ENGLISH-CHINESE BILINGUAL TEXT-INDEPENDENT SPEAKER VERIFICATION

Bin Ma and Helen Meng
Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Hong Kong SAR, China
Email: bma@se.cuhk.edu.hk, hmmeng@se.cuhk.edu.hk

ABSTRACT

This paper describes the development of a text-independent speaker verification (TISV) system for English and Chinese utterances. We have designed and collected a bilingual database that contains spoken responses and commands in short, medium and long durations. The TISV system uses Gaussian mixtures for speaker models. Our experiments indicate that language mismatch between enrolment and verification data leads to significant degradation in verification performance (between 40% to 49%). In order to maximize robustness towards language change in test utterances, speaker models were trained with utterances from both languages. Results indicate that this can effectively close performance degradation gap due to language mismatch as mentioned above.

1. INTRODUCTION

Speaker verification is the process of authenticating the speaker's claimed identity based on his/her input utterances. The technology plays a key role in securing computing for human-centric computer interfaces. These interfaces embrace the user's natural communicative modalities (such as those in human-human communication e.g. speech, hearing, vision, gesture, etc.) at the core of human-computer interaction. Since the user's speech may be acquired easily over the course of a multimodal human-computer interaction, speaker verification offers a non-intrusive means of security with a high degree of usability. This work reports on our first attempt in developing a speaker verification system that forms part of a multimodal, interactive human-computer interface secured with multiple biometrics (augmented with face recognition and fingerprint verification). Hong Kong has a bilingual environment where English and Chinese are commonly used, with Cantonese being the predominant Chinese dialect. Consequently, this work develops a *bilingual* (English and Cantonese) speaker verification system. Our long term goal is to develop combined speaker verification and verbal information verification (VIV) [1] that underlies a human-computer spoken dialog and frequently verifies the speaker's identify to heighten the level of security. As such we strive to develop a *text-independent* system as well. In this context of bilingual, text-independent speaker verification, we investigate the effects of *language mismatch* (between the enrolment and verification speech data) on speaker verification performance.

Much previous work has been conducted in speaker verification [2-6]. Most of the work use databases that are monolingual. Auckenthaler et al. [2] showed that language mismatches between the target speaker and world model lead to major degradations in speaker verification performance, particularly for Mandarin and Vietnamese against target speakers who spoke American English. Qing & Chen [3] reported on a speaker verification system trained on English and Chinese digits/ sentences that show small performance discrepancies between testing on English versus testing on Chinese. In this work, we develop an English-Cantonese bilingual text-independent speaker verification system, and report on performance changes when the system is trained and tested on (i) English utterances only; (ii) Cantonese utterances only; and (iii) combined English and Cantonese utterances.

2. THE CUHK BILINGUAL SPEECH CORPUS

We have designed and collected the CUHK Bilingual Speech Corpus (BSC) to support experimentation and evaluation in this work. Prompts for data collection may ask about personalized information, e.g. "*What is your favorite color?*" or "*What is your favorite food?*" Alternatively, the prompts may ask the speaker to issue a command, e.g. "*Please speak a command to open the door.*" In order to incorporate variability in the recorded utterances (e.g. in lexical choices and lengths) to support text-independence, the speaker is asked to provide short, medium and long answers to each prompt. The speaker is also asked to provide semantically consistent answers in both English and Cantonese. Examples are shown in Tables 1 and 2.

Prompt	*What is your favorite color?* **你最喜欢什么颜色？**
Answers Short Medium Long	 Purple. **紫色。** It's purple. **我喜欢紫色。** My favorite color is purple. **我最喜欢的颜色是紫色。**

Table 1. An example prompt for personalized information and related short/medium/long answers from the CUHK Bilingual Speech Corpus.

Prompt	*Command: Open the door.* **开门。**
Answers Short Medium Long	 Open. **开。** Open the door. **开门。** Please open the door for me. **给我开门。**

Table 2. An example prompt for spoken commands and related short/medium/long answers from the CUHK Bilingual Speech Corpus.

The enrolment and verification data are recorded from 16 speakers (10 males and 6 females) from the university student body. Hence the speakers have similar ages and educational backgrounds. Each speaker participated in three enrolment sessions and one verification session, spaced out with one-week intervals. Compositions of the enrolment/verification data sets are summarized in Table 3.

	Enrolment Set	Verification Set
Prompts for personalized info	10	6
Prompts for spoken commands	18	7
# versions in responses (short, medium long)	3	3
# sessions	3	1
Languages	E, C	E, C
Total # utterances per speaker	504	78

Table 1: Compositions of the enrolment and verification data sets. 'E' refers to English and 'C' to Cantonese.

The speech data were recorded with a SHURE BG1.1 microphone in an office environment. We did not deliberately avoid noises from sources such as computer fans and air conditioning. Some recorded utterances also contain background babbling from other talkers.

3. FRONT-END PROCESSING

The microphone speech in the CUHK BSC is sampled at 16 kHz. The digitized data is then pre-emphasized by computing first-order differences. We compute 14th order LPC coefficients for every 10ms over 25.6ms Hamming windows. The first 12th order LPC cepstral coefficients are converted from these LPC feature coefficients. Combined with the signal's log-energy, there are 13 acoustic feature coefficients. Augmenting this vector with the delta and delta-delta derivative vector gives 39 coefficients in total.

4. THE SPEAKER VERIFICATION SYSTEM

4.1 Gaussian mixture models (GMM)

GMM has been shown to be an effective statistical approach for text-independent speaker verification tasks [2-4]. A speaker's characteristics are modeled by a weighted sum of M component Gaussian densities (see Equation 1) [4]:

$$p(x \mid l) = \sum_{i=1}^{M} w_i b_i(x) \quad (1)$$

where x is a d-dimensional random vector; $b_i(x)$ with $i=1,2..M$ are the component densities (see Equation 2) and w_i with $i=1,2,...M$ are the mixture weights that satisfy the constraint in Equation 3.

$$b_i(x) = \frac{1}{(2p)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-m_i)^t \sum_i^{-1} (x-m_i)\} \quad (2)$$

with mean vector m_i and covariance matrix Σ_i.

$$\sum_{i=1}^{M} w_i = 1 \quad (3)$$

The complete Gaussian mixture density is parameterized by the mean vectors, (diagonal) covariance matrices and mixture weights from all component densities. These parameters collectively represent a speaker's model denoted by:

$$l = \{w_i, m_i, \Sigma_i\} \quad i = 1, \cdots, M \quad (4)$$

These GMM parameters are estimated during training by using the Expectation Maximization (EM) algorithm over five iterations.

During testing, the speaker verification system derives T test vectors $X=x_1, x_2, \ldots x_T$ from the claimant's input speech and computes the GMM log likelihood as shown in Equation 5.

$$P(X \mid l) = \sum_{i=1}^{T} \log p(x_i \mid l) \quad (5)$$

4.2 Cohort normalization

The cohort normalization technique has been shown to be effective in improving speaker verification performance [5]. The general approach is to apply a likelihood ratio test to

an input test utterance X using the claimed speaker model l_c :

$$l(X) = \frac{P(l_c \mid X)}{P(l_{\bar{c}} \mid X)} \quad (6)$$

Applying Bayesian rule and assuming equal prior probabilities, the likelihood ratio in the log domain becomes:

$$\Lambda(X) = \log P(X \mid l_c) - \log P(X \mid l_{\bar{c}}) \quad (7)$$

where $l_{\bar{c}}$ is a model representing all other possible speakers. The likelihood $P(X \mid l_c)$ is directly computed from Equation (5). The likelihood $P(X \mid l_{\bar{c}})$ is usually approximated using a collection of *background* speaker models. The background speaker set consists of K speakers (also known as the *cohort speakers*) who are acoustically closest to the claimant. Cohort normalization involves computing the log-likelihood (see Equation 9) and performing normalization (see Equation 10).

$$\log P(X \mid l_{\bar{c}}) = \frac{1}{K} \sum_{k=1}^{K} \log\{P(X \mid l_k)\} \quad (9)$$

$$P_{norm}(X \mid l_i) = \frac{P(X \mid l_i)}{\frac{1}{K} \sum_{k=1}^{K} P(X \mid l_k)} \quad (10)$$

The process of selecting cohort speakers from the CUHK BSC is as follows – speaker ***A*** is first randomly chosen as the true speaker (or claimant). Then eight speakers are randomly selected to form the *cohort speaker set* for ***A*** and the remaining 7 speakers are regarded as impostors. We then identify the ***K*** (=4) *cohort speakers* from the cohort set that are closest to ***A***. This process is repeated for all the 16 speakers in BSC to compute the overall speaker verification performance. This cohort selection process aims to strike a balance between the selection of cohort versus imposter speakers in order not to inflate the verification performance to real-world unseen levels [6].

4.3 Verification and Evaluation

Applying the cohort normalization, the verification score from Equation (10) can be compared with a threshold q to make the verification decision, as shown in Equation 11.

$$P_{norm}(X \mid l) \begin{cases} > q & \text{Accept} \\ < q & \text{Reject} \end{cases} \quad (11)$$

Errors may include false rejection (FR), where a true speaker is rejected against his own claim; and false acceptance (FA), where an impostor is accepted as the falsely claimed speaker. The standard equal error rate (EER) is an evaluation criterion that combines both by reporting on the levels where FA=FR.

5. EXPERIMENTS

5.1 Baseline Experiment

As a point of reference we begin by testing our GMM-based speaker verification (SV) system on the male speakers of the YOHO corpus [6]. The system gave an EER of 0.08%. This compares with previous experiments [4] that reported EER=0.20% on the same data. While the performance discrepancy may be due to differences in front-end processing, the use of a variance-flooring technique and the use of a speaker-dependent threshold (in our case) versus a global threshold (in [4]); this baseline result seems to indicate that we have developed a GMM SV system that gives reasonable performance. Hence we proceed to experimentation with the CUHK BSC corpus.

5.2 Language Mismatch between Enrolment and Verification

We obtain English speaker models by training only on English enrolment data (252 utterances); and Cantonese speaker models by training only on Cantonese enrolment data (also 252 utterances). The number of mixtures (M) used in the GMM are derived by a K-means algorithm and the value of $M=256$ is empirically chosen.

We obtained SV results by applying the trained English speaker models on the English verification subset (EER=3.98%) and the Cantonese verification subset (EER=5.92%). Comparison shows that language mismatch causes a performance degradation of 49%.

Similarly, we applied the trained Cantonese speaker models on the Cantonese verification subset (EER=4.28%) and the English verification subset (EER=5.98%). In this case, language mismatch causes a performance degradation of 40% (see Figure 1). This suggests that the GMM captures not only speaker characteristics, but are biased by the linguistic characteristics of the enrolment data as well.

We have also tried to build a *pooled* model by training also on 252 utterances, but half of these are randomly selected from the English enrolment subset and the remaining from the Cantonese enrolment subset. SV performance of the *pooled* model on the English and

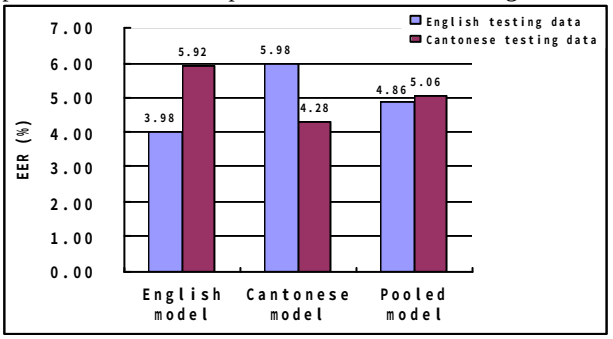

Figure 1: Effect of language mismatch between enrolment and verification data on speaker verification performance.

Cantonese verification subsets are EER=4.86% and 5.06% respectively (see Figure 1). Hence this model is more robust and less sensitive to language changes in the verification data.

In order to maximize the robustness of the trained speaker models for verification of text-independent testing utterances in different languages, we pooled data across the two languages for training speaker models. We have also pooled the testing data across languages. Results are

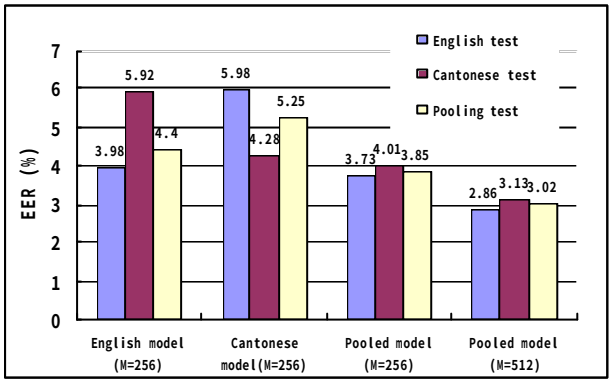

Figure 2: Text-independent speaker verification results based on bilingual training and testing sets.

shown in Figure 2. We used *(M=)* 256 Gaussian mixtures to maintain consistency for comparison with previous results in Figure 1. We also reset (optimized) the number of Gaussian mixtures (*M=512*) empirically since pooling brings increased training data per model. The latter gave the best verification performance (rightmost group of bar-charts in Figure 2) with overall EER=3.02%. This is also the case where performance variations in verification due to different languages in the test set are minimized.

6. CONCLUSIONS

This paper reports on the design and development of a text-independent speaker verification (TISV) system for English and Chinese utterances. We have also designed and collected the CUHK Bilingual Speech Corpus (BSC) to support experimentation and evaluation. The BSC contains spoken responses and commands in short, medium and long durations. The TISV system uses Gaussian mixtures for speaker models. Our experiments indicate that language mismatch between enrolment and verification data leads to significant verification performance degradation (between 40% to 49%). In order to maximize robustness towards language change in test utterances, speaker models were trained with utterances from both languages. Results indicate that this can effectively close performance degradation gap due to language mismatch as mentioned above. The best performance achieved is around EER=3.02% when we pooled both the English and Chinese enrolment utterances of the speaker for training and tested on the verification utterances pooled from both languages. In the future, we will extend this work to cover Mandarin Chinese. We will also explore methods in modelling different languages in the context of multilingual text-independent speaker verification.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] Q. Li, B. H. Juang, Q. Zhou, and C. H. Lee, " Automatic Verbal Information Verification for User Authentication", *IEEE Trans. on Speech and Audio Processing*, pp. 1-10, Sep 2000.

[2] R. Auckenthaler, M. J. Carey, and J. S. D. Mason, " Language Dependency In Text-Independent Speaker Verification", *ICASSP 2001*, May 2001.

[3] X. Qing and K. Chen. " On use of GMM for multilingual speaker verification: An empirical study". *Proceedings of ISCSLP*, pages 263–266, 2000.

[4] D. A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models", *Speech Communication*, Vol. 17, pp. 91-108, 1995

[5] C. S. Liu, H. C. Wang, and C. H. Lee. " Speaker verification using normalized log-likelihood score" , *IEEE Transactions on Speech and Audio Processing*, pp. 56–60, Jan 1996.

[6] J. P. Campbell. "Testing with the YOHO CD-ROM voice verification corpus". *ICASSP-95*, pages 341–344, 1995.