Combining Voice and Gesture for Presenting Data to Remote Audiences

Arjun Srinivasan[*]
Tableau Research

Matthew Brehmer[†]
Tableau Research

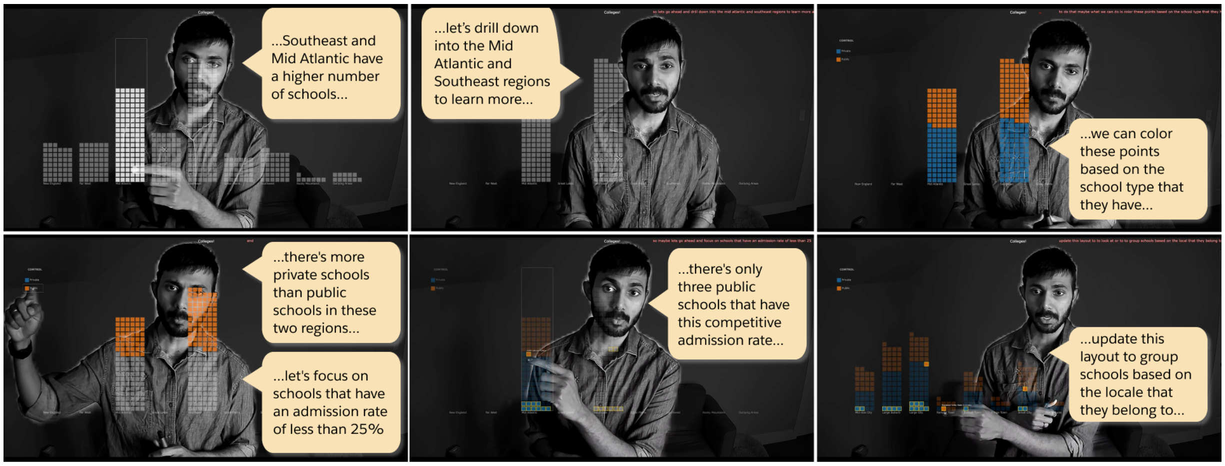
...Southeast and Mid Atlantic have a higher number of schools...
...let's drill down into the Mid Atlantic and Southeast regions to learn more...
...we can color these points based on the school type that they have...
...there's more private schools than public schools in these two regions...
...let's focus on schools that have an admission rate of less than 25%
...there's only three public schools that have this competitive admission rate...
...update this layout to group schools based on the locale that they belong to...

Figure 1: Six moments in a remote presentation about American post-secondary institutions and their admission statistics, in which the presenter appears behind a semi-transparent unit chart composited in the foreground. The presenter ephemerally selects and highlights categories and items in the chart via pointing. Meanwhile, utterances forming part of the presenter's spoken monologue trigger the filtering, sorting, and aggregating of the data. See the supplemental video to watch the 3-minute presentation.

Abstract

We consider the combination of voice commands with touchless bimanual gestures performed during presentations about data delivered via teleconference applications. Our demonstration extends recent work that considers the latter interaction modality in a presentation environment where charts can be composited over live webcam video, charts that dynamically respond to the presenter's operational (i.e., functional and deictic) hand gestures. In complementing these gestures with voice commands, new functionality is unlocked: the ability to precisely filter, sort, and highlight subsets in the data. While these abilities provide presenters with more flexibility in terms of presentation linearity and the capacity for responding to audience questions, imperative voice commands can come across to audiences as stilted or unnatural, and may be distracting.

Index Terms: Human-centered computing—Visualization; Human-centered computing—Human-Computer Interaction (HCI)—Interaction paradigms—Natural language interfaces; Human-centered computing—Human-Computer Interaction (HCI)—Interaction techniques—Gestural input.

1 Introduction and Background

Multimodal interfaces that combine voice and gestural commands have captured our collective imagination for decades [8], though a common reference point continues to be Richard Bolt's *'Put-that-there'* interface (1980) [1], which disambiguates deictic references in speech commands with gestural input (i.e., pointing at a visual element while saying 'this'). In this paper, we draw upon two streams of prior research focusing on such interfaces, those involving presentation applications and those involving data analysis applications.

Leveraging both gesture and voice to control the display of visual aids in a presentation has been an active area of research in recent years, especially since the onset of the COVID-19 pandemic and shifts to remote and hybrid communication in enterprise and educational contexts. While earlier work such as by Fourney et al [3] investigated the potential of controlling the display of projected presentation materials appearing on an adjacent display in co-located synchronous presentations, recent work has considered the potential of compositing presentation visuals and presenter webcam video within a single video frame for remote presentations, thereby elevating the role of gestures beyond interface control; now prominently visible to audiences, these gestures can now also serve a communicative purpose. For instance, Saquib et al [10] allowed presenters to define and perform custom gestures that both trigger and draw attention to the animation and transformation of visual elements appearing in the video's foreground. Liao et al's RealityTalk [6] extended this style of presentation by allowing presenters to utter keywords during a presentation that would trigger the reveal of associated visual assets and overlay them wherever their hands were placed in the video frame. Most recently, Liu et al [7] demonstrated that associated visuals could be retrieved, recommended, and revealed in an on-the-fly manner during an unscripted presentation without pre-registering assets with any keywords ahead of time. Our work builds upon these multimodal presentation interfaces, albeit with a focus on presentations about data with data visualization (i.e., charts) composited in the foreground. While we do require an initial chart specification, we allow the presenter to transform the chart via voice commands in an ad-hoc and unscripted manner.

As for multimodal data visualization interfaces, Srinivasan et al have explored the potential of interleaving voice commands with touch- and pen-based interactions on tablet devices, demonstrating applications for analyzing node-link graphs [14], common statistical charts [11], and unit charts [12]. The last of which, an applica-

[*] e-mail: arjunsrinivasan@tableau.com
[†] e-mail: mbrehmer@tableau.com

tion called DataBreeze, exemplifies actions at the granularity of individual data records, including highlighting, filtering, and sorting.

Our current work considers the applicability of DataBreeze-like voice commands to presentation use cases, replacing pen and touch gestures with mid-air hand gestures visible to a webcam. In particular, we add this functionality to Hall et al's presentation environment [5], one that composites semi-transparent and dynamic charts over live webcam video, which was in turn inspired by Hans Rosling's short documentary films about public health (e.g., [9]).

2 Demonstration

Our three-minute demonstration presentation (illustrated in Fig. 1 and provided as a supplemental video) illustrates the proposed multimodal experience for presenting data to remote audiences. This presentation centers around a dataset of ∼500 American colleges containing both categorical (e.g., Region, Type) and quantitative (e.g., SAT Average, Average Cost) attributes.

To initially orient audiences to the data, the presenter begins by showing a pre-specified unit column chart, where each college appears as one mark and is grouped according to its geographic region. In explaining the distribution of colleges, the presenter points their finger at different columns, highlighting the number of colleges in each region (Fig. 1: top left). Next, the presenter indicates that the Mid-Atlantic and Southeast regions have the most number of colleges, and he presses a button on his remote clicker (or a key on his keyboard) to trigger voice input while continuing his monologue, stating that *"we can see that Southeast and Mid Atlantic have a higher number of schools than the other regions, so let's go ahead and drill down into the Mid Atlantic and Southeast regions to learn more about the schools here."* After uttering this, colleges from regions aside from Mid-Atlantic and Southeast are removed from the unit chart (Fig. 1: top left → center).

Moving on to focus on the differences between college types, the presenter states *"I'm a little curious to see what the distribution of different school types in these densely populated regions is, so to do that, maybe what we can do is color these points based on the school type that they have."* Following this utterance, the initially white marks turn orange or blue depending on if they represent *Public* or *Private* colleges, respectively (Fig. 1: top right). Highlighting the associated college types by pointing at the color legend, the presenter then draws attention to the fact that there are more Private schools than Public schools in the two regions (Fig. 1: bottom left).

Delving deeper into the two shortlisted regions, the presenter then moves on to focus on the competitiveness of colleges using the *Admission Rate* attribute as a measure for the same. Narrowing down to schools that have low admissions rates, the presenter says "*. . . let's go ahead and focus on schools that have an admission rate of less than 25%.*" (Fig. 1: bottom left → center). Pointing at the highlighted colleges, the presenter calls out that there are only three public schools with such competitive admission rates, whereas there are sixteen private schools.

Discussing if the type of area a college is located in could lead to more insight, the presenter then says "*...to see that, let's update this layout to group schools based on the locale that they belong to.*" (Fig. 1: bottom right) This results in the view changing to a new unit column chart with colleges being spatially grouped by their locale type (e.g., Large City, Small Suburb).

3 Discussion

We reflect on our implementation and the experience of preparing and delivering our demonstration presentation to remote audiences during live teleconference calls.

Exploring techniques for processing natural language utterances. Unlike prior systems focusing on visual analysis or visualization specification use cases, utterances used during presentation may not be directed to the system in the form of typical commands or queries [13]. Instead, as illustrated in the scenario (with examples appearing in the previous section), utterances in a presentation context are primarily for the benefit of the audience; they could be more verbose and may interject operation and data references amidst other speculative or hypothetical statements, likely containing hedge words. This raises a question from an utterance interpretation standpoint, as prior systems have predominantly adopted grammar- or machine learning-based techniques that may not comprehend such phrasings.

Designing error correction techniques. Prior systems have leveraged techniques like autocomplete and ambiguity widgets [4] to resolve system errors in analytic scenarios, which is simply an interaction between an individual analyst and a system. However, the use of such techniques (e.g., opening a dropdown to select a different attribute) could disrupt the flow of a presentation from the standpoint of the presenter and the audience alike. As errors inevitably occur when using the modalities of gesture and voice, addressing this challenge requires thinking deeply about the operations that we should and can reliably support through these modalities, as well as the design of more fluid correction and recovery techniques (e.g., a second display to preview and approve resulting visualization changes [2] or including a staging area for visual aids that is visible only to the presenter [7]).

Investigating fluid triggering techniques. In the context of voice input, distinguishing between utterances in cases that the system should process what is being said versus utterances intended solely for the benefit of the audience is a critical challenge. In our current prototype, we use an explicit trigger of a remote clicker to invoke the voice processing module. While this approach was suitable to demonstrate the proposed concept, investigating alternative triggering techniques that do not disrupt the flow of presentation is an interesting design challenge and an open area for future work.

Supporting additional multimodal interaction patterns. Operations in multimodal systems can be supported through a variety of interaction patterns including unimodal (i.e., operations can be performed via an individual modality), sequential (i.e., operations are performed through both modalities in a specific order), or simultaneous (i.e., operations are performed in multiple modalities at the same time). Our initial prototype supports unimodal interaction, and to some extent, a sequential combination of modalities. However, designing more sequential and simultaneous interactions that combine modalities in synergistic ways (e.g., pointing to specify a scope and using voice to issue actions) could help presenters more fluidly perform complex actions such as annotating or filtering within a subset of the data.

Exploring on-the-fly visualization creation. Our current prototype exemplifies performing common interaction operations (e.g., filtering, changing encodings) for a single chart. Besides exploring similar operations for other chart types, another area for future work is to consider if new charts could be created or retrieved in an on-the-fly manner, using a combination of gestures and the presenter's spoken monologue [6, 7].

4 Conclusion

We extended Hall et al. [5]'s gesture-based environment for presenting data to remote audiences by adding a speech-to-text module that captures voice input whenever the presenter presses a remote clicker. The converted text is then parsed and matched to operations (e.g., filtering, sorting, re-coloring), data attributes, and values through a keyword matching approach [11]. This initial foray into multimodal input for remote presentations about data led us to identify several aspects to consider in future research.

References

[1] R. A. Bolt. "Put-that-there" voice and gesture at the graphics interface. In *Proc. ACM Conf. Computer Graphics and Interactive Techniques*, 1980. doi: 10.1145/800250.807503

[2] M. Brehmer and R. Kosara. From jam session to recital: Synchronous communication and collaboration around data in organizations. *IEEE TVCG (Proc. VIS)*, 28(1), 2022. doi: 10.1109/TVCG.2021.3114760

[3] A. Fourney, M. Terry, and R. Mann. Gesturing in the wild: Understanding the effects and implications of gesture-based interaction for dynamic presentations. In *Proc. HCI*, 2010. doi: 10.14236/ewic/HCI2010.29

[4] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proc. ACM UIST*, 2015. doi: 10.1145/2807442.2807478

[5] B. D. Hall, L. Bartram, and M. Brehmer. Augmented chironomia for presenting data to remote audiences. In *Proc. ACM UIST*, 2022. doi: 10.1145/3526113.3545614

[6] J. Liao, A. Karim, S. S. Jadon, R. H. Kazi, and R. Suzuki. RealityTalk: Real-time speech-driven augmented presentation for ar live storytelling. In *Proc. ACM UIST*, 2022. doi: 10.1145/3526113.3545702

[7] X. B. Liu, V. Kirilyuk, X. Yuan, A. Olwal, P. Chi, X. A. Chen, and R. Du. Visual captions: Augmenting verbal communication with on-the-fly visuals. In *Proc. ACM CHI*, 2023. doi: 10.1145/3544548.3581566

[8] C. Noessel and N. Shedroff. *Make It So: Interaction Design Lessons from Science Fiction*. Rosenfeld Media., 2012.

[9] H. Rosling. 200 Countries, 200 Years, 4 Minutes, 2010. BBC Four. https://youtu.be/jbkSRLYSojo.

[10] N. Saquib, R. H. Kazi, L.-Y. Wei, and W. Li. Interactive body-driven graphics for augmented video performance. In *Proc. ACM CHI*, 2019. doi: 10.1145/3290605.3300852

[11] A. Srinivasan, B. Lee, N. H. Riche, S. M. Drucker, and K. Hinckley. InChorus: Designing consistent multimodal interactions for data visualization on tablet devices. In *Proc. ACM CHI*, 2020. doi: 10.1145/3313831.3376782

[12] A. Srinivasan, B. Lee, and J. Stasko. Interweaving multimodal interaction with flexible unit visualizations for data exploration. *IEEE TVCG*, 27(8), 2020. doi: 10.1109/TVCG.2020.2978050

[13] A. Srinivasan, N. Nyapathy, B. Lee, S. M. Drucker, and J. Stasko. Collecting and characterizing natural language utterances for specifying data visualizations. In *Proc. ACM CHI*, 2021. doi: 10.1145/3411764.3445400

[14] A. Srinivasan and J. Stasko. Orko: Facilitating multimodal interaction for visual exploration and analysis of networks. *IEEE TVCG (Proc. InfoVis)*, 24(1), 2017. doi: 10.1109/TVCG.2017.2745219