

Integration of Speaker Recognition into Conversational Spoken Dialogue Systems*

Timothy J. Hazen², Douglas A. Jones¹, Alex Park², Linda C. Kukolich¹, and Douglas A. Reynolds¹

¹Information Systems Technology Group
MIT Lincoln Laboratory
Lexington, Massachusetts, USA
{daj, kukolich, dar}@ll.mit.edu

²Spoken Language Systems Group
MIT Laboratory for Computer Science
Cambridge, Massachusetts, USA
{hazen, alex}@sls.lcs.mit.edu

Abstract

In this paper we examine the integration of speaker identification/verification technology into two dialogue systems developed at MIT: the Mercury air travel reservation system and the Orion task delegation system. These systems both utilize information collected from registered users that is useful in personalizing the system to specific users and that must be securely protected from imposters. Two speaker recognition systems, the MIT Lincoln Laboratory text-independent GMM based system and the MIT Laboratory for Computer Science text-constrained speaker-adaptive ASR-based system, are evaluated and compared within the context of these conversational systems.

1. Introduction

Many conversational dialogue systems can be personalized in order to improve the interaction between the user and the system. Typically, these systems also require a certain level of security in order to prevent unauthorized users from gaining access to sensitive information or services. Speaker identification technology can be a useful tool for helping achieve the goals of these personalized dialogue systems.

In this paper we first discuss the issues of incorporating speaker identification into conversational dialogue systems. In particular, we have examined the integration of speaker recognition capabilities into two existing conversational dialogue systems: the Mercury air travel system [1] and the Orion task delegation system [2]. These systems share the property that a user's identity need not be confirmed immediately at the onset of a conversation. Instead, the system can continue to collect speech samples from the user's interaction until such time that the user requests an action or information that requires security.

This paper then discusses two approaches to speaker identification: a text-independent Gaussian mixture model (GMM) approach developed at MIT Lincoln Laboratory and a speaker adaptive automatic speech recognition (ASR) approach developed at MIT Laboratory for Computer Science. We examine the strengths and weaknesses of these systems in the context of conversational interaction. We then evaluate these systems on data collected from the Mercury system, which has a set of known registered users whose system usage varies from *occasional* to *very frequent*. We also examine the potential benefits of combining these two approaches.

2. Integrating Speaker Recognition in Dialogue Systems

In order to ensure both ease of use and security, it is important to consider the nature of typical dialogues within a system when devising its security strategy. For example, for an air travel reservation system, such as the MIT Mercury system, it may not be necessary to confirm the identity of the user until the user requests the execution of a secure transaction (e.g., *"please book this flight and bill it to my credit card"*). In cases like this, the system could elect to continue on with the conversation, even when it is uncertain of the user's identity, as long as the user is only performing actions that don't require security (e.g., browsing flight options or comparing prices). This strategy would allow the system to collect additional speech data, which could be useful in improving the reliability of the speaker identification.

While security is often the primary consideration when incorporating speaker recognition technology, the issue of convenience should not be overlooked. In some systems, security is desirable but is not as crucial as convenience or ease of use. For example, consider the following simple task delegation request that can be handled by the Orion task delegation system: *"Hi Orion, it's Doug. Call me at 4:30 PM and tell me if United flight 43 is on time."* The cost of allowing an imposter to execute this task is not nearly as severe as the cost of allowing an imposter to illegally bill charges to a user's credit card. Additionally, it is an obvious inconvenience for a known user of the Orion system to have to proceed through a login process in order to make this simple, single utterance request. In this case, it is preferable for the system to automatically determine the user without engaging in a login sub-dialogue. Only when the system is uncertain of the identity of the user should it prompt them for additional security information such as a password. For a system like Orion, an open-set identification approach can be employed, thereby avoiding the use of an explicit login sub-dialogue and increasing the convenience and ease of use of the system.

Another issue that must be addressed when developing a speaker recognition approach for a publicly available conversational system is the potential variability in the amount of training data available for each speaker. A typical system may have a mix of *power users* who call the system frequently and *occasional users* who call less frequently. Under these circumstances, the speaker recognition modeling technique

*This work is sponsored by the Department of Defense under Air Force Contract F19628-00-C-0002 and the DARPA/IPTO Cognitive Systems Initiative Study. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. This research was also supported by an industrial consortium supporting the MIT Oxygen Alliance.

must be designed to perform adequately for the occasional user but be capable of steadily improving as the amount of training data per speaker increases.

Finally, it is important to understand the constraints that may be placed upon users of a dialogue system. For example, the Mercury system utilizes a mixed-initiative approach where, at some points in the dialogue, the system places tight constraints on what the user can say by asking questions like “What is your user name?”, while at other points, the system allows the user to speak in a highly unconstrained manner by prompting the user with queries such as “How can I help you?” Under these circumstances, the optimal speaker recognition strategy may vary (i.e., text-dependent vs. text-independent) depending on the current state of the dialogue.

3. Overview of Speaker Recognition Techniques

In this paper we examine the integration of two types of speaker recognition systems into the dialog system. The first is the MIT Lincoln Laboratory Gaussian Mixture Model Universal Background Model (GMM-UBM) system [3] primarily designed for text-independent recognition. The second system is the MIT LCS Speaker Adaptive ASR-based system [4], whose design assumes the presence of an accurate speech recognizer enabling the use of text-dependent speaker modeling within constrained applications. It is hoped that by combining the two systems we can achieve a better tradeoff between text-dependent and text-independent behavior.

3.1 The GMM Approach

The Lincoln speaker recognition system is designed primarily for text-independent speaker verification tasks. As such, there is no explicit modeling of particular speech sounds, but rather a general distribution is used to implicitly model the underlying sounds in a person’s speech as found in the distribution of acoustic observations (or feature vectors) extracted from the speech signal. Specifically each speaker is represented by a GMM, M^s , over the feature vectors x ,

$$p(x | M^s) = \sum_{i=1}^N p_i^s b_i^s(x),$$

with the parameters, $(p_i^s, \mu_i^s, \sigma_i^s)$, of weights, means and variances. Since a speaker’s training speech will not generally cover all sounds in sufficient instances to adequately train a GMM that is robust to new or unseen sounds, the GMM model parameters are derived by adapting a well-trained speaker-independent background model (also called a universal background model). For example the mean parameter for mixture i of a speaker’s GMM would be estimated as

$$\hat{\mu}_i^s = \lambda_i E(x | i) + (1 - \lambda_i) \mu_i^b$$

where $E(x | i)$ is the expectation of the training data in mixture i , μ_i^b is the background model mean for mixture i , and $\lambda_i = n_i / (n_i + \tau)$ is an adaptation factor for mixture i . In the adaptation factor, n_i is the count of feature vectors in mixture i and τ is a tuning parameter set to 16 in the GMM system.

For this paper, a 2048 mixture background model is used and, based on previous experiments, only mean parameters are adapted in the speaker models. Feature vectors of 38

dimensions are extracted every 10 ms and consist of 19 mel-warped cepstra, derived from the frequency band 300-3300 Hz, and their first order derivatives, estimated with a 5-frame window. To compensate for linear channel effects (possibly time-varying), standard RASTA filtering is applied to the cepstra elements.

For verification, the log likelihood of the input speech utterance is computed against both the background and speaker models, the difference taken and compared to a threshold to decide whether to accept or reject the putative speaker claim.

3.2 The Speaker Adaptive ASR Approach

A spoken utterance can be viewed as a sequence of phonetic events. Although diverse phonetic events can have very different acoustic characteristics, the necessity of text independence precludes most speaker ID systems from using phone-dependent modeling techniques. Thus, text-independent systems operate under the assumption that the phonetic content of the utterance is unknown. In [4], two systems were described that relaxed this assumption by making use of ASR output during speaker recognition.

The speaker adaptive recognition approach uses speaker-dependent speech recognizers to model each speaker. During training, phonetically transcribed enrollment utterances are used to train the speaker-dependent context-dependent phone models for each speaker. During testing, an ASR component generates a phonetic transcription from the test utterance. This transcription is then used by the system to score each segment of speech against a speaker-dependent phone model.

Modeling speakers at the phone level can be problematic because enrollment data sets are typically not large enough to build robust speaker-dependent models for every context-dependent phone model. To compensate for this difficulty, we use an adaptive scoring approach in which the speaker-dependent score is interpolated with a speaker-independent score.

Mathematically, if the word recognition hypothesis assigns each feature vector x to phone j , then the score for speaker S_i is given by

$$\frac{1}{|X|} \sum_{x \in X} \log \left(\frac{\lambda_{i,j} p_{SD}(x | M_j, S_i) + (1 - \lambda_{i,j}) p_{SI}(x | M_j)}{p_{SI}(x | M_j)} \right)$$

where M_j is the model for phone j and $\lambda_{i,j}$ is an interpolation factor given by

$$\lambda_{i,j} = \frac{n_{i,j}}{n_{i,j} + \tau}$$

In this equation, $n_{i,j}$ is the number of training tokens of phone j observed for speaker S_i and τ is a global tuning parameter that is set empirically using a separate development set.

This scoring strategy results in models that capture the detailed phone-level characteristics of a speaker when sufficient training data is available, but relies more on speaker independent models for phones with sparse training data. In other words, the system backs off to a *neutral* score, as provided by the speaker-independent model, when limited training data is available.

4. Experiments

The experiments presented in this paper were conducted using calls collected from users of the Mercury air travel reservation system and the Orion task delegation system. These systems allow new users to register through an enrollment process to create an account that is accessed when the user calls the system. Information stored within the user's account is used to personalize various constraints and models used during the dialogue. Mercury also allows unregistered users to utilize the system anonymously in a speaker independent mode to browse flight schedule information.

When a new call is placed to the system, the caller is first prompted to provide their name or specify that they are not a registered user. If a caller provides a name, the system then prompts for the user's password (which is in the form of a month-day combination). Thus, the first two user utterances within each call made by a regular user are usually constrained to be the user's name and password.

4.1 Mercury/Orion Speech Corpus

For our experiments, the 44 most frequent registered users of Mercury and Orion were selected to represent the set of "known" users. Each of these users spoke a minimum of 48 utterances within the calls representing the training set for our experiments. As would be expected in real-world applications, the amount of training data available for the known users varied greatly based on the frequency with which they used the systems. Of the 44 speakers, 15 had less than 100 utterances available for training, 19 had between 100 and 500 training utterances, and 10 speakers had more than 500 utterances for training. The most frequent user of the system contributed 2550 utterances for his training set. Within the 44 speakers, 21 were females and 23 were males.

In addition to the training data for each known user, an additional set of 20,491 Mercury utterances was used to train the universal background model in the Lincoln Laboratory GMM system. The speech recognition system used by the speaker adaptive ASR approach was trained on over 130,000 utterances collected from a variety of publicly available systems deployed at MIT.

For the test set, all calls made to the Mercury system during a 10-month span were held out for our evaluation set. Only calls containing at least five utterances were included in the evaluation set. Additionally, to remove the issue of accurate speech/non-speech detection, only utterances that were manually determined to contain at least one spoken word were included in the evaluation.

The evaluation set is further broken down into two subsets: a set of 304 calls containing 3705 utterance from members of the set of 44 known users and a set of 238 calls containing 2946 utterances from speakers not in the known speaker set. Each call has a variable number of utterances with an average of 12 utterances per call (stdev=6 utts) where the average utterance duration is 2.3 sec (stdev=1.4 sec). This evaluation data allows us the freedom to evaluate the system in a variety of ways, such as closed-set speaker identification, open-set speaker identification, and speaker verification.

Because our data collection effort was not engineered to achieve any pre-specified goals, the distribution of calls from specific speakers was not controlled by any means other than the users' own desire to use the system. As such, the

distribution of calls over the known set of users is widely varied. Only 21 of the 44 speakers in the known set made calls to Mercury during the period of time in which the evaluation data was collected. Because a small subset of users were highly frequent users of the system, 46% of the calls from known users came from only 3 members of the known speaker set. Despite the uneven distribution of speakers, we feel the evaluation set is reflective of the type of data publicly deployed systems could expect to encounter.

4.2 Experimental Results

For our first experiment, we elected to use an open-set speaker identification paradigm, i.e., for an input utterance (or sequence of utterances) determine whether the speaker is one of a set of enrolled speakers or an unknown (imposter)¹ speaker. This gives rise to three types of errors:

- *False-accept*: the system accepts an imposter speaker as one of the enrolled speakers.
- *False-reject*: the system rejects a true user.
- *Confusion*: the system correctly accepts a true user but confuses him with another enrolled user.

For the speaker recognition systems we plot the trade-off between these errors as a function of a decision threshold. Although more sophisticated decision logic may be deployed, both systems operate by selecting the highest scoring speaker model for an input and comparing its score to a threshold to determine whether to accept or reject.

To simulate two manners in which speaker identification could be used in the Mercury system, we examine the performance on the first utterance of each call (which is the user stating their user name) and over the entire length of the call. This allows us to examine the advantage of delaying the speaker identity decision until the user has selected a flight itinerary and requests that it be "booked". Figure 1 and Figure 2 show the error tradeoff for the GMM and ASR systems for the first utterance and the complete call from the Mercury data. The equal error rate between misses and false alarms is similar with the two systems, at approximately 7% for the first utterance and 5% for the whole call.

The ASR-based system does appear to have an advantage in reduced confusion rate within the known speaker set, Pr(Conf), for the first utterance case, where the ASR-based system's confusion rate is 1.6% (5 errors in 304 calls) while the GMM system has a confusion rate of 5.6% (17 errors in 304 calls). This disparity is likely because the first utterance is constrained to be the user stating their user name. This provides a tight constraint on the phonetic content of this utterance in both the training and evaluation calls, which favors the phonetic modeling of the ASR-based approach.

However, if we examine the in-set confusion rate over the course of the call, the ASR-based system fails to achieve any further improvement while the confusion rate of the GMM system improves to 2.3% (7 errors in 304). While the small number of errors in each case prevents any definitive conclusions, the trend implies that the text-independent GMM may retain a higher level of robustness in the face of the unconstrained data that is typically encountered by the Mercury system after the first two user utterances.

¹ For this data there were no dedicated imposters attempting to break into another's account. Thus effects like an imposter saying a known user's name during login are not evaluated.

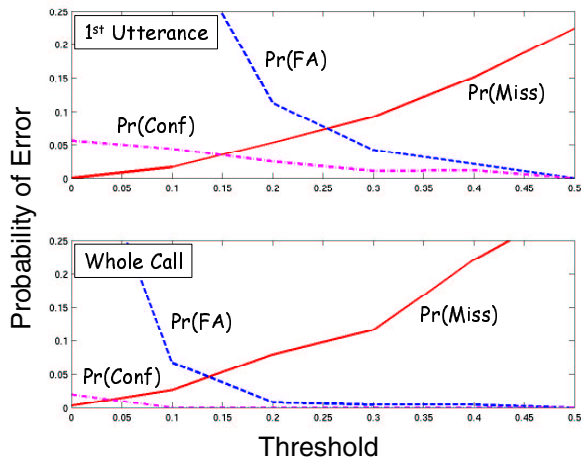


Figure 1 Tradeoff of false-rejection rate, $Pr(Miss)$, against false acceptance rate, $Pr(FA)$, and in-set confusion rate, $Pr(Conf)$, as a function of acceptance threshold for the GMM system for the first utterance only (top) and over the entire call (bottom).

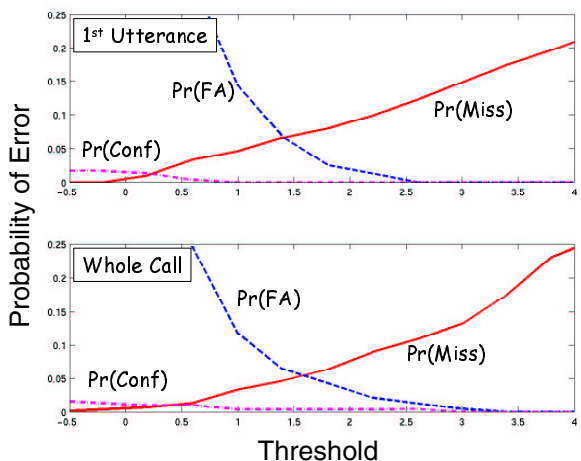


Figure 2 Tradeoff of $Pr(Miss)$ against $Pr(FA)$ and $Pr(Conf)$ as a function of acceptance threshold for the ASR-based system.

To examine the benefit of fusing the two speaker ID techniques, we employed a 10-fold cross-validation paradigm over the test set for determining optimal linear fusion weights. The fusion weights were learned using multi-layered perception (MLP) training with no hidden layers. Initial experiments implementing fusion for the first utterance and whole call cases produced modest but statistically insignificant improvements, likely because the test set size at the call level is small and the number of errors is few. To examine the issue further we examined the performance over all individual utterances in the test data. This evaluation is much harder than the per-call evaluation because many of the utterances in the data are either very short (as little as a single word) or highly spontaneous (unlike the first utterance of each call which is tightly constrained).

Table 3 Performance of the GMM system, ASR-based system and a fused system on the tasks of closed-set speaker recognition and speaker verification.

	Closed-Set ID Error Rate	Verification Equal Error Rate
GMM System	10.4%	6.07%
ASR System	9.5%	6.72%
Fused System	7.3%	4.83%

Table 1 shows the performance of the GMM system, the ASR-based system and the fused system on the task of per-utterance closed-set speaker recognition over the 3705 utterances in the known-speaker set, and on the task of per-utterance speaker verification over the full test. On these tasks, fusion provides clear improvements in performance, demonstrating the advantage of merging our two complementary techniques. In future work we would like to investigate a fusion algorithm that can be adapted to account for contextual factors, such as the local dialogue state, and training data considerations, such as the amount of available training data for the current speaker.

5. Summary

This paper has investigated the issues of integrating speaker recognition into spoken dialogue systems. In some dialogue systems, like the Mercury air travel system and the Orion task delegation system, the constraints of the system allow the final decision on the identity of a user to be delayed until a secure action requiring confirmation of the user's identity is requested. The determination of the speaker recognition operating point and confirmation strategy is highly dependent on the goals of each individual system. Developers must find the optimal trade-off between security and convenience.

We also evaluated two speaker recognition systems, the MIT Lincoln Laboratories GMM-UBM system and the MIT LCS speaker adapted ASR system, on data collected by the Mercury system. Both systems performed comparably on a per-call open-set identification evaluation within the Mercury air travel system achieving an equal error rate of 5% between false rejections of known users and false acceptances of imposters. We have also demonstrated on a closed-set speaker recognition evaluation and a speaker verification evaluation, that the two complementary speaker recognition techniques can be fused to provide additional performance improvements.

As a final note, the authors wish to thank Chao Wang and Stephanie Seneff for their efforts and help on this project.

6. References

- [1] Seneff, S. and Polifroni, J., "Dialogue Management in the Mercury Flight Reservation System", *Satellite Dialogue Workshop of the ANLP-NAACL Meeting*, April 2000.
- [2] Seneff, S., Chuu, C. and Cyphers, D. S., "Orion: From On-line Interaction to Off-line Delegation", in *Proc. of ICSLP*, Beijing, China, October 2000.
- [3] Reynolds, D.A., Quatieri, T.F., and Dunn, R.B., "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing Review Journal* 2000
- [4] Park, A. and Hazen, T. J., "ASR Dependent Techniques for Speaker Identification", in *Proc. of ICSLP*, Denver, CO, September 2002, pp. 1337-1340.