Combining methods to improve speaker verification decision

Dominique Genoud* Frédéric Bimbot# Guillaume Gravier* Gérard Chollet*#

* IDIAP, CP 592, CH-1920 Martigny, #CNRS URA820 F-75634 Paris France

ABSTRACT

The aim of this paper is to describe how the combination of speaker verification algorithms with a priori decision thresholds can improve the overall robustness of a real application. The evaluation is performed in the context of a field application where each client is verified from a 7 digit pin code. This paper demonstrate that it is possible to increase the global performances of the system on combining the result of several algorithms.

1. INTRODUCTION

Moving from laboratory to real applications is often, for speaker verification systems, a very disappointing experience. Among the known problems, the lack of training speech data is a crucial one. Even if very powerful algorithms are available now, problems with a priori threshold settings due to the amount of training data can decrease drastically the performance. Moreover large performance discrepancies can be observed, depending on speaker.

This paper describes a real speaker verification application and the algorithms used,it presents 3 approaches to set a priori thresholds and discusses how to combine methods in order to obtain better performance.

2. THE APPLICATION

The goal of the application described here is to securize an information server [Gen96] based on telephone access. The application server is connected to an ISDN line. The application is divided in two parts: enrolment and access to the service.

During the enrolment phase the speaker is asked to pronounce his name, christian name, address, all the digits from 0 to 9 sequentially, and 5 times his 7 digit personal identification number (PIN). Speech recognition is performed on all the digit sequences in order to time-label the sequences.

During the access phase, the speaker pronounces once his PIN code. A verification process divided in two phases starts then: (1) The digit sequence is recognized using a HMM based speaker independent speech recognizer [Gro93]. (2) the speech sequence is then compared to the speaker references corresponding to the PIN code recognized during the first phase. Depending on the similarity between the reference and the incoming sequence the speaker is accepted or rejected. If the confidence in the decision is too low, the speaker is asked to pronounce a new sentence. This sentence is used to perform a text independent verification which is used to take a final acceptance or rejection decision. In this paper, only the text dependent part using the PIN code is discussed.

3. THE DATABASE USED

The results in this paper are obtained on a database [GC95] composed of 25 speakers recorded over a telephone line in several sessions. During the same session, each speaker had to say, among other sentences in French, 5 times his own 7 digit PIN code and 4 times 10 digit sequences (all the digits from 0 to 9 in different order for each sequence). All these sequences are time-labeled digit by digit using a speech recognizer. Some sub-databases are extracted from this Polycode database. (1) The PolyTD sub-database built with 10 of the 25 speakers, contains the five 7 PIN code utterances of the same session for each speaker. (2) The PolyTDimp sub-database contains the same 10 speakers uttering one 10 digit sequence randomly chosen in all the sessions. In order to simulate impostor access, a given PIN code is extracted from the 10 digit sequences and recomposed using the time labels. (3) The PolyParam sub-database is composed by 7 digit and 10 digit sequence of the 25 speakers database. The sequences used are different from PolyTD and PolyTDimp. This sub-database is used to calculate some constants or determine global thresholds. (4) The PolyTestI sub-database, which also contains the same 10 speakers than in PolyTD, is composed as followed for each speaker: 20 samples of his PIN number pronounced by himself and 9x20 samples of his PIN number constructed from 10-digits sequences pronounced by each of the other speakers. So, in total, there are 200 correct access trials and 1800 impostor trials. The sequences of this sub-database are not the same than PolyParam, PolyTD and PolyTDimp.

4. SPEAKER VERIFICATION METHODS

Three text dependent verification methods were used. These three methods take as input a set of LPC cepstral coefficient with delta and delta-delta coefficient.

4.1. Dynamic Time Warping (DTW)

The DTW algorithm is well known in speech and speaker recognition. It mainly consists in a dynamic comparison between a reference and a test matrix. The algorithm computes a distance between the test and reference patterns [HC78].

4.2. Second Order Statistical Method (SOSM)

In this algorithm a covariance matrix

$$X = \frac{1}{M} \sum_{t=1}^{t=M} X_t X_t^T$$

is generated out of the reference LPCC vector sequence. A covariance matrix Y is created in the same way with the test sequence.

A sphericity measure based on $\mu_{AH}(X,Y)$ [BM94] is performed:

- $\mu_{AH}(X,Y) = \log[\frac{\mathbf{A}}{\mathbf{H}}]$
- $\mathrm{A}(\lambda_1, \lambda_2, \ldots, \lambda_m) = \frac{1}{m} \sum_{i=1}^{i=m} \lambda_i = m^{-1} tr(\mathbf{YX^{-1}})$
- $\mathrm{H}(\lambda_1, \lambda_2, \ldots, \lambda_m) = m(\sum_{i=1}^{i=m} \frac{1}{\lambda_i})^{-1} = m(tr(\mathbf{XY^{-1}}))^{-1}$

No explicit extraction of the eigen values is necessary, the sphericity measure only needs the calculation of the trace $tr(.)$ of the matrix product YX^{-1} or YX^{-1}.

4.3. Hidden Markov Models (HMM)

Two types of HMM [Gok91] are created for each digit (0 to 9): (1) a world model, trained on a database (Swiss French Polyphone [Lan95]) where 300 occurrences of each digit uttered by around 500 different people were extracted. The parameters of this model are estimated by a classic training (Viterbi algorithm and Baum-Welch re-estimation) [Gro93]. This model is speaker independent. (2) a speaker model, which uses as initial parameters the parameters of the world model, and which is re-estimated with the speaker data.

All models have the same HMM left-right one mixture per state structure. Each model has one state per phoneme and one state per phoneme transition [Gra95].

At the time of access, for each digit uttered by the speaker, the log likelihood ratio (LLR)

$LLR_{sw} = \log(L_s) - \log(L_w)$ is computed.

with L_s, L_w being likelihood of the speaker and world models respectively.

5. THRESHOLD SETTINGS

The speaker acceptance or rejection decision in the application is done by comparing the results of the methods (distance or LLR) to a threshold. Three different threshold setting approaches were chosen.

5.1. Speaker independent EER threshold

This threshold is determined by the Equal Error Rate criterion [Ogl94] obtained on PolyParam sub-database (see paragraph 3). It is speaker independent.

5.2. Speaker dependent EER threshold

This threshold is determined for each speaker on the training data PolyTD and PolyTDImp sub-database (see paragraph 3) using also the EER criterion.

5.3. Speaker dependent threshold by FURUI method

Furui demonstrated that in case of few training data, a better threshold can be determined using only impostor access [FUR94]. The threshold determination is divided in two phases : (1) Two constants $C1, C2$, are estimated using the PolyParam (see paragraph 3) sub-database by linear regression. These two constants are speaker independent. (2) During the enrolment phase the speaker dependent threshold is estimated by :

$$Threshold_x = C1(\mu_x - \sigma_x) + C2$$

with μ_x, σ_x the Gaussian parameters of $N(\mu, \sigma)$ estimated on the impostor scores for each speaker x.

6. COMBINING THE DECISIONS

To improve the global response of the application, the decisions given by each method (DTW, SOSM, HMM) were combined. Many possibilities of combining decision are available [The93, Ant95, Das94]. But, depending on the way the decisions are combined, information about the inter-method dependency is necessary.

A weighted majority test is chosen here, as it doesn't need this inter-method dependence information. Each method M takes its own decision d, these decisions are weighted w (normalized between 0 and 1) by the distance between the threshold and the current method score. This can be understood as a confidence in the decision. The weighting function is in this case chosen sigmoidal.

- for all $M_i(d,w), d \in \{0 = reject, 1 = accept\}, w \in [0,1]$, $i = 1, \ldots, n$ with n number of Methods
- if $(\sum_{(d_i=0)} M_i > \sum_{(d_i=1)} M_i)$ $x = 0$, else $x = 1$
- $c = \frac{1}{S} \sum_{(d_i=x)} (d_i = x)(w_i = x)$ with $S = \sum_{(d_i=x)} d_i$
- if $c > Th$, $D = x$, else $D = doubtful$

Afterwards, the confidence mean of the majority methods c is compared to a doubt threshold Th. If this mean is lower than the threshold, a global doubtful decision is taken otherwise the global decision is the decision x of the majority.

7. RESULTS

For the tests we used the PolyTestI sub-database (see paragraph 3). For each of the 10 speakers of the database, 20 true access and 180 impostor access were performed.

Method	FR% (200 tests)	FA% (1800 tests)
DTW	100	0.0
SOSM	51.0	0.0
HMM L/R	28.5	1.05
Combined Decision	33.5	3.05

Table 1: Method Performance with a speaker dependent EER Threshold

Table 1 shows clearly the difficulty to set a proper speaker dependent threshold with few training data. In this case combining methods doesn't improve the final decision. It can also be seen that the threshold is systematically set too low.

Method	FR% (200 tests)	FA% (1800 tests)
DTW	31.5	21.78
SOSM	22.2	30.3
HMM L/R	2.5	5.33
Combined Decision	8.0	3.17

Table 2: Method Performance with a speaker independent EER Threshold

Table 2 shows that the HMM method gives the best results with a speaker independent threshold, due to the fact that a normalization is done with a world model. Here also, combining methods doesn't improve the final decision.

When the threshold is determined by Furui's method. Table 3 shows that each method gives a better score and that combining decisions gives a better final score.

Table 4 give an idea on how the False Acceptance and False Rejection rate evolve when the doubt threshold is modified.

Method	FR% (200 tests)	FA% (1800 tests)
DTW	23.5	7.67
SOSM	14.0	5.28
HMM L/R	5.53	2.72
Combined decision	2.0	2.72

Table 3: Method performance with FURUI threshold

Doubt threshold	FR % (200 tests)	FA % (1800 tests)	Doute% (2000 tests)
0.2	2.0	2.72	0.0
0.5	2.0	2.33	0.83
0.7	2.0	1.11	10.11
0.8	1.0	0.89	20.2

Table 4: Performance with different doubt threshold

The possiblility of a doubtful decision is a way to set a proper level of security in an application.

8. CONCLUSION

The overall performance of a speaker verification system can be improved in combining methods with a proper threshold selection. Our experiments confirm also the robustness of Furui threshold setting method when few training data are available. Combining methods seems to reveal an alternative way to improve the decision taken on complex input systems. The combining algorithm chosen here is quite simple and the research will go on adding more methods, as for example Neural Nets (NN) or Hybride HMM-NN approaches [Haf94], and also in using more sophisticated decision algorithms like non linear fusion with Neural Nets.

9. ACKNOWLEDGMENT

We wish to thanks the Swiss Telecom PTT for their support in creating the speaker verification application, the European Community, and OFES for their support in the Telematics CAVE (CAller VErification), Acts M2VTS (Multi Modal Verification for Tele-services and Security applications) and COST 250 projects.

10. REFERENCES

Ant95. Richard T. Antony. Principles of Data Fusion Automation. Artech House, 685 Canton Street Norwood,MA 02062, 1995.

BM94. Frédéric BIMBOT and Luc MATHAN. Second-order statistical measures for text-independant speaker identification. In ESCA [ESC94], pages 51–54.

Das94. Belur V. Dasarathy. Decision Fusion. IEEE Computer Society Press, Los Alamitos, California, 1994.

ESC94. ESCA, editor. ESCA Workshop on Automatic Speaker Recognition Identification Verification. Bimbot, Chollet Paoloni, Martigny April 1994.

FUR94. Sadaoki FURUI. An overview of speaker recognition technology. In ESCA [ESC94], pages 1–9.

GC95. Dominique Genoud and Gérard Chollet. Polycode a verification database. Technical report, IDIAP, CH-1920 Martigny, 1995.

Gen96. O. Bornet G. Chollet J-L. Cochard A. Constantinescu D. Genoud. Secured vocal access to telephone servers. In IVTTA, editor, IVTTA Proc., Basking Ridge NJ, 1996.

Gok91. A.E. Rosenberg & C.H. Lee & S. Gokoen. Connected word talker verification using whole word hidden markov model. In ICASSP-91, pages 381–384, 1991.

Gra95. Guillaume Gravier. Vérification du locuteur par modèles de markov cachés gauche-droite. Rapport de stage dea, IDIAP, CH-1920 Martigny, 1995.

Gro93. Cambridge University Speech Group. HTK Hidden Markov Model Toolkit. Entropic Research Laboratories Inc., Cambridge, December 1993.

Haf94. Patrick Haffner. A new probabilistic framework for connectionist time alignment. In ICSLP, editor, ICSLP Proc., 1994.

HC78. Sakoe H. and Chiba. Dynamic programing algorithm optimization for spoken word recognition. IEEE Trans. on ASSP, 26(1):43–49, 1978.

Lan95. G. Chollet & J.L. Cochard & A. Constantinescu & Ph. Langlais. Swiss french polyphone and polyvar: telephone speech databases to study intra and inter speaker variability. Technical report, IDIAP, 1995.

Ogl94. John Oglesby. What's in a number ?:moving beyond the equal error rate. In ESCA [ESC94], pages 87–90.

The93. Ph. Thevenaz. Résidu de prédiction linéaire et reconnaissance de locuteurs indépendante du texte. PhD thesis, Université de Neuchâtel, 1993.