Confidence measure based unsupervised target model adaptation for speaker verification

A. Preti[12]*,J.-F. Bonastre*[1]*,D. Matrouf*[1]

[1]LIA, 339 chemin des Meinajaries
84911 Avignon Cedex 9, France
alexandre.preti@univ-avignon.fr
jean-francois.bonastre@univ-avignon.fr
driss.matrouf@univ-avignon.fr

F. Capman[2]*, B. Ravera*[2]

[2]Thales, MMP Laboratory, 160 Bd Valmy
92700 Colombes, France
francois.capman@fr.thalesgroup.com
bertrand.ravera@fr.thalesgroup.com

Abstract

This paper proposes a new method for updating online the client models of a speaker recognition system using the test data. This problem is called unsupervised adaptation. The main idea of the proposed approach is to adapt the client model using the complete set of data gathered from the successive test, without deciding if the test data belongs to the client or to an impostor. The adaptation process includes a weighting scheme of the test data, based on the *a posteriori* probability that a test belongs to the targeted client model. The proposed approach is evaluated within the framework of the NIST 2005 and 2006 Speaker Recognition Evaluations. The links between the adaptation method and channel mismatch factors is also explored, using both Feature Mapping and Latent Factor Analysis (LFA) methods. The proposed unsupervised adaptation outperforms the baseline system, with a relative DCF improvement of 27% (37% for EER). When the LFA channel compensation technique is used, the proposed approach achieves a reduction in DCF of 20% (12.5% for EER).

Index Terms: speaker verification, unsupervised adaptation.

1. Introduction

Gaussian Mixture Model (GMM) based systems are widely used in the field of text-independent speaker recognition [1]. Associated with the background model paradigm (so-called UBM/GMM), they achieve a good level of performance, as shown by the NIST-SRE evaluations [1]. The GMM/UBM is also a key element for discriminative approaches like SVM [2].
A large part of the research efforts has been put in terms of channel compensation during the past years in order to deal with intersession mismatches, which is one of the main causes of degradation for a speaker verification system. A classical solution to this problem is to increase the amount of information used to train the client model by including multiple enrolment sessions, as proven by the performance obtained when multi session are provided in training. However, this solution depends on the availability of such training data and the improvements are not related only to the channel effects, the knowledge on the speaker characteristics is also improved by a larger multi session training set. In order to handle the channel effect problem even if only one session per client is available for training, The Latent Factor analysis (LFA) approach was recently introduced in [3]. LFA follows and extends the GMM-UBM paradigm and achieved a significant improvement during the NIST-SRE 2006 campaign.

[1]http://www.nist.gov/speech/tests/spk/

In unsupervised adaptation, the client model is updated online, using data gathered from the test trials. The different approaches to unsupervised adaptation proposed in the literature rely mainly on a decision step to decide if a test trial belongs to the targeted speaker [4, 5, 6, 7, 8]. If the test trial is considered as client, it is used to either retrain the corresponding client model or to adapt it. The main drawback of such techniques remains the difficulty to set a decision threshold for selecting the trials: the performance gain relies on the number of client test trials detected when a false acceptation, an impostor trial is accepted as a client one, degrades the speaker model. Figure 1 illustrates the client and impostor score distributions for a classical UBM/GMM system computed on the NIST SRE 06 with one session for training and testing. Due to this drawback (hard decision), a previous work on adaptation has shown only small improvement. To explain this result, it is reasonable to say that 1) few client tests obtain a score higher than the threshold and 2) if a client test obtains a score higher than the threshold, it is certainly already well represented in the current client model.

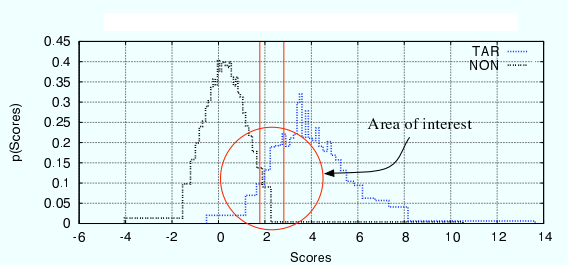

Figure 1: *NIST SRE 05 impostor/target score distribtions*

The work presented in this paper extends previous works [9] which intends to avoid this problem by withdrawing this early decision step. In this approach, the speaker model adaptation is applied to each test trial, even if it does not belong to the client but to an impostor. This "continuous adaptation" method relies on a confidence measure on the tying between the test trial and the target speaker, used to weight each test information during the speaker model adaptation process. We make the asumption that adding a test belonging to an impostor will not damage the target speaker model if it is associated to a low confidence measure. Moreover we hope that some relevant information could be gathered from such impostor data such as channel related information.
This paper is organised as follow. Section 2 is dedicated to the unsupervised adaptation itself. Section 3 concerns the experimental setup while Section 4 presents the experimental results. Finally, Section 5 presents some conclusions and perspectives.

2. Unsupervised adaptation

The work described in this paper intends to avoid the main limitation of classical unsupervised adaptation techniques which rely on a hard decision threshold to decide if a given trial data set should be used for the adaptation. Using a hard decision presents two main disadvantages: to avoid the selection of impostor data for adaptation the thresholds are fixed high, secondly, the selected data is less interesting as it should be already well represented in the current target model. The approach proposed in this paper addresses these problems as no hard decision is used: all the test trials are taken into account during the adaptation phase, according to a confidence measure. The statistics gathered from the test trials are incorporated into the target model using a criteria based on a confidence measure.

2.1. Confidence measure estimation

The confidence measure is the estimation of the *a posteriori* probability of a test trial belonging to a given target speaker. This *a posteriori* probability is computed using two score models, one for the client scores and one for the impostor scores. Each score distribution is modelled by a 12 components GMM learned on a development set.
The confidence measure is then computed using the World MAP (WMAP) approach [9, 10]. WMAP is a simple MAP probability estimator dedicated to speaker recognition as it works on log likelihood ratios, i.e. it takes into account the background model. WMAP also takes into account the prior probabilities of both classes and is defined as follows:

$$P(x=y) = \frac{P(s|x=y).P(x=y)}{P(s|x=y).P(x=y) + P(s|x\neq y).P(x\neq y)} \quad (1)$$

where $P(x=y)$ is the prior probability of a target trial, $P(x\neq y)$ is the prior probability of an impostor trial, $P(s|x=y)$ is the likelihood of the score (LLR) given the target score distribution, $P(s|x\neq y)$ is the likelihood of the score (LLR) given the impostor score distribution.

Note that WMAP outputs a fixed probability equal to the prior probability of a target trial when the observed score is outside the learned target and impostor score distributions, i.e. when the score is very low or very high. Scores used are T-normed [1]. To avoid the problem of the reestimation of the WMAP function after each adaptation step, we use only the initial target model, learned on a single session recording, to compute the score of the test trials.

2.2. Proposed adaptation function

The proposed adaptation function relies on the classical MAP algorithm [1], where only the mean parameters are updated.
The empirical statistics are gathered from all the available data using the EM algorithm (initialized with the background model and maximizing the Maximum Likelihood criterion). The statistics are then combined using the following rules:

- The statistics gathered from the initial voice excerpt used to train the target speaker model is associated to a confidence measure equal to 1;
- The statistics gathered from the different test trials are associated with the corresponding confidence measure;
- The empirical means and the corresponding occupancies are computed for each Gaussian components of the GMM, using all the EM statistics weighted by the corresponding confidence measures.

Finally, the adapted means (μ^i_{map}) for each Gaussian component (i) are computed using the background means (μ^i_{ubm}), the empirical means(μ^i_{emp}) and the occupancy values (n_i) using the MAP formula:

$$\mu^i_{map} = \frac{n_i}{n_i + r}.\mu^i_{emp} + (1 - \frac{n_i}{n_i + r}).\mu^i_{ubm} \quad (2)$$

where r is the MAP relevance factor, fixed to 14 in this work.

3. Tools and Protocols

3.1. Database

All the experiments presented in Section 4 are performed based upon the NIST SRE 2005 and 2006 databases, all trials (det 1), 1conv-4w 1conv-4w, restricted to male speakers only.

This condition consists of 274 and 354 speakers for NIST SRE 2005 and 2006 respectively. Train and test utterances contain 2.5 minutes of speech on average (telephone conversation). The whole speaker detection experiment consists in 13624 tests, including 1231 target tests and 12393 impostors trials for the NIST SRE 2005 database with up to 170 tests by client (50 in average). NIST SRE 2006 database provides 22131 tests, including 1570 target tests with up to about 100 tests by clients (62 in average).The priors used for WMAP are 0.1 for target and 0.9 for impostors. Table 1 provides examples of the cumulative true target trial distributions for four speakers (taken from NIST 05), e.g. true target trials for target x appear three times within the first ten tests, six times between the 10^{th} and 20^{th} (nine times in total) and none between the 20^{th} and the 170^{th}. For target z, there are 7 target trials between the 1^{st} and the 80^{th}, then no more trials are available (denoted by na).

Trials	10	20	50	80	100	120	150	170
Target x	3	9	9	9	9	9	9	9
Target z	1	1	7	7	na	na	na	na
Target y	0	0	0	4	11	na	na	na
Target v	1	8	18	na	na	na	na	na

Table 1: *The cumulative evolution in the number of true target trials according to the number of overall trials (true or impostor) for 4 speakers.*

3.2. Baseline speaker recognition system

The LIA_SpkDet system[2] developed at the LIA lab is used as the baseline in this paper. Built from the ALIZE platform [11], it was evaluated during the NIST SRE'04, SRE'05 and SRE'06 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 parameterization is performed using SPRO[3].
The 512 components UBM is trained on a part of the Fisher corpus[4]. Concerning Tnorm, a cohort of 160 target male speakers of NIST SRE 2004 database has been used. For the front-end processing, the signal is characterized by 50 coefficients including 19 linear frequency cepstral coefficients (LFCC) issued from a filter-bank analysis, their first derivative coefficients, 11 of their second derivative coefficients and the delta energy. An energy-based frame removal is computed before applying a mean subtraction and a variance reduction normalization.
Two different channel normalizations are then applied:

[2] http://www.lia.univ-avignon.fr/heberges/ALIZE/LIA_RAL

[3] http://www.irisa.fr/mestiss/guig/spro

[4] http://papers.ldc.upenn.edu/LREC2004/LREC2004_Fisher_Paper.pdf

- the first process is a feature mapping [12] using three channel conditions (landline, cellular, cordless);
- The second set of features is issued from the Latent Factor Analysis method applied in the feature domain [13]. The adaptation of the feature vector at time frame t, $O(t)$, is obtained by subtracting to the observation feature a weighted sum of the channel compensation offset values:

$$\hat{O}(t) = O(t) - \sum_m \gamma_m(t) \cdot C_m \qquad (3)$$

where $\gamma_m(t)$ is the Gaussian occupation probability, and C_m is the channel compensation offset related to the m-th Gaussian of the UBM model. The eigen channels (matrix U in the model) are learnt via the algorithm given in [3] on the NIST SRE 2004 database, with a rank equal to 40;

The performance is evaluated through classical DET performance curves.

3.3. NIST SRE adaptation mode protocol

The NIST unsupervised adaptation mode allows the update of the target models using the previously seen trial segments (including the current segment) before taking the decision on the current trial segment. It is required to follow the order of the trials in the test protocol. For each test, a score is issued using the test data and the current (adapted) target model.

4. Experiments

This section is dedicated to the experimental results. Results are provided for both 2005 and 2006 NIST databases and for feature mapping and LFA channel compensation techniques.

4.1. NIST SRE 2005 experiments

Figure 2 presents the results for the adapted system and the baseline on the NIST SRE 2005, for feature mapping and LFA channel compensation techniques.

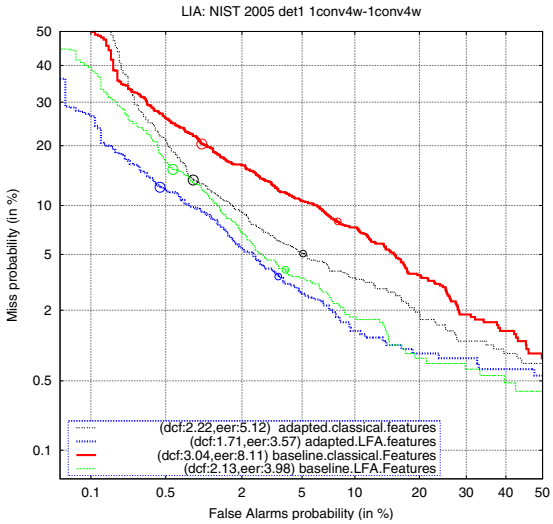

Figure 2: *Results for the adapted/baseline systems, NIST 05*

The results demonstrate the potential of the proposed method as it reaches a significant 27% DCF relative gain (and 37% in terms of EER) with the feature mapping (FM). When the LFA-normalized features are used, the DCF gain is about 20% (and 12.5% for the EER). The gain is smaller in this case, as expected because LFA is known to perform a better channel compensation than FM, reducing the influence of the unsupervised adaptation on the channel effects.

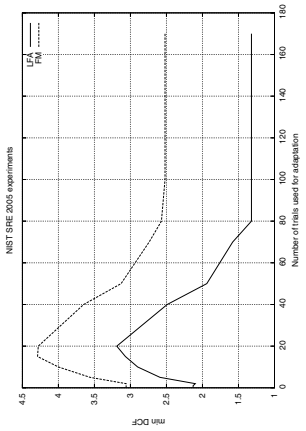

Figure 3: FM

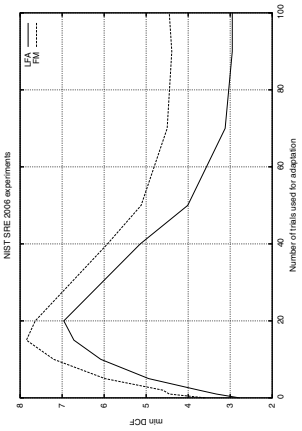

Figure 4: LFA

Step by step DCF for the two different sets of features

In Figure 3, we propose to analyse more precisely the behavior of our method. This figure shows the performance in terms of min DCF of the adapted system for each newly added test trial (n denotes the number of test trials added in the system)[5].The results are very similar for the two channel compensation methods (feature mapping and LFA). The gains are not linear to the number of added trials. The first trials involve a loss in terms of performance, explained by the high percentage of impostor trials compared to target trials proposed by NIST SRE protocols. When enough trials are inserted, the weighting process is able to take advantage of the new information and the unsupervised adaptation brings an improvement. It seems that a minimum quantity of target data should be present before the adaptation could deal with impostor data as the unsupervised adaptation-based system reaches the baseline performance after about 45 added trials. In this case on average four target sessions are added during adaptation. It should be noted that there is not the same number of test trials per speaker. For example, ten or more trials are provided for 274 speakers, fifty or more trials for 118 speakers and 170 trials are available for only three speakers. This fact constrains the positive influence of n parameter as fewer adaptation data is added when n increases.

4.2. NIST SRE 2006 experiments

Results for the adapted system and the baseline are provided in Figure 5, for the NIST SRE 2006 database and using both FM and LFA. The results are disappointing compared to those of NIST05 ones. On this database, the unsupervised adaptation method introduces a significant loss. Three main factors which could explain this unexpected result are discussed bellow.
Firstly, the confidence measure relies on two score distribution models. These models were learned on a separate database (NIST SRE 04) and we have to assess their robustness for NIST06 database compared to NIST05. When looking at the score versus confidence functions produced by the WMAP process for NIST05 and NIST06 we can infer that the problem does

[5]When the target models are updated using a new test trial, the entire test is recomputed, which differs to the NIST protocol where only the current trial and the next trials scores are computed using the new models.

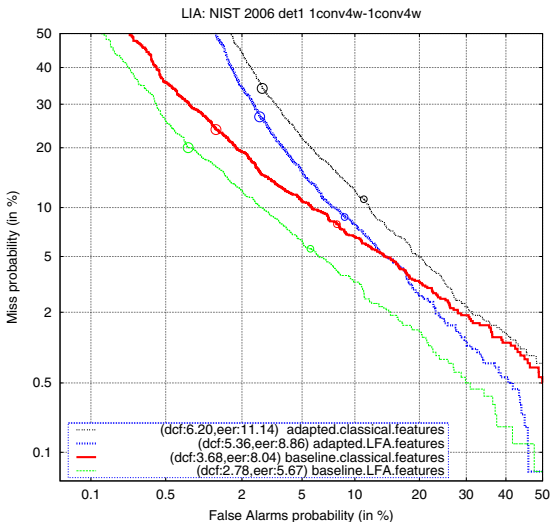

Figure 5: *Results for the adapted/baseline systems, NIST 06*

not come from this part of the process, as the curves are really similar, showing a good generalization power of the confidence estimator.
Secondly, the percentage of impostor trials versus target trials differs between the databases. When about 9.0% of target tests are proposed in NIST05, only about 7.1% are present in NIST06. Even if the difference seems quite small, it gives 21% less target data for 2006 database. Moreover, the number of test trials by target speaker is also smaller for NIST06. In Figure 4, we present the performance of the adapted system versus the number of test trials added to the system. The behavior of the system for NIST06 is very close to the one for NIST05, with a first "loss step" and a second "gain part". This result suggests that the system will be able to bring a gain if enough target trials are proposed.
However when looking at the target and impostor score distributions of the baseline system (cf Figure 1) we observe that a large part of the errors comes from a small percentage of the impostor trials which obtained a very high score. This phenomena is significantly higher for the NIST 2006 than for the NIST 2005 database. It shows the obvious dependence between the baseline system and the adaptation process behavior.

5. Discussion

In this paper we proposed a new method for continuously updating the client models of a speaker recognition system. The main original contribution of the proposed approach is to consider all information gathered from the successive test trials, without deciding formally if a trial belongs to the client or to an impostor. This method relies on the estimation of the trial confidence to belong to a given target model. The confidence measures are used to weight the added data (gathered from the test trials) during a target model adaptation process.
The proposed unsupervised adaptation outperforms the baseline system, with a relative DCF improvement of up to 27% (37% for EER) on NIST 2005 database. When an efficient channel compensation technique is used (Latent Factor Analysis), the improvements are smaller but still significant (20% DCF relative gain and 12.5% for the EER). We noticed that the gain is correlated to the number of target trials presented to the system: a minimum amount of data related to the targeted speaker should be present in order to take advantage of the adaptation.
However the proposed method does not perform well on NIST 2006 database. After a deeper analysis of the unsupervised adaptation behavior, we can conclude that the lack in terms of performance is highly related to the proportion of target and impostor data.
Future work will firstly investigate more deeply the NIST 2006 problem by applying the proposed method on this database with several impostor versus target trial proportions. Secondly, the confidence measure was estimated globally in this work, using all the data gathered from a test trial and for a complete target model. We will try to estimate this confidence independently for each component of the target model and using a segmental view of the test data.

6. References

[1] 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 JASP*, 2004, Vol.4, pp.430-451

[2] W. M. Campbell, D. E. Sturim, D. A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification", Signal Processing Letters, IEEE Volume 13, Issue 5, May 2006 Page(s):308 - 311.

[3] P. Kenny, G. Boulianne, P. Oullet, and P. Dumouchel, "Improvements in factor analysis based speaker verification", In ICASSP, Toulouse, France, 2006.

[4] C. Barras, S. Meignier, J. L. Gauvain, "Unsupervised Online Adaptation for Speaker Verification over the telephone", In Odyssey , Toledo, Spain, 2004.

[5] D. A. van Leeuwen, "Speaker adaptation in the NIST Speaker Recognition Evaluation 2004" In Interspeech, Lisbon, Portugal, 2004.

[6] E.G. Hansen, R.E. Slyh, T.R. Anderson, "Supervised and Unsupervised Speaker Adaptation in the NIST 2005 Speaker Recognition Evaluation", In Odyssey, Puerto Rico, USA, 2006.

[7] A. Preti, J-F. Bonastre, "Unsupervised model adaptation for speaker verification", In ICSLP, Pittsburgh, USA, 2006.

[8] L.P. Heck and N. Mirghafori, "Unsupervised On-Line Adaptation in Speaker Verification: Confidence-Based Updates and Improved Parameter Estimation", Proc. Adaptation in Speech Recognition, Sophia Antipolis, France, August 2001.

[9] A. Preti, J-F. Bonastre, F. Capman, "A continous unsupervised adaptation method for speaker verification", International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering 2006

[10] C. Fredouille, J-F. Bonastre, T. Merlin. "Bayesian approach based-decision in speaker verification". In Odyssey, Crete, Grece, 2001.

[11] J.-F. Bonastre, F. Wils, S. Meignier, "ALIZE, a free toolkit for speaker recognition", In ICASSP, Philadelphia, USA, 2005.

[12] D. A. Reynolds, "Channel Robust Speaker Verification via Feature Mapping", ICASSP, IEEE, Hong Kong, 2003, pp. 53-56.

[13] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, P. Laface,"Channel Factors Compensation in Model and Feature Domain for Speaker Recognition", In Odyssey, Puerto-Rico, USA, 2006.