A Speaker Identification Agent

Luc E. Julia
Larry P. Heck

SRI International
STAR Laboratory
333 Ravenswood Ave.
MENLO PARK, CA 94025 - USA
+1 415 859 42 69
{julia,heck}@speech.sri.com

Adam J. Cheyer

SRI International
Artificial Intelligence Center
333 Ravenswood Ave.
MENLO PARK, CA 94025 - USA

cheyer@ai.sri.com

Abstract. This paper describes a prototype application which combines speaker identification technology and an agent architecture to provide user-definable monitors for incoming voicemail messages. Through a Web-distributable Java user interface, the user may enter requests by using spoken or typed natural language. Multiple distributed agents process the requests, periodically testing the user's voicemail system to identify the composer of the message from a set of selected speakers. When a message meets the conditions specified by the user, agents locate the requester's position and notify him or her of the arrival of the important message by using various communication media (email, fax, telephone, pager). The technology responsible for identifying a speaker from voice is a text-independent method developed at SRI International. Encapsulating this capability as an agent permits plug-and-play reusability in the growing number of applications being developed within the agent-based framework.

1 Introduction

For several years, SRI International has been using agent architectures as a means of developing systems by bringing together technologies such as speech recognition, natural language understanding, planning and heterogeneous database access. In one such application, the Automated Office Assistant [1], ten or more distributed agents provide remote access to conventional applications such as databases, calendars and email, through various input modalities, including handwriting, gesture and speech. Agents implemented in SRI's Open Agent Architecture (OAA) [6] hide the complexities of finding, accessing and combining data to meet a user's requests; users need not know where their requests are being executed, nor how. In addition, the architecture allows independence and cooperation between agents -- as agents dynamically connect to the system, new capabilities become available to the user.

Recently, SRI's speech laboratory developed an original set of methods to identify a speaker by his or her voice [4]. The Speaker Recognition System does not require knowledge or constraints on the spoken text, and is therefore suited for identifying persons from conversational-style voicemail messages. In addition, the system is designed to operate with an open set of speakers, meaning that the system is capable of sorting through known as well as unknown speakers to detect the desired individual. In addition, the Speaker Recognition System is capable of detecting messages, from individuals, that were recorded with different telephone equipment

(e.g., handset type) than was used in training. Finally, the system is capable of operating in real time, independent of the number of speakers known to the system.

This paper describes a first prototype application that combines speaker recognition technology as an OAA agent in the context of an Automated Office Assistant task.

2 The Speaker Identification Demonstration

To demonstrate the capability of the Speaker Identification Agent in the Automated Office Assistant application domain, we imagined the following scenario:

The owner of the user interface show in Fig. 1, Larry Heck, is waiting for an important call from his wife, but he has a number of meetings and cannot wait in his office. After consulting his calendar to see where his day will start, Larry says to his computer: "If voicemail arrives for me from my wife, get it to me immediately", and then he leaves.

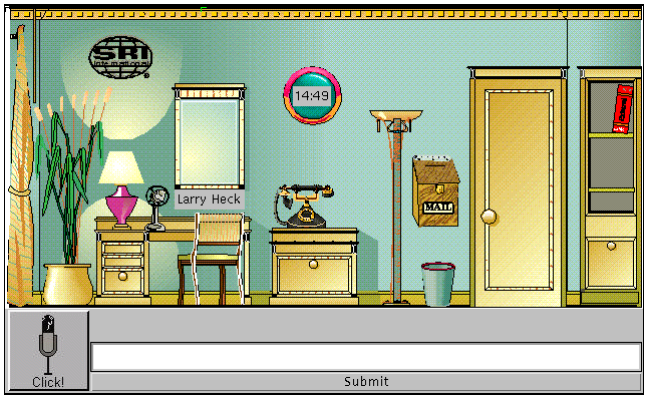

Fig. 1. Office Assistant user interface for the Speaker Identification demonstration

As the user's request is translated and interpreted, a monitor is added to the distributed agents in the system (Fig. 2), indicating that every minute or so, Larry's voicemail should be checked for new messages meeting the specified test conditions.

Later, Larry's wife calls the office, the voicemail system answers and she leaves an important message. Within the following minute, the Voicemail Agent is asked by the application to check Larry's voicemail. A task is delegated by the Facilitator to the Telephone Agent to call the voicemail system's telephone number, and enter Larry's identification and password (retrieved by the Database Agent) as touchtones. Speech recognition and touchtone generation are used to navigate the voicemail system's information space, cycling over all new voice messages. Each new message retrieved by the Speech Recognition agent is analyzed by the Speaker Identification Agent and if the target voice is recognized, the installed trigger fires.

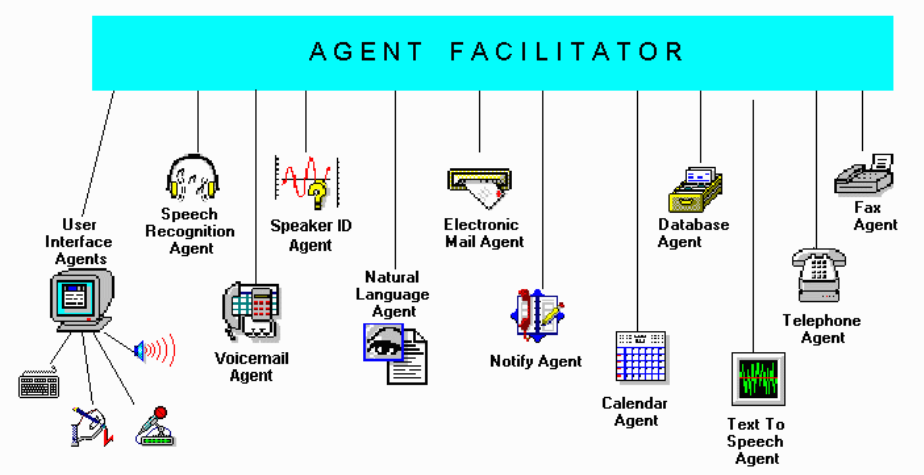

Fig. 2. Software agents used in Speaker Identification application

In the described scenario, when the trigger fires, the Notify Agent, which reasons about the domain of information transfer, attempts to get the message to Larry. In response to a query by the Notify Agent, Larry's location is found by the Calendar Agent to be a certain meeting room, and the phone number of this room is returned by the Database Agent. The Notify Agent then instructs the Telephone Agent to dial the number, ask for Larry Heck (with help from a text-to-speech agent), verify Larry's identity by a touchtone password, and when all is in order, play the message over the phone. If Larry's location is not reachable by a telephone, the agent system can notify Larry of the arrival of the message by fax, email, pager or any other media managed by the dynamic set of agents currently available on the network.

3 Speaker Recognition System

The SRI Speaker Recognition System consists of two main components: a front-end processing component that extracts the most discriminating speech features from the speech waveform, and a classifier that uses the extracted features to decide whether or not the speech is from a particular individual. In this way, the system is similar to a verification system since it makes a binary decision between two hypotheses: "speaker of interest", or "not speaker of interest". However, the difference between this system and a verification system is that the speaker does not claim an identity, but rather the system claims the identity (i.e., claiming the identity of the speaker of interest).

The front-end processing of the speech occurs in several steps. First, the speech is segmented into frames by a sliding 25ms window progressing at a 10ms frame rate. Next, mel-scale cepstral feature vectors are extracted from the speech frames [4]. In this work, we use 17th-order mel-cepstra, with the 0th-order term removed. Finally, the feature vectors are channel equalized with a blind deconvolution. The deconvolution is implemented by subtracting the average cepstral vector from each input phone message. This deconvolution step is necessary to minimize the effect of

training under different conditions than testing (e.g., different handsets, telephone lines).

The classifier for the Speaker Recognition System is based on EM-trained Gaussian mixture models (GMMs). The GMMs are used to represent the acoustic parameter distribution of each claimant speaker,

$$p\left(\vec{x}_t \middle| \lambda_k\right) = \sum_{i=1}^{M} p_i^k b_i^k\left(\vec{x}_t\right)$$

where p_i^k and b_i^k are the mixture weight and the Gaussian density for the i^{th} mixture out of M for speaker k [7]. The average log-likelihood of a claimant speaker given an utterance $X = \left\{\vec{x}_1 \ldots \vec{x}_T\right\}$ is computed as

$$\mathrm{L}\left(X \middle| \lambda_k\right) = \frac{1}{T}\sum_{t=1}^{T} \log p\left(\vec{x}_t \middle| \lambda_k\right)$$

For open-set voicemail speaker detection, a likelihood ratio detector is used that *normalizes* the score of the claimant speaker by the score of a single composite model of all other impostor speakers [4]. In the log-domain, the ratio can be expressed as a difference of terms,

$$\Lambda\left(X|k\right) = \mathrm{L}\left(X \middle| \lambda_k\right) - \mathrm{L}\left(X \middle| \overline{\lambda_k}\right)$$

where $\overline{\lambda_k}$ denotes the composite impostor model. The composite model differs from a cohort modeling approach [5 and 8] in that it is a single, high-order speaker-independent GMM trained with speech from a large number of speakers (speech data from 90 speakers were used). The term *composite* refers to the fact that aspects of many persons' voices are combined into one model rather than using speaker-dependent cohort models. Composite models are much simpler to implement than cohort models, and have been shown to outperform cohort models on the May 1995 and March 1996 NIST Evaluation Corpora from Switchboard [4]. For the voicemail system described in this paper, we empirically determined the best number of mixtures in the composite model to be 2048 Gaussians.

4 Training and Execution

The speaker identification technology requires training to acquire a set of registered "known" users of the system. For our demonstration system, each targeted person recorded 20 sentences and 20 sequences of numbers, for a total of about 4 min of speech. The input text was chosen to provide good phonetic coverage. Once the system has been run over the training data, the Speaker Identification Agent is capable of recognizing voices given as little as 5 s of text-independent speech.

The performance of our system was evaluated in the March and July 1996 NIST speaker recognition benchmarks [4]. The evaluations utilized the Switchboard database, which is a collection of long-distance telephone conversations with unconstrained vocabulary. The test was similar to the voicemail task described in this paper, although somewhat more difficult because less data was used to train the system (2 min). In those evaluations, the SRI system had an equal error rate of 5-8% for 10-s utterances, depending on whether or not the training and testing handsets were matched.

Since the system is implemented by using open-set technology, likelihoods from only one target speaker model (the speaker of interest) and the composite normalizing model must be computed. In this way, the execution time of the system does not depend on the population size of voicemail users. This differs from closed-set approaches, where the execution time grows with the number of speakers known to the system. As a result, the SRI Speaker Recognition System is capable of running faster than real time.

5 Current and Future Work

Because the Speaker Recognition System is now an agent, it will be easy to incorporate it in other applications, especially those that are already designed for spoken or multimodal input within the OAA framework. We are currently applying this within an existing multimodal (pen and voice) map-based application [2]. In this application, no matter what machine or location the user is running the application from, the system will set up the user's preferences according to the identification made during the first interactions. Since this application supports collaborative work, the system will provide a secure way to identify and to give rights to make changes to authorized users according to their levels of security in the system. In addition, to provide secure multimodal interactions, we can use the Speaker Identification Agent to perform speaker verification continuously during interactions with the system. As a user interacts with remote databases through natural combinations of speech, handwriting and drawn gestures, speaker verification is performed on all utterances by using the Speaker Identification Agent set with a high correlation threshold. In parallel, selected handwriting and gesture inputs are analyzed using SigCheck, CIC's dynamic handwriting verification technology [3].

Another application that we are ready to put together is an intelligent answering machine. Based on the fact that people usually talk for more than 5 s when they leave a message, and that the Speaker Recognition System is faster than real time, this application could, once the speaker is recognized, deliver personalized messages to a set of specific callers.

Now that the speaker identification technology seems to be good enough and that we have an actual system to rely on, it opens the field to multiple applications.

6 References

[1] Automated Office Assistant System: http://www.ai.sri.com/cgi-bin/oaa/office.pl

[2] Cheyer A. and Julia L. (1995). Multimodal Maps: An Agent-based Approach. CMC'95 : Eindhoven, Netherlands, pp. 103-113.
[3] CIC Signature Sentinel™: http://www.cic.com
[4] Heck L.P. and Weintraub M. (1997). Handset-Dependent Background Models for Robust Text-Independent Speaker Recognition. Submitted to IEEE Proc. Intern. Conf. on Acoust., Speech, and Signal Proc.: Munich, Germany.
[5] Higgins A., Bahler L. and Porter J. (1992). Speaker verification using randomized phrase prompting. Digital Signal Processing, Vol. 1, pp. 89-106.
[6] Open Agent Architecture™: http://www.ai.sri.com/~oaa
[7] Reynolds D.A. (1995). Speaker identification and verification using Gaussian mixture speaker models. Speech Communications, Vol. 17, pp. 91-108.
[8] Rosenberg A.E., DeLong J., Lee C.H., Juang B.H. and Soong F.K. (1992). The use of cohort normalized scores for speaker verification. IEEE Proc. Intern. Conf. Speech and Signal Proc., pp. 599-602.