Björn Schuller, Philip Cohen, Daniel Sonntag, and Gerasimos Potamianos. 2017. *The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations*. Morgan & Claypool. 403–448 pages.

[32] J Raskin. 2000. Meanings, modes, monotony, and myths. In *The humane interface: new directions for designing interactive systems*. Addison-Wesley, 33–70.

[33] Jonathan C Roberts, Panagiotis D Ritsos, Sriram Karthik Badam, Dominique Brodbeck, Jessie Kennedy, and Niklas Elmqvist. 2014. Visualization beyond the desktop–the next big thing. *IEEE Computer Graphics and Applications* 34, 6 (2014), 26–34.

[34] Hugo Romat, Nathalie Henry Riche, Ken Hinckley, Bongshin Lee, Caroline Appert, Emmanuel Pietriga, and Christopher Collins. 2019. Activeink:(th) inking with data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 42.

[35] Jeffrey M Rzeszotarski and Aniket Kittur. 2014. Kinetica: naturalistic multi-touch data visualization. In *Proceedings of ACM CHI '14*. 897–906.

[36] Ramik Sadana and John Stasko. 2014. Designing and implementing an interactive scatterplot visualization for a tablet computer. In *Proceedings of AVI '14*. 265–272.

[37] Ramik Sadana and John Stasko. 2016a. Designing multiple coordinated visualizations for tablets. In *Computer Graphics Forum*, Vol. 35. 261–270.

[38] Ramik Sadana and John Stasko. 2016b. Expanding selection for information visualization systems on tablet devices. In *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces*. ACM, 149–158.

[39] Sebastian Schmidt, Miguel A Nacenta, Raimund Dachselt, and Sheelagh Carpendale. 2010. A set of multi-touch graph interaction techniques. In *ACM International Conference on Interactive Tabletops and Surfaces*. ACM, 113–116.

[40] Vidya Setlur, Sarah E Battersby, Melanie Tory, Rich Gossweiler, and Angel X Chang. 2016. Eviza: A natural language interface for visual analysis. In *Proceedings of ACM UIST '16*. 365–377.

[41] Vidya Setlur, Melanie Tory, and Alex Djalali. 2019. Inferencing underspecified natural language utterances in visual analysis.. In *Proceedings of ACM IUI '19*. 40–51.

[42] Arjun Srinivasan, Mira Dontcheva, Eytan Adar, and Seth Walker. 2019. Discovering Natural Language Commands in Multimodal Interfaces. In *Proceedings of ACM IUI '19*. 661–672.

[43] Arjun Srinivasan and John Stasko. 2018. Orko: Facilitating multimodal interaction for visual exploration and analysis of networks. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 511–521.

[44] Hariharan Subramonyam and Eytan Adar. 2018. SmartCues: A Multitouch Query Approach for Details-on-Demand through Dynamically Computed Overlays. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 597–607.

[45] Yiwen Sun, Jason Leigh, Andrew Johnson, and Sangyoon Lee. 2010. Articulate: A semi-automated model for translating natural language queries into meaningful visualizations. In *Proceedings of the International Symposium on Smart Graphics*. Springer, 184–195.

[46] John Thompson, Arjun Srinivasan, and John Stasko. 2018. Tangraphe: interactive exploration of network visualizations using single hand, multi-touch gestures. In *Proceedings of AVI '18*.

[47] Andries Van Dam. 1997. Post-WIMP user interfaces. *Commun. ACM* 40, 2 (1997), 63–67.

[48] Minh Tue Vo and Alex Waibel. 1993. Multimodal human-computer interaction. *Proceedings of ISSD* 93 (1993).

[49] Jagoda Walny, Bongshin Lee, Paul Johns, Nathalie Henry Riche, and Sheelagh Carpendale. 2012. Understanding pen and touch interaction for data exploration on interactive whiteboards. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2779–2788.

[50] Web Speech API 2019. `https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API`. (2019).

[51] Haijun Xia, Nathalie Henry Riche, Fanny Chevalier, Bruno De Araujo, and Daniel Wigdor. 2018. DataInk: Direct and creative data-oriented drawing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 223.

[52] Ji Soo Yi, Youn ah Kang, and John Stasko. 2007. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2007), 1224–1231.

[53] Bowen Yu and Cláudio T Silva. 2019. FlowSense: A Natural Language Interface for Visual Data Exploration within a Dataflow System. *IEEE transactions on visualization and computer graphics* (2019).

[54] Emanuel Zgraggen, Robert Zeleznik, and Steven M Drucker. 2014. PanoramicData: Data analysis through pen & touch. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2112–2121.