This article appeared in the March 1997 issue of ACM Transactions on Computer-Human Interaction (Volume 4, Number 1, pages 3–38) and is copyright 1997 ACM.

Note that the text that appeared in printed journal contains very minor typographic and grammatical corrections that do not appear in this version.

SpeechSkimmer:
A System for Interactively Skimming Recorded Speech

Barry Arons

MIT Media Lab

This work was sponsored by Apple Computer, Inc. and Interval Research Corporation.

Authors present address: Barry Arons, Speech Interaction Research, PO Box 14, Cambridge, MA, 02142; e-mail: barons@media.mit.edu.

ABSTRACT

Listening to a speech recording is much more difficult than visually scanning a document because of the transient and temporal nature of audio. Audio recordings capture the richness of speech, yet it is difficult to directly browse the stored information. This paper describes techniques for structuring, filtering, and presenting recorded speech, allowing a user to navigate and interactively find information in the audio domain.

This paper describes the SpeechSkimmer system for interactively skimming speech recordings. SpeechSkimmer uses speech processing techniques to allow a user to hear recorded sounds quickly, and at several levels of detail. User interaction, through a manual input device, provides continuous real-time control of the speed and detail level of the audio presentation. SpeechSkimmer reduces the time needed to listen by incorporating time-compressed speech, pause shortening, automatic emphasis detection, and non-speech audio feedback. This paper presents a multi-level structural approach to auditory skimming, and user interface techniques for interacting with recorded speech. An observational usability test of SpeechSkimmer is discussed, as well as a re-design and re-implementation of the user interface based on the results of this usability test.

KEYWORDS

Speech skimming, audio browsing, speech user interfaces, interactive listening, time compression, speech as data, non-speech audio.

1. INTRODUCTION

Speech is a powerful communications medium that is rich and expressive. Speech is natural, portable, and can be used while doing other things. It is faster to speak than it is to write or type [Gould 1982]; however, it is slower to listen than it is to read. Therefore, recording speech is efficient for the talker, but hearing recorded speech is usually a burden on the listener. Skimming, browsing, and searching are traditionally considered visual tasks that one readily performs while reading a newspaper, window shopping, or driving a car. However, there is no natural way for humans to skim speech information because of the transient nature of audio—the ear cannot skim in the temporal domain the way the eyes can browse in the spatial domain.

This paper describes SpeechSkimmer, a system for skimming speech recordings that attempts to overcome these problems of slowness and the inability to browse audio. SpeechSkimmer uses simple speech processing techniques to allow a user to hear recorded sounds quickly, and at several levels of detail. User interaction through a manual input device provides continuous real-time control over the speed and detail level of the audio presentation.

SpeechSkimmer explores a new paradigm for interactively skimming and retrieving information in speech interfaces. This research takes advantage of knowledge of the speech communication process by exploiting structure, features, and redundancies inherent in spontaneous speech. Talkers embed lexical, syntactic, semantic and turn-taking information into their speech as they have conversations and articulate their ideas [Levelt 1989]. These cues are realized in the speech signal, often as hesitations or changes in pitch and energy.

Speech also contains redundant information; high-level syntactic and semantic constraints of English allow us to understand speech when it is severely degraded by noise, or even if entire words or phrases are removed. Within words there are other redundancies that allow partial or entire phonemes to be removed while still retaining intelligibility.

This research attempts to exploit acoustic cues to segment recorded speech into semantically meaningful chunks. The recordings are then time-compressed to further remove redundant speech information. While there are practical limits to time compression, there are compelling reasons to be able to quickly skim a large speech document. For skimming, redundant as well as non-redundant segments of speech must be removed. Ideally, as the skimming speed increases, the segments with the least information content are eliminated first.

When searching for information visually, we tend to refine our search over time, looking successively at more detail. For example, we may glance at a shelf of books to select an appropriate title, flip through the pages to find a relevant chapter, skim headings to find the right section, then alternately skim and read the text until we find the desired information. To skim and browse recorded speech in an analogous manner the listener must have interactive control over the level of detail, rate of playback, and style of presentation. SpeechSkimmer allows a user to control the auditory presentation through a simple interaction mechanism that changes the granularity, time scale, and style of presentation of the recording.

1.1 Speech as Sound

Along with the meaning of our spoken words, our emotions, and important syntactic and semantic information are captured by the pitch, timing, and volume of our speech. At times, more significance can be transmitted with silence than by the use of words. Such information is difficult to convey in a textual or graphical form, and is best captured in the sounds themselves. Transcripts can be useful for browsing visually, or for electronic keyword searches. However, transcripts are expensive, and automated transcriptions of spontaneous speech, meetings, or conversations are not practical in the foreseeable future [Roe 1993]. A waveform, spectrogram, or other graphical representation can be displayed (see also 3.2), yet this does not indicate what was spoken, or how something was said. Speech needs to be heard.

A graphical user interface may make some speech searching and skimming tasks easier, but there are two reasons for exploring interfaces without a visual display. First, there are a variety of situations where a graphical interface cannot be used, such as while walking, driving, or if the user is visually impaired. Second, an important issue addressed in this research is structuring and extracting information from the speech signal, and the presenting it in an auditory form. Once techniques are developed to process and present speech information that take advantage of the audio channel, they can be applied to visual interfaces.

Early versions of SpeechSkimmer therefore explored moving through speech recordings without a visual display. In the usability test of the system, users requested some graphical feedback to help them navigate through a speech recording. The revised SpeechSkimmer user interface incorporates a small amount of visual feedback, but it still can be used without looking at it.

2. BASE TECHNOLOGIES AND SEGMENTATION

This section introduces the core speech technologies used in the SpeechSkimmer system including time compression of speech, adaptive speech detection, emphasis detection, and segmenting recordings based on pauses and pitch. Readers interested in the user interface design and testing of SpeechSkimmer should skip to section 3.

2.1 Time Compressing Speech

The length of time needed to listen to an audio recording can be reduced through a variety of time compression methods (reviewed in [Arons 1992a]). These techniques allow recorded speech to be sped up (or slowed down) while maintaining intelligibility and voice quality. Time compression can be used in many application environments including voice mail, recordings for the blind, and human-computer interfaces.

2.1.1 Time-Compression Techniques

A recording can simply be played back with a faster clock rate than it was recorded at, but this produces an increase in pitch causing the speaker to sound like Mickey Mouse. This frequency shift results in an undesirable decrease of intelligibility. The most practical time compression techniques work in the time domain and are based on removing redundant information from the speech signal. In the *sampling* method [Fairbanks 1954], short segments[1] are dropped from the speech signal at regular intervals (figure 1). Cross fading, or smoothing, between adjacent segments improves the resulting sound quality.

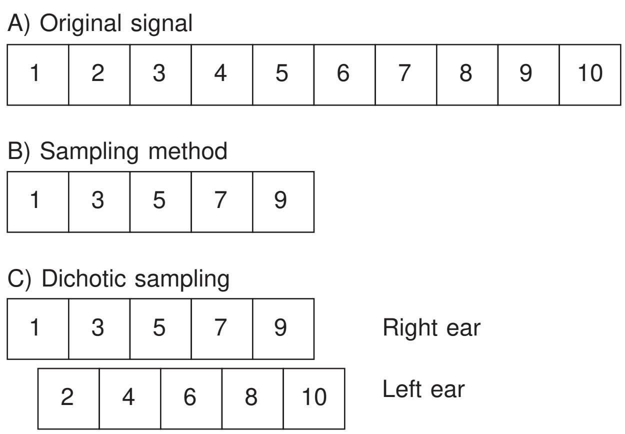

Figure 1. For a 2x speed increase using the sampling method (B), every other chunk of speech from the original signal is discarded (50 ms chunks are used). The same technique is used for dichotic presentation, but different segments are played to each ear (C).

The *synchronized overlap add method* (SOLA) is a variant of the sampling method that is becoming prevalent in computer-based systems [Roucos 1985; Hejna 1990]. Conceptually, the SOLA method consists of shifting the beginning of a new speech segment over the end of the preceding segment (figure 2) to find the point of highest cross-correlation (i.e., maximum similarity). The overlapping frames are averaged, or smoothed together, as in the sampling method. SOLA can be considered a type of selective sampling that effectively removes entire

[1]The segments are typically 30–50 ms; longer than a pitch period, but shorter than a phoneme.

pitch periods. SOLA produces the best quality speech for a computationally efficient time domain technique.

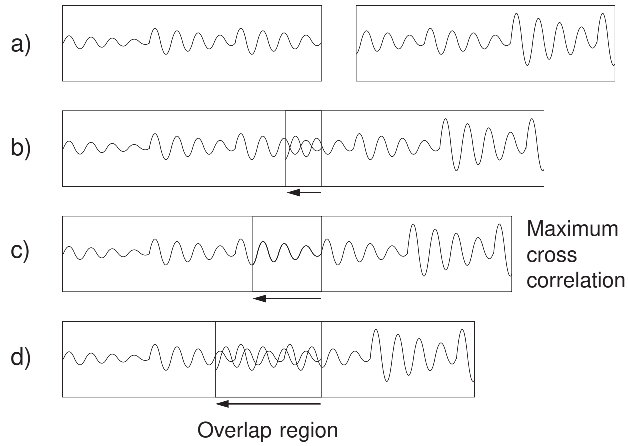

Figure 2. SOLA: shifting the speech segments (as in figure 1) to find the maximum cross correlation. The maximum similarity occurs in case c, eliminating a whole pitch period.

Sampling with dichotic presentation[2] is a variant of the sampling method that takes advantage of the auditory system's ability to integrate information from both ears. It improves on the sampling method by playing the standard sampled signal to one ear and the "discarded" material to the other ear [Scott 1967] (figure 1C). Intelligibility and comprehension increase under this dichotic presentation condition when compared with standard presentation techniques [Gerber 1977].

SpeechSkimmer incorporates several time compression techniques for experimentation and evaluation purposes. All of these speech processing algorithms run in real-time on the main processor of a laptop computer (an Apple Macintosh PowerBook 170 was used) and do not require special signal processing hardware.

2.1.2 Perception of Time-Compressed Speech

Intelligibility usually refers to the ability to identify isolated words. *Comprehension* refers to understanding the content of the material (obtained by asking questions about a recorded passage). Early studies showed that isolated words that are carefully selected and trained can remain intelligible up to 10 times normal speed, while continuous speech remains comprehensible up to about twice (2x) normal speed. Time compression decreases comprehension because of a degradation of the speech signal and a processing overload of short-term memory. A 2x increase in speed removes virtually all redundant information [Heiman 1986]; with greater compression, critical non-redundant information is also lost.

Both intelligibility and comprehension improve with exposure to time-compressed speech. Beasley informally reported that following a 30 minute exposure to time-compressed speech, listeners became uncomfortable if they were forced to return to the normal rate of presentation [Beasley 1976]. Beasley also found that subjects' listening rate preference shifted to faster rates after exposure to compressed speech. Perception of time-compressed speech is reviewed in more detail in [Arons 1992a; Beasley 1976; Foulke 1971].

2.2 Pauses in Speech

Removing (or shortening) pauses from a recording can be used as a form of time compression. The resulting speech is "natural, but many people find it exhausting to listen to because the speaker never pauses for breath" [Neuburg 1978]. In the perception of normal

[2] A different signal is played to each ear through headphones.

speech, it has been found that pauses exert a considerable effect on the speed and accuracy with which sentences were recalled, particularly under conditions of cognitive complexity—“Just as pauses are critical for the speaker in facilitating fluent and complex speech, so are they crucial for the listener in enabling him to understand and keep pace with the utterance” [Reich 1980]. Pauses, however, are only useful when they occur between clauses within sentences—pauses within clauses are disrupting. Pauses suggest the boundaries of material to be analyzed, and provide vital cognitive processing time.

Hesitation pauses are not under the conscious control of the talker, and average 200–250 ms. *Juncture* pauses are under talker control, usually occur and major syntactic boundaries, and average 500–1000 ms [Minifie 1974]. Recent work, however, suggests that such categorical distinctions of pauses based solely on length cannot be made [O’Shaughnessy 1992]. Juncture pauses are important for comprehension and cannot be eliminated or reduced without interfering with comprehension [Lass 1977].

2.2.1 Adaptive Speech Detection

Speech is a time-varying signal; silence (actually background noise) is also time-varying. Background noise may consist of mechanical noises such as fans, that can be defined temporally and spectrally, but can also consist of conversations, movements, and door slams that are difficult to characterize. Speech detection involves classifying these two types of signals. Due to the variability of the speech and background noise patterns, it is desirable to use an adaptive solution that does not rely on arbitrary fixed thresholds [Souza 1983; Savoji 1989]. The most common error made by speech detectors is the misclassification of unvoiced consonants, or weak voiced segments, as background noise.

An adaptive speech detector (based on [Lamel 1981]) was developed for shortening and removing pause, and to provide data for segmentation. Digitized speech files are analyzed in several passes. The first pass gathers energy[3] and zero crossing rate[4] (ZCR) statistics for 10 ms frames of audio. The background noise level is determined by smoothing a histogram of the energy measurements, and finding the peak of the histogram. The peak corresponds to an energy value that is part of the background noise. A value several dB above this peak is selected as the dividing line between speech and background noise. The noise level and ZCR metrics provide an initial classification of each frame as speech or background noise.

Additional passes through the sound data are made to refine this estimation based on heuristics of spontaneous speech. This processing fills-in short gaps between speech segments [Gruber 1982], removes isolated islands initially classified as speech, and extends the boundaries of speech segments so that they are not inadvertently clipped [Gruber 1983]. For example, two or three frames initially classified as background noise amid many high energy frames identified as speech are treated as part of that speech, rather than as a short silence.

This speech detector is fast and works well under a variety of microphone and noise conditions. Audio files recorded in an office environment with computer fan noise and in a lecture hall with over 40 students have been successfully segmented into speech and background noise. See [Arons 1994a] for a review of other speech detection techniques and details of the algorithm.

2.3 Acoustically Based Segmentation

Speech recordings need to be segmented into manageable pieces before presentation. Salient audio segments can be automatically selected from a recording by exploiting properties of spontaneous speech. Segmenting audio and finding its inherent structure is essential for the success of future recording-based systems. “Finding the structure” means locating important or emphasized portions of a recording, and selecting the equivalent of paragraph or new topic

---

[3] Average magnitude is used as a measure of energy [Rabiner 1975].

[4] A high zero crossing rate indicates low energy fricative sounds such as “s” and “f.” For example, a ZCR greater than 2500 crossings/sec indicates the presence of a fricative [O’Shaughnessy 1987]. Note that the background noise in most office environments does not contain significant energy in this range.

boundaries, for the purpose of creating audio overviews or outlines. Ideally, a hierarchy of segments can be created that roughly correspond to the spoken equivalents of sections, subsections, paragraphs, and sentences of a written document.

Several acoustic cues were explored for segmenting speech:

- *Pauses* can suggest the beginning of a new sentence, thought, or topic. Studies have shown that pause lengths are correlated with the type of pause and its importance (see 2.2).
- *Pitch* is similarly correlated with a talker's emphasis and new topic introductions.
- *Speaker identification* for separating talkers in a conversation.

None of these techniques are 100% accurate at finding the important boundaries in speech recordings—they all miss some of the desired boundaries and incorrectly locate others. While it is important to minimize these errors, it is perhaps more important to be able to handle errors when they occur, as no such recognition technology will ever be perfect. SpeechSkimmer addresses using such error-prone cues by providing the user with an interface to navigate in a recording, control what segments get played, and how they are presented, allowing the user to listen to exactly what they want to hear.

2.3.1 Speech Detection for Segmentation

The adaptive speech detector developed for finding and shortening pauses (section 2.2.1) can also be used for segmentation. Since long pauses typically correspond with juncture pauses that occur at important boundaries (section 2.2), the lengths of the pauses in a recording can be used to segment the speech. For example, segmenting a recording with a pause threshold of 0.02 would select the segments of speech that occur just after the longest 2% of the pauses in the recording. Note that a relative, rather than absolute, pause length is used to adapt to the pausing characteristics of the talker.

2.3.2 Pitch-based Emphasis Detection for Segmentation

Pitch[5] provides important information for human comprehension and understanding of speech, and can also be exploited in machine-mediated systems. For example, there tends to be an increase in pitch range when a talker introduces a new topic [Hirschberg 1992; Hirschberg 1986; Silverman 1987] that is an important cue for listeners.

Chen and Withgott [Chen 1992] trained a Hidden Markov Model (HMM, see [Rabiner 1989]) to summarize recordings based on training data hand-marked for emphasis, combined with the pitch and energy content of conversations. They successfully created summaries of the recordings by selecting emphasized portions that were in close temporal proximity. This prosodic approach is promising for extracting high-level information from speech signals. An alternative technique was developed for SpeechSkimmer to detect salient segments and summarize a recording without using statistical models that require large amounts of training data.

A variety of simple pitch metrics were generated and manually correlated with a hand-marked transcript of a 15 minute recording. The metrics were gathered over one second windows of pitch data (100 frames of 10 ms). The number of frames above a threshold, and the standard deviation, were most strongly correlated with new topic introductions and emphasized portions of the transcript. Note that these two metrics essentially measure the same thing: significant range and variability in F0. The "number of frames above a threshold" metric was used in the subsequent development of the algorithm.

Since the range and baseline pitch vary considerably between talkers, it is necessary to adaptively determine the pitch threshold for a given speaker. A histogram of the pitch data is used to normalize talker variability, and a threshold is chosen to select the top 1% of the pitch frames. The number of frames in each one second window that are above the threshold are

---

[5]"Pitch" in this context means the fundamental frequency of voiced speech, and is usually denoted as F0.

counted as a measure of emphasis. The scores of nearby windows are then combined for phrase- or sentence-sized segments of the speech recording.

This pitch-based segmentation technique has been successfully used to provide a high-level summary of speech recordings for a variety of talkers. High scoring salient segments are selected and used by SpeechSkimmer to enable efficient skimming. For further information on the emphasis detector see [Arons 1994b].

2.3.3 Evaluating the Emphasis Detection

Stifelman [Stifelman 1995] compared the segmentation of the emphasis detection algorithm with a hierarchical segmentation based on Grosz and Sidner's theory of discourse structure [Grosz 1986]. Stifelman found that the emphasis algorithm has a relatively high precision (82%), but a low recall (25%), for selecting discourse boundaries in the speech sample tested. This means that the majority of segments selected by the algorithm were good segments in the discourse structure, but that the algorithm did not find all the desired segments.

The discourse structure of a monologue can be thought of as an outline. To extract high level ideas from a recording the major points in the outline are of most interest, rather than those that are deeply embedded. The outermost segments in the discourse structure need to be found for high level skimming or summarization. Figure 3 shows the percentage of segments the emphasis detector selected compared to those manually selected at each level in the discourse hierarchy. A greater proportion of the major points in the discourse structure were found, rather than embedded ones.

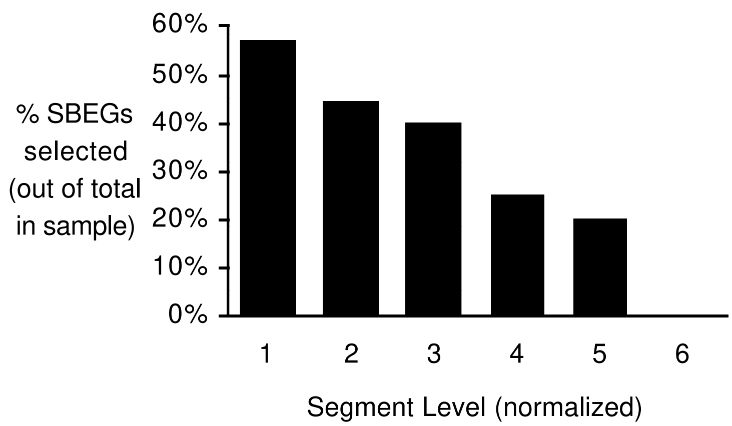

Fig. 3. Percent of segment beginnings selected for each level in the discourse hierarchy (after [Stifelman 1995]).

While there is room for improvement, these results appear promising. The emphasis detection algorithm did select a number of high level points from the discourse hierarchy without too many false alarms (see also 4.2.4). Unfortunately, the algorithm did select some segments that were deeply embedded in the discourse, and the recall rate could be improved. Using this emphasis algorithm as a starting point, it may be possible to improve these scores by tuning the algorithm, or combining it with other acoustic features such as pauses.

2.3.4 Segmentation by Speaker Identification

Acoustically based speaker identification [Reynolds 1995; Kimber 1995] can provide a powerful cue for segmentation and information retrieval in speech systems. For example, when searching for a piece of information within a recording, the search space can be greatly reduced if individual talkers can be identified (e.g., "play only things Marc said"). See section 3.2.1 to see how speaker identification data was used in SpeechSkimmer.

3. SPEECHSKIMMER PROTOTYPE

This section integrates the technologies described in the previous section into a coherent system for interactive listening. A framework is described for presenting a continuum of time compression and skimming techniques. This allows a user to quickly skim a speech recording to find portions of interest, then use time compression and pause shortening for efficient browsing, and then slow down further to listen to detailed information. A multi-level approach to auditory skimming, along with user interface techniques for interacting with the audio and providing feedback is presented.

3.1 Time Compression and Skimming

SpeechSkimmer incorporates ideas and techniques from conventional time compression algorithms, and attempts to go beyond the 2x perceptual barrier typically associated with time scaling speech. These new skimming techniques are intimately tied to user interaction to provide a range of audio presentation speeds. Backward variants of the techniques are also developed to allow audio recordings to be played and skimmed backward as well as forward. Some of the possible time compression and skimming technologies that can be used are shown in figure 4. Corresponding ranges of speed increases for the different classes of techniques are shown in figure 5.

1. Unprocessed (normal)
2. Time-compression
    - Pause removal
    - Pause shortening
    - Sampling
    - SOLA
    - Dichotic sampling or SOLA
    - Combined time compression techniques (e.g., SOLA with pause removal)
    - Backward sampling (for intelligible rewind)
3. Skimming
    - Isochronous skimming (equal time intervals)
    - Speech synchronous skimming. Segmentation based on:
        - Pauses
        - Pitch
        - Energy
        - Speaker identification
        - Word spotting
        - User selected segments
        - Combined segmentation techniques (e.g., pauses, pitch, and energy)
    - Backward skimming

Fig. 4. Techniques of time compression and skimming.

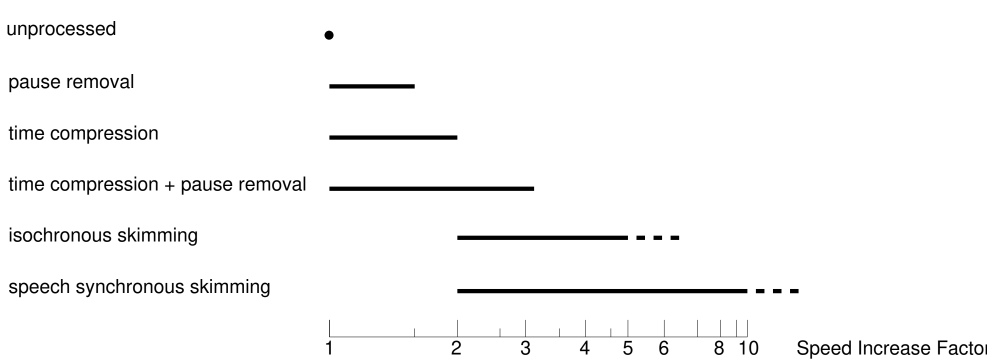

Fig. 5. Schematic representation of the range of speed increases for different time compression and skimming methods.

Time compression can be considered as “content lossless” since the goal is to present all the non-redundant speech information in the signal. The skimming techniques are designed to be “content lossy,” as large parts of the speech signal are explicitly removed. This classification is not based on the traditional engineering concept of lossy versus lossless, but is based on the intent of the processing. For example, isochronous skimming selects and presents speech segments based on equal time intervals. If only the first five seconds of each minute of speech are played, this can be considered coarse and lossy sampling. In contrast, a speech synchronous technique that selects important words and phrases using the natural boundaries in the speech will provide more information content to the listener.

3.2 Skimming Levels

There have been a variety of attempts to present hierarchical or “fisheye” views of visual information [Furnas 1986; Mackinlay 1991]. These approaches are powerful but inherently rely on a spatial organization. Temporal video information has been displayed in a similar form [Mills 1992; Davis 1993; Elliott 1993], yet this primarily consists of mapping time-varying spatial information into the spatial domain. Graphical techniques can be used for a waveform or similar display of an audio signal, but such a representation is inappropriate—*sounds need to be heard, not viewed.* This research attempts to present a hierarchical (or “fish ear”) representation of audio information that *only* exists temporally.

A continuum of time compression and skimming techniques have been designed, allowing a user to efficiently skim a speech recording to find portions of interest, then listen to it time-compressed to allow quick browsing, and then slow down further to listen to detailed information. Figure 6 presents one possible “fish ear” view of this continuum. For example, what may take 60 seconds to listen to at normal speed may take 30 seconds when time-compressed, and only five or ten seconds at successively higher levels of skimming. If the speech segments are chosen appropriately, it is hypothesized that this mechanism provides a summarizing view of a speech recording.

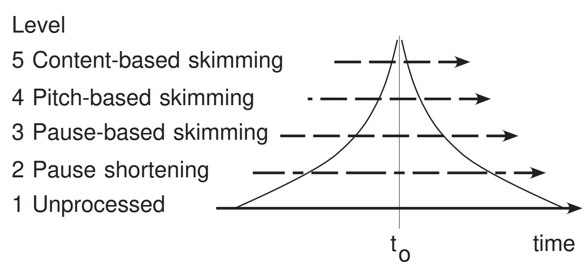

Fig. 6. A hierarchical “fish ear” time-scale continuum. Each level in the diagram represents successively larger portions of the levels below it. The curves represent iso-content lines, i.e., an equivalent time mapping from one level to the next. The current location in the sound file is represented by t_o; the speed and direction of movement of this point depends upon the skimming level.

Four distinct skimming levels have been implemented (figure 7). Within each level the speech signal can also be time-compressed. The lowest skimming level (level 1) consists of the original speech recording without any processing, and thus maintains the pace and timing of the original signal. In level 2 skimming, the pauses are selectively shortened or removed. Pauses less than 500 ms are removed, and the remaining pauses are shortened to 500 ms.[6] This technique speeds up listening yet provides the listener with cognitive processing time and cues to the important juncture pauses (section 2.2).

[6]All thresholds are determined adaptively based on the content of the speech recording.

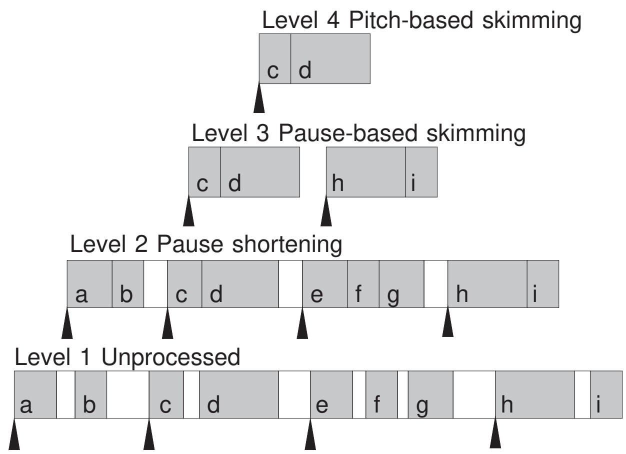

Fig. 7. Speech and silence segments played at each skimming level. The gray boxes represent speech; white boxes represent background noise. The pointers indicate valid segments to go to when jumping or playing backward.

Level 3 is based on the premise that long juncture pauses tend to indicate either a new topic, some content words, or a new talker. For example, filled pauses (i.e., “uhh” or “um”) usually indicate that the talker does not want to be interrupted, while long unfilled pauses (i.e., silences) signify that the talker is finished and someone else may begin speaking [Levelt 1989; O’Shaughnessy 1992]. Thus level 3 skimming attempts to play salient segments based on a simple heuristic: only the speech that occurs just after a significant pause in the original recording is played. For example, after detecting a pause over 750 ms, the subsequent 5 seconds of speech are played (with pauses shortened). Note again that this segmentation process is error prone, but these errors can be overcome by giving the user interactive control of the presentation.

Level 4 is similar to level 3 in that it attempts to present segments of speech that are highlights of the recording. Salient segments for Level 4 are chosen using the emphasis detector (section 2.3.2) to summarize the recording. In practice, either level 3 or level 4 is used as the top skimming level.

It is somewhat difficult to listen to level 3 or level 4 skimmed speech, as relatively short unconnected segments are played in rapid succession. It has been informally found that playing the segments at normal speed (i.e., not time-compressed), or even slowing down the speech, is useful when skimming unfamiliar material. At the highest skimming levels, a short (e.g., 600 ms) pure silence is inserted between each of the speech segments to separate them perceptually. An early version of SpeechSkimmer played recorded ambient noise between the selected segments, but this fit in so naturally with the speech that it was difficult to distinguish between segments.

3.2.1 Alternative Skimming Levels Using Speaker Identification

The SpeechSkimmer system has also been used with speaker identification-based segmentation. A two person conversation was analyzed with speaker identification software [Reynolds 1995] that determined when each talker was active. These data were translated into SpeechSkimmer format such that level 1 represented the entire conversation; jumping took the listener to the next turn change in the conversation. Level 2 played only the speech from one talker, while level 3 played the speech from the other. Jumping within these levels brought the listener to the start of that talker’s next conversational turn.

3.3 Skimming Backward

Besides skimming forward through a recording, it is desirable to play intelligible speech while interactively searching or “rewinding” through a digital audio file [Arons 1991a; Elliott 1993]. Analog tape systems provide little useful information about the signal when it is played completely backward. This is analogous to taking the text “going to the store” and presenting it as the unintelligible “erots eht ot gniog.” Digital systems allow word- or phrase-sized chunks of speech to be played forward individually, with the segments themselves presented in reverse order (resulting in “store, to the, going”). While the general sense of the recording is reversed and jumbled, each segment is identifiable and intelligible. It can thus become practical to browse backward through a recording to find a particular word or phrase. This method is particularly effective if the segment boundaries are at natural pauses in the speech. Note that this technique can also be combined with time-compressed playback, allowing both backward and forward movement at high speeds.

In addition to the forward skimming levels, the speech recordings can also be skimmed backward. Small segments of sound are each played normally, but are presented in reverse order. When level 1 and level 2 sounds are played backward (i.e., considered level –1 and level –2), short segments are selected based upon speech detection, and are played in inverse order. In figure 7 level –1 would play segments in this order: h–i, e–f–g, c–d, a–b. Level –2 is similar, but without the pauses.

3.4 Jumping

Along with controlling the skimming and time compression, it is desirable to be able to interactively jump between segments within each skimming level. If the user decides that the segment being played is not of interest, it is possible to go on to the next segment without being forced to listen to each entire segment [Arons 1991b; Resnick 1992a]. For example, in figure 7 at level 3, segments c and d would be played, then a short silence, then segments h and i. At any time while the user is listening to segment c or d, a jump forward command would immediately interrupt the audio output and start playing segment h. While listening to segment h or i, the user could jump backward, causing segment c to be played. Valid segments to jump to are indicated with pointers in figure 7.

The skimming user interface includes a control that jumps backward one segment and drops into normal play mode (level 1, no time compression). The intent of this control is to encourage high-speed browsing of time-compressed level 3 or level 4 speech. When the user hears something of interest, it is easy to use this control to back up a bit, re-hear the piece of interest, and then continue listening at normal speed.

3.5 Interaction Mappings

Finding an appropriate mapping between an input device and controlling the skimmed speech is subtle, as there are many independent variables that can be controlled. For the SpeechSkimmer prototype, the primary variables of interest are time compression and skimming level, with all others (e.g., pause shortening parameters and skimming timing parameters) held constant.

Several mappings of user input to time compression and skimming level were tried. A two-dimensional controller, such as a mouse, allows two variables to be changed independently. For example, the y-axis is used to control the amount of time compression while the x-axis controls the skimming level (figure 8). Movement toward the top increases the time compression; movement toward the right increases the skimming level. The right half is used for skimming forward, the left half for skimming backward. Moving to the upper right thus presents skimmed speech at high speed.

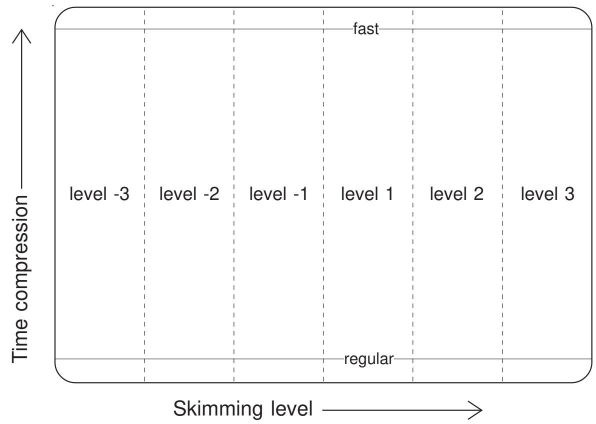

Fig. 8. Schematic representation of two-dimensional control regions. Vertical movement changes the time compression; horizontal movement changes the skimming level.

The two primary variables can also be controlled by a one-dimensional input device. For example, as the controller is moved forward, the sound playback speed is increased using time compression. As it is pushed forward further, time compression increases until crossing a boundary into the next skimming level. Pushing forward within each skimming level similarly increases the time compression (figure 9). Pulling backward has an analogous but reverse effect.

One consideration in all these schemes is the continuity of speeds when transitioning from one skimming level to the next. In figure 9, for example, when moving from fast level 2 skimmed speech to level 3 speech there is a sudden change in speed at the border between the two skimming levels. Depending upon the implementation, fast level 2 speech may be effectively faster or slower than regular level 3 speech. This problem also exists with a 2-D control scheme—to monotonically increase the effective playback speed may require a zigzag motion through skimming and time compression levels.

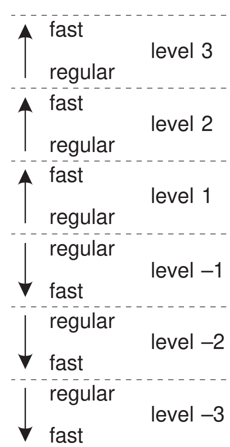

Fig. 9. Schematic representation of one-dimensional control regions.

3.6 Interaction Devices

The speech skimming system has been used with a mouse, small trackball, touchpad, and a joystick in both the one- and two-dimensional control configurations (two independent controls, one for speed and one for skimming level were not tried). A mouse provides accurate control, but as a relative pointing device [Card 1991] it is difficult to use without a display. A small hand-held trackball (e.g., controlled with the thumb) eliminates the desk space required by the mouse, but is still a relative device and is therefore also inappropriate for a non-visual task.

A joystick can be used as an absolute position device. However, if it is spring-loaded (i.e., automatic return to center), it requires constant physical force to hold it in position. If the springs are disabled, a particular position (i.e., time compression and skimming level) can be automatically maintained when the hand is removed (see Lipscomb 1993 for a discussion of such physical considerations). The home (center) position, for example, can be configured to play forward (level 1) at normal speed. Touching or looking at the joystick's position provides feedback to the current settings. However, in either configuration, an off-the-shelf joystick does not provide any physical feedback when the user is changing from one discrete skimming level to another, and it is difficult to jump to an absolute location.

A small touchpad can act as an absolute pointing device and does not require any effort to maintain the last position selected. A touchpad can be easily modified to provide a physical indication of the boundaries between skimming levels. Unfortunately, a touchpad does not provide any physical indication of the current location once the finger is removed from the surface.

3.7 Touchpad Configuration

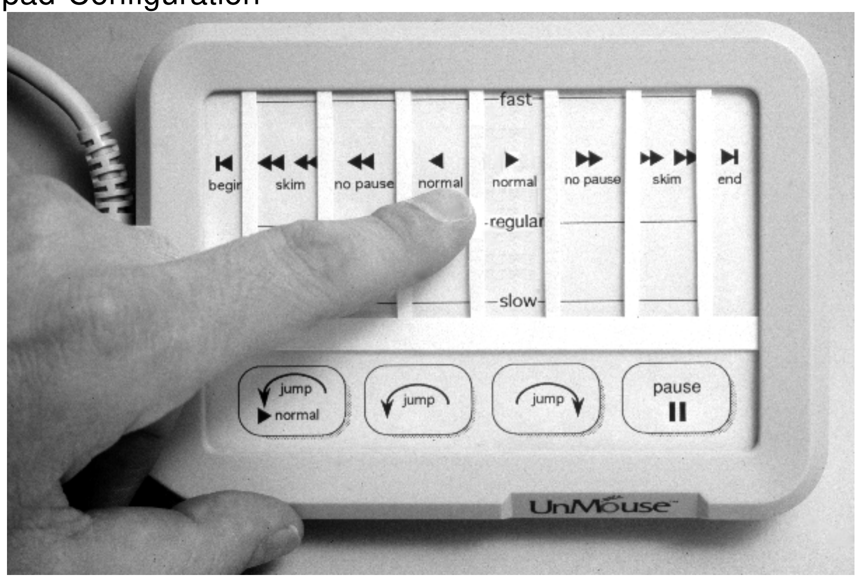

Fig. 10. The touchpad with paper guides for tactile feedback.

The SpeechSkimmer prototype uses a small (7 x 11 cm) touchpad [Microtouch 1992] with a two-dimensional control scheme. Small strips of paper were added to the touch-sensitive surface as tactile guides to indicate the boundaries between skimming regions (figure 10). In addition to the six regions representing skimming levels, two additional regions were added to jump directly to the beginning and end of the sound recording. Four buttons provide jumping and pausing capabilities (figure 11). Note that the template used in the touchpad only contains static information; it is not necessary to look at it to use the system.

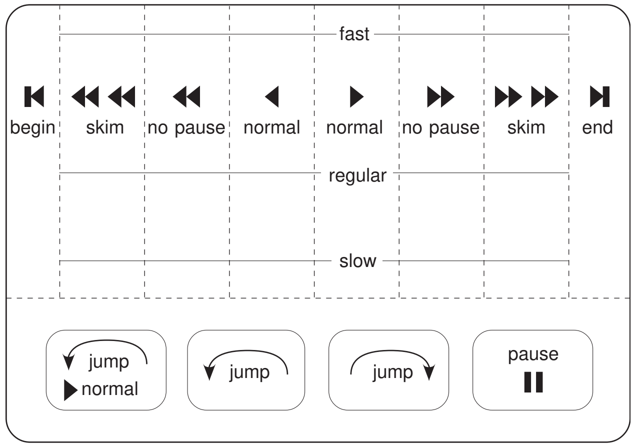

Fig. 11. Template used in the touchpad (a printed version of this fits behind the touch sensitive surface of the pad). The dashed lines indicate the location of the paper guide strips.

The time compression control (vertical motion) is not continuous, but provides a "finger-sized" region around the "regular" mark that plays at normal speed (figure 12). To enable fine-grained control of the time compression [Stifelman 1992], a larger region is allocated for speeding the speech up than for slowing it down. The areas between the tactile guides form virtual sliders that control the time compression within a skimming level (note that only one slider is active at a time).

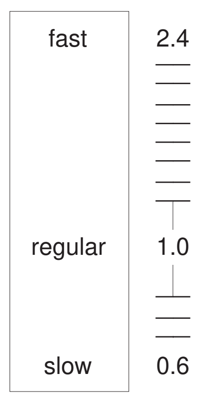

Fig. 12. Mapping of the touchpad control to the time compression range (0.6x to 2.4x).

3.8 Non-Speech Audio Feedback

SpeechSkimmer uses recorded sound effects to provide feedback when navigating [Buxton 1991; Gaver 1989]. Non-speech audio was selected to provide terse, yet unobtrusive navigational cues [Stifelman 1993]. For example, if the user attempts to play past the end of a sound, a cartoon "boing" is played. No explicit feedback is provided for changes in time compression. The speed changes occur with low latency and are readily apparent in the speech signal itself.

When the user transitions to a new skimming level, a short tone is played. The frequency of the tone increases with the skimming level (figure 13). A double beep is played when the user changes to normal (level 1). This acts as an audio landmark, clearly distinguishing it from the other tones and skimming levels.

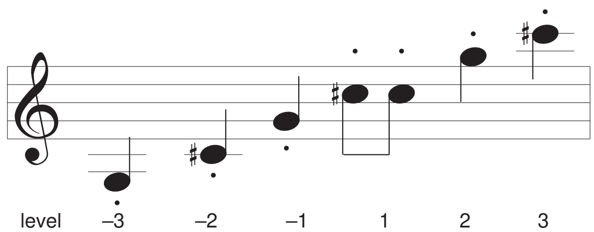

Fig. 13. A musical representation of the tones played at the different skimming levels. Notice the double beep “landmark” for normal (level 1) playing. The small dots indicate short and crisp staccato notes.

A different sound is played when each of the buttons are touched. An attempt was made to create sounds that could be intuitively linked with the function of the button. The feedback played when pausing and un-pausing are reminiscent of a piece of machinery stopping and starting. Jumping forward is associated with a rising pitch while jumping backward is associated with a falling pitch.

3.9 Software Architecture

Each recording is post-processed with the speech detection and emphasis detection algorithms. A single file is created that contains all the segmentation data used for skimming, jumping, and pause shortening.

The run-time application consists of three primary modules: a main event loop, a segment player, and a sound library (figure 14). The skimming user interface is separated from the underlying mechanism that presents the skimmed and time-compressed speech. This modularization allows for the rapid prototyping of new interfaces using different interaction devices. SpeechSkimmer is implemented in a subset of C++, and runs on Apple Macintosh computers.

The main event loop gathers raw data from the user and maps it to the appropriate time compression and skimming ranges for each input device. The event loop sends requests to the segment player to start and stop playback, jump between segments, and set the time compression and skimming levels.

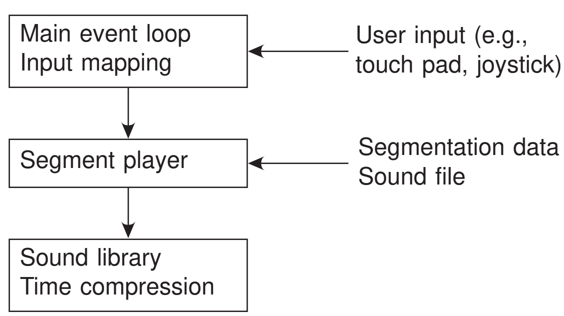

Fig. 14. Software architecture of the skimming system.

The segment player combines user input with the segmentation data to select the appropriate portion of the sound to play. When the end of a segment is reached, the next segment is selected based on the current skimming level and data in the segmentation file. Audio data is read from the sound file and passed to the sound library. The size of the audio data buffers is kept to a minimum to reduce the latency between user input and the corresponding sound output.

The sound library provides a high-level interface to the audio playback hardware (based on the functional interface described in Arons 1992b). Several different time compression algorithms are built into the sound library.

4. USABILITY TESTING

The goal of this test was to find usability problems and successes in the SpeechSkimmer user interface. The usability test was primarily an observational "thinking out loud" study [Ericsson 1984] that is intended to quickly find major problems in the user interface to an interactive system [Nielsen 1993a].

4.1 Method

4.1.1 Subjects

Twelve volunteer subjects between the ages of 21 and 40 were selected from the Media Laboratory environment. Six of the subjects were administrative staff and six were graduate students; eight were female and four were male. None of the subjects were familiar with SpeechSkimmer, but all had experience using computers. Test subjects were not paid, but were offered snacks and beverages to compensate them for their time.

4.1.2 Procedure

The tests were performed in an acoustically isolated room with a subject, an interviewer, and an observer.[7] The sessions were video taped and later analyzed by both the interviewer and observer. A testing session took approximately 60 minutes and consisted of five parts:

1. A background interview to collect demographic information and to determine what experience subjects had with recorded speech and audio. Subsequent questions were tailored based on the subject's experiences. For example, someone who regularly recorded lectures would be asked in detail about their use of the recordings, how they located specific pieces of information in the recordings, etc.

2. A first look at the touchpad. Subjects were given the touchpad (figure 10) and asked to describe their first intuitions about the device. This was done without the interviewer revealing anything about the system or its intended use, other than "it is used for skimming speech recordings." Everything in the test was exploratory, subjects were not given any instructions or guidance.[8] The subjects were asked what they thought the different regions of the device did, how they expected the system to behave, what they thought backward did, etc.

3. Listening to a trial speech recording with the SpeechSkimmer system. The subjects were encouraged to explore and "play" with the device to confirm, or discover, how the system operated. While investigating the device, the interviewer encouraged the subjects to "think aloud," to describe what they were doing, and to say if the device was behaving as they expected.

4. A skimming comparison and exercise. This portion of the test compared two different skimming techniques. A recording of a 40 minute lecture[9] was divided into two 20 minute parts (half of the subjects had attended the lecture when it was originally presented). Each subject listened to both halves of the recording; one part was segmented using the pitch-based emphasis detector (section 2.3.2) the other was segmented isochronously (i.e., at equal time intervals). All SpeechSkimmer controls were active under both conditions; users could change speed, skimming level, jump, and so on, the only difference was in the top level segmentation.

---

[7]Lisa Stifelman conducted the test, the system designer (Arons) observed.

[8]However, if a subject said something like "I wish it did X," and the system did perform that function, the feature was revealed to them by the interviewer through directed questions (e.g., Do you think this device can do that? If so, how do you think you could get it to do it? What do you think that button does?).

[9]Nicholas Negroponte's talk titled *Conflusion: Media in the Next Millennium* presented 10/19/93.

The test was counterbalanced for effects of presentation order and portion of the recording (figure 15).

# of subjects	first presentation		second presentation	
3	pitch-based	part 1	isochronous	part 2
3	isochronous	part 1	pitch-based	part 2
3	isochronous	part 2	pitch-based	part 1
3	pitch-based	part 2	isochronous	part 1

Fig. 15. Counterbalancing of experimental conditions.

When skimming, both of the segmentation techniques provided a 12:1 compression for this recording (i.e., on average five seconds out of each minute were presented). Note that these figures are for normal speed (1.0x), by using time compression the subjects could achieve over a 25:1 time savings.

The subjects first skimmed the entire recording at whatever speed they felt most comfortable. The subjects were asked to judge (on a 7 point scale) how well they thought the skimming technique did at providing an overview and selecting indexes into major points in the recording. The subjects were then given a printed list of three questions that could be answered by listening to the recording. The subjects were asked locate the answer to any of the questions in the recording, and to describe their auditory search strategy. This process was repeated for the second presentation condition.

5. The test concluded with follow-up questions regarding the subject's overall experience with the interaction device and the SpeechSkimmer system, including what features they liked or disliked, what they thought was missing from the user interface, etc.

4.2 Results and Discussion

This section summarizes the features of SpeechSkimmer that were frequently used or most liked by the subjects, as well as areas for improvement in the user interface design.

4.2.1 Background Interviews

All subjects had some experience in searching for recorded audio information on compact discs, audio cassettes, or video tape. Subjects' experience included transcribing lectures and interviews, taking personal notes on a microcassette recorder, searching for favorite songs on tape or compact disc, editing video documentaries, and receiving up to 25 voice mail messages per day. Almost all the subjects referred to the process of searching in audio recordings as time consuming, one subject added that it takes "more time than you want to spend."

4.2.2 First Intuitions

Most subjects found the interface intuitive and easy to use, and were able to use the device without any training. This ability to quickly understand how the device works is partially because the touchpad controls are labeled similarly to consumer devices such as compact disc players and video cassette recorders. While this familiarity allowed the subjects to initially feel comfortable with the device, and enabled rapid acclimatization to the interface, it also caused some confusion since a few of the SpeechSkimmer functions behave differently than on the consumer devices.

Level 2 on the skimming template is labeled "no pause" but most subjects did not have any initial intuitions about what it meant. The label baffled most subjects since current consumer devices do not have pause removal or similar functionality. Some subjects thought that once they started playing in "no pause" they would not be able to stop or pause the playback. Similarly, the function of the "jump and play normal button" was not obvious. The backward play levels were sometimes intuitively equated with traditional (unintelligible) rewind.

4.2.3 Warm-up Task

The recording used in the trial task consisted of a loose free-form discussion, and most subjects had trouble following the conversation. Most said that they would have been able to learn the device in less time if the trial recording was more coherent, or if they were already familiar with the recording. However, subjects still felt the device was easy to learn quickly.

Subjects were not sure how far the jumps took them. Several subjects thought that the system jumped to the next utterance of the male talker when exploring the interface in the trial task (the first few segments selected for jumping in this recording did occur at a change of talker because the pause-based segmentation algorithm was used).

4.2.4 Skimming

Most subjects found that the pitch-based skimming was effective at extracting interesting points to listen to, and for finding information. One user who does video editing described it as "grabbing sound bite material." When comparing pitch-based skimming to isochronous skimming a subject said "it is like using a rifle versus a shotgun" (i.e., high accuracy instead of dispersed coverage). Other subjects said that the pitch-based segments "felt like the beginning of a phrase … [and were] more summary oriented" and there was "a lot more content or keyword searching going on" than in the isochronous segmentation.

A few subjects requested that longer segments be played (perhaps until the next pause), or that the length of the segments could be controllable. One subject said "I felt like I was missing a lot of his main ideas, since it would start to say one, and then jump."

Subjects were asked to rank the skimming performance under the different segmentation conditions. A score of 7 indicates the best possible summary of high-level ideas, a score of 1 indicates very poorly selected segments. The mean score for the pitch-based segmentation was $M = 4.5$ ($SD = 1.7$, $N = 12$); the mean score for the isochronous segmentation was $M = 2.7$ ($SD = 1.4$, $N = 12$). The pitch-based skimming was rated better than isochronous skimming with a statistical significance of $p < .01$ using a t test for paired samples. No statistically significant difference was found on how subjects rated the first versus the second part of the talk, or on how subjects rated the first versus second sound presented.

Most subjects, including the two that did not think the pitch-based skimming gave a good summary, used the top skimming level to navigate through the recording. When asked to find the answers to questions on the printed list, most started off by saying something like "I'll go to the beginning and skim till I get to the right topic area in the recording," or in some cases "I think its near the end, so I'll jump to the end and skim backward."

4.2.5 No Pause

While there was some initial confusion regarding the "no pause" level, if a subject discovered its function, it often became a preferred way to quickly listen and search for information. One subject that does video editing said "that's nice … I like the no pause function.… it kills dead time between people talking … this would be really nice for interviews [since you normally have to] remember when he said [the point of interest], then you can't find where it was, and must do a binary search of the audio track … For interviews it is all audio—you want to get the sound bite."

4.2.6 Jumping

The function of the "jump and play normal" button was not always obvious. However, subjects that did not understand the button found ways to navigate and perform the same function using the basic controls. This button is a short-cut: a combination of jumping backward and then playing level 1 speech at regular speed. One subject had a moment of inspiration while skimming along at a high speed, and tried the button after passing the point of interest. After using this button the subject said in a confirming tone "I liked that, OK." The subject proceeded to use the button several more times and said "now that I figured out how to do that jump normal thing … that's very cool. I like that." It is important to note that after discovering the "jump and play normal" button this subject felt more comfortable skimming at faster speeds. Another subject said “that’s the most important button if I want to find information.”

While most of the subjects used, and liked, the jump buttons, the size or granularity of jumps was not obvious. Subjects assumed that jumping always brought them to the next sentence or topic (in the SpeechSkimmer prototype the granularity of a jump depends on the current skimming level). While using the jump button and the “backward no pause” level, one subject said “oh, I see the difference … I can re-listen using the jump key.”

4.2.7 Backward

Most subjects figured out the backward controls during the warm-up trial, but avoided using them. This is partially attributable to the subject’s initial mental models that associate backward with the conventional rewind of a tape player. Some subjects, however, did find the backward levels useful in locating particular words or phrases that had just been heard.

While listening to the recording played backward, one subject noted “it’s taking units of conversation—and goes backwards.” Another subject said that “it’s interesting that it is so seamless” for playing intelligible segments and that “compared to a tape where you’re constantly shuffling back and forth, going backward and finding something was much easier since [while] playing backwards you can still hear the words.” One subject suggested providing feedback to indicate when the recording was being played backward, to make it more easily distinguishable from forward.

4.2.8 Time Compression

Some subjects thought there were only three discrete speeds and did not initially realize that there was a continuum of playback speeds. A few subjects did not realize that the ability to change speeds extended across all the skimming levels. These problems can be attributed to the three speeds marked on the template (slow, regular, and fast; figure 11). One subject noted that the tactile strips on the surface break the continuity of the horizontal “speed” lines, and made it less clear that the speeds work at all skimming levels. Two subjects suggested using colors to denote the continuum of playback speeds and that the speed labels should extend across all the skimming levels.

Several subjects thought there was a major improvement when listening over headphones. One subject was “really amazed” at how much better the dichotic time-compressed speech was for comprehension than the monotic speech presented over the loudspeaker. Another subject commenting on the dichotic speech said “it’s really interesting—you can hear it a lot better.”

4.2.9 Buttons

The buttons were generally intuitive, but there were some problems of interpretation and accidental use. The “begin” and “end” regions were initially added next to the level 3 and –3 skimming regions on the template to provide a continuum of playback granularity (i.e., normal, no pause, skim, jump to end). Several subjects thought that the begin button should seek to the beginning of the recording and start playing (the prototype seeks to the beginning and waits for user input). One subject additionally thought the speed of playback could be changed by touching at the top or bottom of the begin button.

One subject wanted to skim backward to re-hear the last segment played, but accidentally hit the adjacent begin button instead. This frustrated the subject, since the system jumped to the beginning of the recording and hence lost the location of interest. Note also that along with these conceptual and mechanical problems, the words “begin” and “start” are overloaded and could mean “begin playing” as well as “seek to the beginning of the recording.”

By far the biggest problem encountered during the usability test was “bounce” on the jump and pause buttons.[10] This was particularly aggravating when it occurred with the pause button,

---

[10]Button “bounce” is usually associated with mechanical switches that made several temporary contact closures

as the subject would want to stop the playback, but the system would temporarily pause, then moments later un-pause. The bounce problem was partially exacerbated by the subject's use of their thumbs to touch the buttons. While the touchpad and template were designed to be operated with a single finger for maximum accuracy (as in figure 10), most of the subjects held the touchpad by the right and left sides and touched the surface with their thumbs during the test. This is partially attributable to the arrangement of the subject and the experimenters during the test. Subjects had to hold the device as there was no table for placing the touchpad.

4.2.10 Non-Speech Feedback

The non-speech audio was successful at unobtrusively providing feedback. One subject, commenting on the effectiveness and subtlety of the sounds said "after using it for a while, it would be annoying to get a lot of feedback." Another subject said that the non-speech audio "helps because there is no visual feedback." None of the subjects noted that the frequency of the feedback tone changes with skimming level; most did not even notice the existence of the tones. However, when subsequently asked about the device many noted that the tones were useful feedback to what was going on. The cartoon "boings" at the beginning and end were good indicators of the end points (one subject said "it sounds like you hit the edge"), and the other sounds were useful in conveying that something was going on. The boing sounds were noticed most often, probably because the speech playback stops when the sound effect is played.

4.2.11 Search Strategies

Several different navigation and search strategies were used when trying to find the point in the recording that answered a question on the printed list. Most subjects skimmed (level 3) the recording to find the general topic area of interest, then changed to level 1 (playing) or level 2 (pauses removed), usually with time compression. One subject started searching by playing normally (no time compression) from the beginning of the recording to "get a flavor" for the talk before attempting to skim or play it at a faster rate. One subject used a combination of skimming and jumping to quickly navigate through the recording and efficiently find the answers to the list of questions.

4.2.12 Follow-up Questions

Most subjects thought that the system was easy to use since they made effective use of the skimming system without any training or instructions. Subjects rated the ease of use of the system on a 7 point scale where 1 is difficult to use, 4 is neutral, and 7 is very easy to use. The mean score for ease of use was $M = 5.4$ ($SD = 0.97$, $N = 10$).

Most subjects liked the ability to quickly skim between major points in a presentation, and to jump on demand within a recording. Subjects liked the time compression range, particularly the interactive control of the playback speed. A few subjects were enamored with other specific features of the system including the "fast-forward no pause" level, the "jump and play normal" button, and the dichotic presentation.

One subject commented "I really like the way it is laid out. It's easier to use than a mouse." Another subject experimented with turning the touchpad 90 degrees so that moving a finger horizontally, rather than vertically, changed the playback speed.

Most subjects said they could envision using the device while doing other things, such as walking around, but few thought they would want to use it while driving an automobile. Most of the subjects said they would like to use such a device, and many of them were enthusiastic about the SpeechSkimmer system.

---

before settling to a stable state. The difficulties here are associated with the way the touchpad driver software was configured.

4.2.13 Desired Functionality
In the follow-up portion of the test, the subjects were asked what other features might be helpful for the speech skimming system. For the most part these items were obtained through probing the subjects, and were not mentioned spontaneously.

Some subjects were interested in marking points in the recording that were of interest to them, so they could go back and easily access those points later. A few of the subjects called these "bookmarks."

Some subjects wanted to be able to jump to a particular place in a recording, or have a graphical indicator of their current location. There is a desire, for example, to access a thought discussed "about three-quarters the way through the lecture" by using a "time line" for locating a specific time point.

4.3 Comments on Usability Testing
Informal heuristic evaluation of the interface [Nielsen 1990; Nielsen 1991; Jeffries 1991] was performed throughout the system design. In addition, the test described in section 4.1 was very helpful in finding usability problems. The test was performed relatively late in the SpeechSkimmer design cycle, and, in retrospect, a preliminary test should have been performed much earlier. Most of the problems in the template layout could have been easily uncovered with only a few subjects. This could have led to a more intuitive interface, while focusing on the features most desired by users.

Note that while twelve subjects were tested here, only a few are needed to get helpful results. Nielsen has shown that the maximum cost-to-benefit ratio for a usability project occurs with around three to four test subjects, and that even running a single test subject is beneficial [Nielsen 1993b].

5. REVISING THE SKIMMING INTERFACE

Note again that the usability test was performed without any instruction or coaching of the subjects. It may be easy to fix most of the usability problems by modifying the touchpad template, or through a small amount of instruction.

5.1 Thoughts for Redesign based on the Usability Test
After establishing the basic system functionality, the touchpad template evolved quickly. Figure 16 shows three early templates as well as the one used in the usability test. The "sketch" in figure 17 shows a revised design that addresses many of the usability problems encountered, and incorporates the new features requested. The labels and icons are modified to be more consistent and familiar. Notably, "play" has replaced "normal," and "pauses removed" has replaced the confusing "no pause."

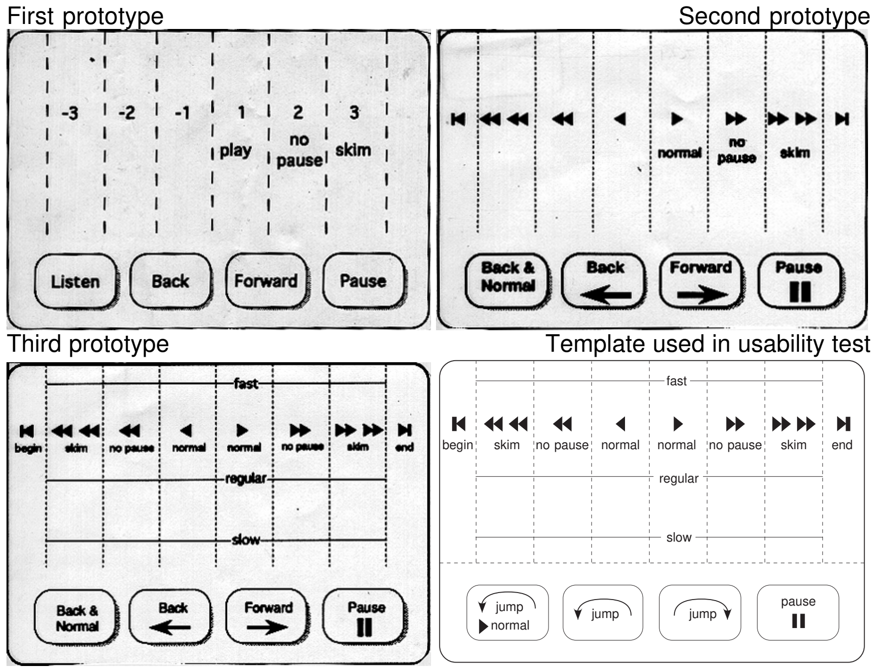

Fig. 16. Early evolution of SpeechSkimmer templates.

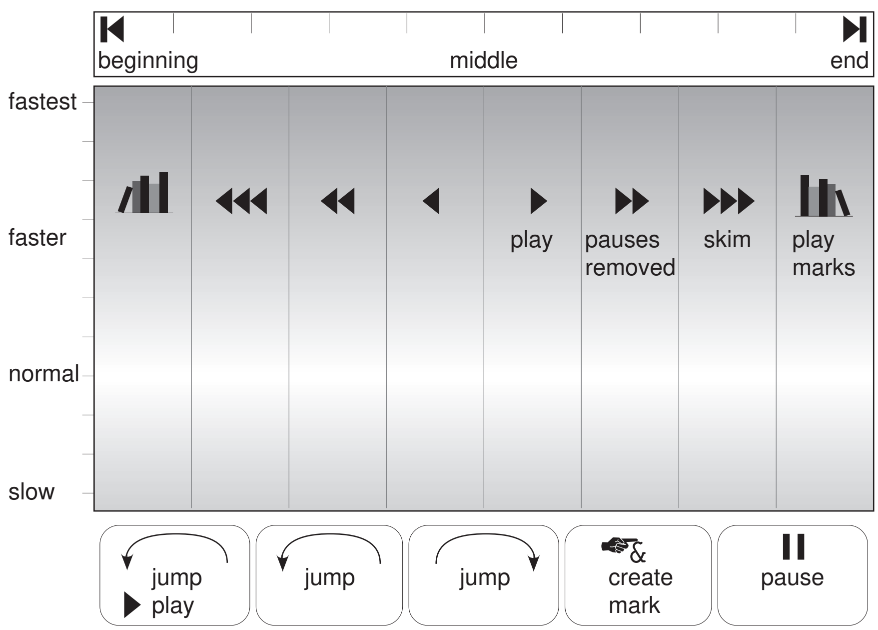

Fig. 17. Sketch of a revised template based on the usability results.

The speed labels are moved, renamed, and accompanied by tick marks to indicate a continuum of playback rates. The shaded background is an additional cue that the speeds extend across all levels. Colors, however, may be more effective than shading. For example, the slow-to-normal range could fade from blue to white, while the normal-to-fastest range could go from white to red, suggesting a cool-to-hot transition.

Bookmarks, as requested by the subjects, can be implemented in a variety of ways, but are perhaps best thought of as yet another level of skimming. In this case, however, the user interactively creates the list of speech segments to be played. In this design a “create mark” button is added along with new regions for playing the user defined segments.

A time line is added to directly access time points within a recording. It is located at the top of the template where subjects pointed when talking about this feature. The time line also naturally incorporates the begin and end buttons, removing them from the main portion of the template and out of the way of accidental activation.

The layout and graphic design of this template is somewhat cluttered, and the “jump and play normal” button remains problematic. However, the intuitiveness of this design, or alternative designs, could be quickly tested by asking a few subjects for their initial impressions.

One of the subjects commented that a physical control (such as real buttons and sliders) would be easier to use than the touchpad. Another approach to changing the physical interface is to use a jog and shuttle control, as is often found in video editing systems. Alternatively, a foot pedal could be used in situations where the hands are busy, such as when transcribing or taking notes.

5.2 A New Interface to SpeechSkimmer

A new user interface based on the results of the usability test, and the design sketched in section 5.1, was implemented using an Apple Newton MessagePad 100 as an input and output device. The MessagePad has a digitizing surface and a graphics display, so it can be used both as an input device, and for presenting status information. The touch sensitive surface works with a stylus or a fingernail (figure 18) rather than the tip of a finger as with the original touchpad. The MessagePad is rotated 90 degrees from its normal orientation into a landscape configuration to provide more screen real estate for the skimming controls.

Fig. 18. An early version of the MessagePad interface.

Why use a touchpad with a display instead of a traditional screen and mouse? A touchpad was originally selected as an input device so that the system could be used without looking at it, or while doing other things. While the MessagePad interface does display a small amount of status information, it can still be used without looking at it (especially when tactile strips are added to the surface). The MessagePad provides an input and display mechanism in a small portable package that is designed to be handheld.

5.2.1 Time Line for Input and Output

A time line (figure 19, top) is used both for displaying the current location within a recording, and for going to a particular time point. The current position in the speech recording is shown using a small vertical bar in the time line. This bar stays synchronized with the recording and moves slowly as the audio plays, acting as a percent done indicator [Myers 1985]. The time line can also be touched to jump to a specific point in the recording.

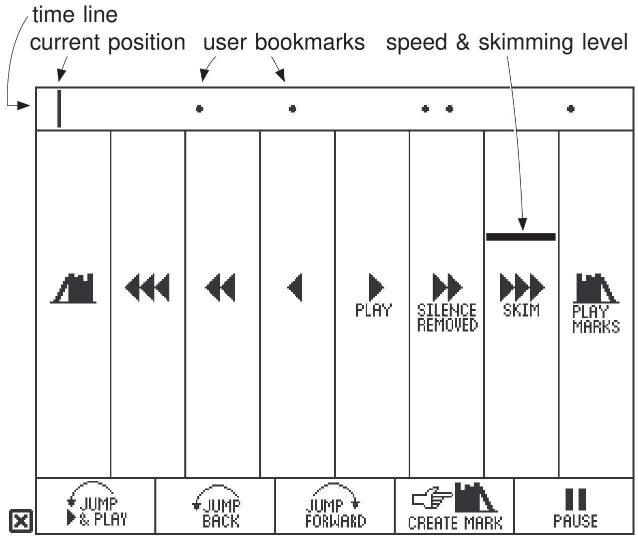

Fig. 19. Screen image from the MessagePad interface.

5.2.2 Bookmarks

A listener can set a personalized bookmark at any point in a recording by touching the "create mark" button (figure 19, bottom). This causes two events to happen. First, a visual indication for a bookmark (a small circle) is added to the time line, allowing the listener to get a sense of their location within the recording. Second, a new speech segment is added to SpeechSkimmer's internal representation of audio to be played at the highest skimming level.

The bookmarks can be accessed manually by touching the circles in the time line, or through the new "play marks" skimming level. This level plays only the segments selected by the user. Thus, this top skimming level represents a user-defined summary of the recording.

5.2.3 Display of Skimming Level and Speed

The playback speed is set by sliding a finger up or down in one of the vertical regions of the MessagePad. The current skimming level and speed are visually indicated by a horizontal bar in one of the slider regions (figure 19). Note that as with the original touchpad interface only one of these virtual sliders can be selected at a time (i.e., only one speed and skimming level is active).

5.2.4 System Architecture

The MessagePad is connected to an Apple Macintosh computer that performs all speech processing and audio playback. A small amount of processing is done on the MessagePad to update the display and translate the raw coordinates from the digitizing tablet into higher level events that are sent to the Macintosh (possibly over a wireless infrared link). Ideally, the entire speech skimming system could be implemented on a portable device such as a MessagePad. However, current generation handheld computers and PDAs (personal digital assistants) have limitations in the areas of audio input and output, sound data storage, and software tools for managing and manipulating audio.

6. RELATED WORK

6.1 Speech Interfaces

SpeechSkimmer draws many of its ideas from earlier speech systems. While SpeechSkimmer automatically structures recordings from information in the speech signal, many of these predecessor systems structure audio through interaction with the user, placing some burden on the creator of the speech data.

Phone Slave [Schmandt 1984] and the Conversational Desktop [Schmandt 1985; Schmandt 1987] explored interactive message gathering, and speech interfaces to simple databases of voice messages. VoiceNotes [Stifelman 1993] investigated the creation and management of a self-authored database of short speech recordings. VoiceNotes investigated many of the user interface issues addressed by SpeechSkimmer in the context of a hand-held computer.

Hyperspeech is a speech-only hypermedia system that explores issues of speech user interfaces, browsing, and the use of speech as data in an environment without a visual display [Arons 1991b]. The system uses speech recognition input and synthetic speech feedback to aid in navigating through a database of recorded speech. Resnick designed several voice bulletin board systems accessible through a touch tone interface [Resnick 1992a; Resnick 1992b]. These systems addressed issues of navigation and provide shortcuts to "skip and scan" among speech recordings.

The "human memory prosthesis" was envisioned to run on a wireless notepad-style computer to help people remember things such as names and reconstruct past events [Lamming 1991]. Information gathered through a variety of sources, such as notetaking, permit jumping to time-stamped points in the audio (or video) stream. Filochat indexed an audio recording with pen strokes on an LCD tablet, allowing the notes to be used to access the recording [Whittaker 1994]. Stifelman's Audio Notebook also synchronizes handwritten notes with an audio recording [Stifelman 1996]. However, rather than writing on a computer screen, the notes are taken with an ink pen in a paper notebook providing a familiar interface. Both handwriting and page turns are used as indices into the audio. Moran et al. captured meetings and indexed them through several notetaking tools [Moran 1996]

6.2 Gisting and Skimming

Word spotting and gisting (obtaining the essence of a message) systems are appealing for summarizing and accessing messages, but have limited domains of applicability; the skimming techniques presented here do not use any domain-specific knowledge and will work across all topics.

Several systems have attempted to obtain the gist of a recording using keyword spotting [Wilcox 1992a; Wilpon 1990] in conjunction with syntactic and/or timing constraints in an attempt to broadly classify a message [Houle 1988; Maksymowicz 1990]. Rose's system takes speech messages and extracts the message category according to a pre-defined notion of topic [Rose 1991]. Similar work has been reported in the areas of retrieving speech documents [Glavitsch 1992; Brown 1996] and editing applications [Wilcox 1992b].

Maxemchuk suggested three techniques for skimming speech messages: using text descriptors for selecting playback points, jumping forward or backward in the message, and increasing the playback rate [Maxemchuk 1980]. Stevens [Stevens 1994] and Raman [Raman 1994] developed systems for reading and browsing math equations and structured documents with a text-to-speech synthesizer. These systems addressed issues of navigating in auditory documents and methods of presenting "auditory glances."

6.3 Segmentation and Display

Kato and Hosoya investigated several techniques to enable fast telephone-based message searching by breaking up messages on hesitation boundaries, and presented either the initial portion of each phrase or high energy segments [Kato 1992; Kato 1993]. Kimber et al. used speaker identification to segment audio recordings that could be browsed with a graphical interface [Kimber 1995]. See [Pfeiffer 1996] and [Hawley 1993] for attempts at automatically analyzing and structuring audio recordings.

Wolf and Rhyne present a method for reviewing meetings based on characteristics captured by a pen-based meeting support tool [Wolf 1992]. They found that turn categories of most interest for browsing meetings were preceded by longer gaps in writing than the other turn types. Several techniques for capturing and structuring office discussions, telephone conversations, and lengthy recordings are described in [Hindus 1993]. This work emphasized graphical representations of recordings in a workstation environment.

7. FUTURE WORK

In addition to the small amount of status information displayed on the MessagePad interface, the skimming system could also take advantage of a full graphical user interface for displaying information. Along with mapping the fundamental SpeechSkimmer controls to a mouse, it is possible to add a variety of visual cues, such as displaying a real-time version of the segmentation information (figure 7), to aid in the skimming process.

Video editing and display systems can also be used with a speech skimming interface. For example, when quickly browsing through a set of video images, only the high-level segments of speech could be played, rather than random snippets of audio associated with the displayed frames. Similarly, a SpeechSkimmer-like interface can be used to skim through the audio track while the related video images are synchronously displayed.

The automatic structuring of spontaneous speech is an important area for future work. Integrating multiple acoustic cues (e.g., pitch, energy, pause, speaker identification) will ultimately produce the most successful segmentation techniques. Word spotting can also be used to provide text tags or summaries for flexible information retrieval. Summarizing or gisting systems will advance as speech recognition technology evolves, but may be most useful when combined with the skimming ideas presented here.

8. CONCLUSION

Speech is naturally slow to listen to, and difficult to skim. This research attempts to overcome these limitations making it easier and more efficient to consume recorded speech. By combining techniques that extract structure from spontaneous speech, with a hierarchical representation and an interactive listener control, it is possible to overcome the time bottleneck in speech-based systems. When asked if the system was useful, one test subject commented “Yes, definitely. It’s quite nice, I would use it to listen to talks or lectures that I missed … it would be super, I would do it all the time. I don’t do it now since it would require me to sit through the duration of the two hour [presentations] …”

This paper presents a framework for thinking about and designing speech skimming systems. SpeechSkimmer allows “intelligent” filtering of recorded speech; the intelligence is provided by the interactive control of the human, in combination with the speech segmentation techniques. The fundamental mechanisms presented here allow other types of segmentation or new interface techniques to be easily plugged in. SpeechSkimmer is intended to be a technology that is incorporated into any interface that uses recorded speech, as well as a stand-alone application. Skimming techniques enable speech to be readily accessed in a range of applications and devices, empowering a new generation of user interfaces that use speech. When discussing the SpeechSkimmer system, one of the usability test subjects put it succinctly: “it is a concept, not a box.”

This research provides insight into making one’s ears as usable as one’s eyes as a means for accessing stored information. Tufte said “Unlike speech, visual displays are simultaneously a wideband and a perceiver-controllable channel” [Tufte 1990, 31]. This work attempts to overcome these conventional notions, increasing the information bandwidth of the auditory channel and allowing the perceiver to interactively access recorded information. Speech is a powerful medium, and its use in computer-based systems will expand in unforeseen ways when users can interactively skim, and efficiently listen to, recorded speech.

ACKNOWLEDGMENTS

Lisa Stifelman provided valuable input in user interface design, assisted in designing and conducting the many hours of the usability test, and helped edit this document. Chris Schmandt provided helpful feedback on the SpeechSkimmer system. Doug Reynolds ran his speaker identification software on my recording. Dorée Seligmann suggested using the MessagePad. Two of the anonymous reviewers helped me focus this paper. Many others have contributed to this work, and have been thanked elsewhere.

REFERENCES

Arons 1991a	Arons, B. Authoring and Transcription Tools for Speech-Based Hypermedia Systems. In *Proceedings of 1991 Conference,* American Voice I/O Society, Sep. 1991, 15–20.
Arons 1991b	Arons, B. Hyperspeech: Navigating in Speech-Only Hypermedia. In *Proceedings of Hypertext (San Antonio, TX, Dec. 15–18),* ACM, New York, 1991, 133–146.
Arons 1992a	Arons, B. Techniques, Perception, and Applications of Time-Compressed Speech. In *Proceedings of 1992 Conference,* American Voice I/O Society, Sep. 1992, 169–177.
Arons 1992b	Arons, B. Tools for Building Asynchronous Servers to Support Speech and Audio Applications. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST),* ACM SIGGRAPH and ACM SIGCHI, ACM Press, Nov. 1992, 71–78.
Arons 1994a	Arons, B. “Interactively Skimming Recorded Speech.” Ph.D. dissertation, MIT, 1994.
Arons 1994b	Arons, B. Pitch-Based Emphasis Detection for Segmenting Speech Recordings. In *Proceedings of International Conference on Spoken Language Processing (Yokohama, Japan, Sep. 18–22),* vol. 4, 1994, 1931–1934.
Beasley 1976	Beasley, D.S. and Maki, J.E. “Time- and Frequency-Altered Speech.” Ch. 12 in *Contemporary Issues in Experimental Phonetics*, edited by Lass, N.J., 419–458. New York: Academic Press, 1976.
Brown 1996	Brown, M.G., Foote, J.T., Jones, G.J.F., Jones, K.S., and Young, S.J. Open Vocabulary Speech Indexing for Voice and Video Mail Retrieval. In *Proceedings of ACM Multimedia 96 (Boston, MA, Nov. 18–22),* ACM, New York, 1996, 307–316.
Buxton 1991	Buxton, W., Gaver, B., and Bly, S. *The Use of Non-Speech Audio at the Interface*. ACM SIGGCHI. Tutorial Notes. 1991.
Card 1991	Card, S.K., Mackinlay, J.D., and Robertson, G.G. “A Morphological Analysis of the Design Space of Input Devices.” *ACM Transactions on Information Systems 9*, 2 (1991): 99–122.
Chen 1992	Chen, F.R. and Withgott, M. The Use of Emphasis to Automatically Summarize Spoken Discourse. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing,* IEEE, 1992, 229–233.
Davis 1993	Davis, M. Media Streams: An Iconic Visual Language for Video Annotation. In *IEEE/CS Symposium on Visual Languages,* Bergen, Norway: Aug. 1993.
Elliott 1993	Elliott, E.L. “Watch-Grab-Arrange-See: Thinking with Motion Images via Streams and Collages.” Master’s thesis, MIT, 1993. Media Arts and Sciences Section.
Ericsson 1984	Ericsson, K.A. and Simon, H.A. *Protocol Analysis: Verbal Reports as Data.* Cambridge, MA: MIT Press, 1984.
Fairbanks 1954	Fairbanks, G., Everitt, W.L., and Jaeger, R.P. “Method for Time or Frequency Compression-Expansion of Speech.” *Transactions of the Institute of Radio Engineers, Professional Group on Audio AU-2* (1954): 7–12. Reprinted in G. Fairbanks, Experimental Phonetics: Selected Articles, University of Illinois Press, 1966.
Foulke 1971	Foulke, E. “The Perception of Time Compressed Speech.” Ch. 4 in *Perception of Language*, edited by Kjeldergaard, P.M., Horton, D.L., and Jenkins, J.J., 79–107. Columbus, OH: Merrill, 1971.
Furnas 1986	Furnas, G.W. Generalized Fisheye Views. In *Proceedings of CHI (Boston, MA),* ACM, New York, 1986, 16–23.
Gaver 1989	Gaver, W.W. “Auditory Icons: Using Sound in Computer Interfaces.” *Human-Computer Interaction 2* (1989): 167–177.

Gerber 1977 — Gerber, S.E. and Wulfeck, B.H. “The Limiting Effect of Discard Interval on Time-Compressed Speech.” *Language and Speech 20*, 2 (1977): 108–115.

Glavitsch 1992 — Glavitsch, U. and Schäuble, P. A System for Retrieving Speech Documents. In *15th Annual International SIGIR ’92,* ACM, New York, 1992, 168–176.

Gould 1982 — Gould, J. “Writing and Speaking Letters and Messages.” *International Journal of Man-Machine Studies 16* (1982): 147–171.

Grosz 1986 — Grosz, B.J. and Sidner, C.L. “Attention, Intentions, and the Structure of Discourse.” *Computational Linguistics 12*, 3 (1986): 175–204.

Gruber 1982 — Gruber, J.G. “A Comparison of Measured and Calculated Speech Temporal Parameters Relevant to Speech Activity Detection.” *IEEE Transactions on Communications COM-30*, 4 (1982): 728–738.

Gruber 1983 — Gruber, J.G. and Le, N.H. “Performance Requirements for Integrated Voice/Data Networks.” *IEEE Journal on Selected Areas in Communications SAC-1*, 6 (1983): 981–1005.

Hawley 1993 — Hawley, M. “Structure out of Sound.” Ph.D. dissertation, MIT, 1993.

Heiman 1986 — Heiman, G.W., Leo, R.J., Leighbody, G., and Bowler, K. “Word Intelligibility Decrements and the Comprehension of Time-Compressed Speech.” *Perception and Psychophysics 40*, 6 (1986): 407–411.

Hejna 1990 — Hejna Jr, D.J. “Real-Time Time-Scale Modification of Speech via the Synchronized Overlap-Add Algorithm.” Master’s thesis, MIT, 1990. Department of Electrical Engineering and Computer Science.

Hindus 1993 — Hindus, D., Schmandt, C., and Horner, C. “Capturing, Structuring, and Representing Ubiquitous Audio.” *ACM Transactions on Information Systems 11*, 4 (1993): 376–400.

Hirschberg 1986 — Hirschberg, J. and Pierrehumbert, J. The Intonational Structuring of Discourse. In *Proceedings of the Association for Computational Linguistics,* 1986, 136–144.

Hirschberg 1992 — Hirschberg, J. and Grosz, B. Intonational Features of Local and Global Discourse. In *Proceedings of the Speech and Natural Language Workshop (Harriman, NY, Feb.23-26),* San Mateo, CA: Morgan Kaufmann Publishers, 1992, 441–446.

Houle 1988 — Houle, G.R., Maksymowicz, A.T., and Penafiel, H.M. Back-End Processing for Automatic Gisting Systems. In *Proceedings of 1988 Conference,* American Voice I/O Society, 1988.

Jeffries 1991 — Jeffries, R., Miller, J.R., Wharton, C., and Uyeda, K.M. User Interface Evaluation in the Real World: A Comparison of Four Techniques. In *Proceedings of CHI (New Orleans, LA, Apr. 28–May 2),* ACM, New York, Apr 1991, 119–124.

Kato 1992 — Kato, Y. and Hosoya, K. Fast Message Searching Method for Voice Mail Service and Voice Bulletin Board Service. In *Proceedings of 1992 Conference,* American Voice I/O Society, 1992, 215–222.

Kato 1993 — Kato, Y. and Hosoya, K. Message Browsing Facility for Voice Bulletin Board Service. In *Human Factors in Telecommunications ’93,* 1993, 321–330.

Kimber 1995 — Kimber, D., Wilcox, L., Chen, F., and Moran, T. Speaker Segmentation for Browsing Recorded Audio. In *CHI ’94 Conference Companion (Denver, CO, May 7–11),* ACM, New York, 1995, 212–213.

Lamel 1981 — Lamel, L.F., Rabiner, L.R., Rosenberg, A.E., and Wilpon, J.G. “An Improved Endpoint Detector for Isolated Word Recognition.” *IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-29*, 4 (1981): 777–785.

Lamming 1991 — Lamming, M.G. Towards a Human Memory Prosthesis. Xerox EuroPARC, technical report no. EPC-91-116 1991.

Lass 1977 — Lass, N.J. and Leeper, H.A. “Listening Rate Preference: Comparison of Two Time Alteration Techniques.” *Perceptual and Motor Skills 44* (1977): 1163–1168.

Levelt 1989 — Levelt, W.J.M. *Speaking: From Intention to Articulation.* Cambridge, MA: MIT Press, 1989.

Lipscomb 1993 — Lipscomb, J.S. and Pique, M.E. “Analog Input Device Physical Characteristics.” *SIGCHI Bulletin 25*, 3 (1993): 40–45.

Mackinlay 1991 — Mackinlay, J.D., Robertson, G.G., and Card, S.K. The Perspective Wall: Detail and Context Smoothly Integrated. In *Proceedings of CHI (New Orleans, LA, Apr. 28–May 2),* ACM, New York, 1991, 173–179.

Maksymowicz 1990 Maksymowicz, A.T. Automatic Gisting for Voice Communications. In *IEEE Aerospace Applications Conference,* IEEE, Feb. 1990, 103–115.

Maxemchuk 1980 Maxemchuk, N. “An Experimental Speech Storage and Editing Facility.” *Bell System Technical Journal 59*, 8 (1980): 1383–1395.

Microtouch 1992 Microtouch Systems Inc. *UnMouse User’s Manual*, Wilmington, MA. 1992.

Mills 1992 Mills, M., Cohen, J., and Wong, Y.Y. A Magnifier Tool for Video Data. In *Proceedings of CHI (Monterey, CA, May 3–7),* ACM, New York, Apr. 1992, 93–98.

Minifie 1974 Minifie, F.D. “Durational Aspects of Connected Speech Samples.” In *Time-Compressed Speech*, edited by Duker, S., 709–715. Metuchen, NJ: Scarecrow, 1974.

Moran 1996 Moran, T., Chiu, P., Harrison, S., Kurtenbach, G., Minneman, S., and van Melle, W. Evolutionary Engagement in an Ongoing Collaborative Work Process: A Case Study. In *Proceedings of the ACM 1996 Conference on Computer Supported Cooperative Work (Boston, MA, Nov. 16–20),* ACM, New York, 1996, 150–159.

Myers 1985 Myers, B.A. The Importance of Percent-Done Progress Indicators for Computer-Human Interfaces. In *Proc. of ACM CHI'85 Conf. on Human Factors in Computing Systems,* 1985, 11–17.

Neuburg 1978 Neuburg, E.P. “Simple Pitch-Dependent Algorithm for High Quality Speech Rate Changing.” *Journal of the Acoustic Society of America 63*, 2 (1978): 624–625.

Nielsen 1990 Nielsen, J. and Molich, R. Heuristic Evaluation of User Interfaces. In *Proceedings of CHI (Seattle, WA, Apr. 1–5),* ACM, New York, 1990.

Nielsen 1991 Nielsen, J. Finding Usability Problems through Heuristic Evaluation. In *Proceedings of CHI (New Orleans, LA, Apr. 28–May 2),* ACM, New York, Apr. 1991, 373–380.

Nielsen 1993a Nielsen, J. *Usability Engineering.* San Diego: Academic Press, 1993.

Nielsen 1993b Nielsen, J. “Is Usability Engineering Really Worth It?.” *IEEE Software 10*, 6 (1993): 90–92.

O’Shaughnessy 1987 O’Shaughnessy, D. Speech Communication: Human and Machine. Reading, MA: Addison-Wesley Publishing Company, Inc., 1987.

O’Shaughnessy 1992 O’Shaughnessy, D. Recognition of Hesitations in Spontaneous Speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing,* vol. I, IEEE, 1992, I521–I524.

Pfeiffer 1996 Pfeiffer, S., Fischer, S., and Effelsberg, W. Automatic Audio Content Analysis. In *Proceedings of ACM Multimedia 96 (Boston, MA, Nov. 18–22),* ACM, New York, 1996, 21–30.

Rabiner 1975 Rabiner, L.R. and Sambur, M.R. “An Algorithm for Determining the Endpoints of Isolated Utterances.” *The Bell System Technical Journal 54*, 2 (1975): 297–315.

Rabiner 1989 Rabiner, L.R. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.” *Proceedings of the IEEE 77*, 2 (1989): 257–286.

Raman 1994 Raman, T.V. “AsTeR: Audio System for Technical Readings.” Ph.D. dissertation, Cornell University, 1994.

Reich 1980 Reich, S.S. “Significance of Pauses for Speech Perception.” *Journal of Psycholinguistic Research 9*, 4 (1980): 379–389.

Resnick 1992a Resnick, P. and Virzi, R.A. Skip and Scan: Cleaning Up Telephone Interfaces. In *Proceedings of CHI (Monterey, CA, May 3–7),* ACM, New York, Apr. 1992, 419–426.

Resnick 1992b Resnick, P. “HyperVoice: Groupware by Telephone.” Ph.D. dissertation, MIT, 1992.

Reynolds 1995 Reynolds, D.A. and Rose, R.C. “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models.” *IEEE Transactions on Speech and Audio Processing 3*, 1 (1995): 72–83.

Roe 1993 Roe, D.B. and Wilpon, J.G. “Whither Speech Recognition: The Next 25 Years.” *IEEE Communications Magazine 31*, 11 (1993): 54–62.

Rose 1991 Rose, R.C. “Techniques for Information Retrieval from Speech Messages.” *The Lincoln Lab Journal 4*, 1 (1991): 45–60.

Roucos 1985 Roucos, S. and Wilgus, A.M. High Quality Time-Scale Modification for Speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing,* IEEE, 1985, 493–496.

Savoji 1989 Savoji, M.H. “A Robust Algorithm for Accurate Endpointing of Speech Signals.” *Speech Communication 8* (1989): 45–60.

Schmandt 1984 Schmandt, C. and Arons, B. “A Conversational Telephone Messaging System.” *IEEE Transactions on Consumer Electronics CE-30*, 3 (1984): xxi–xxiv.

Schmandt 1985 Schmandt, C., Arons, B., and Simmons, C. Voice Interaction in an Integrated Office and Telecommunications Environment. In *Proceedings of 1985 Conference,* American Voice I/O Society, 1985.

Schmandt 1987 Schmandt, C. and Arons, B. “Conversational Desktop (videotape).” *ACM SIGGRAPH Video Review 27* (1987): .

Scott 1967 Scott, R.J. “Time Adjustment in Speech Synthesis.” *Journal of the Acoustic Society of America 41*, 1 (1967): 60–65.

Silverman 1987 Silverman, K.E.A. “The Structure and Processing of Fundamental Frequency Contours.” Ph.D. dissertation, University of Cambridge, 1987.

Souza 1983 de Souza, P. “A Statistical Approach to the Design of an Adaptive Self-Normalizing Silence Detector.” *IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-31*, 3 (1983): 678–684.

Stevens 1994 Stevens, R. and Edwards, A. Mathtalk: The Design of an Interface for Reading Algebra Using Speech. In *Computers for Handicapped Persons: Proceedings of ICCHP '94, Lecture Notes in Computer Science 860,* Springer-Verlag, 1994, 313–320.

Stifelman 1992 Stifelman, L. *A Study of Rate Discrimination of Time-Compressed Speech*, Speech Research Group Technical Report, Media Laboratory, 1992.

Stifelman 1993 Stifelman, L.J., Arons, B., Schmandt, C., and Hulteen, E.A. VoiceNotes: A Speech Interface for a Hand-Held Voice Notetaker. In *Proceedings of INTERCHI (Amsterdam, The Netherlands, Apr. 24–29),* ACM, New York, 1993, 179–186.

Stifelman 1995 Stifelman, L.J. A Discourse Analysis Approach to Structured Speech. In *AAAI Spring Symposium Series. Empirical Methods in Discourse Interpretation and Generation (Stanford, CA),* 1995, 162–167.

Stifelman 1996 Stifelman, L. Augmenting Real-World Objects: A Paper-Based Audio Notebook. In *CHI ’96 Conference Companion (Vancouver, BC, Apr. 13–88),* ACM, New York, 1996, 199–200.

Tufte 1990 Tufte, E. *Envisioning Information.* Cheshire, CT: Graphics Press, 1990.

Whittaker 1994 Whittaker, S., Hyland, P., and Wiley, M. Filochat: Handwritten Notes Provide Access to Recorded Conversations. In *Proceedings of CHI (Boston, MA, Apr. 24–28),* SIGCHI, ACM, New York, 1994, 271–277.

Wilcox 1992a Wilcox, L. and Bush, M. Training and Search Algorithms for an Interactive Wordspotting System. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing,* IEEE, 1992.

Wilcox 1992b Wilcox, L., Smith, I., and Bush, M. Wordspotting for Voice Editing and Audio Indexing. In *Proceedings of CHI (Monterey, CA, May 3–7),* ACM, New York, 1992, 655–656.

Wilpon 1990 Wilpon, J.G., Rabiner, L.R., Lee, C., and Goldman, E.R. “Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models.” *IEEE Transactions on Acoustics, Speech, and Signal Processing 38*, 11 (1990): 1870–1878.

Wolf 1992 Wolf, C.G. and Rhyne, J.R., Facilitating Review of Meeting Information Using Temporal Histories, IBM T. J. Watson Research Center, Working Paper 9/17. 1992.