Constrained Viterbi decoding for embedded user-customised password speaker recognition

Anthony Larcher, Jean-François Bonastre, John S.D. Mason

Constrained Viterbi decoding for embedded user-customised password speaker recognition

AB-105

Anthony Larcher
University of Avignon, LIA
339 chemin des Meinajaries
Avignon, France 84000
anthony.larcher@univ-avignon.fr

Jean-Francois Bonastre
University of Avignon, LIA
339 chemin des Meinajaries
Avignon, France 84000
jean-francois.bonastre@univ-avignon.fr

John S.D. Mason
Speech and Image Group,
Swansea University
Singleton Park
Swansea, UK SA2 8PP
j.s.d.mason@swansea.ac.uk

ABSTRACT

Embedded speaker recognition in mobile devices could involve several ergonomic constraints and a limited amount of computing resources. GMM/UBM systems have proved their efficiency in more classical contexts where good accuracy depends on a relatively large quantity of speech data. The proposed GMM/UBM extension addresses the situations with limited resources and takes advantage from the temporal structure of speech by using client-customised utterances harnessed by a Markov model. New temporal information is then used to enhance discrimination with Viterbi decoding by increasing the gap between client and impostor tests.
Experiments on the MyIdea database are performed when impostors know the client-utterance and also when they do not, highlighting the potential of this new approach. A relative gain up to 64% in terms of EER is achieved when impostors do not know the client utterances and performance is equivalent to the GMM/UBM baseline system in other configurations.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Biometrics

Keywords

Speaker Recognition, Viterbi decoding, Embedded system, Password recognition

1. INTRODUCTION

Classical speaker recognition engines, working on text-independent inputs, allow a high level of performance as shown during NIST evaluation [11]. In a realistic application, the efficiency of these systems could be influenced by several constraints. An embedded application for example often has computational and memory limitations as well as ergonomic aspects; these can mean small quantities of model data and short test audio sequences. State-of-the-art speaker recogntion systems, which often follow the GMM/UBM paradigm [2] depend strongly on the quantity of training data available to enrol a speaker and usually requires a few minutes of speech to achieve good performance [5].
A positive contribution to this problem is to increase the amount of information taken into account by the system by including text dependencies, like in a user-customised password scenario [1]. In this case, the Temporal Structure Information (TSI) gathered from the password helps to compensate for the lack of data due to short duration of audio sequences. In order to model the TSI of speech while achieving statistical modelling, a word recognition system could be combined with a speaker recognition system [7].
An embedded system is often confronted with strongly variable inputs and implementation constraints. Due to this constraint, the acoustic modelling used in the recognition has to be adapted to the environment. Hidden Markov Model (HMM) modelling does not seem well suited as full model adaptation requires a large amount of data and a significant resource consumption. Furthermore the conventional HMM/Viterbi approach does not seem to fit the password-based speaker recognition task. Indeed, the information gathered from the user-customised passwords aims to discriminate clients who pronounce their own password from impostors i) the correct client password, or ii) an unknown utterance. Here we consider particularly the latter case. The Viterbi decoding provides an optimal alignment for client but also impostor test sequences and the maximisation of the likelihood computed for impostor test sequences could lead to a drop in performance. A way to improve discrimination between client and impostor tests is to add a priori information on the password temporal structure. Including such information within the Viterbi decoding aims to degrade scores for impostor accesses.
In this paper, two approaches are associated in an architecture able to deal with the application context constraints. The first is the standard GMM/UBM paradigm, well known for the general acoustic space modelling and its text-independent speaker recognition abilities. The second one is an HMM/Viterbi approach used to take advantages of the text-dependent and TSI aspects by using a Semi-Continuous HMM (SCHMM) [12]. Such a combined system was originally proposed in [3] for speaker recognition and extended to word recognition in [10].
A key point here is the inclusion of an external, additional synchronisation process during testing. The process is a Viterbi decoding and aims to enhance discrimination by using the temporal structure of the pronounced utterance. The aim is to reduce score values when the speaker does not know the client password, by making the Viterbi path sub-optimal. Ideally, the separate information source should be complementary and external to the SCHMM TSI, such as video synchronisation. In this work, we use a word-based synchronisation from an automatic alignment process. Preliminary results of this approach were shown in [8] and [9]; this paper focus on the

constrained Viterbi decoding with more accuracy and shows more precise results.
The specific acoustic architecture is described in Section 2. The constrained Viterbi algorithm is described in Section 3. Section 4 describes the experimental protocol and results. It includes a description of the MyIdea database. Section 5 summarises the benefits of this approach and presents future work direction.

2. A THREE LEVEL ACOUSTIC ARCHITECTURE

The proposed three level architecture shown in Figure 1 is called EBD for Embedded *LIA_SpkDet* and is an extension of the well known GMM/UBM paradigm configured to deal with the user-customised speaker recognition task.

GMM/UBM the architecture of the two first layers of the EBD is similar to a classical GMM/UBM speaker recognition system. A text-independent model of every client is trained by adapting the UBM. This adaptation is described below.

User-customised extension the previous text-independent speaker model is then used to obtain a SCHMM with the goal of harnessing the TSI of the utterance chosen by this speaker. Each state of the SCHMM is trained from a part of that utterance using an iterative Viterbi decoding process. During the test, classical Viterbi decoding is again performed with this SCHMM. Details of the training and test are given below.

Two scores are computed, the first is obtained with only the GMM/UBM modelling and the second is computed with the SCHMM model. These scores are combined to provide a final score for the decision stage.

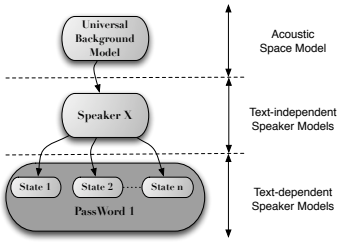

Figure 1: General view of the *EBD* model architecture.

2.1 Training step

The EBD model is trained in three steps, each corresponding to one level of the architecture. The UBM is firstly trained to model the acoustic speech space. It is built off line using a suitably large amount of representative data. It is trained with a classical EM/ML algorithm [4]. The training of the speaker text-independent models consists in adapting the UBM/GMM with the available data pronounced by the client. An energy labelling is performed on the signal and only the frames deemed to be speech are kept. The model is obtained by adapting the UBM using the EM algorithm with the MAP criterion [6].
The third level is to initialise an utterance SCHMM model; the utterance sequence is cut into S segments $\{seg_i\}$ of the same length. Each state i of the SCHMM is adapted from the speaker text-independent model using the speech-labelled frames of seg_i. An EM/MAP algorithm is applied on the weight parameters. Then the SCHMM is optimised using a classical Viterbi algorithm (a new segmentation is achieved by Viterbi and is used to adapt the state models). The number of states of the SCHMM is experimentally determined. Finally, the transition probabilities of the SCHMM are computed using the relative length of each segment.

2.2 Testing step

During a test, the score between the input signal and an utterance SCHMM model is derived from the corresponding Viterbi path. As for the training, the log-likelihood for an input frame is only computed for each Gaussian component of the text-independent model. The system is computational efficient since scores from the original GMM/UBM are used throughout, with only the weightings changed according to state occupancy of the SCHMM.
This scoring process is equivalent in terms of computation to a classical GMM/UBM system producing two scores. The first is obtained with only the text-independent speaker model and the second is computed with the SCHMM model, which itself has two operational modes, namely without further constraints or constrained by the external information as described in Section 3. These two scores are normalised using the log-likelihood of the UBM. They are then combined to give a final score for the decision stage. An empirically-tuned weighted linear combination is used.

3. THE CONSTRAINED VITERBI

Synchronisation points are generated from an external source during both the training and testing phase. These points are used to strongly constrain the Viterbi decoding. This constraint is obtained by allowing or forbidding transitions of the SCHMM corresponding to the synchronisation points (labelled S in Figure 2). In this case, the bottom layer of the EBD system could be compared to a succession of sub-SCHMM. The Viterbi algorithm is then processed from one synchronisation point until the next with the corresponding sub-SCHMM.

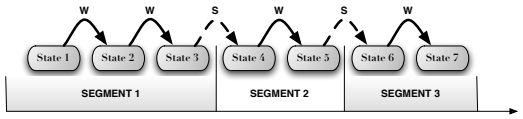

Figure 2: Use of an External Segmentation in the bottom layer of the EBD system to constraint the Viterbi decoding and increase the discriminative power of the EBD approach.

Two types of consequences could be expected from the proposed contraining procedure. The addition of complementary information could improve the training of SCHMM models. Such an improvement could potentially lead to increase client scores during the testing phase.
The second effect, mainly expected when constraining the Viterbi decoding consists in increasing the discrimination power of the EBD approach. This effect is described below.
The external constraint is illustrated on Figure 3. The Viterbi path computed under this constraint is not allowed to pass through the blackened states area. Free zones correspond to the intersection of the synchronisation computed on both the training and the testing utterances. Figure 3 shows paths obtained on a same model for two different password occurrences:

- the occurrence used to train this SCHMM model;
- one occurrence of the same password, pronounced by the client.

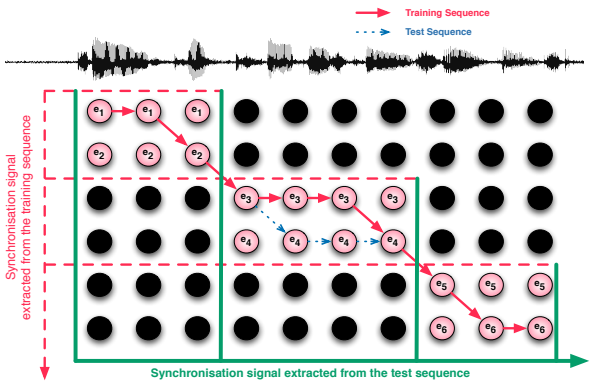

Figure 3: Alignment for client training and test utterances with an external synchronisation.

The temporal structure of a client test utterance is supposed to be close to the temporal structure of the training utterance. With this hypothesis the path computed for the client test utterance goes through the non-forbidden area only. This is illustrated by Figure 3. In this case, the synchronisation constraint has no effect on the Viterbi decoding and the resulting score remains high.

Figure 4 shows two expected alignment path for a given impostor test utterance, one with the synchronisation constraint (the more central one) with or without the synchronisation constraint. The temporal structure of an impostor utterance is assumed to be strongly different from the client password structure. The path resulting from the alignment of this utterance on the client-password SCHMM without applying the external constraint is assumed to go through the forbidden area. Thus, the synchronisation constraint forces the algorithm to find a path through the authorised area. This path is not optimal and the impostor score is therefor lower than the one computed without any constraint.

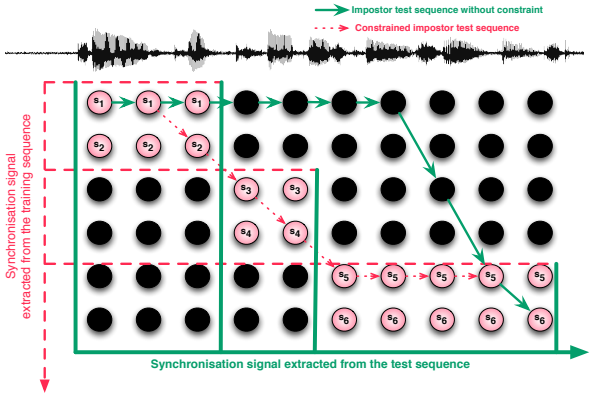

Figure 4: Alignment for impostor utterance with and without external synchronisation.

4. EXPERIMENTS

4.1 The MyIdea database

Experiments are performed on the BIOMET part of the MyIdea database. This database contains audio-video records from 30 male speakers. In this subpart of MyIdea, 25 sentences are recorded in 3 sessions for each speaker. Twelve of these sentences are the same for all the speakers, ten short (about 3 seconds) and 2 long sentences (about 6 seconds). Other occurrences are speaker or session dependent. The three sessions are recorded under controlled acoustic and illumination conditions. MyIdea presents several drawbacks for our work. The recordings are not made in a real environment. The sentence duration variability is limited (2 or 3 seconds for the utterance occurrences), the sentences are too long for a real password dedicated system, and the number of speakers is small for a speaker verification experiment. However it use is reported here with future work aimed at incorporating visual information.

4.2 Experimental protocol

The 30 male speakers are separated into two groups -A and B- each with 15 speakers. Each group is successively used as the Client-set with the others used to train the UBM. The UBM is trained using the whole recorded material of the 15 speakers of the UBM-set.
When using the A-group data set to train the UBM model, the speakers from the B-group are used for enrolment and tests. Due to the small number of speakers available, a jacknifing process is used by training a client model for each available speaker session. Each of the 15 speakers of this Client-set is successively considered as a client for which the 14 other speakers of the Client-set are impostors.
Each client text-independent GMM model is derived from the UBM by using two long sentences and one occurrence of the selected short sentence (around 8 seconds of speech). The utterance-dependent model is trained with the same short sentence occurrence (around 2 seconds of speech). With the jacknifing process, 900 utterance models are trained (10 short sentences, 3 sessions and 30 clients).
The short sentences not used for utterance training are compared to the client model. 1,800 client tests are performed (2 test occurrences for each of the 900 utterances).
Three configurations of impostor tests are proposed. The speaker and utterance models are compared to the 14 impostors who are the remaining speakers of the same group.

UNKNOWN configuration the linguistic content of the impostor test occurrences is different from the training material of client models. Each speaker model is compared to three randomly selected short sentences (one per session) out of the 9 remaining sentences of each of the 14 impostor speakers.

KNOWN configuration the linguistic content of the impostor test sequences is the same as the occurrences used to train the client models. Each utterance model is compared to three randomly selected sentences from each of the 14 other speakers of the Client-set.

ALL configuration the impostor tests are all tests from both the KNOWN and the UNKNOWN configurations.

For both KOWN and UNKNOWN configurations, the number of impostor tests is constant for a given client. Moreover the global number of impostor tests is 37, 800 in those two configurations and 75, 600 in the ALL configuration.

4.3 System configuration

Mel-scaled frequency cepstral coefficients (MFCC) are used, computed every 10ms. An energy labelling is applied to separate the speech frames from the non-speech frames. Acoustic feature frames are 32-dimension vectors, 15 cepstral coefficients, the log-energy and the corresponding Δ coefficients.

In the experimental configuration, the number of components in GMMs is fixed to 256 and all mean parameters of the text-independent speaker models are adapted. Only the 32 most significant weights parameters are adapted for each SCHMM state.

4.4 Results

Experiments are conducted to asses the contributions coming from the three components, GMM/UBM, SCHMM and the constrained Viterbi. The GMM/UBM is regarded as the baseline. Benefits of the SCHMM extension are predicted to come from the TSI and the constrained decoding. The experimental results are presented in Table 1.

Configuration	GMM baseline	EBD system	
		Free	Constrained
UNKNOWN	2.46	1.11	**0.89**
KNOWN	4.00	4.06	**4.07**
ALL	3.22	2.83	**2.83**

Table 1: EER of GMM compared to the EBD system with 20 states per SCHMM, when constraining or not the Viterbi decoding

The first column of Table 1 shows the results of the GMM baseline. Error rates fall from 4.00% to 3.22% and finally $2.46\%$for the KNOWN, ALL and UNKNOWN conditions respectively. These results show that the text-dependency has a real influence on the GMM/UBM performances. Decreasing the variability between the phonetic contents of the training and the test material allows successive improvements in reducing EER.

The main expected advantage of the EBD system using SCHMM compared to a classical GMM/UBM is to incorporate the password-based information like the password itself and the relative TSI. This point is evaluated by the experiments presented in the second column of Table 1. A classical Viterbi algorithm is used in this experiment. Error rates fall from 2.46% to 1.11% by using SCHMM when impostors do not know the client passwords. This result is not confirmed when the impostors know the client utterances (KNOWN) and the performance of the EBD and GMM/UBM systems are equivalent.

In order to evaluate the effect of an external synchronisation, a new experiment is performed. Results of this experiment are presented in the third column of Table 1. As expected, performance of the EBD in the UNKNOWN condition improves when constraining the Viterbi decoding with an external segmentation. In this condition, impostors do not know the client password and pronounce a different utterance whose temporal structure is penalised by the synchronisation constraint.

A further analyse of the evolution of the client and impostor scores is presented in the following paragraphs.

Figure 5 shows the evolution of client score distributions with and without constraining the Viterbi decoding. Distributions of client scores remain equivalent with or without the external information. This additional information seems not to improve the quality of SCHMM models in terms of likelihood with test data. This observation could be explained by the nature of the additional information coming from a phonetical alignment. This information could be over correlated with the acoustic information already exploited by the SCHMM extension. At the same time, these results show that the Viterbi algorithm is still optimal when dealing with client test utterances and does not degrade client scores.

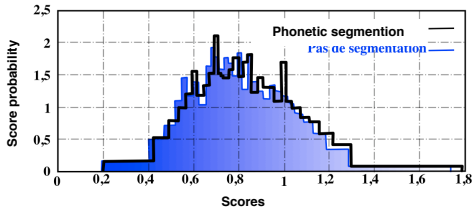

Figure 5: Evolution of the client text-dependent score distributions with or without external synchronisation constraint.

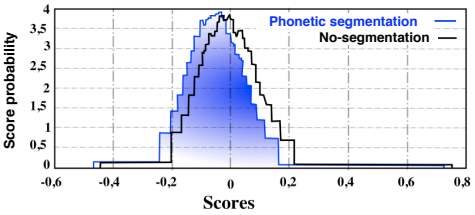

Figure 6: Evolution of the impostors text-dependent score distributions with or without external synchronisation constraint.

The integration of an external source of information in the Viterbi decoding is predicted to degrade the impostor scores.
Figure 6 confirms that impostor scores decrease significantly when adding an external source of information. Furthermore, Figure 7 shows that the effect of the Viterbi constraint affects a huge majority of impostor tests. This observation shows that the synchronisation constraint allows the Viterbi algorithm to provide sub-optimal alignment when the temporal structure of the test utterance is different from the training utterance structure.

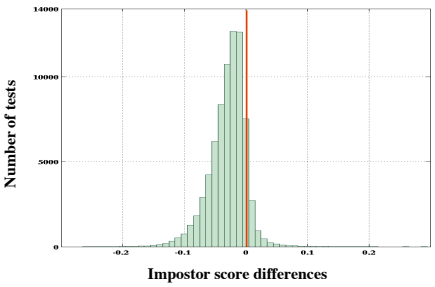

Figure 7: Evolution of the impostor text-dependent scores distributions with or without external synchronisation constraint.

The same experiment performed with impostors knowing the client password (Fig. 8) shows that the effect of the external information is less significant in this configuration due to the similar temporal structure of the utterance. This effect is not sufficient to discriminate speakers by the way they pronounce a lexical content.

5. CONCLUSIONS AND FUTURE WORKS

The approach proposed in this paper is designed for embedded applications. It takes advantages from a GMM/UBM text-independent approach and the HMM/Viterbi speech-recognition power.

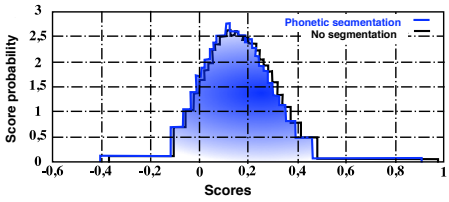

Figure 8: Evolution of the of impostors text-dependent scores distributions with or without external synchronisation constraint.

In addition we propose a version of the Viterbi algorithm constrained by an external source synchronisation. The use of this temporal information in a speaker recognition system leads to improved performance by reducing impostor scores.
Performances of our approach is equivalent to the GMM/UBM baseline system when not considering the linguistic content (example of EER in KNOWN condition, GMM: 4.00, EBD 4.07) whereas the proposed approach outperforms the GMM/UBM when impostors do not know the client utterance (EER in UNKNOWN condition, GMM: 2.46, EBD: 1.11). Furthermore, the external synchronisation allows an additional gain when impostors do not know the client utterance (EER in UNKNOWN condition, EBD without constraint: 1.11, constrained-EBD: 0.89).

Future work will focus on the multi-modality by substituting the phonetic segmentation with temporal information extracted from the video stream. By incorporating this strong constraint in the training and testing phases we aim at increasing the performance and to thwart replay attacks. The EBD approach will be tuned to better balance the speaker and utterance specific information in order to outperform the baseline GMM in every condition. Moreover, given that the first results show the ability of the EBD approach to take advantage of the linguistic content of customised utterances, more tests are to be performed to evaluate the performance of the EBD system with more utterance-variability, for example considering the utterance duration.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] M. F. BenZeghiba and H. Bourlard. User-customized password speaker verification using multiple reference and background models. *Speech Communication*, 48(9):1200–1213, 2006.

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meigner, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds. A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Applied Signal Processing*, 4:430–451, April 2004.

[3] J.-F. Bonastre, P. Morin, and J.-C. Junqua. Gaussian Dynamic Warping (GDW) Method Applied to Text-Dependent Speaker Detection and Verification. In *European Conference on Speech Communication and Technology (Eurospeech)*, Geneva (Switzerland), 2003.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[5] B. Fauve, N. Evans, and J. S. Mason. Improving the performance of text-independent short duration SVM-and GMM-based speaker verification. In *Odyssey Conference - The Speaker and Language Recognition Workshop*, 2008.

[6] J.-L. Gauvain and C.-H. Lee. Maximum a Posteriori estimation for Multivariate Gaussian Mixture Observations of Markov Chains. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 2, pages 291–298, Adelaide (Australia), April 1994.

[7] M. Hebert and L. P. Heck. Phonetic class-based speaker verification. In *European Conference on Speech Communication and Technology (Eurospeech)*, pages 1665–1668, Geneva, September 2003.

[8] A. Larcher, J.-F. Bonastre, and J. S. D. Mason. Reinforced temporal structure information for embedded utterance-based. In *International Conference on Speech Communication and Technology (Interspeech)*, Brisbane (Australia), 2008.

[9] A. Larcher, J.-F. Bonastre, and J. S. D. Mason. Short utterance-based video aided speaker recognition. In *IEEE International workshop on Multimedia Signal Processing (MMSP)*, Cairns (Australia), october 2008.

[10] C. Lévy, G. Linares, P. Nocera, and J.-F. Bonastre. *Mobile Phone Embedded Digit-Recognition*, chapter 7 in Digital Signal Processing for In-Vehicle and Mobile Systems 2, pages 71–84. Springer Sciences, 2006.

[11] M. A. Przybocki, A. F. Martin, and A. N. Le. NIST speaker recognition evaluations utilizing the mixer corpora - 2004, 2005, 2006. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7):1951–1959, 2007.

[12] S. J. Young. The general use of tying in phoneme-based HMM speech recognisers. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 569–572, San Francisco (USA), 1992.