A Wearable Digital Library of Personal Conversations

Wei-hao Lin
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213-3890 USA
+1 412 268 4757
whlin@cs.cmu.edu

Alexander G. Hauptmann
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213-3890 USA
+1 412 268 1448
alex@cs.cmu.edu

ABSTRACT
We have developed a wearable, personalized digital library system, which unobtrusively records the wearer's part of a conversation, recognizes the face of the current dialog partner and remembers his/her voice. The next time the system sees the same person's face and hears the same voice, it can replay the audio from the last conversation in compressed form summarizing the names and major issues mentioned. Experiments with a prototype system show that a combination of face recognition and speaker identification can be effective for retrieving conversations.

Keywords
Personal Digital Library, conversation capture, speaker identification, face recognition, augmented human memory

Real-time information retrieval as a wearable personal memory aid

Our research aims to augment human memory through a personal digital library of experiences. The long-term vision is to allow people to capture and retrieve from a complete audio, video and textual record of their personal experiences and electronic communications. This assumes that within ten years technology will be in place for creating a continuously recorded, digital, high fidelity record of one's whole life in video form [2]. Wearable, personal digital library systems units will record audio, video, GPS and electronic communications. This research fulfills the vision of Vannevar Bush's personal Memex [1], capturing and remembering whatever is seen and heard, and quickly returning any item on request

In the following sections we describe a first implementation of a personal digital library system for remembering people and conversations. There are two modes in the system, collection the conversations for a personal digital library and retrieving summaries of conversations from the personal library.

Personal Memory Collection

The hardware is designed as a wearable device consisting of a miniature 'spy' camera, a cardioid lapel microphone and an omni-directional microphone, all attached to a laptop computer in a backpack. When the system is prompted to collect personal conversations, it attempts to detect the face of the person you are talking to in the video stream, and listen to the conversation from both the close-talking (wearer) audio track and the omni-directional (dialogue partner) audio track.

The close-talking audio is transcribed by a speech recognition system to produce a rough, approximate transcript. The omni-directional audio stream is processed through a speaker identification module. An encoded representation of the face of your current dialog partner, the dialog partner speaker characteristics, and the raw audio of the current conversation is saved to a database.

The audio is further processed through audio analysis (silence removal, emphasis detection) and general speech recognition to efficiently replay only the person names and the major issues that were mentioned in that conversation.

Personal Memory Retrieval

In the retrieval (remembering) mode, the system searches for a face in the video stream and performs speaker identification on the omni-directional audio stream. When either a face is detected and/or a speaker is identified, the face and speaker characteristics will be matched to instances of faces and speaker characteristics stored in the personal memory library database. A linear interpolation, which has been shown effective among different classifier combining strategies [4] is used to combine the probabilities of two face recognition and one speaker identification modules. When a sufficiently high scoring match is found, the system will return a brief summary of the corresponding recorded conversation with the person. Figure 1 shows the process of personal memory retrieval.

Extracting and Retrieving Metadata from Conversations

Face Detection and Recognition

Face detection and matching was used in the CMU NameIt system [5] using the 'eigenface' approach. Meanwhile there have been several commercial systems offering face detection and identification, such as Visionics [7]. In our implementation we are using both the Visionics FaceIt toolkit for face detection and matching as well as, the Schneiderman face detector [6] and 'eigenfaces' [5] for matching similar faces.

Speaker Identification

Speaker identification is done through our own implementation of Gaussian Mixture Models [8]. The speaker identification system also uses the fundamental pitch frequency to eliminate false alarms. About four

seconds of speech are required for reliable speaker identification under benign environmental conditions.

Information Summarization

The audio stream to be summarized is selectively compacted based on power. Similar to the video skims [9] and audio summarization [3], only selected portions of the audio are played. Silences are used to define 'cuts', and low signal-to-noise ratio segments are eliminated. We have also implemented a TF.IDF weighting scheme to rank segments based on transcript words. So far the system tends to play back too much of the conversation, forcing the user to actively interrupt the summary playback especially when the conversation was not with the current dialog partner.

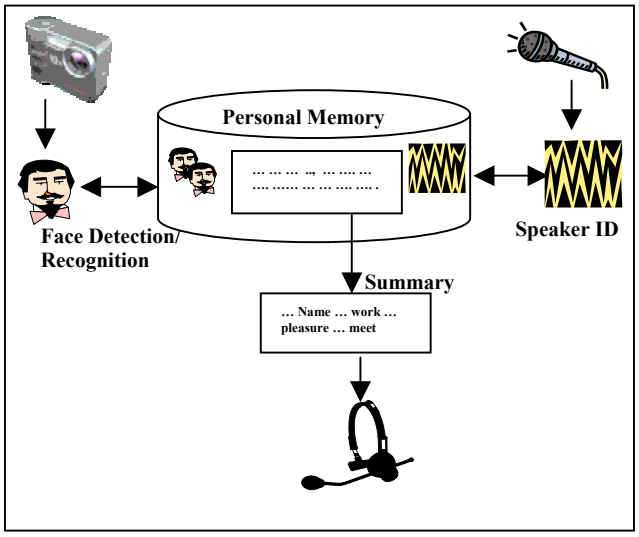

Figure 1. Process for Personal Memory Retrieval

Experimental Data

We collected two conversations each with 22 people while wearing our prototype capture unit. Each conversation was analyzed for faces and speaker audio characteristics as described above. The first of each conversation served as the example inside personal digital library, while the second conversation was used as the query or retrieval prompt to 'remember' the first conversation. We used an average of about five seconds at the beginning of each conversation for creation of the library and as the retrieval query.

We used the average rank as the retrieval metric, i.e. on average, at what rank was the correct conversation found.

Retrieval Results

The results of our experiment in Table 1 show that the Schneiderman/Eigenface detection/recognition method retrieved the correct conversation at an average rank of 3.42 of the 22 possible conversation candidates. The Visionics face recognition system found the correct conversation only at rank 4.5. Speaker identification proved to be slightly more reliable with an average rank of 3.09. A linear combination of the three types of evidence found the correct conversation at rank 2.59. Thus in a library of 22 personal conversations, after listening to 2 conversation summaries, you would likely find the correct one.

Table 1 Average rank of the correctly remembered conversation summary.

Retrieval Method	Average Rank
(Schneiderman + Normalized Eigenfaces)	3.42
(Visionics Face Detection)	4.50
Speaker Identification	3.09
Combined Evidence	2.59

Discussion

The main focus of our system is the integration of multi-modal human experience in a personal digital library. The novelty of the system is in using the face and the audio cues to help remember essential details about the previous meeting with the same person, and automatically creating a personal digital library of information associated with the face, the voice and the words. Eventually an intelligent assistant drawing from an annotated personal history could overcome age and other limits to mental capacity and help recall the details needed in a given situation.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation Information Technology Research under grant number IIS-0121641.

REFERENCES

1. Bush, V., As we may think, Atlantic Monthly, Vol.176, No. 1; pages 101-108, 1945
2. Gray, J., What next? A few remaining problems in Information Technology, *ACM Federated Research Computer Conference*, Atlanta, GA, May 1999.
3. Arons, B.M., *Interactively Skimming Recorded Speech*, Ph.D. Dissertation, MIT, February 1994.
4. J. Kittler, M. Hatef, R. Duin , and J. Matas, On Combining Classifiers, *IEEE Trans. Pattern Analysis Machine Intelligence*, 20(3), 1998
5. Satoh,S., and Kanade,T. NAME-IT: Association of Face and Name in Video. *IEEE CVPR97,* Puerto Rico, 1997.
6. Schneiderman, H. and Kanade, T. Probabilistic Modeling of Local Appearance and Spatial Relationships of Object Recognition, *IEEE CVPR*, Santa Barbara, 1998.
7. Visionics FaceIt, Face Recognition Software, www.visionics.com, 2001
8. Schmidt, M., Golden, J., and Gish, H. "GMM sample statistic log-likelihoods for text-independent speaker recognition," *Eurospeech-9*, Rhodes, Greece, September 1997, pp.855 - 858.
9. Smith, M. and Kanade, T. Video skimming and characterization through the combination of image and language understanding techniques, *IEEE CVPR97,*(San Juan, Puerto Rico, 1997), 775 – 781.