Research Statement

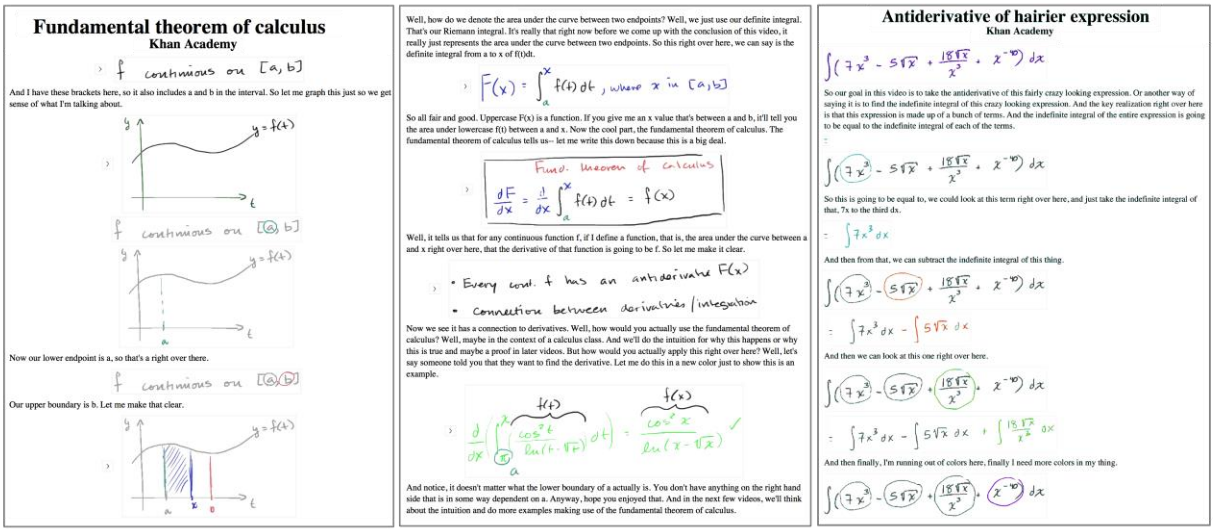
Hijung (Valentina) Shin
http://people.csail.mit.edu/hishin

My goal is to design effective tools to make the creation and manipulation of audiovisual media, such as audio narratives, videos, and animations as natural as writing or sketching. We are witnessing a proliferation of audiovisual media assisted by new types of technologies and communication platforms: podcasts, online lectures, GoPro videos, 360-degree videos and so on. However, major challenges remain from both the consumer and the producers' point of view: simply navigating audiovisual media to search and collect information is often time-consuming, authoring compelling audiovisual media is still a long and tedious task that requires skilled professionals, and collaborating on projects online or offline is unwieldy with existing tools.

The richness and complexity of information in these media are a source of attraction, but without effective tools, they can be obstacles to efficiency. For one, these media blend several modalities such as audio, images, and text. Furthermore, these modalities are interrelated within an extra dimension, time. The abundance of concurrent information makes audiovisual media difficult to digest and manipulate in a meaningful way. In my research, I create intelligent software systems that turn these challenges into opportunities. These interfaces explore approaches to effectively exploit the multidimensional and multimodal data in audiovisual media: 1) harnessing and rearranging spatio-temporal structure, 2) promoting synergy between different modalities, and 3) promoting synergy between automatic algorithms and direct user manipulation. My research combines work on interface, visualization, algorithms and data structures.

Research Themes

Harnessing and rearranging spatio-temporal structure: Audio and visual components in videos are tightly related to each other in space and time. By taking advantage of this relationship, which is different for each media, I have shown how to infer a structure and further repurpose this structure to facilitate specific tasks. My SIGGRAPH Asia work on *Visual Transcripts* illustrates this methodology [1]. *Visual Transcripts* is a novel navigation interface that automatically transforms blackboard-style lecture videos into interactive lecture notes, interleaving a set of figures with hierarchically organized paragraphs of text (Figure 1).



Fundamental theorem of calculus
Khan Academy

› f continuous on $[a,b]$

And I have these brackets here, so it also includes a and b in the interval. So let me graph this just so we get sense of what I'm talking about.

f continuous on $[a,b]$

Now our lower endpoint is a, so that's a right over there.

f continuous on $[a,b]$

Our upper boundary is b. Let me make that clear.

Well, how do we denote the area under the curve between two endpoints? Well, we just use our definite integral. That's our Riemann integral. It's really that right now before we come up with the conclusion of this video, it really just represents the area under the curve between two endpoints. So this right over here, we can say is the definite integral from a to x of f(t)dt.

› $F(x) = \int_a^x f(t)\,dt$, where x in $[a,b]$

So all fair and good. Uppercase F(x) is a function. If you give me an x value that's between a and b, it'll tell you the area under lowercase f(t) between a and x. Now the cool part, the fundamental theorem of calculus. The fundamental theorem of calculus tells us-- let me write this down because this is a big deal.

› Fund. theorem of calculus
$\frac{dF}{dx} = \frac{d}{dx}\int_a^x f(t)\,dt = f(x)$

Well, it tells us that for any continuous function f, if I define a function, that is, the area under the curve between a and x right over here, that the derivative of that function is going to be f. So let me make it clear.

› • Every cont. f has an antiderivative F(x)
• Connection between derivatives / integration

Now we see it has a connection to derivatives. Well, how would you actually use the fundamental theorem of calculus? Well, maybe in the context of a calculus class. And we'll do the intuition for why this happens or why this is true and maybe a proof in later videos. But how would you actually apply this right over here? Well, let's say someone told you that they want to find the derivative. Let me do this in a new color just to show this is an example.

› $\frac{d}{dx}\left(\int_{\pi}^{x} \overbrace{\frac{\cos^2 t}{\ln(t-\sqrt{t})}}^{f(t)}\,dt\right) = \overbrace{\frac{\cos^2 x}{\ln(x-\sqrt{x})}}^{f(x)}$ ✓

a

And notice, it doesn't matter what the lower boundary of a actually is. You don't have anything on the right hand side that is in some way dependent on a. Anyway, hope you enjoyed that. And in the next few videos, we'll think about the intuition and do more examples making use of the fundamental theorem of calculus.

Antiderivative of hairier expression
Khan Academy

$\int\left(7x^3 - 5\sqrt{x} + \frac{18\sqrt{x}}{x^3} + x^{-40}\right)dx$

So our goal in this video is to take the antiderivative of this fairly crazy looking expression. Or another way of saying it is to find the indefinite integral of this crazy looking expression. And the key realization right over here is that this expression is made up of a bunch of terms. And the indefinite integral of the entire expression is going to be equal to the indefinite integral of each of the terms.

$\int\left(7x^3 - 5\sqrt{x} + \frac{18\sqrt{x}}{x^3} + x^{-40}\right)dx$

So this is going to be equal to, we could look at this term right over here, and just take the indefinite integral of that, 7x to the third dx.

$= \int 7x^3\,dx$

And then from that, we can subtract the indefinite integral of this thing.

$\int\left(7x^3 - 5\sqrt{x} + \frac{18\sqrt{x}}{x^3} + x^{-40}\right)dx$

$= \int 7x^3\,dx - \int 5\sqrt{x}\,dx$

And then we can look at this one right over here.

$\int\left(7x^3 - 5\sqrt{x} + \frac{18\sqrt{x}}{x^3} + x^{-40}\right)dx$

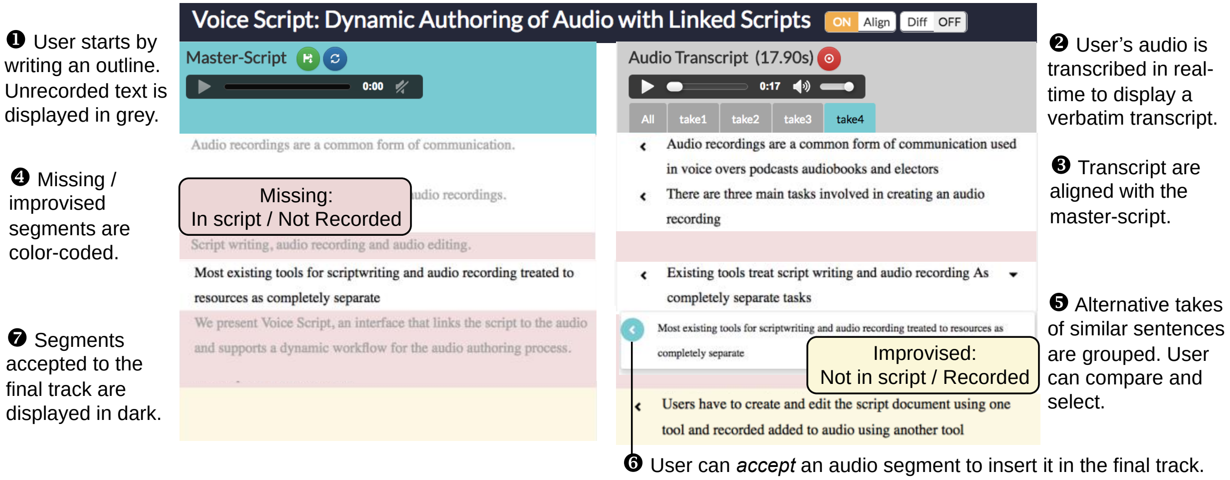
$= \int 7x^3\,dx - \int 5\sqrt{x}\,dx + \int \frac{18\sqrt{x}}{x^3}\,dx$

And then finally, I'm running out of colors here, finally I need more colors in my thing.

$\int\left(7x^3 - 5\sqrt{x} + \frac{18\sqrt{x}}{x^3} + x^{-40}\right)dx$

Figure 1. Examples of *Visual Transcripts* from two different video lectures. Our interface automatically transforms lecture videos into an interactive lecture note, interleaving figures drawn on the board with verbal explanations. Clicking on a figure or text plays the video (inline by animating the figures).

First, I convert the continuous drawings on the board into a discrete set of key illustrations by implementing an automatic segmentation algorithm that takes into account the spatio-temporal arrangement of these drawings. In addition, by analyzing the temporal correspondence between audio and visuals, I automatically classify sentences into two categories: depictive sentences that verbalize what the visuals depict, and explanatory sentences that provide additional information not directly represented in the visuals. Then, to improve skimming and searching, *Visual Transcripts* converts the video from the original temporal layout to a spatial layout, juxtaposing the figures with corresponding explanatory sentences and hiding the depictive sentences for interactive display. I conducted comparative user studies against state-of-the art systems and demonstrate that users prefer our interface for learning and that our interface is effective in helping them browse or search through lecture videos. As this example illustrates, each media has characteristic spatio-temporal structures (e.g., visuals consist of discrete elements, audio and visuals complement each other), and likewise each task can benefit from different spatio-temporal arrangements (e.g., step-by-step delineation of key visual frames, placing related visuals and audio side-by-side). By leveraging innovative insights about both aspects, I aim to develop novel practical tools that improve how users interact with media.



Audio recordings are a common form of communication.

…udio recordings.

Script writing, audio recording and audio editing.

Most existing tools for scriptwriting and audio recording treated to resources as completely separate

We present Voice Script, an interface that links the script to the audio and supports a dynamic workflow for the audio authoring process.

Audio recordings are a common form of communication used in voice overs podcasts audiobooks and electors

There are three main tasks involved in creating an audio recording

Existing tools treat script writing and audio recording As completely separate tasks

Most existing tools for scriptwriting and audio recording treated to resources as completely separate

Users have to create and edit the script document using one tool and recorded added to audio using another tool

❶ User starts by writing an outline. Unrecorded text is displayed in grey.

❷ User's audio is transcribed in real-time to display a verbatim transcript.

❸ Transcript are aligned with the master-script.

❹ Missing / improvised segments are color-coded.

❺ Alternative takes of similar sentences are grouped. User can compare and select.

❻ User can *accept* an audio segment to insert it in the final track.

❼ Segments accepted to the final track are displayed in dark.

Figure 2. *Voice Script*, an interface for authoring speech recordings. (Left) The master-script document unifies script and audio. (Right) Individual audio recordings are transcribed using automatic speech recognition. Our interface automatically segments and aligns the master-script and audio transcripts to facilitate merging and syncing of the script with the audio transcripts.

Promoting synergy between different modalities: Different modalities that compose audiovisual media have different roles and potential to assist user tasks. For example, static illustrations and text are easier to skim, search or edit, whereas animated videos and audio are better at conveying sequential information or tonal nuances. *Visual Transcripts* couple these modalities so that static representations function as an interactive transcript: clicking on a figure or text replays the corresponding part of the video by animating the figures inline. As another example, in my recent work presented at UIST [2], I exploit the link between speech and text to build *Voice Script*, an effective interface for authoring speech recordings (Figure 2). Based on formative interviews with producers, I identified that an essential but tedious part of the authoring process is to go back and forth between the text script and the audio recordings. Traditional tools are suboptimal for this process because they deal exclusively with only one of the modalities such that users typically have to write the script in one application like a word processor, and record and edit the audio in a separate waveform-editing tool. I designed *Voice Script*, which addresses this problem by a new data structure that unifies the script and audio in a single text document. I use automatic speech recognition to transcribe speech into text, and treat both script edits and audio recordings as updates to the same document. Users can work seamlessly on recording speech, including improvisation, and browsing and editing via text. As these examples illustrate, to design optimal interfaces that create a synergy between multiple modalities it is important to rethink data structures and representations for media.

Promoting synergy between automatic algorithms and direct user manipulation: Hybrid approaches that combine automatic algorithms with direct manipulation have been proposed and successfully implemented in many application domains, including authoring of 3D models[3,4], animations [5], and illustrations [6,7]. Automatic algorithms can greatly simplify tedious tasks and allow users to focus on the creative part of the workflow. For instance, one of the most time-consuming tasks in authoring speech recordings is, when there are multiple recordings, to align them and cut and merge them into a final track. In *Voice Script*, I implemented an algorithm to automatically segment the written and recorded texts and align them even when the texts do not match exactly (e.g., due to improvisation) or when they have different natural boundaries such as punctuation or pauses. This allowed users to easily merge sections of recordings as in a text differencing and merging interface, as well as keep track of improvised or unrecorded parts of the script.

In a recent, ongoing project, I am applying the hybrid approach to develop *Live Presentations*, a slide presentation interface that integrates live inking. Writing in real-time makes presentations more engaging [8], but it is difficult to write neatly on the fly. Presenters often write unevenly, or run out of space to write. Sometimes they have to adapt or modify the writing after the fact. I approach this issue by delegating the problem of real-time layout management and beautification through automatic algorithms. Since the system takes care of the dynamic layout of slide contents, the presenter can focus on content delivery. Unlike previous work that treat ink as a separate layer on top of slides, I support interaction between the inked contents and the prepared slide contents.

To apply hybrid approaches to other types of media and tasks, I want to explore effective algorithms that relieve users from painstaking technical tasks as well as data structures that enable smooth transitions between automatically generated outputs and interactive manipulation.

Research Agenda

In my research so far, I made some progress towards designing better tools for manipulating audiovisual media in several specific domains. To make these tools even more effective and to empower users even further, supporting collaboration is essential. I envision a future where novices, experts and distributed online communities will collaborate directly in real-time to author compelling audio recordings, videos, animations, and even new types of media. Achieving this vision will require progress in many areas. Continuing to improve usability for single users and exploring new functionalities to directly support collaboration are both critical. Adobe is an ideal place to pursue this research with all its top-class content creation tools. Especially with the move to Creative Cloud, which makes an important step to support collaboration, it is important to develop the right tools and techniques that can make collaboration even smoother.

In what follows, I put forth several sub-challenges that I am especially excited to pursue through my research, applying the methodologies outlined above but not limited to them: 1) facilitating navigation and editing of temporal media 2) facilitating iterative design with real-time editing, and 3) supporting web-based collaboration on audiovisual media.

Facilitating structured navigation and editing of temporal media: Unlike a photo which can be viewed at once, users have to *navigate* through videos or audios to know their contents. In a single author scenario, authors can rely on their unique knowledge, for example, to navigate to a specific point in an audio that needs to be replaced or to find an alignment with an alternative recording. In a collaborative scenario, users have to navigate and edit media which they have not authored in the first place. This can be especially challenging for temporal media such as videos because of its tight coupling with time. For example, timelines in standard video browsers are suitable for a linear viewing experience, but they don't allow for easy skimming, searching or going back and forth. Editing videos or audio is even more difficult. Even basic operations such as alignment and synchronization are tedious. Exposing and rearranging the structure of the media and taking advantage of static representations can alleviate this problem. In *Visual Transcripts*, I explored a spatial representation of lecture videos to facilitate navigation. In *Voice Script*, I used a text-based representation for spoken audio, previously proposed by other researchers [9,10], and visualized the alignments spatially. In both instances, I use static representations to project the temporal data into a different, spatial layout and make their structure easier to browse and edit.

I believe similar approaches can be extended to a wide range of applications. For example, static representations of animations (similar to *Visual Transcripts* or motion illustrations [6]) could be used to facilitate post-editing, and then reverse-engineered to render edited animations. Recent advances in audio and video analysis [11,12] opened up new ways to look at audiovisual media and manipulate their structure. I want to explore new data structures and representations that take advantage of such algorithms as well as new methods to develop interfaces for a variety of tasks, including authoring animations and composing sound recordings (including but not limited to speech).

Facilitating iterative design with real-time editing: Iteration is a crucial part of any serious authoring process, including collaborative scenarios, where multiple people work together on the design and feedback process. The ability to quickly explore multiple alternatives, give and receive feedback, make or undo changes, and visualize the consequences is critical. Often times these involve combining unstructured data with structured data. For example, in *Voice Script*, I integrate improvised speech (unstructured) with script texts (structured), and in *Live Presentations*, I provide interaction between on-the-fly ink (unstructured) with prepared slide contents (structured). I want to extend this approach to other types of media, for example, integrating sketching or annotations into UX design, or using verbal or vocal annotations in audio authoring. This relates to recent work such as the system for video critiquing presented by Pavel et al.[13], which incorporates spoken comments, mouse interactions and hand gestures. I will study effective low-effort interactions, such as sketching or verbal annotations, develop methods to integrate the output of these interactions with structured data, and investigate ways to support such interactions in real-time.

Supporting web-based collaboration on audiovisual media: Online collaborative editing is a growing theme across a number of domains and we are seeing a rise of new web-based editors such as OnShape for 3D modeling or Prezi for presentations. However, major challenges remain to support web-based collaboration for audiovisual media. In particular, traditional data structures for audio or video were developed for single user, single machine interfaces. By rethinking how we store and represent audiovisual data, I plan to address issues that arise with online collaboration such as conflict resolution, visualization of editing history, and synchronous editing. The ultimate aim is to make online editing of audio and video as smooth as using Google Docs, and to enable new types of collaborative media such as wiki-podcasts or wiki-videos.

In sum, I want to make creating, manipulating and consuming audiovisual media easy and enjoyable. I made a first step towards this goal, by designing novel interfaces that address challenges in several specific domains. To make further steps, my research will focus especially on supporting collaboration, and expand to other types of audiovisual media.

[1] Shin, Hijung Valentina, et al. "Visual transcripts: lecture notes from blackboard-style lecture videos." ACM Transactions on Graphics (SIGGRAPH ASIA), 2015.
[2] Shin, Hijung Valentina, Wilmot Li, and Frédo Durand. “Dynamic Authoring of Audio with Linked Scripts.” Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST), 2016.
[3] Gal, Ran, et al. "iWIRES: an analyze-and-edit approach to shape manipulation." ACM Transactions on Graphics (TOG). Vol. 28. No. 3. ACM, 2009.
[4] Prévost, Romain, et al. "Make it stand: balancing shapes for 3D fabrication." ACM Transactions on Graphics (TOG) 32.4 (2013): 81.
[5] Bai, Yunfei, et al. "Artist-directed dynamics for 2D animation." ACM Transactions on Graphics (TOG) 35.4 (2016): 145.
[6] Chi, Pei-Yu Peggy, et al. "Authoring Illustrations of Human Movements by Iterative Physical Demonstration." Proceedings of the 29th Annual Symposium on User Interface Software and Technology. ACM, 2016.
[7] Xin, Jun, et al. “Energy-Brushes: Interactive Tools for Illustrating Stylized Elemental Dynamics.” Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST), 2016.
[8] Guo, Philip J., Juho Kim, and Rob Rubin. "How video production affects student engagement: An empirical study of mooc videos." Proceedings of the first ACM conference on Learning@ scale conference. ACM, 2014.
[9] Rubin, Steve, et al. "Content-based tools for editing audio stories." Proceedings of the 26th annual ACM symposium on User interface software and technology. ACM, 2013.
[10] Rubin, Steve, et al. "Capture-time feedback for recording scripted narration." Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology. ACM, 2015.
[11] Mysore, Gautham J., Paris Smaragdis, and Bhiksha Raj. "Non-negative hidden Markov modeling of audio with application to source separation." International Conference on Latent Variable Analysis and Signal Separation. Springer Berlin Heidelberg, 2010.
[12] Davis, Abe, et al. "The visual microphone: passive recovery of sound from video." (2014).
[13] Pavel, Amy, et al. "VidCrit: Video-based Asynchronous Video Review." Proceedings of the 29th Annual Symposium on User Interface Software and Technology. ACM, 2016.