AudioWiz: Nearly Real-time Audio Transcriptions

Samuel White
University of Rochester
Rochester, NY 14627
samuel.white@rochester.edu

ABSTRACT

Existing automated transcription solutions filter out environmental noises and focus only on transcribing the spoken word. This leaves deaf and hard of hearing users with no way of learning about events that provide no spoken information such as the sounds produced by a faulty appliance or the barked alert of a dutiful guard dog. In this paper we present AudioWiz, a mobile application that provides highly detailed audio transcriptions of both the spoken word and the accompanying environmental sounds. This approach is made possible by harnessing humans to provide audio transcriptions instead of more traditional automated means. Web-workers are recruited automatically in nearly real-time as dictated by demand.

Categories and Subject Descriptors

K.4.2 [**Computers and Society**]: Social Issues – *Assistive technologies for persons with disabilities*.

General Terms

Design, Experimentation, Human Factors.

Keywords

Assistive technology, sound visualization, audio transcription.

1. INTRODUCTION

Sound permeates every facet of our daily lives. Speech enables communication, beeps or high pitch alarms warn of danger, and the grinding of a disc player suggests its malfunction. But despite its significance, ambient audio information is typically unavailable to deaf and hard of hearing people. Automatic tools to address this problem are being developed but remain inaccurate, expensive, and limited in scope [1, 2]. AudioWiz instead relies on human power harvested from the web to provide transcriptions. This approach affords the ability to analyze audio of nearly any quality because of the excellent ability humans have at discerning useful information.

2. AUDIOWIZ OVERVIEW

The AudioWiz application and service is comprised of two main parts. A client side application that runs on a users device and records audio, and a server side application that handles worker recruitment and job queuing. While a user has the AudioWiz application running on their device, a buffer storing as much as thirty seconds of audio is maintained. As incoming audio is put into the buffer it is also displayed visually on the devices screen (Figure 1). This dynamic visual representation depicts the volume level of all stored audio and scrolls right to left as new audio arrives giving the user a means to detect important audio events visually before deciding if they are worthy of transcription. Once important events are identified, they users presses the "Transcribe It!" button and the audio buffer is compressed and uploaded to the web server for human transcription.

Audio transcriptions are provided by web-workers who are recruited in nearly real-time as required. Workers are not pre-trained to perform audio transcriptions but instead are given only simple instructions on how to complete their task. These directives instruct workers to listen for both significant verbal and nonverbal events. In the absence of perceived significant events workers are to instead asked to describe everything they hear in as much detail as possible. Harnessing human power for transcriptions affords the ability to discern complex contextual and environmental information with minimal guidance, something far beyond the capabilities of existing automated systems. Recruited web-workers are paid one cent for each valid transcription they provide and are barred from attempting to transcribe the same audio more than once.

completed transcriptions are sent back to users in real-time and are immediately displayed on the devices screen. The entire transcription process can be completed in as little as one minute giving users a realistic way to decipher auditory information of nearly any kind.

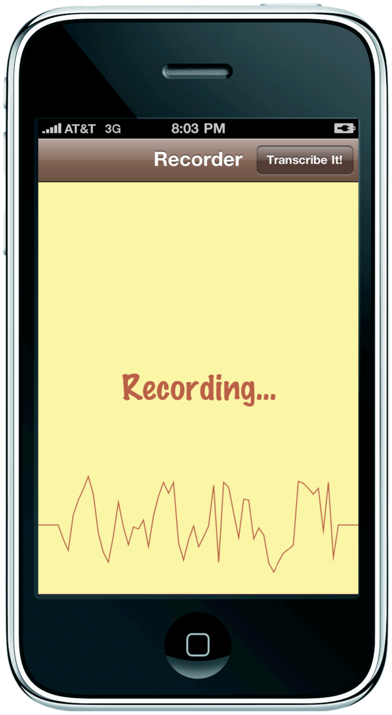

Figure 1: The AudioWiz recording interface running on an iPhone 3GS. Across the lower portion of the screen scrolls a constantly updating visualization of the current thirty seconds of audio available for transcription.

3. IMPLEMENTATION

We chose to develop our device software for Apple's iPhone 3GS. This platform was selected primarily because the device includes a hardware MPEG-4 encoder. Leveraging this hardware encoder allows us to rapidly compress our target audio before transmitting it for transcription thus reducing the delay associated with compressing the audio in software alone. This approach results in significantly shorter transmission times when measured against uncompressed or less compressed audio formats.

The workers at the heart of the application are recruited from Amazon's Mechanical Turk service using an abstraction layer called TurKit [3]. TurKit provides a simple set of APIs that make the process of farming out jobs to human workers trivial. Wrapping TurKit is a set of freely available scripts developed by Jeffrey Bigham at the University of Rochester called quickTurKit. quickTurKit allows AudioWiz to begin recruiting workers the moment an instance of the client application is launched. This way, workers are already available when the first slice of audio is transmitted for transcription.

During runtime incoming audio is visualized on screen. This visualization is constantly refreshing from right to left to accurately display all audio currently stored on the device and available for transcription. Higher peaks in the visualization represent periods of increased auditory activity and volume.

4. CONCLUSION AND FUTURE WORK

We have presented AudioWiz, a novel low-cost solution for providing deaf and hard of hearing users with nearly real-time audio transcriptions. Unlike traditional automated approaches that only provide a transcription of the spoken word, AudioWiz is able to provide information about contextual and environmental audio containing no spoken words whatsoever.

Moving forward we would like to provide an interactive way for users to select just a small slice of audio from the visual representation of all audio stored on the device. We believe doing so would allow us to provide even faster results given that workers could be given smaller chunks of audio to transcribe and possibly even further lower associated costs. Additionally, it may be useful if users provide workers with some description of what events they are trying to detect, such as a doorbell or telephone ring. This way workers could focus their efforts and provide more meaningful responses at faster speeds.

5. REFERENCES

[1] M Ravishankar. Efficient algorithms for speech recognition. PhD thesis, Carnegie Mellon University, 2005. http://citeseerx.ist.psu.edu/viewdoc/download?DOI=10.1.1.72.3560&rep=rep1&type=pdf.

[2] Ho-Ching, F. W., Mankoff, J., and Landay, J. A. 2003. Can you see what i hear?: the design and evaluation of a peripheral sound display for the deaf. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Ft. Lauderdale, Florida, USA, April 05 - 10, 2003). CHI '03. ACM, New York, NY, 161-168. DOI= http://doi.acm.org/10.1145/642611.642641

[3] Little, G., L. Chilton, M. Goldman, and R.C. Miller. TurKit: Human Computation Algorithms on Mechanical Turk. UIST 2010, 2010.