Short Utterance-based Video Aided Speaker Recognition

Anthony Larcher [1], Jean-Francois Bonastre [2], John S.D. Mason [3]

University of Avignon, LIA
339 ch des Meinajaries, BP1228, 84 911 Avignon, FRANCE
[1] anthony.larcher@univ-avignon.fr
[2] jean-francois.bonastre@univ-avignon.fr

Speech and Image Group, Swansea University
Singleton Park Swansea SA2 8PP UK
[3] 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 of the linguistic nature of the speech material into the GMM/UBM framework by using client-customised utterances. Furthermore, the acoustic structure is then reinforced with video 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 47% 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.

I. INTRODUCTION

The efficiency of speaker recognition systems in a realistic application could be influenced 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. Given that limited data can greatly reduce the recognition 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 tend to be assessed on text-independent inputs and follow often the GMM/UBM (Gaussian Mixture Model/ Universal Background Model) paradigm [**?**]. This solution gives a high level of performance as shown during the NIST evaluations [**?**]. Unfortunately, the UBM/GMM 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-customised utterance scenario [**?**]. In this case, the Temporal Structure Information (TSI) gathered from the utterance can help to compensate for 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[**?**], [**?**].

To satisfy application ergonomic constraints and to allow the speaker to chose is own customised-utterance, 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 the 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 [**?**] for speaker recognition and extended to word recognition in [**?**].

Two mains approaches are possible to take into account the bimodal aspect of speech (audio-video) for speaker recognition. Generally, this problem is view as a fusion process between the two modalities. An early fusion at the data level is difficult due to the different nature of the parameters and their asynchronism. Several works were proposed, mainly in speech recognition [**?**] [**?**] and show a performance improvement only when noisy audio data are used. The fusion at the score level is more often proposed due to its simplicity [**?**], but such a fusion process does not take advantage from he temporal joint-

information and it is still costly in terms of computational resources (separate systems are needed). Finally, an interesting alternative to a fusion process consists in a joint decoding processing of both modalities. However the asynchronism aspect of audio-video modalities leads to complex algorithms like in [**?**] and [**?**].
In this paper we propose a such joint-decoding which, thanks to the specific aspects of our system, shows an acceptable level of complexity. In order to reinforce the relaxed synchronisation between states and frames due to the SCHMM structure of the TSI modelling, we propose to embed video information during the audio decoding by adding further time-constraints gathered from a video synchronisation process.
The specific three stage architecture and the enrolment algorithms are described in Section **??** as well as the way of reinforcing the TSI with an video-learnt synchronisation. The experimental protocol and results are described in Section **??** including a description of the audio-video database, MyIdea. Section **??** summarizes the benefits of this approach and presents possible future work directions.

II. DESCRIPTION OF THE APPROACH

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

- **GMM/UBM**, the architecture of the two first layers of the EBD is similar to a classical GMM/UBM speaker recognition system. Speaker text-independent model of every client is trained by adapting the Universal Background Model. This adaptation is described below.
- **SCHMM**, the previous text-independent speaker model is then used to obtain an 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, Viterbi decoding is again performed with this SCHMM. Details of the training and test are given below.
- **Video-Learnt 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 video information is only computed during the train, it could be computed off line. This information is here labelled "video-learnt" to reflect that it is only computed during the training phase.

A. EBD Hierarchical Architecture

Figure **??** 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.

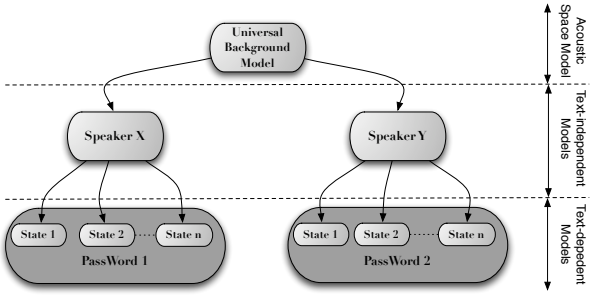

Fig. 1. General view of the *EBD* model architecture.

B. 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 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.
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 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.
One advantage of this process is that all parameters except weights are tied between the states of the SCHMM and the text-independent speaker 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 only involves a weighted sum which is negligible compared to the full log-likelihood computation.

C. 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 **??**.

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

The final speaker matching score corresponds to the log-likelihood computed with the $log\, p(X_{speech}|\lambda)$ only.
As for the training, the log-likelihood for an input frame is only computed for each Gaussian component of the text-independent model. The computation of the log-likelihood of this frame with each state of the SCHMM which involves a linear combination is negligible. This scoring process equivalent in terms of computation to a classical GMM/UBM system produces 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 video-learnt information from the initial training phase. 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.

D. Video-Learnt Synchronisation

Synchronisation points are extracted from a very simple video process. The video stream is first pre-processed to obtain a black-and-white sequence which is the Y component of the sequence resulting of an RGB to YCbCr transformation.
A mono-dimensional temporal signal is issued from this black-and-white video stream in order to estimate the quantity of change between successive frames. Subtractions are processed between the pixels of one image and thus of the following one. The absolute values of pixel subtractions are summed to obtain a value of the discrete temporal signal S. This computational process is described by Equation **??**.

$$S(n) = \sum_{w=0}^{W} \sum_{h=0}^{H} abs(I^n_{(w,h)} - I^{n+1}_{(w,h)}) \quad (2)$$

where W and H are respectively the width and the height of the video images, I^n and I^{n+1} are two consecutive images of the video stream and $I^n_{(w,h)}$ is the value of the pixel (w, h) of the image I^n. $S(n)$ is the discrete temporal signal from which the synchronisation points are extracted.
Local maxima of the signal S are found by applying a sliding window algorithm. These local maxima are stored to become the video-learnt synchronisation points.
These points, which could be generated off-line during the training phase only, are subsequently used during both the training and test phase to strongly constrain the Viterbi decoding. This result is obtained by allowing or forbidding transitions of the SCHMM corresponding to the synchronisation points and labelled V, for video, in Figure **??**. Other transitions, labelled A are computed from the Audio and not modified when adding the video-learnt segmentation. The number of synchronisation points depends on both the speaker and utterance he pronounced.

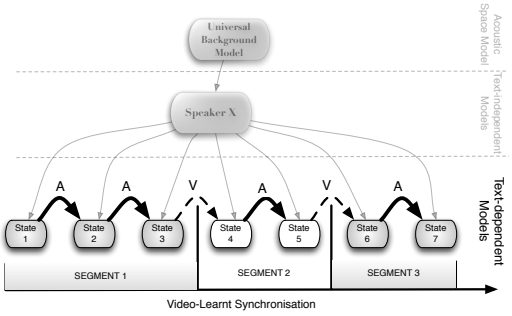

Fig. 2. Use of an Video-Learnt Segmentation in the bottom layer of the EBD system to constrain the Viterbi decoding and reinforce the TSI in the utterance SCHMMs. The V labelled transitions are constrained by the video-learnt segmentation as the A labelled ones still unchanged

III. EXPERIMENTS

A. 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. However, 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.

B. 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 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. 37,800 impostor tests are performed in this configuration.
- **KNOWN configuration** the linguistic content of the 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. 37,800 impostor tests are performed in this configuration.
- **ALL configuration** the tests are from both the KNOWN and the UNKNOWN configurations. The number of impostor tests in this configuration is 75,600.

C. System Configuration

Mel-scaled frequency cepstral coefficients (MFCC) are used here, 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.

The linear-combination of the two scores described in Section **??** is computed with: 0.3 for the score computed with the text-independent speaker model; 0.7 for the final EBD text-dependent score. These coefficients are empirically determined.

D. Results

Experiments are conducted to assess the contributions coming from the three components, GMM/UBM, SCHMM and video-learnt 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 **??**, expressed in terms of equal error rates (EER), 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 video-learnt 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 **??** show the results when the impostors do not know the speaker utterances. Error rates fall from 2.72 down to 2.00 and 0.87 to 0.55 for the 1-occ and 2-occ respectively when SCHMM is used. Increasing the number of states up to 20 allows successive improvements in reducing EER. Unfortunately 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 are confirmed in Table **??** when two utterances occurrences are used for the training (2-occ). 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.

Configuration		GMM baseline	Number of states of the EBD		
			5	10	20
UNKNOWN	1-occ	2.72	2.39	2.22	**2,00**
UNKNOWN	2-occ	0.87	0.67	0.56	**0.55**
KNOWN	1-occ	**4.49**	4.56	4.56	4.72
KNOWN	2-occ	**2.22**	2.64	2.78	2.87
ALL	1-occ	3.72	3.62	3.61	**3.55**
ALL	2-occ	**1.58**	1.91	1.90	1.97

TABLE I
EER OF GMM AND EBD SYSTEMS (WITH DIFFERENT NUMBERS OF STATES) USING ONE OR TWO TRAINING OCCURRENCES IN DIFFERENT TEST CONFIGURATIONS

In order to evaluate the effect of a video-learnt synchronisation, a new experiment is performed. Results of this experiment are presented in Table **??**. The first column is the original GMM baseline and the next two columns give the EBD results as above in Table **??** for the 20 state case, and then with the use of the video-learnt information. As expected, performance of the EBD improves when using video-learnt segmentation. This is particularly highlighted in *1-occ* conditions where the system possesses less speaker specific data. The EER in the

ALL configuration fall from 3.72 down to 3.16. Moreover performance of the EBD system is greatly increased in the case of impostor knowing the utterances (KNOWN configuration), since performance of the EBD becomes equivalent to the baseline GMM/UBM in this configuration (the EER in 1-occ KNOWN configuration are 4.49 for the baseline GMM and 4.44 for the EBD). It seems that the video segmentation provides information on the utterance but also on the speaker. Moreover, the EBD using an video-learnt synchronisation gains more from the increase of training data than the same system without this temporal information.

Configuration		GMM baseline	Video-Learnt Information	
			None	Active
UNKNOWN	1-occ	2.72	2.00	**1.78**
UNKNOWN	2-occ	0.87	0.55	**0.43**
KNOWN	1-occ	4.49	4.72	**4.44**
KNOWN	2-occ	**2.22**	2.87	2.33
ALL	1-occ	3.72	3.55	**3.16**
ALL	2-occ	1.58	1.97	**1.57**

TABLE II
EER OF GMM COMPARED TO THE EBD SYSTEM IN 20 STATE CONDITION, WHEN USING OR NOT A VIDEO-LEARNT SEGMENTATION

IV. CONCLUSIONS AND FUTURE WORKS

The approach proposed in this paper is designed for embedded applications. It takes advantages of a GMM/UBM text-independent approach and the HMM/Viterbi speech-recognition power. In addition we propose to reinforce the Temporal Structure Information modelling by a synchronisation issued from a video stream. The use of the temporal information in a speaker recognition system allows to improve performance particularly when a relatively small quantity of speech data is available for training and test.

Performance 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 4.56) 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 video-learnt synchronisation leads to a gain in all situations, when impostors know or not the client-utterance, and outperforms or is equivalent to the GMM/UBM baseline in all situations.

Future work will focus on the multi-modality by using the video segmentation even during the testing phase. By incorporating this strong constraint in the training and also in the testing phases we aim at increasing 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 all conditions. Moreover, as the first results show the ability of the EBD approach to take advantage of the temporal structure 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.

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/

Mistral is a project supported by the French A.N.R. (Agence Nationale pour la Recherche). http://mistral.univ-avignon.fr/

MoBio is a European Funded Project (FP7-2007-ICT-1). http://www.mobioproject.org/