Reinforced Temporal Structure Information For Embedded Utterance-Based Speaker Recognition

Anthony Larcher[1,2]*, Jean-François Bonastre*[1]*, John S.D. Mason*[2]

[1]Laboratoire d'Informatique d'Avignon (LIA), UAPV, France
[2]Speech and Image group, Swansea University, Wales, UK
{anthony.larcher, jean-francois.bonastre}@univ-avignon.fr, 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. Even if they have proved their efficiency in more classical contexts, GMM/UBM based systems show their limits in such situations, with good accuracy demanding a relatively large quantity of speech data, but with negligible harnessing of linguistic content. The proposed approach addresses these limitations and takes advantage from the linguistic nature of the speech material into the GMM/UBM framework by using client-customised utterances. The GMM/UBM is then reinforced with new temporal information.
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 45% in terms of EER is achieved when impostors do not know the client utterance and performance is equivalent to the GMM/UBM baseline system in other configurations.

1. INTRODUCTION

The efficiency of speaker recognition systems in a realistic application could be affected by several constraints. For example, an application which should be immediately usable will strongly limit the enrolment material and hence could lead to poor recognition accuracy. Since limited data can greatly reduce performance, ergonomic constraints could also impose short test sequences influencing similarly performance. Some other constraints can be the memory and computational resource limitation or the use in variable environments. Embedded systems might well present these conditions.
State-of-the-art speaker recognition engines work on text-independent inputs and follow often the GMM/UBM (Gaussian Mixture Model/ Universal Background Model) paradigm [1]. This solution gives a high level of performance as shown during the NIST evaluations [2]. Unfortunately, the GMM/UBM depends strongly on the quantity of training data available to enrol a speaker in contrast to the context considered here which involves relatively short duration speech material. A solution to this problem is to increase the amount of information taken into account by the system by including text dependencies, like in a user-customized utterance scenario. In this case, the Temporal Structure Information (TSI) gathered from the utterance can help to compensate the short duration of the 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 [3], [4].
To satisfy targeted application constraints, the system should accept all kinds of utterances, especially short duration ones, and also be language-independent. Adding language options when using a phoneme-based word recognition system would seem viable with an appropriate choice of phonemes covering languages. However, this solution could be expensive in terms of storage and computational cost.
Furthermore, an embedded system could be confronted with strongly variable environments. Due to this constraint the acoustic modelling used in the recognition system has to be adapted to the environment and the computational cost of the adaptation has to follow the targeted context resource constraints. HMM adaptation does not seem well suited as it normally requires a relatively large amount of training data.
The solution proposed in this paper tries to associate the well known advantages of a GMM based statistical acoustic model with an original architecture able to deal with the application context constraints and to incorporate external temporal information. It uses the GMM/UBM paradigm for the general acoustic space modelling and its text-independent speaker recognition capabilities. It also involves an HMM/Viterbi approach in order to incorporate the text-dependent and TSI aspects using a Semi-Continuous HMM (SCHMM). Such a combined system was originally proposed in [5] for speaker recognition and extended to word recognition in [6].
In order to reinforce the relaxed synchronisation between states and frames due to the SCHMM structure of the TSI modelling, we propose to add yet further time constraints from a separate synchronisation process. Ideally, the separate synchronisation source should be complementary and external to the SCHMM TSI, such as video synchronisation. In this first work, in order to validate the hypotheses, we use a word-based synchronisation from an automatic alignment process.
The specific three stage architecture and the enrolment algorithms are described in Section 2 as well as the way of reinforcing the TSI with an external synchronisation. The experimental protocol and results are described in Section 3 including a description of the MyIdea database. An estimation of the memory and computational costs is also described in this part. Section 4 summarises the benefits of this approach and presents different future work directions.

2. DESCRIPTION OF THE APPROACH

The proposed approaches, called EBD for Embedded *LIA_SpkDET* takes advantages of three mechanisms which contribute to the overall scoring process.

GMM/UBM the two first layers of the EBD architecture are equivalent to a classical GMM/UBM speaker recognition system. One GMM model is trained to give a text-independent acoustic model of each client. The GMM

adaptation is described below.

SCHMM the previous text-independent speaker model is used to obtain an SCHMM with the goal of harnessing the TSI of the utterance chosen by this speaker. Using an iterative Viterbi decoding process, each state of the SCHMM is trained from a part of that utterance. During the test, Viterbi decoding is again performed with this SCHMM. Details of the train and tests are given below.

External information the goal here is to use further information to assist in the overall verification task by adding additional constraining components. These constraints are computed during the training phase and used to constrain both the training and test Viterbi decoding. As the external information is only computed during the train, it could be computed off line. This information is here labelled external to reflect that it is aimed to come from video in further works. However, assuming the minor constraint that the client has to speak relatively distinctly, a phonetical alignment could also generate such a synchronisation as investigated here.

Two scores are computed, the first is obtained with only the GMM/UBM modelling and the second is computed with the SCHMM model which could be constrained or not by the external information.These scores are combined to give a final score for the decision stage. An empirically-tuned linear combination is used.

2.1. EBD Hierarchical Architecture

Figure 1 shows the architecture of the EBD system. Nodes of this structure are GMM models. The upper layer is the least specialised one. It is a classical UBM which aims at modelling the acoustic space.
The middle layer contains speaker specific text-independent models. These GMMs are obtained by a classical GMM/UBM adaptation process. Each model is derived from the UBM by using the EM algorithm and following the *Maximum A Posteriori* (MAP) criterion. Only the mean parameters are adapted and other parameters are shared with the UBM.
The bottom layer uses the ability of the left-right SCHMM to capture the TSI of the user-customised utterances. Each of the SCHMM states is a GMM derived from the corresponding middle level model. As explained below, the transformation function works only on the weights of the GMMs, the other parameters are taken from the middle layer model.

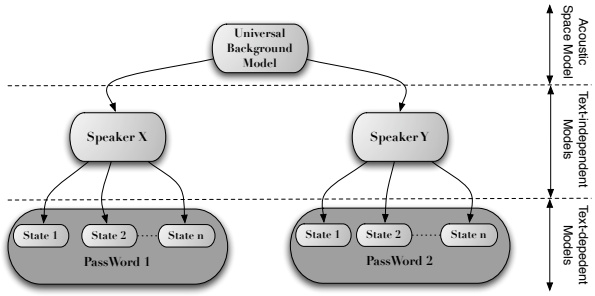

Figure 1: General view of the *EBD* model architecture.

2.2. 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 largest part of the acoustic space. It is built off line using a suitably large amount of representative data. It is trained with a classical EM/ML algorithm. The training of the text-independent speaker models consists in adapting the GMM/UBM 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.
In order 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 speaker model using the speech-labelled frames of seg_i. An EM/MAP algorithm is applied on the weight parameters. As all parameters except weights are tied between the states of the SCHMM and the text-independent model, the log-likelihood for an input frame is only computed for each Gaussian component of the text-independent model. Then the log-likelihood of this frame with each state of the SCHMM involves a weighted sum which is negligible compared to the full log-likelihood computation. 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.

2.3. Testing Step

During a test, the score between the input signal and an utterance SCHMM model is derived from the corresponding Viterbi path. All the input frames are used during this Viterbi decoding phase. The log-likelihood of a frame sequence could be expressed as a sum of two log-likelihood accumulations, one using speech-labelled frames and the other using non-speech-labelled frames, as shown in Equation 1.

$$logp(X|\lambda) = logp(X_{speech}|\lambda) + logp(X_{non-Speech}|\lambda) \quad (1)$$

The final speaker matching score corresponds to the log-likelihood computed with the $logp(X_{speech}|\lambda)$ only. This score is normalised by the log-likelihood computed using the UBM model on the same speech-labelled frames.

2.4. External Synchronisation

Synchronisation points are generated off line from an external source during the training phase. These points are used during both the training and test phase 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 a synchronisation point until the next with the corresponding sub-SCHMM.

3. EXPERIMENTS

3.1. 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

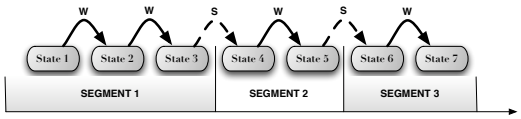

Figure 2: Use of an External Segmentation in the bottom layer of the EBD system to constraint the Viterbi decoding and reinforce the TSI in the utterance SCHMMs.

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.

3.2. 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. Two conditions are defined:

1-occ in this condition, 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).

2-occ this condition is the same as above except that an additional occurrence of the selected short sentence (which is the chosen utterance) is used to train both the text-independent and the utterance models. The number of utterance models is still the same as above.

The number of target trials is condition-dependent. The short sentences not used for utterance training are compared to the client model. 1,800 client tests are performed in the *1-occ* condition (2 test occurrences for each of the 900 utterances) while 900 client tests are performed in the *2-occ* condition (1 test occurrence 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.

3.3. 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 128 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.

3.4. Results

Experiments are conducted to assess the contributions coming from the three components, GMM/UBM, SCHMM and external information. The GMM is regarded as the baseline and the benefits of the other two are predicted to come from TSI. The experimental results presented in Table 1 show performance of the EBD system depending on the number of states, the nature of the impostor tests and the quantity of training data. No external synchronisation is used in these experiments. Performance of the baseline GMM in the same conditions is provided for comparison. It is important to note that the GMM system reflects the middle layer of the EBD system. The first two rows in Table 1 show the results when the impostors do not know the speaker utterances. Error rates fall from 2.72 down to 2.07 and 0.87 to 0.56 for the 1-occ and 2-occ respectively when SCHMM is used. Increasing the number of states up to 15 allows successive improvements in reducing EER. This result is not confirmed when the impostors know the client utterances (KNOWN). A loss of performance is observed when the number of states increases. It seems that, in speaker matching scores, the TSI is dominated by the utterance content information rather than the speaker specific information, *i.e.* the system recognises the utterance instead of the speaker. This results is confirmed by an experiment similar to the previous one but with two occurrences of the speaker utterance used during the training phase which is also presented in Table 1. Indeed, the GMM system performs better than the EBD and seems to gain more from the increase in training data than the EBD except in the UNKNOWN condition where the increase of training data also increases the advantage of the EBD by reinforcing the TSI exploited by the SCHMM models.

In order to evaluate the effect of an external synchronisation, a new experiment is performed. This external information which is here in a lexical constraint is computed using the *LIA SPEERAL Toolkit*. Results of this experiment are presented in Table 2. The first column is the original GMM baseline and the next two columns give the EBD results as above in Table 1 for the 15 state case, and then with the use of the external information. As expected, performance of the EBD improves when using an external segmentation. This is particularly highlighted in *1-occ* conditions wheres the system posses less speaker spe-

Configuration		GMM baseline	Number of states of the EBD		
			5	10	15
UNKNOWN	1-occ	2.72	2.39	2.22	**2.07**
UNKNOWN	2-occ	0.87	0.67	0.56	**0.56**
KNOWN	1-occ	**4.49**	4.56	4.56	4.61
KNOWN	2-occ	**2.22**	2.64	2.78	2.89
ALL	1-occ	3.72	3.62	**3.61**	3.66
ALL	2-occ	**1.58**	1.91	1.90	2.09

Table 1: EER of GMM and EBD systems (with different number of states) using one or two training occurrences in different test configuration

cific data. Moreover, the EBD using an external synchronisation gains more from the increase of training data than the same system without this temporal information.

Configuration		GMM baseline	External Information Condition	
			None	Active
UNKNOWN	1-occ	2.72	2.07	**1.89**
UNKNOWN	2-occ	0.87	0.56	**0.47**
KNOWN	1-occ	4.49	4.61	**4.39**
KNOWN	2-occ	**2.22**	2.89	2.78
ALL	1-occ	3.72	3.66	**3.33**
ALL	2-occ	**1.58**	2.09	1.77

Table 2: EER of GMM compared to the EBD system in 15 state condition, when using or not an external segmentation

3.5. Resource Constraints

An embedded system is likely to be confronted by resource constraints. In order to illustrate the capability of the EBD to deal with limited resources, we compare two systems in terms of memory constraint and computational cost. As the EBD system lies at the border between speaker and isolated-word recognition systems it could be compared to two systems working in parallel. State-of-the-art isolated-word recognition systems are based on HMM and phonemic models. In this sense the two approaches compared to the EBD system are: a phonemic based approach using non-contexual models and a global HMM for each speaker-customised-utterance. The chosen speaker recognition approach is a classical GMM/UBM system.
Table 3 shows a rough approximation of the memory and computational resources of these systems. These results are obtained with the following assumptions:

- 5 speakers are enrolled;
- 2 utterance models per speaker;
- 32 acoustic parameters per feature;
- 128 Gaussian distribution per GMM;
- 15 state-per-utterance-models for the EBD and HMM-based system;
- the acoustic phonemic-model contains 108 emitting states;

This first approximation shows that the structure of the EBD system needs less resources than the compared approaches.

4. 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. In addition we propose to reinforce the TSI modelling by an external source of synchronisation. The use

	Number of parameters	Number of log-likelihood computations
GMM baseline + Phonemes based system	927, 000	442, 000
GMM baseline + Utterance specific HMM	1, 276, 000	614, 000
EBD	**33, 000**	**25, 000**
GMM baseline	29, 000	25, 000

Table 3: Approximation of the memory (in terms of parameters) and computational cost (in terms of log-likelihood computation) of EBD and three baseline systems per input frame.

of the temporal information in a speaker recognition system allows to improve performance particularly when a relatively small quantity of speech data is available.
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: 4.49, EBD 4.61) whereas the proposed approach outperforms the GMM/UBM when impostors do not know the client utterance (EER in UNKNOWN condition, GMM: 0.87, EBD: 0.56). Furthermore, the external synchronisation allows a gain in all six situations.
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, as the first results show the ability of the EBD approach to take advantage of the linguistic content of customised utterances, more tests have to be performed to evaluate the performance of the EBD system with more utterance-variability, for example considering the utterance duration.

5. Acknowledgment

BioBiMo (Biométrie Bimodale sur Mobile) is a project supported by the French A.N.R. (Agence Nationale pour la Recherche). http://biobimo.eurecom.fr/

6. References

[1] 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*, vol. 4, pp. 430–451, April 2004.

[2] 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*, vol. 15, no. 7, pp. 1951–1959, 2007.

[3] M. Hebert and L. P. Heck, "Phonetic class-based speaker verification," in *European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, 2003, pp. 1665–1668.

[4] J. Navratil, U. V. Chaudhari, and S. H. Maes, "A speech biometrics system with multigrained speaker modeling," in *Conference for Natural Speech Processing*, 2000.

[5] 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.

[6] C. Lévy, G. Linares, P. Nocera, and J.-F. Bonastre, *Mobile Phone Embedded Digit-Recognition*. Springer Sciences, 2006, ch. 7 in Digital Signal Processing for In-Vehicle and Mobile Systems 2. [Online]. Available: http://www.lia.univ-avignon.fr/php/publications2.php?page=5&selection=auteur&tableau2=levy