Unsupervised model adaptation for speaker verification

Alexandre Preti, Jean-François Bonastre

LIA, Université d'Avignon
Agroparc, BP 1228
84911 Avignon CEDEX 9, France
{alexandre.preti,jean-francois.bonastre}@univ-avignon.fr

Abstract

This paper deals with unsupervised model adaptation for speaker recognition. Two adaptation schemes are proposed, the first one is based on a test by test model adaptation and the second one proposes a batch mode, where the adaptation is performed using a set of tests before computing the decision score for each of them. The experiments are conducted thanks to the NIST SRE 2005 database. This paper shows clearly the interest of unsupervised model adaptation when enough test data is available (batch mode) and the intrinsic difficulty of an online (test by test) adaptation mode.

Index Terms: Speaker verification, Unsupervised adaptation.

1. Introduction

Gaussian Mixture Model (GMM) systems for speaker recognition have shown robust results for several years and are widely used in speaker recognition applications [4, 5]. The main drawback of a speaker recognition system remains the little amount of data available for enrolling a speaker model, especially when only a one session record is available per speaker. An interesting way of improving performance of such systems is to increase the amount of training data by taking into account information coming from the use of the system, the test data [1, 2, 3]. This task relies on a client model adaptation process, which could be supervised (the system knows if the test record belongs to the target speaker) or unsupervised (the system has no information on the test data). Unsupervised adaptation is more adapted to real working conditions because it does not involve a human decision. But the difficulty of unsupervised speaker adaptation is to decide whether a test segment should be used or not to adapt the involved target model. In this paper we propose two different experimental protocols, the batch and the online modes. These are further described in section 3.3. The Generalized Likelihood Ratio (GLR) is used to select the trials. To assess the use of the GLR, results obtained using a classical LLR are also provided.

Section 2 describes the unsupervised adaptation principle. Section 3 introduces database, tools and protocols used to set up experiments. Experimental results are presented in section 4. Finally, conclusions are given in section 5.

2. Unsupervised adaptation

2.1. Proposed issue

In this work we propose to use information gathered during the system real life (trials) in order to adapt the speaker models. If the test belongs to the target speaker we use it to improve the speaker model . We proposed two experimental protocols. First we use all the selected trial segments involving a target speaker to adapt its model before taking the trial by trial decisions. This protocol will be further named the batch protocol. Then we respect the NIST SRE unsupervised adaptation mode [4], i.e., the client model can be updated using only previous trial segments involved before taking the decision for the current test. This protocol will be further named the online protocol.

In this paper the Generalized Likelihood Ratio (GLR) is used in order to select the trials. If the GLR ratio between the target model and the test model is under a decision threshold, the trial is kept for adapting the target model (see section 3.3).

2.2. Generalized Likelihood Ratio

The GLR is used here to take the decision to adapt or not the client model with the test data. The GLR test is well-suited to decide if two data sets belong to the same speaker or not [6] as it doesn't use any prior information. The GLR is defined as follows :

$$d_{GLR}(u_0, u_1) = \frac{L(u_0|M(U_0))L(u_1|M(U_1))}{L((u_{01})|M(U_{01}))} \tag{1}$$

where u_0 and u_1 are the data set for both the data set 0 and the data set 1 (respectively), $M(U_0)$ and $M(U_1)$ the models estimated via EM ML criterion on the data u_0 and u_1 (respectively) and $M(U_{01})$ the model estimated via EM ML criterion on both the data sets.

In order to assess the choice of the GLR method we compare the GLR-based system with a classical LLR-based system.

The LLR is defined as follows :

$$\Lambda(X|s) = \frac{1}{T}\sum_{t=1}^{T} logp(x_t|\lambda_s) - logp(x_t|\lambda) \tag{2}$$

where $X = \sum_{t=1}^{T} x_t$ denotes the feature vectors, λ_s is the model of the speaker s, and λ is the world model.

In both cases (GLR or LLR-based systems), the decision to use a test trials is taken using only the initial speaker training data set and the current trial. GLR is computed using the initial speaker training data set and the current trial data to keep comparable amount of data (all tests previously selected are not used to take this decision).

A threshold has been determined empirically to take the decision of acceptation.

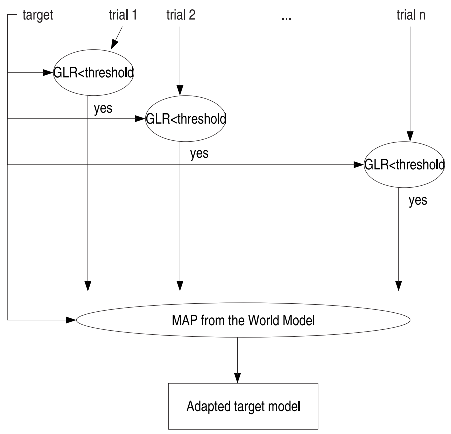

Figure 1: Batch protocol description

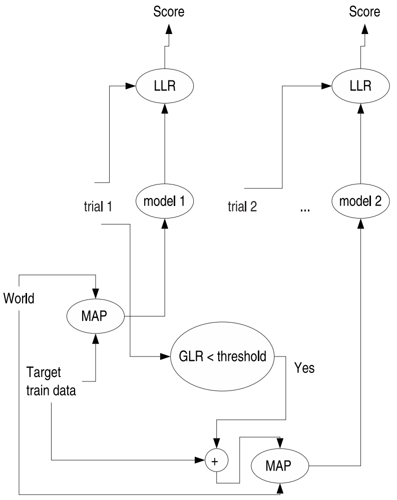

Figure 2: Online protocol description

3. Tools and Protocol

3.1. Database

All the experiments presented in section 4 are performed based upon the NIST 2005 database, all trials (det 1), 1conv-4w 1conv-4w, restricted to male speakers only.

This condition consists of 274 speakers. Train and test utterances contain 2.5 minutes of speech on average (telephone conversation). The whole speaker detection experiment consists in 13624 tests (1231 target tests). These are used for the adaptation set. From 1 to 170 tests are computed by speaker, with an average of 51 tests.

3.2. Baseline speaker recognition system

The LIA_SpkDet system [7] developed at the LIA lab is used as baseline in this paper. Built from the ALIZE platform [8, 9], it was evaluated during the NIST SRE'04 and SRE'05 campaigns, where it obtained good performance for a cepstral GMM-UBM system. Both the LIA_SpkDet system and the ALIZE platform are distributed under an open source licence.

The LIA_SpkDet system is based on classical UBM-GMM and T-Norm approach for likelihood score normalization. The background model used for the experiments is the same as the background model used by the LIA for the NIST SRE 2005 campaign (male only). The training is performed based on NIST SRE 1999 and 2002 databases, and consists in about 1 million of speech frames. For the front-end processing, the signal is characterized by 32 coefficients including 16 linear frequency cepstral coefficients (LFCC) (Filter-bank analysis) and their first derivative coefficients extracted with SPRO [10]. A frame removal based on a three component GMM energy modeling is computed. A mean and variance normalization process is finally applied on coefficients. The world and target models contain 128 Gaussian components. LLR scores are computed using the top ten components. In this paper, for computational time reasons, 128 component models are used (when the best system uses 2048 component models).

3.3. Protocol description

3.3.1. *Batch protocol*

This experimental protocol allows to adapt a target model with all the selected trials involved with it. The new target model is created by using Bayesian adaptation (MAP) from the UBM with a regulation factor of 14. Frames used for MAP adaptation are issued from the train data plus the trials selected according to the GLR criterion (see 2.2). Next, the LLR scores for each test are computed using the adapted speaker model. The performance is evaluated through classical DET performance curves. This protocol is shown in figure 1.

3.3.2. *Online protocol*

The NIST unsupervised adaptation mode allows to update target models using previous seen trial segments (including the current segment) to take the decision on the current trial segment. It is required to follow the order of the trials in the test protocol.

For each test, the adapted model is computed first, then the LLR is calculated to obtain a score. This is decribed in figure 2.

4. Results

4.1. Comparison LLR/GLR

A classical 128 GD experiment without unsupervised adaptation is set in order to compare LLR with GLR. Log Likelihood Ratio (LLR) gives better classification results than GLR (see results in Table 1). GLR introduces respectively 31% and 51% of relative increase for the DCF and the EER.

	DCF	EER
LLR	4.92	9.67
GLR	7.20	19.82

Table 1: *Comparison LLR/GLR on a 128 GD experiment.*

4.2. Batch protocol

Figure 3 shows DET curves of the baseline and the speaker adaptation method whithout score normalization. Without score normalization the speaker adaptation technique provides respectively a 17.9% and 0.7% of relative reduction in DCF and EER (Cf. table 2). The amount of train data for each model is increased by about 340%. It means that 2.4 trials of the 51 (on average) are used to create the target model. Indeed, the target model is now learnt on 8.5 min (on average) instead of 2.5 min. A total of 667 trials on the 13624 are used for speaker adaptation.

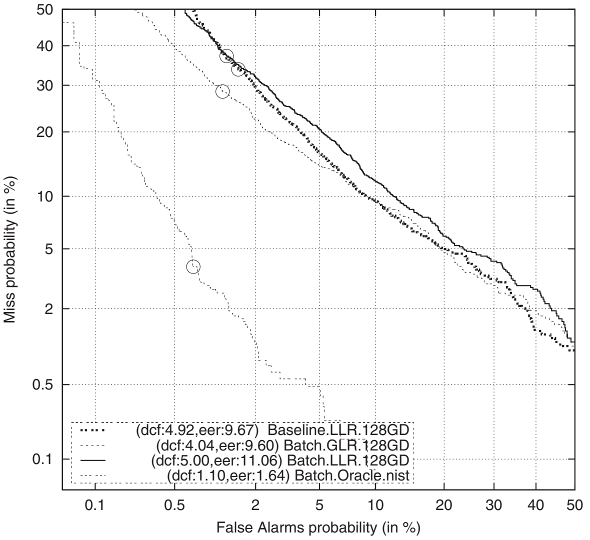

Figure 3: DET curves for batch protocol : GLR, LLR based speaker adaptation, Oracle and baseline (no adaptation performed) (no score normalization)

	DCF	EER
Baseline	4.92	9.67
Batch mode, GLR based speaker Adaptation	4.04	9.60

Table 2: *EER and DCF for the baseline system and the batch adaptation mode without score normalization*

In order to assess the choice of the GLR to select the trials, we have compared the GLR based system with a LLR based system. We have set a threshold to obtain the same amount of trials selected for both the systems. Table 3 presents the results. We can see that the GLR based trials selection outperforms the LLR based trials selection. A loss of relative reduction of respectively 19% and 13% of DCF and EER is observed by using the LLR based trials selection. [1]

Trial selection method	DCF	EER
GLR based	4.04	9.60
LLR based	5	11.06

Table 3: *EER and DCF of (1) the batch speaker adaptation protocol GLR based (2) LLR based (no score normalization)*

4.3. Online protocol

A total of 611 trials on the 13624 are used for speaker adaptation. Up to 2.22 trials are used to adapt the involved speaker model (on average). Train data are now from 2.5 to 8 min instead of 2.5 min (on average). Results are quite close from the baseline with a relative reduction of respectively 3.2% and 2.5% of DCF and EER. As previously, GLR based selection criterion is compared with a LLR based selection criterion. Table 4 presents the results.

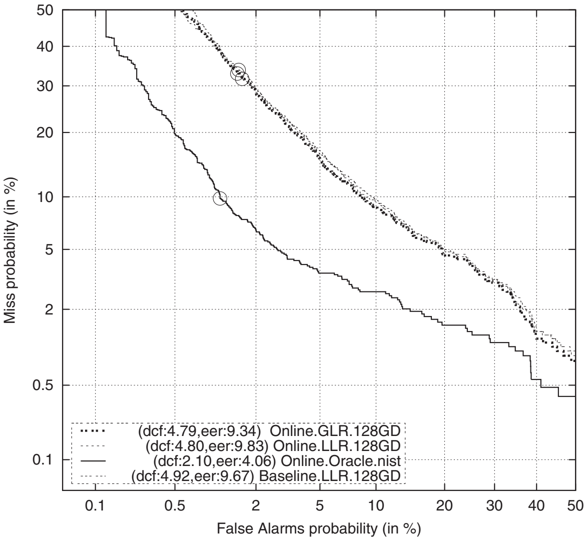

Figure 4: DET curves for the Online protocol GLR based, LLR based, Oracle and baseline (no adaptation performed) (no score normalization)

Trial selection method	DCF	EER
GLR based	4.79	9.34
LLR based	4.80	9.83

Table 4: *EER and DCF of (1) the online adaptation system GLR based (2) LLR based (w/o score normalization)*

We can see that the GLR based trials selection outperforms the LLR based trials selection. A loss of relative reduction of respectively 0.8% and 4% of DCF and EER is observed when using the LLR based trials selection. [1]

4.4. Oracle experiments

To observe the behaviour of such systems in the optimal conditions, an Oracle experiment is set up for the two different protocols (see fig. 3 and 4). In this case there is no error on the decision of adaptation,as we use the database description in order to decide if

[1] Note that LLR gives better classification results (see section 4.1).

a trial belongs to a given speaker. This experiment shows the limits of such adaptation protocols using the NIST SRE database. 1231 tests from the 13624 are known to be target utterances. It means that 4.42 trials segments (on average) by target correspond to the target model involved.

4.5. Comparison between the threshold based adaptation method and the Oracle

Oracle experiments use all the target trials (1231) to process the speaker model adaptation. The online protocol is more restrictive since the amount of data to perform the adaptation is fewer than for the batch protocol. This may explain the performance difference between the two Oracle experiments (see det curves 3 and 4). Moreover we have seen that about a half of these tests were used to adapt speaker model in the two different threshold based adaptation methods. Some mistakes could have been made in the trials selection too. This may explain the gap between the Oracle experiment results and our system results.

4.6. Comparison between the 128 GD GMM (with unsupervised adaptation) and the 2048 GD GMM (without unsupervised adaptation)

The baseline 2048 GD GMM system is gathered from the system described in 3. All the target model are gathered from Bayesian Adaptation of the World Gaussian means only using a regulation factor of 14. Results of our unsupervised adaptation GLR-based batch protocol system are similar to the baseline in term of DCF whereas it is only 128 Gaussian components (see results in table 5. Indeed, increasing the amount of train data could be considered as a way of improving system performance without having to increase GMM complexity.

	DCF	EER
Baseline 2048 GD	4.12	7.98
Bacth protocol 128 GD	4.04	9.60

Table 5: *EER and DCF of (1) the Baseline 2048 Gaussian components (2) Batch protocol adaptation 128 Gaussian components (w/o score normalization)*

5. Discussion

These results show that the batch protocol gives better performance for unsupervised speaker adaptation with a reduction of the DCF for the Oracle experiments from 2.10 for the online protocol to 1.10 for the batch protocol. Obviously the more tests there are the better an unsupervised adaptation system will behave. The proposed system gives quite good results with a relative reduction of 17.9% and 0.7% respectively for the DCF and EER for the batch protocol and 3.2% and 2.5% relative reduction respectively for the DCF and EER for the online protocol. However, these results are far from the improvement seen in the Oracle experiments.

The batch mode system gives results nearest to the Baseline 2048 GD in term of DCF although it is only based on 128 Gaussian components. So, increasing the amount of training data is a way of improving the recognition rate of a speaker recognition system. Finally we showed that the GLR is more suited than the LLR for an unsupervised adaptation system but its computation is more expensive.

6. Future Work

In the work presented here the trial selection does not use the model updated. It always refers from the target model learnt on a one session record. Future work could investigate taking the trials selection decision on updated target models.
Furthermore this selection is threshold based. We are actually working on a system without hard decision (threshold) to adapt speaker model.
Finally we do not applied T-NORM score normalization in our experiments because the impostor cohort is actually trained on 2.5 min data. A cohort trained on different length of data is needeed to apply T-NORM because the target models are trained on a varying length of data. This will be investigated.

7. Acknowledgements

This work was supported by the French *Ministère de la recherche et de l'industrie* under the CIFRE grant number 858/2005 in association with the Thales Communications company.

8. References

[1] C. Barras, S. Meignier, J. L. Gauvain, ”Unsupervised Online Adaptation for Speaker Verification over the telephone”, In Odyssey , Toledo, May-June 2004.

[2] L.P. Heck, N. Mirghafori, ”Online unsupervised adaptation in speaker verification”, Proc. International Conference on Spoken Language Processing, Beijing, China, Oct 2000.

[3] C. Fredouille, J. Marithoz, C. Jaboulet, J. Hennebert, C. Mokbel, and F. Bimbot, ”Behavior of a bayesian adaptation method for incremental enrollment in speaker verification”. In Proceedings ICASSP, 2000.

[4] NIST Speaker Recognition Evaluation campaigns web site, http://www.nist.gov/speech/tests/spk/index.htm

[5] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, D. A. Reynolds, ”A tutorial on text-independent speaker verification”, *EURASIP Journal on Applied Signal Processing*, 2004, Vol.4, pp.430-451

[6] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, ”Clustering speakers by their voices,” in Proceedings of ICASSP, 1998.

[7] LIA_SpkDet system web site, http://www.lia.univ-avignon.fr/heberges/ALIZE/LIA_RAL

[8] ALIZE project web site, http://www.lia.univ-avignon.fr/heberges/ALIZE/

[9] J.-F. Bonastre, F. Wils, S. Meignier, ”ALIZE, a free toolkit for speaker recognition”, *Proceedings of ICASSP05*, Philadelphia (USA), 2005

[10] ”SPRO: a free speech signal processing toolkit”, Guillaume Gravier,http://www.irisa.fr/metiss/guig/spro/