EFFECTS OF DEVICE MISMATCH, LANGUAGE MISMATCH AND ENVIRONMENTAL MISMATCH ON SPEAKER VERIFICATION

Bin Ma[1], Helen M. Meng[1] and Man-Wai Mak[2]

[1]Dept. of Systems Engineering and Engineering Management, The Chinese Univerisity of Hong Kong
[2]Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University
[1]{bma, hmmeng}@se.cuhk.edu.hk, [2]enmwmak@polyu.edu.hk

ABSTRACT

Device, language and environmental mismatch adversely affect speaker verification (SV) performance. We investigate such effects empirically based on the M3 (multi-biometric, multilingual and multi-device) Corpus [1]. Device mismatch (among 3G phone, Pocket PC and a desktop PC plug-in microphone) brings relative performance degradation of roughly 11-fold; language mismatch (between English and Cantonese) brings 288% and environmental mismatch (between office environment and recording studio) brings 162%. In particular, it is found that speaker model adapted from a device-independent UBM by MAP is less sensitive to device mismatched testing. Additionally, durational variations within two-second utterances may cause a relative change of 10-fold in SV performance.

Index Terms— Speaker verification, biometrics corpus, M3 speaker verification evaluation

1. INTRODUCTION

Speaker verification is the process of authenticating the speaker's claimed identity based on his/her input utterances. This technology plays a key role in securing computing for human-centric computer interfaces. In real-time applications, the proliferation of mobile, handheld devices present challenges for speaker verification. For example, mobile use means that speaker verification technically needs to handle a variety of environmental conditions. Also, different audio input devices (e.g., microphones on PDAs or cellphones) may induce significant variations in the quality of captured speech. Some techniques, such as feature mapping [2], speaker model synthesis [3] and handset normalization [4], have been proposed to alleviate this problem. The language uttered may also affect SV performance, as demonstrated in our previous work [5]. The length of testing utterance segments is another factor affecting SV performance. In particular, it has been shown that the EER of SV system is exponentially related to the length of test segment [6]. The current study attempts to qualify such effects based on SV experiments with the M3 speech data, which contains multilingual, multi-device and data for mobile use, as will be elaborated later.

2. THE SPEECH DATA OF M3 CORPUS

The M3 corpus is designed to support research in multi-biometric technologies for pervasive computing using mobile devices. Three kinds of biometrics, three devices, as well as three languages, are included in M3. Our research focuses on the speech data in M3. A brief introduction to M3 speech data is presented in this section.

2.1 Speech data collection setup

During data collection, the multilingual speech data are captured from multiple devices from two recording conditions: an open laboratory and a recording room. The devices include a Pocket PC (PPC), a 3G phone and a desktop PC plug-in microphone. Details are listed in Table 1. The speech data across devices are recorded simultaneously.

Device	Configuration	Format
Pocket PC	Model: HP iPAQ H2200 series	WAV
	Audio: 22kHz, 16 bits mono	
3G phone	Model: NEC C616	WAV
	Audio: 8 kHz, 16 bits mono	
Desktop PC plug-in microphone	Config: Pentium 3 996 MHz 512M	WAV
	Audio: 16 kHz, 16 bits mono	
	Microphone: Shure BG 1.1 cardioid	

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

2.2. Speaker description

We invited subjects from the college community (age range from 20 to 30) to attend the three sessions of M3 data collection, with at least three-week intervals between sessions. The subjects speak English as well as Cantonese and/or Mandarin. We have 32 subjects (23 males and 9 females) who completed all three sessions. They form the *multi-session speaker set*. Another 108 subjects are later invited to provide a single session of data. They form the *single-sessin speaker set*.

2.3. Utterance design

We designed a series of text prompts to elicit the subjects' speech utterances that are appropriate for two purposes. First, the spoken utterances cover both English and Chinese (Cantonese or Mandarin). Second, the utterances are recorded in *short, medium* and *long* forms, while the consistency in the cognitive content is maintained at the same

time. The text prompts fall into three categories: (i) The general set is frequently used in most applications. It contains the alphabet, digits and common commands. (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 (e.g. the subject's horoscope) or opinion (e.g. the subject's favorite food).

2.4. Speech data quality

In order to gauge the quality of the speech data, the NIST SNR tool is used on all speech utterances in M3 corpus. Utterances with SNR value below 10dB were discarded. On average, the recordings of desktop microphone, PPC, and 3G phone have SNR of 29dB, 27dB, and 49dB respectively. Analysis of the SNR values across the M3 speech data reveals that the first session generally has lower speech quality than the second and third. This is because the recording environment of first recording session per subject is open laboratory and that of second and third recording session is a recording room.[1]

3. BASELINE SV SYSTEM

We developed a GMM-UBM SV system [7], which is generally used in text-independent SV task. It is used to establish a preliminary SV benchmark of M3 speech data.

Speech data acquired with different devices have different sampling rates (PPC: 22.05KHz and 3G phone: 8KHz). Hence we resampled these data to conform with the sampling rate of desktop PC speech (16KHz). As silent segments in the recordings do not carry speaker identity information, we used speech activity detection to remove them. After silence removal, we use mel-frequency cepstral coefficient (MFCC) as the main feature vector. 19 MFCCs are computed for every 10ms using a 25.6ms Hamming window. Cepstral mean subtraction (CMS) is applied. The 19-dimensional vector is appended with the delta vectors to give 38 coefficients in all.

Two kinds of speaker models are used. They are the traditional GMM and adapted GMM. The traditional speaker GMM is trained using speaker-specific training data with the EM algorithm. Each speaker GMM uses 256 mixtures and the universal background model (UBM) uses 2048 mixtures. The adapted speaker model is derived by adapting the parameters of the UBM using the speaker's training speech and a form of maximum a posteriori (MAP) estimation [7]. The adaptation approach is to derive the speaker's model by updating the well-trained parameters in the background model via adaptation.

4. EXPERIMENTAL SETUP

Under the GMM-UBM framework, the data usage, front-end processing specific for M3 speech corpus, as well as the SV performance measurement is described in the following.

[1] This arrangement was not by design as we had to move our laboratory from one building to another during the recording process.

4.1. Data partition of M3 speech corpus

We define the data partitioning scheme of M3 as shown in Table 2. Session 2 is used for training and sessions 1 and 3 for testing respectively. For each enrolled speaker, there are 108 true speaker trials. To keep a gender-balanced number of imposter trials, 8 randomly selected male speakers and 8 female speakers are selected from the 32 speakers in the multi-session speaker set (excluding the claimant). These 16 speakers, plus 58 speakers (29 males plus 29 females) of single-session speaker set, are used to impersonate each claimant. Hence, there are 74 (37 males and 37 females) imposters in total. The speech data of 40 speakers (20 males and 20 females) in single-session speaker set is used to train a device-independent universal background model. This set of speakers will not be further used as imposters.

Data sets		Data source	Description
Device-independent UBM (Training data)		40 speakers Single-session speaker set Disjoint with imposters	Languages: - English; Chinese Devices: - PC; PPC; 3Gphone
Enrollment	Device-dependent speaker model (training data)	Multi-session speaker set Session 2 for each speaker	Languages: - English; Chinese Lengths: - Short; medium; long Devices: - PC; PPC; 3Gphone 117 utterances
Verification	True speaker testing data	Matched environmental testing: - Multi-session speaker set - Session 1 of each speaker Mismatched environmental testing: - Multi-session speaker set - Session 3 of each speaker	Languages: - English; Chinese Lengths: - Short; medium; long Devices: - PC; PPC; 3Gphone 108 utterances
	Imposter testing data	Multi-session speaker set - 16 speakers, randomly selected, excluding the claimed speaker Session 2 for each speaker - 58 speakers	Languages: - English; Chinese Lengths: - Short; medium; long Devices: - PC; PPC; 3Gphone

Table 2 Data partitions in the M3 speech corpus.

4.2 Performance measure

The performance of a SV system is usually estimated by two kinds of error measures: false acceptance rate (FAR) and false rejection rate (FRR). False acceptance occurs when the system incorrectly accepts an impostor, and false rejection occurs when the system incorrectly rejects a true speaker. The equal error rate (EER), which is obtained when FAR equals FRR, is used for reporting the experimental results in this work.

5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1. Effect of language mismatch on speaker verification performance

We use the PC speech data of 27 speakers who speak English and Cantonese to investigate the effect of language mismatch between enrollment and verification on SV performance. Cross-language testings are implemented. The

results[2] are shown in Table 3. For enrollment in English, performance degrades from EER 1.58% (with English verification data) to 4.64% (with Cantonese verification data), which is a 194% performance degradation. For enrollment in Cantonese, performance degrades from EER 1.65% (with Cantonese verification data) to 6.42% (with English verification data), which is 288% performance degradation.

EER (%)		Testing languages	
		English	Cantonese
Training languages	English	1.58	4.64
	Cantonese	6.42	1.65

Table 3. SV performances of language-mismatched enrollment and verification cases.

5.2. Effect of environmental mismatch on speaker verification performance

We implement environmental mismatch SV experiments using speaker models trained with Session 2 data (recording room). Testing data include Session 1 (open-lab, i.e., mismatched environment) and Session 3 (recording room, i.e., matched environment). The experimental results are shown in Figure 1.

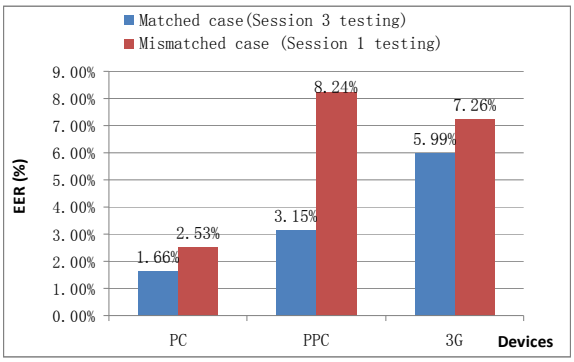

Figure 1. SV performances under match and mismatched recording environments.

We can see that environmental mismatch between enrollment and verification degrades SV performance. For the desktop speech data, environmental mismatch causes 52% degradation (from EER 1.66% to 2.53%). For the PPC data, environmental mismatch causes 162% degradation (from EER 3.15% to 8.24%). For the 3G phone data, environmental mismatch causes 21% degradation (from EER 5.99% to 7.26%). We can see that environmental mismatch between training and testing affects PPC's SV performance most among the three devices. This is because PPC's microphone is more sensitive to the environmental factors than the close-talking microphone on PC and the microphone on 3G phone. 3G phone's SV performance is affected by environmental factors less than PPC's. The insensitivity of the 3G phone to environmental mismatch may be due to built-in noise cancellation, which makes the 3G phone more robust to the environmental variability. In matched environmental test, 3G phone SV performance is the worst among the three devices. This is because wide-band recordings of PC and PPC are capable of capturing more speaker information than narrow-band recordings of 3G phone.

5.3. Effect of the length of testing utterances

We investigate the effect of different lengths of testing utterance on SV performance. Recall that M3 speech data contains short, medium and long response to each text and prompt, e.g., "Apple." (short); "I like apples." (medium); "Hello computer, my favorite food is apple." (long). Respective average durations for short, medium and long utterances are below 1 second, equal to 1 second and 2 seconds in the testing set. Experimental results (in Figure 2) confirm that longer testing utterances give better SV performances. Short testing segment induces SV performance degradation of 10-fold, 233% and 153% (compared with long testing segment) for desktop PC, PPC and 3G phone data respectively. Short utterances can drastically degrade SV performance, because there is sparse speaker information in short testing utterances, compared with longer utterances.

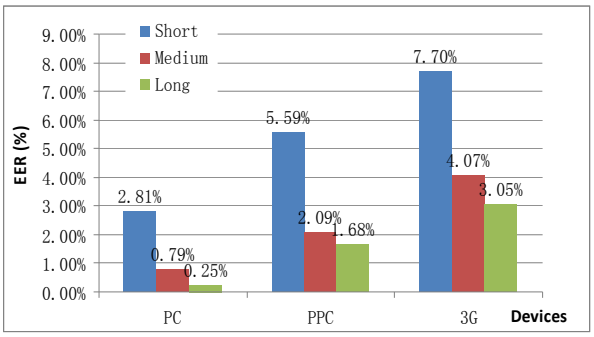

Figure 2. SV experimental results on different length of verification utterances.

5.4. Effect of device mismatch on SV performance

M3 speech data is simultaneously recorded using three devices for each speaker and thus support investigates in device-mismatched SV. The preliminary device-mismatched SV experimental results are shown in Table 4. Speaker models are created by directly applying the EM algorithm to speaker's data or by adapting the UBM using MAP adaptation.

Enrollment device (Session 2)	Verification device (Session 3)		
	PC	PPC	3G
PC (EM / MAP)	1.39/1.82	4.01/3.73	17.96/12.77
PPC (EM / MAP)	4.03/3.09	3.10/3.48	20.70/13.97
3G (EM / MAP)	16.68/13.68	13.43/12.37	5.04/5.63

Table 4. SV performance of device-mismatched enrollment and verification cases, expressed in terms of EER (%).

Each row of Table 4 shows a device-specific speaker model tested with both device-matched and device-mismatched testing data. It shows that device-matched SV performance

[2] During the evaluation, we found that two speakers (018 and 024), who performs to be outliers, greatly biased experimental results. Therefore, they are taken out from the evaluation.

is better than device-mismatched SV. For the PC model, using 3G testing data causes 11-fold performances degradation (from EER 1.39% to 17.96%). For the PPC model, using 3G phone testing data causes performance degradation of 567% (from EER 3.10% to 20.70%). For the 3G phone model, using PC testing data causes the performance degradation of 230% (from EER 5.04% to 16.68%).

We can see that for PC model, using 3G phone testing data causes more performance degradation than using PPC testing data. Similar trends are also observed in the cross testing between PPC and 3G phone. The possible reason is that the original low sampling rate (8kHz) of 3G phone recordings causes its poor performance in device mismatched tests. It is noted that PC and PPC speaker models have information in the 4-8 kHz region, whereas 3G phone recordings and models have nothing there.

The EERs in shadow show an unexpected trend of results. When speaker model is adapted by MAP, for PPC model, PC testing (EER 3.09%) gives lower EER than PPC testing (EER 3.48%). Analysis shows that when PPC model is tested with PPC speech, Speaker 015 has significantly higher EER (15.89%) that raise the average EER of matched PPC-based evaluation. When Speaker 015's PPC model is tested with PC speech, the EER is 4.00%, which is around the average. If Speaker 015 is excluded from the evaluation, the EERs of PPC model tested by PC and PPC recordings are 3.05% and 3.06% respectively. The anomalous trend is lessened. The possible reason of this phenomenon is the speaker model is adapted by MAP, in which the background model used is trained with pooled recordings of three devices. Speaker model's device-dependent characteristics are weakened by being adapted from the device-independent background model. Therefore, PPC model cannot obviously work better on PPC test than on PC test. This explanation is also supported by other device-mismatched tests. MAP adapted speaker models outperform those created by EM algorithm in device-mismatched experiments. However, in device-matched test, adapted models work not as well as EM trained models.

5.5. Lamb-sheep figure of the speakers in M3

"Lamb", "goat" and "sheep" are defined by Koolwaaij et.al in their work [8] to classify speakers in a SV system. Under this classification, a speaker with high FAR is called a lamb (easily imitated), a speaker with high FRR is called a goat (easily rejected) and a speaker with both low FAR and FRR is called a sheep. Adopting these definitions here, we present a lamb-sheep plot to analyze the speech data used in our experiments. In the lamb-sheep figure, the x-axis shows the speaker-dependent FRR and the y-axis shows the speaker-dependent FAR. Thereafter, the speakers can be located in the lamb-sheep figure according to their speaker-specific FRRs and FARs. For example, in the device-matched PC experiment, each speaker's individual FAR and FRR can be calculated with the predefined speaker independent threshold. Figure 3 shows the distribution of the M3 speakers in terms of their SV performances. The "lambs" (represented by asterisks) observed in Figure 3 are exactly those whose SV performances biased the evaluation in our experiments.

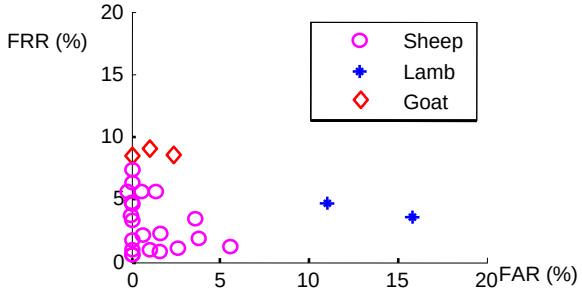

Figure 3. Lamb-sheep figure of device-matched PC testing case.

6. CONCLUSIONS

This paper empirically investigates how device, language and environmental mismatch affect SV performance based on the M3 Corpus. We found that device mismatch causes SV performance degradation of approximately 11-fold; language mismatch causes roughly 3-fold SV performance degradation and environmental mismatch causes 162% SV performance degradation. It is also found that Speaker models adapted by MAP are less sensitive to device mismatch than models created by EM algorithm. A "lamb-sheep" figure is also proposed to help analyze the speech data and speaker model's quality in a SV system.

7. ACKNOWLEDGMENTS

The work described in this paper is supported by the Central Allocation Grant from the Research Grants Council of the Hong Kong SAR (CUHK 1/02C).

8. REFERENCES

[1] H. Meng, et.al, "The Multi-biometric, Multi-device and Multilingual (M3) Corpus," *Proc. MMUA*, May, 2006.

[2] D. A. Reynolds, "Channel robust speaker verification via feature mapping," *Proc. ICASSP'03*, vol.2, pp. 57-61, Apr. 2003.

[3] R. Teunen, B. Shahshahani, and L. Heck, "A Model-based Transformational Approach to Robust Speaker Recognition," *Proc.ICSLP'00*, vol.2, pp.495-498, Oct. 2000.

[4] R. Auckenthaler, M. Carey, and H.loyd-Thomas, "Score Normalization for Text-Independent Speaker Verification systems," *Digital Signal Processing*, vol. 10, pp. 42-54, Jan. 2000.

[5] B. Ma and H. Meng, "English-Chinese Bilingual Text-Independent Speaker Verification," *Proc. ICASSP'04*, Oct. 2004.

[6] M. W. Mak, R. Hsiao, and B. Mak, "A Comparison of Various Adaptation Methods for Speaker Verification with Limited Enrollment Data", *Proc.ICASSP'06*, pp. 929-932, May, 2006.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, Jan. 2000.

[8] J. W. Koolwaaij and L. Boves, "A new procedure for classifying speakers in speaker verifcation systems," *Proc. Eurospeech'97*, pp. 2355-2358, Oct. 1997.