The Audio Notebook
Paper and Pen Interaction with Structured Speech

Lisa Joy Stifelman

Bachelor of Science, Engineering Psychology, Tufts University, 1988
Master of Science, Massachusetts Institute of Technology, 1992

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
at the
Massachusetts Institute of Technology

September 1997

Author:_____
Program in Media Arts and Sciences
August 8, 1997

Certified by:_____
Christopher M. Schmandt
Principal Research Scientist, Media Laboratory
Thesis Supervisor

Accepted by:_____
Stephen A. Benton
Chair, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

The Audio Notebook

Paper and Pen Interaction with Structured Speech

Lisa Joy Stifelman

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning, on August 8, 1997,
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Abstract

This dissertation addresses the problem that a listener experiences when attempting to capture information presented during a lecture, meeting, interview, or conversation. Listeners must divide their attention between the talker and their notetaking activity. A tape recording can capture exactly what and how things are said, but it is time consuming and often frustrating to find information on a tape. This thesis combines user interaction and acoustic processing techniques to enable a listener to quickly access any portion of an audio recording. Audio recordings are structured using two techniques: *user structuring* based on notetaking activity, and *acoustic structuring* based on a talker's changes in pitch, pausing, and energy. By bringing audio interaction techniques together with discourse theory and acoustic processing, this dissertation defines a new approach for navigation in the audio domain.

The first phase of research involved the design, implementation, testing, and use of the Audio Notebook. The Audio Notebook combines the familiarity of taking notes with paper and pen with the advantages of an audio recording. This device augments an ordinary paper notebook, synchronizing the user's handwritten notes with a digital audio recording. The user's natural activity, writing and page turns, implicitly indexes and structures the audio for later retrieval. Interaction techniques were developed for spatial and time-based navigation through the audio recordings. Several students and reporters were observed using the Audio Notebook during a five-month field study. The study showed that the interaction techniques enabled a range of usage styles, from detailed review to high speed skimming of the audio. The study also pointed out areas where additional information was needed to improve the correlation between the user's notes and audio recordings, and to suggest structure where little or none was generated by the user's activity.

In the second phase of research, an acoustic study of discourse structure was performed using a small multi-speaker corpus of lectures. Based on this study, acoustic processing techniques were designed and implemented for predicting the locations of major phrases and discourse segment beginnings. These acoustic structuring techniques were incorporated into the Audio Notebook, creating two new ways of interacting with the audio—*audio snap-to-grid* and topic suggestions. Using phrase detection, the Audio Notebook "snaps" back to the nearest phrase beginning when users make selections in their notes. Topic suggestions are displayed along an audio scrollbar, providing navigational landmarks for the listener. The combination of user activity and acoustic structuring techniques is very powerful. The two techniques complement each other, allowing listeners to quickly and easily navigate through an audio recording and locate portions of interest. Thus, rather than replacing real-world objects like paper and pen, we can successfully augment them, combining the advantages of the physical world with the capabilities of digital technology.

Thesis Supervisor: Christopher M. Schmandt
Principal Research Scientist, Media Laboratory

This work was performed at the MIT Media Laboratory. Support for this research was provided in part by the National Science Foundation under grant number IRI-9523647, AT&T, Interval Research Corporation, News in the Future Consortium, and Digital Life Consortium. The views expressed herein do not necessarily reflect those of the supporting agencies.

Doctoral Dissertation Committee

Thesis Advisor:_____

Christopher M. Schmandt
Principal Research Scientist
Media Laboratory
Massachusetts Institute of Technology

Thesis Reader:_____

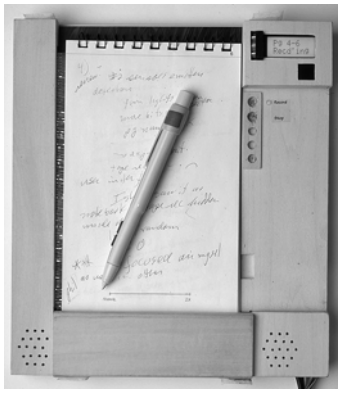
Barbara J. Grosz
Gordon McKay Professor of Computer Science
Division of Engineering and Applied Sciences
Harvard University

Thesis Reader:_____

Nathaniel I. Durlach
Senior Research Scientist
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Contents

1. Introduction



1.1 The Problem

This thesis is motivated by the problem of trying to capture and later review information presented during a lecture, meeting, interview, or conversation. A listener must simultaneously attend to the talker while attempting to write notes about what is said. In our everyday lives we are presented with many situations in which it is desirable to capture a detailed account of a presentation or conversation:

> While attending a lecture, a student attempts to write down what the professor is saying as well as information written on the blackboard. Students must divide their attention between listening to the professor, trying to comprehend the material, and taking notes for later review. The student's task is made even more difficult if the professor speaks rapidly or presents a lot of unfamiliar material.
>
> When a reporter performs an interview, he/she is trying to take notes, listen to the response, and think of the next question based on that response. In addition, notetaking can be distracting to some interviewees. Thus, reporters' notes often have gaps that would need to be filled to enable direct quotes for the story.
>
> When a lawyer conducts an interview, it is often necessary to capture a very accurate account of what is said. Detailed notetaking is essential to service the client's goals, and it may not be possible to re-interview a client to obtain missing information.
>
> At the scene of an accident or crime, a police officer attempts to capture exactly what witnesses say while the incident is still fresh in their minds. The officer may be working under difficult conditions because of the emotion and the usual gathering of onlookers.

In each of these examples, the listener could have made a tape recording of the information. A tape recording can capture exactly what and how things are said, and contains information that cannot easily be described in a transcript or handwritten notes. Imagine reading a transcript of Martin Luther King's "I Have a Dream" speech rather than hearing it. The intense emotion of his speech, the quality of his voice, and subtleties of his accents and pauses would be lost. However, it is time consuming and often frustrating to find information on a tape recording. A user must shuttle between fast forward and rewind to find the portions of interest on the tape. It is difficult to skim through the recording or correlate it with one's handwritten notes.

1.2 Research Overview

The Audio Notebook combines the familiarity of paper and pen with the advantages of an audio recording. The goal is to retain the original audio while allowing a listener to quickly and easily access portions of interest. The Audio Notebook augments a paper notebook, synchronizing the user's handwritten notes with a digital audio recording of the material. The user's natural activity, writing and page turns, implicitly indexes the audio for later retrieval. *Time* is mapped to *space*—the spatial layout of writing in a physical notebook enables rapid navigation in the audio domain. Familiar objects like paper and pen are used for interacting with the audio rather than artifacts left over from analog devices, such as fast forward and rewind controls. Users can *skim* their handwritten notes and the audio recording at the same time.

There are two major phases of the thesis research—user-structured audio and acoustically-structured speech (Figure 1-1). The first phase starts with the design and development of the Audio Notebook. The Audio Notebook exploits a user's notetaking activity to structure an audio recording by providing indices for subsequent retrieval (referred to as *user-structured audio*). As part of the Audio Notebook development, several techniques were designed for interacting with audio using paper and pen. Two versions of Audio Notebook hardware prototypes were developed and studied. A longitudinal user study of the Audio Notebook was performed over a five-month period. This study suggested areas where additional structural information was needed to augment the user's activity, leading to the second major phase of research—*acoustically-structured speech*.

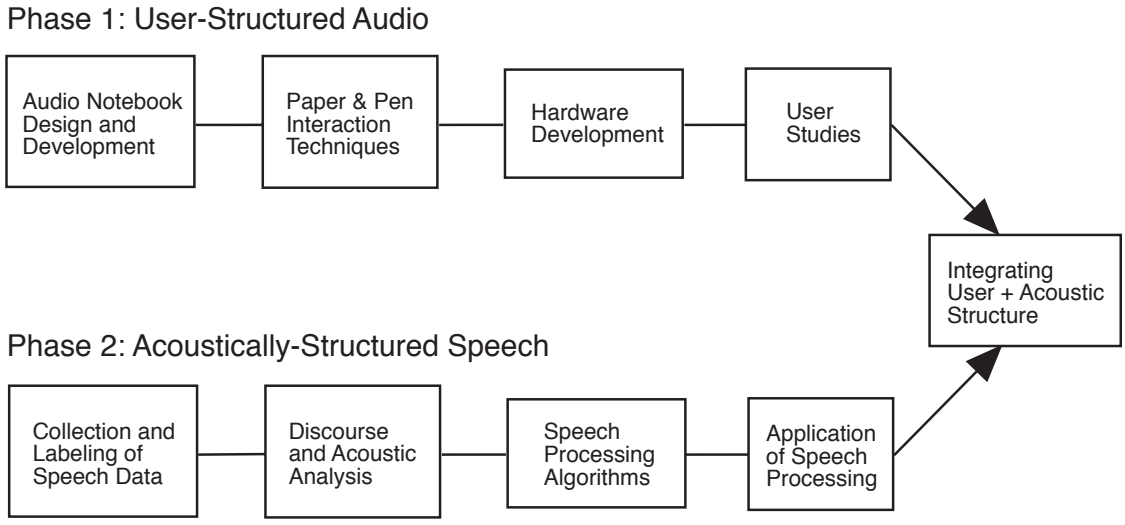


Figure 1-1: Overview of the two major phases of thesis research—User-Structured Audio and Acoustically-Structured Speech.

Phase two of the research involved studying the ways talkers use prosodic cues (e.g., changes in pitch, pausing and energy) to signal structural information such as the start of new phrases and new topics. Based on an acoustic study of discourse structure using a small multi-speaker corpus of lectures, processing techniques were developed for predicting the locations of major phrases and discourse segment beginnings. The approach was to first begin with the user's activity during recording and then to augment this basic structure using speech processing techniques.

Finally, user-structuring and acoustic-structuring of the audio recordings were integrated in the Audio Notebook user interface. Acoustic structuring techniques were exploited for the Audio

Notebook user interface to improve the correlation between the user's notes and the audio recordings, to start playback from phrase beginnings, and to suggest structure where little or none was generated by the user's activity. By bringing audio interaction techniques together with discourse theory and acoustic processing, this dissertation defines a new approach for rapid navigation and skimming in the audio domain.

1.2.1 Speech Data versus Speech-to-Text

People often ask, why don't you just translate the speech recording to text? While a lot of progress has been made in the area of automatic speech-to-text transcription, it remains a challenging research problem, especially for unconstrained, spontaneous speech, recorded under natural conditions (i.e., containing background noise).

There are many reasons to retain the original audio, if it could be made easier to access. Speech is expressive. There is a lot of information contained in a speech recording that is difficult to capture in a text transcript. For instance, the emotion of the speech and prosodic information such as accent and pausing are lost when it is transcribed. This research focuses on working with speech as data and developing user interaction and audio processing techniques that enable listeners to quickly find desired portions of an audio recording.

1.2.2 Interface Design Approach: Why Paper?

Previous systems have correlated an audio recording with notes written on an LCD display [Lamming 1991, Whittaker et al. 1994]. This thesis takes a different approach. Rather than writing on a display, the user writes in an ordinary paper notebook. The goal of the Audio Notebook is to *augment* rather than replace familiar objects like paper and pen [Wellner et al. 1993].

Paper and pen provide a portable, tangible, and flexible way of capturing information [Newman and Wellner 1992], and are still widely used, even at a technical institution like MIT. Paper documents have many advantages over digital ones—a sheet of paper can be quickly torn from a notebook, stuffed in one's pocket for easy access, or handed to a friend. Ideas can be quickly scribbled down on paper. While it is possible to write on the screen of a pen-based computer, such as a Newton, many people find writing on paper faster, more accurate, and more familiar. People like the *feel* of pen on paper, and the texture of the pages.

A physical notebook provides a spatial interface that is not matched on a flat computer screen. People have a spatial memory of where they put things, even in a notebook. Using a physical notebook, users can *leaf* through their notes, randomly moving from one location to another, while computer users are forced to *scroll* through information in a very linear fashion. There have been several attempts to address this problem within the confines of a two-dimensional computer screen [Schmandt 1980, Ginsburg et al. 1996]. In contrast, this thesis takes advantage of the physical affordances of a real notebook, while augmenting it with audio recording and playback capabilities. Ultimately, physical and virtual interfaces will be coupled, taking advantage of the capabilities each has to offer.

1.2.3 Taxonomy of Cues for Structuring Speech

This thesis combines multiple types of cues for indexing and automatically structuring a speech recording. Figure 1-2 provides a taxonomy of cues for segmenting speech recordings considered for use in this thesis (for additional possibilities, see [Arons 1994a]).

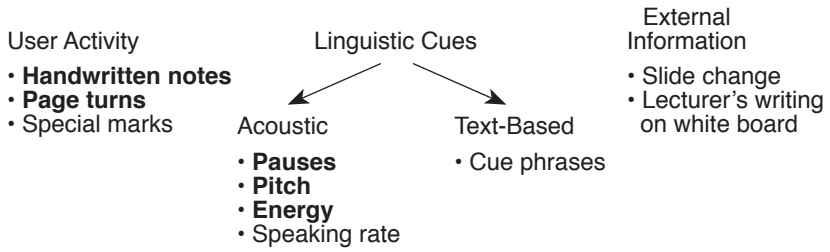


Figure 1-2: A taxonomy of cues considered for segmenting Audio Notebook speech recordings. Cues used by the final Audio Notebook user interface and acoustic processing algorithms are shown in bold.

Segmentation cues are broken down into three categories: implicit cues based on user activity, linguistic (acoustic and text-based) cues for automatically deriving structure, and external information from the environment. For each of the three categories, a number of possibilities are presented followed by a description of which cues are used for the Audio Notebook.

- User activity. This is also referred to as user-structure in this document. The natural activity of the user can provide implicit cues for segmenting a recording [Lamming 1991]. In addition to writing and page turns, special markings such as circles, stars, or lines across a page could also be exploited. For example, circled items in the notes could be used to create a summary of the audio. Without performing handwriting recognition, the digital ink could also be analyzed in a manner similar to a temporal analysis of speech; long pauses in speech carry significant information, by analogy so do pauses in writing [Wolf and Rhyne 1992].

  The Audio Notebook uses handwritten notes and page turns as indices into an audio recording. It is important to emphasize, that no *explicit* user segmentation of audio is used (e.g., button pushes to mark important parts). The natural activity of users can be exploited without interfering with their normal interactions. Analysis of the handwritten notes (special markings and temporal patterns) are areas for future research.

- Linguistic cues. Talkers use prosodic cues (e.g., changes in pitch, pausing, and energy) to convey structural information to a listener. These cues can be exploited in order to automatically derive structural information such as the beginning of new phrases and new topics.

  This thesis focuses primarily on acoustic (as opposed to text-based) cues to discourse structure. Among these, pitch, pause, energy, and speaking rate were studied. Cue phrases were the only text-based feature investigated. The Audio Notebook does not rely on text-based cues because this would require speech-to-text transcription or word spotting.

- External information. There are many changes in the environment (e.g., location, identification of participants) that could be exploited for segmenting an audio

recording. For example, the time point when a lecturer changes slides usually indicates a topic change. Slide changes could be identified using a sound recognition algorithm for identifying page turning noises, or by equipping a slide projector with a mechanism for indicating when a slide is changed. A lecturer's writing on a whiteboard could also be captured and used to index the audio, in a similar manner to the user's writing activity.

External information was not used in this thesis due in part to the amount of infrastructure required. For example, in order to capture a professor's writing during a lecture, an electronic whiteboard would be required. In developing the Audio Notebook the goal was to make a system that would be usable in all environments, and not have to rely on an electronically-equipped classroom or meeting room.

1.3 Thesis Overview

This section provides an overview of the remaining chapters of the thesis.

Chapter 2 provides descriptions of related research. Background areas include: interfaces for capturing, indexing and accessing recorded speech, tangible interfaces that augment paper, and linguistic cues to discourse structure.

Chapter 3 presents the Audio Notebook and user-structured audio. The design and development of two versions of the Audio Notebook are described. Several techniques for interacting with audio using paper and pen are also presented. An early user study of the first Audio Notebook prototype is discussed.

Chapter 4 presents the results of a five-month longitudinal user study of the second Audio Notebook prototype. Results are presented for several students who used the Audio Notebook during one semester, and two reporters who used the Audio Notebook to write a story for publication. This chapter also discusses design changes that were made iteratively during the study.

Chapter 5 describes the development of techniques for acoustically structuring a speech recording. First, an acoustic study of discourse structure for lectures is presented. The collection and labeling of speech data is described along with the theoretical underpinnings and methodology for the study. Two techniques for acoustically deriving structural information are presented: phrase detection and prediction of discourse segment beginnings (i.e., topic introductions). The two techniques are evaluated and results are compared with related work.

Chapter 6 describes the integration of user and acoustic structure in the Audio Notebook user interface. In particular, two new ways of interacting with the audio are presented: (1) audio snap-to-grid using phrase detection and (2) topic suggestions using discourse segment beginning prediction. Lastly, a multi-part approach for correlating notes and audio based on user structure, acoustic structure, and user-system interaction is described.

Chapter 7 provides a summary of the contributions of the thesis research as well as areas for future research.

2. Background

This thesis draws from a number of research areas including speech and audio interaction, speech processing, and acoustic studies of discourse structure. This chapter describes related research in each of these areas.

The first section describes interfaces for capturing, indexing, and accessing speech recordings. Systems that index audio recordings are separated into two categories: (1) systems that use activity information to index audio (and video) recordings; and (2) systems that use acoustic processing techniques to automatically index speech recordings. Very few systems have combined these two techniques. A second important area of related research are interfaces that augment and make use of people's knowledge about everyday objects. In particular, this section focuses on systems that augment paper. The last section of related work describes studies of linguistic correlates of discourse structure. Discourse structure provides a theoretical foundation for this thesis research. This thesis focuses on acoustic cues to discourse structure, and application of these cues for creating *structured* speech recordings.

2.1 Interfaces for Capturing, Indexing, and Accessing Recorded Speech

This section describes interfaces for capturing, indexing, skimming, and browsing recorded speech. Recorded speech can be categorized along many different dimensions, including the length of the speech recorded (short snippets versus long recordings), the type of material recorded (voice mail, lectures, meetings), and the people speaking (self-authored, one speaker, multiple speakers) [Arons 1994a]. In each of the following sections, speech interfaces will be described covering a range of these dimensions.

2.1.1 Spontaneous Capture of Speech

In Monty's thesis on notetaking, she refers to two different types of memory—prospective and retrospective [Monty 1990]. "A note that functions as a prospective reminder reminds one of something to occur in the future" [Monty 1990, 15]. The augmented tape recorder [Degen et al. 1992] and VoiceNotes [Stifelman et al. 1993] are two hand-held interfaces supporting prospective memory. The audio is self-authored, and the recordings are short in length, similar to the duration of voice mail messages.

Degen et al. augmented a conventional tape recorder with two buttons for marking segments of speech while recording [Degen et al. 1992]. Users could mark short recordings with a "to-do" button, or press another button to mark different places of interest in a longer recording. The speech data was then digitized and stored on a Macintosh computer for review. Button presses were displayed in alignment with a time waveform of the speech signal. In an evaluation of this prototype, users expressed the desire to customize the meanings of the marks, for more buttons to uniquely tag audio segments, and the ability to play back the marked segments directly from the device. The augmented tape recorder relied on *explicit* indexing of an audio recording by the user. Section 2.1.2 describes systems that *implicitly* index audio recordings using activity information (e.g., notetaking).

VoiceNotes [Stifelman et al. 1993] is a speech interface for a hand-held voice notetaking device. VoiceNotes allowed users to capture and randomly access spontaneous thoughts, ideas, or things to do in contexts where writing would be difficult (e.g., while driving, walking down the street). While Degen's augmented tape recorder provided buttons for marking the audio, VoiceNotes allowed dynamic creation and naming of categories using speech input. This research explored the concept of a hand-held computer with no keyboard or visual display but a speech interface instead. User-defined categories provided a mechanism for organizing and randomly accessing notes by voice. The interface provided users with random access to their voice notes, dynamic speed control, and customizable speech and non-speech audio feedback. Conversational VoiceNotes [Stifelman 1995] allowed users to tag voice notes with attributes such as importance, data, and time and to use these tags in spoken commands to access subsets of the notes.

Another aspect of VoiceNotes was background recording of speech in a digital "tape-loop." The idea was to support short-term memory by constantly capturing the last several minutes of background speech. The interface, called Capture, was implemented in two forms: as part of the VoiceNotes hand-held interface [Stifelman 1992] and using a graphical user interface [Hindus et al. 1993]. Capture supports *retrospective* memory—"a retrospective reminder stores information from the past" [Monty 1990, 15]. Some other examples of retrospective reminders are class notes and meeting minutes. For the VoiceNotes interface, background recordings were made whenever the user was not intentionally recording notes. The captured speech was automatically segmented based on pauses. Users could then save these segments of speech in one of their lists of voice notes.

2.1.2 Indexing Audio with User Activity

The augmented tape recorder allowed users to explicitly index audio recordings using two buttons for marking the audio [Degen et al. 1992]. In contrast, this section describes systems that implicitly index audio recordings based on the user's activity. Note that these systems generally focus on activity information alone for indexing the audio and very few couple this with automatic processing of the audio recordings or digital ink. In addition, it is surprising that little or no mention is made of the problem of correlating notes and audio, given that listeners do not take notes exactly in synchronization with the talker (see Section 6.3).

The AIR project (Activity-based Information Retrieval) proposed employing user activity (e.g., notetaking, writing on whiteboards, user location) to index multimedia data [Lamming 1991]. The idea was that a collection of events could be used to cross-index the same information, allowing easier retrieval. One prototype called NoTime, indexed an analog video recording using notes written with a stylus on a notepad computer. Filochat also linked writing on an LCD tablet (tethered to a PC) to audio recordings of business meetings [Whittaker et al. 1994]. Professionals who used Filochat to record and review business meetings perceived an improvement in their meeting minutes. The KeyRecorder[1] indexed an audio recording using typewritten notes on a laptop computer. Whether entered with a pen or keyboard, these notes then act as a index into the speech recording—the user can select a portion of the notes and playback the audio recorded during this time. In a similar manner to indexing audio with notetaking activity, NewsTime used closed-captioning information to index audio news recordings [Hindus et al. 1993].

---

[1] I developed this prototype in June of 1992.

Marquee used digital pen strokes for real-time video logging [Weber and Poon 1994]. However, the logging activity is very explicit in comparison to Filochat's use of notetaking activity. For example, rather than timestamping individual pen strokes, users explicitly draw lines across a notetaking area to indicate a "time zone." All notes written within that time zone are indexed to a single time point on the video tape. Users also explicitly create keyword labels by writing them in a keyword area or circling keywords in their notes. The user then manually applies these keyword labels to the time zones during video recording. In a study of Marquee during a brainstorming meeting, users took notes but did not speak during the meeting. The authors stated the need to determine whether or not people can use Marquee to take notes and participate in a meeting at the same time.

Dynomite, like Filochat, indexes audio with notetaking activity on a pen-based computer [Wilcox et al. 1997]. Users can manually assign keywords to pages of notes, or properties like "name" to digital ink selections. This use of keywords is similar to Marquee, however, Dynomite uses text keywords, and Marquee allows users to create keywords from the digital ink by circling information in their notes. The Dynomite system focuses on information retrieval, allowing users to access subsets of their notes based on dates, keywords, and properties. Although audio is recorded continuously, "only portions of audio highlighted by the user are permanently stored" [Wilcox et al. 1997, 186]. This is intended to deal with audio storage limitations, however, discarding portions of the original recording in this manner can be problematic (see Section 3.1).

Wolf and Rhyne developed a shared drawing tool called We-Met which segments a meeting record into meaningful units based on *turns* [Wolf and Rhyne 1992]. A turn is defined as a sequence of activity by one person. The meeting record was segmented using turns proceeded by a pause length of at least one second. This segmentation was intended to support subsequent browsing of the meeting record. Although the We-Met system does not record audio, it provides an interesting use of activity information for automatically segmenting a meeting record.

The systems described so far have focused on the user's notetaking activity. Several other systems have been developed to operate within the infrastructure of an electronically-equipped meeting room or classroom [Abowd et al. 1996] where meeting or lecture notes are also captured.

Researchers at Xerox PARC developed Coral, a "confederation" of tools for capturing, indexing, and *salvaging* collaborative activities [Minneman et al. 1995]. Moran et al. define "salvaging" as a "new kind of activity involving replaying, extracting, organizing, and writing" [Moran et al. 1997, 202]. Here the focus is on collaborative rather than personal use. A meeting room is equipped with a large pen-based electronic display (LiveBoard), video camera, multiple microphones positioned around a meeting table, and a Sun workstation for digitizing the audio and video. During the meeting, one user takes notes on a laptop computer. Notes from the laptop can be "beamed" to the LiveBoard. The audio and video are indexed by all notes written on, or beamed to, the LiveBoard, and by page changes on the LiveBoard.

Extensive studies of the Coral system were performed over two years [Moran et al. 1996, Moran et al. 1997]. The domain of the study is the management of intellectual property (IP) at Xerox PARC. The study focuses on one central user who manages the process and creates reports of the IP meeting results. The manager was provided with a tool called the SalvageStation, for reviewing the meeting recordings and writing his reports. This program, running on a Sun workstation, allowed him to playback the meeting records using notes and whiteboard page

indexes (each page was associated with an IP item for discussion). In one study, the manager's activity was tracked over several "salvaging" sessions [Moran et al. 1997]. Results showed that the manager produced more complete and accurate reports using the audio records of the meetings. The manager developed special annotations made during meetings such as "HA" for hear audio later. Another interesting conclusion was that the salvaging tools were "more valuable—indeed crucial—for dealing with difficult and unfamiliar materials" [Moran et al. 1997, 208].

Another system for recording meetings and presentations is STREAMS [Cruz and Hill 1994]. The STREAMS system uses an elaborate setup with multiple cameras and microphones. The presentations can be indexed with text and audio annotations *after* recording but not during. Forum [Isaacs et al. 1995] is a system for real-time broadcasting of presentations (video, audio, and slides) to computer desktops. The focus of Forum is on real-time interaction versus capture and review of presentations.

2.1.3 Acoustic Indexing

The systems described in the previous section all make use of activity information to index the audio recordings. The following systems index recordings using acoustic information.

SpeechSkimmer [Arons 1997] provides a user with an interactive control for listening to a recording at multiple levels of detail. In this system, the recordings are automatically segmented using speech processing techniques. Time-compression, pause removal, and emphasis detection are used to create a continuum of listening speeds and levels of detail. Several time-compression algorithms were explored including time domain techniques [Fairbanks et al. 1954, Roucos and Wilgus 1985] and a dichotic approach taking advantage of both ears to achieve faster speeds with better comprehension [Scott 1967]. An adaptive pause detection algorithm was developed for finding, shortening, and removing pauses from speech recordings. This algorithm adapts to the background noise of the recording and the length of pauses for a particular speaker (Section 5.4). An emphasis detection algorithm (Section 2.1.3.2) analyses the pitch patterns of the speech in order to extract salient portions. These processing techniques provide a foundation for the SpeechSkimmer user interface (described in Section 2.1.4).

2.1.3.1 Speaker Segmentation

Hindus developed a system for recording telephone conversations called the Listener [Hindus et al. 1993]. This system presents a graphical representation of the conversation while the call is in progress. First, the Listener segments the audio into speaker turns. The system uses two microphones, one connected to the telephone line, and one in the user's office. Using the single speaker recording from the office microphone, the system is able to determine speaker turns. Within each speaker turn, the audio is segmented based on pauses. Users can then select on these speech segments to save them.

Speaker segmentation or indexing, segments a speech recording based on speaker changes without necessarily knowing the speakers' identities [Gish et al. 1991, Wilcox et al. 1994, Roy 1997]. Speaker identification labels the identity of each person in a recording given prior knowledge of the speakers.

Kimber uses speaker segmentation to index audio recordings of meetings [Kimber et al. 1995]. A graphical browsing tool displays a timeline of the speech recording that indicates speaker

changes. Each speaker is displayed in a separate “track”, or the tracks can be combined with speaker changes indicated by color coding.

NewsComm is a hand-held device for accessing audio news on demand [Roy 1996, Schmandt and Roy 1996]. Audio is downloaded from a server to the hand-held device and accessed using button controls. The user interface provides controls for jumping to different portions of the recording depending on the level of detail selected by the user. A contribution of this work is the development of an algorithm for speaker indexing [Roy 1997]. Speaker change and pause length are combined to determine “jump” locations in the recording.

2.1.3.2 Emphasis Detection

One approach to the problem of summarizing and skimming speech has been termed emphasis detection [Chen and Withgott 1992]. This approach uses prosodic cues (e.g., pitch, energy) for finding “emphasized” portions of speech recordings. Chen and Withgott used speech labeled by subjects for emphasis to train a Hidden Markov Model. It was determined that simply identifying the most emphasized words did not provide a useful summary, but by looking across regions of emphasized speech, better results were obtained. The automatically created summary was then compared to summary portions selected by a group of subjects. The agreement between the automatic and manually created summaries was similar to the level of agreement achieved between subjects. In contrast, if summary portions were randomly selected, the agreement score was much lower.

Arons’ emphasis detection algorithm attempts to select salient portions of a discourse [Arons 1994b]. An “emphasis threshold” is determined based on the top one percent of pitch peaks in the recording. Similar to Chen and Withgott, regions of speech above the emphasis threshold are selected for an overview of the recording.

2.1.3.3 Word Spotting

Word spotting is used to locate keywords or phrases in a speech recording. “Word spotting differs from continuous speech recognition in that the task involves locating a small vocabulary of words embedded in arbitrary conversation rather than determining an optimal word sequence taken from a fixed vocabulary” [Rohlicek et al. 1989, 627].

Wilcox et al. integrated word spotting into an audio editing system, allowing users to search for keywords in recordings [Wilcox et al. 1992]. This system uses speaker dependent word spotting. Thus, it is appropriate for applications such as dictation and personal voice notes [Stifelman et al. 1993].

An approach termed “gisting” [Houle et al. 1988, Maksymowicz 1990] has been used to classify speech messages using word spotting, and syntactic and timing constraints. These systems have primarily been used for message retrieval and notification applications, usually attempting to place a message into one of a number of pre-defined categories. Houle [Houle et al. 1988] proposes an automatic gisting system that notifies an operator whenever a high priority message arrives. Houle discusses techniques for filtering the output of a word spotting system (e.g., requiring phrases of two or more words to be spotted within a certain time window) in order to reduce the number of false alarms. It is also proposed that such a system would “summarize” messages for operators.

Denenberg et al. developed a system that analyzes air traffic control communication and creates a "gist" of the activity by automatically identifying and classifying flight information [Denenberg et al. 1993]. However, this system uses continuous speech recognition as opposed to word spotting. The system first segments the speech from background noise, separates the speakers, and outputs a hypothesized word sequence for each segment of speech.

Word spotting was used to index video mail recordings [Brown et al. 1994]. A video mail browser displayed a timeline underneath the video with shaded areas indicating where key words were found. A fixed set of 40 keywords were used for this system. The darkness of the shaded areas indicated the system's confidence. James used a combination of word spotting and information retrieval techniques to allow indexing and searching of BBC radio news reports [James 1996]. James contrasts his system with previous work in topic identification which classifies spoken messages into a pre-defined set of topic categories; James' system allows an unrestricted number of topics. In related work, information retrieval techniques were used for automatic content-based retrieval of BBC broadcast news recordings [Brown et al. 1995]. In this research, text subtitles (i.e., closed-captioning) are used which contain a nearly complete transcript of the broadcasts.

Jabber is a system currently under development that is intended to capture and index video conferences [Kazman et al. 1996]. Meeting participants share audio, video, and computer applications from their workstations. This system uses a combination of speech recognition and information retrieval techniques to index the audio. A continuous speech recognition system (1,000–2,000 word vocabulary) processes the audio, and the recognized words are then grouped into "semantically linked" trees. For searching the meetings, a graphical timeline is created, with buttons for recognized words that can be used to access the audio or video.

2.1.4 Speech Skimming and Browsing

This section describes audio-only systems for skimming and browsing through recorded speech. These systems provide an audio environment without a visual display. The user navigates through the audio streams using tactile input, speech input, or hand gestures.

Several speech-only hypermedia systems have been developed [Muller and Daniel 1990, Arons 1991, Resnick 1993]. Hyperphone was a hypermedia environment for navigating through "voice documents." Voice documents were collections of short segments of text read using text-to-speech output and accessed using speech input [Muller and Daniel 1990]. Hyperspeech allowed users to navigate through a network of digitally recorded interviews using speech input [Arons 1991]. Links between different speech segments (e.g., supporting and opposing views) were hand-crafted. HyperVoice was a telephone-based architecture supporting collaborative generation and access to speech recordings [Resnick 1993]. Callers filled out "telephone forms" when adding information to the system, to structure the recordings for later retrieval. Touch tone input was used to navigate through the system when adding or retrieving information [Resnick 1992].

The SpeechSkimmer user interface provides the user with continuous real-time control over the speed and detail level of a recording [Arons 1997]. Users interactively move between four different levels of detail in the forward or backward direction by sliding their finger along the surface of the touch pad (Figure 2-1). At any level, the user can dynamically change the speed of playback, slowing it down to catch an important point, or speeding it up to quickly listen for a topic of interest. The most detailed level, unprocessed speech, plays the entire recording from

beginning to end. The second level, pause shortened, removes short pauses in the recording and shortens long ones. The third level, pause-based skimming, plays segments of speech following the longest pauses in the recording. At the highest level, pitch-based skimming, emphasis detection is used to select and playback salient portions of the recording as an overview. While listening to these salient sound bites, the user can stop on a topic of interest and listen to it in full detail. Jump controls are also provided, allowing the user to skip forward and back between segments of the recording, the size of the jump based on the level of skimming.

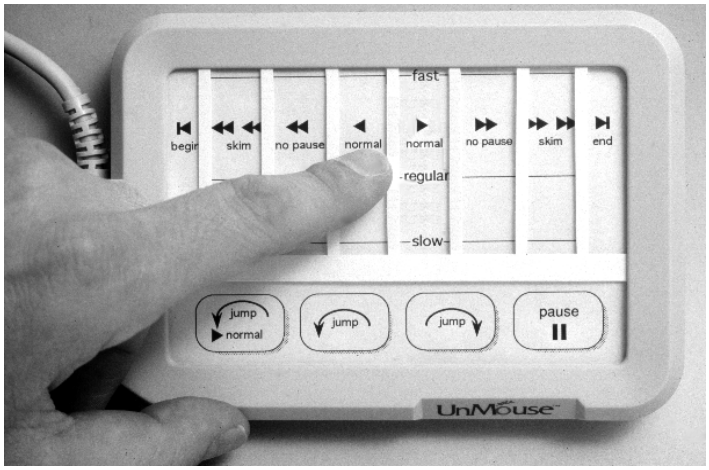


Figure 2-1: The SpeechSkimmer interface uses touch-pad input, providing interactive control of speed and skimming level.

2.1.4.1 Simultaneous Listening Using Spatial Presentation

Another approach to the problem of skimming and browsing audio is simultaneous presentation of multiple audio sources [Arons 1992a, Cohen 1992]. AudioStreamer [Schmandt and Mullins 1995] presents a listener with three simultaneous channels of audio news. The idea is to exploit the "cocktail party effect"—the human's ability to selectively attend to a single talker or stream of audio among a cacophony of others—to reduce the amount of time required to listen [Cherry 1953]. The AudioStreamer system uses speaker indexing [Roy 1997] and closed-captioning information to segment the news recordings. This information is then used to enhance the listening experience, for example, using tones to alert the user when a new story is beginning on a background channel. The premise is that by presenting multiple streams of audio simultaneously, users will be able to focus on one, yet overhear interesting information in another and switch their attention.

Dynamic Soundscape maps a linear audio recording to spatial locations around a listener's head [Kobayashi and Schmandt 1997]. A single audio recording is played while continually changing the location of the sound so that it "orbits" the listener's head. The idea is to take advantage of the listener's spatial memory of where different topics are played. In order to re-play a topic, the listener points to the spatial location where the original audio was played. A new audio stream then starts re-playing the recording from the selected time point. The original audio stream continues to play in the background at a lower volume. Thus, the listener hears multiple streams of audio from the same recording simultaneously.

One issue for the design of a "Dynamic Soundscape" is how fast to move the audio signal around the listener's head. If the signal moves too slowly, there will not be much spatial resolution

between the different parts of the recording. If the signal moves too quickly, the audio recording will wrap around the listener's head many times, so many topics will get played at the same location. Since the goal is to allow listeners to access "topics" at different spatial locations, knowledge of the structure of the recording (e.g., topic change locations) would help to more intelligently determine the spatial mapping of the audio and rate of "orbit" around the listener's head. The Soundscape system also requires a high quality spatial audio system. For example, if the listener experiences front/back reversals (as occur with many headphone-based spatial audio systems), the listener will not be able to correctly identify the location of the topics.

A design issue for both of these spatial skimming interfaces is the ability of the listener to comprehend one sound source while monitoring for interesting topics in several other audio streams. In a study of simultaneous presentation, subjects listened to multiple streams of audio and performed two tasks simultaneously—listening comprehension and target monitoring [Stifelman 1994]. While listening to one passage of speech, the subject had to identify target words in one or two other recordings played simultaneously. The subject's listening comprehension and target monitoring performance decreased significantly as the number of streams increased. Listening comprehension performance dropped from 77% with one stream of audio to 62% with two streams, to 52% with three streams. Target monitoring performance dropped from 63% with two streams of audio to 40% with three streams. However, perfect comprehension and monitoring are not needed for an audio browsing task. The listener can attend to multiple sources while browsing through the recordings, and focus on a single recording once a topic of interest is found. Spatial interfaces for speech skimming should therefore support this kind of task switching, from browsing to more focused listening.

2.2 Tangible User Interfaces

Graphical user interfaces provide people with "direct manipulation" of images on a computer screen. Objects from the physical world (e.g., files, folders) are abstractly represented on the computer screen. In contrast, *tangible* user interfaces augment the physical objects themselves with computational capabilities [Ishii and Ullmer 1997]. "Instead of using computers to enclose people in an artificial world, we can use computers to augment objects in the real world" [Wellner et al. 1993, 26]. One example of particular relevance to this thesis is Durrell Bishop's telephone answering machine [Poynor 1995]. The answering machine outputs colored marbles to represent each voice message. Embodying the audio messages within physical objects makes them easier to manipulate and manage. It also takes advantage of people's familiarity with everyday objects.

This research area has been referred to as ubiquitous computing [Weiser 1991], augmented reality [Feiner et al. 1993], and computer-augmented environments [Wellner et al. 1993]. This section will not attempt to cover all related work in this area but will instead focus on one area of most relevance to the Audio Notebook, interfaces that augment real paper. Perhaps the ultimate marriage between paper and computers is electronic ink, being developed by Joe Jacobson [Comiskey et al. 1997, Negroponte and Jacobson 1997]. In the future, electronic ink may allow people to carry a single book that could automatically change its contents, combining the advantages of paper with the flexibility of the computer.

2.2.1 Augmenting Paper

This section describes three systems that augment paper with computational capabilities, to create a digital desktop, support iterative changes to engineering drawings, and for spatial layout of video. For other examples see [Johnson et al. 1993, Arai et al. 1995, Arai et al. 1997]. In addition, interactive books are discussed and contrasted with the Audio Notebook.

2.2.1.1 The Digital Desk

The Digital Desk takes the opposite approach of conventional computer desktop interfaces [Newman and Wellner 1992, Wellner 1993]. Rather than trying to replicate a physical desktop on a computer screen, the Digital Desk integrates the computer with an actual desktop. “Instead of making the workstation more like a desk, we can make the desk more like a workstation” [Wellner 1993, 88]. The DigitalDesk is an actual desk with a computer display projected onto it, and overhead cameras to track the user’s activity. Electronic images are projected onto paper documents and the desktop. Several example applications were developed, including a calculator and a paint program called PaperPaint.

The DigitalDesk calculator allows users to point to numbers in any printed document to enter them into the calculator. In this way, no transfer from the paper document to the calculator is required. Users can point to the numbers using a pen or finger. An overhead camera is used to capture the numbers and then image processing is performed to recognize the digits. In this prototype, an image of a calculator is projected onto the desktop.

The PaperPaint program applies computer-like cut and paste capabilities to paper documents. PaperPaint allows users to simultaneously work with paper and electronic images and integrate the two. For example, PaperPaint users can simply place a drawing on their desktop, select any portion of it using a stylus, and integrate it with electronic images. Users also continue working with the images within the physical space of their desktop, allowing them to naturally manipulate the physical and electronic drawings with both hands.

2.2.1.2 Engineering Drawings

Ariel is a system which augments blueprints or engineering drawings used by architectural engineers [Mackay et al. 1995]. These blueprints are the central “artifact” used by the engineers for their daily work. The engineers, who are constantly on the move, annotate changes directly on the drawings while on site. These updates are rarely made to the version of the drawings maintained on the computer system. Computer systems remain at the desktop and do not get much use. The Ariel prototype allows an engineer to place a blueprint on a drawing table (actually a large digitizing tablet) and add written and spoken annotations that are simultaneously entered into the computer system. Drawings are recognized using a barcode that the user manually swipes after placing a drawing on the table. Computer menus are projected onto the table, creating a combined physical and virtual interface. The user selects from the menus using a laser pointer that is tracked by an overhead camera.

This system uses the computer to augment the user’s current work artifacts and work habits. However, the prototype implementation does not seem to support a major necessity of the engineers—mobility. It is stated that the engineers are constantly on the move and update drawings while on site. Yet, the engineers must place the drawings on a large work surface outfitted with cameras and a projector in order to annotate them and have these updates tracked by the computer system. In addition, audio annotations made in the field must be transferred from

a cassette recorder to the drawings. Since the engineers did not update drawings with annotations made on site, these audio recordings may never get linked to the drawings.

2.2.1.3 Video Layout

Video Mosaic is a system for spatially laying out video storyboards using paper [Mackay and Pagani 1994]. According to Mackay and Pagani, even experienced video producers still use paper storyboards instead of sophisticated editing systems. Pieces of paper can be quickly organized, easily arranged, and users have spatial memory for where they have placed them. On the other hand, it is easier to search through and reorder a large number of images on a computer. Without a tool like Video Mosaic, there was previously no way to connect the video to these storyboards so all the work had to be redone on the computer. This system combines the best elements from the physical and virtual interfaces, allowing the storyboard to be laid out with paper, and searched and edited on-line.

2.2.1.4 Interactive Books

There are many interactive books available on CD (e.g., the Amanda stories by Amanda Blinn) that combine graphics, audio, and/or video images. The systems described in this section link pictures or pages in physical books with audio recordings. These systems differ from the Audio Notebook in several ways. First, they are *authored* in advance (i.e., audio recordings are manually linked to pictures or pages). In contrast, the Audio Notebook dynamically captures and indexes audio recordings. Secondly, these systems are playback only. The Audio Notebook records, indexes, and plays any audio captured by the user. The Audio Notebook could even be used as a tool for easily generating such interactive books.

Durrell Bishop prototyped a scrapbook which links short audio recordings to pictures in a physical notebook [Poynor 1995]. The system is authored in advance—segments of audio are manually associated with different drawings in a notebook. The user manually tells the system which page they are using by selecting a page number with a stylus. While this may be adequate for a playback-only system, requiring the user to manual select pages is not appropriate for a recording interface (as discussed in Section 3.1).

Many toys are now incorporating audio playback. Electronic greeting cards are now common which record and play short greetings. Golden Books Family Entertainment has developed what it calls “Smart Page” technology. Each page in a storybook has icons on it that a child can touch on to play sounds, music, or voices. These devices are generally limited to playing about 10 seconds of audio, the audio quality is poor, and playback is extremely noisy. A more advanced system is Sega’s Pico. The Pico connects to a TV like a video game, displaying animations on the television screen that a child can control by turning pages in a book, or touching on the screen.

2.3 Linguistic Cues to Discourse Structure

Research in emphasis detection and gisting (as discussed in Section 2.1.3.2) attempts to apply knowledge of the patterns and structure of spoken discourse to problems of summarizing, skimming, and classifying speech data. However, the problem of defining and recognizing discourse structure is not a new one. Although there are many theories of discourse structure [Hobbs 1979, Reichman-Adar 1984, Grosz and Sidner 1986, Polanyi 1988], it is generally agreed that a discourse can be broken into segments [Grosz et al. 1989]. Discourses are commonly said to have a hierarchical structure—segments can contain and/or be embedded in other segments. An

important research endeavor has been to determine the devices used by speakers to indicate structure to listeners. Studies in this area include analysis of intonational and text-based (i.e., lexical) cues to structure. Hirschberg notes that there have been few attempts to combine speech- and text-based approaches to discourse segmentation [Hirschberg 1994].

2.3.1 Acoustic Studies

An important cue to discourse structure is intonation. Pitch range has often been cited as an indicator of topic structure—researchers have noted that people tend to expand their pitch range when introducing new topics, and compress their range when ending or continuing a topic [Brown et al. 1980, Silverman 1987]. Another contributing factor is "final lowering"—the general declination in pitch toward the end of a sentence. Other acoustic cues to structure that have been cited are energy [Brown et al. 1980], pauses [Lehiste 1979, Brown et al. 1980, Chafe 1980, Silverman 1987], speaking rate [Butterworth 1975, Lehiste 1979], and contour type [Brown et al. 1980, Hirschberg and Pierrehumbert 1986].

Several studies have attempted to determine the correlation between intonation and discourse structure. In a study by Hirschberg and Grosz, subjects manually segmented news stories using instructions based on Grosz and Sidner's model of discourse structure [Grosz and Sidner 1986, Hirschberg and Grosz 1992]. Discourse segment beginnings, endings, and embedding relationships between segments were marked by each subject. The correlation between the subjects' hand-marked segmentations and the acoustics was analyzed with respect to pitch range, the change in pitch range from the prior phrase, the change in energy from the prior phrase, speaking rate, prior and subsequent pause lengths, contour type, and type of nuclear accent. The results showed that segment ending phrases were followed by a significantly longer pause than other phrases; segment beginnings were associated with a larger pitch range, greater amplitude, and a longer preceding pause than other phrases. The results also showed that consistent labeling of structure across subjects could be obtained. At least 74% agreement was achieved across subjects in all conditions, and between 77 and 92% agreement for segment beginning determinations.

In more recent work, Hirschberg and Nakatani studied the correlation between prosodic cues and discourse structure using a corpus of read and spontaneously spoken directions [Hirschberg and Nakatani 1996]. This work is discussed in more detail in Section 5.10.4.1 where the findings are compared with the results of an acoustic study of lectures performed for this thesis.

Swertz studied intonational correlates of structure for a corpus of spontaneous narratives [Swertz 1995]. The narratives were descriptions of paintings by two female speakers. Swertz correlated three acoustic cues with *boundary strength*—the proportion of subjects agreeing on the location of a "paragraph" boundary. There was a trend for pitch range, pause duration, and the number of low boundary tones to increase with boundary strength.

Ayers studies structural cues in interactive conversations rather than monologues [Ayers 1994]. This study also analyzes the impact of read versus spontaneous speech on intonational correlates of structure, in particular, pitch range. Two casual conversations between friends are used for the spontaneous speech data. The read versions were obtained by having subjects read portions from single speaker sections of the conversation. Unlike studies using monologues, this analysis takes into account both the topic structure and turn taking divisions in the conversation. Ayers reports that the pitch range for the read speech reflected a hierarchical topic organization more clearly

than for the spontaneous speech. In the spontaneous speech, the pitch range not only increased at the introduction of new topics, but also at the beginning of new or potential turns. Ayers concluded that "in spontaneous speech, corrections and turn management disrupted pitch range cues to topic structure." However, Hirschberg and Nakatani [1996] found that discourse boundaries in spontaneous speech were labeled as, or more, consistently than for read speech. The results of Ayers study may be a reflection of the interactive nature of the conversation (dialogue versus monologue) rather than the spontaneity of the speech.

2.3.2 Text-Based Studies

Until recently, research has focused more on text-based than prosodic cues to discourse structure. In particular, cue phrases are thought to play an important role in indicating structure. "Cue phrases are linguistic expressions such as *now* and *well* that function as explicit indicators of the structure of a discourse" [Hirschberg and Litman 1993, 501]. However, these words have both "discourse" and "sentential" uses. For example, when used as a discourse cue, "now" is said to signal the beginning of a subtopic or the return to a previous topic [Hirschberg and Litman 1993]. However, "now" can also be used as a temporal reference. In order for cue phrases to be useful for segmenting discourse, it is necessary to distinguish discourse from sentential uses.

Hirschberg and Litman have performed extensive studies of cue phrases, exploring both text- and prosodic-based features for disambiguating discourse from sentential uses [Hirschberg and Litman 1993]. In some cases, it is impossible to distinguish between the two given the text alone. Based on a study of the cue word "now", they developed an intonational model of discourse and sentential characteristics. Phrasing was found to be a powerful discriminator—if the potential cue word is in a separate intermediate phrase, the model classifies it as a discourse use. The model was further validated by a study of the cue phrase "well" and a study of multiple cue phrases. Performance of the model for the multi-cue study declined with the addition of the conjunctions "and", "but", and "or" which are difficult to disambiguate. If conjunctions are not considered, the classification success rate increases from 75.4% to 85.3%.

Researchers have also explored the use of information retrieval techniques for locating segment boundaries [Morris and Hirst 1991, Hearst 1993]. Hearst uses a method termed "TextTiling" to segment expository texts and magazine articles into discourse units or tiles. Similarity measurements are obtained based on word repetition and thesaurus information; low similarity values indicate potential discourse boundaries. An evaluation of the algorithm against data segmented by human subjects produced a relatively small number of errors, however, the segmentation is coarse grained and frequently off by a few sentences.

Flammia and Zue [1997] studied a corpus of human-human dialogues taken from telephone conversations between customers and operators of a movie information service. Discourse segmentation was performed by at least seven subjects for 25 different dialogues. The subjects were graduate students with knowledge of computer science and linguistics. Subjects performed a linear discourse segmentation, grouping phrases by their purpose. Subjects were limited to one of five purposes for each segment (e.g., list movies, show times). Phrases were also tagged with one of four possible topics (e.g., location, movie title). The segment purposes, topic tags, and part-of-speech information were used to train a finite-state machine to segment the dialogues. The finite state grammar was only able to segment 56% of the dialogues; of these, the agreement with human annotators had a 59% precision and 66% recall for the training data. The finite state model

did not generalize for previously unseen data. In addition, the finite-state machine generated several possible segmentations for each dialogue and only the one that most closely matched the human annotators' segmentations was used in the evaluation.

2.3.3 Acoustic and Text-Based Studies

Passonneau and Litman [Passonneau and Litman 1993], like Hirschberg and Grosz [Hirschberg and Grosz 1992] studied the reliability of human labeling of discourse segmentation. In Passonneau and Litman's study, subjects segmented narrative monologues, indicating where the speaker finished one "communicative task" and started another one. The average agreement among subjects for the placement of segment boundaries was 73%. The subjects' discourse segmentations were then used in an analysis of linguistic correlates of discourse structure. Three simple algorithms were tested using hand-labeled data—one based on referential noun phrases, one based on cue words, and one using pauses. The cue word and pause algorithms had higher recall (72 and 92% respectively) than the one based on noun phrases (66%) but considerably higher error rates. The cue word algorithm is based on Hirschberg and Litman's findings that discourse uses of cue words tend to occur at the beginning of prosodic phrases [Hirschberg and Litman 1993]. In more recent work, Passonneau and Litman have combined multiple cues for recognizing discourse segment boundaries [Passonneau and Litman 1997]. This work is discussed in more detail in Section 5.10.5.5 and compared with an algorithm for predicting discourse segment boundaries developed for the Audio Notebook.

Several recent studies have analyzed acoustic correlates of discourse structure in human-computer dialogues. Swertz and Ostendorf analyzed data taken from the ATIS (Air Travel Information Service) corpus to determine if speakers prosodically marked the start of new topics [Swertz and Ostendorf 1997]. Note that the ATIS system prints responses to spoken queries on the computer screen, so the "dialogues" are actually a mixture of spoken input and text output. In their study, subjects performing discourse segmentation categorized the purpose of an utterance into one of a fixed number of classes using the text alone (i.e., segmenters did not listen to the speech recordings). They found that discourse segment initial utterances were longer in duration, were preceded by a longer time interval (the time between the display of the system's response and the next query), and had a higher pitch range than non-initial utterances. Swertz and Ostendorf did not find an effect for speaking rate. Two text-based cues were also studied. They found that there were more words in segment initial queries than for non-initial ones. Cue phrases were also studied, but rarely occurred in these human-computer dialogues. However, they did occur more often in segment-initial utterances than in non-initial ones.

3. User-Structured Audio: The Audio Notebook

While attending a lecture or meeting, a listener cannot write down a complete transcript of what is said. Oftentimes listeners may be missing critical information from their notes, want more detail for a particular topic, or need the ability to review the original material. While a cassette recorder can be used to capture a verbatim record, it is difficult to find the desired portions of speech, skim the recording, or correlate it with one's written notes.

This thesis proposes a new device for taking notes and interacting with a speech recording. The device—*a paper-based audio notebook*—is a combination of a digital audio recorder and paper notebook, all in one device. The Audio Notebook *augments* a paper notebook, synchronizing the user's handwritten notes with an audio recording of the material. The user's natural activity, writing and page turns, implicitly indexes the audio. This is referred to as *user-structured* audio in this thesis.

The Audio Notebook provides an innovative and compelling testbed for exploring new audio interaction techniques. This chapter describes two Audio Notebook prototypes, version 1 and version 2. After developing the version 1 prototype, a small observational study of users in real settings was performed. The Audio Notebook did not interfere with the user's normal interactions yet gave reassurance that key ideas could be accessed later. Next, a second Audio Notebook prototype was developed. The goal was to make the design more robust to allow longer term field testing (Chapter 4).

3.1 Interface Design Approach

This thesis defines an approach for *user-structured* audio. This approach follows a user-centered design philosophy of adapting the interface to the user and *not* forcing the user to adapt to the system. The following lists several important design issues and briefly summarizes the approach used in developing the Audio Notebook. As a whole this approach differs from previous work in this area (described in more detail in Chapter 2).

- **Paper versus LCD Display**. When demonstrating the Audio Notebook at the Media Laboratory, a lot of visitors ask, why didn't you implement this on a Newton (or similar screen-based computer)? The design philosophy for using paper and pen versus a computer display were discussed in detail in Section 1.2.2. This design decision is an important part of the user-centered design philosophy. Users are not forced to throw away their current notetaking methods. Even in a highly technical place like MIT, people are still taking notes with paper and pen. Rather than replacing these useful artifacts, this thesis seeks to augment them. Ultimately, physical and virtual interfaces will be coupled [Wellner et al. 1993]. Some tasks can be performed using a physical interface (e.g., notetaking) while others can be performed using a virtual interface (e.g., searching), and others will use a combination of the two.
- **Familiar Objects versus Tape Recorder Controls**. This thesis seeks to create a new, yet more familiar and efficient interface for navigating through an audio recording. Familiar objects are used instead of tape recorder-style controls left over from analog devices. Many

digital audio interfaces simply mimic analog ones, using traditional controls like fast-forward and rewind (or forward and backward by 10 seconds). These controls are hit and miss; users often undershoot or overshoot the desired location. In the Audio Notebook, users can directly manipulate their position in the audio using an audio scrollbar and audio cursor. Users can see their location on the physical control and easily move the cursor to quickly repeat a portion of audio for example. The audio is segmented into manageable chunks based on the user's own page turns. All controls can be operated using the pen. Button controls are activated by dipping the pen inside them. The user can interactively adjust the speed of playback by sliding the pen along a speed control printed on the bottom of the notebook pages.

- **Implicit versus Explicit Indexing**. The Audio Notebook takes advantage of the user's natural activity to index an audio recording for later retrieval. The Audio Notebook does not assume any *explicit* marking of the audio on the part of the user during or after recording. While other systems force the user to mark important parts of the recording in real-time, the philosophy of the Audio Notebook is to free the listener to focus on the talker. The Audio Notebook augments the user's natural activity with automatic structuring of the audio based on acoustic cues (Chapter 5).

  Additional activities during recording (e.g., marking important parts) can burden and distract listeners from their primary task. Just as users forget to set counters on analog tape recorders [Degen et al. 1992] they are likely to forget to press a digital "importance" button[2] while absorbed in an interview, lecture, or meeting. There is a careful balance between the desire to recover pertinent information from a recording, and the amount of work it requires (during or after recording). If this balance is disturbed, than users may abandon the use of the audio recording altogether.

- **Continuous versus Selective Recording**. When first developing the Audio Notebook, some people suggested that audio should only be captured while the user is writing, saving a little before and after the pen strokes are made and discarding the rest. A significant problem with throwing away portions of the original recording is that there is no principled way to determine which part of the audio should really be retained. The listener loses the context and flow of the material when pieces are removed in this manner. In another strategy for saving storage space, users of the Dynomite system [Wilcox et al. 1997] explicitly highlight selected portions to be saved. As described by one of the Audio Notebook users, during an interview or meeting "you don't know what somebody's going to say; you don't know if it's going to be important or garbage." After making a recording, the relative importance of different portions of the material will change over time. For example, archives of past news may retain only partial information (e.g., one set of video and audio clips) and thus, only one viewpoint [Donnelly 1996]. Looking back on past events, the "important" or relevant parts may differ from what was considered important at the time.

  This thesis argues that all of the audio should be captured and saved, while providing a mechanism for the listener to quickly access portions of interest afterwards. This is accomplished using a combination of the user's natural activity and automatic structuring based on acoustic cues. Computer memory and storage are continually decreasing in physical

---

[2] Or make a special "importance" pen stroke.

size and cost. Today, gigabyte removable storage is becoming very common (e.g., Iomega's Jaz drive, Syquest's Syjet) with speeds comparable to built-in hard drives. Over 34 hours of uncompressed telephone quality speech (i.e., 8 bit, 8 kHz mu-law encoded audio) can be stored on a one gigabyte cartridge. Compression techniques can be used to increase this capacity by at least six times while retaining reasonable quality. The important problem is not how much audio to save, but how to manage the large amount of data that will be growing at an ever-increasing rate.

- **Starting versus Ending Points**. The Audio Notebook provides the user with *navigational landmarks* for randomly accessing portions of the audio recordings. The user's activity (writing and page turns) and acoustically-derived structure (phrase breaks and discourse segment beginnings) provide these landmarks. The user can move quickly and fluidly from one landmark to another. An important design decision was to select only navigational starting points, while the user determines the ending point, or where to jump to next. In a study of the SpeechSkimmer user interface, segments of speech were played back as a summary. Listeners often felt that the segments ended too soon, jumping to the next segment of speech before they could fully process what they heard. Scan controls on car radios have a similar problem—they may play too much or too little of the audio.

  The design philosophy of the Audio Notebook is to work with the listener, not automating everything, but instead providing interactive control over the audio. For example, when users make selections in their notes to begin playback, the system determines a starting point in the audio, but the listener can adjust it using the audio scrollbar. The user can then listen to as much of the audio as desired and then quickly jump to another location in the recording (see also Section 5.2).

- **Coarse and Fine-grained Control**. Another design goal is to provide the user with multiple levels of granularity for navigating through an audio recording. When I first proposed an audio scrollbar (Section 3.3.3), people were skeptical and asked "why do you need a scrollbar when the user has access to the audio from their notes?" Other systems have focused solely on the user's notetaking for indexing the audio. However, the granularity of navigation would then be completely dependent upon the density of the user's notes. This goes against a basic goal of the Audio Notebook—to free the listener to focus more on the talker. The user should not be worried about how many notes they are taking. The scrollbar gives users a way to "fine-tune" their position in the audio. In addition, the system makes suggestions of where new topics begin. This can be especially useful for time intervals where there is little or no notetaking activity (see Chapter 6).

- **Page Detection versus Manual Page Selection**. An important design goal is to allow the user to take notes naturally, without feeling encumbered by the device. Therefore, detection of the user's notepad, page number, and page turns must be seamless. If the user was required to manually select the page number, the interface would be moded. Each page would represent a different mode, and the user might forget to change it. The page detection should be transparent, allowing users to turn pages in their notebook as naturally as they would turn pages in any book. A challenge in augmenting physical objects is to take advantage of the user's knowledge about how they work, without overloading their meaning or creating artificial gestural languages.

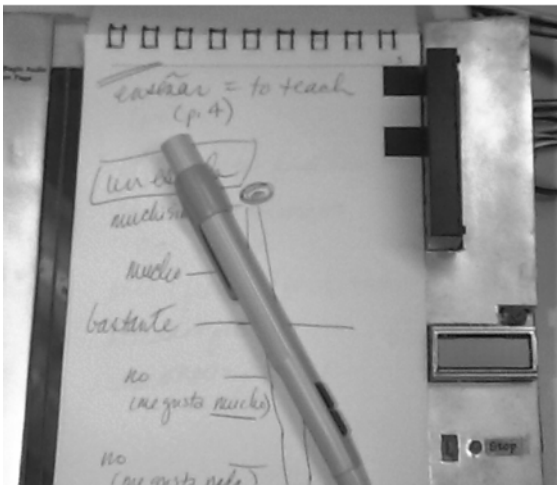
3.2 Functional Requirements

The following is an initial list of requirements that were proposed for the Audio Notebook hardware and software:

- The ability to record, store, and playback at least three hours of audio.
- The ability to change the speed of playback (i.e., time-compression).
- Physical controls (e.g., buttons, sliders, knobs) for user input. These could be soft controls (i.e., implemented as areas on a digitizing tablet or touch pad) or hard buttons.
- Detect when the user is writing on a page.
- Detect which page the user is writing on.
- Detect which notepad the user is writing in.
- Detect when the user turns the page.
- Digitize writing in the notebook.
- Allow the user to point to a particular location on a page to initiate playback.

All of these requirements were met in the first Audio Notebook prototype, except for time-compression control. The storage capacity was increased in the second prototype by using removable storage cartridges. Each cartridge can store up to 12 hours of audio (Section 3.5.1). No form of audio processing was integrated into the Audio Notebook until after a period of user study and field testing. The goal was to start out with *user structuring* of the audio alone, and then to design the audio processing techniques to best augment the user's activity and tasks based on observations of users.

3.3 Audio Notebook Version 1



3.3.1 Architecture of Version 1

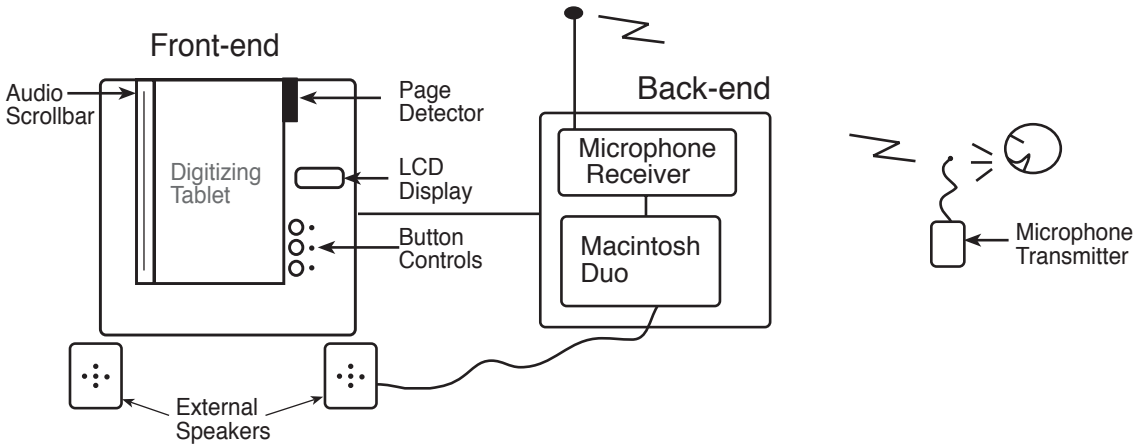


Figure 3-1: Architecture for Audio Notebook prototype version 1.

Figure 3-1 shows the hardware architecture for the first Audio Notebook prototype. The user interacts with the "front-end" of the prototype (Figure 3-2). At the base of the front-end is a digitizing tablet which captures the X-Y location and pressure of the user's pen. The tablet used in this prototype (made by Kurta) is 8.5" x 11" with a 6" x 8" active drawing area. A U-shaped cover made of foam core creates a slot over the tablet's active area for a 5.5" x 9" notepad. The user takes notes using a cordless digitizing pen with an ink cartridge (Figure 3-3). The tablet can sense through a notebook of about 60 pages.

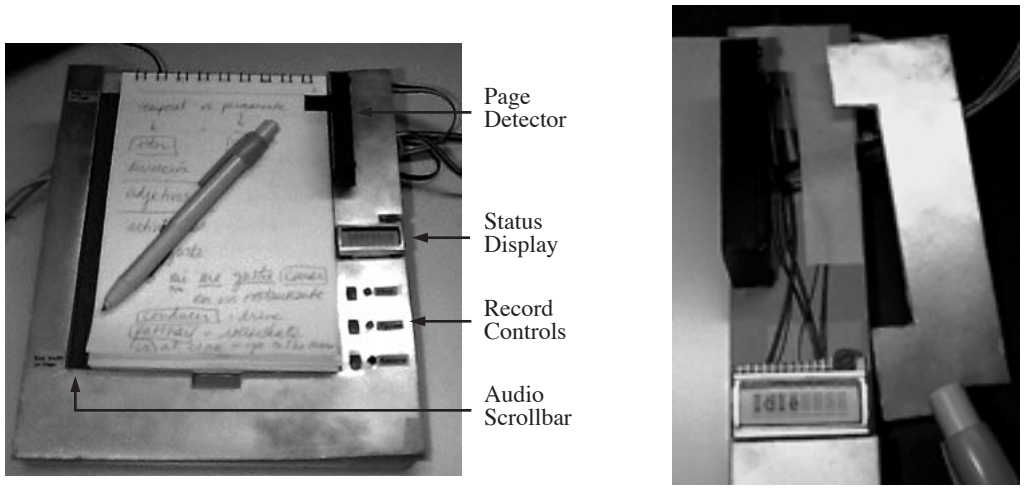


Figure 3-2: Front-end of Audio Notebook version 1. The exterior is made of foam core and painted with a silver finish. The wires to the LCD display and page detector (photo on the right) run underneath the foam core exterior.

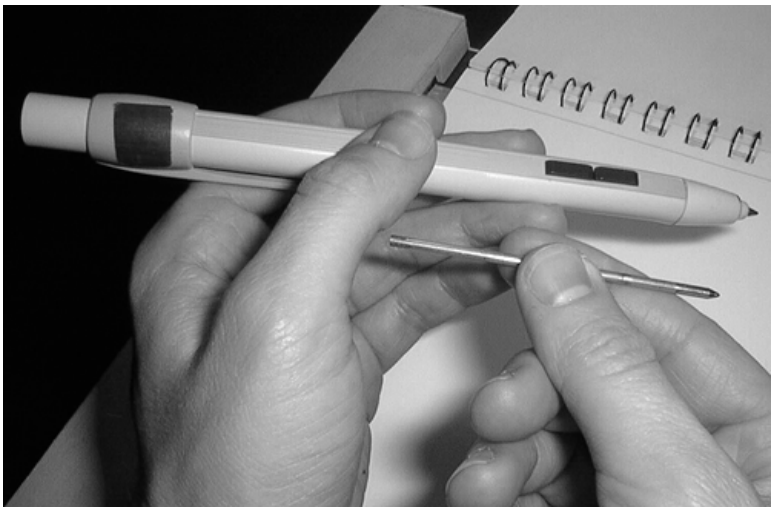


Figure 3-3: Notes are taken with a digitizing pen equipped with an ink cartridge.

The front-end of the prototype contains four user interface controls—an audio scrollbar (Section 3.3.3) and three button controls. The audio scrollbar is located to the left of the area for the notepad. The scrollbar is a made out of a strip of clear plastic that is placed over the digitizing tablet. A groove in the plastic acts as an affordance for the pen. As the user slides his/her pen along the groove, the position of the pen is tracked by the tablet underneath the scrollbar. On the right side of the notepad are three button controls—record, pause, and stop. These controls are placed on top of function keys on the tablet. The user dips the pen inside one of the button controls, activating the associated function key on the tablet. Attached to the bottom of the tablet is a Motorola 6811 microprocessor. The 6811 is used to control the LCD display, status LEDs, and page detection sensors (Section 3.3.1.1). A 2 x 8 character LCD display gives status information about recording and playback.

Figures 3-4 through 3-6 show some early design sketches of the front-end for the Audio Notebook prototype. These drawings suggest a slot for removable storage of the audio recordings. PCMCIA was considered because of its portability but rejected because of the high cost of the

storage cards. In addition, “fish” sensors [Zimmerman et al. 1995] were initially considered for tracking the user’s hand location over the notepad. The idea was to allow a user to point with his/her hand to different locations in their notes. The digitizing tablet was chosen for the final design because of its high resolution and accuracy.

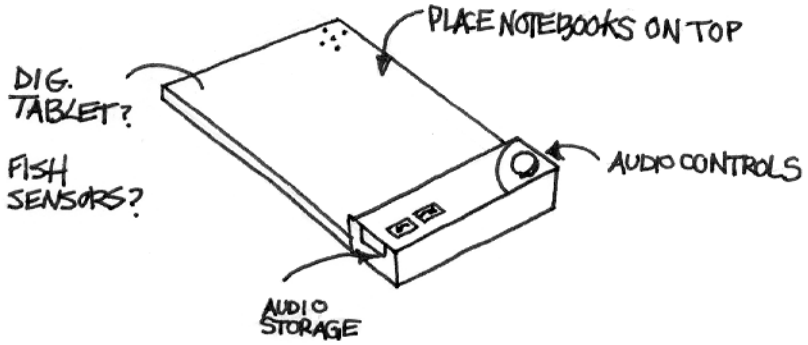


Figure 3-4: Early design sketch of the front-end of the Audio Notebook.

Figure 3-5: Early design sketch of the front-end of the Audio Notebook showing the notepad placed on top of the base unit.

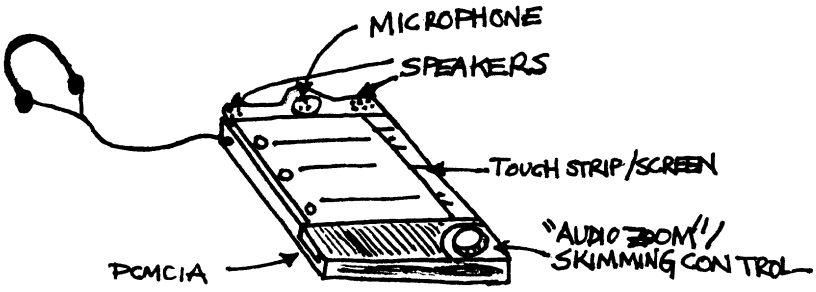


Figure 3-6: Early design sketch of the front-end of the Audio Notebook. The notepad is held in the device using a clip-board-like holder. A microphone and speakers are built into the notepad holder, with audio controls and storage slot at the base of the device.

The “back-end” of the version 1 prototype is composed of a Macintosh Duo, and a Lectrosonics wireless microphone receiver. The front- and back-end of the prototype are connected by a 5 foot tether. This tether contains an ADB (Apple Desktop Bus) connector from the digitizing tablet to the Macintosh, and a serial connector from the 6811 to the Macintosh. The Macintosh handles recording, playback, and storage of pen data. Lastly, the wireless microphone unit is composed of a belt-worn transmitter and a clip-on microphone element. The audio signal is transmitted to the

microphone receiver and then digitized and stored on the Macintosh. The audio is recorded at 8 bits linear[3] 22 kHz. In version 1 of the design, the audio was recorded directly onto the internal drive of the Duo which could store approximately three hours of audio. After a recording was made, the audio was transferred to a desktop machine to make room for the next recording.

3.3.1.1 Page Detection

During recording and playback, the Audio Notebook automatically recognizes the top page of the notepad. Standard bar code readers were too large for use in this prototype and require manual input by pointing or swiping. Therefore, a specialized page detector was developed (Figure 3-7). In version 1 of the design, page numbers were coded using a six bit binary code, and printed along the side of each page using black and white squares. The code is read using six optical sensors that are angled down at the page. An all white code (i.e., no black squares) is used to indicate that there is no notepad in the device. The user can randomly access any page of the notepad—the pages do not have to be turned one at a time.

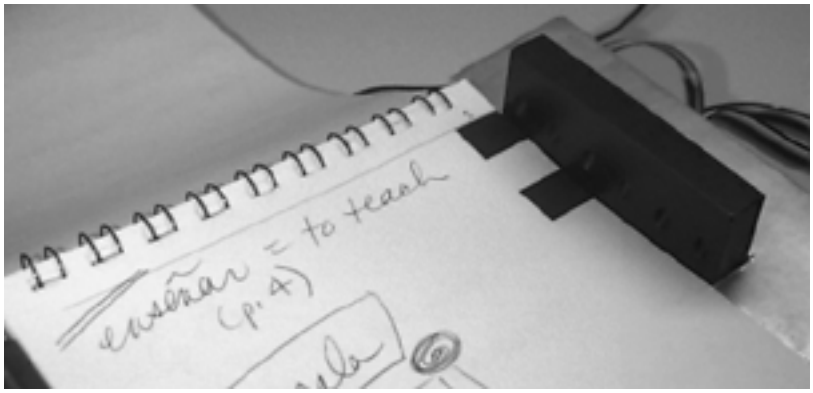


Figure 3-7: Audio Notebook version 1 page detector.

The notepads were printed on an ordinary laser printer using a simple postscript program. They were then cut and bound at a local print shop.

3.3.1.2 Alternate Architectures

Two other methods were considered for implementing the Audio Notebook prototype: (1) using a wireless tether to a desktop or laptop computer; (2) building a standalone device. Alternate method (1) was initially considered as shown in Figure 3-8. This configuration would be useful in a one-to-many recording situation (i.e., when one recording can be used by many listeners) such as recordings of lectures or meetings. For classroom lectures, a single machine could record the audio, while each user creates personal links to the recording through their notetaking activity.

[3]Note that the Macintosh Duo 230 can only record or playback one channel of audio and is only capable of recording with 8 bit linear encoding.

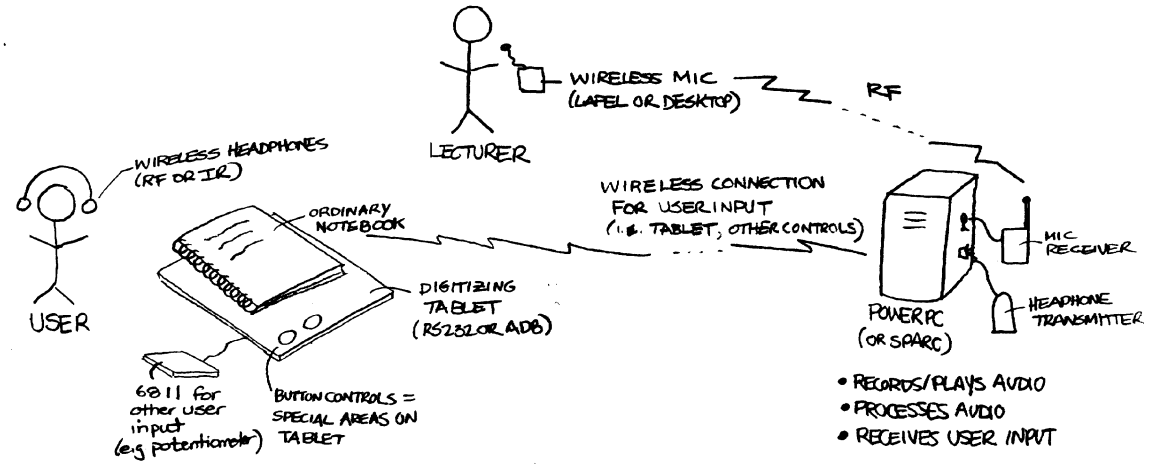


Figure 3-8: One proposal for implementing the Audio Notebook prototype. The audio would be recorded using a wireless microphone worn by the talker. The microphone receiver would be connected to a back-end processor (e.g., PowerPC or Sun Sparcstation) which digitizes the audio. The user interface would have two components: an ordinary paper notebook and an audio notebook recorder/player. The audio notebook recorder/player consists of a digitizing tablet, audio controls, a small microprocessor (e.g., Motorola 6811) for collecting user input, and a mechanism for wirelessly transmitting this input to the back-end processor. User input (pen strokes, button presses) are wirelessly transmitted to the back-end processor.

Both of these methods required hardware development outside the scope of this thesis. The tethered prototype design was selected because it could be rapidly prototyped and was robust enough for user study and field testing (Chapter 4).

3.3.2 Taking Notes and Recording Audio

The user writes in an ordinary paper notebook, while the audio of a lecture or meeting is recorded digitally. The user slides or places the notepad into the device as shown in Figure 3-9). As the user takes notes, the Audio Notebook captures which page the user is writing on, the X-Y location on the page, the pressure of the pen, and a time offset into the audio recording.

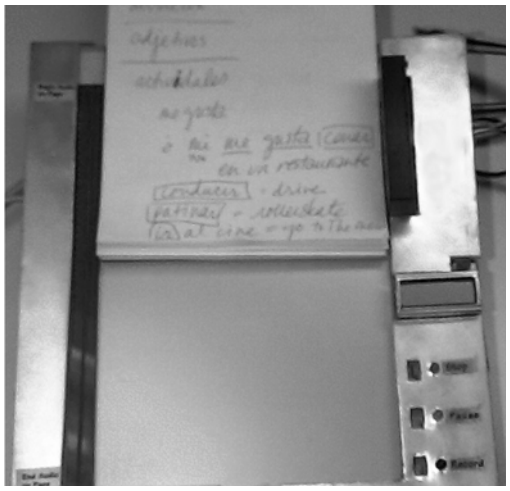
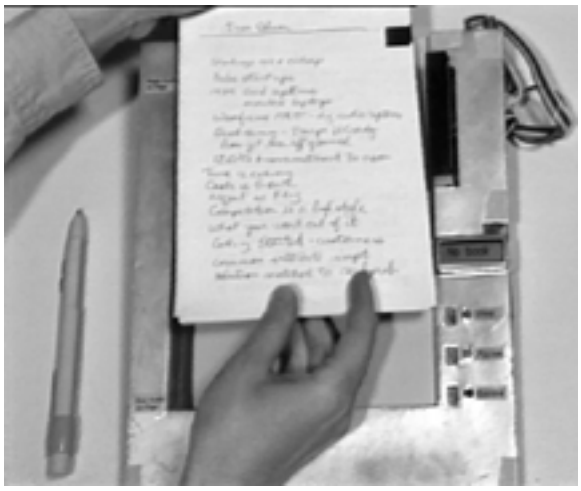
a.  b. 

Figure 3-9: The user slides (a) or places (b) a notepad into the device (version 1 prototype).

The user can start and stop the recording as desired. Button controls for starting and pausing the recording are located on the right side of the notepad, and are activated by dipping the pen inside them (Figure 3-10). All interface elements can be controlled using a digital ink pen or stylus. This creates a very different "look and feel" from a traditional cassette recorder. The idea is to create a more natural and intuitive interface by using familiar objects like pen and paper.

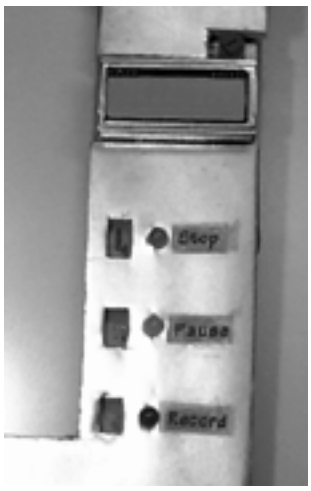


Figure 3-10: Version 1 Audio Notebook audio controls (record, pause, stop) are activated by dipping the pen inside them.

The current page number and state of interaction (e.g., recording status, playback time) are shown on a small LCD display. A study of the VoiceNotes speech interface for a hand-held notetaking device indicated that visual feedback is important for representing state information (e.g., does the system hear what I'm saying?). In addition, while recording a lecture, interview, or meeting, it is important for the device to provide feedback silently. LEDs are also used to indicate the state of interaction since they may be more likely to catch the user's attention when his/her visual focus is on the notepad. In the version 1 design there were three LEDs—red for recording, yellow for paused, and green for stopped.

3.3.3 Reviewing Notes and Audio

After a recording is made, audio can be accessed by space or by time using the following three methods for navigation:

1. **Page Selection**. The spatial layout of writing on a page provides a way of rapidly navigating through a recording. The user can point to a location in his/her notes with the pen and hear the associated portion of the audio recording (Figure 3-11). Playback is triggered by the location and pressure of the pen. The system find the closest X-Y location to the user's selection in the stored pen data for the page of notes. Each stored X-Y coordinate has an associated time point in the audio recording. Playback begins a few seconds prior to this time point to take into account the delay between listening and writing. In the first prototype, this *listening-to-writing offset* was fixed at 2 seconds for all users and recordings. In the final Audio Notebook design, a multi-part approach has been developed for addressing the problem of correlating the user's notes with the audio recordings (see Sections 4.7.2 and 6.3).

Figure 3-11: The user can randomly access different parts of the audio recording by pointing to a location in his/her/notes (version 1 prototype).

2. **Audio Scrollbar.** The audio scrollbar provides a timeline of the audio associated with each page of notes. The audio scrollbar is a physical instantiation of a graphical user interface scrollbar, providing truly *direct* manipulation. The user can quickly navigate through the timeline by dragging his/her pen along the scrollbar (Figures 3-12). Users in a study of the SpeechSkimmer interface [Arons 1997] requested a timeline for navigating through the audio recordings. For example, users wanted the ability to quickly jump to the middle of a recording. For the Audio Notebook, after glancing through a notepad, a user may find a particular topic that he/she needs to review in more detail. If a page of notes is sparse or there is a lot of time between the notes taken, the notes alone will not be an adequate index for reviewing the audio. The scrollbar can provide a finer-grained control over the audio than selection on a page. The timeline represents only one page at a time rather than all the audio recordings in a notebook. An entire notebook can contain 10–12 hours of speech—too much audio to be represented in the scrollbar while retaining high resolution control. Page turns provide an important index for segmenting the audio recording into more manageable chunks.

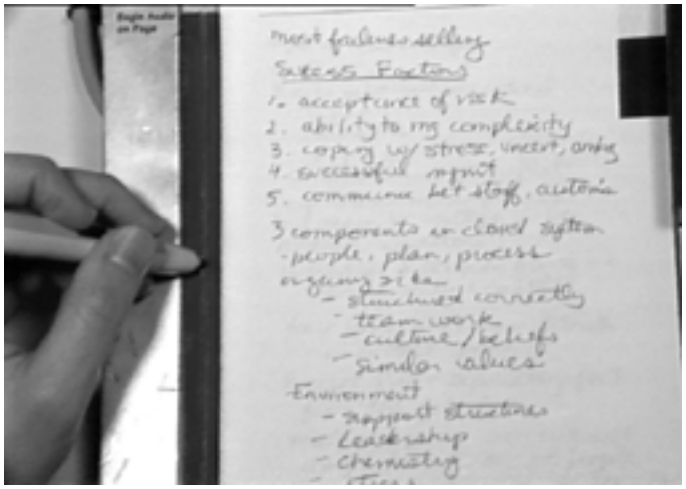


Figure 3-12: The audio scrollbar provides a timeline of the audio associated with each page of notes for fine-grained navigational control (version 1 prototype).

3. **Page Turns**. Another way to navigate through an audio recording is simply by turning the page (Figure 3-13). The user can flip through his/her notepad, find a topic of interest, and go directly to that point in the recording just by turning to that page of notes. Each page has segments of audio associated with it based on when the user was turned to that page when

originally making the recording. Observations of user's notetaking habits indicated that page turns often coincided with topic changes (Chapter 4).

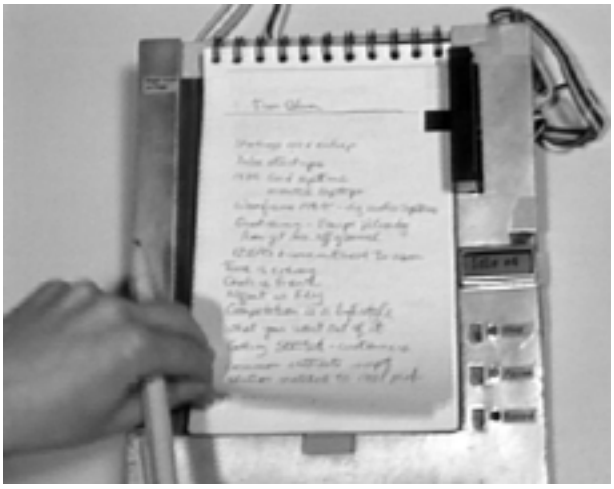


Figure 3-13: Page turns provide important indices into the audio recording. The user can randomly access a topic just by turning to a page in his/her notepad (version 1 prototype).

3.3.4 Ergonomic Design

Careful consideration was taken in the design and organization of all audio controls and interface hardware. The following is a discussion of the design variables considered. The interaction among the design variables is also critical. One design decision impacts another, so these variables cannot be considered in isolation. The design decisions are complex, and tradeoffs had to be made.

- **Notepad Design**. The interface could be designed for a spiral-bound notepad or for sheets of paper held in a clipboard-style design (Figure 3-14). If a spiral-bound notebook is used, it can flip from the top (Figures 3-15 through 3-17) or from the side (Figures 3-18 through 3-19).

  A spiral bound notebook was selected over a clipboard design. A bound notebook provides a spatial interface; users can randomly flip through pages in a bound notepad and may recall a spatial mapping between particular topics and pages in the notepad (e.g., a user could recall a topic was towards the front, middle, or back of a notepad).

  A top-bound spiral notebook (i.e., like a traditional stenopad) was selected over a side-bound one. The notepad needs to stay in place over the tablet's active area, and so that the page codes are aligned with the optical sensors. A top-bound notepad stayed in place more easily than a side-bound one. In the version 1 design, when lying flat on a table, the device is raised up in the back. In this way, the weight of the notepad pushes it forward, and the bottom of the U-shaped cover acts as a bumper, holding the notepad in place. Another advantage of a top-bound notepad is that is takes up less space horizontally on a table. Figure 3-20 shows one person using the device on a table, and the other holding it in his hands.

  Ultimately, it would be desirable to support different styles of notepads for different types of users and preferences. For example, a student may prefer a side-bound spiral notebook, while a reporter may prefer a top-bound one. A lawyer who was interviewed said she prefers taking notes in a yellow, lined, legal-sized pad (Section 3.4).

Note that once a top-bound notepad was selected, this created constraints for the other design decisions, as discussed in the following paragraphs.

- **Page Detection Design and Layout**. The placement of the page detection sensors is dependent on the type of notepad. If the notepad is bound at the side, the page detection sensors can be placed at the top of the device (Figure 3-18). If the notepad is bound at the top, the page sensors must be placed along the bottom (Figure 3-15) or side (Figure 3-16 of the notepad. An advantage of printing the codes at the bottom of the pages is that this space is often left unused. However, if the page detection sensors are at the bottom of the device, they will most likely get in the way of the user's hand. The page sensors are more out of the way on the side of the device. Since a top-bound notepad was selected for this prototype, the sensors were placed on the side of the device.

  An alternative page detector design is to print the codes on the backs of pages and embed the sensors into the base of the device (Figure 3-17 and 3-19). This design was rejected because it forces the user to flip the pages underneath the notepad. This would mean every time a page is turned, the user has to remove the notepad from the device, and fold the notepad in half. For the clipboard design, the page sensor might be embedded in the clip that holds the pages into the device.

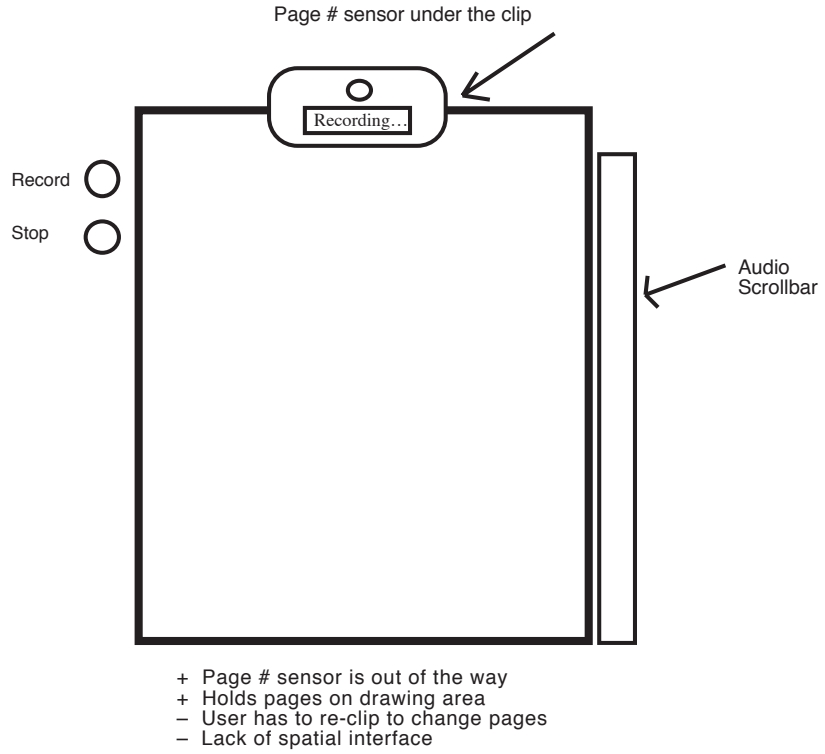
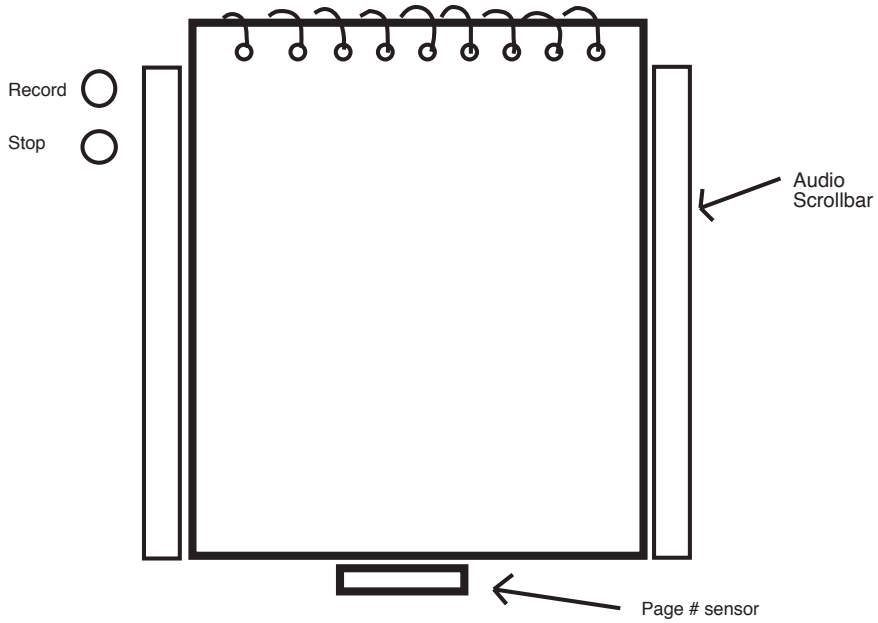
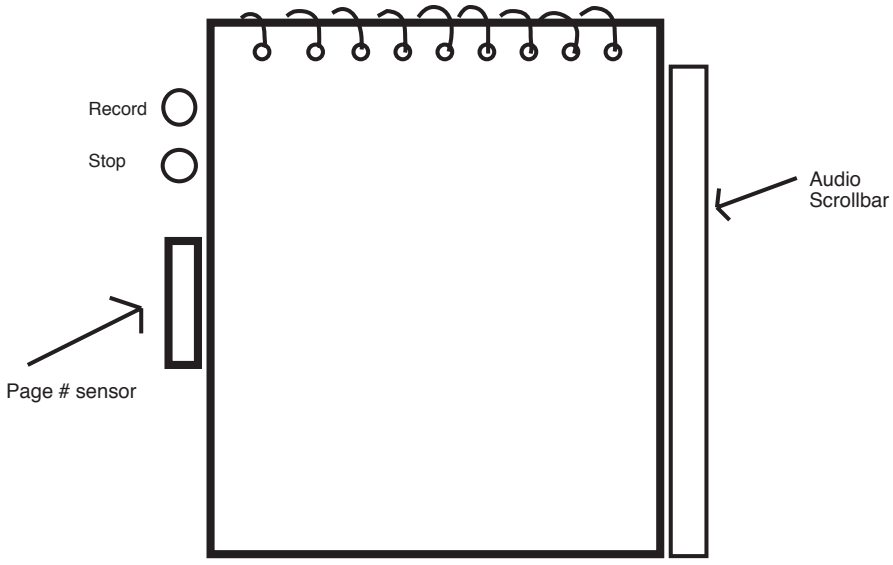


Figure 3-14: This design sketch shows a clip-board style design. The clip doubles as a page detection sensor and paper holder with embedded display.



+ Scroll bar works for right or left handed person
+ Can put other controls at the side of the page
– Page sensor cannot go at the top of the page since the spiral is in the way
– Page sensor in way of hand at bottom of page unless it could be made very flat

Figure 3-15: This design sketch shows a top-bound spiral notepad, page sensors at the bottom, and two scrollbars, one on each side of the device.



+ Can put other controls at the side of the page
+ Page sensor is out of the way more at the side
– Page sensor cannot go at the top of the page where the paper is most likely to be lying flat

Figure 3-16: This design sketch shows a top-bound spiral notepad, page sensors in the middle of the left side, and one scrollbar on the right side of the device.

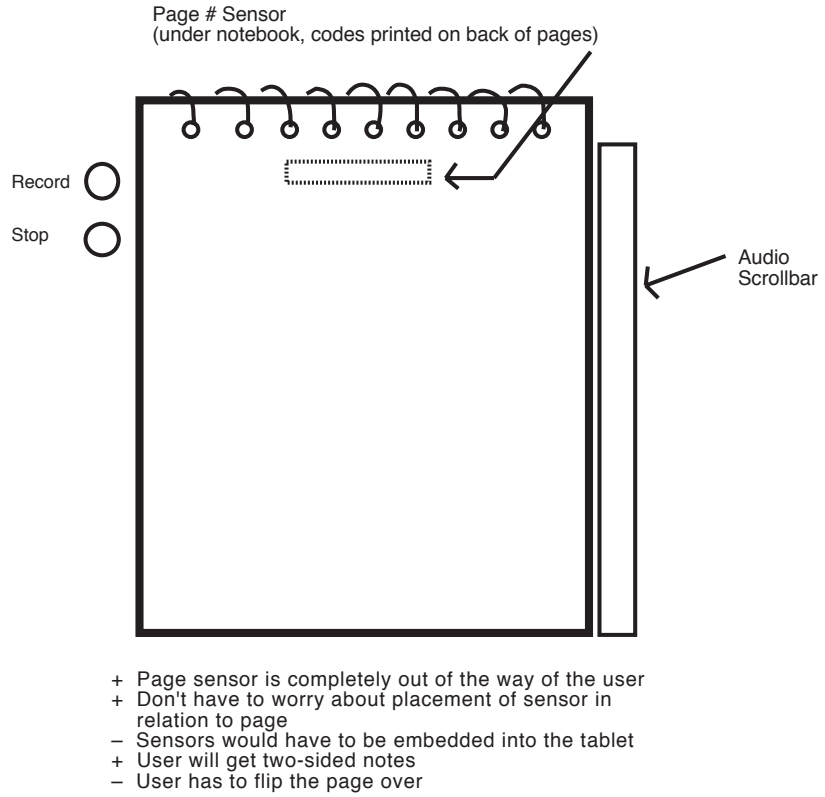


Figure 3-17: This design sketch shows a top-bound spiral notepad with page codes printed on the back of each page, and page sensors embedded in the base on the device.

- **Placement of Audio Scrollbar, Button Controls, and Status Display**. The placement of the controls is also dependent on the type of notepad. If the notepad is bound at the top, the controls can be placed on either side of the notepad; if bound at the side, the controls can be placed at the top or right side of the device. The scrollbar was placed along the side of the notepad to allow it to be as long as possible. The scrollbar can be 8 inches long when it is on the side of the notepad, but it can only be 6 inches long if placed at along the top or bottom. A longer scrollbar allows finer-grained navigational control.

  The scrollbar also had to be placed on the opposite side of the notepad as the page detector. Both the scrollbar and the page detector needed to be directly adjacent to the notepad; the scrollbar needed to be over the tablet's active area, and the page detector had to be adjacent to the codes printed on the side of the pages. The codes were printed on the upper right side of the pages to be the least intrusive to the user's notetaking. Therefore the scrollbar was placed on the left side of the notepad. This leads some people to think the scrollbar has been designed for left-handed users. However, there is a tradeoff; although right-handed users have to reach across the notepad to access the scrollbar, they also have more hand support since they can rest their hand on the notepad. Note that given a side-bound notepad, the scrollbar could be placed at the top of the device and therefore equally usable for either left or right-handed users. However, this would result in a loss of granularity, since there is less space for the scrollbar along the top of the notepad than along the side.

  The placement of the buttons and LCD display were constrained by the design of the tablet. The function keys used to implement the button controls are on one side of the tablet; the

buttons can either be placed on the bottom right or top left side (Figure 3-16) of the device depending on the orientation of the tablet. As shown in Figure 3-16, if the buttons are placed on the top left, the page detector has to be placed toward the middle or bottom of the page. The user would have to reach over the page detector to access the button controls. In addition, the page codes may be obtrusive in the middle (left side) of the page. Therefore, the alternate configuration was selected with the buttons and LCD display on the bottom right below the page detector, and on the opposite side from the scrollbar. In this way the user will not block the display while using the scrollbar or have to reach over the buttons to access it. However, the user must move his/her hand a greater distance to reach the stop button after using the scrollbar.

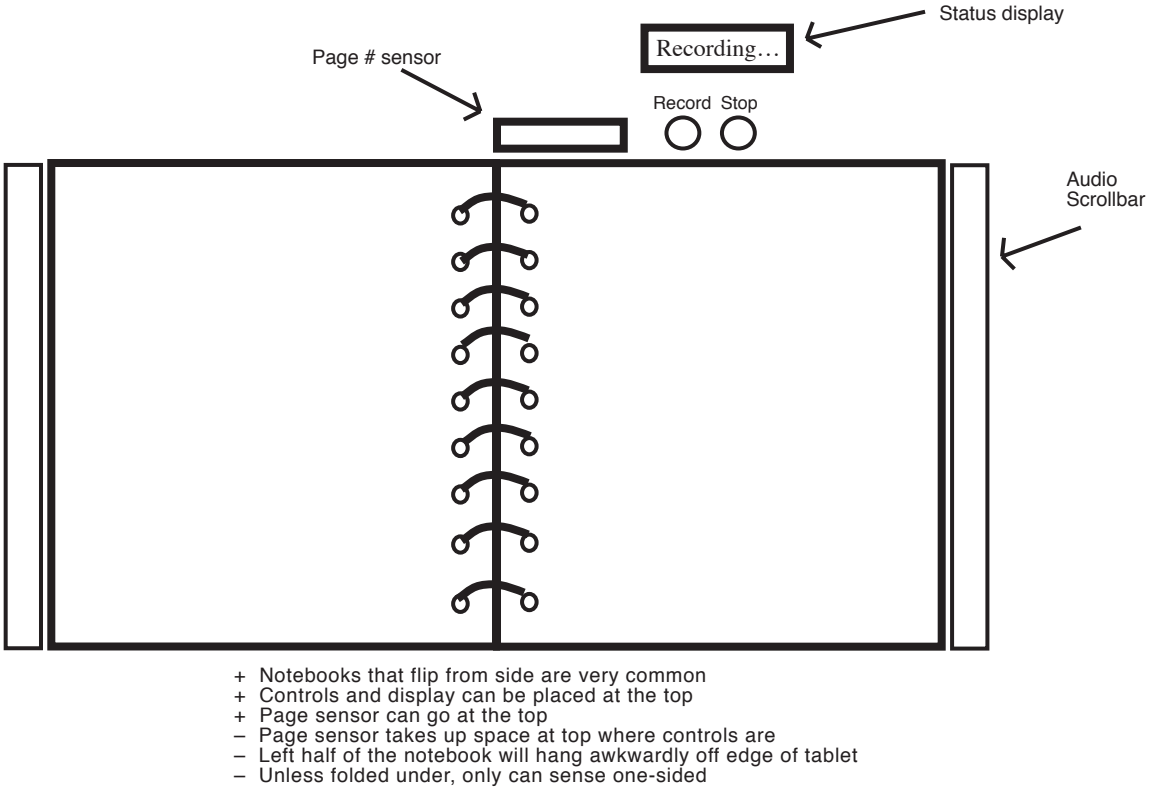
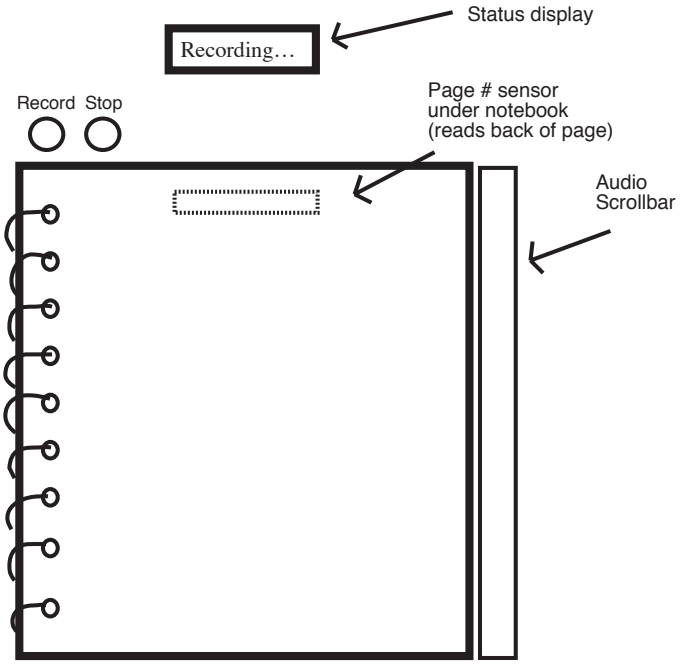


Figure 3-18: This design sketch shows a side-bound spiral notepad with scrollbars for both left and right pages. Page detection sensors, button controls, and display are at the top of the device, above the notepad.



+ Notebooks that flip from side are very common
+ Controls and display can be placed at the top
– Scroll bar can only be placed on the right hand side
– User must turn the left page under in order for the page sensor to work
– The notebook may slide out of place

Figure 3-19: This design sketch shows a side-bound spiral notepad with page codes printed on the back of each page, and page sensors embedded in the base on the device.

Figure 3-20: A top-bound notepad is easily held in place when the device is used on a table-top or held in the user's hands (note that these photos show two subjects from the field study of the version 2 design, which also uses a top-bound style notepad).

- **Finger vs. Pen Control of Scrollbar**. One design decision was whether the scrollbar should be controlled with a finger or pen. In early design sketches (Figure 3-21), finger control of the scrollbar was considered. This has the advantage of allowing two-handed use of the device. The user could hold the pen in one hand, and navigate through the audio using a finger on the other hand. Force sensitive resistors were experimented with for implementing a finger-controlled scrollbar or time-compression control (Figure 3-22). On the other hand, if all interface elements (i.e., scrollbar, buttons) can be controlled using the pen, then the device only requires one hand to operate. The user can write, scroll, and stop playback, all just by using the pen. For this reason, and to simplify the hardware design, pen control was selected. Even this design theoretically allows two-handed use; one subject in the field study used a pen in one hand to write notes and a stylus in the other hand to control the audio scrollbar (Section 4.4).

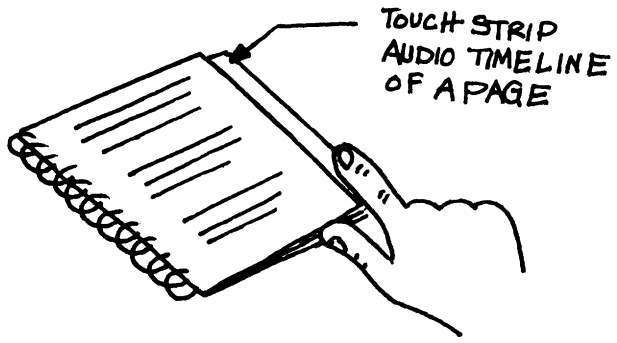


Figure 3-21: Early design sketch of the audio scrollbar.

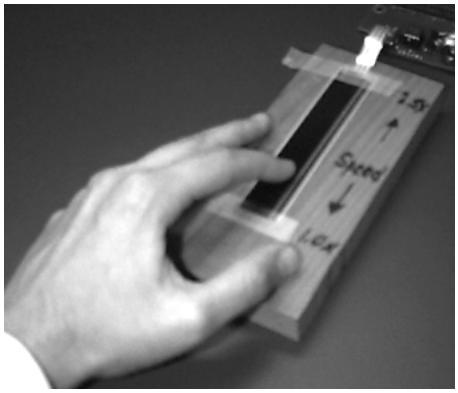


Figure 3-22: A finger-controlled time-compression slider was implemented using a force sensitive resistor as an experiment.

Figure 3-23 shows some ideas for the design of the Audio Notebook pen. For the version 1 design, an off-the-shelf tablet and digitizing pens were used. The pens could be adapted with a stylus or ink cartridge, but it had to be changed manually (i.e., no “clicker” is provided for changing pen points). Therefore multiple pens were ultimately used instead; three different colored pens and one with a stylus.

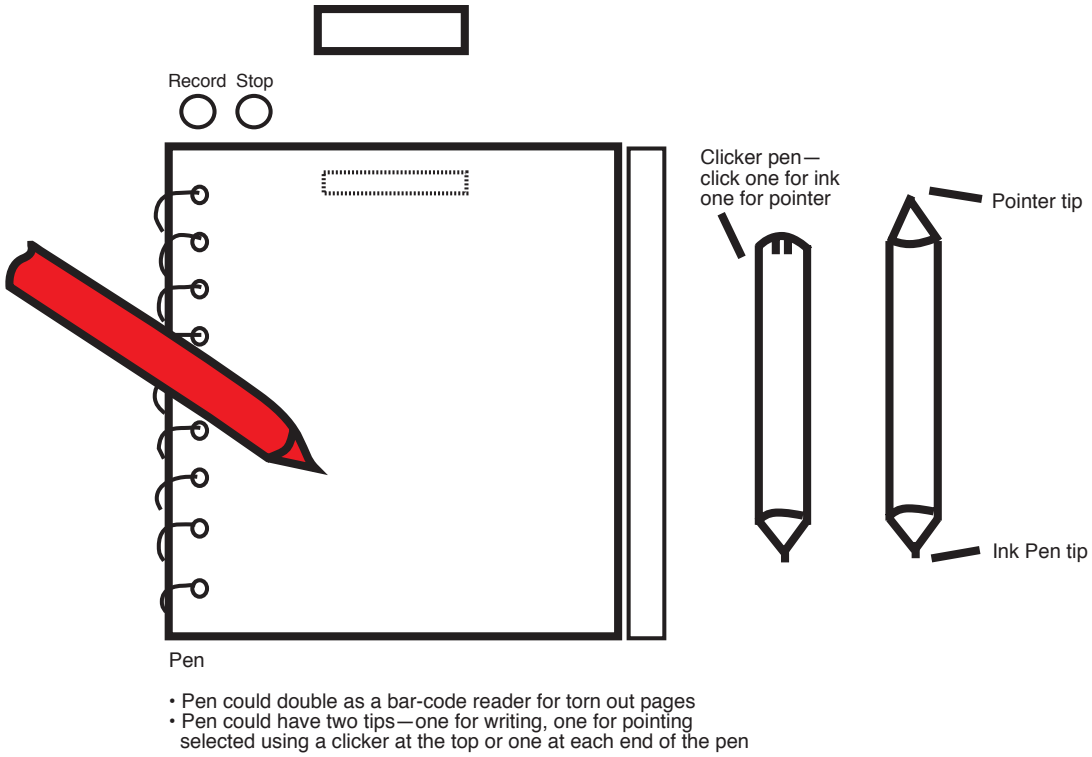


Figure 3-23: This sketch shows ideas for the design of the Audio Notebook pen.

3.4 User Study—Version 1

A small study was performed to observe use of the Audio Notebook prototype in real settings. Users who had a need to take notes and were strongly motivated to review them were selected. The study had two parts—profiles of user's current notetaking habits, and use of the Audio Notebook for taking and reviewing notes.

3.4.1 User Notetaking Profiles

To learn about the notetaking habits of different users and determine if their notetaking style changed when using the Audio Notebook, the following types of users were interviewed: a language student, a science student, a user-studies researcher, and a lawyer.

The lawyer and science student needed detailed notes. The lawyer often attempted to capture a verbatim account; she said that it was often not possible to obtain missing information from a subject after an interview. She also said that if she missed information, this could be seen as a weakness by the client. The lawyer sometimes took notes while talking to a client over the phone. In other instances, another lawyer interviewed the client, while she wrote down both sides of the conversation. She took her notes in a yellow, lined, legal-sized pad, and said she would want the Audio Notebook to work with this kind of notebook.

The computer science student took several pages of densely spaced notes for his classes, especially math. In a recent lecture, he said the professor had spoken continuously for two hours. He said he had no trouble keeping up with the professor. However, he sometimes had difficulty writing things down while the professor was speaking and missed important information. The science student said it was particularly important to "get down" as much as possible in classes

prior to an upcoming exam. He did not review his notes on a regular basis, although he sometimes read them over before or after class. Prior to an exam, he read over all his notes; he said he did not want to miss anything. The science student never recorded a lecture before, but he imagined it would be useful "because you can get all the information… you can't miss anything." However, when asked why he did not use a cassette recorder, he said he never thought of it before, and added that he felt his current method of notetaking was "pretty" efficient; "I get down enough information."

The user-studies researcher recorded his interviews on cassette tape. During an interview, he usually took few notes, jotting down his own thoughts, rather than what subjects said. Sometimes he split his note pages; writing down topics and key phrases from the interview on one side, and his ideas on the other. He did not write anything down verbatim. As soon as possible after "returning from the field," he often wrote a "first impression" of the interview (usually 2–3 pages). He commented that the more time that passed after an interview, the less meaningful the notes were for him. While making the recordings, he did not mark down any time points of interest; "it's guaranteed every time you are going to do that [use the record counter] you forgot to put the counter on." After the interview, he often manually transcribed the tapes entirely. Using the transcripts he wrote a report or summary of the interview to be shared with the other members of his project team. He said that replaying the interview was an important part of the process of creating an account or story, even though the task of transcribing it was so time-consuming. He also commented on the importance of listening to the audio versus reading a transcript— "these are real people you're talking to, it's like hearing it fresh… it gives it that fleshiness."

The language student also took few notes; the classroom sessions involved mostly speaking and interacting with other students and the professor. The classes were held in a small room, and all the students sat around one circular table. For most classes, handouts were given out. The student took some notes on the handouts, and wrote a few other notes on some sheets of paper in her organizer (where she kept her appointments and phone numbers). She said she rarely had difficulty keeping up in class and that she was a good notetaker. She tried using a microcassette recorder but found it a pain to "lug that thing around."

3.4.2 Taking Notes with the Audio Notebook

Three of the users interviewed participated in the second part of the study—use of the Audio Notebook in real settings. Two subjects used the prototype during one of their class sessions, and one during a group meeting. Prior to the notetaking session, each user was only instructed "the Audio Notebook synchronizes your notes with an audio recording." Given only these brief instructions and no prior experience using the device, it was surprising that users altered their notetaking style.

The language student used the Audio Notebook to take notes during one of her Spanish classes. Even though she had never used the Audio Notebook before, she altered her regular notetaking habits during this first usage. The language student took more notes and wrote larger than she usually did. For example, she said "here I wrote 'adjectives' and I never would have bothered to write than down before but now knowing that it would cue the audio later, I wrote this down." She also commented—"I normally scribble in tiny little words on small pieces of paper and here I wrote much larger." Her description of her notetaking activity implies that she was unconsciously annotating points for later access—"I found myself for some reason writing down much more… I

don't know why. I guess so if I wanted to go back and review particular words, I could." In particular, she wanted to be able to review the teacher's pronunciations of different words and phrases.

The user-studies researcher used the Audio Notebook during a meeting with his project team. He had just recently completed a study and did not have any more interviews scheduled at that time, so he recorded a meeting instead. He started out trying to anticipate whenever something important was said by constantly pausing and restarting recording. He thought that he needed to conserve the amount of audio recorded. This activity (starting and stopping the recording) took his attention away from the meeting. He was struggling to determine which parts of the meeting to capture. Part way through the meeting, the researcher changed his strategy, keeping recording on at all times and using his notes to create indices for later access. "I began to realize stopping and starting was futile" he said, "you don't know what somebody's going to say; you don't know if it's going to be important or garbage." This is an example of why it is important not to assume that users will explicitly mark "important" parts of a meeting in real time. No matter how simple the marking actions are, unless they are fully integrated into the user's activity (e.g., notetaking) they will distract users from their primary task (i.e., listening, participating, and taking notes). There were also social implications to the researcher's initial strategy of starting and stopping the recording; one attendee noticed the researcher pausing the recording while he was speaking and said "why are you pausing me!"

The computer science student usually took several pages of notes in every class. However, on the day of class when he used the Audio Notebook, the Professor handed out a quiz review and instructed the students to mark their notes on these sheets of paper. Therefore, the student took very few notes in the audio notepad, creating few indices into the recording. This is an important design challenge for future Audio Notebooks—the user should be able to insert handouts into the device and take notes on these sheets of paper.

3.4.3 Reviewing Notes with the Audio Notebook

Review sessions took place in a laboratory and were video taped. The experimenter (LJS) observed the subjects as they explored the Audio Notebook interface and used it to review their notes and audio recordings. The following two sections provide a detail account of the review sessions for the language student and the user-studies researcher. Since the computer science student mostly took notes on handouts during his class, he did not review the audio.

3.4.3.1 Language Student

The language student used the Audio Notebook to review different words and phrases. Since she used her writing to index the lecture, her notes became a vocabulary lesson; she could select on a word or phrase in her notes and hear it spoken by the Spanish teacher. The language student compared the Audio Notebook to a CD player—"it seems to work like a CD-ROM kind of a disk where you don't have to scan through loads of tape, you can just go where you want and stop where you want." She also commented on the difference between starting and stopping playback with the Audio Notebook versus a tape recorder—"So whereas with a tape it matters were you stop because that determines where you're going to start again, with this kind of system it doesn't seem to matter where you stop because you can start wherever you want."

The language student explored each of the methods for navigating through the audio recording. She found it intuitive to begin playback by selecting with the pen on different locations in her

notes. For the audio scrollbar, the groove in the middle of the scrollbar provided an obvious place to slide the pen. However, the purpose of the scrollbar was not immediately apparent. At first the language student thought it was redundant with the ability to play directly from the page. However, after using the scrollbar for a short period, she commented that it was a good way to "quickly run down a page."

The language student also discovered that turning a page in the notebook triggered playback. If she was not already listening to the audio, she did not want playback to begin automatically on a page turn. She said that she might want to browse through her notes without listening to the audio. Inexperienced users also tend to want tight control over starting and stopping audio playback [Stifelman et al. 1993]. On the other hand, if she was in the middle of listening to the recording and turned the page, she liked the fact that it continued to play.

Another means of playback was suggested by the language student. She wanted to distill a list of "key notes" from writing circled or underlined in her notebook. She said "rather than flip through page by page and tapping on each word… it would be great to just click through the words" using a key note button. She imagined using this feature in conjunction with the other Audio Notebook navigation techniques, not as a replacement. For example, after listening to an item in the key note list, she wanted to be able to elaborate on it if necessary.

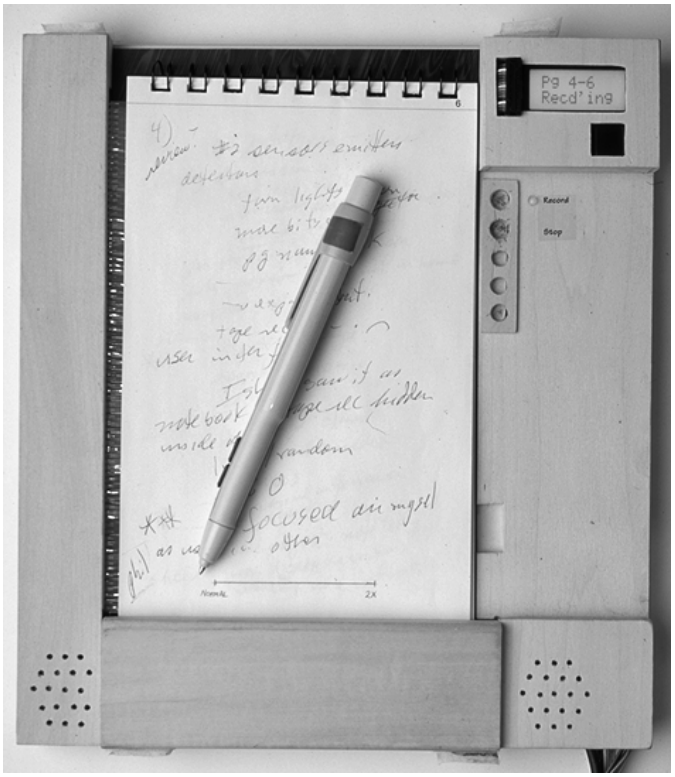
3.4.3.2 User-Studies Researcher

Like the language student, the user-studies researcher also explored each of the methods for navigating through the audio recording. When asked to postulate how he would playback the recording from a particular point in his notes, he guessed that he could press with his finger on his notes. When he was instructed that it worked similarly, but with the pen, he began to select on many different locations in his notes, listening to the audio at each point.

When asked about the audio scrollbar, the researcher noticed the "groove" and began to slide his pen through it, scrolling through the audio. However, he commented that touching places on the page was almost as fast, "assuming you have a reasonably accurate topic record." Like the language student, initially he felt the scrollbar was "slightly redundant" with playback from the notes. After further use, he determined that he could use the page selection to "move roughly to where you want and then use the scrollbar to fine tune it." However, he also noted that once he took his pen away from the scrollbar he lost his place. Displaying a correlation between time and space (e.g., lighting up a point in the scrollbar whenever the user selects on the page) would address this problem. This kind of *audio cursor* would enable the user to "fine tune" the starting point of playback (Section 3.5.4).

Lastly, the user-studies researcher also navigated through the recording by turning the notepad pages. Unlike the language student, it did not bother him that playback began automatically from the time of each page turn.

3.5 Audio Notebook Version 2



3.5.1 Architecture of Version 2

In designing a second version of the Audio Notebook, one goal was to build a more robust prototype for field testing. The prototype needed to run reliably and hold up under the stress placed on it by multiple users over several months.

The overall architecture of the second Audio Notebook is similar to the first, with most design changes focused on the front-end of the prototype (Figure 3-24). One important change to the back-end design, however, is the addition of an Iomega Jaz drive for audio storage. A one gigabyte removable Jaz cartridge is used for each audio notepad, storing up to 12 hours of audio (recorded at 8 bits linear 22 kHz).

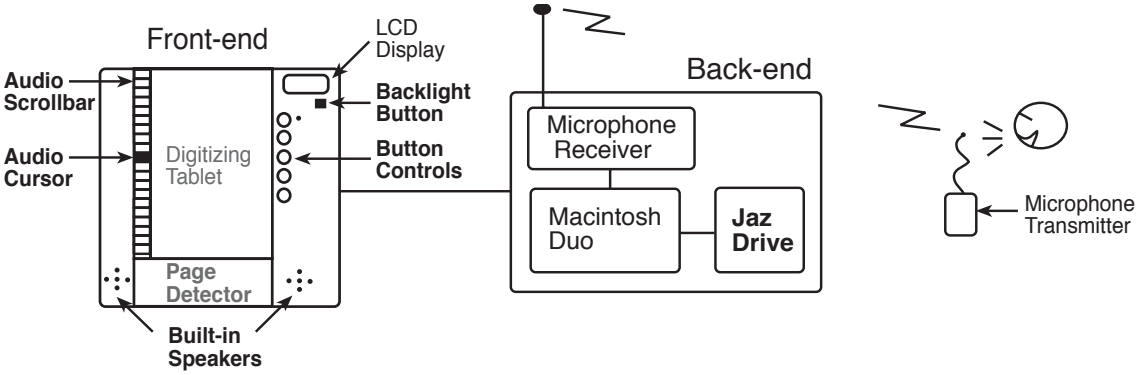


Figure 3-24: Architecture for Audio Notebook prototype version 2. Aspects of the design that were added or modified from version 1 are shown in bold.

The following sections detail the design improvements made to the front-end of the Audio Notebook.

3.5.2 Page Detection

In the first Audio Notebook design, the page detector interfered with the user's handwriting because of its size and location. The page detector was located on the top right of the notepad and was 1" tall. This height was necessary to create the appropriate angle for the optical sensors over the page codes (Figure 3-25). In addition, as the user moved his/her hand across the top of the page towards the detector, the hand or pen would block the sensors, causing page detection errors. Right-handed users would block the sensors with the right side of their hand, while left handed users would shadow the sensors with the pen. In the first design, errors also occurred if the lighting conditions were too bright or too dark. The sensors relied on ambient light to create a reflection off the page, making them sensitive to changes in the light levels.

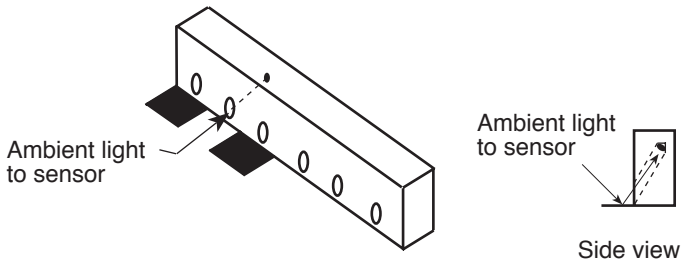


Figure 3-25: In version 1 of the page detection design, the optical sensors were angled down over the page codes.

In the version 2 design, the sensors were placed underneath a handrest at the bottom of the device. In this design, the sensors lie directly over the bottom of the page for accurate code and sensor alignment, and cannot be shadowed or blocked by the user's hand. In addition, the sensors use both an infrared emitter and detector, so they are not reliant on external lighting conditions.

3.5.2.1 Ergonomics

The ergonomic design of the page detector was significantly improved in version 2 of the Audio Notebook. However, there still remain some problems. First, it is somewhat awkward to write on the bottom of the page since the handrest is higher than the surface of the page. The handrest was higher above the notebook surface than originally anticipated because the sensors could not operate properly while resting directly on top of the page codes. Second, the notebook must be removed from the device in order to turn the pages easily. However, users can quickly and easily slide the notepad in and out of the device. In practice, these design problems caused only minor inconveniences for users in the study. Although these issues should be addressed in future designs, the page detection was robust and the design usable for field testing.

3.5.2.2 Foam Core Models

Several foam core models were built for determining how the notepad would slide underneath the handrest into position under the page sensors. Two designs were explored for use with optical sensors that needed to be a fixed distance from the page codes. In the design shown in Figure 3-26, page sensors are mounted underneath a printer-like roller. When a notepad is inserted into the Audio Notebook, the roller rides on top of the notepad, keeping the sensors a fixed distance from the page codes. A problem with this design is that a large diameter roller is needed to allow the notebook to slide smoothly underneath it. The roller needed to accommodate a 50–60 page

notebook was too large to fit under the handrest. A “snow-plow” style design is shown in Figure 3-27. In this design, as the notebook slides under the handrest, the front of the “plow” is angled such that it rides on top of the notebook. In addition, these two designs also shield the optical sensors from external light sources.

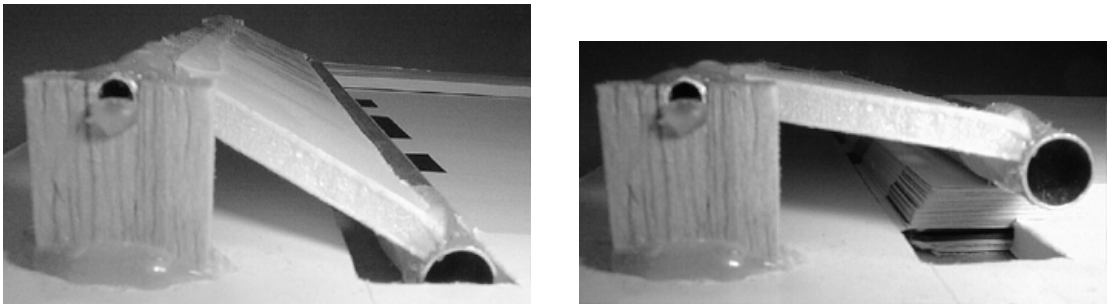


Figure 3-26: In this design, page sensors would be mounted underneath a printer-like roller. The sensors would be the same distance from the paper regardless of the number of pages under the roller. The photo on the left shows one page under the roller, the photo on the right shows 50 pages under the roller.

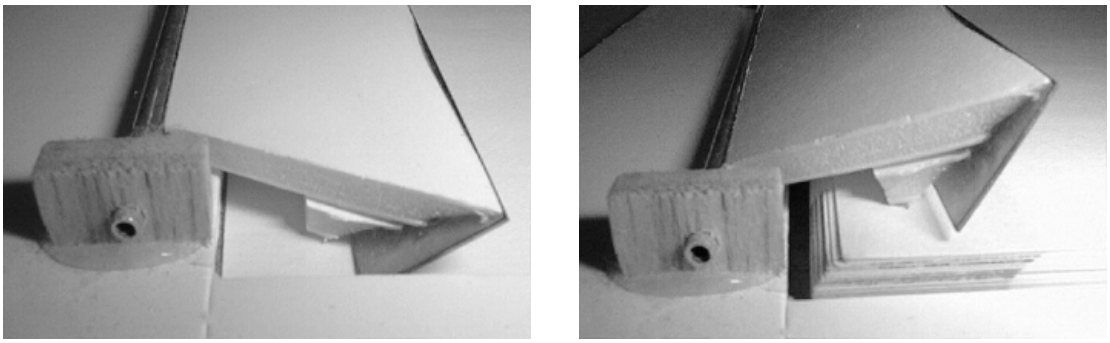


Figure 3-27: In this design, page sensors would be mounted underneath a plow-shaped mechanism. The “plow” is angled so that as the notebook slides under it, the plow ends up resting on top of the notebook. Thus, the sensors would always remain a fixed distance from the paper.

After some further experimentation, optical sensors were found that operated over a wider distance range. Therefore, the sensors no longer needed to be at a fixed distance from the page. The user could simply slide the notepad under the handrest without a mechanism to hold the pages (Figure 3-28). This simplified the design, reducing the number of components, and allowing the user to freely and easily slide the notepad in and out of the device. A working model of the handrest with page detection sensors embedded underneath was incorporated into a foam core mock-up of the Audio Notebook as shown in Figure 3-29.

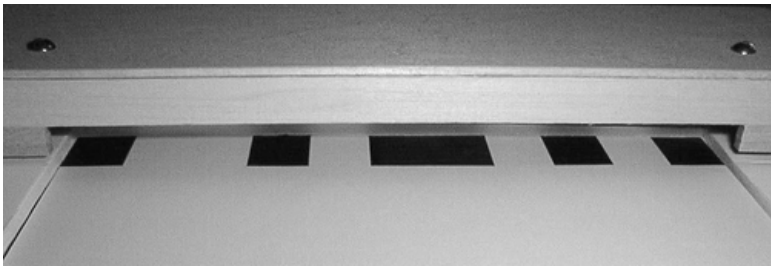


Figure 3-28: In the final design, the handrest is sized to allow a 50–60 page notepad to slide underneath. In this design, the sensors did not need to be a fixed distance from the paper.

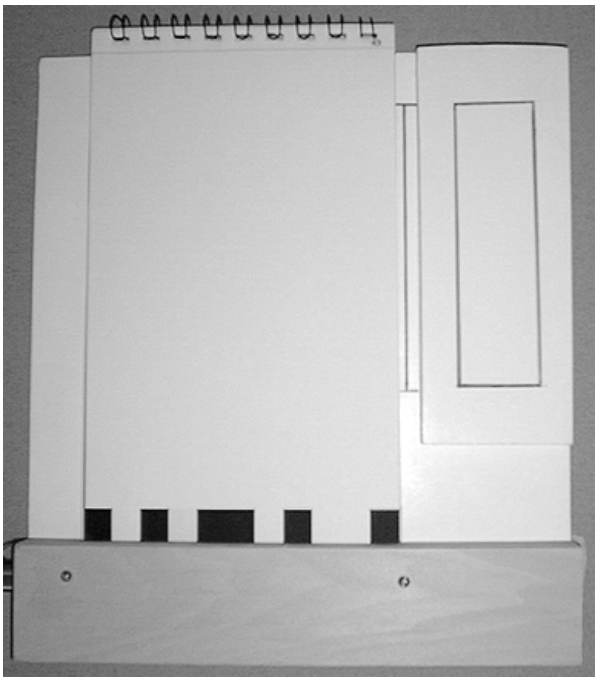


Figure 3-29: A working model of the handrest with page detection sensors embedded underneath was incorporated into a foam core mock-up of the Audio Notebook.

3.5.3 Page Codes

In version 1 of the Audio Notebook, only the page numbers were coded. In order to switch notepads, the associated file name had to be manually selected on the Macintosh back-end processor. In version 2, both the notepad and page numbers are coded on the pages. In this new design, there are 4 bits allotted for coding the notepad and 6 bits for coding pages (Figure 3-30), allowing the device to uniquely identify 16 notepads and up to 64 pages.[4] In addition, a parity bit was added for error checking, further improving the reliability of the page detection.

[4]The actual number of usable pages is 62 since 0 (all white) is used to indicate the absence of a notepad, and 64 (all black) indicates that the notepad is closed (i.e., the black cover is on top).

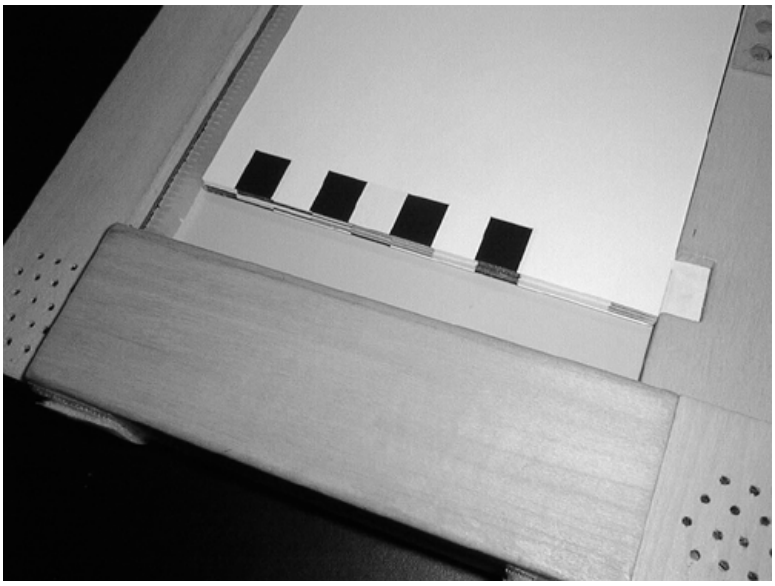


Figure 3-30: In version 2 of the page detection design, both notepad and page numbers are coded on each page. There are 11 bits—4 for notepad, 6 for page number, and 1 for error checking. Page detection sensors located underneath a handrest read the code.

An advantage of a paper notebook over a digital one is the ability to tear out pages and use them separately. By coding both the page and the notepad, pages can now be torn from a notepad while still retaining their identity.

3.5.4 Audio Scrollbar *with* Audio Cursor

In version 1 of the Audio Notebook, the audio scrollbar did not display the user's current position in the audio timeline. Once users removed their pen from the scrollbar, there was no indication of where they left off. In version 2, an *audio cursor* lights up to show the user's current position in the audio scrollbar. The scrollbar is made out of a series of 80 adjacent LEDs (Figure 3-31). The audio cursor is indicated by a green LED which moves along the scrollbar as the audio plays.

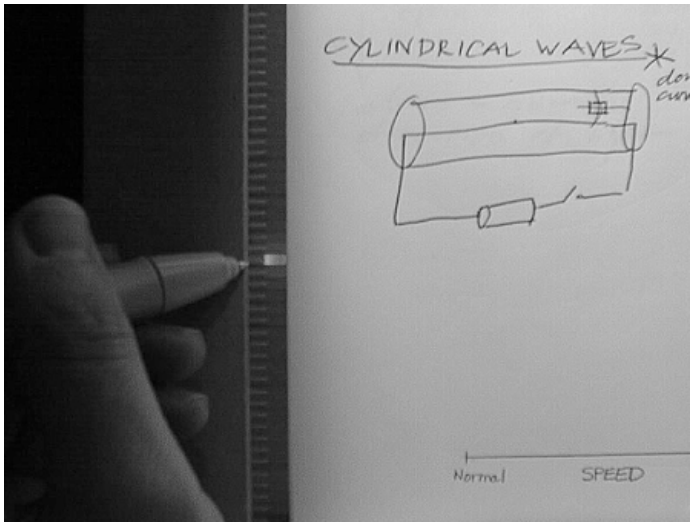


Figure 3-31: The version 2 audio scrollbar acts as both a control and a display. An audio cursor (green LED) shows the user's current position in the timeline for the page of notes.

This new audio scrollbar acts as both a control and a display. The scrollbar now serves three important functions:

1. **Fine-Tuning Playback Start Time**. First, when a user selects somewhere on a page to begin playback, the audio cursor lights up showing the corresponding location in the timeline. The user can then fine-tune the starting point of playback by moving the cursor forward or backward in the timeline (Figure 3-32). For example, a user might backup in the audio to get more context. This also gives the user a space-to-time correspondence—users can select on one note and see when it was written in relation to another.

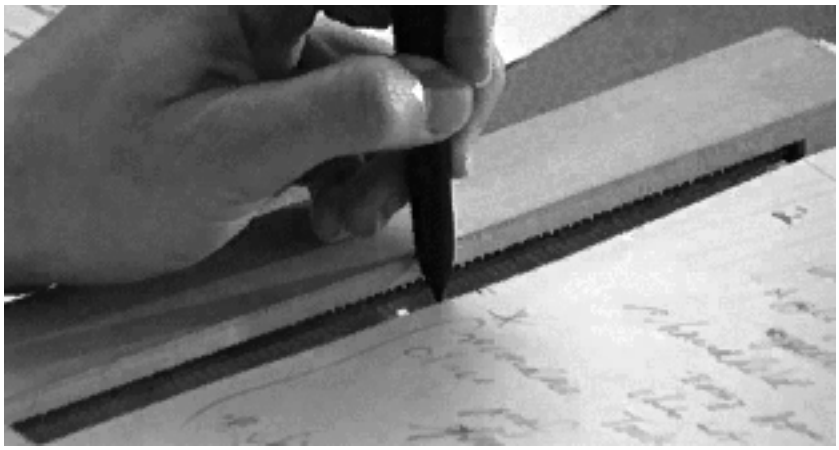


Figure 3-32: In this photo, the user fine-tunes the starting point of playback using the audio cursor and scrollbar.

2. **Fine-Grained Navigational Control**. Second, the audio scrollbar provides the user with fine-grained navigational control. For example, let's say there are two lines of notes written on a page one after the other. These notes could have been written consecutively or 10 minutes apart. In the later case, the user can select on the notes, see their corresponding place in the timeline, and move the audio cursor to navigate in between them.

3. **Audio Information Display**. Third, the scrollbar can be used to display information about the audio. The scrollbar LEDs are multi-colored; they can each be one of three colors—green, red, or yellow-orange. The intention when designing the scrollbar was to use it to display "suggestions" to the listener of places to navigate in the audio recording. For example, the system could display predictions of new topic locations (see Chapter 6).

3.5.5 Button Controls and LEDs

In version 1 of the Audio Notebook there were three button controls—record, pause, and stop, each with an LED indicator beside it. Users in the first study did not distinguish between pausing and stopping the recording. In addition, the LEDs duplicated information shown on the LCD display. In version 2, the design was simplified. In this design, there were only two button controls—record and stop. The stop button can be used to stop recording or playback.

In the version 2 design, one LED was retained to show recording status—a red light indicates that recording is on (Figure 3-33). A recording LED was retained for two reasons. First, if a user accidentally starts or stops recording, the red LED will catch his/her attention more rapidly then the display. In the user study of the first prototype, a bug in the software caused recording to turn off unexpectedly in one case. The user noticed this problem immediately because the record LED

turned off. A change in the message shown on the LCD display is not as noticeable. Secondly, the record LED tells others around the user that recording is taking place. In the first Audio Notebook user study, when the device was used in a meeting, people attending the meeting noticed when recording was on or paused. If they wanted the Audio Notebook user to stop recording at any time, they could be assured that recording was indeed turned off.

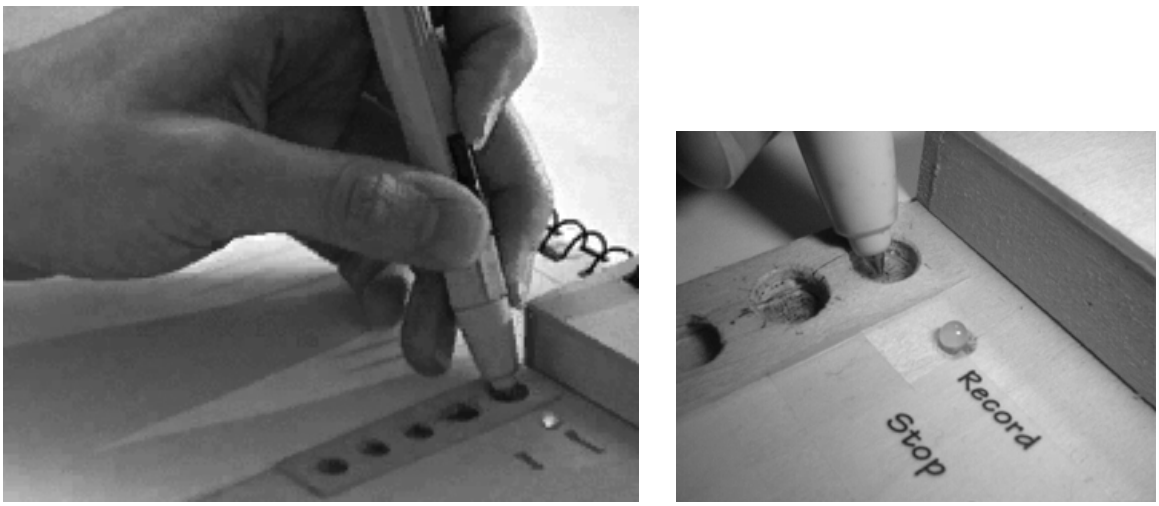


Figure 3-33: In the version 2 Audio Notebook prototype there are two active button controls—record and stop. A red LED indicates when recording is on. Three additional button controls were reserved for future additions to the feature set. Figure 6-7 shows the final design with all five buttons implemented.

4. Longitudinal Field Study

A field study of the second Audio Notebook prototype began in the fall of 1996. In this study, several students and reporters were observed using the Audio Notebook over a five month period. Users were observed during both the capture and review of their notes and audio recordings. The students and reporters used the Audio Notebook for real tasks (e.g., to write a story for publication, to study for an exam) and not artificial tasks performed in a laboratory. Rather than observing users during a single session, this study aimed to observe a small number of users in depth over time.

There were several goals of the Audio Notebook field study. One goal was to determine how the students and reporters made use of the audio recordings given the random access capabilities provided by the Audio Notebook. A second goal was to evaluate the utility of the audio interaction techniques provided by the Audio Notebook. A third goal was to observe the user's notetaking habits, and determine what changes (if any) occurred over a longer period of time than in the first user study.

4.1 Why Review the Audio?

One important issue for this field study was to consider the circumstances in which people would be motivated to review an audio recording of a lecture, meeting, or interview, even given the rapid access capabilities provided by the Audio Notebook.

In the previous user study, a language student, lawyer, and user-studies researcher each had motivations for reviewing the audio. For the language student, after a lecture, the audio allowed her to review the teacher's pronunciation of different words and phrases. For the lawyer, an accurate account of an interview was needed to best service the client's goals.

For this field study, students and reporters are two groups of people with potentially strong incentives for reviewing an audio recording made of a lecture or interview. Students are motivated to review handwritten notes when they are given problem sets or examinations associated with the material. For this reason, both classes chosen for this study included several homework assignments and quizzes given during the course of the semester. For a reporter, the audio is valuable for gathering quotes for a story, or reviewing portions of an interview that need clarification or additional detail.

4.2 Subjects and Procedure

In this field study, the Audio Notebook was used by four students and two reporters. Two of the students used the Audio Notebook regularly in their classes throughout one semester. Their classes met twice each week for 90 minutes. Two additional students in one of the classes used the Audio Notebook to review classes that they were unable to attend. Lastly, two reporters used the device to interview a subject and write a story for publication.

The classroom provided an ideal setting for studying use of the Audio Notebook over time. Two graduate students at the MIT Media Laboratory (referred to as students 1 and 2) used version 2 of

the Audio Notebook; one in a Holography class and the other in a Signals and Systems class. One student used the Audio Notebook for the entire semester. The other student started using the Audio Notebook after initially using a cassette recorder for the first few weeks of class. Prior to the first day of use, all students in the class were informed about the study. The professors agreed to be recorded and to wear a small belt-worn transmitter and lapel microphone. The students were instructed to try the Audio Notebook once or twice and then to decide whether or not to continue using it. Both students requested to continue using the device throughout the semester. After one or more classes, the student scheduled time to review their notes. Students only reviewed notes when they wanted to; they were not required to review each class. In the Signals and Systems class, two students used the Audio Notebook to review lectures that they missed (students 3 and 4). Observations of these users provides some insight into how someone who has not attended the lecture or interview interacts with the Audio Notebook.

In addition to the students, two reporters used the Audio Notebook to perform an interview and write a story for publication. Jack Driscoll, former editor of the Boston Globe and editor-in-residence at the Media Lab, used the Audio Notebook to write a story for the Media Lab newsletter Frames. Jack interviewed the author of this thesis to write a story about the Audio Notebook for a special issue of Frames dedicated to audio research at the lab. The Audio Notebook was also used by *SilverStringer* reporter Don Norris. The *SilverStringers* are a group of retired citizens from Melrose, Massachusetts who are writing their own on-line newspaper called the Melrose Mirror. The Melrose Mirror project was created by Walter Bender, Director of the Media Lab News in the Future Consortium, and is published on the world wide web.

The subjects were given notepads and pens for the Audio Notebook, but otherwise took notes normally. All but one of the review sessions took place in a sound studio at the Media Laboratory; SilverStringer Don Norris reviewed the interview at his home in Melrose. The sound studio was equipped with a camera for filming the subjects as they used the Audio Notebook to review their notes and audio recordings. The experimenter (LJS) stayed in the room with the user, observing his/her use of the device and making notes throughout each session. Each session was also logged on the computer. Figure 4-1 shows the user activity information captured in the log files.

Activity	**Data captured**
Stop button pressed	time of action
Change of speed	time of action, speed selected
Audio scrollbar selection	time of the action, time point in audio
Selection from page	time of action, xy location, time point in audio
Page turn	time of action, page number
Book removal	time of action

Figure 4-1: User activity information captured in a log file during review sessions.

The following four sections provide a detailed account of the observations for each user. The sections begin with a summary of the user's purpose and style of reviewing the audio recordings. Each of the users exhibited different styles of use. The purpose of reviewing the audio differed from one user to the next. Therefore, this study provided the opportunity to observe a variety of usage patterns even though there were only a small number of users.

4.3 Student 1—Rapid Skimming

4.3.1 Usage Summary

Student 1 skimmed quickly through her notes during each session, mainly reviewing material that was unclear during the lecture or not clearly recalled afterwards. This student's total listening time was approximately 1/3 of the original recording time on average. Student 1 skipped around in the audio using spatial navigation and the audio scrollbar, listening for potential sections of the lecture to review. She often used the audio scrollbar as a skimming control by selecting on every few LEDs in the scrollbar. Instead of skipping between two places in the recording and potentially missing information, student 1 wanted to be able to play the audio at a faster rate. After a playback speed control was added to the Audio Notebook interface, student 1 listened to the lectures at the fastest possible speed (2x the original) at all times. Student 1 did not significantly alter her notetaking habits as a result of using the Audio Notebook. She took pride in her notes and wanted them to be very accurate. In addition, student 1's class required a significant amount of notetaking off the blackboard, information not captured by the Audio Notebook.

4.3.2 Pre-Use Interview

Prior to using the Audio Notebook, student 1 was interviewed about her notetaking habits. She had tried several different methods of notetaking as a student. In her most recent method of notetaking, she created her own spiral bound notebooks using graph paper, always dating and numbering each page. Previously she had tried partitioning her notes, leaving a place on each page for adding notes later on. In another notetaking experiment, she put graphs on one side and written notes on the other side of pages. Finally, she bought a computerized tablet—an 8x10 LCD tablet and computer which performed handwriting recognition. She said the device was slow, and did not recognize her handwriting well. After one session, she turned the handwriting recognition off and just used it in a digital ink mode. After each class, she printed the pages. When asked if this was useful, she said "no, it was a pain, and it added an extra step." She tried this device for one month and then switched back to paper notes, stating—

> "There is something to be said about *leafing* through [notes] and you can't replace that… *scrolling* through is not the same as *leafing* through something and actually looking because you might not leaf sequentially; you can leaf randomly. There's something about having the feel of paper too." [Student 1, interview]

4.3.3 Taking Notes

Student 1 felt comfortable using the device in class and liked the fact that it she did not have to sit in the front of the room to record the lecture. She sat in the back row of the classroom during every lecture. Student 1 also did not worry about breaking the device; she treated it very casually, and was not concerned about having a glass of orange juice next to it.

During the first class, she was given one blue pen to write with. She said that not having colored pens made it difficult to draw the diagrams written on the whiteboard and that they looked cluttered in her notes. The professor drew diagrams on the board using several different colored markers. Color is important for a class in optics because different colors are used for rays, waves, and arrows. In addition to using color for diagrams, she said she used color to distinguish headers, topics, and asides (e.g., her own thoughts about something the professor is saying) because it makes the notes look "less cluttered."

4.3.4 Review Sessions

Student 1 used the Audio Notebook to review the classes nine times during the course of the semester (Figure 4-2). The student used two audio notepads containing a total of 24 hours of audio.

Number of classes recorded	16
Number of classes reviewed	13
Number of review sessions	9
Average number of pages of notes per class	4.5
Average amount of audio per page (min)	15.2

Figure 4-2: Usage statistics for student 1.

One or two classes were reviewed in each session, typically within one week of the class. During the nine review sessions, the student reviewed pages of notes and audio associated with 13 of the 16 classes recorded. It is interesting to note that the three classes not reviewed occurred after the second quiz. There was no final in the class, only a final project. This goes back to the question of when users will be motivated to review the audio recordings. For students, there is less motivation to review material when there is no associated problem set or examination. However, classes often build upon one another, with material from one course becoming a pre-requisite for the next. The audio recordings can therefore provide a valuable record of the class, even after it is completed. Student 1 requested to review classes using the Audio Notebook six months after the class had ended.

Student 1 learned to use the Audio Notebook interface very rapidly without help from the experimenter. She had no difficulty playing back the audio by selecting on different locations in her notes. She also began using the audio scrollbar immediately in the first review session. She used it many times to back up and repeat portions of the audio, and also to re-start the audio after stopping playback.

During the third review session, student 1 listened to more of the audio than during the first two. She may have reviewed more of the audio because of the increased time lag between the classes and review session. The first review session occurred just one day after the lecture, and the second three days later. In contrast, in the third review session, the student reviewed two lectures one full week after they were presented. Student 1 also had little sleep prior to one of the classes, so more review was needed. I asked the student if she felt that reviewing the audio was too time-consuming. She said that the time was valuable for reviewing parts of the lecture she did not remember, and that unlike a tape recorder, it saved time because she did not have to re-listen to the entire lecture.

Starting with the fourth review session, a computer log was kept of the student's activity when using the Audio Notebook (Figures 4-3 through 4-5). Figure 4-3 gives the amount of audio reviewed in each session. Note that the amount of audio reviewed includes repeated portions, so it is possible for the time spent listening to exceed the total amount of audio. The fluctuation in the percent audio reviewed is related to the importance of the content and the reason for review. For example, the percent audio reviewed (out of the total for the pages reviewed) in session 6 was only 16.4%. The material reviewed in this session was a guest lecture given by one of the teaching assistants, not the professor. Sessions 7 and 9, for which a large percentage of audio is reviewed (61.4% and 84.0% respectively), occurred just prior to a quiz.

Review Session	Time Spent Listening (min)	Playback Speed	Amount of Audio Reviewed (min)	# Pages Reviewed	Amount of Audio on Pages (min)	Percent Audio Reviewed
4	30.5	1.0	30.5	9	162.0	18.8%
5	50.0	1.0	50.0	5	81.9	61.1
6	13.0	1.0	13.0	3	79.0	16.4
7*	21.3	2.0	42.6	4	69.4	61.4
8	13.3	2.0	26.5	3	70.7	37.5
9*	21.4	2.0	42.9	2	51.0	84.0

Figure 4-3: Usage statistics for the Audio Notebook review sessions. The percent audio reviewed is the amount of audio reviewed divided by the total amount of audio associated with the pages of notes. A * indicates an exam review.

Figure 4-4 shows the percent of audio reviewed in comparison to the listening time required. In session 7, speed control was first introduced. Once speed control was incorporated into the Audio Notebook, the student always listened to the audio recordings at two times the original speed. As shown in the graph, the student's use of speed control resulted in a 50% time savings in review sessions 7–9.

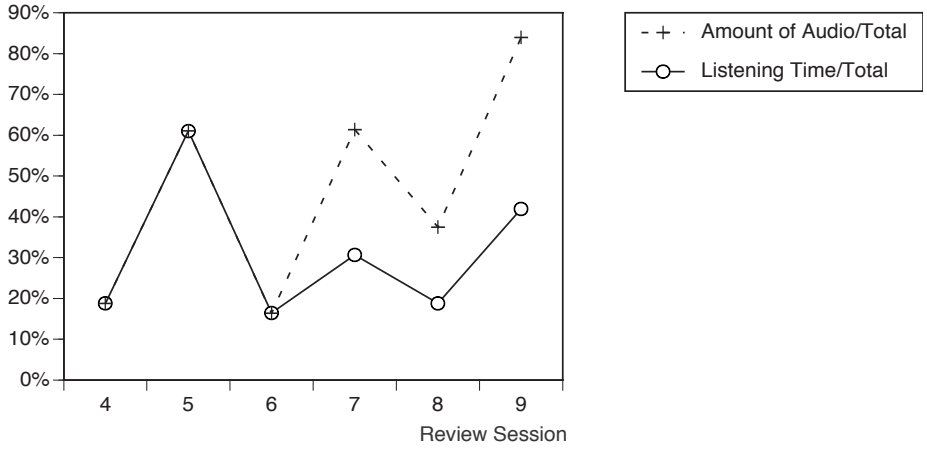


Figure 4-4: The amount of audio reviewed out of the total time of the lecture (percent audio reviewed) in comparison to the listening time required. Using speed control resulted in a 50% time savings in sessions 7–9.

Figure 4-5 gives the number of times each audio navigation control was used in each review session. When I first implemented the audio scrollbar, many people speculated that it was not necessary given spatial navigation. However, the audio scrollbar was used with equal or greater frequency than selection on the page.

Review Session	# Page Selections	# Scrollbar Selections	# Stop Button Presses	# Speed Changes
4	31	28	0	N/A
5	25	53	2	N/A
6	18	34	0	N/A
7*	1	8	0	0
8	4	8	2	0
9*	0	3	1	2

Figure 4-5: Audio navigation techniques used by student 1 for reviewing the audio recordings in each session. A * indicates an exam review.

It is interesting that student 1 listened to the audio very continuously, rarely pausing the recording. The interface supports this style of interaction since the user can jump between different portions of audio using the scrollbar and selection on the page, without pressing the stop button. As can be seen in Figure 4-5, student 1 rarely used the stop button. Note that the audio could also be stopped by removing the notebook from the device. However, in practice, the user generally removed the notebook when turning or flipping through pages, not to pause the audio, except at the completion of a review session.

Student 1 preferred to listen to the audio recordings at 2x the original speed. She did not interactively change the speed while listening to the recordings. The Audio Notebook allows the user to select a default speed preference, so the user is not required to change the speed each time he/she listens to a recording.

4.3.5 Correspondence between Notes and Audio

During one of the first review sessions, student 1 commented that her notes were not "in sync" with the professor as much as she would have liked. Prior to this review session, the experimenter (LJS) had already increased the listening-to-writing offset from 2 to 5 seconds for this user, based on the alignment of some notes and audio from the first lecture. Student 1 said she wanted to take notes off the board more quickly so her notes and the audio would correspond better. In a subsequent review session, the listening-to-writing offset was adjusted as student 1 interacted with the device. At this time, the offset was further increased from 5 to 6.5 seconds. After this adjustment, student 1 found the playback starting times to correspond more closely with the professor.

Student 1 also used the audio scrollbar to fine-tune the starting point of an audio selection; she selected in her notes and then adjusted the starting point forward or backward using the scrollbar. This student and other subjects in the study discovered this use of the scrollbar through exploration of the interface, without explicit instruction. After the listening-to-writing offset was increased, the student rarely adjusted the selection starting times. Given an accurate correlation between the notes and audio, student 1 said it was no longer necessary to paraphrase the professor; her notes were mainly comprised of information from the whiteboard and her own thoughts about the material.

4.3.6 Use of the Audio Recordings

Student 1 used the Audio Notebook to review portions of the lectures that were unclear or missing from her notes. During each review session, she skimmed through the notes and audio by selecting on different areas in her notes and using the audio scrollbar. Once she found a place of interest in the lecture, she often listened to a large amount of the audio without jumping around. She reviewed information that she did not remember clearly, and skipped over everything else.

Many times the student remembered drawing a diagram but did not recall the meaning of it or did not understand it originally. Figure 4-6 shows a page from her notes containing a diagram. Notice the three vertical lines drawn on the bottom right corner of the page. During the review, she selected on these lines to hear the corresponding audio from the lecture. She said that the Audio Notebook was "really helpful because I totally forgot what he was saying in class and if I looked at my notes alone, I wouldn't have known what that meant." In other instances, a particular

phrase written in the notes was unfamiliar afterward. By selecting on the words or graphic and hearing the corresponding audio, she was able to recall their meaning.

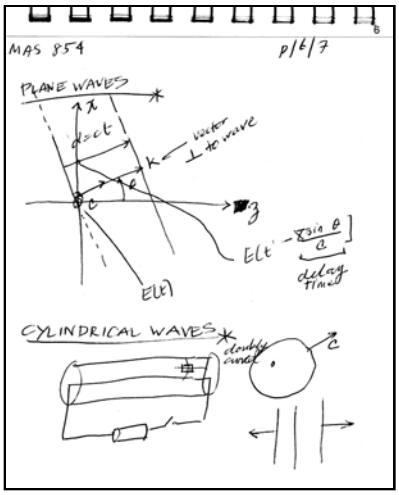


Figure 4-6: Student 1 could not recall the meaning of the three vertical lines at the bottom right corner of this page. She used the Audio Notebook to determine their meaning.

In another example, the student marked her notes with a "?" explicitly indicating missing information for later review (Figure 4-7). By touching on the question marks with the pen, the user is able to review this information. The student commented that things that were not clear in class initially, become clearer upon review using the Audio Notebook.

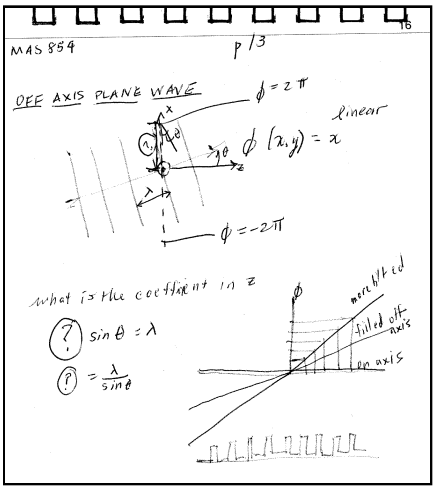


Figure 4-7: A question mark in the notes (bottom left) indicated to review this information later.

In most cases, the student did not review the notes prior to reviewing the audio recording. However, prior to session 7, a review for an upcoming exam, the student read through her notes and put yellow post-it notes on pages needing review. During this session, she listened only to information on the pages with post-its on them and some surrounding information on adjacent pages.

The examples given so far demonstrate how the Audio Notebook was used to review information that was unclear in the notes, or explicitly marked for review during or after the lecture. However, there were other cases in which the discovery of missing information was completely

serendipitous. At one point during a review the student said “I just realized how bad these notes are.” There were a lot of things she was hearing on review that she did not have in her notes. In interviewing subjects about their notetaking, some state that they rarely miss anything. This can be misleading because, like student 1, they may not realize that important information is missing from their notes. In some cases, users may review things they know they missed, but in others, the discovery is serendipitous.

4.3.7 Skimming Using the Audio Scrollbar and Speed Control

Student 1 often used the audio scrollbar to skim through the audio associated with a page of notes. During one review session, the user performed the following actions in sequence:

1. Select with the pen on a location in the notes to begin playback.
2. Use the audio scrollbar to skip ahead in the audio recording.
3. Use the audio scrollbar to skip backwards in the audio recording (in between the original starting point from step 1 and the skip ahead point from step 2).

In step 1, the user finds a location in the notes of interest and starts playback by selecting there. Then, in an attempt to get through the information more quickly, the user jumps ahead using the audio scrollbar. Note that when a user selects on a location in the notes, the audio cursor shows the related position in the scrollbar. The user can then move forward or backward relative to that starting point in the audio. In step 3, the user skips backward again, perhaps feeling she had missed something in the audio. Figure 4-8 gives a pictorial representation of this user activity.

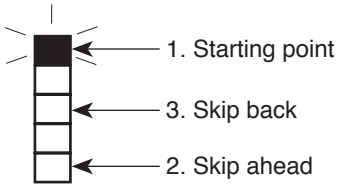


Figure 4-8: Visual representation of audio navigation using the audio scrollbar during one review session.

In other cases, student 1 used an *isochronous skimming* technique to skim through the audio—selecting somewhere in the notes, and then jumping ahead by equal amounts each time using the audio scrollbar. However, she wanted to be able to listen quickly to a section of audio without skipping over portions.

Starting with the seventh review session, a speed control was added, allowing the user to increase the speed of playback up to 2x the original without changing the pitch. The first time student 1 used the speed control, she immediately started listening to the speech at 2x speed without slowing down at any time. She was able to comprehend the speech at this fast rate even though she had no previous experience listening to time-compressed speech. This is most likely because the professor’s speech was familiar to her and she heard the material before. She said “I really like it because you don’t lose anything and it’s easier to listen to this way… when the speech is slow, your mind tends to wander… this way you concentrate on it more.”

In subsequent review sessions, student 1 used the speed control and the audio scrollbar in combination for even faster skimming. She listened to the audio at 2x speed, and used the scrollbar to jump ahead in the recording.

4.3.8 Audio Notebook versus Tape Recorder

Student 1 said that backing up and repeating a portion of a recording, and skipping around was easier using the Audio Notebook than with a tape recorder. She also liked the fact that the Audio Notebook allowed her to look at her notes and listen to the lecture at the same time. This appears to be an important advantage of controlling audio playback directly from the notebook itself—users do not have to divide their attention between their notes and a completely separate playback device. Since users can maintain their visual focus of attention on their notes, they do not lose their place.

Student 1 noted the higher audio quality and lack of background noise in comparison to tape recordings she had made. A student tape recording a class from his/her seat will often pick up a lot of noise from the room and from their own activity (e.g., writing, eating, crumpling of paper, side conversations, etc.). The Audio Notebook prototype uses a high quality wireless microphone to record the audio. Since the professor wears the microphone, the strength of the audio signal remains constant even when his location and orientation vary (e.g., when the professor turns to face the whiteboard).

4.4 Student 2—Detailed Review

4.4.1 Usage Summary

Student 2 used the Audio Notebook very differently from student 1. Student 1 skimmed through the audio and notes as quickly as possible, reviewing only unclear or missing information. Student 2 used each review as a study session, and the Audio Notebook as a "study tool." For student 2, the Audio Notebook provides the opportunity to re-listen to the lecture and replay particular portions to achieve a better understanding of the material. While student 1 did not significantly change her notetaking habits when using the Audio Notebook, student 2 began taking fewer detailed notes, and outlining key points from the lectures instead. Upon review, student 2 added detailed information not written down while originally attending the lecture. In some ways, student 2 relied more heavily on the audio than student 1, because she did not take all the information down in class. On the other hand, student 1 did not add to her handwritten notes, relying instead on the ability to randomly access any portion of the lecture when necessary.

It is interesting to note that the same audio navigation tools—spatial navigation, audio scrollbar, and speed control—were used by each student in different ways. Student 1 used these tools to jump around in the audio, skip over parts not of interest, and locate specific portions for review. These controls allowed student 1 to *skim* quickly through the audio recording. Student 2 used these same audio navigation controls to replay portions of audio, and transcribe quotes of interest. The controls provided student 2 with a mechanism for detailed review rather than high-speed skimming.

4.4.2 Taking Notes

Student 2 altered her notetaking habits when using the Audio Notebook. The student began writing notes that outlined "key points or concepts" presented in class. For example, in one class, she marked down in her notes "key things to observe", as a reminder to review this information. She said these notes were more structured than the notes she was taking previously. During class she would create an "outline" of important information, with bullet items for various concepts. She relied on the Audio Notebook to be able to go back and review this information after the

lecture and fill in the detail where needed. In addition, the student also tried to organize the information by page, attempting to turn pages at topic breaks. She would even "scrunch" information onto a fairly full page, waiting for a good time to turn the page.

This student's reliance on the Audio Notebook was also affected by technical problems. After one class when there was a technical problem with the recording, the student said she might change her notetaking habits in the next class because her confidence was lowered. If the audio was not going to be available, than she would need to take more detailed notes during the lecture. However, during the next lecture, she forgot about the previous problem and still relied on the ability to review the audio afterwards. Once she began using the Audio Notebook, she would feel secure that the material was being captured, and there was no need to write down what the professor said verbatim.

4.4.3 Review Sessions

Student 2 reviewed 7 of the 10 lectures successfully recorded using the Audio Notebook (Figure 4-9). Although there were also 7 review sessions, these do not correspond one-to-one with the lectures. Sometimes the first half of a lecture was reviewed in one session, and the remainder in another. Also, in reviewing for an exam, the student went over material from multiple classes, some already reviewed previously. Two classes were also recorded onto DAT tape when the Audio Notebook was not available. It is interesting that these recordings were never reviewed.

Number of classes recorded	10
Number of classes reviewed	7
Number of review sessions	7
Average number of pages of notes per class	10.2
Average amount of audio per page (min)	8.2

Figure 4-9: Usage statistics for student 2.

Review session 1 lasted 54 minutes (Figure 4-10). This is slightly longer than the total amount of audio associated with the pages reviewed, since some portions of the audio were repeated.

Review Session	Time Spent Listening (min)	Playback Speed	Amount of Audio Reviewed (min)	# Pages Reviewed	Amount of Audio on Pages (min)	Percent Audio Reviewed
1	54.2	1.0	54.2	7	43.5	125%
2	91.8	1.0	91.8	11	65.4	140
3	62.2	1.0	62.2	7	45.4	137
4*	57.1	1.0	57.1	26	170.5	33
5	51.2	1.0	51.2	5	71.9	71
6	14.4	1.0	14.4	8	72.3	20
7	177.3	1.0–2.0	178.5	16	87.7	204

Figure 4-10: Usage statistics for the Audio Notebook review sessions. Note that the percent audio reviewed can be over 100% since it includes repeated portions of the audio. A * indicates an exam review.

During the first review session, student 2 mainly listened to the lecture from start to finish, without skipping portions. The student only made 4 selections in the notes to trigger playback at particular locations (Figure 4-11). However, she used the scrollbar frequently to back up and replay portions of the recording. As the student listened to the audio, she added to her notes. Sometimes she would need to back up and replay a portion to capture more detail for her notes. During this session, she discovered that she did not leave enough room on each page for

additional notes. She said that in the future she would either leave more room on each page or write down the extra notes on separate sheets of paper.

Review Session	# Page Selections	# Scrollbar Selections	# Stop Button Presses	# Speed Changes
1	4	21	3	N/A
2	46	78	12	N/A
3	31	45	7	N/A
4*	46	15	0	N/A
5	42	41	6	N/A
6	20	15	1	N/A
7	42	256	17	38

Figure 4-11: Audio navigation techniques used by student 2 for reviewing the audio recordings in each session. A * indicates an exam review.

As can be seen in Figure 4-10, student 2 usually spent more time reviewing the recording than in attending the original lecture. This is in sharp contrast to the type of high-speed skimming used by student 1. On average, student 1 spent 31% of the original listening time reviewing the audio, whereas student 2 spent 104% on average. The amount of review time exceeded the duration of the original audio, because student 2 often replayed portions of each recording.

In the subsequent review sessions, student 2 started to use selections in her notes more for navigating through the audio. Review sessions 1 and 3 are similar in length (54 and 62 minutes respectively) and amount of audio reviewed (43 and 45 minutes). However, the student used only 4 page selections in session 1, and 31 selections in session 3. In most of the review sessions, student 2 used the scrollbar the same amount or more frequently than the page selections. However, during session 4 when she was reviewing for an exam, student 2 used more page selections than scrollbar selections (46 page selections versus 15 scrollbar selections). In addition, she reviewed only 33% of the audio during this exam review. She was skipping around in the audio using the page selections more often than backing up and re-listening to portions using the scrollbar.

In session 7, speed control was introduced. Although the student changed the speed quite frequently (38 times), very little time savings was achieved. The student would increase the speed for a very short time period (e.g., 10–15 seconds) and then decrease it again.

Figure 4-12 shows the percent audio reviewed over time. Up until the last review session, the time spent reviewing the audio was decreasing. As it got later in the semester, the student had less time to review and had to reduce the listening time. In addition, three of the last four lectures were not reviewed. These were the only three lectures recorded that were not reviewed (see also Section 4.4.5).

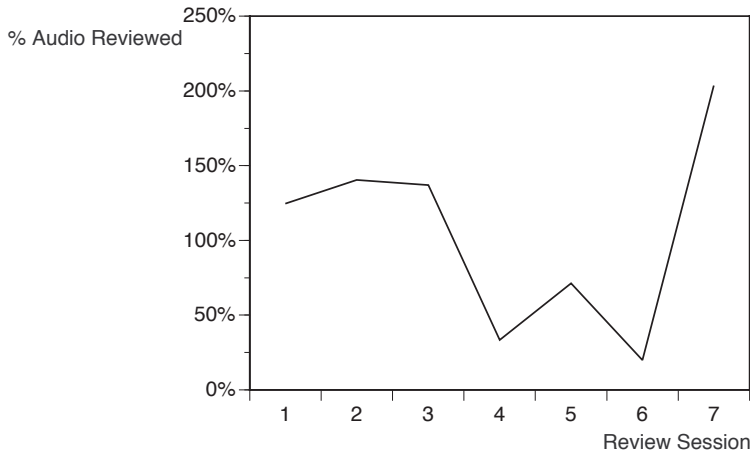
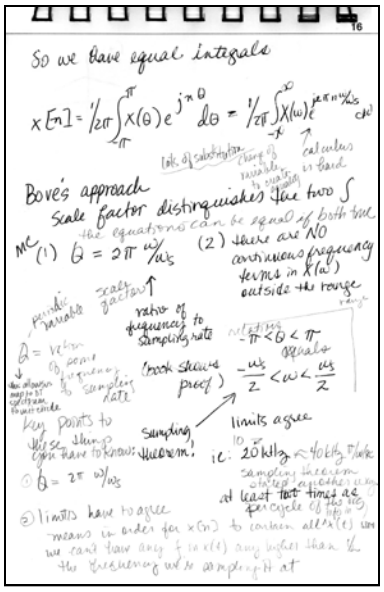


Figure 4-12: Percent audio reviewed over the course of the semester. Note that the percent audio reviewed can be over 100% since it includes repeated portions of the audio.

4.4.4 Use of the Audio Recordings

Student 2 used the Audio Notebook as a "study tool." For this student, reviewing the notes was not just a matter for reviewing missing or unclear information. It was a chance to go over the information a second time to obtain a better understanding. During each review session, the student added information to her notes, sometimes transcribing particular quotes from the professor. Prior to some sessions, to save time, the student would review the notes in advance and mark parts to review with post-it notes. This is similar to student 1's preparation for an exam review session.

Additional notes were added in pencil so "high level" information could be distinguished from this added detail. Figure 4-13a shows an example from one of the notebooks. The writing in gray are the notes added during review. The student also sometimes specially marked something down during class as a reminder to review this information later. For example, on one page (shown in Figure 4-13b), the student wrote "Key things to observe" intending to fill these in upon review. A large space was left in the notes after this phrase for the additional information. During review, more detail was added in this area of the notes. When adding notes, the student sometimes used two hands—simultaneously controlling the audio with one hand, and adding notes with the other. Using her left hand, she adjusted the audio position with the scrollbar, and using her right hand, she took notes.

a. 

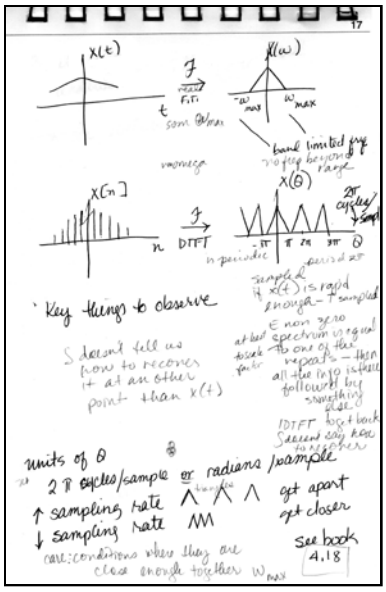
b. 

Figure 4-13: The notebook page on the left shows how notes were added upon review. The additional pencil notes are shown in gray. In the page on the right, the student wrote "key things to observe" and filled in the notes underneath and to the right of this during a review session.

The student viewed the Audio Notebook as a "reminding tool." It allowed her to organize the flow of information in class and then go back and review more specifically what was said at the points "highlighted" by her class notes. These notes then, served as "reminders" of key information to review and learn.

4.4.5 Accessibility of the Device

Monty gives a list of requirements that must be satisfied in order for a note or set of notes "to be useful" [Monty 1990, 32]. The requirements fall into six categories: reminding needs, search needs, browsability, *accessibility*, integration with environment, and protection. One limitation of this study was the accessibility of the device. Students did not have their own Audio Notebook which they could take home with them. Students also had to schedule time with the experimenter to review their notes. For the most part this was not a problem, however, towards the end of the semester, the students did not want to plan a review session in advance with the experimenter. They wanted to be able to review the material as needed, without advance planning.

The question arises, how would the student's use have changed given greater accessibility to the Audio Notebook? Although student 2 did not have constant access to the Audio Notebook, she did have continuous access to her notepads. Therefore, she would transcribe the detailed information during audio review sessions, so that she would have access to all the information, high level and detailed, at all times. Given unlimited access to the Audio Notebook, student 2 may have relied on the audio more, and over time, spent less time adding "detailed" notes since this information could be easily accessed using the Audio Notebook.

4.4.6 Audio Notebook versus Tape Recorder

Student 2 began using the Audio Notebook after the first few weeks of class. Prior to using the Audio Notebook, she tape recorded the classes. When asked to compare using a tape recorder with the Audio Notebook, she said that the Audio Notebook is "200% better than a tape recorder." She found that the Audio Notebook gave her better ability to pause and replay various

portions of the audio recording—she said "I can go where I want to go." She did not use the tapes as a study tool the way she used the Audio Notebook. The tapes were more of a back up in case she missed something important in class. The noisy quality of the audio on the cassette tapes also prohibited using them for detailed review of the material. However, this could be rectified by using a microphone setup like the one used by the Audio Notebook.

4.4.7 Sharing Notes with Other Students

Several students in the Signals and Systems class asked if they could review lectures they were unable to attend. In these cases, the students did not take notes using the Audio Notebook, and used student 2's notes to review the lecture. The students who missed a lecture had already copied notes from another student who did not use the Audio Notebook. So they had two sets of notes—one taken by student 2 using the Audio Notebook and one taken by another student in the class. They kept the second set of notes on the side of the Audio Notebook and added to them as they reviewed the lecture. I asked one of the students if it was awkward to review the class using two different sets of notes. She commented that having both sets was not confusing but helpful; if something was missing from one set of notes, it could often be found in the other. However, she also said that it was frustrating when the student writing in the audio notepad did not take notes during some time intervals of class. In these areas where notes were missing, she would lose her place in the lecture. Recall that student 2 often wrote down key notes only and filled in the details later, and her notes were not written with the intent of sharing them.

Since the students had not attended the lecture, they listened to a large percentage of the audio recording—82.1, 98.1, and 60.2% respectively (Figure 4-14). The first lecture reviewed by student 4 corresponds to the lecture reviewed by student 2 in her seventh review session. While student 2, who attended the original lecture, spent 177.3 minutes listening to the audio, student 4 spent just 58.6 minutes. Student 4 used the time-compression control to save 27.4 minutes or 31.2% time savings. In addition, student 4 listened to most of the audio one time, while student 2 often repeated portions and transcribed them.

Review Session	Time Spent Listening (min)	Playback Speed	Amount of Audio Reviewed (min)	# Pages Reviewed	Amount of Audio on Pages (min)	Percent Audio Reviewed
Student 3	70.6	1.0–2.0	not logged	10	86.0	82.1%
Student 4 Review 1	58.6	1.0–2.0	86.0	16	87.7	98.1
Student 4 Review 2	26.2	N/A	26.2	7	43.5	60.2

Figure 4-14: Usage statistics for student 3 and 4 who missed class.

The audio scrollbar and time-compression appear to be important for users who have not attended the original talk. Both students who had not attended the lecture made frequent use of the audio scrollbar and speed controls (Figure 4-15). This use of the speed control was much more interactive than student 1 who listened to the entire lecture at 2x playback speed. Note that student 3's use of the time-compression was observed but not captured in the log file. Also, in student 4's second review session, the speed control was not available; the notes for the lecture were contained in an older audio notepad that did not have a speed control.

Student 3 often listened to the audio at increased speeds (1.5–2.0x) and then slowed it down when he heard something he wanted to focus on more closely. He used the scrollbar and speed control in combination to jump back and replay a portion at a slower rate. He would also slow the rate to

1.0 when adding to his notes. He listened to the audio very continuously even when writing notes, only using the stop button three times. He also saved time by skipping pages of notes corresponding to introductory material. However, in other cases when he encountered familiar material, he increased the speed to 2.0 rather than skip over it.

Review Session	# Page Selections	# Scrollbar Selections	# Stop Button Presses	# Speed Changes
Student 3	27	29	3	not logged
Student 4–Review 1	12	53	17	53
Student 4–Review 2	15	13	1	N/A

Figure 4-15: Audio navigation techniques used by students 3 and 4. Note that student 3's use of the speed control was observed but not captured in the log file. Also, in student 4's second review session, the speed control was not available.

Like student 3, student 4 made frequent use of the audio scrollbar and time-compression. During her review session she made 53 scrollbar selections and 53 speed changes but only 12 page selections. She said she used two strategies for increasing audio playback. Sometimes she would keep the audio playing fast and then slow it down while taking notes or when a new topic was introduced. In other cases, she would keep the speed at a medium rate for the whole time. Like student 1, she commented that listening faster was better because when listening slowly "your mind wanders." When playing at a fast speed "you spend less time waiting for the professor to spit out the next syllable."

Student 4 used page selections to navigate between two related ideas in the lecture. In one instance, she selected on two locations in the notes, jumping back and forth between them. Thus if two ideas are separated in time, the spatial access allows the listener to hear them together.

Student 4 also wrote on a second set of notes kept on the side of the device. Unlike student 3, student 4 stopped the audio when adding to her notes (17 stop button pressed for student 4 versus three for student 3). When pausing the audio, she would restart from where it left off using the scrollbar and audio cursor.

4.5 Reporter Jack Driscoll—Building a Story Around Quotes

4.5.1 Usage Summary

> Date: Fri, 13 Sep 96 12:25:16
> From: driscoll@media.mit.edu (Jack Driscoll)
> To: lisa@media.mit.edu
>
> Subject: PROOF IN PUDDING
>
> What would you think if the first story written about your interview tool were an interview with you by me in Frames? … jack d.

Jack Driscoll in his role as Editor-in-Residence at the Media Lab, emailed me this message titled "PROOF IN PUDDING" about an upcoming issue of the Media Lab newsletter Frames dedicated to audio research. The idea was for him to interview me using the Audio Notebook and write a story about the device for the December 1996 issue of Frames.

Jack's use of the Audio Notebook provides another perspective on how audio captured by the notebook will be used. Jack used the Audio Notebook primarily to locate and transcribe quotes for his story. This resembles student 2's use of the Audio Notebook in that both users transcribed portions of the audio. However, the two uses are quite different. In Jack's case the search is very directed. He is listening for only the most important "sound bites" to include in his story. He is not carefully studying the material. Also, the accuracy of transcription is vital for a reporter but

not for the student. Jack's notes were also much sparser than the students' notes. His notes are indices of topics and quotes, each line of notes containing only a few words. Jack said that the detail of his notes was also affected by the length of the story; in this case, he was writing a short story of only 200–300 words.

4.5.2 Taking Notes

During the interview, unlike the students, Jack used the notebook in his lap rather than on a table. He had no difficulty taking notes or starting and stopping the recording. Several times he stopped the recording, saying "off the record…" and then asking a question he did not want to record.

Jack's notes are very sparse in comparison to the student's notes. For a reporter performing an interview, it is important to maintain eye contact with the subject, so fewer notes may be taken. Figure 4-16 shows two pages from Jack's notebook. He wrote down key words and phrases rather than a verbatim transcript of what was said. He also used some shorthand annotations.

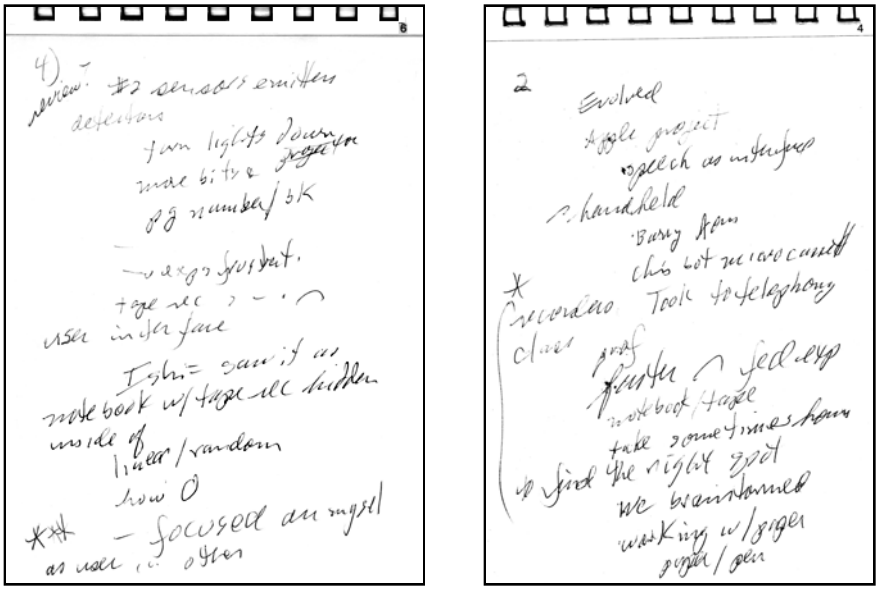


Figure 4-16: Two pages from Jack Driscoll's interview notes. In the middle of the page on the left, the note starting with the words "tape rec" is partially written in shorthand.

4.5.3 Review Session

Prior to the review session, Jack went through his notes and made a few annotations. He put star symbols next to important areas, and marked things to review. While reviewing the audio, he made a few additional annotations using a red pen. These annotations were very simple corrections or additions to his notes. In one case, he added the word "student" over where he had written "blind person" indicating that the note referred to a "blind student."

Jack used the Audio Notebook primarily to select quotes for his story (Figure 4-17). He described two methods for creating a story. In one method, he first outlines the story and determines where quotes are going to be placed before listening to an audio tape of the interview (Section 4.6). However, he said that a well written story is not systematic and that new ideas can come to the writer "like thunderbolts." A second method described by Jack is to select quotes for the story first and then organize the story around the quotes. This is the method Jack used for writing the Frames story.

Length of the interview (min)	54.3
Total time spent reviewing audio (min)	69.5
Percent Audio Reviewed	128%
Number of pages of notes	7
Average amount of audio per page (min)	7.8

Figure 4-17: Usage data for Media Lab Editor-in-Residence Jack Driscoll.

Jack listened to the interview starting from the top of his first page of notes. He listened for interesting quotes to transcribe. Once he heard a quote of interest, he typed it into a document on his laptop computer. He listened to the quote, stopped the audio, and then typed it into his laptop. Then he used the audio scrollbar to backup and replay the quote to check the accuracy of his transcription. Unlike other users in the study, he used the scrollbar and stop button together. Therefore, he said it would be better to have the two controls closer together.

It was interesting that the quotes he selected often were not indexed in his notes. This is evidence that navigation using the page alone is not enough. If a portion of the audio is not indexed on the page, it would be lost to the user without the scrollbar. In this case, using the audio scrollbar, Jack located and transcribed these notes. Something that does not seem significant at the time of the original recording, may become more important later on. Therefore it is vital to provide the user with access to all of the audio.

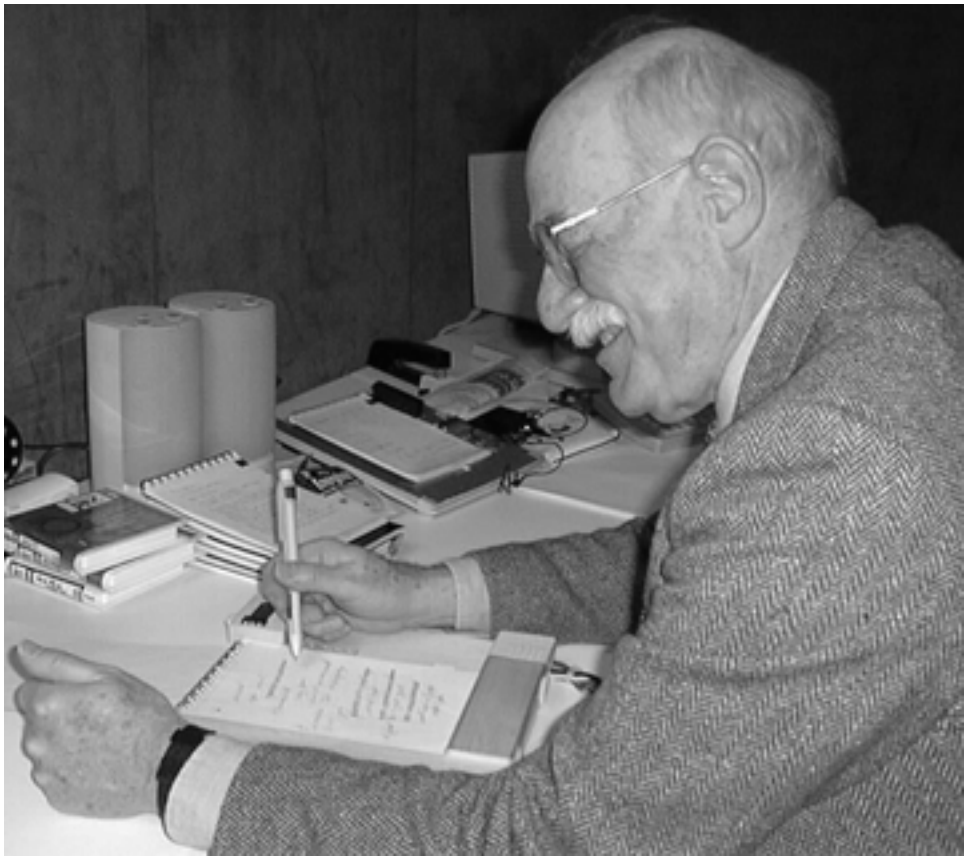
Jack used the scrollbar for navigating through the audio much more frequently then selecting on the page—124 scrollbar selections versus 38 page selections (Figure 4-18). At first, Jack preferred the scrollbar over page selection because he said that when he started the audio from the page, it often started in the middle of a quote. In contrast, he was able to start playback at the beginning of a quote using the audio scrollbar. Through exploration, he discovered that he could save time by first selecting in his notes to begin playback and then adjusting the starting point using the scrollbar.

Page Selections	38
Audio Scrollbar Selections	124
Stop Button Presses	38
Speed Changes	N/A

Figure 4-18: Audio navigation techniques used by Media Lab Editor-in-Residence Jack Driscoll. Speed control was not implemented at the time of this review session.

Jack spent a total of 69.5 minutes using the Audio Notebook to playback the interview. He did not have any experience using the device or training prior to the review session, so he spent about half of this time "familiarizing himself" with the interface. Jack said he also spent part of the time testing the device's capabilities to provide feedback about the design.

4.6 SilverStringer Don Norris—Filling in a Story Outline



4.6.1 Usage Summary

SilverStringer reporter Don Norris was writing a story for the Melrose Mirror about WWII army nurse, Kay Bistany. Jack Driscoll suggested that he use the Audio Notebook to record his interview. The interview was performed in the kitchen of Kay Bistany's brother in Melrose, Massachusetts.

Don was very skeptical about using the Audio Notebook. When I arrived at the private home, he asked several times why he needed the Audio Notebook, and seemed reluctant to use it for the interview. Don said that he used a tape recorder in the past and found it "burdensome." He said that when you take notes, you only write down the parts of interest, whereas a tape recorder captures everything and is difficult to find the parts of interest afterwards. "You can stop writing but the recorder will still keep going." He even used a special transcription playback device with pedals for controlling the audio, but found it difficult to back up and figure out his location in the recording, and the speed up and slow down "made people sound bad."

Don's use of the Audio Notebook was very different from Jack Driscoll's usage. Don took very detailed notes during his interview. He did not rely on the audio recording to capture the information for him. Don's notes also needed to be more detailed than Jack's because he was writing a longer story. While Jack structured his story around the quotes he selected, Don first outlined his story from his handwritten notes. Don reviewed the audio to clarify the information in most cases, rather than to select direct quotes. Don spent less time reviewing the audio recording, in part because his search was more directed than Jack's.

4.6.2 Taking Notes

I told Don that he could take notes as he normally would during an interview and that we could determine after the interview whether or not the Audio Notebook was of any use in writing the story. This is another instance that shows the advantage of using actual paper and pen. Don was not forced to conform to some new way of taking notes. In fact, he probably would have refused to use the Audio Notebook had it been implemented using an LCD display.

Don had no knowledge of the Audio Notebook prior to the interview, but had no difficulty using it. He seemed comfortable with starting and stopping the recording, turning pages, and writing in the notebook. During the interview, Don took six pages of notes in a fairly dense script handwriting (Figure 4-19).

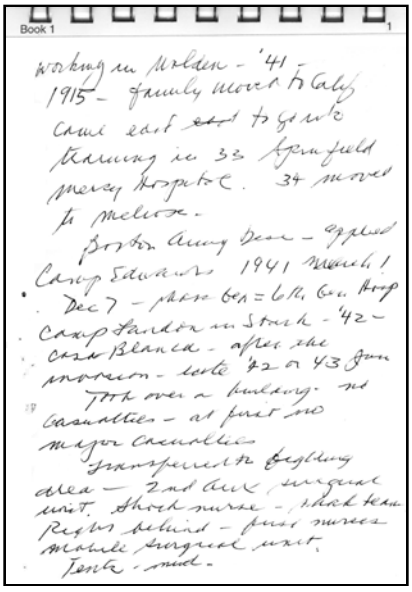


Figure 4-19: A page of notes from Don Norris's notepad.

4.6.3 Review Session

Several weeks after the interview, Don asked to review the material using the Audio Notebook (Figure 4-20). The review session was performed at Don's home in Melrose. Prior to the meeting, Don had typed his handwritten notes into his computer. His goal was to review the audio to clarify several parts of the story. For example, he wanted to verify the timeline of events that Kay presented, and look up some specific names and dates.

Length of the interview (min)	67.7
Total time spent reviewing audio (min)	39.2
Percent Audio Reviewed	57.9%
Estimated time to find and review quotes (min)	25
Number of pages of notes	6
Average amount of audio per page (min)	11.3

Figure 4-20: Usage data for SilverStringer, Don Norris.

The total review session lasted 39 minutes, less than two thirds the time of the original interview. Don spent approximately 25 minutes updating his story, and the remainder of the time playing with the Audio Notebook, exploring the interface. Don used selections on the pages to jump around in the interview. He located an area of his notes corresponding to the information of interest, and started playback from the beginning of a line of notes. He used the scrollbar to jump

forward or backward in the audio, but used selection on the page twice as often for navigating through the interview (Figure 4-21). He found the scrollbar to be more "hit and miss" than selection from the page itself. The correlation between Don's notes and the related audio was very good, making it easy to locate information in this way. The listening-to-writing offset was set to 4 seconds prior to the review session; the same amount that was used for Jack Driscoll's notes (see also Sections 4.7.2 and 6.3).

Page Selections	41
Audio Scrollbar selections	22
Stop Button Presses	7
Speed Changes	0

Figure 4-21: Audio navigation techniques used by SilverStringer, Don Norris.

Don wrote on his typed notes as he reviewed the recording. He used the digitizing pen with an ink cartridge (as opposed to a stylus) to operate the device and add to his notes. This shows an advantage of using an ink pen for interacting with the device. Users can easily alternate between operating the device and adding to their notes.

After using the Audio Notebook, Don commented that it was "better than he imagined it to be." In addition, he thought the quality of the audio recording was extremely good—"the clarity is amazing." Don expected the audio to sound more "scratchy… like an old tape recorder." The audio is only recorded at 8 bits linear, so it appears that a high fidelity microphone setup is more important than the sample size and coding.

4.6.4 Post Review Session Reflections

Don Norris's story about Kay Bistany's WWII nursing experiences can be found at the Melrose Mirror web address—http://silverstringer.media.mit.edu/ss/html/stories/Article90.html. After the story was written, Don sent me the following comments about his experience using the Audio Notebook:

> Lisa:
>
> A few thoughts on recovering lost thoughts in the Kay Bistany interview:
>
> Very seldom have I found it necessary or prudent to use a tape recorder during an interview, both because of time constraints of newspaper production and the potential intimidation of microphones. There are exceptions, such as when the topic of the interview is of a technical or very precise nature—but such was not the case with the Bistany story. Notes scribbled in my own shorthand would suffice.
>
> Later, when you and I went over the interview, I believe you noticed that, having roughed out the story, numerous passages were marked for clarification and rewriting. Kay was modestly reticent about her war experiences, and she was sometimes distracted by other conversations.
>
> I was therefore happy for your Audio Notebook backup. With a rough draft indicating where the story needed attention, we were able to quickly pinpoint some dozen segments of our interview. I did not have to wait for a tape recorder to wind (or rewind) in search of needed material; it was both simple and quick to press a pointer to an entry in my notes, and the Audio Notebook provided access and playback within seconds. The audio was clear and crisp, and the background noises were subdued.
>
> I must admit, had I used a tape recorder, it would have taken some three, possibly four hours to transcribe those ninety minutes of conversation—time definitely wasted from production. For comparison, I believe you and I spent perhaps twenty five minutes bringing the story up to date—much of which time was consumed in my playing with this new toy. Fifteen minutes would have sufficed. It was fascinating, and I hadn't realized its value and utility until after the rather tough interview.

> Kay's story has been published for a week now and has brought nothing but outstanding reviews. Your Audio Notebook not only saved considerable time, but provided opportunity for unusual accuracy. I thoroughly endorse your program and thank you for the opportunity to use what I believe will be part of every reporter's equipment in the future.
>
> Don.

Several months later, Don was interviewing another subject for a story for the Mirror. This time the subject was 99-year-old, former Melrose High School principal and WWI flyer, Harold Poole. Don tape recorded the interview because he thought it would be important to capture the way things were said. In addition, he wanted to pick out sound bites from the interview to put on the web with the story. This time, Don did not have access to the Audio Notebook and used an analog tape recorder instead. After the interview, he wanted to review several portions of the audio recording. After a frustrating time of trying to locate the desired information on the tape, he gave up. He decided the only way to manage the audio would be to transcribe the entire interview. It took him a total of 6 hours to transcribe the approximately 1.5 hour interview of Harold Poole. In contrast, using the Audio Notebook, it only took him 39 minutes to review the 68 minute Bistany interview, quickly locating and taking notes about the portions of interest.

4.7 Iterative Design with Users

During the field study, design changes were made based on observations and feedback from users in the study. This section describes some of these design iterations, what motivated them, and the impact they had on the user's experience. The iterations range from small design changes (e.g., new pens) to additions of new features (e.g., time-compression). Note that student 1 began using the Audio Notebook before other users in the field study, so several changes were motivated by her experiences. Some of the changes were therefore made before other students and reporters used the device.

4.7.1 Colored Pens

Users in the field study were given notepads and a single pen to write with. After the first time using the Audio Notebook, student 1 asked for multiple colored pens. She also initially asked for graph paper but then changed her mind after using the unlined notepad given to her. As discussed in Section 4.3.3, student 1 used multiple colors for drawing diagrams, to mark topic headings, and to separate the information to make it look less "cluttered." She strongly objected to taking notes with only one color. After this request, users were given four pens—one blue, one green, one red, and one with a stylus tip—to use during class and review sessions. The stylus was often used during review sessions, when users did not want to mark up their pages. However, for users who added to their notes while reviewing the audio, an ink pen was used to facilitate easy transitions between operating the controls and writing.

One issue is whether the color information could be useful for segmenting the audio recording. The tablet and pen technology used in the current design do not provide any mechanism for identifying the different pens. However, other mechanisms could be added to code the pens (or distinguish the different colors) if this information proved useful. In practice, users were not consistent in their use of different colors. For example, the color red could be used to draw a heading or for a ray in a graph. Perhaps these two kinds of objects could be distinguished, but headings were also not consistently drawn in a different color than the other notes. Changes in color (regardless of which color is used) can sometimes be meaningful (e.g., a new topic) but not

always. The user could be instructed to use a particular color for header information, but this goes against the design philosophy of the Audio Notebook. It is unclear whether the user's natural activity provides consistent enough use of color for segmentation purposes, but perhaps in combination with other cues (e.g., speaker's prosody, lines drawn across the page) this information could prove useful.

4.7.2 Improving the Correlation between Notes and Audio

The Audio Notebook interface must account for the delay between listening and writing. In the initial design, when the user selected somewhere in the notes to begin playback, the system simply backed up by a constant amount from the time that the pen stroke was made (i.e., the listening-to-writing offset). During initial review sessions, student 1 commented that her notes were not completely in synchronization with the professor (Section 4.3.3). When she selected somewhere in the notes, playback began too late. The user would then adjust the starting point using the audio scrollbar.

A user-specific listening-to-writing offset was then introduced. A profile was created for each user containing the preferred amount of offset. The user-specific offsets ranged from 4–6.5 seconds. For purposes of the field study, the experimenter (LJS) adjusted the offset for the user. However, one could imagine that a slider control could be provided to give the user direct control of the offset. For student 1, the offset was first set by the experimenter and then adjusted interactively with the user. For other users, the experimenter selected a default offset by aligning a single note with the associated audio content.

Some types of notetaking tasks may result in more delay between listening and writing than others. The students in the study had to copy notes from the blackboard, and sometimes had to wait for the professor to move away from the board before copying down the information. In contrast, the reporters could focus entirely on the talker, with no additional information to write down. Student 1 had the longest preferred listening-to-write offset (6.5 seconds). However, student 2's offset was shorter (4.5 seconds). Student 2 used the amount of offset set by the experimenter and did not request to adjust it. Student 2's detailed style of reviewing the audio did not rely on exact synchronization between the notes and audio. Both reporters also used the amount of offset set by the experimenter (4.0 seconds).

When Jack Driscoll used the Audio Notebook to review his interview, he commented that playback often began in the middle of a quote. The associated content was correct, but the starting point was not (see Section 4.5.3). There is no guarantee of where playback will begin when simply backing up by a constant amount, even if this amount is customized for each user. Using a fixed backup amount, playback often begins in the middle or towards the end of a phrase. In these cases, it can be difficult for the listener to follow the content. After the field study, a phrase detection algorithm was implemented (Chapter 5). This algorithm predicts the location of major phrase beginnings. This information was then incorporated into the Audio Notebook; when the user selects somewhere in his/her notes, the audio first backs up by the user-specific listening-to-writing offset, and then "snaps" to the nearest phrase beginning (see Chapter 6 for more detail).

4.7.3 Feedback When Turning Pages

When the user turns to a page in the notebook, the system first loads the pen data for that page of notes. Users must wait until the data is loaded before they can playback the audio. While the

system is loading the pen data, the message “Loading” is printed on one line of the 8 character wide display (the other line displays the page and book number). In practice, users rarely looked at the display. They focused their visual attention on their notes while listening to the audio. As soon as a user turned the page, he/she would begin selecting in the notes. Since they did not notice the message on the display, they became confused about why playback had not started. Although this loading delay is just an artifact of the current software design (i.e., the software can be optimized to eliminate this delay), it brings up an important issue about feedback. It suggests that auditory feedback is needed to present critical state information during playback, rather than visual feedback.

Auditory feedback was then introduced to indicate when the system was loading data. The waiting sound from the television show Jeopardy was used as an auditory icon [Gaver 1986]. The sound would play until loading completed. While one user said “it makes the waiting shorter” another found it annoying after a short time. After a couple of weeks, the sound was turned off. Users now remembered to wait after turning the page without the sound. Later, one of the users missed the sound and wanted it back. She said it made her “feel smart.” When auditory feedback is employed, users should be given control over the amount of feedback, the type of output (speech versus non-speech), and the ability to turn the sound on or off [Stifelman 1995]. In this case, however, speech feedback was avoided, since it could potentially be confused with the speech from the lecture or interview.

4.7.4 Speed Control

In the initial design of the Audio Notebook, no audio processing algorithms were used. The goal was determine the areas where processing was needed, and integrate it based on observations of users. As discussed in Section 4.3.4, student 1 requested the ability to speed up playback after several review sessions. This student used the audio scrollbar to skip around in the audio, but wanted the ability to listen faster without missing anything.

A time-compression control was added to the Audio Notebook interface, allowing users to increase the speed of playback from normal up to 2x the original. The speed of the speech is increased without changing the pitch, so there’s no “mickey mouse” affect. A speed control is printed on the bottom of each notebook page (Figure 4-22). The system detects when the user’s pen is in this control area. Users can interactively increase the speed of the speech during playback by sliding their pen along the speed control. In addition, users can select a default speed preference which is stored in their user profile.[5] There are several advantages to printing the control on the pages. First, this allowed rapid iteration of the design; adding a potentiometer control would have been much more time-consuming since it would have involved modifying the hardware. Second, it creates a paper-like look and feel to the interface; the more hardware controls on the device, the more intimidating it appears to users. One disadvantage is that it takes up space on the page. However, printed at the bottom of the page, the control was outside the primary notetaking space.

---

[5]Note that like the listening-to-writing offset, the speed preference was set for the user by the experimenter for the purposes of the field study.

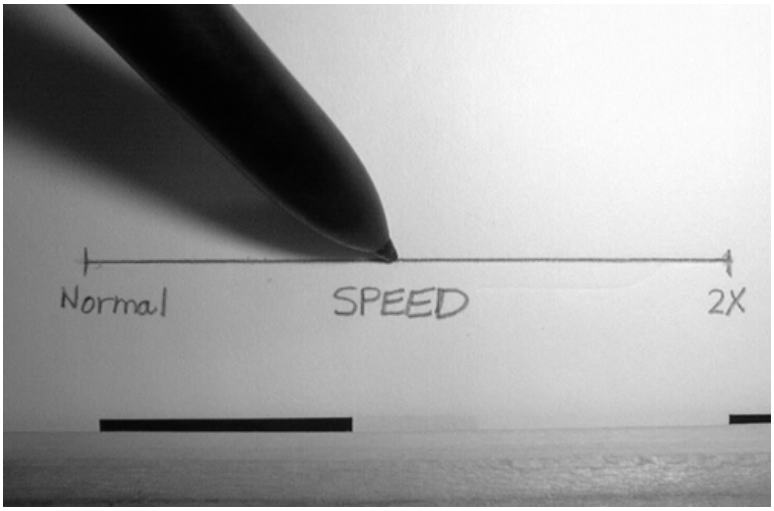


Figure 4-22: A speed control is printed on the bottom of the notebook pages. The user can interactively increase the playback speed from normal to 2x the original by sliding the pen along the control.

The time-compression used by the Audio Notebook is a low-cost time-domain algorithm developed by Arons [Arons 1994a]. The algorithm uses a technique called *isochronous sampling*—removing intervals of speech at regular intervals, and performing linear cross-fades at the boundaries to avoid distortion [Fairbanks et al. 1954]. This algorithm was used by both the VoiceNotes [Stifelman et al. 1993] and SpeechSkimmer systems [Arons 1997]. For a detailed review of time-compression techniques, see [Arons 1992b].

Some users skim through the audio at high speeds, while others use the Audio Notebook to perform a detailed review of the material. For users who review the audio in detail or perform transcription (e.g., reporters transcribing quotes), the ability to decrease the speed should also be provided in future Audio Notebook designs. In addition, given knowledge of the talker's speaking rate over different time intervals of the recording, a more intelligent speed control can be created. Rather than specifying a time-compression amount, the listener can select an overall speaking rate (e.g., 200 wpm) [Lehman 1997]. In this way, slow talkers might be sped up while very fast talkers might be slowed down to attain the user's preferred listening rate.

4.8 Summary of Results

Even though there were only a few users in the field study, their use of the Audio Notebook was very different from one another. This provided the opportunity to study different usage styles and tasks. The field study provides knowledge about how people use audio recordings when they are more accessible and manageable. The audio recordings were used for many different purposes, including:

- to review information that was not clearly understood in the initial presentation
- to clarify information written in the notes
- to fill in missing details, sometimes left out on purpose with the intention of filling them in afterwards
- to find and transcribe quotes.

The interaction techniques supported a range of usage styles, from detailed review to high speed skimming of the audio recordings. It was interesting that the same controls were used for different purposes by different users. For example, the audio scrollbar was used as a skimming control by one user, and for detailed review and repetition of material by another. However, spatial navigation was most often used when listeners were skimming through the information and skipping over portions of the recording; the audio scrollbar was used more often to repeat portions of audio and fine-tune the location in the recording. Spatial navigation was also especially useful for clarifying ambiguous notes because the user could select directly on the information. Navigation by page was useful when students reviewed several different lectures in preparation for an exam. The students would mark pages to review with post-it notes, and then go directly to those portions of the audio recordings by turning to those pages.

Speed control was used both by students who attended the lecture and those that did not. One student always listened to the audio at 2x speed, and did not interactively adjust the rate of playback. She could listen at high speeds without training because she attended the original lecture, and was familiar with the material. For listeners who had not attended the original presentation, speed control was used to achieve time savings without skipping over portions of the audio. These listeners used the speed control very interactively, playing the audio at fast speeds, and then immediately slow down when encountering unfamiliar material. The speed control was also used in combination with the audio scrollbar in two ways: (1) to back up and re-play portions of the audio at slower speeds, and (2) for high speed skimming—jumping around in the audio recording while listening at high speeds.

Of the two students studied over a semester, one changed her notetaking habits more than the other as a result of using the Audio Notebook. Student 1 took pride in her notes, and wanted them to be very accurate. She wrote in different colored pens to keep the information from looking too "cluttered." During class she mainly copied information from the whiteboard and recorded her own thoughts about the material. She said that by using the Audio Notebook, it was no longer necessary to write down exactly what the professor said. Student 2 altered her notetaking more than student 1. She said that her notes were more "structured" than they were before using the Audio Notebook. During class, she would take notes that outlined the material, and then go back and fill in the details during review with the audio. She said that her notes served as "reminders" of information to review.

The audio recordings remained useful over time, even months after the original recordings were made. One student reviewed lectures over six months after the class had ended. She also wanted to listen to lectures that she had not reviewed during the original study. This shows how the importance of the audio recordings can change over time.

Lastly, the field study pointed out areas where additional information was needed to improve the correlation between the user's notes and the audio recordings, and to provide structure where little or none was generated by the user's activity (Section 5.1).

5. Acoustically-Structured Speech

This thesis uses a combination of user activity and acoustically-determined structure to allow rapid navigation through speech recordings. Talkers use prosodic cues (e.g., changes in pitch, pausing, energy) to convey structural information to a listener. This thesis aims to exploit these cues to allow listeners to quickly access desired portions of a speech recording. Audio Notebook users can navigate through audio using a combination of structural and activity indices. Activity information, such as writing and page turns, indexes the speech without knowledge of the underlying structure. By integrating activity indices with knowledge about structure, a more intelligent listening interface can be created.

5.1 Augmenting User Structure

The previous chapter described how a user's activity structures an audio recording by providing indices for subsequent retrieval. However, this activity information can be augmented to further structure an audio recording for easier access. For example, in using the Audio Notebook, reporter Jack Driscoll noted that playback often started in the middle of a phrase when he used spatial navigation. Finding the appropriate starting point in the audio associated with a user's writing is a challenging problem. The issue is that people do not write notes exactly in synchronization with the talker. If audio playback is started from the exact time a note was written, or a constant amount prior to the activity, playback may begin in the middle of a talker's phrase. Knowledge of phrase beginnings and endings can therefore be exploited such that when users make selections in their notes, playback starts from the nearest phrase beginning.

Additional structural information would also be useful for providing navigational landmarks in the audio timeline. For example, when student 1 was trying to skim quickly through a page of notes, she randomly selected on every few LEDs in the audio scrollbar. Given knowledge of where new topics begin, the system could suggest places to navigate in the recording. Thus, rather than randomly jumping from one location to another, topic suggestions would provide a guide for the user's skimming activity. These kinds of navigational landmarks will also be especially important in places where little or no structure was generated by the user's notetaking activity. When listeners must attend very closely to a talker, they may take fewer notes. For example, when a reporter performs an interview, notetaking can be distracting to some interviewees. The interviewer wants to maintain eye contact with the subject, reducing the amount of notetaking activity. Notetaking activity may also decrease when the material is very complex or unfamiliar to the listener. A goal of the Audio Notebook is to free listeners to devote more attention to the talker, so they are not always required to take detailed notes. By providing more accurate and additional structural indices into the audio recording, the system augments the user's activity, helping the user to find the desired portions of audio.

5.2 Structure versus Summary

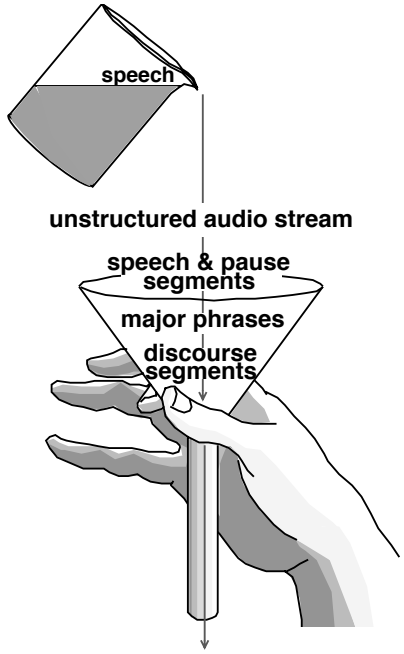
It is important to distinguish the approach of this work from previous work in emphasis detection and audio summarization. This thesis focuses on finding *structure* rather than finding "emphasized" or "important" portions of recordings for a summary. I contend that importance is in the *ear of the listener* (as opposed to the *eye of the beholder*). No single summary can be

created to satisfy all listeners and tasks. This research concentrates instead on locating structural boundaries that a listener can use to navigate through the audio. These boundaries allow the listeners themselves to quickly locate the portions of audio that they believe are most important.

5.3 Levels of Structure

The structure of a text document helps a reader locate information of interest. For example, when people browse through a newspaper, they flip from page to page, scan quickly through the headlines, and jump from paragraph to paragraph. Just as structure is used by the reader of a text document, structure can also be exploited by someone listening to a spoken document. This thesis aims to make it easier for a listener to quickly navigate through a speech recording and find the portions of interest by making use of the underlying structure of the spoken material. A contribution of this thesis is the combination of user activity with acoustically-determined structural information.

This thesis focuses on the following levels of acoustic structure for use in the Audio Notebook interface (as shown in the diagram below): unstructured audio stream, speech and pause segmentation, grouping of speech segments into major phrases, grouping of major phrases into discourse segments.



This diagram shows a funnel. At the top of the funnel is the entire unprocessed speech recording, without any navigational landmarks to assist the listener. The narrowest part of the funnel provides the most time-savings, because it allows the listener to jump from topic to topic.

At each stage, the speech is processed to produce another level of structure. Each level of structure is used as a basis for the next:

- First, the system starts out with an unstructured audio stream.
- Next, the unstructured audio stream is processed to determine segments of speech and non-speech audio (Section 5.4). Non-speech audio includes silence, lip-smacks,

breath noises, marker noise from writing on a whiteboard, and page turning noises for talkers who use slides. Intervals of non-speech audio between speech segments are referred to as *pauses* in this document.

- Next, the speech and pause segmentation is used by an algorithm for predicting major phrase breaks (Section 5.9). Segments of speech are grouped together to form major phrases. Once phrases are identified, these become the primary units for further analysis of the speech.
- Lastly, discourse segment analysis is performed by calculating acoustic features over the duration of each major phrase in a recording, and looking at the changes from one phrase to the next. Major phrases are grouped together into discourse or topic segments.

The following sections describe the acoustic analysis and processing that is performed for each of these levels of structure. Chapter 6 describes how this acoustically-determined structural information is combined with user activity in the Audio Notebook.

5.4 Speech Detection

The first level of processing involves segmenting the recordings into intervals of speech and non-speech audio. This is accomplished using a modified version of Arons' speech detection algorithm [Arons 1994a]. This algorithm uses time-domain measures—energy and zero-crossings—to segment a speech recording into intervals of speech and background noise. The algorithm makes several passes through a recording. On the first pass, average magnitude and zero-crossing rate are calculated over 10 ms intervals. A histogram of the energy is used to determine a background noise level; a threshold several dB above the noise floor indicates speech. On additional passes through the recording, the algorithm uses *fill-in* and *hangover* [Gruber 1982] to smooth between intervals of speech and background noise for increased accuracy. For example, if after an initial pass through the data, there is a short "island" of speech (e.g., under 100 ms) between two segments of background noise, this island is *filled-in*, and the entire interval considered background noise. In addition, intervals of speech are slightly extended or *hung over* to avoid any clipping. This algorithm performs accurately, even when speech is recorded in a noisy environment.

One weakness of this algorithm is that it does not distinguish between speech and non-speech audio. Noises such as page turns and the squeak of whiteboard markers get classified as speech because of their high energy. In contrast, a goal for this thesis is to separate intervals of speech from non-speech audio. For example, if a speaker utters a phrase and then pauses and begins to flip through pages in his/her notes before speaking again, for the purposes of this work, the entire interval between the two spoken phrases should be considered as a between-phrase *pause*. Here we define a *pause* as any interval between two segments of speech. This interval may contain pure silence, breath sounds, lip smacks, coughs, page turns, or other noise. The important point is that the talker is not presenting speech information to the listener during this time interval.

In order to distinguish between speech and non-speech audio, the Arons speech detection algorithm was modified to make use of voicing information. The modified algorithm uses voicing and zero-crossing rate. Using the Entropic speech analysis software, the probability of voicing was calculated for every 10 ms of a recording. The probability of voicing output by Entropic's

get–f0 program is binary, where 0 means unvoiced, and 1 means voiced. Voiced speech is identified using the probability of voicing, and unvoiced fricatives are detected using a zero-crossing rate threshold. The smoothing portion of the algorithm helps to create accurate intervals of speech and non-speech audio. The results of the algorithm were matched against hand-labeled phrase pauses. The automatically generated pauses closely match the hand-labeled data. However, weak unvoiced fricatives, and unvoiced stops are sometimes misclassified. Unvoiced stops were also frequently misclassified by the unmodified Arons' algorithm.

5.5 Acoustic Study: Audio Corpus Collection

In order to study how talkers use acoustic cues to signal structural information, a small corpus of speech was collected. The goal of the study was to determine how these cues can be exploited to automatically segment speech recordings. The study focuses on one of the primary domains for the Audio Notebook—recorded lectures. Related studies in the discourse community have used radio speech [Grosz and Hirschberg 1992, Ostendorf et al. 1995], spontaneous and read directions [Hirschberg and Nakatani 1996], and narrative descriptions about a man picking pears [Passonneau and Litman 1997]. There have been relatively few studies of naturally occurring spontaneous speech that could be applied to a real-world application. Here the goal is to apply the results of the study directly for use in the Audio Notebook interface.

This thesis focuses specifically on two aspects of discourse structure:

- Major phrases boundaries (i.e., intonational phrases)
- Discourse segment (i.e., topic) introductions

5.5.1 Domain Selection

The Audio Notebook may be used to record any kind of audio or speech in any language. However, the acoustic processing research for this thesis is limited to native English speakers and the domain of recorded lectures. There are a number of different dimensions along which different types of recorded speech can be categorized. In selecting a domain for this thesis the following variables were considered:

- Read versus Spontaneous Speech. Past research in discourse and intonation has studied read speech more often than spontaneous speech. In terms of building interactive systems, more study of spontaneous speech is needed. Most everyday activities (i.e., lectures, meetings, conversations) employ spontaneous speech. In addition, some read speech such as news is often spoken using a stylized intonation, making the results less generalizable. For example, some radio announcers accent almost every word, regardless of whether or not they are presenting new or important information to the listener.
- Monologue versus Dialogue. This thesis focuses on the structure of speech for a single talker. A study by Ayers states that acoustic cues to topic structure are disrupted by turn taking cues [Ayers 1994]. However, reliable cues to structure in monologues must first be established before the effect of multiple speakers (e.g., turn taking, holding the floor) can be determined. The analyses used in this thesis take advantage of a theoretical and methodological foundation for discourse

segmentation of monologues developed by Grosz, Hirschberg and Nakatani [Grosz and Hirschberg 1992, Hirschberg and Nakatani 1996].

- Genre. Since the Audio Notebook was used by students (Chapter 4), lectures were selected for the acoustic study. A recording of a lecture is useful to augment written notes, to free the user for more focused listening, and to enable review of examples that could not easily be captured in written form.
- Amount of Organization/Structure. Lectures are generally prepared in advance, and therefore will be more "structured" (i.e., the content may be pre-defined, the topics carefully broken down, the timing worked out in advance) than a casual conversation.

The problem of finding structure in recorded speech is multidimensional. This thesis is not intended to solve the entire problem, but instead will focus on a specific "slice" of the problem—spontaneous monologues in the form of lectures. Even within this slice there is still the factor of individual differences. The intonational contours of a dynamic speaker may be closer to that of a radio announcer than the typical classroom lecturer. While adaptations are made for individual speakers, the goal is to find features that are used by multiple speakers.

5.5.2 Subjects

Three male and three female speakers, all graduate students at MIT, gave talks about their current research. Each speaker was given two $25 gift certificates, one for presenting the talk, and one for segmenting it.

Five subjects labeled the discourse segmentation for each of the six lectures. These subjects received $10/hr for their work.

5.5.3 Procedure

There are two parts to the study: (1) recording of lectures, and (2) discourse segmentation performed by the speaker and several listeners. In part one of the study, a small corpus of spontaneously delivered lectures was collected. Subjects participating in this part of the study gave a short lecture on a topic of their choice. A small audience attended each lecture. Talks were recorded onto digital audio tape. The length of the talks ranged from 5 to 15 minutes. The total corpus contains approximately 60 minutes of speech.

In the second part of the study, another group of subjects labeled the discourse structure for each of the lectures. In addition, the student who gave the talk also segmented his/her own lecture. Each segmenter was given a set of instructions for marking the discourse structure developed by Nakatani et al. [1995]. These instructions are based on Grosz and Sidner's theory of discourse structure [Grosz and Sidner 1986] but have been written for naive subjects (i.e., those who have no knowledge of discourse theory). For example, Grosz and Sidner's *flashbacks* are referred to as "out of order" discourse segments. In addition, *interruptions* are called "out of the blue" segments.[6]

---

[6]In describing these to subjects, I also referred to them as "like totally random."

Each subject labeled the starting and ending points of discourse segments, and indicated a hierarchical topic structure (e.g., segment B is a subsegment of A). Subjects viewed a written transcript and listened to the audio recording of the lecture while performing the segmentation. The transcript was segmented into major phrases (Section 5.7) as the primary unit of analysis. Labelers grouped phrases into segments, and groups of segments into a hierarchical structure based on their interpretation of the speaker's purpose. Each subject was given one discourse sample as a training exercise. After reading the instructions and segmenting the sample, subjects could ask questions about the procedure. Following this training exercise, the segmenters were instructed to work independently and not to discuss the segmentation with anyone. The experimenter did not answer any questions specific to a particular discourse; only the general instructions could be clarified.

Subjects performed the discourse segmentation on a part-time basis, a few hours each week for several weeks. Subjects were encouraged to work at their own pace and to take frequent breaks. Figure 5-1 shows the total number of hours it took for each subject to perform the segmentation. Segmenters spent approximately 2–4 hours on each lecture depending upon it's length and complexity.

Segmenter	Total Time (hr's)
1	23.5
2	26.0
3	24.0
4	20.0
5	27.0

Figure 5-1: The total time spent on discourse segmentation of six lectures by each of the five segmenters.

5.5.4 Test Location and Apparatus

Each talk was given before a small audience in a Media Laboratory conference room. Segmentation was performed using a Sun Workstation in the "Terminal Garden" area on the third floor of the Media Laboratory. Segmenters were provided with headphones for private listening and to block out any background noise. In a few cases, subjects worked in a private office where an additional workstation was located.

The Nota Bene program [Flammia and Zue 1995] was used by subjects performing the discourse segmentation. This tool provides a graphical user interface for labeling discourses. Subjects view the transcript in a scrolling window and can select the portions they want to hear (Figure 5-2). The tool also provides a feature for creating a color coded view of the structure that has been labeled. This helps a subject review the work they have done and make additions or corrections as needed.

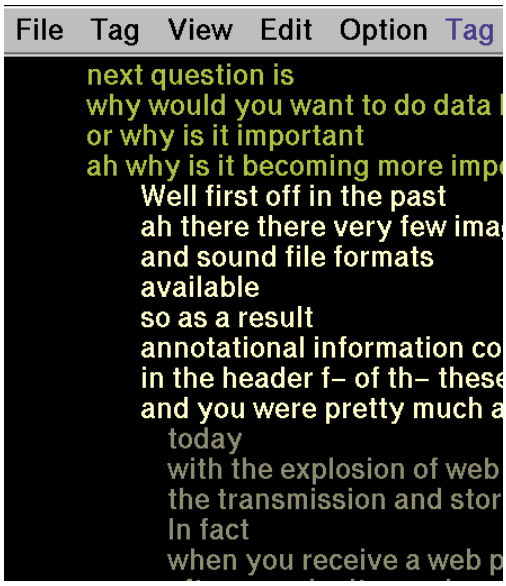


Figure 5-2: This is a screen capture of Nota Bene with one of the subject's segmentations displayed in the window. Colors are used to indicate discourse segments labeled by the subject, and indentations indicate hierarchical structure.

5.6 Theoretical Foundations

For each lecture in the corpus, intonational phrase breaks were manually labeled according to the ToBI labeling standard [Beckman and Ayers]. ToBI represents a collaborative effort by many discourse theorists to establish a prosodic labeling standard [Silverman et al. 1992]. The standard draws from several theories of English prosody, in particular, Pierrehumbert's theory of English intonation [Pierrehumbert 1975, Pierrehumbert and Hirschberg 1990]. This standard has four "tiers" of labels—Tones, Orthographics, Break Indices, and miscellaneous. Two tiers were used in this study—orthographics and break indices. The orthographic tier contains a transcript of the speech recording, time aligned with the speech waveform. Break indices are described in the next section (Section 5.6.1).

The ToBI system uses Pierrehumbert's definitions of intermediate (minor) and intonational (major) phrase breaks. According to Pierrehumbert's theory, an intermediate phrase contains one or more pitch accented syllables followed by a phrase accent, H (high) or L (low). An intonational phrase contains one or more intermediate phrases followed by a boundary tone (H or L). Figure 5-3 gives a pictorial representation of Pierrehumbert's theory.

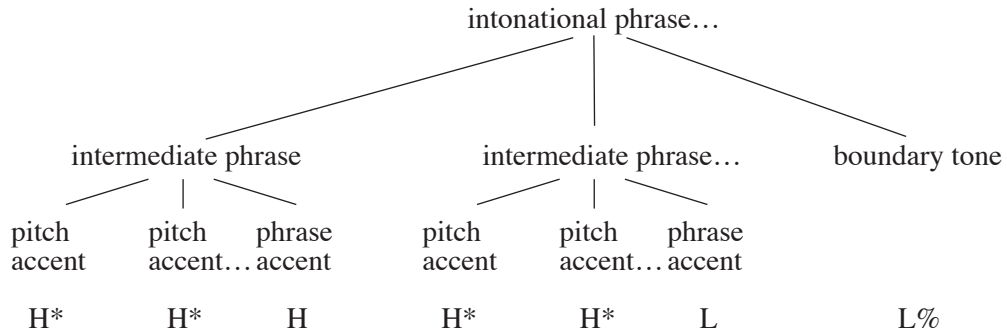


Figure 5-3: Pictorial representation of Pierrehumbert's theory of English intonation.

5.6.1 Break Indices

As defined by the ToBI annotation conventions, *break indices* "represent a rating for the degree of juncture perceived between each pair of words and between the final words and the silence at the end of the utterance" [Beckman and Hirschberg]. ToBI defines four levels of break indices. Below is a brief definition of each type of break index, starting at the highest level:

4 Level 4 represents Pierrehumbert's intonational phrase boundaries or *major* phrase breaks. These are marked by a boundary tone (H or L) that follows the last phrase accent.

3 Level 3 represents Pierrehumbert's intermediate phrase boundaries or *minor* phrase breaks. These are marked by a phrase accent phrase accent (H or L) following the last pitch accented syllable.

2 Level 2 boundaries are marked when a significant disjuncture can be heard but it is not clearly a full phrase boundary. For example, there could be "a strong disjuncture marked by a pause" but with no tonal changes across the boundary normally associated with an intermediate or intonational phrase [Beckman and Hirschberg].

1 Level 1 boundaries are placed between words within a phrase where there is not coarticulation.

0 Level 0 boundaries are used in cases where there is coarticulation between two words; also referred to as "clitics." For example, a 0 break index would be placed between the words "gas shortage" if they where pronounced "gashortage" (with an elision of the /s/ phoneme).

Labeling is performed by a combination of looking at the waveform, pitch track, and sometimes spectrogram, and listening to the audio recording. A ToBI labeler goes through "ear training" exercises by listening to many sample utterances. When labeling a particular audio recording, the labeler also calibrates to the speaker. The disjuncture between words will differ somewhat from speaker to speaker. For example, one of the speakers in the corpus often increased his speech rate at phrase boundaries. While these boundaries exhibited clear tonal changes associated with a phrase boundary (e.g., continuation rise), the speaker oftentimes sounded as if they were rushing from one phrase to the next.

A portion of speech from one of the speakers in the corpus is given in Figure 5-4. In this diagram, there are three windows—one window for the speech waveform, one window for ToBI labels, and a third window containing a pitch track of the recording. The label window shows the two label tiers used in this study, orthographics and break indices. In this example, there is one intonational phrase which contains two intermediate phrases. The level 3 intermediate phrase boundary (following "research") is marked by an H- phrase accent. The phrase ends with an LH% continuation rise. An H- phrase accent or boundary tone often indicates that there is related information to follow [Hirschberg and Pierrehumbert 1986]. In this case, the speaker goes on to say "and the topic I'll be covering today is data hiding in audio." Notice that there is an H* accent aligned with "name", marked by a peak in the F0 contour. Note that these accents are described as an example for readers who are unfamiliar with Pierrehumbert's theory of English intonation. Accents (i.e., the tone tier) were not labeled for this study.

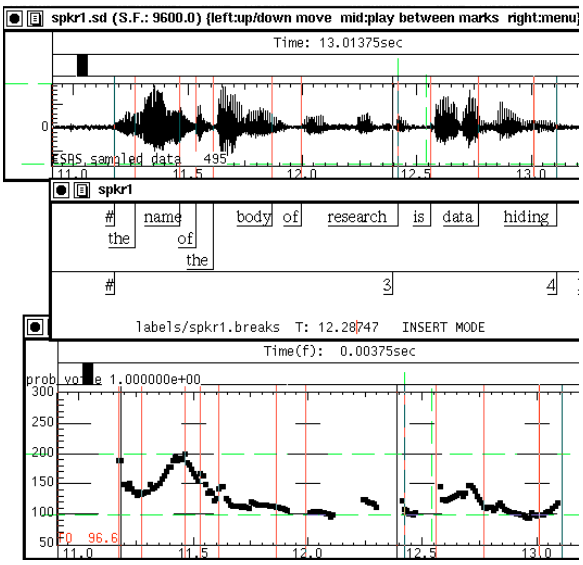


Figure 5-4: Segment of speech from the corpus containing one intonational phrase, and two intermediate phrases.

In a study of labeler consistency for the ToBI standard [Silverman et al. 1992], there was 91% agreement for the placement of phrase boundaries (levels 3 and 4) across 20 labelers of varying experience. There was 81% agreement for distinguishing between level 3 and level 4 boundaries. The lowest consistency was for levels 0–2 (not used in this study). Agreement for these boundaries was 67% across all labelers.

5.7 Speech Labeling

First, each of the six recordings were transcribed for use in the discourse segmentation portion of the study. Recall that Nota Bene displays a transcript to the subjects labeling discourse structure (Figure 5-2). The transcripts include filled pauses such as “um” and “uh” but no comments about the speech (e.g., pause lengths).

Next, intonational phrase breaks were manually labeled according to the ToBI labeling standard, using the Entropic speech analysis software. Labels of major phrase breaks were needed for two main purposes:

1. In order for subjects to code the discourse segmentation, the speech is first manually segmented into major phrases. Using Nota Bene, subjects group phrases into discourse segments. A transcript is provided for Nota Bene, with one major phrase on each line. In addition, phrase beginning and ending times are also needed so that subjects can play the audio associated with any phrase or group of phrases.

2. Manual labels of major phrase breaks are needed for studying acoustic cues associated with major phrase beginnings and discourse segment beginnings. For example, using the manual labels of major phrase beginnings and endings, the duration of the pause preceding each phrase can be calculated. For discourse structure analysis, acoustic features are calculated for each intonational phrase in a recording.

Level 4 break indices (i.e., intonational or major phrases) were marked by one labeler with minimal experience and verified by an expert[7]. Both labelers were trained in the ToBI labeling system; the "verifier" is an expert and founder of the system. Figure 5-5 is a window of speech from one of the recordings, showing how the break index labels appear.

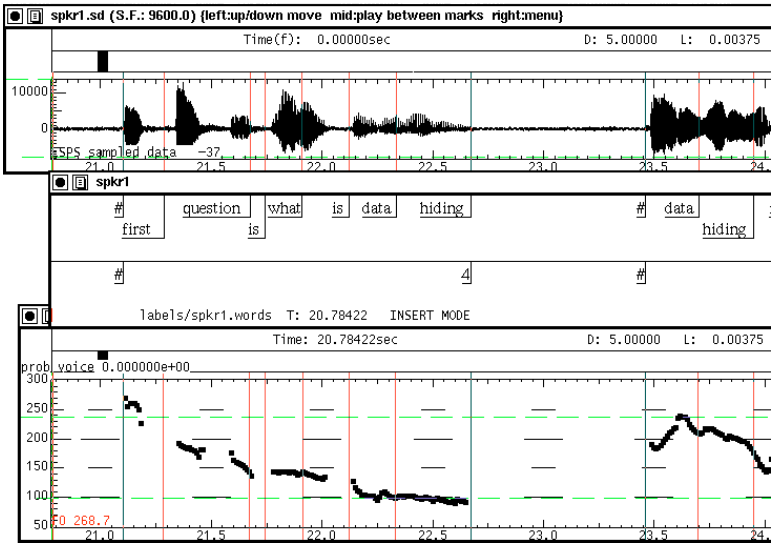


Figure 5-5: A window of speech with level 4 break indices labeled.

In ToBI, the end of phrases are marked with a break index level (3 = minor phrase, 4 = major phrase). The beginning of each phrase is labeled using a "#" symbol. Both the beginning and ending of each phrase needs to be marked in order to calculate the amount of pause time between each phrase. In addition, as shown in Figure 5-5, some of the orthographics tier has been labeled. Although each speech recording was transcribed entirely for use with Nota Bene, only the words corresponding to level 4 break indices and certain content words were time-aligned with the speech waveform[8]. Some words were time-aligned with the speech waveform to make it easier to review the break indices and find portions of the speech when necessary. Some non-speech information was also labeled in the miscellaneous tier (e.g., page turns, coughs).

Figure 5-6 shows a portion of a break index label file (.break) generated by the Waves software from the hand-entered labels.

[7]The verifier checked the break index labels for a portion of each recording, but not the entire discourse.

[8]Note that some speech recognition systems can be used to automatically align a text transcript with a speech recording.

```
signal spkr1
comment created using xlabel Thu Feb 22 15:49:37 1996
#
     0.518063  123 #
     1.772820  123 4
     1.981939  123 #
     2.445346  123 4
     2.511886  123 #
     2.856242  123 4
     3.324400  123 #
     5.389741  123 4
     5.905419  123 #
     9.287080  123 4
    10.004085  123 #
    10.457984  123 4
```

Figure 5-6: Break index label file. The first column contains the time point of the label in seconds, the second is the color, and the third is the label (# = begin phrase, 4 = end of major phrase).

5.7.1 Automatic Pause Labeling

In addition to the hand labels, labels were also automatically generated using speech processing. The ToBI break index labels marking phrase beginnings and endings can be used to calculate the length of the pause between each phrase. However, this does not capture all of the pauses in a recording. Pauses also occur within phrase boundaries. Hand labeling each pause in a recording would be extremely time-consuming. In order to speed up this process, each recording was also processed using the modified version of Arons' speech detection algorithm (Section 5.4). This algorithm segments a speech recording into intervals of speech and non-speech audio. Intervals of non-speech audio between two segments of speech are considered pauses.

All automatically detected pauses were matched against the hand labeled between-phrase pauses. Pauses occurring within phrases were put into a separate label tier from the between-phrase pauses. Figure 5-7 shows a sample window of labeled speech; PS (Pause Start) stands for the starting point of a within-phrase pause, and PE (Pause End) for the end of the pause. A label file like the one shown in Figure 5-6 is generated for these within-phrase pauses.

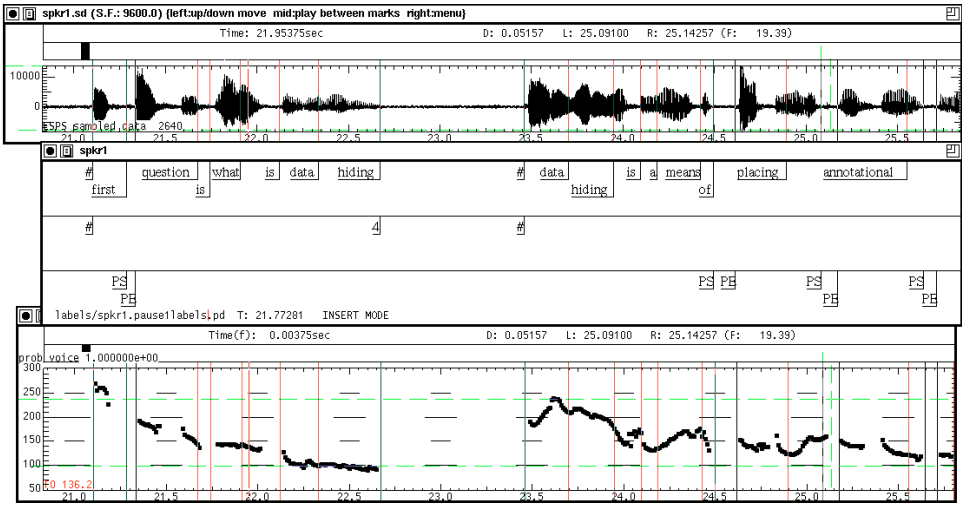


Figure 5-7: A window of speech showing within-phrase pause labels (PS, PE).

Arons' speech detection algorithm sometimes labels an interval of speech as silence when there is weak voiced speech or unvoiced stops. The automatically generated within-phrase pause labels were manually edited to remove these false alarms.

5.8 Evaluation Metrics

This section defines the metrics that are used to evaluate the segmentation algorithms developed in this thesis. Hand labels of features (e.g., phrase or discourse segment beginnings) are compared against predictions made by a segmentation algorithm. For example, if the problem is to automatically predict the location of discourse segment beginnings, then there are two classes: segment beginnings and non-segment beginnings. Segment beginnings are the target class. The number of hits, misses, false alarms, and correct rejections are determined as shown in Figure 5-8. For example, if a segmentation algorithm predicts that a phrase is a segment beginning and it is also hand labeled that way, then this constitutes a hit.

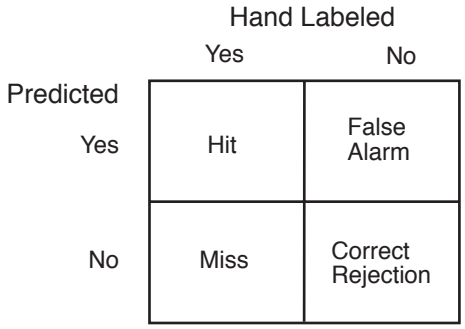


Figure 5-8: Hand-labeled features are compared against predictions made by segmentation algorithms to determine the number of hits, misses, false alarms, and correct rejections.

The following evaluation metrics are then calculated: recall, precision, fallout and error (Figure 5-9). *Recall* (or hit rate) is equivalent to the percent correct identification for the target class (i.e., segment beginnings). *Precision* is the number of correctly identified segment beginnings out of the total number hypothesized. *Fallout* (or false alarm rate) is the number of false alarms out of the total number of items in the non-target class. *Error* is the overall number of errors out of the total number of items in the data set.

$$\text{Recall} \qquad \frac{H}{H + M}$$

$$\text{Precision} \qquad \frac{H}{H + FA}$$

$$\text{Fallout} \qquad \frac{FA}{FA + CR}$$

$$\text{Error} \qquad \frac{FA + M}{H + FA + M + CR}$$

Figure 5-9: Evaluation metrics. H = Hits, M = Misses, FA = False Alarms, CR = Correct Rejections.

5.9 Phrase Beginning Prediction

Major phrase boundaries may be signaled by a pause, or changes in pitch, energy, and phoneme duration, in particular the lengthening of the final syllable in a phrase [Wang and Hirschberg 1992]. This thesis primarily explores the use of pauses as a low cost method of identifying major phrase breaks. The problem is to distinguish between pauses that occur between phrases and those occurring within a phrase. This is a classification problem with one continuous independent variable (pause length) and two categorical classes (within-phrase and between-phrase pauses).

A pause distribution was generated for each recording in the corpus using the break index and within-phrase pause labels (Section 5.7). Each pause in a recording is classified as a within-phrase or between-phrase pause. Figure 5-10 shows a portion of the pause data for one of the speakers. There is a row for each pause in the recording with columns for the start time of the pause, the end time, the pause length, and the type of pause (1 = within-phrase, 2 = between-phrase).

StartTime	EndTime	Length	Class
0	520	520	2
1770	1980	210	2
2450	2510	60	2
2860	3320	460	2
4050	4130	80	1
4380	4470	90	1
5390	5910	520	2
7910	7980	70	1
8640	8710	70	1
9290	10000	710	2
10460	11180	720	2

Figure 5-10: Sample data taken from one of the recordings. The first column is the start of the pause (in milliseconds), the second is the end of the pause, the third is the pause length, and the last is the class of the pause (1 = within-phrase, 2 = between-phrase).

The pause data for each speaker was processed using CART (Classification and Regression Tree Analysis). The implementation of CART used in this thesis was developed by Michael Riley of AT&T Research [Riley 1989] (for a more detailed description, see Appendix A). CART was provided with the pause length and class (the last two columns of data shown in Figure 5-10). CART then selected a pause length threshold to split the data into the two classes that minimizes the number of classification errors. Figure 5-11 shows the tree that was generated for the first speaker in the corpus. This process was repeated for each of the speakers so that speaker-dependent pause thresholds were generated for separating between-phrase from within-phrase pauses. Figure 5-12 shows the pause thresholds determined for each of the speakers in the corpus.

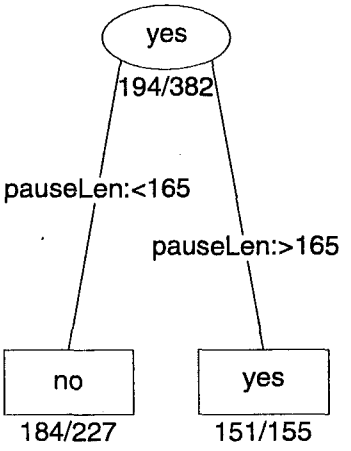


Figure 5-11: Tree generated by the CART program for the first speaker in the corpus. Yes = a between-phrase pause, and No = a within-phrase pause.

Speaker	Pause Threshold (ms)
1	165
2	175
3	95
4	155
5	105
6	105

Figure 5-12: Pause threshold for each speaker in the corpus. This threshold splits the data into two classes, between-phrase and within-phrase pauses, minimizing the classification error.

In splitting two classes using a single feature, the total classification error is minimized at the "crossover" point between the two data sets. Figure 5-13 shows a hypothetical probability density functions for two classes. Two types of error are shown in shaded gray areas—misses and false alarms. The number of total misclassifications, misses plus false alarms, is minimized when a decision boundary is set at the crossover point shown.

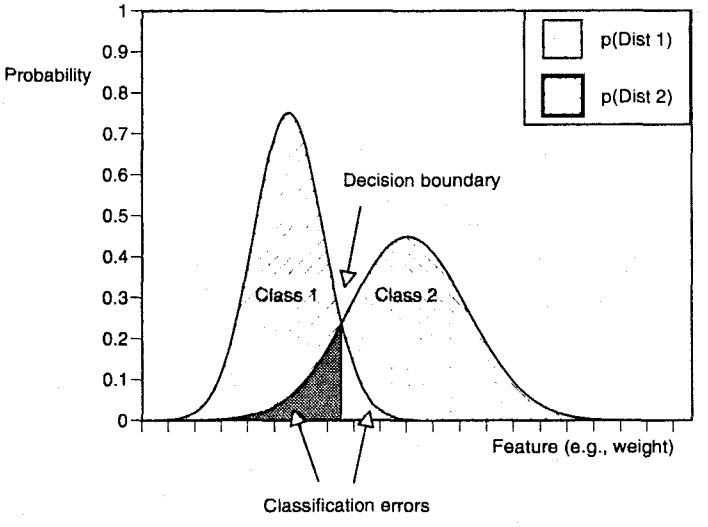


Figure 5-13: When distinguishing between two classes using a single feature (i.e., one dimension), classification error is minimized when a decision boundary is chosen at the cross-over point between the two probability densities.

The goal is to distinguish the between-phrase pauses from those occurring within a phrase. The pause thresholds shown in Figure 5-12 allow identification of a high percentage of between-phrase pauses, while making few false alarms. Figure 5-14 shows the evaluation metrics for identifying between-phrase pauses using the CART-generated pause threshold for each speaker in the corpus.

Speaker	%Recall	%Precision	%Fallout	%Error
1	78	97	2	12
2	81	100	1	14
3	91	99	9	9
4	87	99	1	10
5	89	99	4	10
6	95	99	10	5

Figure 5-14: Evaluation metrics for predicting between-phrase pauses using the CART-generated pause threshold.

The next section analyzes these results by taking a closer look at the pause distributions for the speakers in the study.

5.9.1 Pause Distributions

For each speaker in the corpus, a histogram was created of all the pauses in the recording. Figure 5-15 shows a pause distribution for one of the speakers in the study.

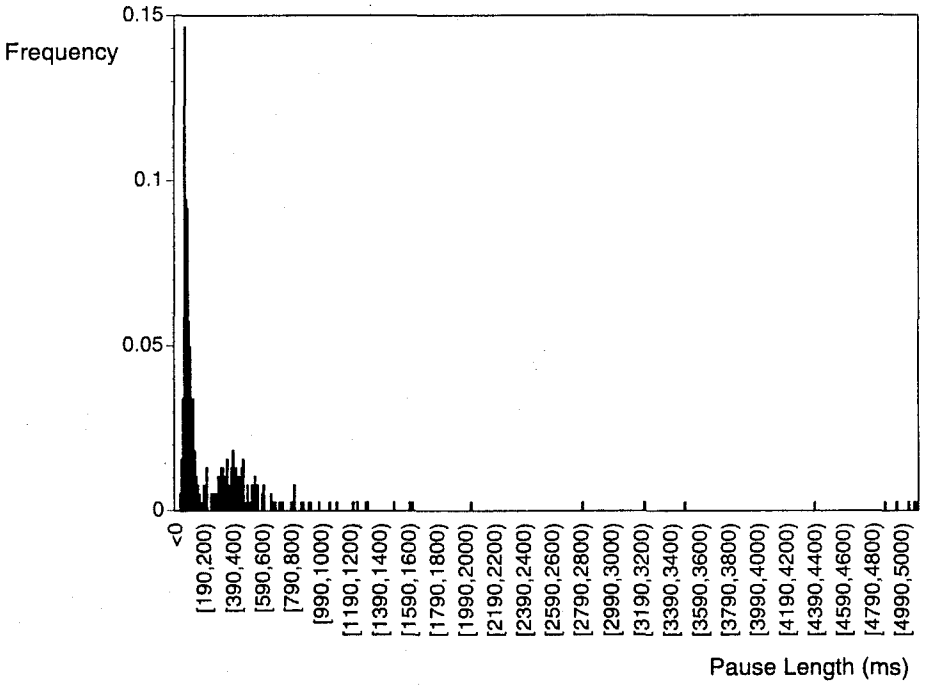


Figure 5-15: Distribution of all pauses for speaker 1.

Duration data such as pause length tends to be shaped like a decaying exponential because it is always greater than zero and has no defined upper limit. The pause length data extends over several orders of magnitude. A log transformation of the raw pauses helps to aggregate the data, particularly over the upper range of the distribution, and reduces sparse data areas (i.e., histogram

bins without any samples). A log transformation also makes the distribution more Gaussian-shaped which has many advantages for data analysis.

Figure 5-16 is a log-transformed pause distribution for speaker 1. The pauses are also separated into two classes—one for the pauses occurring within a phrase (within-phrase pauses) and one for those occurring between phrases (between-phrase pauses). The bottom plot shows a histogram for all of the pauses in the recording; the top plot shows the distribution for the within-phrase pauses alone; the middle plot shows the between-phrase pauses alone.

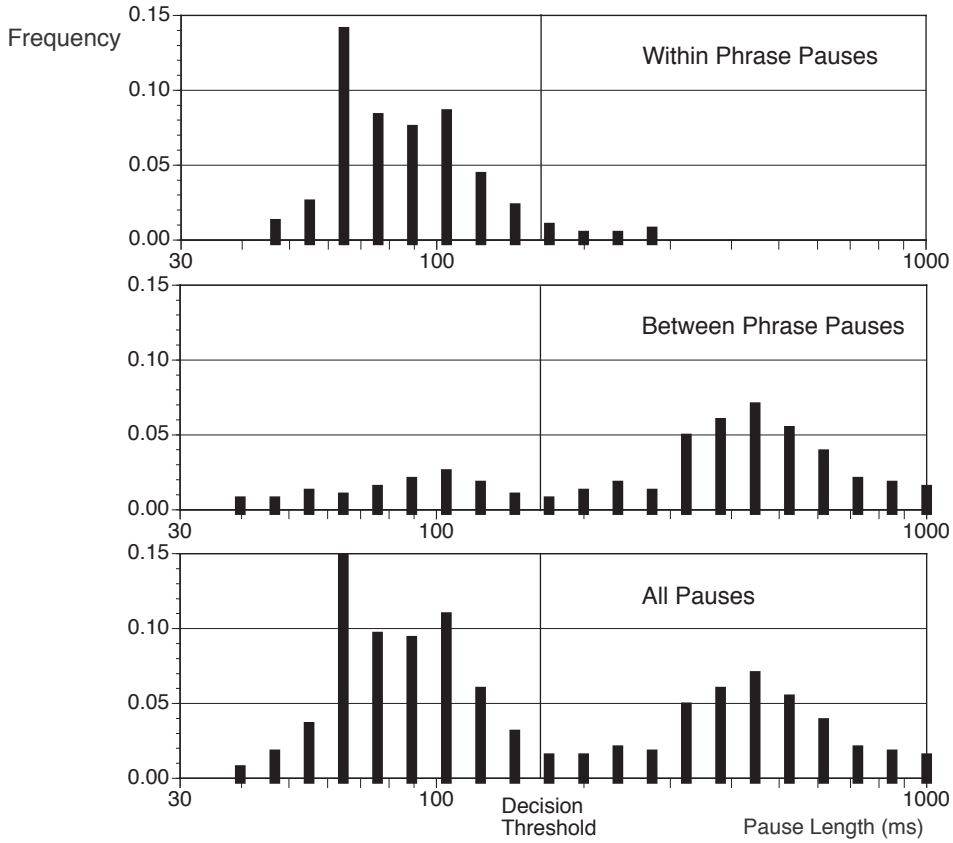


Figure 5-16: Log transformed pause distributions for speaker 1. The bottom plot shows all pauses in the recording. The pauses have been separated into two classes: within-phrase pauses (top plot) and between-phrase pauses (middle plot). For this speaker, the decision threshold that minimizes the number of misclassification errors is at 165 ms.

Given the speaker-dependent pause threshold for distinguishing between-phrase pauses from those occurring within a phrase, the precision for speaker 1 is 97% and the recall is 78% (Figure 5-14). The precision for all the speakers is high because of the limited range of within-phrase pauses (Figure 5-17). For speaker 1, there are only four within-phrase pauses above the threshold of 165 ms (i.e., only four false alarms). In addition, 78–99% recall is achieved using just the single cue of pause length. The percent recall is lower than the precision because between-phrase pauses occur over a wide range (Figure 5-18) and overlap with the high end of the within-phrase distribution.

For this application, precision is preferred over recall. Starting in the middle of a talker's phrase is more detrimental than starting playback one or two phrases early. Subjects in the SpeechSkimmer user study found it difficult to comprehend the speech if playback began in the middle of a phrase

or was abruptly interrupted [Arons 1994a]. Listeners also have difficulty establishing the context of the information if playback starts in the middle of a phrase.

Speaker	Mean	Std. Dev.	Range	Min	Max
1	83	32	210	40	250
2	80	26	150	50	200
3	72	31	180	40	220
4	72	28	150	30	180
5	58	21	100	30	130
6	82	62	220	30	250

Figure 5-17: Within-phrase pause statistics for each speaker in the corpus. Numbers are in milliseconds.

Speaker	Mean	Std. Dev.	Range	Min	Max
1	593	903	5090	30	5120
2	443	446	3830	20	3850
3	546	825	7351	19	7370
4	414	396	2520	40	2560
5	526	692	6740	20	6760
6	431	332	2380	30	2410

Figure 5-18: Between-phrase pause statistics for each speaker in the corpus. Numbers are in milliseconds.

5.9.2 Automatic Speaker Adaptation

The pause length thresholds distinguish between-phrase from within-phrase pauses at a high rate of precision and recall, yet these thresholds are speaker dependent and based on a single lecture from each talker. The painstaking and time-consuming nature of the labeling procedure makes it prohibitive to accomplish for every talker recorded using the Audio Notebook. In addition, a talker's style may differ from lecture to lecture, so a threshold established for one recording, might not be appropriate for another.

The next challenge was to automatically select a threshold for an unknown speaker and an unknown recording. The next two sections will show two methods based on previous research that have limited utility for solving this problem. The third section will describe an approach based on pattern classification techniques which is used to adaptively determine a threshold for each speaker and for each lecture.

5.9.2.1 Fixed Threshold Approach

A simple approach used by many researchers is to select a single fixed threshold for the features in question. This is generally accomplished by pooling the data across multiple speakers. For example, Grosz and Hirschberg combine the data for three radio news stories, and state that discourse segment beginnings can be predicted using "a simple combination of constraints on duration of preceding pause (> 647 ms) and pitch range (< 276 Hz)" [Grosz and Hirschberg 1992]. Although these results are promising for the use of pause and pitch for predicting discourse segment breaks, it is unlikely that these fixed thresholds will be applicable over a wide range of speakers, even if selected from the same domain.

If the pause length data for all the speakers in the corpus is combined and processed using CART, a single fixed threshold of 155 ms is obtained. Note that the speaker dependent thresholds ranged from 90–170 ms. Figure 5-19 shows the evaluation metrics using this fixed threshold for all speakers. The recall is 82%—close to the lowest recall obtained using the speaker dependent thresholds (Figure 5-14). Although the precision is very high, it may be possible to maintain a

high precision while improving the recall using another technique for determining the threshold. Again, it is also unlikely that these results will be applicable across a wider range of speakers with varying pausing styles.

Hits	1328
Misses	293
False Alarms	12
Correct Rejections	538
% Recall	82
% Precision	99
% Fallout	2
% Error	14

Figure 5-19: Evaluation metrics based on a single fixed threshold for all speakers.

5.9.2.2 Frequency-Based Threshold Approach

An approach used by Arons for SpeechSkimmer's pause-based skimming, is to select a threshold based on a percentage of a speaker's pause distribution [Arons 1997]. Segments of speech following pauses above a threshold are selected for playback. For example, the top 1% of a speaker's pause distribution could be used as a threshold for speech segment selection. This is a significant improvement over a fixed threshold (e.g., 1000 ms) since it adapts to each speaker's range. However, Arons varies the percentage based on the desired amount of compression (i.e., the fewer speech segments selected, the more compression) rather than anything characteristic of the speaker.

A similar approach was attempted for adaptively selecting a threshold for distinguishing between-phrase from within-phrase pauses. Figure 5-20 shows the percent cumulative frequency (i.e., percent of the speaker's pause distribution) associated with each pause threshold previously determined from the hand-labeled pause data. Thresholds fall anywhere from 10 to 60% of the speaker's pause range. Therefore, it would be inappropriate to select a single percentage for all speakers.

Speaker	Pause Threshold	%Cumulative Frequency
1	165	59.5%
2	175	42.5
3	95	19.0
4	155	39.5
5	105	25.5
6	105	10.5

Figure 5-20: Pause threshold in milliseconds, and pause threshold converted to a percentage of the speaker's cumulative frequency.

The plot in Figure 5-21 shows that the relationship between pause threshold and percent cumulative frequency does not monotonically increase—a higher pause threshold does not always correspond to a higher percentage of the speaker's distribution.

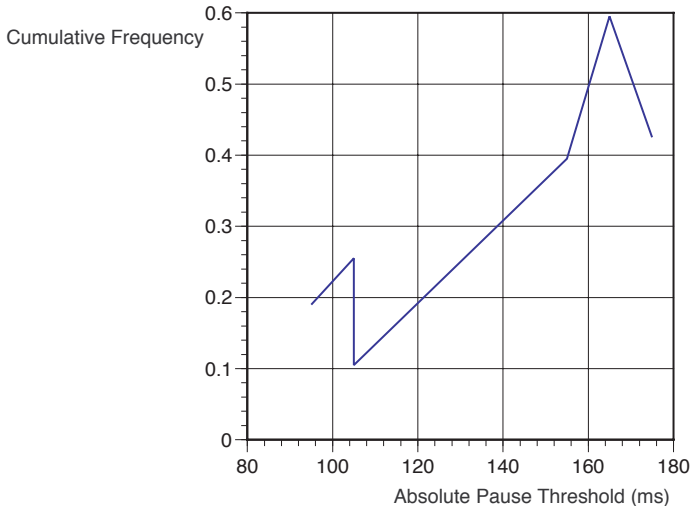


Figure 5-21: This plot shows the speaker-dependent pause thresholds in milliseconds versus the pause thresholds converted to a percentage of the speaker's cumulative frequency for the six speakers in the corpus.

5.9.2.3 Pattern Classification Approach

Figure 5-16 showed the pause distributions for one of the speakers. Since this data was manually labeled, the overall pause distribution can be separated into two classes—within-phrase and between-phrase pauses. Ideally, one would like to be able to start with the complete pause distribution and automatically separate it into these two classes.

This can be accomplished using an unsupervised pattern classification technique. Supervised learning "assumes that the training samples used to design a classifier were labeled to show their category membership" [Duda and Hart 1973, 189]. Unsupervised techniques are needed for unknown speakers and recordings where the classes have not been labeled in advance. The problem of separating between-phrase from within-phrase pauses involves estimating the parameters for a mixture of two Gaussians. Each Gaussian corresponds to one of the two classes—between-phrase and within-phrase pauses. This is an unsupervised learning problem where all of the parameters of the mixtures are unknown, but the classes (between-phrase and within-phrase) are defined. Unsupervised learning is particularly well-suited to classification problems when "collection and labeling of a large data set of sample patterns" is required because this "can be surprisingly costly and time consuming" [Duda and Hart 1973, 189].

An iterative technique known as *EM* (Expectation Maximization) is often employed for iteratively estimating parameters of Gaussian mixtures [Dempster et al. 1977, Bishop 1995]. This technique involves starting with initial estimates for the Gaussian parameters—μ_i (mean), σi^2 (variance), and $P(\omega_i)$ (a priori probability of each class)—based on the training samples. Note that $P(\omega_i)$ is equivalent to the weight of each distribution. For example, for speaker 1, the within-phrase pauses account for 49% of the data, and the between-phrase pauses account for 51% of the data. Using these initial estimates, the parameters are iteratively estimated until convergence is achieved. Figure 5-22 lists the equations for updating the maximum likelihood estimates for the parameters $P(\omega_i)$, μ_i, and σi^2 at each iteration. There are three parameters to be estimated for each of the two classes: $P(\omega_1)$, μ_1, σ_1^2, $P(\omega_2)$, μ_2, and σ_2^2.

For ω_i $(i = 1,\ 2)$ where ω_i is the class,

x_j is the pause length, and N is the number of samples:

$$\hat{P}(\omega_i) = \frac{1}{N}\sum_{j=1}^{N} P(\omega_i|x_j) \qquad (1)$$

$$\hat{\mu}_i = \frac{\sum_{j=1}^{N} P(\omega_i|x_j)x_j}{N\hat{P}(\omega_i)} \qquad (2)$$

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{N} P(\omega_i|x_j)\left[x_j - \hat{\mu}_i\right]^2}{N\hat{P}(\omega_i)} \qquad (3)$$

where according to Bayes Rule:

$$P(\omega_i|x_j) = \frac{p(x_j|\omega_i)P(\omega_i)}{p(x_j)} \qquad (4)$$

$$p(x_j) = \sum_{i=1}^{2} p(x_j|\omega_i)P(\omega_i) \qquad (5)$$

Given gaussian mixtures:

$$P(x_j|\omega_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_j-\mu_i)^2}{2\sigma_i^2}} \qquad (6)$$

Figure 5-22: Equations for iteratively estimating the parameters of the Gaussian distributions for each class using the EM algorithm. $P(\omega_i)$ = a priori probability, $P(\omega_i|x)$ = posteriori probability, μ = mean, and σ^2 = variance.

The estimation procedure is as follows:

1. Initial estimates of the parameters are calculated by averaging the mean, variance, and a priori probability for five out of the six speakers in the corpus. The speaker being tested is left out of the training data. This procedure is known as leave-one-out cross-validation, or jack knifing.

2. This is known as the E-step of the EM algorithm. First, the class-conditional probability densities, $p(x_j|\omega_1)$ and $p(x_j|\omega_2)$, are calculated for j = 1 to N where N is the number of pauses in the recording (equations 5 and 6). The first time through, these calculations use the initial estimates of μ_i and σ_i^2.

   The posteriori probabilities, $P(\omega_1|x_j)$ and $P(\omega_2|x_j)$, are then calculated according to Bayes Rule (equation 4). The first time through, the initial estimates of the a priori probabilities, $P(\omega_1)$ and $P(\omega_2)$, are used.

3. This is known as the M-step of the EM algorithm. The parameter estimates, $\hat{P}(\omega_i)$, $\hat{\mu}_i$, and $\hat{\sigma}_i^2$, are updated (equations 1–3).

4. Go to step 2 and iterate.

Forty iterations were run and the results graphed to validate that the estimates were converging. Figure 5-23 shows the Gaussian estimates and their sum for the speaker data given previously in Figure 5-16. Figure 5-24 shows the Gaussian estimates overlaid on the raw pause histogram data.

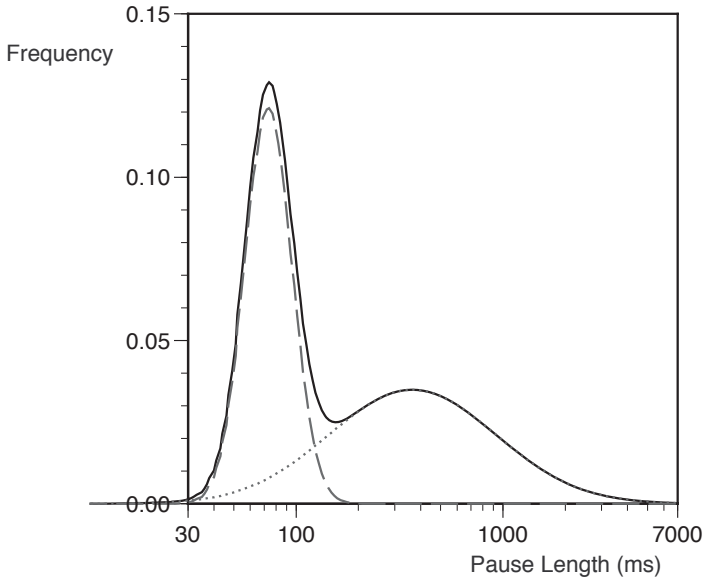


Figure 5-23: This figure shows the Gaussian estimations for each class and the sum of the two estimates for speaker 1. The dashed line is the Gaussian estimate for the within-phrase class of pauses. The dotted line is the Gaussian estimate for the between-phrase class. The solid line is the sum of the two Gaussian estimations.

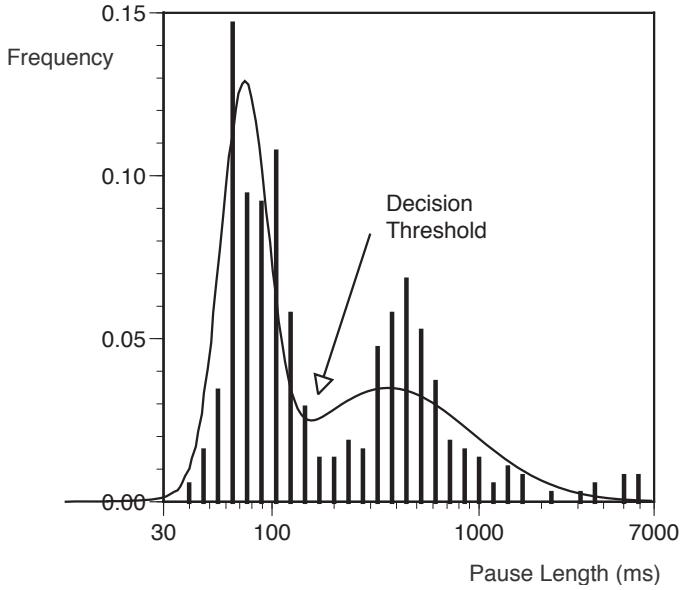


Figure 5-24: This figure shows the sum of the two Gaussian estimations (one for within-phrase pauses and one for between-phrase pauses) overlaid on the raw pause histogram for speaker 1. The local minimum (155 ms) is selected as the decision threshold. This differs by 6% from the threshold determined using the hand-labeled data (165 ms).

Once the parameters are estimated, a threshold can be selected in a number of ways. A common technique is to pick a cross-over point between the two estimated curves (as previously shown in Figure 5-13). When the cross-over point is selected as the decision threshold, the overall number

of misclassification errors is minimized. However, for these distributions, selecting the local minimum of the summed curve more closely approximated the pause thresholds determined using the hand-labeled classes.[9]

The pause thresholds shown in Figure 5-12 will now be referred to as *supervised*, because they were selected based on hand labels of the classes (i.e., each pause was hand-labeled as a within-phrase pause or between-phrase pause). The pause thresholds determined using the automatic estimates of the two class distributions (as shown in Figure 5-24) will be referred to as *unsupervised* because they were determined without hand labels of the classes. The unsupervised thresholds are also adaptive in that they are dynamically estimated based on the distribution of pauses for a particular recording. Figure 5-25 shows the *supervised* pause thresholds based on the hand-labeled data, in comparison to the *unsupervised* automatically estimated pause thresholds.

Speaker	Supervised Threshold	Unsupervised Threshold	% Difference
1	165	155	–6.1
2	175	106	–39.4
3	95	110	15.8
4	155	114	–26.5
5	105	110	4.8
6	105	88	–16.2

Figure 5-25: Comparison of the *supervised* pause thresholds (in milliseconds) selected using the hand-labeled data to the *unsupervised* automatically estimated pause thresholds.

The unsupervised pause threshold for speaker 2 has the largest difference from the supervised threshold. This is partly because the upper tail of the distribution still exhibits characteristics of a decaying exponential rather than a normal shape, even after a log transformation[10] was performed (Figure 5-26, graph on the left). One way to improve the results is to place an upper limit on the pauses considered. For example, all pauses above 1000 ms might automatically be considered as between-phrase pauses and the parameter estimation performed for a distribution of all pauses below this amount. The graph on the right side of Figure 5-26 shows the estimations when pauses above 1000 ms have been eliminated. The resulting unsupervised estimate of the pause threshold changes from 106 ms to 159 ms, much closer to the supervised pause threshold (170 ms). Another alternative would be to use a nonparametric technique to smooth the raw pause histogram and determine a local minimum. Nonparametric techniques do not make assumptions about the shape of the distribution (i.e., Gaussian or otherwise).

---

[9]The data in the upper range of the distribution is skewing the Gaussian estimations somewhat. The pause duration data, even after log transformation, still exhibits characteristics of a decaying exponential rather than a Gaussian distribution. This may be the reason that the local minimum results in a closer estimate of the hand-labeled pause thresholds than the cross-over point between the two Gaussian estimates. An alternative approach would be to use a nonparametric technique for smoothing the pause histogram and selecting a local minimum, rather than assuming a Gaussian-shaped distribution.

[10]As discussed in Section 5.9.1, a log transformation was performed on the pause data in an attempt to aggregate the data over the upper range of the distribution, and make the data appear more Gaussian-shaped.

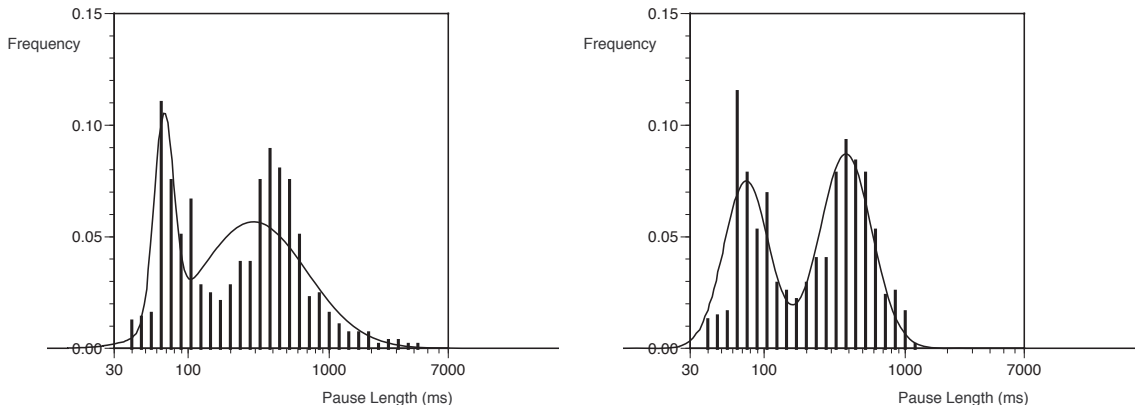


Figure 5-26: The sum of the two Gaussian estimations overlaid on the raw pause histogram for speaker 2. The graph on the left shows the estimates when all pauses are considered. The graph on the right shows the estimates when an upper limit of 1000 ms is placed on the pause data. The unsupervised estimate of the pause threshold (local minimum) is 106 ms in the graph on the left, and 159 ms in the one on the right (closer to the supervised pause threshold of 170 ms).

5.9.2.4 Evaluation of Results

Figure 5-27 shows the resulting evaluation measures when the unsupervised thresholds are used for distinguishing between-phrase pauses from within-phrase pauses. Figure 5-28 shows a comparison of the supervised, unsupervised, and fixed pause thresholds for predicting between-phrase pauses. The unsupervised method results in a higher overall recall (87%) than the fixed threshold (82%) while maintaining a high precision (97%). The evaluation metrics for the unsupervised thresholds closely approximate the results achieved with supervised thresholds determined using the hand-labeled data.

Speaker	%Recall	%Precision	%Fallout	%Error
1	78	97	3	12
2	87	94	14	13
3	86	99	6	13
4	91	97	7	9
5	87	99	3	11
6	96	98	30	6
All	87	97	7	11

Figure 5-27: Evaluation metrics for predicting between-phrase pauses using the unsupervised pause thresholds.

	Threshold Selection		
	Supervised	Unsupervised	Fixed
Recall	86	87	82
Precision	99	97	99
Fallout	2	7	2
Error	11	11	14

Figure 5-28: Unsupervised versus fixed methods for predicting between-phrase pauses in comparison to the supervised thresholds. These evaluation metrics were calculated across all speakers.

Note that these approaches will only detect phrases that are preceded by a pause. Some major phrases are indicated by a change in pitch, energy, or phoneme duration, without a preceding pause. For example, a pitch change such as a continuation rise can mark a phrase break without an associated pause. Figure 5-29 shows the percentage of major phrases preceded by a pause (i.e., pause length > 0 ms) for each of the speakers in the corpus, and all speakers combined.

Speaker	% Phrases Preceded by a Pause
1	82
2	72
3	78
4	90
5	67
6	60
All	73

Figure 5-29: The percent of major phrases preceded by a pause of greater than 0 ms for each speaker in the corpus and all speakers combined.

When the data for all six speakers is combined, 73% of all major phrases are preceded by a pause (Figure 5-29). This represents on upper-limit on the performance of a pause-based algorithm for detecting phrase beginnings for this data set. Using the unsupervised threshold, the phrase detection algorithm identified 87% of the major phrases preceded by a pause, with 97% precision (Figure 5-28). The recall becomes 64% (roughly 87% of 73%) when *all* major phrases are included—major phrases with and without a preceding pause (Figure 5-30). The precision of 97% remains the same. Fallout and error decrease because the total number of potential places to mark a phrase boundary increases. In attempting to recognize all major phrases, a phrase break can be placed after any word boundary. In all other evaluations of the pause-based phrase detection algorithms (Figure 5-28), only phrase boundaries following a pause were considered.

	Algorithm Predictions			
Hand Labels	Major Phrase	Other	Recall	64%
Major Phrase	1412	802	Precision	97
Other	40	7840	Fallout	5
			Error	8

Figure 5-30: Confusion matrix for hand labels of all major phrase breaks (with and without a preceding pause) versus predictions made by the unsupervised phrase detection algorithm. Hits = 1412, Misses = 802, False Alarms = 40, and Correct Rejections = 7840. The algorithm identifies 64% of all major phrases with 97% precision, 5% fallout, and 8% error.

In practice, identifying only those major phrases preceded by a pause proved very useful for the Audio Notebook user interface (Chapter 6). Phrase detection is used in the Audio Notebook to find playback starting points in an audio recording associated with handwritten notes (Section 6.1). A major phrase is not equivalent to a sentence. A major phrase can consist entirely of a cue word like “now” or a filled pause such as “um.” Below is an example from the lecture corpus. There are two major phrases with no pause preceding the phrase boundary (marked with a ‘*’):

> now * when we’re talking about children and news

For the Audio Notebook user interface it would be not be necessary, or even desirable, to break up a quote like the one above into two phrases. The problem is that the highest level phrase boundaries provided for in the ToBI labeling system (level 4, major phrases) are still to low level for the Audio Notebook. By considering only those major phrases preceded by a pause, some of these unwanted phrase boundaries are filtered out.

5.9.2.5 Final Phrase Detection Algorithm Used to Process Audio Notebook Recordings

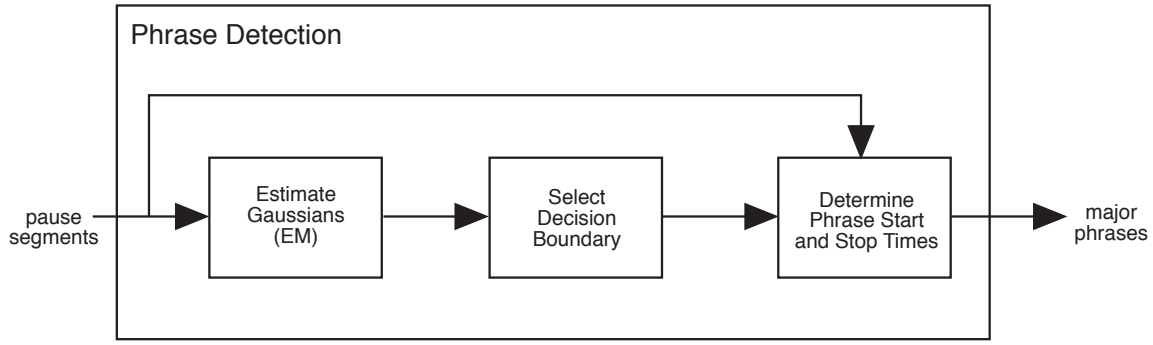


Figure 5-31: Diagrammatic representation of the final phrase detection algorithm.

Figure 5-31 shows the components of the final phrase detection algorithm. The output of the speech detection algorithm serves as input to the phrase detector. The speech detection algorithm outputs speech and pause segment starting and ending times for a recording. The phrase detection algorithm is composed of three major steps:

1. Estimate Gaussians (EM). First, a pause distribution is generated for the speech recording using the pause segments output by the speech detection algorithm. This pause distribution is a mixture of two classes—within-phrase and between-phrase pauses. The classes are not labeled so the problem is unsupervised. The EM algorithm is then used to iteratively estimate the parameters for the two Gaussian distributions (i.e., within-phrase and between-phrase pauses). The output of this step are six parameters: $P(\omega_1)$, μ_1, σ_1^2, $P(\omega_2)$, μ_2, and σ_2^2 (prior probability, mean, and variance for both distributions).

2. Select Decision Boundary. The next step is to select a decision boundary for separating the between-phrase pauses from those occurring within a phrase. The two Gaussian estimates (determined in step 1) are summed. Then, the local minima between the two means is used as the decision threshold. Pauses above this decision threshold are considered between-phrase pauses; those below the threshold are considered within-phrase pauses.

3. Determine Phrase Start and Stop Times. Using the pause segments output by the speech detection algorithm and the decision boundary determined in step 3, phrase start and stop times are generated. Each pause segment that is above the decision threshold indicates a phrase break. The beginning of the pause segment indicates the end of a phrase while the end of a pause segment indicates the start of a phrase. These phrase starting and ending times are then used in the Audio Notebook user interface (Section 6.1). The phrases also serve as units of analysis for segment beginning prediction (Section 5.10).

The phrase detection algorithm was used to process Audio Notebook recordings made during the field study. The Audio Notebook software maintains a separate recorded speech file for each page of the user's notes. The audio is therefore processed on a per page basis. A phrase-beginning pause threshold (i.e., decision boundary for separating the two classes of pauses) is determined for each page of audio. This threshold, in combination with the speech and pause segmentation, is used to determine phrase beginning and ending times (as described in step 3 above).

Figure 5-32 shows the results for four pages of audio taken from a Holography lecture given by Steve Benton. In these diagrams, the sum of the two Gaussian estimations (i.e., within-phrase and

between-phrase distributions) is overlaid on the raw pause histogram for each page of audio. The phrase-beginning pause thresholds range from 122–155 ms for these recordings. Chapter 6 describes how the phrase detection results are used in the Audio Notebook user interface.

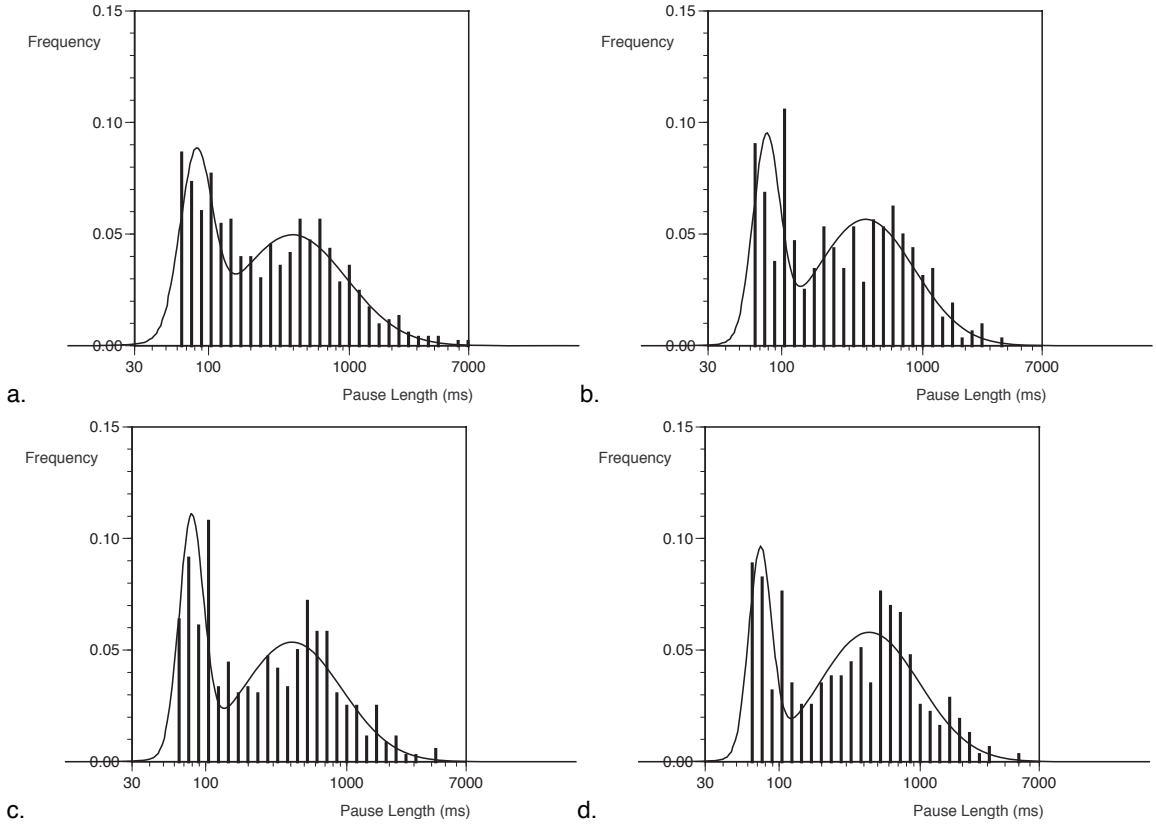


Figure 5-32: Gaussian estimations for four pages of audio taken from a Holography lecture. The estimations are overlaid on the raw pause histogram data. The decision thresholds are as follows: (a) 155 ms, (b) 137 ms, (c) 134 ms, and (d) 122 ms.[11] The amount of audio associated with each page of notes is as follows: (a) 12.2 min, (b) 7.7 min, (c) 8.8 min, (d) 7.6 min.

5.9.2.6 Comparison of Results to Related Work

The previous sections have described the development of a low-cost technique for detecting major phrases in spontaneous speech using pauses. This technique was developed for use in the Audio Notebook for two purposes: (1) to start playback from the nearest phrase beginning correlated with a user's handwritten notes; and (2) to use the major phrases identified as units of analysis for discourse segment beginning prediction. The Audio Notebook phrase detection algorithm was developed and tested using a small corpus of spontaneously spoken lectures. The algorithm achieved 87% recall with 97% precision for detecting major phrases preceded by a pause. The recall is 64% of all major phrases, including those that have no preceding pause (Section 5.9.2.4).

Researchers at Boston University have been working on algorithms for labeling all types of intonational features [Wightman and Ostendorf 1994, Ostendorf and Ross 1997]. Wightman and Ostendorf (hereafter, W&O) used pauses, phrase-final lengthening of phonemes, breaths,

[11] Note that the speech detection algorithm only outputs pauses of greater than or equal to 50 ms.

speaking rate changes, and F0 measures to automatically label major and minor phrase breaks [Wightman and Ostendorf 1994]. In W&O's research, the goal was to automatically label intonational features of speech that has already been transcribed. Their algorithm uses speech recognition to automatically align a transcript of the word sequence with the speech waveform. W&O tested their labeling algorithm using two different corpora of professionally read speech. The first corpus was composed of 70 sentences, each read by four professional radio news announcers. The second corpus of speech was composed of radio news stories read by one female announcer. For the corpus of read sentences, W&O reported 64% correct identification for major phrases with a 6% false detection rate. For the corpus of read news, W&O reported 78% correct identification of major phrases with a 7% false detection rate.

In order to compare W&O's results to the evaluation metrics used in this thesis, the confusion matrices presented in their paper were used to calculate recall, precision, fallout, and error[12] (as shown in Figure 5-33). There is a tradeoff between the number of major phrase breaks detected and the accuracy of the identification. W&O's algorithm achieves the highest recall (78%) when tested on a single radio news speaker, while the Audio Notebook algorithm achieves the highest precision (97%).

Algorithm	Corpus	%Recall	%Precision	%Fallout	%Error
W&O	Read Sentences	64	61	6	10
W&O	Read news	78	77	7	10
Audio Notebook	Spontaneous lectures	64	97	5	8

Figure 5-33: Comparison of evaluation metrics for detecting major phrases between Wightman and Ostendorf's algorithm and the one developed for the Audio Notebook.

There are many differences between the two studies, including the type of speech analyzed, the number of speakers, and the goals. The goal of W&O's algorithm was to automatically label intonational features for a speech recording that had already been transcribed. The goal for the Audio Notebook was to automatically detect major phrase breaks given an audio recording without transcription. In more recent work, Ostendorf and Ross are developing models for recognizing intonational features without a transcript. However, experimental results were not presented for automatically recognizing phrase boundaries [Ostendorf and Ross 1997].

Wightman and Ostendorf were also attempting to recognize many intonational features (e.g., pitch accents, all levels of break indices), while the focus of the Audio Notebook study was on identifying major phrase breaks alone. This thesis offers a simple method for segmenting speech into major phrases, and shows how this can be useful for speech interfaces and applications.

5.10 Segment Beginning Prediction

The next step was to analyze the acoustic correlates of the discourse segmentation marked by the subjects in the study. First, the segmentations must be coded and matched to determine the agreement among subjects. Segment break locations agreed upon by a majority of the coders are then analyzed acoustically. The intonational phrase serves as the unit of acoustic analysis. Acoustic cues across these phrasal units are compared against the agreed upon segment breaks to determine which features best predict the structure.

---

[12]Recall is equivalent to the percent correct identification given by Wightman and Ostendorf, and fallout is equivalent to false detection rate.

5.10.1 Segmentation Coding

Based on a coding scheme developed by Grosz and Hirschberg, each of the intonational phrases in a discourse are classified into one of the five discourse segment categories [Grosz and Hirschberg 1992]:

- Segment initial sister (SIS). The utterance beginning a new discourse segment that is introduced as the previous one is completed (e.g., utterance 4 in Figure 5-34).
- Segment initial embedded (SIE). The utterance beginning a new discourse segment that is a subcomponent of the previous one (e.g., utterance 12).
- Segment medial (SM). An utterance in the middle of a discourse segment (e.g., utterances 5–7).
- Segment medial pop (SMP). The first utterance continuing a discourse segment after a subsegment is completed (e.g., utterance 15).
- Segment final (SF). The last utterance in a discourse segment (e.g., utterance 3).

[1

[1.1

1. Well my name's Jim Smith
2. but whenever I write it it comes out James for some reason but
3. I don't care what you call me.

]1.1

[1.2

4. um I'm uh I'm currently at the Kalamazoo Computer Science Laboratory
5. I've been at Kalamazoo for a long time aside from about a nine month break
6. um I've been there and gotten my my bachelor's my master's
7. um something called an engineer's degree
8. which pretty much makes me a Ph.D. student er otherwise I'd have to leave.

]1.2

[1.3

9. um I work for a uh networking group
10. and I'm sort of a special person in the group because I'm not really what they do
11. except that I'm supposed to be driving their need for this um high-speed ne network

[1.3.1

12. um and I work for Professor Schmidt which I mention here because he came out
13. and and a lot of you got to hear what he had to say
14. and I might repeat a little bit of that

]1.3.1

15. My interests are in speech processing and recognition for uh multimedia applications
16. and again that from my group's perspective they're interested in me as someone who who gives a reason for their for their network.

]1.3

]1

Figure 5-34: A sample portion of a manual discourse segmentation.

These classifications are used to code each subject's discourse segmentation for each of the six speakers in the study. Grosz and Hirschberg combined the first two categories, SIS and SIE into a single category of segment beginning utterances (SBEG). They also considered SBEG plus SMP utterances as a broader class of discourse segment shifts. The SMP utterances exhibited similar acoustic characteristics to SIS and SIE utterances. Therefore, for this analysis, SBEGs are considered as a combination of all three categories: SIS, SIE, and SMP.

5.10.2 Matching Segmentations

Each of the six discourse samples in the corpus were segmented by five labelers. For each recording, the five segmentations were matched to determine *majority boundaries*—segment boundary locations agreed upon by at least four of the five segmenters.[13]

The primary goal is to determine acoustic correlates of discourse segment beginnings (SBEGs). The discourse segmentations were matched in two ways: (1) by comparing the five coders as a group, and (2) by comparing the five coders against the speaker's segmentation of his/her own talk. First, SBEG matches were determined by adding up the number of subjects marking a phrase as either SIS, SIE, or SMP. If at least four out of five subjects marked a phrase in one of these categories, the utterance is considered a segment beginning (Figure 5-35).

Speaker	#SBEGs agreed upon	Total # Phrases	%SBEGs/ Total
1	37	235	15.7%
2	49	561	8.7
3	22	390	5.6
4	21	199	10.6
5	46	576	7.9
6	30	247	12.1

Figure 5-35: The number of SBEG boundaries agreed upon by at least 4 out of the 5 naive coders.

In several studies of this kind, boundaries marked by naive coders have been compared against an expert [Carletta 1996]. In this thesis, subject's segmentations are also matched against the speaker's coding of his/her own lecture. Four out of five coders must agree with the speaker for a match (Figure 5-36). Notice that the percentage of SBEGs agreed upon is higher when the labelers are compared as a group than when they are matched against the speaker's segmentation.

Speaker	#SBEGs agreed upon	Total # Phrases	%SBEGs/ Total
1	32	235	13.6%
2	39	561	6.9
3	14	390	3.6
4	18	199	9.0
5	37	576	6.4
6	23	247	9.3

Figure 5-36: The number of SBEG boundaries agreed upon by at least 4 out of the 5 naive coders with the speaker's segmentation.

5.10.2.1 Agreement Statistics

The amount of agreement among segmenters was assessed using Siegel and Castellan's kappa statistic [Siegel and Castellan 1988]. Carletta proposes this statistic as a standard method of evaluating agreement for classification tasks [Carletta 1996]. This statistic corrects for chance expected agreement (Figure 5-37). For example, if there are two categories that are equally likely, and there are two coders, on average they will agree by chance 50% of the time.

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

Figure 5-37: P(A) is the proportion of times that the segmenters agree. P(E) is the proportion of agreement expected by chance. K = Kappa.

[13]Three out of five would constitute a majority but a stricter agreement criteria of four out of five labelers was used for this analysis.

First, the kappa statistic was calculated for the group of naive coders' agreement on SBEG boundaries (Figure 5-38). There are two categories—SBEG (SIS+SIE+SMP) and non-SBEG (all other phrases in the discourse).

Speaker	P(A)	P(E)	K
1	0.852	0.639	0.59
2	0.848	0.714	0.47
3	0.884	0.775	0.49
4	0.863	0.710	0.53
5	0.860	0.734	0.47
6	0.867	0.688	0.57

Figure 5-38: Agreement measures for the group of 5 naive coders.

These kappa values indicate weak inter-coder reliability. According to Carletta (based on [Krippendorff 1980]), values above 0.8 are considered the most reliable, and those between 0.67 and 0.8 are considered tentative. However, the kappa values do differ significantly from expected agreement due to chance ($p < .01$).

These agreement statistics are based on exact matches between the segmentations—the SBEGs must be marked at the exact same phrase for 4 out of the 5 coders. A problem arises with filled pauses (e.g., um) and cue phrases (e.g., now) that are alone in a phrase. Sometimes coders put the break at the cue or filled pause, sometimes after. If these breaks are considered matches, the agreement improves somewhat (K = 0.59, 0.50, 0.53. 0.57. 0.49, and 0.57 for the six speakers respectively). However, these values still indicate weak inter-coder reliability.

Next, the agreement was assessed using pairwise comparisons between each naive coder and the speaker's segmentation of his/her own talk (Figure 5-39). These values also indicate weak agreement between the speaker and the naive coders. Averaging across speakers gives an overall agreement measure for each coder. Notice that each coder exhibits a similar level of agreement with the speakers on average (mean kappa = 0.51 for three of the coders). Averaging across coders gives an overall agreement measure for each speaker. Speakers 4 and 6 have the highest average agreement, 0.60 and 0.59 respectively, while speaker 3 has the lowest (mean kappa = 0.44).

	Naive Coder						
Speaker	1	2	3	4	5	Mean	Std. Dev.
1	0.59	0.59	0.48	0.49	0.50	0.53	0.055
2	0.46	0.58	0.44	0.35	0.47	0.46	0.082
3	0.47	0.38	0.39	0.59	0.36	0.44	0.095
4	0.47	0.65	0.60	0.58	0.69	0.60	0.083
5	0.44	0.51	0.46	0.43	0.50	0.47	0.036
6	0.60	0.63	0.56	0.62	0.55	0.59	0.036
Mean	0.51	0.56	0.49	0.51	0.51	0.51	
Std. Dev.	0.071	0.099	0.078	0.11	0.11		

Figure 5-39: Paired kappa agreement measures—naive coder versus the speaker's coding of his/her own talk.

5.10.3 Acoustic Features

The primary unit of acoustic analysis is the intonational phrase. All acoustic features used in the analysis are listed in Figure 5-40. Many of these acoustic features were selected following Grosz and Hirschberg's study of acoustic correlates of discourse structure [Grosz and Hirschberg 1992].

Acoustic features are calculated for each phrase in a recording. Energy and fundamental frequency (F0) measures are obtained using Entropic speech analysis software. The Waves getf0 program outputs energy (RMS) and F0 measures (in Hz) for every 10 ms of speech in a recording. Pause durations were obtained from the hand labels of phrase beginning and ending times. For example, priorPause (the duration of the pause preceding each phrase in the discourse) is calculated by taking the starting time of one phrase and subtracting the ending time of the previous phrase. Subsequent pause (subeqPause) is the duration of the pause following each phrase in a recording. For fundamental frequency, several measures are calculated including F0 maximum and F0 average. In addition, speaking rate (in syllables per second) and phrase length are also calculated.

Two types of features are calculated for each phrase in a recording: (1) Absolute and (2) Relative. Absolute measures are used to compare the value of the acoustic features for segment beginning and non-segment beginning phrases. For example, the duration of pauses occurring prior to SBEG phrases is compared to the pauses preceding non-SBEG phrases.

Relative measures are obtained by comparing the acoustic features of one phrase to the previous or next phrase in a recording. Relative features look at the differences between adjacent phrases. For example, ΔF0Avg is calculated by taking the F0 average for one phrase and subtracting the F0 average for the previous phrase. Thus, relative measures indicate how much increase or decrease there is between acoustic features from one phrase to the next.

5.10.4 Correlating Features with Segment Beginnings

For each acoustic feature given in Figure 5-40, the difference between the means for segment beginnings versus non-segment beginnings was tested. Phrases labeled as segment beginnings by at least four out of the five discourse structure coders were used in this analysis. All other phrases were considered non-segment beginnings. This analysis was performed on a speaker dependent basis for each of the six speakers in the lecture corpus.

Figure 5-41 gives a list of acoustic features and indicates where significant differences were found between SBEGs and non-SBEG phrases at the $p < .01$ level of statistical significance. An up arrow (▲) in the table indicates that the mean value of a feature is significantly higher for SBEG phrases than for non-SBEGs. A down arrow (▼) indicates that a feature is significantly lower for SBEG phrases. For relative features, an up arrow indicates a significantly greater increase over the previous phrase for SBEGs than for non-SBEGs; a down arrow indicates a significantly greater decrease over the previous phrase for SBEG phrases. Features exhibiting a significant difference for all six speakers are written in bold in the table.

Given:

p = the number of 10 ms windows in a phrase

q = the number of 10 ms windows in the entire recording

k = the phrase number

$$F0Min = \min\left(f0_1, f0_2, \ldots, f0_p\right)$$

The following metrics were calculated for each phrase in a discourse:

Absolute Metrics:

$$priorPause = startTime_k - endTime_{k-1}$$

$$subseqPause = startTime_{k+1} - endTime_k$$

$$F0Max = \max\left(f0_1, f0_2, \ldots, f0_p\right)$$

$$F0RangeInterval = F0Max_k - F0Min_k$$

$$F0Avg = \frac{\sum_{i=1}^{p} f0_i}{p}$$

$$subseqF0Max = F0Max_{k+1}$$

$$subseqF0RangeInterval = F0RangeInterval_{k+1}$$

$$subseqF0Avg = F0Avg_{k+1}$$

$$rmsAvg = \frac{\sum_{i=1}^{p} rms_i}{p}$$

$$phraseLen = endTime_k - startTime_k \text{ (in ms)}$$

$$rate = \frac{\text{number of syllables in phrase}}{phraseLen/1000} \text{ (in syllables per second)}$$

Relative Metrics:

$$\Delta F0Max = F0Max_k - F0Max_{k-1}$$

$$\Delta F0RangeInterval = F0RangeInterval_k - F0RangeInterval_{k-1}$$

$$\Delta F0Avg = F0Avg_k - F0Avg_{k-1}$$

$$subseq\Delta F0Max = F0Max_{k+1} - F0Max_k$$

$$subseq\Delta F0Avg = F0Avg_{k+1} - F0Avg_k$$

$$ratioRmsAvg = \frac{rmsAvg_k}{rmsAvg_{k-1}}$$

$$\Delta RmsAvg = rmsAvg_k - rmsAvg_{k-1}$$

$$\Delta Rate = rate_k - rate_{k-1}$$

Figure 5-40: Definitions of all acoustic features used in the analysis. These features are analyzed to determine their correlation with SBEG phrases.

Features	Speaker 1	2	3	4	5	6
Absolute						
priorPause	▲	▲	▲	▲	▲	▲
subseqPause		▲*			▼	
F0Max	▲	▲			▲*	▲
F0Range	▲	▲			▲*	▲
F0RangeInterval	▲					▲
F0Avg	▲	▲			▲	▲
subseqF0Max					▲*	
subseqF0Range					▲*	
subseqF0RangeInterval						
subseqF0Avg					▲	
rmsAvg	▲	▲	▲		▲	▲
phraseLen			▼*		▼	
rate						
Relative						
ΔF0Max	▲				▲	▲
ΔF0Range	▲				▲	▲
ΔF0RangeInterval	▲				▼	
ΔF0Avg	▲	▲	▲	▲*	▲	▲
subseqΔF0Max	▼					▼
subseqΔF0Avg	▼	▼*				▼
ratioRmsAvg	▲	▲	▲	▲	▲	▲
ΔRmsAvg	▲	▲	▲	▲	▲	▲
Δrate	▼	▼	▼	▼*	▼*	

Figure 5-41: This table compares mean values of each feature for SBEG and non-SBEG phrases. An ▲ or ▼ arrow symbol in a table cell indicates a significant difference between the means at the $p < .01$ level. The ▲ symbol indicates that the mean is significantly higher for SBEGs than for non-SBEGs. A ▼ indicates that the mean is significantly lower for the SBEG phrases. A * indicates $p < .05$ level of statistical significance.

Three features, priorPause, ΔF0Avg, and ΔRmsAvg, showed significant differences between segment beginnings and other phrases for all six speakers. The pauses prior to segment beginning phrases were significantly longer than those preceding other phrases in the recordings. For segment beginnings, there was also a significantly greater increase over the previous phrase for F0 and RMS average. In other words, speakers paused longer prior to segment beginnings, and increased the average pitch and energy of their speech relative to the previous phrase when uttering segment beginning phrases.

In addition to these acoustic features, cue phrases (e.g., “now”, “so”, “well”, and “okay”) were also analyzed to determine their correlation with discourse segment beginnings (See Appendix B). Although potentially useful, cue phrases could not be used for a segment beginning prediction algorithm, since no lexical information is available for Audio Notebook recordings.

5.10.4.1 Comparison of Results to Related Work

The previous section compared acoustic features associated with segment beginnings with other phrases in the lectures for the six speakers in the corpus. These findings can be compared to Hirschberg and Nakatani's (hereafter, H&N) results for a single speaker from their Boston Directions corpus [Hirschberg and Nakatani 1996]. Boston Directions is a corpus of read and spontaneous monologues of speakers giving subway and walking directions of varying complexity. H&N studied several different aspects of discourse structure, including discourse segment beginnings. Only their results for segment beginnings are relevant to this thesis research.

Figure 5-42 gives a comparison of the results for a subset of the acoustic features studied. In the table, "N/A" indicates features that were not reported on in the H&N study[14]. The second column in the table shows the results for this thesis. The number in parentheses indicates for how many of the six speakers the difference between SBEGs and non-SBEGs was found to be significant. The third column in the table shows the results reported by H&N for one speaker in their directions corpus. The results shown here are H&N's findings for spontaneously spoken directions.[15]

Acoustic Cue	Thesis Results 6 Speakers	H & N 1996 1 Speaker
priorPause	longer (6)	longer
ΔF0Avg	increase (6)	N/A
ΔRmsAvg	increase (6)	N/A
RmsAvg	higher (5)	higher
Δrate	decrease (5)	no difference*
F0Avg	higher (4)	higher
F0Max	higher (4)	higher
ΔF0Max	increase (3)	increase
subseqPause	longer (1), shorter(1)	shorter

Figure 5-42: This table compares the results for this thesis study of six speakers with results reported by Hirschberg and Nakatani [1996] for a single speaker. The table gives a list of acoustic features where significant differences were found between SBEGs and other phrases in the recordings. *H&N found a significant difference for Δrate when comparing SBEGs to segment final phrases but not when comparing SBEGs to all other phrases in the discourse.

There are many similarities in the results for the two studies. Significant differences were found between SBEGs and other phrases for priorPause, RmsAvg, F0Max, and ΔF0Max in both studies. However, there are also several differences in the results. H&N found a significant difference between SBEGs and non-SBEGs for ΔF0Max. In contrast, for this thesis, significant differences were only found for half of the speakers in the lecture corpus (i.e., 3 out of the 6 speakers) for ΔF0Max. However, a significant difference was found for all six speakers in the lecture corpus for

[14]Note that Hirschberg and Nakatani's analysis uses minor phrases (i.e., intermediate phrases) as the primary unit of analysis, while this thesis uses major phrases (i.e., intonational phrases) as the primary unit of analysis.

[15]Hirschberg and Nakatani's results for segment beginnings were almost identical for read and spontaneously spoken directions.

ΔF0Avg (a feature not reported on by H&N). This result may indicate that ΔF0Avg is more reliable across speakers than ΔF0Max. However, this result may have been caused by differences in how F0 maximum was measured in the two studies. H&N manually labeled F0 maximum at the energy peak of the voiced portion of the nuclear-accented[16] syllable. For this thesis, F0 maximum was determined by automatically selecting the highest F0 value for a phrase. H&N compared manual labels of F0 maximum within the nuclear-accented syllable to a simple automatically-selected F0 peak. They found the manual labels of F0 maximum to be a "more robust" measure than a simple F0 peak. The goal for this thesis was to utilize features that could be automatically calculated by an algorithm. Automatically identifying pitch accents is a difficult problem, so the analysis of lectures did not assume knowledge of accented syllables. Therefore, when accent and syllable boundary information are unavailable, ΔF0Avg may be a more robust measure than a simple difference between F0 peaks.

Another difference between H&N's findings and the results for this thesis was for subsequent pause. For this thesis, subsequent pauses were found to be significantly longer for SBEGs than for non-SBEGs for one speaker, and significantly shorter for another speaker. For the other four speakers in the lecture corpus, no significant differences were found for subsequent pause. H&N found that pauses following SBEG phrases were significantly longer than other pauses in the discourse for the single speaker they analyzed.

Thus, these comparisons between SBEGs and other phrases in the lecture recordings were useful for determining potential features for distinguishing SBEGs from non-SBEGs for multiple speakers.

5.10.5 Speaker Independent Segment Beginning Prediction

Based on the previous analysis, the following features were selected for predicting segment beginning phrases: priorPause, ΔF0Avg, and ΔRmsAvg. These three features were selected because there were significant differences between the means for SBEGs and non-SBEGs for all speakers in the corpus (Figure 5-42). For the Audio Notebook, a speaker independent algorithm is needed since it is not practical to obtain hand-labeled training data for every speaker that could possibly be recorded.

Figure 5-43 shows the steps involved in predicting segment beginnings. The inputs to the first step, *Calculate Features*, are F0 and energy measures for every 10 ms of speech, and major phrase starting and ending times. Using this input, three acoustic features are calculated (priorPause, ΔF0Avg, and ΔRmsAvg) for every phrase in a recording.

[16]The nuclear accent is the last accent in an intermediate phrase.

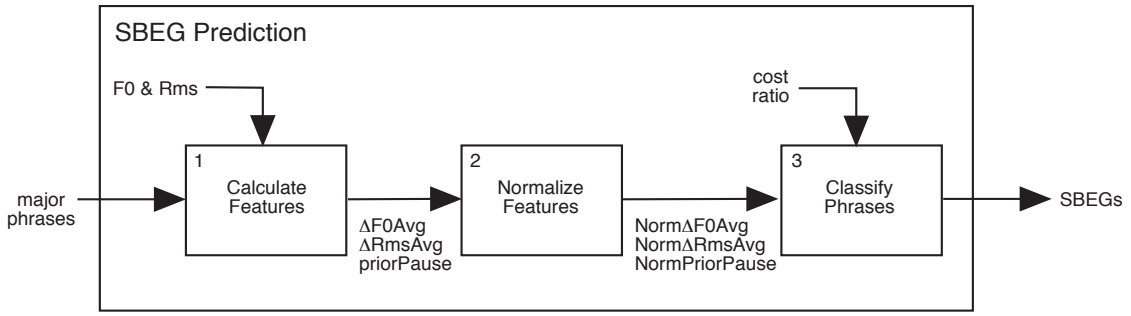


Figure 5-43: Diagrammatic representation of the segment beginning prediction algorithm. For each phrase in a recording, the algorithm outputs a predicted class: SBEG or non-SBEG.

The next two steps in the SBEG prediction algorithm, *Normalize Features* and *Classify Phrases* are discussed in detail in the sections that follow.

Note that prior to developing this SBEG prediction algorithm, a classification and regression tree (CART) analysis was used to investigate different combinations of features for detecting segment beginnings (see Appendix C). However, the results did not prove useful for the Audio Notebook. One limitation encountered was the inability to vary misclassification costs. Sections 5.10.5.3 and 6.2 discuss the importance of varying misclassification costs for application of the SBEG prediction in the Audio Notebook user interface.

5.10.5.1 Normalizing Features

In step 2 of the SBEG prediction algorithm, *Normalize Features*, the three acoustic features—priorPause, ΔF0Avg, and ΔRmsAvg—are normalized (Figure 5-43). This normalization compensates for differences in the features distributions for different speakers, which is necessary to train a speaker independent algorithm. Comparisons between the means for various acoustic features (Section 5.10.4) were performed for each speaker individually. However, in training a speaker independent algorithm, the data for multiple speakers was combined. The goal was to select normalizations that resulted in the most distinction between the two classes (SBEG and non-SBEG) when the data is combined across speakers.

Several different methods of normalizing each acoustic feature were compared using Fisher's criterion. Fisher's criterion provides a measure of the spread between two classes over the spread within two classes (Figure 5-44) [Therrien 1989]. In order to be able to distinguish between two classes given a particular feature, the within-class spread should be small relative to the between-class spread. In other words, the difference between the means of the features for the two classes should be as large as possible, and the variances for each class should be as small as possible. The higher the Fisher's criterion, the better the distinction between the two classes.

$$F = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

Figure 5-44: Fisher's criterion provides a measure of the spread between two classes (numerator) over the spread within two classes (denominator) [Therrien 1989, 103]. μ = mean, σ^2 = variance.

Fisher's criterion was first calculated for the unnormalized features: priorPause[17], ΔF0Avg, and ΔRmsAvg (Figure 5-45). This analysis is performed on the pooled data for all six speakers.

For priorPause, the percent cumulative frequency for a speaker's distribution (0–100%), was selected as one potential normalization. The cumulative frequency is calculated based on a histogram of all the pauses in a speaker's recording. For example, a pause length of 800 ms may have a cumulative frequency of 75% for a particular speaker's pause distribution. As shown in Figure 5-45, this normalization for priorPause (%cumFreqPriorPause), produced a higher Fisher's criterion (1.19) than the unnormalized prior pause length (0.73).

Next, ΔF0Avg was compared to ratioF0Avg, and ΔRmsAvg was compared to ratioRmsAvg (as defined in Figure 5-40). Recall that ΔF0Avg is calculated by taking the F0 average for one phrase and subtracting the F0 average for the previous phrase. An alternative is to take the F0 average for one phrase and *divide* by the F0 average for the previous phrase (ratioF0Avg). The Fisher's criterion for ΔF0Avg (0.55) was higher than for ratioF0Avg (0.51), and the Fisher's criterion for ΔRmsAvg (0.53) was higher than for ratioRmsAvg (0.25). Therefore, deltas were used instead of ratios for the final SBEG prediction algorithm.

	Feature	Fisher's Criterion	Fisher's Criterion Zero-Mean Unit Variance (Second Type of Normalization)
Unnormalized	priorPause	0.73	0.74
Normalized	%cumFreqPriorPause	1.19	1.22
Unnormalized	ΔF0Avg	0.55	0.64
Normalized	ratioF0Avg	0.51	0.57
Unnormalized	ΔRmsAvg	0.53	0.61
Normalized	ratioRmsAvg	0.25	0.42

Figure 5-45: Fisher's criterion for 3 pairs of features, unnormalized and normalized. In addition, a second level of normalization is performed using zero-mean unit variance (shown in the last column).

Next, for each of the six features listed in Figure 5-45, a second type of normalization was performed using zero-mean unit variance. This is also referred to as the normal score. For each feature value, the speaker's mean is subtracted, and the result is then divided by the speaker's standard deviation (Figure 5-46). This helps to normalize for differences in the variances (i.e., spread) for each speaker. The goal is to put the feature values into standard units so they will be more comparable between speakers. As shown in Figure 5-45, the zero-mean unit variance normalization results in an improved Fisher's criterion for all six features.

$$x_{\text{normal score}} = \frac{x - \mu}{\sigma}$$

Figure 5-46: Zero-mean unit variance or normal score is used to normalize the features across speakers. x = feature value, μ = mean, σ = standard deviation.

Based on the results of this analysis, the following three normalized features were selected for classifying phrases as SBEG versus non-SBEG:

---

[17]Note that the log of the pause length is used rather than the raw pause duration. The reasons for this log transformation were discussed in Section 5.9.1.

1. zero-mean unit variance of ΔF0Avg (NormΔF0Avg).
2. zero-mean unit variance of ΔRmsAvg (NormΔRmsAvg)
3. zero-mean unit variance of %cumFreqPriorPause (NormPriorPause).

These three normalized SBEG prediction features are calculated for every phrase in a recording.

5.10.5.2 Classifying Phrases

In step 3 of the SBEG prediction algorithm (Figure 5-43) the three normalized features were used to train a speaker independent classifier for distinguishing the two classes—SBEGs and non-SBEGs. This problem is supervised since the class labels are pre-determined from the discourse segmentation performed by the five labelers. This work was done in collaboration with another Media Lab doctoral student, Giri Iyengar, whose expertise is in the area of pattern classification.

The phrase classifier, implemented by Giri Iyengar, is composed of two steps: (1) feature extraction (and data reduction) and (2) using Gaussian classifiers to form a decision curve for discriminating the two classes.

5.10.5.2.1 Step 1 of Classifying Phrases: Feature Extraction

A two-class feature extraction algorithm was implemented (for more detail see [Therrien 1989, 78–80]). Using the three normalized acoustic features provided, (NormΔF0Avg, NormΔRmsAvg, and NormPriorPause), this technique selects the best plane along which to project the data. The best plane is the one that results in the most separation between the two classes, SBEG and non-SBEG. This feature extraction technique also reduces the number of dimensions (from three to two), similar to principle component analysis.

Figure 5-47 shows the projections for the unnormalized data on the left and the normalized data on the right. In the normalized case, there is less within-class scatter for the SBEG class than for the unnormalized data. In both cases, there is still a lot of overlap between the two classes. Figure 5-48 shows the projections for two individual speakers. For speaker 1 there is more separation between the two classes, and less spread within the classes than for speaker 2.

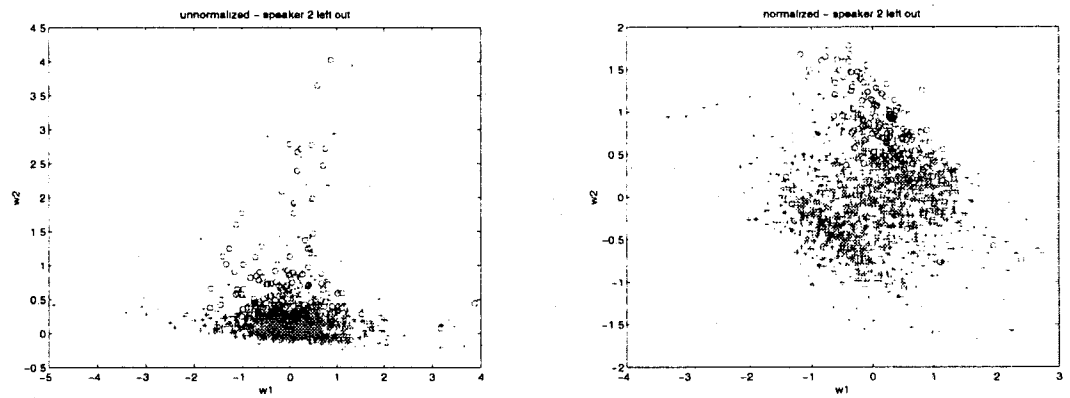


Figure 5-47: Unnormalized data on the left, normalized on the right. These projections were generated using the data for all speakers, leaving out speaker 2. The 'o' symbols are SBEGs (class 1) and the '+' symbols are non-SBEGs (class 2).

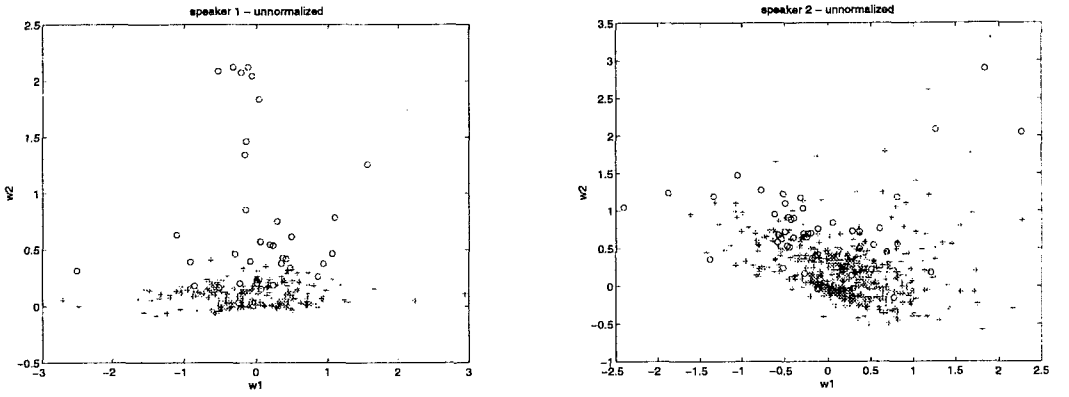


Figure 5-48: Projections for two individual speakers based on the unnormalized data; speaker 1 on the left, speaker 2 on the right. The 'o' symbols are SBEGs (class 1) and the '+' symbols are non-SBEGs (class 2). Note that for speaker 1, there is more separation between the two classes, and less spread within the classes than for speaker 2.

5.10.5.2.2 Step 2 of Classifying Phrases: Gaussian Estimation and Calculation of Decision Curve

In step 1 of the phrase classification, the three acoustic features (i.e., dimensions) were projected onto a plane, reducing the number of dimensions from three to two. In step 2 of phrase classification, the data in the projected plane is modeled as two dimensional Gaussian distributions representing the two classes (SBEG and non-SBEG)

Gaussian classifiers are then used to form a decision boundary. A quadratic decision curve is computed for splitting the data into two classes.[18] The decision curve may be an ellipse, hyperbola, or parabola (for more detail see [Therrien 1989, 95–98]). A decision rule for classifying phrases as SBEG or non-SBEG minimizes *Bayes risk*—using the likelihood of each class, prior probabilities, and misclassification costs (see Section 5.10.5.3). Figure 5-49 shows a decision curve overlaid on the projections for speaker 1.

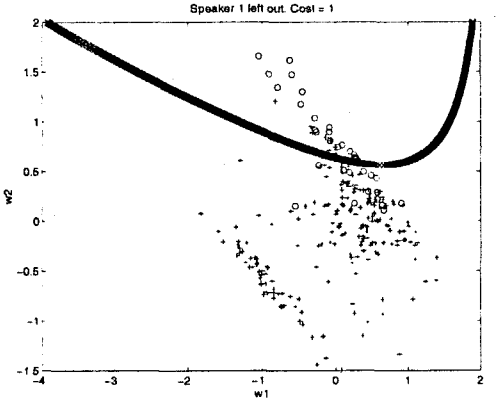


Figure 5-49: This shows the decision boundary overlaid on the projections for speaker 1 (normalized). The Gaussian classifiers have been trained using the other five speakers in the lecture corpus.

5.10.5.3 Varying Misclassification Costs

The decision boundary shown in Figure 5-49 assumes that the cost of misclassifying an SBEG is equal to the cost of misclassifying a non-SBEG. For a two-class decision rule, a cost can be

---

[18] The decision surface is quadratic when the class covariance matrices differ.

assigned to each of the four possible outcomes, where class 1 is an SBEG and class 2 is a non-SBEG:

C11 = Cost of predicting class 1 when it is class 1 (i.e., a hit)

C22 = Cost of predicting class 2 when it is class 2 (i.e., correct rejection)

C12 = Cost of predicting class 1 when it is class 2 (i.e., a false alarm)

C21 = Cost of predicting class 2 when it is class 1 (i.e., a miss)

The phrase classification step of the SBEG prediction algorithm takes a misclassification cost ratio as input (Figure 5-43). As shown in Figure 5-50, the cost ratio is defined as the cost of a miss (i.e., missing an SBEG) over the cost of a false alarm (i.e., falsely identifying a non-SBEG as an SBEG). For example, a cost ratio of 5 defines the cost of a miss as 5 times the cost of a false alarm. Alternatively, a cost ratio of 1/5 defines the cost of a false alarm as 5 times the cost of a miss. The higher the cost ratio, the more SBEGs selected, but the more potential for false alarms.

$$\text{Cost Ratio} = \frac{C_{21}}{C_{12}}$$

Figure 5-50: The misclassification cost ratio specifies the relative cost of the two types of errors, misses in the numerator and false alarms in the denominator. For example, a cost ratio of 5 defines the cost of a miss as 5 times the cost of a false alarm.

It is important to have the ability to vary the relative cost of the two types of errors. A fixed classification cost will not be appropriate for all tasks and all users. For certain tasks, more segment beginning predictions may be needed, even at the cost of more false alarms. In other cases, the user may desire fewer, but more accurate suggestions of segment beginnings. Allowing a variable classification cost ratio provides flexibility for the interface designer, and ultimately, the user (see Section 6.2).

Figure 5-51 shows two different decision boundaries for the same projections (also compare this to the decision boundary shown in Figure 5-49). In the plot on the left, a miss is defined as 5 times the cost of a false alarm (cost ratio = 5). In the plot on the right, a false alarm is defined as 5 times the cost of a miss (cost ratio = 1/5). The decision boundary on the left selects more SBEGs but at the cost of making more false alarms. The decision boundary on the right selects fewer SBEGs with more accuracy.

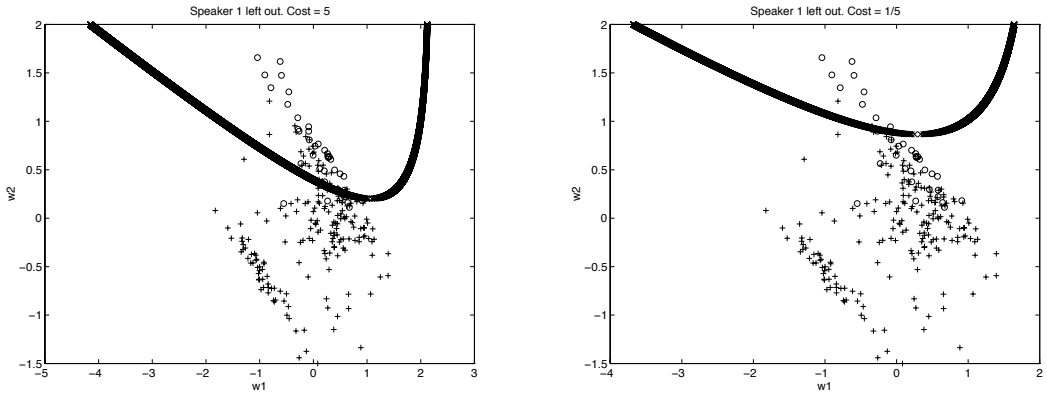


Figure 5-51: These figures show two different decision curves; the plot on the left uses a cost ratio of 5, and the plot on the right uses a cost ratio of 1/5.

The next section presents cross-validated testing results, and shows the impact of varying the misclassification cost ratio on the evaluation metrics.

5.10.5.4 Training and Testing

A leave-one-out cross-validated training and testing procedure was then performed. The classifiers were trained on five of the speakers and tested on a sixth. For this training, the labeled data for five of the speakers was projected onto the computed plane (Section 5.10.5.2) and the two classes were estimated as Gaussians. This procedure was repeated six times, each time training the classifiers using a different combination of the five speakers, and testing on the one left out.

The cross-validated evaluation metrics are given in Figure 5-52. The cross-validation was performed for both the unnormalized and normalized data so the results could be compared. In addition, the worst speaker (speaker 2) was left out to see how much this would improve the results. Speaker 2 had the lowest recall and precision of all the speakers for distinguishing SBEGs from non-SBEGs. As shown in Figure 5-48, there is a large amount of overlap between SBEGs and non-SBEGs for speaker 2 and the spread within each class is large.

	%Recall	%Precision	%Fallout	%Error
Normalized data	59	44	8	12
Unnormalized data	30	51	3	10
Normalized data without speaker 2	68	44	10	12

Figure 5-52: Cross-validated evaluation metrics for the SBEG classifier using a cost ratio of 1. Results are shown for classifying the phrases using the unnormalized features versus the normalized features. In addition, the last row shows the results when the worst speaker (speaker 2) is left out of the data set.

The normalized data resulted in almost twice the recall (59% versus 30%) of the unnormalized data, but with a lower precision (44% versus 51%). There is a tradeoff between recall and precision. However, while there is a 97% improvement in recall with normalization, the precision is only reduced by 14%. Leaving out the worst speaker (speaker 2) increased the recall from 59% to 68% while maintaining the same precision.

Figure 5-53 shows cross-validated evaluation metrics for three different misclassification cost ratios using the normalized data for all speakers. Again, there is a tradeoff between recall and precision. When the cost ratio is set to 5 (i.e., the cost of a miss is 5 times the cost of a false alarm) the recall is 81% correct identification of SBEGs but with only a 30% precision. As the cost ratio is decreased, the precision increases but the recall decreases.

Cost Ratio	%Recall	%Precision	%Fallout	%Error	%SBEGs Selected
5	81	30	21	21	27
1	59	44	8	12	13
1/5	37	53	4	10	7

Figure 5-53: Cross-validated evaluation metrics for the SBEG classifier for three different cost ratios using the normalized features. The last column (%SBEGs selected) gives the percent of phrases hypothesized as SBEG out of the total number of phrases.

Varying the cost ratio also causes the number of phrases identified as SBEGs to change. As shown in the last column of Figure 5-53, the higher the cost ratio, the more SBEGs selected out of the total number of phrases. The more phrases that are classified as SBEGs, the more potential for false alarms. A receiver operating characteristic (ROC) curve shows the tradeoff between the probability of correctly identifying a target (in this case, an SBEG) versus the probability of a

false alarm. Figure 5-54 shows an ROC curve for the SBEG prediction algorithm that was generated using 15 different cost ratios between 1/100 and 100. The higher the ROC curve is in the upper left-hand corner of the graph, the lower the class error probabilities (i.e., probability of a miss or false alarm).

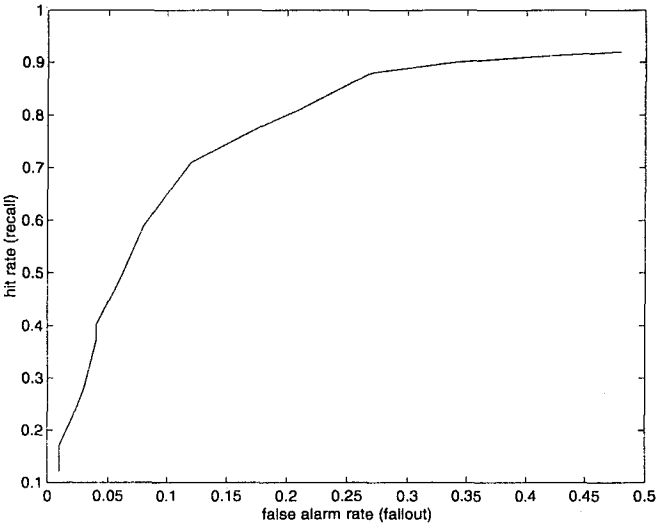


Figure 5-54: Receiver operating characteristic (ROC) curve for the SBEG prediction algorithm. An ROC curve shows the tradeoff between hit rate (recall) and false alarm rate (fallout). This curve was generated using 15 different cost ratios between 1/100 and 100.

SBEGs make up only 10% of the total number of phrases in the corpus. As shown in Figure 5-53, while the fallout is low (4%) given a cost ratio of 1/5, the precision is only 53%. The fallout is low because the number of false alarms is small in relation to the total number of non-SBEGs. Precision gives another measure of accuracy by looking at the number of correct identifications out of the total number of SBEGs hypothesized. Figure 5-55 shows the tradeoff between recall and precision for the same 15 cost ratios used to generate the ROC curve. This provides another way of looking at the tradeoffs between identification and accuracy.

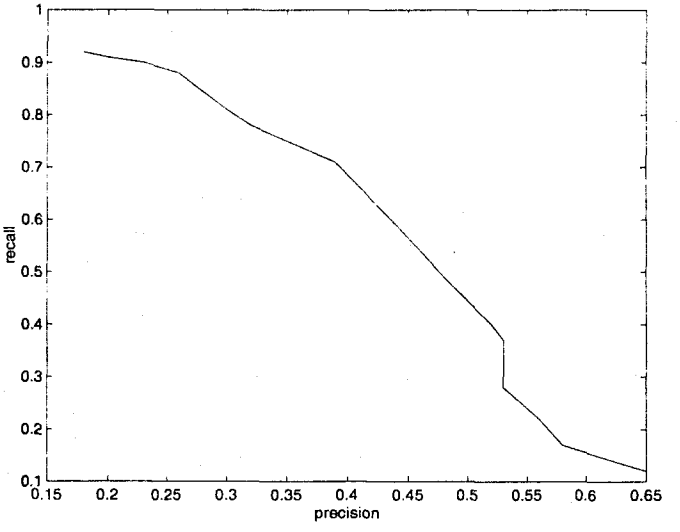


Figure 5-55: Tradeoff between recall and precision for the SBEG prediction algorithm for 15 different cost ratios between 1/100 and 100.

Depending on the user's task, usage style, and level of detail of their notes, more recall or more precision may be needed. The misclassification cost ratio can be manipulated to increase the recall or increase the precision depending on the user's needs. This is discussed in more detail in Section 6.2.

5.10.5.5 Comparison of Results to Related Work

The evaluation metrics for the SBEG prediction algorithm can be compared to Passonneau and Litman's (hereafter, P&L) results for identifying discourse segment boundaries [Passonneau and Litman 1997]. P&L proposed several algorithms for classifying discourse segment *boundaries*. According to P&L, a *boundary* can be placed between any two phrases in a discourse. Therefore, what P&L refer to as segment "boundaries" are similar to what are called segment beginnings in this thesis.

P&L studied a corpus of spontaneous narrative monologues called the Pear Stories. These narratives were originally collected and transcribed by Wallace Chafe [Chafe 1980]. P&L worked from Chafe's transcripts which included annotations of pause durations. P&L used several linguistic features alone and in combination to predict segment boundaries. These features included: referential noun phrases, cue phrases (phrases beginning with cue words such as "now", "so", "and "okay"), pause duration, and sentence-final-contour (indicated by a period or question mark in the transcript).

The evaluation metrics for what P&L refer to as their "best performing algorithm" generated using machine learning techniques are shown in Figure 5-56 [Passonneau and Litman 1997]. These results are comparable to the Audio Notebook SBEG prediction when the cost ratio is set to 1/5. P&L achieved better results when they tested their algorithm using a cross-validated testing a training method (43% recall, 63% precision) as opposed to a hold-out method[19]. P&L's proposed algorithm does not take into account misclassification costs. The Audio Notebook SBEG prediction algorithm allows misclassification costs to be varied so tradeoffs can be made between recall and precision as previously shown in Figure 5-53. Section 6.2 discusses how the SBEG prediction and ability to vary costs are used in the Audio Notebook user interface.

Algorithm	Cost Ratio	%Recall	%Precision	%Fallout	%Error
Audio Notebook	1/5	37	53	4	10
P&L	N/A	39	52	5	11

Figure 5-56: Comparison of results for the Audio Notebook SBEG prediction algorithm (using a cost ratio of 1/5) with one of P&L's algorithms proposed for identifying discourse segment boundaries.

Note that P&L's algorithm was proposed but not implemented. Their algorithm would require speech-to-text transcription, including transcription of cue phrases and referential noun phrases. The SBEG prediction algorithm developed for this thesis was implemented and used to automatically process Audio Notebook recordings, and does not require speech-to-text transcription.

5.10.5.6 Final Segment Beginning Classifier

The final segment beginning classifier was trained using all six speakers in the labeled corpus. The leave-one-out training and testing described in the Section 5.10.5.4 was used to estimate the

[19] A hold-out method of testing involves training an algorithm on one set of data, and testing it on an entirely new set of data that has been held out of the training set.

performance of the classifier and evaluate different cost ratios. The SBEG prediction algorithm is used to process Audio Notebook recordings which are not part of the training set; therefore, all six speakers are used to train the final classifier. This training produces parameters for feature extraction (i.e., projections w1 and w2) and parameters for the Gaussian estimations (prior probabilities, means, and covariance matrices). Inputs to the SBEG classifier are: the stored training parameters, feature vectors containing three normalized acoustic features for each phrase in a recording, and a misclassification cost ratio (Figure 5-50). Recall that the three normalized features are:

- NormΔF0Avg
- NormΔRmsAvg
- NormPriorPause.

The SBEG prediction algorithm is built on top of the speech detection and phrase beginning detection (Figure 5-57). First, a recording is processed by the speech detection algorithm to divide the audio into speech and pause segments. Next, the speech and pause segmentation is used as input to the phrase detection algorithm. The phrase detection algorithm classifies pause segments as within-phrase or between-phrase pauses. Finally, the phrases are used as units of analysis for the SBEG prediction algorithm.

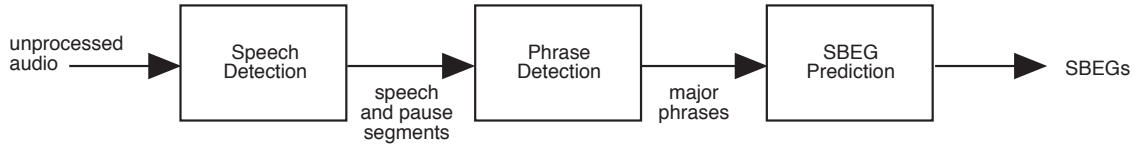


Figure 5-57: Diagrammatic representation of the different levels of acoustic processing, showing how the processes build upon one another.

6. Combining User and Acoustic Structure

The final Audio Notebook user interface design incorporates a combination of user activity and acoustic cues to structure speech recordings. The approach of this thesis was to first begin with the user's structuring of the audio recordings and then to determine how speech processing techniques could be used to augment this basic structure. In Chapter 3, an approach for user-structured audio is presented. The user's natural activity—writing and page turns in a paper notebook—are used to index the audio for later retrieval. In Chapter 4, an acoustic study was presented with three resulting techniques for acoustically structuring a speech recording: speech and pause segmentation, phrase detection, and discourse segment break prediction. This chapter describes how these acoustic structuring techniques are incorporated into the Audio Notebook user interface. Two of these techniques—phrase detection and segment prediction—are used directly by the Audio Notebook interface, while the speech detection is a component of the phrase detection algorithm.

6.1 Audio Snap-to-Grid using Phrase Detection

When a student or reporter takes notes during a lecture or interview, there is a delay between the talker's speech and the user's handwritten notes. The listener assimilates what was said before writing notes. During a lecture, a student may have to wait until the professor moves away from the blackboard before copying down the information. When Audio Notebook users select in their notes to begin playback, the system first determines the closest X-Y location in the stored pen data and the associated time point in the audio recording. Next, to account for the delay between listening and writing, the Audio Notebook uses a *listening-to-writing offset*—the system subtracts the listening-to-writing offset from the playback starting time. During the field study, this listening-to-writing offset could also be personalized for each user. The listening-to-writing offset was stored in a profile for each user along with the user's preferred listening speed.

A problem with using a fixed offset is that playback may begin toward the middle or end of a phrase. When backing up by a fixed amount in the recording, there is no guarantee of finding a coherent starting point in the speech. This makes it difficult for the listener to follow the information presented [Arons 1994a] (see Chapter 4). When reporter and editor Jack Driscoll used the Audio Notebook, he commented that when he selected in his notes, playback often began in the middle of a quote. To overcome this problem, after selecting on the page, he used the audio scrollbar to adjust the starting point to the beginning of the quote. While the audio scrollbar allows a user to fine-tune the playback starting point, it would become tedious if users had to adjust the starting point each time a selection was made.

In graphics programs, users can turn on a grid, so when they are drawing or selecting an object, the mouse "snaps" to the nearest grid point. Given knowledge of major phrase break locations, the Audio Notebook can "snap back" to the nearest phrase beginning when a selection is made—I refer to this as snap-to-grid for audio. The phrase detection algorithm described in Chapter 5 is used to process Audio Notebook recordings. For each recording, the system predicts phrase starting and ending times. Now when users select in their notes to begin playback, the system first

backs up by the user-specific listening-to-writing offset, and then snaps back to the nearest phrase beginning.

Figure 6-1 shows some examples of the correlation between Jack Driscoll's notes and the audio with and without phrase detection. The phrase detection improved the correlation in these examples and many others. Notice that without phrase detection, the beginning of a word was often clipped. In addition, the playback starting point selection was hit or miss without phrase detection; sometimes playback would start near the beginning of a phrase, sometimes in the middle, sometimes toward the end.

Writing Selected in Notebook	Playback without Phrase Detection	Playback with Phrase Detection
Speech as interface	"handheld devices"	"the idea was to look at speech as an interface to handheld devices"
faster ~ fedex	"-an the federal express, the guy in the federal express commercials."	"faster than the federal express, the guy in the federal express commercials"
take sometimes hours	"-imes hours if I really wanted to find the right spot"	"and it was just extremely, it would take sometimes hours if I really wanted to find the right spot"
We brainstormed	"–stormed about it a little bit"	"and we brainstormed about it a little bit"

Figure 6-1: Correlation between notes and audio with and without phrase detection. These examples are taken from a page of Jack Driscoll's notes (Figure 4-16). A leading hyphen indicates that the word was cut off at the beginning.

Note that a spontaneously spoken phrase is not the same as a complete sentence. In the second example in Figure 6-1, without phrase detection, the speech is cut off at the beginning. With phrase detection, the beginning of the phrase is not cutoff; playback begins "faster than the federal express, the guy in the federal express commercials." However, the talker's complete thought was "and the professor spoke… faster than the federal express, the guy in the federal express commercials." An 800 ms pause after the word "spoke" separates this thought into two phrases. Therefore, although the system snaps back to the nearest phrase break, it will not necessarily correspond to a complete thought. In these cases, the user can back up further in the recording using the audio scrollbar. The next section describes how topic suggestions guide the user by providing additional navigational landmarks.

6.2 Topic Suggestions using Segment Beginning Prediction

In addition to phrase detection, discourse segment beginning prediction was integrated into the Audio Notebook user interface. Each prediction represents a potential topic change location. Predictions are displayed along the audio scrollbar as suggestions of places to navigate in the recording. These suggestions provide navigational landmarks for the user.

6.2.1 Early Design Prototype

Figure 6-2 shows an early design prototype exploring the use of topic suggestions for the Audio Notebook interface. This design is based on an early vision of the audio scrollbar. This design prototype, implemented in Macromind Director, is a simple visual and audio animation. In this design, there are tick marks displayed along the audio scrollbar representing the topic suggestions. In the animation, the hand moves down the scrollbar, selecting on several of the

suggestions. When the hand stops on one of the suggestions, a topic beginning phrase starts to play.

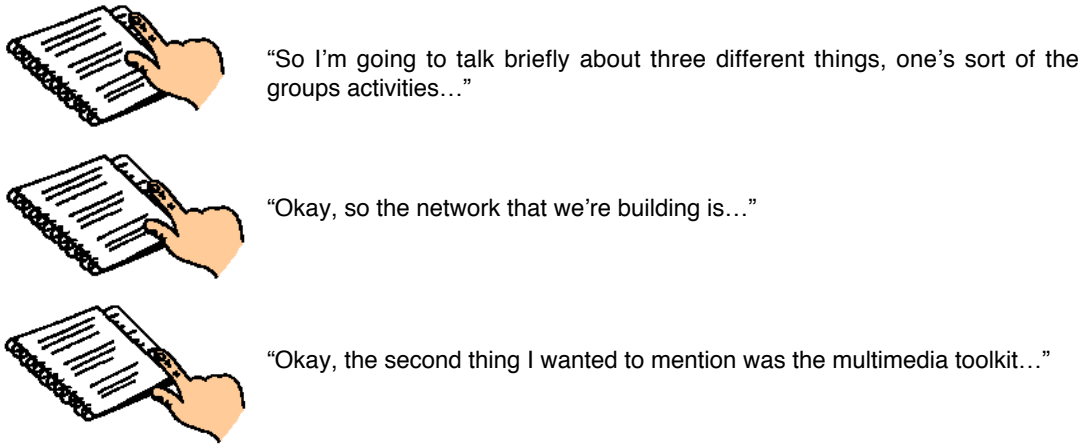


Figure 6-2: This visual and audio animation was an early design prototype exploring the use of topic suggestions. This design was rapidly prototyped in Macromind Director before the Audio Notebook and speech processing algorithms were developed.

6.2.2 Integration with Version 2 Audio Notebook

The audio scrollbar for the version 2 Audio Notebook was designed to be both a control and a display (Section 3.5.4). The scrollbar is composed of 80 multi-color LEDs; each LED can be one of three colors—red, orange, or green. Prior to the integration of the topic suggestions, the scrollbar was only used to display the audio cursor. The audio cursor, displayed using a green LED, indicates the user's current position in the timeline for a page (Figure 6-3).

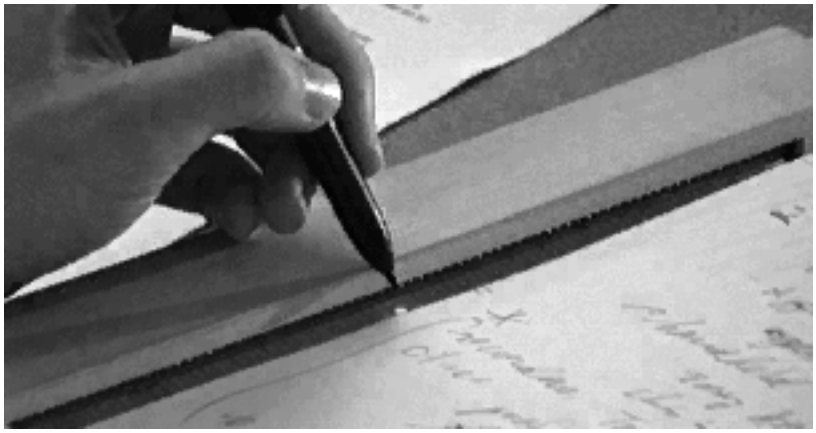


Figure 6-3: The audio cursor is displayed by lighting up a green LED along the audio scrollbar. As the audio plays, the cursor moves down the scrollbar.

With the addition of segment prediction, topic suggestions are now displayed in red along the audio scrollbar (Figure 6-4). This kind of structural information provides navigational landmarks in the timeline for more intelligent navigational control. Rather than blindly jumping to the beginning, middle, or end of the timeline, the user can now jump from one topic suggestion to another. When the user selects on one of the red LEDs, the audio begins to play from the start of the topic beginning phrase (Figure 6-5). When the green audio cursor moves over a suggestion, the LED turns from red to orange.

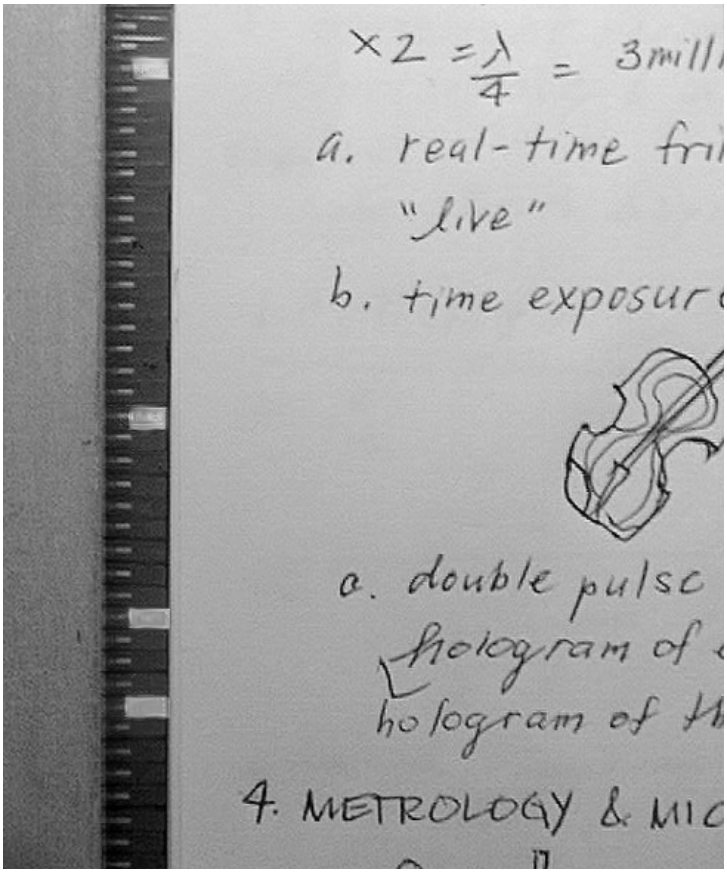


Figure 6-4: Topic suggestions are displayed along the audio scrollbar using red LEDs.

There are many important uses of the topic suggestions. First, topic suggestions index the audio when there is a lack of user activity. Figure 6-5 shows some suggestions for one page of the holography student's notebook. In this example, a large portion of the audio is not indexed because the user did not take notes. Perhaps the user was listening intently to the lecturer, did not believe the information to be important at the time, or had even fallen asleep. The topic suggestions provide structure where none was generated by the user's activity. This is important because the Audio Notebook should free the listener to devote more attention to the talker and not force the user to take verbatim notes.

Secondly, topic suggestions further enhance the correlation between the user's notes and the audio recording. Just as the system "snaps back" to the nearest phrase beginning when the user makes a selection in the notes, the user can also back up to the nearest topic suggestion. In the example shown in Figure 6-6, the user has marked down a topic heading in the notes, but it does not correspond exactly to the beginning of the topic in the audio recording. The user can select on the heading in the notes and then back up to the nearest suggested topic beginning. In this case, when the user selects on the heading titled *4. Metrology and Microscopy*, the audio begins "uh really neat that's called phase conjugation…"; but when the previous topic suggestion is selected, playback begins (as written in the notes) "Okay number 4 is going to be Metrology and Microscopy, and uh as an example of this let me preview something that's going to be uh really neat that's called phase conjugation…"

Figure 6-5: Topic suggestions displayed for one page of the holography student's notebook. The user is selecting on the first suggestion.

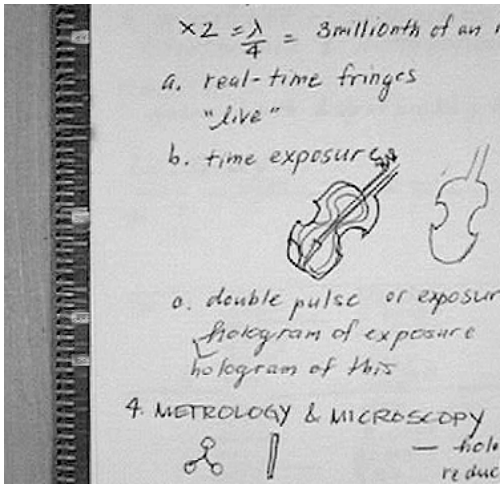


Figure 6-6: Topic suggestions displayed for another page of the holography student's notebook. In this case, a heading written in the notes (4. Metrology and Microscopy) does not correspond exactly with the topic phrase in the audio recording. The user can back up to the nearest topic beginning, as indicated by the LEDs, to find the audio associated with the heading

Topic suggestions also provide a way of quickly skimming through a recording. In the field study, student 1 often used the scrollbar to skim through a page of audio (Section 4.3.4). During the field study, the scrollbar did not display any navigational landmarks (other than the audio cursor) so the user's skimming was random. She selected on every few LEDs in the scrollbar, skipping over several each time, without knowledge of the content. The topic suggestions displayed along the scrollbar now provide a guided way of skimming more efficiently through the recording, allowing the user to jump from one topic suggestion to the next.

6.2.3 Adjusting the Number of Suggestions

The segment predictor will make errors, sometimes missing a topic beginning, other times predicting a topic change where none exists. As described in Section 5.10.5.3, the segment prediction algorithm allows the system to vary misclassification costs. The algorithm uses a cost ratio (Figure 5-50) in computing its decisions. The higher the cost of a false alarm (i.e., predicting a segment beginning where none exists), the fewer topic suggestions predicted. The higher the cost of missing a segment beginning, the more topics predicted. Tradeoffs must be made between identification and accuracy (i.e., recall and precision)—the more topics predicted, the more potential for false alarms. Interaction strategies must be adopted to allow the user to take advantage of algorithm's predictions while coping with potential errors.

The Audio Notebook allows the user to trade off between identification and accuracy. Rather than having the user control the cost ratio directly, the user is given control over the number of suggestions. A script is used to run the algorithm iteratively with different cost ratios to obtain different numbers of suggestions. The script iterates the cost ratio, with targets of 5, 10, and 15 suggestions. The number of suggestions is dependent on the data, so the exact number will vary, falling into one of the following ranges: 5–9 (min), 10–14 (medium), and 15–20 (max). The user can select between these three levels of suggestions using button controls (Figure 6-7). The fewer the number of suggestions, the higher the accuracy.

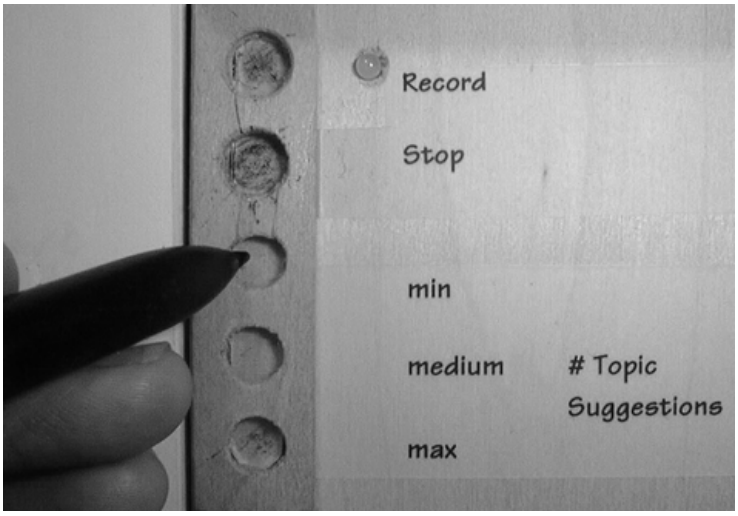


Figure 6-7: The user can select between three levels of topic suggestions; min (5–9 suggestions), medium (10–14), and max (15–20).

Figure 6-5 showed the audio scrollbar display of topic suggestions for a page from a holography student's notebook. In this photo, the number of suggestions is set to the minimum number. These five topic suggestions begin as follows:

1. "Now there are lots– we're going to work our way through the history of this uh definition number one…"
2. "So this is the stuff we're going to be talking about…"
3. "But it turns out that in China for example they're making counterfeit holograms…"
4. "So there are two issues in security…"

5. "Okay okay so that's finishing up chapter one…"

Each of these topic suggestions begin with one of the following cue words—"Now", "So", "But", or "Okay." A cue word at the start of a phrase may signal a topic change. Note that the algorithm selected these phrases using only acoustic analysis, without knowledge of the lexical content of the audio recording.

If the number of suggestions are then increased to the next level (medium), the following additional five topic suggestions are added to the audio scrollbar display:

6. "popular definition– how do we go– but you know we want to go from here to here one of these days and that's what we're going to be talking about…"
7. "And since then we've been working on making holograms that are bigger…"
8. "Certainly we all go to the same conferences but making pretty pictures with holograms is sometimes looked down upon by the more academic holographers…"
9. "Did I miss any important applications anybody's heard of…"
10. "Uh well the uh security, there are two aspects…"

As the number of suggestions are increased, more topic boundaries are located, but the number of false alarms increases as well. Suggestion number 6 appears to be a false alarm, although this is not conclusive because the Audio Notebook recordings have not been manually segmented. In summarizing two definitions of Holography, the professor said "Okay, so this is the first definition and this is the <pause> popular definition…" Topic suggestion number 6 could be improved if it began with "Okay, so this is the first definition…" rather than with "popular definition."

When the number of suggestions is set to the minimum amount, topic beginnings will be missed. In the example shown in Figure 6-6, the topic beginning "Okay number 4 is going to be Metrology and Microscopy," does not appear at the minimum suggestion level. This topic suggestion gets added when the user selects a "medium" number of suggestions. Depending on the style of usage, task, and level of detail of the user's notes, users may want more or fewer suggestions. When the user has not taken notes, or a note does not correlate properly with the audio, he/she may want more suggestions, even at the cost of more false alarms.

6.3 Correlating Notes with Audio—A Multi-Part Approach

A contribution of this thesis is a multi-part approach for correlating a user's handwritten notes with an audio recording. While several researchers have adopted the strategy of indexing time-based media (i.e., audio and video) with user notetaking activity, very little work has been done to address the problem of synchronization. People do not write notes in exact synchronization with the talker. Notes are taken sometime after the information is presented. This thesis uses a combination of user activity, acoustically-derived structure, and user-system interaction to find the starting point in an audio recording associated with a handwritten note.

The strategies used in this thesis for correlating notes and audio are summarized below. Each of these techniques were incorporated into the Audio Notebook interface over several design iterations. The strategies build upon one another, each one further improving the correlation

between notes and audio. Strategies 1 and 3 are automated, while 2, 4, and 5 involve interaction with the user.

1. System-defined listening-to-writing offset (Section 3.3.3). When the user selects on a handwritten annotation to hear the related audio, the system backs up by a constant amount. Previous research in this area only used this approach for dealing with the problem of the delay time between listening and writing.

2. User-personalized listening-to-writing offset (Section 4.7.2). During the field study, the listening-to-writing offset was customized for different users. The time delay between listening and writing differed depending on the user's notetaking style and task. For example, there was more delay time between listening and writing for the students than for the reporters in the study. Students had to copy notes off the blackboard when listening to the professor, while reporters could focus their attention on the interviewee.

3. Snap-to-grid audio using phrase detection (Section 6.1). The system subtracts the listening-to-writing offset (default or user-personalized) from the time a note was originally written. When backing up by this fixed amount, playback may begin in the middle or toward the end of a phrase. To address this problem, the system first subtracts the listening-to-writing offset and then snaps back to the nearest phrase beginning.

4. Backing up to the nearest topic suggestion using segment prediction (Section 6.2). Even after backing up to the nearest phrase beginning associated with a handwritten annotation, playback may still begin in the middle of a topic. The user may need more context to understand the information. In these cases, the user can then back up to the nearest topic suggestion.

5. Fine-tuning with the audio scrollbar (Section 3.5.4). The Audio Notebook provides the user with the "final say" over the correlation between the notes and audio. The user can fine-tune the playback starting point, backing up to get more context, or jumping ahead to skip over topic introductory material. The goal is to for the system to make its "best guess" at the audio starting point associated with a handwritten annotation, while allowing the listener to make adjustments when necessary or desired.

7. Conclusion

This chapter concludes the thesis by summarizing a number of contributions of the research, and presenting some areas for future work.

7.1 Contributions

This thesis makes contributions in several areas of research, including: audio interaction design, the study of acoustic cues to discourse structure, and acoustic processing techniques for structuring speech recordings. The following is a summary of contributions of this thesis research:

- ***User*-structured audio with the Audio Notebook**

  The first major phase of research focused on *user* structuring of audio recordings. This included the design and development of the Audio Notebook—a novel device that combines the familiar interface of a paper notebook with the advantages of an audio recording.

- **Paper and pen interaction techniques for navigation in the audio domain**

  Several user interaction techniques were developed for navigating through an audio recording using paper and pen. These interaction techniques can be used in combination with one another, allowing users to hone in on desired information. Users can leaf through their notebooks to roughly locate an area of interest, then use spatial navigation to pinpoint a particular topic, and lastly use the audio scrollbar to fine-tune the exact location in the recording.

- **Longitudinal study of interaction with the Audio Notebook**

  A five-month field study showed that the Audio Notebook interaction techniques supported a range of usage styles, from detailed review of audio recordings to high speed skimming. This study provided knowledge of how several students and reporters used audio recordings when they were more accessible and manageable. The longitudinal study also pointed out areas where the user's activity alone did not provide enough structure for the listener. This lead to the second major phrase of research—*acoustic* structuring of speech recordings.

- ***Acoustically*-structured speech recordings**

  Discourse theory and methodology provided a framework for developing *acoustic* structuring techniques that were then integrated into the Audio Notebook. Two processing techniques were developed based on an acoustic study of discourse structure for lectures: detection of major phase and discourse segment beginnings.

- **Audio snap-to-grid and topic suggestions with varying costs**

  The acoustic structuring techniques were then integrated into the Audio Notebook, creating new ways of interacting with the audio recordings—audio snap-to-grid and topic suggestions. Using phrase detection, the Audio Notebook "snaps" back to the nearest phrase beginning

when users make selections in their notes. Topic suggestions are displayed along the audio scrollbar, providing navigational landmarks for the listener.

- **Combination of user activity and acoustic cues for structuring speech recordings**

  Finally, the combination of user activity and acoustic structuring techniques is very powerful. The two techniques complement each other, allowing listeners to quickly and easily access portions of interest in an audio recording.

7.2 Future Research

The Audio Notebook provides a framework for interacting with structured speech recordings. Given this framework, the Audio Notebook can be extended in a variety of ways depending upon the information available (e.g., user's activity, speaker's activity, audio, video, speech transcript).

In Section 1.2.3 in the introduction to this thesis, a taxonomy of cues for structuring speech recordings was presented. Cues for indexing and structuring audio recordings were separated into three categories: user activity, linguistic cues, and external information. Further exploration could be done in each of these areas.

First, user activity can be further exploited to implicitly index an audio recording. Features of digital ink could be explored in a similar manner to digital audio. For example, just as pauses in speech were studied in this thesis, pauses in writing can also be analyzed. The amount of writing activity may be important, similar to energy measures for speech. These "prosodic" features of digital ink could be combined with acoustic cues for automatically identifying structural information.

In the second category of cues, linguistic information, the focus for this thesis was on acoustic (as opposed to lexical) cues to discourse structure. Cue phrases (e.g., "okay", "now") were studied but could not be used in the speech segmentation algorithms without speech recognition or word spotting technology. If word spotting technology were available, techniques used for information retrieval in text documents could be applied to the speech recordings [Kazman et al. 1996]. For example, researchers have used similarity measurements based on word repetition and thesaurus information to find discourse boundaries in text [Morris and Hirst 1991, Hearst 1993].

The third category of cues proposed in the taxonomy was referred to as "external information." This includes the activity of the lecturer (as opposed to the listener/notetaker). If an electronic whiteboard is used, the speaker's writing activity serves as an additional index into the audio recordings. If a video recording is used in addition to the audio, then the speaker's gestures can be analyzed in a similar manner to the speaker's prosody.

In combining structural information from multiple sources (e.g., indices from the user's own handwriting, indices from the speaker's handwriting, videotaped information), care must be taken not to overload the user. In designing the Audio Notebook, the seamless integration of multiple sources of structural information supports a variety of strategies and usage styles for navigating through the audio recordings.

7.3 Summary

This thesis has presented a new approach for rapid navigation and skimming in the audio domain. This approach combines user activity and acoustic cues for structuring audio recordings. This structure makes the recordings more accessible and manageable than they have been traditionally, so users can quickly and easily locate portions of interest. The Audio Notebook provided reassurance that key information would be available for later review, yet did not interfere with regular notetaking activity or interactions with others. Rather than replacing real-world objects like paper and pen, we can successfully augment them, combining the advantages of the physical world with the capabilities of digital technology.

Acknowledgments

First I want to thank my advisor, Chris Schmandt, who has taught me a great deal about speech interface technology and design over the past seven years. Chris took my application out of a pile and gave me the opportunity to study at MIT. One of the things that's really nice about having Chris as an advisor is that his door is always open, and he is available to talk without a formal appointment. Our meetings often took place outside his office, usually walking around the Charles. Chris has provided me with many opportunities to meet people in the field, and to gain work experiences outside the lab. He also gave me the opportunity to work with him on an NSF grant proposal, which provided support for this thesis research. I am grateful to Chris for pushing me to get my thesis proposal done quickly! He has given me a lot of freedom in carrying out this thesis research, as well as his support and confidence. I especially want to thank Chris for always believing in me, and for working with me to develop our relationship.

Barbara Grosz not only taught me about discourse theory and computation, but perhaps more importantly, about academic research. She invited me to spend time at Harvard, and treated me like one of her own students. Barbara is a great mentor because she encourages you by asking the right questions and guiding you, but will not tell you exactly what to do. I really enjoyed discussing problems with her and thank her for spending time with me.

When I first read about Durlach's theory of binaural masking level differences, I was really nervous about my oral exams! Nat has always provided his total support and encouragement. Thanks for taking the time to be a part of my committee given your hectic schedule.

Barry Arons provided support and encouragement at every step of the way, more than can be expressed in this simple acknowledgment. Barry and I discussed the idea for a paper-based audio notebook when we were brainstorming about ideas for my master's thesis in 1991. I talked about the idea a lot, but didn't have the opportunity to work on it. After hearing about it so much, Barry finally said "just do it already!" Barry and I spent a lot of time together discussing interface design issues. When I needed someone to build the hardware for the second Audio Notebook prototype, Barry (finally) volunteered. No one else could have done a better job or been better to work with. Barry read all of my thesis drafts and provided detailed comments.

Giri Iyengar collaborated with me on the segment beginning prediction algorithm. Giri also helped me to put together graphs for the thesis and reviewed several drafts written about the SBEG algorithm. Working with Giri made the late stages of thesis work a lot more fun.

Jim Glass provided much needed advice about speech analysis and speech processing. He also reviewed a section of this thesis.

Fred Martin provided advice and assistance during the development of the second Audio Notebook prototype.

Thanks especially to everyone at Interval Research Corporation. My summer internship and fellowship at Interval were invaluable experiences. Thanks to David Liddle, David Reed, Debby Hindus, and Andrew Singer for making this possible. Thanks to Wayne Burdick for working with me during my fellowship and implementing the first Audio Notebook hardware with help from interns Jacob Tuft and Tai Mai. Debby Hindus provided a constant source of support and encouragement along the way. Sue Faulkner and Bud Lassiter helped me to videotape the first Audio Notebook user study and edit a summary tape. Thanks also to Glenn Edens, Jonathan Cohen, Michele Covell, Lee Felsenstein, John Hughes, Bonnie Johnson, Brenda Laurel, Roger Meike, Daniel Shurman, Malcolm Slaney, Hank Strub, Rob Tow, Bill Verplank, and Sean White.

The discourse analysis and processing components of this thesis were inspired by Julia Hirschberg and Christine Nakatani's research with Barbara Grosz on discourse and prosody. Their work provided a methodological foundation for my thesis research. Julia was also my AT&T mentor during my two years as an AT&T Media Lab Fellow. During this time we had

many discussions about my research, and Julia was always available by email to answer my numerous questions. Julia also helped me with ToBI labeling, and checked over my work. I've learned a great deal from Julia and appreciate all the time she spent with me. I also want to thank Christine Nakatani for everything she taught me about discourse and for her support and friendship. Christine invited me to participate in a small reading group at Harvard which helped me to learn more about discourse and prosody.

Hiroshi Ishii, Justine Cassell, Walter Bender, and Mitch Resnick provided support and encouragement along the way. Hiroshi went out of his way to share resources, and to invite discussion about my work and the work in his group. Justine advised me on discourse analysis and thesis writing. Walter helped to give me a different perspective on my work from the standpoint of news in the future, and gave me opportunities to present this research.

One of the best parts about working on the Audio Notebook was having the opportunity to work with Jack Driscoll. Jack is a constant source of ideas and encouragement. He also reviewed portions of this thesis and provided *editorial* comments.

Special thanks to Gwelleh Hsu, Kathi Blocher, Daniel Dreilinger, Joey Berzowski, Daniel Gruhl, Alan Wexelblat, Michelle McDonald, Michele Evard, Susie Wee, and Kevin Brooks.

Steve Benton and Mike Bove let me record their lectures for an entire semester and have students in their classes use the Audio Notebook.

I especially want to thank Don Norris for using the Audio Notebook, and taking the time to write down his thoughts about his experiences.

John Underkoffler helped me to create a video of the Audio Notebook. He spent many hours with me filming and editing and I'm indebted to him for this.

Giovanni Flammia developed the Nota Bene program used for discourse segmentation and provided support for it.

Thanks to Jocelyn Riseberg for advice on statistics.

Thanks to Wendy Plesniak, Michele Evard, Janet Cahn, Bill Gardner, Deb Roy, Brenden Maher, Nick Sawhney, Jordan Slott, and Jim Clemens for their friendship and support during my time at the Media Lab. Thanks to Wendy and Janet for always being there for me when I needed advice.

Thanks to Eric Hulteen for supporting my initial ideas about this work. Eric provided a lot of support when I really needed it.

Thanks to Linda Peterson for all her support. Thanks also to Santina Tonelli, Felice Napolitano, Rebecca Prendergast, Deirdre Fay, and Lena Davis.

Philip Sampson and John Kreifeldt taught me about human-machine interface design at Tufts University. This foundation continues to be a driving force in all of my work.

Thanks to Mike Hawley for nagging me until I agreed to play ice hockey with the FreezeFrames. I don't think he realized what kind of obsession with the sport this would lead to.

Thanks to my teammates on the MIT Women's Ice Hockey team for keeping me sane (and sometimes insane) over the last four years, especially coaches Katia "Tweety" Pashkevitch and EJ MacDonald, assistant coaches Susie Wee and Kirsten Domingo, tape girl Julie Nichols, and fellow "fossil" Ph.D. sufferers: Tory "Herminator" Herman, Esther "Fred" Jesurum, Allison "Moose" Mackay, Aradhana "Smitty" Narula, Raquel "Rocky" Romano, Cynara Wu, and my linemate of four years, Jill Depto.

Thanks to my family for their support and for putting up with me! Thanks to my mom and dad for making me believe I could do anything I set my mind to, and always being there for me.

Appendix A:
Classification and Regression Tree Analysis

Classification and regression tree analysis (CART) was used to examine the value of different acoustic features for predicting structure. An important advantage of using CART is that it can "statistically select the most significant features" while allowing "human interpretation and exploration of their result" [Riley 1989]. The implementation of CART used in this thesis was developed by Michael Riley of AT&T Research. More extensive descriptions of tree-based statistical models are given by Breiman [Breiman et al. 1984].

To explain how CART works, a simple example is shown in Figure A-1. In this example, an attempt is made to classify a team of women hockey players as forwards or defense using their weight. There is one feature—weight, which is a continuous independent variable. There are two classes—forwards and defense, which are categorical dependent variables.

Team Roster

Weight	Position
105	Defense
125	Forward
125	Forward
125	Forward
125	Forward
130	Forward
130	Forward
130	Forward
135	Forward
135	Forward
140	Defense
145	Defense
150	Defense
150	Defense
150	Defense
155	Defense
155	Defense
155	Defense
160	Defense
160	Forward
165	Defense
170	Forward
180	Defense

Figure A-1: Sample data with one feature (hockey player weight) and two classes (forward, defense).

The CART algorithm starts with all of the data to be classified and *splits* the data using one of the features provided. Binary splits of the data are performed iteratively, creating a decision tree. In this example, since there is only one feature, there is only one splitting point in the tree. If more than one feature is provided, the CART program must select the optimal feature to use to split the data at any given point. At a given node, the CART program selects the feature that results in the lowest classification error. Note that since this is done step by step at each node in the tree, the solution is not guaranteed to be globally optimal [Riley 1989].

Figure A-2 shows the tree output for the hockey player classification problem. Since the actual CART program output can be confusing to interpret, more explanatory labels were added for someone unfamiliar with the technique.

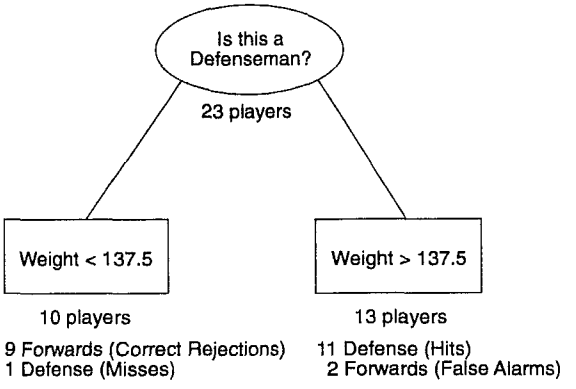


Figure A-2: CART output explained.

Here is a synopsis of the problem and result:

- There are 23 players on this women's ice hockey team.
- 12 of the players are defense, and 11 are forwards.
- The data is split based on weight (since this is the only feature provided).
- CART selects 137.5 lb. as the splitting point.
- There are 13 players weighing > 137.5 lb., and 10 players weighting < 137.5 lb.
- Of the 13 players weighing > 137.5 lb.: 11 are defense, 2 are forwards.
- Of the 10 players weighing < 137.5 lb., 9 are forwards, and 1 is defense.
- In predicting if a player is a defenseman based on her weight:
  11 players are correctly identified as defense (Hits)
  2 players are incorrectly identified as defense (False alarms)
  9 players are correctly identified as forwards (Correct Rejections)
  1 player is incorrectly identified as a forward (Misses)

Figure A-3 shows the actual tree output generated by the CART program.

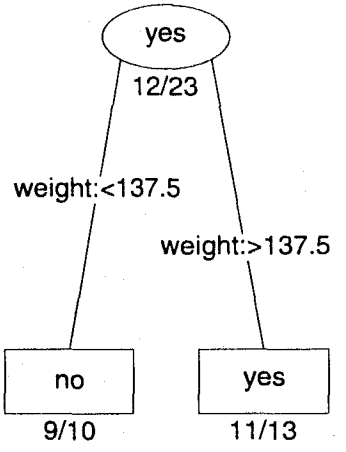


Figure A-3: CART output. Note that the two classes, defense and forward, have been labeled "yes" and "no." Since we want to predict if a player's position is defense, "yes" is equivalent to "plays defense."

Appendix B:
Cue Phrases as Indicators of Segment Beginnings

In addition to the acoustic features described in Section 5.10.3, cue words were also correlated with the discourse segmentation (for background on cue words, see Section 2.3.2). Speakers in the study often combined cue words (e.g., okay so, okay now, so now). Hirschberg and Litman report that "now" is often used in combination with other cue phrases [Hirschberg and Litman 1993]. Many cue phrases are also preceded by the word "and" (e.g., and so, and lastly, and now) or the filled pause "um." In order to account for all of these cases, a basic list of single cue words was first created. Then, a simple context free grammar was used to generate combinations of up to three cue phrases (e.g., "and hopefully then").

This list of potential cue phrases was then automatically matched against each transcript. A record was kept of which phrases began with cue words, and for which phrases the cue words alone comprised the entire phrase. Figure B-1 gives a list of all the cue phrase combinations found in the six discourse samples.

so	and so	now	and now	so now	and so now
okay	okay so	okay now	and say okay	okay then	allright
well	one	well one	and one	but one	two
and two	three	last	the last	and lastly	but anyway
first	the first	well first	first of all	so first of all	now first of all
next	the next	second	the second	secondly	the third
then	so then	and then	again	so again	and again
but again	and then again	also	and also	and then also	basically
so basically	and basically	so overall	and finally	and then finally	so as a result
and as a result	so in conclusion	for example	so for example	say	let's say
so let's say	and hopefully then	and hopefully	in addition	but in addition	meanwhile
though	yet	still	otherwise	further	however
therefore	because				

Figure B-1: Cue phrases used in the six discourse samples.

Major phrases beginning with cue words were then correlated with discourse segment beginnings agreed upon by at least four out of the five segmenters in the study. Figure B-2 shows the results of using cue words that begin a phrase (startsWithCue) to predict SBEGs in the training data (i.e., for each of the six speakers in the lecture corpus).

Speaker	%Recall	%Precision	%Fallout	%Error
1	46	47	10	17
2	65	31	14	16
3	46	35	5	8
4	62	62	5	8
5	78	31	16	16
6	43	50	6	12
All	59	36	11	13

Figure B-2: Evaluation metrics for predicting SBEG phrases using the feature startsWithCue—phrases that begin with cue words.

The results in Figure B-2 show that over all the speakers, 59% of SBEG phrases began with cue words. However, out of all the possible phrases that began with cue words, only 36% were associated with SBEG phrases. This means that while speakers often began SBEG phrases with

cue words, they also began other phrases (non-SBEGs) with cue words. For example, speaker 2 frequently used the cue word "so" as a means of checking with the audience to make sure they were following him. Note that cue phrases could also be used to signal other aspects of discourse structure, such as segment final phrases. However, for this corpus, only a handful of cue phrases were associated with segment final utterances. A limitation of this analysis is the lack of accenting information. According to Hirschberg and Litman's findings, the type of accent on a cue word that begins a phrase helps to distinguish discourse uses (i.e., those that signal structural information) from others.

Hirschberg and Litman found phrasing to be a good discriminator of discourse uses of cue words. Therefore, one might expect cue words in a separate major phrase to be an accurate (i.e., high precision) indicator of SBEG phrases. However, out of the total number of phrases made up of cue words alone, only 47% are associated with segment beginnings.

This section has discussed the correlation between cue phrases and discourse segment beginnings. Further analysis was also performed using combinations of linguistic features—acoustic cues and cue phrases (see Appendix C).

Appendix C:
Speaker Dependent Analysis of Feature Combinations for Predicting SBEGs using CART

Classification and regression tree analysis (CART) was used to analyze combinations of linguistic features for classifying segment beginning phrases. Both acoustic features (Figure 5-40) and cue phrases (Appendix B) were used in this analysis. These features are used to predict the class of each phrase in a recording. There are two possible classes, segment beginning (SBEG) and non-segment beginning.

Using CART, a separate tree was generated for each of the six speakers in the lecture corpus. The goal was to examine each speaker individually before attempting to combine the data across speakers. Trees are then pruned to an "honest" tree length. According to Riley "too large trees may match the training data well, but they won't necessarily perform well on new test data, since they have overfit the data" [Riley 1989, 341]. The CART program outputs cross-validated[20] misclassification rates for different tree lengths. The user then prunes the tree, attempting to minimize both the classification error and tree length. Since the goal is to determine the most significant features associated with segment beginning phrases, all trees were pruned to 10 nodes or fewer.

The structure agreed upon by the group of naive coders is used for this analysis (Figure 5-35 gave the number of SBEGs agreed upon out of the total number of phrases for each speaker). The top feature used by CART for splitting the data (i.e., first split in the tree) is priorPause for all speakers, except one. For speaker 2, the first splitting feature is startsWithCue (i.e., the phrase begins with cue words). The complete tree for speaker 4 is shown in Figure C-1.

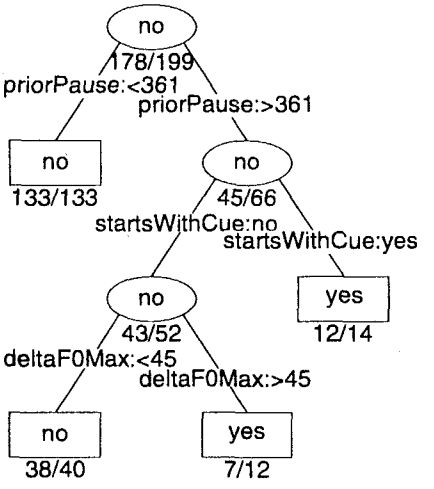


Figure C-1: CART tree for classifying SBEG phrases for speaker 4.

[20]Cross-validation involves growing the tree on a portion of the data (e.g., 9/10) and testing on the remaining data, repeating this (for 9/10 split, repeat 10 times), and then averaging the results.

For speaker 4, all phrases with a prior pause < 361 ms are non-SBEGs. For those phrases with a priorPause > 361, 12 out of 14 starting with a potential cue words (not including the conjunctions "and" and "but") are SBEGs. If the phrase does not begin with cue words, an increase in F0Max over the previous phrase (ΔF0Max) is used to classify the SBEG. This tree identifies SBEG phrases with 90% recall and 73% precision. The recall and precision for all speakers is given in Figure C-2. Note that although CART uses an internal cross-validation, the numbers output on the tree (used to calculate these evaluation metrics) are not cross-validated. Therefore, the results shown in Figure C-2 only indicate how well the tree classified the training data.

Speaker	% Recall	% Precision	Tree Length
1	70	90	2
2	39	86	5
3	68	71	5
4	90	73	4
5	59	96	9
6	70	66	3

Figure C-2: Recall and precision metrics for predicting segment beginning phrases agreed upon by 4 out of the 5 naive coders. The tree length is the number of terminal nodes in the tree after pruning.

Figure C-3 shows the features used by CART for distinguishing SBEGs from other phrases for each speaker. Notice that there are a lot of differences between speakers. The length of priorPause is the only feature used for all speakers. Only four other features are used for more than a single speaker (rmsAvg, ΔF0Avg, Δrate, startsWithCue).

Features	Speaker 1	2	3	4	5	6
Absolute						
priorPause	●	●	●	●	●	●
subseqPause						
F0Max						
F0Range						
F0RangeInterval					●	
F0Avg						
subseqF0Max					●	
subseqF0Range						
subseqF0RangeInterval			●			
subseqF0Avg		●				
rmsAvg		●			●	
phraseLen						
rate						
Relative						
ΔF0Max				●		
ΔF0Range						
ΔF0RangeInterval						
ΔF0Avg			●		●	
subseqΔF0Max						
subseqΔF0Avg						
ratioRmsAvg						
ΔRmsAvg						●
Δrate			●		●	
Categorical						
startsWithCue		●		●		

Figure C-3: Features selected by CART for distinguishing SBEGs from other phrases. SBEGs were agreed upon by at least 4 out of the 5 naive coders.

References

[Abowd et al. 1996] G. D. Abowd, C. G. Atkeson, A. Feinstein, C. Hmelo, R. Kooper, S. Long, N. Sawhney and M. Tani. Teaching and Learning as Multimedia Authoring: The Classroom 2000 Project. In Proceedings of ACM Multimedia 96, pages 187–198. 1996.

[Arai et al. 1997] T. Arai, D. Aust and S. E. Hudson. PaperLink: A Technique for Hyperlinking from Real Paper to Electronic Content. In Proceedings of CHI '97, pages 327–333. ACM, 1997.

[Arai et al. 1995] T. Arai, K. Machii and S. Kuzunuki. Retrieving Electronic Documents with Real-World Objects on InteractiveDESK. In Proceedings of the ACM Symposium on User Interface Software and Technology, pages 37–38. ACM, 1995.

[Arons 1991] B. Arons. Hyperspeech: Navigating in speech-only hypermedia. In Proceedings of Hypertext '91, pages 133-146. ACM, 1991.

[Arons 1992a] B. Arons. A Review of the Cocktail Party Effect. *Journal of the American Voice I/O Society*, 12:35–50, 1992.

[Arons 1992b] B. Arons. Techniques, perception, and applications of time-compressed speech. In Proceedings of AVIOS '92, pages 169-177. American Voice I/O Society, 1992.

[Arons 1994a] B. Arons. Interactively Skimming Recorded Speech. Ph.D. Thesis. Massachusetts Institute of Technology, 1994.

[Arons 1994b] B. Arons. Pitch-Based Emphasis Detection for Segmenting Speech Recordings. In Proceedings of the International Conference on Spoken Language Processing, pages 1931-1934. 1994.

[Arons 1997] B. Arons. SpeechSkimmer: A System for Interactively Skimming Recorded Speech. *ACM Transactions on Computer-Human Interaction*, 4(1):3–38, 1997.

[Ayers 1994] G. Ayers. Discourse Functions of Pitch Range in Spontaneous and Read Speech. OSU Linguistics Department Working Papers, vol. 44, 1994.

[Beckman and Ayers] Guidelines for ToBI labeling http://ling.ohio-state.edu/Phonetics/ToBI/ToBI0.html.

[Beckman and Hirschberg] The ToBI Annotation Conventions http://ling.ohio-state.edu/Phonetics/ToBI/ToBI6.html.

[Bishop 1995] C. M. Bishop. Neural Networks for Pattern Recognition. New York, Oxford University Press, Inc., 1995.

[Breiman et al. 1984] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone. Classification and Regression Trees. Pacific Grove, CA, Wadsworth & Brooks, 1984.

[Brown et al. 1980] G. Brown, K. Currie and J. Kenworthy. Questions of Intonation. University Park Press, 1980.

[Brown et al. 1994] M. G. Brown, J. T. Foote, G. J. F. Jones, K. S. Jones and S. J. Young. Video Mail Retrieval Using Voice: An Overview of the Cambridge/Olivetti Retrieval System. In Proceedings of ACM Multimedia 94 Workshop on Multimedia Database Management Systems, pages 47–55. ACM, 1994.

[Brown et al. 1995] M. G. Brown, J. T. Foote, G. J. F. Jones, K. S. Jones and S. J. Young. Automatic Content-Based Retrieval of Broadcast News. In Proceedings of ACM Multimedia 95, pages 35–42. ACM, 1995.

[Butterworth 1975] B. Butterworth. Hesitation and Semantic Planning in Speech. *Journal of Psycholinguistic Research*, 4:75-87, 1975.

[Carletta 1996] J. Carletta. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[Chafe 1980] W. L. Chafe, Ed. (1980). *The Deployment of Consciousness in the Production of Narrative*. The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production. Ablex Publishing Corp.

[Chen and Withgott 1992] F. R. Chen and M. Withgott. The Use of Emphasis to Automatically Summarize Spoken Discourse. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pages 229-233. IEEE, 1992.

[Cherry 1953] E. C. Cherry. Some Experiments on the Recognition of Speech, with One and Two Ears. *Journal of the Acoustical Society of America*, 25:975-979, 1953.

[Cohen 1992] M. Cohen. Integrating Graphic and Audio Windows. *Presence*, 1(4):488-481, 1992.

[Comiskey et al. 1997] B. Comiskey, J. D. Albert and J. Jacobson. Electrophoretic Ink: A Printable Display Material. In Proceedings of SID 97, 1997.

[Cruz and Hill 1994] G. Cruz and R. Hill. Capturing and Playing Multimedia Events with STREAMS. In Proceedings of ACM Multimedia 1994, pages 193–200. ACM, 1994.

[Degen et al. 1992] L. Degen, R. Mander and G. Salomon. Working with audio: Integrating personal tape recorders and desktop computers. In Proceedings of CHI '92, pages 413-418. ACM, 1992.

[Dempster et al. 1977] A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum Likelihood from Incomplete Data via the *EM* Algorithm. *Journal of the Royal Statistical Society*, B 39(1):1–38, 1977.

[Denenberg et al. 1993] L. Denenberg, H. Gish, M. Meteer, T. Miller, J. R. Rohlicek, W. Sadkin and M. Siu. Gisting Conversational Speech in Real Time. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pages II-131–134. IEEE, 1993.

[Donnelly 1996] R. Donnelly. VP Engineering, ABC Radio Networks. Personal communication, 1996.

[Duda and Hart 1973] R. O. Duda and P. E. Hart. Pattern Classification and Scene Analysis. Stanford Research Institute, Meno Park, California, John Wiley & Sons, 1973.

[Fairbanks et al. 1954] G. Fairbanks, W. L. Everitt and R. P. Jaeger. Method for time or frequency compression-expansion of speech. *Transaction of the Institute of Radio Engineers, Professional Group on Audio*, AU-2:7-12, 1954.

[Feiner et al. 1993] S. Feiner, B. Macintyre and D. Seligmann. Knowledge-Based Augmented Reality. *Communications of the ACM*, 36(7):53–62, 1993.

[Flammia and Zue 1995] G. Flammia and V. Zue. N.b.: A Graphical User Interface for Annotating Spoken Dialogue. In Proceedings of the AAAI '95 Spring Symposium Series. Empirical Methods in Discourse Interpretation and Generation, pages 40–46. 1995.

[Flammia and Zue 1997] G. Flammia and V. Zue. Learning the Structure of Mixed Initiative Dialogues Using a Corpus of Annotated Conversations. In Proceedings of the 5th European Conference on Speech and Communication Technology (EUROSPEECH), 1997.

[Gaver 1986] W. W. Gaver. Auditory Icons: Using Sound in Computer Interfaces. *Human Computer Interaction*, 2:167-177, 1986.

[Ginsburg et al. 1996] A. Ginsburg, J. Marks and S. Shieber. A Viewer for PostScript Documents. In Proceedings of the ACM Symposium on User Interface Software and Technology, pages 31–32. ACM, 1996.

[Gish et al. 1991] H. Gish, M. Siu and R. Rohlicek. Segregation of Speakers for Speech Recognition and Speaker Identification. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pages 2-873–2-876. IEEE, 1991.

[Grosz and Hirschberg 1992] B. Grosz and J. Hirschberg. Some Intonational Characteristics of Discourse Structure. In Proceedings of the International Conference on Spoken Language Processing, pages 429-432. 1992.

[Grosz and Sidner 1986] B. Grosz and C. Sidner. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[Grosz et al. 1989] B. J. Grosz, M. Pollack and C. Sidner. Discourse. In M. Posner, editor, *Foundations of Cognitive Science*, MIT Press, 1989.

[Gruber 1982] J. Gruber. A comparison of measured and calculated speech temporal parameters relevant to speech activity detection. *IEEE Transactions on Communications*, COM-30(4):728–738, 1982.

[Hearst 1993] M. A. Hearst. TextTiling: A Quantitative Approach to Discourse Segmentation. Sequoia 2000 Technical Report 93/24. University of California, Berkeley, 1993.

[Hindus et al. 1993] D. Hindus, C. Schmandt and C. Horner. Capturing, Structuring, and Representing Ubiquitous Audio. *ACM Transactions on Information Systems*, 11(4):376–400, 1993.

[Hirschberg 1994] J. Hirschberg. Linguistic Cues to Discourse Segmentation. Technical Report. AT&T Bell Laboratories, 1994.

[Hirschberg and Grosz 1992] J. Hirschberg and B. Grosz. Intonational Features of Local and Global Discourse Structure. In Proceedings of the DARPA Workshop on Spoken Language Systems, pages 441–446. 1992.

[Hirschberg and Litman 1993] J. Hirschberg and D. Litman. Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, 19(3):501–530, 1993.

[Hirschberg and Nakatani 1996] J. Hirschberg and C. Nakatani. A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues. In Proceedings of ACL-96, 1996.

[Hirschberg and Pierrehumbert 1986] J. Hirschberg and J. Pierrehumbert. The Intonational Structure of Discourse. In Proceedings of the Association for Computational Linguistics, pages 136–144. 1986.

[Hobbs 1979] J. Hobbs. Coherence and co-references. *Cognitive Science*, 3(1):67-82, 1979.

[Houle et al. 1988] G. R. Houle, A. T. Maksymowicz and H. M. Penafiel. Back-End Processing for Automatic Gisting Systems. In Proceedings of AVIOS ’88. American Voice I/O Society, 1988.

[Isaacs et al. 1995] E. A. Isaacs, T. Morris, T. K. Rodriguez and J. C. Tang. A Comparison of Face-To-Face and Distributed Presentations. In Proceedings of CHI ’95, pages 354–361. ACM, 1995.

[Ishii and Ullmer 1997] H. Ishii and B. Ullmer. Tangible Bits: Towards Seamless Interfaces between People, Bits, and Atoms. In Proceedings of CHI ’97, pages 243–241. ACM, 1997.

[James 1996] D. A. James. A System for Unrestricted Topic Retrieval from Radio News Broadcasts. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pages 279–282. IEEE, 1996.

[Johnson et al. 1993] W. Johnson, H. Jellinek, L. Klotz, R. Rao and S. Card. Bridging the Paper and Electronic Worlds: The Paper User Interface. In Proceedings of INTERCHI ’93, pages 507–512. ACM, 1993.

[Kazman et al. 1996] R. Kazman, R. Al-Halimi, W. Hunt and M. Mantei. Four Paradigms for Indexing Video Conferences. *IEEE Multimedia*, 3(1):63–73, 1996.

[Kimber et al. 1995] D. G. Kimber, L. D. Wilcox, F. R. Chen and T. P. Moran. Speaker Segmentation for Browsing Recorded Audio. In Proceedings of CHI ’95, pages 212–213. ACM, 1995.

[Kobayashi and Schmandt 1997] M. Kobayashi and C. Schmandt. Dynamic Soundscape: Mapping Time to Space for Audio Browsing. In Proceedings of CHI ’97, pages 194–201. ACM, 1997.

[Krippendorff 1980] K. Krippendorff. Content Analysis: An Introduction to its Methodology. Sage Publications, 1980.

[Lamming 1991] M. G. Lamming. Towards a Human Memory Prosthesis. Technical Report #EPC-91-116 EPC-91-116. Rank Xerox EuroPARC, 1991.

[Lehiste 1979] I. Lehiste, Ed. (1979). *Perception of Sentence and Paragraph Boundaries*. Frontiers of Speech Research. Academic Press.

[Lehman 1997] S. Lehman. Speaker Rate Detection for Voice Mail Applications. Master’s Thesis. Massachusetts Institute of Technology, 1997.

[Mackay and Pagani 1994] W. E. Mackay and D. S. Pagani. Video Mosaic: Laying Out Time in a Physical Space. In Proceedings of ACM Multimedia 94, pages 165–172. ACM, 1994.

[Mackay et al. 1995] W. E. Mackay, D. S. Pagani, L. Faber, B. Inwood, P. Launiainen, L. Brenta and V. Pouzol. Ariel: Augmenting Paper Engineering Drawings. In Proceedings of CHI '95 Conference Companion, pages 421–422. ACM, 1995.

[Maksymowicz 1990] A. T. Maksymowicz. Automatic Gisting Systems for Voice Communications. In Proceedings of IEEE Aerospace Applications Conference, pages 103–115. IEEE, 1990.

[Minneman et al. 1995] S. Minneman, S. Harrison, B. Janssen, G. Kurtenbach, T. Moran, I. Smith and W. van Melle. A Confederation of Tools for Capturing and Accessing Collaborative Activity. In Proceedings of ACM Multimedia 95, pages 523–534. ACM, 1995.

[Monty 1990] M. L. Monty. Issues for Supporting Notetaking and Note Using in the Computer Environment. Ph.D. Thesis. U.C. San Diego, 1990.

[Moran et al. 1996] T. P. Moran, P. Chiu, S. Harrison, G. Kortenbach, S. Minneman and W. van Melle. Evolutionary Engagement in an Ongoing Collaborative Work Process: A Case Study. In Proceedings of CSCW '96, pages 150–159. ACM, 1996.

[Moran et al. 1997] T. P. Moran, L. Palen, S. Harrison, P. Chiu, D. Kimber, S. Minneman, W. van Melle and P. Zellweger. "I'll Get That Off the Audio": A Case Study of Salvaging Multimedia Meeting Records. In Proceedings of CHI '97, pages 202–209. ACM, 1997.

[Morris and Hirst 1991] J. Morris and G. Hirst. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17(1):21–47, 1991.

[Muller and Daniel 1990] M. J. Muller and J. E. Daniel. Toward a definition of voice documents. In Proceedings of Conference on Office Information Systems '90, pages 174-182. ACM, 1990.

[Nakatani et al. 1995] C. H. Nakatani, B. J. Grosz, D. D. Ahn and J. Hirschberg. Instructions for Annotating Discourses. Technical Report Number TR-21-95. Center for Research in Computing Technology, Harvard University: Cambridge, MA., 1995.

[Negroponte and Jacobson 1997] N. Negroponte and J. Jacobson. Surfaces and Displays. Wired. January 1997, page 212.

[Newman and Wellner 1992] W. Newman and P. Wellner. A Desk Supporting Computer-Based Interaction with Paper Documents. In Proceedings of CHI '92, pages 587–592. ACM, 1992.

[Ostendorf et al. 1995] M. Ostendorf, P. J. Price and S. Shattuck-Hufnagel. The Boston University Radio News Corpus. Technical Report Number ECS-95-001. Boston University ECS Department, 1995.

[Ostendorf and Ross 1997] M. Ostendorf and K. Ross, Ed. (1997). *A Multi-Level Model for Recognition of Intonation Labels*. Computing Prosody. forthcoming, Springer-Verlag.

[Passonneau and Litman 1993] R. J. Passonneau and D. J. Litman. Intention-Based Segmentation: Human Reliability and Correlation with Linguistic Cues. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pages 148–155, 1993.

[Passonneau and Litman 1997] R. J. Passonneau and D. J. Litman. Discourse Segmentation by Human and Automated Means. *Computational Linguistics*, 23(1):103–140, 1997.

[Pierrehumbert 1975] J. Pierrehumbert. The Phonology and Phonetics of English Intonation. Ph.D. Thesis. Massachusetts Institute of Technology, 1975.

[Pierrehumbert and Hirschberg 1990] J. Pierrehumbert and J. Hirschberg. The Meaning of Intonational Contours in the Interpretation of Discourse. In P. R. Cohen, J. Morgan and M. E. Pollack, editor, *Intentions in Communication*, pages 271–311. The MIT Press, 1990.

[Polanyi 1988] L. Polanyi. A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638, 1988.

[Poynor 1995] R. Poynor. The Hand that Rocks the Cradle. I.D. May/June 1995, pages 60–65.

[Reichman-Adar 1984] R. Reichman-Adar. Extended Person-Machine Interface. *Artificial Intelligence*, 22(2):157-218, 1984.

[Resnick 1992] P. Resnick. Skip and Scan: Cleaning up Telephone Interfaces. In Proceedings of CHI ’92, pages 419-426. ACM, 1992.

[Resnick 1993] P. Resnick. Phone-Based CSCW: Tools and Trials. *ACM Transactions on Information Systems*, 11(4):401–424, 1993.

[Riley 1989] M. D. Riley. Some Applications of Tree-based Modeling to Speech and Language. In Proceedings of Speech and Natural Language Workshop, pages 339–352. DARPA, 1989.

[Rohlicek et al. 1989] J. R. Rohlicek, W. Russell, S. Roukos and H. Gish. Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pages 627–630. IEEE, 1989.

[Roucos and Wilgus 1985] S. Roucos and A. M. Wilgus. High quality time-scale modification for speech. In Proceedings of International Conference on Acoustics, Speech, and Signal Processing, pages 493-496. IEEE, 1985.

[Roy 1996] D. K. Roy. NewsComm: A Hand-Held Interface for Interactive Access to Structured Audio. In Proceedings of CHI ’96, pages 173–180. ACM, 1996.

[Roy 1997] D. K. Roy. Speaker Indexing Using Neural Network Clustering Vowel Spectra. *International Journal of Speech Technology*, 1(1):143–149, 1997.

[Schmandt and Mullins 1995] C. Schmandt and A. Mullins. AudioStreamer: Exploiting Simultaneity for Listening. In Proceedings of CHI ’95, pages 218–219. ACM SIGCHI, 1995.

[Schmandt and Roy 1996] C. Schmandt and D. Roy. Using Acoustic Structure in a Hand-Held Audio Playback Device. *IBM Systems Journal*, 35(3&4):453–472, 1996.

[Schmandt 1980] C. M. Schmandt. Some Applications of Three Dimensional Input. Master's Thesis. Massachusetts Institute of Technology, 1980.

[Scott 1967] R. J. Scott. Time Adjustment in Speech Synthesis. *Journal of the Acoustic Society of America*, 41(1):60–65, 1967.

[Siegel and Castellan 1988] S. Siegel and N. J. Castellan. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill Book Company, second edition, 1988.

[Silverman 1987] K. Silverman. The Structure and Processing of Fundamental Frequency Contours. Ph.D. Thesis. Cambridge University, 1987.

[Silverman et al. 1992] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert and J. Hirschberg. TOBI: A Standard for Labeling English Prosody. In Proceedings of the International Conference on Spoken Language Processing, pages 867-870. 1992.

[Stifelman 1992] L. J. Stifelman. VoiceNotes: An Application for a Voice-Controlled Hand-Held Computer. Master's Thesis. Massachusetts Institute of Technology, 1992.

[Stifelman 1994] L. J. Stifelman. The Cocktail Party Effect in Auditory Interfaces: A Study of Simultaneous Presentation. MIT Media Laboratory Technical Report, 1994.

[Stifelman 1995] L. J. Stifelman. A Tool to Support Speech and Non-Speech Audio Feedback Generation in Audio Interfaces. In Proceedings of the ACM Symposium on User Interface Software and Technology, pages 171–179. ACM, 1995.

[Stifelman et al. 1993] L. J. Stifelman, B. Arons, C. Schmandt and E. A. Hulteen. VoiceNotes: A Speech Interface for a Hand-Held Voice Notetaker. In Proceedings of INTERCHI '93, pages 179-186. ACM SIGCHI, 1993.

[Swertz 1995] M. Swertz. Combining Statistical and Phonetic Analyses of Spontaneous Discourse Segmentation. In Proceedings of the XIIIth International Congress of Phonetic Sciences (ICPhS), Vol. 4, pages 208–211. 1995.

[Swertz and Ostendorf 1997] M. Swertz and M. Ostendorf. Prosodic Indications of Discourse Structure in Human-Machine Interactions. submitted manuscript. 1997.

[Therrien 1989] C. W. Therrien. Decision Estimation and Classification: An Introduction to Pattern Recognition and Related Topics. John Wiley & Sons, 1989.

[Wang and Hirschberg 1992] M. Q. Wang and J. Hirschberg. Automatic Classification of Intonational Phrase Boundaries. *Computer, Speech, and Language*, 6:175–196, 1992.

[Weber and Poon 1994] K. Weber and A. Poon. Marquee: A Tool for Real-Time Video Logging. In Proceedings of CHI '94, pages 58–64. ACM, 1994.

[Weiser 1991] M. Weiser. The computer for the 21st century. *Scientific American*, 265(3):94-102, 1991.

[Wellner 1993] P. Wellner. Interacting with Paper on the DigitalDesk. *Communications of the ACM*, 36(7):87–96, 1993.

[Wellner et al. 1993] P. Wellner, W. Mackay and R. Gold. Computer-Augmented Environments: Back to the Real World. *Communications of the ACM*, 36(7):24–26, 1993.

[Whittaker et al. 1994] S. Whittaker, P. Hyland and M. Wiley. Filochat: Handwritten Notes Provide Access to Recorded Conversations. In Proceedings of CHI '94, pages 271–277. ACM SIGCHI, 1994.

[Wightman and Ostendorf 1994] C. Wightman and M. Ostendorf. Automatic Labeling of Prosodic Patterns. *IEEE Transactions on Speech and Audio Processing*, 2(4):469–481, 1994.

[Wilcox et al. 1994] L. Wilcox, D. Kimber and F. Chen. Audio Indexing Using Speaker Identification. In Proceedings of SPIE Conference on Automatic Systems for the Identification and Inspection of Humans, Vol. 2277, pages 149–157. 1994.

[Wilcox et al. 1992] L. Wilcox, I. Smith and M. Bush. Wordspotting for voice editing and audio indexing. In Proceedings of CHI '92, pages 655-656. ACM, 1992.

[Wilcox et al. 1997] L. D. Wilcox, B. N. Schilit and N. Sawhney. Dynomite: A Dynamically Organized Ink and Audio Notebook. In Proceedings of CHI '97, pages 186–193. ACM, 1997.

[Wolf and Rhyne 1992] C. G. Wolf and J. R. Rhyne. Facilitating Review of Meeting Information Using Temporal Histories. Technical Report RC 19811 (80426). IBM Research Division, T. J. Watson Research Center, 1992.

[Zimmerman et al. 1995] T. G. Zimmerman, J. R. Smith, J. A. Paradiso, D. Allport and N. Gershenfeld. Applying Electric Field Sensing to Human-Computer Interfaces. In Proceedings of CHI '95, pages 280–287. ACM, 1995.