Scalable Summaries of Spoken Conversations

Sumit Basu[1], Surabhi Gupta[1,2], Milind Mahajan[1], Patrick Nguyen[1], and John C. Platt[1]

[1]Microsoft Research
One Microsoft Way
Redmond, WA 98052
sumitb@microsoft.com

[2]Stanford University
353 Serra Mall, Gates Building
Stanford, CA 94305
surabhi@cs.stanford.edu

ABSTRACT

In this work, we present a novel means of browsing recorded audio conversations. The method we develop produces *scalable summaries* of the recognized speech, in which we can increase the amount of text continuously with the desired level of detail to best fill the available space. We present an interface in which a user can view an entire conversation in one screen, but can also quickly zoom in to see the full transcript; the corresponding audio can be easily played as well. The scaling is achieved via a combination of topic segmentation and informative phrase selection, where the threshold for informativeness decreases with increasing level of detail. Finally, we evaluate our method and interface against a baseline interface with a user study.

Author Keywords

Speech Summarization, Conversation Summarization, Speech Browsing, Conversation Browsing, Zoomable User Interfaces.

ACM Classification Keywords

H.4.3 [Commuications Applications] – *Information Browsers,*. I.2.7 [Natural Language Processing] – *Text Analysis*, H.5.1 [Multimedia Information Systems] – *Audio Input/Output*, H.5.2 [User Interfaces] – *Graphical User Interfaces*.

INTRODUCTION

Many of the interactions of modern information workers occur in face-to-face meetings, from hallway conversations to formal meetings. Unfortunately, much of the information exchanged in these meetings is lost - while people may take a few notes if pen and paper are handy, vast quantities of detail are forgotten within hours. However, audio recordings are becoming increasingly prevalent - it is becoming far easier to record individual meetings or conversations due to the miniaturization and availability of recording equipment. Many cellphones, for instance, can now make reasonable voice recordings.

The difficulty, of course, is making use of this data - scanning through conversations with a fast-forward button is rarely a good use of an information worker's time. In this work, we present users with a means to quickly get an overview of the content of conversations as well as drill down into specific details where they are interested.

While we think this work will be useful to a broad audience in the long term, there are users in specific job roles for which this could be useful today: journalists, patent attorneys, ethnographers, psychiatrists - basically, all individuals who need to conduct and review conversations as a core part of their work. While our results in this initial work show modest benefits on the relatively short audio segments used in our study (15 minutes each), we expect the advantages of our method will only increase with longer documents.

BACKGROUND

Summarization can be broadly defined as trying to give a shorter, more condensed version of some original document while also preserving meaning. Summarizing speech is far more difficult than traditional text media due to recognition errors, the lack of sentence boundaries, or any other kinds of document cues (paragraph/section boundaries, headings, etc.). Conversations make this even more difficult since there are multiple speakers with (often) unknown speaker changes. The recognizer output is simply a stream of words with no sentences or punctuation, making many traditional NLP techniques such as parsing quite difficult to apply. A small example of the data that we worked with is shown in Figure 1 below.

```
HAVE ALL THE BITS TOGETHER IN ONE PLACE I
WOULD SEEM TO BE REASONABLY REGRETS I MEAN
AND YOU KNOW MADE SENSE SO THIS IS A GOOD
READ ON WHAT'S A REASONABLE INDEED THAT'S
WHAT I JUST FINISHED READING WAS DRY KILL
A I WAS IN UH WHAT WAS THE BRANDS YOU KNOW
THERE YEAH YEAH WE WITH HIM YEAH YEAH I
WAS NEW ORIGINAL UH YEAH I'M NOT A GOOD UH
UH WHAT ANY GOOD OR WAS IT JUST SO
DIFFERENT THAN THAN WHAT THEY THEY WRITE
THESE DAYS
```

Figure 1: Typical output of the speech recognizer for conversational data

In this paper, we present a method and interface for viewing summaries of spoken conversations (or speech documents in general) that have an interactive level of detail. The method

IUI'08, January 13-16, 2008, Maspalomas, Gran Canaria, Spain.
Copyright 2008 ACM 978-1-59593-987-6/ 08/ 0001 $5.00

segments the conversation into topics and shows key words and phrases for the topic arranged in the time order that they appear. We have designed the method for conversations that the user has been a participant in or has otherwise heard already; thus relevant keywords should act as meaningful landmarks as they scan through the text. At the top level view, only a few keywords are visible for each topic; as the user zooms in, she can see more and more words until she sees almost a full transcript.

There has been much prior work on text summarization but far less on speech summarization; the contribution of our work to the state of the art is (1) the scalable nature of the summary (2) the means by which we pick the phrases to be shown (3) the interface which allows the user to examine the speech document at a controllable level of detail.

RELATED WORK

There has been a fair amount of work in the last decade on summarizing speech and audio documents; nearly all of this has centered around broadcast domains such as news. Hirschberg has a nice tutorial which explains the various approaches to speech summarization and their differences/similarities to text summarization in [7]; Zechner [19] also has a good survey of the area. In the interests of space, we will only summarize a few of the major approaches.

Most of the methods thus far have involved first producing a transcript and then working with the text along with some auditory cues. The work of Christensen et al. [4] was developed for news stories and assumes the first few sentences (based on sentence boundary detection) are topic related; it then finds a set of sentences that have high similarity with the topic and low similarity with each other (to encourage diversity) as the summary. Koumpis and Renals [10] employed a classification approach in which they use a combination of lexical and prosodic features to train classifiers at a word level to decide which words should be used in the summary. He et al. [5] involved the users' input in creating summaries of presentations: they used the logs of what content users retrieved as a feature for summarization, which did as well as linguistic/acoustic features. Hori and Furui [8] treated the summarization of the audio via a recognizer like a language translation task, where they scored each word based on its topic significance, linguistic significance (word probability), recognizer confidence, and transition cost, then decoded for a best solution.

There is some recent work by Maskey and Hirschberg [14] which has attempted to bypass the recognition step altogether and produces a summary directly from the audio signal. Their approach was to train classifiers to label the importance of words/sentences using only acoustic features; they found they could make significant gains over a baseline system using acoustic features alone. They were able to achieve even greater performance using a combination of acoustic and lexical cues.

Beyond this work on broadcast and presentation domains, there has been more recent interest in conversational speech recognition and summarization, particularly in the context of meetings – while different from spontaneous conversations, this is far closer to our domain than news. The closest work we are aware of to ours is the DiaSumm system [18], which creates static (fixed scale) summaries of spoken conversations in the meeting domain, using human-generated transcripts. The system could in principle be used on speech recognition data, but the reported results are on manual transcripts with hand-segmented turns. DiaSumm automatically finds topic boundaries in the text; it then uses the cumulative TFIDF (see Section 4) of all words from each turn to rank them against each other. The summary consists of the set of full turns per topic that have the highest TFIDF scores.

Our work differs from these past systems in several ways. First and most importantly, our summaries are scalable and our interface allows for a continuous variation of the level of detail to allow both broad overviews and detailed investigation. Second, we present informative words and phrases instead of entire turns, since we do not assume that it will be possible to accurately find turn information (i.e., the recordings may be monophonic). Finally, our approach is fully automatic and runs directly on the audio file and its speech recognition output. However, while our segmentation approach does make use of prosodic cues, our summarization mechanism does not, and based on the clear benefits shown in the past work this is an area of future work for our summarizer.

Finally, we would like to acknowledge the work of Ben Bederson and his colleagues in the user interface community who have developed the concept of the "Zoomable User Interface," or ZUI (see [1] for their initial work; [9] contains a brief survey of zoomable interfaces). Their pioneering research has explored how zooming could be used as an effective paradigm of helping users navigate large amounts of information by allowing them to see information at different levels of detail. Our approach is inspired by their work, as our domain has the kinds of complexity that led to their efforts. In short, spoken conversations are long, difficult to navigate, and tremendously dense in terms of content; as such, we expected a means for users to be able to move smoothly from an overview or "bird's eye view" to a full transcript could greatly help with browsing and finding information.

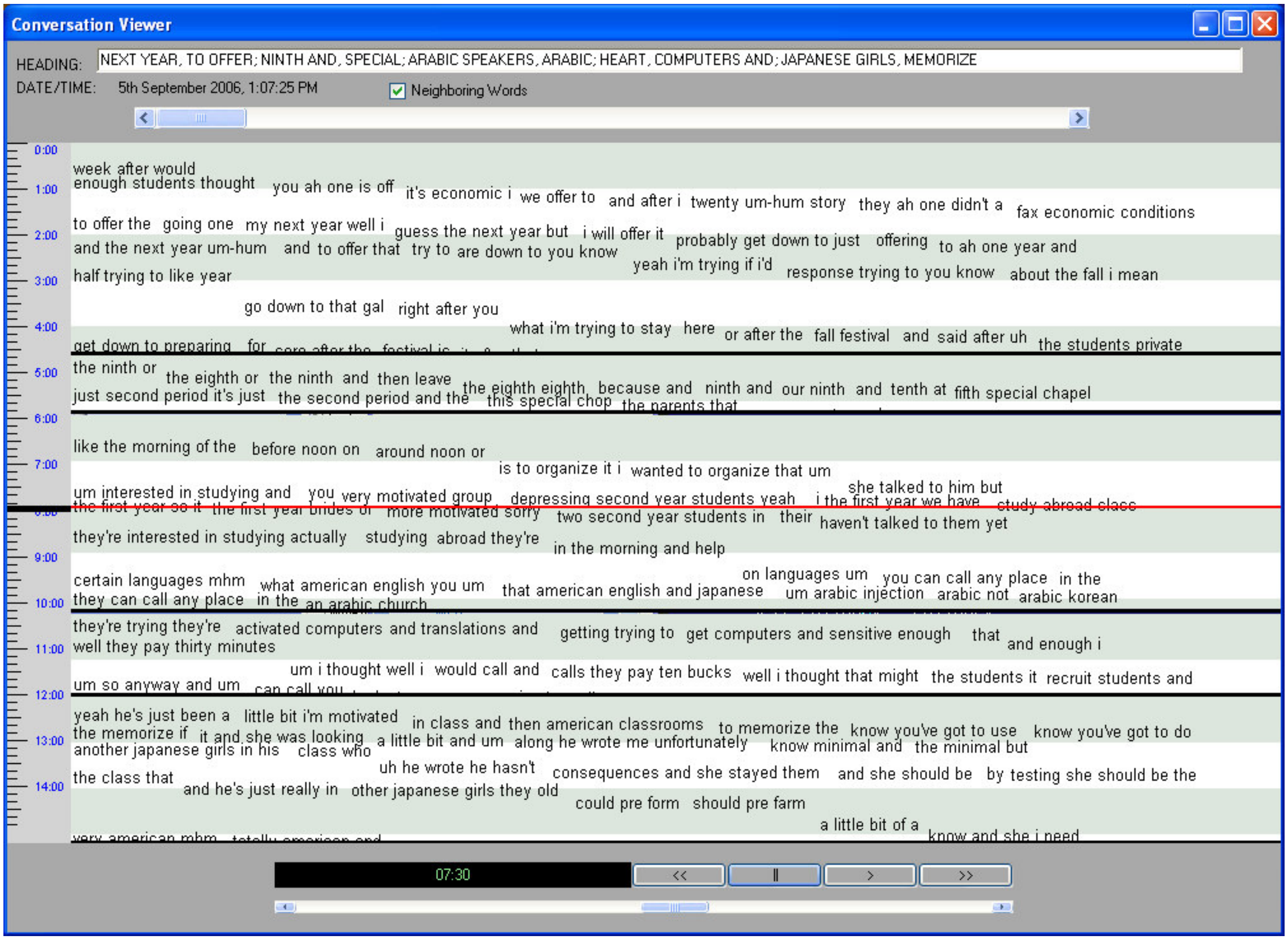
OUR APPROACH

Our approach consists of three components: topic segmentation, scalable key phrase extraction, and the visualization/navigation interface, which we detail in the subsections below.

All of these stages require the speech recognition output; for this purpose we use a baseline system trained on 2000 hours of data (the LDC Switchboard and Fisher corpora, both of which contain conversational speech on telephone channels). It achieves 75% word accuracy on the Fisher corpus, which is on par with other baseline recognizers on this corpus (see, e.g., [12]), and requires 5-6 times real-time to analyze a given audio file.

1.1 TOPIC SEGMENTATION

We tried a variety of standard approaches to topic segmentation, but due to the many recognition errors and lack of sentence/turn boundaries, all achieved fairly poor performance. The classic work in this space is Hearst's TextTiling algorithm [6], which builds a score metric for segmentation based on hand-crafted lexical features. However, we wanted to train the model against data instead of relying on fixed thresholds, as she had optimized it for text corpora. Maybury [15] developed an approach for automatically segmenting broadcast news based on hand-coded models of discourse cues and story structure, which was closer to our domain since it was designed for speech data, but was highly dependent on the news format (intro, overview, interview, etc.) and also not trainable. We also examined the work of Ries [16] on segmenting recognized text from meetings, but this required



HEADING: NEXT YEAR, TO OFFER; NINTH AND, SPECIAL; ARABIC SPEAKERS, ARABIC; HEART, COMPUTERS AND; JAPANESE GIRLS, MEMORIZE

DATE/TIME: 5th September 2006, 1:07:25 PM

week after would
enough students thought you ah one is off it's economic i we offer to and after i twenty um-hum story they ah one didn't a fax economic conditions
to offer the going one my next year well i guess the next year but i will offer it probably get down to just offering to ah one year and
and the next year um-hum and to offer that try to are down to you know yeah i'm trying if i'd response trying to you know about the fall i mean
half trying to like year
go down to that gal right after you
what i'm trying to stay here or after the fall festival and said after uh the students private
get down to preparing for
the ninth or the eighth or the ninth and then leave the eighth eighth because and ninth and our ninth and tenth at fifth special chapel
just second period it's just the second period and the this special chop the parents that
like the morning of the before noon on around noon or
is to organize it i wanted to organize that um
um interested in studying and you very motivated group depressing second year students yeah she talked to him but i the first year we have study abroad class
the first year so it the first year brides of more motivated sorry two second year students in their haven't talked to them yet
they're interested in studying actually studying abroad they're in the morning and help
certain languages mhm what american english you um that american english and japanese on languages um you can call any place in the
they can call any place in the an arabic church um arabic injection arabic not arabic korean
they're trying they're activated computers and translations and getting trying to get computers and sensitive enough that and enough i
well they pay thirty minutes
um i thought well i would call and calls they pay ten bucks well i thought that might the students it recruit students and
um so anyway and um can call you
yeah he's just been a little bit i'm motivated in class and then american classrooms to memorize the know you've got to use know you've got to do
the memorize if it and she was looking a little bit and um along he wrote me unfortunately know minimal and the minimal but
another japanese girls in his class who
uh he wrote he hasn't consequences and she stayed them and she should be by testing she should be the
the class that and he's just really in other japanese girls they old
could pre form should pre farm
a little bit of a know and she i need
very american mhm totally american and

Figure 2: Our conversation browsing interface shown at the top level of zoom, i.e., zoomed-out. At this view, the entire conversation can be seen in one screen. The phrases shown are those that have the highest information content as described in the paper. The vertical position of each phrase corresponds to the time it began in the audio. The black lines are topic boundaries; the alternating horizontal bars are at one-minute intervals. The zoom in/out bar is at the top of the screen. The red line indicates the current position of the audio playback; the zoom control expands around this point. There is an audio playback control at the bottom, as well as a timeline to the left with an indicator of what is being played. The user can also click on any word or the timeline to play the audio starting at that point.

training via manual transcriptions that would then be labeled according to his definition of topic boundaries. Furthermore, he found very low inter-annotator agreement (k=.35) and fairly poor overall performance using such boundaries due to the difficulty of the labeling task.

We thus decided to instead train a topic detection system on news data, where the topic boundaries would be clear, and then apply the result to conversational data, in the hopes that the acoustic and some textual cues would still perform in the new domain. We were also careful not to include features which would be overly specific to the news domain (story structure, etc.). We trained a log-linear model for the probability of any given point being a topic boundary, using word distribution features as well as particular keywords: this is also similar to the work of Beeferman et al. [2]. We also used acoustic cues such as pauses in addition to the textual features, as in the work of Hsu et al. [10]. Setting a threshold on this probability then allowed us to control how many segments would be found; we used a fixed threshold for all of our experiments. A final stage then used heuristic constraints to remove segments that were too short by specifying a minimum segment length.

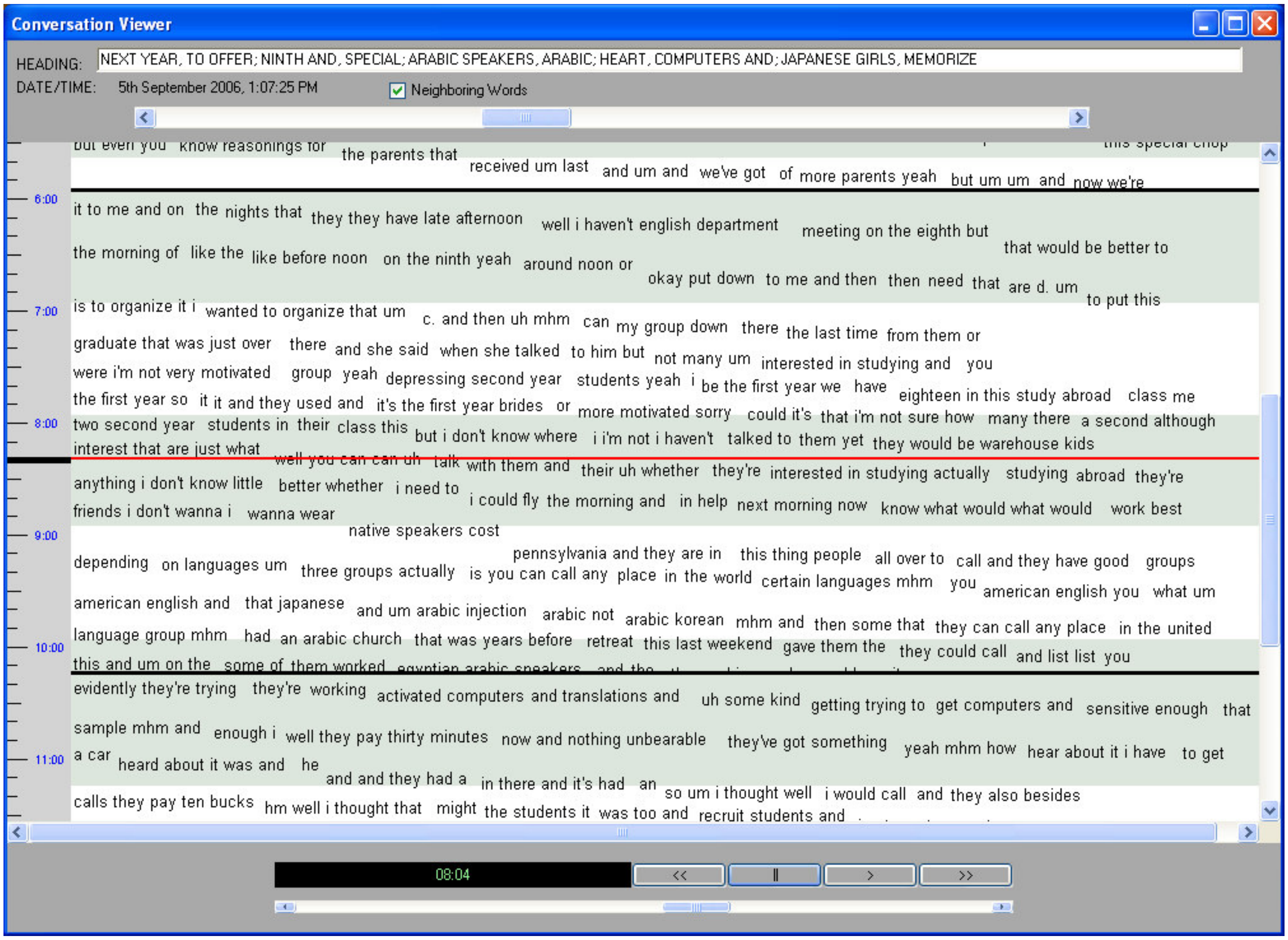
We found that the resulting segmentations were quite effective, both from our own inspection and from the feedback of our subjects, though we have not yet formally evaluated their correctness. In general, we found that we would see some topic boundaries being missed, but oversegmentation was very rare.

1.2 SCALABLE KEY PHRASE EXTRACTION

Given the topic boundaries, we next had to find the relevant keywords given a particular level of detail. We began by finding the TFIDF, or Term-Frequency Inverse Document Frequency measure [17], of each unigram and bigram in a given topic segment. The TFIDF for word w in document d from corpus c is defined as follows:

$$TFIDF(w,d,c) = TF(w,d) * IDF(w,c)$$

where



HEADING: NEXT YEAR, TO OFFER; NINTH AND, SPECIAL; ARABIC SPEAKERS, ARABIC; HEART, COMPUTERS AND; JAPANESE GIRLS, MEMORIZE

DATE/TIME: 5th September 2006, 1:07:25 PM ☑ Neighboring Words

but even you know reasonings for the parents that received um last and um and we've got of more parents yeah but um um and now we're this special chop

6:00 it to me and on the nights that they they have late afternoon well i haven't english department meeting on the eighth but that would be better to
the morning of like the like before noon on the ninth yeah around noon or okay put down to me and then then need that are d. um to put this

7:00 is to organize it i wanted to organize that um c. and then uh mhm can my group down there the last time from them or
graduate that was just over there and she said when she talked to him but not many um interested in studying and you
were i'm not very motivated group yeah depressing second year students yeah i be the first year we have eighteen in this study abroad class me
the first year so it it and they used and it's the first year brides or more motivated sorry could it's that i'm not sure how many there a second although

8:00 two second year students in their class this but i don't know where i i'm not i haven't talked to them yet they would be warehouse kids
interest that are just what well you can can uh talk with them and their uh whether they're interested in studying actually studying abroad they're
anything i don't know little better whether i need to i could fly the morning and in help next morning now know what would what would work best
friends i don't wanna i wanna wear native speakers cost

9:00 depending on languages um three groups actually pennsylvania and they are in is you can call any place in the world this thing people certain languages mhm all over to you call and they have good american english you groups what um
american english and that japanese and um arabic injection arabic not arabic korean mhm and then some that they can call any place in the united
language group mhm had an arabic church that was years before retreat this last weekend gave them the they could call and list list you

10:00 this and um on the some of them worked egyptian arabic speakers and the
evidently they're trying they're working activated computers and translations and uh some kind getting trying to get computers and sensitive enough that
sample mhm and enough i well they pay thirty minutes now and nothing unbearable they've got something yeah mhm how hear about it i have to get

11:00 a car heard about it was and he and and they had a in there and it's had an so um i thought well i would call and they also besides
calls they pay ten bucks hm well i thought that might the students it was too and recruit students and

Figure 3: Our conversation browsing interface at an intermediate level of zoom. Note that the level of detail has expanded to use the available visual space with importance-ranked phrases. The user can move forward and backward in time using the vertical scroll bar, the mousewheel, or by dragging on the canvas.

$$TF(w,d) = \left|\{t_j : t_j = w\}\right| / |d|$$

and

$$IDF(w,c) = \log(|c| / |\{d_k : w \in d_k\}|)$$

In other words, TF refers to the fraction of the terms t in document d that are the word in question, w, and IDF is the *inverse* of the fraction of documents in c that contain at least one instance of w. Intuitively, the TFIDF is a measure of the importance of a term: the TF term represents how frequent the word is in the document, while the IDF term balances that with how common the word is overall; that way a common word like "but" will have a low TFIDF even though it may occur many times in a document, while a less common term like "digestion" can have a high score with only a few occurrences. Note that in our context, the document is the relevant segment of the conversation, and the corpus is the Fisher corpus of conversational speech.

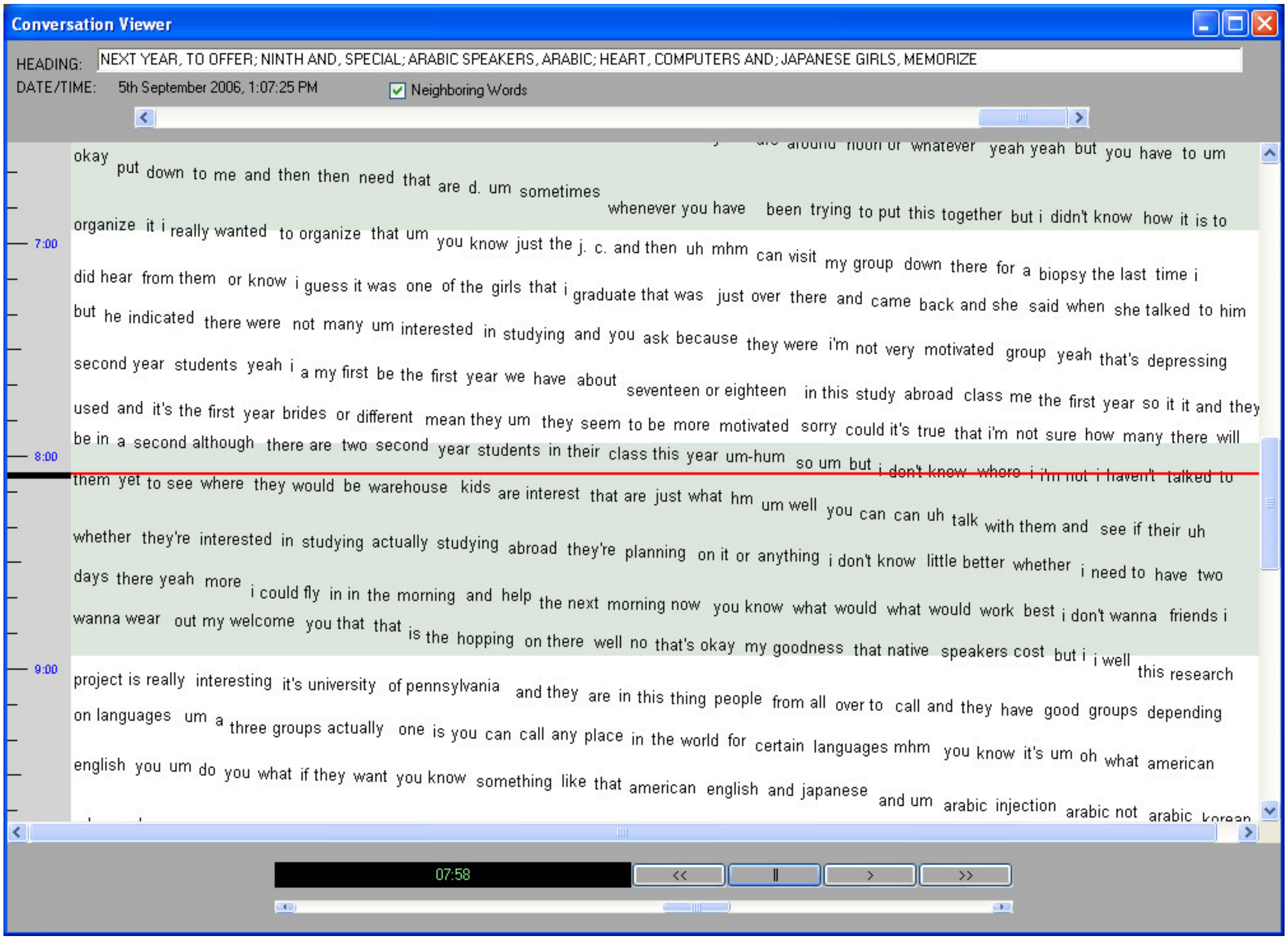
To enable our scalable summaries, we needed a ranked list of unigrams and bigrams in descending TFIDF order. To do this, we first had to normalize the bigram scores against the unigrams. In general, a given bigram is geometrically less likely to occur than a given unigram (i.e., if a word X and word Y each appear with probability 1/N, the sequence XY will occur with probability 1/N^2). Thus for independent terms, the TF of each unigram in the bigram (if they were equally probable) would be the square root of the bigram's TF, while the IDF of each unigram would be half the IDF of the bigram (due to the logarithm). While the terms in the bigram are not truly independent, in practice this scales the data appropriately. We thus take the square root of each bigram's TF and halve its IDF to bring it to the same scale as the unigrams and use these values to create a single, sorted list.

Given this list, we can move our TFIDF threshold to get as many or as few keywords as we want for each segment: we will describe how we make use of this capability in the interface in the sections below.

1.3 CONVERSATION BROWSING INTERFACE

The core of our work is the conversation browsing interface: all of the analysis above is in the interests of being able to quickly browse and find information in the conversation with this tool. We show screenshots of the interface in Figures 2, 3, and 4.

The interface can be zoomed from a minimum point at which the whole conversation is seen in one screen ("zoomed out," Figure 2) to a maximum where the entire transcript is being shown



Conversation Viewer

HEADING: NEXT YEAR, TO OFFER; NINTH AND, SPECIAL; ARABIC SPEAKERS, ARABIC; HEART, COMPUTERS AND; JAPANESE GIRLS, MEMORIZE

DATE/TIME: 5th September 2006, 1:07:25 PM

☑ Neighboring Words

are around noon or whatever yeah yeah but you have to um
okay put down to me and then then need that are d. um sometimes
whenever you have been trying to put this together but i didn't know how it is to
organize it i really wanted to organize that um you know just the j. c. and then uh mhm can visit my group down there for a biopsy the last time i
7:00
did hear from them or know i guess it was one of the girls that i graduate that was just over there and came back and she said when she talked to him
but he indicated there were not many um interested in studying and you ask because they were i'm not very motivated group yeah that's depressing
second year students yeah i a my first be the first year we have about seventeen or eighteen in this study abroad class me the first year so it it and they
used and it's the first year brides or different mean they um they seem to be more motivated sorry could it's true that i'm not sure how many there will
be in a second although there are two second year students in their class this year um-hum so um but i don't know where i i'm not i haven't talked to
8:00
them yet to see where they would be warehouse kids are interest that are just what hm um well you can can uh talk with them and see if their uh
whether they're interested in studying actually studying abroad they're planning on it or anything i don't know little better whether i need to have two
days there yeah more i could fly in in the morning and help the next morning now you know what would what would work best i don't wanna friends i
wanna wear out my welcome you that that is the hopping on there well no that's okay my goodness that native speakers cost but i i well this research
9:00
project is really interesting it's university of pennsylvania and they are in this thing people from all over to call and they have good groups depending
on languages um a three groups actually one is you can call any place in the world for certain languages mhm you know it's um oh what american
english you um do you what if they want you know something like that american english and japanese and um arabic injection arabic not arabic korean

07:58

Figure 4: Our conversation browsing interface at maximum zoom (zoomed-in all the way). At this point, the entire recognition transcript is shown. Note the many errors in the speech recognizer output.

("zoomed in," Figure 4). At each value of zoom, we go through the ranked list of phrases described above. For each phrase, we mark every occurrence within that topic for rendering, as well as its left and right neighboring words to give the user additional context. We then move on to the next term in the ranked list, and so on, until we run out of space. Thus, at each level of zoom, we show as many key phrases per topic as possible in priority order, given the available space.

The amount of space occupied by the words can be determined from the set of marked words i:

$$A_w = \frac{\sum_i width(w_i)}{h}$$

where A_w is the area occupied in pixels by the words, h is the text height, and w_i refers to the i^{th} word. However, it would not be prudent to pack the words in as tightly as possible, particularly when the view is zoomed out, since adjacent marked phrases may have little to do with each other. We thus use an exponential characteristic in filling the available space, so as to fill the space tightly when we are completely zoomed in, but remain relatively sparse at low and intermediate levels:

$$A_w^* \propto \exp(l / l_{\max}) * A_s$$

where A_s is the amount of available screen area, and l and l_{max} refer to the current and maximum zoom levels. The minimum zoom is set to be the entire document so that it can be viewed in a single screen. The maximum zoom level, on the other end of the spectrum, is fixed to a set number of seconds per pixel based on human speaking rates, as opposed to a fraction of the actual document – in this way we remain independent of document length, and it is only the effective step size of the zoom control that changes.

To render the keywords to the interface, we had to choose a balance between consistency and legibility over all zoom levels. Since adjacent keywords could have little to do with each other at low levels of zoom, we wanted to make the time that separated them explicit. However, at high levels of zoom, adjacent keywords will come from adjacent positions in the transcript, and we want the user to be able to read through this in a natural, left-right manner. We achieved this by setting the vertical position of each phrase by the time it occurs in the audio, and the horizontal position as directly to the right of the last rendered phrase. At low levels of zoom (zoomed out), the keywords are thus well separated due to the vertical spacing (see Figure 2); at high levels of zoom, the transcript can be read off left-to-right (Figure 4).

This disadvantages of this approach are that a given word is not at the same horizontal position over all zoom levels, and that even at the highest level of zoom the text will have a diagonal aspect to it, which several subjects complained about being somewhat harder to read. However, we tried several other approaches in our pilot studies, including fixed horizontal positions for words, and the reported approach proved to be the best compromise.

1.4 DESIGN AND USABILITY OF THE INTERFACE

We went through a variety of design iterations for our interface that we tested in many pilot experiments; the version we show here represents what has proved to work best from these experiences.

Figures 2, 3, and 4 show the interface in action at levels of minimum, intermediate, and maximal zoom. The phrases shown are those that have the highest information content as described above. Note that the phrases themselves are not offset from the contextual words by color/font/etc., as that proved to be too distracting to our pilot users.

The vertical position of each phrase corresponds to the time it began in the audio. The black lines are topic boundaries; the alternating horizontal bars are at one-minute intervals. Initially, we used the background color to mark the current segment, but as segments could be quite long, at high levels of zoom the background would be of a uniform color. This would disorient users as the background would remain this single color as they zoomed in and out. The alternating bars instead provide a strong visual indicator of the level of zoom and significantly reduced the disorientation effect. Another possibility would be to color code each turn, assuming turns can be estimated with sufficient accuracy.

The zoom in/out bar is at the top of the screen. The red line indicates the current position of the audio playback and stays current as the audio progresses; the zoom control expands around this point. There is an audio playback control at the bottom with play/pause/skip/etc., as well as a timeline to the left with an indicator of what is being played. The user can also click on any word or the timeline to play the audio starting at that point. Finally, the scroll bar to the right allows for moving forward and backward in time; the user can also move using the mouse scroll wheel or by dragging on the canvas.

EVALUATION

To evaluate our interface, we designed an information retrieval task and compared our scalable interface against a baseline interface that had the same capabilities except for zooming: it simply showed the entire transcript (see Figure 6).

We chose the first half (15 min. each) of two different conversations from the LDC Callhome (American English) database [3], conversations 4112 and 4074, and had the users listen to them on their own. Three to five days later, they used the two interfaces to answer six questions for each audio file. We had ten subjects with varying degrees of technical proficiency: all were familiar with computers, but some were "power users" while some were more casual users. They ranged in job roles from computer scientist to reporter, and ranged in age from mid-twenties to mid-fifties. We randomized all relevant variables: order of presentation of the interfaces, which interface was used for which file, and the order of the questions. Note that we wrote the questions before we had used the interface or seen the transcript for these files, to prevent biasing for the keywords that would be extracted. Furthermore, the questions were designed using different words than those used in the audio as much as possible to force the users to search for content and not words in the questions. For instance, one conversation referred to "burglaries" and things that were "stolen" from one speaker's home. The relevant question asked, "What was taken from the speaker's home in the first break-in?" Furthermore, because of the many recognition errors, most of the questions could not be answered using the recognized text alone; it was typically necessary for the user to find the relevant section and then listen to the audio.

To gather the subject data, we took a formal approach and used an isolated experimentation room where we could observe the subjects' behavior without being in the room via a one-way mirror. Before each segment of the study, the subjects were given an explanation of the interface they were about to use, and had a chance to use each of the navigation controls described above on a separate audio file that was not part of the study.

Once familiar with the interface, the users were instructed to imagine that they were reporters - even if they knew the answer to a question off the top of their heads, they had to find the relevant location in the audio file so that they could get a quote for the article they were writing. When they found the answer, they typed it into the application displaying the questions, which would also log the time it took for them to answer each question.

Afterwards, we interviewed the subjects to learn about their impressions about both interfaces; we report on this in the following section.

RESULTS

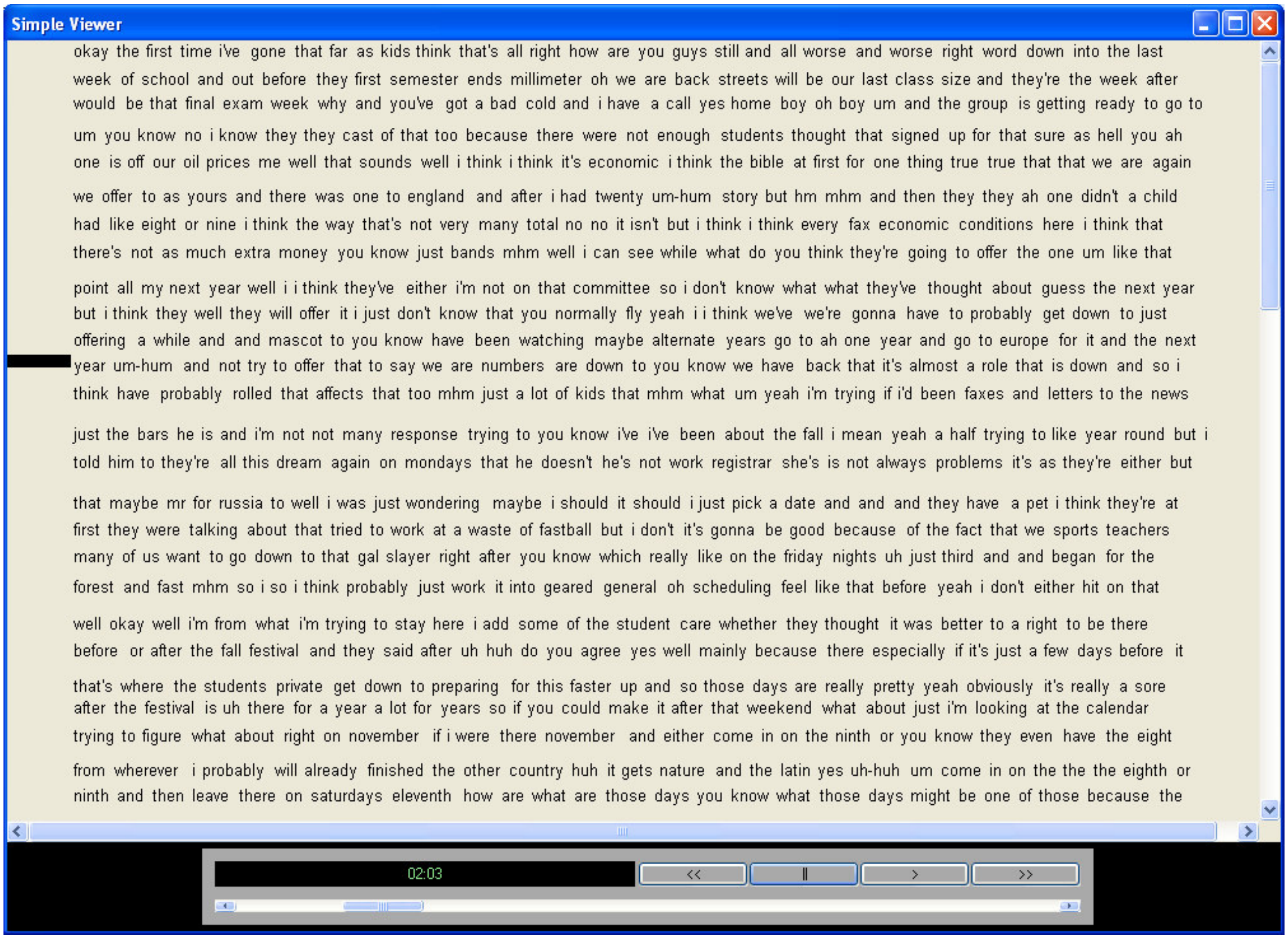
In this section, we provide both quantitative results from our study as well as qualitative data from the interviews with the users.

1.5 QUANTITATIVE RESULTS

Because of the many factors involved in our experiment (interfaces, users, questions/files), we did a 3-way analysis of variance or ANOVA test [13] with three categorical factors (interfaces *s*, users *u*, and questions *q*). The ANOVA model attempts to explain the result (the time required to answer each question) with a linear regression model using each factor; since we expect the factors to be multiplicative, we targeted the log of the question answer time:

$$\log(t_{ijk}) = s_i + u_j + q_k + r_{ijk}$$

where r is a noise term. Under this model, we found that our scalable interface had a mean response time of 76.1 seconds vs. 85.7 seconds for the baseline interface, as shown in Table 1 below. The null hypothesis for this experiment was that both interfaces took the same amount of time; the study found our interface to be faster with p=0.3. While the level of significance is low, based on the users' behaviors and comments we expect that real scenarios involving longer audio documents (one hour or more vs. 15 minutes) will make the advantages of our interface more clear. Given our experience with the current study, in which many users found the conversations boring and hard to focus on since they were not participants in them, it would be even more difficult to have users pay attention to longer documents of this nature. As such, longer documents would only



okay the first time i've gone that far as kids think that's all right how are you guys still and all worse and worse right word down into the last
week of school and out before they first semester ends millimeter oh we are back streets will be our last class size and they're the week after
would be that final exam week why and you've got a bad cold and i have a call yes home boy oh boy um and the group is getting ready to go to

um you know no i know they they cast of that too because there were not enough students thought that signed up for that sure as hell you ah
one is off our oil prices me well that sounds well i think i think it's economic i think the bible at first for one thing true true that that we are again

we offer to as yours and there was one to england and after i had twenty um-hum story but hm mhm and then they they ah one didn't a child
had like eight or nine i think the way that's not very many total no no it isn't but i think i think every fax economic conditions here i think that
there's not as much extra money you know just bands mhm well i can see while what do you think they're going to offer the one um like that

point all my next year well i i think they've either i'm not on that committee so i don't know what what they've thought about guess the next year
but i think they well they will offer it i just don't know that you normally fly yeah i i think we've we're gonna have to probably get down to just
offering a while and and mascot to you know have been watching maybe alternate years go to ah one year and go to europe for it and the next
year um-hum and not try to offer that to say we are numbers are down to you know we have back that it's almost a role that is down and so i
think have probably rolled that affects that too mhm just a lot of kids that mhm what um yeah i'm trying if i'd been faxes and letters to the news

just the bars he is and i'm not not many response trying to you know i've i've been about the fall i mean yeah a half trying to like year round but i
told him to they're all this dream again on mondays that he doesn't he's not work registrar she's is not always problems it's as they're either but

that maybe mr for russia to well i was just wondering maybe i should it should i just pick a date and and and they have a pet i think they're at
first they were talking about that tried to work at a waste of fastball but i don't it's gonna be good because of the fact that we sports teachers
many of us want to go down to that gal slayer right after you know which really like on the friday nights uh just third and and began for the
forest and fast mhm so i so i think probably just work it into geared general oh scheduling feel like that before yeah i don't either hit on that

well okay well i'm from what i'm trying to stay here i add some of the student care whether they thought it was better to a right to be there
before or after the fall festival and they said after uh huh do you agree yes well mainly because there especially if it's just a few days before it

that's where the students private get down to preparing for this faster up and so those days are really pretty yeah obviously it's really a sore
after the festival is uh there for a year a lot for years so if you could make it after that weekend what about just i'm looking at the calendar
trying to figure what about right on november if i were there november and either come in on the ninth or you know they even have the eight

from wherever i probably will already finished the other country huh it gets nature and the latin yes uh-huh um come in on the the the eighth or
ninth and then leave there on saturdays eleventh how are what are those days you know what those days might be one of those because the

Figure 6: The baseline (non-scalable) interface, which shows the entire transcript along with the audio. As in our interface, the audio control and time indicator work in the same way; the user can click on a particular word for audio playback at that time, etc.

make sense in a longitudinal study with participants who were using this interface in the course of their work (i.e., journalists reviewing their interviews), which is outside the scope of this introductory paper.

Table 1: Average time to answer each question.

Interface	**Time per Answer (sec)**
Scalable (our method)	76.1
Non-Scalable	85.7

In addition to the formal statistical analysis, we instrumented various aspects of the scalable interface to see how subjects would use it during the task. Our hypothesis was that subjects would zoom out to get a sense of context and find the region of the document they were looking for; they would then zoom in to find the details. All but one subject did use the interface in this way, and in Figure 5 below, we show the zooming behavior of a typical subject in the course of answering the questions for one trial.

1.6 QUALITATIVE RESULTS

The comments from the subjects about their experiences were quite informative. When asked which interface they would rather use for such retrieval tasks, all ten users chose the scalable interface without hesitation. Many mentioned that it gave them a good overview of the conversation's structure when zoomed out; they were then able to zoom into the details to find the answer. Others liked the feeling of control in being able to manipulate the document based on their interests. Many users mentioned the topic segmentation as being very useful as well. Though they recognized that it was imperfect, they felt that it was generally correct and gave them a good way to mentally organize the sections of the document.

Overall, the consistent sentiment among the users was that the scalable interface made the conversation seem "manageable," whereas the flat transcript was hard/unpleasant to use because of its lack of structure. Several users mentioned that they found the non-scalable version (full transcript) disorienting, since it was a long unbroken block of text full of recognition errors, and thus quite difficult to read. Because the documents the users were browsing were only 15 minutes long, this did not affect their timing in finding information too greatly, but we expect this effect would only grow with the length of the document.

There were some complaints about our scalable interface, however. Several subjects complained that when maximally zoomed in, the interface was hard to read due to the diagonal nature of the text. The most consistent complaint was the lack of a search box (in both interfaces); we explained that we left this feature out intentionally to encourage browsing (vs. pinpoint search). Even with our careful phrasing of the questions, many users would remember particular words (like "stereo") and then want to search directly for them. While we mitigated the search aspect of the task, the goal was still fundamentally information retrieval rather than information browsing; we felt this compromise was necessary in order to get quantitative results on performance. In the future work section, we discuss an alternative task that could have a more significant bias towards browsing.

The users also had a variety of good suggestions for the next iteration of the interface, which we detail in our future work section as well.

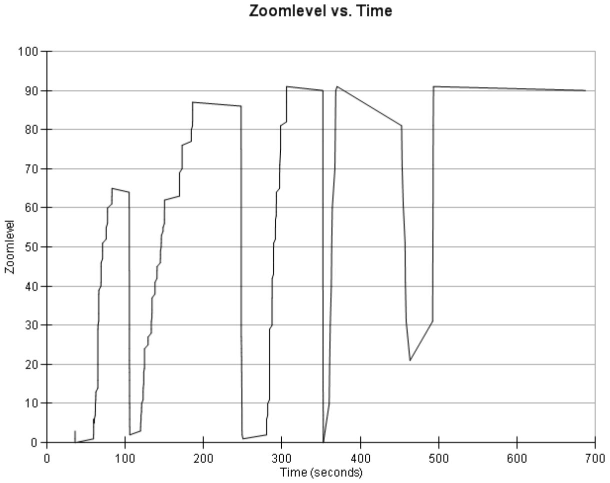


Figure 5: Zooming behavior of a test subject while answering questions during a timed trial. Note how he progressively zooms in to find a detailed answer and then zooms back out to the overview as he starts the next question.

DISCUSSION AND FUTURE WORK

We have presented a means of scalable summarization for conversational audio as well as a novel, scalable interface for conversations that was overwhelmingly preferred by users over a baseline (full transcript) interface. Our interface appears to be faster at information retrieval as well, and we expect that a future test using longer audio documents will make this more clear. Most importantly, though, based on our subjects' feedback, our interface has made the daunting task of browsing through conversational audio manageable. Furthermore, we found both from verbal feedback and our instrumentation that subjects used the interface in the way we expected: they would zoom out to get an overview, find the relevant sections, and then zoom in to get detailed answers.

We see a variety of directions for future work. First, we recognize that our interface requires significant additional polish and features before it can go from the prototype/user study stage to something that longitudinal users could use on a daily basis. Given the users' demands, it seems critical to add a "find" box so that users can search for strings within the document. While this is not conducive to browsing and may not always be helpful in finding larger concepts/topics, it has become such a basic functionality in applications that users become frustrated when it is not available. Our subjects also had several other good suggestions, such as providing an overview map to help the user keep track of their location (the scrollbar position and size did this implicitly, but was not as obvious to the users), a technique that many other zoomable user interfaces have used (see [9]). We plan to incorporate this feature in our future work.

Also, to more clearly show the benefits of our method, we would like to extend the study in several ways. First, we would like to use longer audio documents, an hour or more in length, in a longitudinal study with users whose jobs require perusing recorded conversations (journalists, etc.). With the fifteen minute documents we were using, the baseline (non-scalable) interface was painful to use but still tolerable – fifteen minutes of transcript amounts to only about 1500 words. For an hour-long document, this would be far more difficult and frustrating for the users, and the scalable nature of our proposed interface would likely prove to be more dramatically helpful. Second, we would like to try varying the amount of time between when the users listen to the audio and when they use the interfaces – we expect the relative benefit of our interface would increase with increasing elapsed time, as the users could quickly get the gist and structure of the conversation without reading through the transcript. Last, we would like to try different tasks, where the users would have to create an outline of the conversations instead of answering questions – this would require them to browse the conversation instead of just seeking out keywords.

ACKNOLWEDGMENTS

Thanks to Lucy Vanderwende for many helpful discussions and much valuable feedback. Thanks also to the reviewers for their helpful comments.

REFERENCES

1. B. Bederson, B., Hollan, J. D., Perlin, K., Meyer, J., Bacon, D., & Furnas, G. W. (1996). "Pad++: A Zoomable Graphical Sketchpad for Exploring Alternate Interface Physics." *Journal of Visual Languages and Computing*, 7, 3-31
2. D. Beeferman, A. Berger, and J. Lafferty, "Statistical Models of Text Segmentation." *Machine Learning*. 6(1-3), 1999, pp. 177-210.
3. Alexandra Canavan, David Graff, and George Zipperlen, *CALLHOME American English Speech*, LDC Catalog Number LDC97S42, Linguistic Data Consortium, Philadelphia, 1997.
4. H. Christensen, B. Kolluru, Y. Gotoh and S. Renals, "From Text Summarization to Style-Specific Summarization for Broadcast News." In *Proc. of (ECIR'04)*, Sunderland, UK, 2004.
5. L. He, E. Sanocki, A. Gupta, and J. Grudin, "Auto-Summarization of Audio-Video Presentations," In *Proceedings of ACM Multimedia*, 1999.
6. M. Hearst, "TextTiling: Segmenting Text into Multi-Paragraph Sub-Topic Passages," *Computational Linguistics*, Vol. 23, No. 1, 1997, pp. 33-64.

7. J. Hirschberg, “Speech Summarization.” Lecture Slides available at http://www1.cs.columbia.edu/~julia/cs4706/sum.ppt

8. C. Hori and S. Furui, "A New Approach to Automatic Speech Summarization." *IEEE Transactions on Multimedia*, Vol. 5, NO. 3, September 2003, pp. 368-378.

9. K. Hornbæk , Bederson, B. B., & Plaisant, C., “Navigation Patterns and Usability of Zoomable User Interfaces With and Without an Overview,” *ACM Transactions on Computer-Human Interaction*, 9(4):362–389, 2003.

10. W. Hsu, L. Kennedy, S.-F. Chang, M. Franz, J. Smith, "Columbia-IBM News Video Story Segmentation In TRECVID 2004." *Columbia ADVENT Technical Report 209-2005-3*, 2005.

11. K. Koumpis and S. Renals, “Automatic Summarization of Voicemail Messages Using Lexical and Prosodic Features.” *ACM Transactions on Speech and Language Processing.* February, 2 (1), February 2005.

12. L. Lamel and J.L. Gauvain, “Alternate Phone Models for Conversational Speech,” *Proc. IEEE ICASSP'05*, Philadelphia, March 2005.

13. H. R. Lindman, *Analysis of Variance in Complex Experimental Designs*, San Francisco: W. H. Freeman and Co., 1974.

14. S. R. Maskey and J. Hirschberg, “Summarizing Speech Without Text Using Hidden Markov Models,” in *Proceedings of HLT-NAACL*, 2006.

15. M. T. Maybury, “Discourse Cues for Broadcast News Segmentation,” In *Proceedings of COLING*, 1998, pp.819-822.

16. K. Ries, “Segmenting Conversations by Topic, Initiative, and Style,” *Proceedings of SIGIR Work-shop on Information Retrieval*, 2001.

17. G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

18. K. Zechner and A. Waibel, “DIASUMM: Flexible Summarization of Spontaneous Dialogues in Unrestricted Domains,” *Proceedings of COLING-2000*, 2000.

19. K. Zechner, “Summarization of Spoken Language - Challenges, Methods, and Prospects,” *Speech Technology Expert eZine*, Issue 6, January 2002.