Multi-Modal Face and Speaker Identification for Mobile Devices[‡]

Timothy J. Hazen[1], Eugene Weinstein[2], Bernd Heisele[3], Alex Park[4], and Ji Ming[5]

[1] Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
hazen@csail.mit.edu

[2] Department of Computer Science
New York University
New York, New York, USA
eugenew@cs.nyu.edu

[3] Honda Research Institute USA
Cambridge, Massachusetts, USA
bheisele@honda-ri.com

[4] Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
malex@csail.mit.edu

[5] School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast
Belfast, United Kingdom
j.ming@qub.ac.uk

1 Introduction

In this chapter we discuss the application of two biometric techniques, face and speaker identification , for use on mobile devices. This research has been spurred by the proliferation of commercially-available hand-held computers. Because of their mobility and increasing computational power, these devices are fast becoming a pervasive part of our lifestyle. Even 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

[‡] This is a preprint of Chapter 9 of the book *Face Biometrics for Personal Identification: Multi-Sensory Multi-Modal Systems*, edited by R. I. Hammoud, B. R. Abidi and M. A. Abidi., Springer, Berlin 2007.

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 can 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 system that combines two biometric techniques, speaker identification and face identification, for use with a mobile device. We provide a high level overview of our speaker and face identification technologies in Section 2. Following the description of these technologies, the chapter will focus on the following three research questions:

1. How much improvement in speaker identification performance can be gained by combining the audio and visual biometric information?
2. Can full video information allow for more accurate face identification than single image snapshots?
3. How can speaker identification systems be made more robust to variable environments?

To answer question 1, it has been found that combining speaker and face identification technologies can have a dramatic effect on person identification performance. In one set of experiments, discussed in Section 3, a 90% reduction in equal error rate in a user verification system was achieved when integrating the face and speaker identification systems.

The answer to question 2 is still largely open for debate. However, in preliminary experiments examining the use of static and dynamic information extracted from video, it was found that dynamic information about lip movement made during the production of speech can be used to complement information from static lip images in order to improve person identification. These results are discussed in Section 4.

To answer question 3, degradation in speaker identification rates in noisy conditions can be mitigated through the use of noise compensation techniques and/or missing feature theory. Noise compensation involves the adjustment of acoustic models of speech to account for the presence of previously unseen noise conditions in the input signal. Missing feature theory provides a mech-

anism for ignoring portions of a signal that are so severely corrupted as to become effectively unusable. In Section 5 we examine the use of two techniques for noise robust speaker identification, the posterior union model for handling missing features and universal compensation.

2 Person Identification Technologies

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. It has been found for many text-independent applications that, even when linguistic knowledge is ignored completely, accurate speaker identification based purely on acoustic information can be performed. The standard approach is to extract wide-band spectral feature vectors from the audio signal (in the form of mel-scale cepstral coefficients or MFCCs [2]) at a fixed interval (typically every 10 milliseconds). The full collection of acoustic features from all utterances in an individual's training set are then pooled together and modeled with a single probability density function constructed from a Gaussian mixture model (GMM). Speaker identification is performed by scoring the MFCC feature vectors against the GMMs of enrolled speakers to generate likelihood scores for these speakers [14].

For the problem of speaker verification (i.e., verifying with a *yes* or *no* decision whether a user is who they claim to be), speaker likelihood scores are typically normalized by a *universal background model* which captures the general distribution of speech over all users. Mathematically, the GMM speaker verification score for a set of acoustic feature vectors $\mathbf{x}_1$ through $\mathbf{x}_N$ for purported user S is modeled probabilistically as follows:

$$\sum_{i=1}^{N} \log \frac{\mathrm{p}(\mathbf{x}_i|S)}{\mathrm{p}(\mathbf{x}_i)} \tag{1}$$

Here, $\mathrm{p}(\mathbf{x}_i)$ represents the GMM for the universal background model.

Text-independent systems have proven to work well for some applications. However, when the linguistic content of the message is known text-dependent speaker recognition systems generally perform better. This is because text-dependent systems can tightly model the characteristics of the specific phonetic content contained in the speech signal. 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 pre-specified and can be tightly constrained. In this case, a text-dependent system is preferred.

In our work, we have developed a speaker identification system that uses speaker-dependent speech recognition models to perform the speaker identification process [12, 13]. During training, phonetically transcribed enrollment utterances are used to train context-dependent acoustic-phonetic models for each speaker. During testing, a speaker-independent automatic speech recognition system hypothesizes a phonetic transcription for the test utterance. This transcription is then used by the system to score each segment of speech against each speaker-dependent acoustic-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, an adaptive scoring approach can be used in which the specific acoustic-phonetic models for a speaker can be interpolated with the speaker's text-independent GMM model. This improves the robustness of the approach when limited enrollment data is available. Mathematically, the speaker score for our phonetic approach is modeled probabilistically as follows:

$$\sum_{i=1}^{N} \log \left(\lambda_i \frac{\mathrm{p}(\mathbf{x}_i|u_i, S)}{\mathrm{p}(\mathbf{x}_i|u_i)} + (1-\lambda_i) \frac{\mathrm{p}(\mathbf{x}_i|S)}{\mathrm{p}(\mathbf{x}_i)} \right) \tag{2}$$

Here, a phonetic label u_i is provided from a speech recognition engine for each acoustic feature vector $\mathbf{x}_i$. The interpolation factor λ_i is determined separately for each phonetic unit u_i based on the number of times it appeared in the enrollment data:

$$\lambda_i = \frac{\mathrm{count}(u_i)}{\mathrm{count}(u_i) + K} \tag{3}$$

Here K is a predetermined constant (typically 5 in our systems). The interpolation factor prefers the context dependent model ratio for phone u_i when that phone has been observed often in the enrollment data, but it backs off toward the global GMM approach if u_i is rarely or never seen in the enrollment data.

2.2 Face Identification

Identifying people from images of their face is a widely studied problem. In addition to discussion of this topic in other chapters of this book, a thorough review of the literature on this topic is available in [19]. In this chapter, we discuss only the technologies used in our experiments. The primary face identification framework used in our work is largely based on work originally presented in [7].

Face Detection

Before face identification techniques are applied, the face must first be detected and located within a given image. The Viola-Jones face detection

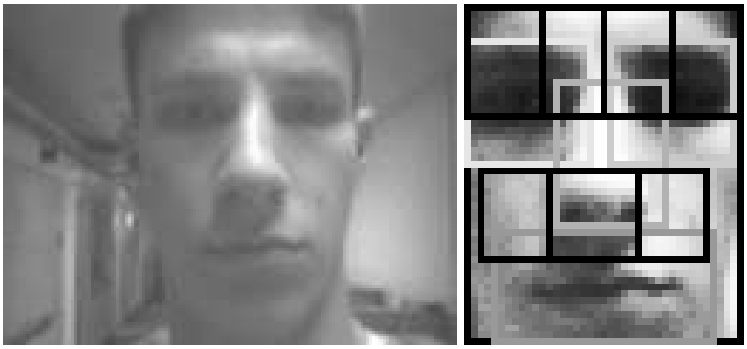

Fig. 1. A sample image and its face detection result with the face component regions superimposed.

algorithm (which is based on a boosted cascade of feature classifiers) is a commonly used approach which we have used as our baseline face detection algorithm [16].

As an alternative, we have also used a fast hierarchical classifier to roughly localize the face in the image [8]. 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 [7] 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 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 then applied to the combined output of each of the 14 component classifiers from each candidate position. The candidate positions that yield the highest output of the second-level classifier are taken to be the detected component positions. 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.

SVM-Based Face Recognition

A common approach to visual feature extraction for face identification is to use an *appearance-based* approach in which the raw image pixels are either used directly or projected into a lower dimension subspace. Large dimension feature vectors can only be used with classification methods which exhibit robustness to the *curse of dimensionality*, e.g. support vector machines (SVMs) [15]. We have used SVMs as the primary classification technique for face identification in our systems.

In our initial work, presented in [6], we used a full face image compressed to 40x40 grey-scale pixels and histogram normalized to adjust the brightness. Improved results were later obtained by extracting appearance-based features from ten of the fourteen component regions found during the face detection process [5]. The ten selected components are similarly scaled in size and normalized, and then combined into a single feature vector which serves as input to the face recognition component.

For face recognition, a one-vs-all SVM classification scheme is used, where one classifier is trained to distinguish each person in the database from all the others. In the training process, the feature vectors corresponding to a person's training images are used as positive examples for the classifier, and the feature vectors extracted from images of all other users 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 is applied to the data before training.

The runtime verification process consists of computing the output score for the purported user's SVM classifier [15]. The scores are zero-centered. In other words, 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.

GMM-Based Face Recognition

In our work on audio-visual speech recognition, we have used appearance-based visual features extracted from the raw images of the mouth region [3]. We have since adapted this approach to person identification using Gaussian mixture models (identical in nature to those used in the speaker identification field). Because probabilistic classifiers, such as the GMM, typically require lower-dimension feature spaces to avoid problems of sparse training data, a dimensionality reducing transform is often required. In experiments discussed in Section 4, we present results on GMM-based person identification using visual information derived only from the lip region of the face.

2.3 Multi-Modal Fusion

In our work, a simple linear weighted summation is employed for the classifier fusion where the weights for each classifier are trained discriminatively (on held-out development data) to minimize classification error. For the combination of face and speech classifiers, only one fusion parameter (the ratio of the

weights of the classifiers) needs to be learned. A simple brute force sampling of different ratios can be used in this case. More complicated techniques (such as gradient descent training) could be used in situations where more than two scores must be fused.

3 Multi-Modal Person ID on a Handheld Device

3.1 Overview

Our initial experiments in multi-modal person identification were performed using iPAQ handheld computers. A login scenario that combined face and speaker identification techniques to perform the multi-biometric user verification process was devised. 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.

Speech data were collected utilizing the built-in electret condenser microphone of the iPAQ. 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, was the handheld device platform used for pervasive computing research in MIT's Project Oxygen [17]. An image of the iPAQ with the expansion sleeve is shown in Figure 2. Because of the computation and memory limitations, the images and audio were 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 the future we expect the computational and memory components of handheld devices to improve such that our systems can be deployed directly on these small handhelds.

3.2 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. For each image collection session, users captured 25 frontal face images in a variety of rooms in our laboratory 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

Fig. 2. The iPAQ handheld computer used in our study, along with two sample images collected in the iPAQ.

the quality of the images, Figure 2 shows two sample images captured during the data collection.

During image collection, the Viola-Jones face detector [16] 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.

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. 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.

A separate set of evaluation data was collected to perform user verification experiments. For this evaluation set, 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. Each utterance/face pair collected from out-of-set impostors was used to generate an impostor example for each of the 35 enrolled users yielding 7000 impostor examples.

3.3 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 identification 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 weight training algorithm to generate the multi-modal fusion weights.

3.4 Face Detection Issues

The performance of a face identification system is affected by the quality of the images it is provided. If the system tightly controls the user and rejects images in which the head is tilted or rotated, the face is contorted in any unusual fashion, etc., then the variance of the data will be reduced and improved performance should be expected. In our work we initially collected facial images within a system running the Viola-Jones face detector. In our evaluations we have used a component-based face detection algorithm which is more conservative in its detection decisions. As a result, a sizable number of images in the training and evaluation data sets were rejected by the component-based face detection algorithm.

To detail the effect of the face detection algorithm upon the face identification results, two experiments were conducted: 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 having users retake snapshot images until the face detection algorithm accepts one.

3.5 Experimental 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 face detection conditions. The results are reported using the equal error rate metric. The equal error rate (EER) is the point in the detection-error tradeoff curve where the likelihood of a false acceptance of an impostor (i.e., a *false alarm*) is equal to the likelihood of false rejection of the true user (i.e. a *miss*).

In the table's forced face detection results, all evaluation sessions are used. However, for the conservative face detection results, 12% of the images were rejected. In this case, the system's verification results were computed using only the 88% of the data that passed the more conservative face detection

Table 1. User verification results expressed as equal error rates (%), over three systems (face only, speaker only, and multi-modal fusion), using two different face detection scenarios.

System	Forced Face Detection	Conservative Face Detection
Face	4.87%	2.57%
Speaker	1.66%	1.63%
Fused	0.66%	0.15%

threshold. Though the speaker identification system is unaffected by the face identification method that is used, the speaker identification equal error rates are different in the two columns because rejection of an image causes the companion spoken phrase from an evaluation login session to also be discarded, thus altering the speaker identification results slightly.

In examining the results, one can see that the face identification system using conservative face detection thresholding shows a nearly 50% improvement in EER performance over the system using forced face detection. Of course, the improved performance does come with a cost: in a deployed system, a user would face the added inconvenience of providing a new snapshot image whenever the face detector rejects an image.

Next, the results show that the speaker identification system is performing better than the face identification component, though the performance is of the same order of magnitude. When the two systems are used in combination, significant improvements are obtained over the use of either modality by itself. When using conservative face detection, the addition of the face identification system to the speaker identification system produced a 90% relative reduction in the equal error rate from 1.63% to 0.15%. Detection error tradeoff (DET) curves for the three systems (when using conservative face detection) are shown in Figure 3. These results demonstrate that highly accurate biometric authentication can be obtained via the multi-biometric approach of combining speaker identification and face identification technology.

4 The Use of Dynamic Lip-Motion Information

When performing face identification, an interesting question to ask is whether full motion video provides any substantive advantage over the use of individual still images. Video has been shown to be useful for face identification by providing a collection of temporally related images. Increased robustness can be obtained using video because face detection results can be interpolated over multiple frames, and results from frames with poor images can be discounted or ignored when considered jointly with other better scoring frames [9]. However, one might wonder if the actual dynamic motion of facial features themselves, such as the motion of the lips when someone is talking,

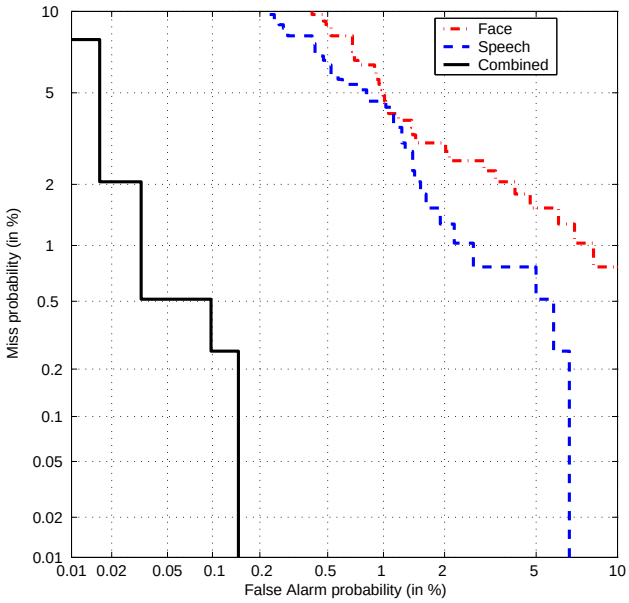

Fig. 3. DET curves for face and speech systems run independently and in combination when tested using impostors unknown to the system and when using the conservative face detection threshold.

can be used to identify a person. Or even more importantly, can the dynamic lip information provide any significant improvement over using only the static information available from the individual frames extracted from the video?

To examine this issue we have performed experiments using the AV-TIMIT video corpus [4]. This corpus was originally collected for use in audio-visual speech recognition experiments. It contains read sentences recorded in a quiet room using a high-quality digital camera for the video and a far-field array microphone for the audio. The first 10 utterances recorded for each user were used to train the face and speaker identification system and five additional utterances were used for our evaluation. In total the corpus contains recordings from 221 different people (yielding $221 \times 5 = 1105$ evaluation utterances). Because the AV-TIMIT corpus being used was recorded in quiet office conditions and the training data comes from the same session as the evaluation data, this person identification task does not represent realistic conditions. To make the task more challenging, our face identification system only uses the lip region portion of the video. Despite the unrealistic conditions of the task, we can still use this corpus to compare different visual features and to examine the effect of fusing audio and visual information.

From each individual frame, the image of the lips are represented using the top components from a principal component analysis (PCA) rotation applied

Table 2. Person identification results from visual lip images using static PCA features, dynamic PCA features, and a fusion of the static and dynamic features.

Lip Image Feature Vector	Person ID Error Rate
48-dimension static PCA features	6.0%
96-dimension static PCA features	3.6%
192-dimension static PCA features	4.1%
48-dimension dynamic PCA features	6.6%
96-dimension dynamic PCA features	7.7%
192-dimension dynamic PCA features	17.8%
48 dimension static PCA features + 48 dimension dynamic PCA features	2.1%
96 dimension static PCA features + 96 dimension dynamic PCA features	2.1%

to a discrete-cosine transform of the image from the lip region (a.k.a. *eigenlips* [1]). We refer to these feature vectors as the *static PCA* features. The first-order time difference between PCA vectors in sequential image frames is used to represent the dynamic changes in lip images. We refer to these feature vectors as the *dynamic PCA* features.

Because statistical classifiers require a tradeoff between the increased specificity from larger feature space dimensionality and the susceptibility of large dimension classifiers to over-training, we have evaluated the system using several different feature vector dimensionalities. We have also constructed feature vectors using static PCA features only, using PCA difference features only, and using a combination of the static and dynamic features. To perform person identification in this system, the individual feature vectors are modeled using a single Gaussian mixture model per speaker. Table 2 shows the closed-set person recognition performance on the 221 person AV-TIMIT corpus using eight different feature vector configurations used in our experiments. The results show that static lip information is more useful than dynamic lip information, but that improvements in person identification can be achieved by combining the static and dynamic information.

Table 3 shows the individual results of the audio-only and visual-only person identification system for closed set person recognition. The table also shows the combined audio-visual result when linearly combining the audio and visual scores. In this case, the optimal weighting of 0.95 for the audio stream and 0.05 for the visual stream yields an error rate of 0.27% (3 errors out of 1105 trials). When ratio of the audio weight to the visual weight is varied between 0.8/0.2 and 0.98/0.02 the person identification is never worse than 0.54% (6 errors out of 1105 trials).

These results on AV-TIMIT demonstrate, once again, the power of combining audio and visual information for person identification. In on-going re-

Table 3. Person identification results for audio-only, visual-only, and audio-visual systems using audio and lip-image information from the AV-TIMIT corpus.

System	Person ID Error Rate
Audio Only	1.2%
Visual Only	2.1%
Audio-Visual	0.27% to 0.54%

search, our group is currently moving beyond systems using the high-quality, single-session AV-TIMIT video, and towards the creation of a system that can handle video collected using commercial-off-the-shelf web cameras and handheld devices.

5 Noise Robust Speaker Identification

As discussed earlier, one of the great challenges of performing speaker or face identification in mobile applications is the possibility of severe variations in the feature measurements due to the environmental conditions (i.e., background noise, lighting conditions, etc.). One technique for addressing this problem is the application of missing feature theory. The basic premise of missing feature theory is that some features of the observation space may be so corrupted that they become useless for the task of person identification and should be ignored. For speaker identification this could involve either temporal corruption (e.g., a brief impulsive noise such as a door slam) or spectral corruption (e.g., a noise in a narrow spectral band such as a police siren). Comparable analogies could also be drawn for face identification (e.g., sudden severe shadows, occlusions of portions of the face, etc.). One could also view the problem of multi-modal fusion within the missing feature theory framework, where either of the audio or visual feature streams could be unreliable at any point in time and ought to be ignored in deference to the more reliable information stream.

In some situations, the corruption may not be so severe that it completely masks all usable information within a feature. In this case, a means of accounting for the corrupting noise in the observation of a feature is more desirable than completely ignoring the feature. In an ideal situation, models for biometric features could be trained from enrollment data collected under all of the corrupting conditions the user may encounter. Unfortunately, this is not feasible for most mobile applications and methods for compensating for unseen conditions must be employed.

In our work we have investigated the problem of robust speaker identification in noisy environments. In particular we have examined a missing feature approach called the *posterior union model*, and a noise compensation technique called *universal compensation*. Though we have not yet extended this work beyond speaker identification experiments, we believe these ideas can be

extended to the problems of face identification and the fusion of multi-modal information.

5.1 The Posterior Union Model

The basic premise behind missing feature theory (as we apply it to speaker identification) is that improved performance can be achieved by utilizing only information about features that can be reliably extracted from the input signal. Thus, if an input signal can be be represented as a collection of independent features $X = \{x_1, x_2, \ldots, x_N\}$, then there exists some optimal subset of uncorrupted features $X_{sub} \subseteq X$, that can be used as the basis for the speaker identification decision. This problem can be expressed probabilistically as

$$[S', X'_{sub}] = \arg \max_{S, X_{sub}} \mathrm{P}(S|X_{sub}) \tag{4}$$

where S represents a specific speaker and X_{sub} represents a specific subset of features from X. The expression seeks to find the most likely speaker S' by jointly maximizing the posterior probability over all speakers and all possible feature subsets X_{sub} in X. Here X'_{sub} is the optimal feature subset found for the most likely speaker S'. Using Bayes' Rule the expression is rewritten as

$$[S', X'_{sub}] = \arg \max_{S, X_{sub}} \frac{\mathrm{p}(X_{sub}|S)\mathrm{P}(S)}{\mathrm{p}(X_{sub})} \tag{5}$$

where $\mathrm{P}(S)$ is generally given a uniform distribution and $\mathrm{p}(X_{sub})$ is a normalizing term that is independent of the speaker S.

The posterior union model (PUM) generalizes the problem by removing the constraint that an exact set of optimal features, X'_{sub}, be found. Instead, for a given number of features M, PUM makes the following assumption:

$$\mathrm{p}(X'_{sub}|S, M) \approx \sum_{X_{sub} \subseteq X_N^M} \mathrm{p}(X_{sub}|S) \tag{6}$$

Here, X_N^M is the collection of all combinations of sets of M features chosen from the full N features in X. The approximation assumes that the sum of $\mathrm{p}(X_{sub}|S)$ over all X_{sub} drawn from X_N^M is dominated by the optimal subset of M features. This reduces the problem to finding the optimal number of reliable features M, but not the exact subset. In practice, individual features are rarely completely reliable or completely unreliable, but somewhere in between. Thus, the use of the union model allows a softer probabilistic decision than forcing features to either be used or discarded. Details of the PUM implementation can be found in [11].

5.2 Universal Compensation

If we consider that individual features may be only partially corrupted, then the missing feature approach should be amended to account for partially corrupted features. The universal compensation technique provides just such a

mechanism. Instead of decomposing the features in X into reliable features that are used and unreliable features that are ignored, the features can be decomposed into subsets containing variable degrees of corruption. In this formulation we can use the expression

$$\mathrm{p}(X|S) = \sum_{l=0}^{L} \mathrm{p}(X_l|S, \Phi_l)\mathrm{P}(\Phi_l|S) \tag{7}$$

where Φ_l represents a level of corruption and X_l represents the specific set of features in X which are corrupted at level Φ_l. In this case the posterior union model can be extended such that it considers the optimal number of features corrupted at each corruption level Φ_l and not just those that are completely clean or completely corrupted. Details of this formulation are found in [10].

In practice for speaker identification tasks, the universal compensation technique is applied by taking clean audio training data and adding noise at variable signal-to-noise ratios to simulate the different corruption levels Φ_0 through Φ_L. We have primarily added white noise to the clean training to simulate the corruption, but different types of noises could be used depending on the expected environments. Models for each speaker at each corruption level are trained. During evaluation on unseen data the posterior union model is used to select the number of features from the full set that optimally match each corruption level.

5.3 Experimental Results

To demonstrate the effectiveness of the posterior union model and universal compensation techniques, we conducted experiments on a handheld-device database collected at MIT. The database was designed to study speaker verification in realistic noisy conditions with limited enrollment data [18]. The database contains 48 enrolled speakers (26 male, 22 female) and 40 impostors (23 male, 17 female), each reciting short ice cream flavor phrases.

In our primary experiments, users enrolled into the system by speaking four examples of a specific phrase into the hand-held device. The enrollment session was conducted in a quiet office environment using an external earpiece microphone. For each enrolled user, speaker identification models were trained from the four enrollment examples. Low-pass filtered white-noise was added to each example at nine different signal-to-noise ratios between 4 and 20 dB (increasing 2 dB every step). This gives a total of ten corruption levels (including the no corruption condition) for the training phase. To evaluate the system, the same enrolled users and the 40 previously unseen impostors recited new evaluation phrases using the same hand-held device. However, the evaluation data were instead collected outdoors next to a noisy street intersection using the internal microphone of the device.

The speaker identification system uses phrase-dependent hidden Markov models to represent each speaker in the enrollment set. The features used to

represent the acoustic information are modeled with sub-band spectral components derived from decorrelated log filter-bank amplitudes collected from 20ms wide time windows sampled every 10ms. In total two energy and time difference values were used to represent the features within 10 different spectral sub-bands. The posterior union model is thus tasked with selecting the optimal number of sub-bands corrupted at each of the 10 different noise corruption levels. Details of the system used in this experiment can be found in [10].

For our experiments, we implemented four different systems all based on the same set of acoustic features:

- BSLN-Cln: a baseline system trained only on the clean office data and tested using the full set of acoustic features.
- BSLN-Mul: the baseline system trained on the full set of clean and artificially corrupted data pooled together to train a single multi-condition model for each speaker.
- PUM: a system trained only on the clean office data but allowed to select the optimal number of reliable sub-band components using the PUM approach.
- UC: a system trained on the clean and artificially corrupted data using the PUM approach to optimally select number of sub-bands matching each corruption level.

The experimental results are shown in Figure 4. The figure shows that a baseline speaker verification system trained in a quiet environment performs quite poorly when it is then used in a noisy environment (next to a noisy street intersection in this case). However, by artificially adding various levels of white noise to the training material, the equal error rate (EER) of the system is reduced from 30.2% to 22.4%. If the posterior union model is used in conjunction with the baseline system, the EER is reduced from 30.2% to 17.2%. Finally, if the PUM is combined with the system trained using varying levels of artificially added noise (i.e., the universal compensation approach), the EER is further reduced to 14.1%. These results show that techniques do exist to improve the robustness of speaker identification even in noisy environments that are mis-matched with the systems training conditions.

6 Summary

In this chapter, we have shown the power of combining face and speaker identification techniques for improved person identification. In Section 3, we demonstrated that a multi-biometric approach can reduce the equal error rate of a user verification system on a hand-held device by up to 90% when combining audio and visual information. In Section 4, we showed that dynamic information captured from a person's lip movements can be used to discriminate between people, and can provide additional benefits beyond the use of

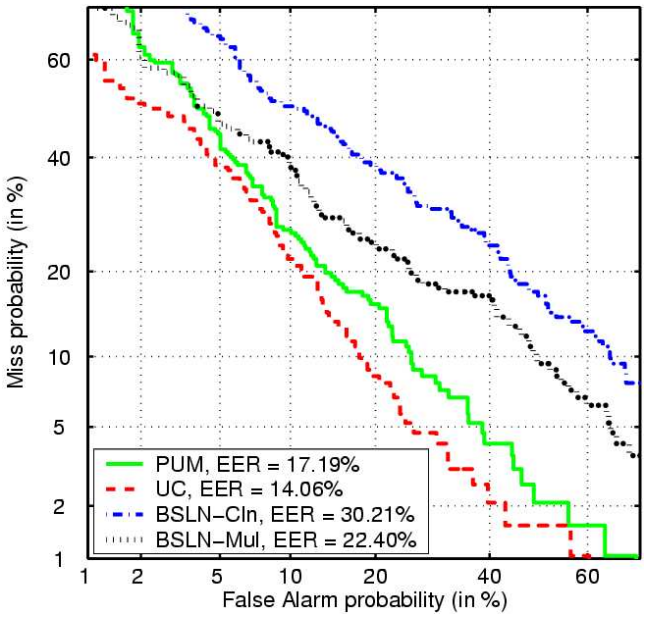

Fig. 4. DET curves comparing four limited enrollment speaker verification systems trained in a clean environment and tested in a mis-matched noisy environment.

static facial features. In Section 5, we addressed the problem of robust speaker identification for hand-held devices and showed the benefits of the posterior union model and the universal compensation techniques for handling corrupted audio data. In future work we plan to extend the use of the posterior union model to different facial feature vectors extracted from images as well as to the multi-modal fusion of different audio and visual features.

Acknowledgments

Portions of this work were supported by the following sponsors: the MIT Oxygen Alliance, ITRI, Intel Corporation, and the Queen's University Belfast Exchange Scheme.

References

1. C. Bregler and Y. Konig. Eigenlips for robust speech recognition. In *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 2, pages 669–672, Adelaide, Australia, Apr. 1998.
2. S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, Aug. 1980.

3. T. Hazen. Visual model structures and synchrony constraints for audio-visual speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 14(3):1082–1089, May 2006.
4. T. Hazen, K. Saenko, C. La, and J. Glass. A segment-based audio-visual speech recognizer: Data collection, development and initial experiments. In *Proc. of Int. Conf. on Multimodal Interfaces*, State College, Pennsylvania, Oct. 2004.
5. T. Hazen, E. Weinstein, R. Kabir, A. Park, and B. Heisele. Multi-modal face and speaker identification on a handheld device. In *Proc. of the Workshop on Multimodal User Authentication*, Santa Barbara, California, Dec. 2003.
6. 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, Nov. 2003.
7. 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*, volume 2, pages 688–694, Vancouver, Canada, July 2001.
8. B. Heisele, T. Serre, S. Prentice, and T. Poggio. Hierarchical classification and feature reduction for fast face detection with support vector machines. *Pattern Recognition*, 36:2007–2017, 2003.
9. S. McKenna and S. Gong. Recognising moving faces. In H. Wechsler, P. Phillips, Bruce. V., F. Soulie, and T. Huang, editors, *Face Recognition: From Theory to Applications*, pages 578–588. Springer-Verlag, Berlin, Germany, 1998.
10. J. Ming, T. Hazen, and J. Glass. Speaker verification over handheld devices with realistic noisy speech data. In *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006.
11. J. Ming and F. Smith. A posterior union model for improved robust speech recognition in nonstationary noise. In *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages 420–423, Hong Kong, Apr. 2003.
12. A. Park and T. Hazen. ASR dependent techniques for speaker identification. In *Proc. of Int. Conf. on Spoken Language Processing*, pages 1337–1340, Denver, Colorado, Sep. 2002.
13. A. Park and T. Hazen. A comparison of normalization and training approaches for ASR-dependent speaker identification. In *Proc. of Int. Conf. on Spoken Language Processing*, Jeju Island, Korea, Oct. 2004.
14. D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, January/April/July 2000.
15. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, Germany, 1995.
16. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 511–518, Kauai, Hawaii, Dec. 2001.
17. 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*, pages 48–54, Zurich, Switzerland, Aug. 2002.
18. R. Woo, A. Park, and T. Hazen. The MIT Mobile Device Speaker Verification Corpus: Data collection and preliminary experiments. In *Proc. of Odyssey, The Speaker & Language Recognition Workshop*, San Juan, Puerto Rico, June 2006.
19. W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, Dec. 2004.