CA^3: Collaborative Annotation of Audio in Academia

Ben Congleton, John Booker, Laurian Hobby, Meg Kurdziolek, Lauren Shupp, Manuel A. Pérez-Quiñones
Virginia Tech
660 McBryde Hall, Blacksburg, VA 24060
(bc, jobooker, lhobby, mkurdziolek, lshupp, perez)@vt.edu

ABSTRACT
We present a collaborative tagging tool for audio streams. We discuss two case studies using this tool: The first case study demonstrates the usefulness of simple tags as metadata. The second case study elaborates issues discovered while allowing students to tag events during a classroom lecture, and methods for aggregating and displaying the collected tags. We conclude with insights into collaborative tagging for data retrieval in content streams.

Keywords
Personal information management, note taking, audio indexing, social metadata, collaborative annotation

Categories and Subject Descriptors
H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems– *Audio input/output*

1. INTRODUCTION
We live in the Information Age. We encounter more information then we can retain or process on a daily basis. We even struggle to retain information in controlled environments, such as classrooms [13] and meetings. To mitigate this lack of innate retention we offload retention to video cameras, personal audio recorders, and personal notes. Recording devices help to create comprehensive records of an event, but lack easy methods of re-finding information in the stored content streams. Note taking makes it easier to revisit important ideas, but affects the ability for the note taker to attend to the information itself.

Use of a tape recorder or other passive capture device creates a complete record of an event, and has a low up front cost; however, searching and processing the complete record is often unintuitive. Revisiting captured events is practical for small amounts of data, but in today's information society the time required for an individual to annotate and search large content streams is prohibitive.

We have developed a tool that allows an audio stream to be collaboratively annotated with tags. For example, in a classroom, students have the ability to tag moments in time as "homework". In this scenario a student who missed class could view all of the instants tagged homework by his classmates, and quickly jump to the segment of the audio with the most homework tags. By leveraging the collective annotation of segments in an audio stream, we can provide aggregated metadata to users to aid retrieving information within an audio stream.

Our tool consists of three main parts, the annotation tool, the content stream viewer, and the backend database.. The annotation tool, dubbed the tagger, creates the metadata. This data is then stored and processed by the back end database, where it is later accessed by the content stream viewer.

In order to evaluate our tool, we focused on the domain of classroom note taking. This environment provided the necessary information overload, and study participants who desired to retain large amounts of information for later recall on exams.

We performed two case studies: one for each part of the tool. In the first, we examine how the content stream viewer can improve performance on academic assignments, while observing user strategies and conducting a basic usability study. We found users to be receptive to the tool, and observed how the fidelity, quality of the content stream, and in-stream search-ability, the ability to search for content within a stream, affected user information seeking behavior.

The second study looked at the feasibility of using collaborative tagging as a method for metadata creation in a classroom setting. We found that while collaborative tagging will be affected by social loafing, even small populations of students can produce useful metadata.

We conclude with a discussion of the effectives of collaborative tagging, and the future work identified as a result of the case studies.

2. RELATED WORK
The need to capture of personal experiences for future use has existed for some time. Information overload creates situations where users can not be expected to both attend to the incoming information, and also make notes about it for future reference [1, 11]. Many systems for assisting memory exist, including note-taking tools [1], and personal memory aids [20]. While these tools employ many methods of searching through audio streams, they lack the ability to quickly filter noise from what is sought.

The college classroom provides an excellent representation of the problems with excessive incoming information, as well as the current approaches to handling it. The problem of attending to the information and processing it while at the same time recording it is quite evident in the classroom, as merely copying down the slides is more than enough to overwhelm some

individuals [18]. Providing prior access to the slides is not enough. Although students no longer have to copy the slides down, they still must mark up the slides with the professors' annotations, while processing the professor's words [13].

If slides are not provided, most of the note taking work focuses on copying the slide, instead of personalizing the notes [19]. In a study of note taking behavior, Wilcox found that individuals will often mark their notes with "todo" or "important" tags. These tags became very important for review later on, and the todo items were frequently the sole items revisited [21]. These tags are the "personal" information that needs to be infused into the data stream to make it meaningful to the individual.

While the slides and the markup created during class are important, what the professor says during a class often contains the most information, and a method to capture this data is desirable. If a useful record of the verbal content from a lecture can be relied upon, students become free to take meaningful notes. Audio is trivial to record, but finding specific content in the raw audio stream is not easy. Minneman, et al. describes several types of metadata that can be used to annotate an experience [16]. The first is intentional tags (such as explicitly marking audio as important), side-effect (events such as page turns, slide changes), derived (processing of audio to identify the speaker, etc), and finally post hoc indices (intentional annotations that occur upon review instead of during the capture are all needed. The first of these, the intentional tags, are similar to the labels used during note taking described by Wilcox [21].

Moran et al. describe using annotations features, but only during post-hoc analysis [17]. Instead of "salvaging" the audio information after the fact, there is a need to create these annotations during the capture of the information. Some of this can be done with automatic real time processing, such as segmentation of audio portions based on other time stamped information [8]. Automatic processing can generate a variety of data [7], but a robust interface for creating new metadata in real time is needed. Mere analysis of the content stream without integrating the personal information from the users is only so useful [9]. Any effective index into a large body of information must contain some sort of personal contextual annotation [12].

Systems exist to automatically segment [7, 15] and classify [5, 14] audio. Researchers have successfully segmented speech by speaker [14, 15], and using pauses and other aspects of speech. Non speech audio can be classified into music, environmental sounds, and silence [5], and classifiers exist that can use smaller audio events such as a cars engine, and road noise, to semantically classify episodes of audio, for example a car chase in an action movie [5]. These systems provide valuable metadata, but do not provide the granularity necessary to help prioritize segments for retrieval in a lecture situation. In a lecture segmentation by change in speaker and pauses in presentation provides useful information, but this metadata does not provide insight into the topic or collective importance of the segment.

One method of generating metadata for a content stream is collaborative annotation [4]. When multiple users annotate the same content stream, users benefit through the resulting annotated content stream. Bargeron et al describe a study of a tool for web based collaborative annotation of audio and video streams; however, they do not address how the annotations can serve as an index into a stream for refinding, and focus mostly on post hoc annotation. Appan et al [3] describe interfaces for collaborative annotation, but focus on static content as opposed to stream content such as an audio or video stream.

Collaborative tagging is "the process by which many users add metadata in the form of keywords to shared content [10]". Collaborative tagging is a form of collaborative annotation, regulated to the creation of keywords rather than complete phrases. These keywords are often ill defined and lead to semantic problems, as not all users follow the same lexicon when creating tags. When annotating data in real-time collaborative tagging can compensate for the tendency of individuals to miss important content during class [11] by allowing these individuals access to the collectively annotated content stream. In short, collaborative tagging is generally considered a more keyword oriented form of collaborative annotation.

3. CA^3

The goal of our prototype was to provide a framework for efficient generation, collection, and usage of collaborative tags that describe audio content. We chose to focus on audio from classroom situations to limit the scope of our development effort and focus our tool's performance on standard repeatable memory retrieval scenario; however, our initial findings can be generalized to other problem domains.

The prototype system consisted of a web based content tagging system, for generation and collection of tags, a highly interactive content stream viewer, that allowed students were use the tags to refind desired information, and a backend data management tool, to maintain the database of collaborative generated tags.

3.1 Web Based Tagging System

Taggers are the most unique aspect of our development effort. A Tagger allows users to mark specific points in time with a keyword. For example, in a lecture a student might tag certain aspects of the lecture as "Important", and future class projects as "Homework." The tag names were deliberately kept to a small number both so that the user was not overwhelmed, and so that students with difference vocabularies would not use different words to describe the same event. The tag, user, and timestamp of the tag are stored in database for future use. The collection of time-stamped tags was then sent to a server to be aggregated. By combining the data in a shared repository, we were able to provide a comprehensive index into the audio, without relying on each user to provide a complete record of the event.

The tagger application itself was a lightweight web program that could run on mobile devices as easily as laptops. To make the tagger application platform independent it was built using PHP and a MySQL backend, and browser detection made it possible for all client platforms to use the same tagging URL.

3.2 Content Stream Viewer

To make use of the collaborative tagging efforts of the class, we created a tool that used those tags as an index into multiple content streams. The content stream viewer allowed students to view multiple streams concurrently (Figure 1) such as synchronized transcripts, video, and audio annotated with user generated tags. The viewer also allowed students to search for specific text in the transcript, navigate the content streams by the indexes created by the tags, and playback the streams as needed.

We chose to develop our application using Macromedia Flash Professional 8 to take advantage of Flash's multimedia capabilities, and to accelerate development of the initial prototype. Making our viewer available on the internet ensures universal access both on portable devices, and in more traditional desktop settings, which has been found to be helpful in other such devices. [2] [6].

The main feature of the playback interface is a timeline view that indicates the current playing position over top of a histogram representing the tags. Time was on the x-axis, and number of tags per bucket on the y-axis. By showing a histogram, instead of raw tags or discrete bookmarks, we are aggregating the collective opinion of the tagging population (the students). This compensates for times when an event is missed completely by some users (others will have tagged it) or when an event is tagged late (there will often be at least one early tagger as well). Users can also identify the tags they created personally, indicated by a star above the histogram bar. Clicking on a section of the histogram will seek to the time in the audio file represented by the cursor's x position.

Another view provided by our system is time-synched text. This was initially implemented with manual transcriptions, but could easily accept any text with timestamps, such as speech-to-text output, once those systems fully mature. Clicking on a section of the text will seek to the audio section corresponding to that particular caption.

4. BACKEND DATABASE

The third component of the CA^3 suite is the backend data management tool. The primary function of the backend system is to ensure that incoming tags and incoming content streams from different sources have synchronized timestamps. Additionally, to support the collaborative tagging aspect of our tool the backend system can consolidate tags from multiple users. Lastly, a web application over top of the database provides a front end to browse the course material.

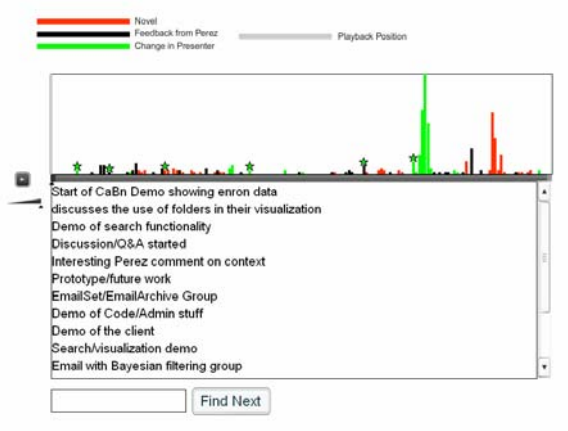

Figure 1 The content stream viewer showing (from top) the legend, histogram, audio progress bar, and synchronized text.

5. CASE STUDY: TNT

5.1 Introduction

Students are expected to remember volumes of information, with many specific details. They take notes, but they are often unorganized and rife with errors. An audio record of the class would solve problems associated with missing important information or making errors during transcription, but organization would still be an issue. Before we developed a method for collaborative tagging, it was important to evaluate if the type of metadata we planned to collect would be useful in the academic domain. This first case study, called Taking Notes Together, examined the efficiency of our simple tags.

5.2 Methods

We designed a content stream viewer for replay of tagged audio and transcripts. We recorded a typical lecture, waited two weeks, and then asked student volunteers to complete a quiz. Students were randomly assigned to use the Content Stream Viewer, their traditional notes, or both.

After completing the quiz, we asked students to complete a short questionnaire to evaluate the usability of the TNT tool, and characterize the navigation strategies employed by students using the tool. The questionnaire consisted of 10 likert style general usability questions, 8 questions related to tool navigation strategies, and 3 questions gauging student interest in collaborative tagging of audio streams.

5.3 Results

5.3.1 Quiz Results

Students performed significantly better on the quiz when solely using our content stream viewer (**Table 1** Incorrect answers by group. TNT decreased errors significantly.).

Table 1 Incorrect answers by group. TNT decreased errors significantly.	
Condition	Total # of incorrect Answers
Own Notes	50
TNT Only	4
TNT+ NOTES	12

5.3.2 Usability study

We found very few differences in the responses from students using only TNT and students using TNT and their own notes. Not surprisingly, students who only used TNT were more likely to find the audio useful (3.75/5) when completing the quiz than students who used TNT and their notes (2.44/5) ($p = 0.006$). Additionally, students who only used TNT (4.28/5) were more likely to desire the ability to tag content from the lecture during class than students using TNT and their notes (3.77/5) ($p = 0.092$).

Both groups of students primarily used the find box (11/16), but almost equally choose the tagged graph (6/16) and transcripts (7/16) as their secondary navigation technique (Table 2).

In general students found TNT to be useful, interesting, stimulating, and easy to use. They felt the navigation and data organization was intuitive, and agreed that they would use TNT to study if it was available for their classes.

Table 2 Pairings of 1st/2nd most used features

Most useful/Second useful	Votes
Search/Transcript	6
Search/Graph	5
Transcript/Search	2
Transcript/Transcript	1
Transcript/Audio	1
Transcript/Graph	1

5.4 Discussion

Both the results from the quiz and the usability results raise interesting questions. It was interesting that the group of students without their notes performed better than those with their notes. This emphasizes the fact that more information does not guarantee an improvement in performance. We suspect that users with their own notes consulted them first, and were less willing to switch to using CA^3 as a backup. The group with no notes of their own would have started with CA^3, which was more accurate (though possibly slower to use). This motivates the need for prioritization of content streams and interface to content streams based on user task and profile.

The results from the usability study were not surprising, however it is important to note that the majority of the tested students used text search to find content in a content stream.

Future work should make use of less reliable computer generated transcripts to more accurately evaluate the usefulness of our tool as an automated memory aid. A study must be done to compare student information seeking behavior with transcripts of varying fidelity. The inherent inequality of certain content streams is obvious; however, future work should elaborate upon how content stream fidelity, and in-stream search-ability affect information seeking behavior.

6. CASE STUDY: TAGGING

6.1 Introduction

With the efficiency of our tagging methodology demonstrated sufficiently, we next investigated methods of collaboratively generating and collecting those tags. This second case study involved using the tagger application during a class to determine what type of tags might be typically generated, as well as ways of viewing that data to best get an index into the audio stream.

6.2 Methods

A graduate level computer science class was asked to use our tagging tools to collaboratively tag events during a series of student presentations. 14 of approximately 30 students in the classroom participated in tagging events using a web-based Tagger. Students were given the option to use either their own laptops or a PDA we provided. We asked students to tag "changes in presenter", "feedback from the professor", and "novel" contributions. The first two tags were objective, and could be tied to specific events in the audio, while the "novel" tag was intentionally ambiguous to provide each student with the ability to define their own meaning for the tag.

6.3 Results

The primary goal of this study was to establish the feasibility of collaborative tagging as a method of adding metadata to a content stream. The 14 students produced a total of 331 tags during a 90 minute class period. A single user was responsible for 223 of these tags, while another outlier created 40 tags. To get a more conservative minimum expectation for tag creation, we removed the two outliers so that the average user created 4 tags and the median user created 1 tag.

To provide an aggregated view of the collected tags, a histogram view was provided as was seen in Figure 1. The histogram itself, only for the changes in presenter tags, can be seen in Figure 2. Finally, a modified histogram showing a weighted score for each tag instead of the total number of tags can be seen in Figure 3. Here, users had their tags adjusted inversely proportional to how many tags user created so that any tag for that user was equal to 1/(# of total tags).

6.4 Discussion

The idea behind the collaborative tagging effort is to help each user get a feel for the collective knowledge of the group. In the un-weighted histogram (Figure 2) we can readily see only one major spike in tag activity, indicating a change in presenter at approximately 2:55PM. This was the result of a single user who, presumably very confidant that an important change in presenters was taking place, generating a large number of tags. This presents a problem, as that spike dwarfs other local maxima. However, it is not appropriate to dismiss those tags as outliers, as they represent valuable information about the audio.

The approach we took was to recognize other outliers as significant sources of information as well, namely those outliers that generated very few tags. The weighted histogram approach lessens the impact of the individual tags for an overactive tagger, while strengthening the tags of a cautious tagger.

In the weighted histogram (Figure 3) we can clearly see 3 presenter changes. The first two tag maxima are the results of two users who only made one tag a piece, and therefore their tags were made to be very influential. The third maximum is still the result of the overactive tagger, but it has been scaled back.

The results show us two important aspects about collaborative tagging. Firstly, even a small number of users who are not very active taggers can generate enough metadata between them to provide insight into the audio stream. Secondly, outliers must be appropriately identified and treated with care. It is possible that future versions of the software should throttle overactive or obviously malicious users, but our weighted histogram approach revealed two extra changes in presenters, without discarding any information.

7. CONCLUSION

Consolidation of content streams creates more comprehensive information resources; however, these more complete information resources do not necessarily improve user performance. The first case highlights how additional information, in this case notes, does not necessarily improve user performance.

It is important to consider the following factors when interacting with a content stream or a series of linked content streams:

- *fidelity* – the quality of the stream as ranked by the information seeker's needs
- *in-stream search-ability* – the ability for users to both navigate and find content in a content stream
- *stream find-ability* – the ability for information seekers to find the stream while searching.

Our initial tagging study shows that users do not create large amounts of metadata individually. Contrary to being a discouraging result, this is a further indication of the need to gather metadata from multiple sources. Other users, environmental aspects (slide changes, etc), and post processing of audio can all integrate to provide an effect metadata index into the audio.

One of the main issues for note taking or other labeling of content streams is the distraction from the stream itself that occurs when the labeling takes place. It is difficult to both create the metadata and to process the information contained in the stream simultaneously. Often labels will be missed, or content not understand as a result. By creating a simple and quick method to attach metadata to a content stream, we reduce the cost of labeling. We reduce the cost of missing a label by distributing the labeling task across all users performing annotation.

8. FUTURE WORK

Our work opens far more doors than it closes. We note the importance of metadata for in-stream search-ability in content streams. We must explore how content streams of various fidelity and in-stream search-ability affect user information seeking behavior. For example, we could provide users with a lower-fidelity transcription, and see how this affects their search strategy. We could also experiment with various fidelities of audio content, and annotated tags. Our work does not even begin to address content stream find-ability, information retrieval literature is the best place to start when approaching this problem, though it may be interesting to explore how better in-stream search-ability improves stream find-ability. Additional metadata content could include timestamped notes (handwritten with a tablet or typed) as well as other environmental data such as slide changes, speaker changes, etc.

We have only completed a preliminary study of the collaborative tagging approach to metadata creation. We need to combine collaborative tagging, with a study similar to that performed by TNT. Such a study provides us the means to test the ability for collaborative tagging to improve information retrieval, and serve as an index into an audio stream. It also will provide us a better understanding of how users can use collaborative tagging of content streams in the classroom.

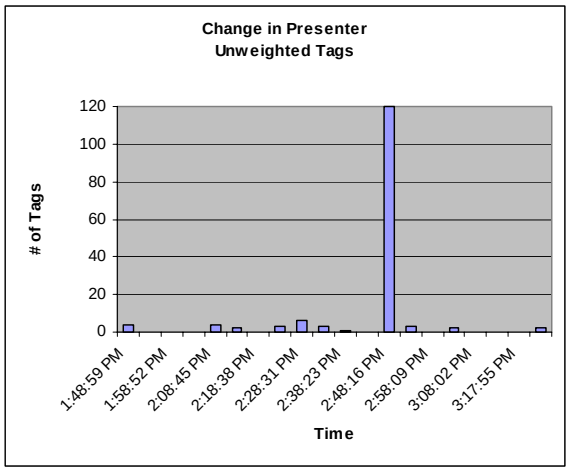

Figure 2 Unweighted Change in presenter histogram

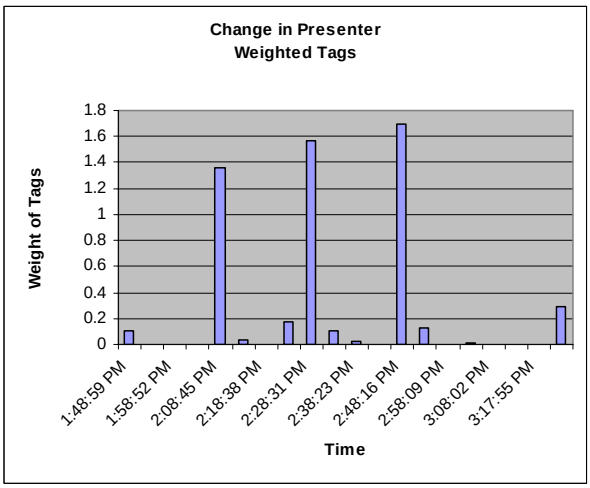

Figure 3 Weighted Change in presenter tags

Future studies will take collaborative tagging of content streams out of the classroom and explore their benefits in less controlled environments.

9. ACKNOWLEDGEMENTS

We thank Dr. Edward Clayton, who was gracious enough to allow us access to his BIT class so that we could evaluate our tool. We also thank our personal information class for their ongoing feedback of our project, as well as their participation in the case studies.

10. REFERENCES

1. Abowd, G.D., *Classroom 2000: an experiment with the instrumentation of a living educational environment.* IBM Systems Journal, 1999. **38**(4): p. 508-530.
2. Abowd, G.D., et al., *Teaching and learning as multimedia authoring: the classroom 2000 project.* Proceedings of the fourth ACM international conference on Multimedia, 1997: p. 187–198.
3. Appan, P., et al., *Interfaces for networked media exploration and collaborative annotation*, in *Proceedings of the 10th international conference on Intelligent user interfaces*. 2005, ACM Press: San Diego, California, USA.
4. Bargeron, D., et al., *Annotations for streaming video on the Web: system design and usage studies*, in *Proceeding of the eighth international conference on World Wide Web*. 1999, Elsevier North-Holland, Inc.: Toronto, Canada.
5. Cheng, W.-H., W.-T. Chu, and J.-L. Wu, *Semantic context detection based on hierarchical audio models*, in *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*. 2003, ACM Press: Berkeley, California.
6. Chiu, P., et al., *NoteLook: taking notes in meetings with digi-tal video and ink.* Proceedings of the seventh ACM international conference on Multimedia (Part 1), 1999: p. 149–158.
7. Ellis, D.P., *Features for segmenting and classifying long-duration recordings of "personal" audio.* In Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing SAPA-04 Jeju, Korea, 2004.
8. Ellis, D.P. and K. Lee, *Minimal-impact audio-based personal archives.* In Proceedings of the the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences, 2004: p. 39-47.
9. Foote, J., *An overview of audio information retrieval.* Multimedia Systems 7, Springer-Verlag, 1999: p. 2-10.
10. Golder, S. and B. Huberman, *The Structure of Collaborative Tagging Systems.* 2005.
11. Grabe, M. and K. Christopherson, *Evaluating the advantages and disadvantages of providing lecture notes: The role of internet technology as a delivery system and research tool.* The Internet and Higher Education, 2005. **8**(4): p. 291-298.
12. Kern, N., et al., *Wearable sensing to annotate meeting recordings.* Personal Ubiquitous Comput., 2003. **7**(5): p. 263-274.
13. Lienhard, J. and T. Lauer. *Multi-layer recording as a new concept of combining lecture recording and students' handwritten notes*. in *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*. 2002: ACM.
14. Lu, L. and H.-J. Zhang, *Speaker change detection and tracking in real-time news broadcasting analysis*, in *Proceedings of the tenth ACM international conference on Multimedia*. 2002, ACM Press: Juan-les-Pins, France.
15. Lu, L. and H. Zhang, *Content analysis for audio classification and segmentation.* IEEE Transactions on Speech and Audio Processing, 2002. **10**(7).
16. Minneman, S., et al. *A confederation of tools for capturing and accessing collaborative activity*. in *MULTIMEDIA '95: Proceedings of the third ACM international conference on Multimedia*. 1995: ACM.
17. Moran, T., et al. *"I'll Get That Off the Audio": A Case Study of Salvaging Multimedia Meeting Records*. in *CHI*. 1997.
18. Stifelman, L., B. Arons, and C. Schmandt. *The audio notebook: paper and pen interaction with structured speech*. in *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*. 2001: ACM.
19. Truong, K., G. Abowd, and J. Brotherton. *Personalizing the capture of public experiences*. in *UIST '99: Proceedings of the 12th annual ACM symposium on User interface software and technology*. 1999: ACM.
20. Vemuri, S., et al., *An audio-based personal memory aid.* Proceedings of Ubicomp 2004: Ubiquitous Computing., 2004: p. 400-417.
21. Wilcox, L., B. Schilit, and N. Sawhney. *Dynomite: a dynamically organized ink and audio notebook*. in *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems*. 1997: ACM.