IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Behavior of a Bayesian adaptation method for incremental enrollment in speaker verification

Corinne Fredouille [a] [b] Johnny Marithoz [c]
C. Jaboulet [a] J. Hennebert [a]
J.-F. Bonastre [b] C. Mokbel [d]
F. Bimbot [e]

IDIAP–RR 00-02

January 2000

published in
ICASSP2000

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] UBS-Ubilab - Banhofstr. 45, CH-8098, Zurich, Switzerland
[b] LIA - BP 1228, 84911 Avignon, France
[c] IDIAP - BP 592, CH-1920 Martigny, Switzerland
[d] Univ. St Joseph - Fac. des Sciences, MarRoukos, Lebanon
[e] IRISA - Campus Beaulieu, 35042 Rennes, France

IDIAP Research Report 00-02

Behavior of a Bayesian adaptation method for incremental enrollment in speaker verification

Corinne Fredouille   Johnny Marithoz   C. Jaboulet   J. Hennebert
J.-F. Bonastre   C. Mokbel   F. Bimbot

January 2000

published in
ICASSP2000

Abstract. Classical adaptation approaches are generally used for speaker or environment adaptation of speech recognition systems. In this paper, we use such techniques for the incremental training of client models in a speaker verification system. The initial model is trained on a very limited amount of data and then progressively updated with access data, using a segmental-EM procedure. In supervised mode (i.e. when access utterances are certified), the incremental approach yields equivalent performance to the batch one. We also investigate on the impact of various scenarios of impostor attacks during the incremental enrollment phase. All results are obtained with the Picassoft platform - the state-of-the-art speaker verification system developed in the PICASSO project.

Contents

List of Figures

1 Introduction

HMM adaptation techniques have been successfully applied in several domains of speech and speaker recognition. These techniques allow supervised or unsupervised adaptation of a recognition system to a particular condition of use, e.g. a particular speaker or a specific environment by adjusting speaker model parameters [3][6][8]. In that case, the initial model is generally estimated using a large amount of data. Adaptation techniques are also used in the context of speaker verification, for estimating a speaker model as an adapted version of a speaker-independent model [10].
The work presented here has been carried out in the context of Work-Package 5 of the European Telematics PICASSO project [1], where robust approaches to text-dependent speaker verification are studied. In the type of applications targeted by the project, a very limited amount of active enrollment data is available (typically, 2 sessions x 2 repetitions). These data provide an unsufficient coverage of the variability of the client's voice and of the variety of conditions of use. To improve the model quality, we investigate an incremental enrollment scheme for adjusting and updating progressively the model with access utterances produced during the actual use of the system, i.e without requiring any specific speech material beside the one uttered by the client while he/she is using the application.
To this aim, we use an HMM adaptation technique, which starts from a model trained on the active enrollment utterances and is then adapted with each new access utterance. This task is performed by an incremental version of the segmental EM algorithm (section 2). It was shown in [9] that this algorithm is a particular case of the Maximum A Posteriori (MAP) adaptation algorithm [3] with adequate choice of the priors depending on the initial data set.
To integrate this incremental approach in our speaker verification system, two points have been studied. Firstly, we have investigated the incremental enrollment in supervised mode, i.e when the client's identity is certified during the adaptation process (section 4). This mode is then experimented with and compared to a classical speaker enrollment. In section 5, the incremental enrollment in unsupervised mode has been studied. For this purpose, several protocols have been defined in order to investigate the behavior of the proposed adaptation technique in the case of impostor attacks. We finally draw a few conclusions and perspectives from this set of experiments.

2 Bayesian adaptation

A first-order HMM λ is defined by a set of Q states, a set of output distributions associated to the states, which we suppose Gaussian with a diagonal covariance matrix $\{N_i(., \underline{\mu}_i, \underline{\underline{\Gamma}}_i); i = 1, \ldots, Q\}$ and a set of transition probabilities between states $\underline{\underline{A}} = \{a_{ij}; i, j = 1, \ldots, Q\}$.
As exposed in [9], the incremental enrollment algorithm proceeds as follows: if a model λ_I is already trained with an initial set of speech data $\underline{\underline{X}}^{(I)}$; and if some new data $\underline{\underline{X}}^{(N)}$ are available to enrich the model parameters, the adaptation procedure by the segmental (Viterbi) EM algorithms yields:

$$\lambda^{opt} = \underset{\lambda}{argmax}[\underset{(S_I, S_N)}{max}\, p(\lambda, S_I, S_N | \underline{\underline{X}}^{(I)}, \underline{\underline{X}}^{(N)})] \tag{1}$$

This procedure optimizes only the model parameters and the new state sequences given the whole data. The optimal state sequences of the initial data are fixed to the same as in the initially trained model. This means that the state sequences corresponding to the initial training data are considered to be always optimal. Eq. 1 can be written:

$$\lambda^{opt} = \underset{\lambda}{argmax}[\underset{S_N}{max}\, p(\lambda, S_I^{opt(\lambda_I)}, S_N | \underline{\underline{X}}^{(I)}, \underline{\underline{X}}^{(N)})] \tag{2}$$

By considering the adaptation of the Gaussian means only, solving this maximization leads to the following re-estimation equation[1]:

$$\underline{\mu}_l^{(i)} = \frac{n_l^I \cdot \underline{\mu}_l^I + n_l^N \cdot \overline{\underline{X}_l}^N}{n_l^I + n_l^N} \tag{3}$$

[1]The experiments reported later are based on the adaptation of Gaussian means only, see [9] for re-estimation equations of covariance matrix and Gaussian weights.

where $n_{i,m}^{(I)}$ (resp. $n_{i,m}^{(N)}$) is the number of feature vectors of the initial (resp. new) data set associated with the m^{th} component of the i^{th} distribution and $\mu_{i,m}^{(I)}$ (resp. $\underline{X}_{i,m}^{(N)}$) is the mean of those initial (resp. new) feature vectors. All the parameters of the initial model are constant and do not depend on the Estimate step of the iteration.

3 Experimental conditions

3.1 Baseline system

The baseline system is the Picassoft platform, i.e a speaker verification system developed in the PICASSO project, around the HTK software package [12], similarly to what was done in the CAVE project [5].
Acoustic features are 12 LPC cepstral coefficients with log-energy, together with their first and second derivatives. Cepstral mean subtraction is applied at the whole utterance level. For client and world modeling, Left-Right HMM word models are used, similarly to the approach in [11]. The topology of the client and world HMMs[2] is fixed to two states per phoneme with one Gaussian distribution per state. The initial client models are estimated with two training sessions.

3.2 PolyVar database

All the experiments have been conducted on a subset of the PolyVar database. PolyVar is a Swiss French database for speaker verification tasks in telephony environments. It contains various items such as read sentences, digits, command words, dates, etc, pronounced by 143 speakers [2].
The task being text-dependent, only a subset of PolyVar has been used. This subset is composed of 17 command words, uttered by speakers during multiple enrollment sessions. The speaker group is composed of 42 females and 52 males, split into three different populations. Two of them, of 19 speakers each (7 females and 12 males) named Population A and B are in turns dedicated to development and evaluation phases. A third population (population W) is used to estimate the world model (28 females and 28 males). In practice, 5 training sessions are available for each word of the vocabulary, for each speaker of Population A and B. In average, 25 repetitions of each word have been uttered by each speaker.

4 Supervised mode

In supervised mode, access data used to adapt a client model are certified as belonging to the correct client.
The experiments presented here aim at comparing the adaptation based incremental enrollment in supervised mode with a classical enrollment. For incremental enrollment, the client model parameters are first estimated on two training sessions. They are afterwards adapted incrementally with one additional training session at a time. Three adaptation steps are used in this paper. For the classical enrollment, the same number of training sessions (i.e five) as for incremental enrollment are used to estimate the client model parameters in batch mode.
Figure 1 depicts the DET [7] curves obtained with incremental enrollment ("12+3+4+5") and classical enrollment on five training sessions ("12345"). For reference, the DET curve obtained with a classical enrollment on the first two training sessions only ("12") is also provided. As expected, it is observed that classical enrollment on five training sessions and incremental enrollment outperform the classical enrollment on two training sessions. This underlines the requirement for large amount of training data for an accurate estimate of speaker model parameters. On the other hand, similar performance is obtained for both classical enrollment and incremental enrollment approaches on five sessions. In

[2]Full word models are used. Parameters are not shared between word HMMs.

this case, the adaptation approach proposed in this paper is able to incrementally adjust client model parameters to new conditions of use while reaching client model quality similar to a classical enrollment approach applied on the same amount of training data used in batch mode.

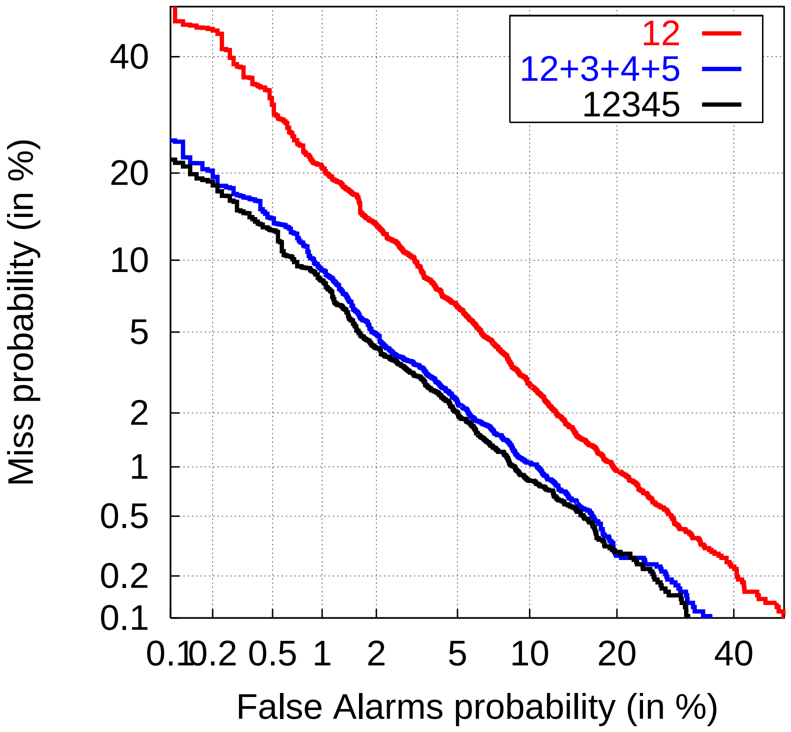

Figure 1: Comparison of supervised incremental enrollment (12+3+4+5) to classical enrollment in batch mode (12345). Performance of the initial model (12) is also represented.

5 Unsupervised mode

Adaptation-based incremental enrollment in unsupervised mode is a more troublesome task. In this context access data used to adapt client model can belong to the client or to an impostor. No a priori certainty can be guaranteed regarding the identity of the speaker and the decision whether to use the data or not for adaptation is left to the system.
In practice, similarly to the decision making process applied for speaker verification, a threshold (which can be specific to the incremental adaptation purpose) is used to decide whether confidence on incoming data is sufficient to use these data or not for adapting the client model. Similarly to a baseline ASV, False Adaptation Acceptance[3] (FAdA) and False Adaptation Rejection[4] (FAdR) can occur. FAdA and FadR rates are defined as:

$$FAdA = \frac{Number\ of\ Impostor\ Adaptation\ Acceptances}{Number\ of\ Impostor\ Access} \tag{4}$$

$$FAdR = \frac{Number\ of\ Client\ Adaptation\ Rejections}{Number\ of\ Client\ Access} \tag{5}$$

[3] An impostor is accepted to adapt a client model.
[4] A client is rejected for incremental enrollment.

5.1 Different scenarios of impostor attacks

In unsupervised mode, it can be reasonably assumed that false adaptation acceptances might cause client model degradation. Therefore, the experiments reported here aim at studying the behavior of the proposed adaptation algorithm against different scenarios of impostor attacks. These scenarios have been designed as follows:

•**Protocol P0.** This protocol is based on client and impostor attempts. These attempts occur randomly while respecting the chronological order within the access set of each client (resp. impostor).

•**Protocol P1.** This protocol aims at simulating massive impostor attacks. Therefore, the attempt list used before in the protocol P0 is reordered to group all the impostor attempts at the beginning. The chronological order is still respected.

•**Protocol P2.** The purpose of this protocol is to simulate attacks of a unique impostor against each client. Four phases are defined and performed for each client:

1. Attacks of a unique impostor: a unique impostor is selected to attack a client. During the attacks, a fixed number of impostor attempts (five in this paper) are carried out against the client model, and they are all supposed successful. They are thus all used for incremental adaptation.

2. Post-attack tests: this phase, based on a **standard test** (with client and impostor attempts without incremental enrollment) aims at evaluating the client model performance after the attacks of the unique impostor.

3. Client adaptation attempts: a series of client attempts (five in this paper) are carried out with incremental enrollment. The client model remains possibly degraded by the initial impostor attacks of step 1. As for step 1, all the client data are used for incremental adaptation.

4. Final test: a new phase of test, similar to the one of step 2, is performed to evaluate the behavior of the client model estimated in step 3.

To keep a consistent number of tests, each client is a potential unique impostor to attack all the other clients. The four phases are repeated until all the clients have been involved in this process.

If compared to a real-life situation, protocol P2 is certainly a worst-case situation, as, in general, a significant proportion of the massive impostor attacks would be rejected.

5.2 Test configurations

To investigate on the behavior of incremental enrollment, experiments have been conducted on protocols P0, P1 and P2. Results of these experiments are illustrated by DET curves.
The same speaker population is involved in the three protocols.
For protocols P0 and P1, the same test data set is used (6478 client access and 11628 impostor access). Only, the order of attempts within this set differs between both protocols (impostor attempts occur first for P1). Therefore, results obtained on both protocols are comparable.
For protocol P2, three separate test data sets are used: a first set for the attacks of a unique impostor, a second one for the client adaptation attempts and a third one for the test phases (post-attack tests and final tests). The total number of access performed over all the phases is 221293 client access and 470934 impostor access.
To evaluate the incremental enrollment on protocol P2, different scoring configurations are proposed:
• "IM+IA": log likelihood ratios computed during both phases of unique impostor attacks and of post-attack tests are used to yield a DET curve.
• "IM+IA+CA": log likelihood ratios computed during both phases of client adaptation attempts and of final tests are used to yield a DET curve.
• "IM": log likelihood ratios computed while running the final test with a classical enrollment i.e with

Protocol	FAdA (%)	FAdR (%)	HTER (%)
P0	6.8	6.8	6.8
P1	13.8	3.4	8.6

Table 1: FAdA, FAdR and HTER for protocols P0 and P1.

speaker models trained on two sessions only are used to yield a reference DET curve.
• "IM+CA": log likelihood ratios computed during both phases of client adaptation attempts and of final tests are used to yield a DET curve. However, as opposed to "IM+IA+CA", no phase of unique impostor attacks has been applied previously. This configuration shows what the ideal situation would look like if no impostor data were used for adaptation (equivalent to a supervised incremental enrollment).
The adaptation threshold (see section 5) is set a posteriori in order to optimize the HTER (arithmetic average of the FAdA and the FAdR). For both protocols P0 and P1, incremental enrollment is tested in normal conditions with an adaptation threshold set up to -1. Conversely, for protocol P2 this threshold is set to $-\infty$ in order to use all the data (stemming from client or impostor) for the adaptation. This $-\infty$ value is chosen in order to estimate the impact of successful unique impostor attacks on incremental enrollment.

6 Results

6.1 Protocols P0 & P1

Figure 2 provides DET curves obtained using incremental enrollment on both protocols P0 and P1 and using classical enrollment on two training sessions used in batch mode.
One can observe that incremental enrollment on both P0 and P1 gives the best performance if compared to the classical enrollment. In terms of EER, P0 outperforms P1. Nevertheless, the difference of DET curves between P0 and P1 is not as larger as expected. Indeed, depending on the massive impostor attacks involved in P1, it could be assumed that performance would degrade drastically. In fact, the careful study of the FAdA and FAdR rates for P0 and P1 (see table 1) tends to show that, as expected, the massive impostor attacks do untune the client model towards a more "acceptant" model but fewer false rejections occur and the HTER is only increased by approximately 30 % relative error.

6.2 Protocol P2

Figure 3 shows the DET curves obtained on protocol P2 and related to the different scoring configurations defined in section 5.2. These curves reveal large performance differences and the following points can be underlined:
•"IM+IA". Five adaptations from data of a unique impostor are sufficient to really degrade client models and to multiply by more than 2 the EER of the reference curve. Therefore, unique impostor attacks are an important issue for the adaptation method.
•"IM+IA+CA". Pursuing with five client adaptations, the client models have worse performance than initial client models based on two training sessions. Degradation due to impostor adaptation is laboriously reversible.

7 Conclusions

This work advocates for the viability of an adaptive approach for speaker model update, provided that the verification based on the initial model is reliable enough to control the proportion of impostor speech in the incremental enrollment scheme. In case of massive successful impostor attacks, the model can get severely untuned. But techniques for monitoring such large deviations can be envisaged and they will be a topic for our future work.

References

[1] F. Bimbot, M. Blomberg, L. Boves, & al, An overview of the PICASSO project research activities in speaker verification for telephone applications, *Eurospeech'99*, pp 1963-1966, Budapest (Hungary), Sept. 1999.

[2] G. Chollet, J.-L. Cochard, A. Constantinescu, C. Jaboulet, P. Langlais, Swiss French PolyPhone and PolyVar: telephone speech databases to model inter- and intra-speaker variability, *Linguistic Databases*, edited by John Nerbonne, pp 117-135, 1997.

[3] J. L. Gauvain, C.-H. Lee, Maximum A Posteriori estimation for multivariate Gaussian mixture observation of markov chains, *IEEE Transactions on Speech Audio Processing*, Vol.2(2), pp 291-298, April 1994.

[4] O. Kimball, M. Schmidt, H. Gish, Speaker verification with limited enrollment data, *Eurospeech'97*, Rhodes (Greece), Sept. 1997.

[5] C. Jaboulet, J. Koolwaaij, J. Lindberg, J.-B. Pierrot, F. Bimbot, CAVE-WP4 generic speaker verification system, *RLA2C'98*, Avignon (France), April 1998.

[6] C.-H. Lee, C.-H. Lin, B.-H. Juang, A study on speaker adaptation of the parameters of continuous density hidden markov models, *IEEE Trans. on ASSP*, Vol.39, num.4, pp 806-814, Apr. 1991.

[7] A. Martin, M. Przybocki, The DET curve in assessment of detection task performance, *Eurospeech'97*, Vol.4, pp 1895-1898, Rhodes (Greece), Sept. 1997.

[8] C. Mokbel, L. Mauuary, L. Karray, D. Jouvet, J. Monné, J. Simonin, K. Bartkova, Towards Improving ASR Robustness for PSN and GSM Telephone Applications, *Speech Communication*, V. 23, n1, pp. 141-159, Oct. 1997.

[9] C. Mokbel, O. Collin, Incremental enrollment of speech recognizers, *ICASSP'99*, Phoenix (USA), March 1999.

[10] D. A. Reynolds, Comparison of background normalization methods for text-independent speaker verification, *Eurospeech'97*, pp 963-966, Rhodes (Greece), Sept. 1997.

[11] A. E. Rosenberg, C.-H. Lee, S. Gokcen, Connected word talker verification using whole word HMM, *ICASSP'91*, Toronto (Canada), 1991.

[12] S. Young, J. Odell, D. Ollason, V. Valtchev, P. Woodland, The HTK book, *HTK 2.1 manual*, 1997.

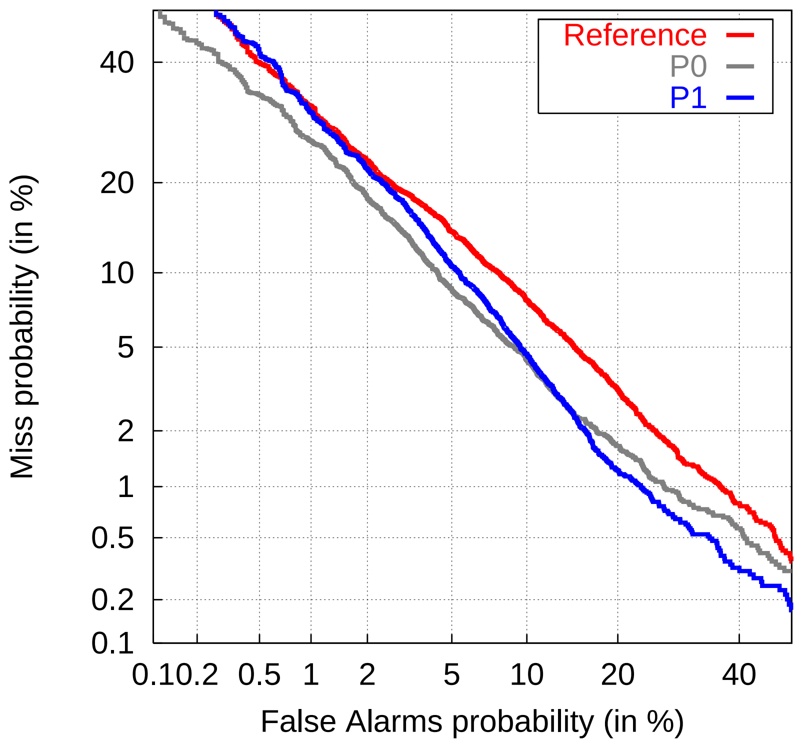

Figure 2: Comparison of a classical enrollment on two training sessions with an incremental enrollment in unsupervised mode under two different protocols P0 and P1.

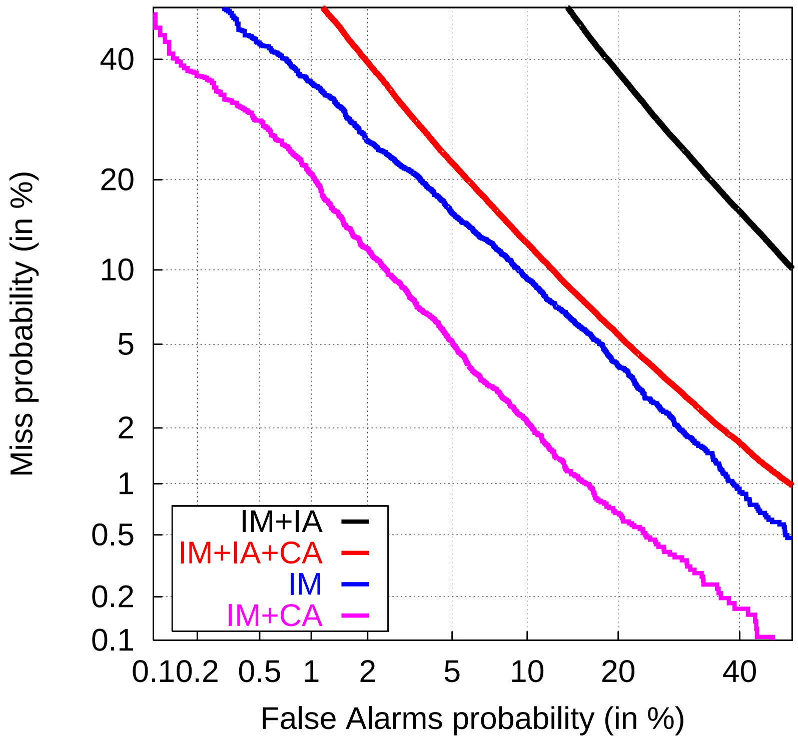

Figure 3: Behavior of incremental enrollment in unsupervised mode under protocol P2 with different scoring configurations.