MULTI-MODAL FACE AND SPEAKER IDENTIFICATION ON A HANDHELD DEVICE

Timothy J. Hazen, Eugene Weinstein, Ryan Kabir, Alex Park

MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139, USA

Bernd Heisele

Honda Research Institute USA, Inc.
Boston, MA 02111, USA

ABSTRACT

In general, most systems for face and speaker identification are tested on high quality data collected in well-lit and quiet environments. In this study, we investigate the application of existing face and speaker identification techniques to the task of user authentication on a handheld device. In this context, the audio/visual capture hardware is of lower quality than equipment typically used in laboratory experiments. Additionally, variable background conditions which can degrade the audio/visual signal may be present. These factors can be expected to harm the performance of the system. Under these circumstances, using a combination of biometric modalities can improve the robustness and accuracy of the person identification task. In this paper, we present our approach for combining both face and speaker identification technologies on a handheld device, and experimentally demonstrate a fused multi-modal system which achieves a 90% reduction in equal error rate over the better of the two independent systems.

1. INTRODUCTION

This paper investigates the integration of two biometric techniques, face and speaker identification, into handheld devices. This research is spurred by the recent increased popularity of commercially-available handheld computers which have allowed computation to become more mobile and pervasive. Formerly specialized devices, such as cellular telephones, now offer a range of capabilities beyond simple voice transmission, such as the ability to take, transmit and display digital images. As these devices become more ubiquitous and their range of applications increases, the need for security also increases. To prevent impostor users from gaining access to sensitive information, stored either locally on a device or on the device's network, security measures must be incorporated into these devices. Face and speaker verification are two techniques that can be used in place of, or in conjunction with, pre-existing security measures such as personal identification numbers or passwords.

Handheld devices offer two distinct challenges for standard face and voice identification approaches. First, their mobility ensures that the environmental conditions the devices will experience will be highly variable. Specifically, the audio captured by these devices could contain highly variable background noises producing potentially low signal-to-noise ratios. Similarly, the images captured by the devices can contain highly variable lighting and background conditions. Second, the quality of the video and audio capture devices is also a factor. Typical consumer products are constrained to use audio/visual components that are both small and inexpensive, resulting in a lower quality audio and video than is typically used in laboratory experiments.

To examine these issues we have developed a prototype system for incorporating two biometric techniques, speaker identification and face identification, into a mobile device. Results of an early evaluation of this system were previously reported in [1]. In our previous study, we evaluated a combined face and speaker identification system within a user verification "login" scenario on an iPAQ handheld computer. The combined system was able to achieve a 50% reduction in the verification equal error rate (EER) over a system using only our speaker identification technology. This large improvement in performance was attained despite the fact that speaker identification system achieved an EER that was 75% smaller than that of the face identification system. This result was surprising because it showed that large improvements could be obtained through the combination of different biometric systems, even when one of the systems was vastly superior in accuracy to the other. In the work conducted in this paper, we improve upon our previous results by replacing our older, simpler face identification system with a newer state-of-the-art system.

The rest of the paper is organized as follows. We first present an overview of our two biometric techniques and the fusion technique for combining them. Next, we discuss the mobile-device paradigm in which we are conducting our experiments and the methods of data collection employed. We follow this with experimental results showing the performance of the two biometric techniques on the data we have

In *Proc. of Workshop on Multimodal User Authentication*, pp. 113–120, December 11-12, 2003, Santa Barbara, California.

collected, both individually and in combination. Finally, we summarize and discuss the results and present plans for future directions of our work.

2. PERSON IDENTIFICATION

2.1. Speaker Identification

Speech has long been recognized as a reasonable biometric for person identification. However, speech is a variable signal whose main purpose is not to specify a person's identity but rather to encode a linguistic message. In systems where the linguistic content of the speech is unknown (e.g. for surveillance tasks), text-independent speaker identification systems are generally used. However, in security applications where the user is cooperative in the attempt to prove his/her identity, the linguistic content of the speech message is typically known and can be tightly constrained. In this case, a text-dependent system can be used. When the linguistic content of the message is known, text-dependent speaker recognition systems generally perform better than text-independent systems because they can tightly model the characteristics of the specific phonetic-content contained in the speech signal.

A common technique used in speech-based person identification is to prompt the user with a randomly generated challenge phrase. During authentication, automatic speech recognition can be used to verify that the spoken utterance matches the prompted utterance. For this type of scenario, it is both reasonable and beneficial to use the automatic speech recognition (ASR) output to leverage the phonetic constraints that give text-dependent systems their advantage. In [2], two techniques were described that use the ASR output during the analysis of the phonetic content from the test utterance.

In our speaker adaptive ASR approach, the system uses speaker-dependent speech recognizers to model each speaker. During training, phonetically transcribed enrollment utterances are used to train context-dependent phonetic models for each speaker. During testing, a speaker-independent 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 each speaker-dependent phonetic model. Modeling speakers at the phonetic level can be problematic because enrollment data sets are typically too small to build robust speaker-dependent models for every context-dependent phonetic model. To compensate for this difficulty, we use an adaptive scoring approach in which the speaker-dependent (SD) score is interpolated with a speaker-independent (SI) score.

Mathematically, if the word recognition hypothesis assigns each feature vector x from the utterance X to phonetic unit j, then the score for speaker S_i, $p(X|S_i)$, 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 phonetic unit 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 examples of phonetic unit j observed for speaker S_i, and τ is a global tuning parameter that is set empirically using a separate development set. The log ratio in the equation generates positive scores when the input speech is a good match to a particular speaker's models and negative scores when the speech is a poor match.

This scoring strategy results in models that capture detailed phonetic-level characteristics for a speaker when sufficient training data is available, but relies more on speaker independent models for phonetic units with sparse training data. Thus, for cases with limited training data, the speaker independent model provides a more *neutral* score. In the limiting case, if no speakers have training data for any of the phones observed in a particular test utterance, then they will all receive the same neutral score of zero, which is an intuitively consistent result.

2.2. Face Identification

The face identification framework used in our work is similar to the one described in [3], but with some differences in detection and classification methods.

2.2.1. Face Detection

A two-step process is used for face detection. First, a fast hierarchical classifier similar to the one described in [4] is applied to the captured image, to roughly localize the face in the image. The region around the face is then cropped out from the larger image, histogram equalized, and scaled to a fixed size.

Next, a component-based face detector [3] is applied to the extracted region to precisely localize the face and to detect facial components. This method first independently applies component detection classifiers to the face image. Each of these support vector machine (SVM) classifiers is trained to detect a particular component, such as a nose, mouth, or left eyebrow. In all, 14 face components are used, and each component classifier is evaluated over a range of positions in the vicinity of the expected location of the desired component. A geometrical configuration classifier is

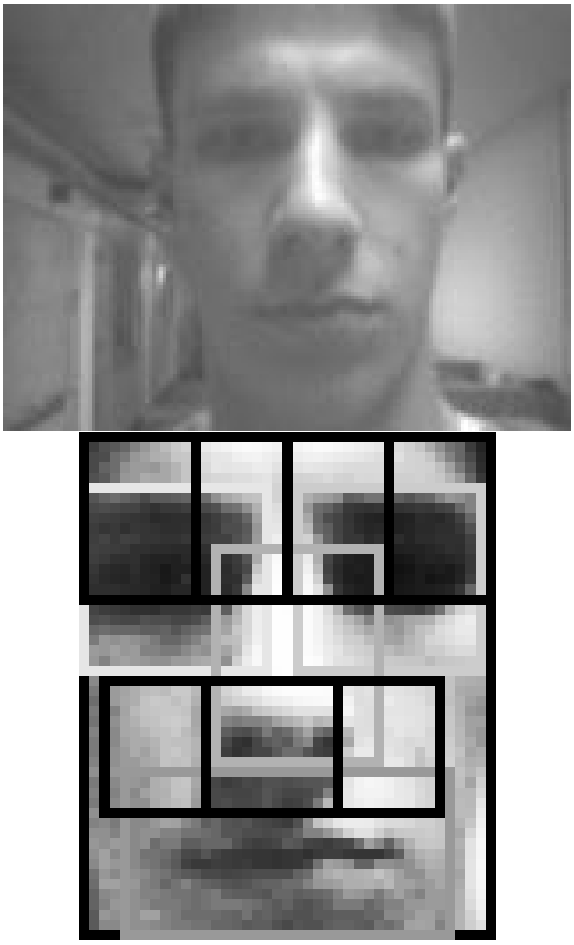

Fig. 1. A sample image and its face detection result with the face component regions superimposed.

then applied to the combined output of each of the 14 component classifiers from each candidate position. The candidate positions that yields the highest output of the second-level classifier are taken to be the detected component positions.

Ten out of the 14 components are used for face recognition. The remaining four are not used because they either overlap heavily with other components, or display few structures of use in distinguishing people from one another. The gray values of the ten selected components are normalized in size and combined into a single feature vector. The feature vector serves as input to the face recognizer. Figure 1 illustrates an enrollment image, as well as its selected face region with the positions of the facial components as detected by our system.

2.2.2. Face Recognition

For recognition, a one-vs-all SVM scheme is used, where one classifier is trained to distinguish each person in the database from all the others [5]. In the SVM training process, for each person's classifier, the feature vectors corresponding to that person's training images are used as positive examples, and the feature vectors corresponding to all the others' images are used as negative examples. The SVM training process finds the optimal hyperplane in the feature space that separates the positive and negative data points. Since the training data may not be separable, a mapping function corresponding to a second-order polynomial SVM kernel function [5] is applied to the data before training.

The runtime recognition process consists of computing the SVM classifier output score for each person's SVM classifier [5]. The scores are zero-centered – that is, a score of zero means the data point lies directly on the decision hyperplane, and positive and negative scores mean the data point lies on the positive and negative example side of the decision hyperplane, respectively. The absolute value of the SVM output is a multiple of the distance from the decision hyperplane, and could be normalized to produce the distance. Thus, a highly positive score represents a large degree of certainty that the data point belongs to the person the SVM was trained for, and a highly negative score represents the opposite. The output scores from all SVM classifiers make up the n-best list that we treat as our face recognition result.

For our face identification task, we collected and tested frontal face image data only. Most state of the art face identification systems attempt to account for rotations in and out of the image plane, and/or occlusions – which would be present in a typical surveillance task. However, for the handheld face identification problem, the user will be cooperating with the identification process; and in general, the user certainly will be looking at the screen of the handheld device as he or she is using it. Thus, accounting for heavily rotated or occluded faces is not important in this project. Generally, rotations or occlusions in face images make the problem of identification more challenging; thus, our problem is easier in this respect. Nonetheless, the variable lighting and background conditions and inexpensive camera present an orthogonal challenge, to ensure the non-triviality of our problem.

2.3. Multi-Modal Fusion

Past work on fusing face and speaker classifiers has generally used very simple combination strategies. Poh and Korczak used a logical AND rule on the results of their independent face and speaker systems [6]. This rule is most useful when the goal is to limit false acceptances, since both classifiers must accept the user in order to produce

an acceptance by the fused-classifier. Brunelli and Falavigna [7] and Kittler *et al* [8] use basic probabilistic combination operators on the outputs from their independent recognizers. Bigün *et al* utilize a Bayesian statistics approach which compensates for biases and interdependencies between different classifiers [9]. An alternative to these statistical fusion approaches is the use of discriminatively trained methods such as decision trees or linear discriminant functions [10].

In our work, a linear weighted summation is employed for the classifier fusion where the weights for each classifier are trained discriminatively on a held-out development set using minimum classification error (MCE) training. The MCE training optimizes the equal error rate of false acceptances and false rejections under the user verification scenario. Because the final decision only requires the combination of two independent classifiers, only one additional parameter (the ratio of the weights of the classifiers) needs to be learned. A simple brute force sampling of the parameter space is used for this MCE training. More complicated techniques (such as gradient descent training) could be used in situations where more than two scores must be fused.

3. EXPERIMENTS

3.1. The Handheld Device

For our experiments we utilized a collection of iPAQ handheld computers. Speech data were collected utilizing the built-in microphone of the iPAQ. Two different models of iPAQs were used, with two different models of off-the-shelf, inexpensive electret condenser microphones. Face data were collected using a 640x480 CCD camera located on a custom-built expansion sleeve for the iPAQ. The iPAQ handheld computer, combined with the custom sleeve, is the handheld device platform used for pervasive computing research in the MIT Oxygen Project [11]. An image of the iPAQ with the expansion sleeve is shown in Figure 2. Because of the current computation and memory limitations of the iPAQ handhelds, the images and audio are captured by the handheld device, but then transmitted over a wireless network to servers which perform the operations of face detection, face identification, speech recognition, and speaker identification. In future work we hope to improve the computational efficiency and memory footprints of our systems so they can be deployed directly on small handheld devices.

3.2. The Login Scenario

Our experiments were conducted using a login scenario that combined face and speaker identification techniques to perform the multi-biometric user verification process. When "logging on" to the handheld device, users snapped a frontal view of their face, spoke their name, and then recited a prompted lock combination phrase consisting of three randomly selected two digit numbers (e.g. "*25-86-42*"). The system recognized the spoken name to obtain the "claimed identity". It then performed face verification on the face image and speaker verification on the prompted lock combination phrase. Users were "accepted" or "rejected" based on the combined scores of the two biometric techniques.

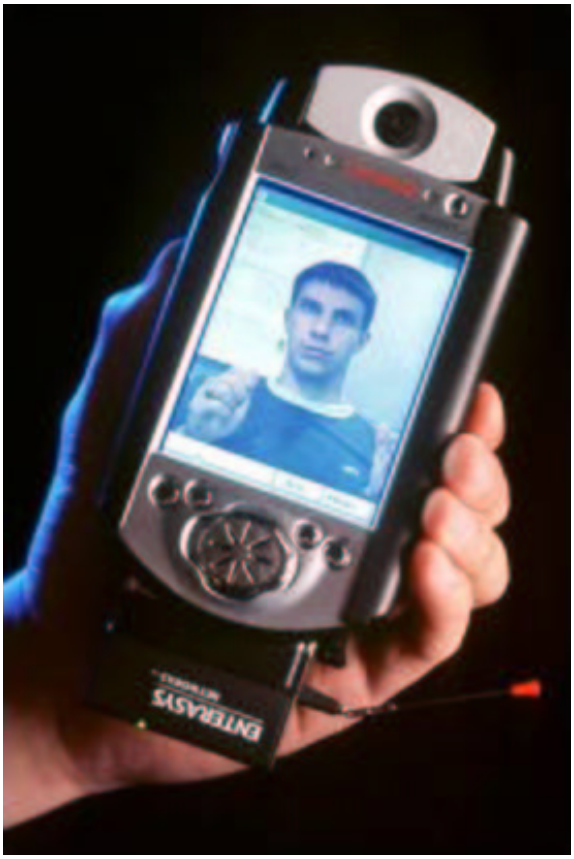

Fig. 2. The iPAQ handheld computer used in this study.

3.3. Data Collection

For our set of "enrolled" users, we collected face and voice data from 35 different people. Each person performed eight short enrollment sessions, four to collect image data and four to collect voice data. For each voice session, each user recited 16 prompted lock-combination phrases. Each image collection session consisted of the user taking 25 frontal face images in a variety of rooms in our lab with different lighting conditions. No specific constraints were placed on the distribution of the locations and lighting conditions; users were allowed to self-select the locales and lighting conditions of their images. To illustrate the quality of the images, Figure 3 shows two sample images captured during

Fig. 3. Two sample images collected on the iPAQ.

the data collection.

During image collection, a fast face detector [12] was applied to each captured image to verify that the image indeed contained a valid face. This face detector occasionally rejected images when it failed to locate the face in the image with sufficiently high confidence. When this occurred the user was instructed to capture a new image. Due to a conservative face detection confidence threshold, no false positives (i.e., images with incorrectly detected faces) were observed from this face detector during the data collection. It is important to note that the face detector used during our data collection was not the same one used in the experiments in this paper.

Each voice and image session was typically collected on a different day, with the time span between sessions often spanning several days and occasionally a week or more. Each enrollment session typically lasted less than 5 minutes with the total enrollment time taking approximately 30 minutes on average. In total this yielded 100 images and 64 speech samples per enrolled user for training. An additional set of four enrollment sessions of audio data (i.e., 64 additional utterances) from 17 of the training speakers was available for development evaluations and multi-modal weight fusion training.

For our evaluation, we collected 16 sample login sessions from 25 of the 35 enrolled users. This yielded 400 unique utterance/face evaluation pairs from enrolled users. We also collected 10 impostor login sessions from 20 people not in the set of enrolled users for an additional 200 utterance/face evaluation pairs from unenrolled people.

We used the evaluation data to perform our user verification experiments. Each utterance/face pair from in-set speakers was used as a positive example of that user. This yielded a total of 400 positive examples for our evaluation. Each utterance/face pair from each in-set user could also be used as an impostor for the other 34 users in the enrolled set. This yielded 13600 impostor examples from in-set speakers. Each utterance/face pair collected from out-of-set impostors was also used to generate an impostor example for each of the 35 users in the enrolled set. This yielded 7000 impostor examples from users not in the enrollment set. In general, it is expected that impostors that have never been observed by the system will generate more classification errors than enrolled users who try to impersonate other enrolled users. This is because the models are trained to discriminate between users observed in the training data and thus may not generalize well to unseen users.

3.4. Training

The face and speaker systems were trained on the enrollment data for the 35 enrolled users. To train the fusion weights, one of the four face enrollment sessions was held out and a development face ID system was trained on the remaining three face sessions. Face identification scores from this held-out set were pairwise combined with speaker identification scores generated for utterances from the existing speaker identification development set. The true in-set examples and in-set impostor examples were provided to the MCE weight training algorithm previously described to generate the multi-modal fusion weights.

3.5. Face Detection Issues

For the experiments presented in this paper, the face detection algorithm used during the evaluation is not the same as the face detection algorithm used during the data collection. The detection algorithm used during the evaluation was specifically tuned to accept facial images that are well suited to the component-based classification method used for face identification. Because this classification method works best with frontal images of faces that are not tilted or contorted, the face detection algorithm was initially tuned such that tilted or contorted faces were rejected. The face detection algorithm used during our data collection was less conservative in its accept/reject decision of a hypothesized

Table 1. User verification results expressed as equal error rates (%), when forcing the face detector to output a detected face, on three systems (face only, speaker only, and multi-modal fusion) under two impostor conditions (known in-set impostors vs. unknown out-of-set impostors).

System	In-set Impostors	Out-of-set Impostors
Face	3.21%	4.87%
Speaker	0.75%	1.66%
Fused	0.24%	0.66%

face in an image. As a result, a sizable number of images in the training and evaluation data sets were rejected by the new face detection algorithm.

Because of the reduced number of images for our evaluation, we could not make a direct comparison with our previous test results. To allow us to make this comparison, we elected to run two experiments, one where the conservative face-detection decisions were used and a second experiment where the face detection algorithm was forced to output a detected face even if the image's detection score fell below the standard acceptance threshold. These two experiments allow us to examine the trade-off between the added gain in accuracy enabled by stricter control in the input facial images, and the potential added inconvenience of requiring users to provide an untilted, uncontorted frontal image.

3.6. Experimental Results

3.6.1. Forced Face Detection Results

Table 1 shows our user verification results for three systems (face ID only, speaker ID only, and our full multi-modal system) under two different impostor conditions (using only known in-set impostors vs. using only unknown out-of-set impostors). This experiment uses a detection threshold which forces the face detector to output a face hypothesis for all of the images, even when the detection confidence score is low. Figure 4 shows the results for the out-of-set impostor evaluation on a detection error trade-off (DET) curve.

Several observations should be made from these results. First, the speaker ID system has an equal error rate (EER) which is three times smaller than that of the face ID system when evaluated with unknown out-of-set impostors. These face ID results are better than our previously reported results in which the face ID system produced an EER which was four times larger than the speaker ID EER.

Next, the combined system has a 60% reduction in EER from 1.66% in the speech only system to 0.66% in the combined system. This is a slightly better improvement than the 50% reduction we had observed in our previous study. This demonstrates that sizable improvements can be obtained when multiple independent biometric techniques are combined even when one biometric technique performs substantially better than others.

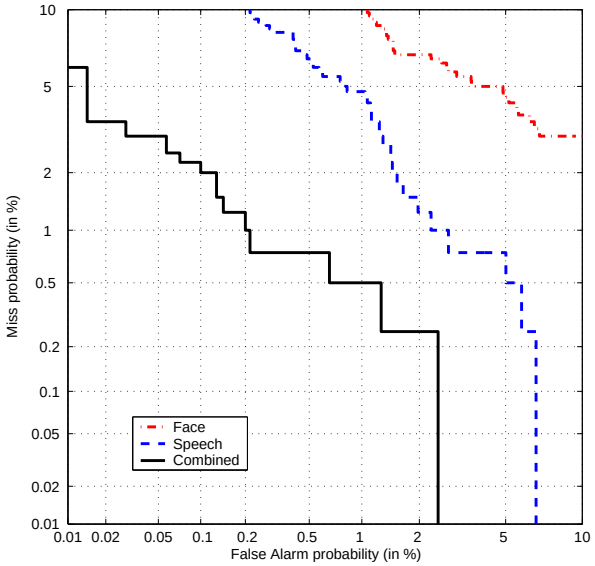

Fig. 4. DET curves for face and speech systems run independently and in combination when tested using impostors unknown to the system and when using a face detector that is forced to output a detected face for each input image.

Finally, it is interesting to note that the combined system achieves an EER of only 0.24% on the in-set impostor experiment. In other words, the EER when using the unknown impostors is 2.75 times greater than the EER of the in-set impostor experiment. This shows the importance of evaluating the system using people that are not part of the training data.

3.6.2. Conservative Face Detection Results

When applying the conservative face detection threshold to the evaluation utterances, 12% of the images were rejected. To evaluate the system under these conditions, the face ID system was first re-trained using the same threshold

Table 2. User verification results expressed as equal error rates (%), when using the conservative face detection threshold on three systems (face only, speaker only, and multi-modal fusion) under two impostor conditions (known in-set impostors vs. unknown out-of-set impostors).

System	In-set Impostors	Out-of-set Impostors
Face	1.66%	2.57%
Speaker	0.77%	1.63%
Fused	0.00%	0.15%

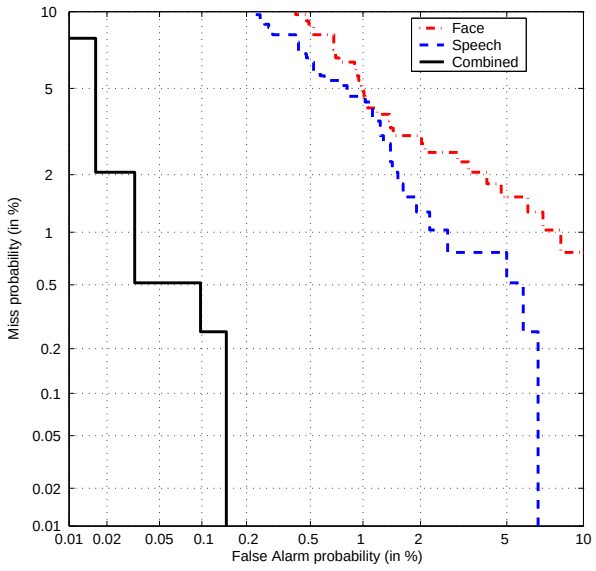

Fig. 5. DET curves for face and speech systems run independently and in combination when tested using impostors unknown to the system and when using the conservative face detection threshold.

for detection. The system's verification results were then re-computed using the 88% of the data that passed the more conservative face detection threshold.

Table 2 shows the equal error rates under these new constraints. The face ID system shows a nearly 50% improvement in EER performance over the forced detection result when the images with poor face detection scores were discarded. When used in conjunction with the speaker ID component, the combined system achieved an EER of only 0.15% when testing with out-of-set impostors. This a sizeable 90% reduction in EER from the speech only system! This combined system also achieved perfect separation between true users and in-set impostors resulting in a 0.0% EER on the in-set impostor experiment. This demonstrates that highly accurate biometric authentification can be obtained if the user is willing accept additional constraints on the verification process that may increase the inconvenience of the system. Unfortunately, because so few errors are observed, due to the limited size of our evaluation set, it is not possible to make any firm claims about the absolute level of the error rate of the system. We plan to increase the size of our evaluation set in future experiments.

3.6.3. Comparison with YOHO Corpus

To examine the degradation that might be experienced when our speaker identification technique is utilized in a mobile environment, we compared the performance of closed-set speaker recognition on the mobile handheld data set against the performance of our system on the tightly constrained YOHO corpus, which uses the same lock combination phrase approach that we employed [13]. It is important to note that the YOHO corpus was collected using a single close-talking telephone handset in a quiet office, and thus does not suffer from the degradations that are present in our mobile devices due to the low quality far-field microphone and the variable background conditions. In [2], it was shown that our system's speaker recognition error rate was 0.31% over YOHO's closed-set of 138 speakers. Using our 400 utterance in-set speaker evaluation set, our system's speaker recognition error rate was 0.25% over our closed set of 35 enrolled speakers (i.e., only one misrecognition in 400 trials). Thus we have achieved roughly the same error rate as on YOHO, but only with a much smaller set of speakers.

4. SUMMARY AND FUTURE WORK

In summary, our initial study in biometric fusion for user verification has demonstrated the benefits of combining face and speaker identification even when one of the biometric techniques has superior performance to the other. A 90% reduction in user verification equal error rate was observed when our speaker identification system was fused with a face identification system. This result was achieved with a system that forces the user to provide a frontal image that can be automatically detected with a high-level of confidence. By adjusting the confidence-level of the face detector, the system can reduce the inconvenience of re-capturing images when the face detector fails, but at the expense of reduced user verification accuracy.

Though this study demonstrated the feasibility of our approach, our current evaluation set is quite small. In future work we plan to expand the size of evaluation set and examine the specific types of errors the system makes. We also plan to investigate the performance of the system under the conditions where impostors are specifically selected based on resemblances of their voice or facial properties (i.e., same gender or ethnicity) to particular enrolled users.

5. ACKNOWLEDGMENTS

The authors wish to thank Dave Dopson and Ken Steele, who helped in the development of the application; and Jane Wu who assisted with data evaluation. This work was supported in part by the MIT Oxygen Alliance.

6. REFERENCES

[1] T. Hazen, E. Weinstein, and A. Park, "Towards robust person recognition on handheld devices using face and speaker identification technologies," in *Proc. of Int.*

Conf. on Multimodal Interfaces, Vancouver, Canada, November 2003.

[2] A. Park and T. Hazen, "ASR dependent techniques for speaker identification," in *Proc. of Int. Conf. on Spoken Language Processing*, Denver, Colorado, September 2002, pp. 1337–1340.

[3] B. Heisele, P. Ho, and T. Poggio, "Face recognition with support vector machines: Global versus component-based approach," in *Proc. of Int. Conf. on Computer Vision*, Vancouver, Canada, July 2001, vol. 2, pp. 688–694.

[4] B. Heisele, T. Serre, S. Prentice, and T. Poggio., "Hierarchical classification and feature reduction for fast face detection with support vector machines," *Pattern Recognition*, vol. 36, pp. 2007–2017, 2003.

[5] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, Berlin, Germany, 1995.

[6] N. Poh and J. Korczak, "Hybrid biometric person authentication using face and voice features," in *Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentification*, Halmstad, Sweden, June 2001, pp. 348–353.

[7] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 955–966, October 1995.

[8] J. Kittler, Y. Li, J. Matas, and M. Sanchez, "Combining evidence in multimodal personal identity recognition systems," in *Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentification*, Crans-Montana, Switzerland, March 1997, pp. 327–334.

[9] E. Bigün, J. Bigün, B. Duc, and S. Fischer, "Expert conciliation for multi modal person authentication systems by Bayesian statistics," in *Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentification*, Crans-Montana, Switzerland, March 1997, pp. 291–300.

[10] A. Ross, A. Jain, and J. Qian, "Information fusion in biometrics," in *Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentification*, Halmstad, Sweden, June 2001, pp. 354–359.

[11] E. Weinstein, P. Ho, B. Heisele, T. Poggio, K. Steele, and A. Agarwal, "Handheld face identification technology in a pervasive computing environment," in *Short Paper Proceedings, Pervasive 2002*, Zurich, Switzerland, August 2002, pp. 48–54.

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of Conf. on Computer Vision and Pattern Recognition*, Kauai, Hawaii, December 2001, pp. 511–518.

[13] J. Campbell, "Testing with the YOHO CD-ROM voice verification corpus," in *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*, Detroit, Michigan, May 1998, pp. 341–344.