VoiceTranscriber: Crowd-powered Oral Narrative Summarization System

Hung-Chi Lee[1] and **Jane Yung-jen Hsu**[1,2]

[1]Department of Computer Science and Information Engineering
[2]Intel-NTU Connected Context Computing Center
National Taiwan University

{d99922020, yjhsu}@csie.ntu.edu.tw

Abstract

VoiceTranscriber is a mobile crowd-powered system for summarizing stories from recorded voices that relies on the human abilities of discrimination and expression. We propose a system for organizing crowd workers to act as transcribers of original recorded voices to an automated speech recognition service. We also evaluate the system's usability in four storytelling processes. Feedback from participants indicates that VoiceTranscriber is easy to use and learn, and that the summarized stories have high factuality.

Introduction

Oral narrative is a commonly used method to share the life experiences of families, whether they are important events or those from everyday life. In particular, storytelling is a method to summarize and preserve the valuable life experiences of the elderly for future generations.

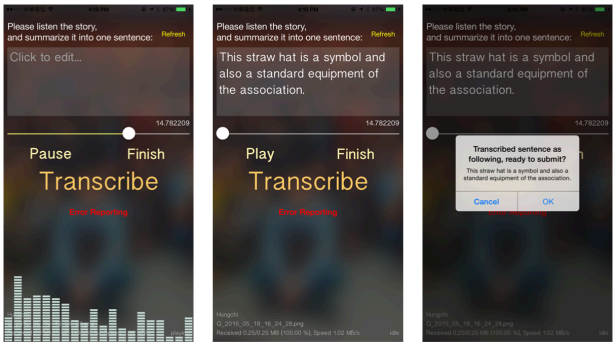
Automated speech recognition (ASR) is a speech-to-text method that has the advantages of low cost and immediate feedback. However, it also has low accuracy in real settings, and makes frequent errors that distort the meaning of the original narrative (Silsbee and Bovik 1996). Another approach is re-speaking, whereby a well-trained person in a controlled environment is connected to a live audio feed and repeats what they hear to an ASR device (Imai et al. 2002). Another technique is to use crowds of people to help in the transcription tasks. LEGION:SCRIBE is a system in which groups of non-experts collectively caption speech in real-time on-demand (Lasecki et al. 2012).

Previous studies have focused on transcribing the narratives precisely. In our study, transcribers exclude information such as expletives or repeated words, while extracting concrete tags and stories from the voice recordings of people sharing their memories.

The core contribution of this paper is the provision of a mobile crowd-powered system for organizing crowd workers to serve as voice transcribers of original recorded voices to an ASR device. In the study, transcribers operated the mobile application VoiceTranscriber to listen to the original voices and then reconstructed the contents of the narrative in text form.

VoiceTranscriber System

VoiceTranscriber is a mobile application for deploying crowd-sourcing transcription tasks to transcribers. There are three steps in the task (as shown in Figure 1). (1) Listening: transcribers can play or pause the voices. The narrative is recent recordings, takes minimal transcription times, and is selected and downloaded by the VoiceTranscriber. (2) Summarizing and re-speaking: after transcribers listen to the original voices, they are asked to summarize what they heard and repeat it in one sentence. Then, the ASR system transforms the sentences spoken by the voices into text. (3) Editing and submitting: the transcribed text is displayed on the screen, where it can be edited or modified with respect to typographical errors, and then submitted to the server.



(1) Listening (2) Summarizing & Re-speaking (3) Editing & Submitting

Figure 1: User interfaces of the VoiceTranscriber.

Re-Speaking for Automated Speech Recognition

Our approach relies on the human ability to understand and discriminate between story elements that are important and those that are not, to filter out noises, and to summarize the main points into short sentences. We considered re-speaking the essence of the story in sentences with a clear voice to be an easier and quicker task than typing the text. In this manner, we can achieve higher ASR accuracy and condense the original stories.

Tag Extraction and Summarization

Having obtained concise sentences from the voice files, we then apply the CKIP Chinese word segmentation system. To extract meaningful tags from sentences, CKIP functions as a Web service and returns segmented text with part-of-speech tags. Our VoiceTranscriber system generates a summarized story for each event. The story is then pieced together from all transcriptions related to the event and have been sorted by a timestamp to correspond each with the original story flow.

Preliminary Experiment

For our preliminary experiment, we recruited four young participants and put them into two non-overlapping pairs for four rounds. In each session, one of participants took the role of storyteller and shared his/her life experiences. At the same time, four transcribers were located in a separate room awaiting the start of their tasks.

Each group spent 30 minutes in their assigned task, after which we conducted semi-structured interviews with the storytellers and the transcribers. By the end of the experiment, participants had shared a total of 27 photos and collected 195 voice files containing 249 sentences and ultimately 1,041 tags—an average of 39 tags for each photo.

Storyteller Interview

After the photo sharing processes were complete, the storytellers were asked to rate the summarized stories on a five-point Likert-like scale, where 5 is highest, against performance indexes such as completeness, factuality, and overall satisfaction with the stories. The results show that storytellers rated the factuality aspect at 4.3, indicating that they believed the transcribers had correctly transcribed the story they had narrated. However, the scores for completeness and satisfaction were lower, at 3.7 and 3.6, respectively (as shown in Figure 2). One storyteller said that *"The satisfaction rating is because the summarized stories are combined sentence by sentence, so the coherence of the story is not good enough."*

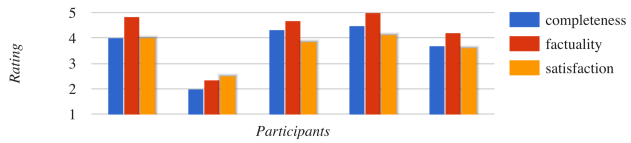


Figure 2: Performance indexes of the summarized stories.

Transcriber Interview

We interviewed the transcribers using two questionnaires. The Computer System Usability questionnaire uses a seven-point Likert-like scale, where 7 is highest. The NASA-Task Load Index (NASA-TLX) is a 21-scale questionnaire, where 0 is the best score. Most transcribers rated the VoiceTranscriber as being easy to use (6) and easy to learn (6). Overall, they rated their satisfaction with the service as good (5). However, when asked about their task load, they reported that the task made a high mental demand (11), required effort (13.75), and was associated with much frustration (10.75). One transcriber stated that *"Because the storyteller may share a long story in the recorded session, but I could not remember so much information at the same time. Consequently, I had to listen to the voices several times in order to try and transcribe them clearly. This made me tired and frustrated."*

Conclusion and Future Work

We proposed a crowd-powered system VoiceTranscriber, to semi-automate the process of summarizing stories and extracting meaningful tags. Usability evaluation results from four storytelling processes, and participant feedback revealed that VoiceTranscriber is easy to use and learn, and that the summarized stories have high factuality. However, some transcribers consider that the VoiceTranscriber tasks are challenging and involve a heavy workload. In the future, we will focus on improving the completeness and coherence of the summarized stories and on reducing the perceived effort and mental demands placed on the transcribers.

Acknowledgments

This work was supported in part by the Ministry of Science and Technology, National Taiwan University, and Intel Corporation under Grants MOST 103-2627-E-002-001, MOST 103-2911-I-002-001, NTU-ICRP-104R7501, NTU-ICRP-104R7501-1 and NTU-ICRP-104R890861.

References

Silsbee, P. L.; and Bovik, A. C. 1996. Computer lipreading for improved accuracy in automatic speech recognition. *Speech and Audio Processing, IEEE Transactions* 4(5): 337-351.

Imai, T.; Matsui, A.; Homma, S.; Kobayakawa, T.; Onoe, K.; Sato, S.; and Ando, A. 2002. Speech recognition with a re-speak method for subtitling live broadcasts. In *Intl. Conf. on Spoken Lang. Processing, ICSLP-2002,* 1757–1760.

Lasecki, W.; Miller, C.; Sadilek, A.; Abumoussa, A.; Borrello, D.; Kushalnagar, R.; and Bigham, J. 2012. Real-time captioning by groups of non-experts. In Proceedings of the 25th annual ACM symposium on User interface software and technology, 23-34.