THE MULTI-BIOMETRIC, MULTI-DEVICE AND MULTILINGUAL (M3) CORPUS

Helen Meng[1], P.C. Ching[1], Tan Lee[1], Man Wai Mak[2], Brian Mak[3], Y.S. Moon[1], Man-Hung Siu[3],
Xiaoou Tang[1], Henry P.S. Hui[1], Andrew Lee[1], Wai-Kit Lo[1], Bin Ma[1] and Eddie K.T. Sio[1]

[1]The Chinese University of Hong Kong (CUHK), [2]The Hong Kong Polytechnic University (HKPolyU),
[3]Hong Kong University of Science and Technology (HKUST)
Email: hmmeng@se.cuhk.edu.hk

ABSTRACT

This paper presents an overview of the M3 (multi-biometric, multi-device and multilingual) Corpus. M3 aims to support research in multi-biometric technologies for pervasive computing using mobile devices. The corpus includes three biometrics – facial images, speech and fingerprints; three devices – the desktop PC with plug-in microphone and webcam, Pocket PC and 3G phone; as well as three languages of geographical relevance in Hong Kong – Cantonese, Putonghua and English. The multimodal user interface can readily extend from desktop computers to mobile handhelds and smart phones which have small form factors. Multimodal biometric authentication can also leverage the mutual complementarity among modalities, which is particularly useful in dynamic environmental conditions encountered in pervasive computing. For example, we should emphasize facial images over speech when verification is performed in noisy acoustic environments. M3 is designed to include variable environmental factors indoors and outdoors, simultaneous recordings across multiple devices to support comparative and contrastive investigations, bilingual text prompts to elicit both application-oriented and cognitive speech data, as well as multi-session data from a fairly large set of subjects.

1. INTRODUCTION

We describe the corpus collection effort in the *M3* project. This project aims to develop human-centric interface technologies that support *secure computing* by *a diversity of users* in *a variety of usage contexts*. Human-centric interfaces embrace the user's natural communicative modalities, such as sight and hearing, at the center of human-computer interaction. These modalities can readily extend from desktop computing to pervasive computing with small-form-factor devices such as mobile handhelds and smart phones. As computing permeates our daily lives, security to computers, networks and content becomes an issue of prime importance. While keys, passwords, security cards, etc. are common in user authentication, they may be easily lost or forgotten especially when one has to keep track of an increasing set of tokens. These developments motivated our research in multimodal user interfaces that are secured with minimally intrusive biometric authentication functions, in the context of pervasive computing. We sampled the functionalities of recent product models of personal digital assistants (PDAs) and smart phones and decided to include three multimodal biometrics: (i) facial images (visual input) captured by the camera; (ii) speech (audio input) captured by the microphone or telephone; and (iii) fingerprint (tactile input) captured by the fingerprint reader. As a concept illustration, we developed a preliminary system with these multi-biometric authentication functionalities. The video demonstration[1] of this system is cast in a scenario where a traveler is making a hotel reservation using his PDA, a PocketPC (PPC). The interaction involves entering his credit card number and verifying his identity using the three biometrics, combined with a decision fusion strategy. This usage context relating to pervasive computing calls for a *suite* of enabling technologies, including: (1) *speaker verification* that verifies the claimant based on his/her voiceprint; (2) *verbal information verification* that verifies verbal message content based on the claimant's personalized / cognitive information*;* (3) *noise-robust speech processing* that counteracts performance degradation (in verification) due to ambient noise; (4) *transducer normalization* that counteracts performance degradation due to varying distortions in handset mismatch between sessions of user enrolment and verification; (5) *robust facial identification* that is insensitive to variations in illumination and face poses; (6) *robust fingerprint authentication* that is insensitive to off-centered placement of fingerprints and varying degrees of finger pressure; (7) *data integrity assurance* that matches lip motions with the speech signal (audio-visual speaker verification) to prevent fraudulent system penetration due to pre-recorded facial images; and (8) *decision fusion strategies* that combine multimodal

[1] The video may be found at: www.se.cuhk.edu.hk/hccl/M3Corpus/BiometricDemoVideo.html.

biometrics and leverage their mutual complementarity to improve overall performance in authentication.

Development of the aforementioned technologies is critically dependent on the availability of appropriate multimodal data. This motivates the creation of the M3 Corpus with the following design objectives:

(a) Maximize the number of subjects based on available resources in order to achieve statistical significance;
(b) Capture multimodal data samples with rich variability in recording conditions, e.g. variations in spoken content, acoustic noise, illumination, face poses, fingerprint placement and pressure, etc., in order to develop robust technologies for pervasive use;
(c) Capture multilingual speech data with geographical relevance to Hong Kong – predominant languages include English and two dialects of Chinese (Cantonese and Putonghua);
(d) Capture audio-visual / multimodal data with multiple devices, such as PocketPCs (PPCs) and 3G (third generation) phones, in order to support investigations in transducer normalization.

It may be of interest to note that by virtue of its design objectives, the M3 Corpus includes multi-device data that is geo-culturally unique to some degree. M3 includes a facial image corpus of Chinese subjects, which is uncommon in facial identification research. Furthermore, the facial images include lip motions that correspond to verbal information spoken in both English and Chinese languages. The multi-device speech recordings also include accented English spoken by native Cantonese or Putonghua speakers.

2. PREVIOUS WORK

As mentioned above, the M3 Corpus is unique in that it targets collection of multimodal biometrics data (speech, facial images, fingerprints) for the usage context of pervasive computing using mobile devices (PPCs and 3G phones); and largely relevant to the geographical context of Hong Kong, i.e., multilingual English/Chinese verbal content and Chinese facial images. However, the M3 collection effort draws heavily from the experiences of previous multimodal data collection efforts. For example, M2VTS (Messer et al. 1998) and XM2VTS (Messer et al. 1999) are for user authentication, with multimodal, audio-visual speech data from 37 and 295 subjects respectively and each subject utters two digit sequences and one sentence. IBM also has an audio-visual speech corpus to support large-vocabulary speech recognition research (Neti et al. 2000). The CUAVE database (Patterson et al., 2002) is a speaker-independent corpus in support of audio-visual speech recognition research, with over 7,000 utterances of isolated as well as connected digits spoken by approximately 50 subjects. CUAVE also include speech recordings from pairs of simultaneous speakers in order to support research in multi-speaker solutions. The AVOZES corpus contains audio-video speech data from 20 Australian English speakers with systematic coverage of phonemes and visemes (Goeke & Millar, 2004). Additionally, aspects of speaking-face data corpus design methodologies are presented in (Millar et al. 2004).

3. DATA COLLECTION SETUP

The data collection setup for M3 is especially designed to capture multimodal, multilingual data from multiple devices from a variety of ambient recording conditions. Details are described in the following:

3.1. Recording Devices

A series of devices are used to record the multimodal data in the M3 corpus. Audio-visual data recording is captured with the Pocket PC (PPC), 3G phone and also a desktop PC with a web-camera and plug-in microphone. We also include a digital camera to record static facial images, as well as an optical sensor for capturing fingerprints. Details of the equipment are listed in Table 1.

Device	Configuration	Format
Pocket PC	Model: HP iPAQ H2200 series	
	Camera: Pretec CompactCamera	mmf
	OS: PocketPC 2003 Premium	
	Video: 232 x 174, 24fps	
	Audio: 22kHz, 16 bits mono	wav
3G Phone	Model: NEC C616	
	Video: 176 x 144, 14.34 fps	mp4
	Audio: 8 kHz, 16 bits mono	wav
PC (speech & AV)	Config: Pentium 3 996 MHz 512M	
	OS: Windows XP	
	Webcam: EagleTec ET-VCCCD	avi
	Video: 320 x 240, 30 fps	
	Audio: 16 kHz, 16 bit mono	wav
	Microphone: Shure BG 1.1 cardioid	
Camera	Model: Nikon Coolpix 5000	jpg
	Resolution: 2056 x 1920	
PC (finger-print)	Config: Pentium 3 996 MHz 512M	bmp
	OS: Redhat Linux Release 9	
	Reader: SecureTouch optical sensor	
	Resolution: 320 x 320 8bit grayscale	

Table 1. Recording devices used in the M3 corpus, together with information on system configurations and data formats.

All these devices are used to capture data from every subject. *Multi-device data* facilitate calibration as well as contrastive and comparative experiments in investigations pertaining to transducer normalization.

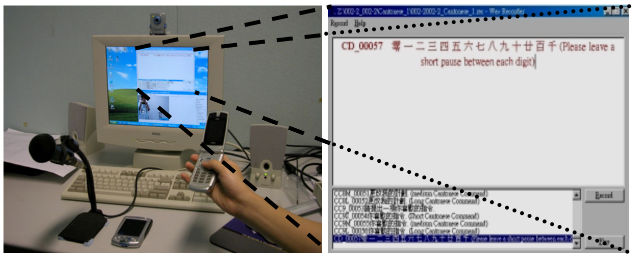
A noteworthy detail is that the recording buffer of the 3G phone has the capacity for approximately one minute of speech. Hence, special care is taken to ensure that the input speech utterance is recorded in its entirety, or else the recording has to be repeated.

3.2. Environments

The usage context of pervasive computing presents a multitude of environmental factors, such as variations in illumination for facial image capture and ambient noise in speech recordings. Therefore, we included both indoor and outdoor recording environments for every subject in the collection of the M3 Corpus.

3.2.1 Indoors

The indoor setting is a *quiet* recording room (~6m^2) with noise insulation linen. Recording is done simultaneously with the desktop PC, PPC and 3G phone (see Figure 1a). During the recording process, an assistant helps position the 3G phone to maintain recording quality. A software program guides the subject with textual prompts for speech utterances (see Figure 1b). The program allows the subject to navigate through the list of prompts by moving the scrollbar, as well as initiate/terminate utterance recording by pressing on the start/stop buttons. A webcam is also positioned on top of the desktop monitor to capture a video of the subject throughout the recording process.



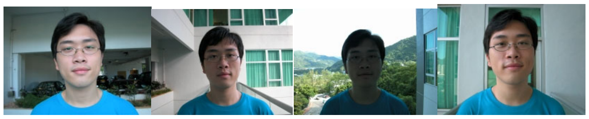
CD_00057 零一二三四五六七八九十廿百千 (Please leave a short pause between each digit)

(a) (b)

Figure 1. Indoor data collection setup (a) placement of recording devices (b) program interface.

3.2.2 Outdoors

The outdoor setting is on a foot-bridge connecting two adjacent academic buildings and facing a car-park. This location was chosen due to its unique lighting condition for achieving illumination variations in visual data collection, as well as possible background noise in audio data collection. Ambient light on the bridge originate mainly from one direction (east). When the subject faces the other three cardinal directions, we can achieve side-lighting from both sides and also back-lighting due to blockage of the buildings (see Figure 2). To avoid poorly illuminated facial images, data collection is avoided on rainy days and during late afternoons.



(a) (b) (c) (d)

Figure 2. Outdoor facial images with the subject facing different directions. The source of illumination is (a) in front; (b) to the left; (c) at the back; and (d) to the right of the subject.

3.3. Designing Bilingual Text Prompts for Speech Utterances

Speech data in the M3 collection aims to support research in speaker authentication based on vocal tract characteristics as well as verbal content in the spoken messages (i.e. verbal information verification). The spoken utterances also need to cover both English and Chinese. Hence we design a series of text prompts to elicit the appropriate speech utterances from the subjects. The text prompts falls largely into three categories: (i) The general set is frequently used in most applications. It includes the English alphabet, digits and common commands, as well as Chinese digits and commands in both dialects,[2] i.e. Cantonese and Putonghua. (ii) A domain-specific set based on possible user requests in the tourism domain. (iii) The cognitive set relates to the subject's personal profile and may be based on fact (e.g. the subject's horoscope) or opinion (e.g. the subject's favorite color). Verbal information verification using the subject's cognitive data is often preferred over spoken passwords (such as a chosen digit string), since cognitive data is easier for the user to remember. The text prompts for cognitive data are composed with reference to many websites that requires setup of personal profiles. These prompts are also parallel across languages. Subjects are asked to provide responses in *short*, *medium* or *long* forms, while maintaining consistency in the cognitive content.

Overall, there are 27 English text prompts, including one for the alphabet, three for digits, 7 for commands, 5 for tourism-related[3] requests and 11 for cognitive data. The Chinese text prompts are largely parallel, except for having no alphabet prompt and a single digit prompt. This is summarized in Table 2.

Text Prompts (categories)	# in English	# in Chinese
General	11	8
Domain-specific (Tourism)	5	5
Cognitive	11	11

Table 2. Number of bilingual textual prompts used to elicit speech utterances in the M3 Corpus.

[2] Unlike English, Chinese has no alphabet. The set of digits we include also vary across languages due to different verbalization methods for counting.

[3] We are developing a multimodal interface for pervasive computing in the tourism domain, where the mobile computer helps tourists with navigation.

3.4. Multimodal Data Recording Procedures

3.4.1 Subjects

We invited over a hundred subjects for M3 data collection at the initial stage. Most of these subjects are from the college community (i.e. from the age group of 20-30) and speak English as well as Cantonese and/or Putonghua. The subjects are invited to attend *three* sessions of recording, with at least a one-month interval in between sessions. These sessions will provide data for enrolment/verification in the form of a training set, development test set and test set. During each session, data collection is conducted *sequentially* for all modalities, namely, fingerprints, static facial images, followed by *simultaneous* recording of face videos (with lip motions) and speech. Since each session involves multiple modalities and multiple devices for every subject, typical session durations are long, spanning 3.5 hours on average per session per subject. Every subject was asked to fill out a form during the first session, to record their answers to the text prompts for cognitive data (as mentioned in the previous section). This step serves to maintain consistency in recorded verbal content lest the subjects forget their answers when they return for subsequent sessions.

The multi-session data collection effort saw some degree of attrition, i.e. certain subjects did not return for subsequent sessions of recording. In the end, we have 39 subjects who completed all three sessions. Another 108 subjects are later invited to become imposter speakers, as will be explained later.

3.4.2 Capturing Fingerprints

For each subject, we record 20 impressions of their right index fingerprint during every session. The subjects are instructed to position their finger centrally with respect to the fingerprint reader, while they view the captured fingerprint on-screen. The multiple impressions capture variability across fingerprints in terms of placement and pressure. At the end of each collection session, an assistant also performs a quick manual check to ensure that all fingerprint images are well positioned with adequate contrast.

3.4.3 Capturing Facial Images and Videos

- **Static facial images**: we use a digital camera to capture callibrative facial images with a plain, blue (uncluttered) backdrop in the *indoor environment*. Five static images are captured per subject per session in order to include face poses with different orientations (see Figure 3). To ensure some degree of consistency in the face poses across subjects, each person was asked to look at a paper cross (see Figure 4) with a hole in the center where the camera lens is placed and with four dots at the terminal end of each beam.

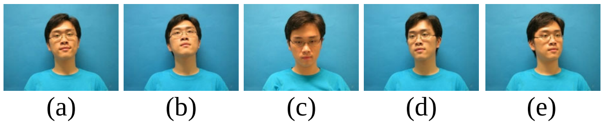


Figure 3. Indoor facial pictures when the subject was asked to look (a) squarely forward (b) upwards (c) downwards (d) to the right (e) to the left.

We also use the digital camera to capture four static facial images in the *outdoor environment*, corresponding to illumination from four different sides (see Figure 2). Hence each subject has 9 static facial images per session – five from indoors and four from outdoors.

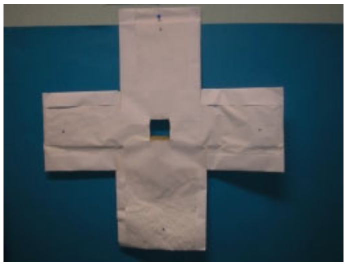


Figure 4. Paper cross with a hole in the center (where the camera lens is placed) and four dots at the terminal end of each beam. This is used to capture the different face pose of subjects in their static facial images.

- **Facial videos**: we capture the full range of face poses in three videos where the subject is asked to perform three head movements – (i) from up to down; (ii) from left to right; (iii) head rotation in a full circle (see Figure 5). These video frames cover most of the face poses in real-life applications when the exact frontal view of the subject's face is guaranteed. For the *indoor setup*, all three devices (webcam, PPC and 3G phone) are used *sequentially*. For the *outdoor setup*, only the PPC and 3G phone are used and the subject is asked to face the illumination source to maximize clarity. All these steps amount to 15 facial videos per subject per session and from *multiple devices*.

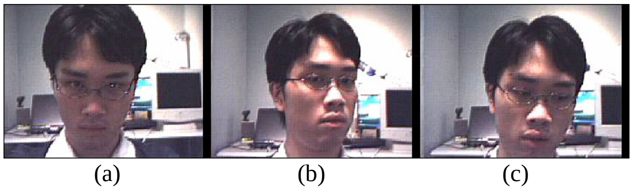


Figure 5. Facial videos with three head movements: (a) up-down; (b) left-right; and (c) full-circle rotation.

3.4.4 Capturing Speech Data (audio-only)

This data aim to support specifically speaker authentication and verbal information verification research. There are two main portions:

- **Multi-session speech data from the 39 "enrolled" subjects**: we shall refer to the subjects who completed all three sessions of data collection as the "enrolled subjects". They have sufficient data for speaker enrolment and verification. This audio-only data are collected in the indoor setup *simultaneously* with all three devices (desktop with microphone, PPC and 3G phone). In each session, a subject is prompted with a total of 22 textual prompts in English and 19 prompts in Chinese, with distributions as shown in Table 3.

Textual Prompts (categories)	# in English	# in Chinese
General	8	5
Domain-specific (Tourism)	5	5
Cognitive	9	9

Table 3. Number of bilingual textual prompts used to elicit speech utterances in the M3 speech-only, indoor sub-corpus.

For each prompt, the subject is asked to provide responses in three forms – short, medium and long. These incorporate a certain level of lexical variability while keeping the key verbal content consistent. Each subject is asked to provide speech utterances in English and one or both of the Chinese dialects.

- **Single-session speech data from 108 subjects:** This is a disjoint set of subjects who may serve as imposters in speaker authentication experiments. Putonghua speech is sparse here due to happenstance – the local community has many more Cantonese speakers than Putonghua speakers. Hence we plan to focus more on the Cantonese dialect in our upcoming research work. Based on the collected data, we generated a master-file that plans the details of each attack, including the imposter speaker, the targeted enrolled speaker, the corresponding textual prompt and whether the verbal content is compromised in the attack.

3.4.5 Capturing Audio-Visual Speech Data

By audio-visual (AV) speech data, we are referring to the recording of a "talking face". There are also two sections of AV speech data:

- **Desktop AV speech data (indoor only)**: we recorded both audio and video of the enrolled speakers uttering their responses to the prompts shown in Table 3. This is done using a desktop PC with a plug-in microphone and a webcam.

- **Mobile device AV speech data (indoor and outdoor)**: these are recorded with both the PPC and 3G phone *sequentially*. Each enrolled subject utters 12 utterances in each language – 8 relating to cognitive prompts and 4 to general prompts. These are spread evenly across all four illumination conditions during outdoor recording, as summarized in Table 4.

Illumination Source	# general prompts	# cognitive prompts
In front	1	2
From left	1	2
From behind	1	2
From right	1	2

Table 4. Distribution of textual prompts along each illumination condition in AV speech recordings using mobile devices in the outdoor setup.

The subject is instructed to hold the mobile device at arm's length (see Figure 6). The approximate distance between the subject's eyes and the device should be around 35cm such that the entire face appears in the captured video.

All utterances are recorded again with mobile devices in the *indoor setup* which has no illumination variations. Parallel data recorded both indoors and outdoors can support contrastive experiments investigating effects of ambient conditions.

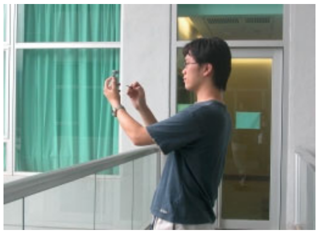


Figure 6. Posture for holding the mobile device for AV speech collection in the outdoor setup.

4. CORPUS COLLECTION, STATISTICS, TRANSCRIPTION AND ORGANIZATION

4.1. Difficulties in the Collection Process

Over the past two years, we devoted much effort in selecting devices, conducting pilot collection sessions and recruiting subjects. We found that collecting fingerprint data is especially difficult because some recruited subjects later decided that they are reluctant to provide their fingerprint data due to privacy concerns.[4] Attrition of subjects (i.e. those who did not return for subsequent sessions of data collection) also presented difficulties in terms of the loss in number of enrolled subjects. Two of our staff members checked the recorded data and pruned data files of very low quality. They also deleted audio segments that recorded nothing but sounds of mouse-clicks (due to the subject's interactions with the text-prompting software). However, if the recorded mouse-clicks are superimposed on speech, the audio segments are preserved. There are also speech files

[4] Four of our enrolled subjects decided that they will not provide images of their fingerprints.

which inevitably recorded the buzz of the 3G phone or the subject's mobile phone while it searched for a channel. In later sessions, subjects were asked to turn off their own mobile phones to reduce this impact. These cases are difficult to avoid and the corresponding speech files are preserved.

4.2. Signal-to-Noise Ratio (SNR) Measurements

In order to gauge the quality of the speech data, we ran the NIST SNR tool and discarded recordings with values below 10dB. Overall the desktop microphone recordings average 30dB, the PPC recordings average 27dB and the 3G phone recordings average 49dB.

4.3. Distribution of Subjects

Table 5 shows the distribution of the 39 enrolled subjects in M3, in terms of gender and languages used.

	Male	Female
Putonghua + Cantonese + English	9	5
Putonghua + English	4	2
Cantonese + English	16	3
Total # enrolled subjects	29	10

Table 5. Distribution of enrolled subjects in terms of gender and language used. An enrolled subject has completed all three sessions of data collection.

As mentioned earlier, we have additionally recruited 108 imposter speakers (58 male and 50 female), each offering a single session of speech data.

4.4. Speech Transcription

The speech data is processed by two transcribers. One of them processed half of the data collected in the first session, while the other processed all the remaining data. During the transcription process, recorded speech is segmented into individual speech utterances and stored as one file per utterance. Disfluencies in the speech recordings of natural speech are transcribed verbatim.

4.5. Corpus Size and Organization

The M3 corpus is organized into 6 sub-corpora, according to the modality type and devices used. This is summarized in Table 6, which also presents the size of each sub-corpus.

A series of information elements are coded in each sub-corpus. They include:

(i) The subject's identification number;

(ii) The session's identification number (3 sessions per subject);

(iii) The recording device used (including the 3G phone, desktop PC, PocketPC, digital camera, fingerprint scanner);

(iv) The language spoken (English, Putonghua, Cantonese);

(v) The utterance type (including numbers/letters, requests/commands, general questions and cognitive questions and answers;

(vi) The length of the spoken input (short, intermediate and long forms);

(vii) The location of the recording (i.e. indoors or outdoors);

(viii) The direction of illumination for outdoor face recordings (i.e. illumination from the front, from the left, from behind and from the right);

(ix) The file type (i.e. MP4 from the 3G phone, mpg from the PocketPC, avi from the desktop and txt from transcriptions)

Multiple Modalities	**Multiple Devices**	**Data Quantity**
Fingerprints	Fingerprint reader	574 MB
Static Facial Images	Digital Camera	1.74 GB
Facial Videos	PC Webcam	4.04 GB
	PPC	521 MB
	3G Phone	75.2 MB
Speech (audio-only)	PC	1.94 GB
	PPC	2.17 GB
	3G Phone	895 MB
Imposter Speech (audio-only)	PC	1.70 GB
	PPC	1.86 GB
	3G Phone	724 MB
Audio-visual Speech	PC Webcam	3.28 GB
	PPC	28.6 GB
	3G Phone	521 MB

Table 6. Sub-corpora in M3, classified according to modality types and devices used in recording.

5. INITIAL R&D RELATED TO THE M3 CORPUS

We are using the M3 Corpus to support research in a suite of enabling technologies for pervasive computing, as described in the introduction of this paper. The current section presents very brief descriptions of our initial progress along several fronts.

5.1. Multi-lingual and Text-Independent Speaker Verification

M3 can support investigations in English-Chinese, text-independent (better described as text-constrained[5]) speaker

[5] Text-constrained may be a more appropriate description for M3 since its lexical coverage is presently constrained to vocabularies describing personal profiles.

verification (SV) technologies. We leveraged our previous work (Ma & Meng, 2004) and applied it to the M3 corpus, in order to establish some preliminary benchmark results. Only desktop microphone recordings are used for now. We compute 12 mel-frequency cepstral coefficients for every 10ms using a 25.6ms Hamming window, together with Cepstral Mean Normalization. The 12-dimensional vector is appended with the delta and delta-delta vectors to give 36 coefficients in all.

An experiment is performed based on the speech data of a subset of the enrolled subjects (39 in all). The first session is used as the training set, the second as the development test set and the third as the test set. We selected bilingual data (English with either Cantonese or Putongha) for training the speaker model for every subject. Silent segments at the head and the end of speech utterances were removed. We also trained a universal background model (UBM) with data in all available languages and from the all the enrolled subjects. A Gaussian Mixture Model is trained and the number of mixtures was tuned with the development test set. We used speaker-dependent thresholds for evaluation with the development test and test sets. Speaker models used 512 mixtures and the UBM used 1024 mixtures. This SV system achieves an EER of 2.20% on the development test set and 2.60% on the test set. Analysis shows that performance on subject with id=15 was particularly poor since there is great mismatch in SNR between his training and test sets. Excluding this subject brings the EER to 2.29% for the test set. If we test the existing trained models on the test data recorded with the PPC and 3G phone (hence with device mismatch between training and test sets), the SV performance (EER) obtained are 6.82% and 27.88% respectively.

5.2. Fuzzy Logic Decision Fusion in Multimodal Biometric Authentication

We conducted a preliminary investigation in the use of fuzzy logic for decision fusion in multimodal biometric authentication that involves speech, fingerprint and facial image verification (Lau et al., 2004). The study was conducted on the pilot collection data which is the pre-cursor to M3. Decision fusion aims to derive synergy across biometrics to achieve overall robustness, e.g. when the ambient illumination is too dark for reliable facial recognition, the overall authentication decision will revert to the other biometrics. In comparison with straightforward fusion techniques such as voting and weighted averaging, the fuzzy logic decision fusion strategy is more adaptive. For example, its fingerprint weighting considers the placement and pressure in input fingerprint; its facial image weighting considers illumination and face-finding confidence in the input facial images; and its speech weighting considers SNR in the input speech. Results indicate that fuzzy logic fusion outperforms straightforward fusion with majority voting by a factor of three.

5.3. Initial Prototype System

We have developed a preliminary multi-biometric authentication system to serve as our concept demonstration (see footnote 1 for the video demonstration). This system is a simplified integration based on a client-server architecture. The PPC client is the HP iPAQ 5550 that can capture facial images with a compact flash camera, speech with the built-in microphone and fingerprint with the built-in fingerprint reader. The multimodal biometric data are encapsulated in a SOAP message and transmitted to the central biometrics server which distributes the multimodal data to the fingerprint authentication server, the facial image verification server, the speaker verification server and the verbal information verification server (VIV). The VIV process is currently based on a keyword-spotting approach for detecting specific content words in the prompted speech message. Authentication results from individual modalities are then transmitted back to the central server, which performs decision fusion using the fuzzy logic framework to generate the finalized, overall authentication output. We plan to update the modules of this prototype system as we continue to develop various technologies in multimodal biometric authentication.

6. CONCLUSIONS AND FUTURE WORK

This paper presents the design, collection and organization of the M3 (multi-biometric, multi-device and multilingual) Corpus. It aims to support research in multi-biometric technologies for pervasive computing using mobile devices. The corpus includes three biometrics – facial images, speech and fingerprints; three devices – namely the desktop PC with plug-in microphone and webcam, Pocket PC and 3G phone; as well as three languages of geographical relevance in Hong Kong – Cantonese, Putonghua and English. The multimodal user interface can readily extend from desktop computers to mobile devices with small form factors. Multimodal biometric authentication can leverage the mutual complementarity among modalities, which is particularly useful in dynamic environmental conditions encountered in pervasive computing. The design goals of M3 are to include variable environmental factors indoors and outdoors, simultaneous recordings across multiple devices to support comparative and contrastive investigations, bilingual text prompts to elicit both application-oriented and cognitive speech data, as well as multi-session data from a fairly large set of subjects. We also presented a suite of technologies for which research investigations can be facilitated by the M3 corpus and provided a brief description about our ongoing research along some of these directions, including multilingual, text-constrained speaker verification, as well as fuzzy logic decision fusion for multi-biometric authentication. We are also working on audio-guided video face segmentation (Li, 2006) and audio-visual speaker verification (Wu et al. 2006). We plan to make the M3 Corpus available to the research community in the hope of facilitating more work in the multimodal biometrics area.

5 ACKNOWLEDGMENTS

This project is partially supported by the HKSAR Government Central Allocation Grant CUHK1/02C. This work is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies. We would also like to thank Mr. C.W. Lau for his contribution as well as the subjects from CUHK, HKUST and HKPolyU for participating in the data collection.

6 REFERENCES

Goecke, R. and Millar, J., “A Detailed Description of the AVOZES Data Corpus,” Proc. of the 10th Australian Int. Conf. on Speech Science & Technology, SST 2004.

Lau, C.W., Ma, B., Meng, H., Moon, Y.S. and Yam, Y., “Fuzzy Logic Decision Fusion in a Multimodal Biometric System,” Proc. ICSLP 2004.

Li, Z. F., “Audio-guided Video Based Face Recognition,” Ph.D. Thesis, The Chinese University of Hong Kong, February 2006.

Ma, B. and Meng, H., “English-Chinese Bilingual Text-Independent Speaker Verification,” Proc. ICASSP 2004.

Messer, K., Matas, J. and Kittler, J., “Acquisition of a large database for biometric Identity Verification,” Proc. BIOSIGNAL 1998.

Messer, K., et al., “XM2VTSDB: The Extended M2VTS Database,” Proc. Audio and Video-based Biometric Person Authentication, AVBPA 1999.

Neti, C. et al., “Audio-Visual Speech Recognition,” Workshop Report, CSLP, Johns Hopkins University, 2000.

Patterson, E. et al., “CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research, Proc. ICASSP 2002.

Millar, J., Wagner, M. and Goecke, R., “Aspects of Speaking-Face Data Corpus Design Methodology,” Proc. ICSLP2004.

Wu, Z., Cai, L. and Meng, H., “Multi-level fusion of Audio and Visual Features for Speaker Identification,” Proc. Int. Conf. on Biometrics, 2006.