BioSec Multimodal Biometric Database in Text-Dependent Speaker Recognition

Doroteo T. Toledano, D. Hernandez-Lopez, C. Esteve-Elizalde, J. Fierrez, J. Ortega-Garcia, D. Ramos and J. Gonzalez-Rodriguez

ATVS Biometric Recognition Group, Universidad Autónoma de Madrid.
Escuela Politécnica Superior, C/ Francisco Tomás y Valiente, 11, 28049 Madrid, SPAIN.
doroteo.torre@uam.es, d.hernandezlopez@uam.es, cristina.esteve@uam.es, julian.fierrez@uam.es, javier.ortega@uam.es, daniel.ramos@uam.es, joaquin.gonzalez@uam.es

Abstract

In this paper we briefly describe the BioSec multimodal biometric database and analyze its use in automatic text-dependent speaker recognition research. The paper is structured into four parts: a short introduction to the problem of text-dependent speaker recognition; a brief review of other existing databases, including monomodal text-dependent speaker recognition databases and multimodal biometric recognition databases; a description of the BioSec database; and, finally, an experimental section in which speaker recognition results on BioSec and other database widely used in speaker recognition are presented and compared, using the same underlying speaker recognition technique in all cases.

1. Introduction to text-dependent speaker recognition

Automatic speaker recognition tries to recognize the speaker that produces a particular speech utterance. Depending on the constraints imposed on the linguistic content of the utterance there are two types of speaker recognition: text-independent speaker recognition in which the linguistic content of the speech recording is unknown by the system and text-dependent speaker recognition where the linguistic content of the speech is known.

In recent years the National Institute of Standards and Technology (NIST) has promoted research in the context of text-independent speaker recognition with the organization of yearly international competitive evaluations (NIST, 2008; Przybocki, Martin & Le, 2006) which have fostered the definition of challenging tasks through a strong effort in the development of publicly available speech databases. Despite its potential applications in interactive voice response systems, the absence of similar competitive evaluations has kept text-dependent speaker recognition at a slower pace of development and the number and extent of the databases for research in this field is more limited. For that reason BioSec is an important contribution in this area.

In the field of text-dependent speaker recognition there are two methods that have been used for years: Dynamic Time Warping (DTW) and Hidden Markov Models (HMMs). DTW is simpler, but less flexible (Ramasubramanian, Das & Kumar, 2006). HMMs on the other hand are more complex, provide more flexibility and at least comparable results, and are the most commonly used technique in text-dependent speaker recognition (Hébert, 2008; Matsui & Furui, 1993; Che, Lin & Yuk, 1996; Bimbot et al., 1997).

Most of the works previously reported for text-dependent speaker recognition using HMMs tend to use a speaker independent set of HMMs and retrain the parameters of these HMMs using Baum-Welch reestimation to produce a speaker-dependent set of HMMs. After these models have been trained, an utterance is verified by performing speech recognition with the speaker independent and the speaker-dependent HMMs and comparing the acoustic scores obtained. Recently other works in the literature (Subramanya et al., 2007; Toledano et al., 2008) have started to modify this method by substituting Baum-Welch retraining by Maximum Likelihood Linear Regression (MLLR) adaptation (Leggetter & Woodland, 1995) of the speaker independent HMMs. This allows to use more complex (and, if properly trained, more reliable) HMMs while keeping the speaker models small (since only the MLLR transformation matrices need to be stored). This is the basic methodology that we have used for the comparison of text-dependent recognition results in this paper. We have avoided using here recent improvements in text-dependent speaker recognition, such as the use of discriminative methods after the MLLR adaptation (Subramanya et al., 2007) or phoneme or state-based T-Normalization (Toledano et al., 2008) because our main interest in this paper is the comparison of different databases for speaker recognition research. Therefore, we preferred to keep our speaker recognition system simple, yet still in line with the current state of the art in speaker recognition research.

2. Other databases for text-dependent speaker recognition

In this section we present several other databases for text-dependent speaker recognition research grouped into two broad categories: unimodal and multimodal databases.

2.1 Other unimodal databases for text-dependent speaker recognition

For years YOHO (Campbell & Higgins, 1994; Campbell, 1995) has been the best known database for evaluation of text-dependent speaker recognition. It consists of 96 utterances for enrolment collected in 4 different sessions

and 40 utterances for test (10 sessions) for each of 138 speakers. Each utterance consists of different combinations of three pairs of digits (e.g. "12-34-56") in English. However, YOHO has several limitations that more modern corpora try to address. For instance, the MIT Mobile Device Speaker Verification Corpus (Woo, Park and Hazen, 2006) has been specifically designed for research on text-dependent speaker verification on realistic noisy conditions.

2.2 Other multimodal databases for biometric recognition

Due to the increasing interest in multimodal biometric recognition (of which text-dependent speaker recognition is just a particular modality), and given that one of the main difficulties in capturing a biometric database is recruiting donors, many of the newly developed biometric databases are multimodal and cover several biometric traits. Some of these databases include speech as a particular modality and can potentially be used for text-dependent speaker recognition research.

Some of the most veteran and widely used biometric databases are XM2VTS (Messer et al., 1999) containing microphone speech and face images of 295 people captured in 4 different sessions, and MCYT (Ortega-Garcia et al., 2003) database including fingerprints and signature of 330 subjects. More recent databases include BIOMET (Garcia-Salicetti et al., 2003), BANCA (Bailly-Bailliere et al., 2003), MYIDEA (Dumas et al., 2005), MBioID (Dessimoz et al., 2007), and M3 (Meng et al., 2006). Other current initiatives in multimodal database collection closely related to the BioSec database are the following (Faundez-Zanuy et al. 2006; Flynn, 2007):

- BiosecurID. This database includes 7 unimodal biometric traits, namely: speech, iris, face, handwriting, fingerprints, hand and keystroking. The database comprises 400 subjects and was acquired in a realistic office-like scenario.

- BioSecure (BioSecure, 2007). This database considers three acquisition scenarios, namely: unsupervised Internet acquisition, including voice, and face; supervised office-like scenario, including voice, finger prints, face, iris, signature and hand; and acquisition in a mobile device, including signature, fingerprints, voice, and face. The database comprises over 1000 subjects for the Internet scenario, and about 700 users the other two.

3. The BioSec database

The BioSec database was acquired under FP6 EU BioSec Integrated Project (Fierrez-Aguilar et al., 2007), and comprises fingerprint images acquired with three different sensors, frontal face images from a webcam, iris images, and voice utterances of 250 subjects.
The speech part of the corpus (the most interesting part for this paper) was recorded at 44 KHz stereo with 16 bits (PCM with no compression) using both a headset and a distant webcam microphone. Each subject utters 4 repetitions of a user-specific keyword consisting of 8 digits both in English and Spanish. Speakers are mainly native Spanish speakers. In addition, every subject says 3 keywords corresponding to other users to simulate informed forgeries in which an impostor has access to the number of a client. The 8 digits were always pronounced digit-by-digit in a single continuous and fluent utterance.

In addition to the increased number of subjects and a more balanced distribution of donors, the BioSec database has several advantages with respect to other well known databases such as YOHO. For instance it allows the simulation of informed forgeries. The BioSec database also allows studies based on age and the combination of BioSec, BiosecurID and BioSecure allows long term (2 year) temporal variability studies, because they have some subjects in common.

4. Experimental results

Text-Dependent speaker recognition experiments have been performed on YOHO and BioSec Baseline using exactly the same techniques to compare the two databases for experimentation in Text-Dependent speaker recognition. One particularity of these experiments is that in all trials the text spoken coincides with the text expected by the system. In this sense, the experiments are more representative of text-prompted systems in which the system asks the user to utter a specific phrase. In all cases the technique used for speaker recognition has been the following: we start with a set of speaker-independent phonetic HMMs that were trained on TIMIT (for English) or ALBAYZIN (for Spanish). Using the enrolment data we adapt (with MLLR) these models to produce speaker-adapted HMMs. We have also tried reestimation

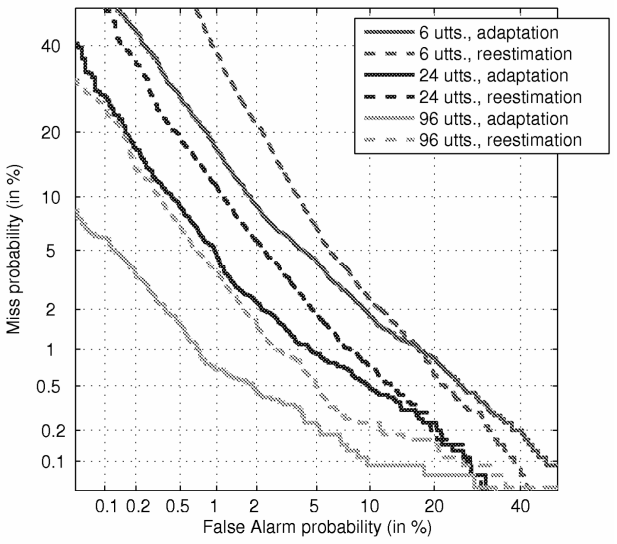

Figure 1: Results (DET curves) obtained on YOHO using MLLR adaptation and Baum-Welch re-estimation using as enrolment material 6, 24 or 96 utterances.

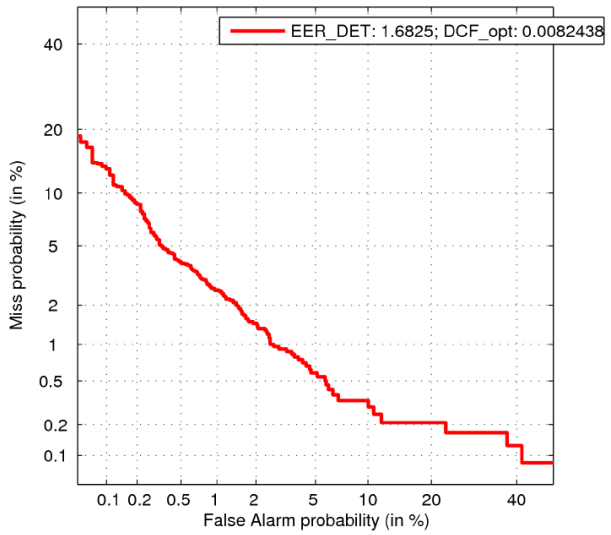

Figure 2: Results (DET curves) obtained on BioSec using MLLR adaptation for Spanish and close-talking microphone.

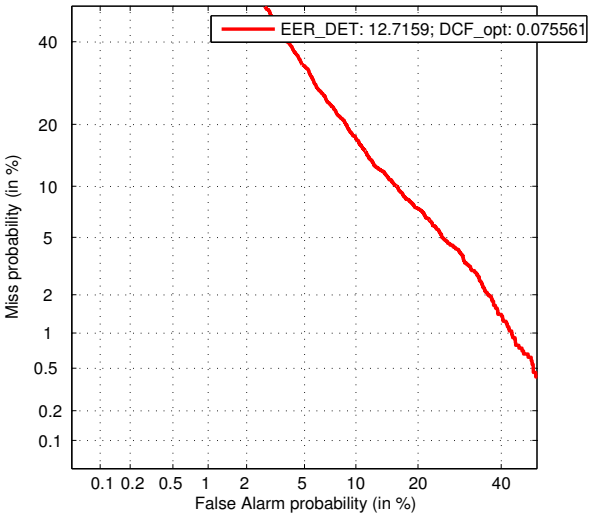

Figure 3: Results (DET curves) obtained on BioSec using MLLR adaptation for English and webcam distant microphone.

with Baum-Welch instead of MLLR adaptation in YOHO, but results were worse, as can be seen in Fig. 1. In the speaker-verification phase we subtract the (log) acoustic scores obtained by the speaker-adapted and the speaker-independent HMMs to obtain a verification score that is more positive to indicate a closer match.

4.1 Results with YOHO

The results presented on YOHO are based on the following experimental protocol: three sets of speaker models are trained using 6 utterances from session 1, the 24 utterances from session 1 or the 96 utterances from the 4 sessions. Speaker verification is performed using a single utterance from the test subset. The target scores are generated by matching each speaker-dependent phone HMM with all the test utterances from that user, leading to a total of 138 x 40 = 5520 scores. The impostor scores are computed by comparing each speaker model with a single utterance randomly selected from those of all other users, which yields 138 x 137 = 18906 trials. For all impostor trials speech is aligned against the actual phonetic content spoken to simulate a text-prompted system in which the impostors know what they have to say. Results obtained with this experimental protocol are presented in Figure 1.

As commented earlier, Fig. 1 shows that for all conditions tested MLLR adaptation in superior to Baum-Welch reestimation. The other important observation is the influence of the amount of enrolment material in performance. It can be seen that using the 96 available utterances for enrolment gives an EER under 1%, while for 6 utterances (which would be much more user-friendly) the EER increases to close to 5%. This result, however, can be lowered to about 3% using score normalization techniques not used in this paper (Toledano et al., 2008).

4.2 Results with BioSec

For these experiments we have considered two subsets of BioSec Baseline (a subset of the BioSec database comprising 2 acquisition sessions from 200 subjects), employing only those utterances that were spoken in Spanish and were captured by the headset microphone and the sentences spoken in English and captured by the webcam microphone. The experimental protocol we have followed is based on the BioSec Baseline core protocol over the specified 150 test subjects so that our results can be easily compared to other results on this corpus (Fierrez-Aguilar et al., 2005). The genuine matchings in this database were performed comparing each of the 4 samples in the first session with the 4 samples from the same user in the second session. This makes a total of 150 x 4 x 4 = 2400 target scores. To generate the impostor matchings, the first sample from the first session was tested against the same sample from the rest of the users, without performing symmetric matches. This leads to a total number of 150 x 149 / 2 = 11175 impostor scores. Results obtained with this experimental protocol are presented in Figures 2-3.

The first surprising fact is that EER in Fig. 2, where we use a single utterance for enrolment, is below 2% while for YOHO using 6 utterances for enrolment the EER is close to 5%. The reason for this surprising performance is the lexical content of the enrolment and test materials: in BioSec the lexical content of the enrolment and target trials is the same (a fixed password assigned to each user), while in YOHO the lexical content differs. Other interesting observation is the huge difference between the curves in Figures 2 and 3. There are two possible causes (which we are currently investigating) for this difference: the channel mismatch (close talking vs. distant webcam microphone) and the non-nativeness of most subjects in English in BioSec.

5. Conclusions

It is usual in research articles to use a database as test bed and compare different algorithms on that database. In text-dependent speaker recognition it has been mainly YOHO the corpus that has served for this purpose. However, YOHO has several limitations that more modern databases overcome. In this sense, researchers willing to use more modern and ample databases can be retracted from using them in order to be able to compare their results to those of other researchers. In this context, it is necessary to have a way of comparing results across different databases. This paper is an attempt to facilitate the use of the BioSec corpora by providing a comparison of text-dependent speaker recognition results across YOHO and BioSec, using exactly the same algorithms and analyizing some of the differences observed in performance on the two databases.

6. Acknowledgements

This work was funded by the Spanish Ministry of Science and Technology under project TEC2006-13170-C02-01.

7. References

Bailly-Bailliere, E., Bengio, S., et al. (2003). The BANCA database and evaluation protocol. In: Proc. of IAPR AVBPA, Springer LNCS-2688, 625-638.

Bimbot F., Hutter H. P., et al. (1997). "Speaker verification in the telephone network: research activities in the CAVE project", in Proc. Eurospeech 1997, pp. 971-974.

BioSecure (2007). Biometrics for Secure Authentication, FP6 Network of Excellence (NoE), IST-2002-507634. (http://www.biosecure.info/).

Campbell J. and Higgins A. (1994). YOHO speaker verification corpus LDC94s16). Available at the LDC website: http://www.ldc.upenn.edu.

Campbell J. P. (1995). "Testing with the YOHO CD-ROM voice verification corpus", in Proc. ICASSP 1995, vol. 1, pp. 341-344.

Che C.-W., Lin Q. and Yuk D.-S. (1996). "An HMM approach to text-prompted speaker verification", in Proc. ICASSP 1996, vol. 2, pp. 673-676.

Dessimoz, D., Richiardi, J., et al. (2007). Multimodal biometrics for identity documents (MBioID). Forensic Science International 167, 154-159.

Dumas, B., Hennebert, J., et al. (2005). MyIdea - Sensors speci¯cations and acquisition protocol. Computer Science Department Research Report DIUF-RR 2005.01, University de Fribourg in Switzerland.

Faundez-Zanuy M., Fierrez-Aguilar J., Ortega-Garcia J. and Gonzalez-Rodriguez J. (2006). "Multimodal biometric databases: An overview", IEEE Aerospace and Electronic Systems Magazine, Vol. 21, n. 8, pp. 29-37, August 2006.

Fierrez-Aguilar J. and Ortega-Garcia J. (2005). "Extended Multimodal Database and Testing Protocol", Deliverable D5.7, BioSec, FP6 IP IST-2002-001766, December 2005.

Fierrez, J., Ortega-Garcia, J., et al. (2007). Biosec baseline corpus: a multimodal biometric database. Pattern Recognition 40, 1389-1392.

Flynn P. J. (2007). "Biometric databases", chapter in A. K. Jain, P. Flynn, A. A. Ross (Eds.), Handbook of Biometrics, Springer, 2007.

Garcia-Salicetti, S., Beumier, C., et al. (2003). BIOMET: A multimodal person authentication database including face, voice, fingerprint, hand and signature modalities. In: Proc. of IAPR AVBPA, Springer LNCS-2688 845-853.

Hébert, M. (2008), "Text-Dependent Speaker Recognition", chapter 37 in Benesty, Sondhi and Huang (Eds.) "Handbook of Speech Processing", Springer.

Leggetter C. J. and Woodland P. C. (1995). "Flexible speaker adaptation using maximum likelihood linear regression", in Proc. Eurospeech 1995, pp. 1155-1158.

Matsui T. and Furui S. (1993). "Concatenated phoneme models for text-variable speaker recognition", in Proc. ICASSP 1993, vol. 2, pp. 391-394.

Meng, H., Ching, P.C., et al. (2006). The multi-biometric, multi-device and multilingual (M3) corpus. In: Proc. MMUA Workshop.

Messer, K., Matas, J., et al. (1999). XM2VTSDB: The extended M2VTS database. In: Proc. of IAPR AVBPA.

NIST (2008). National Institute of Standards and Technology. Speaker Recognition Evaluation Home Page, http://www.nist.gov/speech/tests/spk/index.htm, (accessed Feb. 2008).

Ortega-Garcia, J., Fierrez-Aguilar, J., et al. (2003): MCYT baseline corpus: a bimodal biometric database. IEE Proc. VISP 150, 391-401.

Przybocki M. A., Martin A. F., and Le A. N. (2006). "NIST speaker recognition evaluation chronicles part 2", in Proc. IEEE Odyssey 2006: The speaker and language recognition workshop.

Ramasubramanian V., Das A. and Kumar V. P. (2006). "Text-dependent speaker recognition using one-pass dynamic programming algorithm", in Proc. ICASSP 2006, vol. 1, pp. 901-904.

Subramanya, A.; Zhengyou Zhang; Surendran, A.C.; Nguyen, P.; Narasimhan, M.; Acero, A. (2007). "A Generative-Discriminative Framework using Ensemble Methods for Text-Dependent Speaker Verification" in IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. Volume 4, 15-20 April 2007 Page(s): IV-225 - IV-228.

Toledano D. T., Esteve-Elizande C., Gonzalez-Rodriguez J., Fernandez-Pozo R. and Hernandez-Gomez L. (2008). "Phoneme and Sub-Phoneme T-Normalization for Text-Dependent Speaker Recognition", in Proc. IEEE Speaker and Language Recognition Workshop (Odyssey) 2008.

Woo R. H., Park A. and Hazen T. J. (2006). "The MIT mobile device speaker verification corpus: data collection and preliminary experiments", in Proc. IEEE Odyssey 2006: The speaker and language recognition workshop.