Recording in Thoughts: Recording Speech for the Purpose of Sharing

Neema Moraveji, Pei Yu, Xianjun Huang
Microsoft Research Asia
Beijing, China
neemam@microsoft.com, pei@interaction-ivrea.it, xjhuang@126.com

ABSTRACT
Non-expert users rarely use recorded speech as a means of communication and expression. This paper describes work on a recording interface that reduces the need for editing, so as to encourage the use of speech as a medium. User needs were determined by a diary study and participatory design session. A recording technique, *recording in thoughts*, is presented, aiming to reduce the need for post-production by enabling quick re-recording of individual thoughts and shortening of long segments. Eight producers and 43 listeners compared our technique with a more traditional recording interface, preferring to both create and listen to recordings made with ours. The work concludes with a discussion of next steps for this research.

Author Keywords
Speech recording; user interface; podcast; blog; user study.

ACM Classification Keywords
H.5.2 [Information interfaces and presentation]: User Interfaces

INTRODUCTION

Digital, recorded speech has recently experienced a surge of popularity as a communication medium due primarily to the emergence of digital audio players (DAP) (e.g., Apple's iPod) and the Really Simple Syndication (RSS) delivery mechanism. Attracted to this platform, audio amateurs are publishing recorded speech as a rich medium for expression [2]. This has led to the emergence of thousands of 'podcasts', or public audio journals. As of April 2005, 6 of the 22 million Americans who own DAPs had downloaded podcasts [10], 7000 of which are listed on PodcastAlley.com.

While portable devices are useful for capturing audio [4, 13], creating speech recordings that are coherent enough to share or publish usually requires desktop audio-editing. Even relatively simple consumer tools like Adobe Audition and Audacity require understanding of non-linear editing, waveforms, etc.

While some savvy hobbyists do learn to use desktop editing to create podcasts and audio journals, it is time-consuming and non-trivial. We assert that the daunting prospect of editing keeps interested non-experts away from the medium. This work presents a recording technique that reduces the need for editing and post-production. The technique, called *recording in thoughts*, allows users to a) edit on-the-spot by re-recording individual thoughts and b) create concise entries by quoting other audio programs and adding topical comments. Results of a user study that compared recording in thoughts with a more traditional recording interface show the former was preferred by both producers and listeners.

RELATED WORK

Capturing and utilizing speech is addressed well in the literature, as in [6] and [1]. Degen [4] and Stifelman's [12] work on microcassette recorders are landmark projects dealing with recording notes or ideas for archive or retrieval. However, these systems were not designed to support recording for the purpose of publication.

Beyond inserting an 'Index' for navigation, some digital voice recorders (DVR) offer basic editing directly on the device. On the Philips Pocket Memo 9450 VC, advanced users can delete a portion of recorded audio as well as choose to record in 'Insert' or 'Overwrite' modes.

Several new packages like [7] have recently emerged for amateurs to mimic radio shows. Odeo [9], a desktop recording interface sacrificing editing in the name of simplicity, is currently in private beta.

The literature has not addressed non-expert users who wish to record in mobile contexts for the purpose of sharing.

USER RESEARCH

Diary study

Four university students (one female) identified as potential podcasters participated in a 4-day diary study. All were current bloggers interested in our project description. We gave each a DVR to create audio blog entries.

Participatory-design study

An iRiver iHP-120 DAP with internal microphone was given to a 29 year old male blogger. We asked him to compose a podcast for his friends. For two hours, he recorded while explaining how he wanted to be able to control the device, suggesting industrial and interface design features. As he recorded, he described the edits that he wanted to apply. We then edited his recording to adhere to those edits, and met with him two days later so he could hear his creation.

User research results

Our user research indicated that a traditional DVR is not suitable for recording for the purpose of sharing. Participants were excited about audio's creative possibilities, but disappointed with the difficulty of using the medium. Our participatory design session was promising because the edits prescribed *during the recording process* produced satisfactory results.

- Users often knew, even while recording, whether or not the current recording was interesting or boring.
- Diary subjects were unsatisfied with the product of the DVR because it would require a great deal of editing. It came out "too long", "boring", and "sounded stupid".
- Subjects sometimes wanted to record a long amount of footage but select only a portion to publish.
- The participatory design participant was satisfied and pleasantly surprised with the recording.

RECORDING IN THOUGHTS

In our user research, users were judging their own recordings while creating them. This information is very valuable, yet not useful until post-production. The traditional Record-Pause-Stop-Rewind-Forward paradigm does not afford quick deletion of undesired utterances. As a result, users are forced to either record more and rely on editing or be left with meandering and/or boring content.

We attempted to create a recording interface that matches normal speech habits. In common speech, people string together thoughts to form a complex idea. If one of the thoughts was unclear, one re-explains it along the way, not after all the thoughts are expressed. Similarly, users should be able to quickly re-record an unclear thought, creating a string of coherent thoughts to make up a succinct audio recording, with little to no need for manual post-production. We call this technique *recording in thoughts*.

In the same vein, we felt it necessary to provide an affordance specifically for recording one's thoughts as comments on another's. This is common in blogs, where authors quote and comment on other web pages and news articles. In email, users often quote and reply to individual sentences. This seemed like a logical extension of recording in thoughts.

INTERACTION DESIGN

Our recording interface enables one to a) quickly re-record thoughts during the creation process and b) record succinct commentary of individual quotes. Ideally, a DVR is built purposefully for this interface. For prototyping purposes, however, ours uses a Pocket PC.

For recording, we offer the user two primary buttons: Record and Delete/Recover (Figure 1). The user holds down the Record button to record. When they remove their finger, the recording stops. This implies either "That was good, now I'm collecting my thoughts for the next recording" or "That was a mistake. I want to re-record that thought." For the former, the user collects his/her thoughts while the system is paused, then holds down Record to begin again. For the latter, the user presses the Delete

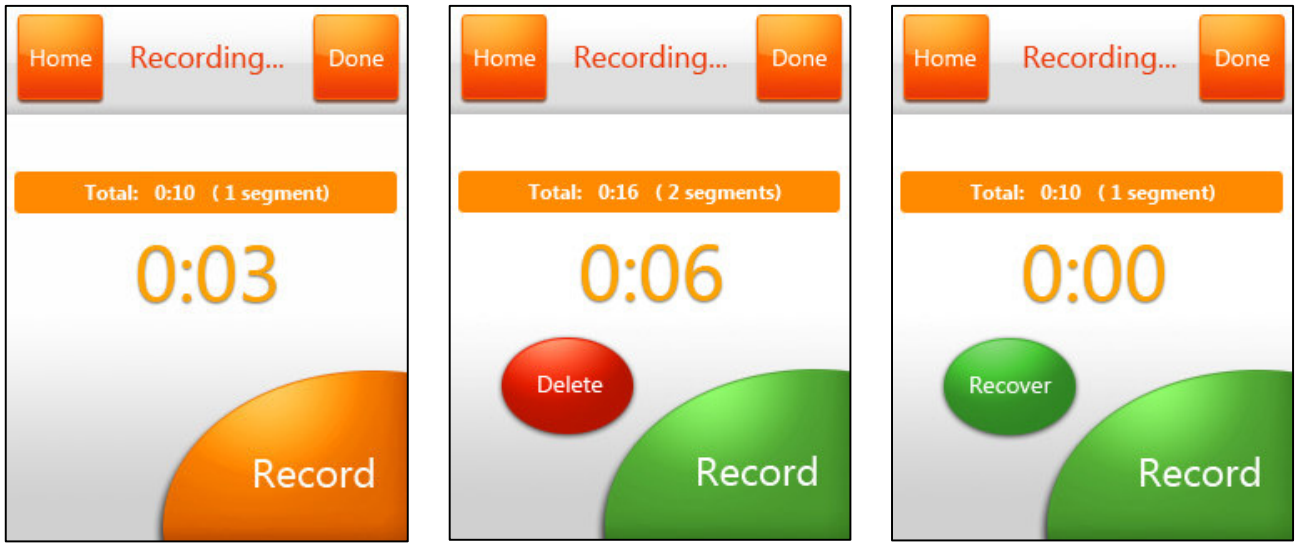

Figure 1. User already recorded one 10 second thought. Left: user holds down "Record" to record a thought (3 seconds so far). Middle: user releases "Record" to signify end of a 6-second thought, total is updated and now user has option to delete it. Right: after deleting, total is updated and user can recover the deleted portion or record directly over it.

button, deleting the current thought. One cannot delete previously-committed thoughts and, immediately after a deletion, the button toggles to Recover.

The commenting action combines quoting with recording a preamble and/or postscript. This enables the user to a) create concise versions of lengthy recordings and b) create focused commentary on noteworthy quotes. The user presses the Quote button at the start and end points of a desired excerpt (Figure 3). Then they see options to record a preamble or postscript and can preview the result, which now contains the excerpt and commentary (if any).

EXPERIMENT

The purpose of our experiment was to observe the reaction of non-expert users to our interface (A) as compared to a traditional interface (B). For B, we used Yoho (Figure 2), the basic version ProTone [11], the most popular Pocket PC audio recording tool on CNet.com. Our hypothesis was that participants and anonymous listeners would find recordings made with A to be more coherent and pleasant.

Eight participants (2 female), average age 28.3, volunteered for the experiment. One had experience listening to podcasts, 6 were blog readers, and 6 owned a DAP. Reading from a script, we presented participants with two interfaces: ours and Yoho. Each was given instructions and conducted a practice recording with each interface. For each of the 2 tasks, subjects recorded once with each device, creating 4 total recordings. Half the participants used A first, the other half used B first.

The first task was to record an audio journal entry. The audience was friends and the topic was "What I like about being in Beijing" for the first device and "What I miss about being home" for the second. The second task was to comment on any portion of the provided audio news programs, "Hurricane First-Hand" [3] for the first device "Technology Report" [8] for the second. Using Likert

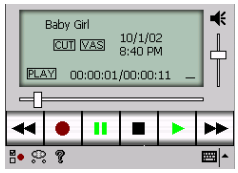

Figure 2. Interface B, Yoho [11].

scales of 7, participants rated recording at the time of creation and, upon completion of all tasks, compared each task's recordings side-by-side.

To determine which recordings anonymous listeners could see as "more coherent and listenable" between the two interfaces, 43 people aged 21-61 listened to 2 pairs of recordings each. That is, each participant had their recordings listened to by at least 5 other people. They had no knowledge of the experiment's purpose.

Results

Table 1 shows the results of subjective ratings at the time of recording; Table 3 shows results of a direct comparison of recordings made with each interface. The 8 participants consistently preferred using recording in thoughts (A) to a popular voice recording interface (B). They also rated A's recordings as more listenable and as requiring less time for editing.

Using interface A, all subjects liked holding the Record button down, 6 did lift their finger up to pause, and 3 used Delete. For Task 1, interface A's recordings were, on average, 35.3 seconds shorter than B's, with the average across both devices being 101.9. For Task 2, 5 participants used the Quote feature. 4 had usability trouble setting the start and end points of the excerpt.

The "Diff" columns of Table 1 show interface A is preferred, and especially for free-form entries. However, for recordings that use quotes, some editing must be done to smooth the transitions to and from the quote. Comparing

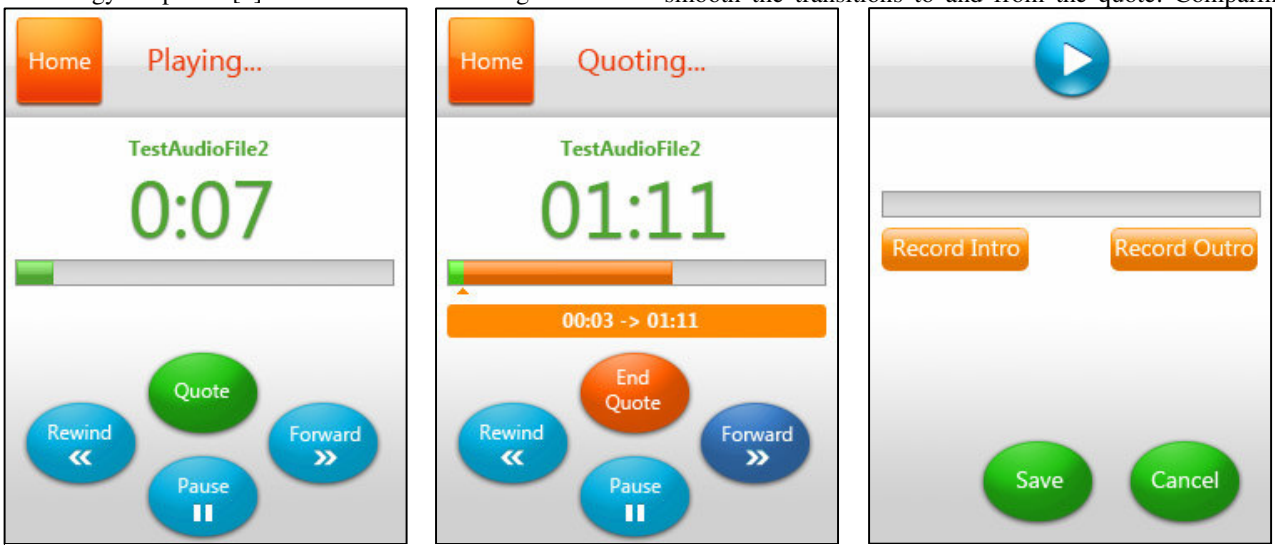

Figure 3. User is playing an audio program. Left: quote button is available. Middle: visual marker shows quote start point and duration. Right: after quoting, user has the option to record preamble or postscript and preview the creation.

Statement	A	B	Diff
Little editing is needed	5.3	4.1	1.2
	4.5	4.6	- 0.1
Organized, structured, not long-winded	4.4	3.1	1.3
	5.8	4.1	1.7
Listenable	5.6	4.0	1.6
	5.1	4.6	0.5
No pressure while recording	5.0	4.6	0.4
	5.6	4.3	1.3

Table 1. Subjective ratings at time of recording for Task 1 (top pair for each statement) and Task 2 (bottom pair).

the recordings side-by-side was less strongly in favor of interface A. However, the producer's satisfaction still remained strong.

When we gave the clips to the 43 anonymous listeners, our initial results were validated. Table 2 shows the result of the listeners' ratings, 84% preferred A for free-form entries while news commentary entries were more evenly split. Commentaries on interface A were often curt, consisting only of a brusque comment followed by a quote while B's were more like free-form entries. For this reason, many listeners found A's to be topical but unorganized. The abrupt change to a second voice, with no introduction, was not always agreeable. Most listeners said that they appreciated the quoted material, which helped to keep recordings on topic.

DISCUSSION

'Recording in thoughts' relieves pressure while discouraging rambling with the combination of holding the Record button down while recording and knowing that one can undo a thought. We expect its effectiveness to increase with practice. The most effective way to use it is may be in short bursts, lifting the finger and putting it back after each sentence.

When commenting on other audio content, quote functionality encourages topical comments and is preferred by producers and listeners alike. However, some improvements should be made to the quote's listening experience. We plan to ease the transitions with automatic processing (fade, inserted silence, audible cue) and to offer a precise method for producers to adjust the trim points of their excerpts. A direct means of including quotations and commentary has thus far been unavailable to creators of audio journals. We believe that enabling this technique, so central to many media, will support an increased range of expression in audio journals.

Recording type	A	B
Free-form	36 (84%)	7
News commentary	24 (56%)	19

Table 2. Preferences of anonymous listeners.

Criteria for comparison	A	B	Diff
Coherence	5.1	4.9	0.2
	5.4	4.9	0.5
Listenability	5.4	4.6	0.8
	5.5	5.0	0.5
Satisfaction	5.5	4.4	1.1
	5.4	4.3	1.1
Time-needed for editing	5.1	4.8	0.3
	5.6	4.5	1.1

Table 3. Side-by-side subjective ratings for each interface.

Next steps

First is to automatically smooth the transitions between thoughts and between recordings and quoted material. Second is to provide "just-enough polish" with audio processing techniques so as to improve the listening experience. Also, we will experiment with methods of 'undoing' parts of the current segment, instead of all of it.

Acknowledgements

We thank Dan Rosenfeld for the great discussions.

REFERENCES

1. Arons, B. Techniques, perception, and applications of time-cornpressed speech. *AVIOS 92*, I/O Society (1992).
2. Chalfonte, B., Fish, R.S., Kraut, R.E. Expressive Richness: Comparison of Speech and Text as Media for Revision. *CHI 91*, ACM Press (1991).
3. CNN.com. Hurricane First Hand. Sept 1, 2005.
4. Degen, L., Mander, R., Salomon, G. Working with audio: integrating personal audio recorders and desktop computers. *CHI 92*, ACM Press (1992).
5. Sawhney, N., Schmandt, C. Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments, *TOCHI* Sept., ACM Press (2000).
6. Hindus, D., Schmandt, C. Ubiquitous audio: Capturing spontaneous collaboration. *CSCW*, ACM Press (1992).
7. MixCast, www.mixcastlive.com
8. National Public Radio podcasts, Sept 13, 2005.
9. Odeo. www.odeo.com. (Private beta)
10. Pew Internet & American Life Project Survey on Podcasting, www.pewinternet.org, April 2005.
11. Pocco Software. www.poccosoftware.com
12. Stifelman L. VoiceNotes: An Application for a Voice-Controlled Hand-Held Computer. *MIT* (1992).
13. Tucker, R.C.F. Speech-as-data technologies for personal information devices. *Personal Ubiquitous Computing*, Springer-Verlag Press (2003).

The columns on the last page should be of approximately equal length.